Akhavan-Hejazi, Hossein; Mohsenian-Rad, Hamed

Article
Power systems big data analytics: An assessment of paradigm shift barriers and prospects

Energy Reports

Energy Reports 4 (2018) 91–100

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Power systems big data analytics: An assessment of paradigm shift barriers and prospects

Hossein Akhavan-Hejazi [a], Hamed Mohsenian-Rad [a,b,*]

[a] Winston Chung Global Energy Center, University of California, Riverside, CA, USA
[b] Department of Electrical Engineering, University of California, Riverside, CA, USA

ARTICLE INFO

Article history:
Received 5 April 2017
Received in revised form 22 November 2017
Accepted 30 November 2017
Available online 20 February 2018

Keywords:
Energy
Big data analytics
Internet of energy
Smart grid

ABSTRACT

Electric power systems are taking drastic advances in deployment of information and communication technologies; numerous new measurement devices are installed in forms of advanced metering infrastructure, distributed energy resources (DER) monitoring systems, high frequency synchronized wide-area awareness systems that with great speed are generating immense volume of energy data. However, it is still questioned that whether the today's power system data, the structures and the tools being developed are indeed aligned with the pillars of the big data science. Further, several requirements and especial features of power systems and energy big data call for customized methods and platforms. This paper provides an assessment of the distinguished aspects in big data analytics developments in the domain of power systems. We perform several taxonomy of the existing and the missing elements in the structures and methods associated with big data analytics in power systems. We also provide a holistic outline, classifications, and concise discussions on the technical approaches, research opportunities, and application areas for energy big data analytics.

1. Introduction

Started in the information technology (IT), Big Data Analytics (BDA) has now found extensive applications in many areas of technology and business intelligence (Chen et al., 2012). Those serving mass consumers are particularly interested in using such tools to understand the current state of their business and track the still-evolving aspects. The electric power industry, interacting with one of the largest customer-serving critical networks is going through some drastic, rapid changes in both business and technical paradigms (Bui et al., 2012; Jaradat et al., 2015; Aiello and Pagani, 2014). Thus, naturally it is presenting limitless opportunities for BDA. Power system Big Data (BD) brings new opportunities such as providing an otherwise non-existing feedback loop, taking actions to correct and enhance planning, and enabling accurate realization of the system states, leading to more informed operations.

In this paper, we aim to overview some fundamental concepts and characterizations of BD and BDA, in the domain of power systems. We address questions such as: What are the attributes of energy data and whether they constitute BD? What are the distinct concepts in BDA related to power systems? What are the challenges in generation, communications, management and analysis of BD? What are the new core theories that furnish BDA in power system domain? What are the barriers to adopt the existing generic BDA tools and platforms for BDA in power systems?

2. Energy big data characterization

Although the term "Big Data" is self-explanatory, it still can be a source of confusion or controversy. For example, what an electric utility may consider BD could be seen as moderate size data for data-centric enterprises. The relativity of BD to the systems that operate based on those data is recognized even within the IT community (Chen et al., 2012; Russom et al., 2011). Nevertheless, a definition often used for BD is a "high-volume, high-velocity and high-variety information asset that requires and demands cost-effective, innovative forms of information collection, storage, and processing for enhanced insight and decision making" (De Mauro et al., 2016). From this definition, the volume, i.e., size of data is not the only factor, as there are other factors too. The so called "three Vs" of BD, see Fig. 1, are described as follows:

- Volume: Many IT-related organizations define BD in terabytes—sometimes petabytes (Cohen et al., 2009; Huang et al., 2014). For instance, the data warehouse of Fox Audience Network (a large advertisement network) holds over 200 terabytes of production data (Cohen et al., 2009). The scope of BD also affects its quantification. For example, as

* Corresponding author.
E-mail address: hamed@ece.ucr.edu (H. Mohsenian-Rad).

https://doi.org/10.1016/j.egyr.2017.11.002

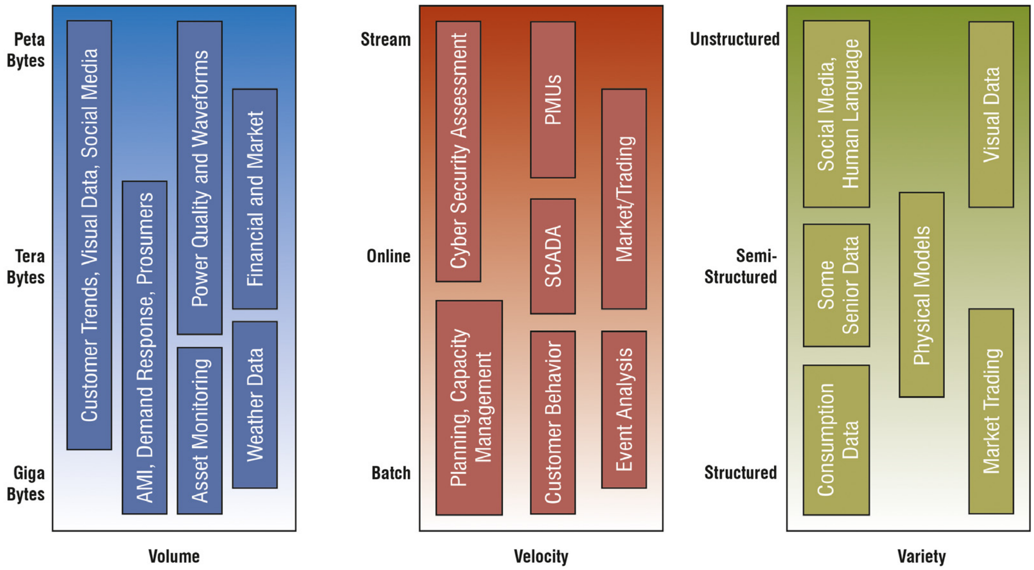

Fig. 1. The 3 V attributes of BD and some examples in the context of power systems.

shown in Fig. 2, one smart meter, with resolution of seconds to minutes generates much fewer data than one phasor measurement unit (PMU), with resolution of milliseconds; yet an advanced metering infrastructure (AMI) may generate large volume of data coming from millions of customers, e.g., the AMI in the New York State with seconds-resolution produces roughly 127.1 terabytes of consumption data per day (Huang et al., 2014).

- Variety: Data now comes from a much greater variety of sources compared to traditional data systems. The so-called structured data (e.g., tables and other data structures of relational databases, record formats of most applications, and the character-delimited rows of many flat files) which still form a majority of data, is now joined by *unstructured data* (e.g., text, voice, and video) and *semi-structured data* (e.g., XML, JSON, RSS feeds, and hierarchical data) (Russom et al., 2011). Examples of energy data are shown in Fig. 2.
- Velocity: The frequency of data generation or the frequency of data delivery is a key attribute. An example is a continuous stream of data, as opposed to once-in-a-while event-triggered data from a sensor. Although the majority of power system sensors are event-triggered, there are also sensors, e.g., PMUs both at transmission and distribution level, that produce data streams at high rates (Shand et al., 2015; Stewart et al., 2014b).

Do we currently face BD in the electric power sector? The answer does not seem to be a clear 'yes' or 'no'. Indeed, the *volume* of data being generated in the power sector has grown tremendously over the past few years due to the deployment of smart grid technologies, e.g., 45.8 million smart meters installed in US by 2013 (Alejandro et al., 2014). The energy data now come from *variety* of sources that span a wide range of locations, types and applications. Additionally, many forms of grid data are generated at high *velocities*. Yet, we may still have reservations to answer 'yes' to the above question, as we explain next.

