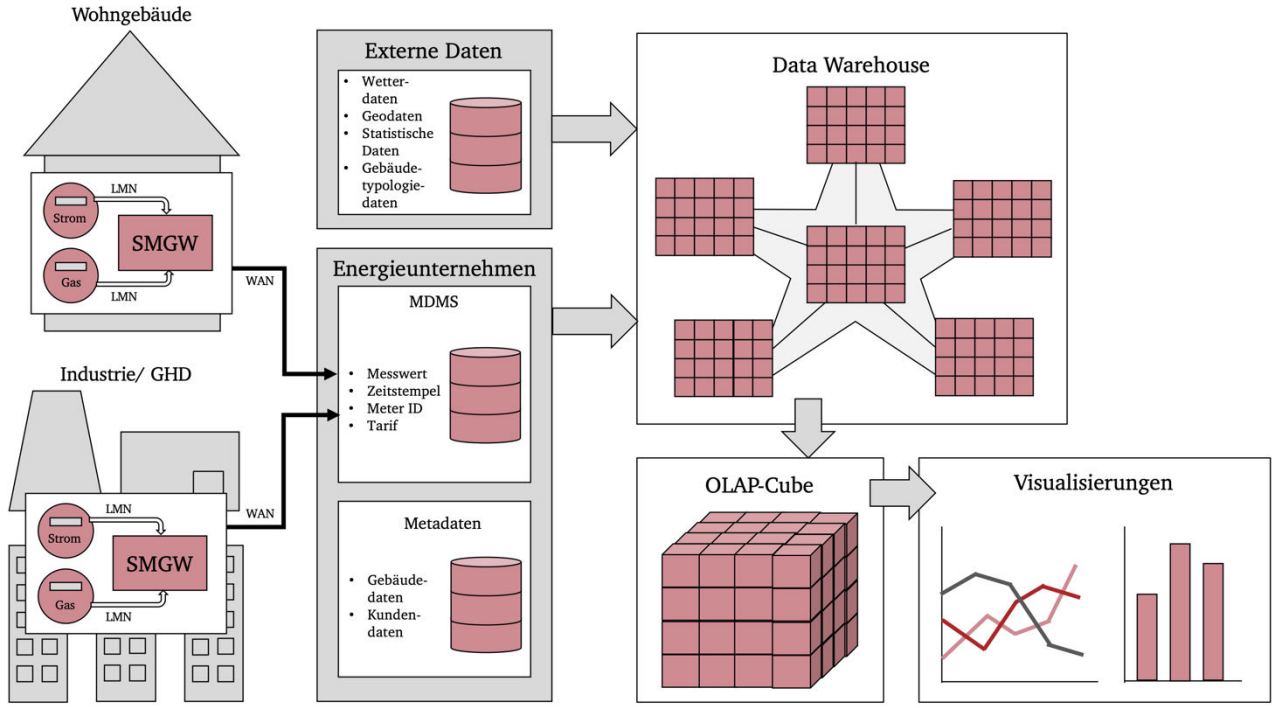
# Untersuchung der Eignung von Online Analytical Processing Cubes für die Integration und Analyse von Energieverbrauchsdaten

**Institut für Numerische Methoden und Informatik im Bauwesen**
**Bachelorarbeit**
**Samuel Achenbach**

TECHNISCHE
UNIVERSITÄT
DARMSTADT




**Untersuchung der Eignung von Online Analytical Processing Cubes für die Integration und Analyse von Energieverbrauchsdaten**
Investigation of the suitability of Online Analytical Processing Cubes for the integration and analysis of energy consumption data

**Vorgelegt von:**
Herr Samuel Achenbach
Matr.-Nr.: 2437189
Studiengang Wirtschaftsingenieurwesen – technische Fachrichtung Bauingenieurwesen (B.Sc.)

**Betreuung durch:**
Prof. Dr.-Ing. Uwe Rüppel
André Hoffmann, M.Sc.

Technische Universität Darmstadt
Fachbereich Bau- und Umweltingenieurwissenschaften
Institut für Numerische Methoden und Informatik im Bauwesen
Juli 2020

## Zusammenfassung

Durch die stetig voranschreitende Ausbreitung von Smart Metern und intelligenten Sensoren im Rahmen des Internet of Things und Smart Home Technologien werden immer mehr Daten generiert. Diese Daten haben das Potential, wichtige Erkenntnisse zu liefern und großen Mehrwert für Endnutzer und verschiedenste Stakeholder aus Wirtschaft, Wissenschaft und Politik zu generieren. Entscheidend dafür ist jedoch, dass die enormen Datenmengen so integriert und verarbeitet werden, dass sie konkrete Antworten auf konkrete Fragen liefern können. Insbesondere energetische Analysen anhand von Smart Meter Daten können viele Faktoren berücksichtigen und eine hohe Komplexität aufweisen. Fraglich ist daher, wie genau die Integration, Modellierung und Analyse der Daten erfolgen soll. Eine mögliche Antwort könnten Online Analytical Processing Cubes aus dem Bereich des Business Intelligence sein. Ursprünglich für die Geschäftsanalytik entwickelt, extrahieren diese analytischen Informationssysteme Daten aus operativen Datenbanksystemen und externen Datenquellen und integrieren diese für Analysezwecke in ein separates Data Warehouse. Dabei werden die Daten multidimensional modelliert, indem sie in zu analysierende Fakten und beschreibende Dimensionen aufgeteilt werden. Die Speicherung der Daten erfolgt dabei in auf Multidimensionalität ausgelegten Datenbankarchitekturen. Im Rahmen dieser Arbeit wurde untersucht, inwiefern Online Analytical Processing Cubes ebenfalls für die Integration und Analyse von Sensordaten, insbesondere Energieverbrauchsdaten, geeignet sind. Zunächst wurden dafür die Grundlagen der Smart Meter Technologie und Online Analytical Processing recherchiert. Anschließend wurde ein Konzept erarbeitet, mit dem die durch Smart Meter gemessenen Energieverbrauchsdaten anhand von mehrdimensionalen Datenbankmodellen und Datenwürfeln gespeichert und mehrdimensional abgefragt werden können. Besonderer Fokus lag in diesem Zusammenhang auf Wetter- und Gebäudetypologiedaten als beschreibende Elemente für energetische Analysen. Anschließend wurde dieses Konzept mit realen Smart Meter Daten umgesetzt. Dafür wurden drei Hypothesen aufgestellt, die durch mehrdimensionale Datenmodellierung untersucht und geprüft wurden. Zur Umsetzung wurden SQLite, R und Python verwendet.

## Abstract

Due to the recent popularity of smart home and internet of things technologies, an increasing number of smart energy meters and sensors are being deployed, which leads to ever more data being generated. This data has the potential to deliver valuable insight and information for users and various stakeholders from business over science to politics. However, in order to fully leverage its potential, it is crucial that the generated sensor data is integrated and processed in a way which can deliver concrete answers to specific questions. In particular, energy analysis based on real smart meter data usually considers a large number of factors and can thus have a high complexity. Hence, the main challenge is finding a proper way for integration, modelling and analysis of sensor data. A possible solution might be the use of Online Analytical Processing Cubes, usually known in the context of business intelligence. Originally developed for business analytics, those analytical information systems extract data from operative and external databases and integrate them in a separate Data Warehouse for analysis. By splitting the data into descriptive dimensions and facts to be analyzed, the data is modelled in a multidimensional way. Therefore, the data is stored in multidimensional database architectures. In this work, it was examined how well Online Analytical Processing Cubes can also be used for the integration and analysis of sensor data, in particular smart meter data. To do so, the first half of this thesis examined the basics of Smart Meters and Online Analytical Processing. Afterwards, a concept was developed to integrate energy consumption data produced by smart meters using multidimensional database architectures and data cubes and to analyze it regarding various dimensions. The main focus here were weather data and building typology and information as describing elements of the energy consumption. To do so, three hypotheses were developed and investigated using multidimensional data modelling. SQLite, R and Python were used for implementation.

## Inhaltsverzeichnis

# 1. Einleitung

Innerhalb der letzten Jahre haben mit digitalen Schnittstellen ausgestattete Sensoren und Geräte im Gebäude- und Immobilienbereich zunehmend an Beachtung, Popularität und Bedeutung gewonnen. Die bekanntesten Beispiele sind hierbei intelligente, digitale Stromzähler, meist und auch im Zuge dieser Arbeit als „Smart Meter“ bezeichnet. Dies sind mit einer als digitale Schnittstelle fungierenden Kommunikationseinheit verbundene Sensoren zur Messung des Energieverbrauchs eines Gebäudes. Jedoch auch im Rahmen der Ausbreitung des Internet of Things (IoT) in den Bereichen „Smart Home“ und „Home Automation“ werden stetig mehr Daten generierende Sensoren installiert. Durch die zunehmende Verbreitung und Nutzung dieser intelligenten Sensoren entstehen enorme Datenmengen, die sinnvoll gespeichert werden müssen. Diese Daten sind jedoch nur von Wert, wenn sie so strukturiert sind und abgefragt werden können, dass sie konkrete Antworten auf konkrete Fragen liefern. Schließlich haben die in den Daten enthaltenen Informationen durch entsprechende Analysen das Potential, vielseitige, interessante Erkenntnisse zu liefern und bisher unerkannte Zusammenhänge aufzudecken.

Insbesondere der Bereich Energiemanagement bietet hier ein hohes Potential für die Nutzung der durch die Sensoren generierten Daten. Durch den Einsatz intelligenter Energieverbrauchszähler für Elektrizität können viele Prozesse automatisiert und optimiert werden, was sich positiv auf alle beteiligten Parteien auswirkt. Bisher hat sich die Nutzung ebendieser Zähler vor allem auf die Messung, Speicherung, Darstellung und Visualisierung von aktuellen und historischen Energieverbräuchen in Gebäuden und die damit aufgezeigten Energieeinsparpotentiale, sowie die Automatisierung von Abrechnungsprozessen bei Stromlieferanten beschränkt. Allerdings bieten Smart Meter für den Bereich Elektrizität auch ein hohes Potential in Bezug auf die Eigenbedarfsoptimierung, die Netzoptimierung und die Steuerung der Elektrizitätsnachfrage. (Zeller, 2015, pp. 81-83)

Zudem steht das Energieversorgungssystem in Deutschland durch den Beschluss der Energiewende, die Verknappung der Ressourcen fossiler Energieträger (Petsch et al., 2012, p. 1), die geringe Akzeptanz von Energieerzeugung durch Atomkraft (Arlt & Wolling, 2011, pp. 14-15) und dem damit verbundenen Atom- und Kohleausstieg vor der Aufgabe, eine stark fluktuierende, dezentrale Stromeinspeisung aus erneuerbaren Energien, insbesondere durch Photovoltaik und Windkraft, besser in das Energieversorgungssystem einzubinden. Dies beinhaltet insbesondere die Aufgabe, das jederzeitige, systemweite Gleichgewicht zwischen Energiebereitstellung und -verbrauch sicherzustellen. (Petsch et al., 2012, p. 1) Aufgrund der immer evidenter werdenden Folgen des Klimawandels hat sich Deutschland außerdem im Rahmen des Pariser Klimaschutzabkommens 2015 dazu verpflichtet, bis zum Jahr 2050 weitgehend treibhausgasneutral zu werden. Um dieses Ziel zu erreichen, spielt neben der bereits erwähnten Erhöhung des Anteils der Stromerzeugung durch erneuerbare Energien auch die Steigerung der Energieeffizienz eine wichtige Rolle. In Bezug auf beide Ansätze bietet der Gebäudesektor ein erhebliches Potential zur Erreichung der Ziele. (Petsch et al., 2012, p. 1) In Bezug auf die Steigerung der Energieeffizienz können allein im Bereich der Privathaushalte nach konservativer Abschätzung bereits ohne preis- oder anreizbasierte Programme Energieeinsparungen von circa 9,5 TWh pro Jahr erreicht werden (wik-Consult, 2006, p. 119), während bezüglich der Integration von erneuerbaren Energien der Haushaltssektor ein theoretisches

Lastverlagerungspotential von bis zu 3,7 GW, beziehungsweise bis zu 20,6 GW bei Berücksichtigung der Wärme- und Kältetechnik, zur besseren Auslastung der vorhandenen Kapazitäten bietet. (Klobasa, 2007, p. 85) Die Smart Meter Technologie wird dabei als einer der Schlüssel zur Nutzung ebendieser Potentiale angesehen. (Petsch et al., 2012, p. 1) Aufgrund der vorhersehbar hohen zukünftigen Relevanz dieses Themenbereiches werden im Rahmen dieser Arbeit daher speziell Sensoren zur Energieverbrauchsmessung, sowie weitere Sensoren, deren Messwerte bezüglich der energetischen Bewertung und Analyse von Relevanz sein könnten, fokussiert. Allerdings können die erarbeiteten Ergebnisse theoretisch auch auf jegliche andere, im Gebäudesektor generierten Sensordaten angewandt werden.

Mit zunehmender Marktdurchdringung der Smart Meter durch ihre bereits geschilderte Schlüsselrolle bei der Erreichung energie- und klimapolitischer Ziele, sowie durch den immensen Wertschöpfungszugewinn, der durch kurze Messintervalle entsteht, werden in Zukunft noch größere, stetig wachsende Datenmengen durch die Messwerte der Sensoren und deren Übertragung entstehen. Diese gilt es sinnvoll zu integrieren. Neben der Integration reiner Messungen aktueller Energieverbräuche, wie beispielsweise bei der Abrechnung von Strom- oder Erdgasverbräuchen mit Lieferanten, die daher eine Art „Transaktionsvorfall“ darstellen, könnten vor allem nachträglich durchgeführte Analysen der zur Verfügung stehenden Datenmengen von Interesse sein. Durch die Verbindung verschiedener Datenquellen aus Smart Metern und anderen IoT Geräten können auf diesem Weg die generierten Daten zu wertvollen Informationen und neuen Erkenntnissen führen. Insbesondere Analysen hinsichtlich zeitlicher, räumlicher oder sonstiger Aspekte könnten hierbei sehr interessant sein. Bezüglich der energetischen Bewertung und Analyse hätte im Gebäudesektor vor allem die Analyse von Energieverbrauchsdaten hinsichtlich der Beziehung zu räumlichen Aspekten, zeitlichen Aspekten, Gebäudetypologien, der technischen Gebäudeausrüstung und Energieeffizienzklassen das Potential, viel Mehrwert zu generieren und wichtige Erkenntnisse für sowohl Politik und Wissenschaft als auch das operative Unternehmensgeschäft von beispielsweise Energieunternehmen, Netzbetreibern oder Elektrizitätsvertriebsunternehmen zu liefern. Neben der Nutzung und Analyse der Daten aus makroskopischer Perspektive, also auf Kreis-, Landes- oder Bundesebene, wäre aber auch eine Nutzung und Analyse auf mikroskopischer Ebene denkbar. Beispielhaft wäre hier die Betrachtung einzelner, mehrstöckiger Bürogebäude, großer Mehrfamilienhäuser oder verschiedenster Nicht-Wohngebäude wie Industriehallen oder Einkaufszentren zu nennen. Insbesondere in den einzelnen Räumlichkeiten installierte Sensoren aus dem Bereich der Gebäudeautomation, sowie an verschiedenen Stellen und Stockwerken installierte Sensoren zur Messung des Raumklimas könnten hier wichtige Erkenntnisse über das Nutzerverhalten in verschiedenen Räumlichkeiten, bisher unerkannte Wärmebrücken, den Sanierungsbedarf bestimmter Gebäudeabschnitte oder unterschiedliche Lastprofile innerhalb des Gebäudes liefern.

Fraglich ist dabei, wie genau die Integration und Analyse der vorhandenen Daten erfolgen soll. Eine mögliche Option könnten Online Analytical Processing Cubes (OLAP-Cubes) aus dem Bereich des Data-Warehousing und Business Intelligence (BI) sein. Hierbei erfolgt zunächst die Integration der Daten durch auf Multidimensionalität ausgelegte, meist sternförmige Datenbankarchitekturen. Anschließend kann der Analyst die Daten mit Hilfe von geeigneten Werkzeugen analysieren, darstellen und visualisieren und so vorher aufgestellte Hypothesen überprüfen. All diese Werkzeuge sind in der Regel

Teil von OLAP-Anwendungen, die speziell für die Analyse von Daten über mehrere Dimensionen hinweg entwickelt wurden. Im Rahmen dieser Arbeit sollen daher OLAP-Cubes und die damit in Zusammenhang stehenden Technologien und Aspekte des Data Warehousing bezüglich ihrer Eignung für die Integration und Analyse der durch Smart Meter und andere IoT Geräte generierten Daten untersucht werden.

Die Arbeit ist dabei in vier wesentliche Teile gegliedert. Die ersten beiden Teile befassen sich mit der Recherche bezüglich der Themengebiete, während in den letzten beiden Teilen, aufbauend auf den zuvor erarbeiteten Ergebnissen, ein Konzept entwickelt und anhand realer Energieverbrauchsdaten umgesetzt und demonstriert wird.

Der erste Teil der Recherche befasst sich dabei mit Smart Metern. Hierbei werden, stellvertretend für weitere Smart Meter, zunächst intelligente Stromverbrauchszähler beleuchtet. Diese werden hinsichtlich der zugrundeliegenden Technologie und Funktion untersucht. Zudem werden die mit einer Nutzung verbundenen Vorteile und Chancen, sowie Nachteile und Risiken für die beteiligten Akteure herausgearbeitet. Anschließend werden die rechtlichen Rahmenbedingungen und die gesellschaftliche Akzeptanz der Technologie erforscht. Bei dem darauffolgenden Blick ins technische Detail werden die vorhandenen Schnittstellen, sowie Format und Inhalt der generierten Daten erörtert, bevor abschließend bereits bestehende Lösungen zur Integration der Daten untersucht werden. Im Anschluss finden auch intelligente Gasverbrauchszähler eine kurze Erwähnung, allerdings wurde sich aufgrund der Komplexität des Elektrizitätsnetzes und dem großen Potential von intelligenten Stromverbrauchszählern dafür entschieden, die für das Konzept relevanten Aspekte anhand der Stromverbrauchszähler zu untersuchen. Smart Meter für Elektrizitätsverbrauchsmessung und Smart Meter für Gasverbrauchsmessung werden im Folgenden synonym auch als Strom Smart Meter oder intelligente Stromzähler, beziehungsweise Gas Smart Meter oder intelligente Gaszähler bezeichnet.

Der Fokus des zweiten Teils liegt auf OLAP-Cubes und den damit zusammenhängenden Data Warehouse Technologien, um die theoretischen Grundlagen der Technologie und ihres potentiellen Nutzens bezüglich der Integration und Analyse der Smart Meter Messdaten zu erarbeiten. Zunächst werden hierfür die Begriffe Business Intelligence und Data Warehouse, sowie deren Geschichte näher erläutert. Anschließend werden die möglichen Datenbankarchitekturen sowie ihre Vor- und Nachteile dargestellt. Danach werden die Grundlagen und Schritte des Prozesses Extract-Transform-Load (ETL) zur Integration der Messdaten in die Datenbank erarbeitet. Als Viertes werden die verschiedenen OLAP-Cube Modelle dargestellt und verglichen, sowie ihre wesentlichen Funktionen erörtert.

Der dritte Teil widmet sich der Erarbeitung eines Konzepts und führt damit die in den vorangegangenen Rechercheteilen erarbeiteten Ergebnisse zusammen. Dabei werden zunächst die Grundlagen des Konzepts vorgestellt. Dementsprechend wird auf verwandte Arbeiten, mögliche Datenquellen, Nutzungsmöglichkeiten und Vorteile sowie Hindernisse und Herausforderungen bei der Umsetzung eingegangen. Im Anschluss wird die technische Umsetzung erläutert. Hier wird insbesondere die Auswahl eines Datenbankmodells durchgeführt und begründet, sowie das Vorgehen bei der Integration, Analyse und Visualisierung der Daten erläutert. Zuletzt werden dann drei

Hypothesen formuliert, die im Zuge einer Demonstration des Modells durch reale Daten geprüft und untersucht werden sollen.

Nach der Erarbeitung und Erläuterung eines Konzepts wird dieses praktisch umgesetzt und die praktische Umsetzung im vierten Teil der Arbeit dokumentiert. Dabei wird zunächst das allgemeine Vorgehen erläutert und dann für jede der drei Hypothesen ein Data Warehouse aufgebaut. Nach der Integration sämtlicher, relevanter Daten werden diese dann entsprechend analysiert und visualisiert. Basierend auf den Analyseergebnissen können dann die Hypothesen überprüft werden. Abschließend werden die Ergebnisse der Arbeit zusammengefasst und kritisch gewürdigt.

# 2. Smart Meter

Der erste Teil der Recherche widmet sich intelligenten Energieverbrauchszählern. Im Zuge dessen werden primär intelligente Stromzähler vorgestellt. Diese werden hinsichtlich der zugrundeliegenden Technologie und Funktion, dem entstehenden Nutzen und der Chancen, sowie der Schwächen und Risiken, der Ausbreitung und gesellschaftlichen Akzeptanz, der Schnittstellen, dem Inhalt und Format der generierten Daten und der bereits bestehenden Integrationsarchitekturen untersucht. Insbesondere die Erkenntnisse durch den Blick ins technische Detail anhand der Erforschung des Datenformats und -inhalts sowie der Schnittstellen gelten im Allgemeinen aber auch für intelligente Gaszähler. Diese Erkenntnisse sind für die Erarbeitung des späteren Konzepts von wesentlicher Bedeutung.

## 2.1. Intelligente Stromzähler

Intelligente Stromzähler, synonym auch als Strom Smart Meter bezeichnet, haben eine besondere Relevanz in Bezug auf die Erreichung der von der Bundesregierung festgelegten Klimaziele und bieten ein hohes Potential zur Unterstützung der Umsetzung der Energiewende. Als eine der Schlüsseltechnologien zur Bewältigung zukünftiger Herausforderungen der Energiewirtschaft wird ihnen eine besondere Bedeutung beigemessen. Dementsprechend sind sie auch wesentlicher Teil der Untersuchungen und erarbeiteten Konzepte im Rahmen dieser Arbeit. Um die in den folgenden Unterpunkten untersuchten Aspekte rund um Strom Smart Meter besser zu verstehen, macht es allerdings Sinn, zunächst den Strommarkt als Ganzes in Betracht zu nehmen, um die wesentlichen Akteure und Beteiligten des Strommarktes und deren Aufgaben darzustellen und herauszuarbeiten, wo und wie genau Strom Smart Meter innerhalb dieses Marktes angesiedelt sind. Dies ist von Relevanz, um später zu verstehen, worin genau die Aufgaben und potentiell realisierbaren Nutzungsmöglichkeiten von Smart Metern als Teil der Wertschöpfungskette des Elektrizitätsmarktes bestehen. Zudem ist Elektrizität als Handelsgut in seinen Eigenschaften in vielerlei Hinsicht sehr besonders und der Elektrizitätsmarkt unterscheidet sich mit seinen Herausforderungen, den verschiedenen Akteuren und der hohen Notwendigkeit von Steuerung wesentlich von anderen Energiemärkten, beispielsweise Fernwärme oder Energiemärkten fossiler Primärenergieträger wie Rohöl oder Erdgas. Diese Eigenschaften sind ebenfalls von Relevanz und werden daher im Folgenden zunächst kurz dargestellt.

### Eigenschaften des Elektrizitätsmarktes

Die Elektrifizierung des Energiesystems begann vor über einhundert Jahren. Seitdem hat Elektrizität als Energieträger weltweit zunehmend an Bedeutung gewonnen und ist heutzutage wesentlicher, nicht mehr wegzudenkender Bestandteil des Energiesystems. Innerhalb der Energiewirtschaft hat sich elektrische Energie zudem immer mehr als universelles Austauschmedium zur Energiewandlung etabliert. Neben der Befriedigung von Grundbedürfnissen durch die Bereitstellung von Wärme und Licht, sichert Elektrizität auch in Verbindung mit Kommunikations- und Informationssystemen grundlegende gesellschaftliche und kulturelle Bedürfnisse. Auch als Produktionsfaktor für die Sektoren Industrie und Dienstleistungen ist Elektrizität unverzichtbar. Damit stellt das Elektrizitätssystem einen Flaschenhals der Volks- und insbesondere der Energiewirtschaft dar. (Zeller, 2015, p. 1)

All diese Belange repräsentieren wesentliche Gründe für die große Bedeutung von Energiepolitik. Diese muss klassischerweise im Trilemma von Wirtschaftlichkeit, Versorgungssicherheit und Umweltverträglichkeit vermitteln. (Zeller, 2015, p. 1) Wirtschaftlichkeit der Energieversorgung ist erforderlich, um „Energiearmut“ (Niebert, 2014, pp. 108-109) zu vermeiden. Eine nicht wirtschaftliche Energieversorgung und die damit verbundenen hohen Stromkosten würden den Industrialisierungsgrad der Volkswirtschaft gefährden und den Konsum von Privathaushalten und Unternehmen reduzieren. Die Versorgungssicherheit erfordert auf kurz- und mittelfristige Sicht die Sicherstellung einer Energieinfrastruktur, die eine zuverlässige und konsequente Stromversorgung ermöglicht. Langfristig muss der Primärenergiebedarf durch technologische und geopolitische Strategien gedeckt sein. (Zeller, 2015, p. 1) Die Umweltverträglichkeit erfordert einen über Generationen hinweg schonungsvollen Umgang mit natürlichen Ressourcen, sowie eine weder Natur noch menschliche Gesundheit gefährdende Stromproduktion. Der Aspekt der ökologischen Nachhaltigkeit besitzt besondere Brisanz durch Themen wie die Endlagerung nuklearer Brennstoffe, die Endlichkeit fossiler Brennstoffe und die mit dem Klimawandel einhergehenden Treibhausgasemissionen. (Zeller, 2015, p. 1) Auch die möglichen gesundheitlichen Folgen einer nicht nachhaltigen Energiewirtschaft sind in den letzten Jahren immer mehr in den Fokus gerückt. (Niebert, 2014, pp. 110-111)

**Besonderheiten von Elektrizität als Handelsgut**

Auch die Eigenschaften von Strom als Handelsware stellen den Markt vor einige Herausforderungen. Die Ware Elektrizität ist ein intangibles, homogenes Gut. (Gleave, 2008, p. 123) Elektrizität ist nicht greifbar und kann nicht direkt wahrgenommen werden. (Zeller, 2015, p. 57) Aus den homogenen Eigenschaften folgt, dass Elektrizität unabhängig von dem verwendeten Erzeugungsverfahren stets in gleicher Qualität geliefert wird und es keine Qualitätsunterschiede zwischen den Anbietern elektrischer Energie gibt. (Gleave, 2008, p. 122) Wird Energie in Elektrizität umgewandelt, muss die Elektrizität genau in dem Zeitraum, in dem sie erzeugt wird, auch genutzt werden. Dadurch ist das bestimmende Element des Strommarktes die Gleichzeitigkeit von Stromerzeugung und Stromverbrauch. (Zeller, 2015, p. 102) Die dadurch bestehende Notwendigkeit zur Synchronisierung von Stromerzeugung und Stromnachfrage ist entscheidend für die statische Wertfolge der Stromlieferung und resultiert aus den hohen Kosten und der begrenzten Verfügbarkeit von direkten und indirekten Speichern von elektrischer Energie. Bei einem Auseinanderfallen, der zu einem Zeitpunkt durch Kraftwerke oder alternative Energieerzeugung in das Elektrizitätsnetz eingespeisten Energiemenge und der zur gleichen Zeit entnommenen Menge, droht Netzinstabilität. Konkret folgt daraus, dass durch die beschränkte Speicherbarkeit von Elektrizität die Elektrizitätsnachfrage zu jedem beliebigen Zeitpunkt immer genau äquivalent zur eingespeisten Elektrizität abzüglich der systembedingten Verluste durch Übertragung und Verteilung sein muss. (Gleave, 2008, p. 121) Hierin liegt eine besondere Herausforderung, da das Gesamtgleichgewicht im Elektrizitätsnetz zu jedem Zeitpunkt gewährleistet sein muss, eine Messung der Verbrauchsseite, also der entnommenen Energie, aber in einem überwiegenden Teil der Fälle nur auf Jahresbasis erfolgt. (Zeller, 2015, p. 69) Eine weitere wesentliche Eigenschaft von Strom besteht darin, dass er in einer Mehrzahl der Anwendung gar nicht oder nur zu höheren Kosten und geringerem Nutzen, beispielsweise im Fall von elektrischer Beleuchtung, substituierbar ist. (Zeller, 2015, p. 68)

**Einteilung der Endverbraucher von Elektrizität**

Die Absatzmarktsegmentierung erfolgt im Elektrizitätsmarkt traditionell nach den Sektoren Privathaushalte, Gewerbe/Handel/Dienstleistungen (GHD) und Industrie. (Zeller, 2015, p. 58) Neben der sektoriellen Segmentierung ist aber auch eine Einteilung nach Verbrauch relevant. Allgemein wird hier zwischen Großkunden und Kleinverbrauchern unterschieden. (Gleave, 2008, pp. 123-124) Die Privathaushalte sind dabei den Kleinverbrauchern zuzuordnen, während Industrieunternehmen überwiegend zu den Großkunden zählen. Der Sektor GHD ist hingegen sehr heterogen aufgestellt und Einzelkunden des Sektors können daher sowohl zu den Großkunden als auch zu den Kleinverbrauchern gehören. Großkunden sind dabei Kunden mit einem Jahresverbrauch von mindestens 100 MWh. (Zeller, 2015, pp. 58-59) Zudem ergibt sich eine weitere Segmentierung des Absatzmarktes hinsichtlich der teils sehr unterschiedlichen Lastprofile und Präferenzen der einzelnen Kundensegmente. (Zeller, 2015, p. 68)

**Die wesentlichen Akteure des Elektrizitätsmarktes**

Der Elektrizitätsmarkt besteht dabei aus verschiedenen Akteuren, die alle mittel- oder unmittelbar in Verbindung stehen. Nach der klassischen Stufentheorie sind die Akteure dabei in die Stufen Erzeugung, Verteilung und Letztverbraucher aufgeteilt. (Gleave, 2008, p. 121) Die Erzeugerseite stellt dabei den Beschaffungsmarkt dar und beinhaltet alle Akteure, die Energie in Elektrizität umwandeln und ins Elektrizitätsnetz einspeisen. Traditionell sind dies vor allem Betreiber von Atom- und Kohlekraftwerken, aber auch Betreiber von Wasserkraftwerken, die auf diese Weise die Grundlast zur Verfügung stellen. Durch den Ausbau erneuerbarer Energien haben über die letzten Jahre hinweg aber auch Betreiber von Anlagen mit alternativer, erneuerbarer Energietechnik an Bedeutung gewonnen. Hierzu zählen beispielsweise Betreiber von Windparks oder großen Photovoltaik-Anlagen (PV-Anlagen). Diese sind in der Stromerzeugung stark saison- und wetterabhängig. Ein weiterer wesentlicher Bestandteil der Erzeugerseite sind deshalb Erzeuger von Regelenergie. Diese erzeugen zu Zeiten hoher Elektrizitätsnachfrage die Elektrizität, die über die derzeit von den anderen Erzeugern produzierbare Energie hinausgeht und halten so die Netzstabilität aufrecht. Ebenfalls beachtenswert ist, dass durch die Beschlüsse im Zuge des EEG auch Letztverbraucher in Form von Privathaushalten oder Unternehmen die Möglichkeit haben, durch die Installation von PV-Anlagen selbst zu Elektrizitätserzeugern zu werden, indem sie überschüssigen Strom ins Stromnetz einspeisen.

Das Gegenstück zum Beschaffungsmarkt sind die Letztverbraucher, die den Absatzmarkt darstellen. Diese können wie bereits zuvor erläutert in Privathaushalte, Industrie und GHD aufgeteilt werden. Die Verbindung zwischen Erzeugungsseite und Verbrauchsseite wird durch die Netzbetreiber geschaffen. Diese betreiben Übertragungs- und Verteilnetze, die als physische Sammelschiene der Elektrizitätswirtschaft fungieren. (Zeller, 2015, p. 2) Als physische Vertriebsmittler vollziehen die Netzbetreiber den physischen Transport der elektrischen Energie und stellen die erfolgreiche Abwicklung der kommerziellen Verträge zur Stromlieferung an die Kunden sicher. Hierzu kaufen sie Leistungen wie Regelenergie zur Netzstabilisierung auf dem Beschaffungsmarkt ein. (Zeller, 2015, p. 3) Ihnen kommt damit eine physikalische Aggregationsfunktion zu. (Zeller, 2015, p. 68) Netzbetreiber können theoretisch weiter unterschieden werden in Betreiber von Übertragungs- und Betreiber von

Verteilungsnetzen. Übertragungsnetze sind Hochspannungsnetze, die elektrische Energie über lange Distanzen hinweg transportieren, während Verteilungsnetze die elektrische Energie über Niedrigspannung an die Kunden verteilen.

Weitere wesentliche Akteure des Elektrizitätsmarktes sind Elektrizitätsvertriebsunternehmen (EVU). Sie sind die Alliierten der Netzbetreiber und schaffen als Aggregator und Mittler die Verbindung zwischen dem Großhandel auf der Beschaffungsseite, den Endverbrauchern auf der Absatzseite und den Netzbetreibern. Ihre wirtschaftliche und vertragliche Aggregationsfunktion besteht darin, Einzelnachfragen der Verbraucher zu bündeln, um die für den Großhandel benötigten Losgrößen der Strombeschaffung zu erreichen. (Zeller, 2015, p. 67) Zudem stellen EVUs eine Verbindung zwischen den regulierten Bereichen der Stromübertragung und -verteilung, für die die Netzbetreiber durch Gebietsmonopole zuständig sind, und den sonstigen, weniger stark regulierten, wettbewerblich organisierten Marktstufen her. Die Hauptfunktion der EVUs besteht hierbei in dem Einkauf und Handel auf der Beschaffungsseite und dem darauffolgenden absatzseitigen Vertrieb an die Endverbraucher. (Zeller, 2015, p. 67) Ein EVU bezieht dafür die Dienstleistungen der Netzbetreiber und ist somit selbst Netzkunde. (Zeller, 2015, p. 68) Im Gegensatz zu den Netzbetreibern sind EVUs allerdings an weniger Regulierungen zur Preis- und Produktgestaltung gebunden und haben einen größeren Handlungsspielraum für Produkt- und Geschäftsmodellinnovationen. (Zeller, 2015, p. 5)

Messstellenbetreiber (MSB) sind ebenfalls Teil des Elektrizitätsmarktes. Ihnen kommt eine Sonderstellung zu, da die Rolle des MSB sowohl von EVUs, Netzbetreibern als auch unabhängigen dritten Dienstleistern übernommen werden kann. (Zeller, 2015, p. 68) Die rechtliche Grundlage hierfür wird in 2.1.4. kurz näher erläutert. Die Hauptaufgabe von MSBs besteht darin, den Energieverbrauch der Endnutzer zu messen und für die Abrechnung bereitzustellen. Die Ablesung kann vor Ort manuell durch den Besuch eines Angestellten des MSBs, durch manuelle Ablesung und Übermittlung an den MSB durch den Kunden selbst oder durch eine Zählerfernauslesung (ZFA) des MSBs bei der Verwendung von Strom Smart Metern durchgeführt werden. Hier wird bereits ein wesentliches, Nutzen stiftendes Merkmal von intelligenten Stromzählern deutlich. Als Schnittstelle zwischen Endverbrauchern, Netzbetreibern und EVUs ist die Messtechnologie somit der Ausgangspunkt für die Implementierung intelligenter Energieverbrauchszähler. (Zeller, 2015, p. 3) Neben den bisher erwähnten Akteuren können noch weitere Akteure im Elektrizitätsmarkt zu finden sein. Zu nennen wären hier beispielsweise Strombörsen, Makler oder Broker der Stromerzeuger. Zudem ist anzumerken, dass ein Unternehmen auch die Aufgaben mehrerer Akteure abdecken. So ist es zum Beispiel möglich, dass ein Unternehmen sowohl die Rolle des Kraftwerksbetreibers als auch die des EVUs einnimmt.

**Historische Herausforderungen des Elektrizitätssystems**

Ähnlich wie alle anderen Märkte lebt auch der Strommarkt davon, Werte und Leistungen für seine Kunden zu generieren. Neben den bisher geschilderten Aufgaben und Herausforderungen der verschiedenen Akteure, die durch die Eigenschaften des Handelsgutes Elektrizität bereits im operativen Tagesgeschäft entstehen und überwiegend auf die Notwendigkeit der Synchronisierung von Angebot und Nachfrage zurückzuführen sind, haben die Akteure auch langfristige Herausforderungen und

Veränderungen zu bewältigen. Diese können sich sowohl auf der Stufe der Erzeugung als auch auf den Stufen Verteilung oder Verbrauch manifestieren und haben durch die Vernetzung des Marktes auch Auswirkungen auf alle anderen Stufen. Zur Bewältigung der so entstehenden Probleme können Lösungen auf allen Stufen des Elektrizitätsmarktes integriert werden. Insbesondere die Steuerung der Stromnachfrage ist hierfür ein wertvolles Werkzeug, auf das häufig zurückgegriffen wird.

Die Stromnachfrage der Endverbraucher wurde bisher im Wesentlichen durch Anreize der EVUs und durch Strompreise gesteuert. Die Ziele der Beeinflussung der Stromnachfrage durch Anreize und Preise haben sich im Zuge der Entwicklung der Elektrizitätswirtschaft allerdings im Laufe der Zeit verändert. (Albadi & El-Saadany, 2008, p. 1990) Mit dem Beginn der Elektrifizierung lag das Ziel der Nachfragesteuerung vor allem auf dem Ausbau und der Verbreitung elektrischer Energie. Ziel war hier also vor allem eine allgemeine Nachfragesteigerung der Elektrizität, die im Substitutionswettbewerb mit Erdgas stand. Dadurch wurden zunehmend dezentrale Gasmotoren und Gasleuchten durch Elektromotoren und Glühbirnen ersetzt. (Wolter & Reuter, 2005, p. 34) Mit dem voranschreitenden Ausbau des Elektrizitätsnetzes gewann dann auch zunehmend die Effizienz an Bedeutung. Neben einer Steigerung der Energieeffizienz von elektrischen Geräten der Endverbraucher, rückte vor allem auch der effiziente Betrieb des Elektrizitätssystems in den Vordergrund. Zur besseren Ausnutzung von Kraftwerken und Stromnetzen wurden deshalb Maßnahmen zur Schaffung einer möglichst gleichmäßigen Stromlast ergriffen. Diese Maßnahmen beinhalteten das Anheben der Grundlast („valley filling“) und eine Verlagerung der Last von Spitzenlastzeiten in Zeiten mit geringer Stromnachfrage („load shifting“). Erreicht wurde eine Steigerung der Grundlast beispielsweise durch die Förderung von Nachspeicherheizungen, welche die zu nächtlichen Schwachlastzeiten verfügbare Elektrizität abnehmen, um einen Wärmespeicher aufzuheizen und so als thermischer Speicher fungieren. (Zeller, 2015, pp. 18-19)

**Zukünftige Herausforderungen des Elektrizitätssystems**

Der Bedarf an elektrischer Energie wurde bis zur Jahrtausendwende fast ausschließlich durch zentrale Großkraftwerke zur Verfügung gestellt. Mit dem voranschreitenden Ausbau von Anlagen zur Erzeugung von Elektrizität durch erneuerbare Energien gewinnen diese immer mehr an Bedeutung und verändern die Zusammensetzung der Stromerzeugung. Dies führt durch die saisonale und wetterbedingte Abhängigkeit der Stromerzeugung durch erneuerbare Energien zu einer zunehmenden Fluktuation der Stromerzeugung und macht die Prozesse zur Sicherung der Netzstabilität schlechter prognostizierbar. Diese Entwicklungen belasten besonders die Übertragungs- und Verteilnetze. (Zeller, 2015, p. 2) In konventionellen Elektrizitätsnetzen wurde zur Sicherung der Netzstabilität vor allem auf die Erzeugerseite zurückgegriffen, die durch die Bereitstellung von Regelenergie und Produktionsanpassungen die Netzstabilität sicherte. Diese Herangehensweise ist mit zunehmender Integration erneuerbarer Energien allerdings nicht mehr ausreichend. (Sirojan, Lu, Phung, & Ambikairajah, 2019, p. 1) Das macht ein Umdenken im Bereich der Laststeuerung erforderlich, da zukünftig nicht mehr eine möglichst gleichmäßige Stromnachfrage, sondern die Anpassung des Stromverbrauchs an die Strombereitstellung, also auch eine zunehmende Steuerung auf der Nachfrageseite, notwendig wird. Die damit einhergehende Flexibilisierung der Stromlast beinhaltet die Steigerung der Last bei Zeiten hoher Stromeinspeisung durch erneuerbare Energien („peak building“)

und die Abschaltung zentral abschaltbarer Lasten im Fall eines Einbruchs der Einspeisung („valley digging“). (Zeller, 2015, pp. 19-20) Die EVUs haben damit die Aufgabe, sowohl die Nachfrage als auch die Erzeugung zu steuern. (Zeller, 2015, p. 6)

Zudem sind auch auf der Nachfrageseite zukünftig einschneidende Veränderungen zu erwarten. Bisher beschränkte sich die gesamte Stromnachfrage im Wesentlichen neben der Stromnachfrage durch die Produktion der Industrie und der Stromnachfrage des Sektors Gewerbe, Handel und Dienstleistungen vor allem auf Haushaltsgeräte, Beleuchtung und Kältetechnik im Gebäudebereich. Neben dem Elektrizitätssektor existieren allerdings darüber hinaus die Energiesektoren Wärme und Mobilität. Der Energiebedarf dieser Sektoren wird aktuell noch zum überwiegenden Teil durch die Verbrennung fossiler Energieträger gedeckt. Dies ist allerdings weder nachhaltig, noch in Anbetracht der klimapolitischen Ziele zielführend. Daher ist eine zunehmende Elektrifizierung der Sektoren Wärme und Mobilität notwendig und zukünftig zu erwarten. Im Sektor Wärme spielen für die Elektrifizierung vor allem Wärmepumpen und Elektroheizungen, die auch als Zusatzheizungen für solarthermische Anlagen ausgelegt sein können, eine wichtige Rolle. (Zeller, 2015, p. 3) Im Bereich Mobilität gewinnen batteriebetriebene Elektroautos immer mehr an Bedeutung und sorgen damit für eine zunehmende Elektrifizierung des Sektors. Fraglich ist nun, inwiefern die damit verbundene, zu erwartende Steigerung der Stromnachfrage durch den gegenläufigen Trend der Effizienzsteigerung bei der Energieumwandlung und dem Betrieb von Geräten kompensiert wird. Manche Quellen gehen insgesamt von einer leichten Stagnation der Stromnachfrage aus. (Zeller, 2015, p. 3) Allerdings würde selbst dieser Fall grundlegende Veränderungen für die Elektrizitätswirtschaft mit sich ziehen, da mit zunehmender Elektrifizierung der Sektoren Mobilität und Wärme die zeitliche Zusammensetzung der Stromnachfrage von deutlich mehr Faktoren abhängen würde, als bisher der Fall. Durch Elektrifizierung der Wärmebereitstellung wäre eine stärkere Korrelation der Stromnachfrage mit dem Außentemperaturverlauf zu erwarten, während durch den Sektor Mobilität die Stromnachfrage, durch das Beladen von Elektroautos bedingt, auch während der Nachtstunden deutlich ansteigen könnte.

**Digitalisierte Elektrizitätssysteme mit intelligenter Verbrauchsmessungsinfrastruktur**

Zur Bewältigung dieser Herausforderungen ist eine Digitalisierung des Elektrizitätssystems erforderlich, die das Elekrizitätssystem intelligent macht. Intelligente Elektrizitätssysteme, international meist als „smart grid“ bezeichnet, sind die Elektrizitätssysteme der Zukunft und bestehen aus vier primären Bausteinen. Diese sind eine Advanced Metering Infrastructure (AMI), Advanced Distribution Operations, Advanced Transmission Operations und Advanced Asset Management. Unter diesen stellt die AMI, eine intelligente Verbrauchsmessungsinfrastruktur, einen ersten Schritt zur Digitalisierung der Elektrizitätssystems dar. Sie besteht aus intelligenten Stromverbrauchszählern, einer Wide Area Communication Infrastruktur, Home Area Netzwerken und Datenmanagementsystemen und ermöglicht so die bidirektionale Kommunikation zwischen den Stufen Erzeugung, Verteilung und Endverbraucher des Strommarktes. (Sirojan et al., 2019, p. 1) Die Integration einer intelligenten Infrastruktur mit Smart Metern ist damit ein zentraler Teil moderner Stromnetze und eine Grundvoraussetzung für die effiziente Steuerung der Stromnachfrage. (Zeller, 2015, p. 6) Die Rolle der MSBs wird mit der zunehmenden Integration von intelligenten Stromzählern um die Aufgabe der Informationsverteilung erweitert. (Zeller, 2015, p. 77) Wie bereits zuvor erörtert,

kann diese Rolle von EVUs, Netzbetreibern oder Dritten übernommen werden. Strom Smart Meter sind dabei auf der Stufe der Endverbraucher, also auf dem Absatzmarkt, installiert und übermitteln die ausgelesenen Stromverbräuche über das integrierte Kommunikationsmodul und die Kommunikationsinfrastruktur der AMI an die für den Messstellenbetrieb zuständige Partei. Dort werden die Daten gespeichert und können im Bedarfsfall an weitere Akteure des Marktes weitergeleitet werden.

Einleitend wurden die wichtigsten Eigenschaften des Elektrizitätsmarktes und des Handelsgutes Strom, die wesentlichen Akteure des Elektrizitätsmarktes, sowie deren Aufgaben und Vernetzungen, als auch die bisherigen und zukünftigen Herausforderungen des Marktes dargestellt. Hieraus wird auch die wichtige Bedeutung und das Potential von intelligenten Stromzählern und der damit verbundenen Kommunikations- und Netzwerkinfrakstruktur deutlich, was gegen Ende dieses Teils bereits kurz beschrieben wurde. Die so geschaffenen Grundlagen sind eine wegweisende Wissensbasis zum Verständnis von Strom Smart Metern und ihren Schnittstellen. Im Folgenden werden daher nun Strom Smart Meter im Detail hinsichtlich ihrer Technologie und Funktion, ihrer Vor- und Nachteile, sowie Chancen und Risiken, ihrer Ausbreitung und Akzeptanz, ihrer Schnittstellen und bestehenden Integrationsarchitekturen, sowie dem Format und Inhalt der von ihnen generierten Daten untersucht.

### 2.1.1. Technologie und Funktionsweise

Nun sollen zunächst die technologischen Grundlagen von Strom Smart Metern und deren Funktionsweise dargestellt werden. Allgemein messen Stromzähler die vom Letztverbraucher genutzte, also dem Netz entnommene Energie. Die Messung erfolgt als die Wirkleistung (in der Regel in Kilowatt) über die Zeit. Die so gemessene Wirkarbeit wird üblicherweise in Kilowattstunden (kWh) angegeben. Prinzipiell gibt es drei verschiedene Arten von Stromzählern, die je nach Wirkprinzip zur Messung der elektrischen Energie und nach Methode der Speicherung und Übertragung der Messwerte unterschieden werden können. Man unterscheidet daher zwischen elektromechanischen und elektronischen Zählern, sowie zwischen digitalen Zählern mit und nicht digitalen Zählern ohne Kommunikationseinheit. (Zeller, 2015, p. 13) Der Ferraris-Zähler kristallisierte sich dabei seit dem Aufschwung der Elektrizitätswirtschaft zu Beginn des 20. Jahrhunderts aufgrund seiner Robustheit und Langlebigkeit als Standardtechnologie heraus und ist bis heute der mehrheitlich genutzte Zähler zur Messung von Privathaushalten. Es handelt sich hierbei um einen elektromechanischen Zähler, der durch das Drehfeldprinzip die Wirkarbeit misst. Hierbei dreht sich durch die bei Wechselstrom hervorgerufene Induktion von Magnetfeldern eine Läuferscheibe. Das so resultierende Drehmoment ist proportional zur Wirkleistung. Zur Ablesung und Speicherung der Messwerte dient das Zahlwerk. Eine Ablesung kann daher lediglich optisch erfolgen. (Zeller, 2015, pp. 13-14) Anfang der 1980er Jahre wurden dann hybride Zähler als Vorstufe moderner elektronischer Zähler entwickelt. Die Zähler koppeln das Triebsystem der Ferraris-Zähler mit einem Mikrorechnersystem. Die Rotation der Triebscheibe wird hierbei über einen Impulsgeber mit der Elektronik verbunden, was die digitale Übertragung der Messwerte ermöglicht. Durch die Fortschritte im Bereich der Leistungselektronik wurden diese jedoch zunehmend durch elektronische Zähler abgelöst, die somit die digitale, intelligente Alternative zu den bis heute weit verbreiteten Ferraris-Zählern bilden. (Zeller, 2015, pp. 14-15)

Diese elektronischen Zähler zur Elektrizitätsverbrauchsmessung stellen die im heutigen Sprachgebrauch als Smart Meter oder intelligente Stromzähler bezeichneten Elektrizitätsverbrauchsmessgeräte dar. Die Nutzung dieser intelligenten Stromzähler wird vor allem durch die technologischen Fortschritte der Informationsinfrastruktur ermöglicht. Sie sind ein wesentlicher und unverzichtbarer Bestandteil eines digitalen Elektrizitätssystems mit fortschrittlicher, intelligenter Verbrauchsmessungsinfrastruktur. Unter intelligenten Netzen, Messsystemen und Endnutzeranwendungen versteht man allgemein die zeitnahe Messung und Steuerung von Energiesystemen unter Verwendung von Informations- und Kommunikationstechnologie. (Zeller, 2015, p. 3) Elektronische Zähler erzeugen durch einen ohmschen Widerstand einen Spannungsabfall, der proportional zur Stromstärke (Wirkleistung) ist. Wenn dieser Widerstand bekannt ist, kann dadurch also der Strom gemessen werden. Alternativ können zur Messung auch Hall-Sensoren verwendet werden. (Zeller, 2015, p. 15) Neben dem Sensor sind elektronische Zähler mit einem Rechenwerk, einem Speichermedium, einer Anzeige und Bedienelementen ausgestattet. Um auch im Falle eines Stromausfalls die sichere Speicherung der Messdaten zu gewährleisten, kommen als Speichermedium nur nichtflüchtige Datenspeicher wie Flashspeicher in Frage. Zudem schreibt das Bundesamt zur Sicherheit in der Informationstechnik ein Sicherheitsmodul vor. Durch das Rechenwerk können Messungen zu verschiedenen Tarifzeiten realisiert werden. Dies ist jedoch auch bereits mit Ferraris-Zählern möglich. (Zeller, 2015, p. 16) Der wesentliche Unterschied zu analogen Ferraris-Zählern besteht deshalb in der Handhabe der Stromverbrauchsmesswerte nach der reinen Messung und Speicherung. Durch das mit dem elektronischen Zähler verbundene Kommunikationsmodul sind Smart Meter in der Lage, die gemessenen Daten über eine Kommunikationsinfrastruktur an andere Parteien weiterzuleiten. Darüber hinaus können Smart Meter auch Daten empfangen. Dies ermöglicht einen bidirektionalen Datenaustausch mit anderen Akteuren des Elektrizitätssystems und macht im Falle eines digitalisierten Elektrizitätsnetzes ebendieses „intelligent“. Das Kommunikationsmodul wird häufig auch als „Gateway“ bezeichnet. Zur Übertragung der Messwerte über größere Entfernungen können verschiedene Technologien zum Einsatz kommen. Die im Haushaltssektor am häufigsten verbreitete Methode ist die Übertragung per digitaler Datenleitung (digital subscriber line DSL) mit 35,9%, gefolgt von der Übertragung per Mobilfunk (global System for mobile communications GSM) mit 23,0% und der Übertragung per Stromleitung (power line communications PLC) mit 22,6%. Letztere wird vor allem in Städten eingesetzt, da die Technologie ohne Einsatz von Signalverstärkern auf eine maximale Entfernung zwischen den Gateways von nur 300 Metern beschränkt ist. Die Übertragung der Daten mithilfe des Telefonnetzes (public switched telephone network PSTN) ist ebenfalls möglich, wird aber nur in 4,1% der Fälle angewandt. (Zeller, 2015, pp. 16-17) Smart Meter zeichnen sich daher nicht nur durch die elektronische Messung des Stromverbrauchs, sondern auch durch das integrierte Kommunikationsmodul, das durch verschiedene Übertragungstechnologien mit dem Elektrizitätsnetz verbunden ist und eine bidirektionale Kommunikation ermöglicht, aus. Nachdem hier bereits die möglichen Übertragungstechnologien geschildert wurden, wird in 2.1.5 näher untersucht, welche möglichen Schnittstellen sich durch diese Vernetzung ergeben.

Die folgende Matrix bietet einen Überblick über die verschiedenen Stromzähler hinsichtlich der Art der Stromverbrauchsmessung und der Existenz eines Kommunikationsmoduls.

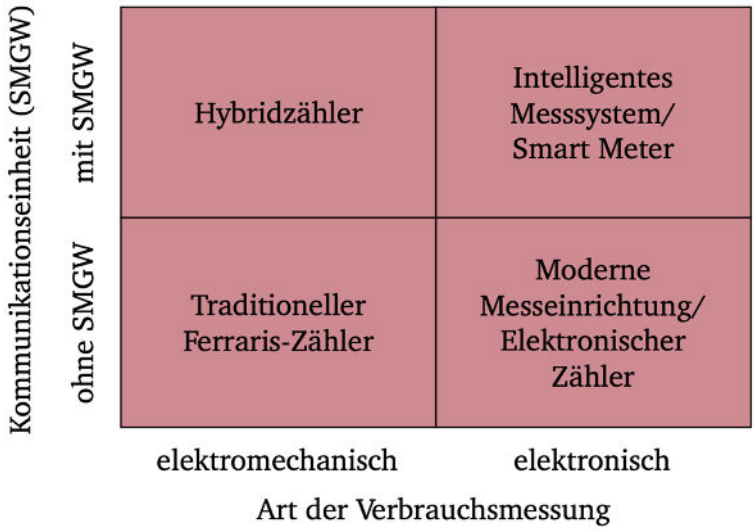



Abbildung 1: Verschiedene Stromverbrauchszähler im Überblick

### 2.1.2. Nutzen und Chancen

Smart Meter können hinsichtlich ihrer Nutzungsweise und Implementierung erhebliche Unterschiede aufweisen. Zudem ist die durch die Nutzung von Smart Metern potentiell realisierbare Wertschöpfung stark von sowohl der Marktdurchdringung und Implementierung weiterer Smart Meter, als auch dem allgemeinen Digitalisierungsgrad des Stromnetzes abhängig. Im Folgenden sollen daher die möglichen Nutzungsmöglichkeiten, Vorteile und Chancen, die durch Smart Meter entstehen, erörtert werden. Angefangen bei dem bereits bei einfacher Installation eines einzelnen Smart Meters möglichen Wertgewinn und Nutzen, werden später vor allem die Potentiale einer flächendeckenden Integration von intelligenten Stromzählern und deren Unterstützung bei der Bewältigung der Herausforderungen des Elektrizitätssystems der Zukunft herausgearbeitet.

Der erste Vorteil von Smart Metern liegt in der so geschaffenen Transparenz und den generierten Informationen rund um den Stromverbrauch und ist unabhängig von der Frequenz der Übermittlung der Messwerte an den MSB und unabhängig von der allgemeinen Marktdiffusion von Smart Metern realisierbar. Durch Smart Meter können Kunden zu jeder Zeit mit nur geringer Zeitverzögerung den aktuellen Stromverbrauch sehen und überwachen. Auf diese Weise können durch einzelne Lastsprünge stromzehrende Haushaltsgeräte oder durch einen hohen Grundlastverbrauch Geräte mit hohem Stand-by-Verbrauch aufgedeckt werden. Insgesamt können die Informationen der Smart Meter die Energieverbrauchsberatung unterstützen. Auch eine automatisierte Energieberatung durch Auswertung der Lastverläufe ist denkbar. Hierfür können auch anonymisierte Nachbarschaftsvergleiche oder Ranglisten genutzt werden. (Zeller, 2015, p. 76) Zudem können nach längerer Nutzung Periodenvergleiche durchgeführt und so Entwicklungen im Stromverbrauch erkannt werden. Pilotprojekte haben diesbezüglich gezeigt, dass die so geschaffene Transparenz und Motivation zum Strom sparen den Stromverbrauch reduzieren kann. (Matsui, Ochiai, & Yamagata, 2014, pp. 164-165) Die Einsparpotentiale durch solche Feedbacksysteme liegen in verschiedenen Studien zwischen 5-15% gegenüber dem bisherigen Verbrauch. (Riester, 2017, p. 25) Auch wenn Periodenvergleiche und die

Messung des aktuellen Stromverbrauchs bereits mit analogen Ferraris-Zählern möglich ist, so ist dies vor allem mit deutlichem Mehraufwand verbunden. Durch Smart Meter können die gemessenen Werte direkt per App und In-house Display ausgelesen werden. (Sirojan et al., 2019, p. 1) Ein besonderer Vorteil liegt hier vor allem in der möglichen Visualisierung der aktuellen und historischen Daten. Gegebenenfalls kann der Stromverbrauch so auch an das, durch die Einspeisung aus erneuerbaren Energien zunehmend variierende Stromangebot angepasst werden. Die auf diese Weise möglichen zeitvariablen Stromtarife bieten dann zu Zeiten mit einem Überangebot an Strom niedrigere Strompreise an, sodass die Verbraucher zu diesen Zeiten die Chance haben, Kosten zu sparen. (Arlt & Wolling, 2011, p. 3) Durch die zusätzliche Nutzung zeitvariabler Tarife können Endverbraucher also nicht nur durch eine allgemeine Reduzierung, sondern auch durch eine Verlagerung des Verbrauchs Stromkosten einsparen. Neben der so geschaffenen Motivation zum Kosten sparen besteht aber auch ein ökologischer Aspekt, denn die Anpassung des Verbrauchs führt indirekt auch immer zu einer Steigerung des Anteils von Strom aus erneuerbaren Energiequellen am Gesamtverbrauch. (Zeller, 2015, p. 97) Im Fall der Installation und Nutzung eigener PV-Anlagen bieten Smart Meter zudem die Möglichkeit der Netto-Messung, also der Einspeisung des selbstproduzierten Stroms ins Netz in Zeiten, in denen der selbst produzierte Strom den eigenen Bedarf übersteigt. Weiteren Nutzen stiften die Möglichkeiten der Energiequalitätsüberwachung, beispielsweise anhand von Spannung und Frequenz, sowie die Aufdeckung von Energiediebstahl. (Sirojan et al., 2019, p. 1) Außerdem können bei zeitvariablen Tarifen neben dem detaillierten historischen Verbrauch auch die historischen Kosten detailliert visualisiert werden. Im Fall von außergewöhnlich hohem Stromverbrauch können Smart Meter den Nutzer zudem durch Warnungen alarmieren. (Sirojan et al., 2019, p. 2)

Neben der geschilderten, durch Smart Meter geschaffenen Transparenz, die für jedwede Endverbraucher hilfreich und vorteilhaft ist, ist die Anwendung von Smart Metern diesbezüglich vor allem im Bereich der Geschäftskunden sehr vielversprechend. Die höhere Transparenz bietet für die Kunden die Möglichkeit, Beratungsdienstleistungen in Anspruch zu nehmen. Diese werden oft als „Smart Consulting“ bezeichnet und können auch automatisiert erfolgen. Die Aufbereitung hoch aufgelöster Stromverbrauchsdaten kann zudem Bestandteil der innerbetrieblichen Energie- und Ressourcenplanung sein. Aus den Verbrauchsdaten können auch Indikatoren zur Unternehmenssteuerung abgeleitet werden. (Zeller, 2015, p. 85) Abweichungen im Lastprofil im historischen Vergleich können auf unterschiedliche Betriebszustände und eventuelle Defekte einzelner Anlagen hindeuten, wodurch diese frühzeitig erkannt werden können. In der Wohnungswirtschaft kann der Stromverbrauch von Leerständen minimiert werden. (Zeller, 2015, p. 85) Auch für Handelsketten besteht viel Potential. Über Benchmark-Vergleiche der einzelnen Niederlassungen können Einsparpotentiale identifiziert und der Vertrieb evaluiert und optimiert werden. (Knab & Konnertz, 2011, pp. 7-8)

Smart Meter ermöglichen ferner durch die mögliche ZFA eine Automatisierung des Abrechnungsprozesses und des operativen Geschäfts von MSBs und EVUs. So muss die Ablesung der Zählerstände nicht mehr manuell und persönlich vor Ort durch einen Angestellten des MSB, durch den Versand von Selbstablesekarten oder durch Übermittlung des Zählerstandes durch den Kunden über ein Online-Portal erfolgen, sondern kann von dem MSB automatisiert durch eine Fernablesung durchgeführt werden. (Petsch et al., 2012, p. 11) Die dadurch eingesparten Kosten bei der Ablesung

und Abrechnung erlauben dann auch kürzere Abrechnungszyklen, beispielsweise monatliche Abrechnungen wie in der Telekommunikationsbranche. Für EVUs verringert sich so das betriebsnotwendige Kapital und EVUs und Kunden profitieren beide von einem geringeren Mehr- oder Mindermengenrisiko. Zudem sinken sowohl das finanzielle Ausfallrisiko als auch die Prozesskosten für das Forderungsmanagement. (Zeller, 2015, p. 78) Somit schaffen intelligente Stromzähler auch günstige Voraussetzungen für die in § 40 Abs. 2 EnWG geforderte Realisierung von unterjährigen Abrechnungen und zeitvariablen Tarifen. Kunden können über das Online-Portal theoretisch für jeden gewünschten Zeitraum eine Abrechnung erhalten. (Petsch et al., 2012, p. 12) Damit entscheidet vor allem die Effizienz des Messstellenbetriebs über die Höhe der Zusatzkosten oder Kosteneinsparungen durch Smart Meter. (Zeller, 2015, p. 95) Weitere Kosteneinsparungen durch die Optimierung und Automatisierung des operativen Geschäfts entstehen zudem bei den Vertriebskosten der EVUs. Sowohl die direkten Kosten für die Stromlieferung als auch die Gemeinkosten des Betriebsablaufs können durch die effizientere Kundenabwicklung und die einfachere Sperrung und Entsperrung von Kunden gesenkt werden. (Zeller, 2015, p. 93)

Zuvor wurde bereits auf die durch Smart Meter möglichen zeitvariablen, dynamischen Stromtarife eingegangen. Diese ermöglichen eine direkte oder indirekte zeitnahe Steuerung der Stromnachfrage. (Zeller, 2015, p. 81) Mit einer durch zeitvariable Tarife gesteuerten Last können erhebliche Einsparungen realisiert werden. Insbesondere kann hier der Bereich Elektromobilität durch die Einführung von Geschäftsmodellen im Bereich Ladestrom profitieren. (Zeller, 2015, p. 161) Zeitvariable Tarife können dabei gegenüber einer Beschaffung über das Standardlastprofil (SLP) für Haushalte einen spürbaren Mehrwert generieren. (Petsch et al., 2012, p. 10)

Wie bereits erwähnt machen Smart Meter eine Netto-Messung des Stromverbrauchs für Kunden mit eigener Stromerzeugung aus PV-Anlagen möglich. Diese an Gebäuden installierten, dezentralen PV-Anlagen haben einen wesentlichen Anteil beim Ausbau erneuerbarer Energien und können einen Teil des eigenen Strombedarfs decken. Die Eigennutzung der so generierten Elektrizität entlastet die Verteilnetze, welche besonders durch die Einspeisung aus dezentralen PV-Anlagen belastet werden. Daher werden Eigenbedarfsquoten für Neuanlagen gefordert und für Bestandsanlagen gefördert. (Zeller, 2015, p. 154) Der Anteil der selbst produzierten Elektrizität zur Deckung des Gesamtstrombedarfs eines Haushalts kann durch Smart Meter in Verbindung mit einer intelligenten Verbrauchssteuerung erhöht werden. Eine solche Eigenbedarfsoptimierung bringt mehrere Vorteile mit sich. Im Wesentlichen ist hier die Kosteneinsparung durch die vermiedenen Nebenkosten der Strombeschaffung, wie Netznutzungsentgelte, Vertriebskosten und Margen der EVUs, zu nennen. Zusätzlich wird so die Autarkie erhöht, was Kundenwünschen nach mehr Selbstbestimmung gerecht wird. Auch der ökologische Effekt durch die Nutzung erneuerbarer Energie wirkt sich positiv aus. (Zeller, 2015, p. 82) Um den Verbrauch besser an die Erzeugung anzupassen, spielen vor allem Prognosen der Stromerzeugung aus erneuerbaren Energien eine wichtige Rolle. (Zeller, 2015, p. 80) Simulationen von Zeller haben jedoch gezeigt, dass durch Anpassung der Last eine Steigerung des Eigenanteils von weniger als 5% möglich ist. Dies ist auf eine hohe Korrelation zwischen dem Standardlastprofil und der Saisonalität der PV-Einspeisung zurückzuführen. (Zeller, 2015, p. 155) Mit zunehmender Elektrifizierung der Sektoren Wärme und Mobilität und zunehmender Marktdurchdringung von zeitvariablen Stromtarifen wäre aber zu erwarten, dass sich die

Auswirkungen von Eigenbedarfsoptimierungen durch Smart Meter deutlich erhöhen, da Saisonalität der PV-Einspeisung und Stromnachfrage nicht länger so stark korreliert wären.

Die bisher genannten Vorteile, die durch die Nutzung von Smart Metern entstehen, lassen sich überwiegend individuell und unabhängig von der weiteren Implementierung von Smart Metern realisieren. Allerdings bieten Smart Meter auch viele Chancen und ein hohes Potential durch einen flächendeckenden Roll-out und die daraus folgende Marktdurchdringung, die sich auf rein individueller Ebene nicht realisieren lassen. Dies ist vor allem auf die zentrale Stellung der Smart Meter als Gateway zwischen Unternehmen und Letztverbrauchern und die Bedeutung der Technologie innerhalb der Stromnetze der Zukunft zurückzuführen. (Zeller, 2015, p. 17) Durch die von Smart Metern generierten Daten kann ein besseres Verständnis und eine bessere Segmentierung des Absatzmarktes geschaffen werden. Durch diese Erkenntnisse über die Letztverbraucher können die beteiligten Unternehmen dann den Beschaffungsmarkt optimieren und besser an den Absatzmarkt anpassen. Dafür ist es zunächst hilfreich, sich die aktuellen Standardprozesse für die verschiedenen Marktsegmente vor Augen zu führen. Die unterschiedlichen Kundensegmente unterscheiden sich teils stark hinsichtlich ihrer Präferenzen und Lastprofile. (Zeller, 2015, p. 68) Für Großkunden der Industrie bieten EVUs aufgrund der hohen Lukrativität der Einzelkunden in der Regel meist an den individuellen Bedarf angepasste Stromtarife an. Das Kundensegment der GHD ist sehr heterogen, was sich auch in der Abrechnung widerspiegelt. Zur Verbrauchsabrechnung von Kleinkunden des GHD Sektors kommen hier zehn unterschiedliche, branchenspezifische Lastprofile zu Anwendung. (Zeller, 2015, p. 62) Für Privathaushalte hingegen wird nur ein SLP verwendet. Dieses unterstellt jedem Kunden aus dem Sektor der Privathaushalte das gleiche Durchschnitts-Profil, welches dann über den Vorjahresverbrauch skaliert wird. Diese SLPs und die dazugehörigen Verfahren wurden im Zuge der Liberalisierung des Messwesens eingeführt und werden von der Mehrzahl der Verteilnetzbetreiber als Basis für die Netznutzungsverträge und Einspeisung der EVUs genutzt. (Zeller, 2015, p. 126) Durch die weitläufige Implementierung von Smart Metern besteht hier die Chance, diese Lastprofile besser zu individualisieren und an die einzelnen Kundenbedürfnisse anzupassen. Im Zuge einer solchen Individualisierung könnten durch die Messdaten neue Dimensionen in die Kundensegmentierung aufgenommen werden, um die Vertriebsprodukte der EVUs besser an die Kunden anzupassen. Durch Clusteranalysen können Kundensegmente mit ähnlichen Lastprofilen zusammengefasst werden. (Ramos & Vale, 2008, p. 7) Auf diese Weise besteht auch für Privathaushalte die Möglichkeit, differenzierte und individuell angepasste Preisangebote wahrzunehmen, die bisher nur für Großkunden angeboten wurden. Das Portfolio-Management der EVUs kann durch die zusätzlichen Informationen über den Verbrauch der Kunden und die bessere Segmentierung der Kunden bessere Preisprognosen erstellen und somit für seine Kunden ein günstigeres Beschaffungs-Portfolio realisieren. (Zeller, 2015, pp. 78-79) Smart Meter Daten könnten es so ermöglichen, Durchschnittsverbräuche und Verbrauchsanpassungen durch direkte oder indirekte Steuerung zukünftig sektorgenau zu antizipieren. Hier konnten in Studien zur Verbrauchsmessung der Privathaushalte bereits Cluster identifiziert werden. (Zeller, 2015, p. 163) Neben einer ausreichend weiten Marktdurchdringung wäre hierfür aber vor allem die Erfassung und Übermittlung der Stromverbrauchsdaten des privaten Sektors in höherer Auflösung, also in kürzeren Messintervallen, notwendig, als dies derzeit bei Privathaushalten zu reinen Abrechnungszwecken durchgeführt wird. Prinzipiell wären möglichst kurze Messintervalle vorteilhaft, allerdings ist bereits die tägliche Erfassung der Lastgänge als positiv für die Beschaffung zu bewerten,

denn die Prognose des zukünftigen Strombedarfs der Privathaushalte auf Basis täglicher Werte als Grundlage für die Beschaffung ist genauer als die traditionelle Beschaffung über einfache SLPs. (Petsch et al., 2012, p. 10) Durch die Anpassung und Segmentierung der Lastprofile durch genauere Daten kann dann eine Beschaffungsmarktoptimierung erreicht werden. EVUs können durch die zusätzlichen Informationen am Beschaffungsmarkt zu günstigeren Zeitpunkten Verträge zur Stromlieferung abschließen, was die Beschaffungskosten senkt. Für EVUs entsteht so allerdings ein Mengenrisiko durch die Prognose des Stromverbrauchs, das bei traditioneller Belieferung nach SLP-Verfahren noch teilweise vom Verteilnetzbetreiber mitgetragen wird. (Zeller, 2015, p. 94) Jedoch ist insbesondere durch eine fortschreitende Elektrifizierung der Sektoren Wärme und Mobilität eine zukünftig weitaus stärkere Segmentierung des Absatzmarktes zu erwarten, was die Beschaffung für die Haushaltskunden über ein einziges SLP ohnehin sehr fragwürdig macht und die Kosteneinsparpotentiale durch Beschaffungsmarktoptimierung nochmals erhöhen dürfte.

Ein weitläufiger Smart Meter Roll-out in Verbindung mit kurzen Messintervallen würde zudem ein großes Potential zur Netzoptimierung bieten. Die zuvor erörterten Potentiale zur Eigenverbrauchsoptimierung, Beschaffungsmarktoptimierung, Anpassung der Lastprofile und automatisierten Steuerung der Stromnachfrage hätten eine Entlastung der Stromnetze zur Folge. Dies würde für die Letztverbraucher geringere Netzentgelte mit sich bringen. Netzentgelte sind die Kosten für den physischen Transport der Elektrizität und bestehen aus den Netzentgelten der Übertragungs- und Verteilungsebene. Sie unterliegen der allgemeinen Regulierung. Durch Eigenbedarfsoptimierung werden die Verteilnetze entlastet, da weniger Energie ins Verteilnetz eingespeist wird, während eine Beschaffungsmarktoptimierung eine Entlastung der Verteilnetze durch die indirekt stattfindende Anpassung an die Einspeisung erneuerbarer Energien verursacht. Diese Entlastung der Netze würde sich für Kunden in geringeren Stromkosten widerspiegeln. (Zeller, 2015, p. 95) Weitere Kosteneinsparpotentiale ergeben sich durch die breitere Datenbasis und die dadurch mögliche bessere Prognose des Stromverbrauchs, sowie die Flexibilisierung der Stromnachfrage, was die Kosten für Mehr- und Mindermengen, den Bedarf der Netzbetreiber an Regelenergie sowie die maximale Netzlast reduziert. Insgesamt ergeben sich so geringere Systemkosten. (Cramton, Ockenfels, & Stoft, 2013) Entscheidend ist hierfür neben einer flächendeckenden Implementierung einer Smart Meter Infrastruktur aber auch, dass die so erreichten Kosteneinsparungen im Netzbetrieb und -ausbau auch tatsächlich von den Unternehmen an die Kunden weitergereicht werden und sich nicht nur als eine Margenerhöhung der EVUs oder anderer Akteure manifestieren.

Zusammenfassend kann also gesagt werden, dass intelligente Stromzähler eine ganze Reihe von Nutzungsmöglichkeiten und Chancen für Letztverbraucher bieten und nennenswerte Potentiale zur Optimierung der operativen und teils auch strategischen Geschäftsprozesse der beteiligten Unternehmen des Elektrizitätsmarktes mit sich bringen. Einige dieser Vorteile, wie die erhöhte Transparenz, die detailliertere Bereitstellung und Visualisierung von Verbrauchsdaten, die Möglichkeit zur Nutzung zeitvariabler Stromtarife und die indirekte oder automatisierte Steuerung der Stromnachfrage, sowie die Eigenbedarfsoptimierung lassen sich aus Sicht der Privathaushalte bereits jetzt realisieren. Die Potentiale zur Beschaffungsmarktoptimierung, Flexibilisierung der Gesamtnachfrage, Netzoptimierung und Segmentierung der Lastprofile, sowie die daraus

resultierenden geringeren Systemkosten, hängen hingegen vor allem von dem Grad der allgemeinen Marktdurchdringung von Strom Smart Metern und der Kürze der Messintervalle ab.

### 2.1.3. Schwächen und Risiken

Auch wenn Strom Smart Meter, wie im vorherigen Teil geschildert, viele Vorteile und Chancen mit sich bringen, birgt die Technologie auch einige Schwächen, Risiken und Bedenken, auf die im Folgenden ebenfalls eingegangen wird. Diese lassen sich hauptsächlich in die entstehenden Kosten für die Kunden, die entstehenden Investitionskosten für die Unternehmen in die IT-Infrastruktur und Datenschutzfragen zusammenfassen.

Wie im vorangegangenen Teil diskutiert, können Smart Meter für Endverbraucher viele Nutzungsmöglichkeiten und Wertsteigerungen bieten. Neben den ökologischen Effekten durch die Reduktion des Energieverbrauchs und die bessere Ausnutzung des Stroms aus erneuerbaren Energien, belaufen diese sich aber im Wesentlichen auf eine Senkung der Energiekosten. Fraglich ist allerdings ob diese Kosteneinsparpotentiale nicht durch die neu entstehenden Kosten durch die Implementierung von Smart Metern neutralisiert oder gar ins Negative umgekehrt werden. Wenn private Haushalte keine PV-Anlage und somit auch kein Bedürfnis nach Eigenbedarfsoptimierung haben und solange Strom Smart Meter die Ausnahme bei Kleinstverbrauchern darstellen, beschränken sich die Kosteneinsparpotentiale durch Smart Meter lediglich auf die Möglichkeit, zeitvariable Tarife zu verwenden, die Motivation, durch die neuen Informationen über den Stromverbrauch Energie zu sparen und die automatisierte Auslesung der Messwerte. Da der Effekt von zeitvariablen Tarifen ohne gleichzeitige automatisierte Steuerung der Haushaltsgeräte zweifelhaft ist und eine zeitnahe Verbrauchskontrolle theoretisch auch mit Ferraris-Zählern möglich ist, stellt sich für Endverbraucher klar die Frage nach den durch Smart Meter entstehenden Zusatzkosten. Derzeit ist eine Implementierung von Smart Metern ohne sonstige Zusatzleistungen aus Kundensicht daher sehr kritisch zu bewerten. Trotz eventueller Einsparungen durch eine Reduktion des Stromverbrauchs und effizientere Abrechnungsprozesse ist es sehr fraglich, ob diese Einsparungen die Zusatzkosten durch die intelligenten Stromzähler übersteigen. Essenziell für eine wirkliche Wertsteigerung aus Kundensicht wäre daher eine flächendeckende Implementierung der intelligenten Stromzähler, sodass die so geschaffene Optimierung der Geschäftsprozesse der Unternehmen des Elektrizitätssektors, die Absatzmarktsegmentierung und die Netzentlastung mit den implizierten Kosteneinsparungen die Stromkosten der Letztverbraucher auch tatsächlich nennenswert senken.

Einen weiteren Kritikpunkt stellen die Möglichkeiten der Fremdkontrolle und des Datenmissbrauchs dar. (Zeller, 2015, p. 75) Der Datenschutz der übermittelten Energieverbrauchsdaten soll jedoch durch notwendige Zertifizierungsverfahren zur Datensicherheit gewährleistet werden, die jeder Anbieter von Strom Smart Meter durchlaufen muss. Für Deutschland setzt das Bundesamt für Sicherheit in der Informationstechnik die technischen Richtlinien fest. (BSI, 2019) Der Kritikpunkt des Datenschutzes ist insbesondere in Verbindung mit dem zuvor genannten Kritikpunkt der Zusatzkosten zu betrachten. Zwar kann der Nachteil der Zusatzkosten durch einen flächendeckenden Roll-out, kurze Messintervalle und die so implizierten Kosteneinsparungen eliminiert werden, die Bedenken rund um den Datenschutz dürften in diesem Fall hingegen sogar weiter zunehmen, was eine wirkliche Verbesserung

der Situation zumindest umstritten macht. Insbesondere mit weiterer Individualisierung der Tarife und der so steigenden Komplexität der Messwerterfassung steigen die Fragen rund um den Datenschutz. Für die Bereitstellung von last- und verbrauchsvariablen Tarifen ist eine viertelstündige Verbrauchsmessung wie im Großkundensektor zudem nicht notwendig und somit datenschutzrechtlich unbegründet. Allerdings gehen dadurch auch die Vorteile bei der Beschaffungsmarktoptimierung verloren. (Petsch et al., 2012, pp. 13-14)

Zusätzlich entstehen bei einer flächendeckenden Implementierung und Messungen in kurzen Intervallen Probleme auf Seiten der Unternehmen. Hier erschweren sowohl die zur vollen Ausschöpfung des Nutzens notwendigen kurzen Messintervalle, als auch das eventuelle Fehlen von adäquater, günstiger Bandbreite zur Kommunikation und Übermittlung der Messwerte, sowie fehlende Ressourcen zur Analyse der immensen Datenmengen eine erfolgreiche Implementierung. Im Fall einer Viertelstundenmessung erfordert die Übertragung all der generierten Daten an einen zentralen Server oder eine Cloud eine hohe Bandbreite und verursacht eine ressourcen-intensive Verarbeitung der Daten. Die Daten können je nach Größe und bestehender IT-Infrastruktur der Unternehmen mit der aktuellen Technologie nur schwer zu verarbeiten sein, sodass die Unternehmen neue Investitionen tätigen müssen. (Petsch et al., 2012, p. 11; Sirojan et al., 2019, pp. 1-2) Neben den Kosten durch die Anschaffung der Smart Meter auf der Kundenseite entstehen so auch auf der Seite der Unternehmen Kosten durch die Anpassung der existierenden Geschäftsprozesse und die Investitionen in die IT-Infrastruktur. (Petsch et al., 2012, pp. 13-14) Neben den Problemen durch die großen entstehenden Datenmengen im Zuge einer hochfrequentierten Verbrauchsdatenerfassung kann zudem die Heterogenität und Inkompatibilität der von Nutzern verwendeten Zählertechnologien zu Herausforderungen führen. Da die Rolle des MSB von EVUs, Netzbetreibern und dritten Dienstleistern übernommen werden kann und daher mittlerweile auch viele verschiedene Anbieter von Smart Metern existieren und verschiedenste Produkte anbieten, kann dies ein erstzunehmendes Problem darstellen. Kommen in einem Netzgebiet verschiedene Zählertechnologien zum Einsatz, muss der MSB ein heterogenes ZFA-System aufbauen, das bei der Fernauslesung mit der Heterogenität der Zähler umgehen kann. Für kommunale MSBs können sich so ernste IT-Strukturprobleme ergeben. Zudem muss bezüglich jeder Zählertechnologie ausgebildetes Personal vorhanden sein, um im Fall von Fehlern bei der Messung oder Übermittlung der Messwerte entsprechend reagieren zu können. (Petsch et al., 2012, p. 11) Solange innerhalb eines Netzes verschiedene Technologien zum Einsatz kommen und die Verfahrensweise der Auslesung je Zählertyp unterschiedlich ist, müssen die verschiedenen notwendigen Prozesse zur Auslesung und Verarbeitung der Daten jederzeit fehlerfrei durchführbar sein. (Petsch et al., 2012, pp. 13-14) Zudem erhöhen individuell angepasste Stromtarife die Anforderungen an die Rechnungslegung und die Visualisierung der Stromverbrauchsdaten und Stromkosten. All die so entstehenden Aufwendungen müssen den Unternehmen wirtschaftlich zumutbar sein. Zusätzliche Unsicherheit bei der Prozessrealisierung entsteht durch ungewisse regulatorische Vorgaben. (Petsch et al., 2012, pp. 13-14)

Weiterhin kann der Zielpluralismus bei der Bildung von zeitvariablen Tarifen und die so entstehenden Zielkonflikte zwischen EVU und Netzbetreiber problematisch sein. Das Interesse der Netzbetreiber liegt in der Vermeidung eines unnötigen weiteren Netzausbaus und der Verringerung der Netzverluste durch die bessere Anpassung der Stromnachfrage an die fluktuierende Einspeisung aus erneuerbaren

Energien und durch Eigenbedarfsoptimierung. Das Hauptziel der EVUs ist hingegen eine Kostenminimierung in der Beschaffung und der Abrechnung von Mehr- und Mindermengen und der Regelenergie. (Petsch et al., 2012, pp. 10-11) Die Regularien lassen jedoch eine Tarifbildung zu Gunsten des Netzbetreibers nicht zu. (Ifland, Exner, & Westermann, 2011, pp. 3-4)

Deutliche, negative Effekte entstehen am ehesten für die Betreiber von Regelenergiekraftwerken. Durch die zunehmende Flexibilisierung der Stromlast und die mögliche Netzoptimierung könnte sich die Auslastung solcher Kraftwerke reduzieren. (Zeller, 2015, p. 87)

### 2.1.4. Rechtliche Rahmenbedingungen und gesellschaftliche Akzeptanz

Ähnlich wie die gesamte Energiewirtschaft, unterliegt auch die Digitalisierung der Energiewirtschaft vielen politischen Regularien und Gesetzen und ist stark von gesetzlich festgelegten Zielvorgaben geprägt. (Riester, 2017, p. 16) Daher ist auch der Ausbau und die Nutzung von Strom Smart Metern durch eine Vielzahl von Gesetzen beeinflusst. Seit dem Beginn dieses Jahrtausends werden intelligente Stromzähler und Messsysteme weltweit zur Messung der dem Netz entnommenen Elektrizität bei Endverbrauchern eingeführt und genutzt. Beinahe jedes Land führt dies auf Basis von Gesetzen aus, die mehr oder weniger restriktiv einen flächendeckenden oder partiellen Ausbau der Technologie vorgeben. Die technische Detaillierung der vorgeschriebenen zu installierenden Systeme kann dabei von Land zu Land sehr unterschiedlich sein. (Herre & Freunek, 2020, p. 196) Insbesondere in Deutschland bestehen durch EU-weite Vorgaben und die deutschlandweit gesetzlich festgelegten Umsetzungsvorschriften ebendieser, sowie durch bundesspezifische Regularien, eine Vielzahl an Gesetzen, die den Ausbau intelligenter Messsysteme beeinflussen. Hinzu kommt, dass diese im Laufe der letzten Jahre mehrmals angepasst und verändert wurden, zudem Strategiewechsel stattfanden und sich der Ausbau allgemein, unter anderem durch Diskussionen rund um den Datenschutz und den tatsächlichen Nutzen der Technologie, sehr verzögert hat. Außerdem hängen die sich daraus ergebenden Implikationen für Endverbraucher von der Höhe des jährlichen Stromverbrauchs und der anlagentechnischen Ausstattung der Gebäude ab. Welche genauen historischen und aktuellen Gesetzesbeschlüsse existieren, die den Ausbau und die Nutzung der Technologie in Deutschland beeinflussen, ist nicht Ziel dieser Arbeit und würde aufgrund des Umfangs dieses Themengebietes den Rahmen dieser Arbeit sprengen. Diese Thesis erhebt daher nicht den Anspruch, eine vollständige und detailgetreue Erörterung dieser Regularien zu sein. Stattdessen sollen im Folgenden die wesentlichen legislativen Treiber und die groben gesetzlichen Rahmenbedingungen bezüglich des Ausbaus der Technologie geschildert werden.

#### Rechtliche Rahmenbedingungen

Auf europäischer Ebene wurden durch die EU Richtlinie 2006/32/EG des europäischen Parlaments und des Rates vom 5. April 2006 über Energieeffizienz und Energiedienstleistungen die Zeile zur Steigerung der Energieeffizienz und der Steuerung der Energienachfrage europaweit definiert. Durch den Einsatz von intelligenten Zählern zur Darstellung des Energieverbrauchs sollen Kunden in die Lage versetzt werden, ihren eigenen Energieverbrauch in Abhängigkeit von persönlichen Bedürfnissen besser zu optimieren. (Petsch et al., 2012, p. 4) Den Startpunkt für den Ausbau der Smart Meter

Technologie in Deutschland markierte dann das Gesetz zur Digitalisierung der Energiewende (GDEW), um der Forderungen der EU nach rechtlichen Grundlagen für den Ausbau in den Mitgliedsstaaten nachzukommen. Das GDEW legt die Rahmenbedingungen für die Nutzung und den Ausbau von Smart Metern in dem Messstellenbetriebsgesetz (MsbG) fest und verpflichtet so die zuständigen MSBs zum Einbau von intelligenten Messsystemen. (Riester, 2017, p. 20) Bezüglich der Smart Meter selbst ist eine bidirektionale Kommunikation von Verbrauchs- und Tarifinformationen, sowie die Möglichkeit zur ZFA und Fernabschaltung verpflichtend vorgeschrieben. Durch die Normierung von Schnittstellen soll zudem die Erweiterung des Systems sichergestellt werden. (Zeller, 2015, p. 12)

Des Weiteren haben sich die Mitgliedsstaaten der EU im Rahmen der EU-Richtlinie 2009/72/EG des europäischen Parlaments und des Rates vom 13. Juli 2009 über gemeinsame Vorschriften für den Elektrizitätsbinnenmarkt neben einer weiteren Öffnung des Elektrizitätsmarktes auch zur Förderung intelligenter Messsysteme und Netze verpflichtet. Die für Deutschland daraus entstehenden Verpflichtungen zum Einbau von Strom Smart Metern werden in § 21 des Energiewirtschaftsgesetzes (EnWG) geregelt. Seit 2010 ist der Einbau von Strom Smart Metern bereits bei Renovierungen und Neuanschlüssen, also insbesondere bei Neubauten, verpflichtend vorgeschrieben. Diese Pflicht zum Einbau wurde mit der Novellierung des EnWG in 2012 auch auf Endkunden mit einem Jahresstromverbrauch von über 6000 kWh ausgeweitet (§ 21c Abs. 1 EnWG) und soll im Zuge von Turnuswechseln umgesetzt werden. Auch für Anlagen zur Kraft-Wärme-Kopplung (KWK) und für PV-Anlagen bestimmter Auslegungen sind intelligente Messsysteme Pflicht. In Kombination damit werden EVUs mit § 40 (5) EnWG dazu verpflichtet, einen Tarif anzubieten, der „Anreize zur Energieeinsparung und Steuerung des Energieverbrauchs“ setzt. (Antic, 2015, p. 14; Petsch et al., 2012, pp. 4-5; Zeller, 2015, p. 12) Weiter verkompliziert werden die Vorschriften zum Ausbau der Technologie zudem durch die Unterscheidung in moderne Messeinrichtungen und tatsächlich intelligente Stromzähler. Während intelligente Stromzähler durch das integrierte Smart Meter Gateway (SMGW) bidirektional mit anderen Parteien des Elektrizitätsnetzes Verbrauchs- und Tarifinformationen kommunizieren können, sind moderne Messeinrichtungen lediglich elektrische Stromverbrauchszähler, die durch die spätere Verbindung mit einem Gateway zu einem echten Strom Smart Meter aufgerüstet werden können. Die aktuelle Gesetzeslage schreibt vor, dass bis 2032 alle Endverbraucher mit einer modernen Messeinrichtung ausgestattet werden sollen und fordert so den Ersatz der herkömmlichen Ferraris-Zähler durch elektronische Zähler mit der Möglichkeit zur Aufrüstung zum Smart Meter. (Riester, 2017, pp. 20-21; Rigoll, 2017, p. 57) Eine solche Aufrüstung ist aber auch weiterhin vorerst nur für Verbräuche über 6000 kWh pro Jahr vorgeschrieben. Somit wird auch zukünftig der Großteil der Privathaushalte nicht von dem Ausbau der Smart Meter Technologie betroffen sein. Es steht den MSBs aber frei, diese Aufrüstung bei den eigenen Kunden auch unterhalb dieser Verbrauchsgrenze unter Einhaltung der Preisvorgaben durchzuführen. Die genauen nach Verbrauchsklassen gestaffelten Einbauzeiträume und Preisobergrenzen sind in § 31 MsbG festgelegt und im Folgenden tabellarisch dargestellt. Zukünftige Abweichungen von dem Plan sowie zeitliche Verzögerungen sind allerdings nicht auszuschließen. (Riester, 2017, p. 21)

| Letztverbraucher (§ 31, Abs. 1, Satz 1-6) | Ab (Zeitraum) | Preis-Obergrenze |
|---|---|---|
| Über 100.000 kWh/ Jahr | 2017 (16 Jahre) | „angemessen“ |
| 50.000 – 100.000 kWh/ Jahr | 2017 (8 Jahre) | 200 € |
| 20.000 – 50.000 kWh/ Jahr | 2017 (8 Jahre) | 170 € |
| 10.000 – 20.000 kWh/ Jahr | 2017 (8 Jahre) | 130 € |
| 6.000 – 10.000 kWh/ Jahr | 2020 (8 Jahre) | 100 € |
| Anlagenbetreiber (§ 31, Abs. 2, Satz 1-4) | Ab (Zeitraum) | Preis-Obergrenze |
| Über 100 kW | 2020 (8 Jahre) | „angemessen“ |
| 30 – 100 kW | 2017 (8 Jahre) | 200 € |
| 15 – 30 kW | 2017 (8 Jahre) | 130 € |
| 7 – 15 kW | 2017 (8 Jahre) | 100 € |

Tabelle 1: Übersicht der Vorgaben des GDEW zu Einbauzeiträumen und Preisobergrenzen nach § 31 MsbG (Riester, 2017, p. 21)

Auch wenn somit für viele der Privathaushalte der Einbau von elektronischen Stromzählern mit Gateway lediglich optional ist, ist ein flächendeckender Rollout das langfristige Ziel. (Riester, 2017, p. 91) Großkunden mit einem Jahresverbrauch von über 100.000 MWh sind neben der Nutzung von Smart Metern auch zu einer registrierten Leistungsabmessung (RLM) verpflichtet. (Zeller, 2015, p. 58) Dabei wird der Elektrizitätsverbrauch in 15-Minuten Intervallen gemessen und an den MSB übermittelt. Die so gewonnenen Informationen sind, neben der so geschaffenen Möglichkeit zur Bildung und Wahrnehmung von individuell zugeschnittenen Lastprofilen und Stromtarifen, aufgrund der Höhe der Stromnachfrage wichtig, um die Stabilität der Verteilnetze sicherzustellen.

