A Review of Big Data role in Smart Grid

Rohit Gupta[1], Dr. K.T. Chaturvedi[2]
[1]*PhD Scholar, UIT RGPV Bhopal*
[2]*Associate Professor, UIT RGPV Bhopal*

***Abstract* - There is always a need to modernize energy delivery of traditional power grids, using intelligent devices and big data technologies, this make them smart. The modernization is performed by deploying equipment such as sensors, smart meters, and communication devices, and by invoking procedures such as real-time data processing and big data analysis. A large data is generated by Smart grids all the time which needs to be analysed. This paper presents definitions and background of smart grid and big data. There are many challenges faced in collecting and analysing this data. In this paper current studies and research developments of big data applications in smart grid are discussed. This paper also summarizes the application leveraged by big data technologies, challenges and opportunities are pointed out in this paper as well.**

***Index Terms* - Big data, Smart grid, Big data analytical applications, Cloud platform, Data mining.**

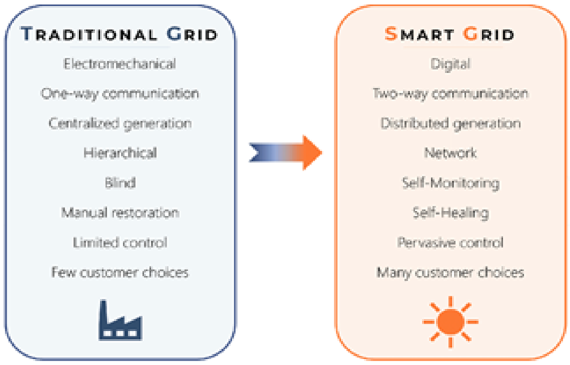
I.INTRODUCTION

Smart Grid (SG) is an important research and development direction in the energy industry. It modifies the conventional power grid by integrating advanced communication and computing methods to improve the entire system control, efficiency, reliability, and safety [1]. Smart grid carries electricity and information between suppliers and consumers, which creates a bidirectional power and information flow system [2]. Many countries have recently adopted smart grid renovation plans [3]. As an example, the ENEL Telegestore project in Italy is the first commercial project utilizing smart grid technology which brings annual savings of approximately 500 million Euros [4, 5]. Smart grids offer several benefits to electric consumers, producers, and operators. SG improves the efficiency, dependability, sustainability, and economics of electric services [6]. Despite its numerous benefits, smart grid is mainly utilized in small regions [6]. There are several roadblocks preventing smart grids from being used in larger regions such as information gathering, storing, processing, and management [7 - 9]. Smart grid requires the capability for processing large volumes of real-time data. For example, in the past, utility companies read meters monthly, but with the Advanced Meter Infrastructure (AMI), meters report data themselves.

II.OVERVIEW OF SMART GRID AND BIG DATA

Smart Grid

The deployment of smart meters and sensors throughout the grid results in massive amounts of data. This includes generation side data (wind farms and photovoltaic plants), consumption side data (residential homes, factories and electric vehicle charging stations), prosumers data (residential photovoltaic panels and vehicle-to-grid) and weather and natural disasters data can be included in the smart grid system. Also, images and video footage could be included to detect physical attacks (California transmission substation sniper attack [23]) or investigate power outages. The smart grid data is considered to be large in volume, high in velocity and wide in variety. The value of this smart grid big data becomes useful when integrated with multi-sourced existing smart grid data in an analytics environment, and can potentially enhance the functionality of the smart grid. Fig. (1) shows the structure of traditional and smart grids [16]. The traditional power grid includes unidirectional transmission, meaning that power flows from power generators to consumers [17]. Smart grid systems, on the other hand include bidirectional transmission, data driven system, and renewable energy resources to offer additional utilities to customers, distributers, and providers [17]. Despite all its benefits, smart grids have difficulty in handling large volume of data within an acceptable time limit and hardware resources.



A. Benefits from big data analytics in smart grid

1)Increasing System Stability & Reliability: Wellbeing is constantly positioned in the first place in the need of force network and can be enveloped two significant angles: dependability and unwavering quality, which can be further sub-isolated into some sub-viewpoints like swaying identification, voltage soundness, occasion recognition and rebuilding, It is commendable pointing that, with the rise of the huge information and progressed information insightful innovation, it is feasible to investigate some new abilities or to improve the obsolete checking and discovery strategies. For example, as revealed in, a swaying in a breeze ranch was identified through PMU however inconspicuous by SCADA, this is a common benefit from the enormous information investigation in the wellbeing of shrewd framework.

2)Increasing Asset Utilization & Efficiency: Practically speaking, the enormous information investigation can expand resource use and efficiency, particularly in better comprehension of the working attributes and actual restrictions of the resources, better approving and adjusting of the models, and better incorporating the inexhaustible assets with huge information instruments. For example, in, the creators utilize the voltages estimated by savvy meters and Geo Information framework (GIS) information to direct transformer exhaustion investigation, which mindful the administrators to upgrade or change the transformer ahead of time. What's more, numerous works have additionally been completed to explore the use of enormous information for model approval and alignment in.

3)Better Customer Experience & Satisfaction: As of late, significant advances have been made for conveying shrewd meters at homes, offices or different premises around the world. The mass rollout empowers simpler charging, misrepresentation identification, admonishing of power outages, shrewd constant evaluating plans, request reaction and efficient energy usage. Be that as it may, all the above applications need high examining rate by the meters and progressed information investigation, just as the data correspondence advances.

Main contributions

Up to this moment, the smart grid and the big data are usually reported separately. However, the analysis of the big data in smart grid is rarely reported. To the best of our knowledge, K Zhou et al. reviewed the big data driven smart energy management in, which mainly illustrates the architecture and industrial applied energy management tools. BA Schuelke-Leech et al. reviewed the opportunities of the big data in electric utilities. However, this work cannot be extended to smart grid. Hu and Vasilakos presented a comprehensive review of the big data application in energy taxonomy and security. In order to get the full utilization of the big data in smart grid, this paper presents a holistically review of the big data issues in smart grid.

In general, the contributions of this paper are manifold and can be summarized as follows:

- It provides a first comprehensive survey covering both smart grid and energy big data analytics. To the best of our knowledge, this is the first attempt to systematically look into the big data issues in smart grid.
- It points out some the latest applications or methods in power system empowered by big data technology. For instance, the data-based approach for fault detection and classification in Micro-grid, which results in a much better performance than the conventional model-based approach; the measurement-based estimation the power flow Jacobian matrix which is proved to be more efficient for inferring the pertinent information about the system topology in near real-time.
- Both implemented data-driven applications by utilities and practical used data analytic methods worldwide have been listed in this paper.

