

# The Smart Grid's Data Generating Potentials

Marco Aiello

Johann Bernoulli Institute for  
Mathematics and Computer Science  
University of Groningen  
Groningen, The Netherlands  
Email: m.aiello@rug.nl

Giuliano Andrea Pagani

Johann Bernoulli Institute for  
Mathematics and Computer Science  
University of Groningen  
Groningen, The Netherlands  
Email: g.a.pagani@rug.nl

**Abstract**—The Smart Grid is the vision underlying the evolution the power grid is currently undergoing. Its pillars are increased efficiency, self-healing, operation automation, and renewable energy integration obtained through real-time control and digitalization of the infrastructure. Thus, an important ingredient—if not the main one—is information technology support for power transmission and distribution. Given the size of the power grid, its pervasiveness, and the need for its availability, it is easy to imagine that any serious ICT infrastructure dealing with it will have to manage a great deal of rapidly forming data. Now the question is whether the amount, diversity, and uses of such data put the smart grid in the category of Big Data applications, followed by the natural question of what is the value of such data. To provide an initial answer to this question, we analyze the current state of data generation of the Dutch grid, its evolution towards a smart grid, and a future realistic scenario. The scenario considered shows that the amount of data generated is comparable to some of today's social media and “classic” Big Data examples.

**Index Terms**—smart grid; power systems; Big Data

## I. INTRODUCTION

THE EVOLUTION of the power grid towards a smart grid is based on a massive deployment of Information and Communication Technology (ICT) in sensing, analyzing and controlling the operations of the power grid, from generation to utilization. This shift to a more information technology-based power grid will require considerable amounts of data to be produced by the sensing equipment, and by the new generation of (smart) meters. Thus, setting a challenge for the current ICT architectures of utilities and electricity distribution companies. If this trend is underway now, it is especially because the diffusion of renewable generation sources at all levels of the power grid calls for timely and precise monitoring of the infrastructure. In addition metering equipment is more affordable and reliable.

If the shift towards a smart digitalized grid is broadly accepted, little is known about the actual data generation potential of the future grid and on the management and utilization of large electricity data streams. The few works in the literature provide only a qualitative analysis of the amount of data that the smart grid is likely to generate with almost no results of quantitative experiences and field tests. To fill such knowledge gap, with the current treatment, we assess the amenability of the Big Data definition to the smart grid considering mainly the volume and velocity features of

smart grid data. We perform such assessment by comparing the amount of data produced and transmitted in today's prominent Big Data examples, coming from the areas of social media and Internet based services. For the comparison, we use information coming from the Dutch power grid. The Netherlands has one of the infrastructures with one the highest availability in Europe with the electrical system being in 2012 99,99486% of the time available [1]. Moreover, The Netherlands plans to have a full rollout of smart meters by 2018 and a standard for the information to be read and exchanged by smart meters is currently being finalized.

The rest of the paper is organized as follows: Section II briefly defines the Big Data concept and offers some examples of today's sources of data. Section III provides a description of the concept of smart grid, while Section IV provides a quantitative analysis of the amount of data generated by a smart grid infrastructure with special focus on the Dutch grid as concrete example. A survey of related work is presented in Section V and the concluding remarks complete the paper, Section VI.

## II. BIG DATA EXAMPLES

Big Data refers to information systems characterized by having to manage high volumes of data, which is rapidly created and that potentially has added value. It is common practice to refer to the 5Vs when talking about Big Data. The 5Vs stand for *Volume*, *Velocity*, *Variety*, *Value* and *Veracity*: big quantity of information that moves fast on a network; data that are diverse and provide relevant facts (implicit) in the information they carry, while being reliable as data sources.

**Communication capacity:** The telecom sector has seen an exponential growth in its infrastructure. The Internet alone in the short time span of 7 years (2000-2007) has increased network utilization by 29 times, with the necessity of transmitting 65 exabytes per year (optimally compressed) [2]. Cisco computed a monthly IP traffic of 43 exabyte for 2012 and lead to an almost 30% increase rate till 2016 reaching 110 exabytes/month and 1.3 zettabytes/year [3].

Prominent examples of Big Data today have to do with social media, where billions of users interact by exchanging data in various forms. Let us consider YouTube and Facebook as representative examples of the trend.

*YouTube:* YouTube is an on-line user-based broadcasting service where every user can watch and publish videos. Some key-facts numbers from YouTube's statistics<sup>1</sup>: more than 1 billion unique users visit YouTube monthly, over 4 billion hours of video are watched each month on YouTube, and 72 hours of video are uploaded to YouTube every minute.

*Facebook:* Facebook is the best-known social platform having recently passed the 1 billion users mark. The statistics concerning the amount of data that Facebook deals with are quite impressive: more than 500 terabytes of new data every day, 300 millions photo uploads, 2.7 billions 'Likes' each day<sup>2</sup>. It is then no wonder that with all these data Facebook requires an Hadoop Distributed File System cluster with more than 100 petabytes of physical space.<sup>3</sup>

Many more examples exist. E.g., computational science projects such as the data obtained from the Large Hadron Collider experiments; the astronomic images collected by (radio) telescopes; data generated on stock markets and used, for instance, for algorithmic trading; private businesses having 1 million transactions per hour, such as, Walmart, or the almost 500 transactions per second managed during peaks at Amazon.

