

Conception of a Big Data platform in context of the energy industry

Benjamin Wagner vom Berg, Marius Brinkmann, Jorge Marx Gómez
Carl von Ossietzky University of Oldenburg
Oldenburg, Germany
benjamin.wagnervomberg@uni-oldenburg.de

In this paper, the idea of an integrated large scalable big data platform for the energy industry as a contribution to EMIS will be described. For this, use cases are identified that could bring forward a sustainable energy consumption and that faces the challenges of a fluctuating energy production. Use cases and requirements for such a big data platform have been elaborated with help of interviews with experts from science and industry. Furthermore existing approaches e.g. from German OFFIS institute were taken in consideration. To strengthen the approach, a software prototype of the data platform was built based on Hadoop and evaluated with the German software company thepeaklab. The prototype collects data of multiple smart meter and realizes the identified use cases for producer and consumer.

Big data, energy, renewables, smart grid, wind power, solar power, Hadoop.

I. INTRODUCTION

The energy sector plays a key role within every agenda for a sustainable development, because a working energy supply is mandatory for today's societies and economies worldwide. But especially the energy sector is seeking for alternatives, because it has to deal with huge environmental impacts on different levels (hazardous nuclear power, carbon dioxide - emissions by fossil burning plants) and upcoming shortages (peak oil). In this context a report of the Intergovernmental Panel on Climate Change (IPCC) proclaims that it is possible to produce 77 % of the worldwide needed energy with regenerative power plants like wind energy, solar and others [4].

The energy industry is strongly changing in technological and political ways since several years. Especially since the nuclear disaster of Fukushima the energy transition with its proclaimed goal to overcome the use of nuclear power and to switch to a major use of renewable energy supplies is a main topic on the German governmental agenda. The energy transition is connected to various problems that are mainly coming from the switch of energy supplies to more sustainable and environment friendly resources like wind and solar power that are connected to specific problems. But also a turn in consumption behavior on customer side to a more energy saving behavior is necessary.

The role of information and communication technologies in the context of sustainable development is discussed frequently.

So far, most studies in this field are dedicated to issues like using resources or energy more efficiently during production processes in industry [6] & [7], leading to the development of corporate environmental management information systems (EMIS) [13]. However, these efficiency gains may be lost by using the corresponding good or service to a greater extent, leading to rebound effects [6], [7] & [12]. Thus, for enabling sustainability by ICT, deep structural change is necessary, including radical changes in consumption patterns and fostering of a sustainable consumption behaviour [14]. Figure 1 shows a newer classification of EMIS by Wagner vom Berg [14] in extension of Marx Gómez [9] that takes not only production-integrated environment protection, but also market-integrated environment protection into account.

Environmental Management Information Systems (EMIS)								
Information- and Reporting- Systems		Eco-Controlling- Systems		Systems of Market-integrated Environment Protection		Systems of Production-integrated Environment Protection		
Govern-ment	So-ciety	KPI-Sys.	Eco-Balance -Sys.	Marketing-oriented Sys.	Consumption-oriented Sys.	Input-oriented Sys.	Process-oriented Sys.	Output-oriented Sys.
Corporate Sustainability Reporting					Environmental Information Management			

Fig. 1. Classification of EMIS [14]

The “systems of market-integrated environmental protection” are a vital complement to EMIS, because in a market-oriented economy, the demand for goods is decisive for the type and quantity of production. This perspective applies ultimately to an earlier stage than the “production-integrated environmental protection”. “Marketing-oriented systems” try to alter consumption behavior by marketing instruments and methods, e.g. by providing information according the environmental impact and benefits of sustainable products. Customers are directly involved within “consumption-oriented systems”, by providing environmental and sustainability-related information for their personal consumption as a decision basis. This information is based on their individual consumption behaviour and for this personal consumption data are necessary.