III.BIG DATA APPLICATIONS IN SMART GRID

Wide area situational awareness

In the real-world scenario of wide area situational awareness application, there are two issues need to be solved: the limitation of the installed PMUs and the latency brought by the decision-making algorithms.

Due to the expensive cost and complicated factors of deploying PMUs in the grid, the number of the synchrophasor sensors is limited and need to be optimally placed. Many optimal PMU placement (OPP) methods are raised, such as mixed-integer programming, model- based OPP, zero-injection bused for further reducing [36], generic algorithm [37], etc. [38–40]. Sodhi et al. [41] proposed an improved OPP framework using five applications viz., improving state estimation, assessing voltage /angular stability, monitoring tie-line oscillations and the availability of communication infrastructure, to assess the potential PMU sites.

For transient issue, the response time is typically inside 100 ms with which programmed assurance gadgets make a move without human choice; while for long haul dependability, control room administrators have sufficient opportunity to become familiar with the circumstance by reproductions or encounters. Notwithstanding, for the circumstance between those two situations, choice in a moderately brief timeframe is fundamental. Regardless of that the clumping handling artificial insight (AI) could help the dynamic interaction, the inactivity suffered by numerical estimations is grievous. Choice tree glances promising in managing moderate information measure, while steam digging is fit for huge information measure for dynamic. A decision tree built from data stream using Hoeffding bound was proposed by Domingos and Hulten [42]. A main tree classifier and a cache-based classifier which can handle high-speed data streams are used to facilitate the intelligent decision making. The stream mining techniques require no model information but achieve on-line SA with reasonable accuracy, processing time and computational resources [43–45].

There are some examples of WASA application. The situational system SMDA (ver5.0) was used for wide-area monitoring and event detection in Hydro-Quebec [46]. NYISO used the real-time and off-line data to display the information on the dashboard which alerts operators of anomalies including voltage drop, transient oscillation, line tripping [7]. Peppanen et al. [47] developed a distribution system state estimation (DSSE) and situational awareness system to monitor Georgia Tech campus distribution system and deployed the 3-D graphical user interface to enhance situational awareness. The data collected by Oklahoma Gas & Electric via PMUs to conduct WASA in Oklahoma and western Arkansas

IV.OTHER APPLICATIONS

Table 1. Big Data Applications in Smart Grids – Methods and Case Studies.

Application	Ref. #	Method(s)	Case Studies
Renewable	[28]	The means of communications through long distance or remote stations using energy efficient cellular communication networks.	Off-grid or standalone base stations powered by local small-scale renewables to not require grid power for communication.
	[29]	Multiple models for current, future, and virtual energy markets used to optimize PV integration into a micro grid.	A 65 solar panel array with 15 kWH energy storage is simulated. The system operation is evaluated without any energy sales, with sales restricted to local users, and sales to
	[31]	An enhanced K-means algorithm, named Time Series Clustering (T.S.C) K-means, combined with Multilayer Perceptron Neural Networks (MLPNN) for solar radiation forecasting.	Several meteorological time-series datasets are used to assess the performance of the proposed T.S.C K-means clustering method and its comparison with other clustering techniques including K-means*, K- means^{++}, K-means, self-organizing map (SOM), fuzzy C-means (FCM), and K-Medoids. Solar radiation datasets from different US states are used to evaluate the accuracy performance of the developed hybrid forecasting method and its comparison with state-of-the-art forecasting techniques.

energy	[32]	A novel time-series based K-means clustering method, named T.S.B K-means, combined with discrete Wavelet Transform (DWT), Harmonic Analysis Time Series (HANTS), and MLPNN for wind power forecasting.	Wind speed, wind power, wind direction, and air temperature data from National Renewable Energy Laboratory (NREL) are used to evaluate the novel clustering and hybrid forecasting methods. A comparative analysis of the proposed hybrid method with other well- established forecasting models including Persistence, New Reference (NR), Adaptive Wavelet Neural Network (AWNN), and Phase Space Reconstruction (PSR) are also performed.
	[33]	A Transformation-based K-means algorithm, named TB K-means, combined with MLPNN for solar radiation forecasting.	Several different datasets are used to evaluate the proposed TB K- means clustering and compare it with different variants of K-means algorithm. Solar radiation time series with different characteristics are used to provide a comparative analysis between the proposed hybrid forecasting and benchmark forecasting models.
	[34]	A novel Game Theoretic Self-organizing Map (GTSOM), combined with Neural gas (NG) and Competitive Hebbian Learning (CHL), DWT and Bayesian Neural Network (BNN) for solar	Historical solar radiation data are used to assess the performance of the hybrid forecasting with the proposed GTSOM and other clustering methods.
Demand response	[39], [40]	An extended framework of the Stackelberg game model for demand response optimization.	Homogeneous and heterogeneous generation supply quantities, generator profit and consumer welfare are evaluated in scenarios with few and many generation units and a large consumer population.
Electric vehicle	[49]	Method of defining a more accurate model of electric consumption by light duty Plug-in Electric Vehicles (PEVs).	Uncontrolled home charging of EVs and uncontrolled "opportunistic" charging at public locations are simulated based on travel survey data.
	[51]	A fuzzy expert method for online management of Evs' charging demand.	An IEEE 38 bus distribution test feeder including charging stations at 4 nodes is simulated. .Different charging solutions/scenarios are implemented on the test system and compared.
	[52]	A sliding horizon-based method for real-time data management and optimal coordination of EV charging with photovoltaic (PV) generation.	A 33 bus system including DG units and EV charging stations is simulated. EV charging coordination and its effect on PV power curtailment is evaluated.
	[55]	A hybrid of Auto Regressive Moving Average (ARMA), Fuzzy C-Means (FCM) clustering, Monte Carlo Simulation (MCS), and Particle Swarm Optimization (PSO) methods for optimal scheduling of Evs to increase the use of PV power for EV charging while providing economic revenues for Evs' participation in V2G services.	A 12 MW PV system with 424 Evs is simulated. A collaborative strategy is developed between the EV aggregators and PV producers to minimize the penalty cost of PV over/under-production by charging the Evs using the PV power in excess of the scheduled output and discharging the V2G power to compensate the PV power under- production. The system performance with and without EV optimal charging/discharging are evaluated and compared.
	[56]	A hybrid of ARMA, FCM clustering, MCS, and Genetic Algorithm (GA) methods for optimal scheduling of Evs to increase the use of wind power for EV charging while providing economic revenues for Evs' participation in V2G services.	A 10 MW wind system with 484 Evs is simulated. A bilateral contract is developed between the EV aggregators and wind producers to use the extra wind power for EV charging and to discharge the V2G power during the periods of wind power deficits. The system performance with and without EV optimal charging/discharging are evaluated and compared.

