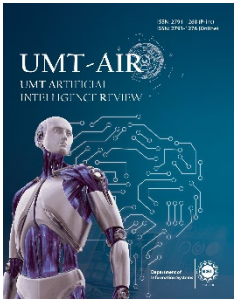

UMT Artificial Intelligence Review (UMT-AIR)
Volume 1 Issue 2, Fall 2021
ISSN(P): 2791-1276 ISSN(E): 2791-1268
Journal DOI: https://doi.org/10.32350/UMT-AIR
Issue DOI: https://doi.org/10.32350/UMT-AIR.0102
Homepage: https://journals.umt.edu.pk/index.php/UMT-AIR

Journal QR Code:

Article: **Big Data Analytics in Smart Power Systems: A Survey Paper**

Author(s): Farhan Ahmad

Affiliation: PRESCON Engineering (Pvt.) Ltd., Pakistan

Article QR:

Farhan Ahmad

Citation: A. Farhan, "Big data analytics in smart power systems: A survey paper," *UMT Artificial Intelligence Review,* vol. 1, pp. 12–26, 2021. https://doi.org/10.32350/UMT-AIR.0102.02

A publication of the
Dr Hassan Murad School of Management,
University of Management and Technology, Lahore, Pakistan

Big Data Analytics in Smart Power Systems: A Survey Paper

Farhan Ahmad[1]*

[1]PRESCON Engineering (Pvt.) Ltd., Pakistan
*Corresponding Author: farhan.ahmad@theprescon.com

ABSTRACT: In the present era, there are various evolutions of technology which bringing to our modernize life. In recent years, technology has brought us many advancements. One of them is integrating big data with smart grid/smart power. In this study, a scientific approach used to help the power system is studied. Additionally, with the help of previously published literature, different survey papers are reviewed to investigate the key challenges of integrating Big Data Analytics (BDA) with smart grid. Subsequently, BDA characteristics are also studied. Next, data analysis techniques and BDA applications in the domain of smart grids are studied. It is followed by a section discussing techniques such as Hadoop and Spark. Their framework is also briefly examined in order to know about their working. The last section provides a conclusion and future directions.

INDEX TERMS: Advanced Metering Infrastructure (AMI), applications of big data, big data analytics (BDA), data architectures, data mining, data privacy, data security, data uncertainty, data volume, Hadoop, Spark, Smart grid, smart power system

I. INTRODUCTION

Technological advancement encourages the use of devices such as sensors, advanced metering infrastructures (AMI), mobile phones, and other digital equipment to produce a large amount of data. At present, data collected from different sources is more than several Giga bytes [1][6]. Big data analytical tools and its applications have huge potential in present day and age, since more and more areas are realizing the need and application of big data.

Present Software and equipment used in the current era are producing large amount of data. For this reason, researchers are working in parallel to design new big data analytical tools and techniques so they are able to process it efficiently big. It is anticipated that big data applications have potential to be used in smart power system [2]. One such example is the use of smart grid and smart power system analytics through big data analytics tools. Presently, modern power grids are incorporated with great innovative equipment and are being used with different equipment, such as for measurement, information, communication and control [2]. Furthermore, advance metering

infrastructure (AMI) technique is also a great revolution in the field of power systems since it is enabling the analysts to observe the trends and consumption in different industries. In addition [2] also discussed phasor measurement units and smart meters that are part of AMI technology. Thus, there are a wide applications of power system that need incorporation of big data analytics to analyse a large amount of data being generated through smart meters and smart sensor devices. Figure 1 shows how large amount of data sets are connected big with the power systems.

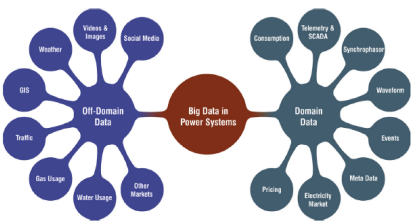

Fig.1. Classification and Examples of Big data types in Power Systems

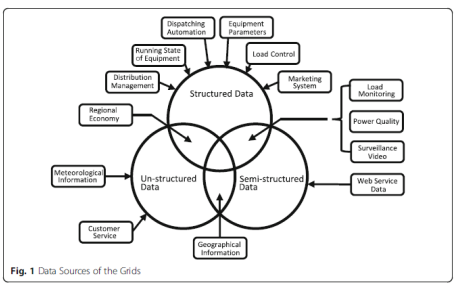

Fig.2. Data Sources of Grids

Smart grid technology is an advanced feedback-based technology that includes two-way power system including information and communication technology (ICT). Conventional power grids involve uni-directional power transfer and is not a feedback-based system. For this reason, authorities are not aware of how much power is being transferred. Thus, this study focused on examined smart power system, which is based on smart communication-based tools such as sensors, to investigate the role of big data and power systems.

