Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions

Bishnu P. Bhattarai\textsuperscript{1}, Sumit Paudyal\textsuperscript{2}, Yusheng Luo\textsuperscript{3}, Manish Mohanpurkar\textsuperscript{3}, Kwok Cheung\textsuperscript{4}, Reinaldo Tonkoski\textsuperscript{5}, Rob Hvospi\textsuperscript{3}, Kurt S. Myers\textsuperscript{3}, Rui Zhang\textsuperscript{3}, Power Zhao\textsuperscript{3}, Milos Manic\textsuperscript{3}, Song Zhang\textsuperscript{3}, Xiaping Zhang\textsuperscript{10} 

\textsuperscript{1}Pacific Northwest National Laboratory, Richland, WA, USA  
\textsuperscript{2}Michigan Technological University, Houghton, MI, USA  
\textsuperscript{3}Idaho National Laboratory, Idaho Falls, ID, USA  
\textsuperscript{4}GE Grid Solutions, Redmond, WA, USA  
\textsuperscript{5}South Dakota State University, Brookings, SD, USA  
\textsuperscript{6}IBM Research, Almaden, CA, USA  
\textsuperscript{7}Oncor Electric Delivery, Dallas, TX, USA  
\textsuperscript{8}Virginia Commonwealth University, Richmond, VA, USA  
\textsuperscript{9}Independent System Operator New England, Holyoke, MA, USA  
\textsuperscript{10}California Independent System Operator, Folsom, CA, USA  

E-mail: Bishnu.Bhattarai@pnnl.gov

Abstract: Big data has potential to unlock novel groundbreaking opportunities in power grid that enhances a multitude of technical, social, and economic gains. As power grid technologies evolve in conjunction with measurement and communication technologies, this results in unprecedented amount of heterogeneous big data. In particular, computational complexity, data security, and operational integration of big data into power system planning and operational frameworks are the key challenges to transform the heterogeneous large dataset into actionable outcomes. In this context, suitable big data analytics combined with 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 frameworks. Detailed information for utilities looking to apply big data analytics and insights on how utilities can enhance revenue streams and bring disruptive innovation are discussed. 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 characterised with larger datasets being generated at an unprecedented rate with localised 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 consist of measurements, along with other non-electrical data (e.g. weather, traffic etc.), if effectively utilised in coordination, will revolutionise the operation of electric power grids. The effective utilisation of data enhances observability of power grids that includes system-wide grid conditions, the behaviour 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 an 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, 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 min and are stored in billing centers. One million smart meters installed in a utility result in 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-synchronised phasors to phasor data concentrators 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 a 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 allow improvisation in...
development of proper business models for the key stakeholders As such, it becomes necessary to process large volume Therefore, big data analytics is to turn a 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 planning and operation can lead to multiple benefits including improved reliability and resiliency, optimised 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 realised 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, and system operators), operational benefits, and a potential path forward to deploy big data analytics in power grids [30].

This study presents insights on big data in the smart grid from several different perspectives – research, electric utilities, and industries perspectives. First, we identify current challenges to transform big data in the 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 realise increased revenues and operational benefits. Furthermore, this study provides insights on how effective integration of big data analytics to utility decision frameworks helps to make the right decision at right time and location by unveiling the interdependencies among various critical infrastructures and operations. The remainder of the paper is structured as depicted in Fig. 3. First, a comprehensive analysis of the big data from utility and industry perspectives is presented in Section 2. Next, in Section 3, key challenges for the 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 the 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, the smart grid is associated with a 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 characterised by a 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, this vast amount of data creates several challenges for data handling, processing, and integration to a 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 optimise all current and future smart grid technologies.

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

Fig. 2 Key characteristics of smart grid big data

Fig. 3 Overall organisation of the paper

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.

As shown in Fig. 2, big data in smart grid are characterised by high volume (in the order of thousands of TB), 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, and 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 organising and storage are typically well established in a generic sense. However, big data analytics is more than just 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 transform big data in the 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 realise increased revenues and operational benefits. Furthermore, this study provides insights on how effective integration of big data analytics to utility decision frameworks helps to make the right decision at right time and location by unveiling the interdependencies among various critical infrastructures and operations.

The remainder of the paper is structured as depicted in Fig. 3. First, a comprehensive analysis of the big data from utility and industry perspectives is presented in Section 2. Next, in Section 3, key challenges for the 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 the 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, the smart grid is associated with a 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 characterised by a 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, this vast amount of data creates several challenges for data handling, processing, and integration to a 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 optimise all current and future smart grid technologies.
future smart grid and source of new business opportunities. (80%) of the electric utilities realise big data analytics as crucial for handling millions of TB data, which continues to increase over time. Recently, Canadian Electric Association has also identified big data analytics as one of the key drivers for grid modernisation to meet their 2050 vision. In addition, Canada has initiated a concept of open data among multiple utilities and service providers in seven Canadian cities in an effort to maximise the value of big data. However, even though utilities recognise 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 the current status of electric utilities in terms of big data implementations. 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 value.