SMART GRID BIG DATA CHALLENGES AND PROPOSED SOLUTIONS

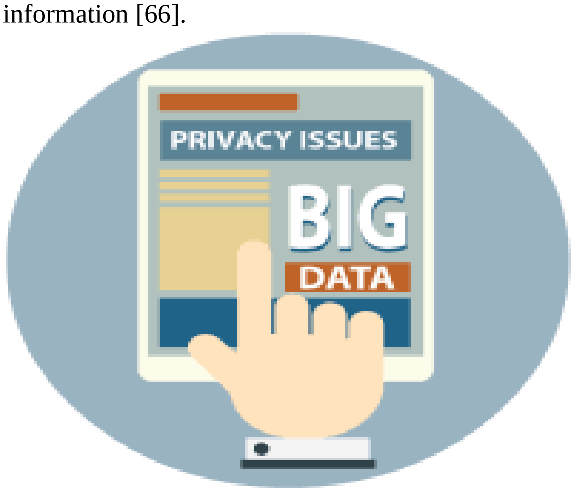
Three main challenges are identified for big data in smart grids: security, quality, and processing location.

BIG DATA SECURITY

The utilization of large information innovation in keen lattices considerably improves the organization network at the cost of expanded security weaknesses [61]. In a major information setting, security openings can be partitioned into three fundamental parts: protection, respectability, and validation.

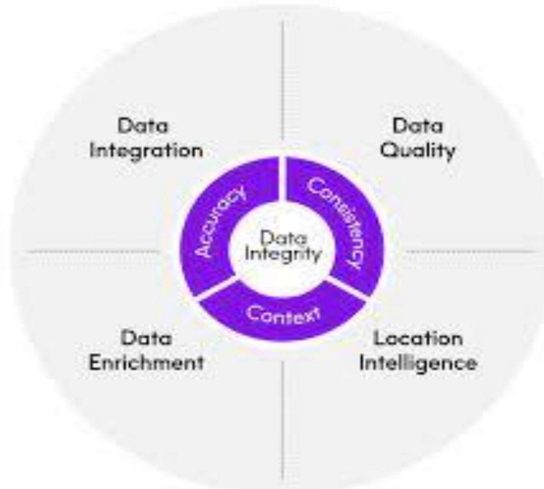
Data Privacy

Shrewd meters can be a fundamental protection concern if their information isn't safely moved and put away [62]. Shrewd meters gather power utilization information of matrix clients. Savvy matrix suppliers break down such information, which gives extraordinary instinct about clients' practices and propensities, to offer keen tweaked administrations [63]. A few strategies have been proposed to dispense with and limit the protection issue. These strategies incorporate yet are not restricted to appropriated steady information assortment strategy [64] and concealing of utilization information inserted data [65]. Since the vast majority of the current arrangements don't think about the tradeoff between expenses of lost protection and information spread (utility), another technique is proposed to fulfill both security and utility prerequisites of brilliant, metered information [66].



Data Integrity

Hazard of trustworthiness assaults is a legitimate concern on the grounds that any infringement of respectability may cause security weaknesses [67]. Client and organization information are generally the objectives for honesty assaults, and any change of such information intrudes on the information correspondence trade and diminishes the whole matrix usefulness [2]. For instance, aggressors can eliminate the more serious level hubs and supplant them with higher likelihood hubs in the force organization, which influences the respectability of information [67].



The information uprightness in shrewd frameworks and energy markets has been widely explored. An investigation introduced the results of virtual offering, which is a strategy for making productive trustworthiness assaulting techniques with no or negligible discovery in energy markets [68]. Another examination showed that information honesty assaults can cause undesirable energy ages and routings, which increment the lattice working expenses [69]. Market incomes and their progressions because of information trustworthiness assaults are utilized as a proportion of foe effect of such assaults [70, 71].

Data Authentication

Clients in savvy networks access the correspondence framework through confirmation, a cycle that checks the client certifications against the records accreditation information base [2]. Confirmation is utilized as an apparatus to distinguish substantial versus non-legitimate personalities inside most of existing security countermeasures [72]. One basic test that keen networks face is message infused assaults. On the off chance that such assaults are not tended to as expected, they can altogether diminish the whole savvy matrix execution [73]. To address such difficulties, a gathering of researchers proposed a confirmation strategy to get keen network information correspondence trade with the utilization of Merkle hash-tree methods [73]. Another examination proposed a protected message verification instrument by incorporating Diffie-Hellman conventions and hash-based message confirmation strategies [74]. Such construction permits shrewd meters inside the savvy lattices to finish shared message confirmation undertakings with insignificant sign trade and dormancy [74]

V.BIG DATA QUALITY

Data quality refers to identifying and to removing the outliers before transferring the data to the system [75]. Energy power consumption data should have high degrees of quality to ensure correct data analysis and ultimately proper decisions. The quality issues of energy consumption data are categorized into noise data, incomplete data, and outlier data [76].

5.1Noise Data

Generally, any data that is difficult to comprehend and/or to decode by computers is considered noise data, which degrades the data quality [76]. In a smart grid context, logical errors and inconsistent energy consumption data are considered noise [77, 78]. Logical errors are defined as the data that violates any given rules and characteristics [79]. For example, if the daily customer energy consumption data includes 25 hours, it is not logical as it exceeds the maximum 24 hours [76]. Moreover, inconsistent data occurs when data does not follow its previously agreed format [80], or it lacks sense when comparing its individual features [81, 82].

5.2 Incomplete Data

As the smart grid data complexity increases, incompleteness is occasionally observed in energy consumption data. Several methods such as delete tuple and data filing are developed to address incomplete data [82]. Delete tuple method simply removes the entire record with incomplete data. However, this method is not appropriate for cases where the data set has several incomplete observations [76]. In such cases, the incomplete data will be filled using advanced algorithms such as average value, artificial value, and regression analysis [82].

5.3Outlier Data

In statistics, if a point of data is considerably distant from other data points, it is called outlier [83]. In energy consumption data, an outlier may be treated as noise and removed. However, they may hold valuable information and therefore, should be detected to preserve the data quality. One method of detection is data quality mining, which is to audit the data to automatically find outliers [84]. In smart grid systems, outliers should be detected, identified, and analyzed as they contain critical information such as power rationing, device failures, and suspicious indicators among others [85].

Big Data Challenges in Smart Grids and Proposed Solutions.

