Hindawi
Complexity
Volume 2018, Article ID 8496187, 21 pages
https://doi.org/10.1155/2018/8496187

Review Article

Complex Power System Status Monitoring and Evaluation Using Big Data Platform and Machine Learning Algorithms: A Review and a Case Study

Yuanjun Guo, Zhile Yang, Shengzhong Feng, and Jinxing Hu

Shenzhen Institute of Advanced Technology Chinese Academy of Sciences, Shenzhen, Guangdong 5108055, China

Correspondence should be addressed to Zhile Yang; zyang07@qub.ac.uk

Received 17 April 2018; Accepted 7 August 2018; Published 20 September 2018

Academic Editor: Ireneusz Czarnowski

Efficient and valuable strategies provided by large amount of available data are urgently needed for a sustainable electricity system that includes smart grid technologies and very complex power system situations. Big Data technologies including Big Data management and utilization based on increasingly collected data from every component of the power grid are crucial for the successful deployment and monitoring of the system. This paper reviews the key technologies of Big Data management and intelligent machine learning methods for complex power systems. Based on a comprehensive study of power system and Big Data, several challenges are summarized to unlock the potential of Big Data technology in the application of smart grid. This paper proposed a modified and optimized structure of the Big Data processing platform according to the power data sources and different structures. Numerous open-sourced Big Data analytical tools and software are integrated as modules of the analytic engine, and self-developed advanced algorithms are also designed. The proposed framework comprises a data interface, a Big Data management, analytic engine as well as the applications, and display module. To fully investigate the proposed structure, three major applications are introduced: development of power grid topology and parallel computing using CIM files, high-efficiency load-shedding calculation, and power system transmission line tripping analysis using 3D visualization. The real-system cases demonstrate the effectiveness and great potential of the Big Data platform; therefore, data resources can achieve their full potential value for strategies and decision-making for smart grid. The proposed platform can provide a technical solution to the multidisciplinary cooperation of Big Data technology and smart grid monitoring.

1. Introduction

Along with the fast installation of computers and communication smart devices, the power industry is also experiencing tremendous changes both in the scale of power grid and in the system complexity. To build up a modern combined energy system of various types of energies including gas, cold, and heat, based on the smart power system, has become a trend of development in the energy industry. As discussed in many literatures [1–3], a modern energy system has several major features: (1) high penetration of new energy resources are supported and utilized effectively; (2) it provides complementation and integration of different types of energies such as electricity, gas, cold, and heat; and (3) an interconnected and relatively open system, distributed resources, and a consumption side are extensively involved. A huge amount of measurement data including production, operation, control, trading, and consumption are continuously collected, communicated, and processed in an amazing speed faster than any period of history [4].

Appropriate and efficient data management and analysis systems are urgently needed to leverage massive volumes of heterogeneous data in unstructured text, audio, and video formats; furthermore, useful information needs to be extracted and shared to meet the fast-growing demands of high-accuracy and real-time performance of modern power and energy systems [5]. Hidden values in power system big data cannot be effectively revealed by means of traditional power system analysis; therefore, Big Data technology and analytics are also in desperate need.

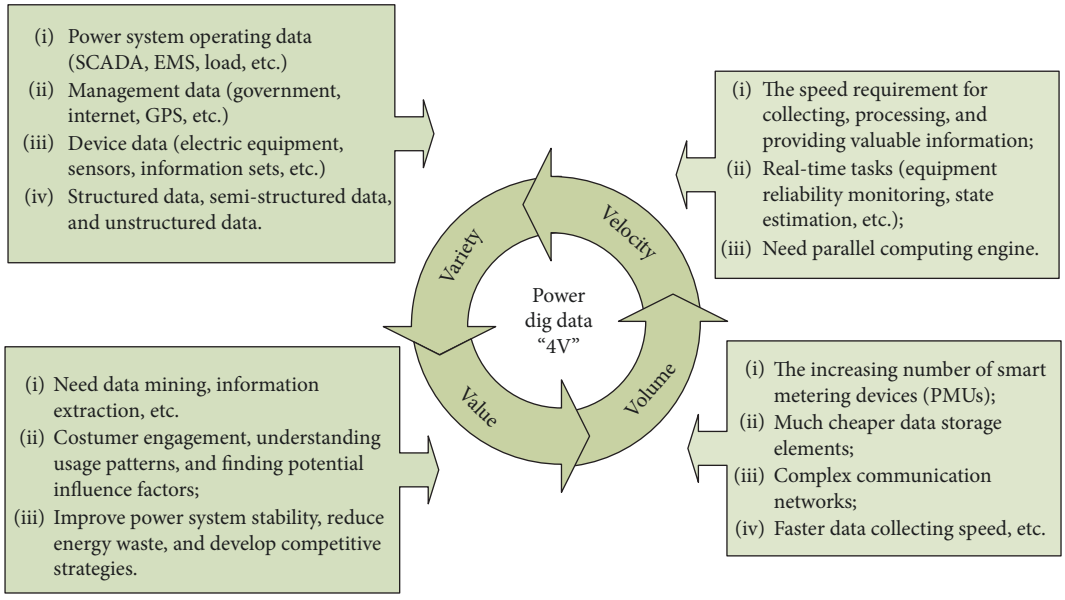


Figure 1: Power system Big Data 4V characteristics.

The Chinese power industry has considerable interests in Big Data analytics associated with power generation and management in order to effectively cope with severe challenges such as limited resources and environmental pollutions, among many others [6]. Actually, Big Data technology has already been successfully applied as a powerful data-driven tool for solving numerous new challenges in power grid, such as price forecasting [7, 8], load forecasting [9], transient stability assessment [10], outlier detection [11], and fault detection and analysis [12], among others [13, 14]. Detailed discussions about Big Data issues and application are reviewed in [15], as well as the insights of Big Data-driven smart energy management in [16]. Major tasks of the architecture for these applications are similar, which focus on two major issues: big power data modeling and big power data analysis.

1.1. Power Grid and Big Data. Supervisory control and data acquisition (SCADA) devices are mainly used in traditional power industries to collect data and to secure grid operations, providing redundant measurements including active and reactive power flows and injections and bus voltage magnitudes [17]. However, the sampling rate of SCADA is slow, and unlike traditional SCADA systems, the phasor measurement unit (PMU) is able to measure the voltage phasor of the installed bus and the current phasors of all the lines connected with that bus. In particular, PMUs are collecting data at a sampling rate of 100 samples per second or higher; therefore, a huge amount of data needs to be collected and managed. To be specific, the Pacific Gas and Electric Company in the USA collects over 3 TB power data from 9 million smart meters across the state grid [18]. The State Grid Corporation of China owns over 2.4 hundred million smart meters, making the total amount of collected data reach 200 TB for a year, while the total number of data in information centers can achieve up to 15 PB. Big Data is also often recognized as challenging in data volume, variety, velocity, and value in many applications [19, 20], and the "4V" characteristics are reflected in the following aspects considering applications in the power system, which is illustrated in Figure 1.

It is possible to get insights from the power system overall Big Data to improve the power efficiency, potentially influence factors of the power system status, understand power consumption patterns, predict the equipment usage condition, and develop competitive marketing strategies. The 4V characteristic can support the whole process of the power system, which is illustrated in Figure 2.

1.2. Challenges. From the above-mentioned research status of Big Data technology and its application in many aspects of the power system, it is easily concluded that Big Data management and analytics are certain development trends of future smart grids. However, there are still challenges that exist in this research area, and strategies and technologies for unlocking the potential of Big Data are still at the early stage of development. First of all, most existing power system utilities are not prepared to handle the growing volume of data, both for data storage and data analytics. On the one hand, traditional machine learning or statistical computing methods are designed for single machines, and an efficient extension of these methods which can be utilized for parallel computing or for large-scale data is urgently needed. On the other hand, most of the analytic methods used in the power system are not suitable to handle Big Data; thus, the gap between Big Data analytics and power system applications still exist, and high-performance computing methods are required. Then, a big hurdle is the lack of an intelligent platform integrating advanced methods for Big Data processing, knowledge extraction and presentation, and support in decision-making. It is believed that the success combination of Big Data technologies and power system analysis will bring

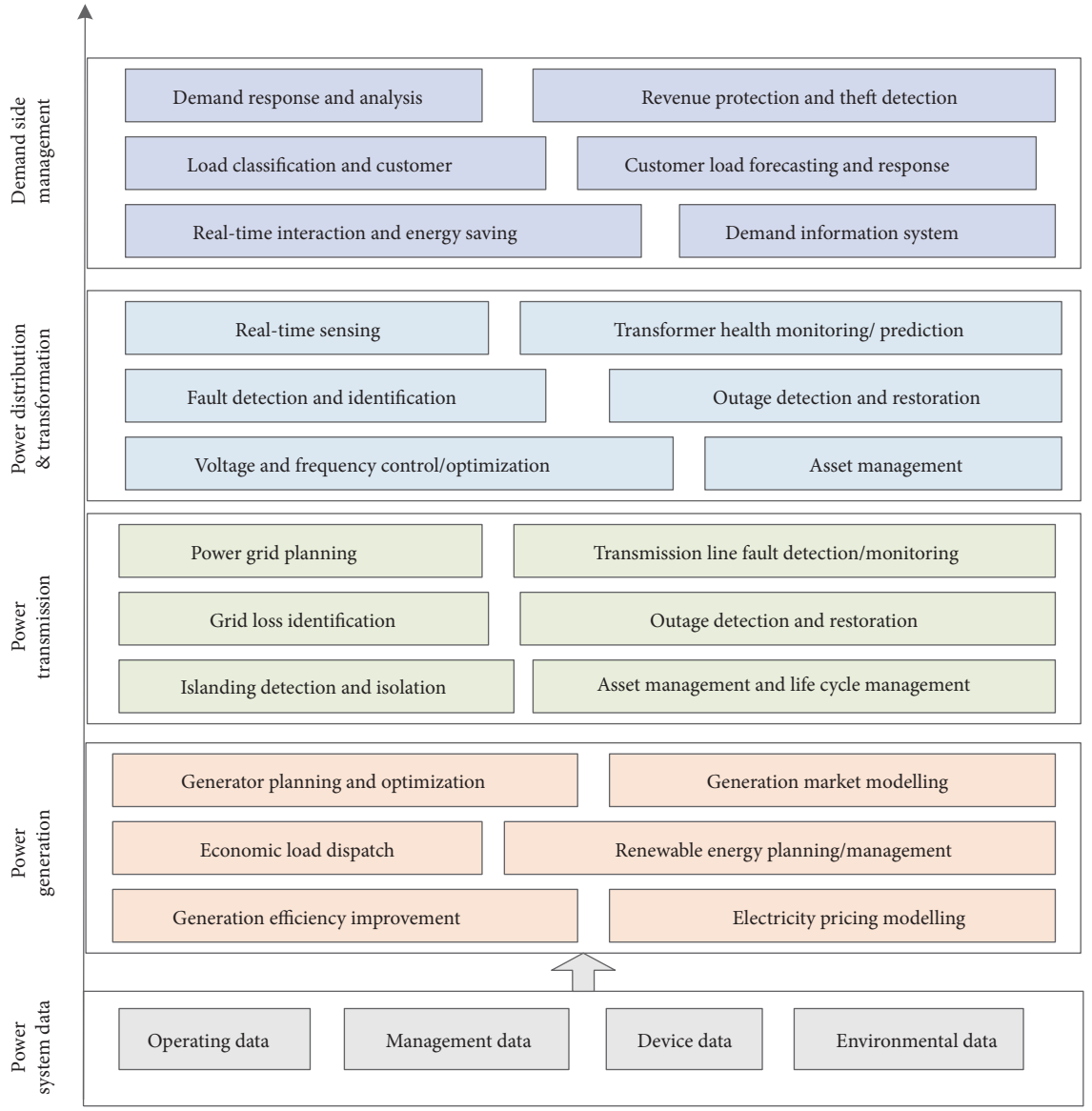


Figure 2: Sketch map of Big Data supporting whole process of the power system.

a number of benefits to the utility grid in the above-mentioned aspects. According to these challenges, this paper will present a novel Big Data platform for complex power system status monitoring and evaluation using machine learning algorithms.