In addition, as electric utilities are heavily regulated organisations, 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 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 for improved reliability and resiliency of the system and tighter boundaries from regulating entities are also steadily forcing the utilities to deploy big data analytics. With big data analytics, electric utilities can exploit the meter resources and obtain various grid services at lower cost. More importantly, big data analytics help to reduce levelised cost of electricity not only by helping to make better investment decision at the right time and right place, but also by unveiling insights and value 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 the 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 the management of big data for providing various grid services to electric utilities and grid operators. Siemens is currently integrating utility operations and data management technologies that could potentially be tapped for grid data analytics. This grid analytics platform can allow utilities to utilise big data for multiple functionalities, including home energy management, grid energy management, and predictive/corrective controls. EnergyIP Analytics has already been used by more than 50 utilities with a total of 28 million installed smart devices.

Similarly, GE has developed an industrial IoT platform, called Predix, to consolidate data from existing grid management systems, smart meters, and grid sensors. In addition, Grid IQ Insight, a cloud-based big data analytics architecture, which utilises Predix platform, is developed to integrate big data analytics to grid applications. Native data collected from Grid IQ Insight are stored in Data Lake that could be tapped for several grid applications. 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

---

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

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

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

### 2.1 Electric utility perspective

An electric utility is a very complex structure having close dependencies and interactions among communications, IoT, and human factors. Recent concerns on increased security and reliability of critical infrastructure are leading to the need of integrated energy system, which integrates various critical infrastructures, including electrical, gas, thermal, and transportation. Therefore, future power grid management systems will be processing overwhelming amounts of heterogeneous data. As illustrated in Fig. 4, individual devices and functional units can generate thousands of TB data annually. Considering a 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 amounts of data.

A recent survey conducted with 1000 electric utility and industry respondents across ten countries depicted that majority (80%) of the electric utilities realise big data analytics as crucial for future smart grid and source of new business opportunities. Recently, Canadian Electric Association has also identified big data analytics as one of the key drivers for grid modernisation to meet their 2050 vision. In addition, Canada has initiated a concept of open data among multiple utilities and service providers in seven Canadian cities in an effort to maximise the value of big data. However, even though utilities recognise 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 the current status of electric utilities in terms of big data implementations. 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 value.

In addition, as electric utilities are heavily regulated organisations, 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 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 for improved reliability and resiliency of the system and tighter boundaries from regulating entities are also steadily forcing the utilities to deploy big data analytics. With big data analytics, electric utilities can exploit the meter resources and obtain various grid services at lower cost. More importantly, big data analytics help to reduce levelised cost of electricity not only by helping to make better investment decision at the right time and right place, but also by unveiling insights and value 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 the utility industry by expanding business volume and revenue streams.

**Fig. 7** A concept of edge
computing, whereby computational intelligence is connected at the edge of the data source, has been 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 employs equipment monitoring and systems expertise to establish end-to-end asset management, business processes for reducing costs, minimising risks, improving reliability, and optimising 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, and data indexing will increase in an 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, and 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 summarises the key challenges, potential impacts, and potential solutions for deploying big data to the 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 the smart grid, data uncertainty will increase in an 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, and 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 summarises the key challenges, potential impacts, and potential solutions for deploying big data to the power grid.

3.3 Data indexing

This section provides insights on their behaviour [68]. Data aggregation 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.

3.3.1 Data privacy: Data privacy of the user is a very critical security concern as the power consumption of consumer normally provides insights on their behaviour [68]. Data aggregation is one of the common approaches 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 unauthorised modification of information. However, due to close interdependencies between power and communication infrastructure, the power industry is also susceptible to increased 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 (P2IDA) 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 synchronisation

With the increasing need for real-time control and communication in the smart grid, time synchronisation is becoming a key concern. Currently, synchronphasors or PMUs provide time synchronised data, which utilise synchronisation based on radio clocks or satellite receivers. Time synchronised 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 a preventive or remedial solution. However, communication, storage, and analysis of streams of data from most of the distribution system devices and customers are currently unsynchronised. As unsynchronised data poses a potential risk of a misleading decision, data should be time synchronised with respect to the same time reference.

3.5 Data indexing

The smart grid data also possess 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 the 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
However, none of the efforts are being made on interoperability among big data analytics platforms, architectures, and grid 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, and 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 the smart grid, there is a need for data sharing and information exchange among different utilities and system operators. Since electric utilities usually do not share data/information with each other, a regulatory framework should be established to facilitate data sharing and unify their efforts. In order to synchronise 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, and 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 a mere three cents [72]. The falling costs of data storage and data management are making 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.

