INL/JOU-17-42462-Revision-0

Big Data Analytics in Smart Grids: State-of-the-Art, Challenges, Opportunities, and Future Directions

Bishnu P. Bhattarai, Yusheng Luo, Rob Hovsapian, Manish Mohanpurkar, Kurt S. Myers, Sumit Paudyal, Reinaldo Tonkoski, Kwok Cheung, Rui Zhang, Milos Manic, Power Zhao, Song Zhang, Xiaping Zhang

June 2019

The INL is a U.S. Department of Energy National Laboratory
operated by Battelle Energy Alliance

INL/JOU-17-42462-Revision-0

Big Data Analytics in Smart Grids: State-of-the-Art, Challenges, Opportunities, and Future Directions

Bishnu P. Bhattarai, Yusheng Luo, Rob Hovsapian, Manish Mohanpurkar, Kurt S. Myers, Sumit Paudyal, Reinaldo Tonkoski, Kwok Cheung, Rui Zhang, Milos Manic, Power Zhao, Song Zhang, Xiaping Zhang

June 2019

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

IET Research Journals

IET Journals
The Institution of Engineering and Technology

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Big Data Analytics in Smart Grids: State-of-the-Art, Challenges, Opportunities, and Future Directions

Bishnu P. Bhattarai [1*], *Sumit Paudyal* [2], *Yusheng Luo* [3], *Manish Mohanpurkar* [3], *Kwok Cheung* [4], *Reinaldo Tonkoski* [5], *Rob Hovsapian* [3], *Kurt S. Myers* [3], *Rui Zhang* [6], *Power Zhao* [7], *Milos Manic* [8], *Song Zhang* [9], *Xiaping Zhang* [10]

[1] *Pacific Northwest National Laboratory, USA*
[2] *Michigan Technological University, USA*
[3] *Idaho National Laboratory, USA*
[4] *GE Grid Solutions, USA*
[5] *South Dakota State University, USA*
[6] *IBM Research - Almaden, USA*
[7] *Oncor Electric Delivery, USA*
[8] *Virginia Commonwealth University, USA*
[9] *Independent System Operator New England, USA*
[10] *California Independent System Operator, USA*
**Bishnu.Bhattarai@pnnl.gov*

Abstract: Big data has a potential to unlock novel groundbreaking opportunities in the power grid sector that enhances a multitude of technical, social, and economic gains. The currently untapped potential of applying the science of big data for better planning and operation of the power grid is a very challenging task and needs significant efforts all-around. As power grid technologies evolve in conjunction with measurement and communication technologies, this results in unprecedented amount of heterogeneous big data sets from diverse sources. In particular, computational complexity, data security, and operational integration of big data into power system planning and operational decision frameworks are the key challenges to transform the heterogeneous large dataset into actionable outcomes. Moreover, due to the complex nature of power grids along with the need to balance power in real time, seamless integration of big data into power system planning and operations is very critical. In this context, big data analytics combined with grid visualization can lead to better situational awareness and predictive decisions. This paper presents a comprehensive state-of-the-art review of big data analytics and its applications in power grids, and also identifies challenges and opportunities from utility, industry, and research perspectives. The paper analyzes research gaps and presents insights on future research directions to integrate big data analytics into power system planning and operational decision framework. Detailed information for utilities looking to apply big data analytics and details insights on how utilities can enhance revenue streams and bring disruptive innovation in the industry are discussed. More importantly, general guidelines for utilities to make the right investment in the adoption of big data analytics by unveiling interdependencies among critical infrastructures and operations are also provided.

1 Introduction

Over the past few years, the adoption of big data analytics in banking [1, 2], health care [3, 4], internet of things (IoT) [5, 6], communication [7, 8], smart cities [9, 10], and transportation [11] sectors have demonstrated huge potential for innovation and business growth. The transition of power grids to 'smart grids' around the world can be characterized with larger datasets being generated at an unprecedented rate with localized integration, controls, and applications. It is highly anticipated that there is a great potential for the application of big data to the current and future power grids [12]. Currently, power grids incorporate all sorts of innovations in measurement, control, communication, and information science to effectively operate electric power systems that deliver affordable, reliable, sustainable, and quality energy to end users. Power grids around the world are also deploying a massive advanced metering infrastructure (AMI) and measurement technologies such as smart meters and phasor measurement units (PMUs) to collect system-wide high-resolution electrical measurements [13–15]. These electrical data comprising of measurements, along with other non-electrical data (e.g., weather, traffic, etc.), if effectively utilized in coordination, will revolutionize the operation of electric power grids. The effective utilization of data enhances observability of power grids that includes system-wide grid conditions, behavior of end users, and renewable energy availability—all crucial information for reliable and economic operation of the electric power grids.

Increased deployment of the measurement devices along with model based data (e.g., simulations) and data from non-electrical sources are resulting in unprecedented amount of widely varying data in electric power grid [16]. A typical distribution utility deals with thousands of terrabytes (TB) of new data every year [17]. As shown in Fig. 1, these data come from various sources including smart meters, PMUs, μPMUs, field measurement devices, remote terminal units (RTUs), smart plugs, programmable thermostats, smart appliances, sensors installed on grid-level equipment (e.g., transformers, network switches), asset inventory, supervisory control and data acquisition (SCADA) system, geographic information system (GIS), weather information, traffic information, and social media [17].

Big data in smart grids are heterogeneous, with varying resolution, mostly asynchronous, and are stored in different formats (raw or processed) at various locations. For example, typical smart meter data are energy consumption collected every 15 minutes and

Fig. 1: Sources of non-electrical and electrical big dataset in smart grids.

are stored in billing centers. One million smart meters installed in a utility results is nearly 3 TB of new energy consumption data every year. Whereas PMUs measure high-resolution voltage and current in the power grid and report at a 30-60 times per second rate as time-synchronized phasors to phasor data concentrators (PDCs) located at the sub-station level or at control centers. PMUs result in nearly 40 TB of new data per year for a typical utility [17]. These big data carry considerable amount of information that enables novel information-driven control algorithms. This in turn can bring revolutionary transformations to the ways grids are planned and operated [18, 19]. Big data in smart grids allows improvisation in existing operation and planning practices at all levels, i.e., generation, transmission, distribution, and end users [17–20]. It enables new opportunities in controlling the grid assets, distributed energy resources (DERs), and end users' energy consumption holistically in real time, which were not possible in conventional grids due to limited measurement and control capabilities.

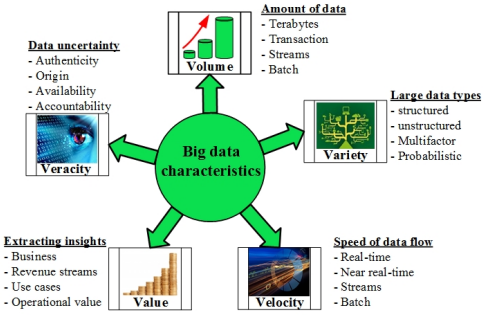

Fig. 2: Key characteristics of smart grid big data.

As shown in Fig. 2, big data in smart grid are characterized by high volume (in the order of thousands of terrabytes), wide varieties (structured/unstructured, synchronous/asynchronous), varying velocity (e.g., real-time, second/minute/hour resolutions), veracity (inconsistencies, redundancies, missing data, malicious information), and values (e.g., technical, operational, economic) [21, 22]. As such, it becomes necessary to process large volume and varieties of both real-time and historical data to extract meaningful information in order to make data-driven decisions [16]. Therefore, big data analytics will play a critical role not only for the efficient operation of future electric grids, but also for the development of proper business models for the key stakeholders (e.g., electric utilities, system operators, consumers, aggregators) [23, 24].

Mega-corporations such as Google, Microsoft, Amazon have matured data-mining and processing tools that allow for quick and easy processing of large amounts of data for a wide variety of applications [25]. Therefore, data organizing and storage are typically well established in a generic sense. However, big data analytics is more than just the data management; it is rather an operational integration of big data analytics into power system decision-making frameworks [26]. Therefore, the key challenge of big data analytics is to turn large volume of raw data into actionable information by effectively integrating into power system operational decision frameworks [27]. Efficient deployment of big data into electric utilities plannign and operation can lead to multiple benefits including improved reliability and resiliency, optimized resource management/operations, improved operational decision, and increased economic benefits to customers, utility, and the system operators [28]. As smart grid data increases exponentially in the future, utilities must envision ever-increasing challenges on data storage, data processing, and data analytics. Even though many electric utilities have realized that deployment of big data analytics is a must and not a choice, for future business growth and efficient operation, implementation of big data analytics in utility framework is lagging [29]. Therefore, there is a need of comprehensive study to investigate current challenges, value proposition to stakeholders (e.g., consumers, utilities, system operators), operational benefits, and potential path forward to deploy big data analytics in power grids [30].

This paper presents insights on big data in smart grid from several different perspectives - research, electric utilities, and industries perspectives. First, we identify current challenges to transform big data in smart grid into actionable information, and then present future directions for its operational integration into utility decision frameworks. In fact, detailed insights to tap currently hidden potential of big data analytics to benefit utility customers, electric utilities, and system operators are presented. Therefore, this study details information and factors to consider for electric utilities and system operators looking to apply big data analytics and provides insights on how utilities can deploy big data analytics to realize increased revenues and operational benefits. Furthermore, this paper provides insights on how effective integration of big data analytics to utility decision frameworks helps to make right decision at right time and location by unveiling the interdependencies among various critical infrastructures and operations.

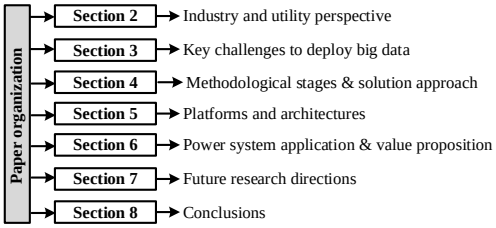

Fig. 3: Overall organization of the paper.

The remainder of the paper is structured as depicted in Fig. 3. First, comprehensive analysis of the big data from utility and industry perspectives is presented in Section 2. Next, in Section 3, key challenges for integration of big data to smart grids are detailed. Potential solutions and methods of big data analytics are detailed in Section 4. Section 5 presents existing big data analytics architectures and platforms suitable for smart grid applications. Next in Section 6, key power system application areas of big data analytics are detailed. Finally, future research directions for big data application to smart grids are presented in Section 7, and the paper is concluded in Section 8.

2 Utility and Industry Perspectives

As depicted in Fig. 1, smart grid is associated with vast amount of data from various sources, including power system operation

(generation, transmission and distribution, customers, services and markets), energy commodity markets (electricity markets, gas, and oil), environment, and weather. Those data are characterized by diversity of its sources, growth rate, spatio-temporal resolutions, and huge volume. It is anticipated that future power grids will generate heterogeneous data at a higher rate than ever. On the one hand, these vast amount of data create several challenges for data handling, processing, and integration to utility decision framework. On the other hand, these large datasets provide significant opportunities for better monitoring, control, and operation of electric grids. In particular, this can help electric utilities to make the system more reliable, resilient, and efficient. Therefore, big data analytics is perceived as a foundation to optimize all current and future smart grid technologies.

2.1 *Electric Utility Perspective*

Electric utility is a very complex structure having close dependencies and interactions among communications, IoT, and human factors [31]. Recent concerns on increased security and reliability of critical infrastructure are leading to the need of integrated energy system, which integrates various critical infrastructure, including electrical, gas, thermal, and transportation [32–34]. Therefore, future power grid management systems will be processing overwhelming amounts of heterogeneous data [35]. As illustrated in Fig. 4, individual devices and functional units can generate thousands of TB data annually. Considering large number of such units (e.g., consumer, sensors, substation) and grid functions (e.g., home energy management, distribution management, DER management), electric utilities have to handle millions of TB data, which continues to increase over time. Therefore, utilities must take a deep dive into what increasing data means to their traditional operations, and have to make necessary strategies to create value from those vast amount of data [36].