A key reservation is that for many power system sensors, including many emerging and state-of-the-art sensors, the majority of data is either *not logged*, or they are *overwritten* very quickly. For example, in most protection relays and related sensors, the data collected is discarded shortly after internal use. Additionally, almost all state-of-the-art power quality sensors[1] tend to store the voltage or current waveforms only a few cycles before and after an *event* is detected (von Meier and McEachern, 2012). If a pre-programmed event is not detected, then no waveform data is automatically stored. Furthermore, the majority of measured data in power systems are intended to be used close to the point that they are generated, and were never intended to be carried to an enterprise data center, as opposed to the BD mentality in the IT sector (Stimmel, 2014). This is partly due to the different design requirements in the power sector, because traditionally, no centralized data storage model satisfied the needs of very low-latency controller systems in many practical power applications. Accordingly, while many of the recently deployed or emerging power data measurement systems lie in the description of BD, the way that they are currently managed does not exactly match the *soul* and *purpose* of BD. Once such hidden data is collected, managed, and analyzed, they will constitute the real BD in power systems. So far, we may have seen only the "Tip of the Iceberg" of the BD in power systems!

The type of data that can eventually form BD in power systems can be classified into *domain* data and *off-domain* data, see Fig. 3. The domain data can be further categorized by their sources. Here are few examples: (1) *Telemetry and SCADA data:* enable continuous flow of measurements on grid equipment status and parameters and other grid variables. The SCADA data can have various sources such as renewable energy resources which generate huge amount of data, such as real-time production, and equipment status. For instance the data from condition monitoring systems of many wind turbines can be utilized in predictive maintenance strategies (Qiu et al., 2016, 2012; Feng et al., 2013; Qiu et al., 2017; Long et al., 2015). (2) *Oscillographic and Synchrophasor data:* make up of voltage and current waveform samples in time or frequency domains that can create a graphical record. (3) *Consumption data:* is most often the smart meter data. (4) *Asynchronous event data:* often come from devices with embedded processors generating messages under a variety of normal and abnormal conditions. (5) *Metadata:* is any data that can describe other data. Grid metadata is highly diverse and may include internal sensor data, calibration data, and other device-specific information. (6) *Financial data:* may

[1] For example refer to PQube sensors at http://www.powersensorsltd.com/PQube.

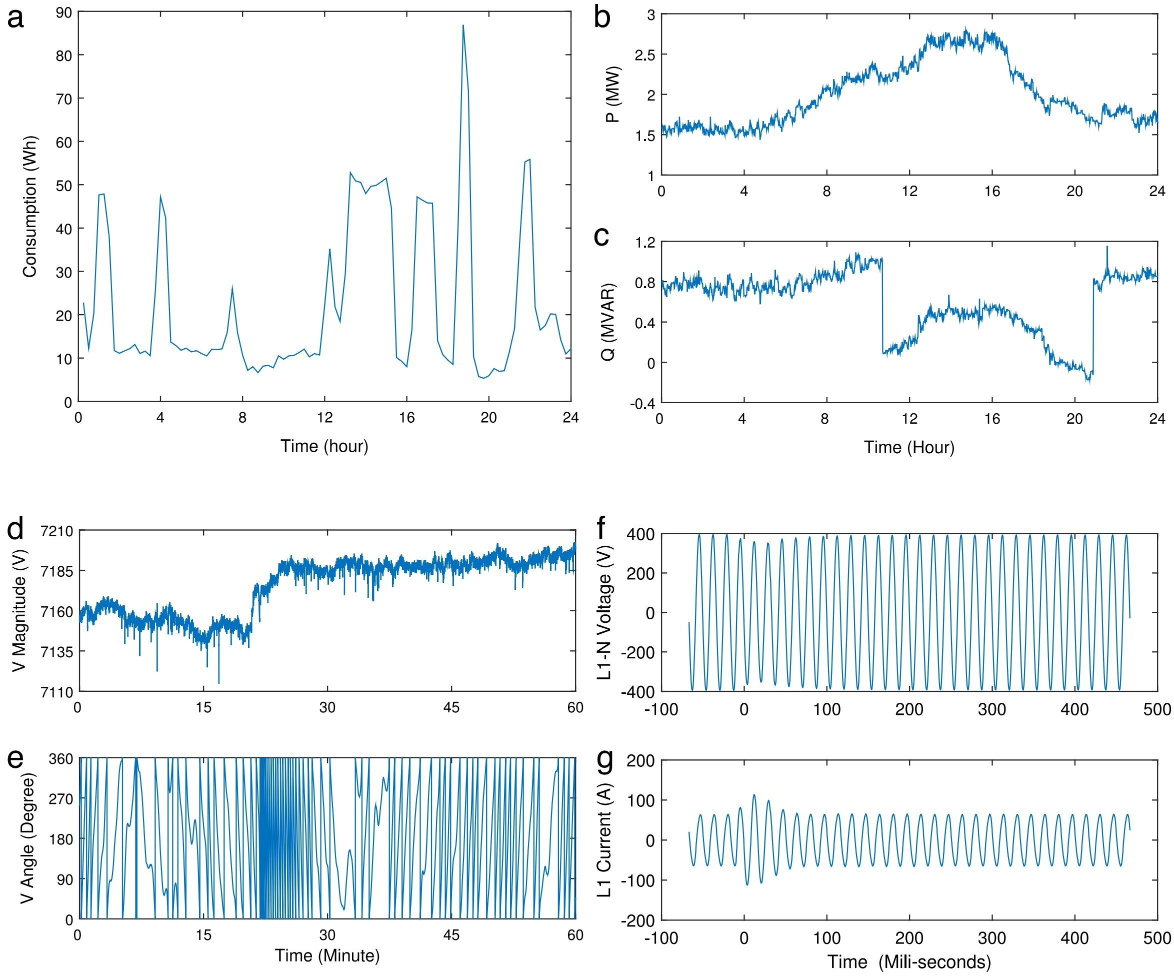

Fig. 2. Examples of power systems data: (a) Smart meter energy consumption readings once every 15 min (Pecon street database); (b) and (c) Substation active and reactive power readings once every 1 min; (d) and (e) PMU voltage magnitude and phase angle readings once every 8.3 ms; (f) and (g) Voltage and current waveform sampling once every 130 microseconds. The daily generated data size ranges from a few Kilobytes for smart meters to several Gigabytes for waveform sensors.

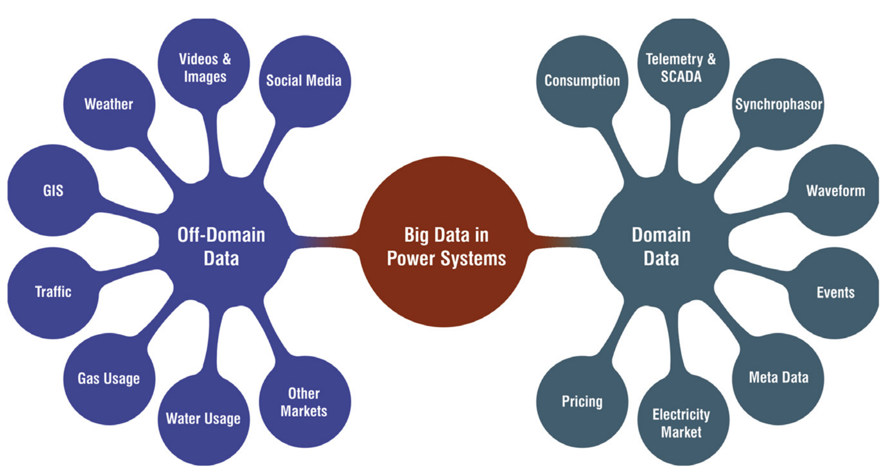

Fig. 3. Classification and examples of BD types in power systems.

include day-ahead and real-time market bids and price data, bilateral transactions, and retail rates.