4 Big data stages and solution approaches

Organising and storing big data, in general, is well understood. Mega-corporations (e.g. Google, Microsoft, and 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.
4.1.2 Data storage: Data storage primarily belongs to data management (e.g. data fusion, data integration, and data transforming) within data repositories [75]. Data storage not only need to manage a 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 and 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 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 a large dataset that can be transformed into actionable outcomes/knowledge. It utilise various algorithms and procedures (e.g. clustering, correlation, classification, categorisation, regression, and 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 behaviours of grid and customers, whereas diagnostic models analyse the operating conditions and decisions made by the grid operators. The diagnostic model is focused on identifying the causes for an event, thereby it is suitable for taking remedial action. As the key objective of data analytics is to provide a prescriptive solution, predictive models are often necessary to forecast operating conditions and future decisions [82]. Prescriptive analysis, on the other hand, is 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.

From the smart grid application perspective, data analytics can be categorised into four broad categories as illustrated in Fig. 11. Event analytics primarily covers diagnosis/detection of the power systems events such as faults and outage managements [83–89]. In addition, event analytics also encompasses a descriptive analysis of prior power system events using various techniques (e.g. classification, filtering, and 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 Fig. 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, 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 applications that falls under the customer analytics include customer classification/categorisation [95, 101], the correlation between consumer behaviour and consumption patterns [89, 97, 99, 102], and demand response (DR) [100, 103].

Please note that data correlation, data classification/categorisation, 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, and 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 the big data domain as well [105, 106].

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

Feature extraction: Feature extraction is one of the important steps of data mining that is intended not only to translate data into meaningful outcomes but also to identify the data attributes networking 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 the smart grid, acquiring and processing all data is technically inefficient from cost, complexity, and storage requirements. The following approaches are designed to make 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]. The key challenge is to find the optimum reduction on a 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 utilises a centralised architecture for data acquisition, analysing, and processing. Such a framework requires a huge exchange of information flow among various intelligent devices within the smart grid [49]. This is inefficient not only from a 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 centralised processor [115]. Recently, edge computing, a method of optimising 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 relieves 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 a 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, parallelisation, 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, cloud computing helps utilities to realise flexibility, agility, and efficiency in terms of saving cost, energy, and resources [29]. Many smart grid applications, including AMI, SCADA, energy management systems, and distribution management systems, can be greatly benefited by the application of cloud computing approach.

4.2.5 Metamodelling: The increase in the complexity of large-scale simulation models often leads to increased run times. Consequently, the simulation of large interconnected networks can benefit from simulation metamodelling to reduce the runtime with acceptable accuracy. Simulation metamodelling is used to build 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 computationally 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 the 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, a 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 standardisation 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 [42], which was developed for a 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 and dashboards). The third layer primarily includes grid applications, while the fourth layer focuses on the visualisation and operational integrations.

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 visualisations, whereas the second layer is designed for analytics and services, thereby consisting of tools and algorithms required for modelling, 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 the 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 leftmost 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.

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).

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 organisation (to ensure data quality for analytical operations), data analytics, decision making (recommendation, alerts, dashboards), and data management (e.g. storage, data security, and governance) [137]. Similarly, horizontal layers include technology platforms and integration layers for operational integration to electric utility operational framework.

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

5.2 Big data platforms

The following subsections present commonly used platforms for big data analytics and compare their performance (Table 2).

5.2.1 Hadoop: Apache Hadoop is an open source framework for storing and processing large datasets using the MapReduce programming model [143]. The Hadoop consists of a 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 the 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 platforms in the electric utility, uses Hadoop for performing data analytics in the 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 (ML), 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 are 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 visualisation 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 the HPC system can process terabytes of data, they are not scalable as horizontal processing platforms. Moreover, initial deployment and scaling costs are higher than 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 fertilisation 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) 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 DR, 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 visualisation, 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 summarise the recent applications sought from the big data in smart grids.

6.1 Energy management related applications

A two-way flow of power and information in the smart grid provides opportunities to small-scale consumers, energy producers, and distribution system operators to take an 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 DR methods, efficient data management framework, and data analytics are critical to enable the energy management for the optimised operation of power grids. Diamantoulakis et al. [152] proposed various steps in extracting information from big data for energy management in smart grids. In particular, this work identifies the 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 ML 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 the energy management of large public bindings in [153]. Deep learning-based household load level load forecasting method was developed in [98], which is one of the inputs needed for household level energy management systems. A big data enabled electric vehicle (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 the smart grid. Utilities use various DR techniques to enhance customers’ active engagement in the grid management [160]. Through a 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 utilise behind the meter energy resources [156]. Big data analytics can dynamically classify and categorise consumer consumption behaviours and electrical characteristics that can help utilities to make better operational decisions [146, 147]. Chelmis et al. [157] developed 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], the authors proposed to use smart meter data and applied the time-based Markov model and clustering algorithms to identify end users’ energy consumption dynamics, which is crucial for the DR tools. Yu et al. [18] identified the significance of high-granular load forecasting and customer consumer behaviour modelling using big data useful for distribution grid operational planning. DR on smart cities utilising big data is developed in [158]. Similarly, the energy consumption pattern in big cities is 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). Chen et al. [161] listed the significance of GIS, global positioning system, and weather 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]. Wang et al. [163] used PMU big data and core vector machine to assess the 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] utilising big data collected from μPMUs. Anomaly detection method on a 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 Visualisation