### III. THE SMART GRID

The term smart grid does not yet have a unique definition [4], rather the various stakeholders and scientific disciplines involved have their own point of view on the area. From a physical and technical perspective, the system has always to satisfy the equilibrium between energy supply and demand in order to keep the correct operations and safety of the system. The information flow related to the operations of the grid is the real innovation of the Smart Grid. The grid will become more and more digital with information recorded by the sensors and the digital meters deployed at users' premises and along the grid and power stations. The benefit is to have enriched information of the performance of the system, its stability, and customer consumption. Another important motivation that drives the modernization of the grid lies in the ability to accommodate more renewable sources [5]. With more and more unpredictable (renewable) power sources, the electrical system needs more flexibility in managing the demand and supply equilibrium. One of the mechanisms to achieve such flexibility is through the use of variable electricity tariffs. These tariffs vary even several times per day and are transmitted to the users that are able to react to them by increasing or decreasing their consumption. For example, on a very windy day the wind energy production surges and in order to keep the balance of the grid, tariffs are lowered to incentivize the use of electricity. Naturally, to have such real-time flexibility, information has to be exchanged with the users both in terms of dynamic tariffs and energy used at a given price to enable the accounting.

An essential component in this scenario is the smart meter. It has the same primary duty of the traditional analog meter, but

<sup>1</sup><http://www.youtube.com/yt/press/statistics.html>

<sup>2</sup><http://goo.gl/2wSzw>

<sup>3</sup><http://goo.gl/lc0x9>

TABLE I: 2012 Dutch smart grid data.

Metering	
Metered customers	7,827,350
Installed smart meters	450,000
Smart meter sampling period (min)	86,400 (by law 2 months)
Smart Devices	
Electric vehicles	6,275
Battery packs	N/A
Intelligent appliances per household	0
Grid Infrastructure	
Nodes HV (380/220kV)	36
Nodes MV/LV	155,000

in addition it has a memory to store information, a processor, a digital clock, it can use multiple tariffs, it is able to measure bidirectional power flows (i.e., consumed and produced energy), and it is connected with a telecommunication link to the utility. All these features let us talk today of Advanced Metering Infrastructure (AMI).

### IV. THE SMART GRID AND BIG DATA

Taking The Netherlands as our case study, we consider the current and a possible future situation for the power grid, specifically considering the amount of generated data. We have chosen The Netherlands because the electrical system is very reliable, with just an average of 27 minutes downtime per year per customer [1], because it has a modern infrastructure with an unbundled energy market, and because it is going fast towards a digitalization of the electrical system.

#### A. Volume and Velocity

Volume and velocity are the most two prominent aspects of the transition towards a smart grid.

*Metering and Smart Buildings:* In the last years in The Netherlands, a compulsory smart meter roll-out had been planned and then canceled due to privacy concerns [6]. Thus, currently smart meters are only partially installed in the country and the approach is first to test the AMI in controlled settings and in pilots before proceeding to a massive roll out, expected to cover the country by 2018.

In our quest for Big Data, we consider the current situation of advancement of the AMI and smart grid in The Netherlands and we compare it to two realistic scenarios of the near future. The current situation is synthesized in Table I. To date, in The Netherlands only about 5% of the meter installations are smart. Furthermore, the data measuring capabilities of the meter is only partially used, since current law mandates that they can only be read once every two months. A number of *smart devices* might also interact and exchange information with the grid about tariffs or the power use. Most notably, electric vehicles, batteries in residential premises and intelligent home appliances contribute to the picture [7]. Weather information represents another essential source in the future smart grid. Knowing the weather in advance allows an home energy management system to forecast how much energy will be produced by user's solar panels or small wind turbines, and the internal needs for heating and cooling.

TABLE II: Near future scenario (e.g., one decade).

Metering	
Metered customers	8,000,000
Installed smart meter	8,000,000
Smart meter sampling period (min)	15
Smart Devices	
Electric vehicles	790,000
Battery packs	45,000
Intelligent appliances per household	10
Grid Infrastructure	
Nodes HV (380/220kV)	45
Nodes MV/LV	158,419

TABLE III: Far future scenario (e.g., four/five decades).