For these reasons information technology seems to be also a key factor for a successful energy transition especially by supporting intelligent energy networks, such as smart grids. Aim of these smart grids is to support different (decentralized) ener-

gy supplies and to synchronize consumption and production of energy [17]. One main task within the establishment of a working smart grid is to build a working and suitable information basis.

The difficulty in dealing with the increasing amount of data that is generated by the connections of producers, consumers and other actors of the energy industry is not only to collect the data, but to process the data by various user groups and to gain findings which add value to the energy services and to the energy system on the whole. Highly fluctuating energy production methods such as photovoltaics or wind energy are dependent on meteorological conditions [16]. The smart grid generates a plurality of data and with their help analytical information can be generated with the purpose to stabilize the energy network. In this connection a big data approach to handle the three “V’s” of data (volume, variety, velocity) seems to be very promising [18]. With the integration of a big data platform for the energy business, the proposed analytical information could be gained and processed. For this, as a first step, future use cases on the consumer and producer side have to be identified and analyzed. These use cases build the basis to identify the requirements for a big data platform that supports smart grids and a more efficient and sustainable energy system on the whole. Such a big data platform for the energy industry is considered as main contribution to EMIS, because it extends traditional EMIS in a valuable way by bringing forward a sustainable energy supply and consumption.

II. BIG DATA AND SMART GRIDS

Since the mid-nineties, the energy supply transformed by reasons of liberalisation together with deregulation of electricity markets and an increasing share in renewable energy production, such as wind and solar energy. One way to face the challenges of the energy transition is to use information and communication technology. Energy actors, such as producers, consumers, energy storage and grid components, are integrated into a data-driven information technology network [1]. Since the production plans of the traditional power plants are controlled by the demand, the use of renewable energy makes the supply challenging. The supply is no longer determined by load profiles but by natural conditions, such as wind and weather. Because wind and solar energy have the greatest potential among renewable energy generation options, they represent the essential methods in the context of renewable energy [17]. There are no facts that are based on practical experiences and that prove the amount of data obtained by smart metering in reality. This is due to the fact that smart grid components, such as photovoltaic systems or energy grids, are not actively monitored and do not produce data that is stored and analysed. For this reason, some future (fictional) scenarios are described in the following, intended to show the use of big data technology. The considerations are made with the help of Prof. Dr. Sebastian Lehnhoff who addresses the current issues of energy computer science at the OFFIS Institute for Computer Science located in Oldenburg, Germany. In OFFIS a research project is realised that aims to collect and process the meter data of the

first German offshore wind farm Alpha Ventus¹ that operates under open sea conditions. Twelve wind turbines provide 5.184 billion values per day¹. Because of this amount of collected data, six terabytes of disk space per year is needed. In the following, a fictitious calculation example represents which amount of data might be expected if the photovoltaic systems in German households got metered. In table 1, the expected amount of data can be seen that could occur in Germany on the basis of a predetermined frequency only by photovoltaic panels on the roofs of private homes.

Location / Interval	1 sec	1 min	15 min
Germany (1.5 mio.)	1036.8 GB	17.28 GB	1.15GB

Tab. 1. Example for the amount of data without communication data

A study by the German Federal Association of Solar Industry Ltd revealed that by the end of 2014 a total of 1.5 million photovoltaic systems were installed in Germany [3]. When considering these value, it should be noted that the installation number of photovoltaic systems in the context of the exponential growth of renewable energy and new buildings increases significantly. It can be seen that due to the amount of data, especially in the context of the integration of additional smart grid components, the use of big data technology is reasonable. In this context a study of Greentech Media from 2012 prognoses that the efforts for big data applications will increase dramatically in the upcoming years (see Fig. 2):