Ein weiterer Meilenstein in der gesetzlichen Regulierung der Energiewirtschaft war darüber hinaus die Liberalisierung und Deregulierung des leistungsgebundenen Energiemarktes im Rahmen der EU-Richtlinien zur Schaffung eines Binnenmarktes. Es ist somit gesetzlich geregelt, dass die Bereiche Erzeugung, Handel und Vertrieb der versorgungswirtschaftlichen Wertschöpfungskette dem freien Wettbewerb unterliegen. Lediglich der Betrieb der Übertragungs- und Verteilnetze bleibt als natürliches Monopol bestehen. (Schweinfurth, 2020, p. 228) Auch der Messstellenbetrieb erfuhr im Zuge dessen eine wettbewerbsorientierte Liberalisierung. Auf Gesetzesebene wurde dies durch die Novellierung des EnWG in 2009 umgesetzt. Nach § 21b EnWG haben Endverbraucher so die Möglichkeit, einen Dritten mit der Abwicklung der Verbrauchsmessung zu beauftragen. (Petsch et al., 2012, pp. 4-5; Zeller, 2015, p. 11) Zuvor lag der Messstellenbetrieb im Verantwortungs- und Aufgabenbereich der Verteilnetzbetreiber. (Schweinfurth, 2020, p. 228)

Mit der Liberalisierung des Messwesens wurde auch der Grundstein für eine marktorientierte Einführung der Smart Meter geschaffen. (Zeller, 2015, p. 21) Diese marktorientierte Einführungsstrategie ist aber auch einer der wesentlichen Gründe für die bisher geringe Marktdurchdringung der Smart Meter Technologie in Deutschland. Ursprünglich sollte im Rahmen der EU-Richtlinie 2009/72/EG von 2009 und der EnWG Novellierung in 2011 eine Ausstattung von mindestens 80% der Endverbraucher mit Smart Metern bis zum Jahr 2020 unter Berücksichtigung einer Wirtschaftlichkeitsbetrachtung forciert werden. (Antic, 2015, p. 13; Petsch et al., 2012, p. 4; Zeller, 2015, p. 12) Bei einer Kosten-Nutzen-Analyse zum Einsatz von Smart Metern, im Zuge eines Gutachtens im Auftrag des Bundesministeriums für Wirtschaft und Energie (BMWi), konnte aber ein

flächendeckender, vorgeschriebener Zählerwechsel zur Erreichung des 80%-Ziels bis 2020 keinen gesamtwirtschaftlichen Vorteil versprechen. (Edelmann & Kästner, 2013, p. 218) Die geringe Marktdurchdringung wird unter anderem daran deutlich, dass bei einer Erhebung zu aktuellen Zahlen der Verbreitung von Smart Metern im Jahr 2012 nur 4% der Haushaltsanschlüsse mit elektronischen Stromzähler, zuvor synonym auch als moderne Messeinrichtung bezeichnet, ausgestattet waren und wiederherum nur etwa 10% dieser auch tatsächlich fernausgelesen wurden. (Zeller, 2015, p. 24)

Alternativ wäre gegenüber der von Deutschland verfolgten marktorientierten Einführungsstrategie ein flächendeckender, forcierter Zählerwechsel möglich, wie er beispielsweise in Italien und Schweden durchgeführt wurde. (Zeller, 2015, p. 25) Insgesamt erfolgt die Umsetzung der EU-Richtlinien zur Markteinführung der Smart Meter Technologie in den europäischen Mitgliedsstaaten auf sehr unterschiedliche Weise. Eine genaue Darstellung der einzelnen Einführungsstrategien würde den Rahmen dieser Arbeit sprengen. Stattdessen werden in der folgenden Darstellung kurz die Ergebnisse einer Studie der österreichischen Energie Agentur aus dem Jahr 2011 zum Stand der Umsetzung in den EU-Mitgliedsstaaten und Norwegen grafisch dargestellt. Die Darstellung erfolgt als Matrix mit den Dimensionen „rechtliche Grundlage“ und „Fortschritte bei der Einführung von Smart Metering“.

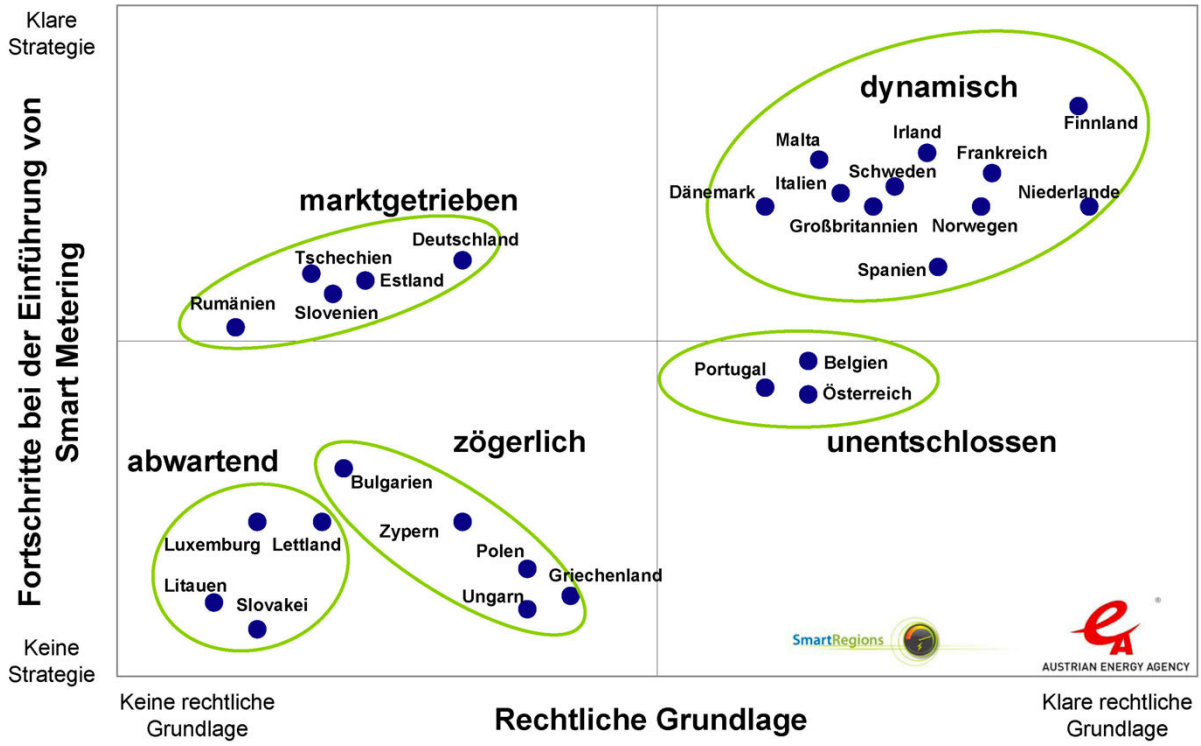



Abbildung 2: Umsetzung des Smart Meter Rollouts in Europa (Antic, 2015, pp. 17-18)

Die unterschiedlichen Einführungsstrategien unterscheiden sich jedoch lediglich hinsichtlich der Geschwindigkeit der Marktdurchdringung. Unabhängig davon, ob eine schleichende, marktorientierte Diffusion als Einführungsstrategie verfolgt wird oder ein flächendeckender Einbau beschlossen wird, werden Smart Meter langfristig gesehen die elektromechanischen Ferraris-Zähler ablösen. (Zeller,

2015, p. 25) Prognosen gehen daher davon aus, dass der Smart Meter Markt weltweit auch in zukünftigen Jahren weiter wachsen wird. (Sun et al., 2015, pp. 468-469)

**Gesellschaftliche Akzeptanz**

Der Erfolg bei der Umsetzung dieser energiepolitischen Regularien sowie die Geschwindigkeit der Marktdurchdringung der Smart Meter Technologie hängt dabei entscheidend von der Akzeptanz auf der Verbraucherseite ab. (Arlt & Wolling, 2011, p. 3) Daher macht es Sinn, die gesellschaftliche Akzeptanz und Einstellung gegenüber der Technologie, eventuelle Assoziationen und die Wünsche und Sorgen der Kunden näher zu untersuchen. In einer Umfrage gaben ungefähr die Hälfte der Teilnehmer an, schon einmal etwas von Smart Metern gehört zu haben. (Riester, 2017, pp. 89-90) Daraus folgt allerdings, dass noch immer die Hälfte der Bevölkerung noch nie etwas von Smart Metern gehört hat. Um einen landesweiten, flächendeckenden Ausbau der Technologie in absehbarer Zukunft zu erreichen, während die gesetzlichen Vorgaben den Einbau für den überwiegenden Teil der Privathaushaltskunden noch immer als optional vorgeschrieben haben und jeder Zweite ebendieser Kunden noch nie etwas von der Technologie gehört hat, muss also noch viel Überzeugungsarbeit geleistet werden. Die Akzeptanz der Installation von einem Smart Meter ist allerdings laut Umfragen relativ hoch. Dies ist jedoch sehr preiselastisch, da die Akzeptanzwerte stark sinken, wenn von jährlichen Zusatzkosten ausgegangen wird, da die Bereitschaft, diese selbst zu tragen, sehr gering ist. (Riester, 2017, pp. 89-90) Die Zahlungsbereitschaft der Stromkunden bezüglich des Einbaus der Smart Meter ist in Umfragen historisch jedoch signifikant gestiegen. In 2009 waren noch weniger als 20 % der Befragten bereit, mehr als 100 Euro für den Einbau auszugeben, während der Wert im Jahr 2010 bereits auf über 50 % anstieg. (Arlt & Wolling, 2011, pp. 26-27) Attraktiv macht den Einbau aus Kundensicht vor allem die mögliche Selbstkontrolle, die Vereinfachung der Abrechnungsprozesse und das Umweltbewusstsein. (Verbraucherzentrale, 2010, pp. 5-6) Ein hoher wahrgenommener Nutzen der Technologie führt anhand der Ergebnisse einer Studie von Riester (2017) zu einer höheren Akzeptanz und Nutzungsabsicht. (Riester, 2017, pp. 89-90) Zudem werden die Akzeptanz und Nutzungsabsicht stark durch die Einfachheit der Nutzung, das Umweltbewusstsein und das wahrgenommene Risiko beeinflusst. Auffällig ist auch, dass die jüngste Altersgruppe die höchste Akzeptanz aufweist und die Akzeptanz und Frauen signifikant höher ist als unter Männern. (Riester, 2017, pp. 89-90) Dieselbe Studie kam aber auch zu dem Ergebnis, dass 39 % aller Assoziationen mit Smart Metern negativ ausfallen. (Riester, 2017, pp. 89-90)

Unter den Endverbrauchern lassen sich bezüglich der Akzeptanz zudem vier Cluster identifizieren. Das erste Cluster bilden die „Interessierten“. Diese stellen das größte Cluster dar und weisen wenige Auffälligkeiten auf. Sie bewerten den Nutzen eher positiv und sind folglich auch eher positiv gegenüber Smart Metern eingestellt. Die Werte bezüglich des wahrgenommenen Risikos und den Faktoren Energiesparverhalten und Umweltbewusstsein liegen im höheren Mittelfeld. (Riester, 2017, pp. 68-69) Die „Umwelt- und Risikobewussten“ bilden das zweite identifizierbare Cluster und weisen ein signifikant höheres Umweltbewusstsein, Energiesparverhalten, aber auch Risikobewusstsein als das Gesamtmittel auf. (Riester, 2017, pp. 68-69) Das dritte Cluster besteht aus den „Befürwortern“. Diese Kundengruppe weist signifikant höhere Werte bei dem wahrgenommenen Nutzen und der Akzeptanz auf als das Gesamtmittel, während der wahrgenommene Nutzen eher gering ist. (Riester, 2017, pp. 68-

69) Das vierte und kleinste Cluster besteht aus den „Ablehnern“. Der wahrgenommene Nutzen und die Akzeptanz der Technologie sind hier besonders gering, während diese Gruppe gleichzeitig das Risiko am höchsten einschätzt. Auffällig ist hier auch, dass dieses Kundensegment zum Großteil aus Männern aus der höchsten befragten Altersgruppe besteht. (Riester, 2017, pp. 68-69)

Allgemein fällt die Beurteilung der zusätzlichen Funktionen und Nutzungsmöglichkeiten von Strom Smart Metern aus Sicht der Endverbraucher eher positiv aus. Die detaillierten Verbrauchsinformationen, die Möglichkeit zur Einführung zeitvariabler Tarife und die Option, bestimmte Geräte durch Smart Meter direkt und automatisiert zu steuern, wird von der Hälfte der Bürger als nützlich, von einem weiteren Drittel zumindest als teilweise nützlich eingeschätzt. (Arlt & Wolling, 2011, pp. 26-27) Kritikpunkte stellen vor allem die anfallenden Kosten, Sicherheitsaspekte, der Wunsch nach besserem Datenschutz, die Häufigkeit der Datenübertragung und die fehlende Transparenz darüber, was mit den Daten passiert, dar. (Riester, 2017, pp. 89-90) Bevorzugtes Feedbacksystem zur Darstellung des Stromverbrauchs ist bei 50,2 % der Kunden das In-Home-Display, gefolgt von der mobilen App mit 31,3 % und dem Online-Portal mit 18,6 %. (Riester, 2017, p. 87) Im Rahmen eines Projekts wurde die zeitnahe Darstellung allerdings nicht immer als notwendig empfunden. (Petsch et al., 2012, p. 10) Können die in 2.1.2 beschriebenen Nutzungsmöglichkeiten von Smart Metern auf Makroebene nicht die erhofften Kosteneinsparungen bei der Stromlieferung erzielen oder machen diese sich aus Kundensicht nicht bemerkbar und werden den Kunden außer einer zeitnahen Darstellung und Informationen über den Stromverbrauch keine Zusatzleistungen geboten, ist in Anbetracht der festgestellten geringen Kundenbereitschaft, entstehende Zusatzkosten selbst zu tragen, also eher fraglich, ob Kunden, die zu der Endverbraucherklasse zählen, für die der Einbau bisher nur optional ist, bezüglich des Einbaus und der Nutzung von Smart Metern selbst die Initiative ergreifen werden.

### 2.1.5. Schnittstellen intelligenter Stromzähler

Aus der zentralen Stellung der Smart Meter als Bindeglied zwischen Endverbrauchern und Unternehmen der Energiewirtschaft wird die Bedeutung dieser Technologie innerhalb der intelligenten, digitalisierten Elektrizitätsversorgungsnetze der Zukunft deutlich. Zum wirklich intelligenten Stromzähler werden Smart Meter aber, wie zuvor bereits erläutert, erst durch das Kommunikationsmodul, das SMGW. Im Gegensatz zu rein elektronischen Zählern, die in Gesetzestexten synonym auch als moderne Messeinrichtungen bezeichnet werden, und elektromechanischen Ferraris-Zählern, sind Smart Meter durch ihr Gateway in der Lage, die ausgelesenen Messwerte und Daten automatisiert zu verarbeiten und an andere Marktteilnehmer zu übermitteln. Eine Kommunikation ist aber bidirektional möglich und kann also auch in die Gegenrichtung verlaufen. Das SMGW bildet den Ausgangspunkt für ebendiese bidirektionale Kommunikation und Übermittlung von Daten und ist gemeinsam mit den Systemen zur Speicherung und Analyse der Daten das Herzstück einer AMI. (McHenry, 2013, p. 834) Die Übertragungstechnologien, die bei der Übermittlung der Daten zum Einsatz kommen können, wurden bereits in 2.1.1 dargestellt. Auf welche Weise und in welcher Reihenfolge die ausgelesenen Elektrizitätsverbrauchsdaten verarbeitet werden wurde jedoch noch nicht erörtert und ist für die Untersuchung von Konzepten und möglichen Lösungen zur Integration und Analyse von

Energieverbrauchsdaten von wesentlicher Bedeutung. Besonders wichtig ist in diesem Zusammenhang eine Erörterung bestehender Schnittstellen eines Smart Meters, über die die Messdaten übermittelt werden können. Im Folgenden soll daher nun ein Blick ins technische Detail geworfen werden, um die genaue Systemarchitektur eines intelligenten Messsystems, die technischen Rollen bei der Kommunikation der Daten und die Schnittstellen des SMGW zu verschiedenen Kommunikationsnetzen zu erläutern.

Die Aufstellung von Standards für ein intelligentes Messsystem und die Normierung von Schnittstellen liegt in Deutschland im Aufgabenbereich des Bundesamtes für Sicherheit in der Informationstechnik (BSI). Dieses hat für Smart Meter ein Schutzprofil aufgestellt, welches die möglichen Bedrohungen eines SMGW in seiner Einsatzumgebung beschreibt. SMGWs müssen die so definierten Mindestanforderungen für entsprechende Sicherheitsmaßnahmen erfüllen und werden auf Basis dieses Schutzprofils geprüft und zertifiziert. Dieses Schutzprofil und die technischen Richtlinien beschreiben in ihren Ausführungen zunächst auch vier technische Rollen, die mit dem SMGW interagieren. Der erste Akteur ist hierbei der Endverbraucher, eine natürliche oder juristische Person, die Elektrizität, Gas, Wasser oder Fernwärme bezieht, beziehungsweise in manchen Fällen auch selbst dezentral produziert. Der Endverbraucher ist der Eigentümer der Messwerte, die er durch eine am SMGW vorgesehene Schnittstelle abrufen kann. (BSI, 2019, p. 13) Die zweite technische Rolle wird von autorisierten externen Marktteilnehmern (EMT) übernommen. Diese können mit dem SMGW eine Kommunikation zum Austausch von Daten aufnehmen. Hierzu zählen zum Beispiel Betreiber von Verteilnetzen, MSBs, Messdienstleister, EVUs und sonstige Dienstleister mit Autorisierung. (BSI, 2019, p. 13) Der SMGW Administrator übernimmt die dritte technische Rolle und ist eine vertrauenswürdige Instanz, die das SMGW konfiguriert, überwacht und steuert. Zu seinem Verantwortungsbereich zählt außerdem die Erstellung und Administration der in das SMGW eingespielten Profile zur Tarifierung, Bilanzierung und Netzzustandsdatenerhebung. Dafür nutzt der SMGW Administrator dieselbe Schnittstelle des SMGW wie die EMTs. (BSI, 2019, p. 13) Der Service-Techniker übernimmt die vierte technische Rolle und kann im Fehlerfall vor Ort auslesend auf das System-Logbuch und weitere Diagnosedaten zugreifen, indem er eine lokale Diagnoseschnittstelle am SMGW nutzt, die die Daten über dasselbe Netz überträgt, das auch die Endverbraucher nutzen. (BSI, 2019, p. 13)

Das SMGW ist also das Zentrum eines intelligenten Messsystems und fungiert als Bindeglied zwischen allen Beteiligten. Zudem dient es aber auch als zentrale Sicherheitskomponente, indem es eine verschlüsselte Datenübertragung zwischen den Marktteilnehmern sicherstellt. Da die Kommunikation bidirektional erfolgt, können nicht nur Informationen übermittelt, sondern auch Befehle von EMTs empfangen und ausgeführt werden. Bei der Übermittlung von Daten zwischen SMGW und EMTs wird grundsätzlich zwischen Push- und Pull-Betrieb unterschieden. Im Push-Betrieb werden die Daten aktiv gesendet, was in IP-basierten Systemen den Regelfall darstellt, während diese im Pull-Betrieb, meist nur im Falle eines Fehlers oder bei spontanen Ereignissen, von der Zentrale angefordert und abgefragt werden. (Riester, 2017, pp. 23-24)

SMGWs nutzen zur Übertragung von Daten drei verschiedene, interoperable Netzwerke. Die Kommunikation zwischen Teilnehmern verschiedener Netzwerke findet dabei ausschließlich über das SMGW statt. Ein direkter bidirektionaler Austausch zwischen den Netzen ist nicht gestattet. (Riester,

2017, pp. 23-24) Abgeleitet von einer solchen Systemarchitektur mit drei verschiedenen Netzwerken muss ein SMGW also mindestens drei physische Schnittstellen bereitstellen. Diese Schnittstellen werden im Folgenden zunächst grafisch dargestellt und dann näher untersucht.

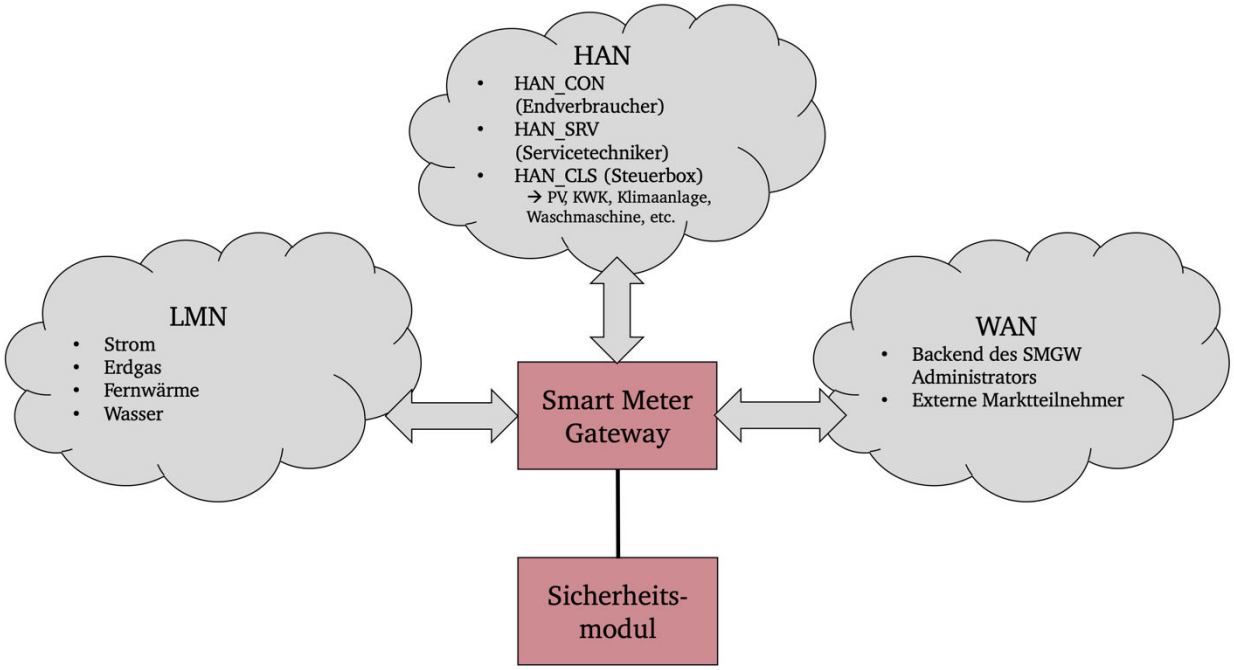



Abbildung 3: Schnittstellen des Smart Meter Gateways (eigene Darstellung nach BSI, 2019, p. 14; Riester, 2017, p. 23)

Das SMGW ist also der Ankerpunkt zwischen LMN, WAN und HAN. Diese Netzwerke haben die folgenden Eigenschaften und Aufgaben:

**Local Metrological Network**

Das Lokale Metrologische Netz, engl. Local Metrological Network (LMN), wird von dem SMGW genutzt, um mit den angebundenen Zählern für Stoff- und Energiemengen eines oder mehrerer Endverbraucher zu kommunizieren. Neben Stromzählern können auch Zähler zur Messung des Gas-, Wasser- oder Fernwärmeverbrauchs zum Einsatz kommen. (BSI, 2019, p. 14) Das SMGW erhält die Messdaten der Zähler ausschließlich über das LMN und ist nach dem Empfang der Daten auch für die Verarbeitung, Speicherung und Versendung der Mess- und gegebenenfalls Netzzustandsdaten verantwortlich. Die Verarbeitung der Messwerte beinhaltet insbesondere die Zeitstempelung und Tarifierung, weshalb das SMGW dem Eichrecht unterliegt. Die angeschlossenen Zähler sind dem SMGW zuvor in Form von entsprechenden Zählerprofilen durch den SMGW Administrator bekannt gemacht worden. Die von den Zählern erhaltenen Daten enthalten die Verbrauchswerte sowie im Fall von PV- oder KWK-Analgen auch die in das Netz eingespeisten Energiemengen der Endverbraucher. Zusätzlich können weitere für den Netzbetrieb relevante Parameter wie beispielsweise Netzspannung, Frequenz oder Phasenwinkel empfangen werden. (BSI, 2019, pp. 15-16) Die Daten der lokalen Zähler werden dabei in regelmäßigen Abständen, beispielsweise alle zwei Sekunden (Antic, 2015, p. 11), über das LMN verschlüsselt und integritätsgesichert vom SMGW empfangen. Nach erfolgreicher

Entschlüsselung und Integritätsprüfung der Messdaten versieht das SMGW diese mit einem Zeitstempel aus der eigenen, systeminternen Uhr des SMGW, ordnet die entsprechende Tarifstufe zu und speichert sie in Messwertlisten. Die Messwerte können dann an die verschiedenen EMTs versandt werden. (BSI, 2019, pp. 15-16)

**Wide Area Network**

Über das Weitverkehrsnetz, engl. Wide Area Network (WAN), kommuniziert das SMGW mit den EMTs und dem SMGW Administrator. (BSI, 2019, p. 14) Das SMGW wird durch Konfigurationsinformationen und Befehle, die es vom SMGW Administrator empfängt und dann verarbeitet, konfiguriert und administriert. Zudem werden im SMGW vom SMGW Administrator Regelwerke in Form von Auswertungsprofilen hinterlegt, die die Weiterverarbeitung der im LMN empfangenen Messwerte vorgeben. Den letzten Schritt der Weiterverarbeitung der Messwerte im SMGW stellt die Übermittlung ebendieser an autorisierte EMTs durch das WAN dar. Bei der Übertragung von nicht abrechnungsrelevanten Messwerten, beispielsweise netzbetriebsrelevanten Parametern, an einen EMT ist es notwendig, die Identität des Endverbrauchers, also die Identität des entsprechenden Stromzählers, nicht offen zu legen. Um dies zu gewährleisten wird die im Datensatz enthaltene Identifikationsnummer des Zählers durch ein Pseudonym ersetzt. Des Weiteren werden die Daten über einen Dritten, den SMGW Administrator, an den Empfänger vermittelt, um die Identität des sendenden SMGW zu verbergen. (BSI, 2019, p. 16)

**Home Area Network**

Im Heimnetzwerk, engl. Home Area Network (HAN), stellt das SMGW drei logische Schnittstellen für die über das Heimnetzwerk verbundenen Instanzen bereit. Eine davon ist ein Controllable-Local-Systems-(CLS-)Interface (IF_GW_CLS), über das die intelligenten Geräte des Endverbrauchers gesteuert werden können. Hierzu zählen z.B. intelligente Haushaltsgeräte wie Waschmaschinen oder Kühlschränke, Klimaanlagen, PV- oder KWK-Anlagen, Nachtspeicherheizungen und Stromunterbrecher. Auf diese Weise können die steuerbaren Komponenten des HAN mit Hilfe des SMGW gesicherte Kommunikationsverbindungen mit EMTs im WAN aufbauen. Das SMGW stellt hierfür Transport-Layer-Security-(TLS-)gesicherte Verbindungen zu CLS und EMT bereit, die es aufeinander abbildet. Konkrete Anwendungsfälle zur Steuerung oder Überwachung der CLS-Komponente, Kommunikationsszenarien und auch die dazu notwendigen Protokolle sind für das SMGW transparent. (BSI, 2019, p. 17) Schaltbefehle von Netzbetreibern als autorisierte EMTs werden dann über das WAN an das jeweilige SMGW übermittelt und dann von da aus über das dahinter liegende lokale Heimnetzwerk an das CLS-Interface weitergeleitet. (Bachor & Freunek, 2020, p. 217) Auf diese Weise kann dann das entsprechende Gerät automatisiert und aus der Ferne gesteuert werden. Die zweite logische Schnittstelle besteht für Service-Techniker (IF_GW_SRV) – die vierte technische Rolle, die dadurch Konfigurationsprofile oder das System-Log einsehen und so bessere Fehlerdiagnosen abgeben können. (BSI, 2019, p. 17) Da auch die Endverbraucher ein Interesse an ihren Stromverbrauchsdaten haben, muss das SMGW auch dafür eine logische Schnittstelle im HAN bereithalten. Endverbraucher können ihre Verbrauchsdaten über die Endverbraucher-Schnittstelle (IF_GW_CON) abrufen. Ein Zugriff auf diese Daten kann immer nur lesend erfolgen und erfordert zuvor eine erfolgreiche Authentifizierung.

Dabei kann ein dediziertes, kryptographisch gesichertes Display oder jedes andere CLS-Gerät im HAN, das kryptographisch gesicherte Datenströme verarbeiten kann, genutzt werden, um die Daten auszulesen und zu visualisieren. (BSI, 2019, p. 17) Die Installation und Nutzung von In-Home-Displays in den Wohnräumen der Endverbraucher ist nützlich, da die eigentlichen Stromzähler häufig außerhalb der tatsächlichen Wohnräume installiert sind. Alternativ zu der Auslesung der eigenen Verbrauchsdaten über das HAN können Endverbraucher die Daten aber auch über das Internet abrufen. Die von den MSBs zu Abrechnungszwecken erfassten Verbrauchsdaten werden dem Endverbraucher in diesem Fall von dem MSB oder EVU über das Internet zur Verfügung gestellt und können über Online-Portale oder mobile Apps abgerufen werden. (Rigoll, 2017, p. 76; Zeller, 2015, pp. 17-18)

Sowohl der Weg, den die Verbrauchsdaten von der Messung bis zur Darstellung und Visualisierung für den Endverbraucher zurücklegen, als auch die Kommunikationsverbindungen zwischen den verschiedenen Instanzen, sind im Folgenden graphisch dargestellt. Die Übertragung im LMN erfolgt über Funk (wM-bus) oder Datenkabel (M-bus). (Zeller, 2015, pp. 17-18) Die Daten werden dann im WAN durch Datenkabel (DSL oder Glasfaserkabel), Mobilfunk, Stromkabel oder Telefonkabel mit EMTs kommuniziert sowie im HAN durch Datenkabel (Ethernet Kabel oder M-bus), Funk (WLAN, wM-bus) oder Stromkabel an die logischen Schnittstellen übermittelt.

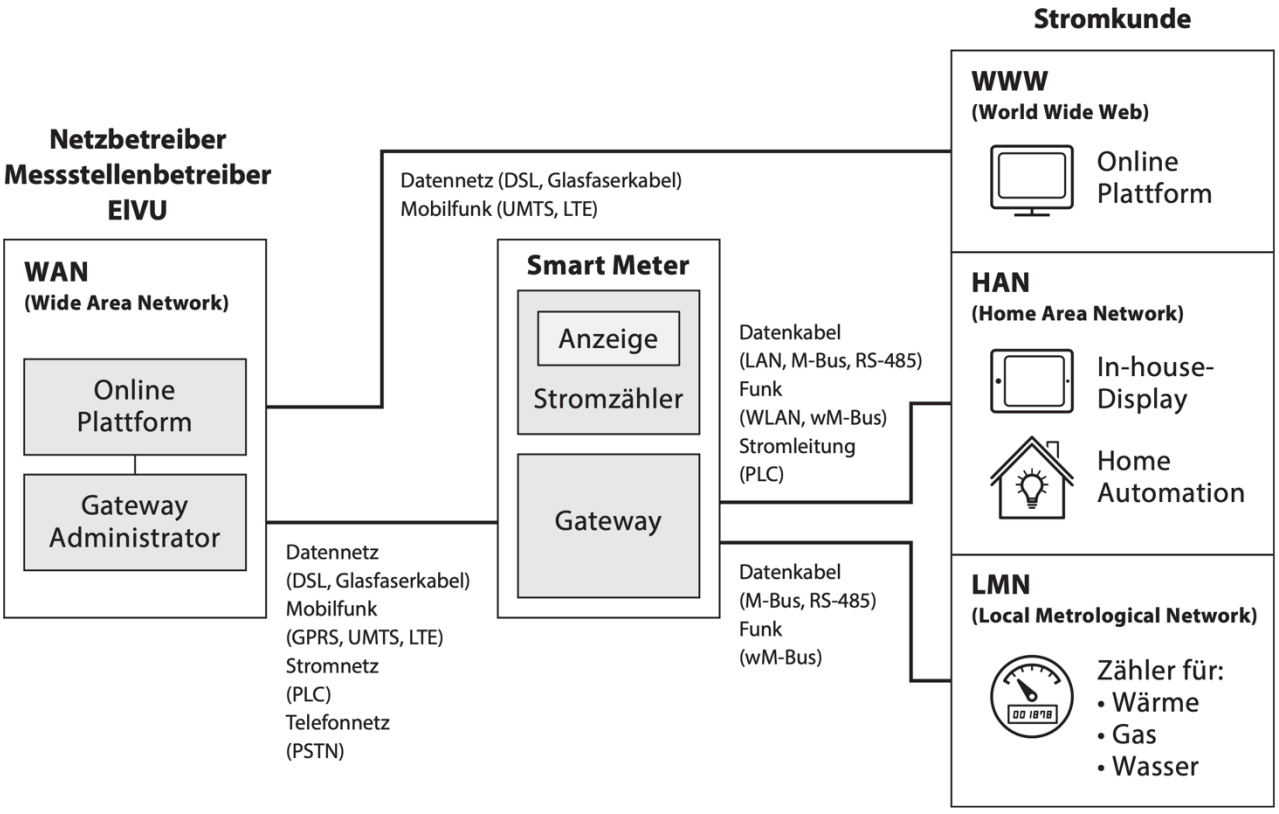



Abbildung 4: Netzwerkanbindungen des Smart Meter Gateways (Zeller, 2015, p. 17)

Zusammenfassend besteht die Hauptaufgabe des SMGW also in dem Empfangen der Messwerte der Zähler über das LMN, der Speicherung und Verarbeitung dieser gemäß der konfigurierten Regelwerke und der anschließenden Versendung. Empfänger können EMTs im WAN oder Instanzen im HAN sein.

Die Messwerte können entweder direkt sternförmig oder indirekt über einen bestimmten Akteur des Marktes an die jeweiligen Adressaten im WAN übermittelt werden. Durch die HAN-Schnittstellen können Service-Techniker und Endverbraucher lesend auf die Daten zugreifen. Für die im HAN angeschlossenen CLS-Komponenten fungiert das SMGW als transparenter Proxy-Server. Die Kommunikation mit CLS und EMTs ist TLS-geschützt und wird im SMGW terminiert. Zudem übernimmt das SMGW die Aufgaben einer Firewall und separiert die verschiedenen Netze voneinander. (BSI, 2019, pp. 14-15)

### 2.1.6. Inhalt und Formatierung der generierten Daten

Um im Rahmen dieser Arbeit die Eignung von OLAP-Cubes zur Integration und Analyse von Energieverbrauchsdaten zu untersuchen, ist neben der im vorherigen Kapitel durchgeführten Untersuchung der Schnittstellen zu den verschiedenen Übertragungsnetzen eines SMGW und den Technologien, die bei der Übertragung zum Einsatz kommen können, auch notwendig zu erörtern, aus was genau diese Energieverbrauchsdaten bestehen. Schließlich muss für den Entwurf von Software-Anwendungen und Konzepten zum Management und der Analyse dieser Daten klar sein, welche Informationen die übermittelten Datensätze genau enthalten. Folglich widmet sich dieses Unterkapitel der Analyse des Inhalts und Formats der Smart Meter Daten.

#### Inhalt der generierten Daten

In Anbetracht der Menge an Faktoren, die den Inhalt der Datensätze beeinflussen und bestimmen können, wird deutlich, dass eine Homogenität der Inhalte der übermittelten Daten durchaus nicht als selbstverständlich hingenommen werden sollte. Finden zur Messung der Stromverbräuche Ferraris-Zähler Anwendung, so sind diese nur in der Lage, die eigentliche Verbrauchsmessung durchzuführen und anhand des Zahlwerks abzuspeichern. Kommt hierfür allerdings ein elektronischer Zähler in Kombination mit einer Kommunikationseinheit (einem SMGW) zum Einsatz, so legt dies die Basis zur Erhebung einer ganzen Fülle nützlicher Informationen. (Schweinfurth, 2020, p. 232) Auf diese Weise wird nicht nur die Tiefe, sondern auch die Breite der Informationen erhöht. Während bei der Nutzung von Ferraris-Zählern eine Analyse der Stromverbrauchsdaten der Endverbraucher praktisch nur auf Jahresbasis oder dem entsprechenden Zeitintervall der manuellen Ablesung möglich ist, ist eine Analyse bei Verwendung von Smart Metern theoretisch in viel höherer Granularität und hinsichtlich spezifischer, auch erst im Nachhinein festgelegter Zeitpunkte oder -räume möglich. Elektronische Zähler können neben der Wirkarbeit als Wirkleistung über die Zeit, also den Stromverbrauch innerhalb einer gewissen Periode, ferner auch Netzzustandsdaten erfassen. Wie bereits in 2.1.5 erläutert, liefern Netzzustandsdaten wichtige Parameter für den Netzbetrieb und die Überwachung der Qualität der übertragenen Elektrizität, beispielsweise Netzspannung, Frequenz oder Phasenwinkel. Da Netzzustandsdaten aber, wie der Name bereits impliziert, vor allem in Echtzeit von Wert und Interesse sind, wird deutlich, dass dieser Mehrwert nur in Verbindung des elektronischen Zählers mit einem SMGW und die folglich geschaffene Möglichkeit der zeitnahen Kommunikation der Daten realisiert werden kann. Im Zuge dieser Arbeit liegt der Fokus allerdings in erster Linie auf den Energieverbrauchsdaten.

Hinsichtlich des genauen Inhalts der Energieverbrauchsdaten bedeutet diese neu geschaffene Fülle an Informationen aber vor allem eine Heterogenität der Datensätze. Welche Inhalte genau übermittelt werden hängt stark von der Verwendung der Daten und dem Zweck der Erhebung ab. Die Inhalte können aber in vielen Fällen in Abhängigkeit von dem Ursprungszweck der Erhebung logisch erschlossen werden. Durch den Verwendungszweck entstehen aber nicht nur verschiedene Ansprüche an die Verfügbarkeit, das Format und die Vollständigkeit der Daten, sondern auch viele Implikationen für das System, das diese Daten verarbeitet. Für eine entsprechende Verwendung der Daten muss bereits bei der Erhebung eine entsprechende Aussagekraft der Datenanzahl und -qualität gewährleistet sein. Zur Sammlung der dezentral verteilten Daten an einem zentralen Ort findet dann eine Übertragung statt. Sowohl Erhebung als auch Übertragung unterliegen in der Realität verschiedensten Rahmenbedingungen, wie Datenschutz und -sicherheit, Übertragungsgeschwindigkeit und der Verfügbarkeit von Kommunikationstechnologien. (Herre & Freunek, 2020, pp. 198-199) Auch die anschließende Bereitstellung der Daten wird maßgeblich von dem Verwendungszweck beeinflusst. Von diesem hängen beispielsweise die Anforderungen an die Datensicherheit und den Datenschutz ab. Je nach Verwendung entscheidet sich, ob die Daten personenbezogen sein müssen oder dies gar nicht sein dürfen oder ob eine Speicherung in einer externen Cloud vertretbar ist. Des Weiteren beeinflusst der Verwendungszweck die benötigte Auflösung und Granularität, die benötigte Anzahl der Daten, die Skalierbarkeit des Speichers, die benötigten Schnittstellen und die Frage, ob der Zugriff auf die Daten in Echtzeit oder verzögert einmal monatlich, beziehungsweise jährlich geschieht. (Herre & Freunek, 2020, pp. 198-199) Die wesentlichen Verwendungszwecke der von Strom Smart Metern generierten Daten und der sich daraus ergebende Inhalt sind im Folgenden beschrieben.

Die erste und bekannteste Nutzung der Smart Meter Daten besteht in der Verwendung zu Abrechnungszwecken. Da die Auslesung der Stromverbräuche durch ZFA des zuständigen MSB erfolgt und das EVU die Abrechnung erstellt, werden hieraus bereits die Mindestanforderungen an den Inhalt der übermittelten Daten deutlich. Neben der Übermittlung des Stromverbrauchs über das festgelegte Zeitintervall in einer geeigneten physikalischen Einheit müssen die Verbrauchsdaten zudem ihren physikalischen Dimensionen, Raum und Zeit, zugeordnet werden. Es muss also klar sein, wann, bzw. für welchen Zeitraum die Messung erfolgte und wo, bzw. bei wem dies geschah. Die Abrechnung muss schließlich unbedingt einem bestimmten Kunden zugeordnet sein. Dabei werden Rechnungen für die bezogene Strommenge immer in größeren Abständen, z.B. monatlich oder vierteljährlich, ausgestellt. Daher besteht in diesem Anwendungsfall kein Bedarf nach zeitlich hochaufgelösten Daten. Die Möglichkeit, jeden Monat einmal Daten zu erheben und zu bilanzieren genügt. (Herre & Freunek, 2020, pp. 198-200) Für Abrechnungszwecke müssen die Datensätze also neben dem Zählerstand mindestens einen Zeitstempel und eine Meter ID oder Kunden ID zur Zuordnung enthalten. Werden von dem Kunden zeitvariable oder dynamische Stromtarife genutzt, muss zudem nach der Tarifierung durch das SMGW auch der zugehörige Tarif übermittelt werden. Darüber hinaus sind für die Abrechnung weitere Metadaten erforderlich. Diese müssen die eigentlichen Informationen über den Kunden enthalten und haben insbesondere im Fall eines Mehrparteienhauses mit zentraler Energiedatenübermittlung eine wichtige Bedeutung. (Rigoll, 2017, p. 77) Eine gemeinsame Übermittlung sämtlicher Metadaten und Kundeninformationen bei jeder Verbrauchserhebung wäre aber aus mehreren Gründen nachteilig und wird daher in der Regel nicht durchgeführt. Die genaue Integration und Verknüpfung der Metadaten werden in 2.1.7 diskutiert.

Besteht die Anwendung jedoch in einer Visualisierung der aktuellen Verbräuche, ist, um eine zeitnahe Darstellung zu gewährleisten, eine zeitlich hohe Auflösung im Minuten- oder Sekundenbereich erforderlich. Verwenden Endverbraucher zur Kontrolle des aktuellen Verbrauchs ein Online-Portal oder eine App, werden ihnen diese Daten von dem MSB über das Internet zur Verfügung gestellt. Daher ist in diesem Fall die aktuelle Wirkleistung oder die gemessene Wirkarbeit, in der Größenordnung von Minuten oder Sekunden aufgelöst, relevant. Auch hier muss eine eindeutige Zuordnung durch die Übermittlung der Meter oder Kunden ID möglich sein. (Herre & Freunek, 2020, p. 198) Da Endverbraucher aber neben dem reinen Verbrauch auch an den entstehenden Kosten interessiert sind, muss auch hier im Fall von zeitvariablen oder dynamischen Tarifen der aktuelle Tarif übermittelt werden.

Werden die Smart Meter Daten hingegen zur Durchführung der Netzplanung genutzt, ist hierfür eine statistisch repräsentative Probe aller Kunden in anonymisierter Form ausreichend. Der Zugriff auf die Daten erfolgt in diesem Fall eher auf Monats- oder Jahresbasis. Jedoch gilt, dass ein Netzbetreiber die Auslastung seines Netzes zielgerechter und genauer planen kann, je feingranularer die Auslastungswerte vorliegen. Um eine gute Netzplanung sicherzustellen, werden die Messdaten daher in Minutenintervallen erhoben. Zusätzlich zu den Verbräuchen sind hier auch verschiedene Parameter über den Netzzustand von Interesse. (Herre & Freunek, 2020, p. 201) Eine Personalisierung der Daten ist aus Datenschutzgründen jedoch unerwünscht. Daher schreibt das BSI in den Richtlinien, wie in 2.1.5 erläutert, für diesen Fall auch eine Pseudonymisierung der Daten vor.

Folglich können, je nach Verwendung der Daten, neben dem reinen Stromverbrauch auch zusätzlich Netzzustands- und Energiequalitätsdaten Inhalt der übermittelten Daten sein. Jede Messung muss dabei zur zeitlichen Zuordnung mit einem Zeitstempel versehen sein. Je nach Anwendung muss zusätzlich eine Personalisierung stattfinden, was in einer zusätzlichen Übermittlung einer Meter ID oder Kunden ID resultiert. Für die Übermittlung der Verbrauchs- und Netzzustandsdaten müssen physikalische Größen verwendet werden. Die wesentlichen physikalischen Größen und die zugehörigen Einheiten sind im Folgenden tabellarisch dargestellt.

| Physikalische Größe | Einheit |
|---|---|
| Elektrische Energie | Kilowattstunden (kWh) |
| Scheinleistung | Voltampere (VA) |
| Wirkleistung | Watt (W) |
| Blindleistung | Var (var) |
| Elektrische Spannung | Volt (V) |
| Elektrische Stromstärke | Ampere (A) |
| Netzfrequenz | Hertz (Hz) |

Tabelle 2: Physikalische Größen und zugehörige Einheiten nach (Rigoll, 2017, p. 82)

Die beschriebene Heterogenität der Datenverwendungszwecke und Datenquellen sowie die Vielzahl der Anbieter von Smart Metern führen zudem dazu, dass auch die Formate der Zeitstempel, mit denen die physikalischen Messgrößen versehen werden sehr heterogen sein können. Die folgend genannten

Zeitstempel sind Beispiele für Formate, die in der Realität zu Anwendung kommen. (Rigoll, 2017, p. 84)

- 2020-04-07T14:25:11.000Z
- 2020-04-07T14:25:11Z
- 2020-04-07T14:25:11+00:00
- 2020-04-07T16:25:11+02:00
- 07.04.2020 16:25:11
- 1586269511

Die ersten vier Zeitstempel entsprechen alle der ISO-Norm 8601 und erlauben ein aufsteigendes Sortieren. Nach dem Jahr, Monat, Tag, Stunde, Minute und Sekunde enthalten sie zudem die verwendete Zeitzone. Der fünfte Zeitstempel findet im Alltag durchaus häufig Verwendung, ist jedoch nicht eindeutig, da beispielsweise nicht klar definiert ist, ob nicht doch das amerikanische Format Monat – Tag – Jahr gemeint ist. Zudem ist unklar, ob die Zeit in Lokal- oder Weltzeit dargestellt ist. Der letzte Zeitstempel liegt in Unixzeit vor, die die verstrichene Zeit in Sekunden seit dem 1. Januar 1970 um 00:00 Uhr UTC angibt. Damit ist dieses Format zwar eindeutig, aber nur maschinell interpretierbar. (Rigoll, 2017, pp. 84-85) Dabei geben alle sechs Zeitstempel die gleiche Zeit in unterschiedlichem Format an.

In der folgenden Abbildung ist ein beispielhafter Schnappschuss realer Smart Meter Daten dargestellt. Die Spalten beinhalten dabei von links nach rechts die Kundennummer, die Zählernummer, den Zeitstempel der Messung, den Verbrauch in den vergangenen 15 Minuten, den zugehörigen Tarif und die Vertragsnummer der Messstelle.

| | V1 | V2 | V3 | V4 | V5 | V6 |
|---|---|---|---|---|---|---|
| 1 | 43179 | 100000001 | 2015-04-01 00:00:00.000 | 2.65900 | Niedertarif | 36146 |
| 2 | 3422 | 100000034 | 2015-04-01 00:00:00.000 | 2.94300 | Niedertarif | 5930 |
| 3 | 3422 | 100000034 | 2015-04-01 00:15:00.000 | 2.66300 | Niedertarif | 5930 |
| 4 | 3422 | 100000034 | 2015-04-01 00:30:00.000 | 0.12500 | Niedertarif | 5930 |
| 5 | 22955 | 100000054 | 2015-04-01 00:00:00.000 | 0.23300 | Niedertarif | 1810 |
| 6 | 3422 | 100000034 | 2015-04-01 00:45:00.000 | 0.11900 | Niedertarif | 5930 |
| 7 | 22955 | 100000054 | 2015-04-01 00:15:00.000 | 0.23500 | Niedertarif | 1810 |
| 8 | 22955 | 100000054 | 2015-04-01 00:30:00.000 | 0.23500 | Niedertarif | 1810 |
| 9 | 3422 | 100000034 | 2015-04-01 01:00:00.000 | 0.15700 | Niedertarif | 5930 |

Abbildung 5: Beispielhafter Schnappschuss realer Smart Meter Daten (Sodenkamp, Hopf, Kozlovskiy, & Staake, 2016, p. 12)

**Länge der Messintervalle**

Wie zuvor geschildert wurde, hängt der Inhalt entscheidend von dem Verwendungszweck ab. Die Informationen der übermittelten Verbrauchsdaten werden neben dem tatsächlichen Inhalt aber auch durch die Länge der Intervalle, in denen diese übermittelt werden, und dem sich daraus ergebenden Grad der Auflösung und Granularität bestimmt. Die Länge der Messintervalle hängt neben dem

Verwendungszweck zusätzlich von gesetzlichen Vorschriften ab. Alle Großkunden, also Stromkunden mit einem Jahresverbrauch über 100.000 kWh, sind zu einer RLM verpflichtet. Dabei werden die Lastgänge obligatorisch im Viertelstundentakt gemessen und übermittelt. (Petsch et al., 2012, p. 6) Je nach Hersteller, Anbieter und Land haben sich im Vergleich verschiedene Messintervalle etabliert. Der US-Amerikanische Anbieter PG&E bietet Smart Meter für Elektrizität und Gas an. Der Elektrizitätsverbrauch wird dabei in Wohngebäuden im Stundentakt, in Nichtwohngebäuden im Viertelstundentakt gemessen. (Sirojan et al., 2019, pp. 1-2) In Australien senden Smart Meter die Verbrauchsdaten in 30-Minuten Intervallen an das System des MSB. (Sirojan et al., 2019, p. 2) Ein für Entwicklungsländer in Afrika entwickelter Smart Meter misst und übermittelt Energiequalitätsinformationen und den eigenen Status in 20-minütigen Intervallen. (Keelson, Boateng, & Ghansah, 2014, p. 43) Piti et al. schlagen in einer Studie 15-Minuten Intervalle zur Steigerung des Energiebewusstseins der Endverbraucher vor und konnten einen wesentlichen Beitrag der Smart Meter zu energieeffizienterem Verhalten nachweisen. (Pitì, Verticale, Rottondi, Capone, & Lo Schiavo, 2017, p. 6) Rodriguez-Diaz et al. haben einen Smart Meter entworfen, der alle 5 Minuten Energieverbrauchsdaten bereitstellt. (Rodriguez-Diaz et al., 2015, p. 2)

**Dateiformat zur Übertragung der generierten Daten**

In der Praxis kommen nicht zuletzt durch die vielen verschiedenen Anbieter von Smart Metern im Zuge der Liberalisierung des Messwesens viele unterschiedliche Dateiformate bei der Übertragung und dem Empfang der Messdaten zum Einsatz. In vielen Fällen werden die Messdaten als Textdatei versandt. (Herre & Freunek, 2020, pp. 199-200) Als Standard etablieren konnten diese sich bisher jedoch nicht. Folglich sind unter anderem auch proprietäre Binärformate, jegliche weiteren kommaseparierten Dateiformate, XML- oder JSON-Dateien üblich. Da jedes dieser Dateiformate eigene Vor- und Nachteile mit sich bringt, ist fraglich, ob sich zukünftig eines davon durchsetzen kann. (Rigoll, 2017, p. 85) Selbst einfache kommaseparierte Dateien können hier schon zu Inkompatibilität und Problemen bei der Integration führen. Zeichensatz, Trennzeichen, Datumsformat, Dezimaltrenner und Einheit der physikalischen Messgröße können theoretisch von System zu System und von Anbieter zu Anbieter anders sein. Zudem kann es sein, dass bei sehr hoher Messfrequenz zur Reduktion des Speicherbedarfs Werte wie Zeitstempel ausgespart und nur in größeren Abständen übermittelt werden, wodurch die Werte dazwischen später interpoliert werden müssen. (Rigoll, 2017, p. 85)

### 2.1.7. Bestehende Architekturen und Systeme zur Integration der Energiedaten

In 2.1.6 wurde unter Anderem erläutert, dass die von Smart Metern generierten und übermittelten Daten rein logisch mindestens drei verschiedene Werte übermitteln müssen. Neben der offensichtlichen Übermittlung des Verbrauchs ist weiterhin auch die Übermittlung der physikalischen Dimensionen des gemessenen Verbrauchs notwendig, um den Verbrauch richtig einzuordnen. Diese physikalischen Dimensionen sind Raum und Zeit. Smart Meter übermitteln diese in der Regel anhand eines Zeitstempels aus der systeminternen Uhr für die Zeit und einer Meter ID oder Kunden ID für die Lokalisierung. Die übermittelten Messdaten müssen also Aufschluss darüber geben, wann, wo, bzw. bei wem, wie viel Elektrizität bezogen wurde. Anhand der in 2.1.5 erklärten Systemarchitektur mit den normierten Schnittstellen wurde zudem bereits deutlich, dass die Verbrauchsmessung zunächst von

dem Zähler per LMN an das SMGW übergeben und dort weiterverarbeitet wird, um dann als vollständig zuordnungsbare Verbrauchsmessung per WAN an EMTs, wie beispielsweise Verteilnetzbetreiber und EVUs, weitergeleitet zu werden. Um die Systemarchitektur eines intelligenten Messsystems jedoch vollständig zu verstehen und um die Wissensgrundlage zum späteren Aufbau eines Konzeptes zur Integration und Analyse der Daten durch OLAP-Cubes zu schaffen, bleibt abschließend noch die Frage zu klären, wie genau die übermittelten Daten dann bei der zuständigen Instanz integriert und abgespeichert und welche Systeme aktuell zum Management dieser Daten genutzt werden. Darauf wird in diesem Unterpunkt eingegangen. Vor der Untersuchung der bestehenden Energiedatenmanagementsysteme (EDMS) und Integrationsarchitekturen bleibt allerdings noch zu klären, wie und durch welche Prozesse die Daten vor der Integration erhalten werden. Insbesondere im Fehlerfall, bei fehlerhafter Übermittlung oder gar keiner Übermittlung der Messdaten, müssen Strukturen existieren, die sich diesem Problem annehmen, bevor die Daten in das EDMS integriert werden.

**Ablauf der Zählerfernauslesung**

Für die Implementierung der zusätzlichen Messdaten in die bestehenden Informations- und Kommunikationssysteme hat sich der Begriff des Meter-Daten-Management (MDM) herausgebildet. Durch die entstehende Datenmenge können allerdings große Herausforderungen entstehen. In Anbetracht bestehender Abrechnungssysteme auf Jahresbasis ergeben durch stündlich gemessene Verbrauchsdaten eines durchschnittlichen Stadtwerkes mit 100.000 Kunden bereits 876 Mio. Datensätze, im Fall von viertelstündiger Messung sogar über 3,5 Mrd. Datensätze. Folglich ist eine Anpassung der innerbetrieblichen Systeme und der IT-Infrastruktur für viele Unternehmen unausweichlich, was auch zu den in 2.1.3 angesprochenen Investitionskosten führt. (Zeller, 2015, p. 77) Die ZFA liegt in der Regel im Aufgabenbereich des Verteilnetzbetreibers. Dieser erfasst die Smart Meter Daten automatisiert durch das eigene ZFA-System. Kommt es bei diesem Prozess zum Fehlerfall, muss entsprechend reagiert werden. Ist die automatisierte ZFA nicht erfolgreich, kann sie zunächst manuell wiederholt werden. Kommt es hierbei erneut zum Fehlerfall, kann ein Kundenbesuch erforderlich sein. Nach erfolgreichem Erhalt der Messdaten werden diese in das Energiedatenmanagement (EDM) des Netzbetreibers übernommen und auf Vollständigkeit geprüft. Fehlen dennoch Werte, ist der Netzbetreiber verpflichtet Ersatzwerte zu bilden. Im Anschluss werden die Daten an das EVU übermittelt, wo sie erneut auf Vollständigkeit geprüft werden. Stellt sich hierbei eine unvollständige Übermittlung oder Versäumnis auf Seiten des Netzbetreibers heraus, werden die nötigen Ersatzwerte vom EVU gebildet. (Petsch et al., 2012, pp. 6-7) Auf diese Weise soll sichergestellt werden, dass die in das EDM übernommenen Messdaten vollständig und aussagekräftig sind. Für die angewandten EDMS sowie auch für den Aufbau eines Konzeptes zur Verwendung von OLAP-Cubes bedeutet dies, dass, falls Verteilnetzbetreiber und EVU ihren Pflichten ordentlich und sachgemäß nachkommen und bei fehlerhafter ZFA reagieren, fehlende Messwerte und die Frage nach dem optimalen Umgang mit diesen nicht zu besonders großen Hindernissen und Herausforderungen führen dürften.

**Die klassische Integrationsarchitektur**

Die Systeme, die dann zur Anwendung kommen, um mit den Energiedaten umzugehen, stellen eine wesentliche Komponente zur Realisierung der Potentiale eines intelligenten Elektrizitätsnetzes dar. (Mikkelsen, Jacobsen, & Terkelsen, 2016, p. 4) Zur Beschreibung der Systeme, die bei dem Umgang mit den erfassten Energiedaten unterstützen, existieren mehrere, teils synonym verwendete Begriffe. Die Zuordnung und Definition der Begriffe wird dadurch erschwert, dass aus Sicht der Energieunternehmen teils ein anderes Verständnis der Begriffe herrscht, als von Seiten der Wissenschaft. (Rigoll, 2017, p. 39) Zusätzlich entstehen Unklarheiten bei dem Vergleich von deutsch- und englischsprachiger Literatur durch Inkonsistenz bei der Begriffsverwendung und die Frage nach der korrekten Übersetzung. Hier sind daher zunächst eine klare Definition und Erläuterung notwendig. Das EDM umfasst die Bereitstellung, Übermittlung, Verarbeitung und Speicherung von energiebezogenen Datensätzen. (Rigoll, 2017, p. 39) Ein EDMS ist dann ein Softwareprodukt oder Workflowsystem, das zum Zweck des EDM eingesetzt werden kann und Bearbeitung und Verarbeitung der Daten ermöglicht. (Rigoll, 2017, p. 39; Schweinfurth, 2020, p. 230) Dieses besteht aus in der Regel aus den Prozesskomponenten Zähl- und Messeinrichtungen, Datenerfassung, Übertragung, Datenspeicherung und Analyse. (Schweinfurth, 2020, p. 232) Im Idealfall sollte ein solches EDMS flexibel und universell einsetzbar sein sowie jeder Marktrolle als Cloud-Lösung zur Verfügung stehen. (Schweinfurth, 2020, p. 230) Darüber hinaus existieren noch Meterdatenmanagementsysteme (MDMS). In Interaktion mit einem EMDS verwendet, übernehmen diese die Aufbereitung und Speicherung der fernausgelesenen Verbrauchsdaten, während das EMDS dann für Prognosen und Analysen, die auf der vom MDMS geschaffenen Datengrundlage basieren, zuständig ist. MDMS kommen also in Kombination mit EDMS zum Einsatz und sollen Letztere entlasten, indem sie die eigentliche Verwaltung der übermittelten Smart Meter Messdaten übernehmen. (Rigoll, 2017, p. 39) Ein MDMS kann in einer erweiterten Form aber auch zusätzlich der Vorhaltung und Pflege gewisser Kundendaten dienen, sowie bei der Marktkommunikation und Interaktion mit anderen Marktteilnehmern unterstützen. (Rigoll, 2017, p. 39) Des Weiteren kommen Head-End-Systeme (HES) zur Anwendung. Diese stellen die erste Stufe der klassischen Integrationsarchitekturen für Smart Meter Daten dar. Im Folgenden wird näher auf das genaue Zusammenspiel der drei Systeme eingegangen.

Zu Beginn überträgt das SMGW die Messdaten automatisiert über das WAN und ein entsprechendes Protokoll an die verarbeitenden Systeme des Verteilnetzbetreibers. In manchen Fällen erfolgt die Übertragung auch über zwischengeschaltete Datenkonzentratoren. (Bachor & Freunek, 2020, pp. 216-217) Das erste verarbeitende System ist das HES, dass die Messdaten durch direkte Kommunikation mit dem SMGW oder über einen Datenkonzentrator empfängt. Das HES kontrolliert die Kommunikation mit sämtlichen Smart Metern und persistiert die eingehenden Daten. Die Speicherung erfolgt nach dem heutigen Stand der Technik in einer relationalen Datenbank. (Bachor & Freunek, 2020, pp. 217, 223; Herre & Freunek, 2020, pp. 205-206) Somit stellt das HES ein System zur Erfassung und Sammlung der Rohdaten aus allen angeschlossenen Zählern dar. Anschließend wird die Übertragung der Daten in das finale Speichermedium vorbereitet. Ferner dienen HES-Anwendungen auch beispielsweise der Administration der unterschiedlichen Firmware der Smart Meter. (Bachor & Freunek, 2020, p. 217; Schweinfurth, 2020, p. 233) Das HES ist dabei im Rechenzentrum des Verteilnetzbetreibers lokalisiert, dem damit die Aufsicht und Verantwortung unterliegt. (Herre &

Freunek, 2020, pp. 205-206) Weitere Geräte sind von der Nutzung des HES ausgeschlossen, da es ausschließlich in Kombination mit Smart Metern verwendet werden kann. (Bachor & Freunek, 2020, p. 223)

Nach der Erfassung und Vorbereitung der Smart Meter Messdaten in dem HES werden die Daten in weitere, nachfolgende Backend-Systeme übertragen und dort weiterverarbeitet. (Bachor & Freunek, 2020, p. 217; Herre & Freunek, 2020, pp. 205-206) Das nächste System, an das die Daten vom HES aus übermittelt werden ist das MDMS. Das HES bietet hierfür eine Message-Queuing-Schnittstelle, die dem MDMS exklusiv zur Verfügung steht. Sie kann nicht von weiteren Anwendungen genutzt werden. (Bachor & Freunek, 2020, p. 223) Die Weiterverarbeitung der Daten erfolgt kaskadierend, d.h. hintereinandergeschaltet und verkettet mit unidirektionaler Verbindung. Dies führt dazu, dass die Messdaten im Gesamtsystem mehrfach und redundant gespeichert werden. (Bachor & Freunek, 2020, p. 217; Herre & Freunek, 2020, pp. 205-206) Auch das MDMS ist ein dem Rechenzentrum des Verteilnetzbetreibers zugeordnetes System. (Bachor & Freunek, 2020, p. 221)

Insgesamt gilt für das System zur Datenintegration, -speicherung und -verarbeitung, dass es auf den massenhaften Import und die Verarbeitung von Informationen aus allen angeschlossenen Smart Metern ausgelegt sein muss. Auch eine Qualitätssicherung und Administration der Messdaten sowie eine zeitnahe und verlässliche Aufbereitung für nachfolgende Anwendungen wie Abrechnungsprozesse sind wichtige Funktionen. Die Messdaten sind nur dann von wirklichem Wert, wenn sie verständliche und relevante Informationen enthalten. Die letzte Komponente des EDM ist daher die Analyse, um sinnvolle Informationen aus den Messdaten zu extrahieren. (Schweinfurth, 2020, p. 233) Die Analyse lässt sich dann über eine anwendungs- oder webbasierte Nutzeroberfläche durchführen, über die die gespeicherten Daten nach Kriterien abgefragt und die Ergebnisse in unterschiedlichen tabellarischen und graphischen Darstellungsformen wiedergegeben werden. Diese Ergebnisse können dann sowohl an eigene Folgeanwendungen als auch an autorisierte EMTs weitergegeben werden. (Schweinfurth, 2020, p. 233) Diese klassische kaskadierende Weiterverarbeitung der Messdaten von HES zu MDMS zu EDMS und eventuell weiteren nachfolgenden Backend-Systemen führt zu einem hohen Transport der Messdaten innerhalb des Unternehmens und zu mehrfacher, redundanter Speicherung. (Herre & Freunek, 2020, pp. 205-206) Daher existieren auch immer mehr Forschungsansätze und Lösungsvorschläge, um die Smart Meter Daten anders zu integrieren. Einige davon werden gegen Ende dieses Kapitels beispielhaft genannt.

**Datenbankanforderungen für eine effiziente Integration und Speicherung der Messdaten**

In der Realität werden Energieverbrauchsdaten abhängig vom Anwendungsfall und geplanten Verwendungszweck sowie von der Art und Größe der Datensätze teilweise in sehr unterschiedlichen Systemen gespeichert. Die Spanne reicht hierbei von einfachen textbasierten Ansätzen bis hin zu verteilten Data-Warehouse-Systemen wie der zuvor dargestellten klassischen Integrationsarchitektur. (Rigoll, 2017, p. 43) Kleinere Datensätze lassen sich auch in Textdateien oder einfachen relationalen Datenbanken speichern. Bei größerem Umfang und bei Bedarf von aufwendigeren Analysen stößt diese Methode jedoch an ihre Grenzen. In diesem Fall sind dann spezialisierte Ansätze notwendig. Über verschiedene Methoden zum Management von Energiedaten existiert vergleichsweise wenig Literatur.

Ein Grund dafür könnte sein, dass Energiedaten häufig gar nicht in dedizierten, für sie bestimmten Datenmanagementsystemen erfasst und verwaltet werden, sondern eher als notwendiges Nebenprodukt der eigentlichen Vorgänge auf dem Energiemarkt in übergeordneten Systemen wie Energiemanagementsystemen gespeichert werden. (Rigoll, 2017, p. 44) In diesem Fall wird den Daten nur beschränke Achtung geschenkt und ihre Erfassung und Speicherung erfolgt in Systemen, die gar nicht für die Daten selbst konzipiert sind. Darüber hinaus sind viele Datensätze noch ausreichend gut mit den üblichen Standardmethoden handhabbar, so dass speziell auf die Daten zugeschnittene Systeme nicht notwendig erscheinen. (Rigoll, 2017, p. 44) Unabhängig davon, wie genau die Energieverbrauchsdaten erfasst und abgespeichert werden, gibt es aber einige wesentliche Grundlagen, die eine effiziente und zielführende Integration und Speicherung der Messdaten erlauben.

Wie bereits erwähnt müssen die Verbrauchsdaten im einfachsten Fall darüber informieren, wann und wo wieviel Energie verbraucht wurde. Daher werden neben der Energieverbrauchsmessung die physikalischen Daten über Raum und Zeit übermittelt. Die räumliche Dimension wird anhand von Meter ID oder Kunden ID übermittelt, während die zeitliche Dimension anhand des Zeitstempels erkennbar ist. Damit sind die Daten eindeutig zuordnungsbar. Allerdings fehlen in diesem Minimalfall selbst für einfachste Verwendungen wichtige Informationen. Beispielsweise enthält die Meter ID wenige bis keine Informationen, wenn sie nicht in einen größeren Kontext eingeordnet werden, beispielsweise die Stadt oder der genaue Haushalt, in dem der Smart Meter installiert ist. Noch deutlicher wird das Problem aber hinsichtlich darauf aufbauender Verwendungen der Daten. Für den Prozess der Abrechnungserstellung sind neben der reinen Zuordnung, um welchen Verbrauch es sich handelt, außerdem die Informationen über den Kunden, an den die Abrechnung adressiert sein soll, essentiell. Hier müssen beispielsweise Kundenname, Kundenadresse, Familienstand oder Alter bekannt sein. Die als Zeitreihe vorliegenden Messdaten müssen also mit weiteren Informationen versehen werden, die die Messdaten genauer beschreiben. Diese beschreibenden Daten werden als Metadaten bezeichnet. (Rigoll, 2017, p. 96) Theoretisch wäre es, um dies zu lösen, möglich, alle benötigten Zusatzinformationen gemeinsam mit den Messdaten zu übermitteln. Dafür würde jeder Messwert mit den entsprechenden Metadaten versehen, was aber aus zwei Gründen nicht optimal wäre. Zum einen erschwert dies die Übertragung, da dadurch jedes Mal viel mehr Informationen übertragen werden müssen. Die zu übermittelnde Dateigröße würde deutlich ansteigen, was die Übertragung verlängern und das Netzwerk stark belasten würde. Das zweite Problem ergäbe sich bei der Abspeicherung der Daten im Backend des Verteilnetzbetreibers. Jeden einzelnen Messwert gemeinsam mit allen zugehörigen Zusatzinformation abzuspeichern würde vielfach zum wiederholten, redundanten Abspeichern der genau gleichen Daten führen und wäre äußert ineffizient. Effiziente Integrationsarchitekturen haben daher gemein, die Zeitreihen- und Metadaten in separaten Datenbanken zu speichern und die Metadaten im Bedarfsfall zu referenzieren. (Rigoll, 2017, p. 96) Da sich die Zeitreihen- und Metadaten in ihren Eigenschaften grundlegend unterschieden, können auch die dafür vorgesehenen Datenbanken sehr voneinander abweichen.

Bei den Zeitreihen kann es sich um sehr große Datenmengen handeln. Sie müssen insbesondere effizient abgespeichert und einfach ausgelesen werden können. Zur Abspeicherung können hierfür sowohl relationale als auch dokumentenorientierte Datenbanken zur Anwendung kommen. Relationale Datenbanken existieren seit mehreren Jahrzehnten, während dokumentenorientierte Datenbanken

noch eine recht junge Technologie sind. Erstere basieren auf einem vorher definierten und festen Datenbankschema und speichern die Messdaten in tabellenähnlicher Form ab. Letztere hingegen weisen kein festes Datenbankschema auf und sind flexibler einsetzbar. Da die Messdaten nicht in tabellarischer Form abgespeichert werden, können nachträgliche Änderungen einfacher möglich sein.(Rigoll, 2017, pp. 96-97) Ein Messpunkt entspricht in einer relationalen Datenbank einer Zeile und in einer dokumentenorientierten Datenbank einem Dokument. Bei einer großen Anzahl an Messungen kann dies aber ineffizient und langsam sein. (Rigoll, 2017, pp. 97-98) Alternativ existieren daher seit geraumer Zeit auch auf Zeitreihen spezialisierte Zeitreihendatenbanken, die eine effiziente Speicherung und viele für Analysen nützliche Operationen auf den Zeitreihen anbieten. Eine weiterer Lösungsansatz besteht in der dateibasierten Speicherung. Neben klassischen kommaseparierten Daten existieren auch auf große Datenmengen zugeschnittene Lösungen wie HDF5. (Rigoll, 2017, pp. 97-98) Die Metadaten enthalten Informationen über die jeweiligen Zeitreihen wie z.B. die physikalische Größe der Verbrauchsmessung oder die Namen der Verbraucher und haben daher im Gegensatz zu den Zeitreihendaten einen vergleichsweise geringen Speicherplatzbedarf. Die Datenbank für die Metadaten muss vor allem eine gute Durchsuchbarkeit der Daten ermöglichen. Auch hier bieten sich relationale oder dokumentenorientierte Datenbanken an. (Rigoll, 2017, pp. 96-97)

**Weitere Integrationsarchitekturen und Lösungsansätze**

Zur Verarbeitung von Smart Meter Daten stehen verschiedene Software-Lösungen zur Verfügung. Bekannte Software-Unternehmen aus dem Bereich der Datenverarbeitung wie IBM oder Oracle sehen in der Bereitstellung von MDMS eine lukrative Geschäftsmöglichkeit und haben daher verschiedene Lösungen auf den Markt gebracht. (Mikkelsen et al., 2016, p. 4) SAP bietet in diesem Zusammenhang auch ganze EDMS wie „SAP for Utilities“ an. Diese weisen eine ganze Palette industriespezifischer Funktionen und Tools auf. Insbesondere enthalten sie Funktionen zur Konsistenzprüfung und Ersatzwertbildung sowie zur Zeitreihenadministration und -darstellung. (Schweinfurth, 2020, pp. 233-234) Schweinfurth bietet hier eine gute Zusammenfassung sämtlicher Funktionen der verschiedenen SAP-Lösungen. (Schweinfurth, 2020, pp. 233-237)

In der Wissenschaft wird jedoch auch an verteilten, dezentralen Systemen geforscht, die zur Speicherung und Verarbeitung der Daten nicht auf einen zentralen Backend-Server setzen. Yi et al. schlagen beispielsweise ein System zum Management und zur Speicherung von Echtzeit Smart Meter Daten vor, dass auf verteile, lokale Speicherung der Messwerte innerhalb der Messinfrastruktur, anstatt auf einen zentralen Backend-Server des entsprechenden Energieunternehmens, setzt und argumentieren, dass die Speicherkapazität von Smart Metern ausreichend groß sei oder entsprechend erweitert werden könne. (Yi, Choi, & Hwang, 2014) Smart Meter lesen dann periodisch die Energieverbräuche aus und speichern diese lokal – wahrscheinlich auf dem integrierten Speichermedium – ab. Die Übermittlung und Kommunikation der Messwerte mit den verschiedenen Akteuren der Energiebranche erfolgt dann über verschiedene Hierarchiestufen. Zunächst übermittelt ein Router die Messwerte an den nächstgelegenen, zuständigen, lokalen Manager. Der lokale Manager aggregiert dann alle Messwerte der entsprechenden Gegend und übermittelt sie an einen Concentrator, der als Gateway das Smart Meter Netzwerk mit dem Internet verbindet. Auf diese Weise kann der Concentrator jegliche, lokal aggregierte Messwerte durch das Internet auf beliebige Server übertragen.

(Yi et al., 2014) Sirojan et al. schlagen die Nutzung von Edge Computing zur verteilten, dezentralen Verarbeitung von Smart Meter Daten vor. (Sirojan et al., 2019) Dabei werden die Daten direkt, durch Ausrüstung der Smart Meter mit entsprechender Hardware, in den Smart Metern selbst verarbeitet und prozessiert. Auf diese Weise wird die Rechenkraft laut der Autoren direkt an die Stelle gebracht, an der sie eigentlich benötigt wird. Dadurch sollen bestehende Einschränkungen von klassischen AMIs umgangen sowie die Datenverarbeitungszeit reduziert werden. Außerdem, so die Autoren, würde dies die benötigte Bandbreite zur Übermittlung der Messdaten reduzieren, die sonst aufgrund der hochfrequenten Datenübertragung sehr groß ist. (Sirojan et al., 2019)

## 2.2. Intelligente Gaszähler

Der überwiegende Teil der Forschung konzentrierte sich bisher auf die Nutzungsmöglichkeiten von intelligenten Stromzählern. (Hurst, Montañez, Shone, & Al-Jumeily, 2020, pp. 7878-7879) Wang et al. bieten hier einen guten Überblick über die verschiedenen Anwendungen und Herausforderungen bei der Nutzung feingranularer Stromverbrauchsdaten. (Wang, Chen, Hong, & Kang, 2018) Auch wenn sie in aller Regel weniger bekannt sind und weniger diskutiert werden, können aber auch intelligente Gaszähler in Gebäuden installiert werden. Da Erdgas aufgrund verhältnismäßig geringer CO2-Emissionen eine wichtige Rolle in der aktuellen und zukünftigen Energieversorgung Deutschlands spielt, haben auch diese Zähler eine nicht unwesentliche Bedeutung. Insbesondere der Heizwärmebedarf von Gebäuden wird in Deutschland aktuell noch zu großen Teilen durch Erdgas gedeckt. (BMWi, 2020) Gaszähler messen den Gasvolumenstrom über die Zeit in Kubikmeter ($m^3$) pro Zeitintervall. Dementsprechend handelt es sich dabei um mechanische Zähler, wobei jedoch verschiedene Technologien zur Messung des Gasvolumenstroms verwendet werden können. Am häufigsten finden in Wohn- und kleineren Nicht-Wohngebäuden Balgengaszähler Anwendung. Weiterhin existieren jedoch auch Drehkolben-, Turbinenrad- und Blendenströmungszähler. Als nicht mechanische Alternative können zudem Ultraschallzähler zum Einsatz kommen. (Sun et al., 2015, p. 467) Da sich Gas jedoch je nach der Temperatur- und Druckverhältnissen stark ausdehnt oder zusammenzieht, muss das Gasvolumen zur Ermittlung des tatsächlichen Energieverbrauchs, beispielsweise für Abrechnungszwecke, zunächst in kWh umgerechnet werden. (Sun et al., 2015, p. 467) Werden diese Zähler mit einer Kommunikationseinheit ausgestattet, können auch diese bidirektional kommunizieren und ihre Messwerte übermitteln. Durch die mechanische Messung sind Gas Smart Meter daher den hybriden Stromverbrauchszählern ähnlich, die den Stromverbrauch ebenfalls elektromechanisch messen und eine Kommunikationseinheit besitzen.

Da Erdgas in erster Linie für die Raumheizung und das Brauchwarmwasser verwendet wird, ist der Gasverbrauch eines Gebäudes weniger komplex als der Stromverbrauch. Während der Stromverbrauch durch sehr viele Faktoren und Haushaltsgeräte beeinflusst wird, hängt der Gasverbrauch zur Deckung der Heizwärmebedarfs überwiegend von der Jahreszeit und den Temperaturen ab. Im Gegensatz zum Stromverbrauch weist er außerdem keine Grundlasten auf. Ferner bestehen viele der Herausforderungen des Elektrizitätsnetzes, wie die Sicherstellung der Netzstabilität oder die Synchronisation von Angebot und Nachfrage, in Gasnetzen nicht, da Erdgas lediglich ein Primärenergieträger ist. Erdgas kann recht einfach transportiert werden und wird, im Gegensatz zu Strom, erst vor Ort durch Verbrennung in Nutzenergie umgewandelt.

Allerdings werden die Eigenschaften von Gas Smart Metern hier nicht näher beleuchtet. Dies hat mehrere Gründe. Zum einen bestehen deutlich größere Literaturbestände bezüglich der Strom Smart Meter. Ferner wurde sich aufgrund der komplexen Anforderungen an das Elektrizitätsnetz, der großen Bedeutung von Elektrizität für die Energieversorgung der Zukunft und der großen Potentiale von Strom Smart Metern dazu entschieden, Strom Smart Meter stellvertretend für sonstige Smart Meter zu untersuchen. Da die Eigenschaften von Strom Smart Metern zuvor bereits sehr ausführlich dargestellt wurden, würde eine zusätzliche, detaillierte Untersuchung von Gas Smart Metern sowohl den Rahmen dieser Arbeit sprengen als auch zu weit von der grundlegenden Fragestellung abweichen. Für die Entwicklung des Konzepts sowie für die Umsetzung sind insbesondere die Schnittstellen, der Inhalt

und das Format der Daten sowie die Integrationsarchitekturen relevant. All diese Punkten wurden bereits detailliert untersucht und gelten prinzipiell ebenso für Gas Smart Meter. Beispielsweise weisen die SMGWs von Gas Smart Metern die genau gleichen Schnittstellen auf und für eine eindeutige Zuordnung der übermittelten Messdaten müssen ebenfalls mindestens ein Zeitstempel und eine Meter ID vorliegen. Hinsichtlich der Ausbreitung weisen Gas Smart Meter jedoch Unterschiede auf. Zwar haben intelligente Gaszähler über die letzten Jahre schrittweise die klassischen Gaszähler ersetzt, bezüglich der Ausbreitung und rechtlichen Situation bestehen jedoch große geographische Unterschiede. Während die Ausbreitung von Gas Smart Metern im Vereinigten Königreich und in Italien rasch vorangetrieben wurde, haben sich Länder wie Spanien und Schweden gegen einen Rollout der Technologie entschieden. (Sun et al., 2015, p. 469)

# 3. Online Analytical Processing Cubes

In Kapitel 2 wurden die Grundlagen von Smart Metern und intelligenten Sensoren erarbeitet und dargestellt. Unter Anderem wurden insbesondere die vorhandenen Schnittstellen sowie der Inhalt und das Format der dadurch generierten und übermittelten Daten erläutert. Darüber hinaus wurden die bestehenden Integrationsarchitekturen und Systeme zum Management der Daten bezüglich des Aufbaus, der Funktionsweise, der Lokalisierung und der wesentlichen Kriterien zur Sicherstellung eines effizienten Ablaufs bei der Integration und Verarbeitung der Daten erörtert. Mit dem Voranschreiten der Ausbreitung intelligenter Energieverbrauchszähler und Sensoren und der folglich entstehenden, enormen Menge an Daten kommt den IT-Architekturen zur Erfassung, Speicherung, Verarbeitung und Aufbereitung zukünftig eine noch wichtigere Stellung zu, als dies aktuell bereits der Fall ist. Unter Anderem stellt sich in Anbetracht der enormen Menge an neu generierten Daten und den darin potentiell enthaltenen Informationen vor allem die Frage nach einer geeigneten IT-Infrastruktur und Technologie zur angemessenen Analyse der Datensätze. Schließlich sind die in den Datensätzen enthaltenen Informationen nur von wirklichem Wert, wenn sie für den jeweiligen Nutzer zugänglich und einfach verständlich sind. Ferner müssen die Daten hinsichtlich einer gewissen Fragestellung abgefragt und untersucht werden können und die gelieferten Ergebnisse müssen auf die tatsächliche Fragestellung Antwort geben. Eine wesentliche Aufgabe solcher Systeme besteht also darin, aus den enormen Datenmengen wirkliche neue, der Fragestellung entsprechende Erkenntnisse und Informationen zu liefern. Darüber hinaus sollten geeignete Visualisierungen der Ergebnisse möglich sein, da diese die Ergebnisse und Zusammenhänge für das menschliche Auge besser kommunizieren als die reine Darstellung von Zahlen. All dies sollte im Idealfall verständlich und unkompliziert, ohne größeren Aufwand sowie ohne längere, zeitliche Verzögerungen durchführbar sein. Neben der Verständlichkeit der gelieferten Analyseergebnisse spielt aber auch die Verständlichkeit und Einfachheit der eigentlichen Datenabfrage eine wesentliche Rolle, da eine das System nutzende Person in der Regel weder tiefere Kenntnisse in der Datenaufbereitung noch in der Datenanalyse oder -visualisierung mit sich bringt. Um das Potential der Datenmengen voll ausschöpfen zu können, ist eine hohe Nutzerfreundlichkeit des Systems also unverzichtbar.

Dieser Herausforderungen sehen sich neben der Energiebranche allerdings noch viele weitere Sektoren gegenübergestellt. Eine Lösung könnte dabei aus dem Bereich des Business Intelligence stammen. In Unternehmen, die Waren oder Dienstleistungen verkaufen, entstehen durch das operative Tagesgeschäft eine Vielzahl an Daten. Beispielsweise werden Waren angeliefert, Waren verkauft oder neue Mitarbeiter eingestellt. Im Zuge der Digitalisierung wurden all diese Daten zunehmend digital abgespeichert und archiviert. Aufgrund der relationalen, objektorientierten Natur der Geschäftsvorfälle und dem Bedarf, die Daten möglichst effizient abzuspeichern und zu späteren Zeitpunkten ändern oder aktualisieren zu können, geschieht dies meist in relationalen Datenbanken. Diese Datenbanken bilden die Relationen und Eigenschaften der verschiedenen Instanzen ab und sind insbesondere darauf ausgelegt, transaktionale Prozesse abzuspeichern. Daher spricht man in diesem Zusammenhang auch von transaktionalen Datenbanken. Im Laufe der Zeit wurde durch die Speicherung der Geschäftsvorfälle in transaktionalen Datenbanken allerdings der Wert dieser Daten für die Überwachung und strategische Planung des Unternehmens deutlich. Eine angemessene Analyse der Daten ist anhand transaktionaler Datenbankstrukturen jedoch nur schwer und mit vielen

Einschränkungen verbunden möglich. Um eine angemessene Analyse durchführen zu können, müssen daher andere Architekturen und Technologien verwendet werden.

Im Zuge dessen entstanden in den 1990-er Jahren die Konzepte BI und Data-Warehousing. Das Konzept des BI verwendet separate, für Analysezwecke ausgelegte Datenbankenstrukturen und errichtet darauf aufbauend, aus in transaktionalen Datenbanken des Unternehmens gespeicherten Daten und sonstigen Datenquellen, ein Data-Warehouse (DW) auf. Basierend auf den Daten des DW können dann mit Hilfe sogenannter Online Analytical Processing Würfel, auch OLAP-Cubes genannt, Daten abgefragt und analysiert werden. Die Kombination aus DW zur Speicherung der Daten und OLAP-Cubes zur Abfrage ebendieser ist speziell auf Multidimensionalität ausgelegt und ermöglicht verschiedenste Analysen wichtiger Unternehmensdaten hinsichtlich verschiedener Dimensionen.

In wie fern DW-Technologien und OLAP-Cubes aus dem Bereich des BI auch zur Integration und Analyse von Smart Meter- und Sensordaten geeignet sind, soll im Rahmen dieser Arbeit untersucht werden. In diesem Kapitel werden daher zunächst die Grundlagen dieses Themas erarbeitet, bevor dann die Ausarbeitung eines entsprechenden Konzepts in Kapitel 4 diskutiert wird. Zuerst werden im Folgenden die Begriffe BI und DW, sowie deren Eigenschaften und Verwendungszwecke, näher definiert. Anschließend werden dann die zum Aufbau eines DW in Frage kommenden Datenbanken diskutiert. Hier wird neben einer Erläuterung der Technologie auch vergleichend auf Vor- und Nachteile eingegangen. Danach wird erörtert, mit Hilfe welcher Prozesse das DW nach erfolgreichem Aufbau der Grundstruktur mit Daten gefüllt wird. Im Anschluss werden dann OLAP-Cubes vorgestellt. Hier werden die wesentlichen Eigenschaften und Anforderungen, der Aufbau, die typischen Funktionen und die verschiedenen Modellierungsverfahren erklärt. Auch auf eine Abgrenzung zu weiteren Informationstechnologien wie Data Mining wird in diesem Zusammenhang eingegangen.