Metering	
Metered customers	9,000,000
Installed smart meter	9,000,000
Smart meter sampling period (min)	5
Smart Devices	
Electric vehicles	3,950,000
Battery packs	135,000
Intelligent appliances per household	20
Grid Infrastructure	
Nodes HV (380/220kV)	60
Nodes MV/LV	178,221

Table II shows a short-medium term scenario for the same objects, when the roll-out of the smart meters will be completed. The 100% penetration is based on the commitment of the European Union to deploy smart meters to at least 80% of all customers (c.f. EU directives 2009/72/EC and 2009/73/EC) and the commitment of the Dutch government. As a sampling period we consider 15 minutes. This is the time interval used in many of the AMI installations and tests (e.g., CenterPoint AMI in Texas.<sup>4</sup> We envision most of the advancements in the adoption of smart devices with almost 800,000 electric vehicles (equal to 10% of the passenger vehicles in 2012), 45,000 battery packs for local energy storage (based on the figures of solar capacity installed till 2011) and the penetration of home intelligent appliances of 10 per metering point (as about half of the average number of appliance per family).

Table III provides a long term vision, say few decades. We consider that the fully digital smart meter infrastructure reaches 9 millions customers throughout the whole country. As sampling period we consider an infrastructure that is closer to real-time measurement with 5 minutes interval. We suppose that smart appliances become the norm, therefore with 20 smart devices on average in each home or office leading to a total of 180 millions devices. The number of electric passenger vehicles reaches 3.95 millions (equal to 50% of the passenger vehicles in 2012), while the battery packs increase threefold. This last assumption is based on the estimate of the increase in distributed generation capacity by three times compared to the 2011 figures.

*The Power Grid:* Complementary to the users and metering of the electricity, there is the transmission and distribution infrastructure. In the bottom of Tables I, II, and III we report

TABLE IV: Data size for various parameters/devices/services in bytes per sample.

Data Use in Metering	Size
Consumption only	193
Consumption/production	245
Consumption/production, instantaneous power (3-phase), and current	530
Consumption/production, instantaneous power (3-phase), current, failures, and gas metering	1,100
Data Use in Smart Devices	Size
Electric vehicle consumption only	193
Electric vehicle consumption/feed-in	245
Electric vehicle consumption/feed-in, instantaneous power (3-phase), and current	530
Electric vehicle consumption/feed-in, instantaneous power (3-phase), current, failures, and gas metering	1,100
Battery	200
Intelligent appliances	200
Data Use in Weather Forecast	Size
Essential weather parameter	13,000
Improved weather parameter	15,000
Advanced weather parameter	20,000

on the current size of the Dutch power grid, size that we do not expect to increase dramatically in the medium-long period. We consider the nodes (e.g., power and transformation station) since these contain the equipment and the sensors that monitor the status of the power assets and the lines. Considering the high voltage, the current number of nodes is provided in the work of Rosas-Casals and Corominas-Murtra [8], whereas the information concerning the medium and low voltage nodes is provided by the various distribution utilities of the Netherlands.<sup>5</sup> For the high voltage grid, we consider a moderate evolution based on the public plans of the transmission operator (TenneT<sup>6</sup>). In considering the evolution, we make the conservative assumption that the ratio between the medium-low voltage nodes and the metered customers is constant and equal to the current ratio. That means that the total number of medium-low voltage nodes is 158,419 and 178,221 in the short term and long term future scenario, respectively.

*Data Size:* The Dutch Smart Meter requirement document [9] defines the size and function of the information sent. Based on this information we consider the user generated data: data from self generation, failure statistics, additional energy metering (e.g., 3-phase installation, gas metering) as shown on the top of Table IV. In the current situation, the amount of data produced by the Dutch smart grid is extremely limited as shown in Figure 1. The most parsimonious case entails just a yearly amount of about 500 megabytes, while measuring the whole set of parameters requires almost 3 gigabytes.

Things change when considering future scenarios where the diffusion of smart meters, electric vehicles and appliances will be pervasive. As shown in the central part of Table IV, we consider four categories of data generation related to the

<sup>4</sup><http://goo.gl/MzGPX>

<sup>5</sup><http://www.energieleveranciers.nl/netbeheerders/overzicht-netbeheerders>

<sup>6</sup><http://www.tennet.eu>

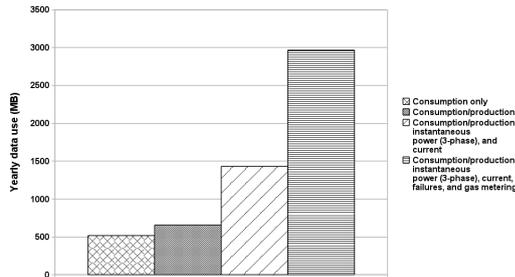


Fig. 1: Current Dutch AMI yearly data generation.