Challenge	Ref.	Solution	Advantage / Disadvantage
	[63]	A regulatory framework equivalent to Health Insurance Portability and Accountability Act (HIPPA) for smart grid privacy and consumer fraud problems	Would provide clear legislative and legal avenues should problems occur / Bureaucracy would not solve some of the problems provided
	[64]	A distributed incremental aggregation framework for smart meters to protect users' privacy by using homomorphic encryption	Unidirectional functionality not allowing for passing information back to a specific unit; Time delay of communication in possible real time operations; Does not look into malicious or fraudulent data acquisition.

Security	[65]	Using a battery connected between the home and the grid so that anyone looking at the power usage will see a battery charging and not the current profiles of the actual items using power	Makes power usage indistinguishable from one day to the next; Overhead of installation and usage and wear and tear costs of a battery system in a home; Difficult to hide high power usage items such as AC, washer, dryer, *etc*.
	[66]	Privacy vs utility: How to get the best of both worlds without sacrificing too much on either side.	Balanced framework / Gives up privacy information of high power item usage as well as the price of the battery
	[67]	Targeted attacks vs random attacks to smart grid: Building faster and more resilient networks to fend off attacks through the communication networks	Faster networks would entail creating a faster protocol to transfer information; Faster connections mean less encryption or protections increasing privacy and attacker problems.
	[69]	Load Redistribution (LR) attacks: Using Multi-start Benders decomposition to find the most damaging immediate attack.	Good attack prevention strategy for this specific type of attack
	[70]	Proposing strategies to detect and localize malicious attacks	Capable of detecting attacks on multiple locations / The number of locations being attacked expands computation.
Quality	[75]	The data mining-based and the state estimation-based electricity consumption outlier data detection methods	Data mining algorithms are faster and better at detecting outliers than traditional methods / Does not account for missing or redundant data.
	[82]	Developing a data mining prototype system (RMINE) for fault diagnosis or system malfunction detection	Capable of obtaining the minimal diagnostic rule set to derive a logical decision in assisting maintenance engineers to diagnose faults
	[86]	Introducing a new class of attacks, called false data injection attacks against monitoring of PMUs or smart grid sensors	N/A
Processing location	[88]	Using embedded neural networks to analyze edge-based load information.	Offers privacy concerns by identifying what is being used in a specific area.
	[89]	Creating a micro grid out of a smart home	Makes a good framework out of the smart home / Lack of intelligent connections to the grid makes it unusable.

1. Techniques used for big data analysis in smart grid

1.1. The platforms for big data analysis

1. HADOOP

Big Data is sort of incomplete without Hadoop and expert data scientists would know that. An open-source Big Data analytics tools, Hadoop offers massive storage for all kinds of data. With its amazing processing power and capability to handle innumerable tasks, Hadoop never allows you to ponder over hardware failure. Though you need to know Java to work with Hadoop, it's worth every effort. Knowing Hadoop will put you ahead in the recruitment race

Pros:

- Hadoop's core strength is its HDFS (Hadoop Distributed File System), which holds all types of data, video, images, JSON, XML and plain texts across the same file system.

- Very useful for research and development purposes.
- Offers easy data access.
- Extremely scalable

Cons:

- Data redundancy can often cause disk space problems.
- For improved efficiency, I/O operations should have been optimized.

Pricing: With the Apache License, this Big Data Analytics tool is free to use.

2. XPLENTY

This cloud-based Big Data Analytics tool for integrating, analyzing and preparing data brings all data sources together. Its intuitive graphical interface allows you with ETL, ELT or replication. Xplenty is a full toolkit to create low-code and no-code data pipelines. It provides solutions for marketing, distribution, and development.

Pros:

- It is a cloud network that is elastic and scalable.
- You can immediately access a range of data stores and a diverse collection of data transformation components.
- By using the rich expression language of Xplenty, you can incorporate complex data preparation functions.
- It offers a customized and flexible API component.

Cons:

- There is no option for monthly subscription.

3. CDH (CLOUDERA DISTRIBUTION FOR HADOOP)

CDH is a complete open-source Big Data Analytics tool and includes Apache Hadoop, Apache Spark, Apache Impala, and many more on its free distribution site. It enables you to acquire, store, manage, discover, model and distribute limitless data.

Pros:

- Complete and accurate distribution.
- The Hadoop cluster is very well managed by the Cloudera Manager.
- Simple to deploy.
- The administration is less complicated.
- High security and administration

Cons:

- Few complicated user interfaces like CM service charts.
- Several suggested installation methods are confusing.

4. R

R is one of the most comprehensive Big Data analytics tools for statistical analysis. The software ecosystem is open-source, free, multi-paradigm, and diverse. The programming languages are C, Fortran, and R. Most extensively used by statisticians and data miners; its use cases include data processing, data manipulation, analysis, and visualization.

Pros:

- The greatest value of R is the immensity of the ecosystem package.
- Unparalleled Graphics and charting features.

5. CASSANDRA

Apache Cassandra is free of cost Big Data analytics tools designed to handle large quantities of data across many commodity servers, offering high availability. The open-source NoSQL DBMS uses CQL (Cassandra Structure Language) to interact with the database.

Pros:

- There is no single failure point.
- It manages huge data really quick.
- It has log-structured storage and linear scalability.

Cons:

- Extra troubleshooting and maintenance work is required.
- It could have boosted the clustering.
- There is no row-level locking feature.

6. KNIME

KNIME is an abbreviation for Konstanz Information Miner, which is an open-source Big Data Analytics tool. It is used for enterprise reporting, integration, data mining, data analytics, and business intelligence. It supports operating systems such as Linux, and Windows X.

Pros:

- Quick to use ETL
- It is very well integrated with other technologies and languages.
- Rich set of algorithms.
- Workflows are highly functional and structured.
- A lot of manual tasks are automated.
- There are no problems with stability.
- Simple to configure.

Cons:

- It covers nearly the whole of RAM.
- Might have enabled graph database integration.

7. DATAWRAPPER

Datawrapper is an open-source Big Data Analytics tool for data visualization. It enables its users to produce clear, accurate, and embedded charts easily. It is broadly used in newsrooms across the world.

Pros:

- Operates exceptionally well on any type of device – smartphone, laptop, or tablet.
- Rapid and interactive responses.
- Excellent export and customization options.

Cons:

- Has limited options for color palettes.

8. MONGODB

MongoDB is a contemporary alternative to databases. It's one of the best Big Data Analytics tools for working on data sets that vary or change frequently or the ones that are semi or unstructured. Some of the best uses of MongoDB include storage of data from mobile apps, content management systems, product catalogs, and more. Like Hadoop, you can't get started with MongoDB instantly. You need to learn the tool from scratch and be aware of working on queries.