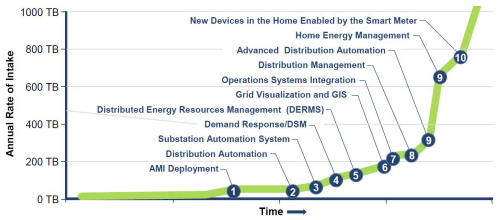

Fig. 4: Pattern of big data volume in electric utilities [35].

A recent survey conducted with 1,000 electric utility and industry respondents across 10 countries depicted that majority (80%) of the electric utilities realize big data analytics as crucial for future smart grid and source of new business opportunities [37]. Recently, Canadian Electric Association has also identified big data as one of the key drivers for grid modernization to meet their 2050 vision [38]. In addition, Canada has initiated a concept of open data set among multiple utilities and service provides in seven Canadian cities in an effort to maximize the value of big data [39]. However, even though utilities recognize big data analytics as an unavoidable task for the future power grids, electric utilities are still reluctant for its implementations. Fig. 5 illustrates an overview of current status of electric utilities in terms of big data implementations [22]. It can be observed that only 20% of the utilities have implemented big data analytics to some extent. However, it is worth mentioning that even those 20% utilities who have implemented big data are tapping only a fraction of potential [37].

In addition, as electric utilities are heavily regulated organizations, they are more focused on system reliability rather than trying a new technology; therefore, they are somewhat reluctant to the implementation of big data analytics. As depicted in Fig. 6, lack of management support, skill shortage, data management issues, and

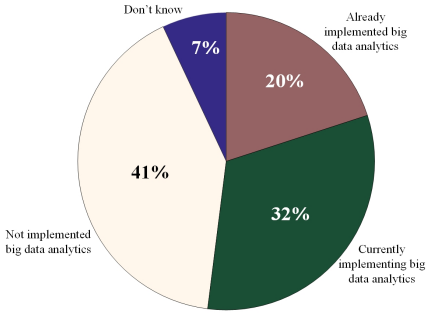

Fig. 5: Current utility status of big data implementations [37].

lack of proper business models are primary factors that are holding the utilities back from the deployment of big data analytics. However, it should be noted that data storage and data management challenges have successfully been addressed in other industries (e.g., banking, IoT). Operational integration of big data to utility decision framework and its value proposition to different stakeholders (e.g., utilities, system operators, aggregators, consumers) and professional training are the key challenges to be considered.

Increasing need of improved reliability and resiliency of the system and tighter boundaries from regulating entities are also steadily forcing the utilities to deploy big data analytics [34]. With big data analytics, electric utilities can exploit behind the meter resources and obtain various grid services at lower cost. More importantly, big data analytics help to reduce levelized cost of electricity (LCOE) not only by helping to make better investment decision at the right time and right place, but also by unveiling insights and value proposition of additional revenue streams (e.g., better participation to energy/power markets, grid services). Therefore, similar to disruptive innovation that big data analytics brought to other industries, it can transform utility industry by expanding business volume and revenue streams.

2.2 *Industry Perspective*

Even though the information technology related companies have achieved substantial success in the field of big data analytics, electrical industries are at the beginning stage to deploy big data. A few industries including Siemens, GE, ABB, OSI-Soft, and so on are developing big data platform and analytics for power grids. An account of a few commercially available platforms is provided here as a sample only and by no means is intended to be exhaustive. Siemens has developed a big data platform, called EnergyIP Analytics, which adds big data to smart grid application and provides insights on management of big data for providing various grid services to electric utilities and grid operators [41]. Siemens is currently integrating utility operations and data management technologies that

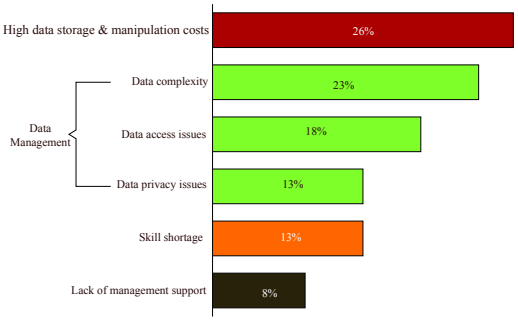

Fig. 6: Current identified barriers for big data implementations [37].

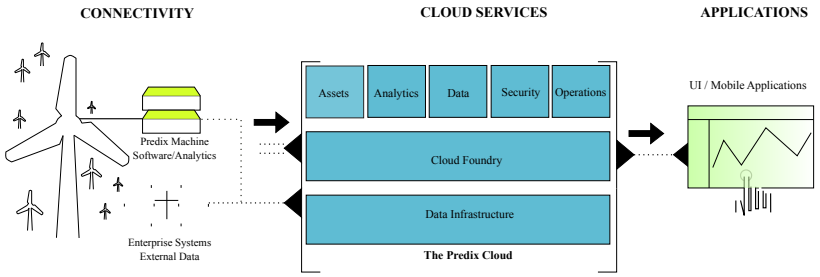

Fig. 7: High-level overview of GE big data analytics platform [40].

could potentially be tapped for grid data analytics. This grid analytics platform can allow utilities to utilize big data for multiple functionalities, including home energy management, grid energy management, and predictive/corrective controls [41]. EnergyIP Analytics has already been used by more than 50 utilities with a total of 28 millions installed smart devices [42].

Similarly, GE has developed an industrial IoT platform, called PREDIX, to consolidate data from existing grid management systems, smart meters, and grid sensors [40]. In addition, Grid IQ Insight, a cloud based big data analytics architecture which utilizes PREDIX platform, is developed to integrate big data analytics to grid applications [43]. Native data collected from Grid IQ Insight are stored in datalake that could be tapped for several grid applications [44]. In fact, these initiatives and developments support multiple grid applications ranging from real-time grid monitoring, distribution automation, home energy management, and ancillary services. As illustrated in Fig. 7, a concept of edge computing, whereby computational intelligence is connected at the edge of the data source, has introduced by GE in its Grid IQ Insight.

ABB is integrating cloud computing and big data analytics intended for future power grid applications. ABB has developed an intelligent big data platform, called ABB Asset Health Center, which provides solutions for processing big data for smart grid applications [45]. In fact, ABB's Asset Health Center embed equipment monitoring and systems expertise to establish end-to-end asset management, business processes for reducing costs, minimizing risks, improving reliability, and optimizing operations across the electric utility [45]. In addition, OSI-Soft PI system, which is one of the most widely deployed database and analytics system, has been contributing to unveil the power of big data analytics to electric utilities. Smart asset management platform has introduced by OSI-Soft for the purpose of real-time monitoring of asset health [46].

The aforementioned industries are offering utilities a way to gain a core understanding of what is the state of grid devices, and developing a launching pad for smart grid big data analytics applications over time. The next step for the industries is to effectively integrate prognosis and diagnosis into big data analytics framework so as to facilitate utilities to provide situational awareness, informed predictive decisions, condition monitoring, health management of critical grid infrastructure, and supporting grid functionalities.

3 Key Challenges for Big Data Analytics

This section presents key challenges in deploying big data analytics to future power grids.

3.1 Data Volume

The amount of data being generated by electric utilities is increasing at an exponential rate. Therefore, big data challenges, such as data storage, data mining, data processing, data querying, data indexing will increase in unprecedented manner in the future. Due to increased deployment of intelligent devices in consumer and their active engagement on different grid services, the data management expands also to the consumer level. Even at the consumer levels, data volume from various devices (e.g., smart meter, electric vehicles, inverters) will be in the order of hundreds of TB [35]. Therefore, effective management of huge volume of data is becoming increasingly challenging issues for utilities. New innovative solutions, such as distributed and scalable computing architecture are necessary [47, 48]. Moreover, dimensionality reduction, a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data, can significantly reduce data complexities [49]. Table 1 summarizes the key challenges, potential impacts, and potential solutions of deploying big data to power grid.

3.2 Data Uncertainty

Data uncertainty is one of the defining characteristics of real-world smart grid data and it stems basically from lack of data or an incomplete understanding of the operational context. Since data quality, which is attributed by accuracy, completeness, and consistency of data, is one of the biggest concerns in smart grid; the quality of utility decision depends entirely on the quality of data. However, since real-world data are highly susceptible to errors due to noises and missing/inconsistent data, data cannot be acquired with 100% certainty. Major causes of data uncertainties and loss of data quality stem from sensor inaccuracies and imprecision, communication latencies/delays, cyber-attack, physical damages of equipment, time unsynchronized data, missing/inconsistent data, noises, etc. Those uncertainties may result from various reasons, for instance, readings of sensors are uncertain because of sensor aging or malicious attacks during data acquisition and control processes. This requires innovative techniques to deal with data mining and data analytics techniques [50]. Probabilistic data analytics and data mining, whereby data uncertainties are modeled as a stochastic process within certain limits, are recently been deployed to deal with data uncertainties in [51]. Similarly, data preprocessing techniques (e.g., data cleaning, data integrity, data conditioning) are often used for identifying and removing noisy data, filling in missing values, resolving redundancies, correcting inconsistencies, and smoothing out noises and outliers [52]. Data cleaning deals with the missing values, smooths out noises, identifies outliers, and corrects inconsistencies within the data.

3.3 Data Security

Smart grid data mostly involve consumer privacy information, commercial secrets, and financial transactions. Therefore, data security (e.g., privacy, integrity, authentication) are very crucial [67].

3.3.1 Data Privacy: Data privacy of user is a very critical security concern as the power consumption of consumer normally provides insights on their behavior [68]. Data aggregation is one of the common approach to address data privacy issues. Different techniques such as distributed aggregation [53], differential aggregation [54], and aggregating with storage [55] are recently developed to address data privacy issues.

3.3.2 Data Integrity: Data integrity is primarily used to prevent unauthorized modification of information. However, due to close interdependencies between power and communication infrastructure, the power industry is also susceptible to increased

Table 1 Summary of Key Challenges to Apply Big Data to Smart Grid

Challenges	Possible Impact	Potential Solution
Data Volume	Need of increased storage and computing resources	Dimensionality reduction, Parallel computing, Edge computing, Cloud computing, pay-per use [47–49]
Data Quality	Lack of complete inofrmation, misleading decision	Probabilistic and storchastic analysis [50, 51], data cleaning (e.g., dealing with missing values, smooth out noises, outliers, and inconsistent data) [52]
Data Security	Vulnerable to malacious attack, compromise consumer privacy and integrity, mislead operational decision and financial transactions	Data anonymization (e.g., data aggregation [53–55], data encryption [56–58], P2DA [59])
Time Synchronization	Mislead operational decision, wrong interpretation of data, bad diagnostic of past events	Synchronize devices based on same radio clocks or satellite receivers [60, 61]
Data Indexing	Computational complexity, long processing time	Deploy new indexing techniques such as R-trees, B-trees, Quad-trees [62–66]
Value Proposition	non-acceptance by stakeholder, delay deployment of big data,	Quantifying both technical and economic values to key stakeholders, namely consumer, system operator, utility.
Standards and Regulation	Interface challenges among various computing, storage, and processing platforms, delayed deployment	Regulatory entity define guidelines about data sharing/exchange, and standards should technically ensure regulatory aspects.

cyber/physical-attacks [69]. Those integrity attacks not only deliberately modify financial transactions, but also severely mislead the utility operational decisions [70]. Privacy-preserving data aggregation (P2DA) scheme can ensure data integrity through a digital signature or a message authentication code [59].