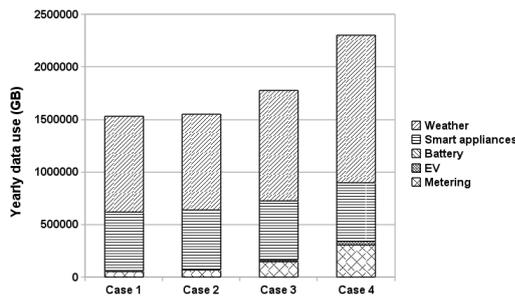


Fig. 2: Near future Dutch smart grid yearly data generation.

electric vehicles. In essence, an electric vehicle is not much different from a house metering point. Another similarity is in the possibility of feeding energy to the grid when required, thus acting as a production unit. In this future scenario, we also consider the interaction with meteorological services that provide weather information and forecast which influences users energy behavior, their energy production units, and the schedule of their appliances. The last three lines of Table IV show the amount of data required by a weather service. The National Oceanic and Atmospheric Administration (NOAA) provides eXtensible Markup Language based information regarding temperature, wind speed, and cloud coverage conditions up to 1 week forecast. The ‘improved’ and ‘advanced’ weather service data are considered for improved weather services with meteorological information with fine grain timescale. This granularity of information is essential for reliable forecasting of local production of energy, and therefore needed in the smart grid.

For the big data estimation, we identify four cases. Each case has increasing data quantities as more units are involved. The generated data can come from:

- *Case 1:* Consumption data from smart metering and electric vehicles; data from batteries and intelligent appliances; essential weather data.
- *Case 2:* Data for consumption/feed-in from smart metering and electric vehicles; data from batteries and intelligent appliances; essential weather data.
- *Case 3:* Data for consumption/feed-in, instantaneous power, and current from smart metering and electric vehicles; data from batteries and intelligent appliances;

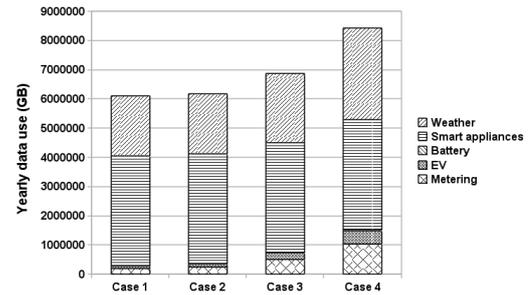


Fig. 3: Far future Dutch smart grid yearly data generation.

rich weather data.

- *Case 4:* Data for consumption/feed-in, instantaneous power, failures from smart metering and electric vehicles; data from gas metering; data from batteries and intelligent appliances; rich weather data.

Considering the near future scenario for full smart grid implementation the amount of data produced yearly for the whole of The Netherlands starts to assume the connotation of “Big Data” (Figure 2). We remark that the sampling frequency is 15 minutes for metering whereas the weather information are provided every hour. In the most conservative scenario (Case 1) the data reaches the amount of 1.5 petabytes. Case 2 is similar in data generation having a total amount again close to 1.5 petabytes. Case 3 is close to 1.8 petabytes due to the increased information recorded and more comprehensive weather information. In Case 4 we note that more than half of the data generated comes from the meteorological information. The actual metering accounts for about 300 terabytes, while the electric vehicles are about 10% of this figure. The high number of smart appliances considered in a full fledged implementation of the smart grid, makes them responsible of the generation of more than half a petabyte of data.

Considering a future scenario with even higher penetration of smart devices and higher sampling rates, the amount of data naturally grows (Figure 3). We assume a sampling period of 5 minutes for metering-related data, and 30 minutes for weather. In the first case (most conservative), the data reaches the 6 petabytes value. Case 2 is almost as data rich as Case 1, whereas Case 3 reaches almost 7 petabytes. In all the cases the source that causes the most of data are smart appliances responsible of a little less than 4 petabytes of data. In Case 4 one reaches the level of about 8.5 petabytes, distributed between smart appliances (almost 3.8 petabytes), weather information (about 3 petabytes), and smart meters (1 petabyte).

Before concluding the analysis of the volume and, to some extent, of the velocity of the data that can come from the future smart grid users, we analyze the data generating potentials of the power grid infrastructure itself. Table V provides the values for the high voltage and the low voltage grid for the current situation and for the near and far future scenarios. For each of the three temporal variants, we consider the percentage of

TABLE V: Data for grid monitoring.

Type of Power Station	% of Station Monitored	Sampling period (min)	Data Size (byte)
Current Scenario			
HV	100	15	12,950
MV/LV	10	1,440	40
Near Future Scenario			
HV	100	5	12,950
MV/LV	50	60	40
Far Future Scenario			
HV	100	1	1,295,000
MV/LV	100	5	4,000

monitored infrastructure, the sampling period and the size of the data.

Today, with very few exceptions, the medium/low voltage stations have manual switching device operations and therefore no data or remote monitoring is in place. The sampling period will shorten in the future and we assume to go from days to 5 minutes. Regarding the data size, we consider the data involved in the query and response of a distribution relay which is 40 bytes following the Modbus standard<sup>7</sup> that is a typical standard for electrical equipment monitor. We assume that in the far future scenario the data will be increased to 4,000 bytes per sample. The assumption is based on the presence of more sensing equipment and richer protocols. Considering the high voltage stations, they are already fully monitored; we consider only a decrease in the period of monitoring from 15 to 5 and 1 minute in the three scenarios. The amount of data required is based on the study of Sanchez *et al.* on the implementation of telecontrol functions for electrical stations using IP technologies, by considering just one picture frame per monitoring function [10]. Even here, we keep the current and the near future value constant and increase 100 folds the situation for the far future since we assume more equipment and richer protocols.