Pros:

- Supports various platforms and technologies.
- No install and maintenance hiccups.
- Robust and cost-effective.

Cons:

- It has a limited analytics resource.

9. LUMIFY

Lumify is one of the open-source Big Data Analytics tools to analyze and visualize large data. This Big Data Analytics tool's key features include full-text search, 2-dimensional and 3-dimensional graphical viewings, automated templates, multimedia analysis, real-time project-or workplace collaboration, to name but a few.

Pros:

- Scalable and secure
- A dedicated full-time development team backs it.
- Supports the cloud-based environment and works excellently with Amazon's AWS.

It is a free Big Data Analytics tool.

10. HPCC

HPCC is an abbreviation for High-Performance Computing Cluster. This open-source Big Data Analytics tool is a complete Big Data solution over a highly scalable supercomputing platform. HPCC is also known as DAS (Data Analytics Supercomputer) and was developed by LexisNexis Risk Solutions. Written in C++ and ECL (Enterprise Control Language), it is based on a Thor architecture that enables data parallelism, pipeline parallelism, and system parallelism.

Pros:

- High performance due to the commodity computing clusters-based architecture.
- Enables parallel data processing.
- Agile, robust and highly scalable.
- Cost-effective and comprehensive

Pricing: It's a free Big Data Analytics tool.

11. STORM

Storm is a cross-platform and open-source Big Data Analytics tool from Apache. Written in Java and

Clojure, Backtype and Twitter are the developers of the storm. Several big brands like Yahoo, Alibaba, and The Weather Channel, to name a few are organizations that use Storm.

Pros:

- There are many applications: real-time analysis, logging, ETL (Extract Transform Load), continuous computation, distributed RPC, machine learning.
- Agile, reliable, and highly scalable.

Cons:

- Difficult to understand and to use.
- Have debugging complexity.

Pricing: It's a free Big Data Analytics tool.

12. RAPIDMINER

Rapidminer is a cross-platform Big Data Analytics tool that provides integrated data science, machine learning, and predictive analysis framework.

Pros:

- Availability of code-optional GUI.
- Well integrated with cloud and APIs.
- Excellent customer support and technical assistance.

Cons:

- Improvements should be made to online data services.

13. QUBOLE

Qubole Data Service is a Big Data Analytics tool that administrates, learns, and optimizes its use independently. This helps the data team to focus on business performance.

Pros:

- Highly flexible and optimized scalability.
- Improved Big Data Analytics adoption.
- Simple to use.
- Accessible worldwide in all AWS domains.

14. TABLEAU

Tableau is a Big Data Analytics tool that offers various integrated solutions that help the world's biggest organizations visualize and understand their data. It provides custom dashboards in real-time and can manage all the data sizes, and can be easily accessed by technical and non-technical professionals. It is one of the best Big Data Analytics tools for data visualization and exploration.

Pros:

- Impeccable Data blending capabilities.
- Provides a bouquet of intelligent characteristics.
- Outstanding and quick support for connection with most of the databases.

Cons:

- Could provide an integrated deployment and migration tool between different table servers and environments.

15. SAMOA

SAMOA is an abbreviation for Scalable Advanced Massive Online Analysis. It is an open-source Big Data Analytics tool for big data stream mining and machine learning. It enables you to build ML algorithms and run them on many DSPEs(Distributed streaming learning devices (distributed stream processing engines).

Pros:

- Simple to use, highly scalable and fast.
- Based on Write Once Run Anywhere (WORA) architecture.

Pricing: It's a free Big Data Analytics tool.

VI.FUTURE OF BIG DATA IN SMART GRIDS

The future of research in big data use in smart grids is diverse. Big data offers many solutions to the bi-directional flow of information as well as processing and analyzing that information. For a smart grid, big data will be a necessity for realizing the best possible solutions for how we as a society should distribute and utilize renewables as well as how to analyze systems for abnormal conditions such as faults or power outages. The future of the smart grid will depend on building these frameworks such that they can be implemented and utilized in a meaningful way. This will include the planning to real time operation for generators and consumers for current practices to those planned for by 2050 [91].

VII.CONCLUSION

This paper presents the definitions and applications of integrating big data technologies in smart grid systems based on current studies and research developments. Several research articles are reviewed to understand the current challenges and solutions of big data applications in smart grids and to identify research gaps. Thus, this survey provides new directions to further investigate such applications and challenges to propose innovative solutions for filling the identified research gaps.

REFERENCES

[1] Chen HH, Chen S, Lan Y. Attaining a sustainable competitive advantage in the smart grid industry of China using suitable open innovation intermediaries. Renew Sustain Energy Rev 2016;62:1083–91.

[2] Botte B, Cannatelli V, Rogai S. The telegestore project in Enel's metering system[C]. In: Proceedings of the international conference and exhibition on electricity distribution. IET; 2005. p. 1–4.

[3] National energy technology laboratory for the U.S. Department of energy, modern

[4] Inovgrid. [Online]; 2010. Available: ⟨http://www.inovcity.pt/en/pages/inovgrid. aspx⟩.

[5] Moma. [Online]; 2012. Available: ⟨http://www.modellstadt-mannheim.de⟩.

[6] Farhangi H. The path of the smart grid. Power Energy Mag IEEE 2010;8(1):18–28.

[7] DOE. Advancement of synchrophasor technology in ARRA Projects; 2016. ⟨https://www. smartgrid.gov/recovery_act/program_publications.html⟩

[8] Yang B, Yamazaki J. Big data analytic empowered grid applications- Is PMU a big data issue? In: Proceedings of the 2015 12th international conference on the European energy market (EEM). IEEE; 2015. p.1–4.

[9] DOE. Summary of the North American synchrophasor initiative (NASPI) activity area; 2015. ⟨https://www.naspi.org/documents⟩.

[10] Yuan J, Shen J, Pan L, et al. Smart grids in China. Renew Sustain Energy Rev 2014;37(3):896–906.

[11] Alahakoon D, Yu X. Smart electricity meter data intelligence for future energy systems: a survey. IEEE Trans Ind Inf 2016;12(1):425–36.

[12] Klump R, Agarwal P, Tate JE, Khurana H, Lossless compression of synchronized phasor measurements. In Proceedings of the IEEE power and energy society general meeting. Minneapolis, MN, USA; Jul. 2010. p. 1–7.

[13] Electric Power Group. Wind farm oscillation detection and mitigation. ⟨https:// www.smartgrid.gov/recovery_act/project_information.html⟩.

[14] Short TA. Advanced metering for phase identification, transformer identification, and secondary modeling. IEEE Trans Smart Grid 2013;4(2):651–8.