In the present study, in this paper, we will analysed role, challenges, and analysis of big data in power systems as well as smart grid analytics. Furthermore, the architecture, tools and techniques that are employed in analysis of data collected from smart power systems will also be analysed. In addition, our paper will review the barriers standing in the way of adopting big data in smart power system applications. Energy security and environmental sustainability are global concerns with immense societal impact. Significant energy assets go toward electricity generation, and the

power sector is expected to grow worldwide [9].

This research provides a survey of big data analytics applications and associated implementation issues. The emphasis is placed on applications that are novel and have demonstrated value to the industry, as illustrated using field data and practical applications. The paper reflects on the lessons learned from initial implementations, as well as ideas that are yet to be explored [10].

This paper conducts a comprehensive study on the application of big data and machine learning in the electrical power grid introduced through the emergence of the next-generation power system the smart grid (SG) [11].

II. BIG DATA ANALYTICS: A NEW THINKING

Big data analytics is a vast domain, covering both broad application of electrical power systems and information technology.

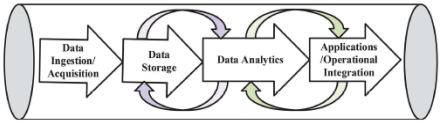

Fig.3. Key Stages of big Data Analytics

It also provides diversified tools and different platforms in order to store, manage and speed process the big data. big Related models are leveraged with the help of different methodologies involved in the big data analytics. Required computational and statistical claims have been always seen as a challenge in the power system, but computational tools and techniques of big data has aided the power system analysts to analyse big data and compute the results to get some useful results [4].

[1] also reported that benchmarks adopted by big data analytics and its solutions are not equal to overall issues. big data solutions vary with the problem to problem. There are various issues that cannot be addressed or solved with the help of tools and techniques used in the big data analytics. However, with recent improvements, power system challenges may be resolved by at least taking a lead from data analytics performed in big data analytics. [1] also reported that BDA has added new diversified dimensions to the solutions of power systems. Hence, big data analytics using multi-purpose tools has helped open a door of new science to the power system domain.

[5] Stated that big data introduced new science to aid the end users to address their problems. They have taken hold on the power system domain very aggressively. They also reported about the use of technology and how big data has been handled efficiently with the help of new tools and techniques developed in the new era of science. This justifies the argument mentioned by [1] that big data has introduced new doors to share knowledge. It also addressed the issues raised in the vast domain of power system. Following figure illustrates the above-mentioned arguments and justifications.

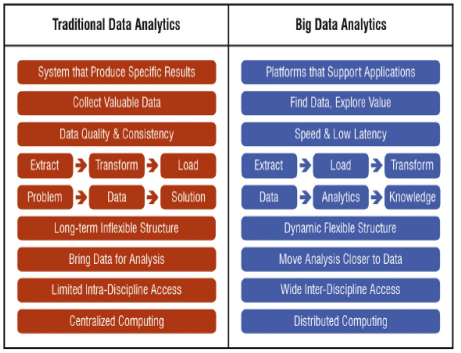

Fig.4. Conceptual Comparison between Traditional Data Analysis and BDA

In addition, literature also reported big data tools and techniques offers diversified solutions including descriptive, diagnostic, corrective, prescriptive, predictive and distributive analytics [1]. Hence, there are lot more benefits of integrating big data tools and solutions with the concerns and issues of power systems.

The next section provides detail on the key challenges that have been addressed with the help of big data.