Based on these projections, we report the expectations for how much data is generated by the Dutch power grid in Figure 4. One notices an important growth for the high voltage related data, though what is remarkable is the important appearance of the distribution grid (medium and low voltage) data. In fact, this is where the smart grid concept will mostly change the current status quo.

### B. Variety

Several standardization bodies and international organizations are working to provide new standardization documents to guarantee the interoperability between electric and electronic equipment in the context of the (smart) grid. To date IEEE alone has more than 100 standardization initiatives. IEC is going in the same direction with more than 100 relevant standards involving the smart grid in its various domains. These standards are not all related to data transmission and information-oriented aspects, since there are also aspects inherited by the electric aspects. In general, the application

<sup>7</sup>[www.modbus.org/](http://www.modbus.org/)

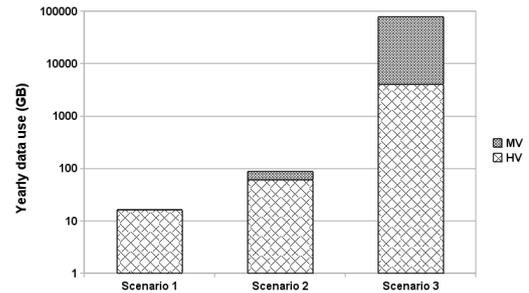


Fig. 4: Dutch power grid data generation.

domains can be divided into different portions where standards and information exchanges can be considered. In the IEC representation are defined: the IEC-61850 that describes the communication networks and systems inside a substation; the IEC-61968 describes the information exchange in the electricity distribution system; the IEC-61970 describes the common information model (CIM) used in the control center to inter-operate with the components and equipments outside the control center. Security aspects of the various communications are addressed in the IEC-62351.

In addition to technical aspects, it is important to consider the data exchange factor related to the market. The efforts of two standardization bodies (IEC and NIST) and State Grid Corporation of China are compared in [11] where similarities and differences in the approach are compared and a probabilistic model on the amount of development is provided. IEC-62325 deals with the market aspects putting the accent on deregulation. The market efforts are also addressed by the OASIS technical committee for the smart grid. Both the OASIS Energy Inter-operation specifications and the OASIS Market Information Exchange provide a framework for standardization where energy interactions are considered as transactions, i.e., energy is an operation comprising a price, a quantity and a date of delivery that can be negotiated on the market.

The home automation setting [12] is also a rich one in terms of data variety. There are many technologies and protocols for home energy saving and many initiatives are underway in relation to the smart grid. Zigbee Alliance Smart Energy Profile provides the protocol stack (based on the IP protocol) to have smart grid application at home such as meter reading, and appliance automation in response to changing energy prices. In order to enable a true interoperability between so many different and heterogeneous services, the tendency is that of using solutions that are based on services available through the Internet [13]. In summary, many standards are appearing with different data specifications and formats.

### C. Veracity

In the broad panorama of the smart grid, the problem of having accurate and meaningful data is essential. Having so many different sources and levels of aggregation requires a fine grained filtering to ensure that data are consistent and

that no compromised or mistaken data are present. In addition to the traditional mechanism of error correction coding [14], one of the issues that are highly debated about the smart grid is the security of a digital power grid infrastructure. There are several possible attacks, ranging from the acquisition of private user data, to tampering the energy bills, to modifying the data recordings of the equipment of the grid, to malware injection into general purpose and embedded computers, modifications of the demand-response mechanism creating imbalances in the production and consumption and therefore compromising the stability of the grid [15]. Therefore accuracy, integrity, confidentiality, and authenticity have to be all satisfied to guarantee the safe operations of the smart grid. The cyber-security aspects of the future power grid are a source of concern. Solutions to the security issues are under investigation for instance by NIST that has established a Cyber Security Working Group on the smart grid topic. In [16] the proposed solution is based on technologies already applied to ICT domains such as public key infrastructure for providing a secure way of authentication and authorization. Encryption based on AES or 3DES algorithms can provide a solution to guarantee privacy. In order to guarantee the trust in a system made of many different components (several thousands to millions) with different technologies and several manufacturers, the solution proposed by Metke and Ekl [16] uses trust anchor security i.e., a series of chains of certification authorities and at the end of the chain a digitally signed certificate for a piece of equipment or device in the smart grid is provided. Among the different security concerns of the smart grid analyzed in [17] some are definitely related to the data: security of devices and their communication interactions, and privacy of the users whose energy consumption is digitally metered.

#### D. Value

It is difficult to give a value to information regarding an infrastructure not yet in place and which is managed as a monopoly in most countries, or at most as a slightly unbundled market. However, we can estimate the value of the smart grid data considering some of the actors involved and the benefits they can achieve with an accessible and rich digital infrastructure.