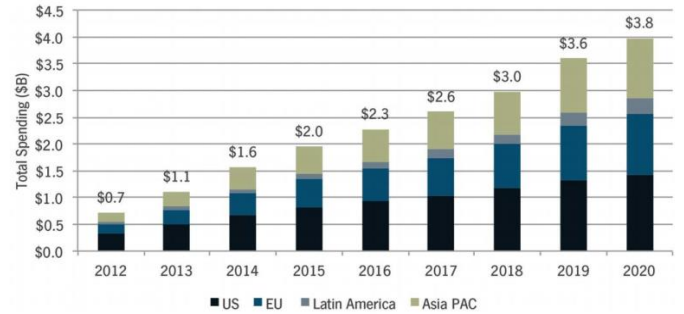


Fig. 2. Expenses worldwide for analytical smart grid applications, 2012-2020

The topic of Big Data is engaged in the collection of large amounts of data and the analysis of these in order to gain useful information. For this, a plurality of different data sources delivers the input in any form. Big Data represents the data-technical basis for decision support systems. Beside the processing of traditional data (structured and a manageable size), Big Data also focuses on unstructured and inconsistent data with a large volume. This makes it possible to transform large amounts of data faster to capture and process data and it can be the enabler to perform real-time analysis on large data sets [18].

¹ www.alpha-ventus.de

III. USE CASE DESCRIPTION

3.1 Load Forecast Automation

The first use case focuses on customers in the context of a smart home. There are smart home devices that should be controlled with the help of load forecasts. The goal is a more efficient energy consumption by means of decentralised energy generation. This approach is based on renewable energy in terms of a coordinated and intelligent use of the energy that is produced with renewable methods. An illustrating scenario is the forecasting based on weather data that results in a strong insolation. For this reason, the smart home automation system guides a washing machine or a heating blade not to turn on at the current time but in three hours. Thus it would be unnecessary to obtain external energy that would be result in higher costs and CO₂-emissions. The times in which own plants produce renewable energy, the price of electricity is low. A user could determine that a device has to be activated if the price of electricity is low. In addition, an interval could be provided in which the device has to do the work. This requires an energy management system as an administrator of the photovoltaic system and consuming devices. The decision to switch on a consumer is done with the help of weather data and forecasts based on them [3]. The ability to influence the occurrence of the energy load with certain mechanisms is associated with the topic of demand side management.

3.2 Selling Energy at Smart Markets

The second use case should handle scenarios dedicated to the supplier side. Because renewable power generators are strongly tied to meteorological circumstances, the generation of energy is heavily volatile. This can cause a temporary oversupply of energy that has been generated by the producers and could be used in different ways. An alternative is the storage of the oversupply and offer it to the grid in times of energy scarcity. The second use case has the goal to market the oversupply generated locally using a smart market. This leads to an optimised utilisation of the energy network and enables consumers to act as an energy supplier. For example, the disposal can be done by means of virtual power plants in which small distributed producers are combined into a larger power plant. This results in a better controllability. There are already some operators of virtual power plants that realise the selling of electricity from small power plants. Next Kraftwerke² in Germany is one of the largest virtual power plants in Europe and allows using smaller units for balancing the fluctuations of solar and wind power. Specifically, for energy trading it is necessary to recognise and analyse trends in pricing. For this, an information system can support private customers that do not have expertise in the energy industry to give recommendations based on data driven knowledge. Also on the control energy market it is possible to provide an oversupply of energy and to find a buyer. Since the future of energy supply will be based on renewable

energy, the purchase and sale of energy in an under- or oversupply situation is of great importance.