III. Key Challenges For Big Data Analysis in Smart Power System

Big data analytical tools have many benefits when used for power systems. However, this study investigates key challenges that power system domain offers to the big data analytics.

i. Data Volume: Volume of the data in the power system domain is referred to as the amount of data being generated by the utilities all around the country [2]. This data appears to be expanding at an exponential rate every day. As technology becomes more intertwined with traditional methods, power system tools, more and more digital data is being produced. It is observed that data is increased to hundreds of TB [2].

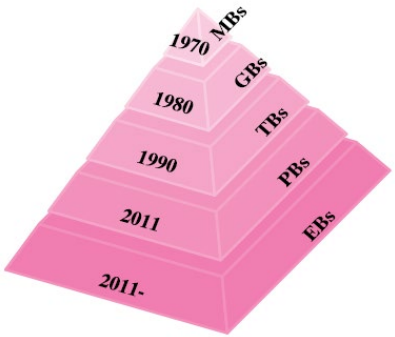

Fig.5. Brief History of Big Data

Fig.6. Attribute of BD and some Data Source in Context of Power System

Data management is now not only confined to the generation end, but also with the customer end, as an outcome of this large-scale data. To analyse the data in which they can see the power flow from the generation end to the customer end. Furthermore, theft control is a lost of power system wherein authorities are interested in monitoring the power that is being consumed in the form of losses and/or theft in our daily lives. Hence, on such a large scale, that is, country wide data management and data analysis, is an issue for the authorities, hence they have a keen interest to integrate new technological tools to handle large data volume.

In this regard, new architecture is being implemented and under the discussion in the literature. Researchers are developing a distribution management strategy, beginning with a distributed and scalable architecture that analyses measurable data in order to efficiently control the distribution end power system [2].

Assessment of data, in addition to computational analysis of the huge volume of data, is also a key issue [1]. Scientists are working on the data sets and dimensionally decreasing them so that the data may be analysed in smaller sections. This would aid the data engineers in analysing the data in an efficient way in order to generate positive results. Also to produce some fruitful suggestions [7].

ii. Data Un-certainty: Data un-certainty deals with the data accuracy and consistency as well as data completeness. The majority of the data in real-time data is not unique and contains null values, or data sets are manipulated. This can cause the data un-certainty issues. Moreover, smart meters that are being installed on the consumer

end to get data in the control centre or in the monitoring room sometimes produce mixed data that need data cleaning. However, due to technical in-competencies, our engineers are not able to get useful results due to which this data becomes wastage [2]. Some of the most frequent reason of conflicting data are inaccurate sensors, communication latencies/delays, physically damaged equipment, unsynchronized time data, missing data and noises. Furthermore, any malicious attack on the data generating equipment can also results in producing uncertain data [2]. [8] also stated that data analysts often face issue of lackness of tractable algebraic equations. This issue makes the relation of simulations and optimizations weaker due to uncertainty links with the simulations of uncertain data. Hence, the desired results are not fetched.

[2] Provide the solutions of the above mentioned problem and reported their findings that by stating that big data uncertainty can be reduced with the help of stochastic processes. This process must be applied within certain limit. Uncertain data should be modelled in a way that within limit stochastic processes deployed will handle the uncertain data and undesired results. Stochastic process is collection of variables that are random in nature. It is defined with the help of formula as: $X=\{Xt:t\in T\}$. This process or phenomena is applied on common probability space in which values are taken in a set that is common. It can be taken as both i.e., discrete or continuous respectively ("Stochastic Processes - an overview | ScienceDirect Topics", 2021). As a result, data uncertainty is a substantial problem in data analysis that must be addressed with prudence as well as using techniques discussed in the literature.

iii. Data Security and Data Privacy: This is the most common issue faced by data engineers not only in the applications of power system but also in the overall applications of IT industry. This is a pertinent factor, and security agencies are working hard to ensure data security and privacy. To protect valuable and significant data, a variety of data security methods are employed. FireHost, StorageGrid Webscale, Keyless SSL, and others are among them. Data security engineers, on the other hand, employ a variety of

other data security solutions to safeguard cloud data as well as data stored on data centre hard drives.

Smart grid related data is often private to customers and cannot show to their customers. However, due to cyber-attacks, it is always a fear that attackers can change or destroy the data. This data includes financial transactions, commercial secrets etc. [2] studied the solutions to the data privacy to report that data aggregation is the most important tool to secure data. In addition, they also quoted that this aggregation can be of different types such as differential and distributed aggregation.