2. Big Data Technologies for Complex Power System Monitoring

With the increasing varieties of data recording devices, much more unstructured power Big Data are being recorded continuously. Some particular data need to be collected or analyzed under different scales or projected to another dimension to describe the data. Therefore, some conflicts between data structure or semanteme need to be solved when projecting or transforming heterogeneous data into a unified form; the uncertainty and dynamics should also be taken into consideration for data fusion. Based on these concerns, the Big Data platform is designed to consist a generalized management model according to the complex logical relations between data objects, representing the data by normalization and extraction of the principal information. Challenges exist in how to design a flexible data management system architecture that accommodates multimode power data. This section introduces the state of the art of Big Data management technologies and data stream and value management.

2.1. State of the Art of Big Data Management Technology. In terms of distributed structure for Big Data management, the most popular designs are Hadoop [21] and Spark [22]. Hadoop was established in 2005, by Apache Software Foundation, with the key technologies of Map/Reduce [23], Google File System (GFS) [24] developed by Google Lab, and unrelational and high-volume data structure Bigtable [25], which have formed a novel computing distribution model. Base on the techniques above, Hadoop and open-

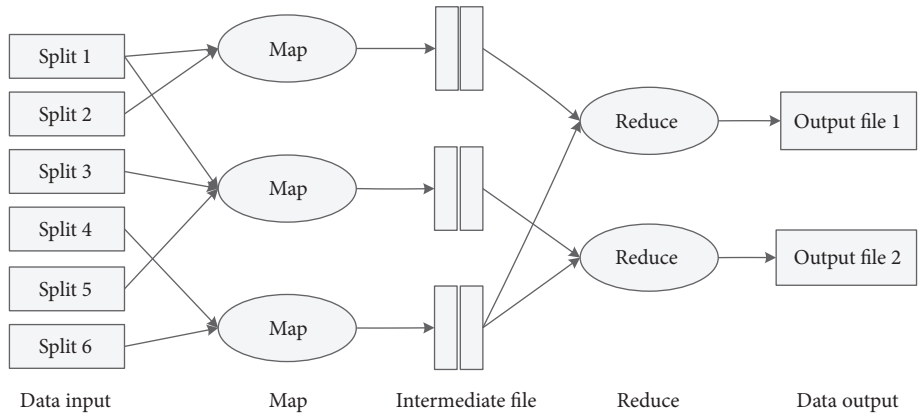


Figure 3: Hadoop MapReduce flowchart.

source projects like Hive and Pig have constituted the entire Hadoop ecosystem [26].

Hadoop, based on the distributed structure idea, enjoys many advantages such as high extensibility and high fault tolerance, and it is able to process heterogeneous massive data at high efficiency and low cost. In the Haddop ecosystem, files stored in HDFS (Hadoop Distributed File System) uses the subordinate structure, which are divided into several blocks; each of them has one or more duplicates distributed on different datanodes, thus the redundancy can prevent data from any loss caused by hardware failures. MapReduce is a programming model and an associated implementation for processing and generating large datasets. The computation can be specified by a map and a reduce function, and the underlying runtime system automatically parallelizes the computation across large-scale clusters of machines; the flowchart is given in Figure 3. With the high concurrent processing way, several computing processes are organized simultaneously, thus the data handling capacity can be increased from terabyte level to petabyte level.

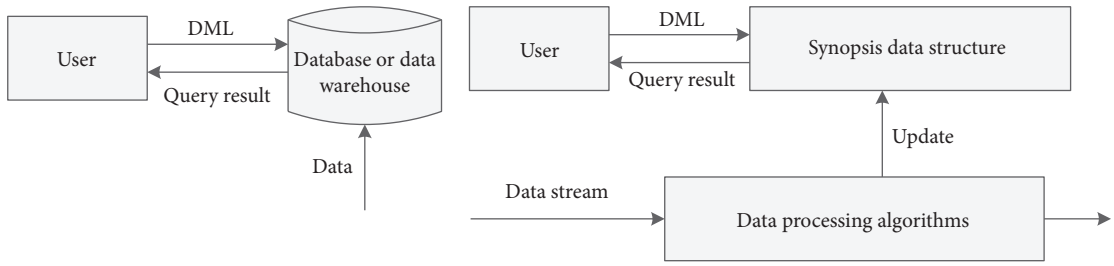
It can be seen that Hadoop technology is able to provide a reliable storage and processing approach; however, there are still limitations due to the Map and Reduce process. For a complex computation process, MapReduce needs a massive amount of Jobs to finish, and the relationships between these Jobs are managed by developers. Moreover, MapReduce is less supportive for interactive data and real-time data processing.

Similar to the computing frame of Hadoop MapReduce, another open-source tool Spark, developed by University of California Berkeley AMP lab, has the same advantages of MapReduce. Further, Spark can keep the intermediate results in RAM rather than write them in HDFS; thus, Spark can be better suitable for recursive algorithms such as data mining and machine learning applications. As a result, Spark is usually applied as a complement to Hadoop.

The key technology to Spark is the Resilient Distributed Dataset (RDD) [27], which is an abstraction to resolving the issue of slower MapReduce frameworks by sharing the data in memory rather than in disks, saving a large amount of I/O operations performed to query the data from disks. Therefore, RDD can greatly improve the recursive operation of machine learning algorithms and the interactive data mining methods.

Recently, a number of Big Data management systems have been developed to handle Big Data issues. For example, four representatives, MongoDB [28], Hive [29], AsterixDB [30], and a commercial parallel shared-nothing relational database system, have been evaluated in [31], with the purpose of studying and comparing Big Data systems using a self-developed microbenchmark and exploring the trade-offs between the performance of a system for different operations versus the richness of the set of features it provides. In terms of Big Data platform and tools that are suitable for power system and smart grid utilities, main contributions are made by leading IT companies like IBM [32], HP [33], and Oracle [34]. A number of IBM cases are done in order to improve the energy efficiency. For example, Vestas increases wind turbine energy production using a Big Data solution to more accurately predict weather patterns and pinpoint wind turbine placement [35]. CenterPoint Energy applies analytics to millions of streaming messages from intelligent grid devices enabling it to improve electric power reliability [36]. In the meantime, some newly established small technology companies, like C3 IoT [37], Opower [38] which has been acquired by Oracle in June 2016, Solargis [39], and AutoGrid [40], are doing Big Data analytics research and development according to the electricity market demand.

The large Internet companies in China, namely, Baidu [41], Aliyun [42], and Tencent [43], are all developing Big Data platform, tools, and applications according to their own business. For example, Baidu has been first in the world to open its Big Data engine to the public, which consists of key technologies of Big OpenCloud, Data Factory, Baidu Brain, and others. In this way, Baidu has won the prior opportunities to cooperate with the government, organizations, manufacturing companies, medical services, finance, retail, and education fields. Other companies like Inspur [44], Huawei [45], and Lenovo [46] also provide hardware from computer servers and storages to the Big Data analytic software, which have laid a good foundation for the development of the Big Data platform.



(a) Traditional data processing model (b) Data stream processing model

Figure 4: Comparison between the traditional data processing model and the data stream processing model.

2.2. Data Stream and Value Management. One of the most important ways to form Big Data is real-time data streaming, which is recorded continuously with time series. The data stream can be limitless, bringing a critical challenge for the data management system to store and process the streaming data. A definition was first proposed by Guha and Mcgregor in [47] that streaming data is considered to be an ordered sequence which can only be read one or a few times. Therefore, data stream management technology is the key issue to handling Big Data storage and processing.

Figure 4 shows the comparison between the traditional data processing model and the data stream processing model. For the traditional database, data storage is statical and not queried or updated often. Users send data manipulation language (DML) statements as queries, and the system will return the results after searching in the database. Therefore, there are inevitable I/O exchanges generated which will slow down the searching efficiency. For real-time processing of large amounts of streaming data, the traditional approach cannot meet the requirement. On the contrary, only synopsis data structure is stored instead of storing the entire dataset, and the data volume is much less and simpler to query compared to the traditional model.

The early research and design of the Big Data stream management system was only for single task applications. In order to handle streaming data with multiple tasks, the continuous query language was first proposed by Terry in Tapestry [48] in 1992, mainly used for filtering E-mails and the bulletin board system. Then it was followed by Mark Sullivan of Bell Labs in 1996, who designed a real-time monitoring tool named Tribeca [49] for the application of network surveillance. Tribeca was able to provided a limited number of continuous query languages and query operations. NiagaraCQ [50] was cooperatively developed by the Oregon Graduate Institute and University of Wisconsin, which support continuous query language and monitoring of durable and stable datasets in the entire wide-area network. In addition, Viglas and Naughton from the same project proposed a rate-based optimization on the issues of data streaming query speed [51]. In order to meet the requirements of data stream applications, a general data stream management is needed, and the official concept of a data stream management system was proposed in [52].

Nowadays, the most popular general data stream management system can be summarized as follows: Aurora [53], which was developed by the Massachusetts Institute of Technology, University of Brown, and Brandeis University, has a simple but special frame and can be used especially for data streaming surveillance based on a key technology of trigger networks. Aurora has a good balance on accuracy, response time, resource utilization, and practicability, but with a drawback of a simple query approach using the load shedding technique. TelegraphCQ [54], developed by the University of California Berkeley, is mainly used for sensor networks, which comprise a front end, a sharing storage, and a back end. The data stream in a constantly changing and unpredictable environment can be adaptively referred in any query. However, the approximate query mechanism will be neglected when the resource is insufficient. STREAM [55], developed by Stanford University, is the prototype system based on relational database. Under the circumstances of limited resources, STREAM can extend the searching language and execute the queries with high efficiency; thus, STREAM has a better performance on the continuous query. Other very famous data stream management systems are also released to cope with data stream challenges, such as Storm by Twitter [56], Data Freeway by Facebook [57], Samza by LinkedIn [58], TimeStream by Microsoft [59], and Gigascope by AT&T [60].

Data value in power systems can provide guidance towards data acquisition, data processing, and data application. Data valuation can be determined by several factors, including data correlation, data fidelity, and data freshness [61]. To be specific, data correlation can be considered from two aspects: one is how it is related with power dispatch, fault evaluation, and risk assessment; the other one is the correlation within the data itself, where the data value will be higher when the correlation is higher. Data fidelity refers to the conformance of the collected data to the real data situation. Defects of collected data always exist due to the sampling rate, noise, and data acquisition equipment from different devices across the entire grid; thus, the real data situation may not be revealed. At last, data freshness is also an important factor which determines the data value, especially in power systems where most data is streaming data, which is recorded without interrupt.

3. Analytical Tools and Methods for Power System Big Data

3.1. Big Data Analytical Open-Source Tools. Data analysis approaches such as machine learning play an important role

Table 1: Open-source/free software of Big Data machine learning method brief descriptions.

