14th International Conference on Protection & Automation in Power System
Amirkabir University of Technology (Tehran Polytechnic), Tehran
December 31, 2019 - January 1, 2020

Amirkabir University of Technology
(Tehran Polytechnic)

Attributes of Big Data Analytics for Data-Driven Decision Making in Cyber-Physical Power Systems

Jalal Moradi
Young Researchers and Elite Club
Khomeinishahr Branch, Islamic Azad University
Isfahan, Iran
sj.moradi@iaukhsh.ac.ir

Hossein Shahinzadeh
IEEE Member, Department of Electrical Engineering
Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
h.s.shahinzadeh@ieee.org

Hamed Nafisi
Department of Electrical Engineering
Amirkabir University of Technology
Tehran, Iran
nafisi@aut.ac.ir

Mousa Marzband
Department of Electrical Engineering
Northumbria University
Newcastle, United Kingdom
mousa.marzband@northumbria.ac.uk

Gevork B. Gharehpetian
Prof., Department of Electrical Engineering
Amirkabir University of Technology
Tehran, Iran
grptian@aut.ac.ir

***Abstract*— Big data analytics is a virtually new term in power system terminology. This concept delves into the way a massive volume of data is acquired, processed, analyzed to extract insight from available data. In particular, big data analytics alludes to applications of artificial intelligence, machine learning techniques, data mining techniques, time-series forecasting methods. Decision-makers in power systems have been long plagued by incapability and weakness of classical methods in dealing with large-scale real practical cases due to the existence of thousands or millions of variables, being time-consuming, the requirement of a high computation burden, divergence of results, unjustifiable errors, and poor accuracy of the model. Big data analytics is an ongoing topic, which pinpoints how to extract insights from these large data sets. The extant article has enumerated the applications of big data analytics in future power systems through several layers from grid-scale to local-scale. Big data analytics has many applications in the areas of smart grid implementation, electricity markets, execution of collaborative operation schemes, enhancement of microgrid operation autonomy, management of electric vehicle operations in smart grids, active distribution network control, district hub system management, multi-agent energy systems, electricity theft detection, stability and security assessment by PMUs, and better exploitation of renewable energy sources. The employment of big data analytics entails some prerequisites, such as the proliferation of IoT-enabled devices, easily-accessible cloud space, blockchain, etc. This paper has comprehensively conducted an extensive review of the applications of big data analytics along with the prevailing challenges and solutions.**

Keywords*— Big data analytics, Machine learning and data mining, Future power systems, Artificial intelligence, Clustering and classification, Smart grids.

I. Introduction

In recent years, the emerging concept of the applications of big data analytics in the area of smart grids has been propounded. The term “big data” refers to the procedure of massive data acquisition, processing, and analyzing using artificial intelligence techniques such as data mining, machine learning, and neural networks for some purposes such as clustering, pattern recognition, classification, feature selection, or time series forecasting, etc. Power systems are progressively changing toward smart and digitalized grids, which have a higher level of autonomy and complexity. The penetration of large-scale renewable energy sources (RES) is accelerated, which alludes to the increase in the level of uncertainty in power systems. The pervasiveness of microgrids and active distribution networks with mounting penetration of virtual power plants (VPPs) in demand-side is an ever-growing trend in the current power systems. The coordination of interconnected microgrids has immense complexity that requires elaborate controlling schemes to be dealt with. Regard to the proliferation of microsources such as roof-top solar panels, small-scale wind turbines, and a variety of small-scale storage facilities as well as incremental penetration of electric vehicles (EVs), particularly with vehicle-to-grid (V2G) capability, distribution networks are evolving to AC/DC mesh grids. The ongoing technological developments in distributed technologies such as internet of things (IoT), particularly in intelligent electronic devices (IEDs), sensors, and actuators, have been led to the production of a large body of data that can be employed for various purposes in power systems. The penetration of EVs can immensely impact on the operation of power systems soon. The discovery of consumption and feed-in patterns of EVs and incorporation of them requires intelligent computational tools. The existence of a huge quantity of microsources in a restructured environment will be led to local markets that must handle microtransactions. This matter conveys the necessity of the use of blockchain and cryptocurrency and indicates that a large body of data will be generated that can be explored to find useful patterns for better operation of future energy grids [1]. These abovementioned features are some of the underlying attributes of future power systems, which imply future smart electrical grids that differ from the current power grids because of mounted decentralization, enhanced monitoring capabilities, expanded high-frequency communication, and diversified generation sources. The expansion of smart grid infrastructures, the proliferation of IoT, and increasing the penetration of RES are the major thrusts in the ongoing research conducting in the area of future power systems [2]. Fig. 1 illustrates the different areas of applications of big data analytics in power systems and the interconnections.