Advanced visualisation is one of the key application areas of big data analytics that can improve the overall assessment of smart grids. Big data analytics with the visualisation technologies is used for monitoring real-time power system status as well as accurate grid connectivity information. Conventionally, various visualisation techniques such as single line diagram, two-dimensional (2D), and 3D charts/plots were used for grid visualisation. However, due to the increased number of variables and their interdependencies, advanced visualisation techniques are often required for the big data visualisation in the smart grid. Scatter diagram, parallel coordinate, and Andrew curve in combination with real-time monitoring can resolve the problem of high-dimensional data visualisation [147]. Commercial tools, such as real-time dynamics monitoring system (RTDMS), are available for visualisation using PMU big data [151]. RTDMS provides several visualisation 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 the availability of large dataset from various sensors and intelligent devices across the grid, the system will be more visible, thereby having better and more accurate state estimation. However, due to the introduction of a large number of active nodes, power system optimisation problems become mix-integer, non-linear, and non-convex, thereby making the system computationally challenging [167]. Through the improved state estimation realised by using big data, we can analyse large datasets (e.g. number, type, and 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 utilise a large mix of voltage regulation resources (e.g. smart inverters, solid state transformers, on-load tap changers, voltage regulators, static synchronous compensators (STATCOMs)) on the feeder. The coordination of these resources will require real-time monitoring and predictive tools to optimise the utilisation of these resources and lead to reduced operational costs and increase the power quality and reliability of the system. Peppanen et al. [169] proposed a model calibration of distribution feeders based on big data collected from AMI and photovoltaic micro-inverters. The authors of [170–172] used a data-driven approach to estimate the behind-the-meter solar power, which is generally not visible from control centers. A PMU-based state evaluation method is developed in [173].

6.5 Applications to cyber-physical systems

Since the smart grid is a critical infrastructure, any cyber or physical vulnerabilities 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 malicious attacks. However, most of the existing power systems were not designed by accounting cyber-security. Unlike the 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, the 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 malicious attacks and prevent the system from huge damages.

7 Future research directions

As mentioned in the preceding sections, big data analytics in the smart grid is more than just the technical challenges of handling
big data. Due to the very complex nature of the electrical grid, it has close interdependencies with other critical infrastructure (e.g. transportation, gas, water, heating, and IoT). The following are future directions to effectively deploy big data in the 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, and 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, and gas sectors). Regulatory compliance as a whole may need an extensive overhaul to accommodate the impact of big data applications and also the cyber-security aspects of such applications. First, technical standards should be established to maximise the value of big data as well as to ensure data exchanges among different entities are feasible and meaningful. Subsequently, a regulatory framework should also be established 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 exist no standardised architectures and platforms for deploying big data analytics to the 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-organising, 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 Utilisation 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 utilise 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 utilise these heterogeneous big data set, which could uncover crucial hidden information otherwise not possible from electrical measurements only. A database such as Pecan Street Dataport [180] and GE Data Lake [44] would be a lot valuable to the 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 the 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 harmonisation 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

Owing to the 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 metamodelling, dimensionality reduction, and edge computing should be done to reduce the computational and communication burdens [7].

### 7.7 Integration with advanced visualisation

Existing smart energy big data analytics schemes do not incorporate visualisation 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 visualisation 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 visualisation together with automated operation provides directive information to the operators and avoids the need of an intuitive decision. Therefore, co-design of smart grid big data analytics and advanced visualisation 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 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, and 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 the 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.

---

IET Smart Grid

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
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 the utility industry is contingent upon the 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 the smart grid. More importantly, the 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 a 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 study presented a comprehensive state-of-the-art review of big data analytics for smart grids. First, utility and industry perspectives on the current state of big data implementation in the power system are 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 the 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 study provides detailed information and items to consider for utilities looking applying big data analytics to, and detailed insights on how utilities can utilise big data analytics to develop new business models and revenue streams. Furthermore, this study will unveil interdependencies among various critical infrastructures and help utilities to make the right investment and operational decisions at the right time and right locations.

9 References