3.3 Benchmarking of Energy Production

To achieve a balanced energy mix, it is important to compare different energy production methods with each other. The third use case will compare the various energy production methods, for example, based on solar, wind or fossil resources. The comparison provides insights into the question of what energy is more efficient than another and allows synergies. The use case is especially important for the transfer market, which results from the energy transition. The transition market will offer fossil energy parallel to renewable energies until the renewable energies will be provide the whole amount of energy needed. There will be a rating of the energy mix which can serve as a marketing tool. The profiling for customers is made possible with the help of an assessment of the sustainability of energy supplies. For example, the consumption of renewable energy will be assessed as a sustainable usage and accordingly result in an environmentally aware profile. If a customer consumes fossil energy this will have a negative impact on his profile. This has a great importance for customers that want to use resources ecologically sensible and have a high awareness of sustainability. For companies a certification is conceivable that identifies them as a particularly sustainable actor. For this, the price and the time of use is taken into account. For example, an energy consumption in times of low sunlight lowers the performance of photovoltaic systems and therefore creates the need to consume a fossil energy extraction. Such a consumption provides full flexibility in real time, since a user can set how sustainable the current energy consumption should be. For example, this mechanism could be realised with a regulator in a smart gateway. Thus the application serves as a motivation to make use of environmentally sustainable power by the comparison of production methods and presentation as a real-time monitoring.

3.4 Main Requirements for the Data Platform

The main requirements for the data platform were created by taking the 3-V concept for big data into account and consider the results of the use case analysis. Basically, there are two functional requirements: Data submission and data query. The specialty of the software system to be created, and therefore the essential requirements, are caused by the high volume of data of many different data sources and the diversity of the data. For this reason, the 3-V were taken as a part of the requirements described in the context of the energy industry.

- Interface for the query of data as a web service
To enable services and smart grid scenarios using the database, it must be possible to query energy domain related data by means of a platform-independent web interface. Using the stored data, IT service provider, energy provider and other smart grid actors can offer

² <https://www.next-kraftwerke.de>

new services and implement the requirements of the changing conditions of the energy industry.

- **Interface for the submission of data as a web service**
The data platform will be a data sink for potentially all actors within the energy industry. For this reason, clients such as smart meters or weather services should be able to submit data over the internet using a platform-independent web service. It is important that the clients have no dependencies on the data platform or vice versa, so that a complete decoupling of the participants and data platform takes place.
- **Collection of the large volume of data from many external data sources (volume)**
The intelligent energy grid equips all devices with communication technology, so that the large volume of data is created that has to be collected. Additionally, many transactions will happen to control the devices and processes. For this, a fail-safe strategy with appropriate performance is necessary in order to receive the data of the smart grid participants and to transmit them to the data platform. This collection process is to be allocated as a part of the 3-V model category “volume”.
- **Realisation of a high data timeliness (velocity)**
In order to early detect trends in the energy network and offer highly flexible energy services, high data timeliness is important. An evidence for this are the very short intervals when sending data by through the smart grid participants. To meet the requirements of “velocity” as part of the 3-V classification, the IT infrastructure of the data platform must ensure this high data timeliness with the help of data collection methods and backup of the collected data.
- **Processing of poly structured data (variety)**
Because smart meters use different protocols and a variety of data sources exist, the data platform has to be able to process poly structured data. Since there are no uniform standards and protocol specifications yet, different data by various smart meter should be expected. As part of the 3-V classification, the “variety” is a feature that is of high importance in the context of intelligent energy grids.

IV. CONCEPTION FOR A BIG DATA APPLICATION IN SMART GRIDS

In figure 3, the overall concept of the data platform for the energy industry is presented in form of a system architecture. The concept was created by using the derivation and adaption of existing big data reference architectures, such as the Lambda architecture [1]. This approach can be implemented by any company that takes part on the energy industry and there is no enforcement to use certain technologies. Therefore, the differ-

ent elements of the architecture can be implemented with any products or supplemented by existing systems. The illustrated generic system architecture enables a save import and export of data and data types in the context of distributed systems.