[15] Overholt P, Kosterev D, Eto J, et al. Improving reliability through better models: using synchrophasor data to validate power plant models. IEEE Power Energy Mag 2014;12(3):44–51.

[16] Luan W, Peng J, Maras M, et al. Smart meter data analytics for distribution network connectivity verification. IEEE Trans Smart Grid 2015;6(4), [1–1].

[17] Berrisford AJ. A tale of two transformers: an algorithm for estimating distribution secondary electric parameters using smart meter data. Electrical and computer engineering (CCECE). In: Proceedings of the 2013 26th annual IEEE Canadian conference. IEEE; 2013. p. 1–6.

[18] Hajnoroozi AA, Aminifar F, Ayoubzadeh H. Generating unit model validation and calibration through synchrophasor measurements. IEEE Trans Smart Grid 2015;6(1):441–9.

[19] Bolognani S, Bof N, Michelotti D. et al. Identification of power distribution network topology via voltage correlation analysis. In: Proceedings of the IEEE conference on decision and control; 2013. p. 1659–64.

[20] Erlinghagen S, Lichtensteiger B, Markard J. Smart meter communication stan- dards in Europe – a comparison. Renew Sustain Energy Rev 2015;43:1249–62.

[21] Zhou K, Yang S. Understanding household energy consumption behavior: the contribution of energy big data analytics. Renew Sustain Energy Rev 2016;56:810–9.

[22] Tuballa ML, Abundo ML. A review of the development of Smart Grid technologies. Renew Sustain Energy Rev 2016;59:710–25.

[23] Haider HT, See OH, Elmenreich W. A review of residential demand response of smart grid. Renew Sustain Energy Rev 2016;59:166–78.

[24] Moretti M, Djomo SN, Azadi H, et al. A systematic review of environmental and economic impacts of smart grids. Renew Sustain Energy Rev 2017:68.

[25] Ali SM, Jawad M, Khan B, et al. Wide area smart grid architectural model and control: a survey. Renew Sustain Energy Rev 2016;64:311–28.

[26] Manyika J, Chui M, Brown B, et al. Big data: the next frontier for innovation, competition, and productivity. Analytics 2011.

[27] Naimi AI, Westreich DJ. Big data: a revolution that will transform how we live, work, and think. Am J Epidemiol 2014;17(9):181–3.

[28] Zhou K, Fu C, Yang S. Big data driven smart energy management: from big data to big insights. Renew Sustain Energy Rev 2016;56:215–25.

[29] Schuelke-Leech BA, Barry B, Muratori M, et al. Big data issues and opportunities for electric utilities. Renew Sustain Energy Rev 2015;52(2):937–47.

[30] Hu J, Vasilakos A. Energy big data analytics and security: challenges and opportunities. IEEE Trans Smart Grid 2016;7(5):2423–36, [2016, 7(5):1-1].

[31] Mishra DP, Samantaray SR, Joos G. A combined wavelet and data mining based intelligent protection scheme for microgrid. IEEE Trans Smart Grid 2016;7(5):2295–304.

[32] Chen YC, Wang J, Dominguez-Garcia AD, et al. Measurement-based estimation of the power flow Jacobian matrix. IEEE Trans Smart Grid 2016;7(5):2507–15.

[33] Endsley MR, Connors ES. Situation awareness: State of the art[C]. In: Proceedings of the IEEE power & energy society general meeting-conversion & delivery of electrical energy in century. IEEE; 2008. p. 1–4.

[34] Aminifar F, Fotuhi-Firuzabad M, Shahidehpour , et al. A probabilistic multistage PMU placement in electric power systems. IEEE Trans Power Deliv 2011;26(2):841–9.

[35] Gopakumar P, Chandra G, Reddy M, et al. Optimal placement of PMUs for the smart grid implementation in Indian power grid – a case study. Front Energy 2013;7(3):358–72.

[36] Xu B, Abur A. Observability analysis and measurement placement for systems with PMUs. In: Proceedings of the IEEE PES power systems conference and exposition. Vol. 2; October 2004. p. 943–6.

[37] Milosevic B, Begovic M. Nondominated sorting genetic algorithm for optimal phasor measurement placement. IEEE Trans Power Syst 2003;18(1):69–75.

[38] Western interconnection synchrophasor program; June 2014. ⟨https://www. smartgrid.gov/ project/western⟩.

[39] Huang J, Wu N, Ruschmann M. Data-availability-constrained placement of PMUs and communication links in a power system. IEEE Syst J 2014;8(2):483–92.

[40] Li Q, Cui T, Weng Y, Negi R, et al. An information-theoretic approach to PMU

[41] Sodhi R, Srivastava SC, Singh SN. Multi-criteria decision-making approach for multistage optimal placement of phasor measurement units. IET Gener Transm Distrib 2011;5(2):181–90.

[42] Domingos P, Hulten G. Mining high-speed data streams. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000. p. 71–80.

[43] Dahal N, Abuomar O, King R, et al. Event stream processing for improved situational awareness in the smart grid. Expert Syst Appl 2015;42(20):6853–63.

[44] Ma S, Fang S, Yuan D, et al. The design of power security system in smart home based on the stream data mining. Adv Data Min Appl 2014:716–24.

[45] Omitaomu OA. A control chart approach for representing and mining data streams with shape-based similarity; 2014.

[46] Basu C, Agrawal A, Hazra J. et al. Understanding events for wide-area situational awareness. In: Proceedings of the innovative smart grid technologies conference. IEEE; 2014. p. 1–5.

[47] Peppanen J, Reno MJ, Thakkar M, et al. Leveraging AMI data for distribution system

model calibration and situational awareness. IEEE Trans Smart Grid 2015;6(4):2050–9.

[48] Schweppe FC, Wildes J. Power system static-state estimation,part I: exact model.IEEE Trans Power Appl Syst . 1970;PAS-89(1):120–5.

[49] Zhao J, Zhang G, Das K, et al. Power system real-time monitoring by using PMU- based robust state estimation method. IEEE Trans Smart Grid 2015;7(1):1.

[50] Pignati M, Zanni L, Sarri S. et al. A pre-estimation filtering process of bad data for linear power systems state estimators using PMUs. In: Proceedings of the power systems computation conference; 2014. p. 1–8.

[51] Xie L, Chen Y, Kumar PR. Dimensionality reduction of synchrophasor data for early event detection: linearized analysis. IEEE Trans Power Syst 2014;29(6):2784–94.

[52] DOE. Project information; 2016. ⟨https://www.smartgrid.gov/recovery_act/ project_information.html⟩.

[53] Dahal OP, Brahma SM, Cao H. Comprehensive clustering of disturbance events recorded by phasor measurement units. IEEE Trans Power Deliv 2014;29(3):1390–7.