During recent years many scholars have addressed some of the application of big data in power systems. This subject is broad, and the presented works in the literature have not provided a comprehensive review to cover all categories and challenges. The previous works usually have covered some areas rather than a broad point of view. In [3], a review is presented about the various usages of big data in distributions level. Some authors have introduced the role of big data in transient stability assessment [4]. A similar study is also conducted on the security and resilience assessment in large-scale grids using big data techniques [5]. Many researchers have also addressed the applications of big data and cloud computing in the areas of wide-area measurement, wide-area monitoring, and wide-area protection [6,7]. Some authors have also proposed common big data techniques used in smart grids [8-10]. Some authors have investigated the role of big data analytics in demand response modeling [11]. The applications of big data in renewable power generation are also addressed in [12,13]. An example of the use of big data in electricity markets and the restructured environment is presented in [14]. The authors in [15], have proposed a concise review, and they explained the state-of-the-art opportunities and future direction.

In the extant study, a new well-designed classification of this topic is done, while many details are elaborately discussed. The categorization of subjects is made based on a holistic point of view, which provides a better overview of the topic. In each part, aside from open challenges and barriers, the possible solutions and prospects are also suggested. In other words, this paper presents a fully-fledged review of the applications of big data analytics in modernized cyber-physical power systems for data-driven decision making. In the next section, the terminology of big data is elucidated. In the third section, the algorithms and methods proposed for big data handling are introduced. In the fourth section, the possible challenges faced or to be faced in the smart grid implementation using big data analytics have been investigated. In section 5, an elaborate description of various applications and initial requirements pertaining to the deployment of big data analytics for better exploitation of resources and more flexible, secure, reliable, environmentally sustainable, and economical operation is presented. Ultimately, the conclusions are presented in the end.

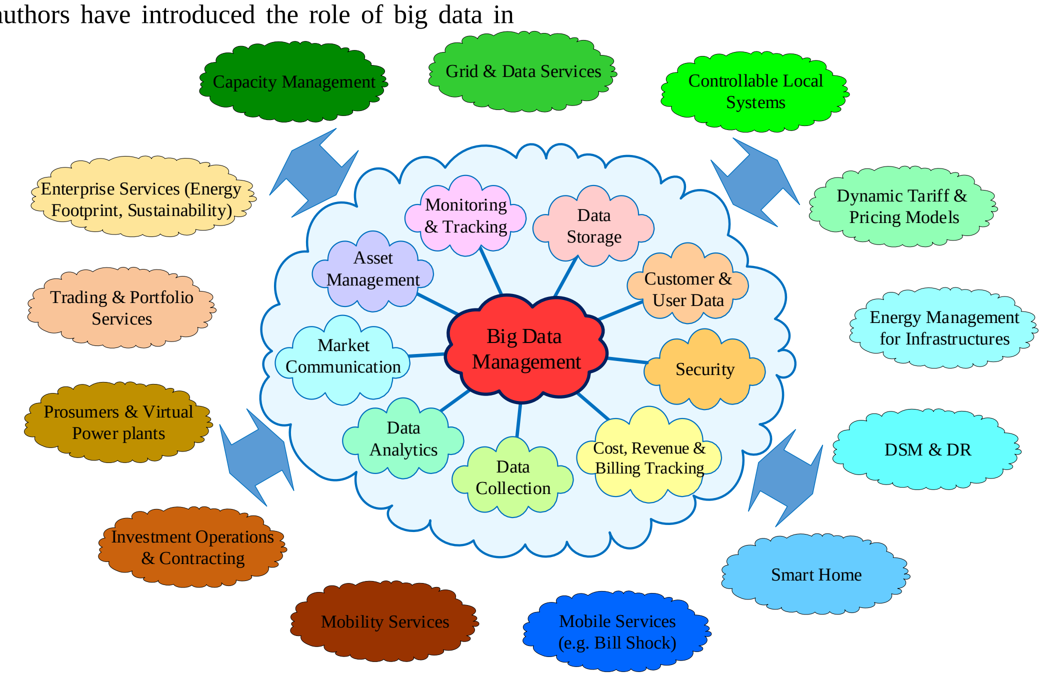

Figure 1. Big data management structure.