Name	Date	Developer	Brief descriptions
Octave	1993	James Rawlings, University of Wisconsin-Madison; John Ekerdt	A high-level language for numerical computations; suitable for solving linear and nonlinear problems; mostly compatible with Matlab, batch-oriented language [64].
Weka	1994	University of Waikato	Can be applied directly or called from a self-developed Java code and well-suited for developing new machine learning schemes [65].
R	1996	Ross Ihaka, Robert Gentleman	A language and environment for statistical computing and graphics; provides more than 70 packages of statistical learning algorithm; highly extensible [66].
Shogun	1999	Soeren Sonnenburg and Gunnar Raetsch	It provides a wide range of unified machine learning methods; easily combines multiple data representations, algorithm classes, and general purpose tools; rapid prototyping of data pipelines and extensibility of new algorithms [67].
http://AForge.net	2008	Andrew Kirillov, Fabio Caversan	It is an open-source C# framework in the fields of Computer Vision and Artificial Intelligence; image processing, neural networks, genetic algorithms, fuzzy logic, machine learning, robotics, etc. [68].
Mahout	2009	Grant Ingersoll, Apache Software Foundation	It is an environment for quickly creating scalable machine learning applications; a framework to build scalable algorithms; has mature Hadoop MapReduce algorithms; suitable for Scala + Apache Spark, H2O, and Apache Flink [69].
MLlib	2009	UC Berkeley AMPLab, The Apache Software Foundation.	It is the Spark implementation of machine learning algorithms; easy to write parallel programs; and has potential to build new algorithms [70].
scikit-learn	2010	David Cournapeau, Matthieu Brucher, etc.	It is built on NumPy, SciPy, and matplotlib in Python environment; accessible, reusable in various contexts, and with simple and efficient tools [71].
Orange	2010	Bioinformatics Lab, University of Ljubljana, Slovenia	It is a data visualization and data analysis software; has interactive workflows with a large toolbox and a visualized process design based on Qt graphical interface [72].
CUDA-Convnet	2012	Alex Krizhevsky	It is a machine learning library with a built-in GPU acceleration; has been written by C++; with the CUDA GPU processing technology by NVidia [73].
ConvNetJS	2012	Andrej Karpathy, Stanford University	It is a JavaScript library for training deep learning models in the browser; is able to specify and train convolutional networks; comprises an experimental reinforcement learning module [74].
Cloudera Oryx	2013	Sean Owen, Cloudera Hadoop Distribution	It provides simple real-time large-scale machine learning and predictive analytics infrastructure; is able to continuously build/update models from large-scale data streams and query models in real time [75].

in power systems as algorithms can be trained using historical data collected over time, providing useful information for system operators. As historical data is collecting at an increasing speed with large volume, effective machine learning approaches are urgently needed in discovering valuable information and providing to power system operators. Big Data is stored in a distributed way on multiple computers; thus, it is not appropriate for all machine learning methods to process. Moreover, if data analytics needs to be finished on a single computer, it may be too large to fit into the main memory. Most traditional libraries/tools, such as R [62], Weka [63], and Octave [64], implemented machine learning algorithms in a single-threaded fashion by design and are not able to analyze large volumes of distributed data. More recently, advanced modern Big Data processing platforms are designed and implemented with parallel machine learning algorithms in order to achieve high efficiency. First of all, this section gives a comprehensive literature survey of state-of-the-art machine learning libraries and tools for Big Data analytics in Table 1.

From Table 1, it can be seen that along with the rapid development of the computer technology, a hot favorite of

developing machine learning library/tools started in the early 1990s. In almost a decade, the research trend moved forward to distributed and large volumes of data from the traditional single machine algorithm design. Octave is the earliest developed machine learning package, performing numerical experiments using a language that is mostly compatible with Matlab. Similarly, Weka was also developed by universities, which makes this free software suitable for academic use by integrating general purpose machine learning packages. In particular, R has been widely used in both academia and industry due to the comprehensive statistical computing and graphics software environment. As mentioned above, Octave, Weka, and R are designed for single-threaded computing and thus are not able to handle large volumes of power system data.

In recent years, the R community has developed many packages for Big Data processing. For example, the *biglm* package [76] is able to perform linear regression for large data, and the *bigrf* [77] package provides a Random Forest algorithm in which trees can be grown concurrently on a single machine, and multiple forests can be built in parallel on multiple machines then merged into one. Another group of R packages, such as *hive* [78], focus on providing interfaces between R and Hadoop, so that developers can access HDFS and run R scripts in the MapReduce paradigm.

Among the oldest, most venerable of machine learning libraries, Shogun was created in 1999 and written in C++, but is not limited to working in C++. In terms of supported language, Shogun can be used transparently in such languages and environments: as Java, Python, C#, Ruby, R, Lua, Octave, and Matlab, thanks to the SWIG library [79]. Another machine learning project designed for Hadoop, Oryx comes courtesy of the creators of the Cloudera Hadoop distribution. Oryx is designed to allow machine learning models to be deployed on real-time streamed data, enabling projects like real-time spam filters or recommendation engines.

3.2. Machine Learning and Statistical Processing Methods

3.2.1. Machine Learning Algorithms. Besides the powerful open-source algorithms or tools mentioned above, machine learning and statistical processing methods are also applied to support handling various issues of the power data. Basic machine learning algorithms are embedded in different open-source libraries/tools. Table 2 gives a comprehensive study and comparison.

There are many benefits for the modern power system since machine learning algorithms have been applied in many aspects of power systems successfully. Firstly, system stability and reliability have been remarkably increased. Many literatures have reported impressive experimental results of various machine learning algorithms with applications in oscillation detection, voltage stability, fault or transient detection and restoration, islanding detection and restoration, postevent analysis, etc. [80–83]. With the emergence of the Big Data analytics and smart grid technology, the above-mentioned monitoring and detection methods have been greatly improved, and an increasing number of novel approaches are being studied. For instance, real-time identification of dynamic events using PMUs is proposed in [84]; based on data-driven and physics models, security of power system protection and anomaly detection are greatly improved, thanks to the rich synchrophasor data.

Secondly, power equipment utilization and efficiency are greatly increased. In the power industry, the issues of waste of equipment resources are difficult to handle, and data resource is independent, thus it is impossible to evaluate the exact status of each asset. Big Data analytics can provide better validation and calibration of the models, eliminate the independence of data resources, and help operators understand the operating characteristics and life cycles of the equipments. For example, a data-driven approach for determining the maintenance priority of circuit breakers is introduced in [85]; the proposed method can consider both equipment-level condition monitoring parameters and system-level reliability impacting indices; thus, the maintenance priority list can be generated.

Thirdly, Big Data visualization can help operators improve situation awareness and assist decision-making. Machine learning and data analytics only produce numerical results or two-dimensional charts and diagrams, which need operators with professional skills or experience to give accurate and timely decision. A Big Data platform with 3D visualization in [86] manages massive power Big Data with multimode heterogeneous characters, showing the tripping lines and affected areas based on a 3D environment. Thus, the operators can make quicker and more reliable decisions and take possible preventive actions under the circumstance of thunder and lightning weather.

3.2.2. Statistical Processing Control Methods. Statistical processing control methods originally are applied in industrial quality control, employing statistical methods to monitor and control a process based on historical and online data. In our early work, some data-driven methods based on linear principal component analysis (PCA) [87] were applied in power system data analysis [88], setting up a distributed adaptive learning framework for wide-area monitoring, capable of integrating machine learning and intelligent algorithms in [89]. In order to handle power system dynamic data and nonlinear variables, dynamic PCA [90] and recursive PCA [91] were also developed to improve the model accuracy. It is worth mentioning that linear PCA is unable to handle all process variables due to the normal Gaussian distribution assumption imposed on them, and many extensions using neural networks have been developed [92, 93]. To address the challenges of handling the redundant input variables, obtaining higher model accuracy, and utilizing non-Gaussian distributed variables, an improved radial basis function neural network model-based intelligent method is also proposed in the early work [94]. The neural input selection is based on a fast recursive algorithm (FRA) [95, 96], which was proposed for the identification of nonlinear dynamic systems using linear-in-the-parameter models. It is possible to utilize optimization methods in order to get more accurate models by tuning algorithm-specific parameters, such as particle swarm pptimization (PSO), genetic algorithm (GA), differential evolution (DE), artificial bee colony (ABC), and ant

Table 2: Comparisons of open-source machine learning tools/algorithms for Big Data.

Category	Algorithm	Open source/free software						
		Weka	R	Shogun	Mahout	MLib	Orange	Oryx
Classification	Logistic regression	√		√	√	√	√	
	(Complementary) naive Bayes	√		√	√	√	√	
	Decision tree	√				√	√	
	Neural networks	√		√				
	SVM	√		√		√	√	
	Random forest	√	√				√	√
	Hidden Markov models			√	√			
Regression	Linear regression	√	√		√	√	√	
	Generalized linear models		√			√		
	Lasso/ridge regression		√		√		√	
	Decision tree regression	√				√		
Clustering	k-means	√		√	√	√	√	√
	Fuzzy k-means				√	√		
	Gaussian mixture model (GMM)					√		
	Streaming k-means					√		
Collaborative filtering	Alternating least squares (ALS)		√		√	√		√
	Matrix factorization-based				√			
Dimensionality Reduction	Singular value decomposition (SVD)			√	√	√		
	Principal component analysis	√	√	√		√		
Optimization primitive	Stochastic gradient descent)					√	√	
	Limited-memory BFGS (L-BFGS)			√		√		
Feature extraction	TF-IDF					√		
	Word2Vec					√		
Frequent pattern mining	FP growth	√				√		
	Association rules	√				√	√	

colony optimization (ACO), among other heuristic methods. The proper tuning of the algorithm-specific parameters is a very crucial factor which affects the performance of the above-mentioned algorithms. The improper tuning of algorithm-specific parameters either increases the computational effort or yields the local optimal solution. In our early work [97–99], teaching-learning-based optimization (TLBO) has been utilized for training an RBF neural network battery model. The TLBO method does not have any algorithm-specific parameters and significantly reduces the load of tuning work.

These methods mentioned above can be programmed and integrated as part of the analysing engine to support the processing of the power Big Data. Therefore, the data processing engine can support overall system operation and control by building a dynamic, global, and abstract power data model, based on which consequences are inferred and decisions are made. A detailed method comparison can be found in Table 3.

The fundamental assumption for many standard data-driven methods such as PCA, PLS, and LDA is that the measurement signals follow multivariate Gaussian distributions. As introduced in Table 3, PCA and PLS have similar principals to extract latent variables, but they perform in different ways. PCA tries to extract the biggest variance from the covariance matrix of the process variables, while PLS attempts to find factors or latent variables (LVs) to describe the relationship of output and input variables. PCA and LDA are also closely related in finding linear combinations of variables to explain data. However, LDA deals with the discrimination between classes, while PCA deals with the entire data samples without considering the class structure of the data. Similar to PLS, SIMs require both the input process data and the output data to form input-output relations. A brief comparison among the above-discussed basic data-driven methods is given in Table 4.

The issues of Gaussian distribution assumption on data, requirement of input-output relationships, the number of principal components or latent variables, and the computational complexity for these methods are compared in this table. In addition, LDA is comparable with PCA and the datasets should be well documented in order to

Table 3: An overview of state-of-the-art intelligent processing methods.

Category	Method	Descriptions	Applications
Standard	Principal component analysis (PCA)	PCA summarizes the variation in a correlated multiattribute data to a set of uncorrelated components, a linear combination of the original variables.	Pattern recognition [100], dimension reduction [101], feature extraction [102], process monitoring [103].
	Partial least squares (PLS)	PLS can find the fundamental relations between two data matrices, and latent variables are needed to model the covariance structure in these spaces.	Power load forecasting [104], performance evaluation of power companies [105], etc.
	Linear discriminant analysis (LDA)	LDA finds a linear combination of features that characterizes or separates two or more classes of objects or events.	Face recognition [106], feature selection for power system security assessment [107].
	Subspace identification methods (SIM)	SIMs are powerful tools for identifying the state space process model directly from data.	Power oscillatory state space model [108], power system stability analysis [109], etc.
Time-varying	Recursive PCA	RPCA is a generalization of PCA to time series; the eigenvector and eigenvalue matrices are updated with every new data sample.	Voltage stability moitoring [110], power system fault location detections [111].
	Dynamic PCA	DPCA includes dynamic behavior in the PCA model by applying a time lag shift method while retaining the simplicity of model construction.	Industrial monitoring [112, 113], dynamic economic evaluation of electrical vehicles [114].
Nonlinear	Kernel PCA/PLS	KPCA is first to map the input space into a feature space via nonlinear mapping and then to compute the PCs in that feature space.	Power equipment assessment [115], real-time fault diagnosis [116], power system monitoring [117], etc.
	Neural networks	Neural networks are computational models that can be used to estimate or approximate unknown nonlinear functions.	Dimension reduction [118, 119], voltage stability assessment [120], fault location detection [121], etc.
Non-Gaussian	Independent component analysis (ICA)	ICA decomposes multivariate signals into additive subcomponents which are independent non-Gaussian signals.	Fault detection [122], power quality monitoring [123], and estimation [124].
	Gaussian mixture models (GMM)	GMM describe an industrial process by local linear models using finite GMM and Bayesian inference strategy.	Power flow modeling [125], power load modeling [126].
	Support vector data description (SVDD)	SVDD defines a boundary around normal samples with a small number of support vectors.	Classification, process monitoring [127], oscillation modes detection [128], etc.