[54] Chen Y, Xie L, Kumar PR. Power system event classification via dimensionality reduction of synchrophasor data. In: Proceedings of the IEEE, sensor array and multichannel signal processing workshop; 2014. p. 57–60.

[55] IEEE Std C37.118.1™‒2011. IEEE standard for synchrophasor measurements for power systems; 2011. p. 1–61.

[56] IEEE Std C37.118.2™‒2011. IEEE standard for synchrophasor measurements for power systems; 2011. p. 1–53.

[57] Khan M, Li M, Ashton P. et al. Big data analytics on PMU measurements. In: Proceedings of the international conference on fuzzy systems and knowledge discovery; 2014.

[58] Kezunovic M, Xie L, Grijalva S. The role of big data in improving power system operation and protection. In: Proceedings of the bulk power system dynamics and control – IX optimization, security and control of the emerging power grid (IREP), 2013 IREP symposium; 2013. p. 1–9.

[59] Meier R, Cotilla-Sanchez E, Mccamish B, et al. Power system data management and analysis using synchrophasor data. Technol Sustain IEEE 2014:225–31.

[60] Khan AR, Mahmood A, Safdar A, et al. Load forecasting, dynamic pricing and DSM in smart grid: a review. Renew Sustain Energy Rev 2016;54:1311–22.

[61] Selakov A, Cvijetinović D, Milović L, et al. Hybrid PSO–SVM method for short- term load forecasting during periods with significant temperature variations in city of Burbank. Appl Soft Comput 2014;16(3):80–8.

[62] Deihimi A, Orang O, Showkati H. Short-term electric load and temperature forecasting using wavelet echo state networks with neural reconstruction. Energy 2013;57(3):382–401.

[63] Zhu L, Wu QH, Li MS. et al. Support vector regression-based short-term wind power prediction with false neighbors filtered. In: Proceedings of the international conference on renewable energy research and applications. IEEE; 2013. p. 740–4.

[64] Rocio Cogollo M, Velasquez JD. Methodological advances in artificial neural networks for time series forecasting. IEEE Lat Am Trans 2014;12(4):764–71.

[65] Huang SC, Lu CN, Lo YL. Evaluation of AMI and SCADA data synergy for distribution feeder modeling. IEEE Trans Smart Grid 2015;6(4), [1–1].

[66] Berrisford A J. A tale of two transformers: an algorithm for estimating distribution secondary electric parameters using smart meter data; 2013. p. 1–6.

[67] Kwac J, Rajagopal R. Data-driven targeting of customers for demand response. IEEE Trans Smart Grid, IEEE Trans Smart Grid 2016;7(5):2199–207.

[68] Peppanen J, Reno MJ, Broderick RJ, et al. Distribution system model calibration with big data from AMI and PV inverters. IEEE Trans Smart Grid 2016;7(5):2497–506.

[69] Pan S, Morris T, Adhikari U. Developing a hybrid intrusion detection system using data mining for power systems. IEEE Trans Smart Grid 2015;6(6):134–43.

[70] Bayrak G, Kabalci E. Implementation of a new remote islanding detection method for wind–solar hybrid power plants. Renew Sustain Energy Rev 2016;58:1–15.

[71] Li C, Higuma K, Watanabe M, et al. Monitoring and estimation of interarea power oscillation mode Based on application of CampusWAMS. Turk J Agric - Food Sci Technol 2008;3(1).

[72] Dasgupta S, Paramasivam M, Vaidya U, et al. PMU-based model-free approach for real-time rotor angle monitoring. IEEE Trans Power Syst 2015;30(5):2818–9.

[73] Liu Y, Zhan L, Zhang Y, et al. Wide-area measurement system development at the distribution level: an FNET/GridEye example. IEEE Trans Power Deliv 2015;31(2), [1–1].

[74] Chai J, Liu Y, Guo J. et al. Wide-area measurement data analytics using FNET/ GridEye: a review[C]. In: Proceedings of the power systems computation conference; 2016. p. 1–6.

[75] Kamwa J, Beland G, Trudel. et al. Wide-area monitoring and control at hydro- québec: past, present and future. In: Proceedings of the power engineering society general Meeting; 2006.

[76] e-terraphasorpoint 3.0. ⟨http://www.alstom.com/Global/Grid/Resources/ Documents/Automation/NMS/e-terraphasorpoint%203.0.pdf⟩.

[77] Siemens.⟨http://www.engerati.com/resources/efficient-network-integration-renewable-energy-resources-distribution-level⟩; 2015.

[78] NARI.⟨http://www.naritech.cn/html/jie166.shtm⟩; 2015.

[79] Chang V, Wills G. A model to compare cloud and non-cloud storage of big data.Future Gener Comput Syst 2016;57:56–76.

[80] Singh M, Ali A. Big data analytics with Microsoft HDInsight in 24h, Sams Teach Yourself: Big Data, Hadoop, and Microsoft Azure for better business intelligence; 2016.

[81] Rusitschka S, Eger K, Gerdes C. Smart grid data cloud: a model for utilizing cloud computing in the smart grid domain. In: Proceedings of the IEEE international conference on smart grid communications; 2010. p. 483–8.

[82] Birman KP, Ganesh L, Renesse RV. et al. Running smart grid control software on cloud computing architectures. In: Proceedings of the the workshop on computational needs for the next generation electric grid; 2012.

[83] Baek J, Vu QH, Liu JK, et al. A secure cloud computing-based framework for big data information management of smart grid. IEEE Trans Cloud Comput 2015;3(2):233–44.

[84] Simmhan Y, Prasanna V, Aman S, et al. Cloud-based software platform for big data analytics in smart grids. Comput Sci Eng 2013;15(4):38–47.

[85] Khan Mukhtaj. Hadoop performance modeling and job optimization for big data analytics. London: Brunel University; 2015.

[86] Markovic DS, Zivkovic D, Branovic I, et al. Smart power grid and cloud computing. Renew Sustain Energy Rev 2013;24(1):566–77.

[87] Fang B, Yin X, Tan Y, et al. The contributions of cloud technologies to smart grid. Renew Sustain Energy Rev 2016;59:1326–31.

[88] Apache Hadoop. ⟨http://hadoop.apache.org/⟩; 2016.

[89] Zhang Y, Chen S, Wang Q, et al. i2MapReduce: incremental MapReduce for mining evolving big data. IEEE Trans Knowl Data Eng 2015;27(7):1.

[90] Xing EP, Ho Q, Dai W. et al. Petuum: a new platform for distributed machine learning on big data. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2015. p. 1–1.