3.3.3 Data Authentication: Smart grid data requires authentication as a basis to distinguish legitimate and illegitimate identity. Data authentication is not only necessary to preserve user privacy, but also to ensure data integrity [56]. Therefore, authentication including encryption, trust management, and intrusion detection are important security mechanisms that can prevent, detect, and mitigate network attacks [57]. Different techniques such as data encryption and signature generation are normally used for data authentication and security management in smart grids [58].

3.4 Time Synchronization

With the increasing need of real-time control and communication in smart grid, time synchronization is becoming a key concern. Currently, synchrophasors or PMUs provide time synchronized data, which utilize synchronization based on radio clocks or satellite receivers. Time synchronized data allows analysts to draw meaningful connections between events and aids both forensic analysis of past events, near real-time situational awareness, and informed predictive decisions [60]. Forensic determination of a sequence of past events (e.g., what actually tripped, what was the initiating event) and real-time situational awareness of the grid's health can be very powerful to provide preventive or remedial solution. However, communication, storage, and analysis of streams of data from most of the distribution system devices and customers are currently unsynchronized. As unsynchronized data poses potential risk of misleading decision, data should be time synchronized with respect to same time reference.

3.5 Data Indexing

The smart grid data also posses issues on data indexing and query processing. The existing methods use generic tools such as SQL server and SAP for query purposes; however, these may not suffice from smart grid application point of view, particularly if real-time applications are sought from the big data. Therefore, advanced data indexing and query-processing algorithms will play critical roles in smart grid big data analytics. State-of-the-art data indexing techniques including variants of R-trees, B-trees, and Quad-trees would definitely be useful for efficiently indexing the big data in smart grids [62–66].

3.6 Standards and Regulation

There are a few standards information models and communication protocols (e.g., IEC 61850, IEC 61850-90-7, IEC 61970/61968, IEEE 1815, IEEE 2030.5) for smart grid interoperability [71]. However, none of the efforts are being yet made on interoperability among big data analytics platforms, architectures, and grid operations frameworks. Instead, different utilities are implementing big data analytics with different storage, computing, processing platforms. Such diversified use of protocols, architectures, and platforms for big data analytics will not only limit its potential, but also delay the adoption of big data analytics to power grid [8]. Therefore, to take full advantage of big data application to smart grid, there is need of data sharing and information exchange among different utilities and system operators. Since electric utilities usually do not share data/information with each other, regulatory framework should be established to facilitate data sharing and unify their efforts. In order to synchronize the efforts from utility, industry, and academia, there is a strong need to build standards for big data analytics architecture, platforms, and interoperability.

3.7 Business Models and Value Proposition

To successfully deploy big data analytics in smart grids, proper business models should be developed [25]. Even though other industries (e.g., Google, Facebook, Amazon) disruptively transformed their business via big data analytics, electric utilities are still in the initial stage. The business models should be justified on the basis of market opportunity/volume, required investment, and values to different stakeholders. Recent research has estimated the value of the global utility data analytics market at a cumulative $20 billion between 2013 and 2020, growing to nearly $4 billion a year by 2020 [72]. This shows huge market potentials for big data analytics to electric utilities.

As shown in Fig. 8, the Utility Analytics Institute has predicted that data-related costs are continuously decreasing. Over the past 30 years, the cost to store data has been cut in half every 14 months or so [72]. For instance storing a gigabyte of data in 1995 cost about $11,200, by 2000 it was $11, and today costs mere three cents [72]. The falling costs of data storage and data management is making the real-time data collection and storing economically feasible, thereby providing significant opportunities for utilities to make successful business models. However, utilities require a clear understanding of where long term economic and technical values of big data lie, and should develop proper business models for all stakeholders, including utilities, system operators, and customers.

Fig. 8: Global utility analytics spending [72].

4 Big Data Stages and Solution Approaches

Organizing and storing big data in general is well understood. Mega-corporations (e.g., Google, Microsoft, Amazon) have mature data-mining and processing tools that allow quick and easy processing of large amounts of data. However, data management is more than just the technical challenges of data handling. Instead, data analytics should be effectively integrated into utility strategies, operational frameworks, and decision-making process. The following sections detail methodological stages and solution approaches for big data analytics.

4.1 Big Data Methodological Stages

As shown in Fig. 9, key steps for big data analyses include data acquisition, data storage, data analytics, and operational integration, which are described in detail in the following subsections.

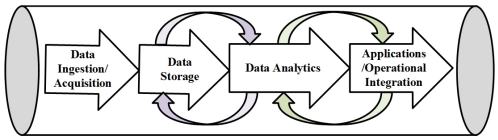

Fig. 9: Key stages of big data analytics.

4.1.1 Data Acquisition: Data acquisition primarily deals with collection of data from multiple heterogeneous sources in different formats and features. Since power grid data often contains private information and personal behaviors of consumers, data confidentiality and security are critical aspects within the data accessing and transmitting. In order to ensure the data confidentiality and security during data acquisition, data encryption-decryption, and aggregation-disaggregation approaches are generally employed [73]. Those approaches preserve the sensitivity and privacy within the data and often restrict unauthorized access of data [74].

4.1.2 Data Storage: Data storage primarily belongs to data management (e.g., data fusion, data integration, data transforming) within data repositories [75]. Data storage not only need to manage large amount of widely varying data collected in different forms/formats, but also have to deliver data to multiple analytics platforms having different requirements (e.g., temporal/spatial resolutions, formats). Recently, data-centric storing and routing technologies have widely been employed for big data storage, whereby data is defined and routed referring to their names instead of the storage node's address [76]. Each data object has an associated key and each working node stores a group of keys. This makes the data storage flexible and scalable. A novel approach for effective storage of time-series data is proposed to reduce the computation expense [77].

4.1.3 Data Analytics: Data analytics is designed to identify hidden and potentially useful information and patterns within large dataset that can be transformed into an actionable outcomes/knowledge. It utilizes various algorithms and procedures (e.g., clustering, correlation, classification, categorization, regression, feature extraction) to extract valuable information from the dataset [78–81]. Depending on the potential use cases, data analytics involves one or more of the descriptive, diagnostic, predictive, and prescriptive analytics. As shown in Fig. 10, descriptive models are often used to describe operational behaviors of grid and customers, whereas diagnostic models analyze the operating conditions and decisions made by the grid operators. The diagnostic model is focused on identifying the causes for an event, thereby is suitable for taking remedial action. As the key objective of data analytics is to provide preventive solution, predictive models are often necessary to forecast operating conditions and future decisions [82]. Prescriptive analysis, on the other hand, are designed for providing longer term insights to utilities in making strategic operational and investment planning. Please note that Section 6 provides details of potential applications of big data in various smart grid and power system applications. Therefore, the following paragraphs provide a brief overview of key smart grid applications corresponding only to selective data analytics techniques.

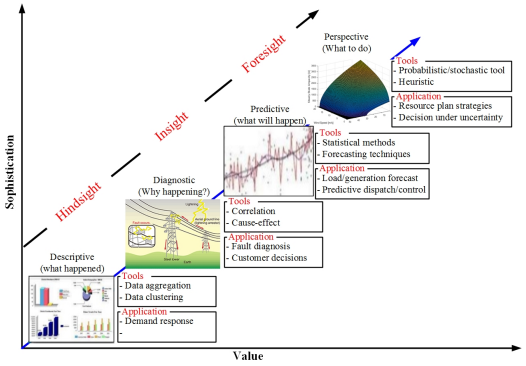

Fig. 10: Summary of different analytics techniques and their key applications.

From the smart grid application perspective, data analytics can be categorized into four broad categories as illustrated in Figure 11. Event analytics primarily covers diagnosis/detection of the power systems events such as faults and outage managements [83–89]. In addition, the event analytics also encompass descriptive analysis of prior power system events using various techniques (e.g., classification, filtering, correlation) [83–85, 89]. Detection of abnormal operating conditions including fault detection [83–85], system outage detection [86–88], detection of malicious attacks [84], and theft of electricity [89] are some of the key application areas for event analytics.

State and operational analytics primarily include a combination of diagnostic, predictive, and perspective analytics. As illustrated in Figure 11, the key power system application of the state analytics includes state estimation [83, 90], system identification [86, 91, 92], and grid topology identifications [90, 93–96]. Similarly, the key power system applications for operational analytics include energy/load forecast [97–99], energy management and dispatch of resources [87, 88, 96, 100]. Similarly, customer analytics also includes one or more of the descriptive, diagnostic, predictive, and perspective analytics depending on the specific applications and use cases. The key power system application that falls under the customer anlaytics include customer classification/categorization [95, 101], correlation between consumer behavior and energy consumption patterns [89, 97, 99, 102], and demand response [100, 103].

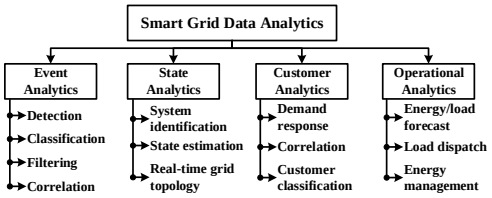

Fig. 11: Application specific analytics applied to smart grid.

Please note that data correlation, data classification/categorization, and pattern recognition are commonly used algorithms for the aforementioned smart grid analytics (as shown in Fig. 11). The following section briefly highlights those algorithms.

Data Correlation: Correlation is a well-known statistical technique to determine relationship and compatibility among different datasets. As smart grid data are closely related to various factors (e.g., grid events/disturbances, weather, grid operations, electricity prices), correlation analysis provides key insights on data and their interdependencies [104]. Conventionally, data correlation has widely been used for forecasting and planning of power systems. However, with the emergence of big data, the correlation analysis has been focused on big data domain as well [105, 106].

Data Classification and Categorization: Data classification is the process of organizing data into meaningful categories so as to make it easy to find and retrieve information. In smart grid, data are normally categorized on the basis of time, importance, and privacy requirement [107]. Artificial neural network and self organizing mapping are most commonly used models for data classification and categorization in smart grid big data [108]. In addition, K-means, hierarchical clustering, Fuzzy C-means are often implemented for data categorization [109, 110].

Feature Extraction: Feature extraction is one of the important step of data mining that is intended not only to translate data into meaningful outcomes, but also to identify the data attributes affecting those features [111]. As large volumes of data from the sensors and intelligent devices installed around the smart grid often contain noise, incompleteness, and redundancies, feature extraction play critical roles [112].

4.2 Potential Solutions for Big Data Analytics

Due to large volume and variety of data in smart grid, acquiring and processing all data is technically inefficient from cost, complexity, and storage requirements. The following approaches are designed to make the big data analytics efficient and effective for smart grid applications.

4.2.1 Dimensionality Reduction: Dimensionality reduction is one of the effective techniques used to provide a reduced and representative version of large dataset [47, 113]. Key challenge is to find the optimum reduction on dataset that can provide the same information as the original dataset [113]. Some literature has proposed online dimensionality reduction on synchrophasor measurements using random projection approach [114]. Even though the random projection is simple, scalable, and provides faster execution, it has not been sufficiently explored in power systems.

4.2.2 Distributed and Edge Computing: Conventional power system utilizes a centralized architecture for data acquisition, analyzing, and processing. Such framework requires huge exchange of information flow among various intelligent devices within the smart grid [49]. This is inefficient not only from communication perspective, but also from data storage, security, and data handling perspectives. Therefore, the future power grid should implement distributed computing and data mining architecture to reduce the computational burden at the centralized processor [115]. Recently, edge computing, a method of optimizing computing performance by processing data at the edge of the network near the data source, has been gaining attention in big data applications [7, 116]. Edge computing primarily relives the communication bandwidth needed between the data source and central processing system, whereas distributed computing reduces data handling burdens by parallel processing of the information [117, 118]. Recently, some literature presents distributed data analysis and control techniques for various applications including load prediction and volt-var control [115, 119]. Distributed and edge computing make the solution scalable, less affected by peer failures, require less computational burden, and reduced communication resources [6, 76].