Table 4: A brief comparison among basic data-driven methods.

	PCA	PLS	LDA	SIM
Gaussian distribution	√	√	√	
Input-output relationship		√		√
Number of principal components	√			
Number of latent variables		√		
Computational complexity	Low	Medium	Medium	Medium

offer detailed information about the normal operating condition and faulty cases. SIM does not impose any special assumption on the process data since it only investigates the input-output relationship, and different threshold computation methods are available for Gaussian and non-Gaussian distributed data. The number of PCs and LVs are important design parameters in PCA and PLS methods, which can affect modeling performance. The main computation burden comes from performing SVD on the covariance matrix of different dimensions; thus, the standard PCA has lower computational cost over other basic methods.

Table 5: Comparisons of the non-Gaussian data methods.

Method	Data assumption	Parameters	Disadvantages
ICA	Can be described as a linear combination of non-Gaussian variables	Number of ICs	(1) High computational cost (2) Hard to determine the control limit
GMM	Can be described by local linear models	Multiple parameters in the model	(1) Complicated to train the models (2) Hard to determine the number of local models
SVDD	No strict assumption of data distribution	Kernel parameters in the model	(1) Hard to tune the kernel parameters (2) Trade-off between accurate boundary and low false alarm control limit

For time-varying process methods, recursive and adaptive methods are able to track slow-varying processes with a stable model structure. However, the model updating may be carried out randomly if no appropriate updating scheme is available. Meanwhile, dynamic process monitoring methods are easy to implement in practice, but the number of dynamic steps significantly affects the monitoring results and the window size is difficult to be determined.

Compared to linear monitoring methods, nonlinear approaches can be used in much wider applications due to the flexibility of nonlinear functions, which can model nonlinear relationships between variables. Especially for the kernel methods, various nonlinearities can be modelled by introducing different kernel functions. Similarly, neural networks are also capable of modeling any kind of nonlinearity theoretically. However, there are still some drawbacks; for example, the structure of the neural networks is difficult to determine and the training of the network parameters is also computationally demanding. A similar issue exists to kernel-based methods and an appropriate kernel parameter tuning method is needed, and the selection of a kernel function is not a trivial issue. A new approach to tackle the issues of representing nonlinear behavior as well as the non-Gaussian distributed variables is urgently needed.

For non-Gaussian distributed data, the basic methods cannot perform well due to the Gaussian distribution assumption. Alternatively, ICA, GMM, and SVDD are three most widely used and promising methods for non-Gaussian process monitoring. Although these methods were developed separately, they are actually highly related to each other. Sometimes, these methods can even be combined, and they are also capable of handling more than only one data characteristic. For example, ICA is used to describe the measurement signals as a linear combination of non-Gaussian variables, while GMM has a similar assumption that the process dataset can be described by several local linear models. Moreover, the calculation of control limits for ICA-based non-Gaussian process monitoring involves kernel density estimation, which is commonly used for SVDD. Detailed comparative advantages and disadvantages of these methods are listed in Table 5.

4. A Real-System Case

In this paper, a Big Data platform integrated with data management and analytical engine is proposed as a real-system case study. This platform was designed to meet the special condition of power grid in South China, such as large-scale, complex geographical and weather conditions and AC/DC mixed operation over long distances. Big Data technologies are applied to this power network to assist with condition monitoring and state estimation of the transmission and distribution systems, collecting multiplatform power data and realizing high-efficiency processes and analysis of data from the power grid at different levels.

4.1. The Framework of Electric Power Big Data Platform. The framework of the electric power Big Data platform consists of database, data interface, Big Data Management system, analytic engine with various machine learning tools and algorithms, and application and 3D visualization modules; a detailed structure is given in Figure 5. The first challenge is to set up an efficient database for the large volume of multisource heterogeneous power data which are collected through different sources. A traditional power system database is designed to store structured data files using tables; thus, the size of storage is limited and the data operation efficiency is low. For Big Data platforms, various data are collected, for example, operational data collected from the production management system and energy management system, real-time data recorded from an online monitoring system and equipment monitoring system, and other forms of heterogeneous data of weather files, geography information, images, and video data. In terms of data status, historical data, real-time data, and data streaming are all needed for Big Data processing and analysis. This platform integrates several data storages according to each data structure, so that the platform can provide useful and timely information to assist decision-making by processing large amounts and different data structures with high efficiency. All the information and knowledge can be integrated to provide strategies for system operation and evaluation, system inspection, and status estimation for power equipments and the entire power grid.

In order to efficiently manage and store the multisource Big Data, this paper proposes a special data acquisition structure. For various databases, SQOOP is a tool designed for efficiently transferring bulk data between HDFS and structured datastores such as relational databases (MySQL, Oracle). For messages between databases and the platform, MQTT (message queuing telemetry transport) is chosen as part of the data interface. MQTT is well known as an "Internet of Things" connectivity protocol, and it was designed as an extremely lightweight published/subscribed

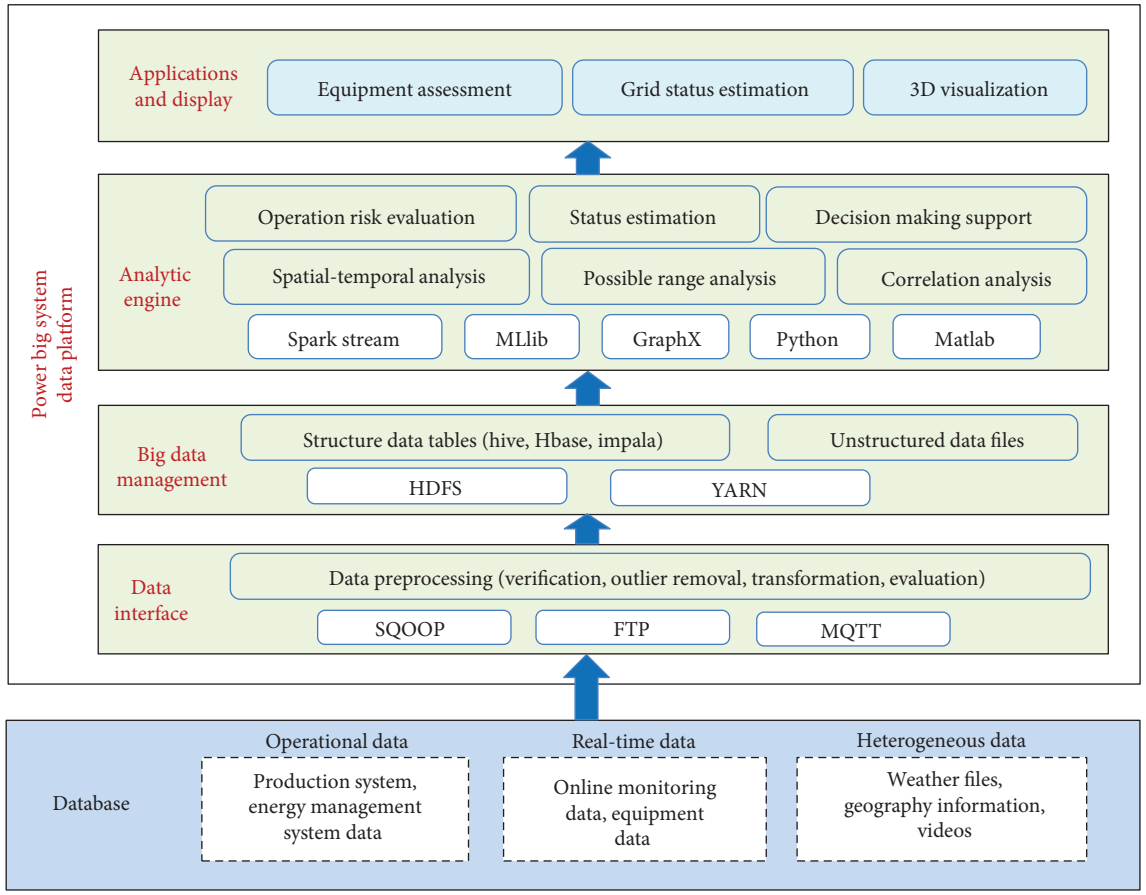


Figure 5: Big Data processing and analysing platform for electric power system condition monitoring.

messaging transport. For files such as documents and working logs of each equipment, transmission lines, and substations, FTP (file transfer protocol) is the common tool to transfer through the Internet to the platform.

Based on this data interface, power system data collected by smart devices can be managed in real time. Data preprocessing, including data verification, outlier removal, transformation, and evaluation process, can be realized to provide a solid and practical database for the analysis procedure. Moreover, other relative unstructured data such as weather condition, lighting and storms, geography information, and human activities (local population, age distribution, professionals, behavior and active pattern, internet sentiment, and so on) can be connected to a certain extent with the power load, power generation, consumptions, electricity market, and so on. These data sources mentioned above are impossible to be processed and analyzed simultaneously through the traditional way; only this novel approach using Big Data to deal with the challenges can establish a more comprehensive knowledge model of the city power grid.

4.2. High-Performance Analytical Engine. To effectively manage the Big Data is only the first step; the key issue is to set up an analytic engine with high efficiency. Based on the functional modules and the need for power system applications, this particular analytic engine can provide with several practical functions, such as operation risk evaluation, status estimation, and decision-making support. The detailed structure of the Big Data computational engine is given in Figure 6.

This analytical engine integrates a number of open-source basic algorithm packs and self-developed algorithms. The open-source algorithm packs mentioned in Section 3 have been developed and tested by researchers and companies for many years. For example, Apache Spark, a fast and general engine for large-scale data processing, can be used interactively with Scala, Python, and R shells. Many powerful computing libraries are integrated in Apache Spark, such as numerical computing tool NumPy, science computing tool SciPy, data analysing library Pandas, scalable machine learning library MLlib [70], API for graphs and graph-parallel computation GraphX [129], and so on. In addition, this platform has combined an interactive developing and operating environment IPython and Jupyter [130]. Effective power grid decision-making depends critically on anlytic methods in the platform. Therefore, effective methods for the real-time exploitation of large volumes of power data are needed urgently. Robust data analytics, high-performance computation, efficient data network management, and cloud computing techniques are critical towards the optimized operation of power systems.

For self-developed algorithms, spatial-temporal correlation analysis is able to mine both the strong and weak connections among the numerous variables in a power grid, by

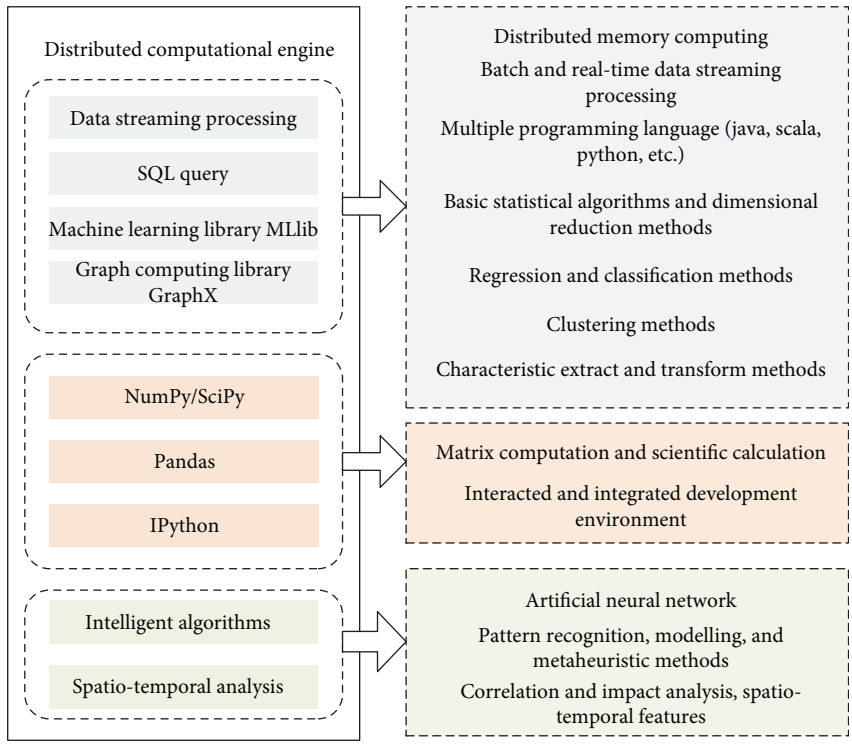


Figure 6: Structure of Big Data platform computational engine.

setting up a power system spatial-temporal model and a data-driven model based on the process history database. Modelling methods are provided, including artificial neural networks, linear and nonlinear analysis methods, Gaussian-based kernel methods, regression and classification methods, and clustering methods. Pattern recognition methods for spatial-temporal correlations are provided, and the spatial proximity weights, time delay, and correlation effect are calculated and quantized [131]. This idea is suitable for analysing the consumption behaviors of citizens in different locations and time, as well as the effect on power transmission lines by the power grid surroundings including geographic information, weather variations, human activities, and road vehicles and traffic situations [132]. A knowledge base of interconnected factors within the entire city grid can be set up for analysis and predictions.