Traditionally, power grid operation relies also on different forms of *off-domain* data, i.e., the data that is not specific to or necessarily intended for the power sector. For instance, the weather data, the data from the National Lightning Detection Network, and GIS data are currently used to enhance power system operations at different levels and time-scales, cf., (Chow et al., 2011; Paoli et al., 2010; Cummins et al., 1998). There are many forms of existing or emerging off-domain data that are yet to be exploited for the power grid operations and energy enterprise. Examples include traffic data, social media data, trade indices, and image and video streams, cf., (Huang et al., 2015; Moreno-Munoz et al., 2016). Essentially, there are no limits on the possibilities of the intelligence brought to power systems from all sources of data, that can collectively ultimately create BD.

3. Big data analytics: Re-thinking and re-structuring

There are several essential features that have come together to introduce the new practice of BDA. First, the BD itself has emerged in many sectors. Second, major advances have arisen in both hardware and software tools and platforms, cf., Hadoop[2] and Spark,[3] to increase affordability of massive data acquisition, communication, and storage. Accordingly, we now not only have the need for BDA but also the tools to do so, e.g., in form of predictive analytics (domain and off-domain data forecasting), data mining and machine learning (classification, regression, clustering), artificial intelligence (cognitive simulation, expert systems, perception, pattern recognition), statistical analysis, natural language processing, and advanced data visualization, cf. (Cohen et al., 2009; Slavakis et al., 2014; Bertsekas and Tsitsiklis, 1989; Chen et al., 2014; Zaki and Ho, 2000). Note that, the majority of these new tools and techniques have *discovery/exploratory natures*. That is, they do not require us to pre-determine what we expect to look for or see in the data. Finally, an important change that has distinguished the practice of BDA today is in the viewpoint towards data, as BD is now viewed as an important "asset" (Russom et al., 2011; Stimmel, 2014). Many traditional and dominant viewpoints and practices towards data are now questioned and redefined.

The new approach in viewing the data has some essential distinctions, see Fig. 4, with what used to be the accepted practice in enterprise data warehousing (EDW) and data management systems, including those of the current practice across the electric power industry:

(a) Data itself plays a centric role in BDA. Instead of building *systems* that manipulate certain data to reach certain foreseen objectives, the new paradigm requires establishing *platforms* upon BD to enable different possible (yet *unknown*) objectives to be pursued (Russom et al., 2011).

(b) The establishment of the above new platforms also means that more types and variety of data are stored even *before* their application and value are fully understood, whereas, under traditional data management, *data is often not kept, unless the value and application of the data is foreseen.*

(c) Given the ubiquity of data in the BDA structures, an EDW must keep pace by collecting data, *regardless of data quality refinements*. Traditional EDW approaches often do not incorporate data unless it is first carefully *cleansed and integrated* (to identify bad or missing data or to convert into a desirable format), e.g., through the traditional Extract–Transform–Load (ETL) procedures. In contrast, in the era of BD, the overhead of perfectly integrating new data sources into an architected data environment is substantial, and can hold up access to data. Therefore, under the BDA platform, an alternative ELT approach is considered, where transformation is done *after* loading, relying on DBMS transformation scripts which are capable of *parallel* execution (Stimmel, 2014).

(d) Data warehousing orthodoxy is based on long term, careful, and inflexible design and planning. However, a modern EDW must be *highly flexible* to allow analysts to ingest and in turn produce data at a rapid pace and in often changing structures. This requires a database whose physical and logical contents can be in continuous evolution (Cohen et al., 2009).

(e) Traditional EDWs have a high sensitivity in assigning the data to the right functions and right users. They are highly hierarchical in the sense that many data are associated to very limited set of functions, divisions, and users. The costs and security concerns, motivated to allow shared access only on a need basis. In contrast, the driving force of *exploring* as many unknown values and applications for BD requires broader access to data. This involves creating data management and security systems that are flexible in establishing and managing the access levels and sharing of the data among other functions, divisions, and users of the same enterprise or third parties (Stimmel, 2014; Hu and Vasilakos, 2016).

(f) Modern data analysis involve increasingly sophisticated methods that go well beyond the rollups of traditional information technology. Traditional EDW sometimes provides only certain statistics of data to user. Even if the statistics is indeed useful, analysts often need to know the exact derivation methods of such secondary data, and more importantly modify and re-purpose them if needed. Yet, the tools and algorithms that EDW provides are often black-boxed. The modern data warehouse should serve both as deep data repository and as sophisticated algorithmic runtime engine (Russom et al., 2011; Cohen et al., 2009).

(g) The traditional *centralized* approaches in EDWs, make them mission-critical, expensive resources, used for executive decision-making. However, due to a number of factors, e.g., the lower price of commodity clusters and the growing massive-scale data sources, there is now a new trend towards collecting and leveraging data in multiple *organizational units*, i.e. *data decentralization*.

In summary, the true interpretation and significance of BDA is achieved once we understand, adapt, and incorporate the new data management environment. Accordingly, the data management and computation structures and tools in the electric power sector are yet to go through an evolution in order to adapt to the new concepts and applications in BDA.

4. Big data analytics: A new science

The science of BDA is not limited to information technology to develop new platforms and tools to store, manage, and speed-process BD. It rather involves a wide range of methodologies developed across multiple disciplines to leverage related models and concepts. The adaption of BDA in different benchmark problems, particularly in power systems, is not always a homogeneous process. BDA often adds a new dimension to many problems.

Traditional analytical methodologies in power systems are typically largely model-based. In contrast, methodologies developed in the IT sector are often purely data driven. A *hybrid* of data and model-based approaches seem to be most effective in many energy benchmark problems. Additionally, there is often a need for technical trade-offs between the details of the models and the dimensionality of the data. *Distributed optimization methods* are a prime example of the methodologies that are now aimed to incorporate BD in a variety of applications in power systems (Slavakis et al., 2014; Bai et al., 2015; Xie et al., 2012). The context of many distributed optimization problems are unique and require customized methods in order to adapt to that particular application.

Broadly speaking, BDA methods aim to achieve a type of enhancement in knowledge or decision, thus, they can be characterized by the intelligence they aim to bring to power industry; these include descriptive, diagnostic, corrective, predictive, prescriptive, adaptive, and distributed analytics (Stimmel, 2014; Kezunovic et

[2] http://hadoop.apache.org/.
[3] http://spark.apache.org/.