## 3.1. Business Intelligence und Data Warehousing

### 3.1.1. Business Intelligence

Unter Business Intelligence versteht man die entscheidungsorientierte Sammlung, Aufbereitung und Darstellung geschäftsrelevanter Informationen. (Li, 2010, p. 4) BI-Prozesse sind dabei im Allgemeinen durch die Aspekte Datensammlung, Datenaufbereitung, Informationsdarstellung, Entscheidungsorientierung und Fokus auf geschäftsrelevante Informationen definiert. Im Zuge der Datensammlung beinhaltet BI die Steuerung von Zugriffen auf Datenbanken. Die Aufbereitung dieser gesammelten Daten stellt dann das eigentliche Kernstück von BI dar. Mit Hilfe mathematischer Verfahren und Analysetechniken sollen aus den vorhandenen Rohdaten Informationen gewonnen werden. Neben der Sammlung und Aufbereitung der Daten stellt BI darüber hinaus die Darstellung der richtigen und relevanten Informationen in einer geeigneten Struktur für die unterschiedlichen Benutzergruppen sicher. (Li, 2010, p. 4) Dabei dient BI stets dazu, Entscheidungsgrundlagen zu verbessern. Da das Speichern großer Datenmengen für Unternehmen aber aufwändig und teuer ist, gilt für BI-Lösungen stets der Grundsatz, so viele geschäftsrelevante Informationen wie nötig, aber so wenige wie möglich zu verwenden. Zusammenfassend besteht das Ziel von BI also darin, Entscheidungsgrundlagen zu verbessern, die Transparenz von Unternehmensprozessen zu erhöhen und bisher unerkannte Zusammenhänge von isolierten Informationen aufzuzeigen. (Li, 2010, pp. 4-5) Der

Wert von BI liegt dabei in der Generierung wichtiger Informationen und der Schaffung guter Handlungsgrundlagen. Realisiert wird dieser Wert dann in Form von höherer Profitabilität des Unternehmens durch bessere Entscheidungen und optimierte Unternehmensprozesse. Dies bedeutet folglich aber auch, dass BI von wenig Wert ist, wenn die neu geschaffene Informationsbasis und die entdeckten Zusammenhänge nicht durch entsprechende Handlung realisiert werden. Daher sollten Führungskräfte durch BI in die Lage versetzt werden, genau die spezifischen Informationen zu erhalten, die für eine optimale Entscheidung relevant sind, sodass darauf aufbauend spezifische Handlungen durchgeführt werden können. BI-Lösungen müssen deshalb die entsprechenden Tools, Methoden und Prozesse beinhalten, die eine Transformation von Rohdaten in praktisch umsetzbares Wissen ermöglichen. (Prakash & Prakash, 2018, p. 20)

Den Grundbaustein hierfür bildet zunächst die Abfrage und Aufbereitung der Rohdaten aus Datenbanken. In großen Unternehmen stoßen die klassischen, einfachen Abfragewerkzeuge dabei jedoch schnell an die Grenzen ihrer Leistungsfähigkeit. Auch die Performance der Systeme wird bei großen Datensätzen zum zentralen Thema. (Li, 2010, p. 1) BI-Lösungen lösen diese Herausforderungen über eine Veränderung, weg von Integrationsarchitekturen und Datenabfragen, die auf transaktionale Prozesse ausgelegt sind, hin zu für analytische Prozesse optimierten Systemen. BI-Systeme beinhalten daher Werkzeuge zur analytischen Verarbeitung von Daten, was als Online Analytical Processing, kurz OLAP, bezeichnet wird. Die Nutzung solcher OLAP-Tools ermöglicht es, Daten in mehrdimensionalen Tabellenstrukturen flexibel und schnell zu analysieren und auszuwerten. Als Grundlage für solche OLAP Anfragen muss ein Data-Warehouse aufgebaut werden, das die Daten aus den Quellsystemen integriert und strukturiert bereitstellt. Aufbauend auf diesem DW können dann Reporting-Anwendungen oder verschiedenste OLAP-Werkzeuge, beispielsweise OLAP-Cubes zum Einsatz kommen. (Li, 2010, p. 1) BI ist in diesem Sinne also OLAP zur Unterstützung bei der Entscheidungsfindung im Unternehmen und ermöglicht eine effiziente und verständliche Analyse und Auswertung von Daten. (Li, 2010, p. 5) Das Gesamtsystem hinter BI ist im Folgenden grafisch dargestellt. Sämtliche darin enthaltenen Teilprozesse werden in den folgenden Kapiteln vorgestellt und erläutert.

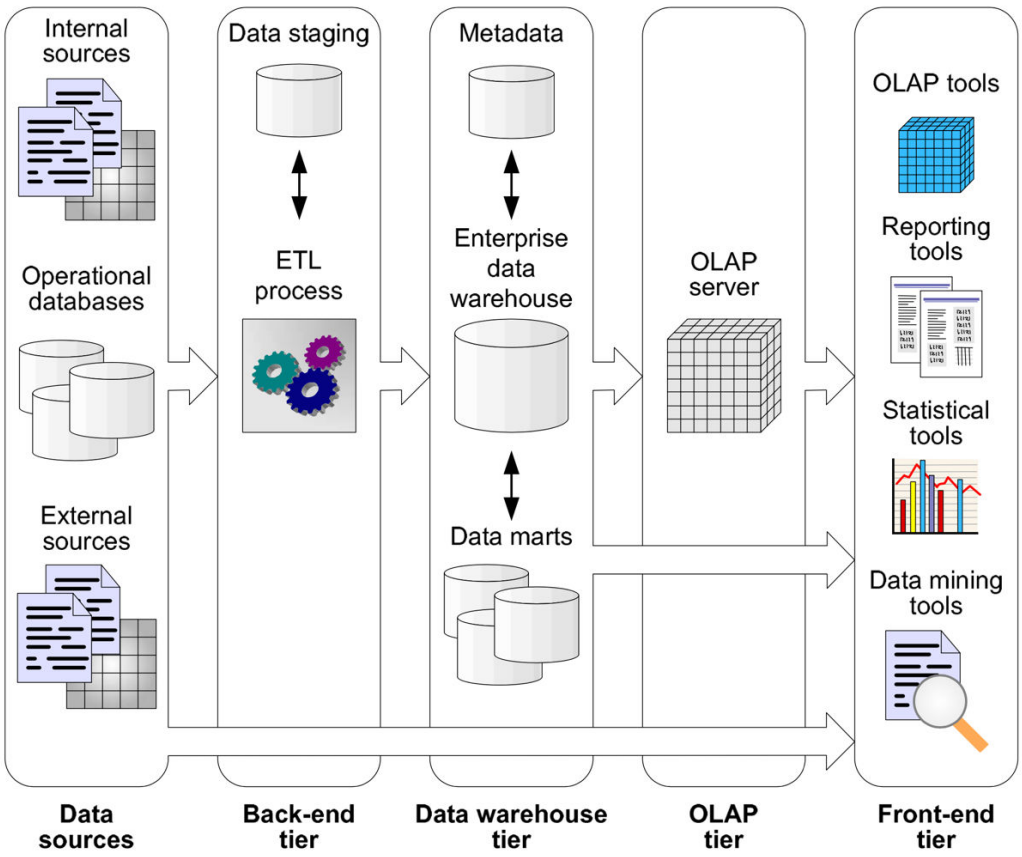



Abbildung 6: Business Intelligence System (Malinowski & Zimanyi, 2008, p. 56)

### 3.1.2. Data Warehousing

BI-Systeme bestehen also im Wesentlichen aus einem Data-Warehouse und Werkzeugen zur analytischen Verarbeitung der darin gespeicherten Daten, indem das DW den Grundbaustein des Systems und die zentrale Anlaufstelle der OLAP-Werkzeuge bildet und über OLAP Anfragen eine Analyse, Auswertung und Visualisierung der Daten realisiert wird. Dem DW kommt daher in BI-Systemen eine zentrale Rolle und folglich eine sehr wichtige Bedeutung zu. Nicht zuletzt ist keine Diskussion von BI möglich, ohne dass der Begriff des DW fällt – manchmal werden die beiden Begriffe sogar quasi synonym verwendet. Deshalb sollen in diesem Abschnitt nun die wesentlichen Eigenschaften eines DW definiert werden.

Seit den 90-er Jahren ist der Begriff des DW zunehmend als die Grundlage für die Extraktion und Aufbereitung brachliegender Unternehmensdaten zu strategisch relevanten Informationen bekannt geworden. Das DW bildet die Datenbasis für analytische Informationssysteme, die Führungskräfte eines Unternehmens bei der strategischen Unternehmensplanung und -analyse unterstützen. (Schlenker, 1998, p. 6) Ein DW stellt somit eine zentrale Stelle dar, an der die Daten gesammelt, gespeichert und für Analysen bereitgestellt werden. Entscheidend ist, dass es sich bei einem DW um ein von den operativen Datenbeständen getrenntes, separates Datenhaltungssystem handelt. Da die darin enthaltenen Daten über OLAP-Werkzeuge abgefragt werden, liegt die Hauptaufgabe von DWs in der Bereitstellung der Daten zu Analysezwecken. (Li, 2010, p. 6) Die wesentliche Datenquelle des DW

stellen dabei die operativen Unternehmensdaten dar. Diese werden normalerweise nach Abschluss des entsprechenden Geschäftsvorfalls archiviert und sollen durch BI-Systeme als Ressource in Form von Informationen für die strategische Planung dienen. (Schlenker, 1998, p. 6) Im Zuge des DW Aufbaus werden als entscheidungsrelevant eingestufte Daten aus operativen Datenbanken, aber auch aus externen Quellen, gesammelt, bereinigt und in eine DW-Struktur transformiert. Diese Extraktion der Daten aus den operativen Datenbanken, gefolgt von der Aufbereitung und Speicherung in einem DW, ist die Grundvoraussetzung für spätere Analysen. Die Aufbereitung und das Laden der Daten in das DW erfolgt in bestimmten Zeitabständen, beispielsweise täglich oder monatlich. Folglich ist die Datenbasis des DW nicht immer in einem aktuellen Zustand und für den Zeitraum bis zur nächsten Aktualisierung konsistent. (Schlenker, 1998, p. 7) Eine DW-Datenbank lässt sich zudem in kleinere „Data Marts“ aufteilen, die eine Art Mini-DW für bestimmte Geschäftsbereiche darstellen können. Das gesamte DW oder die einzelnen Data Marts speisen dann die verschiedenen OLAP-Systeme, die zur Analyse der Daten zur Anwendung kommen. (Schlenker, 1998, p. 7)

Folglich bestehen wesentliche Unterschiede zwischen einem DW und den operativen Datenbanken des Unternehmens, die dazu dienen, die täglichen Geschäftsvorfälle des Unternehmens in Form von Transaktionen zu verarbeiten. (Schlenker, 1998, p. 6) Der Hauptunterschied zwischen einer operativen Datenbank und einem DW liegt in dem Ziel, auf das sie ausgerichtet sind. Während ein DW OLAP unterstützt, dient eine operative Datenbank der transaktionalen Verarbeitung von Daten und enthält die Daten aller Geschäftsvorfälle des Unternehmens. Operative Datenbanken können daher als eine Art „Gegenstück“ zu DWs, die der Analyse der Daten dienen, gesehen werden. Transaktionale Vorgänge werden wiederum als Online Transaction Processing (OLTP) bezeichnet. Solchen operativen Informationssystemen liegt meist eine normalisierte, relationale Datenbank zugrunde, was eine effiziente, für das Tagesgeschäft optimierte Verarbeitung von Transaktionsdaten ermöglicht. (Li, 2010, p. 19; Schlenker, 1998, p. 6) Wenn also beispielsweise eine Bestellung nachträglich geändert wird, ist durch OLTP stets die aktuelle, modifizierte Datenlage bereitgestellt. (Prakash & Prakash, 2018, p. 20) Im Gegensatz dazu besteht der Sinn von einem DW und OLAP in der Analyse der Daten, der Unterstützung bei Entscheidungsprozessen und dem intuitiven Verständnis der gelieferten Abfrageergebnisse. (Li, 2010, p. 19) Um dies zu ermöglichen, stellt ein DW eine Sammlung von Daten aus verschiedenen Datenbanken, Dateien, E-Mails, Spreadsheets und sonstigen Datenquellen zum Zeitpunkt T dar. Da die Daten eines DW im Gegensatz zu einer operativen Datenbank, die mit jeder neuen Transaktion aktualisiert wird, nur in vorgegebenen Abständen aktualisiert werden, ist es häufig der Fall, dass die Daten des DW weniger aktuell sind als die einzelnen Datenquellen des Unternehmens. (Prakash & Prakash, 2018, p. 20) Die folgende Tabelle gibt zusammenfassend einen Überblick über die wesentlichen Unterschiede zwischen OLTP und OLAP.

| Kriterium | OLTP | OLAP |
|---|---|---|
| Funktion | Zuverlässige Datenverarbeitung von operativen Geschäftsdaten, Bereitstellung transaktionaler Informationen | Unterstützung bei der Entscheidungsfindung, Bereitstellung analytischer Informationen |
| Endanwender | Büroangestellte | Manager, Führungskräfte |
| Daten | Operativ, aktuell, detailliert, flüchtig | Historisch, analytisch verdichtet |
| Datenbankdesign | Prozess- bzw. transaktionsorientiert | objektorientiert |
| Datenbankschema | normalisiert | multidimensional |
| Datenzugriff | Schreiben, lesen, löschen, aktualisieren | Lesen, skalieren (anhängen) |
| Datenvolumen | Gigabyte | Gigabyte bis mehrere Terabyte |
| Redundanz der Daten | Niedrig (normalisiert) | Hoch (de-normalisiert) |
| Aktualität der Daten | Nur aktuelle Daten | Nur historisch verdichtete Daten |
| Behandlung der Zeit | Keine explizite Modellierung der Zeit | Dimension „Zeit“ für komplexe Analysen und Aggregationen |
| Antwortzeitverhalten | Kurz bei schreibendem Zugriff, kein konstantes Abfrageverhalten | Je nach Modell konstantes Abfrageverhalten möglich oder nicht |
| Verwendungshäufigkeit | Sehr hoch | Niedrig bis mittel, nur für Analysezwecke |
| Systemauslastung | Eher konstant | Viele Auslastungsspitzen |

Tabelle 3: Unterschiede zwischen OLTP und OLAP nach (Li, 2010, p. 19; Malinowski & Zimanyi, 2008, p. 42)

Zusammenfassend kann also gesagt werden, dass ein DW für die analytische Informationsverarbeitung etwa die gleiche Stellung einnimmt, wie ein normalisiertes, relationales Datenbanksystem für die operative, transaktionale Informationsverarbeitung. Aufgrund dessen sollte der Entwicklung eines DW im Unternehmen entsprechende Aufmerksamkeit zukommen. (Schlenker, 1998, p. 8)

Aus den aufgezeigten Unterschieden zu den üblichen transaktionalen Datenbanksystemen lassen sich die wichtigsten Anforderungen an ein DW herleiten. Da die Sicht einer Führungskraft auf das Unternehmen multidimensional ist, sollte das DW die Daten multidimensional anordnen und halten, um eine einfache und intuitive Analyse zu ermöglichen. (Schlenker, 1998, p. 8) Die Analyseanfragen sollten alle mit annähernd konstanter, möglichst hoher Geschwindigkeit vom System durchgeführt werden. Um dem Benutzer optimale Bedingungen zu bieten, sollte das DW also kurze Antwortzeiten haben. (Schlenker, 1998, p. 8) Durch die Speicherung historischer Daten aus langen Zeitabschnitten im DW, muss aufgrund der großen Datenmenge und der speicherintensiven Struktur das Datenbankmanagementsystem (DBMS) in der Lage sein, diese Datenmengen zu verwalten und zu verarbeiten. (Schlenker, 1998, p. 8) Im Zuge strategischer Analysen sind oftmals größere Zeiträume

über mehrere Dimensionen hinweg von Interesse. Das DW muss daher auf flexible Abfragen großer Datenmengen optimiert sein. (Schlenker, 1998, p. 8)

Laut Inmon, einem der Begründer der DW-Technologie in den 90-er Jahren, wird ein DW durch die folgenden Eigenschaften charakterisiert. Ein DW ist eine themenorientierte, integrierte, zeitbezogene und nichtflüchtige Sammlung von Daten, um Managemententscheidungen zu unterstützen. (Inmon, Welch, & Glassey, 1997; Li, 2010, p. 7; Prakash & Prakash, 2018, p. 20; Schlenker, 1998, p. 8) Dadurch soll ein DW die ideale Plattform für die Analyse historischer Daten sein. (Schlenker, 1998, p. 8) Themenorientiert bedeutet, dass die Daten entscheidungsrelevant sind und in Dimensionen geordnet und zusammengefasst sind. Die Entstehung der Daten ist dabei nicht von Relevanz, da ausschließlich die Verwendung der Daten von Interesse ist. (Li, 2010, p. 7; Schlenker, 1998, p. 9) Integriert meint, dass alle eine Dimension betreffenden Daten aus den OLTP-Systemen bereinigt und zusammengefasst werden, sodass keine Inkonsistenzen und Überschneidungen mehr existieren. Durch konsistente Bezeichner und einheitliche Strukturen wird Struktur- und Formatvereinheitlichung gewährleistet. (Li, 2010, p. 7; Schlenker, 1998, p. 9) Die Herausforderung besteht hier insbesondere in der Verknüpfung von Daten aus operationalen und externen Systemen. (Malinowski & Zimanyi, 2008, p. 41) Zeitbezogen heißt, dass immer ein Zeitraum angegeben werden muss, über den die Daten ausgewertet werden sollen. Die Daten im DW stellen somit eine Reihe von zeitlichen Schnappschüssen dar. Der Datenhorizont beträgt häufig 5-10 Jahre, um Trends entsprechend erkennen und herausfiltern zu können. Im Gegensatz zu zeitpunktbezogenen OLTP-Systemen sind DWs also zeitraumbezogen. (Li, 2010, p. 7; Schlenker, 1998, p. 9) Ferner halten transaktionale Datenbanken die Daten nur solange, wie diese für das operative Tagesgeschäft relevant sind. Dies entspricht meist nur wenigen Monaten. (Malinowski & Zimanyi, 2008, pp. 41-42) Unter nichtflüchtig versteht man, dass die Daten nach dem korrekten Laden und abgesehen von wenigen Korrekturen nicht mehr geändert, sondern nur noch ergänzt werden, da sich nur so Veränderungen und Trends zeitraumbezogen erfassen lassen. Der Zugriff auf das DW erfolgt deshalb nur lesend. (Li, 2010, p. 7; Schlenker, 1998, p. 9)

Der strategische Informationsbedarf hat im Gegensatz zum funktionalen Charakter operativer Informationen einen multidimensionalen Charakter und setzt sich aus einer größeren Anzahl von Faktoren zusammen. Um Schlussfolgerungen ziehen zu können, müssen die Daten aus mehreren Dimensionen zusammengeführt und verdichtet werden. Traditionelle OLTP-Systeme können solche Analysen nur sehr schwer und zeitintensiv durchführen, da das zugrunde liegende Datenbankmodell die Dimensionalität nicht berücksichtigt. (Schlenker, 1998, pp. 6-7) Darüber hinaus sind die typischen Erstellungs- (create), Aktualisierungs- (update) und Löschoperationen (delete) der klassischen relationalen DBMS für die Anwendungen und Funktionen des DW überwiegend irrelevant. Aus diesen Gründen werden die Daten aus den OLTP-Systemen in ein DW integriert, was eine multidimensionale Ansicht der Daten ermöglicht. (Prakash & Prakash, 2018, pp. 20-21; Schlenker, 1998, pp. 6-7) Eine multidimensionale Datenstruktur, auf dessen Grundlage das DW aufgebaut wird, besteht dabei immer aus den drei grundlegenden Objekten Fakten, Dimensionen und Regeln. Fakten stellen die grundlegenden Daten und Unternehmenskennzahlen, die analysiert werden sollen, dar, während die Dimensionen die verschiedenen Parameter repräsentieren, entlang derer die Fakten analysiert werden sollen. Sowohl Fakten als auch Dimensionen haben ihre eigenen Attribute. (Prakash & Prakash, 2018, pp. 20-21; Schlenker, 1998, p. 9) Fakten sind das Ergebnis unternehmerischer Tätigkeit und stellen

häufig Key Performance Indicators (KPIs), also aussagekräftige Unternehmenskennzahlen, beispielsweise Umsatz, Menge, Deckungsbeitrag oder Rentabilität, dar. Dimensionen schlüsseln die Fakten entsprechend auf und können z.B. Kunden, Produkte, Regionen oder Zeiträume sein. Einige Dimensionen können außerdem weiter zu Hierarchien verdichtet werden. So können Produkte zu Produktgruppen, Kunden zu Kundengruppen, Städte zu Regionen oder Monate zu Jahren zusammengefasst werden. (Prakash & Prakash, 2018, pp. 20-21; Schlenker, 1998, p. 9) Regeln sind Vorschriften, die vorgeben, wie eine Kennzahl berechnet werden soll. Beispielsweise kann der Umsatz aus der abgesetzten Menge mal dem Preis pro Stück errechnet werden. (Schlenker, 1998, p. 9) Wie genau diese multidimensionalen Datenstrukturen eines DW aussehen können, wird im folgenden Kapitel diskutiert.

## 3.2. Datenbankarchitekturen

Ein DW stellt letztendlich eine separate, speziell angelegte Datenbank dar. Eine Datenbank wird dabei hauptsächlich durch ihre Struktur und ihren Aufbau bestimmt. Um eine auf die Verwendung der Daten zugeschnittene und effiziente Integration und Speicherung der Daten aus den Quellsystemen zu gewährleisten, spielt die Wahl der Datenbankarchitektur somit eine zentrale Rolle. Die ausgewählte Datenbankarchitektur stellt nicht nur das Rückgrat des gesamten analytischen Informationssystems dar, sondern hat auch Auswirkungen auf später wählbaren Möglichkeiten der Datenabfrage. Um ein DW aufzubauen, haben sich mehrere Datenbankstrukturen etablieren können. Alle diese Architekturen sind darauf ausgelegt, die Multidimensionalität der Daten abzubilden und sind für mehrdimensionale Datenanalysen optimiert. Der wesentliche Unterschied besteht hier in der Wahl von relationalen oder multidimensionalen Datenbankarchitekturen, also der Frage, wie genau die Multidimensionalität bei der Speicherung der Daten umgesetzt wird. (Malinowski & Zimanyi, 2008, p. 49) Im Folgenden werden zunächst verschiedene relationale Datenbankmodelle erläutert und anschließend multidimensionale Architekturen beleuchtet.

An dieser Stelle sei erwähnt, dass die genaue Definition und Abgrenzung sowie die Eigenschaften insbesondere von multidimensionalen Datenbanken nicht immer konsistent und teilweise schwerer verständlich sind. Leider kommt es in Anbetracht verschiedener Literaturquellen manchmal zu Widersprüchen, insbesondere bezüglich der verschiedenen Vor- und Nachteile sowie der Frage nach einer etablierten Grundtechnologie von DWs, die sich nicht zwangsläufig durch zeitliche Unterscheide und Entwicklungen, neue Innovationen oder Marktdurchdringungen erklären lassen. Um die Unterschiede zwischen relationalen und multidimensionalen Architekturen zu verstehen, ist es daher hilfreich, sich vorab noch einmal das zusammenhängende BI-System vor Augen zu führen. Viele der wesentlichen Unterschiede lassen sich nämlich auch logisch erschließen. Wie bereits erwähnt dient das DW der Sammlung, Speicherung und Bereitstellung von Daten zu Analysezwecken, wodurch das DW den Grundstein des BI-Systems bildet. Der eigentliche Zugriff auf die Daten sowie die Analyse und Auswertung erfolgt dann anhand von OLAP-Anwendungen, denen das DW die Daten bereitstellt, indem es lesenden Zugriff bietet. Entscheidend ist also, dass dies ein zweistufiger (oder mehrstufiger) Prozess ist, der sich in Speicherung und Abfrage unterteilen lässt. Daher gibt es Architekturen zur Speicherung und zur Abfrage der Daten. Ein wesentliches Werkzeug zur Abfrage und Analyse der Daten stellen OLAP-Cubes dar, die in 3.4 behandelt werden. Auch diese können sowohl als relationales als auch multidimensionales Modell konzipiert werden. Darüber hinaus gibt es weitere mögliche Modelle und Mischformen. Erfolgt die Speicherung multidimensional, so muss auch die Abfrage durch OLAP-Anwendungen multidimensional erfolgen. Kommen zur Speicherung relationale Datenbanken zum Einsatz, erfolgt die Abfrage meist ebenfalls relational, jedoch sind auch multidimensionale Modelle möglich. (Schlenker, 1998, p. 12) Die Wahl der DW-Architektur hat somit direkten Einfluss auf die Wahlmöglichkeiten für das Modell der OLAP-Anwendung. Ein wirkliches Verständnis des Zusammenhangs und der Unterschiede ist daher nur in Verbindung mit den Ergebnissen aus 3.4 möglich. 3.4 baut somit teilweise auf den Ergebnissen dieses Kapitels auf.

### 3.2.1. Relationale Datenbankarchitekturen

Eine wesentliche Voraussetzung zum erfolgreichen Einsatz von BI ist eine gut strukturierte Datenbank, in der alle benötigten Informationen und Kennzahlen abgelegt werden können. (Li, 2010, pp. 5-6) Eine Möglichkeit, eine solche Datenbank anzulegen, sind relationale Datenbanken. Diese Datenbanken kommen bereits in OLTP-Systemen zum Einsatz und verwenden relationale Datenbankmanagementsysteme (RDBMS). (Schlenker, 1998, p. 13) Daten werden in diesen Datenbanken in Form von Tabellen gespeichert. Diese Tabelle können dann in Relation zueinander gesetzt werden, daher der Name relationale Datenbanken. Aus vorhandenen Tabellen können auf diese Weise auch neue Tabellen erzeugt werden. Die Interaktion mit der Datenbank, also die Veränderung und Abfrage der gespeicherten Inhalte, erfolgt durch eine Structured Query Language (SQL). (Li, 2010, pp. 5-6) Jedes Attribut einer Tabelle wird dabei als eine separate Spalte dargestellt, während die entsprechenden Werte innerhalb der zur Spalte gehörigen Zeilen gespeichert werden. Auch wenn relationale Datenbanken die etablierte Basistechnologie von operativen Systemen darstellen, sind die klassischen Architekturen für Analysezwecke ungeeignet, da sie die Daten zur Optimierung des Speicherbedarfs in hoch normalisierter Form abspeichern. Bei der Ausführung komplexer Datenanfragen, die das Verbinden vieler relationaler Tabellen und das Aggregieren von großen Datenvolumina erfordern, wie es bei Analysen der Fall ist, stoßen klassische Architekturen daher schnell an ihre Grenzen. (Malinowski & Zimanyi, 2008, p. 41) Um die mehrdimensionale Natur der Analysen in einem relationalen System nachzubilden werden deshalb spezielle Datenbankarchitekturen benötigt. Die Organisation der Tabellen erfolgt daher in Form von Stern- oder Schneeflockenschemata oder als Mischform oder Erweiterung dieser beiden. (Gutiérrez, 2010, p. 28; Li, 2010, p. 20; Malinowski & Zimanyi, 2008, p. 50) Diese Datenbanken enthalten nach der Befüllung mit Daten dann nur die Daten, die auch für die Analyse von Nutzen sind. Welche Tabellen benötigt werden und nach welchem Schema sie angeordnet werden hängt dabei von den Anforderungen und geplanten Analysen ab. (Li, 2010, p. 47) Die Speicherung der Daten erfolgt je nach Schema in zweiter oder dritter Normalform. (Meier, Kaufmann, & Kaufmann, 2016, p. 202) Dies wird im Folgenden detailliert beleuchtet.

#### Sternschema

Das bekannteste Schema ist das Sternschema, engl. star schema. Dieses unterscheidet zwischen zwei verschiedenen Arten von Tabellen, Fakten- und Dimensionstabellen. Die Faktentabelle enthält die Fakten und Kennzahlen, die analysiert werden sollen. In den Dimensionstabellen sind dann die beschreibenden Informationen zu den Fakten gespeichert, also die Dimensionswerte, hinsichtlich derer die Fakten analysiert werden sollen. Ein Sternschema besteht aus einer Faktentabelle und mehreren Dimensionstabellen, abhängig von der Anzahl der Dimensionen. Ordnet man die Tabellen grafisch so an, dass die Faktentabelle im Zentrum steht und sich die Dimensionstabellen um die Faktentabelle herum verteilen, ergibt sich dadurch ein sternförmiges Gebilde, das die Analogie zu einer Sternform naheliegt und dem Schema seinen Namen verleiht. (Li, 2010, pp. 20-21; Meier et al., 2016, p. 197) Um eine Zuordnung der Dimensionswerte zu den entsprechenden Fakten zu ermöglichen, enthält jede Dimensionstabelle eine Spalte mit Primärschlüsseln, über die die Dimensionswerte eindeutig identifizierbar sind. Zur Zuordnung der Fakten zu den entsprechenden Dimensionen enthält die

Faktentabellen dann mehrere Spalten für die entsprechenden Fremdschlüssel, die auf die zugehörigen Dimensionstabellen und Werte verweisen. (Jahnke, Groffmann, & Kruppa, 1996, p. 7) Die Faktentabelle hat dabei theoretisch keinen Bedarf nach einer eigenen Spalte für die Primärschlüssel, da die Werte der Faktentabelle direkt analysiert werden und die Faktentabelle lesend auf die Informationen der zugehörigen Dimensionstabellen zurückgreift. Stattdessen ergibt sich der Primärschlüssel der Faktentabelle quasi als eine Zusammensetzung der entsprechenden Fremdschlüssel, also der Primärschlüssel der einzelnen Dimensionstabellen. Die Primärschlüssel der Dimensionstabellen sind also als Fremdschlüssel Komponenten des Primärschlüssels der Faktentabelle und es resultiert jeweils ein mehrdimensionaler Primärschlüssel für die einzelnen Fakten und Kennzahlen. Die Fakten der Faktentabelle hängen sozusagen voll funktional von der Menge der Primärschlüssel der Dimensionstabellen ab. (Farkisch, 2011, p. 28; Li, 2010, p. 21; Meier et al., 2016, p. 198) Die Dimensionstabellen eines Sternschemas können dabei Redundanzen aufweisen, insbesondere wenn sie verschiedene Spalten für die einzelnen Hierarchiestufen haben, um Aggregationen zu ermöglichen. Gibt es in der Tabelle für die zeitliche Dimension also beispielsweise verschiedene Hierarchiestufen für verschiedene Granularitäten, so wird in einer Zeile über alle Spalten für die Hierarchiestufen hinweg die größte Aggregationsstufe mehrfach abgespeichert. Enthält die Tabelle beispielsweise die Spalten Primärschlüssel, Jahr, Monat und Tag, so werden die Informationen über das entsprechende Jahr in allen drei Spalten (Jahr, Monat und Tag) abgespeichert, um eine eindeutige Zuordnung auch auf niedrigster Hierarchiestufe (Tag) zu ermöglichen. Die Redundanz bedeutet, dass die Datenmodelle in de-normalisierter Form abgebildet werden. Das heißt, dass es pro Dimension nur eine Tabelle gibt und die Relationen nur bis zur zweiten Normalform normalisiert sind. (Farkisch, 2011, pp. 27, 29; Malinowski & Zimanyi, 2008, p. 50)

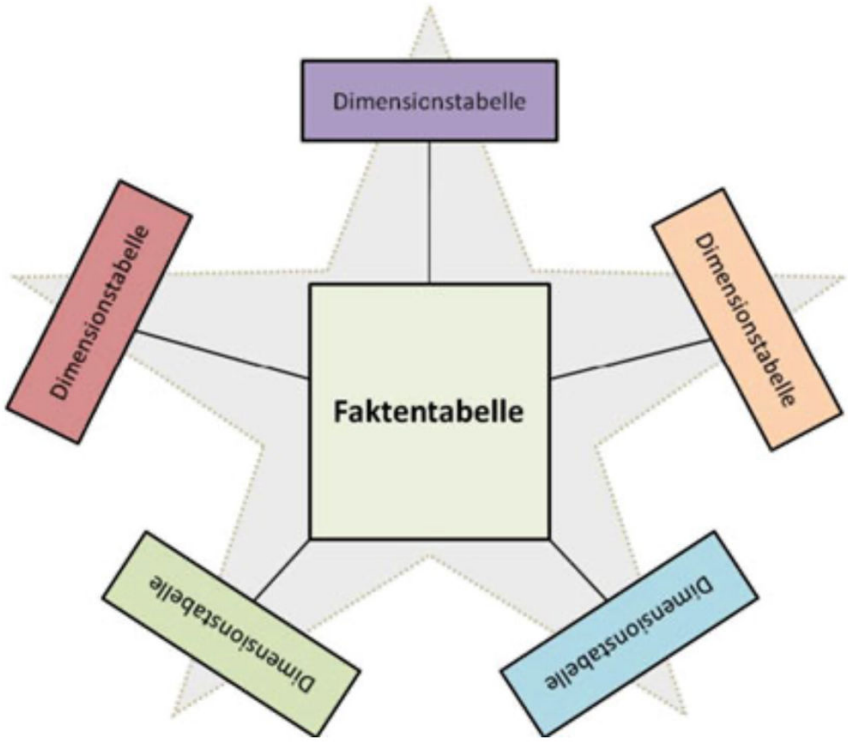



Abbildung 7: Sternschema (Farkisch, 2011, p. 28)

Die Vorteile, die das Sternschema bietet, gliedern sich in allgemeine Vorteile durch die relationale Struktur und konkrete Vorteile gegenüber anderen relationalen Strukturen wie dem Schneeflockenschema. Zunächst werden hier nur die allgemeinen Vorteile dargestellt. Die konkreten Vor- und Nachteile gegenüber dem Schneeflockenschema werden im Unterkapitel zum Schneeflockenschema diskutiert. RDBMS können mit großen Datenmengen umgehen und alle Vorteile von relationalen Datenbanken nutzen. (Gutiérrez, 2010, p. 28) Es handelt sich um eine seit Jahrzehnten weitläufig etablierte Technologie mit ausgereiften, gut erforschten Produkten für verschiedenste Anwendungen. Mit relationalen Datenbanken kann durch die standardisierte Abfragesprache SQL einfach interagiert werden, sie haben eine ausgereifte Metadatenverwaltung und sind für Mehrbenutzerbetrieb ausgelegt. (Schlenker, 1998, p. 13) Darüber hinaus punkten sie durch die ihre gute Skalierbarkeit, die einfache Pflege der Daten und in Bezug auf die Fragen nach Sicherheit, Backups und Datenwiederherstellung. (Farkisch, 2011, p. 27) Durch die Verwendung des Starschemas ist außerdem ein intuitives Verständnis der verschiedenen Dimensionen und Zusammenhänge zwischen Fakten und Dimensionswerten gewährleistet. Die geringe Anzahl an Dimensionstabellen ermöglicht neben einem einfachen Verständnis des Datenmodells auch eine geringe Anzahl von Join-Operationen bei der Abfrage und bedeutet in Kombination mit der Skalierbarkeit durch die Speicherung in Tabellenform einen geringen Aufwand bei der Datenpflege. (Li, 2010, p. 21)

Dem stehen allerdings auch einige Nachteile gegenüber. Ein wesentlicher Nachteil besteht darin, dass SQL keine speziellen Konstrukte zur Datenabfrage nach multidimensionalen Kriterien und zur Durchführung multidimensionaler Operationen bietet. Stattdessen müssen diese mit herkömmlichen SQL-Befehlen nachgebildet werden. Durch das Formulieren dieser SQL-Befehle entstehen umfangreiche Abfragegebilde, um die verschiedenen Tabellen zu verknüpfen und zu durchsuchen. Die Abfragebefehle können immer nur paarweise abgearbeitet werden (Join für Join) und müssen daher mehrfach durchlaufen werden. Dies verschlechtert die Performance des Systems, verursacht langsame Abfragegeschwindigkeiten und verhindert durch das abfragespezifische Leistungsverhalten die Realisation konstanter Antwortzeiten. (Gutiérrez, 2010, p. 13; Schlenker, 1998, p. 28) Insbesondere bei sehr umfangreichen Dimensionstabellen kann das System daher an seine Grenzen stoßen. Die manuelle Formulierung von SQL-Befehlen ist zudem aufwendig und fehleranfällig. (Meier et al., 2016, pp. 198-199) Die im DW üblichen hierarchischen Strukturen innerhalb einer Dimension sind anhand des Sternschemas nicht ideal darstellbar und Aggregationen werden nicht explizit unterstützt. Ferner stellt sich die Frage, wie genau das Wechseln zwischen verschiedenen Hierarchiestufen und eine Änderung der Struktur der Dimensionsanordnung (Pivoting) in SQL umgesetzt werden soll. (Meier et al., 2016, pp. 198-199; Schlenker, 1998, p. 13) Allgemein muss daher die Umsetzung der typischen OLAP-Funktion (siehe 3.4) bei Verwendung von RDBMS durch spezifische Software „erkauft“ werden. Hersteller haben allerdings versucht darauf zu reagieren, indem sie ihre Softwarepaletten um geeignete Werkzeuge angereichert und den SQL-Standard zur Formulierung dieser Operationen erweitert haben. (Meier et al., 2016, pp. 198-199)

Beispielhaft ist im Folgenden ein Sternschema zur Modellierung von Kaufaufträgen dargestellt.

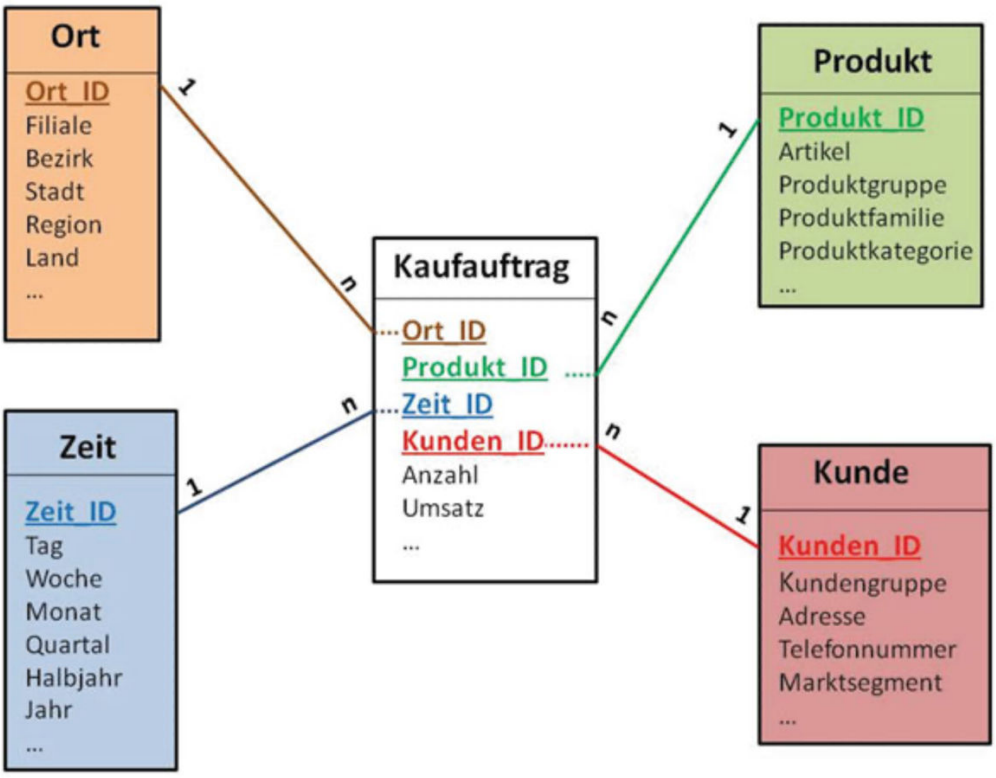



Abbildung 8: Sternschema für Kaufaufträge (Farkisch, 2011, p. 29)

**Schneeflockenschema**

Das zweite bekannte relationale Schema ist das Schneeflockenschema. Das Schneeflockenschema ist eine Erweiterung des Sternschemas. Das bedeutet, dass die Grundarchitektur des Sternschemas mit einer Faktentabelle in der Mitte und den rundherum angeordneten Dimensionstabellen bestehen bleibt, die einzelnen Dimensionstabellen jedoch auf weitere Dimensionstabellen, auch Satellitentabellen genannt, verweisen können. Durch diese Weiterverzweigung des Datenmodells entsteht die Form einer Schneeflocke, was dem Schema seinen Namen verleiht. Auf diese Weise enthalten die Dimensionstabellen nicht mehr alle Dimensionswerte, sondern nur die Daten für die entsprechende Dimensionshierarchie. (Li, 2010, p. 22) Die Satellitentabellen stellen daher eine Art Verdichtungstabellen zur Abbildung der hierarchischen Struktur der Dimensionen dar. (Jahnke et al., 1996, pp. 7-8) Jede Hierarchiestufe einer Dimension wird auf diese Weise in einer eigenen Tabelle gespeichert. Aufgrund dessen erfolgt die Speicherung der Daten bei Verwendung des Schneeflockenschemas normalisiert, in der dritten Normalform. Folglich liegen im Gegensatz zum Sternschema keine Redundanzen mehr vor, da die höchste Aggregationsstufe (beispielsweise das Jahr bei der Dimension Zeit) nicht auch in jeder darunter liegenden, feineren Hierarchiestufe abgespeichert werden muss. Dies führt zu einer kleineren und besser strukturierten Datenmenge. Ein Schneeflockenschema ist also ein in dritter Normalform vorliegendes Sternschema. (Farkisch, 2011, p. 30; Li, 2010, p. 22; Malinowski & Zimanyi, 2008, p. 50) Bei Schneeflockenschemata kommt den Schlüsselbeziehungen also eine besondere Bedeutung zu. Zu jedem Fremdschlüssel einer funktional bestimmenden Relation muss ein passender Primärschlüssel in der funktional abhängigen Relation existieren. Andersherum muss jeder Wert eines Primärschlüssels einer Satellitentabelle einer

Dimension auch tatsächlich von einem Fremdschlüssel der bestimmenden Relation referenziert werden. So wird beispielsweise bei der Dimension Zeit sichergestellt, dass zu jedem Tag zugehöriges Jahr existiert und dass es kein Jahr vorkommt, in dem es keine Tage gibt. (Farkisch, 2011, p. 31)

Im konkreten Fall bedeutet dies also, dass die Faktentabelle im Mittelpunkt die Fakten und Kennzahlen sowie die Primärschlüssel der Dimensionstabellen der ersten Ebene, die direkt auf die Faktentabelle verweisen, als Fremdschlüssel abspeichert. Die Dimensionstabellen der ersten Ebene weisen dann neben ihrem Primärschlüssel, den Dimensionswerten der niedrigsten, am höchsten aufgelösten Hierarchiestufe (z.B. Sekunden in der Dimension Zeit) und eventuellen weiteren Informationen noch eine weitere Spalte für Fremdschlüsselwerte auf. Diese Fremdschlüssel verweisen auf die Primärschlüssel der angebundenen Satellitentabelle der zweiten Ebene. Diese enthält dann erneut neben dem Primärschlüssel, den Dimensionswerten der nächst höheren Aggregationsstufe (z.B. Minuten in der Dimension Zeit) und eventuellen weiteren Informationen eine Fremdschlüsselspalte, die auf die nächst höhere Ebene verweist. Allerdings gibt es neben dieser ersten Variante der Anordnung, die der Darstellung der Dimensionshierarchien dient, noch eine weitere Möglichkeit der Anordnung. Es ist ebenfalls möglich, dass eine Dimensionstabelle einer Stufe auf gleich zwei verschiedene Satellitentabellen verweist und daher auch zwei Fremdschlüsselspalten enthält. Dies kann beispielsweise dann der Fall sein, wenn nicht nur Hierarchien gespeichert werden. In einer Dimensionstabelle für die Zeit könnte beispielsweise die Tabelle neben den Dimensionswerten der aktuellen Hierarchiestufe (z.B. Tag) und den Verweisen auf die nächsthöhere Hierarchiestufe (z.B. Monat) zusätzlich auf eine weitere Tabelle verweisen, welche die Wochentage abspeichert. Würde in diesem Fall ein stattdessen ein Sternschema verwendet, so müssten die Wochentage entweder in eine separate Dimensionstabelle oder in die gleiche Dimensionstabelle, was die Wiedergabe der Aggregationsstufen stören würde.

Als relationales Schema weißt das Schneeflockenschema die gleichen Vor- und Nachteile relationaler Datenmodelle auf, die auch das Sternschema aufweist. Allerdings gibt es konkrete Unterschiede zu dem Sternschema, die hier näher beleuchtet werden sollen. Der Vorteil des Schneeflockenschemas gegenüber des Sternschemas liegt vor allem darin, dass die Daten normalisiert abgespeichert werden. Dadurch erleichtert sich die Pflege der Daten, da beispielsweise neue Datenwerte einfacher ergänzt werden können. Ein weiterer Pluspunkt ist, dass durch die Normalisierung keine Dimensionswerte mehr redundant abgespeichert werden müssen, was insbesondere bei großen Tabellen den Speicherbedarf erheblich reduzieren kann. (Malinowski & Zimanyi, 2008, p. 50) Dem stehen allerdings teils gravierende Nachteile gegenüber, was dazu führt, dass für DW und OLAP-Anwendungen in der Regel dennoch Starschemata bevorzugt werden. (Farkisch, 2011, p. 31) Durch die zusätzlichen Verzweigungen ist das Datenmodell weniger intuitiv und schwerer verständlich als das intuitive, einfache Sternschema. Auch die Übersichtlichkeit des Datenmodells wird dadurch reduziert. Ferner führen die Satellitentabellen zu noch komplizierteren Abfragesequenzen als dies beim Sternschema ohnehin schon der Fall sein kann. Dies erhöht neben der Fehleranfälligkeit bei der Erstellung von Abfragen weiterhin die ohnehin schon kritische Antwortzeit des Systems. Beide Argumente sind auf die zusätzlich notwendigen, aufwendigen JOIN-Operationen zurückzuführen, die bei der Abfrage benötigt werden, um auf die Werte der höheren Aggregationsstufen zuzugreifen. Dimensionshierarchien werden im Sternschema hingegen durch die verschiedenen Spalten der Dimensionstabelle

repräsentiert, was das Erstellen und Modifizieren von Hierarchien erleichtert. (Farkisch, 2011, p. 31; Malinowski & Zimanyi, 2008, p. 50) Die Normalisierung der Dimensionstabellen bei Schneeflockenschemata erschwert also nicht nur die Ausformulierung von Analyseanfragen, sondern kann auch die Leistungsfähigkeit des Systems durch erhöhte Antwortzeiten reduzieren. Darüber hinaus ist eine Nichteinhaltung der Normalisierung wie bei Sternschemata im Rahmen von entscheidungsunterstützenden Systemen nicht so kritisch, da sich die Daten nur selten verändern, wenn überhaupt ergänzt werden. Der so verursachte zusätzliche Speicherbedarf durch die Redundanzen ist ebenfalls eher unkritisch, da die de-normalisierten Dimensionstabellen im Vergleich zur ohnehin normalisiert vorliegenden Faktentabelle vergleichsweise klein sind. (Farkisch, 2011, p. 31)

Werden die zuvor dargestellten Kaufaufträge anstelle eines Sternschemas durch ein Schneeflockenschema modelliert, so ergibt sich die folgende Datenbankstruktur.

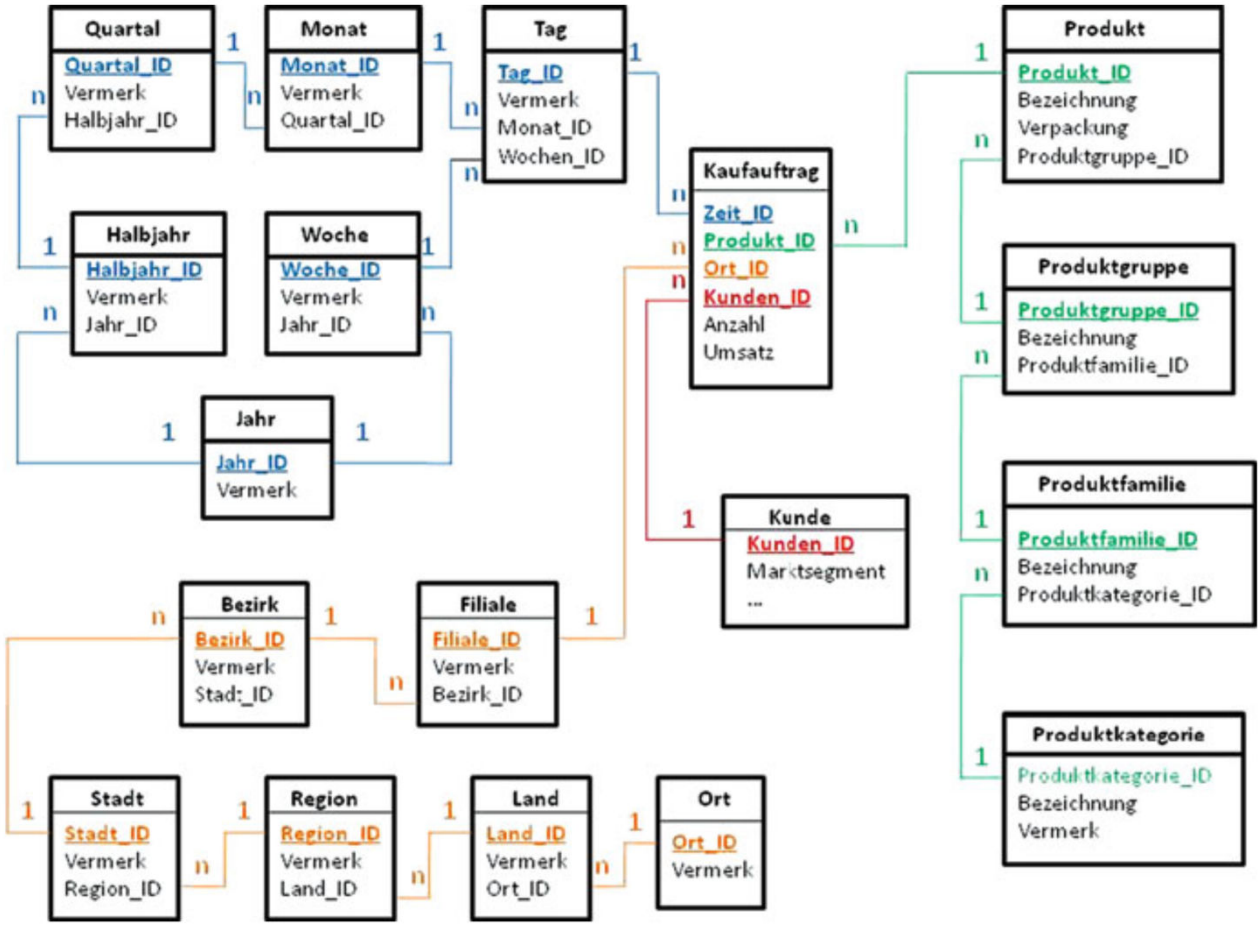



Abbildung 9: Schneeflockenschema für Kaufaufträge (Farkisch, 2011, p. 30)

### Mischschema

Ferner existieren Mischschemata, engl. auch als starflake schema bezeichnet. Solche Mischformen sind prinzipiell ein Sternschema, bei dem jedoch einzelne Dimensionen stattdessen als Schneeflockenschema modelliert werden. Diese Formen stellen also eine Kombination aus Stern- und

Schneeflockenschema dar, bei denen einige Dimensionen normalisiert sind, andere hingegen nicht. (Malinowski & Zimanyi, 2008, p. 50) Dieser mögliche Wechsel zwischen Speicherung der Dimensionen in der zweiten oder dritten Normalform kann aus Gründen der Performance- oder Effizienzsteigerung zur Anwendung kommen. Beispielsweise empfiehlt sich bei den sich häufig ändernden Dimensionstabellen eine Normalisierung, um den Pflegeaufwand zu reduzieren, während bei sich nur selten ändernden Dimensionen (beispielsweise Adressen oder Filialen) weiterhin von dem Performancegewinn durch eine De-Normalisierung profitiert werden kann. (Farkisch, 2011, p. 32)

**Galaxie-Schema**

In einigen Fällen kann es vorkommen, dass eine Kennzahl der Faktentabelle nicht über alle Dimensionen hinweg einen Sinn ergibt. Aus diesem Grund besteht die Möglichkeit, ein Stern- oder Schneeflockenschema zu einem Galaxie-Schema, engl. auch als galaxy schema oder constellation schema bezeichnet, zu erweitern. Ein Galaxie-Schema weißt mehrere Faktentabellen auf, deren Fakten aber nur auf Dimensionstabellen verweisen, die für diese Fakten einen Sinn ergeben. Auf diese Weise können sich mehrere Faktentabellen eine gemeinsame Dimensionstabelle „teilen“. (Farkisch, 2011, p. 32; Malinowski & Zimanyi, 2008, p. 50; Schlenker, 1998, p. 10) Um dies gleich an einem im Zuge dieser Arbeit relevanten Beispiel zu verdeutlichen, wäre beispielsweise denkbar, dass ein DW mit den Stromverbrauchswerten verschiedener Privathaushalte inklusive der durch private PV-Anlagen produzierten Strommenge und der von Windkraftanlagen produzierten Strommenge aufgebaut wird. Die Messwerte der Haushalte und Windkraftmenge stellen in diesem Fall die Fakten dar, ergeben jedoch nicht immer über die gleichen Dimensionen einen Sinn. Während die Messwerte der Privathaushalte über die Dimension der Energieeffizienzklassen von Gebäuden einen Sinn ergeben, ist dies für Windkraftanlagen nicht der Fall. Daher sollten Messwerte der Windkraftanlagen in einer separaten Faktentabelle gespeichert werden, die nicht auf die Energieeffizienzdimension verweist. Hingegen beinhalten beide Faktentabellen eine zeitliche Dimension und können sich daher unter Umständen die zeitliche Dimensionstabelle teilen. Erfolgt die Speicherung der räumlichen Dimension in Form von geographischen Koordinaten, können sich beide Faktentabellen auch hier eine Dimensionstabelle teilen, während dies nicht möglich ist, sollten die räumlichen Koordinaten der Privathaushalte stattdessen als Adressen gespeichert werden, da Windkraftanlagen in der Regel keine klassischen „Adressen“ aufweisen.

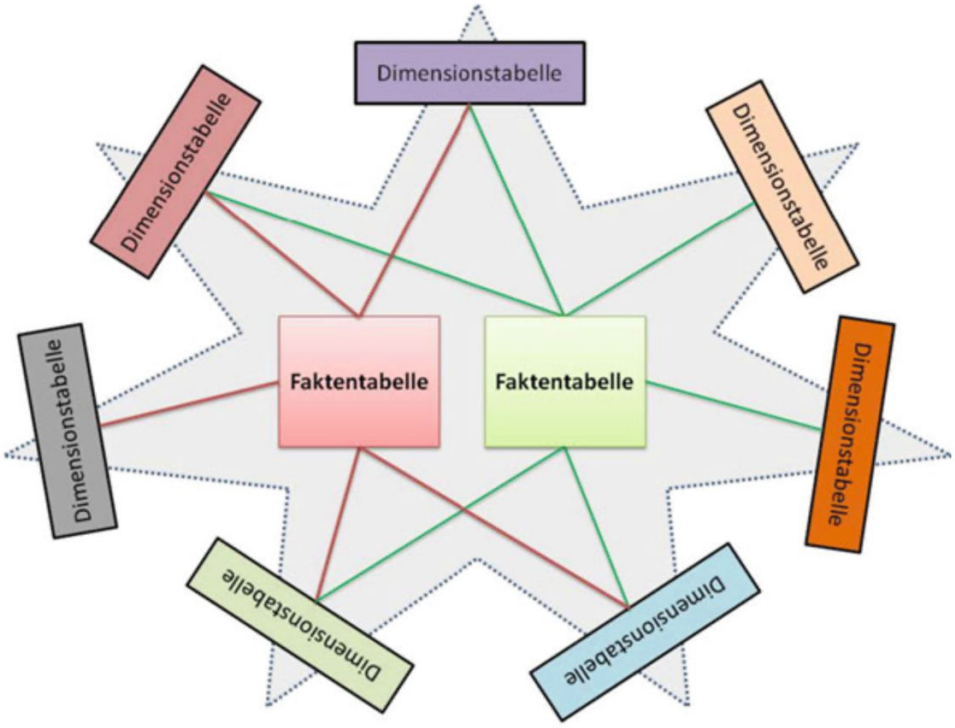



Abbildung 10: Galaxie-Schema (Farkisch, 2011, p. 32)

### 3.2.2. Multidimensionale Datenbankarchitekturen

Neben der Speicherung der Daten in relationalen Tabellen, die so modelliert sind, dass die Multidimensionalität der Daten wiederspiegelt wird, besteht ferner auch die Möglichkeit, die Daten in multidimensionalen, nicht-relationalen Datenbanksystemen zu speichern. Multidimensionale Datenbanken sind noch nicht so lange auf dem Markt wie relationale Datenbanken. Dafür sind sie speziell auf die Bedürfnisse multidimensionaler Datenhaltung und -analyse zugeschnitten. Die Speicherung der Daten in diesen Modellen erfolgt nicht in relationalen Tabellen, sondern in multidimensionalen Arrays. Diese Struktur entspricht im Prinzip direkt dem logischen Abbild einer mehrdimensionalen Datenmatrix, wodurch sich ein intuitives Verständnis ergibt. (Schlenker, 1998, p. 14) Die Daten werden voraggregiert in einer multidimensionalen, speziell angepassten, nicht-relationalen Matrixstruktur gespeichert. Die physische Speicherung erfolgt in Arrays. Der Zugriff auf einen entsprechenden Datenwert erfolgt über dessen Position, also die speziellen Koordinaten des Datenpunktes in der Matrix. Bei der Modellierung der Datenmatrix können verschiedene Aggregationen im Voraus berechnet und abgespeichert werden. Somit können die Daten multidimensionaler Datenbanken später leichter durch OLAP-Anwendungen abgefragt und analysiert werden, da das multidimensionale Modell alle typischen OLAP-Operationen ohne Umwege implementieren und durchführen kann. (Farkisch, 2011, p. 27; Gutiérrez, 2010, p. 27; Kappelhoff, 2016, p. 82; Malinowski & Zimanyi, 2008, p. 49) Multidimensionale Datenbankmanagementsysteme (MDBMS) haben eine auf Analysen zugeschnittene Abfragesprache, wodurch sich die Abfrageperformance gegenüber relationalen Systemen mit SQL-basierter Abfrage deutlich verbessert. (Schlenker, 1998, p. 14)

In Abhängigkeit von der Anzahl der Dimensionen lässt sich gut vor Augen führen, wie genau die Daten in einem MDBMS vorliegen. Im eindimensionalen Fall liegen die Daten schlicht als Reihe bzw. Vektor vor, also als einfaches Array. Ein bekanntes Beispiel wäre hier eine einfache Zeitreihe. Im zweidimensionalen Fall liegen die Daten dann bereits als zweidimensionale Matrix bzw. Tabelle vor. Im dreidimensionalen Fall erweitert sich das Modell auf eine dreidimensionale Matrix bzw. einen Würfel. Im Fall von n Dimensionen liegen die Daten dann in Form eines n-dimensionalen Würfels vor. Aufgrund der Würfelform ist es bereits an dieser Stelle nicht sonderlich überraschend, dass diese Art der Datenspeicherung neben relationalen Stern- oder Schneeflockenschemata eine häufig verwendete Datenbankarchitektur für bestimmte OLAP-Cube Technologien darstellt. Auf die genauen Zusammenhänge zwischen den Datenbankarchitekturen und den resultierenden OLAP-Cube Technologien zur Abfrage der Daten wird in Teil 3.4 näher eingegangen. Solche Datenwürfel können je nach Situation entweder eher dünn oder sehr dicht besiedelt sein. Dies hängt davon ab, wie viele der insgesamt möglichen Kombinationen von Dimensionswerten tatsächlich Datenwerte bzw. Fakten aufweisen. Wenn beispielsweise nicht alle Produkte zu allen Zeiten in allen Filialen verkauft werden, ergibt sich ein eher dünn besiedelter Datenwürfel. In der Praxis sind die meisten Würfel im Rahmen von BI-Anwendungen eher dünn besiedelt. (Malinowski & Zimanyi, 2008, p. 44)

Im Kontext eines MDBMS stellt eine Dimension eine geordnete Liste von Dimensionswerten, welche die Werte der Dimensionselemente und die Klassifikationsknoten höherer Aggregationsstufen enthalten, dar. Die Dimensionswerte umfassen somit sämtliche Ausprägungen einer Dimension. Die Ordnung der Liste ermöglicht das Suchen und den Zugriff auf die Dimensionswerte. Anhand dieser Ordnung wird jedem Dimensionswert eine festgelegte ganze Zahl als Ordnungszahl zugewiesen, wodurch der Dimensionswert eindeutig identifizierbar ist. (Farkisch, 2011, pp. 32-33) Durch die m Dimensionswerte einer bestimmten Dimension lässt sich der Datenwürfel bezüglich dieser Dimension in m unterschiedliche, parallele Ebenen aufteilen. Der Schnittpunkt von n Ebenen eines n-dimensionalen Datenwürfels identifiziert so genau eine einzelne Zelle innerhalb des Datenwürfels. In dieser Zelle ist dann die zugehörige Kennzahl abgelegt. (Farkisch, 2011, pp. 32-33) Die Indizes dieses multidimensionalen Arrays sind also die Koordinaten der verschiedenen Zellen. Anhand dieser Indizes werden die Koordinaten der Zelle eindeutig bestimmt und ein Zugriff auf die Zelle ermöglicht. Dabei wächst die Anzahl der Zellen proportional mit der Anzahl der Dimensionen und zugehörigen Dimensionswerten. Anhand der mehrdimensionalen Koordinaten der Zelle wird dann anhand einer Rechenformel ein genauer Index einer Zelle errechnet. Der Index einer Zelle z mit den Koordinaten $(x_1, x_2, \dots, x_n)$ lässt sich folgendermaßen ermitteln: (Farkisch, 2011, p. 34)

$$\text{Index}(z) = x_1 + (x_2\text{-}1) * |D_1| + (x_3\text{-}1) * |D_1| * |D_2| + \dots + (x_n\text{-}1) * |D_1| * |D_2| * \dots * |D_{n\text{-}1}|$$

Zur Veranschaulichung kann man sich beispielhaft einen dreidimensionalen Würfel mit jeweils drei möglichen Dimensionswerten vorstellen. Ein solcher Würfel weist 3 * 3 * 3 = 27 verschiedene Zellen auf. Jede dieser Zellen hat die drei Koordinaten (x, y, z), da ein dreidimensionaler Würfel drei Achsen aufweist. Die Zelle vorne links unten hat daher die Koordinaten (1,1,1), während die gegenüberliegende Zelle hinten rechts oben die Koordinaten (3,3,3) aufweist. Anhand der zuvor aufgeführten Formel errechnen sich dann für erstere Zelle, Respektive letztere Zelle die Indizes wie folgt:

Index(1,1,1) = 1 + (1-1) * 3 + (1-1) * 3 * 3 = 1
Index(3,3,3) = 3 + (3-1) * 3 + (3-1) * 3 * 3 = 27

Somit erhält jede Zelle einen entsprechenden Index zwischen 1 und 27 und ist damit eindeutig identifizierbar. Durch reine Multiplikation der Koordinaten wäre keine eindeutige Zuordnung der Zellen möglich. (2,3,2) würde dann beispielsweise zu dem gleichen Index führen wie (3,2,2). Anhand dessen ist es in einem MDBMS für den Benutzer und den Rechner einfach, Zellen zu lokalisieren, da ihre Positionen bekannt sind. Für den Zugriff auf die Daten ist lediglich die Berechnung der Indizes der Zellen anhand des oben vorgestellten Algorithmus erforderlich. (Farkisch, 2011, p. 35)

MDBMS haben gegenüber RDBMS verschiedenste Vor- und Nachteile. Solche Datenbankmodelle bieten den Vorteil, dass sie speziell an Analysezwecke und die Fragestellungen angepasst sind. Sie spiegeln die multidimensionale Natur der Daten direkt n-dimensional wider. Dies gestaltet neben der Modellierung der Daten auch die Schnittstelle zu den Abfragewerkzeugen, die Endbenutzer zur Analyse verwenden, einfacher, da keine Umsetzung von flachen, relationalen Tabellen auf eine mehrdimensionale Würfelsichtweise mehr notwendig ist. (Jahnke et al., 1996, p. 7) Die typischen Funktionen zur Handhabung der multidimensionalen Informationen sind durch die Speicherung der Daten als n-dimensionale Arrays und den sich so ergebenden n-dimensionalen Würfel dem System bereits inhärent, während diese Funktionalitäten bei Verwendung relationaler Datenmodelle erst noch durch aufwändige Programmierung hinzugefügt werden müssen. Dass die vorliegenden operativen Daten aus den OLTP-Systemen in relationaler Modellierung vorliegen und daher zunächst aufwändig extrahiert, umgewandelt und aufbereitet werden müssen, ist in dem Sinne kein aussagekräftiges Argument, als dass die operativen Daten aus Performance- und Modellierungsgründen auch bei Verwendung eines RDBMS zunächst in ein spezielles, geeignetes Datenmodell überführt werden müssen, beispielsweise ein Sternschema. (Jahnke et al., 1996, pp. 7-8) Der wichtigste Vorteil dieser Datenmodelle liegt allerdings in der Abfrageperformance. Durch die Art der Datenspeicherung und des Zugriffs sind bei der Datenabfrage kürzere Antwortzeiten möglich. Insbesondere bei dicht besiedelten Datenwürfeln ermöglichen diese Systeme einen kompakteren, effizienteren Weg, Daten, die in mehreren Dimensionen verkettet sind, zu verwalten und auszuwerten. Gegenüber RDBMS haben MDBMS also einen Performancevorsprung durch die bereits im Voraus multidimensional organisierten Daten und die effizienten Aggregationen. (Farkisch, 2011, p. 35; Gutiérrez, 2010, pp. 27-28; Kappelhoff, 2016, p. 82; Malinowski & Zimanyi, 2008, p. 50)

Trotz dieser teils vielversprechenden Vorteile und der naheliegenden Vermutung, dass diese Datenmodelle aufgrund der intuitiven Modellierung wohl am besten für mehrdimensionale OLAP-Anfragen und OLAP-Cubes geeignet sein dürften, ist die Verwendung von MDBMS als Basis für das DW auch mit nicht zu vernachlässigbaren Nachteilen verbunden. Es existiert keine standardisierte Abfragesprache, wie dies mit SQL bei RDBMS der Fall ist, wodurch auch kein einheitlicher Funktionsumfang der verschiedenen Systeme gegeben ist. Jedes System ist anders aufgebaut und strukturiert. (Schlenker, 1998, p. 14) Durch die verschiedenen Abfragesprachen und die vergleichsweise Neuheit der Datenbanktechnologie stellt sich zudem die Frage, wie gut diese Systeme sich als Zusatz der betrieblichen Informationsinfrastruktur einsetzen lassen, da neue Administrationskenntnisse im Unternehmen geschaffen werden müssen. (Jahnke et al., 1996, p. 7) Da

diese Bedenken allerdings auf relativ alte Literaturquellen zurückzuführen sind, stellt sich die Frage, inwiefern diese auch noch zur heutigen Zeit gerechtfertigt sind. Unabhängig von möglichen Standardisierungen, der Etablierung bestimmter Systeme und der Weiterentwicklung im Laufe der letzten Jahre, lassen sich allerdings auch einige, dem System inhärente Nachteile identifizieren. Im Gegensatz zu RDBMS sind MDBMS schlechter und aufwändiger skalierbar und die Datenmengen, die sich mit MDBMS verwalten lassen, sind geringer als bei Verwendung eines RDBMS. (Farkisch, 2011, p. 35; Schlenker, 1998, p. 14) Gleichzeitig führen MDBMS aufgrund des zusätzlichen Aufwands durch die Vorausberechnung der Aggregationen für jede mögliche Dimensionskombination zu einem höheren Speicherplatzbedarf als RDBMS. (Kappelhoff, 2016, p. 82; Malinowski & Zimanyi, 2008, pp. 49-50) Weitere Probleme entstehen insbesondere im Fall von dünn besiedelten Datenwürfeln. Durch die Modellierung der Daten als n-dimensionaler Würfel wird im Fall von dünner Besiedelung viel Rechenzeit dadurch verloren, dass Nullen oder NULL-Werte aufaddiert werden. (Gutiérrez, 2010, p. 28) In RDBMS hingegen werden die Daten in verschiedenen Tabellen gespeichert, die bei der Abfrage durchlaufen werden, wodurch lediglich tatsächlich existierende Daten aggregiert werden – und keine zusätzlichen NULL-Werte. Auch die Ergänzung, Erweiterung und Aktualisierung der Datensätze gestaltet sich bei Speicherung der Daten als multidimensionale Arrays schwierig. Zur Erstellung eines solchen n-dimensionalen Würfeln müssen jedes Mal alle Dimensionskombinationen und Datenaggregationen der Datensätze berücksichtigt und neu berechnet werden. Eine dynamische Ergänzung des Datenwürfels zur Beantwortung spontaner Datenabfragen ist daher nicht praktikabel. (Gutiérrez, 2010, p. 28) RDBMS sind wesentlich ausgereifter und besser skalierbar, sie können größere Datenmengen verwalten und verursachen weniger Probleme bei der Ergänzung und Aktualisierung der Daten. Insbesondere kontinuierliche Updates der Daten im Betrieb sind MDBMS immer noch ein Problem, da die notwendigen Backup- und Widerherstellungsmechanismen vergleichsweise nur rudimentär vorhanden sind. (Farkisch, 2011, p. 36) Auch das Fehlen von festgelegten Zugriffsmethoden und standardisierten Programmschnittstellen führt dazu, dass die Informationen ausschließlich durch die bereitgestellten Tools der jeweiligen Hersteller abgerufen werden können. Die Speicherung der Daten als mehrdimensionale Arrays verursacht zudem einen erhöhten Aufwand bei der Verwaltung der Daten, da Änderungen an den Dimensionswerten jedes Mal eine komplette Restrukturierung und Neuberechnung des Datenwürfels bewirken. Vor angemessener Adressierung dieser Probleme spricht deshalb noch einiges gegen die Verwendung dieser spezialisierten Datenbanksysteme. (Farkisch, 2011, p. 35)

Unabhängig davon, ob die Multidimensionalität der Daten durch ein RDBMS oder ein MDBMS modelliert werden soll, liegen die Daten aus den OLTP-Systemen zunächst ohnehin in relationaler Form vor und müssen erst noch integriert werden. Die Frage, wie die Umwandlung einer flachen OLTP-Datenstruktur auf Basis relationaler Tabellen in eine mehrdimensionale Darstellung der Daten erfolgen soll, kann also in beiden Fällen nicht umgangen werden. Die Integration der operativen Daten aus OLTP-Systemen in ein MDBMS gestaltet sich jedoch auf jeden Fall schwieriger. Insbesondere ab dem vierdimensionalen Fall sind n-dimensionale Arrays und Datenwürfel für das menschliche Auge nicht mehr vorstellbar, während diese Mehrdimensionalität durch relationale Sternschemata unabhängig von der Anzahl der Dimensionen intuitiv und einfach darstellbar ist. Um die Daten in einem mehrdimensionalen Array abzuspeichern, müssen die entsprechenden Zellen zunächst nach dem oben genannten Algorithmus indexiert werden. Durch diese Indexierung sowie die Berechnung

und Zuordnung der Kennzahlen zu den entsprechenden Zellen werden viel Speicherplatz und Rechnerleistung benötigt. Um den Index, der sich aus den Dimensionswerten ergibt, berechnen zu können, müssen zudem alle Dimensionswerte vorliegen, d.h. aus den Datenquellen durch Abfrage extrahiert und zunächst geordnet werden. Die direkte Berechnung der Zellenindexe anhand der Dimensionswerte ist auf Basis von OLTP-Systemen sehr zeitaufwändig und unübersichtlich, da jeder einzelne Dimensionswert erkannt und dann unter Umständen anhand zusätzlicher JOIN-Operationen abgefragt werden muss. OLTP-Systeme sind nicht auf diese Multidimensionalität ausgelegt, was die Extraktion der Daten erschwert. Um dies zu vermeiden könnten die Daten alternativ zunächst in einer zusammenhängenden Tabelle zusammengeführt werden, die jede Kennzahl mit ihren zugehörigen Dimensionswerten kombiniert. Genau das soll allerdings aufgrund der hierarchischen Dimensionsstrukturen, der Größe der Faktentabelle und dem damit einhergehenden Speicherplatzbedarf durch relationale Schemata wie dem Sternschema vermieden werden. Die grundlegende Frage ist also eher, ob die operativen Daten überhaupt auf direktem Weg multidimensional gespeichert werden können oder ob der multidimensionalen Speicherung ohnehin zunächst eine relationale Integration der Daten in Form eines Stern- oder Schneeflockenschemas vorausgehen sollte, auf Basis dessen dann erneut ein multidimensionales Datenmodell aufgebaut wird. Für beide Fälle, sowohl für eine Abfrage auf Basis relational oder multidimensional gespeicherter Datenmodelle, als auch für Mischformen dieser beiden, existieren verschiedene OLAP-Cube-Modelle, die in 3.4 näher erläutert werden. Neben der Frage, ob die Überführung der Daten aus OLTP-Systemen in ein MDBMS tatsächlich ohne weiteres möglich ist oder ob nicht doch zunächst eine Integration der Daten in ein relationales Modell wie ein Sternschema notwendig ist, führt ein MDBMS weiterhin zu Problemen, wenn Dimensionsattribute vorliegen, die sich in ihrer Natur nicht zu Hierarchiestufen aggregieren lassen. Dieser Fall stellt sogar direkt ein Ausschlusskriterium für MDBMS dar, da diese die Daten in n-dimensionalen Datenwürfeln modellieren, der sich logischerweise wiederherum aus kleineren Würfeln mit höherer Granularität zusammensetzt. Dieser Nachteil blieb in den untersuchten Literaturquellen bisher überraschenderweise unerwähnt.

## 3.3. ETL-Prozesse zur Datenaufbereitung

Nachdem im vorangegangenen Unterkapitel die zur Auswahl stehenden Datenbankarchitekturen beleuchtet wurden, stellt sich nun die Frage, wie diese Datenbank mit Daten gefüllt werden soll. Der Prozess der Integration der Daten aus den Quellsystemen in das aufgebaute DW gliedert sich in die drei Schritte Extraktion, Transformation und Laden der Daten, engl. Extract-Transform-Load (ETL). Die genauen Schritte dieses ETL-Prozesses werden hier näher erläutert.

Im Zuge der Extraktion werden die Daten aus den Quellsystemen, also aus operativen Datenbanken, Dateien, Spreadsheets oder sonstigen Quellen extrahiert, um sie in einem gemeinsamen Datenspeicher zu laden. Diese Extraktion kann automatisiert in vorher festgelegten, periodischen Zeitabständen, ereignisgesteuert oder auf manuelle Anfrage durch einen Benutzer geschehen. (Li, 2010, p. 7; Malinowski & Zimanyi, 2008, p. 56)

Transformation beschreibt die Weiterverarbeitung der extrahierten Daten als Vorbereitung für das endgültige Laden der Daten in das DW. Ziel ist hier die Überführung sämtlicher Daten in einheitliche Formate, unter anderem durch die Anpassung der Datentypen, Datumformate, Maßeinheiten, Kodierungen und Kombination bzw. Separation entsprechender Attributwerte. Durch den Transformationsprozess sollen die operativen Daten also in betriebswirtschaftlich interpretierbare Daten umgewandelt werden. Der Transformationsprozess ist der aufwändigste und wichtigste Schritt des ETL-Prozesses und lässt sich in die Teilprozesse Filterung, Harmonisierung, Aggregation und Anreicherung aufteilen. (Li, 2010, pp. 7-8; Malinowski & Zimanyi, 2008, pp. 56-57)

Als erster Schritt der Transformation wird eine Filterung der Daten vorgenommen. Im Rahmen der Filterung werden die für Analysezwecke benötigten Daten, die aus heterogenen unternehmensinternen und -externen Quellen stammen, selektiert, zwischengespeichert und von Mängeln befreit. Aufgrund der Zielvorgaben eines DW sind nämlich nur die Daten zu verwenden, die auch für Analysen benötigt werden und bei Unternehmensentscheidungen unterstützen können. Zunächst werden die extrahierten Daten in speziell dafür vorgesehene Arbeitsbereiche (staging areas) des DW eingestellt. Anschließend werden diese Daten dann von syntaktischen und semantischen Mängeln befreit. (Li, 2010, p. 8)

Den zweiten Schritt der Transformation stellt die Harmonisierung dar, im Zuge derer die gefilterten Daten verknüpft werden. Dabei gibt es zwei Arten der Harmonisierung, die syntaktische und die betriebswirtschaftliche Harmonisierung. Die Notwendigkeit einer syntaktischen Harmonisierung entsteht aufgrund der meist hohen Heterogenität der operativen und externen Daten. Anhand von Transformationsregeln werden hier insbesondere Schlüsseldisharmonien der extrahierten Daten bereinigt, sowie Schwierigkeiten aufgrund von unterschiedlich kodierten Daten, Synonymen und Homonymen gelöst. Schlüsseldisharmonien entstehen durch die Verwendung unterschiedlicher Zugriffsschlüssel in den operativen Datenbanken und stellen Unverträglichkeiten der Primärschlüssel der extrahierten und gefilterten Daten dar. Um dies zu verhindern werden daher in diesem Fall in diesem Teilprozess neue, künstliche Primärschlüssel erzeugt, die zusätzlich die Primärschlüssel aus den operativen Systemen als Fremdschlüssel mitführen können. (Li, 2010, pp. 8-9) Probleme können zudem entstehen, wenn Daten unterschiedlich kodiert sind, da sie über identische Attributnamen mit identischer Bedeutung verfügen, gleichzeitig jedoch unterschiedliche Domänen bzw. Wertebereiche

aufweisen. Die Wahl einer eindeutigen Domäne und die Verwendung entsprechender Zuordnungs- und Mapping-Tabellen lösen dieses Problem. (Li, 2010, p. 9) Attribute mit der gleichen Bedeutung und Domäne, jedoch unterschiedlichen Namen und Bezeichnungen werden als Synonyme bezeichnet. Die Lösung dieses Problems erfolgt durch eine eindeutige Festlegung auf eine der Bezeichnungen und die Überführung aller weiteren Attributbezeichnungen in die gewählte Bezeichnung. Homonyme hingegen weisen denselben Attributnamen, jedoch unterschiedliche Bedeutungen auf. In diesem Fall müssen die Attributnamen neu vergeben werden, um eine Unterscheidung zu ermöglichen. (Li, 2010, p. 9) Neben der syntaktischen Harmonisierung wird für eine optimale Überführung der Quelldaten in analytische Informationssysteme auch eine betriebswirtschaftliche Harmonisierung benötigt. Hierbei werden die betriebswirtschaftlichen Kennzahlen abgeglichen und die benötigte Granularität flächendeckend festgelegt und sichergestellt. Insbesondere Währungen müssen nach spezifischen Kriterien vereinheitlicht und homogenisiert werden, da sonst keine wirklichen Vergleiche der Daten möglich sind. Liegen die Daten in unterschiedlichen Detaillierungsgraden vor, muss eine einheitliche Granularität festgelegt und erzeugt werden. Je detaillierter die Daten vorliegen, desto mehr Informationen können sie für das Management liefern. (Li, 2010, p. 9)

Als dritter Schritt des Transformationsprozesses werden dann Aggregationen durchgeführt. Diese erweitern die gefilterten und harmonisierten Daten um eine Verdichtungsstruktur. Alle Einzeldaten müssen hierbei über Aggregationsalgorithmen zusammengefasst werden. Zur Berechnung bestimmter Kennzahlen kann es nötig sein, auf bestimmten Ebenen Vorsummierungen durchzuführen und diese in den Datenbestand zu übernehmen. Im Zuge des Aggregationsprozesses werden daher Dimensionshierarchien entwickelt und in die Dimensionstabellen übernommen. Kleinere Elemente können dadurch von solchen Dimensionstabellen auf eine höhere Stufe aggregiert werden. (Li, 2010, pp. 9-10)

Als vierten und letzten Schritt der Transformation werden die aggregierten Daten angereichert. Hier erfolgen die eigentliche Berechnung und Integration betriebswirtschaftlicher Kennzahlen, die dann in den Datenbestand aufgenommen werden. Die so gewonnenen Informationen sind von großem Interesse und Nutzen für die Endanwender des entscheidungsunterstützenden Systems. Je nach Anwender können unterschiedliche Kennzahlen errechnet und bereitgestellt werden. Während ein Vertriebsleiter eher Interesse an Rentabilitätsraten und Gewinnen auf Filialebene haben könnte, hätte ein Produktmanager beispielsweise mehr Interesse an Deckungsbeiträgen einzelner Produkte. (Li, 2010, p. 10)

Nach Abschluss des Transformationsprozesses sind die Daten bereit, in das DW geladen zu werden. Das Laden der Daten umfasst die Extraktion der Daten aus eventuellen staging areas, in denen die Daten während des Transformationsprozesses abgelegt wurden, den Transport und die Integration in das Zielsystem, das DW. (Li, 2010, p. 10) Damit sind die Daten in dem Datenbanksystem des DW gespeichert und stehen für Abfragen und Analysen zur Verfügung. Der Ladeprozess wird dabei jedes Mal durchgeführt, wenn die Daten in dem DW aktualisiert werden sollen. Diese Aktualisierungen sind notwendig, da gute Entscheidungen nur auf Basis möglichst aktueller Daten getroffen werden können. Je nach Bedürfnis und Anwendung können unterschiedliche Aktualisierungsfrequenzen – angefangen bei monatlich bis hin zu beinahe Echtzeit – festgelegt werden. (Malinowski & Zimanyi, 2008, p. 57)

Der schrittweise Ablauf des ETL-Prozesses inklusive der Teilprozesse ist im Folgenden übersichtlich zusammengefasst und grafisch dargestellt.

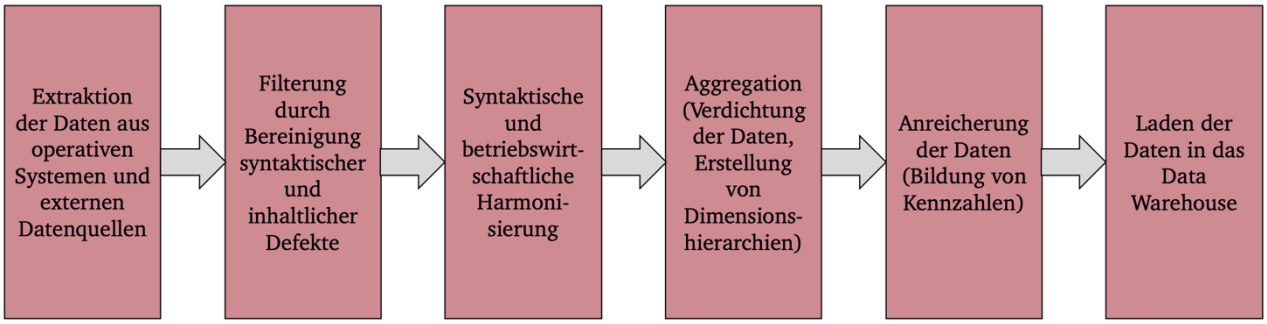



Abbildung 11: Zusammenfassung des ETL-Prozesses

## 3.4. Online Analytical Processing Cubes

Nachdem die Datenbankarchitektur des DW ausgewählt und aufgebaut, sowie die Daten aus den operativen Systemen und externen Datenquellen extrahiert, transformiert und ins DW geladen wurden, stehen diese für die Abfrage und Analyse bereit. Auf die vorangegangenen Schritte aufbauend ist also entscheidend, anhand welcher Systeme diese Daten abgefragt, ausgewertet, dargestellt und visualisiert werden können. OLAP-Werkzeuge bieten durch OLAP-Cubes eine hierauf speziell zugeschnittene Lösung an. OLAP-Cubes spielen im Rahmen dieser Arbeit als Abfrage- und Darstellungswerkzeug eine wesentliche Rolle und sollen im Folgenden näher erläutert werden.

### 3.4.1. Grundregeln und Definitionen

Allgemein ist Online Analytical Processing eine Datenbanktechnologie, die dazu dienen soll, verschiedensten Anwendern einen schnellen, direkten, interaktiven und verständlichen Zugriff auf Datenbestände zu ermöglichen. Um seiner Informations- und Entscheidungsunterstützungsfunktion gerecht zu werden, ist OLAP auf die Verarbeitung mittlerer bis großer, für die Analyse aufbereiteter Datenmengen ausgerichtet. (Li, 2010, p. 11) Die Abfrage erfolgt durch mehrdimensionale Werkzeuge, die im Zusammenhang mit den DW-Technologien des BI in den 1990er-Jahren Bekanntschaft errungen. Der Begriff des OLAP wurde insbesondere durch Veröffentlichungen von Codd im Jahr 1993 geprägt, indem er 12 Grundregeln für eine adäquate, mehrdimensionale Analyse formulierte und diese im Jahr 1995 auf 18 erweiterte. Diese besagen unter anderem eine mehrdimensionale konzeptionelle Sichtweise auf die Daten, Zugriffsmöglichkeiten für mehrere Anwender, eine getrennte Speicherung von OLAP- und Basisdaten, eine Unterscheidung zwischen Nullen und fehlenden Werten, stabile Antwortzeiten und unbegrenzte Dimensionen und Verdichtungsstufen. (Li, 2010, pp. 12-13) Aufgrund des Umfangs und der Komplexität dieser Regeln sowie dem Vorwurf, diese seien zu sehr auf bestimmte Produkte zugeschnitten, wurde daher ferner die weitaus griffigere Definition „FASMI“ entwickelt. Auf diese soll hier kurz näher eingegangen werden.

„FASMI“ ist eine Abkürzung für „Fast Analysis of Shared Multidimensional Information“. Damit sind die wesentlichen Grundregeln und Anforderungen an OLAP-Werkzeuge klar definiert. „Fast“ schreibt vor, dass einfache Abfragen innerhalb weniger Sekunden bereitgestellt sein sollen. Auch die Antwortzeit bei komplexeren Abfragen sollte sich noch im Sekundenbereich befinden. „Analysis“ besagt, dass ein OLAP-System der reinen Analyse dient und somit ein Anwender in der Lage sein sollte, auch ohne aufwändige Programmierung komplexe Datenabfragen und Analysen durchzuführen. Das System sollte dabei mit jeder Art von Geschäftslogik und statistischer Analyse umgehen können. „Shared“ beschreibt dann die Mehrbenutzerfähigkeit der Systeme. Diese schließt ein, dass das System über entsprechende Zugriffsschutzmechanismen und Sperrmechanismen für den Schreibzugriff verfügt. Mit „Multidimensional“ ist dann die Mehrdimensionalität der Analysen gemeint. Die Multidimensionalität ist das Schlüsselkriterium von OLAP-Datenbanken und folglich auch von OLAP-Werkzeugen. Über die Kombination verschiedener Dimensionen und Hierarchisierungen lässt sich der logische Aufbau von Organisationen und Verfahren sehr gut und realitätsnah darstellen und folglich analysieren. „Information“ beschreibt den primären Verwendungszweck solcher Systeme als Informationssystem und -werkzeug. Um dies zu gewährleisten müssen alle für die Analyse relevanten Informationen vorhanden und abrufbar sein. (Farkisch, 2011, p. 26; Li, 2010, p. 13)

### 3.4.2. Die Grundidee von Online Analytical Processing Cubes

Wie der Name bereits suggeriert, versuchen OLAP-Cubes, bzw. OLAP-Würfel, die Analyseergebnisse in Form eines Würfels nachzubilden. Im Kapitel zu den möglichen Datenbankarchitekturen wurde bereits näher erörtert, dass multidimensionale Analysen aus zwei wesentlichen Faktoren – Fakten und Dimensionswerten – bestehen. Ähnlich wie beispielsweise das Sternschema nutzen OLAP-Würfel daher ein jedermann bekanntes Objekt, das dazu dienen soll, die Fakten und Kennzahlen mit verschiedenen Dimensionswerten zu verknüpfen und hinsichtlich ebendieser Dimensionswerte grafisch darzustellen. Im Fall von Eindimensionalität existiert also nur eine Dimension, über die die Fakten analysiert werden. Die einzelnen Fakten sind dementsprechend als einfacher Vektor, einfaches Array, bzw. einfache Reihe darstellbar. Ein bekanntes Beispiel hierfür wäre eine klassische Zeitreihe, die jedem Wert einen bestimmten Zeitpunkt zuordnet. Im zweidimensionalen Fall gibt es zwei Dimensionen mit entsprechenden Dimensionswerten, sodass sich bei der Darstellung der Fakten über die Dimensionen hinweg nun eine Matrix bzw. eine Tabelle, deren Achsen den verschiedenen Dimensionswerten entsprechen, ergibt. Ab dem dreidimensionalen Fall resultiert dann die erwähnte Würfelform. Im dreidimensionalen Fall entsteht so ein typischer Würfel, während allgemein bei n Dimensionen ein n-dimensionaler Würfel entsteht. Solche Würfel sind menschlich allerdings nur bis zur dritten Dimension vorstellbar.