This proposed distributed computational engine is the key element of the entire Big Data platform; many functional modules can be developed based on these open-source tools. It is believed that this novel approach will gradually change the traditional way of power system analysis and operation, which is also the only efficient way to realize future smart grids with high level of automation and intelligence.

4.3. 3D Visualization. The geographic information system (GIS) has been widely used in electric power systems [133, 134], which is vital for improving the operation efficiency of the electric power system. It can maintain, manage, and analyze power data and integrate power network models, maps, and related data in a solution for desktops, webs, and mobile devices. Most power GIS systems mainly adopt a two-dimensional map as the visualization model. However, 2D GIS has significant limitations in terms of presentation and analysis of geospatial and power data, and it is difficult to display panorama information of power running status. The proposed Big Data platform adopts a web-based visualization method based on Cesium and 3D City Database (3DCityDB) [135] to construct a three-dimensional panorama electric power visualization system, which is given as in Figure 7.

The 3D models of electric tower, line, equipment, and geographical entity (buildings, roads, etc.) will be visualized in Cesium scene and managed by a Cesium manager. In the server side, Java Servlet and JavaServer Pages for power-related data processing functions reside in Tomcat which directly communicate with web client and process client requests. The two-dimensional map requests will be submitted to the Geoserver, while three-dimensional map requests will be processed by a 3DCityDB web feature service. 3DCityDB is a free open-source package consisting of a database scheme and a set of software tools to import, manage, analyze, visualize, and export virtual 3D city models according to the CityGML standard. In this architecture, 3DCityDB has two important tasks: one is to convert a two-dimensional electric map model to a three-dimensional model and save into the PostgreSQL database, and the second is used to provide a three-dimensional web feature service for a power system client based on Cesium.

Based on the model calculation and Big Data analytical engine, the visualization of spatial information and power system applications can be realized in the way of providing services. Thus, the power system equipments and power grid can be merged together with GIS and revealed on the map, as well as the environmental factors. Therefore, many demands

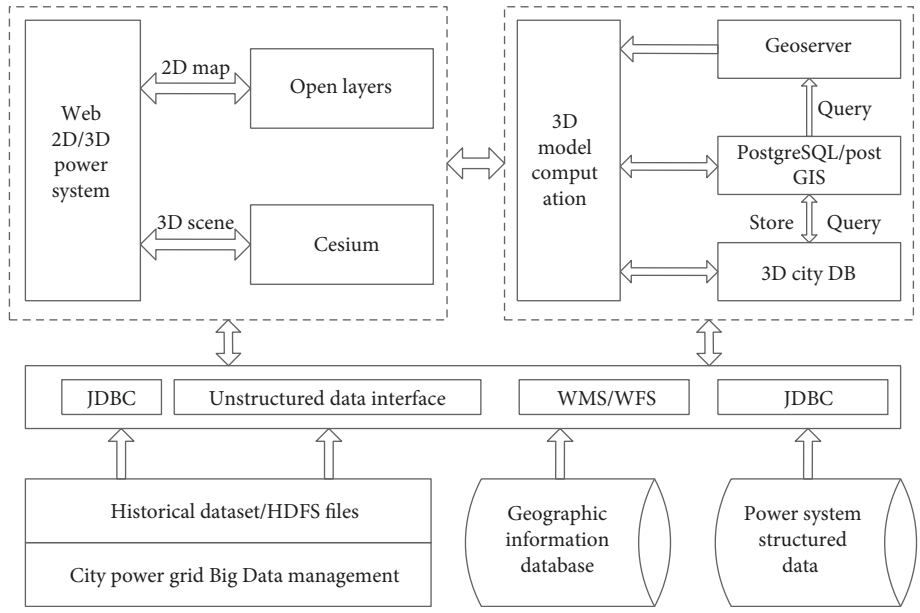


Figure 7: Framework of the 3D display system.

of power grid visualization can be reached, including real-time monitoring, analysis, and decision-making, among others. The development of the 3D visualization system can provide an optimal way of presenting the huge amount of information and improve the situation awareness of system operators as well as the novel explanation of newly appeared information; thus, the accuracy of decision-making for the entire power system can be greatly increased.

5. Application Study

5.1. Development of Power Grid Topology and Parallel Computing Using CIM Files. Power element data, connections, and their status are stores as common information model (CIM) files in the power system, which are significant for power system analytics. The first step is to extract the connectivity between each electric point as data to be stored in the relational database. For most of the analytic methods, the above-mentioned CIM file extraction is applied to fit in the relational database. However, a topology analysis needs plenty of correlation analyses between multiple and complex tables; it is hard to meet the demand of real-time and fast-speed processing requirement. The proposed platform in this paper develops a fast-processing scheme for the power grid topology setup; thus, the analysis can be realized with high efficiency. The diagram is given in Figure 8.

The proposed platform detects any update of CIM files which were transmitted into the FTP end, load new data into memory, and correlate with other structured data using Spark SQL, generating a preprocessed data table. After that, a fast search according to "physical-electrical-physical" rules in the power grid is applied to set up a topology of the grid. The whole process is realized based on the Spark SQL database and parallel computation; thus, the analysis efficiency is greatly improved, thanks to the fast and parallel correlation analysis. Under this framework, many tasks can be done easily including analysis result extraction, power grid topology setup, power system branch model calculation, and "bus-branch" model analysis and other functions. Therefore, this platform is able to provide a database and analytical engine for power grid large-scale parallel computation, real-time status analysis for smart grids, and other useful applications.

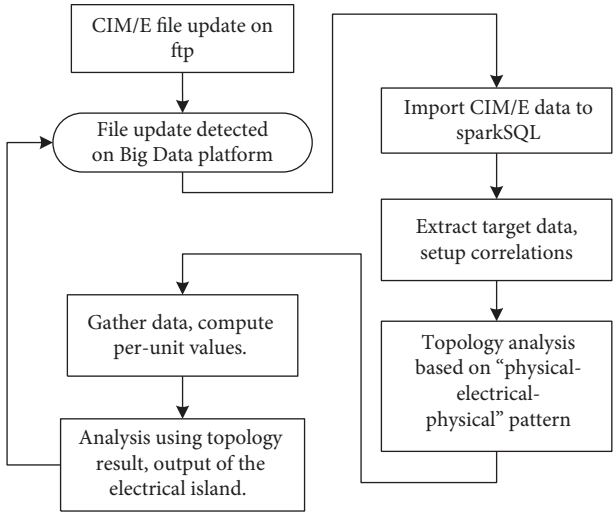


Figure 8: Fast-speed analysis flowchart for CIM files.

5.2. High-Efficiency Load-Shedding Calculation. The calculation of load-shedding in the power system can quantify how much loss the real system is undergoing after equipment failure in an objective way; thus, it can measure the operation risks and provide significant information for decision-making of equipment reconditioning or replacement. The actual reduction of load-shedding for different types on each electrical point is needed for the calculation; thus, it is very

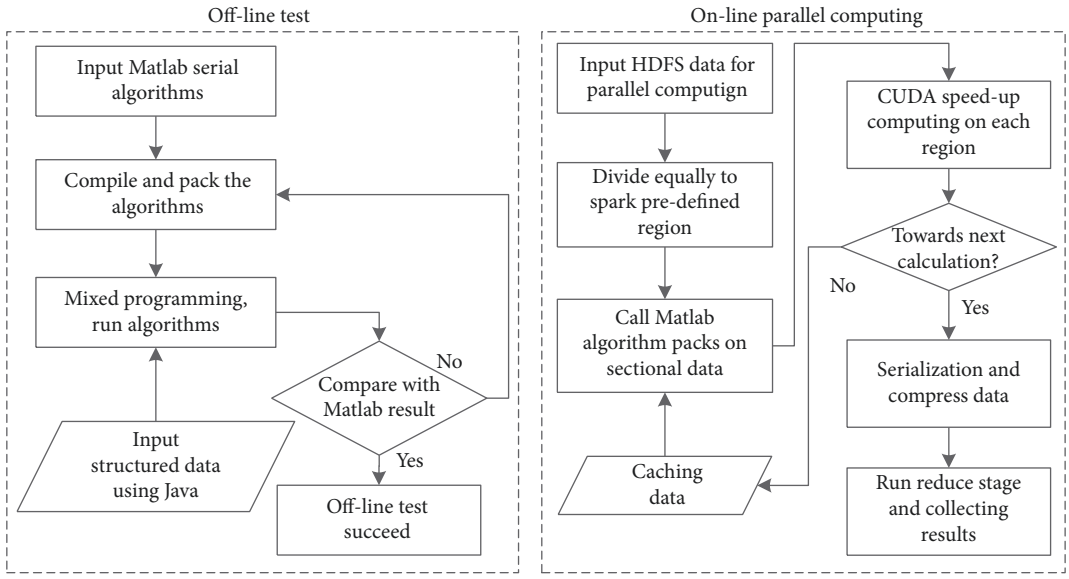


Figure 9: Fast-speed analysis flowchart for CIM files.

Table 6: The comparisons of parallel computing with single machine results.

Machines	Model	Memory	Cores	Executor	Time
Machines	YARN	60G	30	30	11 min
Single	Local	200G	20	20	2.5 hours

time-consuming to calculate power grid risk evaluation with plenty of predefined fault scenarios. In the proposed platform, a calculation scheme based on Spark and Compute Unified Device Architecture (CUDA) is applied, as shown in Figure 9.

The complete load-shedding scheme contains two stages: offline test stage and online parallel computing stage. The computation tasks are firstly divided into different working regions on Spark, then Matlab algorithms are packed and called, and further processing of each computation task is transmitted to working threads on every division, where parallel computing is realized. After that, results at each step are collected progressively; thus, the risk evaluation tasks for multiple scenarios can be finished. For real-system cases, a total number of 6000 scenario files with 1.2 GB size are calculated according to the flowchart given in Figure 9, and the comparison with calculation time on a single machine is given in Table 6.

It can be easily seen that parallel computing is able to solve the problem of low efficiency when risk evaluation in multiscenarios is taken in the power system. The load level of each electrical point can be monitored dynamically, and the topology change of power grid due to any system maintaining or drop out of multiple power system units can also be calculated with high efficiency, therefore, the computation time is greatly shortened.

5.3. Power System Transmission Line Tripping Analysis Using 3D Visualization. With the support of the Big Data platform, transmission line trip records, power quality data, weather data, and other related data can be collected, in order to monitor and analyze the transients. In addition, a three-dimensional visualization system is developed to merge together all the analysis results with geographic, landforms, and even weather conditions, then display in a very intuitive way. Therefore, situation awareness of system operators is greatly enhanced. Two main tasks are introduced in this section: firstly, the correlation between line trips and power transients is analyzed by employing statistical methods, especially the distribution patterns of line tripping and power quality voltage dips against the lasting time. Secondly, the interconnection rules among line tripping, weather condition, voltage dips, and voltage swells and other disturbances are exploited.

In order to analyze the correlation between transmission line trips and voltage dips, multisource data is needed, consisting of (1) transmission line tripping data, recording tripping time, fault description, fault type, and so on, and (2) voltage disturbance data, including monitoring location, disturbance type, happening time, lasting time, and magnitude. The first step to analyze the transients is data fusion, combining two sets of data according to the unified time tags, and the preprocess diagram is given in Figure 10.