4.2.3 High Performance Computing: Modern electric grids require real-time monitoring, control, and operation of large number of resources. As most of the real-time operation and control applications require fast data processing, we need high performance computing (HPC) to be able to integrate big data analytics to utility control and operation [120]. Even though the computational capacity of the HPC has increased significantly in the past few years, HPC based computation is still not economically viable to several applications [121]. As such, data analytics based on task parallelism can provide economic and efficient solutions for power system computational issues [122].

4.2.4 Cloud Computing: Cloud computing approach is a promising solution for computation intensive grid applications because it uses computational resources based on demand [123]. Cloud computing has distinct advantages, such as scalability, flexibility, distributed computing, parallelization, fast retrieval of information, interoperability, virtuality, and extensibility. Recently, cloud computing has been applied to energy conscious scheduling in smart grid [124–126]. ISO New England has successfully deployed this concept on Amazon Web Services [127]. The deployment of cloud computing to smart grid brings several benefits, including increased fault tolerance and security due to multi-location data backup [128]. Moreover, the cloud computing helps utilities to realize flexibility, agility, and efficiency in terms of saving cost, energy, and resources [29]. Many smart grid applications, including advanced metering infrastructure, SCADA, energy management system, and distribution management systems, can be greatly benefited by application of cloud computing approach.

4.2.5 Metamodeling: The increase in complexity of large scale simulation models often lead to increased run times. Consequently, the simulation of large interconnected networks can benefit from simulation metamodeling to reduce the runtime with acceptable accuracy. Simulation metamodeling is to build a model of a simulation models in order to reduce the run times. The suitability of the model is evaluated based on the required computational expense, reliability, and accuracy. Typically this evaluation uses Bootstrap error and the predicted residual error sum of squares statistic to efficiently compute the standard error and bias [129]. The implementation of such algorithms and the software environment are extremely important to develop computational efficient and accurate models [129]. Metamodels can be applied for energy and market forecast in power systems, and smart grid simulations [130–132].

5 Big Data Architecture and Platforms

This section describes common architecture and platforms for big data analytics, and presents insights on application of those architectures and platforms in power systems.

5.1 Big Data Architecture

Currently, there are no standard big data analytics architectures developed for power grid applications [133]. Therefore, clear understanding of big data architecture is required to identify how big data integrates with the existing power system control and operational architecture, what are the essential characteristics of big data environment, how they differ from traditional computational environments, and what scientific, technological and standardization challenges are needed to deploy big data solutions [134]. The following subsections describe common big data analytics architectures.

5.1.1 General Electric Grid IQ Insight Architecture: Grid IQ insight is a big data analytics architecture which works based on the foundation of PREDIX data analytics platform. In fact, PREDIX is an industrial IoT based platform [40] which was developed for variety of applications including power system. Grid IQ Insight is a cloud based horizontal architecture consisting of four layers as shown in Fig. 12. The bottom most layer is basically a physical layer which consists of utility assets, operational systems, and external data, whereas the second layer is primarily a cloud based API and utility specific data layer (e.g., analytics, dashboards). The third layer primarily includes grid applications, while the forth layer focuses on the visualization and operational integrations.

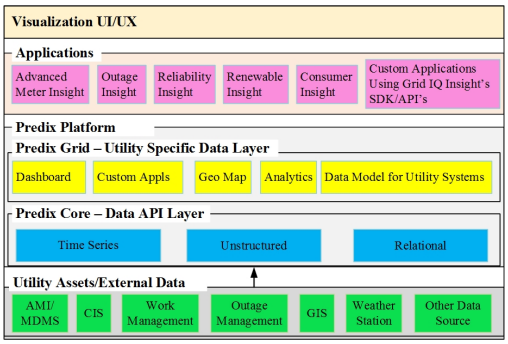

Fig. 12: GE Predix platform for big data analytics [40].

5.1.2 Booz Allen Hamilton Architecture: The Booz Allen Hamilton is a cloud based horizontal reference architecture consisting of four layers [135]. The top layer deals with the human insights and action, and establishes data interfaces and visualizations, whereas the second layer is designed for analytics and services, thereby consists of tools and algorithms required for modeling, analysis, and simulations of data. The third layer deals with data management and is designed to deal with all heterogeneous data sources. The bottom layer is infrastructure layer which stores and manages smart grid data.

5.1.3 IBM Big Data Architecture: IBM big data architecture is a four vertical layered reference architecture, where the left most layer deals with data sources, and the second layer consists of big data platforms and capabilities [136]. Similarly, the third layer deals with data analytics and customer insights, and the last layer is designed to integrate data analytics results for various operations. Lockheed Martin energy data analytics architecture, as shown in Fig. 13, is an example of IBM vertical reference architecture. However, unlike the case of IBM architecture, the Lockheed Martin architecture has only three layers.

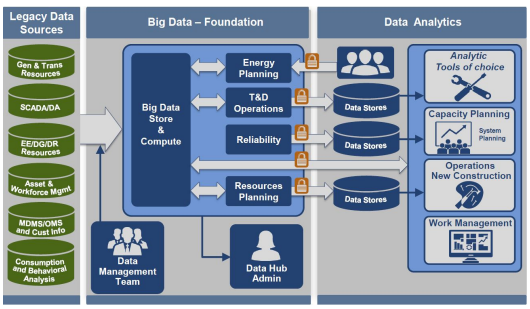

Fig. 13: IBM based Lockheed Martin big data analytics architecture for smart grid applications [35].

5.1.4 SAP Big Data Architecture: This is a combination of horizontal and vertical reference architectures developed by SAP [134]. As shown in Fig. 14, vertical layers include data sources and data ingestion, while horizontal layers include applications, real-time data accelerated analytics, and data management (e.g., storage, data processing and deep analytics).

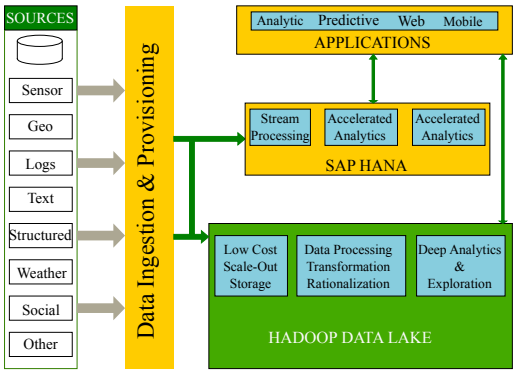

Fig. 14: SAP reference architecture for big data processing [134].

5.1.5 ORACLE Big Data Architecture: As shown in Fig. 15, Oracle big data architecture also consists of horizontal as well as vertical layers. The vertical layers include data sources, data acquisition, data organization (to ensure data quality for analytical operations), data analytics, decision making (recommendation, alerts, dashboards), and data management (e.g., storage, data security, governance) [137]. Similarly, horizontal layers include technology platforms and integration layers for operational integration to electric utility operational framework.

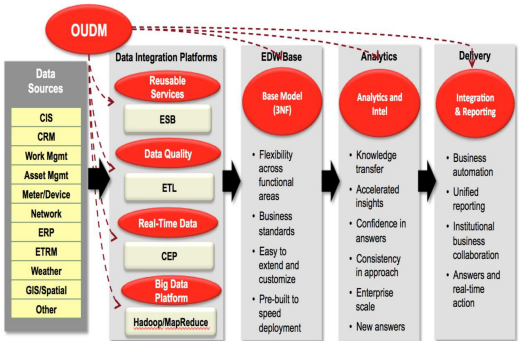

Fig. 15: Oracle big data analytics reference architecture [137].

It is worth mentioning that there are several other architectures (e.g., Big Data Ecosystem Reference Architecture, GPUMKLIB big data Architecture Framework, National Big Data Reference Architecture) developed for big data analytics in IoT sectors. Different variations of those reference architectures are being implemented on power industry, however, standard big data architectures for power grids have not yet been developed.

5.2 Big Data Platforms

The following subsections presents commonly used platforms for big data analytics and compare their performance in Table 2.

Table 2 Comparison of various big data analytics platforms.

Platform	Data Scaling	Scalability	Fault Tolrance	I/O Performance	Application
Hadoop	Horizontal	Yes	Yes	Limited	Batch Processing [138–140]
Spark	Horizontal	Yes	Yes	Moderate	Batch and real-time processing [141]
Storm	Horizontal	Yes	Yes	Moderate	Real-time processing [141]
Drill	Horizontal	Yes	Yes	Good	Interactive analytics [142]
HPC	Vertical	Limited	Yes	Very good	Batch, stream, and interactive [121, 122]

5.2.1 Hadoop: Apache Hadoop is an open source framework for storing and processing large datasets using MapReduce programming model [143]. The Hadoop consists of storage part (known as hadoop distributed file systems (HDFS)) and processing part (known as MapReduce programming model) [138–140]. Primarily, Hadoop splits files into large blocks and distributes them across nodes so as to process data in parallel. Due to distributed storage structure, HDFS not only ensures high availability, but also high fault tolerance against hardware failures. OSI-Soft, which is one of the widely used database and data analytics platform in electric utility, uses Hadoop for performing data analytics in PI system.

5.2.2 Spark: Spark is a fast, in-memory, open-source big data processing engine which is designed to overcome the disk I/O limitations of Hadoop [144]. Spark can perform in-memory computations and allow the data to be cached in memory, thereby eliminating the Hadoop's disk overhead limitation for iterative tasks [141]. Spark is a general engine for large-scale data processing which is up to 100 times faster than Hadoop MapReduce when the data can fit in the memory and up to 10 times faster when data resides on the disk.

5.2.3 STORM: Apache Storm is also an open source distributed real-time computation system, that can reliably process unbounded streams of data [141, 145]. It is scalable, fault-tolerant, and easy to set up and operate, thereby having several use cases, including real-time analytics, online machine learning, and real-time computation.

5.2.4 Apache Drill: Apache Drill is an open source software framework that supports data-intensive distributed applications for interactive analysis of large-scale datasets [142]. Drills is able to scale 10,000+ servers and process petabytes of data and trillions of records within seconds. In addition, Drill can discover schemas on-the-fly, thereby delivering self-service data exploration capabilities on data stored in multiple formats in files or databases. Drill can seamlessly integrate with several visualization tools, thereby making big-data platform interactive.

5.2.5 High Performance Computing: HPC is a vertical scale up platform for big data processing which consists of a powerful machine with thousands of cores. Due to high quality hardware implementation, fault tolerance in HPC systems is not problematic as hardware failures are extremely rare [121]. Even though HPC system can process terabytes of data, they are not scalable as horizontal processing platforms. Moreover, initial deployment and scaling costs are higher compared to other horizontal scale-out platforms [122].

6 Application of Big Data in Smart Grids

In smart grids, the big data coming from several sources carry valuable information, and the cross fertilization of the heterogeneous data sources can unlock several novel applications beneficial to all the stakeholders, i.e., electric utilities, grid operators, customers, etc., for planning and operational decisions. The big data has potential to *a)* improve reliability and resiliency of power grid, *b)* deliver optimum asset management and operations, *c)* improve decision making by sharing information/data, and *d)* to support rapid analysis of extremely large data sets for performance improvement. However, the current trend in smart grid is that the smart meter big data is primarily used for demand response, load forecasting, baseline estimation, and load clustering type of applications [146–150], while the application of PMU big data is focused mainly on transmission grid visualization, state estimation, and dynamic model calibration [61, 83, 151]. Fig. 16 shows some of the potential applications of big data in smart grid useful for various stakeholders. Next, we summarize the recent applications sought from the big data in smart grids.