II. Big data analytics definition

The term data refers to any type of sensed output from any source, whereas big data denotes a massive body of data that is too complex, big, and overwhelming to be processed and analyzed using traditional methods. Hence, the volume, velocity, value, and variety are regarded as four primary attributes of big data. These attributes convey the big size of data, the wide diversity of types of data (strings, numbers, texts, encrypted data, etc.), the importance of data and analytics, and real-time data acquisition. If traditional methods have adequate time to process the data, they eventually can analyze any large volume dataset. However, the ever-growing continuous acquisition of data severely hedges the data processing that it reveals the importance of velocity and precludes traditional methods to be used [16]. Besides, the term analytics denotes the systematic computational analysis of data or statistics. In all facets of data processing from any system, the amount of data is not important, but what really matters is what will be done with these data and how to make sense out of data. In other words, the value of data must be extracted, and the required information must be extrapolated. The necessity of online real-time decision-making is growingly and broadly recognized in various sectors of everyday life and particularly in industry, which requires a large reliable computation burden as well as a highly equipped environment. The application domain of big data is phased in for health care sector, transportation sector, financial systems, search engines, social networking, to name just a few. Big data helps to acquire a better perception of complicated systems, such as power systems, to assist sustainable development socially, economically, and environmentally [17].

III. Algorithms for big data processing

The most prevalent methods for dealing with big data analytics are machine learning and deep learning methods. Machine learning refers to the deployment of artificial intelligence (AI) to teach a machine (a computer system) by exploring patterns and discovering inferences among unclassified training data without the use of explicitly programmed instructions. The term “machine learning” was first coined by Arthur Samuel in 1595. There are different types of learning algorithms, which are employed in machine learning analytics, such as supervised and unsupervised learning, reinforcement learning, feature learning, anomaly detection, association rule learning, etc. The generality of supervised learning expresses a mathematical model made up of appropriate inputs and the desired outputs. Classification algorithms, support vector machine (SVM), and regression algorithms are recognized as some common types of supervised learning. In unsupervised learning methods, a mathematical model is built based on eligible inputs and regardless of desired outputs in order to extrapolate patterns and structure. This method is normally employed to explore and discover the existing structures in a dataset. Clustering and dimensionality reduction are two well-known models of unsupervised learning. Pattern recognition is also a model that can be implemented both supervised or unsupervised. Reinforcement learning methods denote a category of machine learning that uses positive and negative feedback to improve the cognition of a system in a dynamic environment, and it is particularly used for decision-making and automatic controlling.

Knowledge discovery and data mining is another field of data science that it has a large and ambiguous degree of overlap with machine learning. However, they have conceptual differences in terms of meaning, history, tasks, origin, implementation, nature, applications, abstraction, techniques, and scopes. The most distinct difference between them is that data mining is used to extract knowledge from a big dataset, while machine learning is used to export a new algorithm from a big dataset that can operate virtually the same as the original system does. The goal of data mining methods is to extract the rules from existing data, while the goal of machine learning is to teach a computer (machine) to learn and perceive the given rules. Data mining requires human interference and creativity, while machine learning is automated, self-learned, once design self-implemented, and needless of human efforts in order to train a machine to do a task intelligently. It can be alleged that machine learning can be applied in a vast area, while data mining has a more limited scope. As a more comprehensive definition, machine learning works based on known properties that are learned from the training data, while data mining delves into the discovery of previously unknown properties. Data mining employs machine learning algorithms, although it follows different goals; machine learning, on the other hand, also uses data mining algorithms as a preprocessing step or as "unsupervised learning" to improve learner accuracy. Big data has also intimate ties to statistics and optimization. For example, some learning problems are formulated to minimize an objective function, which expresses the discrepancy between real actual values and the predicted values from the model being trained [18].