[91] Shyam R, Bharathi GHB, Sachin KS, et al. Apache spark a big data analytics platform for smart grid. 21; 2015. p. 171–8.

[92] Apache Spark. ⟨http://spark.apache.org/⟩; 2016.

[93] Zhou D, Guo J, Zhang Y, et al. Distributed data analytics platform for wide area synchrophasor measurement systems. IEEE Trans Smart Grid 2016;7(5):2397–405.

[94] Li D. Machine learning aided decision making and adaptive stochastic control in a hierarchical interactive smart grid. Dissertations & theses - gradworks; 2014.

[95] Wu X, Zhu X, Wu GQ, et al. Data mining with big data. IEEE Trans Knowl Data Eng 2014;26(1):97–107.

[96] Al-Jarrah OY, Yoo PD, Muhaidat S, et al. Efficient machine learning for big data: a review. Big Data Res 2015;2(3):87–93.

[97] Khazaei J, Fan L, Jiang W, et al. Distributed prony analysis for real-world PMU data. Electr Power Syst Res 2016;133:113–20.

[98] Shaker H, Zareipour H, Wood D. A data-driven approach for estimating the power generation of invisible solar sites. IEEE Trans Smart Grid 2016;7(5):2466–76.

[99] Shaker H, Zareipour H, Wood D. Estimating power generation of invisible solar sites using publicly available data. IEEE Trans Smart Grid 2016;7(5):2456–65.

[100] Taieb SB, Huser R, Hyndman RJ, et al. Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. IEEE Trans Smart Grid 2016;7(5):2448–55.

[101] Green RC, Wang L, Alam M. Applications and trends of high-performance computing for electric power systems: focusing on smart grid. IEEE Trans Smart Grid 2013;4(2):922–3.

[102] Partridge M, Calvo RA. Fast dimensionality reduction and simple PCA. Intell Data Anal 1998;2(1–4):203–14.

[103] Poekaew P, Champrasert P. Adaptive-PCA: an event-based data aggregation using principal component analysis for WSNs. In: Proceedings of the international conference on smart sensors and application. IEEE; 2015.

[104] Nuchprayoon S. Electricity load classification using K-means clustering algorithm. In: Proceedings of the Brunei international conference on engineering and technology. IET; 2014. p. 1–5.

[105] Yang H, Zhang J, Qiu J, et al. A practical pricing approach to smart grid demand response based on load classification. IEEE Trans Smart Grid 2016, [1–1].

[106] Mu FL, Li HY. Power load classification based on spectral clustering of dual scale. In: Proceedings of the IEEE international conference on control science and systems engineering. IEEE; 2014.

[107] Ye M, Liu Y, Xiong W. et al. Voltage stability research of receiving-end network based on real-time classification load model. In: Proceedings of the IEEE international conference on mechatronics and automation; 2014. p. 1866-70.

[108] Che JX, Wang JZ. Short-term load forecasting using a kernel-based support vector regression combination model. Appl Energy 2014;132(11):602–9.

[109] Liu N, Tang Q, Zhang J, et al. A hybrid forecasting model with parameter optimization for short-term load forecasting of micro-grids. Appl Energy 2014;129:336–45.

[110] Rahman S. Formation and analysis of a rule-based short-term load forecasting algorithm. In: Proceedings of the IEEE (Institute of electrical and electronics engineers), (USA). 78:5(5); 2015. p. 805–16.

[111] Hernández L, Baladrón C, Aguiar JM, et al. Artificial neural networks for short-term load forecasting in microgrids environment. Energy 2014;75(C):252–64.

[112] Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs. Proc IEEE 2015;104(1):11–33.

[113] Drumond LR, Diaz-Aviles E, Schmidt-Thieme L. et al. Optimizing multi-relational factorization models for multiple target relations. In: Proceedings of the ACM international conference; 2014. p. 191–200.

[114] Schweppe F, Wildes J. Power system static-state estimation, part I, II, III. IEEE Trans Power Appl Syst . 1970;PAS-89(1):120–35.

[115] Xu X, He X, Ai Q, et al. A correlation analysis method for power systems based on random matrix theory. IEEE Trans Smart Grid 2015.

[116] Yang Q. Cross-domain data fusion. Computer 2016;49(4), [18–18].

[117] Diou C, Stephanopoulos G, Panagiotopoulos P, et al. Large-scale concept detection in multimedia data using small training sets and cross-domain concept fusion. IEEE Trans Circuits Syst Video Technol 2011;20(12):1808–21.

[118] Tsai CC, Chang-Chien LR, Chen IJ, et al. Practical considerations to calibrate generator model parameters using phasor measurements. IEEE Trans Smart Grid 2016:1–11.

[119] Hana S. SAP AG, in-memory database, computer appliance. Vertpress; 2012.

[120] Kennedy, Stephen James MCP. M.C.P. Massachusetts Institute of Technology Transforming big data into knowledge: experimental techniques in dynamic visualization. Massachusetts Institute of Technology; 2012.

[121] Cuzzocrea A. Privacy and security of big data: current challenges and future research perspectives. In: Proceedings of the international workshop on privacy & secuirty of big data. ACM; 2014. p. 45–7.

[122] Bertino E. Big data – security and privacy. In: Proceedings of the IEEE international congress on big data. IEEE; 2015.
[123] Depuru SSSR, Wang L, Devabhaktuni V. Smart meters for power grid: challenges, issues, advantages and status. Renew Sustain Energy Rev 2011;15(6):2736–42.
[124] Kabalci Y. A survey on smart metering and smart grid communication. Renew Sustain Energy Rev 2016;57:302–18.
[125] Hashem IAT, Yaqoob I, Anuar NB, et al. The rise of "big data" on cloud computing: review and open research issues. Inf Syst 2015;47(47):98–115.
[126] Jagadish HV, Gehrke J, Labrinidis A, et al. Big data and its technical challenges. Commun Acm 2014;57(7):86–94.
[127] Schweppe FC, Rom DB. Power system static-state estimation, part II: approximate model. IEEE Trans Power Appl Syst 1970;PAS-89(1):125–30.
[128] Schweppe FC. Power system static-state estimation, part III: implementation. IEEE Trans Power Appl Syst 1970;PAS-89(1):130–5.
[129] Gol M, Abur A. A fast decoupled state estimator for systems measured by PMUs. IEEE Trans Power Syst 2015;30(5):2766–71.
[130] Jokar P, Arianpoo N, Leung VCM. Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans Smart Grid 2016;7(1):216–26.
[131] Shuai Z, Sun Y, Shen Z, et al. Microgrid stability: classification and a review. Renew Sustain Energy Rev 2016;58:167–79.