6.1 Energy Management Related Applications

Two-way flow of power and information in smart grid provides opportunities to small scale consumers, energy producers, and distribution system operators to take active part in grid management and ancillary services. In order to support energy management in real-time, we have to efficiently and intelligently process large volumes of data in smart grids [78]. Improved forecasting tools for energy resources and loads, improved demand response (DR) methods, efficient data management framework, and data analytics are critical to enable the energy management for the optimized operation of power grids. Reference [152] proposed various steps in extracting information from big data for energy management in smart grids. In particular, this work identifies need of methods for dimensionality reduction of data (e.g., Random Projection method), algorithms that can extract load patterns from large-scale data set (e.g., K-means and ANNs), design of machine learning algorithms for improved forecasting, design of data compression for low memory requirements, development of scalable and distributed computing architecture for real-time performance, and so on. Big data is used in energy management of large public bindings in [153]. Deep learning based household level load forecasting method is developed in [98], which is one of the inputs needed for household level energy management systems. A big data enabled EV charging scheme is proposed in [154].

DR, which is the key component of any energy management tools, is one of the drivers of big data analytics in smart grid. Utilities use various DR techniques to enhance customers' active engagement in grid management [155]. Through large amount of data obtained from smart meter and home devices, utilities not only can get near real-time information of consumption, but also can develop proper incentives and operational strategies to better utilize behind the meter energy resources [156]. Big data analytics can dynamically classify and categorize consumer consumption behaviors and electrical characteristics that can help utilities to make better operational decisions [146, 147]. Reference [157] develops methods to cluster energy customers based on time-series data collected from smart meters with an objective to identify suitable customers for DR programs. In [78], authors proposed to use smart meter data and applied time-based Markov Model and clustering algorithms to identify end users' energy consumption dynamics, which is crucial for the DR tools. Reference [18] identifies the significance of high-granular load forecasting and customer consumer behavior modeling using big data useful for distribution grid operation and planning. Demand Response on smart cities utilizing big data is developed in [158]. Similarly, energy consumption pattern in big cities are identified using big data techniques in [159].

6.2 Improvement of Smart Grid Reliability and Stability

In [160], data collected from Twitter is used to identify and locate the power outage, which could help enhance power system reliability. This is an interesting application of big data techniques applied to smart grids based on data collected from social media (non-electrical data). Reference [161] listed significance of GIS, GPS, and weather

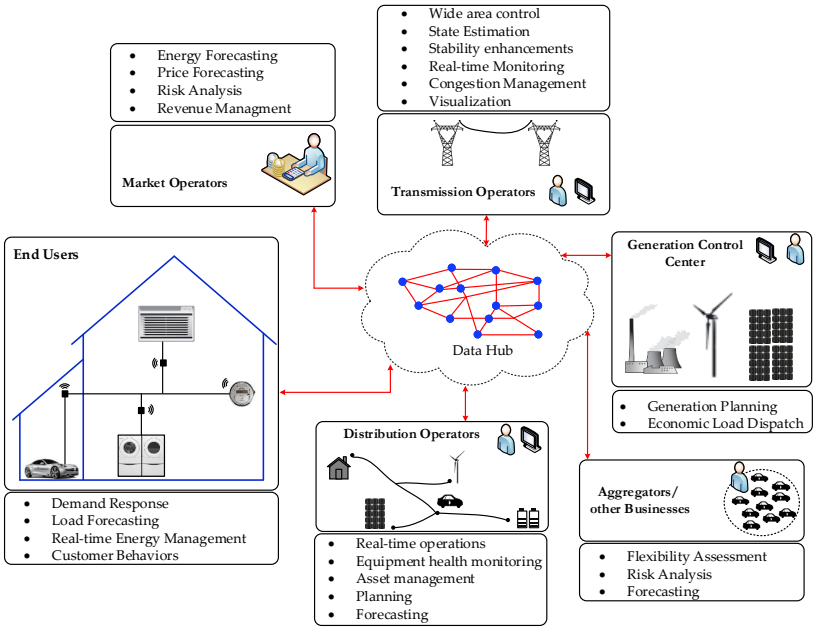

Fig. 16: Some of the potential applications of big data analytics in smart grids.

data in outage management. Application of SCADA big data for voltage instability detection is discussed in [162], which seems promising over traditional snapshot approach. Similarly, PMU big data could be used for stability margin prediction [162] and real-time asset health monitoring [61]. Reference [163] used PMU big data and Core vector machine to assess transient stability margin. PMU based data-driven mode oscillation detection is proposed in [164]. A PMU-based fault location technique is proposed in [85]. The big data methods proposed in [84, 160–164] help improve reliability and stability of power grids. An event detection application is developed in [165] utilizing big data collected from μPMUs. Anomaly detection method on power grid is developed in [166], which is based on big data collected from smart meters. In addition, big data can greatly benefit applications such as transmission constraint management or generator performance monitoring for improving market and operational efficiency.

6.3 Visualization

Advanced visualization is one of the key application area of big data analytics that can improve the overall assessment of smart grids. Big data analytics with the visualization technologies is used for monitoring real-time power system status as well as accurate grid connectivity information. Conventionally, various visualization techniques such as single line diagram, 2D, and 3D charts/plots were used for grid visualization. However, due to increased number of variables and their interdependencies, advanced visualization techniques are often required for the big data visualization in smart grid. Scatter diagram, parallel coordinate, and Andrew curve in combination with the real-time monitoring can resolve the problem of high dimensional data visualization [147]. Commercial tools, such as Real Time Dynamics Monitoring System (RTDMS), are available for visualization using PMU big data [151]. RTDMS provides several visualization options including dashboard display for situational awareness, voltage angle contour plots, voltage magnitude plot, frequency plot, oscillatory mode plot, etc.

6.4 Parameter/State Estimation

Parameter and state estimations are essential for power system planning, operation, and control. Estimations are used for several applications including operational resource planning, real-time system monitoring, and resilient control design against cyber- and/or physical-attacks [167]. The availability of huge amount of data within the smart grid framework provides challenges as well as opportunities for state estimation. Due to availability of large dataset from various sensors and intelligent devices across the grid, system will be more visible, thereby having better and more accurate state estimation. However, due to introduction of large number of active nodes, power system optimization problems become mix-integer, nonlinear, and non-convex, thereby making the system computationally challenging [167]. Through the improved state estimation realized by using big data, we can analyze large datasets (e.g., number, type, sequences) of post-contingency conditions and take corrective actions against a set of predefined contingencies [27, 168]. For instance, the trend in Volt/VAr regulation is to utilize a large mix of voltage regulation resources (e.g. smart inverters, solid state transformers, on-load tap changers, voltage regulators, STATCOMs) on the feeder. The coordination of these resources will require real time monitoring and predictive tools to optimize the utilization of these resources and lead to reduced operational costs and increase the power quality and reliability of the system. Reference [169] proposed a model calibration of distribution feeders based on big data collected from AMI and photovoltaic micro-inverters. References [170–172] used data-driven approach to estimate the behind-the-meter solar power, which are generally not visible from control centers. A PMU based state evaluation method is developed in [173].

6.5 Applications to Cyber-Physical Systems

Since smart grid is a critical infrastructure, any cyber or physical vulnerabilties could lead to widespread impacts. Conventionally power system planner used to perform contingency analysis to provide resiliency under sudden disturbances against system faults and/or natural disasters [174]. Due to close interdependencies between power and communication infrastructure, the future grids subject

to increased risk of malacious attacks. However, most of the existing power system were not designed by accounting cyber-security. Unlike random nature of equipment fault/failure probability distribution, cyber-attacks are normally coordinated and deliberately targeted to most critical components of the energy system. Such structured attacks can lead to cascading failures in the system. Therefore, tight cyber-physical coupling is necessary to extend power system security into both cyber and physical attacks [175–178]. Integration of big data analytics provides an excellent opportunity to timely identify such malacious attacks and prevent the system from huge damages.

7 Future Research Directions

As mentioned in the preceding sections, big data analytics in smart grid is more than just the technical challenges of handling big data. Due to very complex nature of electrical grid, it has close interdependencies with other critical infrastructure (e.g., transportation, gas, water, heating, IoT). The following are future directions to effectively deploy big data in electric utility.

7.1 Interoperability

Even though there are a few standard information models for smart grid interoperability (e.g., IEC 61850, IEC 61850-90-7, IEC 61970/61968, IEEE 1815, IEEE 2030.5), there is no standard information models to describe interoperability among various big data analytics platforms, architecture, and their operational integrations with utility decision frameworks. Furthermore, storage, usage, dissemination, and sharing of data with utility operational frameworks are not unified. Interoperability between various cloud computing service vendors is necessary. Therefore, extensive R&D is needed to develop interoperability among different devices, network operations, data analytics platforms, big data architecture, data repository, and information models.

7.2 Need of Standards and Regulatory Frameworks

Currently, there are no established standards and regulatory frameworks for sharing data among utilities, weather corporation, and other energy systems (e.g., transportation, oil, gas sectors). Regulatory compliance as a whole may need an extensive overhaul to accommodate the impact of big data applications and also the cybersecurity aspects of such applications. First, technical standards should be established to maximize the value of big data as well as to ensure data exchanges among different entities are feasible and meaningful. Subsequently, regulatory framework should also be establish to bind the entities with legal rules and regulations in terms of data sharing. In addition, an impartial third party is also needed in order to make fair estimation and justification of the costs associated with big data deployments for regulated markets and different entities. Therefore, efforts from professional communities should be invested in establishing standards for data sharing among platforms/architectures, and identifying the elements of regulatory frameworks to bind utilities in deploying big data.

7.3 Big Data Architectures/Platforms

Currently, there exists no standardized architectures and platforms for deploying big data analytics to smart grid. Most of the present big data platforms in utility industries rely on cloud computing. As storing and processing of big data within the smart grid requires efficient platforms that are scalable, self-organizing, and adaptive – one of the key solutions is to deploy efficient distributed platforms, such as Hadoop, Cassandra, and Hive [179] that are appropriate for big data analytics. Therefore, holistic and modular energy big data analytics architectures, as well as corresponding computational platforms, are needed to address current barriers within smart grid big data analytics.

7.4 Utilization of Heterogeneous Data

Existing big data applications in smart grids are based on single data type, primarily smart meter or PMU data. However, future applications shall utilize multiple sources of big data (such as data weather, traffic, oil and gas industry, social media, etc.), which can help in assessing the dependence of critical infrastructure on power grids. Therefore, data hubs should be created and be readily accessible to advance resiliency of critical infrastructures. Future grid applications shall utilize these heterogeneous big data set, which could uncover crucial hidden information otherwise not possible from electrical measurements only. A database like Pecan Street Dataport [180], and GE datalake [44] would be lot valuable to research community to uncover interdependencies among the critical infrastructure.

7.5 Integration with Real-time Control, Operation, and Certification

Most of the existing big data deployments to electric utilities have been used for system monitoring and operational planning. However, this is limiting the scope of the big data analytics to electric utility industry. Big data analytics should be integrated into real-time control [48] and operational module so as to provide real-time situational awareness and informed predictive decisions. However, processing of massive data in real time has inherent computational and scalability issues; therefore, these should be the research focus moving forward.

With the diversity of big data applications to the electric utilities, it will be certainly tedious to generate certification programs and operator training certifications to ensure compliance with standards and regulations. Certification mechanism and institutes need harmonization of big data applications which is currently a tremendous void in the electric utility business. The translation of big data applications to electric utility reliability and resilience requirements also needs to be studied and suitable mechanisms of reporting have to be developed and deployed with reasonable confidence. Finally, the ownership of data across multiple ownership models and also customer privacy need to be understood and established under the regulatory framework.