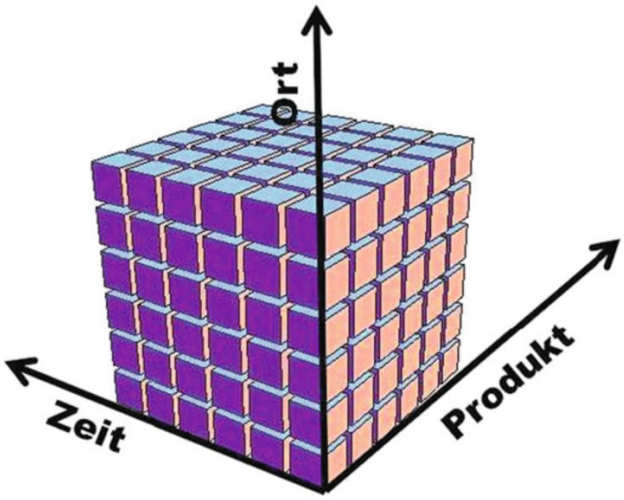



Abbildung 12: OLAP-Cube (Farkisch, 2011, p. 14)

Anhand dieser Herleitung der Verwendung einer Würfelstruktur wird bereits der Vorteil und Sinn dieser Darstellung deutlich. Durch den Würfel soll die multidimensionale Datenstruktur zum Ausdruck gebracht werden. Die Achsen des Würfels stellen die verschiedenen Dimensionen dar und enthalten die Dimensionselemente. (Li, 2010, p. 17) Im dreidimensionalen Fall stellen also beispielsweise alle Kanten in Richtung der X-Achse Dimension 1 und die Kanten in Richtung der Y- und Z-Achse Dimension 2, Respektive Dimension 3 dar. Die entsprechenden Dimensionswerte bzw. -elemente ordnen sich also entlang dieser Achsen, dem Koordinatensystem des Würfels, an, während der Würfel selbst dann die Fakten und Kennzahlen beinhaltet. Ferner lassen viele Dimensionen sich weiter in Hierarchiestufen ordnen. Beispielsweise kann eine geographische Dimension aus den drei

Hierarchieebenen „Kontinent – Land – Stadt“ bestehen, während in der zeitlichen Dimension die Hierarchieebenen „Jahr – Monat – Tag – Zeitpunkt“ denkbar wären. Diese Hierarchieebenen werden auch als Ausprägungen der Dimension bezeichnet und tragen zur Strukturierung des Datenmodells bei. Folglich werden Dimensionsdaten auch als qualitative Daten bezeichnet. (Li, 2010, p. 16) Die verschiedenen Hierarchieebenen führen zu einer Baumstruktur der Dimensionswerte, bei welcher der Knotenpunkt der obersten Hierarchieebene alle Dimensionswerte der Dimension enthält. Ein wichtiger Vorteil der Darstellung der Daten als Würfel liegt in der guten Möglichkeit, den Datensatz hinsichtlich verschiedener Dimensionskombinationen zu untersuchen und zu explorieren. So sind unterschiedliche Sichten auf die Unternehmensdaten möglich. (Li, 2010, p. 17)

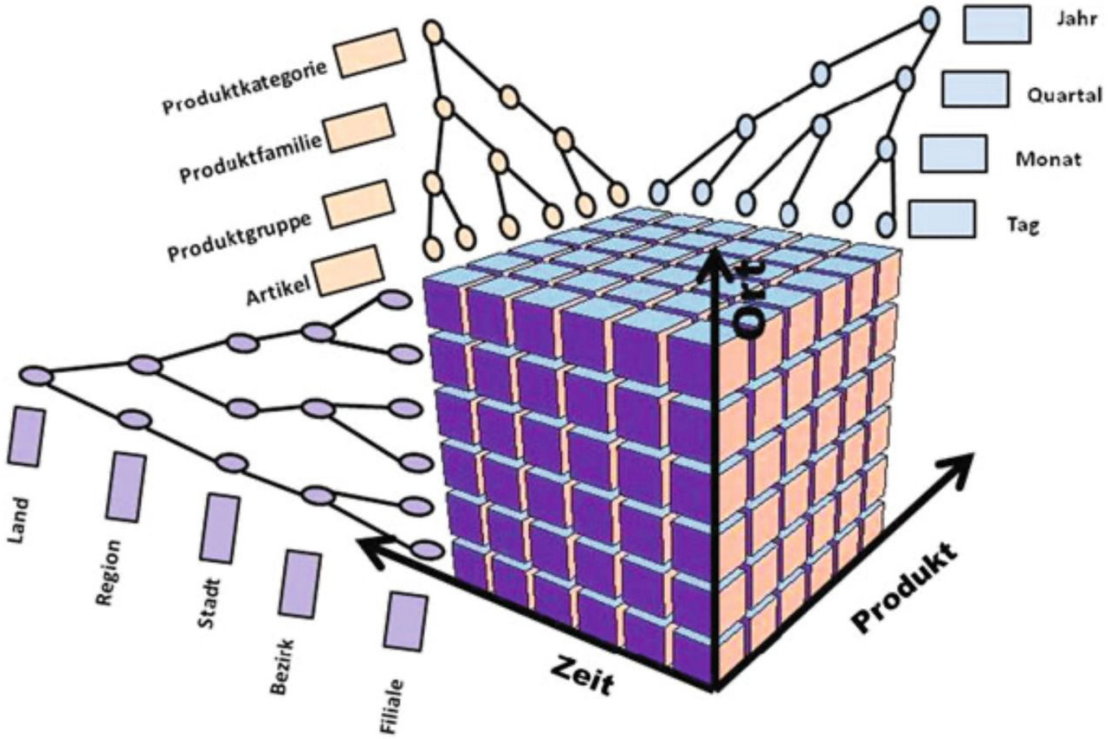



Abbildung 13: OLAP-Cube mit Hierarchieebenen (Farkisch, 2011, p. 18)

Legt man einen bestimmten Dimensionswert für alle Achsen fest, ergibt sich eine genau bestimmte Zelle innerhalb des Würfels, die dann den entsprechenden Wert, den Fakt, enthält. Eine solche Zelle entspricht also theoretisch ebenfalls einem Würfel, wodurch sich letztendlich der gesamte Würfel aus mehreren kleineren Würfeln zusammensetzt, die sich widerherum aus noch kleineren Würfeln zusammensetzen, bis die feinste Granularität erreicht ist. Daran wird auch der Unterschied zwischen Fakten und Kennzahlen deutlich. Fakten sind die Informationen, die über verschiedene Dimensionen hinweg analysiert werden sollen, während Kennzahlen im Zuge des ETL-Prozesses angereicherte Fakten darstellen. Fakten sind quantitative Daten, stellen in einem Würfel die tiefste Informationsebene dar und können über alle Dimensionen hinweg aggregiert werden. Werden ebendiese im Zuge des ETL-Prozesses angereichert, indem sie mit entsprechenden Dimensionswerten verrechnet werden, oder im Zuge von Abfragen im DW über festgelegte Dimensionen aggregiert, ergeben sich daraus entsprechende Kennzahlen. Ändern sich die Dimensionskombination oder die

Dimensionswerte, ändert sich also auch die Kennzahl. Zur Aggregation können verschiedene Aggregationsfunktionen wie Summierung, Durchschnitt, Minimum oder Maximum genutzt werden. (Li, 2010, pp. 15-16) Theoretisch wären auch fortgeschrittene Funktionen wie Median, Varianz, Standardabweichung, Schiefe, Kurtosis oder Quantile denkbar. Beinhaltet eine Kennzahl nicht-numerische Werte wie Text- oder Wahrheitswerte, kommen im Allgemeinen meist Auszähl-Funktionen zum Einsatz. (Li, 2010, pp. 15-16) Insgesamt sind durch Einschränkungen und Definitionen der Dimensionswerte im n-dimensionalen Fall $2^n$ Projektionen möglich, um die Daten zu projizieren und zu aggregieren. Im dreidimensionalen Fall wären dies also $2^3 = 8$ Projektionen. Diese ergeben sich aus einer normalen Projektion auf alle drei Achsen, dem Aggregat einer einzelnen, spezifischen Zelle, den drei verschiedenen Projektionen auf eine der drei Ebenen XY, XZ oder YZ durch Spezifizierung eines Dimensionswertes und den drei verschiedenen Projektionen auf eine der Achsen X, Y oder Z durch Spezifizierung von zwei Dimensionswerten. (Li, 2010, pp. 17-19)

### 3.4.3. Wesentliche Funktionen von Online Analytical Processing Cubes

Aus dem zuvor beschriebenen Aufbau und Konzept von OLAP-Cubes werden bereits die wesentlichen Funktionen ersichtlich. Zunächst wird der Würfel anhand einer Abfrage der Daten des DW entsprechend mit Werten gefüllt. Darauf aufbauend gibt es verschiedene Funktionen, die ein Navigieren des Würfels und damit ein Explorieren der Daten ermöglichen. Diese Funktionen werden zur Navigation, aber auch zur Datenauswahl und -anordnung innerhalb von Datenwürfeln und über Datenwürfel hinweg verwendet. (Schlenker, 1998, p. 10) Sämtliche Funktionen bauen auf der multidimensionalen Sichtweise auf die Daten auf und nutzen die verschiedenen Möglichkeiten der Kombination von Dimensionen und der Hierarchiestufen, um dem Anwender ein möglichst einfaches Navigieren und eine gute Untersuchung der Daten zu ermöglichen. Auf diese können Anwender neue Zusammenhänge und Trends entdecken, aus verschiedenen Sichten auf die Daten schauen und diese analysieren sowie an interessanten Punkten näher ins Detail schauen. (Li, 2010, p. 23)

Zu Beginn wären zunächst die häufig Verwendung findenden Funktionen Roll-Up und Drill-Down zu nennen. Diese dienen dazu, innerhalb der Dimensionshierarchien zu navigieren. Beim Drill-Down wird an einer bestimmten Stelle, also einem bestimmten Dimensionswert, näher ins Detail geschaut, indem an dieser Stelle die Granularität erhöht wird. Folglich werden hierbei die Daten verfeinert, da das Aggregationsniveau auf die entsprechend nächsttiefere, detaillierte Verdichtungsstufe gesenkt wird. Im Ergebnis werden die Daten also genauer dargestellt. (Farkisch, 2011, p. 39; Li, 2010, p. 23; Malinowski & Zimanyi, 2008, p. 49) Dies kann beispielsweise angewandt werden, wenn eine bestimmte Stelle einer Dimension verdächtige oder außergewöhnliche Werte aufweist, die detaillierter untersucht werden sollten. Beispielsweise könnten für einen bestimmten Wert eines Monats dadurch stattdessen alle einzelnen Werte der Tage dieses Monats angezeigt werden. Voraussetzung ist aber, dass nicht bereits die niedrigste und feinste Hierarchieebene erreicht ist. Das Roll-Up stellt dann, wie der Name bereits suggeriert, die inverse Operation zu einem Drill-Down dar, da hier die Daten einer Dimension auf eine höhere, weniger fein granulare Hierarchieebene aggregiert werden. In einer Zeitdimension würden also beispielsweise an Stelle aller Werte der einzelnen Tage eines Monats alle diese Werte zu einem einzelnen Wert für den gesamten Monat aggregiert. Es wird also von den detaillierten Daten wieder auf die verdichteten Daten übergegangen und die Granularität verringert

sich. (Farkisch, 2011, pp. 39-40; Li, 2010, p. 23; Malinowski & Zimanyi, 2008, p. 49) Auch hier ist wieder logische Voraussetzung, dass nicht bereits die höchste, am meisten verdichtete Hierarchieebene erreicht ist.

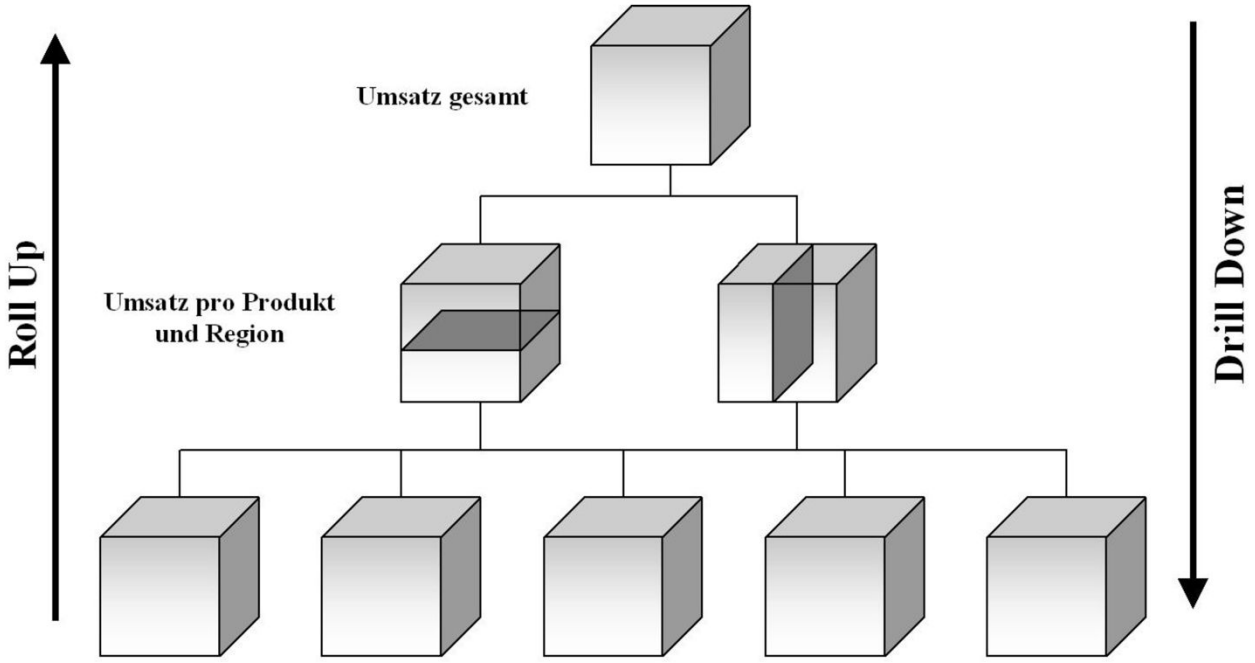



Abbildung 14: Drill-Down- und Roll-Up-Operationen (Li, 2010, p. 24)

Ferner sind zudem die Funktionen Slice und Dice sehr wichtig für die Navigation innerhalb des Würfels. Diese Funktionen stellen Selektionsoperationen dar und ermöglichen durch Auswahl und Beschränkung der Dimensionswerte die Bildung von horizontalen oder vertikalen Ebenen innerhalb des Würfels (Slice) und die Bildung von Subwürfeln bis hin zu einzelnen Zellen (Dice). Als Ergebnis einer Slice-Operation entsteht in einem dreidimensionalen Würfel, wie der Name schon sagt, eine Scheibe, die dem Datenwürfel an einer vorher spezifizierten Stelle entnommen wird. Dies stellt eine der im vorherigen Kapitel erwähnten, möglichen Projektionen der Daten dar. Das Ergebnis ist eine festgelegte Ebene innerhalb des Würfels, die durch Beschränkung einer Dimension auf einen spezifischen Wert erreicht wird. Die Daten liegen dann als zweidimensionale Datenmatrix vor. (Li, 2010, p. 23; Malinowski & Zimanyi, 2008, p. 49) Dice-Operationen hingegen erzeugen einen mehrdimensionalen Ausschnitt des Würfels, also einen Subwürfel. Hierfür werden mehrere Dimensionen auf ein bestimmtes Intervall von Dimensionswerten eingeschränkt. Dieser neu entstehende Datenraum kann dann extrahiert oder weiterverarbeitet werden. Zudem ist auch eine Kombination der Operationen Slice und Dice mit den Operationen Drill-Down oder Roll-Up möglich. (Li, 2010, pp. 23-24)

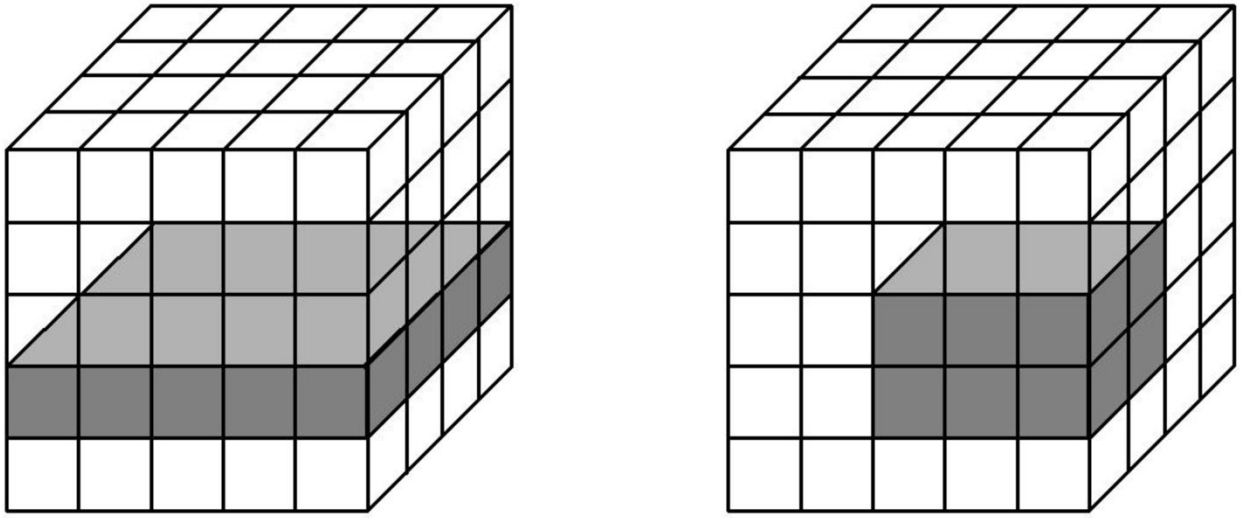


Abbildung 15: Slice- und Dice-Operationen (Li, 2010, p. 24)

Die dritte wesentliche Funktion ist das Pivoting. Unter Pivoting versteht man das Drehen des Würfels durch Änderung der Zuordnung der Dimensionen zu den Achsen des Koordinatensystems des Würfels. Auf diese Weise können die Daten aus einer neuen Sicht heraus betrachtet und analysiert werden. (Li, 2010, p. 24) Dem Anwender wird auf diese Weise mehr Flexibilität bei der Analyse gewährleistet. Pivot wird manchmal synonym auch als Rotate bezeichnet. (Farkisch, 2011, p. 39)

Neben diesen Basisoperationen gibt es weitere, fortschrittliche Operationen, die in Spezialfällen Anwendung finden können. Hierzu gehören insbesondere die Operationen Drill-across und Drill-through. Drill-across kann in dem Spezialfall verwendet werden, dass mehr als ein Datenwürfel betrachtet wird. Dementsprechend beinhalten die Abfragebefehle mehr als eine Faktentabelle und einen Datenwürfel. So können die Kennzahlen des einen Würfels anhand der Dimensionen eines anderen Würfels analysiert werden. Diese Operation erfordert daher, dass die beiden Würfel mindestens eine gemeinsame Dimension zur Verknüpfung aufweisen. (Farkisch, 2011, pp. 40-41; Malinowski & Zimanyi, 2008, p. 49) Die Drill-through Operation erlaubt, von den Daten der niedrigsten, detailliertesten Ebene des Würfels zu den Daten in den operativen Systemen, aus denen der Würfel aufgebaut wurde, zu wechseln. Dies kann beispielsweise erforderlich sein, um Ausreißer in den Daten zu untersuchen. (Malinowski & Zimanyi, 2008, p. 49)

### 3.4.4. Online Analytical Processing Cube Architekturen

Zuvor wurden bereits die Grundregeln, die Grundidee und die typischen Funktionen von OLAP-Würfeln untersucht. Für die tatsächliche Verwendung von OLAP-Würfeln stellt sich nun aber insbesondere die Frage, welche verschiedenen Architekturen und Datenmodelle für OLAP-Würfel existieren. Betrachtet man den gesamten Prozess von OLAP, wird deutlich, dass bereits bei vorherigen Schritten des Prozesses Wahlmöglichkeiten bestanden, nämlich bei der Wahl der Datenbankarchitektur des DW. Hier bestand insbesondere die Wahl zwischen relationalen Datenbankmodellen wie Sternschemata oder Schneeflockenschemata und multidimensionalen Datenbankmodellen. Die Wahl

des OLAP-Cube-Modells baut daher auf der Wahl der Datenbankarchitektur auf. Folglich sollte dieses Kapitel auch in Kombination mit Kapitel 3.2 betrachtet werden. Je nach gewählter Datenbankarchitektur stehen verschiedene OLAP-Cube-Modelle zur Auswahl. Der wesentliche Unterschied besteht bei OLAP-Cube-Modellen in der Wahl von relationalem oder multidimensionalem OLAP. Darüber hinaus gibt es aber auch Mischformen, die als hybride OLAP bezeichnet werden, sowie Spezialformen, wie Desktop-OLAP. Diese sollen im Folgenden näher erläutert werden.

**Relationales OLAP**

Die erste Wahlmöglichkeit besteht in relationalem OLAP (ROLAP). Bei ROLAP werden die Daten, die für die Analyse benötigt und verwendet werden in relationalen Datenbanken, also beispielsweise Sternschemata gespeichert. Die Daten liegen in einer auf Mehrdimensionalität ausgelegten relationalen Datenbank vor und werden dann von dem Abfragewerkzeug, den OLAP-Cubes, abgefragt. Da relationale Datenbanken SQL-Sprachen zur Interaktion verwenden, erfolgt auch die Abfrage der Daten durch SQL-Befehle. (Malinowski & Zimanyi, 2008, p. 49) Die SQL-Anweisungen müssen hinsichtlich der geplanten Analyse und den gewollten Dimensionen und Dimensionseinschränkungen formuliert werden. Diese zur Abfrage der Daten aus der Datenbank notwendigen SQL-Befehle werden zur Laufzeit speziell für die entsprechend angefragte Analyse erzeugt. Die Ergebnisse der Abfrage werden dann als Aggregationstabelle von der Datenbank an das Abfragewerkzeug übergeben, welches daraus den dem Fall entsprechenden OLAP-Würfel aufbaut. (Li, 2010, p. 25) Bei ROLAP-Modellen werden also für jede Abfrage zur Laufzeit die entsprechenden Daten aus dem DW abgefragt und ein entsprechender Würfel erzeugt. Wenn also eine bestimmte Analyse der Datenbestände gefragt ist, wird hierfür ein spezieller OLAP-Würfel erzeugt. Werden dann in Anbetracht der Ergebnisse bestimmte Erkenntnisse klar, die weitere Analysen und typische OLAP-Cube-Funktionen wie einen Drill-Down, ein Vertauschen der Dimensionen oder eine Selektierung durch Festlegung einer Dimension auf einen bestimmten Wert erfordern, wird der SQL-Befehl entsprechend angepasst und diese Daten erneut vom DW abgefragt und aus der ausgegebenen Tabelle ein neuer Würfel aufgebaut. Der wesentliche Vorteil dieses System besteht darin, dass es sehr flexibel ist, eine sehr große Anzahl an Dimensionen und Ausprägungen verarbeiten kann und einfach und schnell zu implementieren und verstehen ist, da keine aufwändigen Vorkalkulationen nötig sind und der Würfel immer für den speziell angefragten Analysefall aufgebaut wird. Der wesentliche Nachteil besteht allerdings darin, dass diese Systeme dem Nutzer keine gleichbleibenden Antwortzeiten gewährleisten können, da je nach Abfrage deutlich mehr Dimensionen verwendet werden und somit auch mehr Tabellen verknüpft und durchsucht werden müssen. (Gutiérrez, 2010, p. 28; Li, 2010, p. 25) Darüber hinaus müssen für jede Anfrage die Daten erneut abgefragt und ein neuer Würfel aufgebaut werden, selbst wenn es dabei um nur kleine Änderungen der Dimensionswerte handelt.

**Multidimensionales OLAP**

Alternativ dazu können auch multidimensionale OLAP-Modelle (MOLAP) verwendet werden. Diese Modelle verwenden einen entgegengesetzten Weg, indem sie alle möglichen Ausgestaltungen des Würfels bereits im Voraus berechnen. MOLAP-Modelle berechnen also bereits vor den eigentlichen Abfragen für Analysen alle möglichen Kennzahlen und Aggregationen über alle möglichen

Dimensionskombinationen und Hierarchieebenen hinweg und erstellen somit praktisch bereits vorher einen alle Werte umfassenden Würfel, aus dem dann bei der Abfrage nur noch ein der Abfrage entsprechender Subwürfel erzeugt werden muss. (Gutiérrez, 2010, p. 27) Die zugrundeliegende Datenbankarchitektur für MOLAP-Modelle ist daher meist ebenfalls eine multidimensionale. (Malinowski & Zimanyi, 2008, p. 49) Wie bereits in Kapitel 3.2 erläutert wurde, speichern solche Datenbankmodelle die Fakten in mehrdimensionalen Arrays, wodurch sich direkt ein n-dimensionaler Würfel ergibt, dessen Werte über die Koordinaten der Achsen, welche die Dimensionswerte beinhalten, eindeutig zugeordnet werden können. Somit erfolgt die eigentliche Erzeugung des Würfels praktisch schon direkt in der Datenbank selbst und bei einer Abfrage müssen lediglich die Dimensionswerte und Hierarchieebenen entsprechend festgelegt werden. Durch diese Umsetzung ist die Abfrageperformance von MOLAP-Modellen unschlagbar performant. (Malinowski & Zimanyi, 2008, p. 50) Beispielsweise sind insbesondere Drill-Down oder Roll-Up Operationen problemlos und praktisch ohne Zeitverzögerung möglich. (Gutiérrez, 2010, p. 27) Dies wirkt sich allerdings negativ auf die Implementierung aus, da letztendlich sämtliche Verzögerung durch Berechnungen von der Laufzeit auf die Aufbauphase vorverlegt werden. Insbesondere bei OLAP-Systemen, die in regelmäßigen Abständen aktualisiert werden, wirkt sich dies negativ aus, da die Systeme nur mit größerem Aufwand skalierbar sind. Schließlich muss bei jeder noch so kleinen Änderung oder Erweiterung von Fakten oder Dimensionswerten des Würfels der gesamte Würfel mit allen Dimensionskombinationen und -beschränkungen neu berechnet werden. (Gutiérrez, 2010, p. 28) Dementsprechend können mit MOLAP-Modellen keine so großen Datenmengen verarbeitet werden, wie mit ROLAP-Modellen. (Malinowski & Zimanyi, 2008, p. 50) Außerdem ist die Auswahl von MOLAP nicht empfehlenswert, wenn der Würfel nur sehr dünn besetzt. In diesen Fällen wird viel Rechenzeit für die Aggregation von Nullwerten verschwendet. (Gutiérrez, 2010, p. 28) Ferner gelten in diesem Zusammenhang auch die bereits in Kapitel 3.2 erwähnten Vor- und Nachteile der zugrundeliegenden Datenbankarchitekturen, in Anbetracht derer relationale Datenbanksysteme etwas mehr Vorteile bieten. Theoretisch sind MOLAP-Modelle aber auch auf Basis von relationalen Datenbankarchitekturen möglich. In diesem Fall wären die Daten im DW beispielsweise in Form eines Sternschemas hinterlegt und das Abfragewerkzeug würde durch Abfrage aller Werte und Kombinationen vor der Laufzeit einen vollständig aggregierten und verdichteten Würfel aufbauen.

**Hybrides OLAP**

Darüber hinaus gibt es hybride OLAP-Modelle (HOLAP), die versuchen, die Vorteile beider Datenmodelle, ROLAP und MOLAP, miteinander zu kombinieren und sowohl die Möglichkeit der multidimensionalen als auch der relationalen Speicherform im Stern- oder Schneeflockenschema nutzen. (Li, 2010, p. 25) In der Praxis werden daher meist häufig abgefragte Dimensionskombinationen im Voraus berechnet, während alle weiteren Abfragen dann zur Laufzeit selbst erzeugt werden. Für hohe Aggregationsebenen wird die multidimensionale Modellierung genutzt, während für feingranulare, detaillierte Abfragen das ROLAP-Modell verwendet wird. In solchen Systemen sollte daher der Großteil der Daten in einem relationalen Datenbankmodell gespeichert werden. Sind die Daten des multidimensionalen Modells nicht ausreichend, um die Abfrage durchzuführen, wird in diesen Fällen auf die detaillierten Daten aus dem relationalen Modell zurückgegriffen. (Gutiérrez, 2010, p. 30; Malinowski & Zimanyi, 2008, p. 50) Unter Beachtung der

Anzahl an Dimensionen kann in diesen Modellen der Füllgrad des Datenwürfels errechnet werden, ab dem eine Speicherung in mehrdimensionalen Arrays effizienter als eine relationale Speicherung ist. (Farkisch, 2011, pp. 36-37) Der Vorteil kann hier darin bestehen, dass in den wichtigen Fällen die sehr gute Abfrageperformance von MOLAP-Modellen verfügbar ist, während die Systeme gleichzeitig flexibel und leicht skalierbar bleiben und große Datenmengen verarbeiten können. (Gutiérrez, 2010, p. 30) Die Nachteile bestehen aber zum Einen darin, dass sich dadurch die Komplexität des Gesamtsystems erhöht, was einer zügigen Implementierung im Weg stehen kann. (Li, 2010, p. 25) Zum Anderen stellt sich außerdem die Frage, wie genau festgelegt werden soll, welche Dimensionskombinationen und Ausgestaltungen des Würfels relevant sind, bzw. relevant genug sind, um diese im Voraus zu berechnen.

**Desktop OLAP**

Ferner findet in der Literatur auch die Spezialform des Desktop-OLAP (DOLAP) Erwähnung. DOLAP-Systeme benötigen und verwenden im Gegensatz zu allen anderen OLAP-Modellen kein Server-Backend, um OLAP-Abfragen durchzuführen, sondern laden stattdessen alle für die Analyse notwendigen Daten aus den zur Verfügung stehenden Datenquellen auf den Client und bereiten diese dort multidimensional auf. (Farkisch, 2011, p. 37) Der wesentliche Nachteil solcher Systeme besteht darin, dass die Übertragung einen hohen Netzwerkverkehr verursacht und die Verarbeitung der Daten hohe Anforderungen an die vorhandene Hardware stellt. Zudem können unterschiedliche Konfigurationen der Geräte die Wartbarkeit des Systems stark beeinträchtigen. (Li, 2010, p. 25)

**Gegenüberstellung der verschiedenen Modelle**

| **Kriterium** | **ROLAP** | **MOLAP** | **HOLAP** | **DOLAP** |
|---|---|---|---|---|
| Antwortzeit | Nicht linear | Linear | MDB-Teil linear, sonst nicht linear | |
| Technisch möglicher Aktualisierungs-zyklus | Echtzeit | Zyklisch | Echtzeit | Echtzeit |
| Komplexität | Mittel | Hoch | Hoch | Mittel |
| Datenzugriff | SQL – nur lesend | API – lesend und schreibend | API – lesend und schreibend | Proprietär – nur lesend |
| Resultierender Netzwerkverkehr bei Zugriffen | Mittel | Gering | Mittel | Hoch |

Tabelle 4: Gegenüberstellung der verschiedenen OLAP-Modelle nach (Li, 2010, p. 25)

### 3.4.5. Abgrenzung zu Data Mining

Sobald das DW aufgebaut und die relevanten Datenbestände vollständig in das DW geladen sind, stehen die Daten zur Abfrage und Analyse bereit. Mit OLAP-Cubes wurde bereits ein typisches

Werkzeug für die Abfrage und Analyse dieser Daten vorgestellt. Allerdings gibt es noch weitere Abfragewerkzeuge und Technologien über OLAP hinaus, die auf das DW oder bestimmte Data Marts zugreifen können und die den von den Datenbanken bereitgestellten Daten weiterverarbeiten. Hier wären insbesondere Reporting-Tools und Data Mining zu nennen. Während Reporting sich insbesondere auf die Erstellung von Berichten und die Bereitstellung von Geschäftsdaten für das Management bezieht, können hinsichtlich der Eigenschaften und dem Zweck von Data Mining in Abgrenzung zu OLAP-Analysen Unklarheiten bestehen. Daher sollen hier die beiden Begriffe abgegrenzt und deren Unterschiede näher erläutert werden.

Der Begriff „Data Mining“ ist ein Sammelbegriff für Techniken und Methoden, um aus bestehenden Datenbeständen neue, bis dato unbekannte Informationen zu gewinnen. Die hierfür eingesetzten Technologien umfassen unter anderem Klassifikationsalgorithmen, Cluster-Analysen, Entscheidungsbäume, Faktoranalysen, neuronale Netze und sonstige Algorithmen des maschinellen Lernens. Durch diese mathematischen Verfahren sollen bisher nicht erkannte Zusammenhänge aufgedeckt werden. (Li, 2010, p. 6) Allerdings können auch durch OLAP-Analysen bisher nicht erkannte Zusammenhänge und Trends aufgedeckt werden. Die Frage ist also, inwiefern sich diese beiden Konzepte unterscheiden.

Wie sich herausstellt, weisen OLAP und Data Mining eine Reihe von Unterschieden auf. OLAP-Cubes bieten nicht nur Standardabfragen der Datenbestände an, sondern liefern durch wichtige Funktionen wie Slice und Dice oder Drill-Downs weitere Werkzeuge, um komplexere Anfragen und Analysen zu handhaben und Vergleiche zu erstellen. Data Mining Tools gehen jedoch darüber hinaus, da sie dem Endanwender keine Analyseergebnisse liefern, sondern Ergebnisse, von denen der Endanwender noch nicht wusste, dass er danach sucht. (Li, 2010, p. 6) Der Hauptzweck von OLAP liegt darin, dem Benutzer statistische Belege für vorher aufgestellte Hypothesen zu liefern. Im Rahmen solcher Analysen und Untersuchungen und aufgrund der mehrdimensionalen Anordnung und graphischen Darstellung der Werte können allerdings weitere Zusammenhänge und Entwicklungen erkennbar werden. Der Unterschied des Data Mining besteht hierbei darin, dass Data Mining diese Zusammenhänge selbstständig aus den Daten heraus bereitstellt und diese folglich nicht zufällig im Rahmen sonstiger Analysen aufgedeckt werden. Zudem verwenden OLAP-Werkzeuge multidimensionale Strukturen, um Analysen durchzuführen, während solche multidimensionalen Strukturen für Data Mining nicht zwingend erforderlich sind. (Li, 2010, p. 6)

# 4. Konzept

In Kapitel 2 wurden die Grundlagen von intelligenten Energieverbrauchszählern ausführlich erläutert. Insbesondere die grundlegende Technologie und Funktionsweise, Nutzen und Chancen sowie Schwächen und Risiken der Technologie, die Ausbreitung und gesellschaftliche Akzeptanz, Schnittstellen, das Format und der Inhalt der generierten Daten sowie die bestehenden Integrationsarchitekturen wurden hier untersucht. In Kapitel 3 wurden dann Business Intelligence, Data Warehousing und Online Analytical Processing vorgestellt und untersucht. Hier wurde auch auf die verschiedenen Datenbankarchitekturen, den ETL-Prozess zur Datenintegration, verschiedene OLAP-Modelle und die typischen OLAP-Funktionen eingegangen. Der Grund dieser ausführlichen Recherche liegt in der Bedeutung dieser beiden Bereiche für die Entwicklung eines Konzepts zur Integration und Analyse von Energieverbrauchsdaten durch OLAP-Cubes. Folglich widmet sich dieses Kapitel der Entwicklung eines Konzeptes, mit dem die von Smart Metern gemessenen und anhand des SMGW übermittelten Strom- oder Gasverbrauchsdaten mit Hilfe von Lösungen aus dem BI Bereich ausgewertet werden können. Dementsprechend verbindet dieses Kapitel die beiden vorangegangenen Recherchekapitel und kombiniert die dort erarbeiteten Ergebnisse für eine konkrete, praktische Umsetzung.

## 4.1. Integration und Analyse von Energieverbrauchsdaten durch OLAP

Das Ziel dieser Arbeit ist die Untersuchung der Eignung von Online Analytical Processing Cubes für die Integration und Analyse von Energieverbrauchsdaten. Bisher wurden nach einer Einleitung in das Thema dieser Arbeit vor allem die Grundlagen der Themengebiete Smart Meter und OLAP behandelt. In diesem Teil sollen diese beiden Themengebiete zusammengeführt werden, um aufbauend auf den bisherigen Ergebnissen ein Modell zu entwickeln, das verschiedene Energieverbrauchsdaten sinnvoll integriert und für Analysen bereitstellt. Die Themenbereiche Smart Meter und OLAP existieren bisher weitestgehend nebeneinander und sind ausschließlich in völlig verschiedenen Zusammenhängen und Bereichen von Bedeutung. Im Rahmen dieser Arbeit soll daher geprüft und demonstriert werden, ob die existierenden Lösungen aus dem Bereich des BI, die im Zusammenhang mit der Geschäftsanalytik und Entscheidungsunterstützung in Unternehmen bereits seit mehreren Jahren etabliert sind und großflächig zur Anwendung kommen, auch für die Integration und Analyse von Sensordaten und durch Smart Meter gemessene Energieverbrauchsdaten im Speziellen geeignet sind. Der Energieverbrauch von Immobilien hängt von vielen verschiedenen Faktoren ab und hat somit ähnlich wie Geschäftsdaten eine multidimensionale Struktur. Die Anwendung etablierter analytischer Informationssysteme aus dem Business-Bereich hätte daher das Potential, die typischen Vorteile wie eine multidimensionale Modellierung und Analyse, eine gute Abfrageperformance, eine übersichtliche Struktur, die Möglichkeit, Daten über verschiedene Hierarchiestufen zu aggregieren und gruppieren, und eine vermiedene Belastung der operativen Systeme durch komplexe, rechenintensive Analysen, auch im Zuge von Analysen von Sensor- und Energieverbrauchsdaten zu nutzen. Diese Arbeit hat dabei nicht das Ziel, eine vollständige, kommerziell verwendbare Software zu erschaffen. Stattdessen liegt der Fokus auf einer ersten Untersuchung der Eignung dieser Technologie und einer demonstrativen Umsetzung des Konzepts anhand konkreter Beispiele und Analysen von Hypothesen. Aufbauend auf dieser Arbeit könnten weiterführende Arbeiten sich dann mit einer Automatisierung und Zusammenführung des Gesamtprozesses anhand von Software mit dem Ziel der Erschaffung einer

professionell verwendbaren Anwendung auseinandersetzen. Diese Arbeit hingegen soll durch ein erstes „Proof of Concept“ den Grundstein für eventuelle weiterführende Arbeiten legen.

Das Konzept umfasst sämtliche Schritte von der Messung der Daten bis zur letztendlichen Visualisierung der Ergebnisse. Dementsprechend ergibt sich eine logische Reihenfolge der Prozesse, die notwendig sind, um durch die von Smart Metern erhobenen Daten Analysen durchzuführen und Hypothesen zu prüfen, beginnend mit der Messung der Energieverbräuche und endend mit einer ansprechenden Visualisierung der Daten, auf deren Basis die Ergebnisse interpretiert und die Hypothesen überprüft werden können. Der Gesamtprozess untergliedert sich in die folgenden vier wesentliche Teilprozesse.

1. Auswahl und Extraktion von relevanten Daten aus verschiedenen Datenquellen
2. Integration der Daten in ein Data Warehouse
3. Multidimensionale Abfrage der Daten aus dem Data Warehouse
4. Visualisierung der abgefragten Daten

Jeder dieser Teilprozesse besteht dabei aus weiteren Teilschritten, die durchlaufen werden müssen, um den Teilprozess erfolgreich abzuschließen. Diese Teilschritte der einzelnen Prozesse werden im Folgenden näher erläutert.

**Auswahl und Extraktion von relevanten Daten aus verschiedenen Datenquellen**

Für eine erfolgreiche und aussagekräftige Analyse kommen viele verschiedene Datenquellen zum Einsatz. Die Daten der einzelnen Datenquellen können dabei von der ursprünglichen Erhebung der Daten, über eine eventuelle temporäre Speicherung in diversen Systemen oder eine dauerhafte Integration in spezielle Datenbanken, bis hin zur Auswahl und Extraktion der Daten aus den Datenquellen, teilweise lange Wege zurücklegen. Die extrahierten Daten werden dann für den Aufbau eines DW als Basis für multidimensionale Analysen genutzt und weiterverarbeitet. Die wesentliche Grundlage, der im Zuge dieser Arbeit durchgeführten Analysen, sind Energieverbrauchsdaten, die von intelligenten Energieverbrauchszählern gemessen werden. Diese werden über verschiedene Dimensionen hinweg analysiert. Die verschiedenen Dimensionen weisen ebenfalls verschiedene Werte, auch als Dimensionselemente bezeichnet, auf und können aus verschiedensten Datenquellen stammen.

Zunächst werden die Energieverbrauchsdaten betrachtet. Jede Immobilie benötigt im Zuge einer entsprechenden Bewirtschaftung Energie. Diese Energie wird beispielsweise benötigt, um Räumlichkeiten zu beleuchten, zu beheizen oder zu kühlen oder um das Trinkwasser für Küchen und Bäder zu erwärmen. Ferner kann auch die Konditionierung der Raumluft über die thermische Konditionierung hinaus, beispielsweise durch maschinelle Lüftung oder Be- und Entfeuchtung von Zuluft, einen zusätzlichen Energieverbrauch verursachen. Außerdem benötigen die verschiedenen Haushaltsgeräte Energie, um den Bewohnern ein möglichst angenehmes Dasein zu ermöglichen. Hierzu zählen insbesondere die typischen Küchen-, Kommunikations- und Unterhaltungsgeräte wie Backöfen, Kochstellen, Computer oder Fernseher, aber auch Waschmaschinen, Trockner und etwaige Wasserpumpen. Diese Energie kann auf verschiedene Weisen zur Verfügung gestellt werden. Der

überwiegende Teil aller Haushaltsgeräte deckt den Energiebedarf durch Elektrizität, während sich in allen Bereichen der Wärmebereitstellung bisher meist die Umwandlung fossiler Brennstoffe wie Erdgas zu Wärme durch Verbrennung etabliert hat. Auch wenn die absolute Anzahl an Geräten, die mit Strom versorgt werden, deutlich höher ist als die Anzahl der mit Erdgas versorgten Geräte, so hat nichtsdestotrotz insbesondere in Deutschland der Gasverbrauch meist einen deutlich höheren Anteil am gesamten Endenergieverbrauch eines Gebäudes. Dies ist insbesondere darauf zurückzuführen, dass durch die klimatischen Verhältnisse Raumheizung deutlich häufiger benötigt wird als Raumkühlung. Folglich sind die wesentlichen relevanten Energieverbräuche im Gebäudesektor der Gasverbrauch und der Elektrizitätsverbrauch.

Werden Gebäude dabei mit intelligenten Energieverbrauchszählern ausgestattet, so messen diese zunächst je nach Zähler mechanisch, elektromechanisch oder elektronisch den entsprechenden Energieverbrauch. Dieser wird dann über das LMN an das SMGW übermittelt. Das SMGW kommuniziert diese dann über das WAN an das entsprechende Unternehmen. Im Fall von Elektrizität ist dies in der Regel der zuständige MSB, dessen Rolle von Verteilnetzbetreibern, EVUs oder Dritten übernommen werden kann. Im Fall von Gas ist dies der Gaslieferant. Die so übermittelten Energieverbrauchsdaten werden dann unter anderem zu Abrechnungszwecken in das MDMS des Energieunternehmens integriert und dort gespeichert. Diese MDMS stellen in gewisser Weise das Äquivalent zu den typischen transaktionalen Datenbanken von Unternehmen dar, die im klassischen BI und DW von Relevanz sind. Diese Systeme sind auf das effiziente Management operativer Daten abgestimmt und bieten sowohl schreibenden als auch lesenden Zugriff. Auch wenn diese OLTP-Systeme in einer typischen BI-Implementierung in aller Regel dafür da sind, Geschäftsdaten mit finanziellem Bezug, also Käufe, Bestellungen, Lieferungen, Lagerbestände oder Kundeninformationen, zu verwalten, so weisen sie dennoch viele Parallelen zu den MDMS von Energieunternehmen auf. Da diese unter anderem insbesondere im Zuge von Abrechnungen von Relevanz sind und Energie ebenfalls ein Gut ist, das finanziellen Wert hat, ist auch hier ein finanzieller Bezug gegeben. Noch deutlich wichtiger ist aber darüber hinaus die Tatsache, dass auch diese Systeme operative Daten integrieren und verwalten. Um diese Daten anhand von OLAP-Cubes zu analysieren, müssen diese daher zunächst aus den OLTP-Systemen extrahiert werden. Anschließend können sie dann in ein separates DW integriert werden, dass auf Multidimensionalität ausgelegt ist und ausschließlich lesenden Zugriff auf die Daten für Analysezwecke bietet.

Darüber hinaus müssen für multidimensionale Analysen Daten bereitgestellt werden, die die Dimensionen widerspiegeln, hinsichtlich derer die Energieverbrauchsdaten analysiert werden sollen. Diese können je nach zu prüfender Hypothese und Analysezweck sehr unterschiedlicher Natur sein. Grundsätzlich untergliedern diese Daten sich in interne und externe Daten, wobei intern die Unternehmenssicht beschreibt. Interne Daten sind daher weitere Daten, die in den Systemen und Datenbanken des Energieunternehmens selbst vorliegen. Je nach Analysezweck können sehr unterschiedliche Daten von Interesse sein. Allerdings wäre denkbar, dass in vielen Fällen besonders Kundendaten und sonstige Daten, die für den operativen Betrieb und die Energielieferung von Relevanz sind, auch für Analysen von hohem Wert wären. Die Kundendaten könnten beispielsweise den Namen und Familienstand des Kunden, sowie das jährliche Einkommen und die Anzahl der Haushaltsmitglieder beinhalten, während ebenfalls denkbar wäre, dass weitere Daten über das

Gebäude oder den Haushalt selbst vorliegen. Diese könnten zum Beispiel das Baujahr, die Energieeffizienzklasse auf dem Energieausweis oder die Nutzfläche beinhalten. Alle diese Daten sind häufig Metadaten, also beschreibende Daten, und werden daher meist separat vorgehalten. In Unternehmen existieren daher meist weitere Metadatenbanken. Auch diese können als Datenquellen für OLAP-Analysen dienen. In diesem Fall müssten auch die Metadaten zunächst aus den Metadatenbanken des Unternehmens extrahiert werden.

Die zweite Art von Dimensionsdaten sind externe Daten. Wie der Name bereits suggeriert sind dies Daten von außerhalb des Unternehmens, die nicht aus den Systemen des Energieunternehmens stammen. Solange diese einen Bezug zum Energieverbrauch von Gebäuden haben oder haben könnten, sind hierfür praktisch alle möglichen Datenquellen denkbar. Welche Datenquellen hier bestehen und für Analysen relevant sein könnten wird in 4.1.2 diskutiert. Grundsätzlich lässt sich der erste Teilprozess daher in Auswahl und Extraktion der Energieverbrauchsdaten, also der Fakten, aus den MDMS des Unternehmens, der internen beschreibenden Daten als Dimensionsdaten aus den Metadatenbanken des Unternehmens und der externen beschreibenden Daten als Dimensionsdaten aus unternehmensexternen Datenquellen zusammenfassen.

Daraus lässt sich bereits die Relevanz dieser Arbeit für die allgemeine Integration und Analyse von Sensordaten ableiten. Letztendlich stellen Smart Meter Daten Sensordaten dar, da die Messeinrichtungen den Energieverbrauch über verschiedene Sensoren messen. Allerdings können auch weitere Sensordaten zur Analyse der Energieverbrauchsdaten zur Anwendung kommen, indem die Energieverbrauchsdaten hinsichtlich dieser Sensordaten analysiert werden. Die weiteren Sensordaten würden dann Dimensionen darstellen. Ein typisches Beispiel hierfür wären beispielsweise Temperatursensordaten. Folglich können Sensordaten sowohl selbst analysiert werden als auch Analysen unterstützen und können daher sowohl in die Faktentabellen als auch in die Dimensionstabellen einfließen.

**Integration der Daten in ein Data Warehouse**

Sind die Daten ausgewählt und aus den bestehenden Systemen extrahiert, können diese in ein DW integriert werden. Hierbei stellt sich zunächst die Frage, ob bereits ein DW existiert, das um diese neuen Daten erweitert, also skaliert, oder ob ein neues DW aufgebaut werden soll. Da der zweite Fall komplexer ist, wird dieser hier dargestellt. Der erste Fall ist ohnehin durch die letzten Schritte des zweiten Falls abgedeckt. Zunächst muss für den Aufbau eines DW eine geeignete Datenbankstruktur gewählt werden. Die verschiedenen zur Auswahl stehenden Datenbankarchitekturen wurden bereits in Kapitel 3 ausführlich vorgestellt. Anschließend hängt das weitere Vorgehen davon ab, ob ein neues DW aufgebaut oder ein bestehendes aktualisiert bzw. erweitert wird. Letztendlich gibt es zwar typische, etablierte Vorgehensweisen bei der Integration der Daten in das DW, allerdings sind diese kein Zwang und es kann je nach Situation, Verwendungszweck und bestehenden, zur Unterstützung bereitstehenden Software-Lösungen unterschiedlich vorgegangen werden. Wichtig ist im Ergebnis nur, dass die Daten am Ende sauber in das DW integriert sind. Der wesentliche bestimmende Faktor bei der Integration der Daten in das DW ist neben der Wahl der Datenbankstruktur der ETL-Prozess. In dem Fall, dass das OLAP-System von einem Energieunternehmen selbst verwendet wird, beispielsweise um

das operative Geschäft mit Hilfe analysierter historischer Daten zu optimieren, und das DW in festgelegten, regelmäßigen Abständen um neue Daten erweitert und skaliert wird, empfiehlt es sich dringend, diesen ETL-Prozess durch ein Software-Tool zu automatisieren bzw. zumindest zu vereinfachen. Wird das DW für eine sehr spezielle, einmalige Analyse aufgebaut und sind die Daten sehr heterogen, kann die Integration auch manuell erfolgen. Die Vorgehensweise bei ETL-Prozessen wurde bereits in Kapitel 3 erläutert. Wie konkret bei der demonstrativen, praktischen Umsetzung im Rahmen dieser Arbeit vorgegangen wurde, wird in Kapitel 4 im Zuge der Analyse der Hypothesen dargestellt.

**Multidimensionale Abfrage der Daten aus dem Data Warehouse**

Sobald die Daten aus den Datenquellen extrahiert und erfolgreich in das DW integriert wurden, stehen die Daten für Analysezwecke und Abfragen bereit. Über das DBMS der verwendeten Datenbank oder über entsprechende Software-Tools mit Zugriff auf die Datenbank können die Daten dann multidimensional abgefragt werden. Dabei können, wie bei OLAP-Modellen typisch, sämtliche, für die Analyse relevanten Dimensionen ausgewählt und verwendet und die entsprechenden Dimensionswerte speziell auf die Fragestellung abgestimmt werden. Ferner bietet die hierarchische Struktur der Dimensionen die Möglichkeit, Daten auf unterschiedlichen Hierarchieebenen zu aggregieren und die Aggregationen bezüglich bestimmter Hierarchieebenen zu gruppieren. Diese Hierarchisierungsmöglichkeiten und die damit verbundenen Chancen, die Daten zu explorieren und mehr oder detailliert zu betrachten, stellen einen wesentlichen Vorteil von OLAP-Systemen gegenüber anderen Systemen dar. Das Ergebnis der Abfrage ist dann die Ausgabe einer Tabelle mit den aggregierten Fakten, die hinsichtlich der bei der Abfrage festgelegten Dimensionswerten und Hierarchieebenen gruppiert sind. Diese stehen dann für die Visualisierung bereit.

**Visualisierung der abgefragten Daten**

Nachdem das DBMS des DW die Abfrage erfolgreich durchgeführt hat, indem es alle Werte berechnet, aggregiert und als Tabelle ausgegeben hat, können diese mit Hilfe geeigneter Software visualisiert werden. Theoretisch ist es bereits möglich, Erkenntnisse aus der ausgegebenen Tabelle zu gewinnen, indem die verschiedenen Werte der Tabelle abgelesen und verglichen werden. Dies ist aber nicht nur mühsam und fehleranfällig, sondern auch wenig ansprechend und nicht aussagekräftig. Visualisierungen sind für einen erfolgreichen Abschluss des Analyseprozesses, eine angemessene Dokumentation der Analyseergebnisse und das Verständnis der ausgegebenen Werte deshalb unverzichtbar.

Je nach Anwendung kommen hierfür verschiedene Graphen und Diagramme in Frage. Um die Daten als Graphen und Diagramme visualisieren zu können, sind meistens noch weitere Formatierungen der Daten notwendig. Außerdem stellt sich die Frage, wie mehrdimensionale Datensätze anhand von Diagrammen und Graphen dargestellt werden können. Diagramme und Graphen können theoretisch nur eindimensionale Daten darstellen, also einen Vektor oder Array, da sie die abhängige Variable auf der Y-Achse in Abhängigkeit von der unabhängigen Variablen auf der X-Achse darstellen. Ein klassisches Beispiel wäre hier die Darstellung des Energieverbrauchs über die Zeit, wobei der

Energieverbrauch in kWh auf der Y-Achse und die Zeit auf der X-Achse dargestellt wird. Diese Beschränkung auf eindimensionale Datensätze – eindimensional meint in diesem Fall die Existenz von nur einer unabhängigen Variablen, die die abhängige Variable beschreibt, also nur eine einzige Dimension, hinsichtlich der die Fakten dargestellt werden – gilt ebenfalls für alle weiteren Darstellungsformen wie Säulen- oder Kuchendiagramme, die Werte oder Anteile bestimmter Dimensionswerte darstellen.

Diese Beschränkung auf eine Dimension durch die Beschränkung auf X- und Y-Achse im zweidimensionalen Fall, da die Fakten bereits die Y-Achse belegen, lässt sich allerdings umgehen. Eine Abhilfe schafft hier die Möglichkeit, mehrere Graphen im gleichen Koordinatensystem darzustellen, die alle einen bestimmten Dimensionswert repräsentieren. Auf diese Weise wird die erste Dimension durch die Werte der X-Achse widergegeben, während die zweite Dimension durch die Legende, welche die verschiedenen Graphen beschreibt, dargestellt wird. Dies setzt allerdings in der Regel voraus, dass die zweite Dimension überschaubar viele Werte aufweist. Werden beispielsweise mehrere Dutzend Graphen in das gleiche Koordinatensystem gezeichnet, führt dies zu großer Unübersichtlichkeit und verhindert eine einfache Interpretation. Eine Lösung dieses Problems ist die Darstellung der Linien der Graphen, indem diese „aufeinandergestapelt“ werden, in Form eines sogenannten „stacked plot“. Dies ermöglicht eine eindeutigere Darstellung der verschiedenen Linien, da es nicht zu Überschneidungen kommen kann. Diese Art der Darstellung kann für die Visualisierung von Energieverbrauchsdaten besonders geeignet sein, da beispielsweise die verschiedenen Energieverbräuche einzelner Sektoren in Summe ohnehin den Gesamtenergieverbrauch ergeben und durch Stapelung der Linien übereinander auf der Y-Achse auch gleich der Gesamtverbrauch ablesbar ist. Ein Nachteil dieser Darstellungsform besteht allerdings darin, dass durch die Stapelung die absoluten Werte jeder Linie oberhalb der untersten nur durch Umrechnung ablesbar sind, da für den tatsächlichen Wert einer bestimmten Linie die Differenz aus dem Wert der relevanten Linie und dem der darunterliegenden Linie gebildet werden muss. Ferner stößt auch diese Darstellungsform bei einer sehr großen Anzahl an Dimensionswerten an ihre Grenzen.

Aufbauend darauf kann dann eine dritte Dimension dargestellt werden, indem schlicht weg für jeden Dimensionswert der dritten Dimension ein eigener Graph gebildet wird. Im Zuge können sogenannte „Subplots“ gebildet werden, die untereinander oder nebeneinander angeordnet werden. Allerdings gilt auch hier die Einschränkung, dass es nicht zu viele Dimensionswerte geben sollte, da für jeden Dimensionswert ein einzelner Graph gebildet wird. Allerdings besteht natürlich die Möglichkeit der Selektion, indem nur bestimmte, besonders aussagekräftige Graphen ausgewählt und dargestellt werden. Insgesamt sind durch das zweidimensionale Koordinatensystem, die Stapelung der Linien unter Verwendung einer Legende und das Erstellen von Subplots also drei verschiedene Dimensionen darstellbar.

Der Gesamtprozess des erarbeiteten Konzeptes inklusive aller Teilprozesse ist im Folgenden abschließend grafisch dargestellt. Dabei wird insbesondere die Verbindung der beiden Rechercheteile deutlich, da erstmals sowohl die Ergebnisse des ersten und des zweiten Rechercheteils gemeinsam und in Kombination Anwendung finden. Die in den Rechercheteilen geschaffene Wissensbasis zahlt sich bei Erarbeitung des Konzeptes besonders aus. Wie dargestellt werden die gemessenen

Energieverbrauchsdaten über die definierten Kommunikationsschnittstellen für den Datentransfer an die bestehenden MDMS der Energieunternehmen übermittelt. Diese haben dabei das im zweiten Kapitel erläuterte Format. Bis dorthin handelt es sich um bereits bestehende, etablierte Prozesse. Ab diesem Punkt greift dann das erarbeitete Konzept und die Teilprozesse, die im Rahmen dieses Konzeptes vorgestellt wurden, werden durchlaufen. Folglich ist ab dort das Grundwissen über OLAP-Cubes und BI von großer Relevanz. Das Gesamtkonzept ist in der folgenden Abbildung dargestellt.

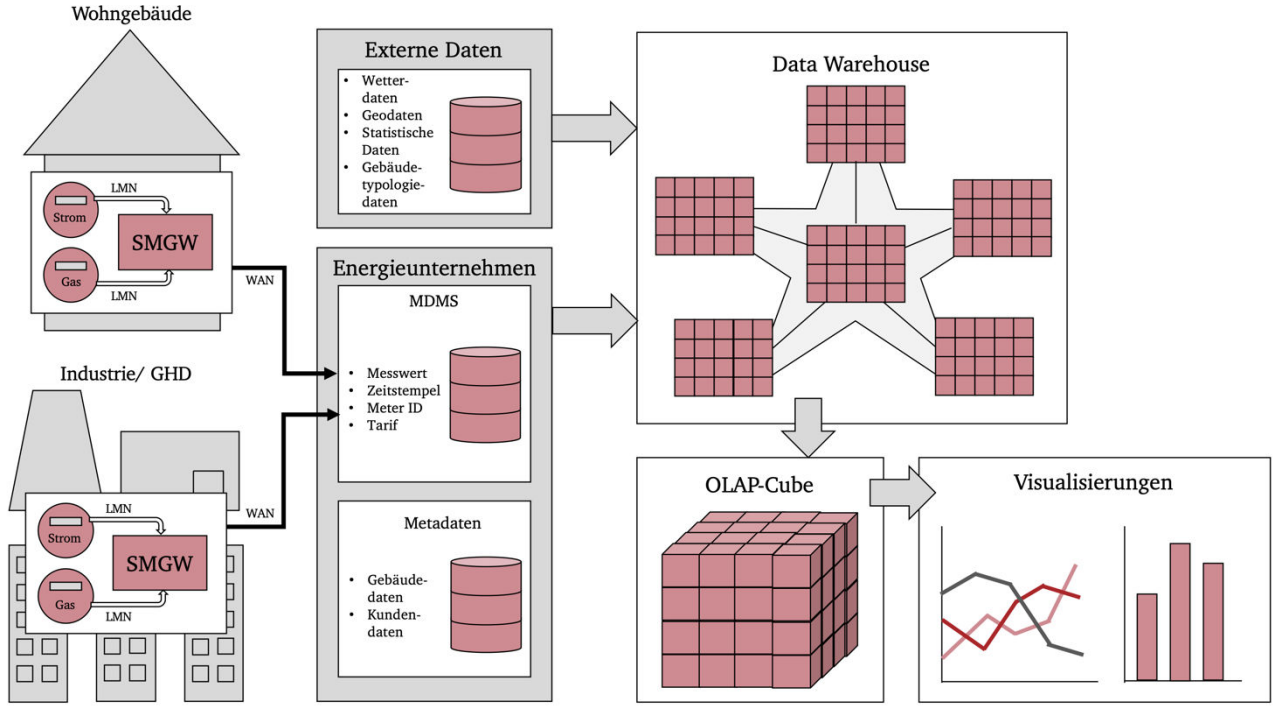



Abbildung 16: Zusammenfassung des Konzepts

### 4.1.1. Verwandte Arbeiten

In diesem Zusammenhang ist besonders interessant, ob es bereits ähnliche oder vergleichbare Arbeiten in diesem Themengebiet gibt. Zum Zeitpunkt der Anfertigung der Arbeit ist dem Autor allerdings trotz ausführlicher Recherche bisher keine vergleichbare Arbeit bekannt, die ein Konzept zur Integration und Analyse von Energieverbrauchsdaten durch OLAP-Cubes erarbeitet hat. In diesem Sinne dürfte diese Arbeit die Erste auf diesem Themengebiet sein. Allerdings existieren verschiedene Arbeiten, die Ähnlichkeiten oder Verbindungen mit diesem Thema aufweisen. Dementsprechend sind teilweise Vergleiche der Konzepte und der Umsetzung möglich. Diese Arbeiten sollen hier kurz näher vorgestellt werden.

Dobson et al. stellen eine Referenzarchitektur und ein Modell für die Integration und Verwaltung von Sensordaten vor, das die Daten möglichst lange aufbewahren und für Abfragen und Analysen bereithalten soll, um der Frage nach einer geeigneten Integration rasant wachsenden Menge an Sensordaten im Rahmen des IoT gerecht zu werden. (Dobson, Golfarelli, Graziani, & Rizzi, 2018) Dafür nutzen sie die bekannten Technologien und Architekturen aus dem Bereich des BI und

verwenden sternförmige Datenbanken als Grundlage des DW, um die mehrdimensionalen Daten dann durch OLAP abzurufen. Nachdem die wesentlichen Grundlagen von BI, DW und OLAP vorgestellt wurden, werden unter anderem bestehende Software-Lösungen erwähnt. Anschließend wird die vorgeschlagene Referenzarchitektur vorgestellt. Obwohl der etablierte Prozess der Integration der Daten in das DW erwähnt wird, wird nicht auf mögliche Visualisierungen der Daten eingegangen. Auch das konkrete Vorgehen beim Aufbau des DW wird nicht behandelt. Stattdessen liegt der Fokus auf einer guten Konzeption und Beschreibung der Verbindungen zwischen den verwendeten Sensoren, den Anwendern und Stakeholdern und der Kommunikation zwischen den verschiedenen Objekten des Systems. Hauptziel dieses Konzepts ist die Verbindung der im DW gespeicherten historischen Sensordaten mit Echtzeitsensordaten. Je nach Situation und Anwendung sollen die Daten aus dem DW in Kombination mit Echtzeitdaten verwendet werden und so weitere Informationen für Analysen und die Steuerung liefern. Ferner soll das DW zur Langzeitintegration und -speicherung der Sensordaten dienen. Demonstrativ werden dann zwei Fallbeispiele behandelt, in denen die Referenzarchitektur Anwendung finden könnte. Der erste Fall befasst sich mit der Überwachung der Luftqualität durch Sensorsysteme, der zweite Fall hingegen mit dem Risikomanagement von Erdrutschen durch Sensorsysteme. (Dobson et al., 2018)

Auch in einer weiteren Arbeit von Soewito et al. wird ein OLAP-Modell im technischen Bereich angewandt, um Wasserbildungsdaten und potentiell auch Daten von Öl- und Gasquellen zu analysieren. (Soewito, Isa, & Gunawan, 2018) Im Zuge der Untersuchungen werden die gemessenen Daten aus Wasserquellen und Bohrungen in ein DW integriert. Hierbei werden die räumlichen Informationen in eine hierarchische Form mit drei Hierarchieebenen überführt. Die höchsten Hierarchieebene stellen die Becken bzw. Täler dar, die mittlere Hierarchieebene die einzelnen Bohrfelder darin und die unterste Hierarchieebene die einzelnen Bohrlöcher. Aufbauend auf dem DW können die Daten dann durch OLAP-Cubes abgefragt werden, die insbesondere durch Drill-Down und Roll-Up Operationen eine gute Analyse ermöglichen. Aufbauend darauf kommen dann auch Data Mining Technologien zum Einsatz und die Daten werden entsprechend visualisiert. Der ETL-Prozess zur Integration der Rohdaten in das DW wird anhand des etablierten Open-Source Software-Tools Pentaho durchgeführt und wird nicht näher erläutert. Das OLAP-Modell selbst weist eine sehr einfache Struktur auf, da es ausschließlich eine Dimensionstabelle, die räumliche Dimension, enthält. (Soewito et al., 2018)

Erwähnung findet BI aber auch in tatsächlichem Bezug zu Smart Meter Daten. Mikkelsen et al. haben im Rahmen einer Arbeit ein Konzept eines MDMS entworfen, das als Web Service umgesetzt wird. (Mikkelsen et al., 2016) Dieses soll die verschiedenen benötigten Funktionen eines MDMS beinhalten und insbesondere die verschiedenen Datenlatenzen von Smart Meter Daten berücksichtigten. Dementsprechend soll es sowohl Echtzeitanalysen von Netzzustandsdaten als auch Analysen von operativen und historischen Daten für Abrechnungszwecke und den Netzbetrieb ermöglichen. Im Zuge der Darstellung der verschiedenen Stufen der Datenlatenz findet hier auch BI Erwähnung. Allerdings spielt BI für das entwickelte Konzept der Arbeit keine Rolle. Stattdessen erfolgt eine klare Trennung zwischen technischen Daten und Geschäftsdaten, also finanziellen Daten, und BI findet ausschließlich im letzteren, also klassischen Fall, Anwendung. Folglich würde BI in diesem Fall zwar von Energieunternehmen eingesetzt, allerdings nur zur klassischen Analyse der Geschäftszahlen und nicht

auch für technische Analysen der Smart Meter Daten. (Mikkelsen et al., 2016, p. 6) Dementsprechend wird von Mikkelsen et al. auch kein Konzept zum Aufbau eines Data Warehouse oder zur Datenmigration erarbeitet.

Das einzige bekannte Beispiel, in dem OLAP-Systeme in Bezug auf Smart Meter Daten Anwendung finden, stammt von Nagesh et al. Diese vierseitige Arbeit stellt eine Architektur für ein Energiemanagementsystem vor, das der Laststeuerung und -verlagerung dienen soll, indem es zu Spitzenlastzeiten den Stromverbrauch von Haushaltsgeräten reduziert. (Nagesh, Krishna, & Tulasiram, 2010) Damit sollen unter anderem günstigere Energiekosten realisiert werden. Dafür werden zunächst die verschiedenen Komponenten der Architektur erläutert und dann das Konzept anhand einer Fallstudie demonstriert. Das Konzept verwendet dabei sowohl OLTP- als auch OLAP-Systeme. Die OLTP-Systeme sind dabei für die eigentliche Laststeuerung und die sonstigen Hauptaufgaben des Energiemanagements zuständig, während die OLAP-Systeme genutzt werden, um den Endverbrauchern detaillierte Informationen über den Energieverbrauch bestimmter Haushaltsgeräte zu liefern. (Nagesh et al., 2010) Allerdings erfolgt in dieser Arbeit keine konkrete Umsetzung. Darüber hinaus wird nicht auf das Konzept und die Technologie hinter OLAP eingegangen. Insbesondere werden keine Datenbankmodelle vorgestellt oder ausgewählt. Die Verwendung von OLAP-Systemen bezieht sich lediglich auf die Bereitstellung der Energieverbräuche der Haushaltsgeräte. OLAP wird dabei lediglich aufgrund der Fähigkeit, große Datenmengen zu verarbeiten, verwendet. Die typischen OLAP-Funktionen und Vorteile wie die multidimensionale Modellierung, Hierarchieebenen, verschiedenste Aggregationen und OLAP-Cube-Operationen bleiben somit weitgehend unbeachtet und ungenutzt. Die Architektur des Systems ist im Folgenden graphisch dargestellt. Das Konzept dieser Arbeit wurde jedoch unabhängig von der von Nagesh et al. vorgeschlagenen Architektur entwickelt.

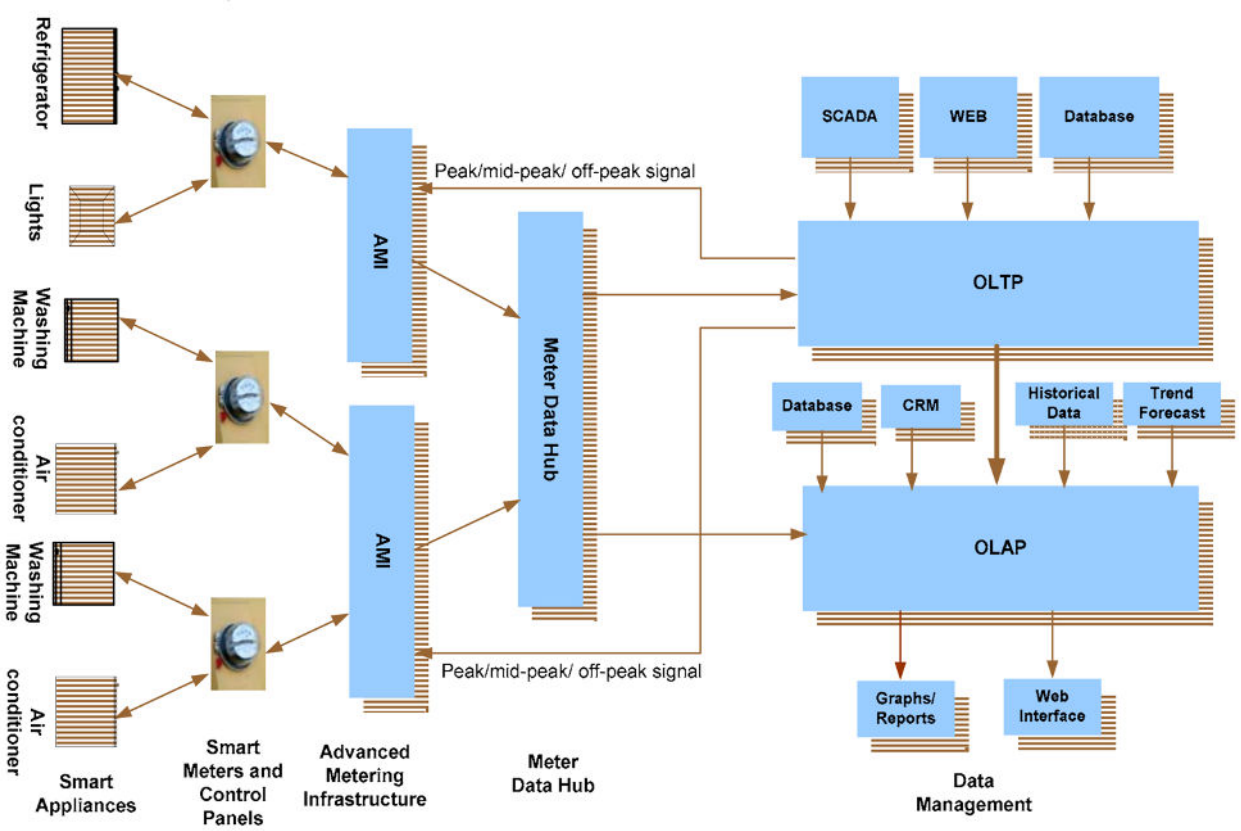



Abbildung 17: Konzept zur Verwendung von OLAP-Cubes (Nagesh et al., 2010, p. 1)

### 4.1.2. Mögliche Datenquellen

Nachdem das Konzept und das allgemeine Vorgehen erläutert wurden, stellt sich außerdem die Frage, welche Datenquellen im Zuge von Analysen genutzt werden können, also aus welchen Datenquellen das DW aufgebaut und gespeist werden soll. Es wurde bereits erläutert, dass die verschiedenen Datenquellen auf der gröbsten Stufe in unternehmensinterne und -externe Datenquellen aufteilen. Da auf Multidimensionalität ausgelegte Datenbanken immer aus Fakten und Dimensionen bestehen, ist zudem fraglich, welche Daten dieser Datenquellen als Fakten und welche als Dimensionswerte in die Analysen eingehen. Prinzipiell kommen als Dimensionswerte alle Werte in Frage, die einen Einfluss auf den Energieverbrauch eines Gebäudes haben könnten. Letztendlich hängt die Wahl der Dimensionen immer vom Analysefall und der zu prüfenden Hypothese ab. Dennoch sollen an dieser Stelle mögliche Datenquellen und deren beispielhafte Verwendung vorgestellt werden.

Der gemessene Energieverbrauch wird über das SMGW an den MSB übermittelt und von dem dortigen MDMS verarbeitet. Die übermittelten Daten enthalten neben dem gemessenen Energieverbrauch während des entsprechenden Intervalls ferner mindestens einen Zeitstempel, eine Meter ID und die Tarifierungsinformationen. Weitere Informationen über den Kunden und eventuell über das Gebäude oder den Haushalt liegen in Metadatenbanken des Unternehmens vor. Zunächst sollten die Fakten ausgewählt werden. Die übermittelten Energieverbräuche der verschiedenen Gebäude oder Wohnungen während der entsprechenden Messintervalle stellen die Fakten dar und werden daher je nach gewählter Datenbankarchitektur in die Faktentabelle oder als Fakten der mehrdimensionalen Arrays importiert. Ferner können diese durch Verrechnung mit bestimmten Dimensionswerten zu Kennzahlen angereichert werden. Allerdings stellt sich dafür zunächst die Frage nach den Datenquellen für die Dimensionen.

Die übermittelten Zeitstempel können dann aus dem MDMS extrahiert und als Dimension verwendet werden. Sie stellen die zeitliche Dimension dar, die in praktisch jeder OLAP-Analyse unabhängig vom Untersuchungsgegenstand zum Einsatz kommen dürfte. Aufbauend auf dem Zeitstempel und der zeitlichen Dimension, die den genauen Zeitpunkt der Messung bzw. das Messintervall angibt, lassen sich dann weitere zeitliche Dimensionen aufbauen. Beispielsweise können aufbauend darauf die zeitlichen Dimensionen Wochentag oder Tageszeit erzeugt werden. Diese können weitere Informationen und Zusammenhänge hinsichtlich des Energieverbrauchs liefern, die auf sich wiederholenden Mustern basieren. Alle zeitlichen Dimensionen basieren daher grundlegend auf dem übermittelten Zeitstemple und sind dem System daher bereits inhärent. Folglich werden alle zeitlichen Dimensionen aus internen Datenquellen aufgebaut.

Wie bereits erwähnt, beruht ein wesentlicher Teil des Gesamtenergieverbrauchs eines Gebäudes auf der thermischen Konditionierung der Räumlichkeiten. Wie groß die Notwendigkeit einer thermischen Konditionierung ist, hängt daher von der Differenz der thermischen Verhältnisse außerhalb des Gebäudes und der gewollten Verhältnisse innerhalb des Gebäudes ab. Folglich dürften deshalb auch Wetter- und Klimadaten eine wesentliche Rolle für energetische Analysen spielen. Wetterdaten werden in der Regel durch über das gesamte Land verteilte Wetterstationen des Wetterdienstes erhoben und als historische Werte in den Datenbanken des Wetterdienstes gespeichert. In Deutschland ist dafür der

Deutsche Wetterdienst zuständig. In der Regel sind alle diese über ein Online-Portal kostenlos verfügbar und frei zugänglich und können daher problemlos heruntergeladen und in das DW integriert werden. Folglich ergeben Wetterdaten eine weitere mögliche Dimension für die Analysen, die aus externen Datenquellen stammt. Theoretisch wäre es aber je nach Situation und Anwender auch denkbar, dass der Anwender selbst eigene Wetterstationen betreibt und die Daten daher in internen Systemen vorliegen. Dies könnte zum Beispiel bei großen Energieunternehmen der Fall sein.

Neben den klimatischen Bedingungen hängt der Energieverbrauch außerdem entscheidend von dem Gebäude selbst ab. Beispielsweise könnten das Alter des Gebäudes, die Energieeffizienz, die technische Gebäudeausstattung und die Art des Gebäudes, also z.B. ob es sich um ein Wohn- oder Nichtwohngebäude handelt, einen Einfluss auf den Energieverbrauch haben. Alle diese verschiedenen Informationen lassen sich allgemein als Gebäudetypologie zusammenfassen. Sie stellen neben den Wetterdaten eine zweite, sehr wichtige Dimension für energetische Analysen dar. Allerdings erweist sich die Frage nach den Datenquellen für gebäudetypologische Informationen im Vergleich zu Wetterdaten deutlich schwieriger, da sich nicht klar und einheitlich sagen lässt, ob, wie, in welchem Ausmaß, zu welchem Detaillierungsgrad und wie einheitlich und konsequent Gebäudetypologiedaten verfügbar sind. Zunächst stellt sich hier die Frage, ob und in wie fern Gebäudetypologiedaten bereits intern in Metadatenbanken verfügbar sind. Falls es sich bei dem entsprechenden Unternehmen um einen Energielieferanten, beispielsweise ein EVU, handelt, wäre denkbar, dass bereits bei Vertragsabschluss gewissen gebäudebezogene Daten erhoben wurden. Allerdings ist zweifelhaft, ob solche Datenquellen ausreichend detailliert und, falls der Liefervertrag schon länger besteht, noch immer aktuell sind. In diesem Zusammenhang sei jedoch erwähnt, dass im Fall von Strom Smart Metern durch CLS und die damit einhergehende automatisierte Fernsteuerung von Haushaltsgeräten über das WAN und HAN bereits diverse Informationen über die elektrische Gebäudeausstattung bestehen dürften, bzw. einsehbar wären. Indes gibt es theoretisch weitere, externe Datenquellen, die zusätzlich zu oder anstelle von Metadatenbanken des Energieunternehmens verwendet werden können, um Gebäudetypologiedaten zu gewinnen. Eine weitere Möglichkeit bestünde darin, solche Daten von den Eigentümern anzufragen oder ggf. käuflich zu erwerben. Da gebäudetypologischen Daten keinen Personen-, sondern einen Gebäudebezug haben, können diese sehr gut anonymisiert erhoben und gespeichert werden. Diese Herangehensweise könnte großflächig durch flächendeckende Kundenumfragen durchgeführt werden, eventuell auch gegen eine gewisse Vergütung. Ferner wäre es auch möglich, solche Daten nicht direkt von den Eigentümern und Kunden, sondern von Dritten zu erwerben, die solche Umfragen für eigene Zwecke durchgeführt haben.

Eine weitere mögliche Datenquelle könnten in diesem Zusammenhang Daten aus Geoinformationssystemen (GIS) darstellen. Diese stehen von mehreren bekannten Unternehmen meist über Web-Portale zur freien Verfügung und können dafür für jedweden Ort verwendet werden. Eventuelle bestehen für die relevanten Regionen sogar professionelle GIS-Modellierungen, die im Rahmen von Stadtplanungen, Baulandentwicklungen oder sonstigen Projekten erstellt wurden und verwendet können. Diese Systeme können selbst im ersten Fall wichtige Informationen liefern. Beispielsweise können durch GIS-Daten Bebauungsdichten, Versiegelungsgrade, Gebäudegrößen und somit Nutzflächen, Gebäudetypen oder topographische Informationen erkannt werden. Des Weiteren wäre es auch möglich, Informationen über die umliegende Gegend der Gebäude zu gewinnen,

beispielsweise die Art des Baugebiets anhand von Größe und Art der umliegenden Straßen und Erschließungsflächen oder anhand der Topographie oder einen Wohlstandsfaktor des Baugebiets anhand der Bebauungsdichte und Gebäudegrößen. Selbst scheinbar triviale Informationen wie die Dach- und Wandfarben oder -materialien wären so zu erheben und könnten aufgrund der Abhängigkeit von Reflektion und Transmission solarer Wärmeeinstrahlung von der Oberflächenfarbe für Analysen eine Rolle spielen. All diese Daten könnten je nach Anwendungsfall interessant sein und entweder in Gebäudetypologiedaten umgewandelt werden oder in Form einer anderen Dimension, beispielsweise der Dimension Wohlstand der Bewohner, in die Analysen eingehen.

Darüber hinaus könnten Kundendaten, also Informationen über die Anzahl der Bewohner, die Berufe oder das Einkommen der Bewohner oder deren Einstellungen gegenüber Energiesparmaßnahmen und somit deren Nutzerverhalten von Relevanz sein. Die für Abrechnungszwecke relevanten Kundendaten sind bereits in Metadatenbanken gespeichert. Da diese jedoch meist nicht besonders detailliert und für energetische Analysen eher weniger relevant sind, könnten auch Kundendaten stattdessen oder unterstützend durch Umfragen erhoben werden.

Die letzte hier erwähnte, mögliche Datenquelle stellen statistische Daten von Gemeinden, Städten, Regionen oder Ländern dar. Solche Daten haben per Definition keine hohe Granularität, da sie beispielsweise verschiedene Einzelwerte verrechnen, um einen Durchschnitt oder Quantile einer Verteilung zu berechnen. Jedoch könnten diese Daten bei größeren Aggregationen, beispielsweise der Analyse von ganzen Quartieren, viele Erkenntnisse liefern. Ferner lässt sich beispielsweise über statistische Daten eines bestimmten Stadtteils auch etwas über das zu erwartende Einkommen, die soziale Klasse oder das Bildungsniveau von Bewohnern eines bestimmten Gebäudes, also die Erwartungswerte und Abweichungen davon, innerhalb dieses Stadtteils aussagen. Statistische Daten haben dabei den besonderen Vorteil, dass sie nicht personenbezogen und damit nicht sensibel sind, was etwaiger Datenschutzbedenken vermeidet. Für energetische Analysen könnten insbesondere Daten über das Einkommen, den Wohlstand, die Eigentumsverhältnisse oder das Bildungsniveau von Bewohnern relevant sein.

Die verschiedenen, denkbaren Datenquellen und deren Zusammenspiel bei dem Aufbau eines DW sind im Folgenden graphisch am Beispiel einer sternförmigen Datenbank dargestellt.

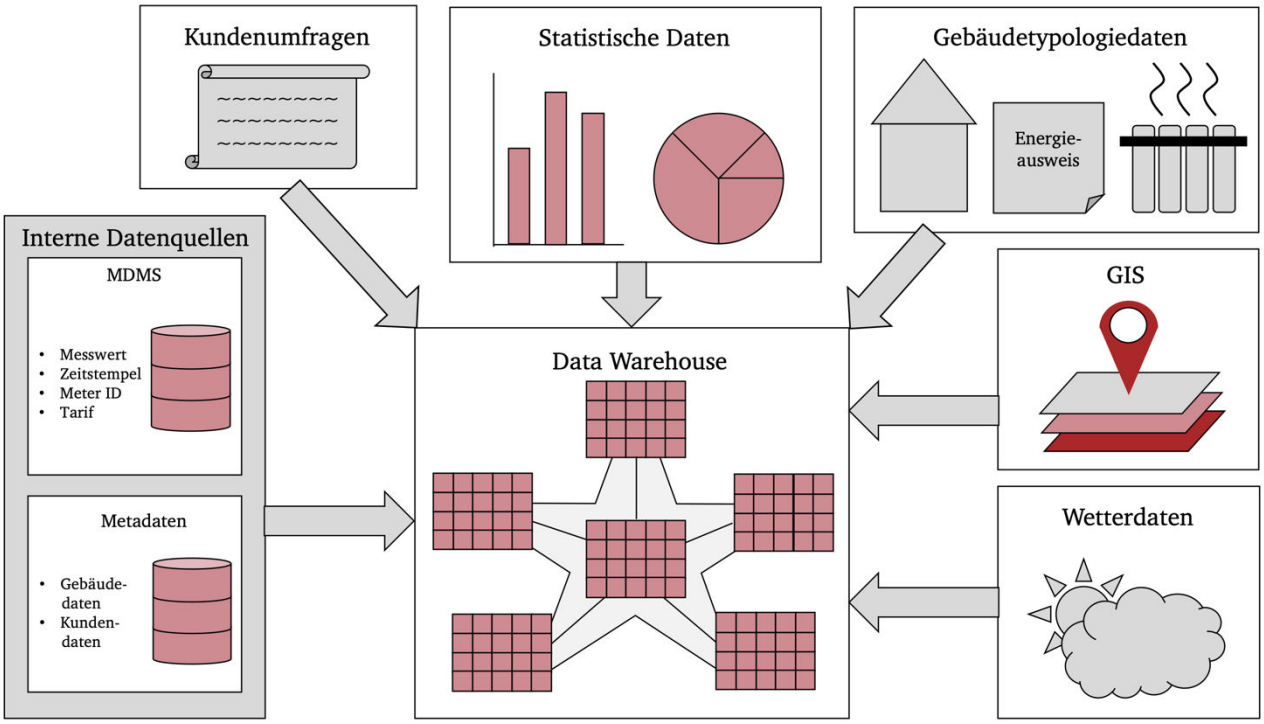



Abbildung 18: Datenquellen für das Data Warehouse

### 4.1.3. Anwendungsmöglichkeiten und Vorteile

Selbstverständlich stellt sich bei neuen Konzepten ebenfalls die Frage nach der Nützlichkeit, den Vorteilen und den Anwendungsmöglichkeiten, die durch eine Umsetzung realisiert werden könnten. Prinzipiell lassen sich die Vorteile durch die Verwendung von OLAP-Cubes für die Integration und Analyse von Energieverbrauchsdaten in zwei Kategorien aufteilen. Zum einen wird Mehrwert durch die Durchführung von Analysen und die so gewonnenen Erkenntnisse bzw. die verschiedenen Anwendungsmöglichkeiten für entsprechende Anwendergruppen und Nutzer geschaffen. Dieser Mehrwert könnte durch eine konkrete Umsetzung des Konzepts als vollständige Software-Lösung realisiert werden. Zum anderen existieren aber auch Vorteile durch die Technologie selbst, also Vorteile im Vergleich zu sonstigen Analyseprozessen und -technologien. In diesem Kapitel werden ausschließlich die denkbaren Anwendungsmöglichkeiten aus Sicht verschiedener Anwendergruppen und Stakeholder erörtert. Die konkreten technischen Vorteile durch die gewählte Umsetzung gegenüber sonstigen Analysetechnologien werden stattdessen während der Umsetzung in Kapitel 4 und abschließend im Fazit diskutiert.

Alle hier genannten Anwendungsmöglichkeiten basieren auf eigener Konzeption. Daher ist nicht auszuschließen, dass noch weitere Anwendungsmöglichkeiten denkbar wären. Entscheidend für die Kreation von wirklichem Mehrwert des Konzeptes ist, dass neue Nutzungsmöglichkeiten und Anwendungsszenarien geschaffen werden, die über die der bereits existierenden Systeme hinaus gehen. Eine reine Filterung, Visualisierung und Analyse der Energieverbrauchsdaten als schlichte Zeitreihendaten ist beispielsweise bereits mit den meisten existierenden EDMS oder MDMS möglich. Eine Anwendung des Konzeptes sollte hier also neue, darüberhinausgehende Funktionalitäten bieten.

Nichtsdestotrotz sollte bedacht werden, dass es sich bei dieser Arbeit lediglich um ein erstes Proof of Concept handelt und kein Anspruch auf die Entwicklung einer marktreifen, kommerziellen Anwendung erhoben wird.