Challenges	Possible impact	Potential solution
data volume	need of increased storage and computing resources	dimensionality reductic parallel computing, ed computing, cloud computing, pay-per us [47–49]
data quality	lack of complete information, misleading decision	probabilistic and stocha analysis [50, 51], dat cleaning (e.g. dealing w missing values, smooth noises, outliers, and inconsistent data) [52
data security	vulnerable to malicious attack, compromise consumer privacy and integrity, mislead operational decision and financial transactions	data anonymisation (e data aggregation [53–5 data encryption [56–58 and P2DA [59])

Fig.7. Summary of key challenges to apply big data to smart grid

IV. Big Data Characterisitics in Smart Grid

Big data in smart grid can be categorized with the help of Vs of data. Smart grid data can be categorized as volume, velocity, and variety. In addition to these Vs, there are some other Vs such as 4Vs, 5Vs, 6Vs, 7Vs and 8Vs.

Talking about the Vz of data being produced by smart grids, these are the following:

a. **Volume:** Volume of data as talked in the paper earlier is amount of data generated. This data is in bulk and getting bulky day by day. Data processing, data storage, and/or data backups are all issues with this data feature. This can be done with the help of latest tools such as Hadoop etc. where data is stored in the multiple locations [6] .

b. **Velocity:** It is referred to as the speed of the data. Within this context, we may state that smart grids in power system are continuously generating a vast amount of data with no significant time. This information is extremely difficult to keep track of over time. Also, along with storing, data backups are important. According to prior literature, sampling rate of 4 time

per hour, 35.04 billion records were fetched which is equivalent to the 2920 TB of the data [6].

c. **Variety:** In the past when traditional systems were deployed in the power system, data types were not much but with the advancement in the technology, data types and variety are increasing and there are different types of data being fetched from the number of devices such as sensors and advanced meters. These sensors and meters are producing multi type of data such as image data, video-based data, graphical data etc. It is also seen in the real life that most of the sensors are producing graphical data which is interpret by the data engineers in the data centres. This data is then interpreted as per requirements and then necessary actions are performed based on the information received [6].

d. **Veracity:** It is about the correct data received from the devices. Sometimes the data received from data sources such as sensors is invalid or imprecise data that cannot be computed. Such data has null values or not reliable values that can be processed. So, such data before processing is cleaned through data exploratory techniques and then useful data is processed or otherwise such data is neglected and not reliable. However, errors in the devices exist which can be eradicated with the help of engineering standards [6].

e. **Value:** This is about the value of the data or the useful information that is fetched from the data sources. Utilities are getting large amount of data at their data centres but somehow in our country it is seen that utility engineers have not much knowledge to produce useful results from such data. So extracting valuable results from the received data is the part of big data characteristics in smart grid/power system [6].

V. BIG DATA ARCHITECTURES

In this section of study, we will understand about the integration of big data with the existing control of power system, smart grid/smart power architectural control, big data environment and how smart grid integrated big data solutions are different from the traditional computational environments.

It is seen that there are no standard architectures in the big data [2]. The following subsections explain large data architectures. big

IQ insight Architecture:

It works on the basis of foundations of PREDIX data analytics platform [2]. It is a cloud-based architecture horizontal in nature. It has four different layers. Below is the image showing detailed four different layers.

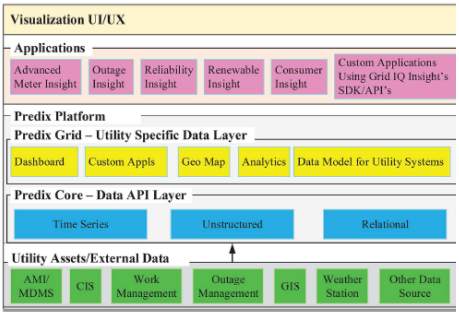

Fig.8. PREDIX Platform for Big Data Analytics

In this platform, the layer in the bottom is physical layer. This physical layer consists of operational systems, utility assets and some external data from the various data sources. Furthermore, layer 2 is specific for utility data and it is based on the cloud-based API. In addition, third layer involves applications specific to grid and lastly fourth layer includes integration with operations and the visualizations.