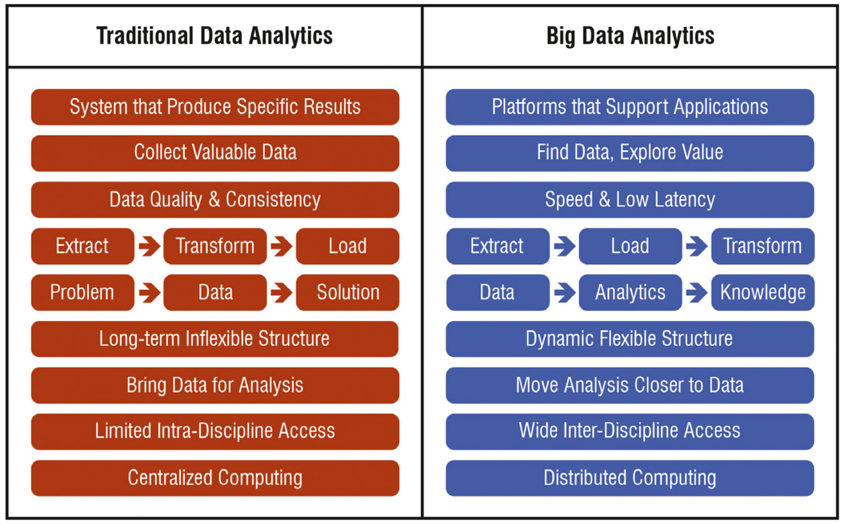

Fig. 4. A conceptual comparison between traditional data analytics and BDA.

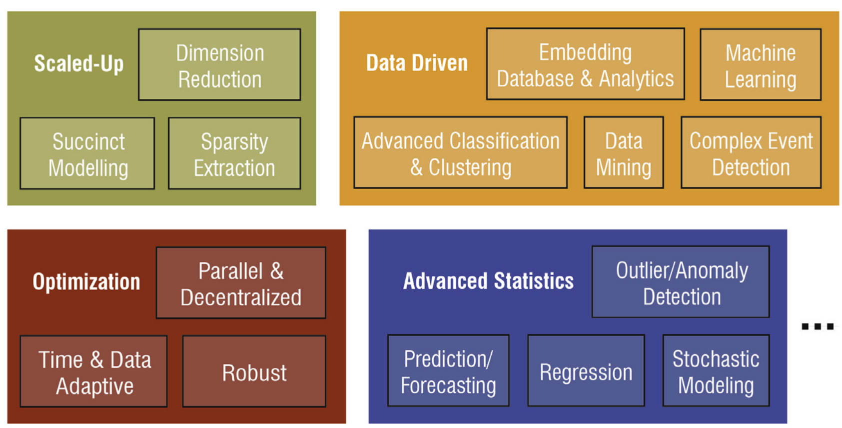

Fig. 5. Example methodologies in BDA to be expanded, enhanced, customized for BDA in power systems.

al., 2013). The efforts for developing many of these intelligent methods has already begun, see Fig. 5, and they are expected to grow in power community. Some of these efforts are good indications to familiarize us with flavor of possible BD methodologies:

(a) **Complex Event Processing:** It often operates in real time. It can be used in conjunction with a variety of other analytics and in a wide range of applications, e.g., on power quality, stability, etc., and are still evolving to handle streaming BD, cf. Andersen et al. (2015), Stewart and von Meier (2016) and Stewart et al. (2014a).

(b) **Data Mining and Computation within Databases:** Many common data mining methods, e.g., clustering, classification, association rules, concern themselves with assigning individual points to cohorts (class labels or cluster IDs). In BDA, these methods are designed to support flexible programming environments, which could be brought to BD scenarios, e.g., via extensible SQL or MapReduce algorithms. For example, data-parallel signal processing tools are developed within the database such as Ordinary Least Squares, Conjugate Gradient, Mann–Whitney U Testing, and general purpose tools, e.g., matrix multiplication and bootstrapping (Cohen et al., 2009). Such tools can also be exploited in power system analysis. For example, the tools such as BTrDB and DISTIL, allow ultra-fast processing on large volume of data for the distribution-level PMU (aka. μPMU) devices by bringing computation directly to the databases (Andersen and Culler, 2016; Arnold et al., 2017). BTrDB and DISTIL framework, provide a novel set of primitives, especially fast difference computation and rapid, low-overhead statistical queries. Thus, this provides capabilities such as locating transients in data spanning months, almost instantly. This concept is also leveraged in multiple applications that include more advanced rapid computations on the μPMU data (Ardakanian et al., 2017; Jamei et al., 2017). The work in Ardakanian et al. (2017) for instance, develops an extension to this framework that allows for on-line event detection and localization in unbalanced three-phase distribution systems.

(c) **Integrating Statistical Packages with Databases:** Databases often provide fairly limited statistical functionality. It is therefore standard practice to extract portions of a database into desktop statistical software package, e.g. SAS, Matlab or R. However, for large datasets, this means sampling the database to form an extract, which loses details. A better approach is to tightly integrate statistical computation with a massively parallel database (Choi et al., 1996), e.g., through ScaLAPACK.[4]

4 http://www.netlib.org/scalapack/.

(d) **Massive Parallelism:** As the number of the sources and devices that generate and operate based on data increases, methodologies are developed to distribute the intelligence and decision making process across many *local* devices. There are efforts, such as the works in Wytock and Kolter (2014), Parson (2014), Batra et al. (2017) and Sossan et al. (2017) on developing methodologies that obtain intelligence from *aggregation* and *disaggregation* of the data coming from sources such as AMI devices, PMUs, and other grid sensors. For example, Wytock and Kolter (Wytock and Kolter, 2014), developed a framework that is particularly suited for energy disaggregation on large volumes of AMI data, based on contextually supervised single-channel source separation. The application of this framework allows to provide itemized energy usage of different activities within households, from the aggregated whole-house power signal, in scales as large as thousands homes. The work in Parson (2014) also focuses on disaggregating the household's total electricity consumption into its contributing appliances using unsupervised training methods. This non-intrusive appliance load monitoring approach combines general appliance knowledge with the smart meter data to perform disaggregation. Neighborhood non intrusive load monitoring is another valuable application developed based on AMI data disaggregation which is very suited for large sets of coarse meter data (Batra et al., 2017). The method incorporated in this application finds homes *similar* to the *test home* and merges information obtained from them to estimate appliance energy usage. It matches every home with a set of *neighbors* that have direct sub-metering infrastructure, i.e. power meters on individual circuits or loads. The application of disaggregation can also be advantageous for distributed energy resources data such as solar generation. For solar systems, it can help estimating unmonitored PV generation and load profiles at downstream of a distribution feeder, by using the measurements on aggregated power flow at the substation/ feeder head, and the global horizontal irradiance data (Sossan et al., 2017; Kara et al., 2017).

(e) **Models that Scale Up:** The curse of dimensionality challenges many of the traditional analytic methods in the era of BD. Therefore, these well-established methods need to be revisited and incorporate alternative algorithms that are capable of scaling up with huge dimensions. There have been recent successful efforts to address modeling challenges in certain power system problems, e.g., in *nonlinear AC optimal power flow* (OPF) problems, by exploiting the sparsity of the system matrices, decomposing data into low rank and low variation components, or reformulating models to avoid failing on BD, cf., (Slavakis et al., 2014; Wu and Shahidehpour, 2010; Lavaei and Low, 2012; Madani et al., 2017).

5. Barriers to adopt big data analytics in power systems

Many energy technology developers as well as regulatory organizations worldwide have targeted extensive data analytics as one of the main solutions to modernize the electricity grid, and addressing the challenges introduced in the transformation of energy needs. For instance, the US Department of Energy (DoE) identifies communication, sensing, and data analytics as one of the four focus areas to solve the technical issues with aggressive integration of renewable resources (DE-FOA-0001495, 2016). They state that these efforts are essential to enable solar generation to grow from 1% of the current electricity supply mix to about 14% by 2030, as projected in the DOE SunShot Vision Study (Sunshot vision study, 2012), and to achieve improved reliability, resilience, affordability, and flexibility.