Die Ergebnisse von Energieverbrauchsanalysen, die anhand von OLAP-Cubes erlangt werden können, sowie die daraus folgenden Schlüsse könnten aus makroskopischer Sicht für verschiedene Interessengruppen, insbesondere Politik, Wissenschaft und Wirtschaft, von Interesse sein. So bieten die Informationen für die Politik eine gute Datengrundlage und die Möglichkeit, beispielsweise den Erfolg bei der Umsetzung der Klimaziele besser nachzuverfolgen. Insbesondere wäre es durch die Analyse der Daten nach den verschiedenen Dimensionen und die Verknüpfung der gemessen Energieverbrauchsdaten mit der Gebäudetypologie möglich, bisher unerkannten Sanierungsbedarf bei gewissen Gebäudetypen oder Gebäuden mit gewisser technische Ausstattung, einem bestimmten Baujahr oder in bestimmten geographischen Regionen aufzudecken und die Energieeffizienz auf gebäudetypologischem, anlagentechnischem oder räumlichem Niveau zu vergleichen. Durch die so gewonnen Erkenntnisse könnten leichter neue, maßgeschneiderte Anreiz- und Subventionsprogramme geplant werden, sowie später deren Erfolg nach der Durchführung besser gemessen werden. Durch die große Anzahl an Dimensionen, die durch OLAP-Cubes für Analysen herangezogen werden können, sind viele verschiedene Definitionen miteinander kombinierbar, sodass nicht nur Erkenntnisse aus geographischer oder technischer Sicht entstehen, die für Anreizprogramme relevant wären, sondern auch demographische Zusammenhänge aufgedeckt werden können. Beispielsweise wäre denkbar, dass bisher nur bestimmte soziale oder demographische Bevölkerungsschichten Subventionsangebote zur Gebäudesanierung mit dem Ziel der Steigerung der Energieeffizienz wahrgenommen haben, dass die Stromnachfrage bestimmter Nichtwohngebäudetypen einen außergewöhnlich stark schwankenden Tagesverlauf aufweist, oder, dass Gebäude eines bestimmten Baujahrzeitraums zu wärmeren Jahreszeiten außergewöhnlich hohe Gasverbräuche haben, was auf ineffiziente Heizungsanlagen hinweist. Die Analyse der Daten anhand von OLAP-Cubes bietet hier den enormen Vorteil der übersichtlichen, intuitiven Datenstruktur, was Aggregationen hinsichtlich bestimmter Hierarchieebenen ermöglicht. Durch die Modellierung der Energieverbrauchsdaten als n-dimensionaler Würfel können die Daten sehr gut exploriert und aus unterschiedlichen Perspektiven betrachtet werden. Insbesondere auf Bundesebene entstehen bei solch flächendeckenden Implementierungen und Analysen enorme Datenmengen, was auf multidimensionale Analysen sehr großer Datenmengen optimierte Datenbanken und Abfragewerkzeuge praktisch voraussetzt.

Anwendung könnte das Konzept aus politischer Sicht aber nicht nur auf Bundes-, sondern auch auf Kreisebene finden. Zum Beispiel könnten Kommunen mit dem Ziel, den Anteil erneuerbarer Energien an der eigenen Stromversorgung und somit auch den eigenen Autarkiegrad zu steigern, anonymisierte, historische Stromverbrauchsdaten des Gebäudesektors und historische Einspeisedaten eigener Windkraftanlagen in Form eines Galaxie-Schemas modellieren, indem die Smart Meter Daten und die Daten der Windkraftanlagen in jeweils separate Faktentabellen integriert werden, die sich bestimmte Dimensionstabellen teilen. Über eine Analyse hinsichtlich des zeitlichen Verlaufs und verschiedenen Gebäudetypologien könnten dann z.B. Gebäudetypen entdeckt werden, die ein hohes Potential zur Lastverlagerung bieten. Auch ließen sich so wertvolle Erkenntnisse für die Planung zukünftiger Anlagen zur Stromerzeugung aus erneuerbaren Energien gewinnen.

Auch für die Anwendung im Zuge der Wissenschaft bestünde ein großes Potential. Durch die Fähigkeit, sehr große Datenmengen, die aus sehr feingranularen Daten bestehen, schnell zu analysieren, viele verschiedene Dimensionen heranzuziehen und zu kombinieren und klare Beschränkungen der Dimensionswerte festzulegen, um die Ergebnisse bestmöglich vergleichen zu können, wären OLAP-Cubes auch optimal für technische Analysen geeignet. Neben den klassischen Energieverbrauchsdaten wäre hier auch denkbar, dass OLAP-Cubes zur Analyse von sehr feingranularen, historischen Netzzustandsdaten genutzt werden können. Da wissenschaftliche und technische Analysen häufig sehr komplex sind, viele verschiedene Datenquellen hinzuziehen, genaue Beschränkungen und Festlegungen der zu verwendenden Daten erfordern und große Datenmengen nutzen, bietet die Modellierung durch OLAP-Cubes sehr viele Vorteile. Zudem wäre es möglich, nur ein gemeinsames DW aufzubauen, auf das alle Wissenschaftler Zugriff haben, um Analysen durchzuführen. Sind die Daten einmal integriert und wird ein ansprechendes Software-Tool zur Datenabfrage verwendet, benötigen die Endanwender dank der intuitiven und übersichtlichen Modellierung zudem nur ihre technische Expertise bezüglich der Fragestellung, jedoch keine tieferen Programmierkenntnisse. Aus wissenschaftlicher Sicht interessante Fragestellungen bezüglich des Energieverbrauchs wären z.B. das Nutzerverhalten. Mit steigender Energieeffizienz wird das Nutzerverhalten immer wichtiger. Beispielsweise spielt das Lüftungsverhalten eine größere Rolle, da bei geringeren Wärmedurchgangskoeffizienten der Gebäudehülle die Transmissionswärmeverluste sinken und somit die Lüftungswärmeverluste einen größeren Anteil am Gesamtheizwärmebedarf haben. Aufgrund der so steigenden Auswirkungen des Nutzerverhaltens auf den Heizwärmebedarf stellt sich daher die Frage, ob die gesetzlichen Normen zur energetischen Bilanzierung dieses Nutzerverhalten ausreichend genau widerspiegeln. Durch die neu gewonnenen Erkenntnisse hinsichtlich verschiedener Nutzerverhalten könnten dann entsprechende Tabellenwerte und Korrekturfaktoren überprüft und ggf. angepasst werden.

Auch für die Wirtschaft, insbesondere die Energiewirtschaft selbst, bestünden Anwendungspotentiale. Durch eine Analyse des Gasverbrauchs hinsichtlich des Wetters, der Region und verschiedener Gebäudetypologien und -größen auf Basis allgemein verfügbarer, anonymisierter Gas Smart Meter Daten könnten sich beispielsweise aus Sicht eines Unternehmens im Bereich des Heizungsbaus wichtige Erkenntnisse für die Produktentwicklung oder den Vertrieb ergeben. Durch die Entwicklung speziell zugeschnittener Produkte könnten so anhand der OLAP-Analysen entdeckte Marktlücken geschlossen werden. Aus Sicht eines Netzbetreibers oder Kraftwerkbetreibers könnten solche analytischen Informationssysteme und die darin enthaltenen historischen Daten in Kombination mit Echtzeitdaten eine wichtige Rolle für die Optimierung des operativen Geschäfts spielen. Aktuelle Ladestände von Batterien und Elektroautos aus Echtzeitdaten könnten auf diese Weise mit Wetterprognosen und Verbrauchsprognosen für die nächsten Stunden, basierend auf OLAP-Analysen, kombiniert werden, um die Kraftwerks- und Netzsteuerung zu optimieren oder beispielsweise anhand von historischen Verbrauchsquantilen für den entsprechenden Zeitraum sicherstellen, dass die Stromnachfrage mit einer bestimmten Wahrscheinlichkeit nicht höher ist als das -angebot. Die historischen Daten könnten durch die einfache Bedienung und guten Antwortzeiten schnell und einfach an den Bedarfsfall angepasst werden. Die Vorteile durch die intuitive Abfrage und das einfache Verständnis der Ergebnisse würden auch Anwendern aus der Wirtschaft zugutekommen.

Ferner wären neben solchen Anwendungen auf Makroebene aber auch Anwendungsmöglichkeiten auf Mikroebene denkbar. OLAP-Cubes könnten beispielsweise für die energetische Analyse des Immobilienbestandes verschiedener Unternehmen verwendet. Naheliegende Beispiele wären hier die Verwendung von OLAP zur langfristigen Speicherung und Analyse historischer Strom-, Gas- und Wasserverbräuche der Immobilien eines Immobilienentwicklers oder Analyse des Energieverbrauchs der Immobilien von Kaufhausketten, Lebensmitteldiscountern, Universitäten oder Rechenzentren. Diese haben aufgrund der verwendeten Lüftungs- und Kühltechnik in der Regel besonders hohe Stromverbräuche, sodass eine Analyse hinsichtlich zeitlicher, räumlicher, anlagentechnischer und klimatischer Bedingungen wesentliche Einsparpotentiale aufdecken könnte. Neben der Verwendung solcher analytischen Systeme als Unterstützung des energetischen Immobilienmanagements könnten die Energieverbrauchsdaten aber im Bereich der Produktionsanalyse verwendet werden, da beispielsweise in großen Industriekonzernen Elektrizität selbst oft ein wesentlicher Produktionsfaktor ist. All diese Anwendungsbeispiele haben den gemeinsamen Vorteil, dass die Verwendung von OLAP-Analysen hier ihr volles Potential ausschöpfen. Zum einen können weitere, eigene Sensordaten, wie Temperatursensoren oder Sensoren, die das Kundenaufkommen messen, als Dimensionswerte verwendet werden, zum anderen bestehen nicht die üblichen Einschränkungen und Bedenken bezüglich des Datenschutzes, da in diesem Fall der Eigentümer der Daten selbst die Analysen durchführt. Ferner können die Messintervalle beliebig kurz gewählt werden und es liegen keine sensiblen, persönlichen Daten wie bei Privathaushalten vor. Solche individuellen Anwendungsfälle dürften hinsichtlich der Situation in Deutschland bislang das größte Potential aufweisen. Hier wäre auch sogar eine kommerzielle Vermarktung etwaiger Software-Lösungen denkbar.

Im spezifischsten Fall wäre das Konzept aber auch auf einzelne Immobilien anwendbar. Bei größeren Mehrfamilienhäusern oder geschäftlich genutzten Hochhäusern könnten durch die OLAP-Cubes die Energieverbrauchsdaten der verschiedenen Wohnungen oder Büroräume analysiert und verglichen werden. In einzelnen Haushalten könnten so auch die Energieverbräuche einzelner Haushaltsgeräte gegenübergestellt werden. Würden beispielsweise über die Räumlichkeiten verteilt verschiedene Temperatur- oder Luftfeuchtigkeitssensoren installiert, könnten so auch technische Analysen bezüglich der Energieeffizienz und etwaiger, bisher unentdeckter Wärmebrücken durchgeführt werden. Die intuitive Aufteilung der Daten in Fakten und Dimensionen sowie die Bildung von Hierarchiestufen würden auch in solchen Fällen viele Vorteile bieten.

Insgesamt gibt es also viele Anwendungsmöglichkeiten, die von einer einfachen Nutzung von OLAP-Cubes für spezielle Analysefälle bis hin zu einer dauerhaften Nutzung für regelmäßige Analysen, Kontrollen, Vergleiche und eine langfristige Speicherung sämtlicher Daten. Welche Vorteile von OLAP-Cubes realisiert werden können, hängt dabei stark vom Anwendungsfall ab. Beispielhafte Anwendungsmöglichkeiten auf Makro- und Mikroebene wurden in diesem Kapitel bereits vorgestellt, jedoch sind der Anwendung theoretisch kaum Grenzen gesetzt. Auch Makroebene wäre zudem theoretisch denkbar, dass ein gemeinsames DW aus den weitestgehend anonymisierten oder pseudonymisierten Smart Meter Daten vieler verschiedener Energieversorger aufgebaut wird, auf dessen Server die verschiedenen Akteure aus Politik, Wissenschaft und Wirtschaft über ein Web-Portal zugreifen können. Wie für OLAP typisch, könnten auch hier verschiedene Zugriffsrechte vergeben

werden, sodass beispielsweise Wissenschaft und Politik Zugriff auf sensiblere Daten haben, während Wirtschaft und Öffentlichkeit nur nicht sensible Daten abfragen können.

### 4.1.4. Herausforderungen und Schwierigkeiten bei der Implementierung

Die durch die Analyse der Daten anhand einer OLAP-Cube Anwendung gewonnenen Erkenntnisse hätten das Potential, enormen Nutzen und Mehrwert für viele Bereiche der Gesellschaft zu generieren. Allerdings wird die volle Ausschöpfung ebendieses Potentials aktuell noch durch eine Reihe von Hürden und Regularien behindert. Zu nennen wären hier vor allem die bisher schleichende Diffusion und geringe Marktdurchdringung der Smart Meter Technologie, Datenschutzbeschränkungen und Bedenken um die Privatsphäre der Nutzer von Smart Metern, die bisher nicht oder kaum vorhandene Übermittlung und Integration von Daten über die Gebäudetypologien und sonstigen Sensordaten für die Analyse, sowie die sowohl sehr heterogenen, als auch zu langen Messintervalle. Insgesamt gilt, dass sowohl mit steigender Marktdurchdringung als auch mit steigender Granularität in Form von kürzeren Messintervallen - beispielsweise stündlich oder 15-minütig – immer mehr Daten generiert und damit auch immer mehr Informationen gewonnen werden können. Daher wäre im Rahmen der Konzeption dieser Arbeit prinzipiell eine Verkürzung und Vereinheitlichung der Messintervalle, sowie eine möglichst weitreichende Diffusion der Smart Meter Technologie wünschenswert und würde eine deutlich größere Abschöpfung der Potentiale des Konzeptes gewährleisten, als dies aktuell möglich ist.

Allerdings ist davon auszugehen, dass die Datenmenge und die Wertigkeit und Nützlichkeit der daraus gewinnbaren Erkenntnisse in keinem linearen, sondern eher in einem degressiven Zusammenhang stehen. Dies gilt sowohl für die Diffusion, also die Ausbreitung, der Smart Meter Technologie, als auch für die Länge der Messintervalle. Auch wenn eine völlige Marktdurchdringung, also die Ausstattung jedes Haushalts und jeder Immobilie mit einem Smart Meter, wünschenswert und optimal für die Durchführung von Analysen wäre, so könnte bereits eine nicht vollständige Marktdurchdringung ausreichende Sicherheit zur Allgemeingültigkeit und Anwendung der Ergebnisse der Analyse gewährleisten. Wie in vielen anderen Bereichen der deskriptiven Statistik, könnte auch in diesem Fall bereits eine relativ geringe prozentuale Marktdurchdringung eine ausreichend große Stichprobe repräsentieren. Ab einem gewissen Grad der Marktdiffusion von Smart Metern generiert die weitere Ausstattung von Haushalten mit Smart Metern zudem nicht mehr so viele neue Erkenntnisse wie noch zu Beginn der Fall. Mit fortschreitender Marktdurchdringung nimmt die Wertigkeit der neuen Erkenntnisse daher ab. Allerdings ist insbesondere bezüglich räumlicher und gebäudetypologischer Analysen eine gleichmäßige, möglichst räumlich, bzw. gebäudetypologisch unabhängige, relative Marktdurchdringung notwendig. Dieser degressive Zusammenhang von Datenmenge und Erkenntnisgewinn dürfte sich auch bezüglich der Granularität der Daten einstellen. Auch wenn prinzipiell gilt, dass mit kürzeren Messintervallen noch mehr Daten zur Verfügung stehen und sich daraus noch mehr und noch detailreichere Analysen durchführen lassen, so ist insbesondere bezüglich des Konzeptes und der durchgeführten Analysen im Rahmen dieser Arbeit eine Schranke bei den Messintervallen definierbar. Zur Verdeutlichung des degressiven Zusammenhangs können sich hier vergleichend die Messintervalle monatlich, stündlich, 15-minütig und minütig vor Augen geführt werden. Während bei monatlicher Messung und Übermittlung der Verbrauchswerte lediglich die monatlichen Gesamtverbräuche analysieren lassen, geben stündliche Messintervalle deutlich mehr

Aufschluss über den Energieverbrauch, da sich so auch Schwankungen unter der Woche und während eines einzelnen Tages erkennen lassen. Werden stattdessen 15-minütige Messintervalle realisiert, werden Energieverbräuche noch zeitnaher dargestellt und es lassen sich eventuell neue Zusammenhänge, beispielsweise der Stromnachfrage mit Strompreisen, bestimmten Haushaltsgeräten oder thermischen Speichern, aufdecken, die bei stündlicher Messung nicht erkannt wurden. Allerdings stellt sich bereits an dieser Stelle die Frage, ob die neu gewonnen Informationen eine Vervierfachung der Datenmenge und die damit einhergehende Belastung von Kommunikationssystemen und Datenbanken rechtfertigen. Wird darüber hinaus sogar im Minutentakt gemessen und übertragen, wird der beschriebene degressive Zusammenhang gut deutlich. Mit Ausnahme der Übertragung von Netzzustandsdaten ist sind solch kurze Messintervalle fraglich, da im Vergleich zu 15-minütigen Messintervallen zwar die 15-fache Datenmenge generiert wird, wodurch aber nicht die 15-fache Menge an Informationen, Trends und Zusammenhängen gewonnen werden kann. Daher wäre eine 15-minütige Messung, wie sie im Zuge der RLM bereits etabliert ist, vollkommen ausreichend.

Zudem könnte es, je nach Interesse und Anwendung, bereits ausreichend sein, wenn nur eines der beiden Kriterien bestmöglich erfüllt ist. Beispielsweise wäre vorstellbar, dass für Analysen aus Sicht der Politik aus im Folgenden näher erläuterten Gründen vor allem eine möglichst hohe Marktdurchdringung erforderlich wäre, während für Analysen im Rahmen der Wissenschaft oder Wirtschaft vor allem kurze Messintervalle von vorrangigem Interesse wären. In beiden Fällen wäre das jeweils andere Kriterium nachrangig und daher eine geringere Marktdurchdringung, bzw. längere Messintervalle hinnehmbar. In Anbetracht der Vorteile, die kurze Messintervalle für Kunden bieten, da auf diese Weise günstigere, zeitvariable Tarife und Kostensenkungen durch Netzentlastung realisierbar wären, besteht zudem Hoffnung, dass kürzere Messintervalle sich in Zukunft auch im Sektor der Privathaushalte etablieren könnten. Allerdings können selbst ohne hohe zeitliche Auflösung der Daten dank der weiteren Dimensionswerte viele Erkenntnisse gewonnen werden. So kann auch die Untersuchung von Monats- oder Jahresverbräuchen unter Betrachtung von statistischen Daten wie Einkommen, soziale Klasse oder Bildungsniveau oder geographischen Daten wie Bebauungsdichte, Geländegefälle oder Höhenlage interessante Ergebnisse liefern. Bezüglich der Marktdurchdringung ist zudem klar, dass der Ersatz traditioneller, analoger Zähler durch Smart Meter nicht eine Frage des Ob, sondern des Wann ist.

Neben einer bisher geringen Marktdurchdringung der Smart Meter Technologie und einer geringen Granularität der Smart Meter Daten stellt außerdem der Datenschutz ein wesentliches Hindernis für eine erfolgreiche Umsetzung und Anwendung dar und erschwert das volle Abschöpfen der vorhandenen Potentiale. Bei Energieverbrauchsdaten handelt es sich um sehr sensible Daten. Das GDEW beinhaltet deshalb neben den technischen Richtlinien und Vorgaben des BSI auch Regelungen zum Datenschutz und der Datenkommunikation in intelligenten Energienetzen. Diese Regelungen besagen unter anderem, dass personenbezogene Daten nur von berechtigen Stellen erhoben, verarbeitet und genutzt werden dürften. Dazu zählen MSBs, EVUs, Netzbetreiber und jede Partei, die über eine Einwilligung des Verbrauchers verfügt. Allgemein ist die Erhebung, Verarbeitung und Nutzung der Daten nur zur Erfüllung der Verträge oder mit Einwilligung des Verbrauchers gestattet. (Riester, 2017, p. 31) Die Daten müssen dann verschlüsselt übertragen werden, wobei personenbezogene Daten zu anonymisieren und zu pseudonymisieren sind. Außerdem ist dem MSB

gestattet, personenbezogene Daten in erforderlichem Umfang zu erheben. Diese müssen allerdings gelöscht werden, sobald eine Speicherung für die Aufgabenerfüllung nicht mehr erforderlich ist. (Riester, 2017, p. 31)

Zudem unterliegen die Daten neben den Vorgaben des GDEW den allgemeinen Bestimmungen des deutschen Datenschutzes, der die drei Hauptprinzipien der Erforderlichkeit, Zweckbindung sowie Datenvermeidung und Datensparsamkeit vorgibt. Aufgrund des Grundsatzes der Erforderlichkeit dürfen personenbezogene Daten nur erhoben und verarbeitet werden, wenn sie zur Erreichung eines Zwecks, z.B. der Erfüllung von Verträgen oder Behördenaufgaben, notwendig sind. Alle sonstigen, nicht erforderlichen Verarbeitungen von Daten erfordern die Einwilligungen der betroffenen Personen. (Riester, 2017, pp. 31-32) Der Grundsatz der Zweckbindung besagt, dass die erhobenen Daten nur für den während der Erhebung vorgesehenen Zweck verarbeitet werden dürfen. Nach dem dritten Grundsatz der Datenvermeidung und -sparsamkeit sollten nur wenig personenbezogene Daten wie möglich erhoben, verarbeitet und genutzt werden. Dementsprechend sind Datenverarbeitungsprozesse so zu entwickeln, dass sie mit möglichst wenig Daten auskommen und alle personenbezogenen Daten weitestgehend in anonymisierter und pseudonymisierter Form verarbeitet werden. (Riester, 2017, p. 32) Grundsätzlich gilt im Allgemeinen, dass jede Art der Datenverarbeitung einem Verbot mit Erlaubnisvorbehalt unterliegt. Folglich dürfen nach deutschem Datenschutzrecht personenbezogene Daten nur erhoben, verarbeitet und genutzt werden, solange eine Rechtsvorschrift dies erlaubt oder anordnet oder der Betroffene seine wirksame Einwilligung ausgedrückt hat. (Riester, 2017, p. 32)

Um die durch Datenschutzgesetze und Datenschutzbedenken bestehenden Hindernisse bei der Umsetzung zu umgehen, bestehen verschiedene Herangehensweisen und Lösungen. Zunächst sei hier erwähnt, dass etwaige Ängste und Bedenken von Verbrauchern häufig auf subjektiver Wahrnehmung und fehlender Aufklärung statt auf objektiven Mängeln basieren. Durch die notwendigen Zertifizierungsverfahren für Smart Meter, die vorgeschriebenen Pseudonymisierungen und Verschlüsselungen der Daten sowie die technischen Richtlinien des BSI kommt dem Datenschutz in solchen Systemen bereits eine sehr hohe Stellung zu. Hier läge die Aufgabe also in vielerlei Hinsicht vor allem in besserer Aufklärung und Informationsbereitstellung gegenüber Verbrauchern. Weiterhin gelten sämtliche Bedenken bezüglich der Privatsphäre ohnehin nur für den privaten Sektor. Da Privathaushalte aber beispielsweise einen eher geringen Anteil an der gesamten Stromnachfrage des Gebäudesektors haben, wären auch bei Verzicht auf die Nutzung der Verbrauchsdaten von Privathaushalten bereits viele Erkenntnisse zu gewinnen. Daher besteht insbesondere für Analysen des geschäftlichen Sektors ein hohes Potential. Ferner gilt außerdem, dass in den meisten Anwendungsfällen Daten über die Personen und Verbraucher selbst gar nicht wirklich von Interesse sind. Die Frage danach, wer genau in welchem Zeitraum wie viel Energie verbraucht hat ist daher nicht nur datenschutzrechtlichen Gründen sehr schwierig, sondern auch für die Analysen meist gar nicht relevant. Stattdessen ist die meist die Frage nach dem Wo, nicht dem Wer, also die Frage nach räumlichen Aspekten, sowie Informationen des Gebäudes der Verbraucher selbst relevant. Dabei handelt es sich also gar nicht um direkt personenbezogene, sondern nur um indirekt personenbezogene Daten. Wichtig ist in diesem Fall, dass aus den übermittelten räumlichen und gebäudetypologischen Daten keine genauen Rückschlüsse auf individuelle Verbraucher und Personen möglich sind. Gebäudetypologiedaten könnten daher anonymisiert oder pseudonymisiert und ohne

rückverfolgbaren Verbraucherbezug übermittelt und gespeichert werden. Bedenken bezüglich der Nutzung räumlicher Daten wie Adressen, die eine genaue Rückverfolgung ermöglichen würden, können umgangen werden, indem Adressen in ausreichend geringer Auflösung übermittelt werden. Beispielsweise könnten anstatt der genauen Adresse nur der entsprechenden Stadtteile, die Gemeinde oder der Landkreis übermittelt werden.

Für alle sonstigen, weiterhin bestehenden Datenschutzprobleme bestehen zudem mögliche technische Lösungen. Rigoll erwähnt beispielsweise die Möglichkeit, die verschiedenen verbrauchs- und personenbezogenen Daten mit Gefährdungskennzahlen zu versehen, die Aufschluss über die Sensibilität der Daten und mögliche Gefährdungen der jeweiligen Verbraucher, also der Eigentümer der Daten, geben. (Rigoll, 2017, pp. 163-179) Diese Gefährdungskennzahlen können basierend auf der zeitlichen Auflösung der Daten, dem durch die Daten abgedeckten Gesamtzeitraum, dem Alter und der Art der Daten vergeben werden. So steigt die Sensibilität der Daten mit einer höheren zeitlichen Auflösung, da mehr Informationen über die Verhaltensmuster der Verbraucher erhoben werden. (Rigoll, 2017, pp. 164-167) Auch durch längere abgefragte Zeiträume steigt der Gefährdungsgrad, da sich aus kürzeren Ausschnitten von Energieverbrauchsdaten weniger Rückschlüsse ziehen lassen als aus längeren. (Rigoll, 2017, pp. 167-170) Bei Datenabfragen beeinflusst auch das Alter der Daten die potentiellen Auswirkungen der Datenabfragen auf die Privatsphäre der betroffenen Nutzer. Mit steigendem Alter sinkt die Gefährdung der Nutzer. Insbesondere Daten, die bereits mehrere Jahre alt sind, lassen kaum noch Rückschlüsse auf aktuelle Lebensgewohnheiten der Verbraucher zu. (Rigoll, 2017, pp. 170-173) Hinsichtlich der verschiedenen Datenkategorien existieren ebenfalls verschieden große Risiken für die Privatsphäre von Nutzern. Entscheidend ist hier, ob es sich um Verbrauchsdaten von Haushaltsgeräten handelt, die eine direkte Interaktion mit dem Nutzer erfordern oder nicht. Ein Elektro- oder Gasherd oder ein Fernseher wird in der Regel nur in Anwesenheit des Nutzers verwendet, während Kühlschränke, Klimageräte oder Umwälzpumpen nur sehr bedingte Rückschlüsse auf das Nutzerverhalten ermöglichen. (Rigoll, 2017, pp. 173-177) Diese vier verschiedenen Kennzahlen können durch entsprechende festgelegte Gewichtungen auch zu einer einzigen Gefährdungskennzahl zusammengefasst werden. (Rigoll, 2017, pp. 177-179) Um den Bedenken um die Privatsphäre der Verbraucher gerecht zu werden, wäre es dann möglich, OLAP-Werkzeuge, mit denen die Daten aus dem DW abgefragt werden, mit einer Funktion auszustatten, die sämtlichen Datenanfragen vor der Ausführung eine Gefährdungskennzahl zuweist. Ist diese zu hoch und damit die Privatsphäre der Kunden zu sehr gefährdet, müssten Anpassungen der Anfrage durch eine Erhöhung der Hierarchieebenen bestimmter Dimensionen durchgeführt werden, sodass die Daten auf eine weniger detaillierte Ebene aggregiert werden. Durch eine Reduktion der zeitlichen Auflösung oder eine Einschränkung des abgefragten Zeitraums würden die abgefragten Daten dann eine geringere Gefährdung darstellen. In diesem Zusammenhang wäre auch denkbar, dass für höhere Granularitäten wie genaue Adressen oder Tagesverläufe verschiedene Zugangscodes oder Datenschlüssel eingegeben werden müssen, die nur bestimmten, vertrauenswürdigen Instanzen vorliegen.

Da die meisten OLAP-Anfragen die Daten aber ohnehin auf höhere Niveaus aggregieren, sind aus Sicht der Analyseergebnisse keine Datenschutzbedenken nötig. Kritisch ist daher in erster Linie die Integration und Speicherung der Daten in das DW, da die Daten dadurch prinzipiell schon auf individueller Ebene vorliegen. Zum Schutz der Kunden wäre hier auch das Hinzufügen eines

Störterms, also eines Rauschens, denkbar, das in der Summe Null und damit den Gesamtverbrauch nicht verzerrt, dafür aber Rückschlüsse auf Lebensgewohnheiten verhindert. Rigoll schlägt zudem vor, Datenbanken, die Energieverbrauchsdaten beinhalten, mit einer Funktion auszustatten, die direkte Anfragen für bestimmte Daten an Kunden ermöglicht, sodass Kunden die Einwilligung zur Erhebung und Nutzung dieser Daten unkompliziert erteilen oder ablehnen können. Im Fall einer Ablehnung bestünde dann die Möglichkeit, Gründe für die Ablehnung zu nennen, wodurch das Unternehmen die Anfrage entsprechend anpassen kann. Auf diese Weise wäre eine Art Verhandlungsprozess möglich, der die Aussagekraft der Daten auf ein für die Kunden akzeptables Niveau reduziert, die Einwilligung der Kunden wäre garantiert und es würde klare Transparenz bezüglich der Datenverarbeitung herrschen. (Rigoll, 2017, pp. 8-9)

All das stimmt bezüglich der Realisierung und Abschöpfung des Potentials der Anwendung optimistisch. Fraglich ist bei makroskopischer Anwendung allerdings weiterhin, ob insbesondere Gleichmäßigkeit der Marktdurchdringung von Smart Metern über verschiedene Gebäudetypologien hinweg vorausgesetzt werden kann. Ausschlaggebend für die konzipierten Nutzungsmöglichkeiten im Rahmen dieser Arbeit ist grundlegend nur, dass die installierten Smart Meter fernausgelesen werden, also auch tatsächlich ein SMGW besitzen, und in geringen Messintervallen Daten generieren.

## 4.2. Technische Umsetzung des Konzepts

Sollen OLAP-Cubes und BI-Technologien für die Analyse von Energieverbrauchsdaten verwendet werden, sind für eine technische Umsetzung viele verschiedene Varianten denkbar. Da darüber hinaus der Gesamtprozess von der Datenauswahl bis hin zur Visualisierung aus mehreren Teilprozessen besteht, ist es zudem möglich, für jeden Teilprozess eine speziell für den Anwendungsfall geeignete technische Umsetzung zu wählen. Theoretisch stehen sämtliche Umsetzungsmöglichkeiten, von einer manuellen Umsetzung sämtlicher Teilprozesse, über die Nutzung bestehender Programm-Suiten zur intuitiven, automatischen Umsetzung der Teilprozesse, bis hin zur Erstellung einer hochentwickelten Software-Anwendung, die sämtliche Teilprozesse ausführen und verknüpfen kann und damit den Gesamtprozess ohne sonstige Unterstützung durchführt, zur Auswahl. Zudem ist es möglich, für jeden eigenständigen Teilprozess zu entscheiden, ob dieser manuell „von Hand“, mit Hilfe einer bestehenden Software-Anwendung oder durch Entwicklung einer eigen entwickelten, speziell zugeschnittenen Anwendung erfolgen soll.

Im einfachsten Fall könnten das DW manuell in einer SQL-Sprache aufgebaut und die Daten integriert werden. Anschließend würden die Daten durch SQL-Befehle abgefragt und die Ergebnistabelle als CSV-Datei exportiert. Diese CSV-Datei könnte dann in Microsoft Excel importiert werden und die Werte der Tabelle könnten mit Hilfe bestehender Funktionalitäten des Programms durch Pivot-Tabellen oder sonstige Lösungen exploriert, untersucht und visualisiert werden. Ferner wäre darauf aufbauend auch eine komplexere, anspruchsvollere Umsetzung möglich, indem die Formatierung und Visualisierung der Analyseergebnisse durch Programmierung und spezielle Skripte erfolgt. Im Gegensatz dazu wäre es aber auch möglich, eine vollständige, eigene Anwendung zu entwickeln, die sämtliche Schritte dieses Prozesses so automatisiert, dass alle Prozesse von der Integration bis zur Visualisierung einfach und intuitiv über eine Benutzeroberfläche durchgeführt werden können. Über integrierte ETL-Tools für die Datenintegration, Benutzeroberflächen für die Datenabfrage und Werkzeuge zur Visualisierung wäre der Gesamtprozess auch von Anwendern ohne tiefere Programmierkenntnisse durchführbar. Solche hochentwickelten Anwendungen würden durch eine aufgebaute Datenbankverbindung für den integrierten Datenbankzugriff dafür sorgen, dass der Transfer von Dateien und Daten über mehrere Programme hinweg vermieden wird und keine ständigen Importe und Exporte der Daten notwendig sind. In der Praxis ist es aber aufgrund der Komplexität des Gesamtprozesses und der verschiedenen Anforderungen der einzelnen Teilprozesse meistens der Fall, dass für alle Teilprozesse ein separat entwickeltes Software-Tool besteht und somit der Datentransfer zwischen den Anwendungen weiterhin erfolgen muss.

Zudem stellt sich im Falle manueller Umsetzung auch die Frage, welche Programme und Programmiersprachen verwendet werden sollen. Für die Integration der Daten in das DW und deren Abfrage ist dafür zunächst eine geeignete Datenbanksprache zu wählen. Auch für die Visualisierung können verschiedene Programmiersprachen gewählt werden. Im Zuge der Umsetzung im Rahmen dieser Arbeit haben sich darüber hinaus weitere Programme zur Durchführung und Unterstützung der ETL-Prozesse und Datenmigrationen zwischen den verschiedenen Programmen als notwendig und nützlich herausgestellt. Hierzu zählen insbesondere Texteditoren, CSV-Editoren, Tabellenkalkulationsprogramme und Statistikprogramme.

Da es sich bei der Umsetzung in dieser Arbeit um ein erstes Proof of Concept handelt und bisher nur kaum bis gar keine Literatur über die Nutzung von OLAP-Analysen in Verbindung mit Energieverbrauchsdaten besteht, wurde sich dazu entschieden, die praktische Umsetzung manuell, als „von Hand“, unter Verwendung speziell für den Fall zugeschnittener Skripte, durchzuführen. Soweit möglich, werden sämtliche Schritte also durch Programmierung automatisiert, allerdings wird noch keine zusammenhängende Anwendung inklusive einer Benutzeroberfläche entwickelt. Die manuelle Umsetzung hat zudem den Vorteil, dass der zugrundeliegende Gesamtablauf mit allen Teilprozessen sehr gut und klar dokumentiert wird. Würde hier stattdessen eine bestehende Software verwendet, würde die Umsetzung zudem eher in eine Anleitung zur Nutzung ebendieser ausarten, anstatt diese zugrundeliegenden technischen Abläufe zu demonstrieren.

### 4.2.1. Auswahl der Datenbankarchitektur

Als Datenbankarchitektur für das DW werden Sternschemata verwendet. Diese sind aus verschiedenen Gründen am besten für die Analysen im Rahmen dieser Arbeit geeignet. Der erste Vorteil besteht darin, dass die Daten durch relationale Tabellen modelliert werden. Dies hat zur Folge, dass für die Speicherung, Bearbeitung und Abfrage der Daten etablierte, weitläufig bekannte, relationale DBMS zum Einsatz kommen, was mit all den in Kapitel 3 geschilderten Vorteilen dieser DBMS gegenüber multidimensionalen DBMS verbunden ist. Zudem bieten Sternschemata eine einfach verständliche, intuitive Struktur zur Datenmodellierung und sind dadurch sehr übersichtlich. Da im Rahmen von wissenschaftlichen und technischen Analysen nicht selten sehr viele verschiedene Dimensionen berücksichtigt werden und jede Dimension durch eine separate Dimensionstabelle modelliert wird, kann das DW dadurch ohnehin bereits aus sehr viele Tabellen bestehen. Würde an dieser Stelle stattdessen das Schneeflockenschema gewählt, würden sich noch deutlich mehr Tabellen ergeben, was die Übersichtlichkeit drastisch reduziert, da eine einzige Dimension durch mehrere Tabellen modelliert wird, die zudem alle verknüpft werden müssen. Der gleiche Vorteil gilt auch bezüglich späterer Datenabfragen, da diese bei komplexen Analysen bereits bei Verwendung des Sternschemas sehr lang und damit fehleranfällig sein können. Die Integration der Daten in DW geht im Falle einer sternförmigen Modellierung ebenfalls einfacher und schneller als im Falle einer schneeflockenförmigen Modellierung. Da die Datenquellen für Energieverbrauchsanalysen, wie sich herausstellen wird, sehr heterogen sein können, gestaltet sich der ETL-Prozess zudem bereits komplex und aufwendig genug und sollte nicht durch die Verwendung von Schneeflockenschemata weiter verkompliziert werden. So ist es unter anderem bei Sternschemata sehr einfach, Dimensionsdaten abschließend zu überprüfen, da diese alle in einer einzigen Tabelle gespeichert sind. Zudem sind auf Sternschemata basierende DWs im Zuge von Datenaktualisierungen sehr gut und einfach vertikal skalierbar.

Der zweite Vorteil besteht bezüglich der Datenabfrage. Schneeflockenschemata und multidimensionale Datenbankarchitekturen stellen bezüglich der Antwortzeiten, der Komplexität der Abfragebefehle und des Speicherplatzbedarfs praktisch genaue Gegenpole dar. Sternschemata sind hinsichtlich dieser Kriterien genau dazwischen angeordnet und stellen eine gute Mischung aus beiden Extremen dar, indem sie einen guten Kompromiss aus Performance und Speicherplatzbedarf bieten und gleichzeitig am besten skalierbar sind. Sternschemata weisen bessere Antwortzeiten als Schneeflockenschemata auf, da weniger Tabellen kombiniert und durchlaufen werden müssen. Gegen multidimensionale

Datenbankmodelle spricht zudem, dass für wissenschaftliche und energetische Analysen in aller Regel im Voraus Hypothesen aufgestellt werden und der Fokus damit nicht so sehr auf dem Explorieren der Daten besteht, sondern auf der Prüfung genau festgelegter Fragestellungen. In Anbetracht der Vorteile des Sternschemas ist es daher hinnehmbar, dass Abfragen erst zur Laufzeit durchgeführt und nicht vorausberechnet werden. Da das DW ein von den operativen Informationssystemen getrenntes System darstellt und die Analysen meist anhand von Cloud Servern durchgeführt werden, werden die operativen Systeme für das Tagesgeschäft des Unternehmens ohnehin nicht durch die Analysen belastet. Ferner ist auch bei Verwendung relationaler Datenbankmodelle die Abfrageperformance meist ausreichend gut. Es lassen sich im Gegensatz zu multidimensionalen Datenmodellen lediglich keine konstanten Antwortzeiten realisieren. Eventuell auftretende Wartezeiten bei den Datenabfragen sind rein objektiv betrachtet definitiv hinnehmbar. Stattdessen stellt sich sogar eher die Frage, ob der Mehraufwand durch Integration und im Voraus berechnete Aggregationen sämtlicher Abfragemöglichkeiten und Kombinationen von Dimensionswerten, wie sie bei multidimensionalen Datenmodellen der Fall sind, gerechtfertigt ist, nur um konstantere, eventuell nur wenige Sekunden kürzere Antwortzeiten zu realisieren.

Als Datenbanksprache wird SQLite verwendet. SQLite bietet den Vorteil, dass es auf den meisten Geräten bereits vorinstalliert ist und es sich um eine weit verbreitete und ausgereifte Datenbanksprache handelt. Zudem werden Datenbanken in SQLite lokal als Datenbank-Datei auf der Festplatte gespeichert und nicht, wie beispielsweise im Fall von MySQL, auf Servern gehostet. Dies vereinfacht auch schlicht die Dokumentation der Umsetzung im Zuge dieser Arbeit. Der große Vorteil von SQLite besteht aber darin, dass mit SQLiteStudio eine sehr praktische, benutzerfreundliche, graphische Benutzeroberfläche zur Verfügung steht. Da die Umsetzung manuell erfolgt, führt die Verwendung der GUI zur Bearbeitung der Datenbank zu großen Zeiteinsparungen, da die Strukturen von Tabellen schneller erstellt oder verändert werden können. Außerdem vereinfacht und beschleunigt dies Importe und Exporte von Dateien in und aus der Datenbank. Ferner bietet SQLite Schnittstellen für alle etablierten Programmiersprachen, wodurch Software-Anwendung eine Datenbankverbindung aufbauen und auf die Datenbank zugreifen können.

### 4.2.2. Integration der Datensätze durch ETL-Prozesse

Der ETL-Prozess erfolgt zu Demonstrations- und Dokumentationszwecken manuell „von Hand“. Dieser stellt einen wesentlichen Teil des gesamten OLAP-Prozesses dar, da ein erfolgreicher Aufbau der Datenbank und eine erfolgreiche Integration aller relevanten Werte nicht nur eine größere Herausforderung sein kann, sondern auch den Grundstein für alle weiteren Schritte legt. Nur bei einer sauberen Extraktion, Transformation und Integration der Daten in das DW können ohne Probleme Analysen durchgeführt werden. Sobald der ETL-Prozess erfolgreich abgeschlossen ist und die Daten sauber im DW gespeichert wurden, steht einer erfolgreichen mehrdimensionalen Analyse praktisch kaum noch etwas im Weg.

Alle Dateien bestimmter Zwischenstufen des Transformationsprozesses werden zunächst in einer Staging Area gespeichert. Als Staging Area kann theoretisch sowohl die OLAP-Datenbank selbst fungieren, indem z.B. temporäre Tabellen erstellt werden, die der Transformation der Daten dienen

und aus denen dann die finalen Tabellen erzeugt werden, als auch eine separate Datenbank für den ETL-Prozess aufgebaut werden, deren fertige Tabellen dann exportiert und in die OLAP-Datenbank importiert werden. Ferner ist auch aber auch schlicht ein einfacher Ordner mit den jeweiligen Dateien bestimmter Zwischenstufen des Transformationsprozesses nutzbar. Wird die OLAP-Datenbank oder eine separate Datenbank als Staging Area genutzt, wird die Transformation über SQL-Befehle durchgeführt. Erfolgt der ETL-Prozess extern, also nicht in einer Datenbank, kann die Transformation durch Tabellenkalkulationsprogramme wie Microsoft-Excel, CSV-Editoren oder Texteditoren durchgeführt werden.

Im Zuge der Umsetzung wird die Transformation sowohl extern durch etwaige Programme und Editoren als auch intern in der OLAP-Datenbank nach Import der CSV-Dateien in entsprechende Tabellen durchgeführt. Ob die Verarbeitung intern oder extern erfolgt, hängt dabei vom Einzelfall ab richtet sich nach der Geschwindigkeit und Einfachheit der Varianten. Da zur Prüfung aller Hypothesen ein neues DW aufgebaut werden und sich um die erstmalige Integration der Daten handelt, ist eine teilweise Durchführung des ETL-Prozesses innerhalb der OLAP-Datenbank unproblematisch. Wichtig und zu bedenken ist hier aber, dass dies im Fall der Skalierung eines bestehenden DW, also der Aktualisierung der Datenbank durch Integration weiterer Daten, durchaus zu Problemen führen könnte. In solchen Fällen würde es sich daher dringend empfehlen, den gesamten ETL-Prozess klar vorzuverlegen, sodass die Daten vor dem Import in die OLAP-Datenbank bereits vollständig transformiert sind und damit eine klare Trennung zwischen Staging Area und OLAP-Datenbank herrscht. Außerdem wurden im Zuge des ETL-Prozesses einige Formatierungen und Transformationen der Daten durch R- oder Python-Skripte automatisiert. Diese werden auf den speziellen Anwendungsfall zugeschnitten programmiert und können etwaige, durch Dateninkompatibilitäten, verschiedene Granularitäten oder fehlende Werte hervorgerufenen Probleme lösen.

### 4.2.3. Analyse und Visualisierung der Daten

Sobald die Daten erfolgreich aufbereitet und in das DW geladen wurden, stehen diese für Abfragen und Analysen bereit. Durch die relationale Modellierung der Daten im DW erfolgt auch die Abfrage relational. Somit handelt es sich um ein ROLAP-Modell. Die Gründe für die Wahl eines relationalen Modells wurden bereits bei der Auswahl der Datenbankarchitektur genannt. Durch die vorher aufgestellten Hypothesen stehen die relevanten Kennzahlen und Dimensionswerte bereits vor der Analyse fest und es ist nicht sinnvoll, alle weiteren möglichen Kombinationen und Aggregationen zu berechnen. Stellt sich durch die Untersuchung der Abfrageergebnisse beispielsweise heraus, dass bestimmte Stellen einen Blick ins Detail erfordern, werden die Abfrage entsprechend angepasst und die Daten hinsichtlich der neuen Angaben aggregiert und ausgegeben. Als dies geschieht zur Laufzeit.

Die Abfrage erfolgt durch SQL-Befehle. Das Grundgerüst der Abfragebefehle ist dabei immer gleich und muss lediglich an den entsprechenden Fall angepasst werden. Abfragebefehle haben in der Regel die Form (SELECT … FROM … WHERE … GROUP BY). Allerdings ist auch die Einbindung weiterer Operationen wie (ORDER BY) möglich. Durch den „SELECT“-Befehl werden alle Spalten der Tabellen angegeben, die ausgegeben werden sollen. Im Zuge dessen werden auch die gewollten Aggregationsoperationen spezifiziert und festgelegt. So kann beispielsweise die Summe, der

Durchschnitt oder das Minimum oder Maximum des Verbrauchs ausgegeben werden. „FROM“ legt gibt dann an, aus welchen Tabellen diese gewählten Spalten stammen. Durch den Befehl „WHERE“ werden gleich zwei Aufgaben übernommen. Zum Einen werden die verschiedenen Tabellen über die Primärschlüssel und Fremdschlüssel verknüpft. Daher muss in diesem Teil der Abfrage angegeben werden, welche Fremdschlüsselspalten der Faktentabelle mit welchen Primärschlüsselspalten der Dimensionstabellen verknüpft werden sollen. Zum Anderen werden hier aber auch Einschränkungen der Dimensionswerte der gewählten Dimensionen festgelegt. So können beispielsweise einzelne Dimensionswerte oder ein Intervall von Dimensionswerten ausgewählt oder auch spezielle Dimensionswerte von der Analyse ausgeschlossen werden. Um den „WHERE“-Befehl den Anforderungen entsprechend zu erweitern, kommt dieser Befehl meist in Kombination mit den Operationen „AND“ oder „OR“ zur Anwendung. Im Fall von „AND“ müssen beide Kriterien gleichzeitig erfüllt sein, im Fall von „OR“ nur eines der beiden. Somit stellt der „WHERE“ in der Regel den längsten Teil der SQL-Abfrage dar. Abschließend werden die ausgewählten Daten dann über „GROUP BY“ gruppiert. Dabei werden die Spalten angegeben, hinsichtlich derer die Werte aggregiert und gruppiert werden sollen. Über die Angabe der Spalten wird somit anhand der Hierarchieebene das gewünschte Aggregationsniveau angegeben. In der Folge werden die einzelnen Werte der Faktentabelle für jeden gewählten Dimensionswert der Hierarchieebene aggregiert. Hier können mehrere Spalten ausgewählt werden, wodurch im Ergebnis eine mehrdimensionale Aggregation der Werte der Faktentabelle entsteht. Prinzipiell gilt für die Anordnung der Spalten bei der Aufzählung, dass diese von links nach rechts aggregiert werden. Dementsprechend ist die letzte (rechte) Spalte die, über die als letztes aggregiert wird. Es empfiehlt aus Gründen der Interpretation, hier die Spalte mit den meisten Dimensionswerten zu wählen. Der Output ist dann eine Tabelle mit den entsprechend aggregierten und gruppierten Daten.

Diese Tabelle kann dann weiterverarbeitet und visualisiert werden. Für eine Visualisierung der Ergebnisse durch Graphen und Diagramme oder die Darstellung der Daten als Würfel oder Matrix müssen allerdings noch einige Formatierungen vorgenommen werden, da der Output selbst als Tabelle, nicht als n-dimensionales Array oder Würfel erfolgt. Diese Tabelle beinhaltet als Spalten die ausgewählten Spalten und als Zeilen die aggregierten Fakten und die dazugehörigen Dimensionswerte. Die Visualisierung kann dann durch eigenständige Programmierung oder Verwendung bestehender Visualisierungs-Tools umgesetzt werden. In dieser Arbeit wird die Visualisierung in Python durchgeführt. Python bietet dafür die zwei mächtigen Bibliotheken Pandas und Matplotlib. Pandas dient der Datenverarbeitung, kommt häufig im Data Science Bereich zum Einsatz und ist auf die Verarbeitung von Datenreihen und -vektoren getrimmt. Insbesondere bietet es besonders viele Funktionen zur Verarbeitung und Manipulation tabellenartiger Datenstrukturen wie Matrizen oder Zeitreihen. Matplotlib ist eine Bibliothek für die Visualisierung von Daten und wird im häufig in der Wissenschaft verwendet. Zudem besitzen die typischen DataFrame- und Series-Objekte von Pandas eine eigene Plot-Funktion, wodurch die beiden Bibliotheken gut und intuitiv zusammenarbeiten.

Der Output der SQL-Abfrage wird zunächst als CSV-Datei gespeichert. Diese CSV-Datei wird dann von Pandas eingelesen, für die Visualisierung entsprechend formatiert, sodass die Daten sinnvoll auf X- und Y-Achse widergegeben werden können, und dann graphisch dargestellt. Anstelle der Datenmigration und dem damit verbundenen Ex- und Import der Ergebnisse als CSV-Datei wäre es aber auch möglich,

die Abfrage über eine aufgebaute Datenbankverbindung direkt von der Visualisierungsanwendung aus durchzuführen. Python bietet hier eine SQLite 3 Bibliothek, über die eine Datenbankverbindung aufgebaut und SQL-Befehle übermittelt werden können. Auf diese Weise könnten alle weiteren Prozessschritte nach dem Aufbau des DW von einem einzigen Skript durchgeführt werden, wodurch die Entwicklung eines allumfassenden Abfragewerkzeugs möglich wäre. In dieser Arbeit wurde sich aber aus Test- und Dokumentationsgründen für die erste Variante entschieden. Diese ist meist sogar schneller umsetzbar, da keine Datenbankverbindung konfiguriert werden muss. Zudem liegt der Output der SQL-Abfrage so übersichtlich als CSV-Datei bereit und kann einfach an Dritte weitergegeben werden.

## 4.3. Demonstration des Modells durch Analyse von Hypothesen

Um eine praktische Umsetzung des Konzeptes zu demonstrieren werden drei verschiedene Hypothesen aufgestellt, die anhand einer OLAP-Analyse geprüft werden sollen. Für jede Hypothese werden Datenquellen ausgewählt, ein DW aufgebaut und die Daten abgefragt und visualisiert. Die Hypothesen sind so geordnet, dass die Komplexität der Analyse und die Anforderungen an das DW mit jeder Hypothese steigen.

### 4.3.1. Hypothese 1

**Der monatliche Stromverbrauch von Nichtwohngebäuden ist über das Jahr hinweg konstanter als der monatliche Stromverbrauch von Wohngebäuden.**

Es ist zu erwarten, dass der Stromverbrauch einer Immobilie bzw. eines Haushaltes in den Wintermonaten höher ist als in den Sommermonaten. Da die Wintermonate deutlich weniger Tageslicht, weniger Sonnenstunden und kältere Außentemperaturen aufweisen, dürfte der Stromverbrauch durch Beleuchtung und eventuelle elektrische Warmwasserbereitung in den Wintermonaten höher sein als in den Sommermonaten. Außerdem wäre denkbar, dass der Stromverbrauch durch Unterhaltungselektronik in den Wintermonaten ebenfalls höher ist, da die Menschen durch weniger Tageslicht, schlechteres Wetter und kältere Temperaturen mehr Zeit im Haus verbringen. Fraglich ist nun, inwiefern dieser Mehrverbrauch in den Wintermonaten durch Mehrverbräuche in den Sommermonaten kompensiert wird, sodass der monatliche Stromverbrauch über das Jahr hinweg weitestgehend konstant ist und kaum Schwankungen aufweist. Solche etwaigen, sommerlichen Mehrverbräuche dürften in erster Linie auf Klima- und Raumlufttechnik zurückzuführen sein, da Zu- und Abluftventilatoren von Lüftungsanlagen, Kompressoren von Kältemaschinen, Klimaanlagen oder Entfeuchtern und Pumpen für die Zirkulation von Kältemitteln allesamt Elektrizität in mechanische Energie umwandeln. Solche mechanischen Systeme zum Raumluftaustausch und zur Raumluftkonditionierung kommen zumindest in Nord- und Westeuropa meist nur in kommerziell genutzten Nichtwohngebäuden wie in Banken, in Büros, im Lebensmittelhandel oder in Kaufhäusern zum Einsatz. Die meisten Privathaushalte nutzen stattdessen zum Raumluftaustausch und zur Kühlung der Räumlichkeiten im Sommer in der Regel noch manuelle Lüftung. Maximal kommen in solchen Fällen einzelne Ventilatoren oder kleinere Klimaanlagen zum Einsatz. Dementsprechend ist anzunehmen, dass solche sommerlichen Mehrverbräuche die winterlichen Mehrverbräuche nur in Nichtwohngebäuden kompensieren. Der so verursachte konstantere Verlauf des monatlichen Strombedarfs könnte in Nichtwohngebäuden zudem durch allgemein höhere Grundlasten im Vergleich zu Privathaushalten verstärkt werden, die den Einfluss der klimatischen Verhältnisse auf den Gesamtstromverbrauch verringern. Eine höhere Grundlast könnte in Nichtwohngebäuden z.B. durch die konstantere Verwendung von Beleuchtung über die Jahreszeiten hinweg sowie über eine höhere Anzahl stromverbrauchender Geräte wie Arbeitscomputern, deren Stromverbrauch unabhängig von der Jahreszeit ist, entstehen. Diese Hypothese soll anhand einer OLAP-Analyse überprüft werden.

Als Datenquelle zur Überprüfung dieser Hypothese werden die Strom Smart Meter Daten von einem Pilotprojekt aus Irland verwendet. Das Pilotprojekt „Electricity Customer Behaviour Trial“ wurde von Juli 2009 bis einschließlich Dezember 2010 von der irischen Commission for Energy Regulation durchgeführt. Während dieser eineinhalb Jahre wurden über 5000 irische Privathaushalte und Geschäfte mit Strom Smart Metern ausgestattet, um die Auswirkungen von Smart Metern auf den Stromverbrauch zu untersuchen und Informationen für eine Kosten-Nutzen-Analyse eines landesweiten Smart Meter Rollouts zu gewinnen. Dabei wurde der Stromverbrauch in 30-Minuten-Intervallen gemessen. Die Messwerte und Meter IDs liegen dabei in anonymisierter Form vor und beinhalten keine geographischen Informationen, sodass eine Rückverfolgung auf individuelle Personen ausgeschlossen ist. Neben den Messdaten wurden ferner vor Beginn und nach Abschluss des Pilotprojekts Umfragen, unter anderem mit Fragen zu den Gebäuden und dem Nutzerverhalten, durchgeführt, deren Ergebnisse ebenfalls zur Verfügung stehen. Darüber hinaus liegen Informationen über die entsprechenden Tarife und die Art der Gebäude vor. Die Daten wurden freundlicherweise von dem Irish Social Science Data Archive zur Verfügung gestellt.

### 4.3.2. Hypothese 2

**Der Gasverbrauch eines Privathaushaltes hängt primär von der Außentemperatur und nicht dem Nutzerverhalten ab.**

Erdgas hat sich weltweit als ein wichtiger Energieträger für die Bereitstellung des Energiebedarfs von Gebäuden etabliert. Dabei wird Erdgas primär für die Wärmebereitstellung zur Raumkonditionierung verwendet. Allerdings kann Erdgas auch Teile des Energiebedarfs eines Haushaltes über den Heizwärmebedarf hinaus decken. Das bekannteste Beispiel ist hier die Wärmebereitstellung zur Brauchwarmwassererwärmung. Ferner können aber auch Kochstellen, Herde oder optische Feuer mit Gas betrieben werden. Für die Transmissionswärmeverluste durch die Gebäudehülle ergibt sich aus bauphysikalischer Sicht rechnerisch ein linearer Zusammenhang zwischen der Außentemperatur und dem Heizwärmebedarf und somit auch dem Gasverbrauch in kWh. Indessen hat neben der Außentemperatur aber auch das Verhalten der Nutzer einen Einfluss auf den Gasverbrauch. So hängt der Gasverbrauch ebenfalls von der für den persönlichen Komfort gewählten Soll-Temperatur der Räumlichkeiten, dem Lüftungsverhalten, dem Trinkwarmwasserverbrauch und der Sparsamkeit der Bewohner ab. Die Frage ist nun also, inwiefern dieser lineare Zusammenhang zwischen Außentemperatur und Gasverbrauch durch das Nutzerverhalten beeinflusst und verzerrt wird. Während die Energieeffizienz des Gebäudes zwar ebenfalls einen großen Einfluss auf den Gasverbrauch hat, hat diese jedoch keine Konsequenzen für den linearen Zusammenhang zwischen Außentemperatur und Gasverbrauch, da der sich Gasverbrauch lediglich insgesamt erhöht oder verringert und somit der lineare Zusammenhang weiterhin bestehen bleibt. Werden aber beispielsweise von den Nutzern in bestimmten Monaten des Jahres während der Abendstunden über längere Zeit optische Feuer betrieben, in den Übergangsmonaten höhere Soll-Temperaturen der Räume als in den Wintermonaten gewünscht oder wird während der Wintermonate besonders viel besonders warmes Trinkwasser verbraucht, so würde dies den linearen Zusammenhang zwischen Außentemperatur und Gasverbrauch verzerren. Die Hypothese ist jedoch, dass solches Nutzerverhalten

nur einen sehr geringen Einfluss auf diesen Zusammenhang hat und der Gasverbrauch eines Haushaltes weiterhin primär von der Außentemperatur abhängt. Dies soll anhand einer OLAP-Analyse untersucht werden.

Zur Überprüfung der Hypothese werden die Gasverbrauchsdaten aus einem weiteren Smart Meter Pilotprojekt aus Irland verwendet. Das Pilotprojekt „Gas Customer Behaviour Trial“ wurde von Dezember 2009 bis einschließlich Mai 2011 von der irischen Commission for Energy Regulation durchgeführt. Während dieser eineinhalb Jahre wurden knapp 1500 irische Privathaushalte mit Gas Smart Metern ausgestattet, um die Auswirkungen von Smart Metern auf den Gasverbrauch zu untersuchen und Informationen für eine Kosten-Nutzen-Analyse eines landesweiten Gas Smart Meter Rollouts zu gewinnen. Auch hier wurde der Gasverbrauch in 30-minütigen Intervallen gemessen. Die Messwerte und Meter IDs liegen wieder in anonymisierter Form vor und beinhalten keine geographischen Informationen, um eine mögliche Rückverfolgung auf individuelle Personen zu vermeiden. Auch in diesem Pilotprojekt wurden vor Beginn und nach Abschluss des Pilotprojekts Umfragen, unter anderem mit Fragen zu den Gebäuden, der Ausstattung des Haushaltes und der Einstellung der Bewohner, durchgeführt, deren Ergebnisse ebenfalls vorliegen. Auch diese Daten stammen aus dem Irish Social Science Data Archive und wurden auf Anfrage zur Verfügung gestellt. Darüber hinaus werden als externe Daten die Wetterdaten des irischen Wetterdienstes verwendet.

### 4.3.3. Hypothese 3

**Ein hohes Haushaltseinkommen steigert den Gasverbrauch, da es weniger Anreize für energiesparendes Nutzerverhalten bietet.**

Das Verhalten der Bewohner kann einen großen Einfluss auf den Energieverbrauch eines Gebäudes haben. Durch ineffizientes Lüftungsverhalten, permanente Beheizung der Räumlichkeiten auch in Abwesenheit oder einen hohen Trinkwarmwasserverbrauch kann ein Gebäude deutlich mehr Energie verbrauchen als eigentlich notwendig. Gleichzeitig wird das Nutzerverhalten mit steigender Energieeffizienz des Gebäudes aber immer wichtiger, da beispielsweise die Lüftungswärmeverluste durch die reduzierten Transmissionswärmeverluste einen größeren Anteil am Gesamtheizbedarf des Gebäudes haben. Aufgrund dieses Zusammenhangs ist das Nutzerverhalten für die Realisierung eines energieeffizienteren, klimaneutralen Gebäudebestands von besonderer Relevanz und es stellt sich folglich insbesondere die Frage, welche Faktoren das Nutzerverhalten beeinflussen. Neben Umwelt- und Klimaschutzaspekten stellen vor allem Kosteneinsparungen einen wesentlichen Anreiz für ein energiesparendes Nutzerverhalten dar. Dieser Anreiz ist besonders dann von Relevanz, wenn sich die Kosteneinsparungen auch tatsächlich bemerkbar machen. Insbesondere in einkommensschwachen Haushalten können Energiekosteneinsparungen durch ein energiesparenderes und effizienteres Nutzerverhalten einen großen Unterschied auf das verbleibende Budget des Haushaltes machen. In einkommensstarken Haushalten hingegen dürften diese Kosteneinsparungen jedoch nur geringere Effekte haben, da die Energiekosten aufgrund des höheren Lebensstandards einen geringeren Anteil an den monatlichen Gesamtkosten des Haushaltes haben dürften. Die Frage ist daher also, inwiefern das Einkommen das Nutzerverhalten und den Energieverbrauch beeinflusst und ob ein höheres

Haushaltseinkommen auch gleichzeitig einen verschwenderischen Verbrauch von Energie bedeutet. Diese Hypothese soll anhand einer OLAP-Analyse untersucht werden. Als Datengrundlage dienen in diesem Fall erneut die Gas Smart Meter Daten des Pilotprojekts aus Irland.

# 5. Umsetzung

## 5.1. Grundlegendes Vorgehen bei der Umsetzung

Prinzipiell wird bei der Umsetzung nach dem folgenden Schema vorgegangen. Dieses Schema kann als Richtlinie und Vorlage für die Integration und Analyse von Energieverbrauchsdaten durch OLAP-Cubes verstanden werden.

1. Festlegung der Grundlagen
   1.1. Aufstellung einer Hypothese
   1.2. Auswahl der Daten aus operativen Systemen und externen Datenquellen
   1.3. Auswahl der zu verwendenden Software oder Programmiersprache(n)
   1.4. Auswahl der Datenbankarchitektur
2. Aufbau des Data Warehouse
   2.1. Aufbau der Faktentabelle und Integration der Messdaten
   2.2. Transformation und Aufbereitung der Dimensionsdaten in einer Staging Area
   2.3. Aufbau der Dimensionstabellen und Integration der Dimensionsdaten
   2.4. Verknüpfung von Fakten- und Dimensionstabellen bzw. Durchführung letzter Formatierungen
3. Abfrage der Daten aus dem Data Warehouse
4. Visualisierung der Abfrageergebnisse
   4.1. Formatierung und Aufbereitung der Abfrageergebnisse für die Visualisierung
   4.2. Auswahl der Visualisierungsart und Darstellungsform
   4.3. Visualisierung der Daten
5. Interpretation der Daten und Prüfung der Hypothese

Während der Untersuchung, Darstellung und Interpretation der Abfrageergebnisse können sich zudem neue Fragestellungen ergeben. Diese Erkenntnisse könnten dann beispielsweise eine neue Sicht auf die Daten, einen Blick ins Detail, eine Aggregation der Daten auf eine höhere Hierarchieebene oder eine Einschränkung, Erweiterung oder Veränderung der herangezogenen Dimensionswerte erforderlich machen. In solchen Fällen muss dann die Abfrage entsprechend angepasst werden. Folglich wird dann zu Schritt 3 zurückgekehrt und der Prozess von dort aus erneut durchlaufen.

## 5.2. Hypothese 1

**Der monatliche Stromverbrauch von Nichtwohngebäuden ist über das Jahr hinweg konstanter als der monatliche Stromverbrauch von Wohngebäuden.**

Gegenstand der Analyse sind also die Stromverbrauchsdaten, die hinsichtlich verschiedener Dimensionen analysiert werden sollen. Die gemessenen Stromverbräuche stellen somit die Fakten der Analyse dar und werden daher in der Faktentabelle gespeichert. Um die monatlichen Stromverbräuche über das Jahr hinweg zu analysieren, wird außerdem eine zeitliche Dimension benötigt. Zudem müssen Informationen über die Gebäudetypen hinzugezogen werden. Dies geschieht durch Hinzufügen einer Dimension „Gebäudetyp“. Ein wesentlicher Unterschied bezüglich des Stromverbrauchs besteht im Vergleich von Wohn- und Nichtwohngebäuden zudem in der Frage, ob es sich um Tage unter der Woche oder um Wochenenden handelt. Während die Privathaushalte auch am Wochenende bewohnt werden und die Benutzer weiterhin Strom verbrauchen, eventuell durch längere Anwesenheit im Gebäude sogar mehr als unter der Woche, sind Geschäfte an Wochenenden, zumindest an Sonntagen, in der Regel geschlossen und dürften daher außer diverser benutzerunabhängiger Grundlasten, beispielsweise durch Kühlschränke, keine Stromverbräuche aufweisen. Um dies zu berücksichtigen wird ebenfalls die Dimension Wochentag berücksichtigt. Abschließend wird außerdem zu Demonstrationszwecken die Dimension Ort hinzugefügt. Da die Smart Meter Daten anonymisiert und ohne Ortsangaben vorliegen, wurden für diese Dimension Werte nach Belieben vergeben.

### 5.2.1. Aufbau des Data Warehouse und Integration der Daten

#### Faktentabelle (FactTable)

Zunächst muss die Faktentabelle erstellt werden. Die Dateien mit den Messwerten der Smart Meter bestehen dabei aus drei Spalten – eine Spalte für die gemessenen Stromverbrauchswerte und zwei Spalten für eine räumliche und zeitliche Zuordnung der Messungen in Form einer Meter ID und einer Time ID. Diese zwei Spalten für die „physikalisch“ Zuordnung der gemessenen Werte werden, wie sich zeigen wird, Grundlage für die Integration und Verknüpfung weiterer Dimensionen sein und können ebenfalls als Fremdschlüssel benutzt und angesehen werden. Folglich wird für den erfolgreichen Import der Dateien eine Faktentabelle mit den drei entsprechenden Spalten erstellt.

```
CREATE TABLE FactTable (
    MeterID INTEGER,
    TimeID INTEGER,
    Usage DOUBLE
);
```

Anschließend können dann die einzelnen Dateien importiert werden. Diese setzen sich aus 6 verschiedenen Dateien zusammen, die nach den Meter IDs gegliedert sind. File 1 enthält dabei alle Daten der Meter IDs 1000 – 1999, File 2 alle Daten der Meter IDs 2000-2999 und so weiter. Allerdings

enthält File 6 die Daten aller Meter IDs größer als 6000, also von 6000 bis 7444, und ist somit größer als die anderen Dateien. Zunächst wurden die Dateien für den Import als CSV in die DW Staging Area gespeichert. Allerdings wäre bei SQLiteStudio in der Regel auch ein Import in Textformat möglich. Dabei trat jedoch stets ein recht gravierender Fehler beim Import der Daten auf. Der erste Wert der Zahl der ersten Spalte (Meter ID) wurde nicht importiert. Wurden die Dateien unverändert inklusive der ersten Zeile, die keine Spaltenüberschrift darstellt, importiert, wurde der Wert der ersten Spalte und ersten Zeile vollständig importiert, bei allen restlichen Zeilen jedoch der erste Wert der ersten Spalte nicht mit importiert. Wurde manuell in der Datei eine Kopfzeile für die Spaltenüberschriften eingefügt und die Datei ohne die erste Zeile importiert, wurden alle Werte der ersten Spalte inklusive des Wertes der ersten Zeile ohne den ersten Wert importiert. Nach langem Ausprobieren und vielen Versuchen, das Problem durch andere Dateiformate, das Weglassen oder Einfügen von Spaltenüberschriften oder verschiedene Datentypen für die erste Tabellenspalte zu lösen, konnte jedoch trotzdem keine Lösung für dieses Problem gefunden werden. Die nach Wochen, also nach der Zeit und nicht der Meter ID, geordneten Daten der Gas Smart Meter, die für die zweite Hypothese verwendet werden, ließen sich hingegen problemlos importieren, genauso wie alle anderen, nicht nach den Strom Smart Meter IDs geordneten Dateien. Der Unterschied zwischen den beiden Datensätzen besteht neben der Tatsache, dass die Strom Smart Meter Daten nach Meter ID in der ersten Spalte gegliedert sind, während die Gas Smart Meter Daten nach der Time ID in der ersten Spalte aufgeteilt sind vor allem darin, dass die Spalten der Datensätze unterschiedliche Trennzeichen verwenden. Die Strom Smart Meter Daten nutzen Leerzeichen (whitespace), während die Gas Smart Meter Daten Kommata (comma) verwenden. Durch die schiere Größe der Dateien war es allerdings auch nicht möglich, die angewandten Trennzeichen anhand eines Texteditors durch andere Trennzeichen zu ersetzen, da so viele Millionen Stellen des Dokumentes betroffen sind, dass übliche Texteditor bei dem Versuch regelmäßig abstürzen. Ferner trat bei dem Import von aus den Strom Smart Meter Daten erzeugten Subdateien im CSV Format, bei denen ebenfalls die Time ID in der ersten Spalte steht, genau das gleiche Problem auf und es wurde erneut der erste Wert der ersten Spalten nicht importiert.

Dieses Problem musste daher stattdessen durch eine nachträgliche Anpassung der Werte der ersten Spalte nach dem Import in die Faktentabelle umgangen werden. Es wurde manuell in die Datei eine Kopfzeile eingefügt, die dann beim Import ignoriert wurde, wodurch alle Werte der ersten Spalte gleichmäßig ohne die erste Zahl importiert wurden. Anschließend wurden diese weggelassenen Werte dann wieder hinzugefügt, indem die Zahlen entsprechend aktualisiert wurden, sodass die richtigen, ursprünglichen Werte herauskommen. Dies kann über die UPDATE-Funktion gelöst werden. Wichtig ist, dass diese Aktualisierung direkt nach dem Import der einzelnen Datei durchgeführt wird, da ansonsten mehrere Meter IDs bereits unter gleichem Wert importiert werden und später nicht mehr zuzuordnen sind. Beispielsweise würden dann die Meter IDs 1500, 2500, 3500, 4500 und 5500 alle als 500 importiert und wären nicht mehr unterscheidbar. Jede Datei wird nach dem Import daher direkt aktualisiert. Für File 1 sieht dies folgendermaßen aus:

```
UPDATE FactTable SET MeterID = MeterID + 1000;
```

Die Werte der ersten Spalte von File 4 können beispielsweise folgendermaßen korrigiert werden:

```
UPDATE FactTable SET MeterID = MeterID + 4000;
```

Die Herangehensweise führt jedoch bei File 6 erneut zu einem Problem, da File 6 wie bereits erwähnt als einzige Datei die Messwerte von mehr als 1000 Meter IDs enthält, nämlich die Werte der Meter IDs 6000 bis 7444. Dies führt durch das Importproblem dann dazu, dass die Meter IDs 6000 – 6444 und 7000 – 7444 alle nur als die letzten drei Ziffern importiert und folglich vermischt werden. Beispielsweise werden die Meter IDs 6100 und 7100 beide als 100 importiert und die zugehörigen Messwerte können nicht mehr auseinandergehalten und zugeordnet werden. Auch nach vielen Versuchen, das Problem zu beheben, konnte keine Lösung gefunden werden. Es ließen sich ebenfalls keinerlei Unreinheiten in der zugrundeliegenden Datei erkennen. Aufgrund dessen wurde entschieden, File 6 nicht für die Analyse heranzuziehen. Durch die erfolgreiche Korrektur aller weiteren Importdateien stehen allerdings auch so noch knapp 5000 verschiedene Meter IDs zur Verfügung, was immer noch ausreichend sein dürfte.

Nach dem Import der Dateien in die Faktentabelle empfiehlt es sich, die Spalten der Datei neu anzuordnen, um eine bessere Übersicht zu gewährleisten. So sollten alle Spalten für Kennzahlen und Faktenwerte nebeneinander angeordnet sein. Gleiches gilt für alle Fremdschlüsselspalten, die auf die Dimensionstabellen verweisen. In diesem Fall wurde entschieden, die Kennzahlenspalten links und die Fremdschlüsselspalten rechts anzuordnen.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                          FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage   DOUBLE,
    MeterID INTEGER,
    TimeID  INTEGER
);

INSERT INTO FactTable (
                          Usage,
                          MeterID,
                          TimeID
                      )
                      SELECT Usage,
                             MeterID,
                             TimeID
                        FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

**Dimension Zeit (DimTime)**

Als nächstes wird als erste Dimension die zeitliche Dimension integriert. Die zeitliche Dimension basiert auf den vermerkten Messzeitpunkten der einzelnen Messdaten und ist daher in diesem Beispiel bereits den Dateien mit den Messdaten und somit auch der Faktentabelle inhärent. Die Angabe des Messzeitpunktes erfolgt als fünfstelliger Code, bestehend aus einer Mischung aus Julian Day Format und Tagesstundenkodierung. Das Datum wird durch die ersten drei Ziffern über eine Art Julian Day Format angegeben, dass die Anzahl der vergangenen Tage seit dem 31.12.2008 beschreibt. Somit

handelt es sich um kein richtiges Julian Day Format, da der Starttag dieses Formates mehrere Tausend Jahre zurück liegt. „001“ würde somit dem 01.01.2009 entsprechen. Die Angabe des Tageszeitpunktes erfolgt dann durch die letzten zwei Ziffern in Form einer Codierung von 1 bis 48. Diese repräsentiert die 48 Messungen innerhalb eines Tages. „01“ entspricht dabei dem Messintervall 00:00:00 Uhr bis 00:29:59 Uhr, usw. Die Codierung der Messzeitpunkt ist für Menschen nicht lesbar und muss für eine angemessene Interpretation und die Erstellung von Hierarchieebenen in einen lesbaren Zeitstempel umgewandelt werden. Da die Transformation dieser Codierung in einen Zeitstempel aufgrund der Art der Codierung etwas komplexer und SQLite viele spezielle Befehle für Zeitformate bietet, erfolgt die Transformation der Daten direkt in Datenbank durch SQLite.

Zunächst wird die Dimensionstabelle für die Zeit erstellt.

```
CREATE TABLE DimTime (
    PK_Time  INTEGER  PRIMARY KEY,
    Date STRING,
    Time STRING,
    DateTime DATETIME
);
```

Anschließend werden der Tabelle als Primärschlüssel die Time IDs der Faktentabelle zugewiesen. Da einzelne Time IDs in der Faktentabelle mehrmals vorkommen (einmal für jede Meter ID) müssen diese dafür entsprechend gruppiert werden. Außerdem werden die Time IDs bereits in die jeweiligen Codierungen für das Datum und die Tageszeit aufgeteilt und in die entsprechenden Spalten eingefügt. Dies geschieht über die „substr“-Funktion von SQLite.

```
INSERT INTO DimTime (PK_Time, Date, Time)
SELECT TimeID, substr(TimeID, 1, 3), substr(TimeID, 4)
FROM FactTable
GROUP BY TimeID;
```

Darauffolgend wird das Datum von dem Julian Day Format in ein klassisches, lesbares Datumsformat umgewandelt. Hierbei muss zudem der spezifische Startzeitpunkt berücksichtigt werden.

```
UPDATE DimTime SET Date = date(Date + julianday('2008-12-31'));
```

Danach wird auch die Stundenkodierung in ein lesbares Format umgewandelt. Um die beiden einzelnen Spalten mit dem Datum und der Tageszeit später zu einem richtigen Zeitstempel kombinieren zu können, müssen dabei die in SQLite möglichen Datetime-Formate berücksichtigt werden. Diese sehen beispielsweise folgendermaßen aus: „YYYY-MM-DD HH:MM:SS“. Jede einzelne Codierung wird dann durch die zugehörige Zeitangabe ersetzt.

```
UPDATE DimTime SET Time = '00:00' WHERE Time = 1;
UPDATE DimTime SET Time = '00:30' WHERE Time = 2;
UPDATE DimTime SET Time = '01:00' WHERE Time = 3;
…
UPDATE DimTime SET Time = '23:30' WHERE Time = 48;
```

Nach erfolgreicher Decodierung können die Strings für das Datum (Date) und die Tageszeit (Time) dann zusammengeführt und in ein richtiges Datetime-Format umgewandelt werden. Der so entstehende Zeitstempel wird dann in die vierte Spalte der Dimensionstabelle eingefügt.

```
UPDATE DimTime SET DateTime = datetime(Date || ' ' || Time);
```

Somit sind die Time IDs erfolgreich in Zeitstempel decodiert. Diese Zeitstempel stellen dann die niedrigste Hierarchieebene der Zeitdimension dar, da sie die maximale Granularität aufweisen. Anschließend kann die Zeitdimensionstabelle dann weiter transformiert werden. Die Hilfsspalten Date und Time haben ihren Zweck für die Decodierung der Time IDs erfüllt und können daher gelöscht werden. Anstelle dessen werden die Spalten für die verschiedenen Hierarchiestufen Datum (Date), Monat (Month), Quartal (Quarter) und Jahr (Year) hinzugefügt. Die Tabelle muss also entsprechend angepasst werden.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                          FROM DimTime;

DROP TABLE DimTime;

CREATE TABLE DimTime (
    PK_Time  INTEGER  PRIMARY KEY,
    Year     STRING,
    Quarter  STRING,
    Month    STRING,
    Date     DATE,
    DateTime DATETIME
);

INSERT INTO DimTime (
                        PK_Time,
                        DateTime
                    )
                    SELECT PK_Time,
                           DateTime
                      FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Die Spalten der verschiedenen Hierarchiestufen werden dann aufbauend auf und ausgehend von der Stufe mit der maximalen Granularität, den Zeitstempeln, mit Werten gefüllt.

```
UPDATE DimTime SET Date = date(DateTime);
UPDATE DimTime SET Month = strftime('%Y %m', DateTime);
UPDATE DimTime SET Year = strftime('%Y', DateTime);
UPDATE DimTime SET Quarter = '2009 3' WHERE Month BETWEEN '2009 07' AND '2009
09';
UPDATE DimTime SET Quarter = '2009 4' WHERE Month BETWEEN '2009 10' AND '2009
12';
UPDATE DimTime SET Quarter = '2010 1' WHERE Month BETWEEN '2010 01' AND '2010
03';
…
UPDATE DimTime SET Quarter = '2010 4' WHERE Month BETWEEN '2010 10' AND '2010
12';
```

Nach dem Abschluss aller Integrations- und Transformationsvorgänge und dem vollständigen Aufbau der Dimensionstabelle Zeit wurden jedoch sowohl bei Abfragen als auch beim weiteren Aufbau der Datenbank Fehler im Datensatz deutlich. Die Faktentabelle selbst enthält zwar keine NULL Werte, jedoch wird bei einer Abfrage der Zeitdimensionstabelle nach den Primärschlüsseln und den dazugehörigen Zeitstempeln deutlich, dass nicht jeder Primärschlüssel erfolgreich in einen Zeitstempel decodiert werden konnte.

```
SELECT PK_Time, DateTime
FROM DimTime
WHERE DateTime IS NULL;
```

In Folge der vorangegangenen Abfrage wurde dann deutlich, dass 284 der 26010 verschiedenen Zeitwerte als Zeitstempel NULL Werte darstellen. Dies war bereits beim Aufbau der Zeitdimensionstabelle verdächtig, da das Pilotprojekt über 536 Tage lief und täglich 48 Messungen durchgeführt wurden, wodurch maximal 25728 verschiedene Zeitwerte möglich wären, jedoch 26010 vorliegen. Bei näherer Untersuchung stellte sich heraus, dass für Meter ID 1208 und 5221 Time IDs existieren, deren Stundenkodierung über die 24 Tagesstunden hinausgeht. Dies ist für insgesamt 6 aufeinanderfolgende Tage der Fall. An den Tagen 297 bis 302 weisen die beiden Meter IDs Stundenkodierungen von 1 bis 95 anstatt 1 bis 48 auf. Zudem liegen für die Stundenkodierung 48 zwei unterschiedliche Messungen vor. Dies gilt es zu bereinigen. Da alle weiteren Messungen dieser beiden Meter IDs jedoch keine Fehler aufweisen besteht kein Grund, alle Werte dieser Smart Meter zu löschen. Aufgrund der doppelt vorliegenden Werte bei Stundenkodierung 48 fehlt jedoch nach dem Löschvorgang für diese beiden Smart Meter an 6 Tagen der letzte Messwert. Ferner existieren die fehlerhaften Time IDs 66949 und 66950, die ebenfalls eine über 24 Tagesstunden hinausgehende Kodierung aufweisen. Die logisch darauffolgenden Zeitwerte 67001 und 67002 existieren jedoch ebenfalls, wodurch sich die Vermutung, dass die fehlerhaften Werte ausnahmsweise anders kodiert sind, nicht bestätigt. Von dieser fehlerhaften Time ID sind dieses Mal allerdings alle Meter betroffen. Folglich müssen sowohl die Zeitdimensionstabelle als auch die Faktentabelle von diesen fehlerhaften Werten und den dazugehörigen Zeilen bereinigt werden. Vor der Eingabe eines DELETE Befehls empfiehlt es sich allerdings abschließend noch einmal die Anzahl der Zeilen, die durch den Befehl gelöscht werden, zu überprüfen. Für die Dimensionstabelle Zeit ergeben sich die folgenden Befehle.

```
SELECT count(*) FROM DimTime WHERE DateTime IS NULL;
```

Diese Abfrage gibt als Ergebnis 284 aus, was genau der zuvor händisch errechneten Fehlermenge entspricht. Somit können diese Daten problemlos gelöscht werden.

```
DELETE FROM DimTime WHERE DateTime IS NULL;
```

Für die Faktentabelle müssen mehrere einzelne DELETE Befehle eingegeben werden, um die betroffenen Time IDs und Meter IDs genau zu identifizieren. In der Faktentabelle selbst liegen nämlich keine NULL Werte vor, da die fehlerhafte Kodierung und die Redundanzen erst durch die Dekodierung in Zeitstempel in der Zeitdimensionstabelle erkannt wurden.

```
DELETE FROM FactTable WHERE MeterID = 1208 AND TimeID = 29748 OR MeterID = 5221
AND TimeID = 29748;
DELETE FROM FactTable WHERE MeterID = 1208 AND TimeID = 29848 OR MeterID = 5221
AND TimeID = 29848;
DELETE FROM FactTable WHERE MeterID = 1208 AND TimeID = 29948 OR MeterID = 5221
AND TimeID = 29948;
DELETE FROM FactTable WHERE MeterID = 1208 AND TimeID = 30048 OR MeterID = 5221
AND TimeID = 30048;
DELETE FROM FactTable WHERE MeterID = 1208 AND TimeID = 30148 OR MeterID = 5221
AND TimeID = 30148;
DELETE FROM FactTable WHERE MeterID = 1208 AND TimeID = 30248 OR MeterID = 5221
AND TimeID = 30248;
```

Die doppelt vorliegenden, spezifischen Werte der Meter IDs 1208 und 5221 sind nun gelöscht. Alle weiteren fehlerhaften Werte lassen sich dann einfach daran identifizieren, dass ihre Stundenkodierung über 48 hinaus geht. Folglich können sowohl die verbliebenen fehlerhaften Werte der Meter IDs 1208 und 5221 als auch die allgemein für alle Meter IDs fehlerhaften spezifischen Time IDs 66949 und 66950 mit dem gleichen Befehl auf einmal gelöscht werden.

```
DELETE FROM FactTable WHERE substr(TimeID, 4) BETWEEN '49' AND '95';
```

Dies betrifft insgesamt 19524 Zeilen. Nun liegen in keiner Tabelle mehr fehlerhaften Werte vor.

**Dimension Gebäudetypologie (DimBuildType)**

Als nächstes wurde die Dimension Gebäudetypologie aufgebaut. Die Gebäudetypologie ist eine räumliche, keine zeitliche Dimension und wird daher anhand der Meter ID der Gebäude zugeordnet und verknüpft. Ein Teil der Daten über die Gebäudetypologien der einzelnen Gebäude liegt als eigene Datei vor, während der andere Teil in den Umfrageergebnissen dokumentiert ist. Daher erfolgt in diesem Fall die Transformation und Aufbereitung nicht in der Datenbank, wie dies aufgrund der Zeit-Funktionen von SQLite und dem Aufbau der Dimensionstabelle anhand der Time IDs aus der Faktentabelle bei der Zeitdimension der Fall war, sondern außerhalb der Datenbank. Insgesamt wurde in dem Pilotprojekt zwischen drei verschiedenen Arten von Gebäuden unterschieden: „Residential“, „SME“ und „Other“. Alle Gebäude der Art „Residential“ sind Privathaushalte, während „SME“ für Geschäftsgebäude steht, also beispielsweise Dienstleister, Bauunternehmen, soziale Einrichtungen oder Werkstätten. Der Gebäudetyp „Other“ wird in der Dokumentation des Pilotprojektes nicht näher erläutert und ist auch nicht Teil der Umfragen. Jedoch könnte davon ausgegangen werden, dass es sich dabei vielleicht um öffentliche Gebäude wie Bibliotheken oder Bildungseinrichtungen handelt. Darüber hinaus werden in den Umfragedaten der Privathaushalte und Geschäftsgebäude Informationen über die genaue Art des Gebäudes, Respektive des Geschäfts gegeben.

Im ersten Schritt wird daher die Excel Tabelle mit den Informationen über die Gebäudetypologien und Tarife aufbereitet. Neben der ersten Spalte mit den Meter IDs wird eine zusätzliche Spalte eingefügt. Die Gebäudetypologien sind anhand einer separaten Tabelle auf dem Excel Blatt kodiert. Dabei ist jedem der drei übergeordneten Gebäudetypen eine Nummer zugeordnet. Da die reine Kodierung für einen Endanwender nicht lesbar ist, muss stattdessen der eigentliche Wert eingefügt werden. Über eine SVERWEIS-Funktion werden die dekodierten Werte eingefügt. Neben der ersten Spalte mit den Meter

IDs gibt so die zweite Spalte den entsprechenden dazugehörigen Gebäudetyp an. Darüber hinaus gibt es eine weitere, feinere Hierarchieebene, die neben dem übergeordneten Gebäudetyp, also Wohngebäude, Geschäftsgebäude oder Sonstige, weitere Informationen zu den jeweiligen Untertypen enthält. Diese Informationen wurden aus den Umfrageergebnissen gewonnen und müssen daher zunächst in die Datei integriert werden. Für Wohngebäude stehen die Untertypen Apartment, Doppelhaus, Einfamilienhaus, Reihenhaus, Bungalow und keine Angabe zur Auswahl. Für Geschäftsgebäude liegen keine Gebäudedaten im Sinne der Wohngebäudetypen vor. Stattdessen liegen Informationen über die Art der Nutzung, also die Art des Geschäfts, das darin betrieben wird, vor. Insgesamt stehen sieben verschiedene Kategorien zur Auswahl. Um auch diese weitere Hierarchiestufe zu integrieren, wurden aus den jeweiligen Excel-Tabellen mit den Umfrageergebnissen zunächst alle nicht relevanten Spalten gelöscht, sodass nur noch die Spalten mit den Meter IDs und den Ergebnissen der spezifischen Frage zum Gebäudetyp übrig bleiben. Diese Ergebnisse sind sowohl für Wohn- als auch für Geschäftsgebäude kodiert. Die Kodierungen erfolgen anhand der Aufschlüsselungen in den Word-Dokumenten mit den Fragen und Antwortmöglichkeiten der Umfragen. Die beiden Spalten mit den Meter IDs und Untertypen wurden dann in das bestehende Excel Dokument kopiert. Weiterhin wurden die zugehörigen Tabellen mit den Aufschlüsselungen der Kodierung erstellt. Über eine Kombination der Funktionen SVERWEIS, WENN UND WENNFEHLER konnten dann alle Werte entsprechend in eine einzige, gemeinsame Tabelle zusammengeführt werden. Auch wenn für alle Meter IDs in der Excel-Datei angegeben war, ob es sich um Wohngebäude, Geschäftsgebäude oder Sonstige handelt, so lagen in den Umfrageergebnissen für einige dieser Meter IDs keine Angaben über die zugehörigen Untertypen vor. In diesem Fall wurde über die WENNFEHLER-Funktion ein "Unknown", also "unbekannt" eingefügt. Die drei finalen Spalten MeterID, BuildType und BuildTypeDetail wurden dann als CSV-Datei gespeichert. Diese CSV-Datei beinhaltet dann die bereits vollständig transformierten Gebäudetypologien und steht für den Import in das DW bereit.