For analysis of voltage dips at different voltage levels of 110 kV and 10 kV, the voltage dip recordings are divided into four kinds, including voltage dips caused by line trips at 110 kV and 10 kV, not by line trips at 110 kV and 10 kV. Taking 10 kV voltage level for example, the scatter plot is generated and shown in Figure 11.

In this figure, each symbol represents a transient event, with duration as the x-axis and magnitude as the y-axis. In order to separate the transient events by their causing reasons, the blue dot represents the voltage dip caused by line trips while the red x shows that the occurring voltage dip was not due to line trips, both at the 10 kV voltage level. The x-axis has taken the logarithm for the purpose of

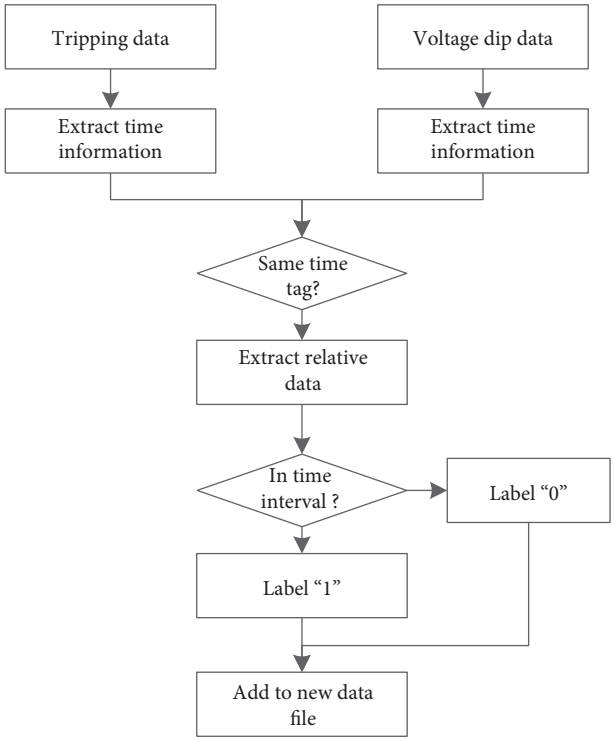


Figure 10: Diagram of data files fusion preprocess.

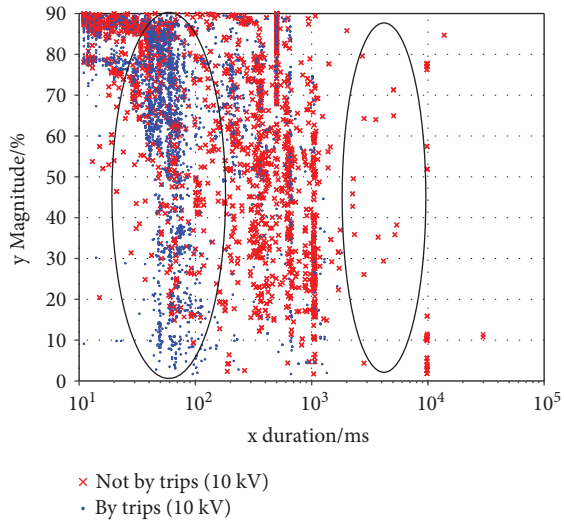


Figure 11: Scatter plot of voltage dips and breakdowns against duration under 10 kV voltage level (half logarithmic axis).

showing the distribution more clearly. Generally speaking, it can be seen from Figure 11 that at the 10 kV level, the lasting time of voltage dips caused by line trips is less than that caused by other reasons, as shown in the left ellipse, with duration around 100 ms. And the voltage dips caused by other reasons last for a longer time, as enclosed in the right ellipse.

The scattered points only show the distribution of durations against magnitudes of the voltage dips. It is necessary to combine substation coordinates, maps, and other geographic information with these transients; thus, the transmission line status and the affected substation can be shown in terms of voltage dip magnitudes and durations. Therefore, the possible influence of transmission line trippings to the substations can be visualized to system operators. The Big Data platform employs a 3D simulation display system, using data from the management layer as well as the model output directly from the computing engine, including 3D models of power line and electric equipment, 3D building models, geospatial data, and power attribute data. Geospatial data as a 3D virtual environment can show geographic objects (e.g., roads, bridges, and rivers) around the electricity network. The generated 3D virtual environment with power transmission line situation is given in Figure 12.

In Figure 12, the green line represents the normal operational transmission line, while the red lines are with the appearance of the line trips. In order to show the voltage transient status, a cylinder with blue color shows the voltage dip magnitude, and the pink cylinder is the duration, and the name of the affected transmission lines is shown in the floating red tags above the cylinders. Therefore, the affected area can be directly visualized through the 3D virtual environment, and the dynamic change of the power grid operational status is easier to control for the system operators. If any transient happened, actions can be taken in time to prevent any enlargement of the accident.

6. Discussion and Conclusion

This paper reviewed both the issues of Big Data technologies for power systems and employed a Big Data platform for power system monitoring and evaluation analysis. Based on the review of Big Data management technology and analytical tools and machine learning methods, a case study of the proposed novel Big Data platform for a power system is given with three application cases introduced. The framework of the power system Big Data platform consists of database collecting power data from all different parts across the grid, data interface, Big Data management system integrating different management technologies, analytic engine with various machine learning tools and algorithms, applications, and 3D visualization modules for further optimizing the strategy and decision-making assistance.

Based on the various power data sources, the proposed platform has integrated different data interfaces and distributed data storage according to the data structure; thus, the platform is able to handle traditional structured data, semi-structured data, and unstructured data simultaneously. For the analytical engine, both open-source tools and self-developed models are integrated as modules. In our early work, intelligent processing methods have been proved to be able to handle linear, dynamic, nonlinear, and non-Gaussian distributed variables by setting up accurate and efficient models. This has enabled the decision-making subsystem to focus on generating an optimized equipment maintenance strategy and providing a global view for situation awareness and information integration.

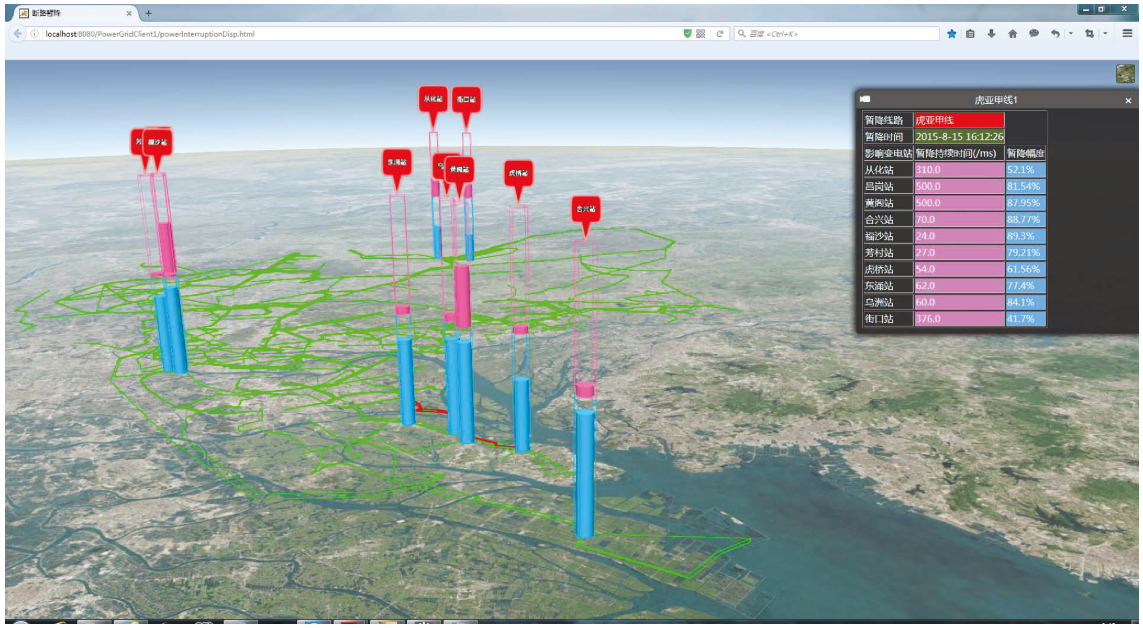


Figure 12: 3D display of voltage dips and breakdown transmission lines with geographic information.

In order to demonstrate the effectiveness of the proposed platform, three real-system cases are introduced including development of power grid topology and parallel computing using CIM files, high-efficiency load-shedding calculation, and power system transmission line tripping analysis using 3D visualization. These cases are all realized based on the proposed Big Data platform; the key issue in case one is to extract the connectivity between each electric point from different databases. It is suitable to process high-volume and multimode heterogeneous data using multiple data storage methods in the proposed Big Data platform with very high efficiency. In case two, with a proposed parallel computing scheme based on Spark and CUDA, load-shedding calculation in power systems under different scenarios can be realized in a very fast-speed way, and a comparison between single-machine and multiple-machine parallel computing is given, which demonstrated the high efficiency of the scheme. A highlight in the third case is the utilization of a Big Data platform with the 3D visualization system. With the help of the Big Data virtual environment, the affected transmission lines and areas can be directly detected, with detailed dynamic information of line tripping time, location, duration, and causes. With the help of the 3D visualization system, digital results become more valuable and situation awareness of system operators is greatly enhanced, which is a reliable way to improve the safety and reliability of the entire power grid.

As mentioned in this survey, the development of future smart grid will towards a huge and complex energy system, which is deeply integrated with traditional power and renewable energies, as well as the powerful information and communication systems. The energy system also represents three levels or perspectives of the entire objective existence: physical energy level, industrial information level, and human society level. Under this big picture, more researchers are focusing on novel dimensions. For newly developed machine learning and data mining tools, deep learning, transfer learning, and multidata fusion methods are receiving extensive attention in recent years. Deep learning integrates supervised and unsupervised learning, with multiple hidden layer artificial neural network structures, and is capable of extracting abstract conceptions from data. While transfer learning makes a break through fundamental assumptions of the statistical learning theory, it can improve learning accuracy by utilizing the correlated data with different distributions. Multidata fusion technique is capable of analysing heterogeneous datasets collecting from different data sources; thus, it can extract more useful information.

By applying the above-mentioned new methods and technologies, more research directions and topics gradually appear. Firstly, the load prediction and modeling problem is the earliest application of data mining and analytics. Along with the fast installations of smart meters, much more precisely load modeling can be achieved by utilizing the equipment data and electrical measurements at both transient and steady states. More machine learning methods are available for load prediction and modeling, including feed-forward artificial neural network, SVMs, recurrent artificial neural networks, and regression trees, among others. Secondly, the fusion and merging analysis of the power system and transportation system can be done along with the increasing number of electrical vehicles. Considering the load data from charging stations, traffic flow and transportation network, on-board GPS tracks of electrical vehicles, and other data related to the driving and charing behaviors, a research on the driving and charging behavior characteristics is achievable. Closely related to that, the electricity market prediction and simulation is another possible hot topic, which can also be applied in many aspects such as evaluation of market shares for the individual power company,

investment income for power generation, and decision-making for power market mechanism design.

In conclusion, this paper has demonstrated a glance of the crossover and merging of the latest Big Data technology and smart grid technology. There are still many researchworks to do in the future. From all the application aspects, Big Data technology for human behavior in panorama mode has a great and long-term potential in real-time future smart grid and energy system, even the city planning, pollution abatement, transportation planning, and other useful applications.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (51607177, 61433012, and U1435215), Shenzhen Science and Technology Innovation Commission application demonstration project (No. KJYY20160608154421217), and China Postdoctoral Science Foundation (2018M631005).

References

[1] Y. XUE, "Energy internet or comprehensive energy network?," *Journal of Modern Power Systems and Clean Energy*, vol. 3, no. 3, pp. 297–301, 2015.

[2] Y. Xue and Y. Lai, "Integration of macro energy thinking and big data thinking part one big data and power big data," *Automation of Electric Power Systems*, vol. 40, no. 1, pp. 1–8, 2016.