In order to adopt extensive advanced data analytics in power systems, the collection, communication and management of the data in this domain need to transform as well. BDA and BD structures are tightly linked together. That is, the new generation of advanced analytics is highly aware of the constraints and requirements of the data systems, and leverages the capabilities of those systems and structures closely. In return, the new systems and structures for data collection, communication, and data management in power systems, should be designed by targeting the BDA requirements and to provide improvements in analytics performance. Accordingly, redesigning the data structures and upgrading the data management systems in the power sector can furnish a foundation to expand and enhance BDA for more effective monitoring, control, and operation planning, hence, to improve grid performance. It is essential though to thoroughly identify the requirements, challenges, and features of BDA that are specific to power sector, for developing and reforming such data systems. There are barriers/steps which need to be overcome/taken in power grids to enable and facilitate BDA:

(a) **Addressing Discarded Data:** As pointed out earlier, the traditional approach to power system data management encompass only the most crucial data that are immediately needed to supervise and oversee predesigned operations and applications. The drivers for this approach in the past, has been the inability and/or high cost of transferring or managing many deployed sensors in data warehousing systems, as well as presuming no value for data that do not have a specified application(s). In contrast, the BDA involves extracting new values from the data that are collected not specifically for a defined purpose. For example, the authors in Shahsavari et al. (2017a, b, c) develop applications for diagnostics of fault events, and distribution system equipment malfunction based on high-resolution voltage and current measurements of the distribution-level PMUs that are deployed remotely further up- or down-stream in the electric distribution grid. Accordingly, even though the measurements are not collected originally for such applications, the exploratory BDA creates new values, for example by an opportunity to expedite maintenance on malfunctioning device to avoid early failure. Exploratory analytics based on SCADA alarm log data are also performed in Qiu et al. (2016, 2012), Feng et al. (2013), Qiu et al. (2017) and Long et al. (2015) to propose various novel applications such as failure detection of wind turbines and even quantifying the turbine gearbox's fatigue life. Accordingly, these are just few example works that provide evidence that the data in power systems should not be collected as need basis and the discarded data should be addressed.

(b) **Addressing Siloed Data:** The access to data is still one of the greatest challenges for analysts in many sectors. Based on a recent survey (Press, 2016), the data science practitioners have asserted that 80% of the time is involved in acquiring and preparing data. The data silos, the isolated stockpiles of data that are prohibitively restricted from any non-intended use or user, is a great barrier to BDA development. In power systems, siloed data poses as even a greater challenge. Data availability across entities aside, different divisions within entities such as distribution system operators, or transmission operators do not have access to data from each others' division. Some of the drivers could be related to structural shortcomings. Considering natural limits on resources, applications and databases had been optimized for their main function and specific teams, and the incentives of individual teams often not sufficient to encourage data sharing. In addition to drivers that are shared with other sectors, the sensitivity over power system operational data could also be traced to cyber-physical security. The physical system has shown vulnerable in the past and the ways that such vulnerability could be exploited and the extent of associated damages and costs are not fully known. For instance, the BlackEnergy malware caused extensive power outage in Ukraine in 2016. BlackEnergy was discovered on the Internet-connected SCADA systems, affecting General Electric Cimplicity, Advantech/Broadwin WebAccess, and Siemens WinCC (ICS-CERT, 2014). Until the causes for restrictions are not fully addressed, it is likely that these data silos continue the restrictions on access, and impede BDA in power systems. Nevertheless, the issue could be tackled by identifying methods and approaches that provide the means to remove at least certain types of data silos.

(c) **Supporting Real-time Analytics:** Many applications of energy BDA are with respect to automated operation/ control, requiring real-time data collection and real-time actions. Many applications of energy BDA are with respect to automated operation control, requiring real-time data collection and real-time actions. This requirement is drastically different from many BDA applications in the IT sector, since power grid is a critical infrastructure. Real-time monitoring, control, and operation of power systems are identified by many government authorities such as the US DOE, to be essential in support of modern power transmission and distribution grids (DE-FOA-0001495, 2016). They can particularly facilitate the integration of greater amounts of distributed energy resources (DERs), such as solar generation systems residing at the grid edges. Real time monitoring and control of DERs can address the challenges associated with such resources and alleviate issues such as reverse power flow (back injection) in distribution feeders and into substations, excessive operation of distribution system's primary voltage control equipment, and reconfiguration of protection equipment to handle bi-directional power flows and still trip for system faults correctly, without nuisance or sympathetic tripping. However, the systems that would monitor and control millions of such edge resources and devices utilizing various BDA, entail a revision of approach in generation, transfer, and management of the data from the real-time measurements and operational set points. If the control signals are needed to be dispatched to many DERs across distribution feeders in fractions of seconds to respond to variations of these resources and hence the grid conditions, a fast and reliable communication would be a great challenge. The existing communications networks that must support such real-time access to data often have unreliable performance, in terms of response time, bandwidth and latency. Thus, the systems that generate, manage, or utilize the data in monitoring and control should be designed by accounting for such constraints and to provide fail-safe functionalities. Also methods and technologies should be employed to decrease the concentration of data exchanged for real-time actions, to shift from a centralized control scheme to a distributed and/or hierarchical architecture and to decrease the sensitivity of controllers to faulty data. For instance, approaches could be developed to proactively manage very large distributed energy resource populations using only a few measurement points for input through predictive state estimation and a few carefully selected control nodes in order to provide system-wide monitoring and control using a small fraction of the active devices on the grid. Many efforts, e.g., several projects as part of the DOE Sunshut Energize initiative (DE-FOA-0001495, 2016), are initiated already to enable new control and monitoring technologies based on real-time analytics and optimization for next-generation power distribution systems with massive numbers of DERs, which can integrate real-time distributed control with system-wide energy management (Bernstein and Dall'Anese, 2017; Kroposki et al., 2017; Guo et al., 2017; Guggilam et al., 2017). Utilities may need to upgrade their communication systems and to employ advanced network designs that support service differentiation, e.g., to distinguish delivering of critical protection-relay data from non-critical billing data. BDA and data management systems must become aware of system limitations in transferring different data. Finally, the network design must balance the overhead on the system with the speed needed for various signals. For example, cyber-security requirements will introduce more latency into the signal path. Industry standard communication protocols introduce overhead, as well.

(d) **Coexistence of Centralized and Distributed Data Management:** As advocated by the IT sector, the common approach to BDA is to migrate from centralized data management to distributed data management systems. The purpose is to reduce the cost and overhead of the data and systems integration. At the first glance, this may seem to fit well in the context of power systems to extend the data structure by establishing distributed file management systems, because BD in power systems is naturally distributed, with respect to geographical regions and the variety of operators that overlook the data. However, the transformation or retirement of the traditional centralized systems is neither possible nor effective. Thus, supporting a coexistence and coordination among the existing centralized and the future distributed architectures is essential to enable BDA in power grids.

(e) **Balancing Integrated and Disintegrated Systems:** Another challenge is the extent of disintegration; should we disintegrate the data management/analytic system to every corner and every device across the power grid? A fully integrated and a fully disintegrated design are both likely to be ineffective in power systems. While the disperse nature of data sources in power grid calls for disintegrated data management, it also poses as a barrier due to the communications and cyber-security limits. Therefore, reaching to a right tradeoff between integrated or disintegrated systems is challenging.