Zunächst wird dafür die Dimensionstabelle erstellt.

```
CREATE TABLE DimBuildType (
    PK_BuildType    INTEGER PRIMARY KEY,
    BuildType       STRING,
    BuildTypeDetail STRING
);
```

Dann wird die zuvor erwähnte CSV-Datei in diese Tabelle importiert. Allerdings muss noch eine letzte Transformation vorgenommen werden, die durch Excel nicht ohne weiteres durchführbar war. Auf der niedrigsten Hierarchieebene stehen nun die Untertypen der drei Gebäudetypologien. Allerdings geben diese Untertypen bei alleiniger Betrachtung keinen Aufschluss auf die übergeordnete Gebäudetypologie. Diese Zuordnung ist nur bei Kenntnis des Datensatzes und durch eigenständige Herleitung möglich. Nach dem Import der Datei sollten daher die Spalten so zusammengeführt werden, dass die Untertypen die übergeordnete Gebäudetypologie mitführen. Daher wird die untere Hierarchieebene so aktualisiert, dass sie auch die Werte der oberen Hierarchieebene beinhaltet und die typische, de-normalisierte Form einer Dimensionstabelle eines Sternschemas aufweist.

```
UPDATE DimBuildType SET BuildTypeDetail = BuildType || '/' || BuildTypeDetail;
```

Aufgrund der zuvor erwähnten Probleme beim Import der Messwerte der Meter 6000 bis 7444 empfiehlt es sich anschließend, die nicht relevanten Meter IDs zu löschen. Dies reduziert den Speicherplatzbedarf und verbessert die Performance.

```
DELETE FROM DimBuildType WHERE PK_BuildType > 5999;
```

Wird die Dimensionstabelle Gebäudetypologie (DimBuildType) auf diese Weise erzeugt, ist es möglich, allen Messwerten eine entsprechende Gebäudetypologie zuzuordnen. Somit kommt es bei Datenabfragen zu keinerlei technischen Problemen oder Fehlern. Allerdings entspricht diese Art der Datenintegration theoretisch nicht genau der OLAP-Definition. Die Verwendung der Meter IDs als Primärschlüssel ist in dem Sinne nicht sauber, als dass die Anzahl der Zeilen der Dimensionstabelle eigentlich gleich der Anzahl an Dimensionswerten sein sollte. Nur so wäre auch eine Speicherung der Daten als n-dimensionales Array sauber möglich. Die Dimension selbst weist eigentlich nur 15 verschiedene Dimensionswerte auf, in der hier erstellten Dimensionstabelle werden jedoch sämtliche Meter IDs aufgezählt, sodass die Anzahl der Zeilen der Anzahl der Meter IDs entspricht. Eigentlich müssten, anstatt die Meter IDs als Primärschlüssel zu verwenden, 15 neue Primärschlüssel erzeugt werden, beispielsweise durch automatisches Hochzählen. Bei einer solchen Umsetzung ist jedoch die anschließende Zuordnung der Primärschlüssel zu den Messwerten der Faktentabelle komplizierter und aufwendiger, bietet dafür aber eine bessere Abfrageperformance und einen geringeren Speicherplatzbedarf. Dennoch wurden der Aufbau der Dimensionstabelle hier zu Demonstrationszwecken auf die einfache, schnellere, jedoch etwas unsaubere Art durchgeführt. Eine sauberere Art, dies zu lösen, wird jedoch während des Aufbaus der DWs für die Hypothesen 2 und 3 dargestellt. Diese verwendet Hilfstabellen, um daraus sowohl die eigentliche Dimensionstabelle aufzubauen als auch deren Primärschlüssel der Faktentabelle als Fremdschlüssel zuzuordnen.

**Dimension Wochentag (DimWeekday)**

Die Wochentage ergeben sich aus dem Datum. Um aus einem vorgegebenen Datum einen zugehörigen Wochentag zu extrahieren, existiert in SQLite eine Funktion. Dies erspart viel Arbeit und verhindert, dass ein eigenes Programm geschrieben werden muss, dass die Daten einließt, mit einer Schleife durchläuft, das Datum aus dem Zeitstempel extrahiert und entsprechend eines Algorithmus die zugehörigen Wochentage zuordnet. Folglich erfolgt die Erstellung und Transformation dieser Dimension erneut im DW selbst.

Die Datenquelle für die Wochentage ist also die Zeitdimensionstabelle DimTime. Wichtig für die spätere Zuordnung ist allerdings auch, dass die Wochentage einem entsprechenden vollständigen Zeitstempel zugeordnet werden können, da nur so später eine Zuordnung der Wochentage zu den Messwerten der Faktentabelle möglich ist. Es existieren nur sieben verschiedene Wochentage. Daher wird diese Dimensionstabelle nur aus drei Spalten für die Primärschlüssel, die Teile der Woche (Arbeitstag oder Wochenende) und die Wochentage als String und sieben Zeilen, eine für jeden Wochentag, bestehen. Dies benötigt so gut wie keinen Speicherplatz und wirkt sich sehr positiv auf die Abfragegeschwindigkeit aus. Daher wäre es ungünstig, zur Zuordnung die Primärschlüssel der Zeitdimension DimTime zu verwenden, da diese die über 25.000 Zeitstempel des Gesamtzeitraums des Pilotprojektes abdecken. Die Dimensionstabelle kann aufgrund ihrer geringen Größe daher theoretisch

von Hand mit Werten gefüllt werden. Alternativ können die Zeitstempel in Wochentage umgewandelt und nach Wochentagen gruppiert werden. In diesem Fall wird die zweite Variante angewandt.

Im ersten Schritt wird die Dimensionstabelle DimWeekday erzeugt.

```
CREATE TABLE DimWeekday (
    PK_Weekday INTEGER PRIMARY KEY,
    PartOfWeek STRING,
    Weekday    STRING
);
```

Im zweiten Schritt wird diese dann mit Hilfe der Dimensionstabelle DimTime mit Werten gefüllt. Zunächst werden dabei die Primärschlüssel erzeugt. SQLite bietet für die Umwandlung von Zeitformaten die Funktion „strftime“.

```
INSERT INTO DimWeekday (PK_Weekday)
SELECT strftime('%w', Date)
FROM DimTime
GROUP BY strftime('%w', Date);
```

Im dritten Schritt werden dann aufbauend auf den Primärschlüsseln die zwei weiteren Spalten mit Werten gefüllt.

```
UPDATE DimWeekday SET Weekday = 'Sunday' WHERE PK_Weekday = 0;
UPDATE DimWeekday SET Weekday = 'Monday' WHERE PK_Weekday = 1;
UPDATE DimWeekday SET Weekday = 'Tuesday' WHERE PK_Weekday = 2;
…
UPDATE DimWeekday SET Weekday = ‘Saturday’ WHERE PK_Weekday = 6;
UPDATE DimWeekday SET PartOfWeek = 'Workweek' WHERE PK_Weekday BETWEEN 1 AND 5;
UPDATE DimWeekday SET PartOfWeek = 'Weekend' WHERE PK_Weekday = 0 OR PK_Weekday
= 6;
```

**Dimension Ort (DimLocation)**

Zu Demonstrationszwecken wird als weitere Dimension auch der Ort hinzugefügt. Da die Smart Meter Daten anonymisiert sind und keine Ortsangaben enthalten, handelt es sich dabei aber um fiktive Werte. Dafür wird eine Datei verwendet, in der bereits die geordneten Meter IDs vorliegen. Anschließend werden diesen dann fiktive Werte zugeordnet. Die Zuordnung kann beispielsweise über Excel und WENN-Funktionen, SVERWEIS-Funktionen oder Zellenverweise gelöst werden. Anschließend müssen die entsprechenden Spalten dann als CSV-Datei gespeichert werden. Es wurde entschieden, zufällige Städte nach den Meter IDs geordnet hinzuzufügen. Als Meter IDs liegen Werte von 1000 bis 5999 vor. Für jeden tausender Schritt wird eine neue irische Stadt hinzugefügt. So entstehen 5 verschiedene Städte.

- Meter IDs 1000-1999: Dublin
- Meter IDs 2000-2999: Cork
- Meter IDs 3000-3999: Galway
- Meter IDs 4000-4999: Limerick
- Meter IDs 5000-5999: Waterford

Diese Datei steht dann für den Import in das DW bereit. Hierfür muss wieder eine Dimensionstabelle erzeugt werden.

```
CREATE TABLE DimLocation (
    PK_Location INTEGER,
    City        STRING
);
```

Anschließend kann der Import erfolgen. Für diese über die Meter IDs verknüpfte Dimension gelten jedoch die gleichen Anmerkungen wie bereits für die Dimension DimBuildType.

**Verknüpfung der Faktentabelle mit den Dimensionstabellen**

Sind alle Dimensionstabellen erfolgreich erstellt und transformiert, muss in einem letzten Schritt noch die Faktentabelle mit den Dimensionstabellen verknüpft werden. Dies geschieht, indem die Faktentabelle über Fremdschlüsselspalten die Primärschlüssel der Dimensionstabellen den eigenen Fakten zuordnet und diese dann als Fremdschlüssel speichert. Im ersten Schritt müssen diese Fremdschlüsselspalten der Faktentabelle daher zunächst erzeugt werden. Die Spalten MeterID und TimeID der Faktentabelle können dabei bereits als Fremdschlüsselspalten für die Dimensionstabellen DimTime und DimBuildType verwendet werden. Sie sollten dafür lediglich entsprechend umbenannt werden. Des Weiteren müssen die Fremdschlüsselspalten für DimWeekday und DimLocation erzeugt werden.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                        FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage       DOUBLE,
    BuildTypeFK INTEGER,
    TimeFK      INTEGER,
    LocationFK  INTEGER,
    WeekdayFK   INTEGER
);

INSERT INTO FactTable (
                          Usage,
                          BuildTypeFK,
                          TimeFK,
                      )
                      SELECT Usage,
                             MeterID,
                             TimeID
                        FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Anschließend können den neu erzeugten Fremdschlüsselspalten dann die entsprechenden Fremdschlüssel zugeordnet werden. Im Fall der Dimension Ort (DimLocation) sind der Fremdschlüssel die Meter IDs und folglich gleich der bereits existierenden Fremdschlüssel der Dimension

DimBuildType. Die Fremdschlüssel der Dimension Wochentag basieren auf der zeitlichen Dimension (DimTime) und können daher auch direkt aus den bereits existierenden Fremdschlüsseln für die Zeitdimension erzeugt bzw. extrahiert werden. Dadurch, dass die Fremdschlüssel anhand eines anderen, bereits bestehenden Fremdschlüssels erzeugt werden, wird verhindert, dass die Primärschlüssel der Dimensionstabelle Wochentag über aufwendige Tabellenverknüpfungen und Verweise den Fakten zugeordnet werden müssen. Aufgrund der großen Faktentabelle wäre dies sehr rechenintensiv. Möglich ist dies dadurch, dass die Primärschlüssel der Dimensionstabelle Wochentag ebenfalls bereits auf Basis der Time IDs, also der Primärschlüssel der Dimensionstabelle Zeit erzeugt wurden. Dieser Vorgang wird dann schlicht wiederholt, um diese Primärschlüssel erneut in der Faktentabelle zu erzeugen, nur dass diese diesmal Fremdschlüssel darstellen.

```
UPDATE FactTable SET LocationFK = BuildTypeFK;
UPDATE FactTable SET WeekdayFK = strftime('%w', (substr(TimeID, 1, 3) +
julianday('2008-12-31')));
```

Nachdem sämtliche Dimensionstabellen erstellt, mit den entsprechenden Werten gefüllt, um Hierarchiestufen erweitert, Fehler in den Daten und nicht relevante Daten entfernt sowie die Primärschlüssel der Dimensionstabellen an die Faktentabelle übergeben wurden, sollte auch die Faktentabelle abschließend überprüft und eventuell angepasst werden. Darüber hinaus sollten insbesondere bei schlechter Kenntnis des Datensatzes und fehlendem Wissen über eventuelle Inkonsistenzen oder Fehler in den Daten entsprechende Definitionen und Bedingungen festgelegt werden, um ebendiese Fehler aufzudecken. Besonders das Verknüpfen von Fakten- mit Dimensionstabellen anhand der Primär- und Fremdschlüssel kann aufgrund des Prozesses rund um das Hinzufügen und Speichern der entsprechenden Fremdschlüssel in der Faktentabelle sehr fehleranfällig sein. Folglich ist es wichtig, die Primärschlüssel der Dimensionstabelle sorgfältig zu wählen. Hierbei besteht ständig der Konflikt zwischen Einfachheit, Übersichtlichkeit und Schnelligkeit der Datenintegration auf der einen Seite und Sicherheit und Sauberkeit auf der anderen Seite. Je nach Datenlage und Dimension kann es sich empfehlen oder ist es sogar zwingend erforderlich, dass entsprechende Primärschlüssel neu generiert werden. Ein neu erzeugter, automatisch hochzählender Primärschlüssel ist eindeutig und leicht verständlich, während bei der Verwendung einer bereits existierenden Datenspalte als Primärschlüssel darauf geachtet werden muss, dass diese eindeutig sind, keine NULL Werte enthalten und tatsächlich als Identifikationsnummer fungieren können. Dementsprechend muss diese Spalte ausschließlich Integer, also ganzzahlige Werte enthalten. Werden allerdings die Informationen der als Primärschlüssel erdachten, bereits existierenden Spalte weiterhin für Analysen benötigt und dennoch ein neuer, automatisch hochzählender Primärschlüssel erzeugt, so kann dies insbesondere bei großen Dimensionstabellen den Speicherplatzbedarf bedeutend erhöhen. In allen Dimensionstabellen wurden die entsprechenden Spalten bereits als Primärschlüsselspalten definiert. Im Folgenden wird daher abschließend nur noch die Faktentabelle konfiguriert. Alle Spalten mit Fremdschlüsseln wurden bereits entsprechend benannt und sollten nun auch als solche definiert werden. Die existierenden Spalten MeterID und TimeID wurden entsprechend in BuildTypeFK und TimeFK umbenannt. Eindeutigkeits-Bedingungen („unique-constraints“) sind allerdings nur für die Primärschlüssel der Dimensionstabellen relevant und hier bereits durch die Definition der Spalten als Primärschlüsselspalten abgedeckt. Eine Dimensionstabelle und die Faktentabelle weißen eine 1 zu n Beziehung auf, was bedeutet, dass ein einzelner Wert der Dimensionstabelle mehrmals in der

Faktentabelle vorkommen kann. Beispielsweise weißt eine Meter ID mehrere Messwerte auf und zu einem spezifischen Zeitpunkt (sprich Zeitstempel) werden von mehreren Smart Metern Energieverbräuche gemessen. Eine zusätzliche Definition der Fremdschlüsselspalten der Faktentabelle als eindeutig („unique“) wäre daher fehlerhaft. Stattdessen sollte aber sichergestellt werden, dass kein Fremdschlüssel den Wert NULL hat. Der Befehl zur endgültigen Formatierung der Faktentabelle sieht daher folgendermaßen aus.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                          FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage       DOUBLE,
    BuildTypeFK INTEGER REFERENCES DimBuildType (PK_BuildType)
                        NOT NULL ON CONFLICT FAIL,
    TimeFK      INTEGER REFERENCES DimTime (PK_Time)
                        NOT NULL ON CONFLICT FAIL,
    LocationFK  INTEGER REFERENCES DimLocation (PK_Location)
                        NOT NULL ON CONFLICT FAIL,
    WeekdayFK   INTEGER REFERENCES DimWeekday (PK_Weekday)
                        NOT NULL ON CONFLICT FAIL
);

INSERT INTO FactTable (
                           Usage,
                           BuildTypeFK,
                           TimeFK,
                           LocationFK,
                           WeekdayFK
                       )
                       SELECT Usage,
                              BuildTypeFK,
                              TimeFK,
                              LocationFK,
                              WeekdayFK
                         FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Das DW wäre damit vollständig erstellt und stünde für alle möglichen Analysen und Abfragen zur Verfügung.

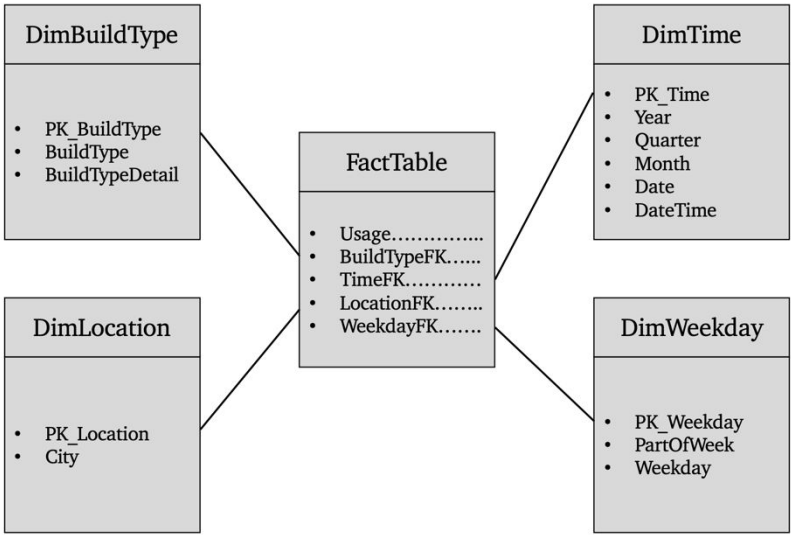



Abbildung 19: Data Warehouse für Hypothese 1

### 5.2.2. Abfrage der Daten aus dem Data Warehouse

Nun können die Daten des DW abgefragt werden. Für die Untersuchung der Analyse ist in erster Linie der Verlauf des monatlichen Stromverbrauchs über das Jahr hinweg relevant. Daher sollten die Verbrauchsdaten hinsichtlich jedes Monats aggregiert werden. Ferner ergibt sich aus der Hypothese auch, dass zwischen den verschiedenen Gebäudetypen unterschieden werden soll. Folglich sollten bei der Aggregation der Verbrauchsdaten auch die Gebäudetypen berücksichtigt werden. Zudem empfiehlt sich eine Unterscheidung zwischen den Tagen unter der Woche und dem Wochenende, da die Nichtwohngebäude am Wochenende nicht oder nur bedingt genutzt werden. Der Verbrauch wird dann, um Vergleichbarkeit zu gewährleisten, als Durchschnitt sämtlicher, der Aggregationsgruppe zugehöriger Verbrauchsdaten ausgegeben. Da unterschiedlich viele Wohn- und Nichtwohngebäude existieren wäre eine Ausgabe der Summe zwar auch in Ordnung, weil in erster Linie an dem Verlauf selbst besteht, es würde sich jedoch nichts über die Höhe der Stromverbräuche der verschiedenen Gebäudetypen im gegenseitigen Vergleich aussagen lassen. Um diese Vergleichbarkeit zu gewährleisten, beschreibt der Durchschnitt also den durchschnittlichen Stromverbrauch eines 30-minütigen Messintervalls für sämtliche dazugehörigen Gebäude. Ausgegeben wird daher der Durchschnittsverbrauch, der Teil der Woche, der Gebäudetyp und der Monat (inklusive des Jahres). Außerdem muss die Faktentabelle mit den entsprechenden Dimensionstabellen verknüpft werden. Dies geschieht über die Bedingung, dass die Fremdschlüssel der Faktentabelle den Primärschlüsseln der Dimensionstabelle entsprechen. Die finale Datenabfrage hat dann die folgende Form.

```
SELECT avg(Usage), PartOfWeek, BuildType, Month
FROM FactTable, DimWeekday, DimBuildType, DimTime
WHERE FactTable.WeekdayFK = DimWeekday.PK_Weekday
AND FactTable.BuildTypeFK = DimBuildType.PK_BuildType
AND FactTable.TimeFK = DimTime.PK_Time
GROUP BY DimWeekday.PartOfWeek, DimBuildType.BuildType, DimTime.Month
```

Das Abfrageergebnis wurde dann als CSV-Datei exportiert.

### 5.2.3. Visualisierung der Abfrageergebnisse

Nun können die Daten visualisiert werden. Die Aufbereitung und Visualisierung der Abfrageergebnisse erfolgt anhand der Bibliotheken Pandas und Matplotlib in Python. Im ersten Schritt müssen dafür zunächst die Bibliotheken importiert werden. Das Einlesen der generierten CSV-Datei erfolgt dann durch Pandas, indem die Datei in einen Data Frame, eine zweidimensionale Datenmatrix, eingelesen wird. Um die Messwerte zu visualisieren müssen diese jedoch nochmals formatiert werden. Die SQL-Abfrage gibt die Daten als Tabelle mit den vier Spalten Durchschnittsverbrauch, Teil der Woche, Gebäudetyp und Monat aus. Um die Daten zu visualisieren, müssen diese also zuerst in eine relationale Form gebracht werden, sodass Werte für die X- und Y-Achse des Diagramms vorliegen. Da Interesse an dem Verlauf des Stromverbrauchs über die Zeit besteht, dienen die Monate als X-Achse, während der Durchschnittsverbrauch auf der Y-Achse dargestellt wird. Folglich werden beim Import der CSV-Datei die Monate als Index festgelegt. Dabei sollten diese explizit als Datumswert deklariert werden.

```
# Bibliotheken importieren
import matplotlib.pyplot as plt
import pandas as pd

# Datei einlesen
df = pd.read_csv('/dateipfad/dateiname.csv', index_col = 'Month',
parse_dates=['Month'])
```

Die erste Dimension wird also anhand der X-Achse dargestellt. Die zweite Dimension mit den Gebäudetypen wird dann repräsentiert, indem die Verbrauchsdaten der verschiedenen Gebäudetypen in dieses Koordinatensystem eingezeichnet werden. Um zudem die dritte Dimension, die Teile der Woche, zu berücksichtigen, wird für jeden Dimensionswert dieser Dimension ein separates Diagramm erstellt. Dadurch entsteht ein Diagramm für die Tage unter der Woche und ein Diagramm für das Wochenende. Für die Datenaufbereitung bedeutet dies, dass zwei neue Data Frames aufgebaut werden müssen, die in der ersten Spalte die Indexwerte und in allen weiteren Spalten die Verbrauchswerte der verschiedenen Gebäudetypen aufweisen. Zuerst werden die Verbrauchsdaten der Arbeitstage untersucht. Daher gilt es zunächst, die Daten aus dem vorliegenden Data Frame zu filtern und zu einem neuen Data Frame zu verknüpfen.

```
# Messwerte der Wochentage und Gebäudetypen filtern
commercial = df.loc[(df.BuildType == 'Commercial') & (df.PartOfWeek ==
'Workweek'), 'avg(Usage)']
residential = df.loc[(df.BuildType == 'Residential') & (df.PartOfWeek ==
'Workweek'), 'avg(Usage)']
other = df.loc[(df.BuildType == 'Other') & (df.PartOfWeek == 'Workweek'),
'avg(Usage)']

# Gefilterte Werte verknüpfen
usage = pd.concat([commercial, residential, other], axis=1)
usage.columns = ['Commercial', 'Residential', 'Other']
```

Anschließend können die Verbrauchsdaten visualisiert werden. Dies geschieht zunächst durch aufeinandergestapelte Graphen der verschiedenen Gebäudetypen.

```
# Messwerte der Wochentage visualisieren
plt.style.use('seaborn-white')
usage.plot.area(figsize=(10,7.5))
plt.title('Average monthly electricity consumption', fontsize = 15,
fontweight='bold')
plt.ylabel('kWh per 30 minutes', fontsize=12, fontweight = 'bold')
plt.xlabel('Date', fontsize=12, fontweight='bold')
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.show()
```

Das Ergebnis sieht dann folgendermaßen aus.

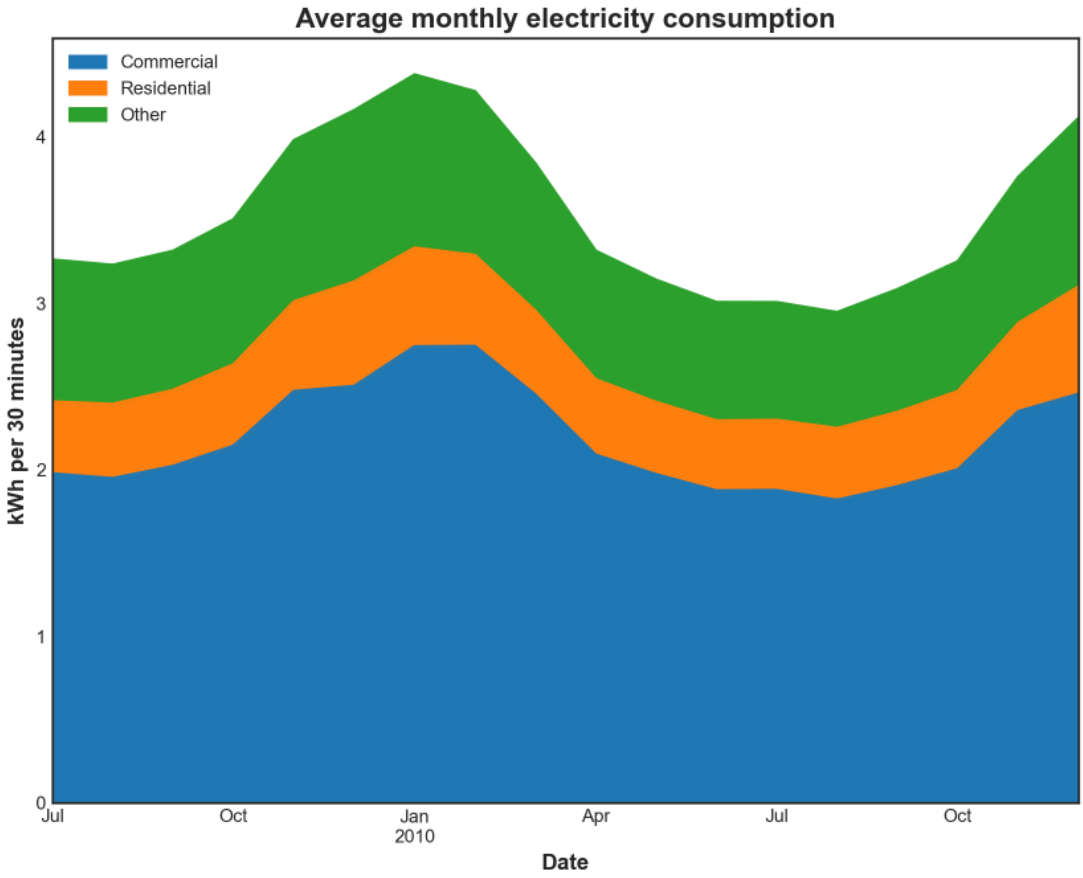



Abbildung 20: Durchschnittlicher monatlicher Stromverbrauch an Arbeitstagen

Es ist sofort erkennbar, dass der Durchschnittsverbrauch der Geschäftsgebäude an Arbeitstagen etwa vier- bis fünfmal und damit deutlich höher ist als Durchschnittsverbrauch der Wohngebäude. Der Durchschnittsverbrauch der sonstigen Gebäude liegt dabei zwischen diesen beiden Gebäudetypen. Zudem ist ein klarer saisonaler Verlauf des gesamten Stromverbrauchs zu erkennen. Dabei ist der Stromverbrauch in den Wintermonaten klar höher als der Stromverbrauch in den Sommermonaten. Allerdings trifft dies nicht nur auf den Gesamtverbrauch, also die Summe aller Einzelverbräuche, sondern prinzipiell auch auf jeden einzelnen Gebäudetyp zu. Allerdings lassen sich in dieser Darstellung aufgrund der Stapelung keine wirklichen Unterschiede in dem saisonalen Verlauf der verschiedenen Gebäudetypen erkennen, sondern lediglich, dass diese alle prinzipiell einem saisonal verlaufenden Stromverbrauch folgen. Allerdings sticht bei genauer Betrachtung ein Detail ins Auge. Im Dezember 2009 weist der Stromverbrauch der Geschäftsgebäude eine „Delle“ auf. Gleichzeitig wird der Gesamtverbrauch von Geschäfts- und Wohngebäuden jedoch nicht beeinflusst unter verläuft weiterhin gleichmäßig, da der Verbrauch der Wohngebäude in diesem Monat ansteigt und somit die „Delle“ ausgleicht. Dies dürfte auf die Weihnachtsfeiertage zurückzuführen sein, an denen die Geschäfte prinzipiell geschlossen haben und die Menschen stattdessen zuhause sind. Insbesondere die Weihnachtsbeleuchtung und die Feiertagsgeschehnisse dürften in diesem Monat den Stromverbrauch der Wohngebäude nach oben treiben, was in den Daten gut widergespiegelt wird. Es ist dementsprechend in der Grafik erkennbar, dass ein Teil des Stromverbrauchs der Geschäftsgebäude

praktisch exakt von den Wohngebäuden „übernommen“ wird. Zunächst werden für eine Untersuchung der saisonalen Abweichungen die Standardabweichungen der Stromverbräuche der verschiedenen Gebäudetypen errechnet und dargestellt.

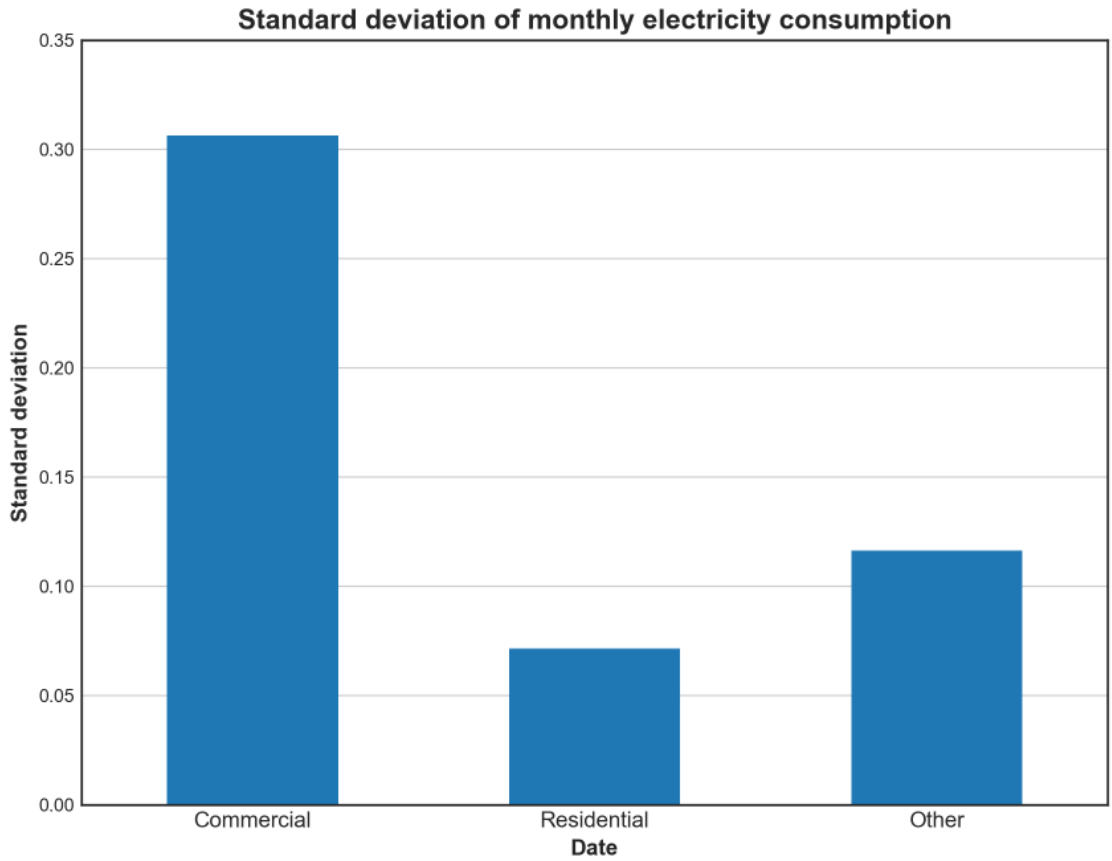



Abbildung 21: Standardabweichung des monatlichen Stromverbrauchs an Arbeitstagen

Es fällt auf, dass Nichtwohngebäude die höchste Standardabweichung aufweisen. Unüberlegt könnte man daher darauf schließen, dass der Stromverbrauch der Nichtwohngebäude über das Jahr hinweg am meisten schwankt und die Hypothese somit widerlegt wäre. Allerdings ist dieser Unterschied lediglich auf den insgesamt höheren Durchschnittsverbrauch zurückzuführen. Bei reiner Betrachtung der Diagramme und einem graphischen Vergleich scheint es, als ob alle Gebäudetypen in etwa die gleichen Schwankungen aufweisen, da die Höhe der Standardabweichung immer etwa proportional zur Höhe der Stromverbräuche ist. Daher empfiehlt es sich, die Daten zu normieren, um diese besser vergleichbar zu machen. Dafür wird ein neuer Data Frame aufgebaut, der alle Werte normiert, indem er diese durch den maximalen Wert des jeweiligen Gebäudetyps teilt. So ergeben sich immer Werte zwischen 0 und 1 und der Verbrauch wird nicht mehr in kWh, sondern als prozentualer Vergleich zu dem maximalen Verbrauch dargestellt.

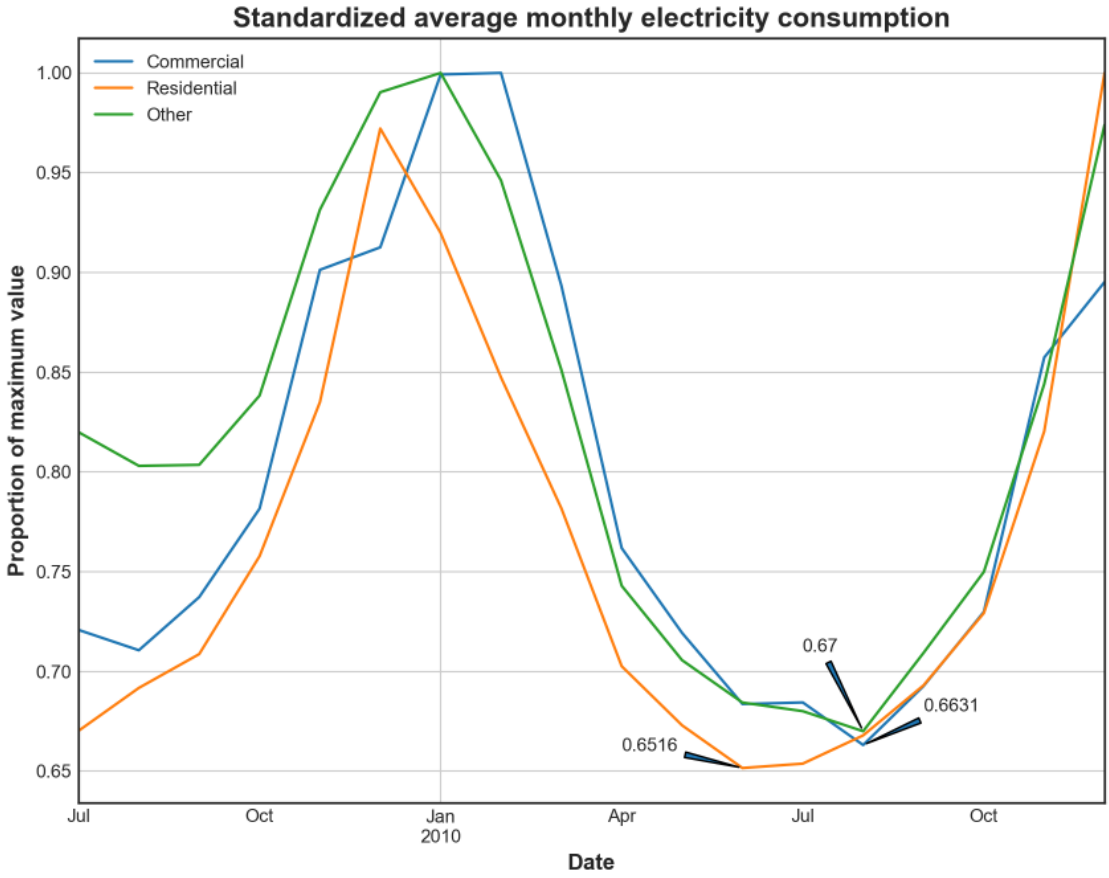



Abbildung 22: Normierter monatlicher Stromverbrauch an Arbeitstagen

Dabei ist erkennbar, dass die monatlichen Stromverbräuche sämtlicher Gebäudetypen prinzipiell dem genau gleichen Muster folgen. Unterschiede sind nur bei genauerer Betrachtung erkennbar. Zunächst fällt erneut der Verlauf in den Wintermonaten auf. Hier bestätigt sich die Vermutung aus der ersten Darstellung, da die Geschäftsgebäude und sonstigen Gebäude ihren maximalen Stromverbrauch im Januar und Februar haben, während die Wohngebäude im Dezember am meisten Strom verbrauchten. Auch ist hier erneut klar die „Delle“ der Geschäftsgebäude im Dezember erkennbar. Der Stromverbrauch der Wohngebäude ist im Dezember hingegen außergewöhnlich hoch. Interessant ist nun auch, in welchen Monaten die geringsten Stromverbräuche existieren und wie hoch diese im Vergleich zum maximalen Stromverbrauch sind. Die Wohngebäude weisen ihren minimalen Stromverbrauch im Juni auf. Dieser entspricht 65,16% des maximalen monatlichen Stromverbrauchs. Die Nichtwohngebäuden haben ihren minimalen Stromverbrauch mit 66,31% des maximalen Stromverbrauchs im August. Diese könnte auch damit zusammenhängen, dass der August ein beliebter Sommerurlaubsmonat ist. Auch der minimale Stromverbrauch der sonstigen Gebäude liegt im August. Dieser beträgt 67% des maximalen Stromverbrauchs. Somit bestehen zwar Unterschiede in den saisonalen Schwankungen der Stromverbräuche, diese sind jedoch kleiner Natur.

Für eine möglichst gute Analyse der saisonalen Abweichungen der Stromverbräuche wäre zudem ebenfalls interessant, ebendiese Abweichungen aus Sicht des Mittelwertes zu betrachten. Dies kann

anhand einer Normalisierung der Werte geschehen, indem zunächst die Differenz aus den monatlichen Verbrauchswerten und dem Durchschnitt dieser gebildet wird und diese dann zur Normierung durch die Standardabweichung geteilt wird.

```
# Messwerte der Wochentage normalisieren
com_norm = []
for i in usage['Commercial']:
    com_norm.append((i-
(usage['Commercial'].mean()))/(usage['Commercial'].std()))

res_norm = []
for i in usage['Residential']:
    res_norm.append((i-
(usage['Residential'].mean()))/(usage['Residential'].std()))

oth_norm = []
for i in usage['Other']:
    oth_norm.append((i-(usage['Other'].mean()))/(usage['Other'].std()))

# Neuen DataFrame aus normalisierten Werten aufbauen
usage_norm = pd.DataFrame({'Commercial': com_norm, 'Residential': res_norm,
'Other': oth_norm}, index=usage.index)
```

Dies ergibt folgende Darstellung.

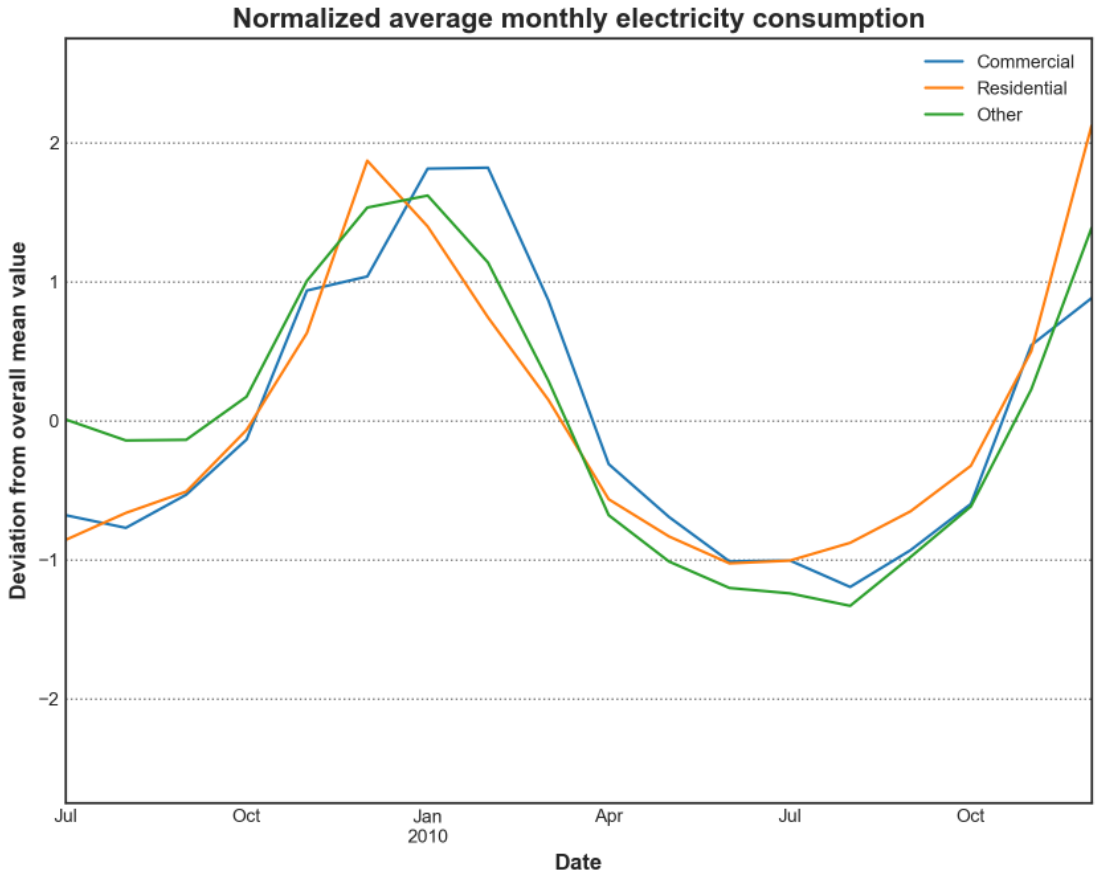



Abbildung 23: Normalisierter monatlicher Stromverbrauch an Arbeitstagen

Hier ergibt sich nämlich eine weitere interessante Erkenntnis. Die Abweichung des Stromverbrauchs nach oben ist deutlich größer als die Abweichung nach unten. Teilweise ist diese sogar beinahe doppelt so hoch. So weichen die monatlichen Stromverbräuche der Sommermonate um kaum mehr als eine Standardabweichung nach unten vom Gesamtdurchschnitt ab, während die der Wintermonate teilweise um fast zwei Standardabweichungen höher liegen als der Gesamtdurchschnitt. Dies deutet darauf hin, dass sich der Stromverbrauch in den Wintermonaten exponentiell erhöht und somit ein exponentieller und kein linearer Zusammenhang zwischen dem Stromverbrauch und den klimatischen Verhältnissen besteht. Gründe dafür könnten niedrigere Außentemperaturen, weniger Tageslicht, weniger Sonnenstunden und schlechteres Wetter in den Wintermonaten sein, was zusätzlich dazu führt, dass Bewohner mehr Zeit innerhalb des Hauses verbringen.

In diesem Zusammenhang sollten nun noch die monatlichen Stromverbräuche der Wochenenden untersucht werden. Das Vorgehen ist dabei genau gleich. Der einzige Unterschied besteht darin, dass nun nach dem Wochenende, nicht nach den Arbeitstagen, gefiltert wird.

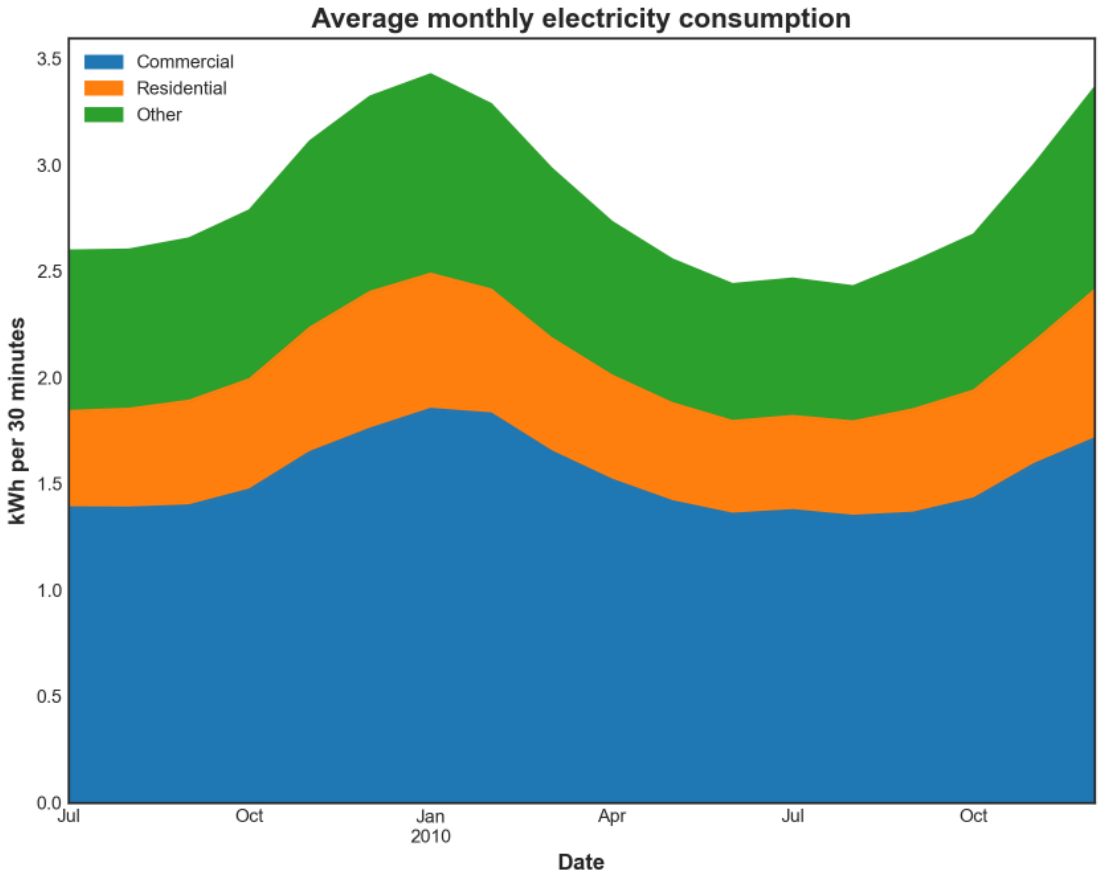



Abbildung 24: Durchschnittlicher monatlicher Stromverbrauch an Wochenenden

In diesem Fall sticht im Vergleich zu den Arbeitstagen gleich ins Auge, dass der Verbrauch der Geschäftsgebäude am Wochenende deutlich niedriger ist. Somit ist der Unterschied des Durchschnittsverbrauchs zu den Wohngebäuden nicht mehr so groß. Allerdings weisen die Geschäftsgebäude auch weiterhin den höchsten Durchschnittsverbrauch auf, gefolgt von sonstigen Gebäuden und zuletzt den Wohngebäuden. Dies spiegelt sich ebenfalls in der Standardabweichung wider, die nun für die Geschäftsgebäude deutlich geringer ist.

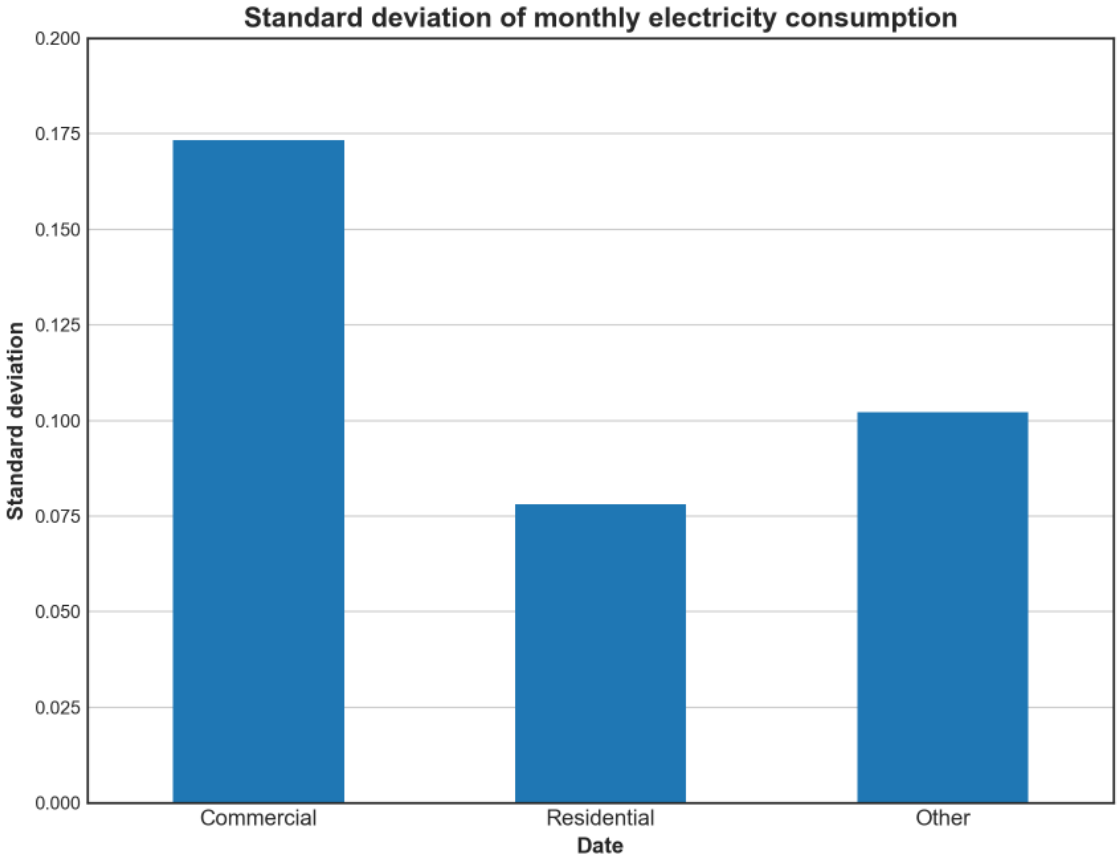



Abbildung 25: Standardabweichung des monatlichen Stromverbrauchs an Wochenenden

Um auch hier für eine eigentliche Untersuchung der saisonalen Schwankungen nicht auf reine graphische Vergleiche setzen zu müssen, werden erneut sämtliche Werte durch den Maximalwert normiert.

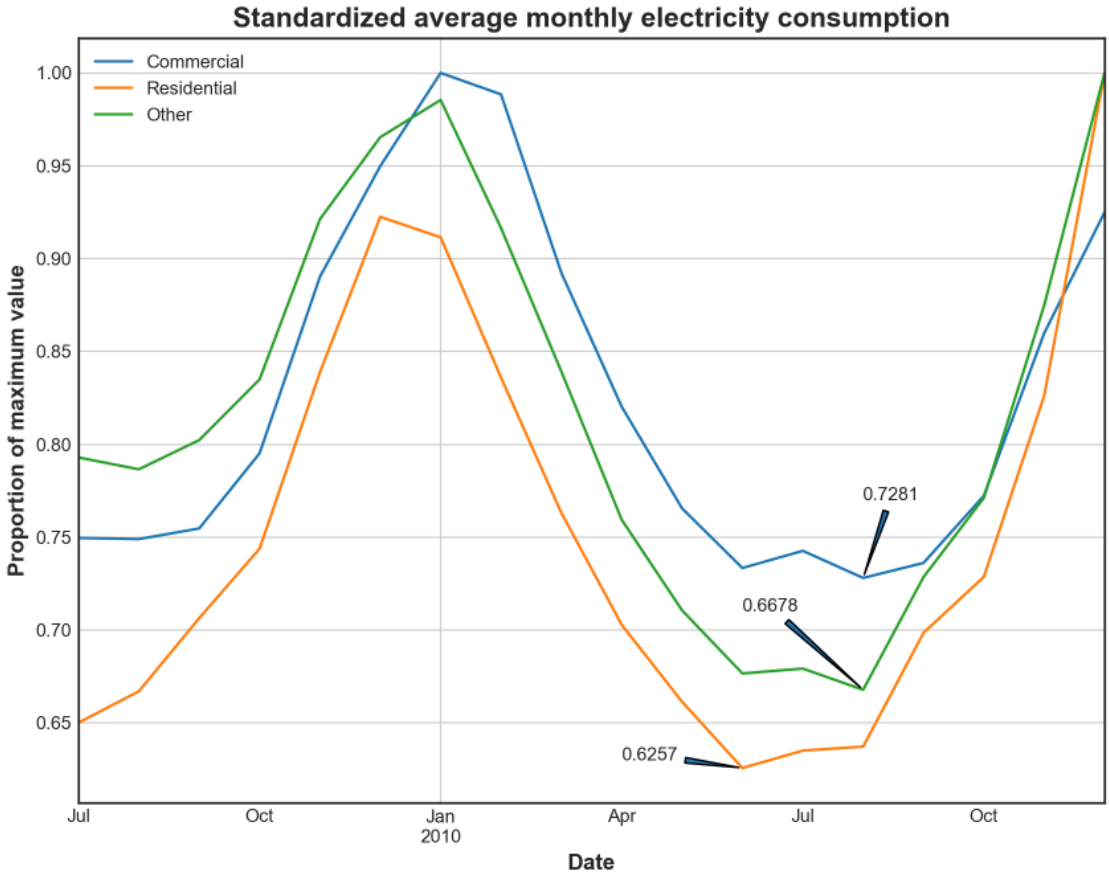



Abbildung 26: Normierter monatlicher Stromverbrauch an Wochenenden

Durch Normierung ergibt sich in diesem Fall ein deutlicheres Bild und es werden klare Unterschiede zwischen den Gebäudetypen erkennbar. So ist der Verbrauch der Geschäftsgebäude deutlich konstanter als jener der Wohngebäude. Während die Maximal- und Minimalwerte sämtlicher Gebäudetypen weiterhin in den genau gleichen Monaten auftreten, weist der Stromverbrauch der Geschäftsgebäude deutlich geringere saisonale Schwankungen auf. Dieser beträgt im August 72,81% des Maximalverbrauchs. In Wohngebäuden hingegen beträgt der Minimalverbrauch nur 62,57% des Maximalverbrauchs. Sonstige Gebäude liegen mit einem Minimalverbrauch von 66,78% genau dazwischen. Eine Normalisierung der Werte suggeriert auch in diesem Fall einen exponentiellen Verlauf des Stromverbrauchs.

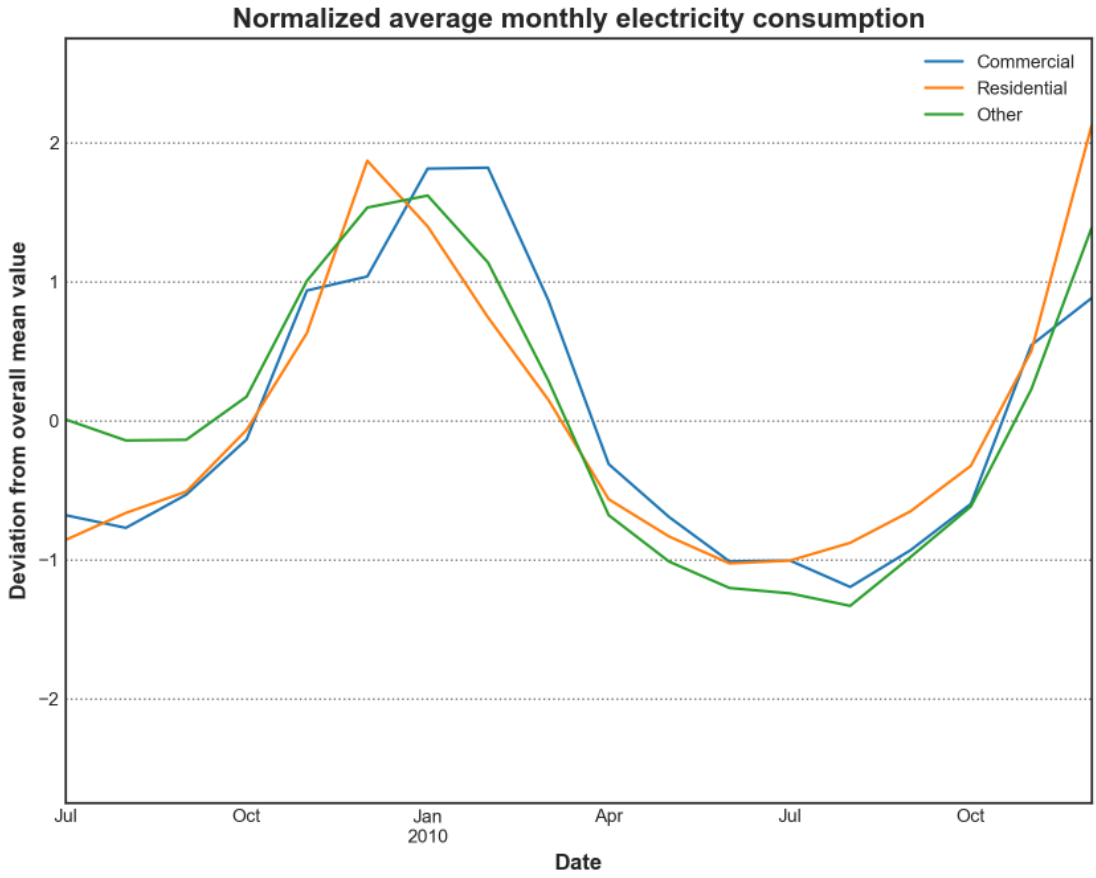



Abbildung 27: Normalisierter monatlicher Stromverbrauch an Wochenenden

Während der Stromverbrauch der sonstigen Gebäude recht gleichmäßig nach oben und unten abweicht, weichen die Stromverbräuche der Wohn- und Geschäftsgebäude auch an Wochenenden zur kalten Jahreszeit deutlich weiter nach oben ab, als sie zur warmen Jahreszeit nach unten abweichen.

### 5.2.4. Prüfung der Hypothese durch Interpretation der Ergebnisse

Zur Prüfung der Hypothese wurden umfangreiche Analysen durchgeführt. Allgemein gilt zunächst festzuhalten, dass die angenommenen Unterschiede prinzipiell deutlich subtiler waren als ursprünglich vermutet wurde. Dementsprechend lässt sich die Hypothese weder völlig eindeutig belegen noch völlig eindeutig widerlegen. Im Allgemeinen stimmt jedoch, dass der monatliche Stromverbrauch von Nichtwohngebäude über das Jahr hinweg konstanter ist als der von Wohngebäuden. Entscheidend ist hier allerdings, ob Arbeitstage oder Tage des Wochenendes betrachtet werden. An Tagen unter der Woche ist dieser Unterschied zwischen den Gebäudetypen mit circa 1% minimal, während er an Wochenenden mit 10% recht signifikant ist. Würde die beiden Ergebnisse gewichtet zusammenfassen, wäre eindeutig, dass definitiv ein Unterschied, wenn auch kein großer, zwischen den Gebäudetypen besteht. Ferner hat die Analyse interessante Erkenntnisse und Informationen bezüglich der Monate mit den minimalen und maximalen Stromverbräuchen, sowie des Einflusses der Weihnachtsfeiertage auf

den Stromverbrauch geliefert. Auch wurde deutlich, dass der Stromverbrauch in den Wintermonaten in allen Gebäudetypen exponentiell ansteigt.

## 5.3. Hypothese 2

**Der Gasverbrauch eines Privathaushaltes hängt primär von der Außentemperatur und nicht dem Nutzerverhalten ab.**

Für die Untersuchung dieser Hypothese werden die Gas Smart Meter Messdaten verwendet. Auch in diesem Fall stellen die Gas Smart Meter Daten die Fakten dar und werden in der Faktentabelle gespeichert. Um eine Analyse bezüglich der Hypothese durchzuführen werden zunächst Wetterdaten mit Temperaturwerten benötigt. Diese stellen eine der wesentlichen Dimensionen der Analyse dar. Eine reine Analyse des Gasverbrauchs hinsichtlich der Außentemperatur ist allerdings für eine wirkliche Prüfung der Hypothese nicht ausreichend. Der Gasverbrauch weißt über den Tag starke Schwankungen auf und hat in den Morgen-, Mittags- und Abendstunden in der Regel klare Spitzen. Im Gegensatz zum Stromverbrauch weist der Gasverbrauch zudem keine Grundlasten auf, wodurch viele Zeitintervalle existieren, während derer der Gasverbrauch gleich null ist. Für eine Analyse hinsichtlich der Außentemperatur sollte daher zudem die Tageszeit berücksichtigt werden. Dadurch soll gewährleistet sein, dass immer nur Intervalle mit vergleichbarer Nachfrage verglichen werden. Würde die Unterscheidung nach Tageszeiten nicht stattfinden, könnte dies die Ergebnisse stark verzerren. Geprüft werden soll, ob tatsächlich ein linearer Zusammenhang zwischen Außentemperatur und Gasverbrauch besteht. Ohne Beachtung der unterschiedlichen Tageszeiten würden die Gasverbräuche lediglich hinsichtlich der Außentemperatur aggregiert. Da während der Nachtstunden die geringste Gasnachfrage besteht, gleichzeitig Nachstunden aber häufig die kältesten Außentemperaturen aufweisen dürften, könnte dies dazu führen, dass der Gasverbrauch der kältesten Außentemperaturen sogar niedriger ist als der zu wärmeren Zeiten, was jedoch lediglich auf Nachfrageschwankungen während des Tages zurückzuführen wäre. In solchen Fällen würde die Hypothese dann zu Unrecht verworfen. Um so etwas zu vermeiden wird ebenfalls die Dimension Tageszeit herangezogen. Ferner muss auch die technische Gebäudeausstattung, bzw. die Nutzung dieser beachtet werden. Nur so kann der Effekt des Nutzerverhaltens berücksichtigt werden. Beispielsweise spielt eine Rolle, ob die Trinkwarmwassererwärmung elektrisch oder durch Gas erfolgt oder ob gasbetriebene Kochstellen existieren und falls ja, wie lange diese täglich betrieben werden. Bezüglich dessen wurden Umfragewerte erhoben, die ebenfalls bereitstehen. Diese geben unter anderem an, ob mit Gas geheizt wird oder nicht, wie genau das Trinkwarmwasser erwärmt wird, welche Art von Kochstellen vorliegen, ob optische Feuer vorliegen und falls ja, wie lange diese täglich betrieben werden und ob gasbetriebene Trockner existieren und falls ja, wie lange diese täglich in Gebrauch sind. Auch hier wird zuerst eine Datenbank als Grundlage des DW erstellt und aufgebaut. Als erstes wird erneut die Faktentabelle angelegt und mit den Messwerten beladen. Anschließend werden darauf aufbauend die Dimensionstabellen erstellt, mit Werten gefüllt und in Dimensionshierarchien gegliedert. Abschließend wird dann die Faktentabelle mit den Dimensionstabellen verknüpft.

### 5.3.1. Aufbau des Data Warehouse und Integration der Daten

#### Faktentabelle (FactTable)

Die Gas Smart Meter Daten sind nach der Zeit, nicht nach der Meter ID geordnet. Insgesamt lief das Pilotprojekt 78 Wochen, also 1,5 Jahre. Jede Datei enthält die Daten einer einzelnen Woche, sodass insgesamt 78 Dateien bestehen. Diese lagen aus ausführbare Unix Datei vor und wurden daher zunächst als CSV-Dateien gespeichert. Auf diese Weise sind die Daten direkt als Tabelle dargestellt. Die erste Spalte enthält die Meter ID, die zweite die Zeitangabe und die dritte den Gasverbrauch in kWh. Bei diesem Pilotprojekt bestanden keine Fehler im Datensatz und auch keinerlei Probleme beim Import der Messwerte, sodass der Import sehr schnell und einfach durchgeführt werden konnte. Zunächst wird die Faktentabelle erstellt.

```
CREATE TABLE FactTable (
    MeterID INTEGER,
    TimeID  INTEGER,
    Usage   DOUBLE
);
```

Die CSV-Dateien können dann problemlos importiert werden.

#### Dimension Zeit (DimTime)

Wie auch bei den Strom Smart Meter Daten ist die Zeit als eine 5-stellige Zahl angegeben, bei der die ersten drei Zahlen die Anzahl der Tage seit dem 31. Dezember 2008 und die letzten zwei Zahlen die Tageszeit als Zahl von 1-48 kodiert angeben. Alle für zeitliche Dimensionen benötigten Informationen können auch hier wieder direkt aus der kodierten Zeit ID gewonnen werden. Das Vorgehen beim Aufbau der Zeitdimensionstabelle ist dabei identisch mit dem Vorgehen in Hypothese 1. Daher wird dies hier nicht noch einmal näher erläutert, sondern stattdessen nur die SQL-Befehle angegeben.

```
CREATE TABLE DimTime (
    PK_Time  INTEGER  PRIMARY KEY,
    Date     STRING,
    Time     STRING,
    DateTime DATETIME
);

INSERT INTO DimTime (PK_Time, Date, Time)
SELECT TimeID, substr(TimeID, 1, 3), substr(TimeID, 4)
FROM FactTable
GROUP BY TimeID;

UPDATE DimTime SET Date = date(Date + julianday('2008-12-31'));

UPDATE DimTime SET Time = '00:00' WHERE Time = 1;
UPDATE DimTime SET Time = '00:30' WHERE Time = 2;
UPDATE DimTime SET Time = '01:00' WHERE Time = 3;
…
UPDATE DimTime SET Time = '23:30' WHERE Time = 48;
```

```
UPDATE DimTime SET DateTime = datetime(Date || ' ' || Time);

CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                        FROM DimTime;

DROP TABLE DimTime;

CREATE TABLE DimTime (
    PK_Time  INTEGER  PRIMARY KEY,
    Year     STRING,
    Quarter  STRING,
    Month    STRING,
    Date     DATE,
    DateTime DATETIME
);

INSERT INTO DimTime (
                        PK_Time,
                        Month,
                        Date,
                        DateTime
                    )
                    SELECT PK_Time,
                           Date,
                           Time,
                           DateTime
                      FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;

UPDATE DimTime SET Date = date(DateTime);
UPDATE DimTime SET Month = strftime('%Y %m', DateTime);
UPDATE DimTime SET Year = strftime('%Y', DateTime);
UPDATE DimTime SET Quarter = '2009 4' WHERE Month BETWEEN '2009 10' AND '2009
12';
UPDATE DimTime SET Quarter = '2010 1' WHERE Month BETWEEN '2010 01' AND '2010
03';
UPDATE DimTime SET Quarter = '2010 2' WHERE Month BETWEEN '2010 04' AND '2010
06';
…
UPDATE DimTime SET Quarter = '2011 2' WHERE Month BETWEEN '2011 04' AND '2011
06';
```

**Dimension Tageszeit (DimDaytime)**

Die Dimensionswerte der Tageszeit sind der Time ID bereits inhärent. Zunächst wird die Dimensionstabelle erstellt. Diese weist neben einer Spalte für die Primärschlüssel eine Spalte für das Messintervall und eine Spalte für die Hierarchieebene Tageszeit auf. Dabei wird ein Tag in die vier Tageszeiten Nacht (0 bis 6 Uhr), Morgen (6 bis 12 Uhr), Nachmittag (12 bis 18 Uhr) und Abend (18 bis 24 Uhr) eingeteilt. In Anlehnung an die Analysen von Hurst et al. spiegelt dies die verschiedenen Verbrauchsintervalle eines Tages recht gut wider (Hurst et al., 2020, p. 7880) und gewährleistet zudem gleichzeitig gleichmäßig lange Zeitintervalle.

```
CREATE TABLE DimDaytime (
    PK_DT     INTEGER PRIMARY KEY,
    TimeOfDay STRING,
    HourOfDay TIME
);
```

Danach werden zuerst die Primärschlüssel der Tabelle anhand der Time ID der Faktentabelle erzeugt. Im Anschluss daran werden dann die Primärschlüssel in das genaue Messintervall und die Tageszeit dekodiert.

```
INSERT INTO DimDaytime(PK_DT) SELECT substr(TimeID, 4) FROM FactTable GROUP BY
substr(TimeID, 4);

UPDATE DimDaytime SET HourOfDay = time('00:00') WHERE PK_DT = 1;
UPDATE DimDaytime SET HourOfDay = time('00:30') WHERE PK_DT = 2;
UPDATE DimDaytime SET HourOfDay = time('01:00') WHERE PK_DT = 3;
…
UPDATE DimDaytime SET HourOfDay = time('23:30') WHERE PK_DT = 48;

UPDATE DimDaytime SET TimeOfDay = 'Night' WHERE HourOfDay BETWEEN '00:00:00'
AND '05:30:00';
UPDATE DimDaytime SET TimeOfDay = 'Morning' WHERE HourOfDay BETWEEN '06:00:00'
AND '11:30:00';
UPDATE DimDaytime SET TimeOfDay = 'Afternoon' WHERE HourOfDay BETWEEN
'12:00:00' AND '17:30:00';
UPDATE DimDaytime SET TimeOfDay = 'Evening' WHERE HourOfDay BETWEEN '18:00:00'
AND '23:30:00';
```

**Dimension Wetter (DimWeather)**

Die Wetterdaten spielen für die Analyse eine entscheidende Rolle und wurden von der Webseite des irischen Wetterdienstes heruntergeladen. Als Wetterdaten liegen umfangreiche Messdaten für Niederschlagshöhe, Temperatur, Kühlgrenztemperatur, Taupunkt, Dampfdruck, relative Luftfeuchtigkeit, durchschnittliche stündliche Windgeschwindigkeit, dominante Windrichtung, Sonnenstunden, Bewölkung und sonstige Parameter vom 01.01.1990 bis zum 01.04.2020 in CSV-Format vor.

Wenn Wetterdaten als Dimension dienen sollen, stellt sich zunächst natürlich die Frage, welche Messdaten genau oder welche Kombination von Messdaten am besten verwendet werden sollten. Für die Analyse des Stromverbrauchs könnten mehrere Messwerte, nicht nur die Außentemperatur, von Relevanz sein. Beispielsweise dürfte im Rahmen der Beleuchtung vor allem der Bewölkungsgrad, die Sonnenstunden, aber eventuell auch die Wolkenhöhe oder die Niederschlagshöhe einen Einfluss auf den Stromverbrauch haben. Daher stellt sich in solchen Fällen insbesondere in Hinblick auf den Aufbau von Hierarchieebenen die Frage, welcher Wert oder welche Werte für die Wetterdimension verwendet werden sollten. Denkbar wäre hier zum einen, für jeden relevanten klimatischen Parameter eine eigene Dimensionstabelle aufzubauen. Alternativ wäre es auch möglich, sich auf einen, im besten Fall den aussagekräftigsten Parameter, festzulegen, also den, der den größten Einfluss auf den Energieverbrauch hat. Dieser könnte vorab anhand einer Regressionsanalyse identifiziert werden. Über die dabei errechneten Signifikanzniveaus und die Werte der Parameter ließen sich dann allerdings

nicht nur statistisch nicht relevante Parameter eliminieren oder der einflussreichste Faktor auswählen, sondern auch alle klimatischen Parameter zu jeweils einer einzigen Kennzahl zusammenfassen. Beispielsweise könnte die Außentemperatur mit 60%, die Luftfeuchtigkeit mit 20%, die Windstärke mit 10% und der Bewölkungsgrad ebenfalls mit 10% gewichtet werden. Gas wird im Gegensatz dazu jedoch meist primär nur zur Wärmebereitstellung, also die Raumheizung und Trinkwassererwärmung verwendet. Der Gasverbrauch dürfte daher zum größten Teil von der Außentemperatur und den damit zusammenhängenden Transmissionswärmeverlusten durch die Gebäudehülle abhängen. Im Folgenden wird daher auch ausschließlich der Gasverbrauch betrachtet.

Wetterdaten stellen unter all den im Rahmen dieser Arbeit erwähnten Dimensionen die einzige dar, die theoretisch sowohl räumlich als auch zeitlich verknüpft wird. Bei allen anderen Dimensionen erfolgt die Zuordnung entweder räumlich anhand der Meter ID – beispielsweise im Fall von Gebäude-, Geo- oder statistischen Daten – oder zeitlich anhand der Time ID, wie dies bei allen zeitlichen Dimensionen der Fall ist. Da die Daten anonymisiert vorliegen und keine Ortsangaben beinhalten, ist eine räumliche Zuordnung in diesem Fall allerdings nicht möglich. Es ist lediglich bekannt, dass es sich um Wohnhäuser in Irland handelt. Da Irland aber weder geografisch besonders groß, noch das landesweite Klima besonders heterogen ist, wurden und konnten daher nur die Wetterdaten einer einzigen, stellvertretenden Wetterstation verwendet werden. Dabei wurde die Wetterstation am Flughafen von Dublin ausgewählt. Alternativ wäre es auch möglich, stattdessen einen Durchschnitt aus den Wetterdaten aller in Irland verfügbaren Wetterstationen zu bilden. Dies wäre jedoch mit erheblichem Mehraufwand verbunden, dürfte aufgrund der Homogenität des Klimas des Landes jedoch kaum zu Unterschieden führen und wäre daher wenig sinnvoll. Folglich findet die Zuordnung der Temperaturdaten nur über die zeitliche Dimension, also die Time ID, statt.

Zunächst empfiehlt sich eine vorläufige Bearbeitung der Daten in MS-Excel. Ein Tabellenblatt kann 1.048.576 Zeilen enthalten. Da die Datei 265.178 Zeilen hat, kann die Datenmenge noch mit Excel verarbeitet werden. Im Zuge der vorläufigen Bearbeitung werden zunächst die nicht relevanten Zeilen gelöscht, sodass nur die Daten des relevanten Zeitraums vom 01.12.2009 bis zum 30.05.2011 bestehen bleiben. Falls jedoch kein Excel oder ein geeigneter CSV-Editor vorhanden wären, wäre es alternativ aber auch möglich, in SQL zunächst eine Tabelle mit allen vorhandenen Spalten zu erstellen, die CSV-Datei zu importieren und dann über SQL Befehle die nicht benötigten Zeilen und Spalten zu löschen.

Bei der Integration und Zuordnung der Temperaturdaten traten dann jedoch zwei recht gravierende Probleme auf. Ein Problem bestand darin, dass die Temperaturdaten in stündlichen Intervallen gemessen wurden. Diese müssten stattdessen halbstündig vorliegen, da Smart Meter Messdaten dies auch tun und die gleiche Granularität benötigt wird. Dies kann allerdings über die Annahme, dass der Verlauf der Außentemperatur stetig ist, gelöst werden. Die fehlenden Werte dazwischen könnten also als der Durchschnitt des vorherigen und folgenden Wertes errechnet werden. Um dies umzusetzen, wurde ein Python-Skript geschrieben, dass die Temperaturdaten als Array in CSV-Format einließt, dieses durchläuft und anhand eines Algorithmus die Zwischenwerte berechnet. Diese werden dann an der Zwischenstelle eingefügt und nach Abschluss des Algorithmus das Ergebnisarray in eine neue CSV-Datei geschrieben. Als Ergebnis wurden so die Wetterdaten durch das Python-Skript auf 30-minütige

Intervalle verfeinert. Die Temperaturdaten liegen somit in gleicher Granularität vor wie die Gas Smart Meter Daten.

```
import csv

values = []
i = 0
j = 1

with open('/dateipfad/dateiname.csv') as tempdata:
    new_csv_reader = csv.reader(tempdata)
    for row in new_csv_reader:
        values.append(float(row[0]))

while j < len(values):
    newvalue = (values[i] + values[j])/2
    values.insert(j, newvalue)
    i = i + 2
    j = j + 2

file = open('/dateipfad/dateiname.txt', 'w')
for x in values:
    file.write(str(x))
    file.write('\n')
file.close()
```

Dann müsste eine Dimensionstabelle erstellt werden. Da die bereits alle nicht relevanten Zeilen gelöscht wurden und die verbleibenden genau dem gesamten Zeitraum der Smart Meter Daten entsprechen, ist auch die Zeitstempelspalte nicht mehr notwendig. Diese entspricht aufgrund der Kodierung der Time ID ohnehin nicht den bereits in der Faktentabelle gespeicherten Fremdschlüsseln der Dimension Zeit. Die Frage ist nun, was für ein Primärschlüssel für die Wetterdaten verwendet werden soll, bzw. wie genau die Temperaturwerte den entsprechenden Gas Smart Meter Daten zugeordnet werden können. Wie zuvor erläutert, hängen die Wetterdaten in diesem Beispiel nur von der Dimension Zeit und nicht von der Dimension Raum ab, da keine geographische Zuordnung der anonymisierten Messdaten möglich ist und innerhalb Irlands keine großen klimatischen Unterschiede bestehen dürften. Die Zuordnung erfolgt also über die Dimension Zeit. Folglich ist für eine entsprechende Zuordnung die bereits existierende Fremdschlüsselspalte TimeID der Faktentabelle, bzw. die Primärschlüsselspalte der Zeitdimensionstabelle DimTime relevant. Liegen die Daten in der gleichen Granularität und über den genau gleichen Zeitraum vor, hätten die Dimensionstabellen für die Zeit und für die Wetterdaten also genau die gleiche Anzahl an Zeilen. Dadurch wäre eine entsprechende Zuordnung der Temperaturdaten zu den Smart Meter Daten sehr einfach unter Nutzung der Dimensionstabelle Zeit möglich, indem eine Wetterdimensionstabelle mit den Spalten für die Primärschlüssel und Temperaturdaten erzeugt und mit den Temperaturdaten sowie den Primärschlüsseln der Dimensionstabelle Zeit DimTime gefüllt wird.

Allerdings gab es dabei das zweite grundlegende Problem. Das Pilotprojekt lief vom 01.12.2009 bis zum 30.05.2011, also über 546 Tage. Insgesamt liegen aber nur die Messdaten von 540, bzw. für manche Intervalle nur von 539 Tagen vor. Anstatt den eigentlichen 26208 Zeilen (546 Tage * 48 Stunden pro Tag), existieren jedoch nur 25918 Zeilen. Wären diese Daten fehlerhaft und lägen daher

als NULL Werte vor, würde dies kein so großes Problem darstellen, da diese NULL Werte direkt über SQLite Befehle abgefragt und ausgegeben werden können. Da diese Daten aber stattdessen schlichtweg nicht vorhanden sind, gestaltet es sich insbesondere durch die Kodierung schwierig, diese zu erkennen. Bevor eine richtige Zuordnung der Temperaturdaten zu den Smart Meter Daten erfolgen kann, müssen diese fehlenden Zeitpunkte jedoch identifiziert werden, sodass die Temperaturdaten dieser Zeitpunkte entsprechend ebenfalls entfernt werden können und somit die Temperaturdaten und die Primärschlüssel die gleiche Zeilenanzahl aufweisen, wodurch eine automatisierte Zuordnung der Temperaturdaten anhand der Primärschlüssel der Zeitdimensionstabelle DimTime möglich wäre. Um diese fehlenden Zeitwerte möglichst effizient, ohne aufwendige Im- oder Exporte von Dateien in und aus SQLiteStudio und ohne Zusatzsoftware zu identifizieren, wurde dies in SQLiteStudio gelöst.

Zunächst wurde eine Hilfstabelle erzeugt und mit Werten gefüllt. Die erste Spalte ist ein Indexwert, der über die „Auto Increment“-Konfiguration für Primärschlüssel automatisch beim Befüllen anderer Spalten der Tabelle erzeugt wird. Dieser dient als Vergleichskriterium und spiegelt die Zeilennummer wider. Die zweite Spalte wird mit dem Sub-String der Time ID in Julian Day Format gefüllt und gibt Auskunft über die vergangenen Tage seit dem 31.12.2008. Als Datenquelle dient hier der Primärschlüssel der Zeitdimensionstabelle DimTime, da die Daten dort bereits nach DateTime geordnet vorliegen. Da allerdings nur die Tage betrachtet werden sollen, müssen diese Daten vorher zunächst erneut gruppiert werden, damit nicht jeder Wert 48 Mal hintereinander abgespeichert wird.

```
CREATE TABLE Time_Values (
    [Index] INTEGER PRIMARY KEY AUTOINCREMENT,
    Day     INTEGER,
    Diff    INTEGER
);

INSERT INTO Time_Values (Day) SELECT substr(PK_Time, 1, 3) FROM DimTime GROUP
BY substr(PK_Time, 1, 3);

UPDATE Time_Values SET Diff = Day - [Index];
```

Die erste Zeile der Indexspalte hat dann den Wert 1 und die erste Zeile der Spalte Day den Wert 335, da der 335. Tag des Jahres 2009 der erste Tag des Pilotprojektes war. Die Differenz aus beiden Werten wird in der dritten Spalte Diff gespeichert. Diese ist am Anfang 335 – 1 = 334. Durch die Tage ohne Messwerte während des Pilotprojekts und die dadurch fehlenden Time IDs entstehen dann Sprünge in dem Verlauf der Werte der Spalte Day. Fehlt eine Time ID, steigt der Wert der nächsten Zeile um 2 statt um 1 an, während der Zeilenwert der Indexspalte immer nur um 1 ansteigt. Folglich erhöht sich die Differenz an diesen Stellen. Über eine Ausgabe aller Stellen mit einer gewissen höheren Differenz können dann die fehlenden Time IDs identifiziert werden.

```
SELECT * FROM Time_Values WHERE Diff > 334;
```

Die Datenausgabe zeigt, dass für den Tag 616 keine Messwerte vorliegen.

```
SELECT * FROM Time_Values WHERE Diff > 335;
```

Auch für die Tage 630, 631 und 632 liegen keinerlei Messwerte vor.

```
SELECT * FROM Time_Values WHERE Diff > 338;
```

Die Tage 869 und 870 fehlen ebenfalls.

Darüber wurde in Anbetracht der Zeilenanzahl beim Formatieren der Julian Day Formate in Datetime-Formate deutlich, dass zusätzlich an jeweils einem spezifischen Tag um 00:30 Uhr und 01:00 Uhr keinerlei Messwerte vorliegen. Neben den bisher genannten Tagen, an denen sämtliche Messwerte fehlen, fehlen also weiterhin noch zwei weitere spezifische Zeitpunkte. Um diese zu identifizieren, muss also der genaue Tageszeitpunkt, nicht nur das Datum betrachtet werden. Auch in diesem Fall wird Julian Day Format aufgrund der anschaulichen Indexierung verwendet. Da aufgrund der Zeilenanzahl im Zuge der Filterung nach genauen Uhrzeiten die entsprechenden fehlenden Uhrzeiten bekannt sind, können auch diese Werte ähnlich identifiziert werden, indem zusätzlich eine Tageszeit spezifiziert wird.

```
CREATE TABLE Time_Values (
    [Index] INTEGER PRIMARY KEY AUTOINCREMENT,
    Day     INTEGER,
    Diff    INTEGER
);

INSERT INTO Time_Values (Day) SELECT substr(PK_Time, 1, 3) FROM DimTime WHERE
substr(PK_Time, 4) = '02';

UPDATE Time_Values SET Diff = Day - [Index];

SELECT * FROM Time_Values WHERE Diff > 334;
…
```

Zur Identifikation des zweiten fehlenden Zeitpunktes wurde ebenso vorgegangen. Es stellte sich heraus, dass neben den Tagen 616, 630, 631, 632, 869 und 870 auch an Tag 816 für die Messintervalle 00:30:00 – 00:59:59 und 01:00:00 – 01:29:59 keinerlei Messwerte bestehen.

Nachdem die fehlenden Time IDs identifiziert wurden, können die entsprechenden Time IDs mit den dazugehörigen Temperaturdaten, die anhand des Python-Skripts generiert verfeinert wurden, zusammengeführt werden. Dies kann in Excel erledigt werden. Zunächst werden die Primärschlüssel der Zeitdimensionstabelle DimTime in eine CSV-Datei exportiert. Diese wird dann in Excel importiert. Die zuvor entdeckten fehlenden Time IDs werden dann durch entsprechend programmierte Excel-Funktionen automatisch hinzugefügt. Auf diese Weise liegen die Daten über den ganzen Zeitraum ohne Lücken vor, woraus folgt, dass die Anzahl der Time IDs nun genau der Anzahl der Werte des Arrays mit den Temperaturdaten entspricht. Folglich muss nur das Array mit den Temperaturdaten in eine weitere Spalte der Exceldatei geladen werden. Die Temperaturdaten sind dann den entsprechenden Time IDs automatisch zugeordnet. Anschließend können die Zeitpunkte, in denen keine Smart Meter Daten vorliegen, einfach identifiziert und die Temperaturdaten dieser Zeitpunkte gelöscht werden. Dies ist zwar nicht unbedingt nötig, reduziert aber den benötigten Speicherplatz und erhöht die Abfragegeschwindigkeit, da bei der Suche im Zuge einer Abfrage weniger Zeile durchlaufen werden müssen. Hauptsächlich wurden die zuvor erwähnten Schritte aber benötigt, um den

Wetterdaten ein entsprechendes Zeitintervall zuordnen zu können. Als Ergebnis liegen dann nur noch die relevanten Temperaturdaten sowie die zugehörigen Time IDs als Primärschlüssel ebendieser vor. Die Time IDs sind der Faktentabelle ja bereits als Fremdschlüssel der Zeitdimensionstabelle bekannt. Die Dimension Zeit und Wetter verwenden somit dann letztendlich die gleichen Primärschlüssel. Die fertige Tabelle wird anschließend als eine CSV-Datei exportiert. Vor dem Import in SQLite besteht allerdings noch ein weiteres kleines Problem. Die Fließkommazahlen der Temperaturdaten wurden für die Verarbeitung in Excel mit Komma getrennt dargestellt. Die Trennung erfolgt in SQLite aber durch einen Punkt, da es sich um "Double" oder "Float" Datentypen handelt. Daher muss die CSV-Datei vorher nochmals mit einem geeigneten Texteditor geöffnet und alle Kommata durch Punkte ersetzt werden. Anschließend steht die Datei für den Import in die SQLite Datenbank bereit. Dies kann aber auch in der entgegengesetzten Richtung zu Komplikationen führen, da Excel durch Punkte getrennte Fließkommazahlen häufig als Datumformate interpretiert und darstellt.