[3] Y. Xue and Y. Lai, "Integration of macro energy thinking and big data thinking: part two applications and exploration," *Automation of Electric Power Systems*, vol. 40, no. 8, pp. 1–13, 2016.

[4] A. Gandomi and M. Haider, "Beyond the hype: big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[5] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, p. 1, 2017.

[6] Chinese Society of Electrical Engineering, "Chinese white paper on the development of large power data," pp. 1–10, 2013.

[7] A. A. Munshi and Y. A.-R. I. Mohamed, "Big data framework for analytics in smart grids," *Electric Power Systems Research*, vol. 151, pp. 369–380, 2017.

[8] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Zomaya, "Robust big data analytics for electricity price forecasting in the smart grid," *IEEE Transactions on Big Data*, p. 1, 2017.

[9] D. Wang and Z. Sun, "Big data analysis and parallel load forecasting of electric power user side," *Proceedings of the Csee*, vol. 35, no. 3, pp. 527–537, 2015.

[10] B. Wang, B. Fang, Y. Wang, H. Liu, and Y. Liu, "Power system transient stability assessment based on big data and the core vector machine," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2561–2570, 2016.

[11] W. Alves, D. Martins, U. Bezerra, and A. Klautau, "A hybrid approach for big data outlier detection from electric power scada system," *IEEE Latin America Transactions*, vol. 15, no. 1, pp. 57–64, 2017.

[12] Y. Zhao, P. Liu, Z. Wang, L. Zhang, and J. Hong, "Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods," *Applied Energy*, vol. 207, pp. 354–362, 2017.

[13] J. Baek, Q. H. Vu, J. K. Liu, X. Huang, and Y. Xiang, "A secure cloud computing based framework for big data information management of smart grid," *IEEE Transactions on Cloud Computing*, vol. 3, no. 2, pp. 233–244, 2015.

[14] S. J. Plathottam, H. Salehfar, and P. Ranganathan, "Convolutional neural networks (cnns) for power system big data analysis," in *2017 North American Power Symposium (NAPS)*, pp. 1–6, Morgantown, WV, USA, September 2017.

[15] S. Sagiroglu, R. Terzi, Y. Canbay, and I. Colak, "Big data issues in smart grid systems," in *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pp. 1007–1012, Birmingham, UK, November 2016.

[16] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: from big data to big insights," *Renewable & Sustainable Energy Reviews*, vol. 56, pp. 215–225, 2016.

[17] G. N. Korres and N. M. Manousakis, "State estimation and bad data processing for systems including pmu and scada measurements," *Electric Power Systems Research*, vol. 81, no. 7, pp. 1514–1524, 2011.

[18] Pacific Gas and Electric Company, "Pacific gas and electric," 2013, https://www.pge.com/.

[19] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid – a review," *Renewable & Sustainable Energy Reviews*, vol. 79, pp. 1099–1107, 2017.

[20] J. Zhu, E. Zhuang, J. Fu, J. Baranowski, A. Ford, and J. Shen, "A framework-based approach to utility big data analytics," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 2455–2462, 2016.

[21] The Apache Software Foundation, "The apache hadoop," 2005, http://hadoop.apache.org/index.html.

[22] The Apache Software Foundation, "The apache spark," 2000, http://spark.apache.org/.

[23] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[24] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, p. 29, 2003.

[25] F. Chang, J. Dean, S. Ghemawat et al., "Bigtable: a distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, no. 2, pp. 1–4, 2008.

[26] J. Yates, J. D. Mcgregor, J. E. Ingram, and J. Yates, "Hadoop and its evolving ecosystem, in: International Workshop on Software," in *5th International Workshop on Software Ecosystems (IWSECO)*, pp. 57–68, Potsdam, Germany, June 2013.

[27] R. B. Ray, M. Kumar, and S. K. Rath, *Fast Computing of Microarray Data Using Resilient Distributed Dataset of Apache Spark*, Springer International Publishing, 2016.

[28] K. Chodorow, *MongoDB: The Definitive Guide*, O'Reilly Media, Inc., 2013.

[29] A. Thusoo, J. S. Sarma, and N. Jain, "Hive - a petabyte scale data warehouse using hadoop," in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 996–1005, Long Beach, CA, USA, March 2010.

[30] S. Alsubaiee, K. Faraaz, E. Gabrielova et al., "Asterixdb: a scalable, open source bdms," *Proceedings of the VLDB Endowment*, vol. 7, no. 14, pp. 1905–1916, 2014.

[31] P. Pirzadeh, M. J. Carey, and T. Westmann, "Bigfun: a performance study of big data management system functionality," in *2015 IEEE International Conference on Big Data (Big Data)*, pp. 507–514, Santa Clara, CA, USA, November 2015.

[32] International Business Machines, "Ibm energy and utilities," 2015, April 2018, http://www-935.ibm.com/industries/energy/case_studies.html.

[33] A. Joiner, "Big data changes everything," 2014, April 2017, http://h20435.www2.hp.com/t5/HP-Software/Big-Data-is-changing-everything/ba-p/100623#.V3JBok9Z_hV.

[34] ORACLE, "Leverage big data and analytics," 2014, April 2018, https://www.oracle.com/industries/utilities/electricity/index.html.

[35] International Business Machines, "Ibm energy and utilities," 2017, August 2018, https://www.ibm.com/industries/uk-en/energy/.

[36] International Business Machines, "Ibm energy and utilities, centerpoint energy," 2017, August 2018, https://www.ibm.com/industries/uk-en/energy/case-studies.html.

[37] C3IoT, "C3 iot platform," 2009, April 2018, http://c3iot.com/products/#energy_grid.

[38] Opower, "Elevate your customer experience," 2007, April 2018, http://www.opower.com/.

[39] Solargis, "Accurate and efficient solar energy assessment," 2010, April 2018, http://solargis.info/.

[40] AutoGrid, "Turning data into power," 2011, December 2017, http://www.auto-grid.com/.

[41] Baidu, "Baidu Big Data," 2011, December http://bdp.baidu.com/.

[42] Aliyun, "Aliyun data ide," 2011, December 2017, https://data.aliyun.com/product/ide?spm=a2c0j.7906235.header.11.ntdqdP.

[43] Tencent, "Tencent big data," 2009, December 2017, http://bigdata.qq.com/.

[44] Inspur, "Inspur," 2009, December 2017, http://www.inspur.com/.

[45] Huawei, "Fusioninsight," 2015, December 2017, http://e.huawei.com/cn/products/cloud-computing-dc/cloud-computing/bigdata/fusioninsight.

[46] Lenovo, "Lenovo thinkclouds," 2016, December 2017, http://appserver.lenovo.com.cn/Lenovo_Series_List.aspx?CategoryCode=A30B03.

[47] S. Guha and A. Mcgregor, "Stream order and order statistics: quantile estimation in random-order streams," *SIAM Journal on Computing*, vol. 38, no. 5, pp. 2044–2059, 2009.

[48] D. Terry, D. Goldberg, D. Nichols, and B. Oki, "Continuous queries over append-only databases," *ACM SIGMOD Record*, vol. 21, no. 2, pp. 321–330, 1992.

[49] M. Sullivan, "Tribeca: A Stream Database Manager for Network Traffic Analysis," in *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases*, Mumbai (Bombay), India, September 1996.

[50] J. Chen, D. J. Dewitt, F. Tian, and Y. Wang, *NiagaraCQ: a Scalable Continuous Query System for Internet Databases*, ACM, 2000.

[51] S. D. Viglas and J. F. Naughton, "Rate-based query optimization for streaming information sources," in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data - SIGMOD '02*, pp. 37–48, Madison, Wisconsin, June 2002.

[52] A. Arasu, S. Babu, and J. Widom, "The cql continuous query language: semantic foundations and query execution," *VLDB Journal*, vol. 15, no. 2, pp. 121–142, 2006.

[53] D. Carney, U. Çetintemel, M. Cherniack et al., "Monitoring streams — a new class of data management applications," in *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*, pp. 215–226, Hong Kong SAR, China, August 2002.

[54] J. M. Hellerstein, M. J. Franklin, S. Chandrasekaran et al., "Adaptive query processing: technology in evolution," *IEEE Data Engineering Bulletin*, vol. 23, pp. 7–18, 2000.

[55] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data streams systems," in *PODS '02 Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 1–16, Madison, Wisconsin, June 2002.

[56] A. Toshniwal and D. Taneja, "Storm @ twitter," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, ACM*, pp. 147–156, Snowbird, Utah, USA, June 2014.

[57] Z. Shao, "Real-time analytics at facebook," 2015, http://www-conf.slac.stanford.edu/xldb2011/talks/xldb2011_tue_0940_facebookrealtimeanalytics.pdf.

[58] C. Riccominig, "How linkedin uses apache samza," 2014, http://www.infoq.com/articles/linkedin-samza.

[59] Z. Qian, Y. He, C. Su et al., "Timestream: reliable stream computation in the cloud," in *Proceedings of the 8th ACM European Conference on Computer Systems - EuroSys '13*, pp. 1–14, Prague, Czech Republic, April 2013.

[60] C. Cranor, T. Johnson, O. Spataschek, and V. Shkapenyuk, "Gigascope: a stream database for network applications," in *Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD '03*, pp. 647–651, San Diego, California, June 2003.

[61] R. Meier, E. Cotilla-Sanchez, B. Mccamish, and D. Chiu, "Power system data management and analysis using synchrophasor data," in *2014 IEEE Conference on Technologies for Sustainability (SusTech)*, pp. 225–231, Portland, OR, USA, July 2014.

[62] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of Computational & Graphical Statistics*, vol. 5, no. 5, pp. 299–314, 1996.

[63] G. Holmes, A. Donkin, and I. H. Witten, "Weka: a machine learning workbench," in *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, pp. 357–361, Brisbane, Queensland, Australia, December 1994.

[64] J. W. Eaton, "gnu octave," 2014, January 2018, http://www.gnu.org/software/octave/.

[65] R. R. Bouckaert, E. Frank, M. A. Hall et al., "WEKAâˆ'Experiences with a Java Open-Source Project," *Journal of Machine Learning Research*, vol. 11, no. 5, pp. 2533–2541, 2010.

[66] F. Morandat, B. Hill, L. Osvald, and J. Vitek, "Evaluating the design of the r language - objects and functions for data analysis," in *Proceedings of the 26th European Conference on Object-Oriented Programming*, pp. 104–131, Beijing, China, June 2012.

[67] S. Sonnenburg, G. Tsch, S. Henschel et al., "The shogun machine learning toolbox," *Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, 2010.

[68] L. M. Surhone, M. T. Tennoe, and S. F. Henssonow, *AForge.NET*, Betascript Publishing, 2010.

[69] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in Action*, Manning Publications Co., 2011.

[70] X. Meng, J. Bradley, B. Yavuz et al., "Mllib: machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2015.

[71] F. Pedregosa, G. Varoquaux, and E. Duchesnay, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[72] J. Demšar, T. Curk, A. Erjavec et al., "Orange: data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[73] A. Krizhevskey, "Cuda-convnet," 2012, April 2018, http://code.google.com/p/cuda-convnet/.

[74] A. Karpathy, "Convnetjs:deep Learning in your browser," http://cs.stanford.edu/people/karpathy/convnetjs/.

[75] S. Owen, "Cloudera oryx: Simple real-time large-scale machine learning infrastructure," 2014, https://github.com/cloudera/oryx.

[76] T. Lumley, "biglm: bounded memory linear and generalized linear models," 2014, http://cran.r-project.org/web/packages/biglm/index.html.

[77] L. B. A. Lim and A. Cutler, "bigrf: Big random forests: classification and regression forests for large data sets," 2014, http://cran.rproject.org/web/packages/bigrf/index.html.

[78] S. T. I. Feinerer, "hive: Hadoop interactive," 2014, http://cran.rproject.org/web/packages/hive/index.html.

[79] D. M. Beazley, "Swig: an easy to use tool for integrating scripting languages with c and c++," in *4th Annual Tcl/Tk Workshop*, Monterey, CA, July 1996.