Other than this, there are various architectures of big data that are working in the smart grid/smart power applications. Some of them are: Booz Allen Hamilton architecture, IBM big data architecture, SAP big data architecture etc.

Following figures of the above-mentioned big data architectures.

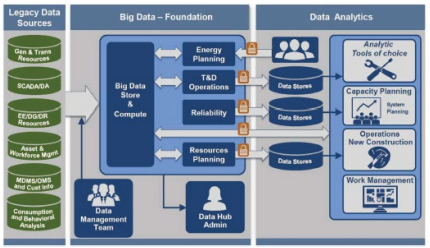

Fig.9. IBM-based Lockheed Martin Big Data Analytics Architecture for Smart Grid Applications

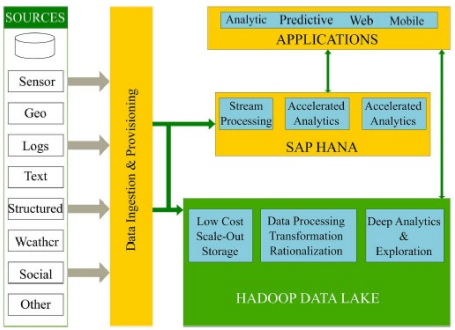

Fig.10. SAP Reference Architecture for Big Data Processing

VI. Analysis Techniques and Solution Approaches

With the advancements, there are number of big data analysis techniques. These techniques and

tools are used to perform necessary data analysis.

a- Hadoop:

It is an open-source framework which is used to process large amount of data. This tool is useful for data storing and processing techniques. It used MapReduce technique to process the large data sets. Its storage part is known ~~has~~ as HDFS (Hadoop Distributed File System). It works on the basis of distribution of large files into small blocks and divides them to different nodes to process. This tool has many benefits as it provides high availability as well as high tolerance against various failures of the hardware [2]. There are number of tools that are based on the Hadoop framework but utilities are mostly using OSI-Soft which is based on the Hadoop. OSI-Soft is being used in the data analytics departments in the utilities to analyze the data in the real time in the smart grid/smart power applications.

b- Spark:

It is an open-source big data processing engine. It is reliable and fast. It is an in-memory engine which is designed to make sure that disks do not suffer I/O limitations. It allows the data to stay in the cache and hence has benefit over Hadoop, that is, disk over head limitation. It is designed for large scale data processing. It is 100 times faster than Hadoop MapReduce. Also, it is 10 times faster when the data on the disk resides [2].

In addition, there are various data analytical tools that are illustrated in the below table. In detail description is shown in the tabular form as shown in figure 11

Platform	Data scaling	Scalability	Fault tolerance	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]

Fig.11. Summary of Big Data Analytical Tools

Category	Algorithm	description
Supervised Learning	Decision tree	A non-parametric method with a tree-like method whose leaves represent class labels and branches represent conjunctions of features
	Naive Bayes	A probabilistic method based on Bayes theorem with the assumption of independence between every pair of features
	Support vector machine classifier	An algorithm to find a separating hyperplane between the two classes by mapping the labeled data to a high-dimensional feature space
	K Nearest Neighbor	A non-parametric method based on the minimum dissimilarity between new items and the labelled items in different classes
	Random Forest	An algorithm consisting of a collection of simple tree predictors independently for the estimation of the final outcome
Unsupervised Learning	K-means	An unsupervised learning method with a given number of clusters to sort the data based on the average value of data in each group as the centroid
	K-medoids	An unsupervised learning method similar to k-means by assigning the centroid of each group with an existing data point instead of the average value
	Hierarchical Clustering	An alternative approach which aims to build a hierarchy of clusters in a dendrogram without a given number of clusters
	DBSCAN	A density-based clustering algorithm to identify clusters with specific shape in distribution
	Expectation-Maximization	An iterative way to approximate the maximum likelihood estimates for model parameters
Correlation	FP-Growth Algorithm	An efficient method for mining the complete set of frequent patterns with a special data structure named frequent-pattern tree with all the association information reserved
	Apriori Algorithm	A classical data analytics algorithm to discover the potential association rules among frequent items
Dimensionality reduction	Principal Component Analysis	An orthogonal transformation of data with a new coordinate system with the greatest variance projected to the first coordinate
	Self-organizing Map	A type of artificial neural network for a low-dimensional representation of the training data space
	Random Matrix	An algorithm which reveal potential regulations with high order matrices for massive data by eigenvalue analysis