Nun kann sich letztendlich der eigentlichen Integration der Wetterdimension in das DW gewidmet werden. Dafür wird eine Dimensionstabelle erstellt.

```
CREATE TABLE DimWeather (
    PK_Weather  INTEGER PRIMARY KEY,
    Temperature DOUBLE
);
```

Die CSV-Datei kann anschließend problemlos importiert werden. Da die Temperaturwerte jedoch als Fließkommazahlen dargestellt sind und es somit kaum gleiche Dimensionswerte gibt, machen Aggregationen hinsichtlich der Außentemperatur auf diese Weise noch keinen Sinn. Stattdessen sollten auch hier Hierarchieebenen hinzugefügt werden. Die erste übergeordnete Hierarchieebene rundet die Fließkommazahlen auf eine Ganzzahl. Auf diese Weise wird die Anzahl der Dimensionswerte bereits recht stark reduziert. Eine zweite, noch höhere Hierarchieebene legt dann zudem Temperaturintervalle fest. Diese hängen von den maximalen und minimalen Außentemperaturen ab. Die Intervallgröße kann theoretisch frei gewählt werden. In Wüstenregionen oder Ländern mit großen Temperaturunterschieden sollten eher größere Intervalle, in milden Klimaregionen wie Irland eher kleinere Intervalle gebildet werden. Die minimale Außentemperatur während des Zeitraums betrug minus 11.5°C, die maximale Außentemperatur betrug plus 23°C. Daher werden für die Intervalle Fünferschritte, beginnend bei minus 10°C und endend bei plus 20°C festgelegt. Im ersten Schritt wird die Dimensionstabelle um die Spalten der neuen Hierarchieebenen erweitert. Im zweiten Schritt werden diese Spalten dann mit Werten gefüllt.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                        FROM DimWeather;

DROP TABLE DimWeather;

CREATE TABLE DimWeather (
    PK_Weather    INTEGER PRIMARY KEY,
    TempIntervall STRING,
    TempRound     INTEGER,
    Temperature   DOUBLE
);

INSERT INTO DimWeather (
                           PK_Weather,
                           Temperature
```

```
                        )
                        SELECT PK_Weather,
                               Temperature
                          FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;

UPDATE DimWeather SET TempRound = round(Temperature);

UPDATE DimWeather SET TempIntervall = '<= -10' WHERE Temperature <= -10;
UPDATE DimWeather SET TempIntervall = '-10 < X <= -5' WHERE Temperature > -10
AND Temperature <= -5;
UPDATE DimWeather SET TempIntervall = '-5 < X <= 0' WHERE Temperature > -5 AND
Temperature <= 0;
UPDATE DimWeather SET TempIntervall = '0 < X <= 5' WHERE Temperature > 0 AND
Temperature <= 5;
UPDATE DimWeather SET TempIntervall = '5 < X <= 10' WHERE Temperature > 5 AND
Temperature <= 10;
UPDATE DimWeather SET TempIntervall = '10 < X <= 15' WHERE Temperature > 10 AND
Temperature <= 15;
UPDATE DimWeather SET TempIntervall = '15 < X <= 20' WHERE Temperature > 15 AND
Temperature <= 20;
UPDATE DimWeather SET TempIntervall = '>= 20' WHERE Temperature >= 20;
```

Nach Ausführung der SQL-Befehle ist der Integrationsvorgang trotz zahlreicher Hindernisse erfolgreich abgeschlossen.

**Gebäudebezogene Dimensionen**

Der überwiegende Teil der möglichen Dimensionen wird entweder anhand der zeitlichen oder räumlichen Dimension verknüpft. In der vorangegangenen Hypothese wurde nur eine, in der Analyse tatsächlich zur Anwendung kommende Dimension über die räumliche Dimension verknüpft – die Gebäudetypologie. Das Problem bei der Verknüpfung über die räumliche Dimension ist häufig, dass sich viele Meter bzw. Gebäude die gleichen Merkmale teilen. So gibt es beispielsweise viele Gebäude in der gleichen Stadt, viele Gebäude, die zu den Wohngebäuden zählen oder viele Gebäude mit der gleichen Energieeffizienzklasse oder Gebäudeausstattung. Dies hat zur Folge, dass die Meter ID als Verknüpfung und Zuordnung dient, diese jedoch nicht die eigentlichen Werte widerspiegelt, da nicht jede einzelne Meter ID einzigartige Dimensionswerte aufweist. Stattdessen sollte im Idealfall jede Dimensionstabelle, die über die Meter ID verknüpft wird, eigentlich nur genau die verschiedenen möglichen Dimensionswerte und entsprechende Primärschlüssel zur Identifikation enthalten. Gibt es beispielsweise mehrere Tausend verschiedene Gebäude und folglich Meter IDs und 10 verschiedene Städte, in denen diese angesiedelt sein können, so ist es nicht zielführend, in der Dimensionstabelle jede einzelne Meter ID und die zugehörige Stadt aufzuführen. Stattdessen sollte die Dimensionstabelle nur 10 Zeilen mit den Städtenamen und den entsprechenden Primärschlüsseln als Identifikationsnummer enthalten. Außerdem sollte jeder Messung in der Faktentabelle dann für die Dimension der Lage oder Adresse der entsprechende Primärschlüssel der Stadt anstelle der Meter ID zugeordnet werden. Diese Art der Zuordnung ist allerdings deutlich aufwendiger, da temporäre Hilfstabellen zum Einsatz kommen müssen, um den Messwerten der Faktentabelle die richtigen Fremdschlüssel zu übergeben. Allerdings bringt dies den großen Vorteil mit sich, dass die

Dimensionstabellen deutlich kleiner werden, was die Abfragegeschwindigkeit, insbesondere bei Anfragen, die auf mehrere, auf der Meter ID aufbauende Dimensionen zugreifen, deutlich erhöht.

In der vorangegangenen Analyse wurde nur die Dimension Gebäudetyp über die Meter ID abgefragt. Aufgrund des Aufwands, der immensen Größe der Faktentabelle und der für die Zuordnung benötigten Rechenzeit wurde im ersten Beispiel darauf verzichtet und stattdessen jede Meter ID einzeln in der Dimensionstabelle aufgeführt. In diesem Beispiel aber werden neben den bisherigen, über die Zeit verknüpften Dimension Wetter, Zeit und Tageszeit vor allem besonders viele Gebäudedaten abgefragt, die über die Meter ID zugeordnet werden. In diesem Fall sind das alle Dimensionen mit Informationen über die gasbetriebene, technische Gebäudeausrüstung. Insgesamt sind dies fünf verschiedene Dimensionen, die allesamt durch die beiliegenden Umfragen erhoben wurden. Diese sind zusätzlich zur Gasheizung bestehende Heizungstypen (DimAdditionalHeating), die verwendeten Technologien zur Trinkwassererwärmung (DimWaterHeating), die Art der Kochstelle (DimCooker), optische Feuer und deren Verwendung (DimFireEffect) und gasbetriebene Trockner und deren Gebrauch (DimDryer). Würden alle diese Dimensionen abgefragt und würden dabei sämtliche Dimensionstabellen die einzelnen Meter IDs mit den zugehörigen Dimensionswerte aufzählen, könnte dies die Performance sehr verschlechtern, da für jede Zeile der Faktentabelle erneut die langen Dimensionstabellen durchsucht werden müssen. Zudem erhöht sich der Aufwand einer sauberen Integration nicht linear, da alle Gebäude bezogenen Dimensionen über die Meter ID verknüpft werden und daher eine große temporäre Hilfstabelle für alle diese Dimensionen gemeinsam verwendet werden kann. Dafür werden alle einzelnen Dimensionen zunächst vorformatiert und -verarbeitet und dann anschließend in eine große Hilfstabelle zusammengeführt. Da in den Umfrageergebnissen sämtlichen Meter IDs die entsprechenden Dimensionswerte zugeordnet wurden, ist eine Verknüpfung zu einer großen Hilfstabelle problemlos möglich.

Zusätzlich zu einer Gasheizung können in Gebäuden noch weitere Heizungen zum Einsatz kommen. Diese können beispielsweise der Unterstützung an besonders kalten Tagen oder der Absicherung bei Ausfällen dienen oder je nach Situation im Wechsel betrieben werden. Zu diesen weiteren Heizungstypen zählen zum Beispiel elektrisch betriebene Heizungen, also klassische Elektroheizungen aber auch Wärmepumpen, klassische Holzöfen, offene Feuerstellen oder erneuerbare Energien wie Solarthermie. Welche zusätzlichen Heizungstypen im Gebäude zur Anwendung kommen wurde im Zuge der Umfrage erhoben. Da hier theoretisch mehrere Antworten möglich sind, wurde dies binär erhoben, indem für jeden einzelnen Heizungstyp separat hinterlegt ist, ob dieser im Gebäude zur Anwendung kommt oder nicht. Diese Informationen sind für die Analyse von wesentlichem Interesse, da unterstützende, weitere Heizungstypen den Gasverbrauch signifikant reduzieren können und sich je nach Steuerung und Nutzung auf die Korrelation zwischen Außentemperatur und Gasverbrauch auswirken können. Zunächst wurden daher alle relevanten Spalten aus den Umfrageergebnissen extrahiert. Nun stellt sich die Frage, wie genau die verschiedenen binär kodierten Heizungstypen zusammengefasst werden sollen. Theoretisch könnte für jeden Heizungstyp eine extra Dimensionstabelle aufgebaut werden. Dies wäre aber nicht nur sehr aufwendig, sondern würde auch den Speicherplatzbedarf stark erhöhen, die Abfragegeschwindigkeit reduzieren, zu langen, komplizierten, fehleranfälligen Abfrageformulierungen führen und die Datenbank sehr unübersichtlich gestalten. Zudem dürften bei dem überwiegenden Teil der Gebäude keine Zusatzheizungen zum

Einsatz kommen, da es sich um Wohngebäude handelt. Falls dennoch Zusatzheizungen zum Einsatz kommen, dürfte es sich in der Regel kaum um mehr als zwei handeln, da somit insgesamt bereits drei verschiedene Heizungstypen zur Raumkonditionierung innerhalb eines Gebäudes Anwendung fänden. Folglich werden die Daten so zusammengeführt, dass alle Anwendung findenden Kombinationen der Zusatzheizungen dargestellt werden. Da die Vorverarbeitung und Formatierung der Datei in Excel erledigt wird, erfolgt die Zusammenführung anhand der Zahlenkodierung. Eine einfache Addition des Binärcodes ist nicht zielführend, da dann die einzelnen Kombinationen nicht mehr auseinander zu halten wären. Insgesamt gab es die 6 Antwortmöglichkeiten Elektrizität, offene Feuer, andere feste Brennstoffe, Erneuerbare Energien, Sonstige und keine davon. Der Wert der ersten Spalte (elektrisch betriebene Heizung: ja oder nein) wird daher mit 100.000 multipliziert, der der zweiten Spalte mit 10.000, der der dritten Spalte mit 1.000, usw. Dann werden die Werte der Spalten aufsummiert. Anhand dieser Summe ist dann die genaue Kombination identifizierbar. Jede relevante Zahl wird entsprechend in einer weiteren Tabelle dekodiert und diese dekodierte Beschreibung dann über eine SVERWEIS-Funktion zugewiesen. So besteht für jede Meter ID die entsprechende Kombination aus Zusatzheizungen. Nach der Zuordnung der zum Einsatz kommenden Zusatzheizungen zu jeder Meter ID stellte sich heraus, das für einige Gebäude bzw. Meter IDs keine Angaben vorlagen. In diesen wenigen Ausnahmefällen wurde der Fehlerwert 999.999 abgespeichert. Wie sich bei näherer Betrachtung der Umfrage herausstellte, waren diese Fälle nicht auf eine fehlerhafte Dokumentation, sondern die Tatsache, dass in diesen Gebäuden keine Gasheizungen existieren, zurückzuführen. Nach endgültigem Aufbau des DW sollten daher sämtliche Gebäude ohne Gasheizungen aus den entsprechenden Tabellen entfernt werden.

Auch für die Warmwasserbereitung können verschiedene Technologien zum Einsatz kommen. In der Umfrage wurde daher auch erhoben, ob die Warmwasserbereitung durch Gas oder sonstige Technologien erfolgt. Insgesamt standen hier 9 Antwortmöglichkeiten zur Verfügung - zentrales Heizsystem, elektrische Tauchsieder, elektrische Durchlauferhitzer (direkt vor Wasserausguss), gasbefeuerte Durchlauferhitzer (direkt vor Wasserausguss), gasbefeuert, ölbefeuert, mit festen Brennstoffen erhitzter Wassertank, erneuerbare Energie (bspw. Solarthermie) oder Sonstige. Relevant ist in diesem Zusammenhang zwar in erster Linie, ob die Warmwasserbereitung mit Gas erfolgt oder nicht, aber auch welche Technologie genau zum Einsatz kommt. Ähnlich wie bei den Zusatzheizungen zur Raumkonditionierung, bei denen ebenfalls relevant ist, welche Technologie diese weitere(n) Heizung(en) nutzt/nutzen, da je nach Typ deutlich mehr oder weniger des gesamten Heizbedarfs durch Gas gedeckt werden muss, kann auch bezüglich der Warmwasserbereitung die Technologie deutliche Unterschiede mit sich bringen. Beispielsweise können durch die Verwendung von Solarthermie zur Warmwasserbereitung in den Sommermonaten in vielen Fällen sonstige verwendete Technologien obsolet werden, während in den Wintermonaten in der Regel immer unterstützend weitere Technologien zur Warmwasserbereitung Anwendung finden müssen. Folglich waren auch im Rahmen der Umfrage hier mehrere Antwortmöglichkeiten möglich. Genau wie bei den existierenden Zusatzheizungen sind auch diese Ergebnisse binär kodiert, in dem jede Antwortmöglichkeit in separaten Spalten aufweist, ob die entsprechende Technologie verwendet wird oder nicht. Auch hier wird dies auf die zuvor erwähnte Weise in Excel zunächst gefiltert, zusammengeführt und formatiert, sodass die verschiedenen Kombinationen entstehen. Allerdings ist in diesem Fall primär relevant, ob die Trinkwassererwärmung vollständig oder teilweise durch Gas erfolgt oder nicht. Falls für die

Trinkwassererwärmung keinerlei gasbefeuerte Systeme zur Anwendung kommen, ist für die Analyse nicht relevant, anhand welcher sonstigen genauen Technologien dies stattdessen erfolgt. Daher werden sämtliche Kombinationen ohne gasfeuerte Systeme schlicht als Sonstige zusammengefasst.

Auch ist für eine Analyse des Gasverbrauchs hinsichtlich des Wetters relevant, ob durch eventuelle Gasherde oder gasbetriebene Kocher ein zusätzlicher Gasverbrauch entsteht. Welche Art von Kochstelle vorliegt wurde ebenfalls erhoben. Hierzu gab es in der Umfrage insgesamt 3 Fragen. Zunächst wurde gefragt, welche Art von Kochstelle vorliegt. Hierbei standen die Antworten Gaskocher, Gaskocher mit Elektroherd oder -ofen, Elektroherd, Öl befeuerter Herd oder mit Festbrennstoffen befeuerter Herd zur Auswahl. Allerdings ist für die Analyse nur interessant, ob ein Gas betriebener Herd vorliegt oder nicht. Alle weiteren Antworten können daher als Sonstige zusammengefasst werden. Liegt ein Gasherd vor, wurde zusätzlich erhoben, ob einer, zwei oder mehr vorliegen und wie lange am Tag diese insgesamt betrieben werden. Da in fast allen Haushalten nur ein Gasherd vorliegt und wesentlich wichtiger ist, wie lange diese durchschnittlich pro Tag betrieben werden anstatt wie viele genau vorliegen, werden hier nur die Ergebnisse der ersten und dritten Frage berücksichtigt. Hieraus lässt sich eine Dimensionshierarchie aufbauen. Die oberste Ebene gibt lediglich an, ob ein Gasherd vorliegt oder nicht, während die untere Ebene angibt, wie lange diese/r täglich betrieben werden. Die Formatierung und Dekodierung erfolgte ebenfalls in Excel.

Zusätzlich kann der Gasverbrauch durch gasbefeuerte, optische Feuer steigen. Da diese Feuer den Gasverbrauch steigern, in der Regel aber lediglich der Optik dienen und meist kaum einen, bzw. nur einen vernachlässigbaren Beitrag zur Raumheizung leisten, sollten diese als Dimension betrachtet werden und somit Teil der Analyse sein. Ähnlich wie auch die weiteren herangezogenen Umfrageergebnisse bezüglich weiterer Gasverbräuche, die nicht direkt mit der Raumkonditionierung und Heizung verbunden sind, gibt jedoch auch diese Nutzung von Gas eine gewisse Abwärme an die Räumlichkeiten ab und stellt somit aus Bilanzierungssicht eine interne Wärmequelle dar. Dementsprechend ist es denkbar, dass diese Wärmequellen den Gasverbrauch zur Raumheizung etwas senken könnten. In der Umfrage wurden drei Fragen dazu gestellt. Zum einen, ob gasbefeuerte optische Feuer vorliegen oder nicht und falls ja, wie viele vorliegen und wie lange jedes davon durchschnittlich pro Tag im Einsatz ist. Auch hier lässt sich eine Hierarchie aufbauen. Die obere Ebene gibt die Anzahl der optischen Feuer von keines bis mehr als 2 an, während die untere Ebene angibt, wie lange pro Tag jedes davon genutzt wird. Die Formatierung und Dekodierung erfolgten in Excel.

Zuletzt können auch Trockner Gas verbrennen, um Wärme zu gewinnen. Wenn ein Haushalt einen gasbefeuerten Trockner verwendet, wirkt sich dies ebenfalls auf den Gasverbrauch aus. Daher wird die Dimension Trockner für die Analyse herangezogen. Im Zuge der Umfrage wurde erhoben, ob solche Trockner vorliegen oder nicht und falls ja, wie viele vorliegen und wie lange jeder dieser pro Tag betrieben wird. Da allerdings in allen Fällen ohnehin, falls solche Trockner verwendet werden, nur ein Trockner vorliegt, lässt sich dies in die Hierarchien „gasbefeuerter Trockner: Ja oder Nein“ und „Einsatzdauer pro Tag“ zusammenfassen. Folglich ergibt sich auch hier eine Hierarchiestruktur mit Ersterem als obere Ebene und Letzterem als untere Ebene. Die Formatierung und Dekodierung erfolgten in Excel.

Anschließend wurden alle formatierten, aufbereiteten Tabellen dieser fünf Dimensionen in eine große Tabelle zusammengeführt. Als Ergebnis entstand eine große Hilfstabelle, aus der dann später die kleinen, nach den Dimensionswerten geordneten Dimensionstabellen aufgebaut und die entsprechenden Dimensionswerte den Messwerten der Faktentabelle übergeben werden können. Diese wurde dann zunächst als CSV-Datei exportiert und stand dann für den Import in das DW bereit.

Dafür wurde zunächst eine entsprechende Tabelle DataIntegration im DW aufgebaut.

```
CREATE TABLE DataIntegration (
    MeterID          INTEGER PRIMARY KEY,
    HeatingID        INTEGER,
    HeatingType      STRING,
    WaterHeatingID   INTEGER,
    WaterHeatingType STRING,
    CookerID         INTEGER,
    CookerType       STRING,
    CookerUsage      STRING,
    FireID           INTEGER,
    FireAmount       STRING,
    FireUsage        STRING,
    DryerID          INTEGER,
    Dryer            STRING,
    DryerUsage       STRING
);
```

Nach dem erfolgreichen Import liegt die Hilfstabelle dann in der Datenbank vor. Nun können mit Hilfe dieser Tabelle die weiteren Dimensionstabellen aufgebaut werden. Vorher sollten allerdings noch die Primärschlüssel der Dimensionen Zusatzheizungen und Trinkwassererwärmung überarbeitet werden. Aufgrund der ursprünglichen Dummy-Codierung und der gewählten Formel zur Zusammenfassung dieser weisen die bestehende Primärschlüssel nämlich große Sprünge auf. Im Anschluss an den Import werden daher die bestehenden Primärschlüssel manuell über UPDATE Befehle nachbearbeitet. Da die Tabellen nur wenige Zeilen haben ist dies ohne Probleme möglich. Bei größeren Tabellen würde es sich allerdings empfehlen, die Tabelle um eine zusätzliche Spalten zu erweitern, die als automatisch hochzählende Primärschlüssel definiert werden. Die Spalten mit den "alten" Primärschlüsseln könnten dann gelöscht werden. Der Primärschlüssel 999.999 für Fehlerwerte bleibt allerdings weiterhin bestehen.

```
UPDATE DataIntegration SET HeatingID = 2 WHERE HeatingID = 10;
UPDATE DataIntegration SET HeatingID = 3 WHERE HeatingID = 100;
…
```

Somit ist die Hilfstabelle vollständig integriert und kann zum Aufbau der Dimensionstabellen und später zur Verknüpfung dieser mit der Faktentabelle genutzt werden.

**Dimension Zusatzheizungen (DimAdditionalHeating)**

Zunächst wird die Dimensionstabelle DimAdditionalHeating erstellt.

```
CREATE TABLE DimHeatingType (
    PK_Heating  INTEGER PRIMARY KEY,
    HeatingType STRING
);
```

Anschließend kann diese mit den Werten aus der Hilfstabelle gefüllt werden, indem diese nach den einzelnen Dimensionswerten gruppiert werden.

```
INSERT INTO DimHeatingType (PK_Heating, HeatingType)
SELECT HeatingID, HeatingType
FROM DataIntegration
GROUP BY HeatingID;
```

**Dimension Trinkwassererwärmung (DimWaterHeating)**

Auch für die Dimension Trinkwassererwärmung muss zunächst eine Tabelle erstellt werden. Anschließend wurde auch diese mit den gruppierten Dimensionswerten aus der Hilfstabelle gefüllt. Da diese Schritte für alle restlichen Dimensionstabellen identisch sind, werden nachfolgend nur noch die SQL-Befehle für diese Schritte angegeben.

```
CREATE TABLE DimWaterHeating (
    PK_WaterHeating  INTEGER PRIMARY KEY,
    WaterHeatingType STRING
);

INSERT INTO DimWaterHeating (PK_WaterHeating, WaterHeatingType)
SELECT WaterHeatingID, WaterHeatingType
FROM DataIntegration
GROUP BY WaterHeatingID;
```

**Dimension Kochstellen (DimCooker)**

```
CREATE TABLE DimCooker (
    PK_Cooker  INTEGER PRIMARY KEY,
    CookerType STRING,
    CookerUsage STRING
);

INSERT INTO DimCooker (PK_Cooker, CookerType, CookerUsage)
SELECT CookerID, CookerType, CookerUsage
FROM DataIntegration
GROUP BY CookerID;
```

Für die untere Hierarchieebene müssen dann jedoch noch, wie für Dimensionstabellen im Sternschema typisch, die Werte aus beiden Spalten zusammengeführt werden, sodass die untere Ebene den Wert der oberen Ebene mitführt und bei der Betrachtung der unteren Dimensionswerte deutlich ist, zu welchen Dimensionswerten der oberen Hierarchieebene diese gehören. Dieser Schritt muss auch für alle verbleibenden Dimensionstabellen durchgeführt werden.

```
UPDATE DimCooker SET CookerUsage = CookerType || '/' || CookerUsage;
```

### Dimension optische Feuer (DimFireEffect)

```
CREATE TABLE DimFireEffect (
    PK_Fire    INTEGER PRIMARY KEY,
    FireAmount STRING,
    FireUsage  STRING
);

INSERT INTO DimFireEffect (PK_Fire, FireAmount, FireUsage)
SELECT FireID, FireAmount, FireUsage
FROM DataIntegration
GROUP BY FireID;

UPDATE DimFireEffect SET FireUsage = FireAmount || '/' || FireUsage;
```

### Dimension Trockner (DimDryer)

```
CREATE TABLE DimDryer (
    PK_Dryer   INTEGER PRIMARY KEY,
    Dryer      STRING,
    DryerUsage STRING
);

INSERT INTO DimDryer (PK_Dryer, Dryer, DryerUsage)
SELECT DryerID, Dryer, DryerUsage
FROM DataIntegration
GROUP BY DryerID;

UPDATE DimDryer SET DryerUsage = Dryer || '/' || DryerUsage;
```

### Verknüpfung der Faktentabelle mit den Dimensionstabellen

Um die Dimensionstabellen mit der Faktentabelle zu verknüpfen, muss die Faktentabelle zunächst um die entsprechenden Fremdschlüsselspalten erweitert werden.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                          FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage          DOUBLE,
    MeterFK        INTEGER,
    TimeFK         INTEGER,
    DaytimeFK      INTEGER,
    WeatherFK      INTEGER,
    HeatingTypeFK  INTEGER,
    WaterHeatingFK INTEGER,
    CookerFK       INTEGER,
    FireEffectFK   INTEGER,
    DryerFK        INTEGER
);

INSERT INTO FactTable (
                          Usage,
                          MeterFK,
```

```
                        TimeFK
                    )
                    SELECT Usage,
                           MeterID,
                           TimeID
                      FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Anschließend werden die Primärschlüssel der einzelnen Dimensionstabellen den Messwerten der Faktentabelle als Fremdschlüssel zugeordnet. Für die zeitlich verknüpften Dimensionen DimDaytime und DimWeather sind die Primärschlüssel dem bereits existierenden Fremdschlüssel der Zeitdimensionstabelle DimTime inhärent und können daher direkt aus diesem erzeugt werden. Der Primärschlüssel der Wetterdimension ist mit dem der Zeitdimension identisch und der Primärschlüssel der Tageszeitdimension ist gleich dem letzten Teil des Primärschlüssels der Zeitdimension, nämlich der Stundenkodierung. Für die anhand der Meter ID zuzuordnenden Dimensionen erfolgt die Zuordnung der Dimensionswerte zu den Messwerten anhand der Hilfstabelle, die für sämtliche Meter IDs die zugehörigen Dimensionswerte enthält.

```
UPDATE FactTable SET DaytimeFK = substr(TimeFK, 4);
UPDATE FactTable SET WeatherFK = TimeFK;

UPDATE FactTable SET HeatingTypeFK = (SELECT HeatingID FROM DataIntegration
WHERE FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET WaterHeatingFK = (SELECT WaterHeatingID FROM
DataIntegration WHERE FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET CookerFK = (SELECT CookerID FROM DataIntegration WHERE
FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET FireEffectFK = (SELECT FireID FROM DataIntegration WHERE
FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET DryerFK = (SELECT DryerID FROM DataIntegration WHERE
FactTable.MeterFK = DataIntegration.MeterID);
```

Danach sind allen Messwerten die entsprechenden Dimensionswerte zugeordnet. Die Hilfstabelle hat ihren Zweck zur Erstellung der Dimensionstabellen und Verknüpfung dieser mit der Faktentabelle erfüllt und kann daher gelöscht werden. Vor einer Nutzung des DW für Analysen sollten jedoch nicht relevante und fehlerhafte Daten gelöscht werden. Dies verringert nicht nur den Speicherplatzbedarf, sondern führt auch zu kürzeren Abfragen, da beispielsweise NULL Werte bei Abfragen nicht explizit ignoriert werden müssen. Für 197 Meter IDs existieren keine Umfrageergebnisse. Folglich liegen für diese Gebäude keine Dimensionswerte der gebäudebezogenen Dimensionen vor und sie liefern für die Analyse der Hypothese keinerlei Wert, da sich nichts über die weiteren Gasverbräuche über den der Raumheizung hinaus aussagen lässt. Daher sollten die Messdaten der entsprechenden Meter IDs aus der Faktentabelle gelöscht werden.

```
DELETE FROM FactTable WHERE HeatingTypeFK IS NULL;
```

Darüber hinaus nutzen wie zuvor erwähnt nicht alle Gebäude Gas zum Heizen. Dabei handelt es sich nach vorangegangener Löschung nur noch um 21 Gebäude. Diese Werte der Dimension Zusatzheizungen dieser Gebäude wurden im Zuge der Formatierung in Excel als 999.999 kodiert. Daher können auch diese betroffenen Werte direkt über die eine Ausgabe der als fehlerhaft kodierten Fremdschlüssel der Dimension Zusatzheizungen gelöscht werden.

```
DELETE FROM FactTable WHERE HeatingTypeFK = 999999;
```

Nachdem alle Tabellen fertig formatiert und von Unreinheiten und nicht relevanten Werten befreit wurden, sollte zur abschließenden Überprüfung der Faktentabelle entsprechende Bedingungen und Definitionen festgelegt werden. Insbesondere sollten alle Fremdschlüsselspalten als solche definiert werden. Darüber hinaus kann die Spalte mit den Meter IDs (MeterFK) gelöscht werden, da diese Dimension selbst nicht für Analysen verwendet wird und ihren Zweck bereits erfüllt hat. Sie diente lediglich der Zuordnung der gebäudebezogenen Dimensionswerte zu den Messwerten der Faktentabelle.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                         FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage          DOUBLE,
    TimeFK         INTEGER REFERENCES DimTime (PK_Time)
                           NOT NULL ON CONFLICT FAIL,
    DaytimeFK      INTEGER REFERENCES DimDaytime (PK_DT)
                           NOT NULL ON CONFLICT FAIL,
    WeatherFK      INTEGER REFERENCES DimWeather (PK_Weather)
                           NOT NULL ON CONFLICT FAIL,
    HeatingTypeFK  INTEGER REFERENCES DimHeatingType (PK_Heating)
                           NOT NULL ON CONFLICT FAIL,
    WaterHeatingFK INTEGER REFERENCES DimWaterHeating (PK_WaterHeating)
                           NOT NULL ON CONFLICT FAIL,
    CookerFK       INTEGER REFERENCES DimCooker (PK_Cooker)
                           NOT NULL ON CONFLICT FAIL,
    FireEffectFK   INTEGER REFERENCES DimFireEffect (PK_Fire)
                           NOT NULL ON CONFLICT FAIL,
    DryerFK        INTEGER REFERENCES DimDryer (PK_Dryer)
                           NOT NULL ON CONFLICT FAIL
);

INSERT INTO FactTable (
                          Usage,
                          TimeFK,
                          DaytimeFK,
                          WeatherFK,
                          HeatingTypeFK,
                          WaterHeatingFK,
                          CookerFK,
                          FireEffectFK,
                          DryerFK
                      )
                      SELECT Usage,
                             TimeFK,
```

```
                            DaytimeFK,
                            WeatherFK,
                            HeatingTypeFK,
                            WaterHeatingFK,
                            CookerFK,
                            FireEffectFK,
                            DryerFK
                        FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Nun ist das DW vollständig aufgebaut und kann seine Daten für Analysen und Abfragen bereitstellen.

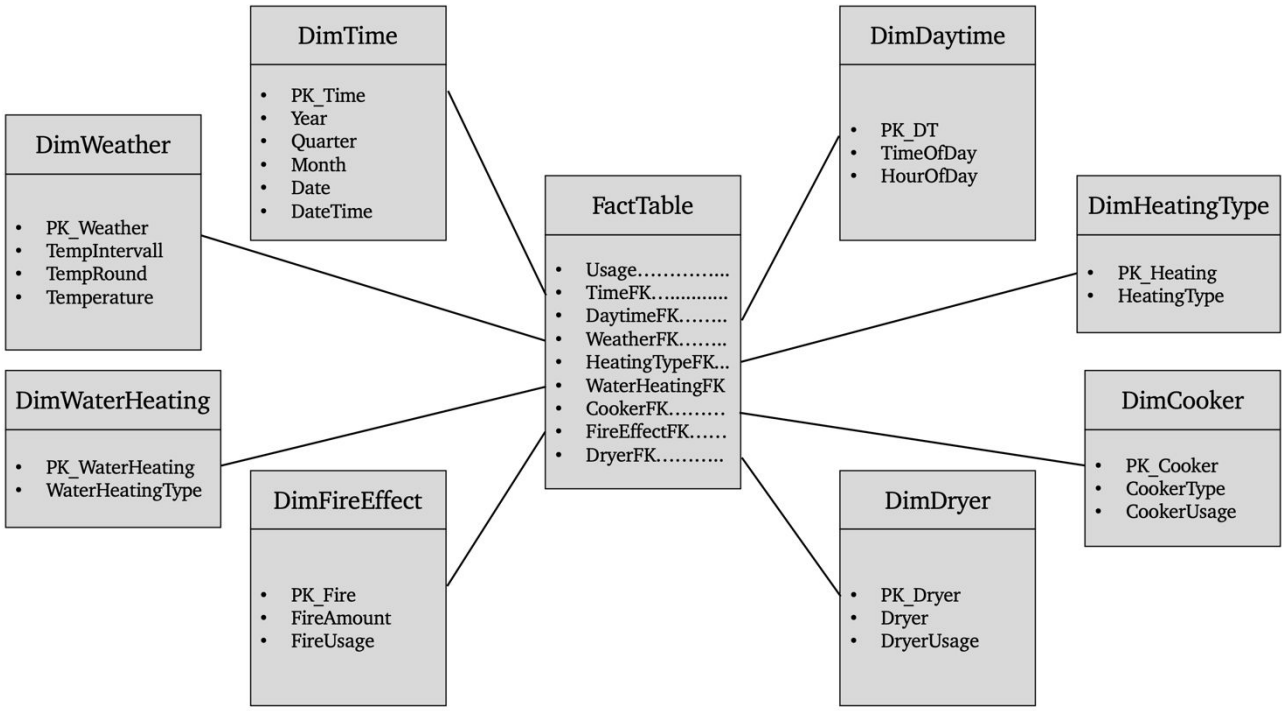



Abbildung 28: Data Warehouse für Hypothese 2

### 5.3.2. Abfrage und Visualisierung der Daten aus dem Data Warehouse

Um die Hypothese zu prüfen, werden insgesamt vier verschiedene Abfragen durchgeführt. Dabei wird jeweils einer der Gasverbrauchsfaktoren untersucht. Alle weiteren Faktoren werden auf die gleichen Werte gesetzt, um Vergleichbarkeit zu gewährleisten. Besonders Interesse besteht bei dieser Analyse an dem Gasverbrauch in Abhängigkeit von der Außentemperatur. Daher wird der Gasverbrauch hinsichtlich der gemessenen Außentemperaturen visualisiert. Folglich stellen die Außentemperaturen den Index dar. Anschließend werden dann die verschiedenen Gebäudeausstattungen, die einen Einfluss auf den Gasverbrauch haben, in diesem Koordinatensystem dargestellt. Um zusätzlich die verschiedenen Tageszeiten zu berücksichtigen, wird für jede der vier Tageszeiten Morgen, Nachmittag, Abend und Nacht ein eigenes Diagramm erstellt. In der ersten Abfrage wird dabei der Gasverbrauch hinsichtlich der verschiedenen Zusatzheizungen dargestellt. Dabei werden die drei am häufigsten vorkommenden Werte verwendet, um möglichst viele Gebäude in die Analyse miteinzubeziehen und

die Ergebnisse dadurch signifikanter und weniger anfällig für Ausreißer zu machen. Diese sind ausschließlich Gasheizung, Gasheizung und Elektroheizung (zentral oder Speicher) und Gasheizung und offene Feuer. Zudem werden alle weiteren Faktoren auf die gleichen Werte gesetzt. Auch hier gilt, dass dies möglichst häufig vorkommende Werte sein sollten. Dementsprechend erfolgt die Trinkwassererwärmung durch Gas und es kommen keine Gasherde, optische Feuer oder gasbefeuerte Trockner vor. Der Gasverbrauch wird dabei wieder als Durchschnitt angegeben, sodass die Anzahl der Gebäude, die in die Aggregationen eingehen, keinen Einfluss auf das Ergebnis hat. Gruppiert und aggregiert werden die Werte dementsprechend hinsichtlich der Tageszeit, dem Heizungstyp und der gerundeten Außentemperatur.

```
SELECT avg(Usage) AS Usage, WaterHeatingType, CookerType, FireAmount, Dryer,
TimeOfDay, HeatingType, TempRound
FROM FactTable, DimWaterHeating, DimCooker, DimFireEffect, DimDryer,
DimDaytime, DimHeatingType, DimWeather
WHERE FactTable.WaterHeatingFK = DimWaterHeating.PK_WaterHeating
AND FactTable.CookerFK = DimCooker.PK_Cooker
AND FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.DryerFK = DimDryer.PK_Dryer
AND FactTable.DaytimeFK = DimDaytime.PK_DT
AND FactTable.HeatingTypeFK = DimHeatingType.PK_Heating
AND FactTable.WeatherFK = DimWeather.PK_Weather
AND (DimHeatingType.HeatingType = 'None of these' OR DimHeatingType.HeatingType
= 'Electricity (central or storage)'
OR DimHeatingType.HeatingType = 'Open fires')
AND DimWaterHeating.WaterHeatingType = 'Gas'
AND DimCooker.CookerType = 'Other cooker'
AND DimFireEffect.FireAmount = 'No fires'
AND DimDryer.Dryer = 'No'
GROUP BY TimeOfDay, HeatingType, TempRound
```

Das Vorgehen ist dabei mit dem der vorangegangenen Hypothese identisch. Die Ausgabetabelle wird als CSV-Datei gespeichert und als ein Data Frame in Python importiert. Anschließend werden die Werte entsprechend gefiltert und ein neuer Data Frame zur Darstellung aufgebaut und visualisiert.

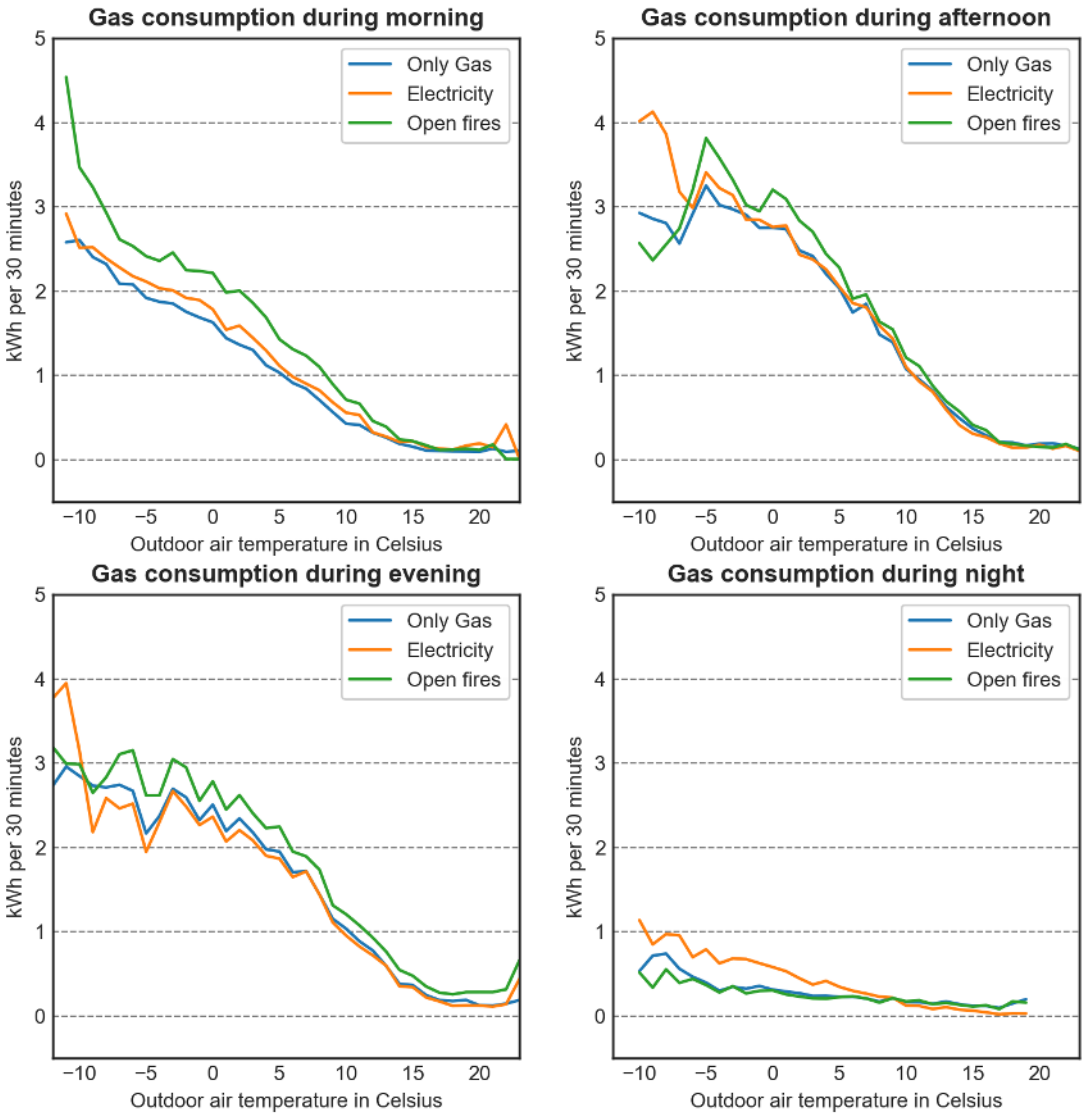



Abbildung 29: Monatlicher Gasverbrauch nach Zusatzheizungen

Es fällt sofort auf, dass für alle Tageszeiten ein ganz klarer Zusammenhang zwischen Außentemperatur und Gasverbrauch erkennbar ist. Der Gasverbrauch steigt mit sinkender Außentemperatur linear an. Dieser lineare Anstieg beginnt ab einer Temperatur von 15°C und setzt sich meist bis zu den minimalen Außentemperaturen fort. Dabei wird der Gasverbrauch auch bei höheren Temperaturen nie wirklich gleich Null, was darauf hindeutet, dass auch zu wärmeren Zeiten manchmal Gas verbraucht wurde. Aufgrund der Dimensionseinschränkungen ist die einzig mögliche Quelle dieses Gasverbrauchs bei höheren Temperaturen, die eigentlich keine Raumbeheizung mehr erfordern, das Brauchwarmwasser. Die Trinkwassererwärmung aller dieser Gebäude erfolgt nämlich durch Gas, während alle sonstigen Gasverbrauchsquellen eliminiert wurden. Zudem fällt klar ins Auge, dass der Gasverbrauch zu allen Tageszeiten außer der Nacht in etwa gleich hoch ist. In den Nachtstunden

hingegen ist der Gasverbrauch deutlich geringer. Dieser Unterschied ist sehr groß, was jedoch zu erwarten war.

Auffällig ist auch, dass praktische keine Unterschiede zwischen den verschiedenen Zusatzheizungen bestehen. Sowohl die Höhe als auch der Verlauf des Gasverbrauch sind praktisch immer genau identisch. Lediglich bei kalten Außentemperaturen liegen teils größere Schwankungen und Abweichungen vor. Dabei sollte jedoch bedacht werden, dass an beiden Enden der X-Achse kaum noch Messwerte vorliegen, da diese Stellen verhältnismäßig extreme Außentemperaturen aufweisen, die nicht häufig vorkommen. Diese Stellen sind daher besonders anfällig für Ausreißer, fehlerhafte Messwerte oder den Zufall. Da keine stärkeren oder schwächeren Steigungen des Verlaufs und keine wirklichen Unterschiede bei der Höhe des Durchschnittsverbrauchs erkennbar sind, deutet dies darauf hin, dass offene Feuer und Elektroheizungen vergleichsweise wenig zur Deckung des Heizwärmebedarfs beitragen und dieser hauptsächlich durch Gas gedeckt wird. Die sonstigen Zusatzheizungen haben also kaum einen Einfluss auf die Höhe oder den Verlauf des Gasverbrauchs.

Allerdings fallen bei einem Blick ins Detail drei Dinge auf. Der Gasverbrauch bei der Konfiguration Gasheizung und offene Feuer ist immer vergleichsweise hoch. Dies könnte damit zusammenhängen, dass Feuerstellen insbesondere in alten Gebäuden vorkommen, die eher schlechtere Baustandards hinsichtlich der Energieeffizienz aufweisen. Daher könnte in diesen Gebäuden allgemein ein höherer Heizwärmebedarf bestehen. Kämen diese Gebäude zusätzlich eher in ländlichen Regionen vor, könnten sie zudem schlichtweg größere Nutzflächen aufweisen, was den Heizwärmebedarf ebenfalls weiter erhöhen würde. Zweitens fällt auf, dass der Gasverbrauch bei ebendieser Konfiguration in Morgenstunden im Vergleich zu den anderen Tageszeiten besonders hoch ist. Grund dafür dürfte sein, dass offene Feuer meist nur in den Abendstunden genutzt werden, jedoch so gut wie nie morgens. Daher muss der Heizwärmebedarf in den Morgenstunden und vormittags ausschließlich durch Gas gedeckt werden. Darüber hinaus fällt auf, dass sich in den Nachtstunden der Verlauf des Gasverbrauchs bei der Konfiguration Gas- und Elektroheizung sehr von denen der anderen Konfigurationen unterscheidet, da er eine größere Steigung aufweist. Dies ist sehr interessant, da es auf die Verwendung typischer Nachtspeicherheizungen hindeuten könnte. Schließlich sind die Elektroheizungen als Zentralheizungen oder Speicher definiert. Elektroheizungen stoßen bei kalten Temperaturen leicht an ihre Grenzen. Daraus könnte der starke Anstieg des Gasverbrauchs bei kalten Temperaturen unter 0°C folgen. Bei wärmeren Temperaturen können solche Heizungen jedoch einen guten Teil des Heizwärmebedarfs decken, was der Grund dafür sein könnte, dass der Gasverbrauch dieser Konfiguration im Gegensatz zu den sonstigen Konfigurationen bei wärmeren Temperaturen gleich null ist.

Anschließend wurden die Werte erneut anhand der Maximalwerte normiert und dargestellt. Diese Art der Darstellung brachte jedoch keine wesentlichen neuen Erkenntnisse, da sich die Gasverbräuche bereits absolut kaum voneinander unterscheiden.

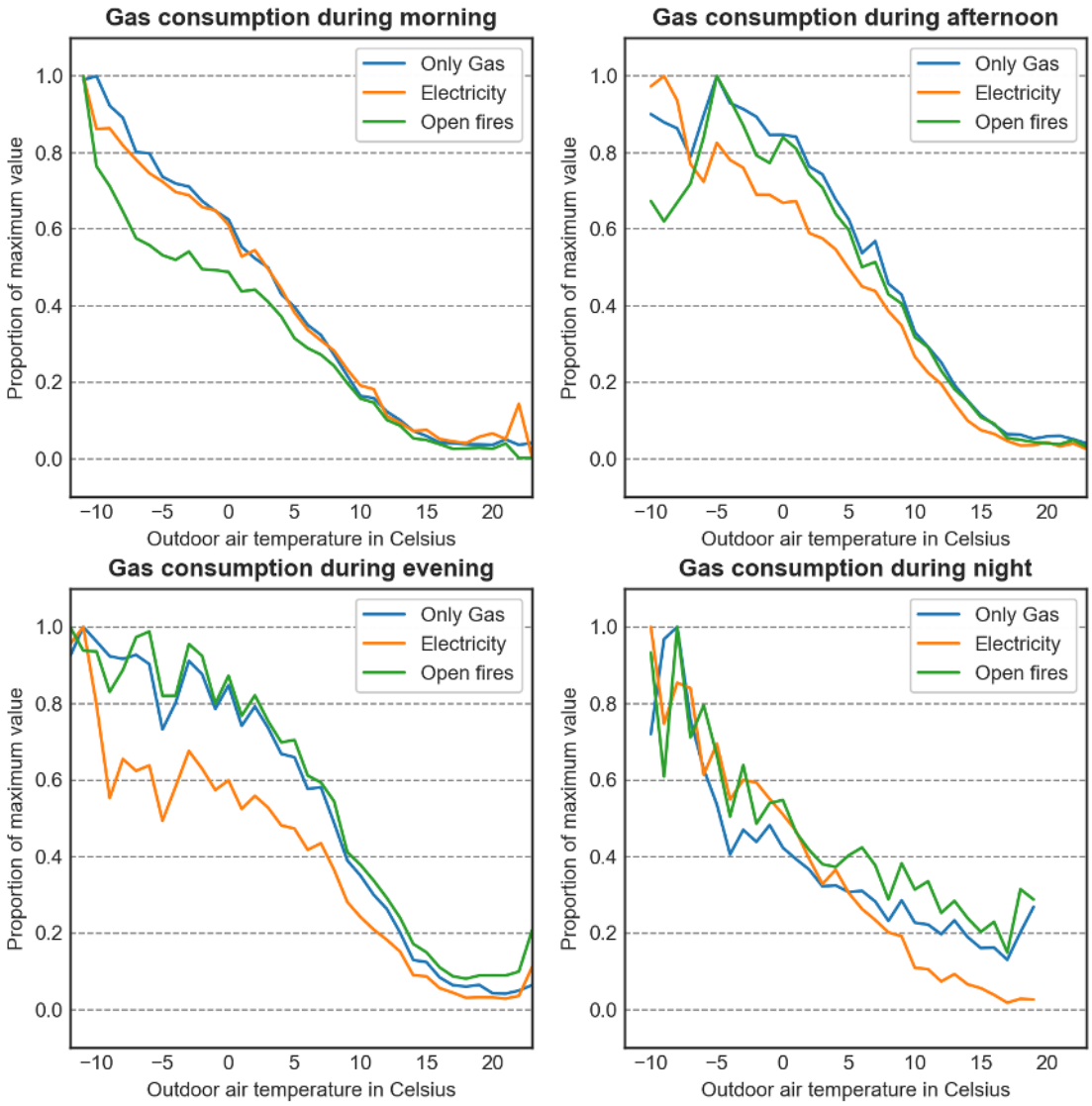



Abbildung 30: Normierter monatlicher Gasverbrauch nach Zusatzheizungen

Als nächstes wurde dann der Gasverbrauch hinsichtlich der Konfigurationen für die Trinkwassererwärmung analysiert. Dieser Fall beinhaltet damit bereits einen höheren Einfluss des Nutzerverhaltens als die Analyse der Zusatzheizungen, da letztere häufig automatisiert anhand von Thermostaten gesteuert werden. Das Vorgehen bei der Analyse ist dabei genau gleich. Dieses Mal wird die Analyse jedoch so geändert, dass die Dimension Zusatzheizungen konstant gehalten und die Dimension Trinkwassererwärmung verändert wird. Alle weiteren Gasverbrauchsquellen werden weiterhin auf Werte gesetzt, die diese eliminieren. Für die Raumheizung kommt ausschließlich Gas zum Einsatz, um die Ergebnisse nicht zu verfälschen und da zudem bei dieser Konfiguration die größte Anzahl an Gebäude in die Analyse eingeht. Als mögliche Konfigurationen der Trinkwassererwärmung werden ein zentrales Heizsystem, Gas, Gas plus elektrische Tauchsieder und Sonstige herangezogen.

Sonstige beinhalten in diesem Fall sämtliche Konfigurationen, die kein Gas verwenden. Dementsprechend wird in diesen Gebäuden lediglich zu Raumheizungszwecken Gas verbrannt. Zunächst wurden die Daten abgefragt und als CSV-Datei exportiert.

```
SELECT avg(Usage) AS Usage, HeatingType, CookerType, FireAmount, Dryer,
TimeOfDay, WaterHeatingType, TempRound
FROM FactTable, DimHeatingType, DimCooker, DimFireEffect, DimDryer, DimDaytime,
DimWaterHeating, DimWeather
WHERE FactTable.HeatingTypeFK = DimHeatingType.PK_Heating
AND FactTable.CookerFK = DimCooker.PK_Cooker
AND FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.DryerFK = DimDryer.PK_Dryer
AND FactTable.DaytimeFK = DimDaytime.PK_DT
AND FactTable.WaterHeatingFK = DimWaterHeating.PK_WaterHeating
AND FactTable.WeatherFK = DimWeather.PK_Weather
AND (DimWaterHeating.WaterHeatingType = 'Central heating system' OR
DimWaterHeating.WaterHeatingType = 'Gas'
OR DimWaterHeating.WaterHeatingType = 'Gas, Electric (immersion)' OR
DimWaterHeating.WaterHeatingType = 'Other')
AND DimHeatingType.HeatingType = 'None of these'
AND DimCooker.CookerType = 'Other cooker'
AND DimFireEffect.FireAmount = 'No fires'
AND DimDryer.Dryer = 'No'
GROUP BY TimeOfDay, WaterHeatingType, TempRound
```

Auch hier werden die Gasverbräuche wieder hinsichtlich der Außentemperatur dargestellt. Dabei wird erneut für jede Tageszeit ein eigenes Diagramm erstellt.

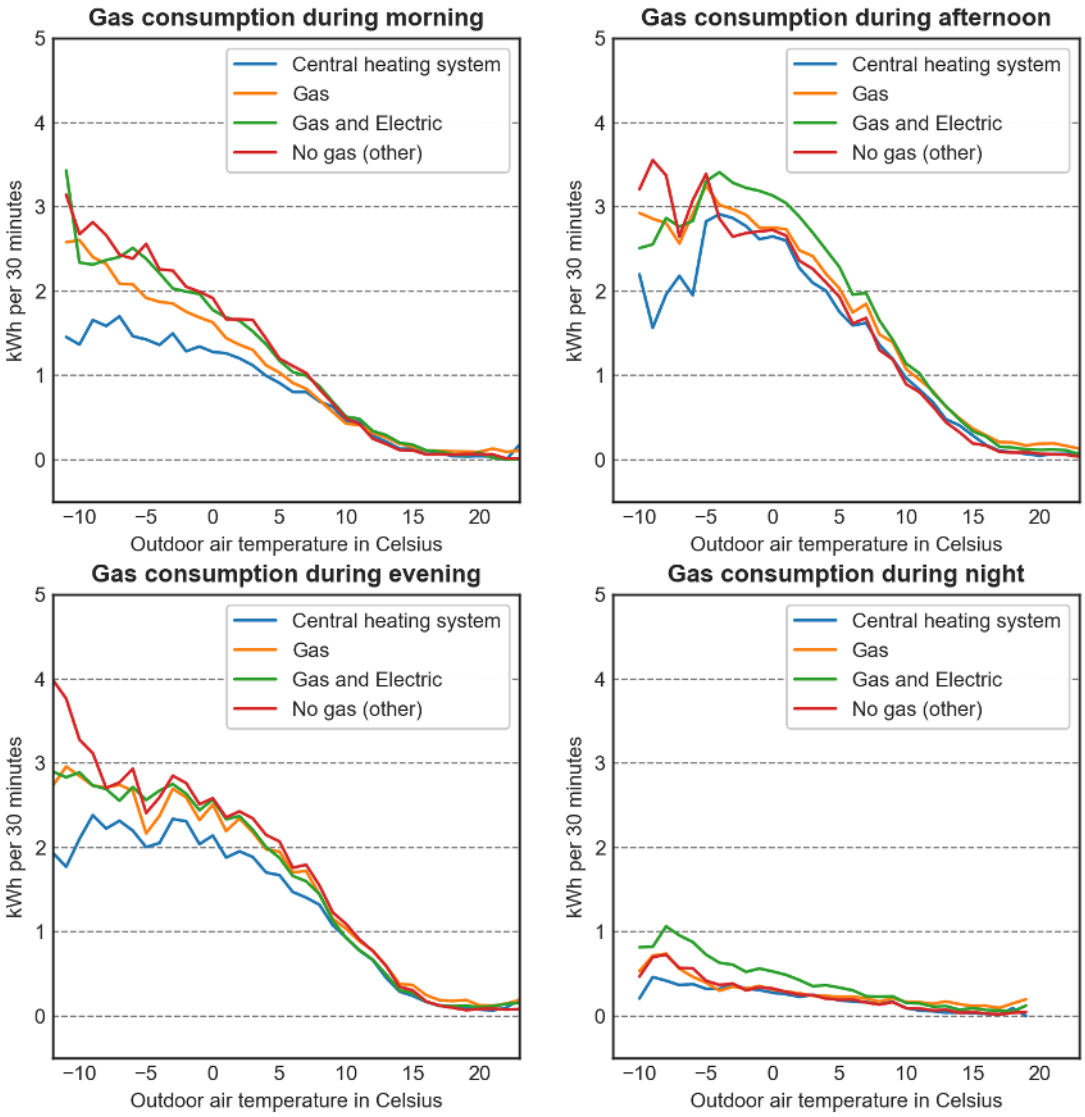



Abbildung 31: Monatlicher Gasverbrauch nach Trinkwassererwärmung

In Anbetracht der Ergebnisse lassen sich keine wirklichen Muster oder Zusammenhänge bezüglich der Trinkwassererwärmungskonfigurationen erkennen. Die Art der Trinkwassererwärmung hat somit keinen erkennbaren Einfluss auf den gesamten Gasverbrauch. Die Ergebnisse hinsichtlich der Außentemperatur sind allerdings auch in diesem Fall wieder eindeutig, da der Gasverbrauch weiterhin sehr stark von der Außentemperatur abhängt. Dabei ergibt sich der gleiche Zusammenhang wie bereits bei der Analyse der Heizungskonfigurationen. Der Gasverbrauch steigt ab Temperaturen unter circa 15°C mit sinkender Außentemperatur linear an. Bei kälteren Außentemperaturen liegen jedoch teils starke Streuungen der Werte vor. Dies ist auch in diesem Fall am ehesten auf die geringe Anzahl der vorliegenden Messwerte bei diesen Temperaturen, kombiniert mit der Tatsache, dass aufgrund der

Temperatur gleichzeitig ein hoher Heizwärmebedarf besteht, zurückzuführen. Diese beiden Faktoren könnten theoretisch zu den teils großen Schwankungen führen.

In einer letzten Analyse soll dann der Einfluss von gasbetriebenen Kochstellen auf den Gasverbrauch untersucht werden. Diese Dimension weißt einen sehr großen Einfluss des Nutzerverhaltens auf und ist die einzige Gasverbrauchsquelle, die prinzipiell unabhängig von der Jahreszeit und den Außentemperaturen sein dürfte, da in der Regel zu jeder Jahreszeit etwa gleichmäßig viel gekocht wird. Bei dieser Dimension liegen neben der Angabe, ob gasbetriebene Kochstellen oder nicht und ob diese durch Elektroöfen unterstützt werden oder nicht, auch Angaben über die durchschnittliche tägliche Nutzungsdauer vor. Zu erwarten wäre daher, dass in Haushalten mit Gasherden, in denen täglich und rund um das Jahr viel gekocht wird, der Gasverbrauch bei höheren Temperaturen höher ist als in sonstigen Haushalten und der Verlauf des Gasverbrauchs über die Außentemperatur dementsprechend eine etwas geringere Steigung aufweist. Wäre dies der Fall, wäre zumindest erstmalig verdeutlicht, dass das Nutzerverhalten einen gewissen Einfluss auf den Gasverbrauch und dessen Zusammenhang mit der Außentemperatur haben kann. Dieses Mal werden die Dimensionen Zusatzheizungen und Trinkwassererwärmung konstant gehalten. Für die Kochstellendimension wird definiert, dass lediglich Haushalte mit gasbetriebenen Kochstellen berücksichtigt werden. Aggregiert wird dann hinsichtlich der täglichen Nutzungsdauer ebendieser Kochstellen.

```
SELECT avg(Usage) AS Usage, HeatingType, WaterHeatingType, FireAmount, Dryer,
TimeOfDay, CookerUsage, TempRound
FROM FactTable, DimHeatingType, DimWaterHeating, DimFireEffect, DimDryer,
DimDaytime, DimCooker, DimWeather
WHERE FactTable.HeatingTypeFK = DimHeatingType.PK_Heating
AND FactTable.WaterHeatingFK = DimWaterHeating.PK_WaterHeating
AND FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.DryerFK = DimDryer.PK_Dryer
AND FactTable.DaytimeFK = DimDaytime.PK_DT
AND FactTable.CookerFK = DimCooker.PK_Cooker
AND FactTable.WeatherFK = DimWeather.PK_Weather
AND DimCooker.CookerType = 'Gas cooker'
AND DimHeatingType.HeatingType = 'None of these'
AND DimWaterHeating.WaterHeatingType = 'Gas'
AND DimFireEffect.FireAmount = 'No fires'
AND DimDryer.Dryer = 'No'
GROUP BY TimeOfDay, CookerUsage, TempRound
```

Vor der Visualisierung wurden die Verbrauchswerte erneut anhand der Maximalwerte normiert, in der Hoffnung, dass sich so mehr über den Verlauf des Gasverbrauchs über die Außentemperatur hinweg aussagen lässt. Leider lassen jedoch anhand der Ergebnisse keine wirklichen Schlussfolgerungen ziehen.

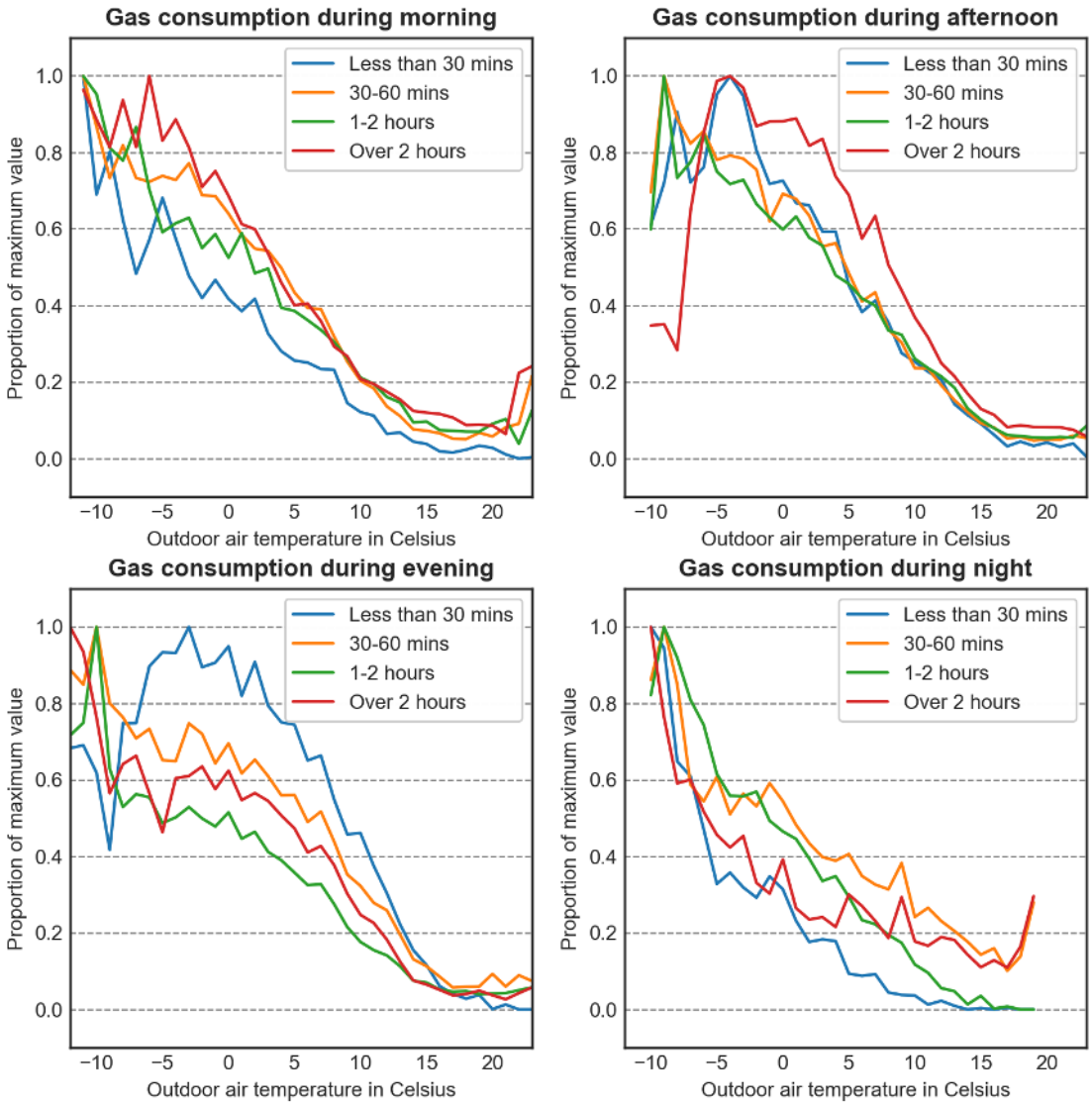



Abbildung 32: Normierter monatlicher Gasverbrauch nach Kochstellen

Das grundlegende Problem dürfte in diesem Fall dreierlei sein. Zum Einen hat die bereits bekannte, recht hohe Streuung der Verbrauchswerte bei sehr niedrigen Temperaturen, basierend auf den wenigen Messwerten, zur Folge, dass der Maximalwert des Verbrauchs bei unterschiedlichen Temperaturen, nicht nur der niedrigsten Temperatur, vorkommen kann. Dies kann der Verlauf der Graphen sehr verzerren. Zum Anderen dürfte der Gasverbrauch von Kochstellen zur kalten Jahreszeit vergleichsweise unbedeutend für den Gesamtverbrauch sein, da ein sehr hoher Heizwärmebedarf besteht. Um die Einflüsse möglichst gut zu untersuchen würde es daher mehr Sinn machen, den Fokus auf wärmere Außentemperaturen zu legen. Drittens macht es wenig Sinn, Unterschiede der täglichen Nutzungsdauer von Gaskochstellen auf den Gasverbrauch in den Nacht- oder Morgenstunden zu untersuchen, da zu diesen Zeiten ohnehin keine Nutzung stattfinden dürfte. Wird die Kochstelle z.B.

täglich für eine Stunde genutzt, so kann dies große Auswirkungen auf den Gasverbrauch dieses spezifischen Zeitraums haben. Da ein Tag jedoch 24 Stunden hat, dürften die Auswirkungen auf den Gasverbrauch des gesamten Tages deutlich geringer sein. Folglich macht es mehr Sinn, anstatt der vier Tageszeiten beispielsweise vier verschiedene Halbstundenintervalle in der Mittagszeit zu betrachten. So wird daher in der abschließenden Analyse verfahren. Die Tageszeit wird auf das Zeitintervall von 12 – 14 Uhr festgelegt. Ferner werden nur die Verbrauchsdaten bei Außentemperaturen über 0°C betrachtet. Um den Fokus auf den Verlauf des Verbrauchs über die Temperatur hinweg zu legen, werden sämtliche Messwerte auch in diesem Fall normiert.

```
SELECT avg(Usage) AS Usage, HeatingType, WaterHeatingType, FireAmount, Dryer,
HourOfDay, CookerUsage, TempRound
FROM FactTable, DimHeatingType, DimWaterHeating, DimFireEffect, DimDryer,
DimDaytime, DimCooker, DimWeather
WHERE FactTable.HeatingTypeFK = DimHeatingType.PK_Heating
AND FactTable.WaterHeatingFK = DimWaterHeating.PK_WaterHeating
AND FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.DryerFK = DimDryer.PK_Dryer
AND FactTable.DaytimeFK = DimDaytime.PK_DT
AND FactTable.CookerFK = DimCooker.PK_Cooker
AND FactTable.WeatherFK = DimWeather.PK_Weather
AND DimCooker.CookerType = 'Gas cooker'
AND DimDaytime.HourOfDay BETWEEN '12:00:00' AND '13:30:00'
AND DimHeatingType.HeatingType = 'None of these'
AND DimWaterHeating.WaterHeatingType = 'Gas'
AND DimFireEffect.FireAmount = 'No fires'
AND DimDryer.Dryer = 'No'
GROUP BY HourOfDay, CookerUsage, TempRound
```

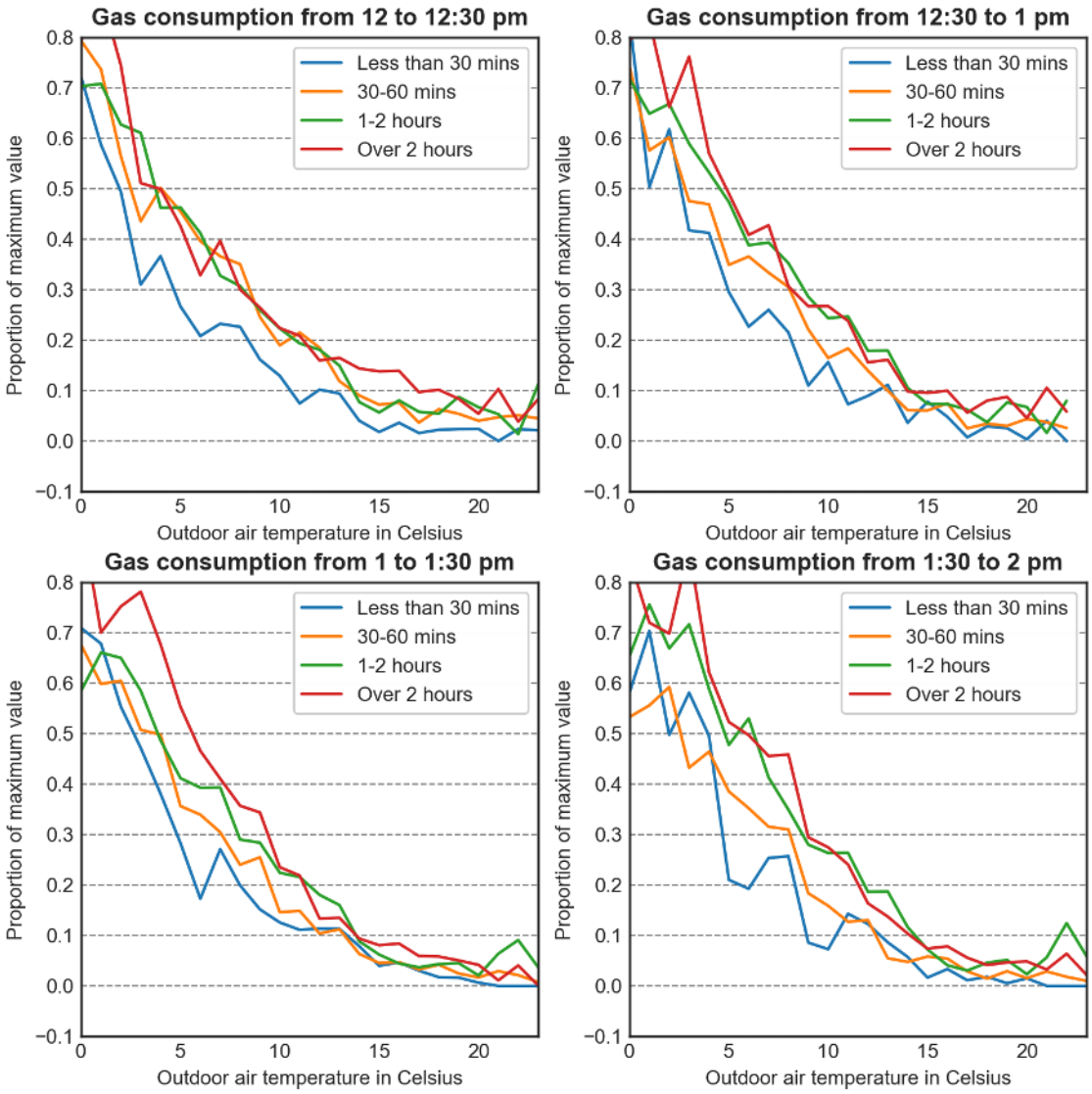



Abbildung 33: Normierter monatlicher Gasverbrauch nach Kochstellen zur Mittagszeit

Das Ergebnis ist hier deutlich anschaulicher und aufschlussreicher. Der Zusammenhang zwischen Außentemperatur und Gasverbrauch bleibt auch weiterhin eindeutig bestehen. Insgesamt hat die Nutzung der Gaskochstellen keinen erkennbaren Einfluss auf die Steigung dieses Verlaufs. Allerdings ist sehr eindeutig erkennbar, dass das Nutzerverhalten einen Einfluss auf die Höhe des Gasverbrauchs hat. So ist bei jeder Außentemperatur zu jedem Messintervall gegeben, dass der Gasverbrauch bei über zweistündiger Nutzung der Kochstelle höher ist als bei maximal 30-minütiger Nutzung. Dies ist daran zu erkennen, dass in jedem Diagramm an jeder Stelle die rote Linie (täglich über 2 Stunden) über der blauen Linie (weniger als 30 Minuten) liegt. Darüber hinaus bilden diese beiden Linien sogar über weite Teile eine Art Grenze auf jeder Seite und schließen somit die dazwischenliegenden Werte ein. Somit lässt sogar an vielen Stellen eine klare Rangordnung des Energieverbrauchs erkennen, die der

logischen Schlussfolgerung entspricht, dass eine längere Nutzung der Kochstelle den Gasverbrauch erhöht. Die gleiche Analyse wurde ebenfalls in den Abendstunden durchgeführt, allerdings waren die Ergebnisse hier deutlich weniger eindeutig. Daraus lässt sich schließen, dass die Gaskochstellen in der Regel insbesondere zur Mittagszeit betrieben wurden.

### 5.3.3. Prüfung der Hypothese durch Interpretation der Ergebnisse

Die Hypothese wurde recht umfangreich geprüft. Jede einzelne Untersuchung hat dabei klar ergeben, wie stark der Gasverbrauch von der Außentemperatur abhängt. Dabei ergab sich immer ein linearer Zusammenhang, beginnend ab Temperaturen unter 15°C. Dieser lineare Zusammenhang konnte weder wesentlich durch die Konfiguration der Heizungen noch der Trinkwassererwärmung beeinflusst werden. Allerdings ließen sich stattdessen viele kleinere, aber interessante Details aufdecken. Insgesamt gilt auch für diesen Fall, ähnlich wie bei Hypothese 1, dass die Ergebnisse deutlich subtiler sind als sich vermuten lassen würde. Nichtsdestotrotz sind diese durch eine genauere Betrachtung und eventuelle Einschränkungen der Achsenwerte bei der Darstellung gut erkennbar. Auch wurde deutlich, dass der Gasverbrauch in den Nachtstunden deutlich geringer ist. Hinsichtlich des Nutzerverhaltens hat sich gezeigt, dass zu Zeiten mit geringerem Heizwärmebedarf das Nutzerverhalten sehr wohl einen Einfluss auf den Gasverbrauch haben kann. Allerdings war selbst dann der lineare Zusammenhang zwischen Außentemperatur und Gasverbrauch unveränderlich, da sich lediglich der Gasverbrauch insgesamt erhöht hat. Damit ist die aufgestellte Hypothese eindeutig bestätigt. Der Gasverbrauch eines Wohngebäudes ist, laut der dieser Analyse zugrundliegenden Daten, eindeutig und in erster Linie von der Außentemperatur abhängig. Sonstige Faktoren wie das Nutzerverhalten können zwar einen Einfluss haben, beeinflussen den grundlegenden Zusammenhang und den Einfluss der klimatischen Verhältnisse jedoch nur in geringem Maße. Um die Einflüsse des Nutzerverhaltens aufzudecken, müssen zudem genaue Zeitpunkte mit geringem Heizwärmebedarf und intensiver Nutzung betrachtet werden.

## 5.4. Hypothese 3

**Ein hohes Haushaltseinkommen steigert den Gasverbrauch, da es weniger Anreize für energiesparendes Nutzerverhalten bietet.**

Im dritten Beispiel soll also der Energieverbrauch hinsichtlich des Einkommens untersucht werden. So soll festgestellt werden, ob ein hohes Haushaltseinkommen zu einem höheren Gasverbrauch führt als ein geringeres Haushaltseinkommen, da durch ein hohes Einkommen dem Energiesparen weniger Bedeutung beigemessen wird. Ergebnis der Analyse wäre also die Antwort auf die Frage, ob wohlhabendere Haushalte eher verschwenderischer mit Energie umgehen als ärmere Haushalte. Diese Fragestellung unterscheidet diese Hypothese von den Hypothesen 1 und 2. In den Hypothesen 1 und 2 lag ein großer Fokus auf dem Verlauf der Energieverbräuche unter bestimmten Charakteristika. In dieser Hypothese liegt der Fokus jedoch auf dem direkten Vergleich des gesamten Energieverbrauchs über einen bestimmten Zeitraum zwischen verschiedenen Einkommensklassen. Die Aufgabe besteht daher insbesondere darin, die gemessenen Gasverbräuche vergleichbar zu machen, da neben dem Nutzerverhalten noch eine Reihe weiterer Faktoren Einfluss auf den Gasverbrauch haben. Deshalb müssen, um nur die Effekte des Nutzerverhaltens, bzw. Einkommens zu untersuchen, alle sonstigen Einflussfaktoren auf die gleichen Werte gesetzt werden. Dazu zählen unter anderem die bereits in Hypothese 2 verwendeten Informationen über die technische Gebäudeausstattung. Darüber hinaus muss auch die Energieeffizienz der Gebäude hinzugezogen werden. Würde diese ignoriert, würden wohlhabendere Haushalte in der Analyse davon „profitieren“, dass sie eventuell schlichtweg in neueren, energieeffizienteren Gebäuden wohnen. Darüber hinaus muss aber auch die Wohnfläche berücksichtigt werden. Schließlich bedeutet ein höherer Gasverbrauch aufgrund einer schlichtweg größeren, zu beheizenden Nutzfläche nicht gleich ein weniger energiesparendes Nutzerverhalten. Folglich besteht in dieser Analyse im Gegensatz zu den vorangegangenen Analysen kein Interesse an dem bloßen Energieverbrauch in kWh, sondern an dem Energieverbrauch pro Fläche, also kWh/m$^2$. Dementsprechend müssen die gemessenen Gasverbräuche mit der Nutzfläche skaliert werden. Die Faktentabelle des DW muss daher angereichert werden. Die letzte Dimension, die für die Analyse relevant ist, ist dann der Wohlstand des Haushaltes. Es bestand der Versuch, verschiedene Informationen mit Bezug zur finanziellen Lage der Haushalte durch eine Cluster-Analyse zu einem einzigen Wohlstandsfaktor zusammenzufassen. Da die Ergebnisse jedoch nicht ausreichend zufriedenstellend waren, wurde zusätzlich eine Dimension für das jährliche Haushaltseinkommen hinzugezogen.

### 5.4.1. Aufbau des Data Warehouse und Integration der Daten

#### Faktentabelle (FactTable)

Da die vorherige Datenaufbereitung für diese Hypothese aufgrund der Anreicherung der Faktentabelle recht rechenintensiv ist und Hurst zeigt, dass eine Betrachtung des gesamten Zeitraums nicht zwangsläufig zur mehr Ergebnissen führt (Hurst et al., 2020, pp. 7881-7882), werden diesmal nur einige ausgewählte Wochen verwendet. Insgesamt werden je vier Wochen aus den Jahres Winter,

Frühling und Sommer verwendet, um zusätzlich die saisonalen Unterschiede zu berücksichtigen. Dabei handelt es sich um die Wochen 1 bis 4 im Dezember 2009, die Wochen 17-20 für während des Frühjahres 2010 und die Wochen 37-40 während des Spätsommers 2010. Damit wären sowohl eine kalte, eine gemäßigte und eine warme Jahreszeit abgedeckt. Im ersten Schritt wird die Faktentabelle erstellt.

```
CREATE TABLE FactTable (
    MeterID INTEGER,
    TimeID  INTEGER,
    Usage   DOUBLE
);
```

Anschließend wurden die Dateien in die Tabelle importiert.

**Dimension Zeit (DimTime)**

Die Erstellung und Aufbereitung der Zeitdimension wurde in den vorangegangenen Beispielen bereits ausführlich dargestellt. Da das zugrundeliegende Zeitformat identisch ist, ist das Vorgehen bei der Transformation und Integration der Zeitdimension identisch. Daher werden die durchzuführenden Schritte hier nicht erneut erwähnt. Diese können bei Bedarf bei Hypothese 1 und 2 eingesehen werden.

**Dimension Gasverbrauchsquellen**

Wie bereits bei Hypothese 2 sollten auch hier die unterschiedliche Gebäudeausrüstung berücksichtigt werden. Es werden erneut die Zusatzheizungen, die Trinkwassererwärmung, die Kochstellen, optische Feuer und gasbetriebene Trockner einbezogen. Diese Daten wurden bereits für Hypothese 2 aufbereitet und liegen daher schon für den Import bereit. Da das Vorgehen identisch ist, wird dies hier nicht noch einmal im Detail erläutert. Allerdings wird auch hier zunächst wieder eine Hilfstabelle erzeugt, in die die aufbereiteten Werte importiert werden. Anschließend wurden auch in diesem Fall die Primärschlüssel der Dimensionen Zusatzheizungen und Trinkwassererwärmung aufgrund der Binärkodierung überarbeitet. Anschließend wurden die einzelnen Dimensionstabellen erstellt und anhand der Hilfstabelle mit Werten gefüllt.

**Dimension Energieeffizienz (DimEER)**

Die Energieeffizienz der Gebäude sollte ebenfalls durch eine spezielle Dimension berücksichtigt werden. In den Umfragen wurden Fragen zur Energieeffizienzbewertung (Energy Efficiency Rating) des jeweiligen Wohngebäudes gestellt. Diese könnten theoretisch für die Analyse verwendet werden. Allerdings wurde von den meisten Teilnehmern des Pilotprojektes angegeben, dass ihr Gebäude kein offizielles Energy Efficiency Rating (EER) aufweist oder dass ihr Gebäude ein EER hat, dieses ihnen jedoch nicht bekannt ist. Von den mehreren Tausend Gebäuden wiesen daher nur eine Handvoll Gebäude ein bekanntes EER auf. Würde dieses dennoch verwendet, stünden nur noch sehr wenige Gebäude zur Auswahl. Da bei diesen jedoch vielleicht sonstige, für die Analyse relevante Informationen fehlen, wäre es möglich, dass letztendlich gar keine Gebäude mehr übrigblieben, für die

sowohl alle Werte existieren als auch identisch sind. Allerdings wurden in der Umfrage weitere Daten erhoben, die mit der Energieeffizienz des Gebäudes zusammenhängen. Die Verwendung dieser Werte führt jedoch zu zwei Problemen. Zum Einen ist keine dieser Angaben ein wirklicher Ersatz für ein EER, da es sich immer nur um sehr spezielle Angaben handelt. Beispielsweise ist der Anteil mehrfachverglaster Fenster zwar relevant für die Energieeffizienz, repräsentiert jedoch nur einen bestimmten Teil aller Parameter, die einen Einfluss auf die Energieeffizienz haben. Gleiches gilt beispielsweise für die Dämmung des Dachbodens. Daher stellt sich die Frage, welcher dieser Faktoren am ehesten als Ersatz für das EER verwendet werden sollte. Theoretisch wäre es natürlich möglich, alle Faktoren mit Energieeffizienzbezug einzeln einzubeziehen. Dadurch würden jedoch sehr viele verschiedene Dimensionstabellen entstehen, das DW würde sehr unübersichtlich und die Abfragen würden sehr lang. Das zweite Problem dieser Herangehensweise besteht durch fehlende Daten. Fast alle dieser Faktoren weisen für eine teils nicht unwesentliche Anzahl von Gebäuden keine Angaben auf. Durch die Notwendigkeit, alle diese Faktoren in die Analyse miteinzubeziehen, würden am Ende nur die Gebäude übrigbleiben, die für alle Faktoren Werte aufweisen. Dementsprechend würde bereits allein durch die Energieeffizienzdimension der überwiegende Teil der vorliegenden Daten eliminiert. Werden dann weitere Dimensionen eingeschränkt, könnten unter Umständen gar keine Werte mehr übrigbleiben.

Daher erfolgte der Versuch, diese einzelnen Faktoren durch eine Cluster-Analyse zu verschiedenen Clustern zusammenzufassen, sodass diese letztendlich durch einen einzigen Wert repräsentiert werden. Als Ergebnis entsteht dann zwar keine Kennzahl, jedoch ist über die Werte eines einzelnen Clusters erkennbar, wo die Energieeffizienz eines Gebäudes auf einer Skala von schlecht bis gut angesiedelt ist. Auf diese Weise wird das Problem durch die vielen fehlenden Angaben in den Umfrageergebnisse umgangen und es muss lediglich eine Dimensionstabelle DimEER erstellt werden. Für die Energieeffizienz wurde das Alter des Gebäudes, das Alter der Heizung in Kategorien, der Anteil mehrfach verglaster Fenster, die Isolierbekleidung des Warmwasserboilers, die Dämmung des Dachbodens, die Dämmung der Außenwände und das Inspektionsintervall des Heizkessels in Kategorien berücksichtigt. Sonstige relevante Umfrageergebnisse lagen nicht vor. Diese wurden zunächst in Excel dekodiert und dann vorformatiert, um Unreinheiten oder unterschiedliche Einheiten zu beseitigen. Anschließend wurden die Werte in eine gemeinsame Tabelle zusammengefasst und als CSV-Datei exportiert.

Zur Durchführung einer Cluster-Analyse wurde R verwendet. R ist ein kostenlos verfügbares, umfangreiches Statistikprogramm und bietet durch Bibliotheken viele verschiedene Cluster-Algorithmen. Einige der Faktoren der Energieeffizienz liegen in binärer Form vor. Bekannte Clusteralgorithmen wie K-Means können jedoch ausschließlich metrische Werte miteinbeziehen. Stattdessen wurde zur Berechnung der Distanzen zwischen den einzelnen Elementen der Gower-Koeffizient verwendet. Dieser kann neben metrischen und ordinalskalierten Werten auch Binärwerte clustern. Im ersten Schritt wurde zunächst die CSV-Datei in R importiert. Im zweiten Schritt wurde dann unter Verwendung des Gower-Koeffizienten als Distanzmaß die Distanzmatrix aufgebaut. Im dritten Schritt wurden dann alle Elemente zu Klassen verschmolzen. Hierfür stehen die drei Linkage-Verfahren Single-Linkage, Average-Linkage und Complete-Linkage zur Verfügung. Die Verschmelzung der Elemente zu Klassen wurde mit allen drei Verfahren durchgeführt. Im Anschluss wurde anhand des

Korrelationskoeffizienten zwischen der Distanzmatrix und der kophenetischen Distanzmatrix das am besten geeignete Verfahren identifiziert und ausgewählt. Im Anschluss sollte dann die Klassenanzahl bestimmt werden, also wie viele Cluster genau gebildet werden sollen. Dazu wurden zunächst die einzelnen Verschmelzungsniveaus berechnet. Das Mojena-Kriterium besagt, dass die Anzahl der Klassen der Anzahl der standardisierten Verschmelzungsniveau größer 2,75 entsprechen sollte. Daher werden alle Verschmelzungsniveaus anhand des Durchschnitts und der Standardabweichung standardisiert und geprüft, wie viele davon größer als 2,75 sind. Dieses Kriterium wird von 41 Verschmelzungsniveaus erfüllt, sodass 41 Cluster gebildet werden sollten. Anschließend wird für jede dieser 41 Klassen der Median sämtlicher Faktoren errechnet, da der Median robuster gegenüber Ausreißern ist als das arithmetische Mittel. Basierend auf den Ergebnissen wird eine Datei erzeugt, die für jede Meter ID das zugehörige Cluster angibt und später als Hilfstabelle für die Zuordnung der Cluster zu den Messdaten in der Faktentabelle dienen wird. Ferner wurde eine Datei erzeugt, die alle Cluster und die zugehörigen Werte der einzelnen Faktoren aufzählt und als Dimensionstabelle dienen soll. Der gesamte R-Code ist im Folgenden dargestellt.

```
> # Bibliothek "Cluster" installieren und verwenden
> install.packages("cluster")
> library(cluster)

> # Datei einlesen
> # Spalte 1 beinhaltet die MeterIDs als Bezeichnungen der Zeilen
> X <- read.table("/dateipfad/dateiname.csv", header=T, row.names=1, sep=";")

> # Distanzmatrix nach Gower-Koeffizient berechnen
> # Spalte 4 und 6 stellen binäre Merkmale dar
> D <- daisy(X, metric="gower", stand = F, type=list(symm = c(4,6)))

> # Clusteranalyse nach den den Verfahren Complete, Single und Average
> Cmp <- hclust(D,method="complete")
> Sng <- hclust(D,method="single")
> Avg <- hclust(D,method="average")

> # Kophenetische Distanzmatrix für die verschiedenen Verfahren berechnen
> D_Cmp <- cophenetic(Cmp)
> D_Sng <- cophenetic(Sng)
> D_Avg <- cophenetic(Avg)

> # Kophenetische Korrelationskoeffizienten als Gütemaß berechnen
> cat("Compelete-Linkage:", cor(D,D_Cmp), ", Single-Linkage:", cor(D,D_Sng), ",
Average-Linkage:", cor(D,D_Avg))
Compelete-Linkage: 0.6658854 , Single-Linkage: 0.5977059 , Average-Linkage:
0.7273952

> # Average-Linkage wählen, da es die höchste Korrelation aufweist

> # Verschmelzungsniveaus berechnen
> a <- Avg$height

> # Verschmelzungsniveaus standardisieren, um nach Mojena-Kriterium die
Gruppenanzahl zu bestimmen
> (a-mean(a))/sd(a)>2.75

> # Es sollten 41 Gruppen gebildet werden
> Cluster <- cutree(Avg, k = 41)
```

```
> # Cluster zu der Tabelle hinzufügen
> X <- cbind(X, Cluster)

> # Durchschnitt und Median der Werte der einzelnen Gruppen bestimmen
> mean <- aggregate(X[, 1:7], list(X$Cluster), mean)
> median <- aggregate(X[, 1:7], list(X$Cluster), median)
> print(mean)
> print(median)

> # Beide Tabellen als Datei exportieren.
> write.table(X, file="/dateipfad/EER_Cluster.csv", sep=";", eol="\n")
> write.table(median, file="/dateipfad/EER_Cluster_Values.csv", sep=";",
eol="\n")
```

Zunächst wurde die Hilfstabelle zur späteren Zuordnung erstellt und die Datei importiert.

```
CREATE TABLE EERIntegration (
    MeterID INTEGER PRIMARY KEY,
    Cluster INTEGER
);
```

Im nächsten Schritt wurde die Dimensionstabelle für die Energieeffizienz erstellt und die Datei mit den errechneten Clustern und dazugehörigen Faktorwerten importiert.

```
CREATE TABLE DimEER (
    PK_EER           INTEGER PRIMARY KEY,
    EnergyEfficiency STRING,
    GasUsageDuration STRING,
    Windows          STRING,
    WaterCylinder    STRING,
    AtticInsulation  STRING,
    WallInsulation   STRING,
    BoilerService    STRING
);
```

Die einzelnen Werte der Spalten sollten dann in eine Spalte zusammengefasst werden, sodass anhand dieser Spalte alle Werte erkennbar sind und Messdaten hinsichtlich dieser Werte gruppiert werden können. Alle Spalten außer der Primärschlüsselspalte und dieser können dann gelöscht werden. Die Dimensionstabelle ist damit fertig integriert.

```
UPDATE DimEER SET EnergyEfficiency = EnergyEfficiency || ", " ||
GasUsageDuration || ", " || Windows || ", " || WaterCylinder || ", " ||
AtticInsulation || ", " || WallInsulation || ", " || BoilerService;
```

**Dimension Wohlstand (DimWealth)**

Als weitere Dimension sollte der Wohlstand der Haushalte betrachtet werden. Wohlstand hängt von vielen verschiedenen Faktoren ab und kann nicht durch eine einzige Zahl erhoben werden. Stattdessen wurden in den Umfragen mehrere Fragen mit Bezug auf den Wohlstand gestellt. Da auch hier eine Integration sämtlicher einzelner Faktoren als eigene Dimension nicht sinnvoll und zielführend wäre, wurde auch hier versucht, die Haushalte hinsichtlich der einzelnen Werte in Klassen

zusammenzufassen. Dafür wurde erneut eine Cluster-Analyse durchgeführt. Die dafür verwendeten Faktoren sind die Art der Beschäftigung – also ob beispielsweise selbstständig, angestellt, arbeitslos, Arbeit suchend, im Ruhestand oder Sonstige – des Hauptverdieners oder der Hauptverdienerin, die soziale Klasse, die Eigentumsverhältnisse des Wohngebäudes, das Bildungsniveau und das Jahreseinkommen. Zudem wurde weiterhin die zur Verfügung stehende Wohnfläche pro Bewohner errechnet und hinzugezogen. Auch bei vielen dieser Faktoren fehlte jedoch ein nicht unbedeutender Teil der Angaben. Diese Faktoren beinhalten zwar keine Dummy-Variablen und damit Binärwerte, dafür aber eine Reihe kategorialer Werte wie die Art der Beschäftigung. Da kategoriale Werte nur sehr schwierig oder in Kombination mit sonstigen Werten gar nicht in Distanzberechnungen einbezogen werden können, bestand der Versuch, diese in ordinalskalierte Werte umzuwandeln, indem untersucht wurde, ob die Ausprägungen in eine Rangordnung überführt werden können. Dies ist insofern teilweise möglich, als dass aus Sicht des Wohlstandes beispielsweise das Mieten einer Wohnung weniger gut ist, als Eigentümer der Wohnung ohne ausstehenden Hypothekenkredit zu sein. Nachdem alle Werte in Excel dekodiert und formatiert wurden sowie alle kategorialen Faktoren bestmöglich in ordinalskalierte Faktoren umgewandelt wurden, entspricht das Vorgehen in R genau dem zuvor geschilderten Vorgehen bei der Klassifizierung der Energieeffizienz.

Anschließend wurden wieder Hilfs- und Dimensionstabelle erstellt, die entsprechenden Dateien importiert und in der Dimensionstabelle alle Werte in einer Spalte zusammengefasst.