[80] G. Marchesan, M. R. Muraro, G. Cardoso, L. Mariotto, and A. P. de Morais, "Passive method for distributed-generation island detection based on oscillation frequency," *IEEE Transactions on Power Delivery*, vol. 31, no. 1, pp. 138–146, 2016.

[81] X. Xu, Z. Yan, M. Shahidehpour, H. Wang, and S. Chen, "Power system voltage stability evaluation considering renewable energy with correlated variabilities," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3236–3245, 2018.

[82] J. Liu, N. Tai, and C. Fan, "Transient-voltage-based protection scheme for DC line faults in the multiterminal VSC-HVDC system," *IEEE Transactions on Power Delivery*, vol. 32, no. 3, pp. 1483–1494, 2017.

[83] M. Sahraei-Ardakani, X. Li, P. Balasubramanian, K. W. Hedman, and M. Abdi-Khorsand, "Real-time contingency analysis with transmission switching on real power system data," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 2501-2502, 2016.

[84] S. Brahma, R. Kavasseri, H. Cao, N. R. Chaudhuri, T. Alexopoulos, and Y. Cui, "Real-time identification of dynamic events in power systems using PMU data, and potential applications—models, promises, and challenges," *IEEE Transactions on Power Delivery*, vol. 32, no. 1, pp. 294–301, 2017.

[85] J. Zhong, W. Li, C. Wang, and J. Yu, "A rankboost based data-driven method to determine maintenance priority of circuit breakers," *IEEE Transactions on Power Delivery*, vol. 33, no. 3, pp. 1044–1053, 2018.

[86] Y. Liu, Y. Guo, Z. Yang, J. Hu, G. Lu, and Y. Wang, "Power system transmission line tripping analysis using a big data platform with 3d visualization," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, Honolulu, HI, USA, November 2017.

[87] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[88] Y. Guo, K. Li, and D. Laverty, "A statistical process control approach for automatic anti-islanding detection using synchrophasors," in *2013 IEEE Power & Energy Society General Meeting*, pp. 1–5, Vancouver, BC, Canada, July 2013.

[89] K. Li, Y. Guo, D. Laverty, H. He, and M. Fei, "Distributed adaptive learning framework for wide area monitoring of power systems integrated with distributed generations," *Energy and Power Engineering*, vol. 5, no. 4, pp. 962–969, 2013.

[90] Y. Guo, K. Li, and D. M. Laverty, "Loss-of-main monitoring and detection for distributed generations using dynamic principal component analysis," *Journal of Power and Energy Engineering*, vol. 2, no. 4, pp. 423–431, 2014.

[91] Y. Guo, K. Li, D. M. Laverty, and Y. Xue, "Synchrophasor-based islanding detection for distributed generation systems using systematic principal component analysis approaches," *IEEE Transactions on Power Delivery*, vol. 30, no. 6, pp. 2544–2552, 2015.

[92] A. Kheirkhah, A. Azadeh, M. Saberi, A. Azaron, and H. Shakouri, "Improved estimation of electricity demand function by using of artificial neural network, principal component analysis and data envelopment analysis," *Computers & Industrial Engineering*, vol. 64, no. 1, pp. 425–441, 2013.

[93] A. Onwuachumba and M. Musavi, "New reduced model approach for power system state estimation using artificial neural networks and principal component analysis," in *2014 IEEE Electrical Power and Energy Conference*, pp. 15–20, Calgary, AB, Canada, November 2014.

[94] Y. Guo, K. Li, Z. Yang, J. Deng, and D. M. Laverty, "A novel radial basis function neural network principal component analysis scheme for pmu-based wide-area power system monitoring," *Electric Power Systems Research*, vol. 127, pp. 197–205, 2015.

[95] K. Li, J.-X. Peng, and G. W. Irwin, "A fast nonlinear model identification method," *IEEE Transactions on Automatic Control*, vol. 50, no. 8, pp. 1211–1216, 2005.

[96] K. Li, J.-X. Peng, and E.-W. Bai, "Two-stage mixed discrete-continuous identification of radial basis function (rbf) neural models for nonlinear systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 3, pp. 630–643, 2009.

[97] Z. Yang, K. Li, Q. Niu, Y. Xue, and A. Foley, "A self-learning tlbo based dynamic economic/environmental dispatch considering multiple plug-in electric vehicle loads," *Journal of Modern Power Systems and Clean Energy*, vol. 2, no. 4, pp. 298–307, 2014.

[98] Z. Yang, K. Li, Q. Niu, and Y. Xue, "A comprehensive study of economic unit commitment of power systems integrating various renewable generations and plug-in electric vehicles," *Energy Conversion and Management*, vol. 132, pp. 460–481, 2017.

[99] Z. Yang, K. Li, Q. Niu, and Y. Xue, "A novel parallel-series hybrid meta-heuristic method for solving a hybrid unit

commitment problem," *Knowledge-Based Systems*, vol. 134, pp. 13–30, 2017.

[100] Q. Shao and C. J. Feng, "Pattern recognition of chatter gestation based on hybrid pca-svm," *Applied Mechanics and Materials*, vol. 120, pp. 190–194, 2011.

[101] M. D. Farrell and R. M. Mersereau, "On the impact of pca dimension reduction for hyperspectral detection of difficult targets," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 2, pp. 192–195, 2005.

[102] L. I. Kuncheva and W. J. Faithfull, "Pca feature extraction for change detection in multidimensional unlabeled data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 69–80, 2014.

[103] Q. Jiang, X. Yan, and B. Huang, "Performance-driven distributed pca process monitoring based on fault-relevant variable selection and bayesian inference," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 1, pp. 377–386, 2016.

[104] R. Zhang, W. Cai, L. Ni, and G. Lebby, "Power system load forecasting using partial least square method," in *2008 40th Southeastern Symposium on System Theory (SSST)*, pp. 169–173, New Orleans, LA, USA, March 2008.

[105] W. Zheng and H. Wang, "Organizational performance evaluation of power supply with partial least-squares regression," in *2011 IEEE 18th International Conference on Industrial Engineering and Engineering Management*, pp. 161–163, Changchun, China, September 2011.

[106] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data — with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[107] C. A. Jensen, M. A. El-Sharkawi, and R. J. Marks, "Power system security assessment using neural networks: feature selection using fisher discrimination," *IEEE Transactions on Power Systems*, vol. 16, no. 4, pp. 757–763, 2001.

[108] R. Eriksson and L. Soder, "Wide-area measurement system-based subspace identification for obtaining linear models to centrally coordinate controllable devices," *IEEE Transactions on Power Delivery*, vol. 26, no. 2, pp. 988–997, 2011.

[109] C. Luo and V. Ajjarapu, "Invariant subspace based eigenvalue tracing for power system small-signal stability analysis," in *2009 IEEE Power & Energy Society General Meeting*, pp. 1–9, Calgary, AB, Canada, July 2009.

[110] J. Yang, W. Li, T. Chen, W. Xu, and M. Wu, "Online estimation and application of power grid impedance matrices based on synchronised phasor measurements," *IET Generation, Transmission & Distribution*, vol. 4, no. 9, p. 1052, 2010.

[111] A. H. Al-Mohammed and M. A. Abido, "A fully adaptive PMU-based fault location algorithm for series-compensated lines," *IEEE Transactions on Power Systems*, vol. 29, no. 5, pp. 2129–2137, 2014.

[112] E. L. Russell, L. H. Chiang, and R. D. Braatz, "Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, no. 1, pp. 81–93, 2000.

[113] R. Srinivasan, C. Wang, W. K. Ho, and K. W. Lim, "Dynamic principal component analysis based methodology for clustering process states in agile chemical plants," *Industrial & Engineering Chemistry Research*, vol. 43, no. 9, pp. 2123–2139, 2004.

[114] M. Chen and L. X. Guo, "The synthetic evaluation method of the dynamic performance and economic performance of battery electric vehicle based on principal component analysis," *Applied Mechanics and Materials*, vol. 215-216, pp. 1259–1262, 2012.

[115] W. Sun and G. Ma, "Condition assessment of power supply equipment based on kernel principal component analysis and multi-class support vector machine," in *2009 Fifth International Conference on Natural Computation*, pp. 485–488, Tianjin, China, August 2009.

[116] J. Ni, C. Zhang, and S. X. Yang, "An adaptive approach based on KPCA and SVM for real-time fault diagnosis of HVCBs," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1960–1971, 2011.

[117] Z. Weiqing, S. Fengqi, X. Zhigao, Q. Zongliang, and Z. Jianxin, "An investigation on system anomaly source diagnosis using kpca-fpsdg," in *2012 Asia-Pacific Power and Energy Engineering Conference*, pp. 1–4, Shanghai, China, March 2012.

[118] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[119] S. Theodoridis and K. Koutroumbas, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

[120] D. Q. Zhou, U. D. Annakkage, and A. D. Rajapakse, "Online monitoring of voltage stability margin using an artificial neural network," *IEEE Transactions on Power Systems*, vol. 25, no. 3, pp. 1566–1574, 2010.

[121] M.-R. Mosavi and A. Tabatabaei, "Traveling-wave fault location techniques in power system based on wavelet analysis and neural network using gps timing," *Wireless Personal Communications*, vol. 86, no. 2, pp. 835–850, 2016.

[122] Y. Zhang and S. J. Qin, "Fault detection of nonlinear processes using multiway kernel independent component analysis," *Industrial & Engineering Chemistry Research*, vol. 46, no. 23, pp. 7780–7787, 2007.

[123] M. Ruiz-Llata, G. Guarnizo, and C. Boya, "Embedded power quality monitoring system based on independent component analysis and svms," in *The 2011 International Joint Conference on Neural Networks*, pp. 2229–2234, San Jose, CA, USA, July 2011.

[124] C. Uzunoglu, M. Ugur, F. Turan, and S. Cekli, "Amplitude and frequency estimation of power system signals using independent component analysis," in *2013 21st Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, Haspolat, Turkey, April 2013.

[125] G. Valverde, A. T. Saric, and V. Terzija, "Stochastic monitoring of distribution networks including correlated input variables," *IEEE Transactions on Power Delivery*, vol. 28, no. 1, pp. 246–255, 2013.

[126] R. Singh, B. C. Pal, and R. A. Jabr, "Statistical representation of distribution system loads using gaussian mixture model," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 29–37, 2010.

[127] X. Liu, L. Xie, U. Kruger, T. Littler, and S. Wang, "Statistical-based monitoring of multivariate non-gaussian systems," *AICHE Journal*, vol. 54, no. 9, pp. 2379–2391, 2008.

[128] X. Liu, D. McSwiggan, T. B. Littler, and J. Kennedy, "Measurement-based method for wind farm power system oscillations monitoring," *IET Renewable Power Generation*, vol. 4, no. 2, p. 198, 2010.

[129] R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica, "Graphx: a resilient distributed graph system on spark," in *First International Workshop on Graph Data Management*

Experiences and Systems - GRADES '13, pp. 1–6, New York, June 2013.

[130] F. Pérez and B. E. Granger, "IPython: a system for interactive scientific computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 21–29, 2007.

[131] L. Yin and S.-L. Shaw, "Exploring space–time paths in physical and social closeness spaces: a space–time gis approach," *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 742–761, 2015.

[132] Z. Yang, K. Li, and A. Foley, "Computational scheduling methods for integrating plug-in electric vehicles with power systems: a review," *Renewable & Sustainable Energy Reviews*, vol. 51, pp. 396–416, 2015.

[133] M. Jahangiri, R. Ghaderi, A. Haghani, and O. Nematollahi, "Finding the best locations for establishment of solar-wind power stations in middle-east using gis: a review," *Renewable & Sustainable Energy Reviews*, vol. 66, pp. 38–52, 2016.

[134] M. A. Anwarzai and K. Nagasaka, "Utility-scale implementable potential of wind and solar energies for Afghanistan using gis multi-criteria decision analysis," *Renewable & Sustainable Energy Reviews*, vol. 71, pp. 150–160, 2017.

[135] B. He, W. X. Mo, J. X. Hu, G. Yang, G. J. Lu, and Y. Q. Liu, "Development of power grid web3d gis based on cesium," in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pp. 2465–2469, Xi'an, China, October 2016.