(f) **Customized Data Management Systems to Cope with Fast Data:** The requirements for coping with the speed of data and processes in power grids sometimes even extend those in the IT sector. The sampling rate of certain power sensor devices, such as PMUs, are so high, and the time window of some processes is so tight that the generic commercial database systems such as SQL or HDFS are not sufficient. Therefore, an advanced practice of BDA will involve different domain specific data management systems to cope with its speed requirements. An example of such custom-developed platforms is the BTrDB[5] developed for management of distribution-level PMU measurements. BTrDB is constructed to provide both higher sustained throughput for raw inserts and queries, as well as advanced primitive that accelerate the analysis of the expected 44 quadrillion datapoints per year per server (von Meier and McEachern, 2012; Andersen et al., 2015; Andersen and Culler, 2016; Arnold et al., 2017). The principles and design of this database are also applicable to a large variety of time-series types. BTrDB also provides

5 http://btrdb.io/.

the foundation for monitoring and visualizing the distribution grid behaviors, in near real-time using the DISTIL framework. DISTIL framework enables agile development of scalable analysis pipelines with strict guarantees on result integrity despite asynchronous changes in data or out of order arrival.

The above list gives just a few representative challenges that we will face in realizing the BDA in power grids. Many other potential challenges are yet to be discovered.

6. Opportunities and applications in power systems

Many application areas of BDA in power systems are not unveiled yet. Nevertheless, several horizons of opportunities are expected already (Bui et al., 2012; Huang et al., 2014; Stimmel, 2014; Kezunovic et al., 2013; Moreno-Munoz et al., 2016; Hu and Vasilakos, 2016; Yu et al., 2015; Yin et al., 2013):

(a) **Enhanced Demand Response:** Demand response, so far, has largely been the province of large utilities and large customers, due to the hindrance in management and value proposition for the operation of huge number of small loads. However, BDA will finally enable utilities, or rather the cloud-based platforms employed by utilities, to look simultaneously at the consumption patterns of millions of users and rapidly determine which customers would be willing to participate in a DR event, how much these customers will charge for participation, and how much will actually be saved. Several approaches and tools are developed already based on BDA that can significantly enhance demand response. For example, Visualization and Insight System for Demand Operations and Management (VISDOM) is a platform for interpretation and extracting actionable information from huge samples of smart meter data in a given utility service area or geographic region (Kwac et al., 2014; Borgeson et al., 2015; Arguelles and Iglesias, 0000; Damström and Gerlitz, 2016). VISDOM is a collection of smart meter data analysis algorithms and visualization tools designed to address the challenge of interpreting patterns in energy data in support of utility energy efficiency and demand response programs. As one of it features, VISDOM allows to filter, sort, and assemble subsets of consumers using feature criteria and to display attributes of filtered customers using generic (i.e. histogram and scatter plot) or task specialized (i.e. cumulative sum and load shape) interactive visualizations (Borgeson et al., 2015).

(b) **Disaggregation and Fine Granularity Forecast:** By employing BDA, forecasts can be issued per-customer for millions of customers every few minutes to fine-tune predictions for power load across an entire region, in specific geographic areas, or along particular distribution feeder branches. The ability to forecast every meter, transformer, and feeder will improve forecast quality and save in operating costs. For example, forecasting tools can be utilized in various applications, such as electricity market prices, or customer response (Kwac and Rajagopal, 2016; Balar et al., 2013; Gama and Rodrigues, 2007; Kezunovic et al., 2013). Clustering techniques have been proposed for determining natural segmentation of customers and identification of temporal consumption patterns (Kwac and Rajagopal, 2016; Balar et al., 2013). Price forecasting methods based on big price data have been proposed utilizing algorithms such as Grey Correlation Analysis, combination of Kernel function and Principle Component Analysis, and Support Vector Machine (Gama and Rodrigues, 2007). Architectures are also proposed, e.g. in Kezunovic et al. (2013), based on an online clustering algorithm and utilizing neural-network based predictive models.

(c) **Utilizing Domain and Off-Domain Data for Fault/Outage Detection:** As intelligence is extended down into distribution feeders, much data can be generated and used for outage detection and power restoration to help utilities improve on critical indices. Smart meters record electricity usage for billing, measure end-of-line voltage and, in the case of an outage, emit a last gasp as they lose power. With the widespread use of Social Media, similar/complimentary information may be obtained from data sources such as customers' mobile tweets, that are linked to an address. For example, analytics such as neural networks, or fuzzy logics have been already utilized for calculating the fault thresholds in real-time in order to overcome some of the challenges in the traditional protection schemes, particularly for scenarios that fault is nonstationary, contains fundamental frequency components, or DC offsets (Kezunovic et al., 2013; Vasilic and Kezunovic, 2005). Accordingly, outage detection and power restoration will be expedited.

(d) **Operations-Planning Convergence:** This so-called convergence refers to the ability of a utility to realize the future conditions of power system with high probability and high accuracy. Operational planning refers to preparation for how weather, load, and generation conditions may change in the next minutes, hours, and days. This is difficult to achieve without systematic data management and accurate modeling. There are various reasons for this convergence gap, e.g. diverse models, diverse data sources and data formats, and inefficient data management tools, which all can be overcome with the unified methods and systematic data management. The challenges in this area are very broad. Some of the applications and approaches that are needed in this area include, but are not limited to: stochastic analysis, predictive analysis, dynamic scenario building for "what-if studies", and machine learning techniques to predict power system behaviors, capability to automatically recognize the changes in network topology due to planned and unplanned switching events, and ultimately enabling the power system to operate as interconnected subnetworks, that can be combined or separated, to achieve optimal power flows and improved reliability (DE-FOA-0001495, 2016). Data-driven approaches, statistical analyses, and stochastic optimization methods have been proposed, e.g. in Hejazi and Mohsenian-Rad (2017), DallAnese et al. (2017, 2015), to plan for the stochastic behaviors of the distributed renewable resources, or to enable predictive stochastic dispatch and operation of energy storage devices and residential photovoltaic inverters under forecasting uncertainties.

(e) **Equipment Monitoring and Life Extension by Predictive Fault Detection:** Predictive maintenance requires detailed information on the condition of equipment. Dedicated condition monitoring systems, and or personnel to perform interval testing have been highly costly for power system equipment that are diversely located on different parts of the grid. With BDA methods however, the close monitoring, degradation detection, and early failure prevention on many system assets can be achieved. Additionally, the data from existing monitoring systems, that are deployed for system performance assessment, can be used to additionally serve for equipment condition monitoring. For example, the authors in Qiu et al. (2016, 2012), Feng et al. (2013), Qiu et al. (2017) and Long et al. (2015) develop applications for facilitating accurate wind turbine failure detection based on the turbine SCADA data. Note that, all large utility scale turbines have a standard SCADA system principally used for performance monitoring. The authors also showed in a SCADA signal case study on a 2 MW class variable-speed wind turbine that by monitoring gearbox oil temperature rise, power output and rotational speed, a gearbox planetary stage failure could be predicted and detected. Developing similar applications for power system equipment monitoring, based on BDA could immensely decrease the O&M cost of the system and the overall investments on new assess.

(f) **Feature Extraction and Advanced Visualization:** The current automated reasoning tools often suffer in accuracy and time complexity when applied to data with many unnecessary attributes. However, BDA methods can help in distinguishing between those which are useful and should be used in decision making, and those which are not. Visual tools combined with operator-feedback machine learning will assist in making such distinctions.

In summary, the foundation of the emerging and potential applications of BDA in power systems are expected to be built upon four new features: (1) Higher volumes of data become available, which enables more robust statistical or data mining analysis that allows increased process accuracy and enhanced control. (2) New types of data are emerging, which allows for creating new feedback loops for planning and operation. (3) Data can now be better managed, which provides clear, actionable information. (4) Advanced analytics is emerging, which allows uncovering new facts about systems to support complex decision-making.