```
CREATE TABLE WealthIntegration (
    MeterID INTEGER PRIMARY KEY,
    Cluster INTEGER
);

CREATE TABLE DimWealth (
    PK_Wealth   INTEGER PRIMARY KEY,
    Wealth      STRING,
    SocialClass STRING,
    Ownership   STRING,
    Education   STRING,
    Income      STRING,
    Size        STRING
);

UPDATE DimWealth SET Wealth = Wealth || ", " || SocialClass || ", " ||
Ownership || ", " || Education || ", " || Income || ", " || Size || "m2";
```

Anschließend können erneut alle Spalten der Dimensionstabelle außer der Primärschlüsselspalte und der Spalte Wealth gelöscht werden. Die Ergebnisse der Cluster-Analyse waren in diesem Fall jedoch nicht ausreichend zufriedenstellend. Da viele Angaben der einzelnen Faktoren fehlen, enthalten auch viele Faktoren der einzelnen Cluster, die auch hier wieder als Median sämtlicher zugehöriger Werte errechnet wurden, keine Angaben. Somit lässt sich nicht ausreichend viel über den Wohlstand der Haushalte eines einzelnen Clusters aussagen. Es liegen schlichtweg für die meisten Wohlstandsfaktoren zu wenige Angaben vor. Neben den vielen fehlenden Angaben könnte zudem die Modellierung von Kategorien als Rangordnungen problematisch gewesen sein.

**Dimension Einkommen (DimIncome)**

Daher wurde für die Analyse zusätzlich die Dimension Einkommen betrachtet. Bei den späteren Abfragen wurde dann auch diese Dimension anstelle der Dimension Wohlstand verwendet. In der Umfrage wurde nach dem Gesamtjahreseinkommen des Haushaltes vor Steuern gefragt. Dabei gab es die Möglichkeit, das Einkommen genau anzugeben, stattdessen eine Kategorie anzugeben oder keine Angabe zu machen. Obwohl viele Haushalte eine Angabe verweigert haben, liegen dennoch ausreichend viele Werte vor. Die Aufbereitung der Daten erfolgte in Excel, gestaltete sich jedoch etwas umständlich, da die Daten teils sehr unterschiedlich angegeben wurden. Einige Haushalte gaben das Wocheneinkommen, andere das Monatseinkommen oder das Jahreseinkommen an. Diese wurden allesamt auf das Jahreseinkommen umgerechnet. Zudem wurde das Einkommen manchmal vor Steuern, manchmal nach Steuern angegeben. Alle Angaben wurden auf das Einkommen vor Steuern umgerechnet, da die Mehrheit der Werte bereits vor Steuern angegeben waren. Dafür wurden die irischen Steuerklassen berücksichtigt. Ferner gab es einige Ausreißer. Um die Analyseergebnisse nicht zu verfälschen wurden sämtliche ungewöhnlich hohe Jahreseinkommen nicht berücksichtigt, da anzunehmen war, dass die Angaben fehlerhaft sind. Um Gleichheit herzustellen, wurden außerdem alle direkt angegebenen Jahreseinkommen einer der vorgegebenen Kategorien zugeordnet, da eine Zuordnung in die umgekehrte Richtung logischerweise nicht möglich wäre. Insgesamt bestehen dann die folgenden fünf Einkommenskategorien:

- Jahreseinkommen unter 15.000 Euro
- Jahreseinkommen zwischen 15.000 und 30.000 Euro
- Jahreseinkommen zwischen 30.000 und 50.000 Euro
- Jahreseinkommen zwischen 50.000 und 75.000 Euro
- Jahreseinkommen über 75.000 Euro

Genau genommen wurden alle diese Formatierung bereits für die Cluster-Analyse der Dimension Wohlstand vorgenommen. Aufgrund der Vielzahl der Faktoren wurden diese Schritte jedoch nicht bei der Dimension Wohlstand, sondern im Rahmen dieser Dimension erwähnt. Die Datei wird dann als CSV-Datei gespeichert und kann in das DW importiert werden.

Zunächst wurde eine Hilfstabelle mit den Meter IDs und den zugehörigen Werten der Dimension Wohlstand erstellt.

```
CREATE TABLE IncomeIntegration (
    MeterID    INTEGER PRIMARY KEY,
    IncomeCode INTEGER,
    Income     STRING
);
```

In diese kann dann die CSV-Datei importiert werden. Aufgrund der Excel-Formatierung sind fehlende Werte als “ “ kodiert und sollten daher nachträglich als explizite NULL Werte deklariert werden.

```
UPDATE IncomeIntegration SET IncomeCode = NULL WHERE IncomeCode = "";
UPDATE IncomeIntegration SET Income = NULL WHERE Income = "";
```

Dann kann die Dimensionstabelle Income erstellt und mit den gruppierten Werten der Hilfstabelle gefüllt werden.

```
CREATE TABLE DimIncome (
    PK_Income INTEGER PRIMARY KEY,
    Income    STRING
);

INSERT INTO DimIncome (PK_Income, Income)
SELECT IncomeCode, Income
FROM IncomeIntegration
WHERE IncomeCode IS NOT NULL
GROUP BY IncomeCode;
```

Damit ist der Integrationsvorgang der Einkommensdimension abgeschlossen.

**Gasverbrauch pro Fläche**

Nachdem das grundlegende Sternschema mit den für die Hypothese relevanten Dimensionstabellen aufgebaut ist, wird die Faktentabelle angereichert. Im Rahmen der Anreicherung werden neue Spalten hinzugefügt, die Kennzahlen beinhalten, die weitere oder aussagekräftigere Informationen liefern als die bisherigen Spalten für Fakten. In Bezug auf Energieverbrauchsdaten gibt es besonders eine relevante Kennzahl, den Verbrauch pro Fläche. Dieser ist aussagekräftiger über den eigentlichen Verbrauch und die Energieeffizienz eines Gebäudes, da er den Energieverbrauch mit der Nutzfläche normiert. Nur so sind Gebäude und Energieverbräuche wirklich miteinander vergleichbar, da der Energieverbrauch, insbesondere für die Raumkonditionierung, von der zu beheizenden, bzw. zu kühlenden Fläche, die sich aus den definierten Systemgrenzen ergibt, abhängt.

Im vorliegenden Fall wurden im Rahmen einer vor dem Pilotprojekt durchgeführten Umfrage die Daten zur Gebäudenutzfläche erhoben. Diese sind mit allen anderen Umfrageergebnissen in einer Excel bzw. CSV-Datei gespeichert. Zunächst werden alle für den Fall nicht relevanten Spalten herausgelöscht, sodass nur noch die Meter ID und die Angaben zu den Nutzflächen übrigbleiben. Im Zuge der Umfrage konnte die Nutzfläche entweder in Quadratmeter oder Quadratfuß (square Feet) angegeben werden. Dies muss vor der Verwendung der Daten normiert werden. In Excel wurden daher vorerst alle in square Feet vorliegenden Daten in Quadratmeter umgerechnet. Sind keine geeigneten Programme vorhanden oder kommt es beim Export oder Import zu Problemen, besteht alternativ aber auch die Möglichkeit, die Bearbeitung in SQLite vorzunehmen. Außerdem muss auch in diesem Fall bedacht werden, alle eventuellen Kommata in Fließkommazahlen, die durch die Verarbeitung mit Excel oder anderen Programmen entstanden sind, vor dem Import in das DW mit Hilfe eines Text Editors in Punkte umzuwandeln, damit diese mit den SQLite Datentypen konform sind.

Nach der Aufbereitung der Nutzflächendaten kann die CSV-Datei in das DW importiert werden. Dafür muss zunächst eine Tabelle erstellt werden.

```
CREATE TABLE FloorArea (
    MeterID     INTEGER,
    SquareMeter INTEGER
);
```

Fehlende Angaben zur Nutzfläche sind mit 999.999.999 kodiert. Diese repräsentieren NULL Werte und sollten daher in solche umgewandelt werden. Insgesamt sind 691 der 1365 vorliegenden Wohngebäuden ohne Angabe der Nutzfläche.

```
UPDATE FloorArea SET SquareMeter = NULL WHERE SquareMeter = 999999999;
```

Ferner scheinen einige fehlerhafte Werte und Ausreißer vorzuliegen. Diese könnten auf bewusst fehlerhafte Angaben, ein Falschverstehen der Fragestellung, indem beispielsweise die Grundstücksfläche statt der Nutzfläche angegeben wurde, oder eine fehlerhafte Eintragung der Werte zurückzuführen sein. Teilweise sind die notierten Nutzflächen so groß, dass sie sich auch nicht auf eine fehlerhafte Angabe der Bezugsgröße zurückzuführen wären, indem Beispielsweise die Nutzfläche in Quadratmeter gemeint war, jedoch Quadratfuß als Einheit angegeben wurde. Alle Werte wurden bereits vorher in Quadratmeter umgerechnet. Daher wurden sämtliche Werte über 500 $m^2$ Nutzfläche als Ausreißer definiert und durch NULL Werte ersetzt. Dies betraf 35 Gebäude.

```
UPDATE FloorArea SET SquareMeter = NULL WHERE SquareMeter > 500;
```

Nun kann die eigentliche Anreicherung erfolgen. Zunächst wurde die Faktentabelle um eine Spalte für den Verbrauch pro Quadratmeter erweitert.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                          FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage      DOUBLE,
    UsagePerM2 DOUBLE,
    MeterFK    INTEGER,
    TimeFK     INTEGER
);

INSERT INTO FactTable (
                          Usage,
                          MeterFK,
                          TimeFK
                      )
                      SELECT Usage,
                             MeterID,
                             TimeID
                        FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Anschließend wurde jeder Messwert durch Teilung durch die Nutzfläche normiert, sodass der Verbrauch pro Quadratmeter angegeben ist.

```
UPDATE FactTable SET UsagePerM2 = Usage/(SELECT SquareMeter FROM FloorArea
WHERE FactTable.MeterFK = FloorArea.MeterID AND FloorArea.SquareMeter IS NOT
NULL);
```

In der Folge sind die Messwerte der Faktentabelle erfolgreich normiert und die Tabelle mit den Nutzflächen hat ihren Zweck erfüllt und kann daher gelöscht werden.

**Verknüpfung der Faktentabelle mit den Dimensionstabellen**

Nun kann die Faktentabelle mit den Dimensionstabellen verknüpft werden, indem jedem Messwert der Faktentabelle die entsprechenden Dimensionswerte zugeordnet werden. Im ersten Schritt wurde die Faktentabelle um die Fremdschlüsselspalten erweitert.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                        FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage          DOUBLE,
    UsagePerM2     DOUBLE,
    MeterFK        INTEGER,
    TimeFK         INTEGER,
    HeatingTypeFK  INTEGER,
    WaterHeatingFK INTEGER,
    CookerFK       INTEGER,
    FireEffectFK   INTEGER,
    DryerFK        INTEGER,
    EERFK          INTEGER,
    WealthFK       INTEGER,
    IncomeFK       INTEGER
);

INSERT INTO FactTable (
                          Usage,
                          UsagePerM2,
                          MeterFK,
                          TimeFK
                      )
                      SELECT Usage,
                             UsagePerM2,
                             MeterFK,
                             TimeFK
                        FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Im zweiten Schritt werden dann die Primärschlüssel der Dimensionstabellen als Fremdschlüssel übergeben und gespeichert.

```
UPDATE FactTable SET HeatingTypeFK = (SELECT HeatingID FROM DataIntegration
WHERE FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET WaterHeatingFK = (SELECT WaterHeatingID FROM
DataIntegration WHERE FactTable.MeterFK = DataIntegration.MeterID);
```

```
UPDATE FactTable SET CookerFK = (SELECT CookerID FROM DataIntegration WHERE
FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET FireEffectFK = (SELECT FireID FROM DataIntegration WHERE
FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET DryerFK = (SELECT DryerID FROM DataIntegration WHERE
FactTable.MeterFK = DataIntegration.MeterID);

UPDATE FactTable SET EERFK = (SELECT Cluster FROM EERIntegration WHERE
FactTable.MeterFK = EERIntegration.MeterID);

UPDATE FactTable SET WealthFK = (SELECT Cluster FROM WealthIntegration WHERE
FactTable.MeterFK = WealthIntegration.MeterID);

UPDATE FactTable SET IncomeFK = (SELECT IncomeCode FROM IncomeIntegration WHERE
FactTable.MeterFK = IncomeIntegration.MeterID);
```

Damit wurden die Dimensionswerte entsprechend zugeordnet und die Hilfstabellen für die Erstellung der Dimensionstabellen und Verknüpfung dieser mit der Faktentabelle können gelöscht werden. Nun gilt es, nicht relevante und fehlerhafte Werte aus der Faktentabelle zu löschen. Zunächst sollten auch hier die Messwerte von Gebäuden, die nicht mit Gas heizen, gelöscht werden. Weitere Einschränkungen des Datensatzes ergeben sich aber auch durch fehlende oder fehlerhafte Nutzflächenangaben und fehlende oder fehlerhafte Angaben des Einkommens.

Insgesamt liegen Messwerte von 1493 verschiedenen Smart Metern vor. Umfrageergebnisse existieren jedoch nur für 1365 Smart Meter. Ferner ist die Menge der 1365 Meter IDs, für die Umfrageergebnisse existieren, nicht vollständig Teil der Menge der 1493 Meter IDs, für die Messwerte vorliegen. Das heißt, dass einige Meter IDs Umfrageergebnisse, aber keine Messwerte aufweisen, während andere hingegen Messwerte aufweisen, jedoch keine Umfrageergebnisse. Folglich ist die Anzahl an Meter IDs, für die Messwerte, jedoch keine Umfrageergebnisse vorliegen, 197 – also nicht lediglich die Differenz aus beiden Mengen. Dieser Zustand wurde jedoch nicht in dem zur Verfügung gestellten Datensatz vermerkt.

Zunächst werden daher, wie schon zur Untersuchung von Hypothese 2, alle Gebäude gelöscht, für die keine Umfrageergebnisse vorliegen. Dies ist wie erwähnt für 197 Gebäude der Fall.

```
DELETE FROM FactTable WHERE HeatingTypeFK IS NULL;
```

Anschließend werden dann alle Werte von Gebäuden gelöscht, die nicht mit Gas heizen. Die Fremdschlüssel der Dimension Heizungstyp dieser Gebäude sind mit 999.999 kodiert. Wie zuvor entspricht dies 21 weiteren Gebäuden.

```
DELETE FROM FactTable WHERE HeatingTypeFK = 999999;
```

Die Anzahl der vorliegenden Gebäude ist somit bereits von 1493 auf 1275 gesunken. Werden dann alle Gebäude ohne entsprechende Angabe der Nutzfläche gelöscht, sinkt die Anzahl weiter auf 598. Werden darüber hinaus alle Gebäude ohne Angabe des Einkommens gelöscht, bleiben letztendlich nur

noch 483 Gebäude übrig. Diese beinhalten dann jedoch alle für die Analyse relevanten Angaben. Theoretisch könnte auch hier auf eine Löschung der Werte verzichtet werden und stattdessen auf eine Anpassung der Abfragebefehle zurückgegriffen werden. Dies macht die Formulierung der Abfragen länger und fehleranfälliger. Um eine vollständige, saubere und korrekte Integration und Speicherung der Daten zu gewährleisten wurde entschieden, alle fehlerhaften Werte der Faktentabelle vollständig zu löschen. Dies ermöglicht Abfragen ohne explizite Behandlung von eventuellen NULL Werten und gewährleistet ein "sauberes" DW. Ferner reduziert eine solche "Säuberung" des DW den benötigten Speicherplatzbedarf und steigert die Abfrageperformance, da die Faktentabelle kleiner wird.

Gleichzeitig birgt diese Entscheidung allerdings auch Nachteile. Der wesentliche Nachteil besteht in der Reduktion der vorliegenden Werte. Dadurch, dass beispielsweise Werte ohne Informationen zum Einkommen des Haushaltes gelöscht werden, gehen weniger Werte in die Analyse ein. Dies hat zweierlei Folgen. Sollten in diesem DW weitere Analysen durchgeführt werden – beispielsweise, da eine erste Analyse neue Erkenntnisse liefert und bisher ungeahnte Zusammenhänge aufdeckt, die näher untersucht werden sollen – und ist für diese Analysen beispielsweise die Dimension Einkommen nicht relevant, so liegen dennoch nur Werte vor, die Informationen über das Einkommen beinhalten. Wären also fehlende und fehlerhafte Werte dieser Dimension nicht gelöscht worden, stünden für diese Analysen mehr Werte zur Verfügung. Insgesamt legt eine solche "Säuberung" das DW also genau auf die Hypothese aus, kann es im Gegenzug aber auch weniger flexibel für eventuelle sonstige Abfragen machen. Bei diesem speziellen Beispiel ergibt sich jedoch noch ein weiteres Problem. Übrig bleiben am Ende noch die Messwerte von 483 Gebäuden. Dies scheint für eine Analyse zwar ausreichend groß, jedoch muss bedacht werden, dass auf Basis dieser Werte für die mehrdimensionale Analyse noch weitere Einschränkungen gemacht werden. Insbesondere müssen die Verbräuche vergleichbar gemacht werden, indem die weiteren Gasverbräuche einheitlich sind, also beispielsweise die gleiche Technologie zur Trinkwassererwärmung, die gleichen Kochstellen und die gleichen Trocknertypen zum Einsatz kommen. Jede weitere Dimensionseinschränkung verringert die verwendete Menge an Gebäuden weiter, sodass am Ende nur die Gebäude übrigbleiben, die am alle Kriterien erfüllen. Werden selten vorkommende Dimensionswerte abgefragt kann dies also dazu führen, dass nur sehr wenige Gebäude oder unter Umständen sogar gar keine Gebäude existieren, die alle Kriterien erfüllen. Insbesondere im Fall weniger Gebäude stellt sich dann die Frage nach der statistischen Signifikanz der Analyseergebnisse. Dies wurde in diesem Fall wie bereits bei Hypothese 2 dadurch abgefangen, dass für die Analyse die Dimensionswerte gewählt wurden, die am häufigsten vorkommen. Diese sind meist durch Einfachheit charakterisiert. Im Fall der Zusatzheizungen sind beispielsweise insbesondere die Kategorien "keine weiteren Heizungen", "offene Feuer" und "Elektroheizungen (zentral oder Speicher)" häufig vertreten, während spezielle Kombinationen verschiedener Heizungstypen eher selten zu finden sind.

Des Weiteren wird in diesem Zusammenhang die Relevanz der vorher durchgeführten Cluster-Analyse deutlich. Wären stattdessen alle Faktoren als einzelne Dimensionen berücksichtigt worden, würden nicht nur deutlich mehr Dimensionstabellen existieren als ohnehin schon der Fall, sondern es würden sich auch die Abfragebefehle deutlich verlängern und verkomplizieren, die Performance würde sinken, da noch mehr Tabellen verknüpft und durchlaufen werden müssen, und es würden kaum noch verwendbare Werte übrig bleiben, da jede einzelne Dimension ihre individuellen fehlerhaften Werte

aufweist. Würden alle diese Dimensionen einzeln eingeschränkt, gäbe es daher kaum bis gar keine Gebäude mehr, die alle Kriterien erfüllen.

```
DELETE FROM FactTable WHERE UsagePerM2 IS NULL;
DELETE FROM FactTable WHERE IncomeFK IS NULL;
```

Zum Abschluss werden dann noch alle Fremdschlüsselspalten als solche definiert und festgelegt, dass diese keine NULL Werte aufweisen dürfen.

```
CREATE TABLE sqlitestudio_temp_table AS SELECT *
                                          FROM FactTable;

DROP TABLE FactTable;

CREATE TABLE FactTable (
    Usage          DOUBLE,
    UsagePerM2     DOUBLE,
    MeterFK        INTEGER,
    TimeFK         INTEGER REFERENCES DimTime (PK_Time)
                           NOT NULL ON CONFLICT FAIL,
    HeatingTypeFK  INTEGER REFERENCES DimHeatingType (PK_Heating)
                           NOT NULL ON CONFLICT FAIL,
    WaterHeatingFK INTEGER REFERENCES DimWaterHeating (PK_WaterHeating)
                           NOT NULL ON CONFLICT FAIL,
    CookerFK       INTEGER REFERENCES DimCooker (PK_Cooker)
                           NOT NULL ON CONFLICT FAIL,
    FireEffectFK   INTEGER REFERENCES DimFireEffect (PK_Fire)
                           NOT NULL ON CONFLICT FAIL,
    DryerFK        INTEGER REFERENCES DimDryer (PK_Dryer)
                           NOT NULL ON CONFLICT FAIL,
    EERFK          INTEGER REFERENCES DimEER (PK_EER)
                           NOT NULL ON CONFLICT FAIL,
    WealthFK       INTEGER REFERENCES DimWealth (PK_Wealth)
                           NOT NULL ON CONFLICT FAIL,
    IncomeFK       INTEGER REFERENCES DimIncome (PK_Income)
                           NOT NULL ON CONFLICT FAIL
);

INSERT INTO FactTable (
                         Usage,
                         UsagePerM2,
                         MeterFK,
                         TimeFK,
                         HeatingTypeFK,
                         WaterHeatingFK,
                         CookerFK,
                         FireEffectFK,
                         DryerFK,
                         EERFK,
                         WealthFK,
                         IncomeFK
                     )
                     SELECT Usage,
                            UsagePerM2,
                            MeterFK,
                            TimeFK,
                            HeatingTypeFK,
```

```
                            WaterHeatingFK,
                            CookerFK,
                            FireEffectFK,
                            DryerFK,
                            EERFK,
                            WealthFK,
                            IncomeFK
                        FROM sqlitestudio_temp_table;

DROP TABLE sqlitestudio_temp_table;
```

Nun steht das DW für Abfragen bereit.

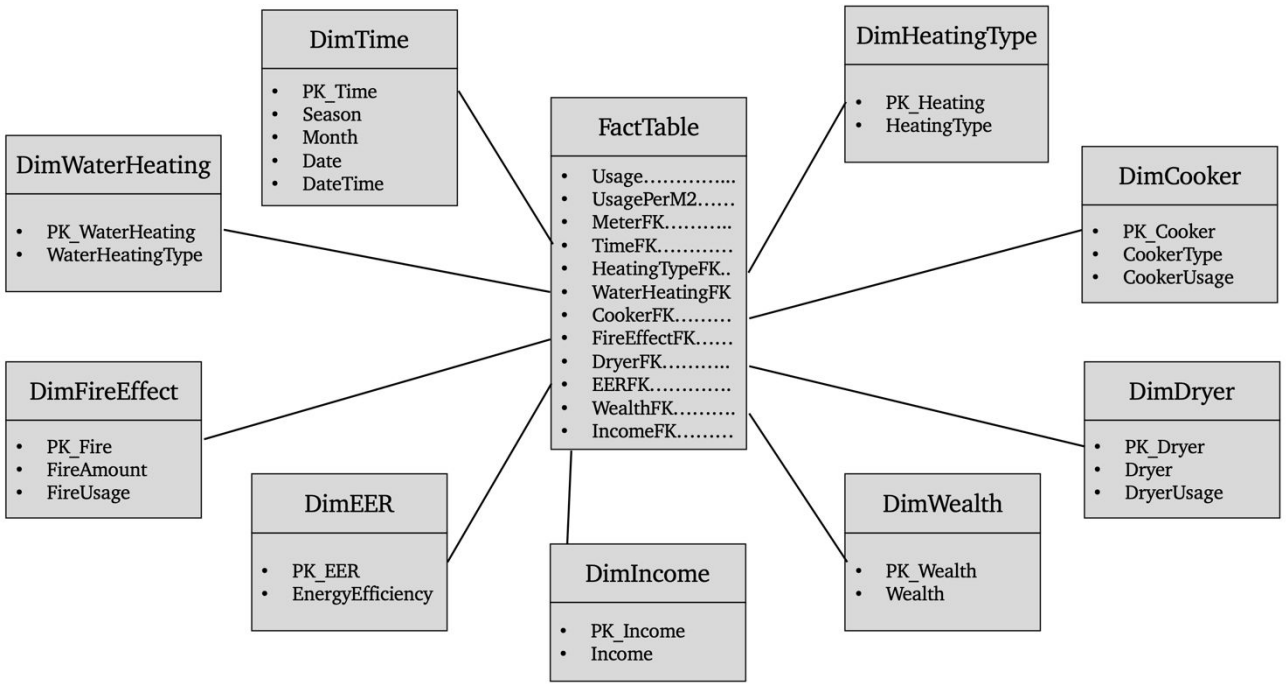



Abbildung 34: Data Warehouse für Hypothese 3

### 5.4.2. Abfrage und Visualisierung der Daten aus dem Data Warehouse

Die Abfrage der Daten zur Untersuchung dieser Hypothese unterscheidet sich in zwei wesentlichen Punkten von den Abfragen der vorangegangenen Hypothesen. Die zugrundeliegende Fragestellung erforderte bereits, dass die Gasverbrauchsdaten mit der Nutzfläche normiert werden, um so tatsächlich nur das Haushaltseinkommen und das damit zusammenhängende Nutzerverhalten zu berücksichtigen. Andernfalls könnten Faktoren wie die Wohnfläche, die eventuell auch mit dem Haushaltseinkommen zusammenhängen könnte, Einfluss auf die Ergebnisse haben und diese verfälschen. Daher besteht dieses Mal explizites Interesse an der Summe des Gasverbrauchs pro Quadratmeter. Da aber erstmals die Summe des Verbrauchs abgefragt wird, ergibt sich der zweite wesentliche Unterschied. Letztendlich besteht kein Interesse an der Gesamtsumme des Gasverbrauchs aller zugehörigen Haushalte, sondern an der durchschnittlichen Summe aller zugehörigen Haushalte. Würden schlichtweg alle zugehörigen Werte addiert, würde das Endergebnis zum Großteil schlichtweg von der

Anzahl der Gebäude abhängen und das Ergebnis würde stark verfälscht. Stattdessen soll für jedes einzelne Gebäude des Gesamtverbrauch ermittelt und dann der Durchschnitt aller Gesamtverbräuche gebildet werden, sodass die Werte der verschiedenen Einkommensgruppen miteinander verglichen werden können. In den vorangegangenen Hypothesen wurde hingegen schlicht der Durchschnitt abgefragt, sodass dieses Problem nicht bestand. Für diesen konnten einfach alle zugehörigen Werte zusammenaddiert und durch die Anzahl geteilt werden. Sollen allerdings zunächst die Summen der einzelnen Haushalte berechnet und dann der Durchschnitt aus ebendiesen gebildet werden, so ist dies durch eine einfache SQL-Abfrage nicht möglich. Dementsprechend werden in diesem Fall zunächst die Messwerte hinsichtlich der einzelnen Gebäude aggregiert. Die Bildung des Durchschnitts dieser Werte erfolgt dann in Python. Somit erfolgt für diese Hypothese die Berechnung in zwei Schritten.

Um die Einflüsse des Einkommens auf den Gasverbrauch zu untersuchen, muss zunächst die Dimension Einkommen herangezogen werden. In einer ersten Untersuchung werden zunächst alle Gebäude, unabhängig von der Energieeffizienz und der technischen Gebäudeausrüstung, analysiert. Dabei wird neben dem Einkommen auch der Einsatz optischer Feuer untersucht. Da optische Feuer eine Art „Luxusgut“ darstellen, könnten diese in Kombination mit dem Haushaltseinkommen ebenfalls einen Einfluss haben. Außerdem werden die Gasverbräuche nach den drei verwendeten Jahreszeiten gegliedert. Dadurch ergibt sich folgende Abfrage.

```
SELECT sum(UsagePerM2) AS Usage, MeterFK, FireAmount, Season, PK_Income
FROM FactTable, DimFireEffect, DimTime, DimIncome
WHERE FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.TimeFK = DimTime.PK_Time
AND FactTable.IncomeFK = DimIncome.PK_Income
GROUP BY FireAmount, Season, PK_Income, MeterFK
```

Nach dem Import der Ergebnistabelle werden in Python zunächst die Werte nach den Jahreszeiten und der Verwendung optischer Feuer gefiltert. Anschließend werden diese weiter in die Einkommensklassen aufgeteilt und für jede Einkommensklasse der Durchschnitt aus den zugrundeliegenden Gebäuden ermittelt. Dann können die Werte visualisiert werden.

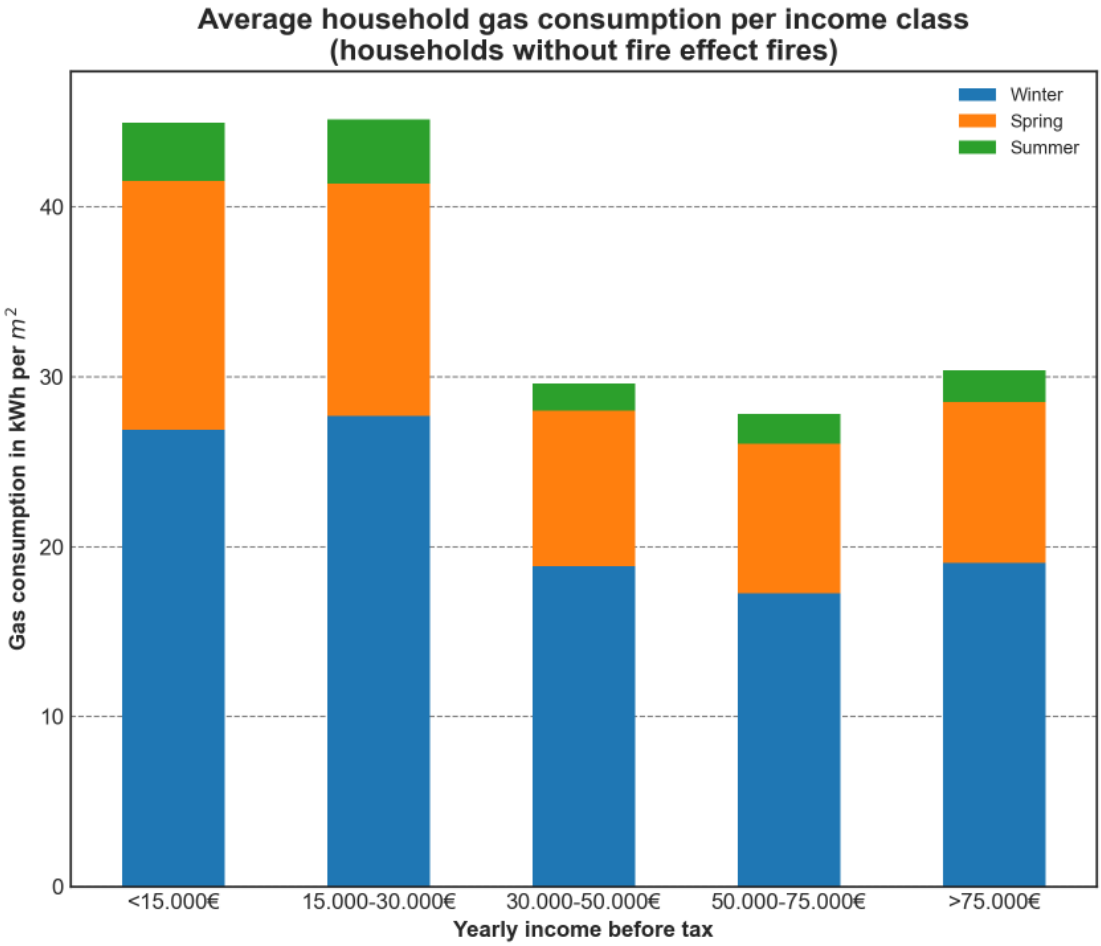



Abbildung 35: Durchschnittlicher Gasverbrauch pro Haushalt nach Einkommensklassen ohne optische Feuer

Für die Haushalte ohne optische Feuer ergibt sich ein unerwartetes Ergebnis. Anstatt dass der Gasverbrauch pro Fläche, wie in der Hypothese angenommen, konsequent mit dem Haushaltseinkommen ansteigt, ergibt sich stattdessen ein klarer Schnitt zwischen den Einkommensklassen 2 und 3, wobei die Haushalte mit höherem Einkommen einen deutlich geringeren Gasverbrauch aufweisen. Der Gasverbrauch der weniger wohlhabenderen Haushalte ist über 50% höher als der der wohlhabenden Haushalte. Dabei weisen die Gruppen 1 und 2 sowie die Gruppen 3, 4 und 5 jeweils beinahe identische Verbräuche auf. Außerdem wird in jeder Gruppe der klare Unterschied zwischen den Jahreszeiten deutlich, da der Gasverbrauch im Winter sehr hoch und im Sommer sehr niedrig ist. Für die Haushalte mit optischen Feuern ergibt sich allerdings ein anderes Bild.

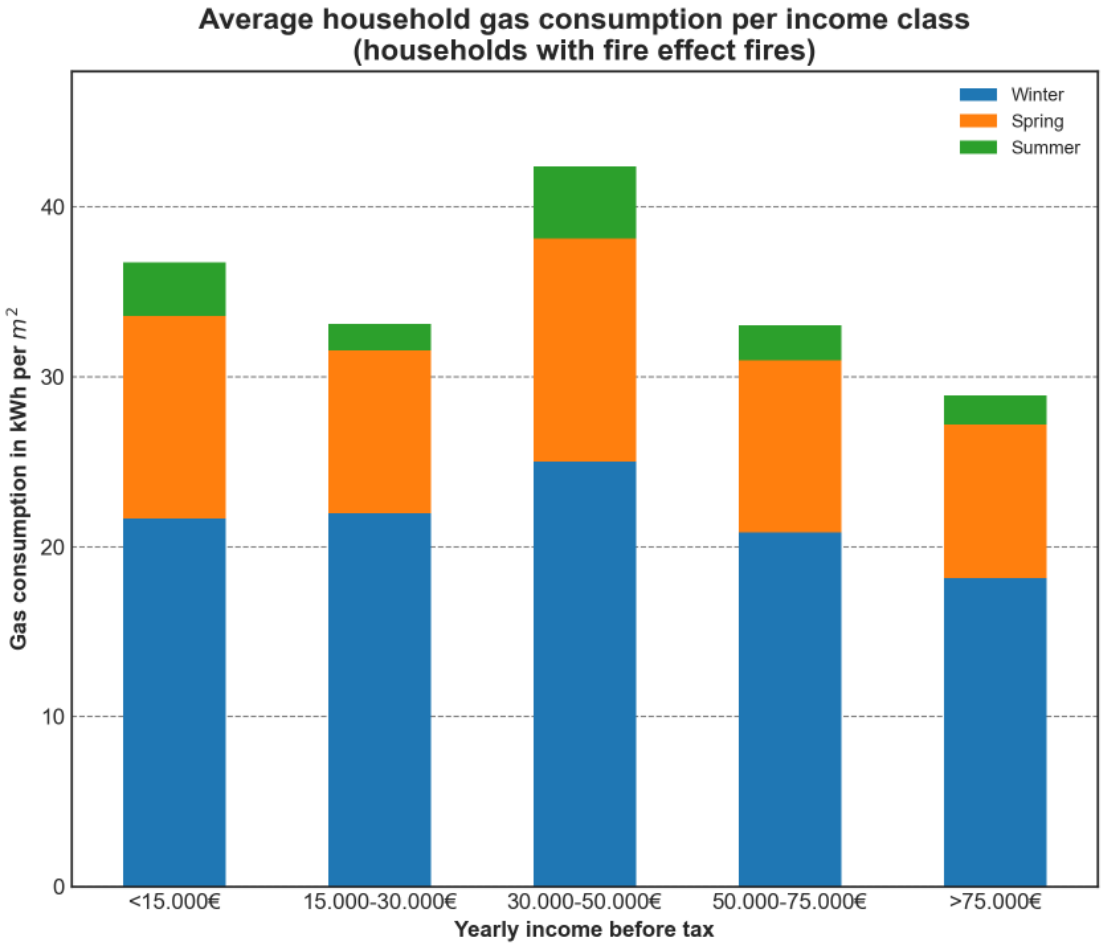



Abbildung 36: Durchschnittlicher Gasverbrauch pro Haushalt nach Einkommensklassen mit optischen Feuern

In diesem Fall ergibt sich kein klarer Schnitt und die Gasverbräuche der einzelnen Einkommensgruppen sind homogener. Auch wenn es auch in diesem Fall eine Gruppe gibt, deren Gasverbrauch 50% höher ist als der einer anderen Gruppe, so lässt sich hier kein eindeutiges Muster erkennen. Allerdings weist auch dieser Fall eine endgegenläufige Tendenz auf und der Gasverbrauch nimmt prinzipiell mit zunehmendem Einkommen ab. Auch die Haushalte mit optischen Feuern weisen klare, saisonale Unterschiede auf. Fraglich ist nun, ob die wohlhabenderen Haushalte von neueren, energieeffizienteren Gebäuden profitieren.

Um dies zu überprüfen sollten weiterhin Energieeffizienzklassen berücksichtigt werden. Die Cluster 4 und 7 sind sich in der energetischen Gebäudeausstattung sehr ähnlich, weisen jedoch unterschiedliche Baujahre auf. Die Gebäude in Cluster 4 sind zwischen 1935 und 1979 errichtet worden, während die Gebäude in Cluster 7 im Jahr 2005 oder später, also direkt vor der Durchführung des Pilotprojekts, erbaut wurden. Folglich wird bei dieser Abfrage die Energieeffizienz beachtet und nur Gebäude der Cluster 4 und 7 verwendet.

```
SELECT sum(UsagePerM2) AS Usage, MeterFK, PK_EER, PK_Income
FROM FactTable, DimEER, DimIncome
WHERE FactTable.IncomeFK = DimIncome.PK_Income
AND FactTable.EERFK = DimEER.PK_EER
AND (DimEER.PK_EER = 4 OR DimEER.PK_EER = 7)
GROUP BY MeterFK
```

Die Visualisierung führt auch in diesem Fall auf interessante Ergebnisse.

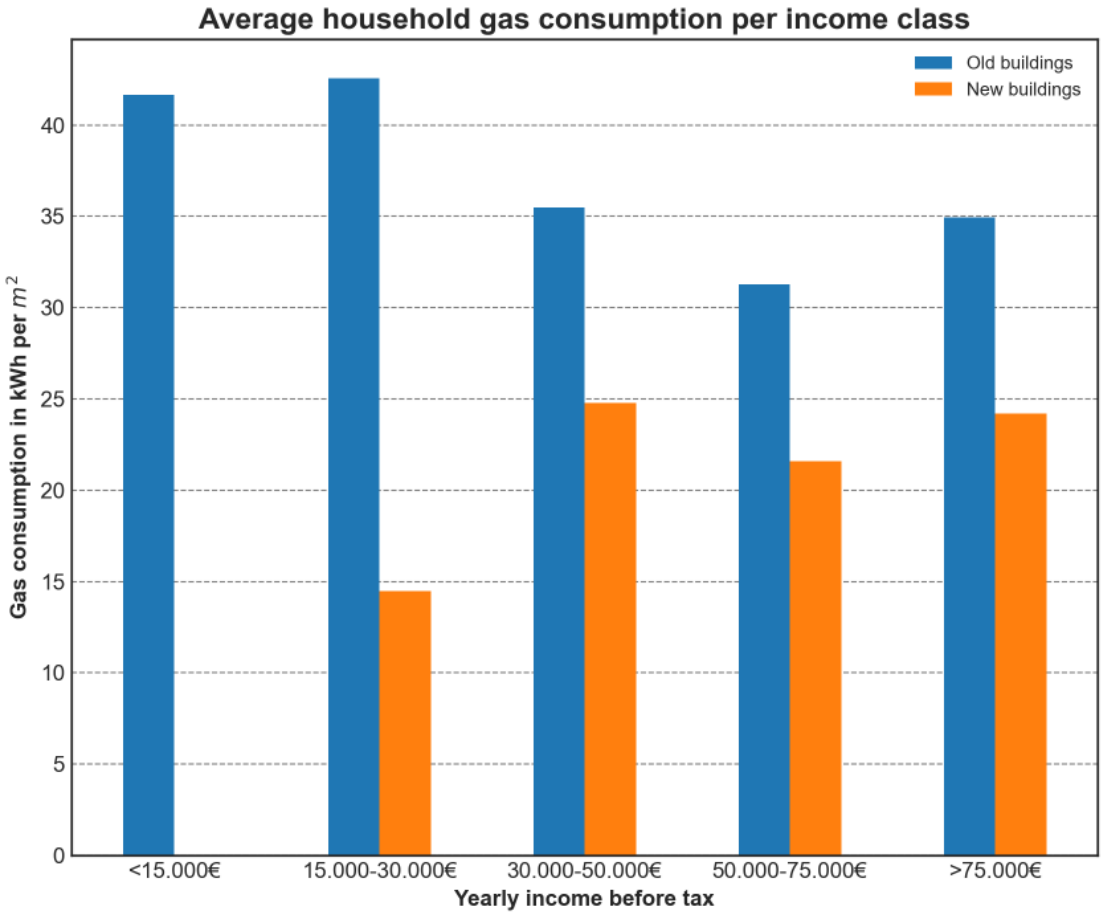



Abbildung 37: Durchschnittlicher Gasverbrauch pro Haushalt nach Gebäudealter

Für die alten Gebäude ergibt sich ein ähnliches Muster wie für die zu Beginn untersuchten Gebäude ohne optische Feuer. Auch bei dieser Gruppierung weisen die niedrigen Einkommensklassen 1 und 2 einen höheren Verbrauch auf. Für die neuen Gebäude ergibt sich jedoch ein entgegenläufiger Trend. Hier steigt der Energieverbrauch mit zunehmendem Einkommen an. Somit lässt sich bisher keine allgemein gültige Aussage bezüglich des Einkommens machen. In einem letzten Schritt wurde daher nochmals versucht, durch eine Festlegung sämtlicher Dimensionswerte auf konstante Werte, eventuelle sonstige Einflussfaktoren zu eliminieren, um so zu eindeutigen Ergebnissen zu kommen. Dies führt allerdings zu dem inhärenten Problem, dass die bereits eher geringe Anzahl an Gebäuden weiter verringert wird und am Ende kaum noch Gebäude für die Abfrage herangezogen werden. Daher wurde

Wert darauf gelegt, alle Dimensionswerte so festzulegen, dass möglichst viele Gebäude diese erfüllen. Bezüglich der Energieeffizienz weisen die Cluster 1 und 4 viele Gebäude auf, welche die sonstigen Einschränkungen erfüllen. Zunächst wird dabei Cluster 1 untersucht.

```
SELECT sum(UsagePerM2) AS Usage, MeterFK, HeatingType, WaterHeatingType,
CookerType, FireAmount, Dryer, PK_EER, Season, PK_Income
FROM FactTable, DimHeatingType, DimWaterHeating, DimCooker, DimFireEffect,
DimDryer, DimEER, DimTime, DimIncome
WHERE FactTable.HeatingTypeFK = DimHeatingType.PK_Heating
AND FactTable.WaterHeatingFK = DimWaterHeating.PK_WaterHeating
AND FactTable.CookerFK = DimCooker.PK_Cooker
AND FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.DryerFK = DimDryer.PK_Dryer
AND FactTable.EERFK = DimEER.PK_EER
AND FactTable.TimeFK = DimTime.PK_Time
AND FactTable.IncomeFK = DimIncome.PK_Income
AND DimHeatingType.HeatingType = 'None of these'
AND DimWaterHeating.WaterHeatingType = 'Other'
AND DimCooker.CookerType = 'Gas cooker'
AND DimFireEffect.FireAmount = '1 fire'
AND DimDryer.Dryer = 'No'
AND DimEER.PK_EER = 1
GROUP BY MeterFK, Season
```

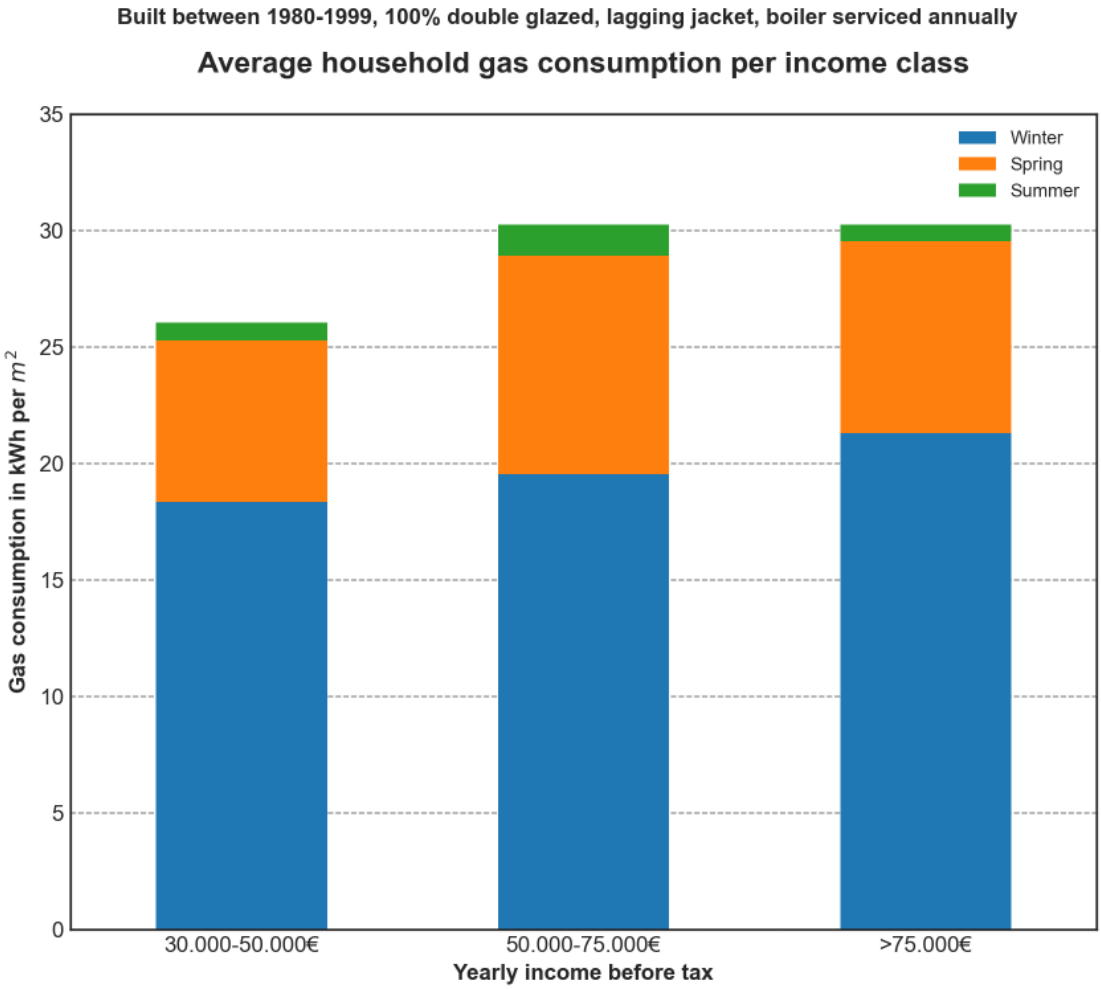



Abbildung 38: Durchschnittlicher Gasverbrauch pro Haushalt für Cluster 1

Für diese Gebäude steigt der Energieverbrauch mit dem Einkommen klar an. So ist der Energieverbrauch der Einkommensklassen 4 und 5 deutlich höher als der der Einkommensklasse 3. Allerdings weisen die Einkommensklassen 4 und 5 quasi identische Verbräuche auf, sodass die Visualisierung suggeriert, dass der Gasverbrauch ab einem Jahreseinkommen von 50.000 Euro vor Steuern nicht weiter ansteigt. Auch hier sind klar die saisonalen Effekte erkennbar. Abschließend wird dann Energieeffizienzcluster 4 analysiert.

```
SELECT sum(UsagePerM2) AS Usage, MeterFK, HeatingType, WaterHeatingType,
CookerType, FireAmount, Dryer, PK_EER, Season, PK_Income
FROM FactTable, DimHeatingType, DimWaterHeating, DimCooker, DimFireEffect,
DimDryer, DimEER, DimTime, DimIncome
WHERE FactTable.HeatingTypeFK = DimHeatingType.PK_Heating
AND FactTable.WaterHeatingFK = DimWaterHeating.PK_WaterHeating
AND FactTable.CookerFK = DimCooker.PK_Cooker
AND FactTable.FireEffectFK = DimFireEffect.PK_Fire
AND FactTable.DryerFK = DimDryer.PK_Dryer
AND FactTable.EERFK = DimEER.PK_EER
AND FactTable.TimeFK = DimTime.PK_Time
AND FactTable.IncomeFK = DimIncome.PK_Income
AND DimHeatingType.HeatingType = 'None of these'
```

```
AND DimWaterHeating.WaterHeatingType = 'Gas'
AND DimCooker.CookerType = 'Other cooker'
AND DimFireEffect.FireAmount = 'No fires'
AND DimDryer.Dryer = 'No'
AND DimEER.PK_EER = 4
GROUP BY MeterFK, Season
```

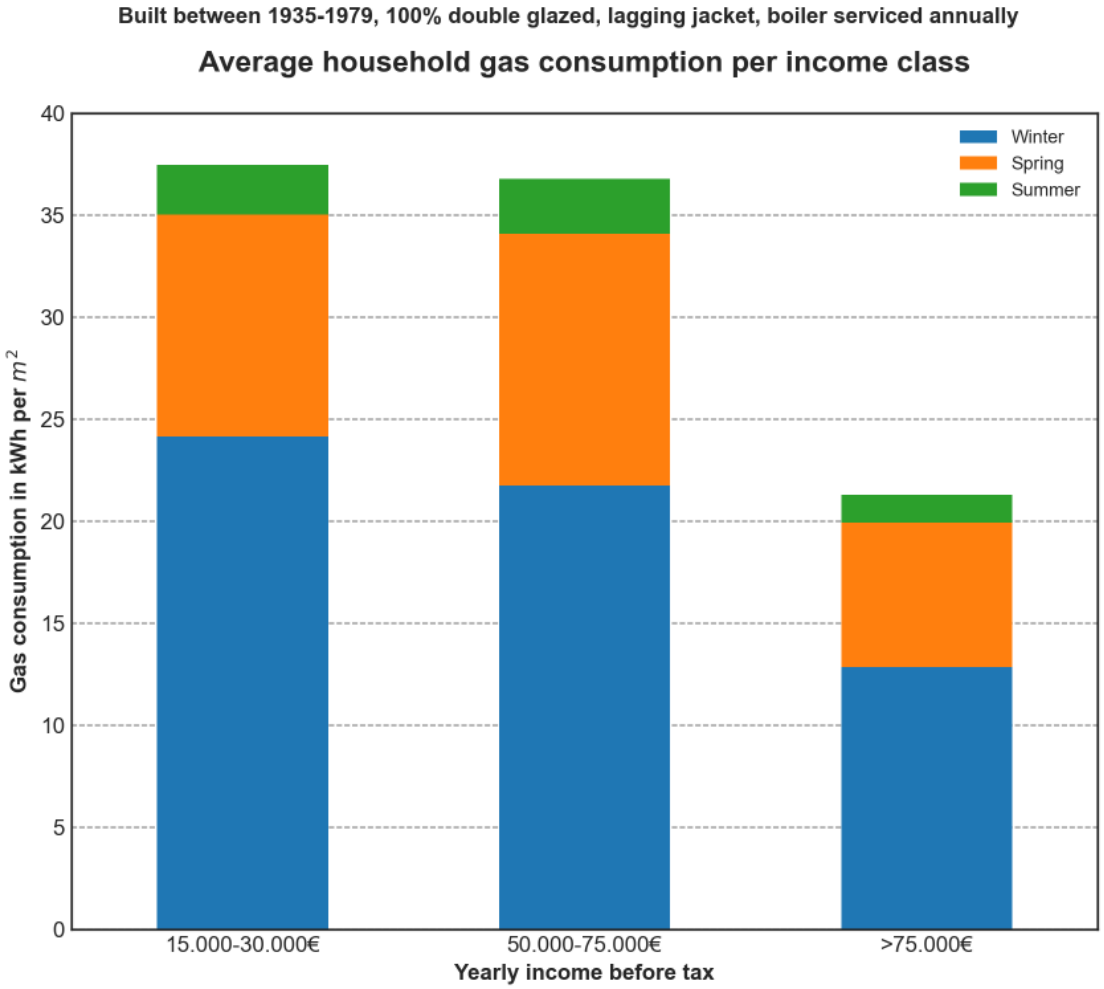



Abbildung 39: Durchschnittlicher Gasverbrauch pro Haushalt für Cluster 4

Für diesen Fall ergibt sich jedoch wieder ein völlig gegenläufiges Bild. Wie in den überwiegenden Fällen bisher fällt hier der Gasverbrauch bei steigendem Einkommen. Allerdings weist hier theoretisch lediglich die Einkommensklasse 5 einen sehr kleinen Verbrauch auf, da der Verbrauch von Einkommensklasse 4 nur minimal geringer ist als der von 2. Die saisonalen Effekte stechen auch hier direkt ins Auge.

### 5.4.3. Prüfung der Hypothese durch Interpretation der Ergebnisse

Die Hypothese, dass der Gasverbrauch eines Wohngebäudes mit steigendem Haushaltseinkommen ansteigt, lässt sich anhand der Analysen nicht eindeutig belegen. In einigen Fällen war dies zwar der Fall, in den überwiegenden Fällen und insbesondere unter Einbezug sämtlicher vorliegender

Messwerte fällt der Gasverbrauch jedoch mit steigendem Einkommen. Durch die Normierung des Gasverbrauchs mit der Nutzfläche wurden zudem mögliche Einflüsse durch unterschiedliche Gebäudegrößen ausgeschlossen. Insgesamt ist die Hypothese damit nicht nur nicht bestätigt, sondern sogar eher widerlegt, da sich stattdessen meist ein gegenläufiges Muster einstellt und die Ergebnisse suggerieren, dass ab einem Jahreseinkommen von 30.000 Euro vor Steuern der Gasverbrauch sinkt. Dennoch gibt es auch einzelne Gebäudegruppen, für die sich diese negative Korrelation nicht bestätigt, da die Verbräuche der Einkommensgruppen homogener sind oder sich diese sogar mit steigendem Einkommen klar erhöhen. Letztendlich lässt sich keine allgemein gültige Aussage treffen, die Analyse aller Werte gemeinsam suggeriert jedoch einen negativen Zusammenhang zwischen Einkommen und Gasverbrauch. Über die Tatsache, woran dies genau liegen könnte, lässt sich objektiv betrachtet nur spekulieren. Denkbar wäre jedoch schlichtweg, dass die Bewohner von Haushalten mit geringerem Einkommen einfach öfter und länger zuhause sind und daher mehr Gas verbrauchen. Diese Vermutung ist sehr einleuchtend, da ein jährliches Haushaltseinkommen von weniger als 30.000 Euro vor Steuern nicht besonders viel ist. Dementsprechend dürfte es sich bei solchen Haushalten am ehesten um Menschen im Ruhestand, Arbeitslose bzw. Sozialhilfeempfänger oder sozial schwache Familien mit mehreren Kindern handeln. Diese drei Gruppen sind per Definition die meiste Zeit zuhause. Eventuell erhalten diese Gruppen darüber hinaus auch finanzielle, staatliche Unterstützung und haben daher weniger Anreize, sich energiesparend zu verhalten.

# 6. Fazit

## 6.1. Zusammenfassung der Ergebnisse

Im Rahmen dieser Arbeit wurde untersucht, wie gut OLAP-Cubes und auf Multidimensionalität ausgelegte Datenbanken für die Integration und Analyse von Smart Meter Daten geeignet sind. Dabei wurde neben einer ausführlichen Recherche über die beiden Themengebiete Smart Meter und OLAP-Cubes ein Konzept zur Verknüpfung dieser Themengebiete entworfen und dieses in drei Beispielen anhand von realen Daten umgesetzt und demonstriert. Dabei widmete sich der erste Rechercheteil intelligenten Energieverbrauchszählern. Aufgrund der Komplexität des Elektrizitätsnetzes, der großen Bedeutung von Elektrizität für die Energieversorgung der Zukunft und den vielseitigen Potentialen von intelligenten Stromverbrauchszählern wurden dabei insbesondere Strom Smart Meter fokussiert. Für diese wurden, stellvertretend für sonstige intelligente Verbrauchszähler, die Funktionsweise, die Vor- und Nachteile, die aktuellen Rahmenbedingungen, die Schnittstellen, das Format und der Inhalt der generierten Daten sowie die bestehenden Integrationsarchitekturen untersucht. Im zweiten Rechercheteil lag dann der Fokus auf OLAP-Cubes und Data Warehousing aus dem Bereich des Business Intelligence. Dieses Themengebiet umfasst die Bereiche Datenbankarchitekturen, Integration der Daten in die Datenbank und Abfrage und Darstellung der Daten. Folglich wurden im zweiten Teil zunächst die Begriffe Business Intelligence und Data Warehousing erläutert, bevor dann verschiedene Datenbankmodelle vorgestellt und näher untersucht sowie der ETL-Prozess zur Integration der Daten und OLAP-Cubes inklusive der möglichen Operationen zur Abfrage und Visualisierung der Daten erläutert wurden.

Anschließend wurden diese beiden Teilbereiche zusammengeführt, indem ein Konzept erarbeitet wurde, dass ebendiese mehrdimensionalen Datenbankstrukturen und die Modellierung von Daten als mehrdimensionale Würfel nutzt, um von Smart Metern generierte Energieverbrauchsdaten zu integrieren und zu analysieren. Auf diese Weise können historische Energieverbrauchsdaten in ein Data Warehouse integriert und dort dauerhaft für spätere Analysen und Untersuchungen vorgehalten werden. Bestehen dann für verschiedene Anwender aus Wirtschaft, Wissenschaft oder Politik bestimmte Fragen bezüglich der Energieverbräuche von Immobilien, können die Daten aus dem Data Warehouse abgefragt und für Analysen bereitgestellt werden. Durch die multidimensionale Struktur und Modellierung können zudem weitere externe Datenquellen integriert und für etwaige Analysen hinzugezogen werden. Im Zuge dieser Arbeit wurden mehrere denkbare, externe Datenquellen vorgestellt, die für energetische Analysen relevant sein könnten. Dazu zählen beispielsweise Wetterdaten. Da es sich sowohl bei Smart Meter Daten als auch bei Wetterdaten letztendlich um Sensordaten handelt, wurde durch das Konzept und die Umsetzung zudem demonstriert, dass diese Art der mehrdimensionalen Modellierung aus dem Bereich des Business Intelligence auch große Potentiale für die Integration und Analyse von Sensordaten bietet. Dabei wurde im Rahmen der Arbeit insbesondere demonstriert, dass ebendiese Sensordaten sowohl selbst Gegenstand der Analyse sein, als auch zur Beschreibung von Daten verwendet werden können. Dementsprechend können Sensordaten sowohl Teil der Fakten, die analysiert werden sollen, als auch Teil der Dimensionen, die diese beschreiben, sein und daher sowohl in die Faktentabelle als auch in Dimensionstabellen eingefügt werden. Ferner wurden die potentiellen Nutzungsmöglichkeiten dieses Konzepts, sowie die wesentlichen Herausforderungen und Hindernisse bei dessen Realisierung und Umsetzung diskutiert.

Bezüglich der Nutzungsmöglichkeiten konnten viele Potentiale auf Mikro- und Makroebene, sowohl für den öffentlichen als auch für den privaten Sektor, konzipiert werden. Diese Potentiale stehen jedoch aktuell noch einer Reihe von Hindernissen gegenüber. Dazu zählen insbesondere der Datenschutz und Privatsphärebedenken, die geringe Marktdiffusion der Smart Meter Technologie, die Frage nach der Granularität und Messfrequenz sowie die Frage, ob die Formate und Inhalte der gemessenen Energieverbrauchsdaten verschiedener Anbieter ausreichend homogen sind, um diese automatisiert zu integrieren. Gleichzeitig wurden jedoch auch mehrere Wege aufgezeigt, diese Probleme zu lösen, zu umgehen oder einzudämmen. Der große Pluspunkt an dieser Stelle ist zudem, dass für eine Realisierung keine wirklichen technischen Einschränkungen oder Hindernisse bestehen, die nicht bereits aus bestehenden Anwendungen von OLAP-Cubes aus dem Bereich der Geschäftsanalytik bekannt wären.

Zum Abschluss wurde das Konzept dann demonstrativ umgesetzt. Dafür wurden drei verschiedene Hypothesen aufgebaut, die anhand von OLAP-Analysen geprüft wurden. Die Umsetzung erfolgte dabei händisch, unter gleichzeitiger Verwendung von Automatisierungsskripten, falls notwendig. Als Datenbanksprache wurde SQLite verwendet. Die Visualisierung der Daten erfolgte in Python. Anhand der OLAP-Analysen konnten interessante und wertvolle Erkenntnisse gewonnen, sowie Annahmen bestätigt oder widerlegt werden. Für alle Fragestellungen war eine multidimensionale Modellierung erforderlich. Die genauen Ergebnisse der Analysen wurden in Kapitel 5 ausführlich diskutiert. Insgesamt hat sich das Konzept im Zuge der demonstrativen Umsetzung als machbar erwiesen. Dabei wurden mehrere Vor- und Nachteile der Verwendung von OLAP-Cubes zur Durchführung der Analysen deutlich.

Es hat sich anhand der manuellen Umsetzung klar gezeigt, dass der ETL-Prozess bei der Umsetzung den Großteil der Zeit beansprucht. Schwierigkeiten während des ETL-Prozesses waren jedoch, zumindest im Rahmen der Analysen dieser Arbeit, meist weniger auf eine hohe Komplexität, sondern eher auf den schieren Umfang der durchzuführenden Schritte und Formatierungen zurückzuführen. Insbesondere die Aufbereitung der Daten anhand verschiedener Programme und der damit einhergehende Datentransfer zwischen diesen Programmen können zeitaufwendig sein. Dabei ist jedoch auch zu beachten, dass der ETL-Prozess in den Fällen dieser Arbeit aufgrund der verwendeten Datenquellen verhältnismäßig umständlich und komplex war. Dies hing insbesondere mit den verwendeten Umfragedaten zusammen, für die sich bereits die Extraktion aufgrund der Kodierung und der Notwendigkeit, mehrere Dateien zu verknüpfen, aufwendig gestaltete. Weiterhin hat sich insbesondere gezeigt, wie aufwendig die Verknüpfung externer Datenquellen wie den Wetterdaten mit internen Datenquellen wie den Smart Meter Messdaten sein kann, wenn unterschiedliche Granularitäten bestehen und gleichzeitig Werte in einem oder beiden Datensätzen fehlen oder sonstige Inkonsistenzen bestehen. In diesem Fall wurde dies aufgrund der komplexen, doppelten Kodierung der Time ID besonders verstärkt. Wäre hier stattdessen beispielsweise ein Unix-Zeitformat verwendet worden, wäre die Zuordnung deutlich einfacher und schneller möglich gewesen. Bezüglich der Integration und mehrdimensionalen Modellierung von Sensordaten wie Smart Meter Daten wurde gezeigt, dass sämtliche relevanten Dimensionen entweder einer physikalisch räumlichen oder zeitlichen Dimension entsprechen. Für die spätere Verknüpfung der Smart Meter Messdaten mit ihren entsprechenden Dimensionswerten ergibt sich dadurch, dass die Verknüpfung entweder anhand

räumlicher oder zeitlicher Aspekte erfolgt, sprich die Messdaten entweder anhand ihrer Zeit ID oder ihrer Meter ID verbunden werden. Eine Ausnahme stellen hier theoretisch Wetterdaten dar, da diese, bei Betrachtung eines größeren geographischen Raumes mit unterschiedlichen klimatischen Verhältnissen, sowohl von der räumlichen als auch der zeitlichen Dimension abhängen und somit anhand der Zeit ID und der Meter ID verknüpft werden müssen. Die Verknüpfung zeitbasierter Dimensionen gestaltet sich dabei einfacher als die Verknüpfung raumbasierter Funktionen. Dies ist darauf zurückzuführen, dass sämtliche Zeitdimensionen wie die genauen Zeitstempel, Monate, Jahre, Tageszeiten oder Wochentage der Zeit ID bereits inhärent sind und folglich sämtliche Zeitdimensionstabellen direkt durch bestimmte Gruppierungen oder Aggregationen anhand der vorliegenden Zeit IDs erzeugt und verknüpft werden können. Die Integration und Verknüpfung räumlicher Dimensionen wie Regionen, statistischen Daten, GIS-Daten, Gebäudetypologien, Energieeffizienzausweisen und der technischen Gebäudeausrüstung gestaltet sich jedoch etwas schwieriger. Dies liegt zum einen an der Tatsache geschuldet, dass diese Dimensionen der Meter ID nicht direkt inhärent sind, also nicht ohne weiteres aus dieser erzeugt werden können, wie dies bei der Zeit ID der Fall ist. Stattdessen müssen in solchen Fällen Metadatenbanken oder externe Datenquellen hinzugezogen werden. Diese werden dann lediglich anhand der Meter ID den entsprechenden Messwerten zugeordnet. Das zweite Problem stellt in diesem Zusammenhang die Redundanz der Dimensionswerte ebendieser Datenquellen dar. Während für die Zeit ID jeder Zeitstempel einen eigenen Dimensionswert darstellt und die Zeitdimensionstabelle dementsprechend einer Auflistung sämtlicher Zeitstempel entspricht, ist dies für die Meter ID nicht der Fall. Mehrere Gebäude können sich die gleichen Dimensionswerte teilen. Beispielsweise beinhaltet eine Stadt mehrere Gebäude oder eine bestimmte Gebäudeausstattung kommt in mehreren Gebäuden vor. Um die Daten als n-dimensionaler Würfel zu modellieren, sollten die auf der Meter ID basierenden Dimensionstabellen daher nicht eine Auflistung sämtlicher Meter IDs, sondern lediglich sämtlicher Dimensionswerte darstellen. Daher erfolgte die Integration und Verknüpfung räumlicher Dimensionen durch Hilfstabellen, die dazu dienten, die Dimensionstabellen zu erstellen und die Dimensionswerte den entsprechenden Messwerten der Faktentabelle zuzuordnen. Sobald der ETL-Prozess gänzlich abgeschlossen und das Data Warehouse aufgebaut ist, stellen alle darauf aufbauenden Abfragen und Visualisierung des weiteren Prozesses keine wesentlichen Hindernisse mehr dar.

Schlussfolgernd kann also die Eignung von OLAP-Cubes für die Integration und Analyse von Energieverbrauchsdaten bewertet werden. Prinzipiell hat die Umsetzung gezeigt, dass die Nutzung von OLAP-Cubes in diesem Zusammenhang nicht nur möglich ist, sondern auch viele Potentiale und Nutzungsmöglichkeiten bietet. Allerdings hat sich ebenfalls herausgestellt, dass Prozesse wie die Dekodierung, Vorformatierung und Extraktion von Daten aus Datenquellen wie Umfragen bereits sehr zeitaufwendig sein können. Aufgrund des aufwendigen ETL-Prozesses für eine saubere Integration und Aufbereitung der Daten kann es je nach Heterogenität und Komplexität der Fragestellung und Datenquellen daher fraglich sein, ob die Nutzung von OLAP-Cubes für Analysen tatsächlich lohnenswert ist. Liegen eher kleinere Datensätze vor, die lediglich einmalig untersucht werden sollen und auch anhand anderer Methoden analysiert werden können, könnten solche Methoden in diesen Fällen OLAP-Analysen und dem teils zeitaufwendigen Aufbau eines Data Warehouse vorzuziehen sein. Beispielsweise könnten die Daten in solchen Fällen schlicht als eine große Tabelle zusammengeführt werden, die direkt in Excel anhand von Pivot-Tabellen oder in Python anhand der Pandas Bibliothek

verarbeitet und visualisiert wird. OLAP-Cubes und multidimensionale Datenbankmodelle sind aber prinzipiell sehr gut für energetische Analysen und die Analyse von Sensordaten geeignet, da sich die relationale, mehrdimensionale Struktur durch Sternschemata und multidimensionale OLAP-Abfragen sehr gut nachbilden lässt. Das volle Potential kann OLAP jedoch erst entfalten, wenn wirklich sehr große Datenmengen vorliegen, diese automatisiert anhand von Skripten, Algorithmen oder ganzen Software-Lösungen transformiert und in das Data Warehouse integriert werden können und Hierarchien und Dimensionskombinationen eine entscheidende Rolle zukommt. In solchen Fällen dürften OLAP-Analysen aufgrund ihres Speicherplatzbedarfs, ihrer Abfrageperformance und den zahlreichen Möglichkeiten, die Daten zu modellieren, visualisieren und explorieren, allen alternativen Lösungen deutlich überlegen sein. Insbesondere die Darstellung von Hierarchieebenen und die dadurch möglichen Aggregationen stellen einen entscheidenden Vorteil dieser Technologie dar. Ferner ist ein Data Warehouse auch sehr gut für die Integration von Sensor- und Energieverbrauchsdaten, insbesondere mit dem Ziel der Langzeitspeicherung, geeignet. Kommen als zugrundliegendes Datenbankmodell Sternschemata zum Einsatz, so lassen diese sich vertikal gut und einfach skalieren und stellen einen guten Kompromiss aus Speicherplatzbedarf und Performance dar, da nicht alle denkbaren Aggregationen und Kennzahlen im Voraus berechnet werden und die de-normalisierte Speicherung der Dimensionswerte nur die minimalen Tabellenverknüpfungen erfordert.

Die Entscheidung, ob OLAP-Cubes oder sonstige Analyseverfahren und -technologien zum Einsatz kommen sollten, sollte in erster Linie davon abhängen, ob die Untersuchung der Datensätze nur einmalig erfolgen soll. Allgemein gilt, dass die Verwendung von OLAP-Cubes für Analysen zwar viele vorangehende Schritte und recht viel Arbeit im Voraus zum Aufbau eines Data Warehouse erfordert, dies jedoch später durch sehr schnell, einfach und intuitiv durchführbare Analysen und Abfragen belohnt wird. Daher sollte dann zur Analyse von Energieverbrauchsdaten ein Data Warehouse aufgebaut werden, wenn ein Interesse an der dauerhaften Speicherung und Bereitstellung ebendieser Daten für mehrdimensionale Analysen besteht und neu generierte Daten in regelmäßigen Abständen hinzugefügt werden sollen. Diese Art der Anwendung von OLAP-Cubes entspricht somit genau dem ursprünglich vorgesehenen Zweck dieser Technologie in der Geschäftsanalytik.

## 6.2. Kritische Würdigung

Das Konzept hat sich anhand der Demonstration als machbar und nützlich erwiesen. Der Gesamtprozess von dem Aufstellen einer Hypothese bis hin zur Visualisierung der Analyseergebnisse konnte erfolgreich durchgeführt werden und hat viele interessante Erkenntnisse geliefert. Allerdings bestehen auch bestimmte Kritikpunkte.

Die Transformation und Aufbereitung der Daten erfolgte teilweise außerhalb der Datenbank anhand verschiedener Programme und teilweise innerhalb der Datenbank durch SQLite. Auf diese Weise konnte für jeden speziellen Fall die am besten geeignete Software ausgewählt und genutzt werden. Da ohnehin für jede Hypothese ein gesondertes Data Warehouse aufgebaut wurde und dies manuell geschah, führte diese Herangehensweise zu keinen Problemen. Allerdings stellt sich die Frage, wie diese Prozesse automatisiert ablaufen könnten. Wird das Data Warehouse in regelmäßigen Abständen automatisiert mit neuen Messdaten gespeist, kann diese Herangehensweise problematisch sein. In diesem Fall ist es nicht empfehlenswert, dass die Transformation und Aufbereitung der neuen Messdaten innerhalb der bestehenden Datenbank mit den bereits integrierten Werten geschieht. Stattdessen sollte dieser Transformationsprozess vollständig vorverlagert werden, sodass die bestehenden Daten in der Datenbank nach dem Import nicht mehr verändert werden. So sollte beispielsweise eine neue Datenbank für die reine Durchführung des Transformationsprozesses der neuen Messdaten deklariert werden. Dies würde den gesamten ETL-Prozess außerhalb der finalen Datenbank erledigen und die fertig transformierten Daten müssten abschließend nur noch in die finale Datenbank geladen werden. So wäre eine klare Trennung zwischen Staging Area und Data Warehouse gewährleistet.

Eine weitere Frage ergibt sich bezüglich der Zeitdimensionen. Für Analysen, die viele zeitliche Dimensionen, wie beispielsweise bei Hypothese 2 der Fall, und eine sehr große Faktentabelle aufweisen, könnte stattdessen die Verwendung eines Mischschemas aus Stern- und Schneeflockenschema, also eines „Starflake“-Schemas, anstelle des reinen Sternschemas besser geeignet sein. Dieses würde die normale zeitliche Dimension mit allen Hierarchieebenen in einer einzigen Tabelle zusammenfassen, jedoch alle weiteren zeitlichen Dimensionen wie Wochentag oder Tageszeit mit der normalen zeitlichen Dimension anstatt der Faktentabelle verknüpfen und die Primärschlüssel der sonstigen zeitlichen Dimensionen als Fremdschlüssel in der Haupttabelle der Zeitdimension speichern. Auf diese Weise müssten zwar mehr Tabellen verknüpft werden und die Zeitdimensionstabelle hätte aufgrund der gespeicherten Fremdschlüssel mehr Spalten, die Faktentabelle würde dafür aber um ebendiese Spalten reduziert. Da die sonstigen zeitlichen Dimensionstabellen wie Wochentag oder Tageszeit nur sehr wenige Zeilen aufweisen, dürfte dies keinen großen Unterschied bezüglich der Antwortzeiten und Performance machen, würde gleichzeitig aber den Speicherplatzbedarf des Data Warehouse drastisch reduzieren. Dies ist darauf zurückzuführen, dass die Faktentabelle meist deutlich größer als die Zeitdimensionstabelle ist, da die Größe der Faktentabelle in der Regel gleich der Größe der Zeitdimensionstabelle multipliziert mit der Anzahl der Meter IDs ist. Aufgrund der deutlich höheren Zeilenanzahl der Faktentabelle hätte die Spaltenreduktion der Faktentabelle einen weitaus höheren Einfluss auf den Speicherplatzbedarf als die Erhöhung der Spaltenanzahl der Zeitdimensionstabelle.

Alle letzter Kritikpunkt wäre an dieser Stelle ebenfalls zu nennen, dass in der hier dargestellten Umsetzung eigentlich keine wirkliche Darstellung der Ergebnisse als n-dimensionaler Würfel stattfand. Aufgrund der Wahl eines relationalen Datenbankmodells erfolgte die Ausgabe der Abfragen als einfache Tabelle. Diese wurde bei der Umsetzung dann direkt für die Visualisierung formatiert, sodass der Schritt der Würfelbildung in gewissem Maße übergangen wurde. Da allerdings die zu prüfende Hypothese bereits von Anfang an klar war und das Explorieren der Daten nicht an erster Stelle stand, war dies in diesen Fällen kein Problem. Wäre der Prozess anhand bestehender Software durchgeführt worden, wäre die Würfelbildung automatisch erfolgt.

## 6.3. Weiterführende Untersuchungen

Weiterführende Untersuchungen könnten sich dann in erster Linie mit der Frage der Automatisierung dieses Prozesses und der Entwicklung entsprechender Software-Lösungen beschäftigen. Dabei könnten sowohl Programme zur Automatisierung des DW-Aufbaus als auch Programme zur Abfrage und Visualisierung der Daten entwickelt wären. Theoretisch wäre es auch möglich, eine allumfassende Anwendung für den Gesamtprozess zu entwickeln. Da das Konzept bereits als möglich bewiesen wurde, bestünde zudem die Frage, welches Interesse an einer solchen Anwendung bei potentiellen Nutzergruppen besteht und in welchem Umfeld eine erste Implementierung denkbar wäre. In diesem Zusammenhang müsste auch geklärt werden, wie sich ein solches Konzept am besten in die bestehenden Systeme von beispielsweise Energieunternehmen eingliedern lassen würde.

## Literaturverzeichnis


Albadi, M. H., & El-Saadany, E. F. (2008). A summary of demand response in electricity markets. *Electric power systems research, 78*(11), 1989-1996.

Antic, K. (2015). Sicherheit und Datenschutz im Smart Grid.

Arlt, D., & Wolling, J. (2011). Energiebewusstsein 2011: Ergebnisse einer repräsentativen Bevölkerungsumfrage in Thüringen zu energiebezogenen Einstellungen und Verhaltensweisen.

Bachor, M., & Freunek, M. (2020). IoT-Lösungen als Alternative zum klassischen Smart Metering. In *Realisierung Utility 4.0 Band 2* (pp. 215-226): Springer.

BMWi. (2020). Erdgasversorgung in Deutschland [Internet] Zitiert am 16.06.2020. URL: https://www.bmwi.de/Redaktion/DE/Artikel/Energie/gas-erdgasversorgung-in-deutschland.html.

BSI. (2019). Technische Richtlinie BSI TR-03109-1 - Anforderungen an die Interoperabilität der Kommunikationseinheit eines intelligenten Messsystems. *https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/TechnischeRichtlinien/TR03109/TR03109-1.pdf?__blob=publicationFile&v=3*.

Cramton, P., Ockenfels, A., & Stoft, S. (2013). Capacity market fundamentals. *Economics of Energy & Environmental Policy, 2*(2), 27-46.

Dobson, S., Golfarelli, M., Graziani, S., & Rizzi, S. (2018). A reference architecture and model for sensor data warehousing. *IEEE Sensors Journal, 18*(18), 7659-7670.

Edelmann, H., & Kästner, T. (2013). Kosten-Nutzen-Analyse für einen flächendeckenden Einsatz intelligenter Zähler. *Editorial office of the Bundesministerium für Wirtschaft und Energie: Berlin, Germany*.

Farkisch, K. (2011). *Data-Warehouse-Systeme kompakt: Aufbau, Architektur, Grundfunktionen*: Springer-Verlag.

Gleave, S. (2008). Die Marktabgrenzung in der Elektrizitätswirtschaft. *Zeitschrift für Energiewirtschaft, 32*(2), 120-126.

Gutiérrez, A. G. (2010). *Applying OLAP Pre-Aggregation Techniques to Speed Up Aggregate Query Processing in Array Databases.* IRC-Library, Information Resource Center der Jacobs University Bremen,

Herre, J., & Freunek, M. (2020). Intelligente Messsysteme–Alternativen zum Smart Meter Rollout. In *Realisierung Utility 4.0 Band 2* (pp. 195-213): Springer.

Hurst, W., Montañez, C. A. C., Shone, N., & Al-Jumeily, D. (2020). An Ensemble Detection Model Using Multinomial Classification of Stochastic Gas Smart Meter Data to Improve Wellbeing Monitoring in Smart Cities. *IEEE Access, 8*, 7877-7898.

Ifland, M., Exner, N., & Westermann, D. (2011). *Appliance of direct and indirect demand side management.* Paper presented at the IEEE 2011 EnergyTech.

Inmon, W. H., Welch, J. D., & Glassey, K. L. (1997). *Managing the data warehouse*: John Wiley & Sons, Inc.

Jahnke, B., Groffmann, H.-D., & Kruppa, S. (1996). On-Line Analytical Processing (OLAP): Entscheidungsunterstützung von Führungskräften durch mehrdimensionale Datenbanksysteme.

Kappelhoff, M. (2016). Machbarkeitsanalyse über den Aufbau eines Enterprise Data Warehouse auf Basis von Apache Hadoop.

Keelson, E., Boateng, K., & Ghansah, I. (2014). A smart retrofitted meter for developing countries. *International Journal of Computer Applications, 90*(5).

Klobasa, M. (2007). *Dynamische Simulation eines Lastmanagements und Integration von Windenergie in ein Elektrizitätsnetz auf Landesebene unter regelungstechnischen und Kostengesichtspunkten.* (Doktorarbeit). ETH Zurich,

Knab, S., & Konnertz, L. (2011). Smart Energy–branchenübergreifende Exploration eines entstehenden Marktes (Smart Energy–Cross-Industry Exploration of an Emerging Market)(German). *et-Energiewirtschaftliche Tagesfragen, 61*(5), 8-12.

Li, X. (2010). Einsatz von OLAP in der Messtechnik der T-Mobile.

Malinowski, E., & Zimanyi, E. (2008). Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications. *Verlag Berlin Heidelberg: Springer*.
Matsui, K., Ochiai, H., & Yamagata, Y. (2014). Feedback on electricity usage for home energy management: A social experiment in a local village of cold region. *Applied energy, 120*, 159-168.
McHenry, M. P. (2013). Technical and governance considerations for advanced metering infrastructure/smart meters: Technology, security, uncertainty, costs, benefits, and risks. *Energy Policy, 59*, 834-842.
Meier, A., Kaufmann, M., & Kaufmann, M. (2016). *SQL-& NoSQL-Datenbanken*: Springer.
Mikkelsen, S. A., Jacobsen, R. H., & Terkelsen, A. F. (2016). *DB&A: An Open Source Web Service for Meter Data Management.* Paper presented at the 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE).
Nagesh, D. R., Krishna, J. V., & Tulasiram, S. (2010). *A real-time architecture for smart energy management.* Paper presented at the 2010 Innovative Smart Grid Technologies (ISGT).
Niebert, K. (2014). Die nachhaltigen Zwillinge. *Bodenbender, W., Schwendy, A., Steinke, J. & Buder, N. TUP Theorie und Praxis der Sozialen Arbeit*, 102-112.
Petsch, M., Nissen, V., Termer, F., Flachsenberger, I., Schorcht, H., Warweg, O., . . . Bretschneider, P. (2012). *Der Einfluss von Smart Metern auf die Geschäftsprozesse kommunaler Energieversorger* (Vol. 2).
Pitì, A., Verticale, G., Rottondi, C., Capone, A., & Lo Schiavo, L. (2017). The role of smart meters in enabling real-time energy services for households: The Italian case. *Energies, 10*(2), 199.
Prakash, D., & Prakash, N. (2018). *Data Warehouse Requirements Engineering*: Springer.
Ramos, S., & Vale, Z. (2008). *Data mining techniques application in power distribution utilities.* Paper presented at the 2008 IEEE/PES Transmission and Distribution Conference and Exposition.
Riester, J. (2017). Energie 4.0–Die Digitalisierung der Energiewirtschaft - Eine empirische Untersuchung zur verbraucherseitigen Akzeptanz der Smart Meter Technologie und Implikationen für deren Vermarktung.
Rigoll, F. (2017). *Nutzerorientiertes Energiedatenmanagement*: KIT Scientific Publishing.
Rodriguez-Diaz, E., Palacios-García, E. J., Savaghebi, M., Vasquez, J. C., Guerrero, J. M., & Moreno-Munoz, A. (2015). *Advanced smart metering infrastructure for future smart homes.* Paper presented at the 2015 IEEE 5th International Conference on Consumer Electronics-Berlin (ICCE-Berlin).
Schlenker, U. (1998). Datenmodellierung für das Data Warehouse. Vergleich und Bewertung konzeptioneller und logischer Methoden.
Schweinfurth, H. (2020). Energiedatenmanagement–EDMS, Big Data, Smart Data. In *Realisierung Utility 4.0 Band 2* (pp. 227-239): Springer.
Sirojan, T., Lu, S., Phung, B., & Ambikairajah, E. (2019). *Embedded Edge Computing for Real-time Smart Meter Data Analytics.* Paper presented at the 2019 International Conference on Smart Energy Systems and Technologies (SEST).
Sodenkamp, M., Hopf, K., Kozlovskiy, I., & Staake, T. (2016). Smart-Meter-Datenanalyse für automatisierte Energieberatungen (" Smart Grid Data Analytics")-Schlussbericht.
Soewito, B., Isa, S. M., & Gunawan, F. E. (2018). *OLAP Analysis of Water Formation Data.* Paper presented at the 2018 International Conference on Information Management and Technology (ICIMTech).
Sun, Q., Li, H., Ma, Z., Wang, C., Campillo, J., Zhang, Q., . . . Guo, J. (2015). A comprehensive review of smart energy meters in intelligent energy networks. *IEEE Internet of Things Journal, 3*(4), 464-479.
Verbraucherzentrale, B. (2010). Erfolgsfaktoren von Smart Metering aus Verbrauchersicht. *Verbraucherzantrale Bundesverband eV*.
Wang, Y., Chen, Q., Hong, T., & Kang, C. (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid, 10*(3), 3125-3148.

wik-Consult. (2006). Potenziale der Informations- und Kommunikations-Technologien zur Optimierung der Energieversorgung und des Energieverbrauchs (eEnergy).

Wolter, D., & Reuter, E. (2005). *Preis-und Handelskonzepte in der Stromwirtschaft: von den Anfängen der Elektrizitätswirtschaft zur Einrichtung einer Strombörse*: Springer-Verlag.

Yi, G., Choi, S.-J., & Hwang, K.-i. (2014). A Light-Weight Metering File System for Sustainable Real-Time Meter Data Management. *Sustainability, 6*(9), 6351-6361.

Zeller, M. (2015). *Analyse und Simulation von Geschäftsmodellen für Elektrizitätsvertriebsunternehmen: Untersuchungen für die Implementierung von Smart Metern*: Universitätsverlag der TU Berlin.

## Abkürzungsverzeichnis

| | |
|---|---|
| AMI | Advanced Metering Infrastructure |
| BI | Business Intelligence |
| BMWi | Bundesministerium für Wirtschaft und Energie |
| BSI | Bundesamt für Sicherheit in der Informationstechnik |
| CLS | Controllable-Local-Systems |
| DBMS | Datenbankmanagementsystem |
| DOLAP | Desktop OLAP |
| DW | Data-Warehouse |
| EDM | Energiedatenmanagement |
| EDMS | Energiedatenmanagementsystem |
| EER | Energy Efficiency Rating |
| EMT | Externe Marktteilnehmer |
| EnWG | Energiewirtschaftsgesetz |
| ETL | Extract-Transform-Load |
| EVU | Elektrizitätsvertriebsunternehmen |
| GDEW | Gesetz zur Digitalisierung der Energiewende |
| GHD | Gewerbe/ Handel/ Dienstleistungen |
| GIS | Geoinformationssystem |
| HAN | Home Area Network |
| HES | Head-End-System |
| HOLAP | Hybrides OLAP |
| IoT | Internet of Things |
| KPI | Key Performance Indicator |
| KWK | Kraft-Wärme-Kopplung |
| LMN | Local Metrological Network |
| MDBMS | Multidimensionales Datenbankmanagementsystem |
| MDM | Meter-Daten-Management |
| MDMS | Meterdatenmanagementsysteme |
| MOLAP | Multidimensionales OLAP |
| MSB | Messstellenbetreiber |
| MsbG | Messstellenbetriebsgesetz |
| OLAP | Online Analytical Processing |
| OLAP-Cubes | Online Analytical Processing Cubes |
| OLTP | Online Transaction Processing |
| PV-Anlagen | Photovoltaik-Anlagen |
| RDBMS | Relationales Datenbankmanagementsystem |
| RLM | Registrierte Leistungsmessung |
| ROLAP | Relationales OLAP |
| SLP | Standardlastprofil |
| SMGW | Smart Meter Gateway |
| SQL | Structured Query Language |
| TLS | Transport-Layer-Security |
| WAN | Wide Area Network |
| ZFA | Zählerfernauslesung |

## Abbildungsverzeichnis

## Tabellenverzeichnis

**Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 und § 23 Abs. 7 APB TU Darmstadt**

Hiermit versichere ich, Samuel Achenbach, die vorliegende Bachelor-Thesis gemäß § 22 Abs. 7 APB der TU Darmstadt ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
Mir ist bekannt, dass im Falle eines Plagiats (§38 Abs.2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.
Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung gemäß § 23 Abs. 7 APB überein.

Datum: Unterschrift:

17.07.2020 S. Achenbach