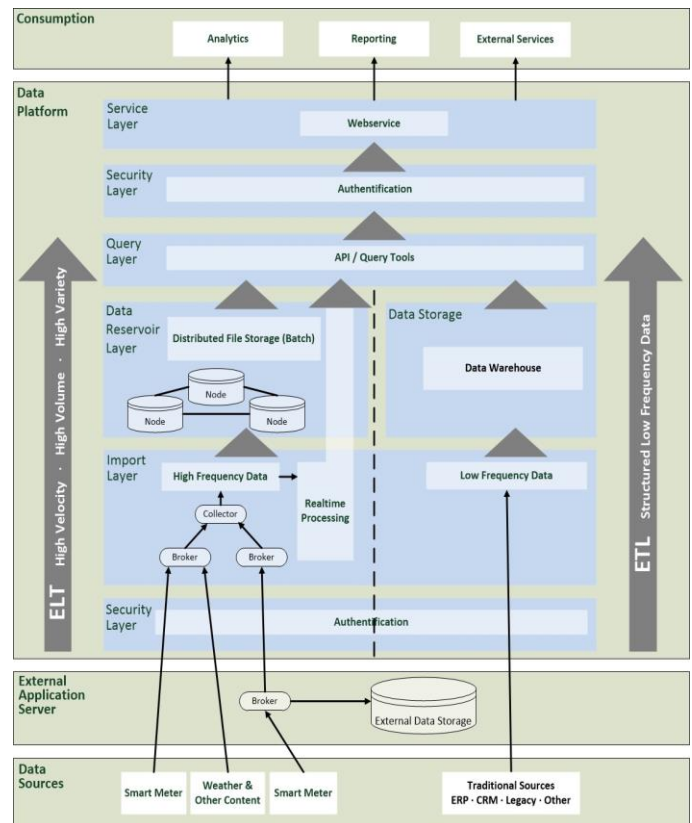


Fig. 3. System architecture of the data platform

The access will be realised by an independent layer that is easily expandable and therefore new systems can be integrated easily. The system architecture also allows the scalable and distributed processing of intensive processes with large amounts of data and storage on different clusters. As data sources, multiple systems can serve because the data platform features an extensible import layer. The customisation and extensibility of the import layer is important because there are currently no fixed standards in the domain of smart grids. The data platform aims to be extendable for future technologies.

The data platform is generally accessible by authorised users with the help of a web interface for querying and submitting data that is protected by the Security Layer. The challenge of the security layer is the use of a stateless authentication because not every communication event can be authorised due to the increased data traffic.

The Import Layer is the layer which integrates the different data as basis for decision-making with help of the data platform. The specificity of the smart grid needs to be regarded which leads to a variety of data sources in form of smart meters. After data were received using the Import Layer, the data must be stored and processed in a database. For this, the characteristic of the high volume of data is important and needs to be taken into account. In order to process the data volume, the

use of big data technology at the level of the Data Reservoir Layer is necessary. A parallel file system is recommended that utilises a shared-nothing architecture.

In an energy industry company exists a high variety of data that is used and has been generated by existing systems, such as ERP or CRM. This data is stored at the level of the Data Storage Layer. This master data, as well as low-frequency structured data as well as high-frequency unstructured data can be persisted in the data memories of the big data platform. It is rather an unrealistic scenario to assume all of the data storage of the energy company will be integrated into a central big data platform. For this reason, it is important for the data platform to provide interfaces and processes for the storage and management of traditional data in relational databases.

As part of big data there are two models of the data query that will be considered at the level of the Query Layer. The first model represents the collection of data similar to traditional batch processes. The second model allows data collection in nearly real time. Depending on the particular process of a data platform, one or both models need to be implemented. The data platform for the energy industry needs both models. To query and analyse a large amount of historical data, the dataset must be loaded with the help of data views on different data tables. For the detection of trends, the data platform needs a model with the help of real-time views.

The data platform must be able to provide data for external systems in order to realise the use cases. It is conceivable that analysis models are implemented and the results are sent in response to a client's request. However, another scenario for a provider of energy analysis is the interrogation of raw data and doing analysis on them. The result could be offered to participants of the smart grid. Therefore, the aim of the Service Layer is the abstraction of platform-internal modules and logic. For this purpose, a platform-independent solution is needed that allows a machine-to-machine interaction over a network, typically the internet.