The end user can obtain value by energy saving and appropriate planning of its energy use. This can be done only with an automation system in place and would be even more profitable with a dynamic tariff system, as shown in [7]. The value can be even greater if a totally open energy market existed where end-users could trade their excess and stored energy freely with any other partner, a scenario we depicted in [18].

The value for the energy utilities will be manifold. First, the AMI provides accurate measure of power consumed thus a more accurate billing for the end user. Second, the AMI and the data flow of a smart meter helps to discover and fight electricity thefts and have better revenues. Worldwide estimations of electricity thefts account for 25 billion dollars [19]. Third, energy utilities with smart grid data are able to early diagnose problems (especially at distribution level) in

the network and provide a fast solution, which results in less customer interruptions and therefore less penalties for service interruption. Fourth, and most importantly, precise knowledge of energy flows allows utilities to better plan the generation of energy and save money, since energy production costs are not a linear function of the amount, but rather a convex one. Fifth, the inclusion of renewables requires a better prediction of the distributed generation to keep the demand-response balance.

With the smart grid concept, there is also the possibility that new actors can find value in the power grid (Big) data. New companies are emerging (e.g., Opower) that work for the utilities in smart meters data analytics, to motivate the customer in energy conservation, dynamic tariffs usage, and energy efficiency through automation. Furthermore, several benefits are achievable through smart appliances. In fact, they enable totally new scenarios of use of the home environment. The home appliances of the next generation will receive a configuration by the user who sets the preferences of the usage time and tariffs limits to be used [20]. These appliances will interact with the home energy management system, with the utilities that might directly control them, and with the manufacturer that receives logs about the appliance life conditions. Electric vehicles will also interact with the utilities, to provide a balancing capacity for the network, therefore they will participate in the demand response functionality receiving energy tariff information, remotely provide the battery status to the user and the log of anomalies to the vehicle manufacturer. Utilities and energy services companies can take advantage of energy analytics to develop new tailor made products. The approach will be similar to the personalized offers that Amazon or other on-line retailers offer to the customers having a more detailed set of information concerning the products that the customer likes or visits frequently. Utilities will provide energy offers based on the energy consumption patterns of customers, on the appliance usage and charging time of the electric vehicle.

#### V. RELATED WORK

The smart grid is a new industrial and research theme open to exploration. The impact of data production and management is still under investigation. The few related works on the topic consider the amount of data generated to be huge, but very few provide actual numbers not to mention field experiences.

The direction towards an electricity sector that is more complex is synthesized by Rusitschka *et al.* [21]. The idea is to exploit cloud and a service-oriented approaches for processing power grid data. The paper does not provide an estimation of the amount of data that the smart grid is going to generate. The authors take for granted that the amount will be big: "The smart grid will be the largest increase in data any energy company has ever seen." [21].

Parikh *et al.* [22] investigate the appropriate wireless technology to be used for the smart grid. The various solution proposed range between WiMAX and cellular for long distance communication to the wireless LAN, ZigBee and Bluetooth to more short distance applications. The review is interesting,

though the paper only considers the theoretical bandwidth of the technologies without considering the data size and sampling requirement of the smart grid.

The important problem of privacy in metering is addressed in [23] where the authors propose an anonymization mechanism for smart metering data. The approach considers two types of metering services a *high frequency* and a *low frequency*. The solution proposed appears effective and able to actually solve the main privacy concerns. The authors only refer to high frequency of sampling without giving any quantitative aspects on the amount of data. Another interesting aspect to be investigated is the amount of overhead in data size necessary to guarantee privacy.

The authors of [24] propose a new communication infrastructure to deal with the enhanced amount of data that will be generated by the smart grid. The infrastructure is a distributed one. The authors explicitly state that the new smart grid infrastructure will generate a significant amount of data given the increased sampling frequency of the grid sensing equipment (e.g., phasor measurement units) and that the current infrastructure is not ready to handle them. The authors do not provide any quantitative value concerning the data generated by the current or the future grid. Only a table shows the main characteristic of the sampling rate of part of the today's sensing and actuating infrastructure.

The concept of a distributed control for the power grid in the future evolution of the grid is illustrated also in [25]. The authors consider that the system will evolve from a mechanical-electrical control system to a fully electronics based one. They also state that more computation will take place locally on the grid with agent-based technologies operating within substations. Again, no sizing concerning the communication infrastructure to enable this additional information exchange between the substations is provided, thus difficult to have a quantitative picture of the Big Data of the smart grid.

In order to test the new scenarios of the data generation of the smart grid, EDF has estimated in its French network 35 millions smart meters and a sampling frequency of 10 minutes to have a total amount of data about 120 terabytes/year [26]. In that report, the authors show how a solution based on Hadoop and an accurate optimization in the data modeling, partitioning, and compression can improve the performance of the system that manages these data.