Fig.1.2 Data Analytics Algorithms

VII. Applications Of Big Data Analytics In Power Systems

Despite developments, there are still just a few big data applications in the smart grid industry. Advanced sensors and smart metres are smart grid applications that generate large amounts of data for analysis. Besides this, RTUs and SCADA installed on the smart grids are producing bulk amount of data that is being monitored in the data control centres [6].

Furthermore, enhanced demand response is also an application of the big data in smart grid. It enables utilities to identify consumption trends and, as a result of this observation, to distribute power appropriately. In real life example, with the help of careful monitoring, utility people are able to identify the peak load and off-peak load demand in various areas and on the basis of observations made, it is easy for them to distribute energy and install assets as per requirements to meet the demand response [1].

In addition, [1] also discussed the disaggregation and load forecasting, which is further sub-application. With the help of disaggregation and load forecasting, utilities can manage power load across the entire country which is very much helpful for the entities to manage the power system operations. In addition, forecasting also includes financial forecasting which is also economically very much helpful for the country.

Another very important application of the BDA in smart grid/smart power system is predictive fault detection and asset management. Generally, distribution transformers in small areas burn due to overloading or some other technical issues. BDA application in the reliability management helps the entities to protect the equipment by proper monitoring the load of specific area and hence install the equipment as per requirement in the specific area.

Hence, there are many other applications that are not yet aiding the people real-time, but still people are being encouraged to conduct research and integrate the smart grid with the big data analytical operations in order to make systems smarter enough to benefit people.

VIII. Conclusion And Future Directions

It is evident that BDA played a vital role in the development of smart grid. It was determined that we can optimize the power system reliability by integrating

BDA with power systems. In addition, our grid equipment is managed with the help of BDA. Furthermore, we are also able to manage the demand load during the peak and off-peak times. Additionally, with the help of advancements, new tools and techniques that are more reliable are continuously being introduced. For example, Spark has more benefits than Hadoop. Thus, in this survey paper, we reviewed various aspects of the relation of the association between smart grid and BDA. It was concluded that this integration is much more beneficial and will automate the power systems more in the coming era.

REFERENCES

1. H. Akhavan-Hejazi and H. Mohsenian-Rad, "Power systems big data analytics: An assessment of paradigm shift barriers and prospects," *Energy Reports,* vol. 4, pp. 91-100, 2018.
2. B. P. Bhattarai, S. Paudyal, Y. Luo, M. Mohanpurkar, K. Cheung, R. Tonkoski*, et al.*, "Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions," *IET Smart Grid,* vol. 2, pp. 141-154, 2019.
3. C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," *Computer science review,* vol. 17, pp. 70-81, 2015.
4. S. Halford and M. Savage, "Speaking sociologically with big data: Symphonic social science and the future for big data research," *Sociology,* vol. 51, pp. 1132-1148, 2017.
5. A. Munshi and A.-R. M. Yasser, "Big data framework for analytics in smart grids," *Electric Power Systems Research,* vol. 151, pp. 369-380, 2017.
6. Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," *Energy informatics,* vol. 1, pp. 1-24, 2018.
7. Y. Zomaya and Y. C. Lee, *Energy-efficient distributed computing systems*: John Wiley & Sons, 2012.

A. Züfle, G. Trajcevski, D. Pfoser, and J.-S. Kim, "Managing uncertainty in evolving geo-spatial data," in *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, 2020, pp. 5-8.

8. Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou*, et al.*, "Cloud-based software platform for big data analytics in smart grids," *Computing in Science & Engineering,* vol. 15, pp. 38-47, 2013.

9. M. Kezunovic, P. Pinson, Z. Obradovic, S. Grijalva, T. Hong, and R. Bessa, "Big data analytics for future electricity grids," *Electric Power Systems Research,* vol. 189, p. 106788, 2020.

10. E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: A review," *Ieee Access,* vol. 7, pp. 13960-13988, 2019.