V. PROTOTYPE

Using the prototype implementation, the evaluation of the concept is made and the realisability of the use cases is approved. In order to realise a prototype, the individual elements of the concept must be implemented with specific technologies. For the technology of distributed data storage, the Hadoop framework is selected. By being able to implement the processes of the Hadoop³ framework on any server and be stuck together with the help of a configuration, the distributed file system is justified to adapt the volume of the smart grid.

This part of the work is particularly important because of the specificity in the energy sector. The decision for the integration tool has been made for the software product Apache Flume⁴. The Service Layer was implemented as a custom software module in Java. It was implemented as a REST web ser-

vice using the Spring framework to receive the HTTP requests and respond accordingly. The Query Layer has been implemented with the module Apache Hive of the Hadoop framework. Apache Hive is a data warehouse which enables greater query datasets in distributed data stores.

The evaluation was made using a mobile app called "moon" that was developed in a cooperation of thepeaklab and university of Oldenburg. The app served as a gateway interface and has implemented various functions and use cases using the data platform. The following functions are implemented using the data platform and "moon":

- Presentation of the current energy consumption
- Presentation of historical energy consumption for a time interval
- Obtaining a flexible price based on grid capacity for a certain amount of energy
- Presentation of the current energy price based on grid factors
- Presentation of the current grid factors

A total of eight smart meter have sent data to the data platform in a laboratory environment. The smart meter have captured the various energy consumption values of individual devices and measuring sections in minute intervals.

VI. CONCLUSIONS AND OUTLOOK

In this work a conception for a big data platform for the energy industry was proposed. A first prototype was implemented in cooperation with the software company thepeaklab.

However, there are still many open questions. First of all the identified use cases has to be evaluated in a deeper way by quantitative research or by testing under real life conditions. For this the big data platform has to be implemented and used by an energy provider or a suitable company or institution.

The project "ENERA – Energie vernetzen"⁵ gives the framework for these future works. ENERA is a transdisciplinary project with many partners from science and industry under leadership of the German energy provider EWE⁶. The project is funded by the German government within the program "Schaufenster Windenergie" and will start end of 2016. The overall project volume is 260 mio. €. Thepeaklab⁷, Carl von Ossietzky university and OFFIS⁸ are partners in the project. The department "Very Large Business Applications" of Carl von Ossietzky university Oldenburg will work on Big Data applications within ENERA based on the here proposed conception and framework.

With this big data approach for the energy domain a new contribution to EMIS was done. Both in the category of systems for production-integrated environment protection and for systems of marketing-integrated protection.

³ Hadoop is a software framework for distributed storage and processing of very large data sets on multiple cluster.

⁴ Apache Flume is a distributed and reliable software product that transfers and aggregates a large amount of data.

⁵ www.enera.de

⁶ www.ewe.de

⁷ www.thepeaklab.com

⁸ www.offis.de

However, a deeper elaboration according to contribution and classification in EMIS has to be done. Suitable preliminary works where EMIS 2.0 and the association to Green Information Systems (Green IS) are described [5, 10] are already done. Further big data approaches with a link to EMIS are possible and already in work [11]. Also in different areas like the mobility domain where the Oldenburg research group “RAP-ID” [15] already developed a prototype for traffic data to support the mobility management in cities based on SAP HANA⁹.

ACKNOWLEDGMENT

This work was fundamentally supported by Sebastian Lehnhoff, Professor for energy informatics at the university of Oldenburg and OFFIS institute. Furthermore by the Oldenburg software company thepeaklab.