The IBM white paper on Big Data and smart grid provides some quantitative insights on the amount of data that utilities are going to deal in a smart grid future [27]. With a smart meter infrastructure and a 15 minutes sampling period, IBM forecasts a 3000 fold increase in the amount of data compared to the current monthly metering situation. Another quantitative example that is provided in the report to show the goodness of the IBM data management solution shows that in a 31-day period the total amount of storage required in a 100 millions meter scenario was less than 4 terabytes. The report explicitly emphasizes that the volume of data will not reach the same amount as for traditional data intensive industries, though in the power systems utilities world the projected

amount could be overwhelming. The report also emphasizes that in addition to the *velocity* characteristic of the smart grid data in collection, processing, and use, the utility will have to deal with the *variety* of the data to handle from power control system data, to surveillance videos, to geographical and meteorological data, to social media mining.

The literature provides good examples of how the power grid will evolve. For instance, in [28] we look at the topological aspects of the distribution grid through the lenses of complex network analysis [29]. In general, all the works agree that additional amount of computation will be required to monitor the supplementary information coming from more sensors deployed in the grid. However, the studies never provide quantitative information concerning the amount of data that AMI or data gathered at substation level will produce. Only few studies try to provide indication of the communication infrastructure required, but in a coarse way.

## VI. CONCLUDING REMARKS

Given the current state of affairs, the smart grid, or better said AMI, is not amenable to be referred to as an example of Big Data, not in the Netherlands, not in other countries. However, when more and more houses and businesses will be equipped with smart meters and the sampling period will be reduced, then the Dutch smart grid will become closer to be a Big Data system. If we compare the absolute numbers of the Big Data examples that we have analyzed in Section II, then there is no competition with the annual amount of data generated by social media or video repositories. However, to make the comparison fair considering Facebook, for example, if we account the the traffic generated by the more than 8 millions Dutch users of Facebook,<sup>8</sup> they are responsible for about 1.6 petabytes a year. We indeed note that the amount of data produced by the future Dutch smart grid is similar to the amount of data that the Dutch users produce each year on Facebook.

The numbers provided in this paper should be taken with a grain of salt. For instance, the Dutch smart meter standard is not yet finalized, therefore the amount of data and information metered could change, thus increasing or decreasing the data required. One aspect that we have not considered in our investigation is the overhead required by the communication infrastructures (i.e., the extra data in addition to the application layer that need to be considered) that could be a substantial burden. Further, all the data containing user and privacy related information will need to be protected by an additional layer of security (e.g., encryption) that require even more bits to be transmitted and stored. On the other hand, we have not considered any compression level possible for the smart grid data, but the figures here presented are raw data. Compression both in the communication and storage of data has achieved remarkable results [30]. For the protocols to control the grid we have considered one for the high

<sup>8</sup><http://goo.gl/HyKzxQ>

voltage (IEC 60870-5 [10]) and one for medium/low voltage (Modbus) among many others available and in use.

Utilities are likely to have to face a number of data related issues in the short and medium term. In the implementation of big-scale AMIs utilities will deal with ICT challenges that are not the core of their business. In addition to a secure telecommunication infrastructure to transfer the metered data and other information, the utilities will have to deal with the storage and management of that data. Of course this is not a big problem from a technical perspective, similar solutions exist in the financial sector. However, the smart grid will require a modernization of the utilities that in addition to the energy orientation will have to become also information oriented. The power industry from a fully analog business where consumer data was recorded annually with pen and paper is going to become the most data intensive industry with enormous quantity of data generated every day.

In this paper, we have made projections and calculations on how the relatively small grid of The Netherlands could become a Big Data generator similar to Facebook. Certainly, sensing more data along the grid to monitor its performance is essential for the utilities and provides an improvement in their operation, it will also ease the billing process and it will help the inclusion of renewable sources. Interesting challenges are not only in the management of the future smart grid data, but also on how to extract value from this data and on novel business models based on the availability of Big Data.

#### ACKNOWLEDGEMENTS

We thank Frank Blauw for useful comments on a previous version of the article. The work is supported by the Dutch National Research Council under the NWO Smart Energy Systems programme, contract no. 647.000.004. Pagani is supported by University of Groningen with the Ubbo Emmius Fellowship 2009 and IBM PhD Fellowship 2013-14.