7.6 Advanced Computational Analytics

Because of huge volume of smart grid data, distributed and parallel intelligence is normally needed to effectively address data computation and handling challenges. Since distributed computing and parallel intelligence are effective for addressing local grid issues and challenges, they need some sort of coordination to preserve global visibility. Therefore, effective distributed intelligence and coordination algorithms should be developed. R&D on advanced approaches, such as metamodeling, dimensionality reduction, edge computing should be done to reduce the computational and communication burdens [7].

7.7 Integration with Advanced Visualization

Existing smart energy big data analytics schemes do not incorporate visualization as an integral part. As the key benefit of big data analytics is to help utilities in taking actions based on real-time situational intelligence obtained from the data analytics, integration of advanced visualization with data analytics is needed. Since most of the current analytics are informative and instructive, it requires the grid operators to take intuitive decisions. Integration of advanced visualization together with automated operation provides directive information to the operators and avoids the need of intuitive decision. Therefore, co-design of smart grid big data analytics and advanced visualization mechanisms can produce a seamless integrated framework which can reduce security risk and help to take effective decisions.

7.8 Advancements in Algorithms

Comprehensive analysis of big data for exploiting buried information and correlations among varied data sources is very difficult.

Therefore, advanced artificial intelligence technique such as deep machine learning (ML) is essential not only to exploit fine-grained patterns within the data, but also to make the decision process less reliance on human interference [181]. However, due to increased deployments of intelligent devices in electric grids, and interdependencies of electrical network with other critical infrastructure (e.g., gas, water, transportation), smart grid data will continue to grow in volume, variety and veracity. Therefore, scalability of the ML models is very critical. Moreover, since timely and accurate capture of hidden information is key to the operation of electrical infrastructure, accuracy and computational efficiency play key roles. Therefore, future R&D efforts on ML should focus on scalability, computational efficiency, and accuracy.

7.9 Value Proposition to Different Stakeholders

For electric utility industry seeking to implement big data solution, a structured business model is necessary to fulfill financial goals and requirements of all stakeholders. Since the success of big data analytics in utility industry is contingent upon active participation of electric utilities, customers, and system operators, identifying revenue streams and development of proper business models are critical to the success of big data deployment to smart grid. More importantly, cost associated with the adoption of big data to all stakeholders should be justified and accepted across the broad stakeholders that includes policy makers, regulators, utilities, and the consumer. Therefore, future research should focus on techno-economic studies to quantify technical and economic values of big data to the electric utilities, system operators, and customers. In addition, workforce training will be required for data analysis interpretation as well as to better understand the capability and limitations of these tools. Thus, access to data will just provide return to utility investments when professionals fully understand the capabilities and tools available, requiring changes in undergraduate and graduate curricula to include data science topics for future power engineers.

8 Conclusion

This paper presented a comprehensive state-of-the-art review of big data analytics for smart grids. First, utility and industry perspectives on current status of big data implementation in power system is presented. Key technical, security, and regulatory challenges for deploying big data to smart grid are identified. Value proposition of big data analytics to key stakeholders (e.g., consumers, electric utilities, and system operators) is described with respect to operational integration of big data to utility's decision frameworks. In addition, future research directions for deploying big data analytics to the power grid are discussed from academia, utility, and industry perspectives. This paper provides detailed information and items to consider for utilities looking to apply big data analytics to, and details insights on how utilities can utilize big data analytics to develop new business models and revenue streams. Furthermore, this study will unveil interdependencies among various critical infrastructure and help utilities to make right investment and operational decisions at right time and right locations.

9 References

1 J. Q. Trelewicz, "Big data and big money: The role of data in the financial sector," *IT Professional*, vol. 19, no. 3, pp. 8–10, 2017.
2 E. I. Lab, "Big data in banking for marketers how to derive value from big data," *White Paper*.
3 U. Srinivasan and B. Arunasalam, "Leveraging big data analytics to reduce healthcare costs," *IT professional*, vol. 15, no. 6, pp. 21–28, 2013.
4 M. M. Islam, M. A. Razzaque, M. M. Hassan, W. N. Ismail, and B. Song, "Mobile cloud-based big healthcare data processing in smart cities," *IEEE Access*, vol. 5, pp. 11 887–11 899, 2017.
5 M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob, "Big IoT data analytics: Architecture, opportunities, and open research challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.
6 M. Satyanarayanan, P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, and B. Amos, "Edge analytics in the internet of things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24–31, 2015.
7 S. K. Sharma and X. Wang, "Live data analytics with collaborative edge and cloud processing in wireless iot networks," *IEEE Access*, vol. 5, pp. 4621–4635, 2017.
8 X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 674–686, March 2017.
9 Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016.
10 P. Ta-Shma, A. Akbar, G. Gerson-Golan, G. Hadash, F. Carrez, and K. Moessner, "An ingestion and analytics architecture for iot applied to smart city use cases," *IEEE Internet of Things Journal*, 2017.
11 K. Wedgwood and R. Howard, "Big data and analytics in travel and transportation," *IBM Big Data and Analytics White Paper*, 2014.
12 T. Hong, "Big Data Analytics: Making the Smart Grid Smarter [Guest Editorial]," *IEEE Power and Energy Magazine*, vol. 16, no. 3, pp. 12–16, May 2018.
13 A. Bose, "Smart transmission grid applications and their supporting infrastructure," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 11–19, June 2010.
14 A. P. S. Meliopoulos, G. Cokkinides, R. Huang, E. Farantatos, S. Choi, Y. Lee, and X. Yu, "Smart grid technologies for autonomous operation and control," *IEEE Transactions on Smart Grid*, vol. 2, no. 1, pp. 1–10, March 2011.
15 G. T. Heydt, "The next generation of power distribution systems," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 225–235, Dec 2010.
16 W. Hou, Z. Ning, L. Guo, and X. Zhang, "Temporal, functional and spatial big data computing framework for large-scale smart grid," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2018.
17 K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 215–225, 2016.
18 N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, "Big data analytics in power distribution systems," in *Proc. IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Feb 2015, pp. 1–5.
19 Y. J. Kim, M. Thottan, V. Kolesnikov, and W. Lee, "A secure decentralized data-centric information infrastructure for smart grid," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 58–65, November 2010.
20 T. F. Garrity, "Getting smart," *IEEE Power and Energy Magazine*, vol. 6, no. 2, pp. 38–45, March 2008.
21 O. Zinaman, M. Miller, A. Adil, D. Arent, J. Cochran, R. Vora, S. Aggarwal, M. Bipath, C. Linvill, A. David *et al.*, "Power systems of the future," *The Electricity Journal*, vol. 28, no. 2, pp. 113–126, 2015.
22 J.-P. Dijcks, "Oracle: Big data for the enterprise," *Oracle White Paper*, 2012.
23 C. L. Stimmel, *Big data analytics strategies for the smart grid*. CRC Press, 2014.
24 S. Chen, Z. Wei, G. Sun, K. W. Cheung, and D. Wang, "Identifying optimal energy flow solvability in electricity-gas integrated energy systems," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 846–854, April 2017.
25 S. Callahan. Big data: The future of energy and utilities. https://www.rdmag.com/article/2015/10/big-data-future-energy-and-utilities. Accessed: 2015-10-05.
26 Z. Dong, P. Zhang *et al.*, *Emerging techniques in power system analysis*. Springer, 2010.
27 J. Hu and A. V. Vasilakos, "Energy big data analytics and security: challenges and opportunities," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2423–2436, Sep. 2016.
28 P. Haase, "Intelligrid: A smart network of power," *EPRI journal*, no. Fall, pp. 26–32, 2005.
29 T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J.-M. Pierson, and A. V. Vasilakos, "Cloud computing: Survey on energy efficiency," *Acm computing surveys (csur)*, vol. 47, no. 2, p. 33, 2015.
30 H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. Zhang, "Energy big data: A survey," *IEEE Access*, vol. 4, pp. 3844–3861, 2016.
31 R. Hebner. Nanogrids, microgrids, and big data: The future of the power grid. http://spectrum.ieee.org/energy/renewables/nanogrids-microgrids-and-big-data-the-future-of-the-power-grid. Accessed: 2017-03-31.
32 sungard, "Big data - challenges and opportunities for the energy industry," White paper, 2013.
33 Z. Asad and M. A. R. Chaudhry, "A two-way street: Green big data processing for a greener smart grid," *IEEE Systems Journal*, vol. 11, no. 2, pp. 784–795, June 2017.
34 SWECO, "Smart grid and big data analytics," Technical report, 2015.
35 S. Pancholi, "Solving big data challenges us electric utility industry," PES general meeting presentation, 2014.
36 J. R. Johnson. (2014) How four u.s. utilities are tackling big data. http://www.energycentral.com/c/iu/how-four-us-utilities-are-tackling-big-data.
37 C. Consulting, "Big data blackout: Are utilities powering up their data analytics?" Technical report, 2015.
38 C. E. Association, "Electric utility innovation toward vision 2050," Technical report, 2015.
39 H. Dong, G. Singh, A. Attri, and A. El Saddik, "Open data-set of seven canadian cities," *IEEE Access*, vol. 5, pp. 529–543, 2017.
40 GE, "Predix: The industrial internet platform," White paper, November, 2016.
41 SIEMENS, "Energyip - a flexible, scalable platform for mdm and more," http://w3.usa.siemens.com/smartgrid/us/en/smart-metering/energyip-mdms-platform/pages/energyip.aspx.
42 Siemens. Siemens energyip application platform: Maximize the return on your smart grid investment. http://w3.siemens.com/smartgrid/global/en/products-systems-solutions/smart-metering/emeter/pages/energyip-platform.aspx.
43 G. Electric. (Accessed: 2017) The role of big data visualization and analytics in the utility industry. http://www.electricenergyonline.com/show_article.php?mag=92&article=750.
44 GE. (Accessed: 2014) Grid iq insight: Translating data to actionable intelligence for empowered decision making.