REFERENCES

- [1] C. Aichele, O. D. Doleski, Smart Market - Vom Smart Grid zum intelligenten Energiemarkt, Springer Vieweg, Wiesbaden, 2014.
- [2] Alpha Ventus, Alpha Ventus Technik, <http://www.alpha-ventus.de/>, 2015.
- [3] BSW-Solar, Statistische Zahlen der deutschen Solarstrombranche (Photovoltaik), http://www.solarwirtschaft.de/fileadmin/media/pdf/2015_4_BSW_Solar_Faktenblatt_Photovoltaik.pdf, 2015.
- [4] O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, C. v. Stechow, Eds, Summary for Policy Makers, in IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2011.
- [5] M. Gräuler, F. Teuteberg, T. Mahmoud, J. Marx Gómez, Anforderungspriorisierung und Designempfehlungen für Betriebliche Umweltinformationssysteme der nächsten Generation – Ergebnisse einer explorativen Studie, in D. C. Mattfeld, S. Robra-Bissantz, Eds, MKWI 2012, 2012.
- [6] L. M. Hilty, Information and Communication Technologies for a more Sustainable World, in Information and Communication Technologies, Society and Human Beings: Theory and Framework, D. Haftor and A. Mirijamdotter, Eds, IGI Global, Hershey, New York, 2010, pp. 410-418.
- [7] L. M. Hilty, W. Lohmann, E. Huang, Sustainability and ICT – An overview of the field. *notizie di POLITEIA* 27, 104, 13-28, 2011.
- [8] Miele, Intelligent verbunden. <http://www.miele.de/de/m/intelligent-verbunden-1955.htm>, 2012.
- [9] J. Marx Gómez, Betriebliches Umweltinformationssystem. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/informations_systeme/Sektorspezifische-Anwendungssysteme/Umwelt_informationssystem%2C-betriebliches/index.html/?searchterm=buis, 2008.
- [10] J. Marx Gómez, B. Rapp, A. Solsbach, T. Mahmoud, A. Memari, J. Bremer, Projekt IT-for-Green: Umwelt-, Energie- und Ressourcenmanagement mit BUIS der nächsten Generation, in: *Ökonomikus, Fachschaftszeitschrift Wirtschafts- und Rechtswissenschaften (WiRe)*. Sommerausgabe, 2011, pp. 18-20.
- [11] J. Marx Gómez, Big Data: Fundamentals, Methods and Technologies; Idea Group Inc., Hershey, expected for Dezember 2016.
- [12] N. Paech, Nachhaltigkeit als marktliche und kulturelle Herausforderung, in *Nachhaltige Zukunftsmärkte - Orientierungen für unternehmerische Innovationsprozesse im 21. Jahrhundert*, K. Fichter, N. Paech and R. Pfriem, Eds. Metropolis-Verlag, Marburg, 2005, pp. 57-94.
- [13] C. Rautenstrauch, Betriebliche Umweltinformationssysteme. Grundlagen, Konzepte und Systeme, Springer, Berlin Heidelberg, Science, 1989.
- [14] B. Wagner vom Berg, Konzeption eines Sustainability Customer Relationship Management (SusCRM) für Anbieter nachhaltiger Mobilität, Shaker Verlag, Berlin, 2015.
- [15] A. Sandau, J. Halberstadt, J. Marx Gómez, D. Stamer, B. Wagner vom Berg, B., Model of mobility demands for future short distance public transport systems, *Proceedings Conf-IRM*, Capetown, 2016.
- [16] H.-G. Servatius, U. Schneidewind, D. Rohlfing, Smart Energy. Wandel zu einem nachhaltigen Energiesystem, Springer-Verlag Berlin Heidelberg, 2012.
- [17] C. Kunz, A. Müller, D. Saßning, Smart Grids für die Stromversorgung der Zukunft, Optimale Verknüpfung von Stromerzeugern, -speichern und -verbrauchern, Agentur für Erneuerbare Energien e.V, Berlin, 2012.
- [18] P. Baron, Big Data für IT-Entscheider: Riesige Datenmengen und moderne Technologien gewinnbringend nutzen, Carl Hanser Verlag, München, 2013.

⁹ SAP HANA is a in-memory database solution that is offered as an appliance of hard- and software by SAP.