#### REFERENCES

- [1] Netbeheer Nederland, "Betrouwbaarheid van elektriciteitsnetten in nederland - resulta ten 2012," Netbeheer Nederland, Tech. Rep. RM-ME-13L10440006, 2013.
- [2] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, 2011.
- [3] Cisco Systems, "Cisco Visual Networking Index: Forecast and Methodology, 2011-2016," Cisco Systems, Tech. Rep., 2012.
- [4] M. G. Morgan, J. Apt, L. B. Lave, M. D. Ilic, M. Sirbu, and J. M. Peha, "The many meanings of "smart grid"," Carnegie Mellon University, Tech. Rep., 2009.
- [5] National Energy Technology Laboratory, "A system view of the modern grid," U.S. Department of Energy - Office of Electricity Delivery and Energy Reliability, Tech. Rep., 2007.
- [6] C. Cuijpers and B.-J. Koops, "Smart metering and privacy in europe: Lessons from the dutch case," in *European Data Protection: Coming of Age*, ser. S. Gutwirth et al. (eds). Springer, 2013, pp. 269–293.
- [7] I. Georgievski, V. Degeler, G. Pagani, T. A. Nguyen, A. Lazovik, and M. Aiello, "Optimizing energy costs for offices connected to the smart grid," *Smart Grid, IEEE Transactions on*, vol. 3, no. 4, pp. 2273–2285, 2012.
- [8] M. Rosas-Casals and B. Corominas-Murtra, "Assessing European power grid reliability by means of topological measures," *Trans. of Ecology and the Environment*, no. 121, pp. 515–525, 2009.
- [9] Netbeheer Nederland, "Dutch smart meter requirements v. 4.0," Netbeheer Nederland, Tech. Rep., 2011.
- [10] G. Sanchez, I. Gomez, J. Luque, J. Benjumea, and O. Rivera, "Using internet protocols to implement iec 60870-5 telecontrol functions," *Power Delivery, IEEE Transactions on*, vol. 25, no. 1, pp. 407–416, 2010.
- [11] X. Miao, X. Chen, X. ming Ma, G. Liu, H. Feng, and X. Song, "Comparing smart grid technology standards roadmap of the iec, nist and sgcc," in *Electricity Distribution (CICED), 2012 China International Conference on*, 2012, pp. 1–4.
- [12] E. Kaldeli, E. U. Warriach, A. Lazovik, and M. Aiello, "Coordinating the web of services for a smart home," *ACM TWEB*, vol. 7, no. 2, p. 10, 2013.
- [13] G. A. Pagani and M. Aiello, "Service orientation and the smart grid state and trends," *Service Oriented Computing and Applications*, vol. 6, no. 3, pp. 267–282, 2012.
- [14] W. W. Peterson and E. J. Weldon, *Error-correcting codes*. The MIT Press, 1972.
- [15] H. Khurana, M. Hadley, N. Lu, and D. Frincke, "Smart-grid security issues," *Security Privacy, IEEE*, vol. 8, no. 1, pp. 81–85, 2010.
- [16] A. Metke and R. Ekl, "Security technology for smart grid networks," *Smart Grid, IEEE Transactions on*, vol. 1, no. 1, pp. 99–107, 2010.
- [17] M. HADLEY, N. Lu, and A. DEBORAH, "Smart-grid security issues," *IEEE Security and Privacy*, vol. 8, no. 1, pp. 81–85, 2010.
- [18] N. Capodieci, G. Cabri, G. A. Pagani, and M. Aiello, "An agent-based application to enable deregulated energy markets," in *36th Annual IEEE Computer Software and Applications Conference, COMPSAC 2012*, 2012, pp. 638–647.
- [19] S. S. ahnd Sreenadh Reddy Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007 – 1015, 2011, special Section on Offshore wind power planning, economics and environment. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S030142151000861X>
- [20] C. Warmer, K. Kok, S. Karnouskos, A. Weidlich, D. Nestle, P. Selzam, J. Ringelstein, A. Dimeas, and S. Drenkard, "Web services for integration of smart houses in the smart grid," *Grid-Interop-The road to an interoperable grid, Denver, Colorado, USA*, pp. 17–19, 2009.
- [21] S. Rusitschka, K. Eger, and C. Gerdes, "Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE Int. Conf. on*, 2010, pp. 483–488.
- [22] P. Parikh, M. Kanabar, and T. Sidhu, "Opportunities and challenges of wireless communication technologies for smart grid applications," in *Power and Energy Society General Meeting, 2010 IEEE*, 2010, pp. 1–7.
- [23] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, 2010, pp. 238–243.
- [24] Y.-J. Kim, M. Thottan, V. Kolesnikov, and W. Lee, "A secure decentralized data-centric information infrastructure for smart grid," *Comm. Mag., IEEE*, vol. 48, no. 11, pp. 58–65, 2010.
- [25] S. Amin and B. Wollenberg, "Toward a smart grid: power delivery for the 21st century," *Power and Energy Magazine, IEEE*, vol. 3, no. 5, pp. 34–41, 2005.
- [26] L. dos Santos, A. G. da Silva, B. Jacquin, M. Picard, D. Worms, and C. Bernard, "Massive smart meter data storage and processing on top of hadoop," in *BigData Workshop, VLDB 2012*, 2012.
- [27] IBM, "Managing big data for smart grids and smart meters," IBM Corporation, Tech. Rep., 2012.
- [28] G. A. Pagani and M. Aiello, "Towards decentralization: A topological investigation of the medium and low voltage grids," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 538 –547, 2011.
- [29] —, "The power grid as a complex network: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 1, pp. 2688–2700, 2013.
- [30] K. Sayood, *Introduction to data compression*. Morgan Kaufmann, 2012.