7. Conclusion

Although the energy domain data have been growing immensely, a majority of power system data is yet to be exploited. Many of energy domain legacy measurement devices and data management systems are based on the traditional concept of enterprise data warehousing, whereas under Big Data Analytic (BDA) approach some of its key fundamentals have been reconsidered. We identify several barriers and steps that need to be overcome/taken in power grids to enable and facilitate BDA. We also classify several core concepts, theories, and methods that could be leveraged in energy BDA. Finally, we outline several of BDA research and application horizons in power system domain.

Acknowledgments

This work is supported in part by Winston Chung Global Energy Center, NSF grants 1462530 and 1405330, and DoE grant EE0008001.

References

DE-FOA-0001495, 2016. Enabling extreme real-time grid integration of solar energy (energise), Tech. rep., DEPARTMENT OF ENERGY (DOE).

Sunshot vision study, February 2012. Tech. rep., Online access: http://energy.gov/eere/sunshot/sunshot-vision-study.

ICS-CERT, 2014. Ongoing sophisticated malware campaign compromising ics (update b), Tech. rep., Department of Homeland Security.

Aiello, M., Pagani, G.A., 2014. The smart grid's data generating potentials, in: 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, pp. 9–16.

Alejandro, L., Blair, C., Bloodgood, L., Khan, M., Lawless, M., Meehan, D., Schneider, P., Tsuji, K., 2014. Global market for smart electricity meters: Government policies driving strong growth. US International Trade Commission - Office of Industries.

Andersen, M.P., Culler, D.E., 2016. Btrdb: Optimizing storage system design for timeseries processing, in: Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST 16).

Andersen, M.P., Kumar, S., Brooks, C., von Meier, A., Culler, D.E., 2015. Distil: design and implementation of a scalable synchrophasor data processing system. In: 2015 IEEE International Conference on Smart Grid Communications (SmartGridComm). IEEE, pp. 271–277.

Ardakanian, O., Yuan, Y., Dobbe, R., von Meier, A., Low, S., Tomlin, C., 2017. Event detection and localization in distribution grids with phasor measurement units. arXiv preprint arXiv:1611.04653.

Arguelles, D., Iglesias, R., Visual filtering of large energy consumption datasets by leveraging usage clusters.

Arnold, D.B., Roberts, C., Ardakanian, O., Stewart, E.M., 2017. Synchrophasor data analytics in distribution grids. In: Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2017 IEEE. IEEE, pp. 1–5.

Bai, Y., Zhong, H., Xia, Q., Kang, C., Xie, L., 2015. A decomposition method for network-constrained unit commitment with ac power flow constraints. Energy 88, 595–603.

Balar, A., Malviya, N., Prasad, S., Gangurde, A., 2013. Forecasting consumer behavior with innovative value proposition for organizations using big data analytics. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, pp. 1–4.

Batra, N., Singh, A., Whitehouse, K., 2017. Neighbourhood nilm: A big-data approach to household energy disaggregation. arXiv preprint arXiv:1511.02900.

Bernstein, A., Dall'Anese, E., 2017. Real-time feedback-based optimization of distribution grids: A unified approach. arXiv preprint arXiv:1711.01627.

Bertsekas, D.P., Tsitsiklis, J.N., 1989. Parallel and Distributed Computation: Numerical Methods, Vol. 23. Prentice hall, Englewood Cliffs, NJ.

Borgeson, S., Flora, J.A., Kwac, J., Tan, C.-W., Rajagopal, R., 2015. Learning from Hourly Household Energy Consumption: Extracting, Visualizing and Interpreting Household Smart Meter Data, in: International Conference of Design, User Experience, and Usability. Springer, pp. 337–345.

Bui, N., Castellani, A.P., Casari, P., Zorzi, M., 2012. The internet of energy: A web-enabled smart grid system. IEEE Netw. 26 (4).

Chen, H., Chiang, R.H., Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. MIS Q. 36 (4), 1165–1188.

Chen, Y., Xie, L., Kumar, P.R., 2014. Power system event classification via dimensionality reduction of synchrophasor data. In: Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 IEEE 8th. IEEE, pp. 57–60.

Choi, J., Demmel, J., Dhillon, I., Dongarra, J., Ostrouchov, S., Petitet, A., Stanley, K., Walker, D., Whaley, R.C., 1996. Scalapack: A portable linear algebra library for distributed memory computers design issues and performance. Comput. Phys. Comm. 97 (12), 1–15.

Chow, C.W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B., 2011. Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed. Sol. Energy 85 (11), 2881–2893.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C., 2009. Mad skills: New analysis practices for big data. Proc. VLDB Endow. 2 (2), 1481–1492.

Cummins, K.L., Krider, E.P., Malone, M.D., 1998. The us national lightning detection network/sup tm/and applications of cloud-to-ground lightning data by electric power utilities. IEEE Trans. Electromagn. Compat. 40 (4), 465–480.

DallAnese, E., Baker, K., Summers, T., 2017. Chance-constrained ac optimal power flow for distribution systems with renewables. IEEE Trans. Power Syst. 32 (5), 3427–3438. http://dx.doi.org/10.1109/TPWRS.2017.2656080.

DallAnese, E., Dhople, S.V., Johnson, B.B., Giannakis, G.B., 2015. Optimal dispatch of residential photovoltaic inverters under forecasting uncertainties. IEEE J. Photovolt. 5 (1), 350–359. http://dx.doi.org/10.1109/JPHOTOV.2014.2364125.

Damström, J., Gerlitz, C., 2016. Classification of power consumption patterns for swedish households using k-means.

De Mauro, A., Greco, M., Grimaldi, M., 2016. A formal definition of big data based on its essential features. Libr. Rev. 65 (3), 122–135.

Feng, Y., Qiu, Y., Crabtree, C.J., Long, H., Tavner, P.J., 2013. Monitoring wind turbine gearboxes. Wind Energy 16 (5), 728–740.

Gama, J., Rodrigues, P.P., 2007. Stream-based electricity load forecast. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer, pp. 446–453.

Guggilam, S.S., Zhao, C., DallAnese, E., Chen, Y.C., Dhople, S.V., 2017. Primary frequency response with aggregated ders. In: American Control Conference (ACC), 2017. IEEE, pp. 3386–3393.

Guo, Y., Baker, K., DallAnese, E., Hu, Z., Summers, T., 2017. Stochastic optimal power flow based on data-driven distributionally robust optimization. arXiv preprint arXiv:1706.04267.

Hejazi, H., Mohsenian-Rad, H., 2017. Energy storage planning in active distribution grids: A chance-constrained optimization with non-parametric probability functions. IEEE Trans. Smart Grid.

Hu, J., Vasilakos, A.V., 2016. Energy big data analytics and security: Challenges and opportunities. IEEE Trans. Smart Grid 7 (5), 2423–2436.

Huang, Y., Warnier, M., Brazier, F., Miorandi, D., 2015. Social networking for smart grid users, in: 2015 IEEE 12th International Conference on Networking, Sensing and Control, pp. 438–443. https://doi.org/10.1109/ICNSC.2015.7116077.

Huang, Z., Luo, H., Skoda, D., Zhu, T., Gu, Y., 2014. E-sketch: Gathering large-scale energy consumption data based on consumption patterns, in: Big Data (Big Data), 2014 IEEE International Conference on, IEEE, pp. 656–665.