https://www.gegridsolutions.com/uos/catalog/grid-iq-insight.htm.
45 ABB, "Using smart grid data to power end-to-end asset management," White paper, 2011.
46 M. Chavero, "New smart asset management strategies in tso industry enabled bya real-time data infrastructure," White paper, 2016.
47 L. Xie, Y. Chen, and P. R. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 2784–2794, Nov 2014.
48 D. Zhou, J. Guo, Y. Zhang, J. Chai, H. Liu, Y. Liu, C. Huang, X. Gui, and Y. Liu, "Distributed data analytics platform for wide-area synchrophasor measurement systems," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2397–2405, Sept 2016.
49 A. Y. Zomaya and Y. C. Lee, *Energy efficient distributed computing systems*. John Wiley & Sons, 2012, vol. 88.
50 R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data," in *Proc. SIGMOD international conference on Management of data*, 2003, pp. 551–562.
51 S. Tsang, B. Kao, K. Y. Yip, W. Ho, and S. D. Lee, "Decision trees for uncertain data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 64–78, Jan 2011.
52 K. Wagstaff, "Machine learning that matters," *arXiv preprint arXiv:1206.4656*, 2012.
53 F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *Proc. First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2010, pp. 327–332.
54 G. Kalogridis, C. Efthymiou, S. Z. Denic, T. A. Lewis, and R. Cepeda, "Privacy for smart meters: Towards undetectable appliance load signatures," in *Proc. First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2010, pp. 232–237.
55 V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proc. SIGMOD International Conference on Management of data*, 2010, pp. 735–746.
56 H. Liu, H. Ning, Y. Zhang, and L. T. Yang, "Aggregated-proofs based privacy-preserving authentication for v2g networks in the smart grid," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1722–1733, Dec 2012.
57 D. S. Markovic, D. Zivkovic, I. Branovic, R. Popovic, and D. Cvetkovic, "Smart power grid and cloud computing," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 566–577, 2013.
58 J. Qi, A. Hahn, X. Lu, J. Wang, and C.-C. Liu, "Cybersecurity for distributed energy resources and smart inverters," *IET Cyber-Physical Systems: Theory & Applications*, vol. 1, no. 1, pp. 28–39, 2016.
59 D. He, N. Kumar, S. Zeadally, A. Vinel, and L. T. Yang, "Efficient and privacy-preserving data aggregation scheme for smart grid against internal adversaries," *IEEE Transactions on Smart Grid*, 2017.
60 S. Sciacca. (Accessed: 2012-09-18) Big data and the need for improved time synchronization standards. http://m.csemag.com/articlepage/big-data-and-the-need-for-improved-time-synchronization-standards/8c0cd0612438905a2e12ab5fb7e4dad4.html.
61 J. Zhao, G. Zhang, K. Das, G. N. Korres, N. M. Manousakis, A. K. Sinha, and Z. He, "Power system real-time monitoring by using pmu-based robust state estimation method," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 300–309, Jan 2016.
62 Y. Tao and D. Papadias, "The mv3r-tree: A spatio-temporal access method for timestamp and interval queries," in *Proc. of Very Large Data Bases Conference (VLDB), 11-14 September, Rome*, 2001.
63 D. Pfoser, C. S. Jensen, Y. Theodoridis *et al.*, "Novel approaches to the indexing of moving object trajectories," in *VLDB*, 2000, pp. 395–406.
64 J. Tayeb, Ö. Ulusoy, and O. Wolfson, "A quadtree-based dynamic attribute indexing method," *The Computer Journal*, vol. 41, no. 3, pp. 185–200, 1998.
65 R. K. V. Kothuri, S. Ravada, and D. Abugov, "Quadtree and r-tree indexes in oracle spatial: a comparison using GIS data," in *Proc. SIGMOD international conference on Management of data*, 2002, pp. 546–557.
66 I. Kamel and C. Faloutsos, "Hilbert R-tree: An improved R-tree using fractals," Tech. Rep., 1993.
67 M. Yigit, V. C. Gungor, and S. Baktir, "Cloud computing for smart grid applications," *Computer Networks*, vol. 70, pp. 312–329, 2014.
68 S. Rusitschka, K. Eger, and C. Gerdes, "Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain," in *Proc. First International Conference on Smart Grid Communications (SmartGridComm)*, 2010, pp. 483–488.
69 A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1244–1253, Sep. 2013.
70 S. Ruj and A. Pal, "Analyzing cascading failures in smart grids under random and targeted attacks," in *Proc. IEEE 28th International Conference on Advanced Information Networking and Applications (AINA)*, 2014, pp. 226–233.
71 M. McGranaghan, D. Houseman, L. Schmitt, F. Cleveland, and E. Lambert, "Enabling the integrated grid: leveraging data to integrate distributed resources and customers," *IEEE Power and Energy Magazine*, vol. 14, no. 1, pp. 83–93, 2016.
72 D. J. Leeds, "The soft grid 2013-2020: Big data & utility analytics for smart grid," *GTM Research*, 2012.
73 S. Tonyali, K. Akkaya, N. Saputro, and A. S. Uluagac, "A reliable data aggregation mechanism with homomorphic encryption in smart grid ami networks," in *Proc. IEEE 13th Annual Consumer Communications & Networking Conference (CCNC)*, 2016, pp. 550–555.
74 W. Zhu and Q. Guo, "Data security and encryption technology research on smart grid communication system," in *Proc. IEEE Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 2016, pp. 175–178.
75 S. K. Bansal, "Towards a semantic extract-transform-load (etl) framework for big data integration," in *Proc. IEEE International Congress on Big Data (BigData Congress)*, 2014, pp. 522–529.
76 Y. Wang, Q. Deng, W. Liu, and B. Song, "A data-centric storage approach for efficient query of large-scale smart grid," in *Proc. IEEE Ninth Web Information Systems and Applications Conference (WISA)*, 2012, pp. 193–197.
77 M. Tahmassebpour, "A new method for time-series big data effective storage," *IEEE Access*, vol. 5, pp. 10 694–10 699, 2017.
78 Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016.
79 D. Huang, H. Zareipour, W. D. Rosehart, and N. Amjady, "Data mining for electricity price classification and the application to demand-side management," *IEEE Transactions on Smart Grid*, vol. 3, no. 2, pp. 808–817, June 2012.
80 S. Singh and A. Yassine, "Mining energy consumption behavior patterns for households in smart grid," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2018.
81 G. Sheng, H. Hou, X. Jiang, and Y. Chen, "A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 695–702, March 2018.
82 B. Gu and V. S. Sheng, "A robust regularization path algorithm for ν -support vector classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 5, pp. 1241–1248, May 2017.
83 M. Pignati, L. Zanni, P. Romano, R. Cherkaoui, and M. Paolone, "Fault Detection and Faulted Line Identification in Active Distribution Networks Using Synchrophasors-Based Real-Time State Estimation," *IEEE Transactions on Power Delivery*, vol. 32, no. 1, pp. 381–392, Feb 2017.
84 H. Jiang, X. Dai, D. W. Gao, J. J. Zhang, Y. Zhang, and E. Muljadi, "spatial-temporal synchrophasor data characterization and analytics in smart grid fault detection, identification, and impact causal analysis," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2525–2536, Sep 2016.
85 M. U. Usman and M. O. Faruque, "validation of a pmu-based fault location identification method for smart distribution network with photovoltaics using real-time data," *IET Generation, Transmission & Distribution*, vol. 12, no. 21, pp. 5824–5833, 2018.
86 Z. S. Hosseini, M. Mahoor, and A. Khodaei, "AMI-enabled distribution network line outage identification via multi-label SVM," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5470–5472, Sep 2018.
87 A. Ahmed, M. Awais, M. Naeem, M. Iqbal, W. Ejaz, A. Anpalagan, and H. Kim, "multiple power line outage detection in smart grids: Probabilistic bayesian approach," *IEEE Access*, vol. 6, pp. 10 650–10 661, 2018.
88 Y. Jiang, C.-C. Liu, M. Diedesch, E. Lee, and A. K. Srivastava, "outage management of distribution systems incorporating information from smart meters," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 4144–4154, Sep. 2016.
89 P. Jokar, N. Arianpoo, and V. C. Leung, "electricity theft detection in ami using customers' consumption patterns.," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.
90 A. M. Prostejovsky, O. Gehrke, A. M. Kosek, T. Strasser, and H. W. Bindner, "distribution line parameter estimation under consideration of measurement tolerances," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 726–735, Apr. 2016.
91 M. A. Azzouz and E. F. El-Saadany, "multivariable grid admittance identification for impedance stabilization of active distribution networks.," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1116–1128, May 2017.
92 F. Wenli, Z. Xuemin, M. Shengwei, H. Shaowei, W. Wei, and D. Lijie, "vulnerable transmission line identification using ISH theory in power grids," *IET Generation, Transmission & Distribution*, vol. 12, no. 4, pp. 1014–1020, 2017.
93 M. Babakmehr, M. G. Simões, M. B. Wakin, and F. Harirchi, "compressive sensing-based topology identification for smart grids," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 532–543, April 2016.
94 G. Cavraro and V. Kekatos, "graph algorithms for topology identification using power grid probing," *arXiv preprint arXiv:1803.04506*, 2018.
95 S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "identifying topology of low voltage distribution networks based on smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5113–5122, Sep. 2018.
96 Y. Weng, Y. Liao, and R. Rajagopal, "distributed energy resources topology identification via graphical modeling," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2682–2694, July 2017.
97 M. Chaouch, "clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves.," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411–419, Jan 2014.
98 H. Shi, M. Xu, and R. Li, "deep learning for household load forecasting–a novel pooling deep rnn," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271–5280, Sep. 2018.
99 R. Gulbinas, A. Khosrowpour, and J. Taylor, "segmentation and classification of commercial building occupants by energy-use efficiency and predictability," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1414–1424, May 2015.
100 A. Naeem, A. Shabbir, N. U. Hassan, C. Yuen, A. Ahmad, and W. Tushar, "understanding customer behavior in multi-tier demand response management program," *IEEE Access*, vol. 3, pp. 2613–2625, 2015.
101 D. He, L. Du, Y. Yang, R. Harley, and T. Habetler, "front-end electronic circuit topology analysis for model-driven classification and monitoring of appliance loads in smart buildings," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 2286–2293, Dec 2012.
102 Y. Wang, Q. Chen, C. Kang, Q. Xia, and M. Luo, "sparse and redundant representation-based smart meter data compression and pattern extraction," *IEEE*