Jamei, M., Scaglione, A., Roberts, C., Stewart, E., Peisert, S., McParland, C., McEachern, A., 2017. Anomaly detection using optimally-placed pmu sensors in distribution grids. IEEE Trans. Power Syst. PP (99), 1–1. http://dx.doi.org/10.1109/TPWRS.2017.2764882.

Jaradat, M., Jarrah, M., Bousselham, A., Jararweh, Y., Al-Ayyoub, M., 2015. The internet of energy: Smart sensor networks and big data management for smart grid. Procedia Comput. Sci. 56, 592–597. http://dx.doi.org/10.1016/j.procs.2015.07.250.

Kara, E.C., Roberts, C.M., Tabone, M., Alvarez, L., Callaway, D.S., Stewart, E.M., 2017. Towards real-time estimation of solar generation from micro-synchrophasor measurements. arXiv preprint arXiv:1607.02919.

Kezunovic, M., Xie, L., Grijalva, S., 2013. The role of big data in improving power system operation and protection. In: Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium. IEEE, pp. 1–9.

Kroposki, B.D., Dall-Anese, E., Bernstein, A., Zhang, Y., Hodge, B.S., 2017. Autonomous energy grids: Preprint, Tech. rep., National Renewable Energy Lab. (NREL), Golden, CO (United States).

Kwac, J., Flora, J., Rajagopal, R., 2014. Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid 5 (1), 420–430.

Kwac, J., Rajagopal, R., 2016. Data-driven targeting of customers for demand response. IEEE Trans. Smart Grid 7 (5), 2199–2207.

Lavaei, J., Low, S.H., 2012. Zero duality gap in optimal power flow problem. IEEE Trans. Power Syst. 27 (1), 92–107. http://dx.doi.org/10.1109/TPWRS.2011.2160974.

Long, H., Wang, L., Zhang, Z., Song, Z., Xu, J., 2015. Data-driven wind turbine power generation performance monitoring. IEEE Trans. Ind. Electron. 62 (10), 6627–6635.

Madani, R., Lavaei, J., Baldick, R., Atamtürk, A., 2017. Power system state estimation and bad data detection by means of conic relaxation, in: Proceedings of the 50th Hawaii International Conference on System Sciences.

von Meier, A., McEachern, A., 2012. Micro-synchrophasors: A promising new measurement technology for the ac grid, Tech. rep., CIEE.

Moreno-Munoz, A., Bellido-Outeirino, F., Siano, P., Gomez-Nieto, M., 2016. Mobile social media for smart grids customer engagement: Emerging trends and challenges. Renew. Sustain. Energy Rev. 53, 1611–1616.

Paoli, C., Voyant, C., Muselli, M., Nivet, M.-L., 2010. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol. Energy 84 (12), 2146–2160.

Parson, O., 2014. Unsupervised training methods for non-intrusive appliance load monitoring from smart meter data, Ph.D. thesis, University of Southampton.

Pecon street database, retrived from http://www.pecanstreet.org/.

Press, G., March 2016. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says, Forbs Magazine.

Qiu, Y., Chen, L., Feng, Y., Xu, Y., 2017. An approach of quantifying gear fatigue life for wind turbine gearboxes using supervisory control and data acquisition data. Energies 10 (8), 1084.

Qiu, Y., Feng, Y., Sun, J., Zhang, W., Infield, D., 2016. Applying thermophysics for wind turbine drivetrain fault diagnosis using scada data. IET Renew. Power Gener. 10 (5), 661–668.

Qiu, Y., Feng, Y., Tavner, P., Richardson, P., Erdos, G., Chen, B., 2012. Wind turbine scada alarm analysis for improving reliability. Wind Energy 15 (8), 951–966.

Russom, P., 2011. Big data analytics, TDWI Best Practices Report, Fourth Quarter, pp. 1–35.

Shahsavari, A., Farajollahi, M., Stewart, E., von Meier, A., Alvarez, L., Cortez, E., Mohsenian-Rad, H., 2017a. A data-driven analysis of capacitor bank operation at a distribution feeder using micro-pmu data. In: Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2017 IEEE. IEEE, pp. 1–5.

Shahsavari, A., Farajollahi, M., Stewart, E., Roberts, C., Mohsenian-Rad, H., 2017. A data-driven analysis of lightning-initiated contingencies at a distribution grid with a pv farm using micro-pmu data, in: Proc. of IEEE PES North American Power Symposium, Morgantown, WV.

Shahsavari, A., Sadeghi-Mobarakeh, A., Stewart, E.M., Cortez, E., Alvarez, L., Megala, F., Mohsenian-Rad, H., 2017b. Distribution grid reliability versus regulation market efficiency: An analysis based on micro-pmu data. IEEE Trans. Smart Grid 8 (6), 2916–2925.

Shand, C., McMorran, A., Stewart, E., Taylor, G., 2015. Exploiting massive pmu data analysis for lv distribution network model validation, in: 2015 50th International Universities Power Engineering Conference (UPEC), pp. 1–4. http://dx.doi.org/10.1109/UPEC.2015.7339798.

Slavakis, K., Giannakis, G., Mateos, G., 2014. Modeling and optimization for big data analytics:(Statistical) learning tools for our era of data deluge. IEEE Signal Process. Mag. 31 (5), 18–31.

Sossan, F., Nespoli, L., Medici, V., Paolone, M., 2017. Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers. arXiv preprint arXiv:1706.04821.

Stewart, E.M., Kiliccote, S., McParland, C., Roberts, C., Arghandeh, R., von Meier, A., 2014a. Using micro-synchrophasor data for advanced distribution grid planning and operations analysis. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA.

Stewart, E.M., Kiliccote, S., Shand, C.M., McMorran, A.W., Arghandeh, R., von Meier, A., 2014b. Addressing the challenges for integrating micro-synchrophasor data with operational system applications, in: 2014 IEEE PES General Meeting—Conference Exposition, pp. 1–5. http://dx.doi.org/10.1109/PESGM.2014.6938994.

Stewart, E.M., von Meier, A., 2016. Phasor Measurements for Distribution System Applications. John Wiley & Sons, Ltd, . http://dx.doi.org/10.1002/9781118755471.sgd087.

Stimmel, C.L., 2014. Big Data Analytics Strategies for the Smart Grid. Auerbach Publications.

Vasilic, S., Kezunovic, M., 2005. Fuzzy art neural network algorithm for classifying the power system faults. IEEE Trans. Power Deliv. 20 (2), 1306–1314.

Wu, L., Shahidehpour, M., 2010. Accelerating the benders decomposition for network-constrained unit commitment problems. Energy Syst. 1 (3), 339–376. http://dx.doi.org/10.1007/s12667-010-0015-4.

Wytock, M., Kolter, J.Z., 2014. Contextually supervised source separation with application to energy disaggregation., in: AAAI, pp. 486–492.

Xie, L., Choi, D.-H., Kar, S., Poor, H.V., 2012. Fully distributed state estimation for wide-area monitoring systems. IEEE Trans. Smart Grid 3 (3), 1154–1169.

Yin, J., Sharma, P., Gorton, I., Akyoli, B., 2013. Large-scale data challenges in future power grids. In: 2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE). IEEE, pp. 324–328.

Yu, N., Shah, S., Johnson, R., Sherick, R., Hong, M., Loparo, K., 2015. Big data analytics in power distribution systems. In: Innovative Smart Grid Technologies Conference (ISGT), 2015 IEEE Power & Energy Society. IEEE, pp. 1–5.

Zaki, M.J., Ho, C.-T., 2000. Large-Scale Parallel Data Mining. Springer Science & Business Media.