Transactions on Power Systems, vol. 32, no. 3, pp. 2142–2151, May 2017.
103 Z. Sui, M. Niedermeier, and H. de meer, "tai: a threshold-based anonymous identification scheme for demand-response in smart grids," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3496–3506, July 2018.
104 A. Ukil and R. Zivanovic, "Automated analysis of power systems disturbance records: Smart grid big data perspective," in *Proc. IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, 2014, pp. 126–131.
105 R. Zhang, "Big data analytics for smart grid-forecast, predict for a smarter grid."
106 R. Pandey, M. Dhoundiyal, and A. Kumar, "Correlation analysis of big data to support machine learning," in *Proc. IEEE Fifth International Conference on Communication Systems and Network Technologies (CSNT)*, 2015, pp. 996–999.
107 K. le Zhou, S. lin Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103 – 110, 2013.
108 M. Macedo, J. Galo, L. De Almeida, and A. d. C. Lima, "Demand side management using artificial neural networks in a smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 41, pp. 128–133, 2015.
109 A. Monti and F. Ponci, "Power grids of the future: Why smart means complex," in *Proc. IEEE Complexity in Engineering*, 2010, pp. 7–11.
110 A. Sancho-Asensio, J. Navarro, I. Arrieta-Salinas, J. E. Armendáriz-Íñigo, V. Jiménez-Ruano, A. Zaballos, and E. Golobardes, "Improving data partition schemes in smart grids via clustering data streams," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5832–5842, 2014.
111 X. Tong, C. Kang, and Q. Xia, "Smart metering load data compression based on load feature identification," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2414–2422, Sep. 2016.
112 J. C. S. de Souza, T. M. L. Assis, and B. C. Pal, "Data compression in smart distribution systems via singular value decomposition," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 275–284, Jan 2017.
113 A. D. Martins and E. C. Gurjão, "Processing of smart meters data based on random projections," in *Proc. IEEE Innovative Smart Grid Technologies Latin America (ISGT LA)*, 2013, pp. 1–4.
114 N. Dahal, R. L. King, and V. Madani, "Online dimension reduction of synchrophasor data," in *Proc. IEEE Transmission and Distribution Conference and Exposition (T&D)*, 2012, pp. 1–7.
115 E. P. Xing, Q. Ho, W. Dai, J. K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar, and Y. Yu, "Petuum: A new platform for distributed machine learning on big data," *IEEE Transactions on Big Data*, vol. 1, no. 2, pp. 49–67, June 2015.
116 A. D'Elia, F. Viola, F. Montori, M. D. Felice, L. Bedogni, L. Bononi, A. Borghetti, P. Azzoni, P. Bellavista, D. Tarchi, R. Mock, and T. S. Cinotti, "Impact of interdisciplinary research on planning, running, and managing electromobility as a smart grid extension." *IEEE Access*, vol. 3, pp. 2281–2305, 2015.
117 N. Mohamed, S. Lazarova-Molnar, I. Jawhar, and J. Al-Jaroodi, "Towards service-oriented middleware for fog and cloud integrated cyber physical systems," in *Proc. IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, 2017, pp. 67–74.
118 K.-K. Nguyen and M. Cheriet, "Virtual edge-based smart community network management," *IEEE Internet Computing*, vol. 20, no. 6, pp. 32–41, 2016.
119 R. Mallik, N. Sarda, H. Kargupta, and S. Bandyopadhyay, "Distributed data mining for sustainable smart grids," *Proc. of ACM SustKDD*, vol. 11, pp. 1–6, 2011.
120 A. Akusok, K.-M. Björk, Y. Miche, and A. Lendasse, "High-performance extreme learning machines: a complete toolbox for big data applications," *IEEE Access*, vol. 3, pp. 1011–1025, 2015.
121 K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Proc. IEEE 26th symposium on Mass storage systems and technologies (MSST)*, 2010, pp. 1–10.
122 R. C. Green, L. Wang, and M. Alam, "Applications and trends of high performance computing for electric power systems: Focusing on smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 922–931, June 2013.
123 S. Bera, S. Misra, and J. J. Rodrigues, "Cloud computing applications for smart grid: A survey," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 5, pp. 1477–1494, May 2015.
124 Y. C. Lee and A. Y. Zomaya, "Energy conscious scheduling for distributed computing systems under different operating conditions," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 8, pp. 1374–1381, Aug 2011.
125 M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1017–1025, June 2013.
126 Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna, "Cloud-based software platform for big data analytics in smart grids," *Computing in Science & Engineering*, vol. 15, no. 4, pp. 38–47, 2013.
127 F. Ma, X. Luo, and E. Litvinov, "Cloud computing for power system simulations at iso new england–experiences and challenges," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2596–2603, Nov 2016.
128 B. Fang, X. Yin, Y. Tan, C. Li, Y. Gao, Y. Cao, and J. Li, "The contributions of cloud technologies to smart grid," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 1326–1331, 2016.
129 C. Harvey, S. Rosen, J. Ramsey, C. Saunders, and S. K. Guharay, "Computationally and statistically efficient model fitting techniques," *Journal of Statistical Computation and Simulation*, vol. 87, no. 1, pp. 123–137, 2017.
130 A. Khosravi, S. Nahavandi, and D. Creighton, "Quantifying uncertainties of neural network-based electricity price forecasts," *Applied energy*, vol. 112, pp. 120–129, 2013.
131 S. Schütte, S. Scherfke, and M. Tröschel, "Mosaik: A framework for modular simulation of active components in smart grids," in *Proc. IEEE First International Workshop on Smart Grid Modeling and Simulation (SGMS)*, 2011, pp. 55–60.
132 E. M. Stewart, S. Kiliccote, C. Shand, A. McMorran, R. Arghandeh, and A. von Meier, "Addressing the challenges for integrating micro-synchrophasor data with operational system applications," in *Proc. IEEE PES General Meeting| Conference & Exposition*, 2014, pp. 1–5.
133 F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, "Nist cloud computing reference architecture," *NIST special publication*, vol. 500, no. 2011, p. 292, 2011.
134 S. M. Borodo, S. M. Shamsuddin, and S. Hasan, "Big data platforms and techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 1, no. 1, pp. 191–200, 2016.
135 S. Mishra, "Survey of big data architecture and framework from the industry," *NIST Big data Public Working Group*, 2015.
136 M. Ferguson, "Architecting a big data platform for analytics," *A Whitepaper prepared for IBM*, vol. 30, 2012.
137 B. Data, "Analytics reference architecture," *An Oracle White Paper September*, 2013.
138 M. Vaidya and S. Deshpande, "Distributed data management in energy sector using hadoop," in *Proc. IEEE Bombay Section Symposium (IBSS)*, 2015, pp. 1–6.
139 Z. Niu, B. He, and F. Liu, "JouleMR: Towards cost-effective and green-aware data processing frameworks," *IEEE Transactions on Big Data*, vol. 4, no. 2, pp. 258–272, June 2018.
140 A. Pal and S. Agrawal, "An experimental approach towards big data for analyzing memory utilization on a hadoop cluster using hdfs and mapreduce," in *Proc. IEEE First International Conference on Networks & Soft Computing (ICNSC)*, 2014, pp. 442–447.
141 S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, Z. Liu, K. Nusbaum, K. Patil, B. J. Peng *et al.*, "Benchmarking streaming computation engines: Storm, flink and spark streaming," in *Proc. IEEE International Parallel and Distributed Processing Symposium Workshops*, 2016, pp. 1789–1792.
142 Apache. (Accessed: 2017-06-17) Apache drill - schema-free sql for hadoop, nosql and cloud storage. http://drill.apache.org/.
143 Hadoop. (Accessed: 2019-01) What is apache hadoop. http://hadoop.apache.org/.
144 D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 8, 2015.
145 Apache. (Accessed: 2019-01) http://https://flink.apache.org/.
146 S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar, "Dependable demand response management in the smart grid: A stackelberg game approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 120–132, March 2013.
147 J. Kwac and R. Rajagopal, "Demand response targeting using big data analytics," in *Proc. IEEE International Conference on Big Data*, 2013, pp. 683–690.
148 Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid*, pp. 1–1, 2018.
149 S. N. Fallah, R. C. Deo, M. Shojafar, M. Conti, and S. Shamshirband, "Computational Intelligence Approaches for Energy Load Forecasting in Smart Energy Management Grids: State of the Art, Future Challenges, and Research Directions," *Energies*, vol. 11, no. 3, 2018.
150 A. Tureczek, P. S. Nielsen, and H. Madsen, "Electricity Consumption Clustering Using Smart Meter Data," *Energies*, vol. 11, no. 4, 2018.
151 A. Agarwal, J. Balance, B. Bhargava, J. Dyer, K. Martin, and J. Mo, "Real time dynamics monitoring system (rtdms) for use with synchrophasor technology in power systems," in *Proc. IEEE Power and Energy Society General Meeting*, July 2011, pp. 1–8.
152 P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," *Big Data Research*, vol. 2, no. 3, pp. 94–101, 2015.
153 V. Marinakis, H. Doukas, J. Tsapelas, S. Mouzakitis, A. Sicilia, L. Madrazo, and S. Sgouridis, "From big data to smart energy services: An application for intelligent energy management," *Future Generation Computer Systems*, 2018.
154 Y. Cao, H. Song, O. Kaiwartya, B. Zhou, Y. Zhuang, Y. Cao, and X. Zhang, "Mobile Edge Computing for Big-Data-Enabled Electric Vehicle Charging," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 150–156, March 2018.
155 A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for healthcare applications," *IEEE Access*, 2017.
156 Y. Simmhan, S. Aman, B. Cao, M. Giakkoupis, A. Kumbhare, Q. Zhou, D. Paul, C. Fern, A. Sharma, and V. K. Prasanna, "An informatics approach to demand response optimization in smart grids," City of Los Angeles Department, Tech. Rep., 2011.
157 C. Chelmis, J. Kolte, and V. K. Prasanna, "Big data analytics for demand response: Clustering over space and time," in *Proc. IEEE International Conference on Big Data (Big Data)*, Oct 2015, pp. 2223–2232.
158 A. Jindal, N. Kumar, and M. Singh, "A unified framework for big data acquisition, storage, and analytics for demand response management in smart cities," *Future Generation Computer Systems*, 2018.
159 R. Perez-Chacon, J. M. Luna-Romera, A. Troncoso, F. Martinez-Alvarez, and J. C. Riquelme, "Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities," *Energies*, vol. 11, no. 3, 2018.
160 H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2516–2524, Sept 2016.
161 P.-C. Chen, T. Dokic, and M. Kezunovic, "The use of big data for outage management in distribution systems," in *Proc. International Conference on Electricity Distribution (CIRED) Workshop*, 2014.
162 M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in *2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid*, Aug 2013, pp. 1–9.
163 B. Wang, B. Fang, Y. Wang, H. Liu, and Y. Liu, "Power system transient stability assessment based on big data and the core vector machine," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2561–2570, Sept 2016.

164 T. Jiang, Y. Mu, H. Jia, N. Lu, H. Yuan, J. Yan, and W. Li, "A novel dominant mode estimation method for analyzing inter-area oscillation in china southern power grid," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2549–2560, Sept 2016.
165 Y. Zhou, R. Arghandeh, and C. J. Spanos, "Partial Knowledge Data-Driven Event Detection for Power Distribution Networks," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5152–5162, Sep. 2018.
166 R. Moghaddass and J. Wang, "A Hierarchical Framework for Smart Grid Anomaly Detection Using Large-Scale Smart Meter Data," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5820–5830, Nov 2018.
167 F. Capitanescu, J. M. Ramos, P. Panciatici, D. Kirschen, A. M. Marcolini, L. Platbrood, and L. Wehenkel, "State-of-the-art, challenges, and future trends in security constrained optimal power flow," *Electric Power Systems Research*, vol. 81, no. 8, pp. 1731–1741, 2011.
168 A. J. Ardakani and F. Bouffard, "Identification of umbrella constraints in dc-based security-constrained optimal power flow," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 3924–3934, Nov 2013.
169 J. Peppanen, M. J. Reno, R. J. Broderick, and S. Grijalva, "Distribution system model calibration with big data from ami and pv inverters," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2497–2506, Sept 2016.
170 H. Shaker, H. Zareipour, and D. Wood, "Estimating power generation of invisible solar sites using publicly available data," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2456–2465, Sept 2016.
171 H. Shaker, H. Zareipour, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2466–2476, Sept 2016.
172 X. Zhang and S. Grijalva, "A data-driven approach for detection and estimation of residential pv installations," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2477–2485, Sept 2016.
173 L. Chu, R. Qiu, X. He, Z. Ling, and Y. Liu, "Massive Streaming PMU Data Modelling and Analytics in Smart Grid State Evaluation based on Multiple High-Dimensional Covariance Test," *IEEE Transactions on Big Data*, vol. 4, no. 1, pp. 55–64, March 2018.
174 K. J. Ross, K. M. Hopkinson, and M. Pachter, "Using a distributed agent-based communication enabled special protection system to enhance smart grid security," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1216–1224, June 2013.
175 M. Touhiduzzaman, A. Hahn, and A. Srivastava, "a diversity-based substation cyber defense strategy utilizing coloring games," *arXiv preprint arXiv:1802.02618*, 2018.
176 C. Vellaithurai, A. Srivastava, S. Zonouz, and R. Berthier, "Cpindex: cyber-physical vulnerability assessment for power-grid infrastructures," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 566–575, March 2015.
177 A. Ahmed, V. Krishnan, S. Foroutan, M. Touhiduzzaman, A. Srivastava, Y. Wu, A. Hahn, and S. Sindhu, "cyber physical security analytics for anomalies in transmission protection systems," in *2018 IEEE Industry Applications Society Annual Meeting (IAS)*. IEEE, 2018, pp. 1–8.
178 M. Touhiduzzaman, A. Hahn, and A. Srivastava, "arcades: Analysis of risk from cyber attack against defensive strategies for power grid," *IET Cyber-Physical Systems: Theory & Applications*, 2018.
179 M. Mayilvaganan and M. Sabitha, "A cloud-based architecture for big-data analytics in smart grid: A proposal," in *Proc. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2013, pp. 1–4.
180 Pecan Street Dataport. https://dataport.cloud/.
181 X.-W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," *IEEE access*, vol. 2, pp. 514–525, 2014.