Deep learning, which is also called as hierarchical learning, is a subset of machine learning that is basically based on artificial neural networks (ANNs). Deep learning has multiple applications, such as computer vision and speech recognition, etc. ANNs consist of multiple consecutive layers to progressively extract latent features from raw inputs. In order to ease the computation in deep learning methods, especially for systems with complex structures, the dimensionality reduction, and numerosity reduction techniques are suggested. Time-series forecasting and classification is an intriguing topic used for various applications in the power system from large-scale to local-scale. One of the facets, which makes the operation of power systems challenging to be tackled, is the fact that the monitoring has to be done on streams of raw and processed data that have a time-series evolution. An accurate forecast helps the operator to better cope with disturbances [19].

IV. Challenges confronting the deployment of big data analytics in future power systems

In future power systems, a wide variety of challenges exists which compromise the reliability, security, and economy of supply. The main challenge correlated to the deployment of big data in power systems stems from the existence of a large number of operational constraints, the complex structures in electricity markets and clearance mechanisms, as well as computation time restrictions for many types of optimization in power systems. Some conducted researches indicate that the practical big data incorporation still has some hurdles in terms of storage, sharing, visualization, computation time, and computation burden. Firms that are prone to use big data analytics tools would like to employ less expensive processing and storage alternatives.

During the recent decade, many governments and private companies have largely invested in data management tools. These tools usually follow the cycle of data preparation, analysis, validation, collaboration, reporting, access, and retrieving in order to deal with different types of large-scale optimization problems with significant uncertainty of input parameters. These input data are gathered from various sources in several ways, mainly to solve a forecasting, classification, or optimization problem. The important point that should be noticed is that the classical optimization methods are not designed and able to handle the large data size properly.

There are some applied methods for dealing with such cases. Unconstrained optimization problems is a type of large-scale case in power system operation, which needs a high memory due to the existence of many variables. In this case, some penalty terms are included in the objective function. Even though some techniques are developed to solve such problems, the conjugate gradient method is suggested to deal with large-scale problems. Some problems in power system operation may have non-smooth functions (discontinuous or non-differentiable), which have abrupt bends in their graph. Many approaches are introduced to deal with small-scale problems, whereas these methods show weakness in scalability. However, the Bundle method is proposed to tackle large-scale cases. This method has two types of diagonal bundle method (D-Bundle) and limit memory bundle method (LMBM). In addition, some problems in the area of reliability, adequacy, risk assessment, state estimation, power markets, and long-term expansion planning have logistic functions, which require logistic optimization techniques. Such problems usually have a large data set of historical records that need big data techniques to be dealt with. Moreover, in some cases, the objective function or some of the constraints do not meet the convexity condition. Many real cases in power system operation and planning have non-convex nature. In such a case, particularly when a large data set is involved, it is very hard to find the global optimum point and the local optimum solutions are also inevitably acceptable [20]. The economic load dispatch (ELD) problem with consideration of the valve-point effect is case in point. Swarm intelligence algorithms are good tools to obtain local optimum solutions with relatively agile performances. Such algorithms usually follow the local random search techniques to seek for the best answer in the search space. Besides, some problems in practice have several objectives that should be maximized or minimized at the same time. Such problems are referred to as multi-objective problems, in which some functions conflict with each other. In such a circumstance, there are Pareto optimal solutions. For a big data set, not only the best accuracy must be achieved, but also the computation time is also important. Hence, some distributed optimization techniques such as alternating direction method of multipliers (ADMM), which is a kind of spectral partitioning approach, is suggested in order to obtain an answer close to a local optimum. Furthermore, during the recent two decades, many researchers and decision-makers have confronted with problems pertaining to optimization programming for real cases in terms of memory, convergence speed, and accuracy because the problem has comprised of thousands or millions of variables. Many approaches have been suggested to tackle high-dimensional optimization problems. Among these methods, metaheuristic methods sound more time-efficient with a fair computing performance. A myriad of scholars has used metaheuristic algorithms for the purpose of optimization is power systems. Many others have improved the existing methods, and some others have innovated new metaheuristic techniques that demonstrate better or the same performances. However, there are still vast open research fields, challenges, and issues for improvement. Among these methods, evolutionary algorithms have the fame of being powerful techniques. However, as the number of variables increases, the performance of optimizer is deteriorated. A scant improvement in the result can be concluded in outstanding economic saves in practical cases in power systems. In such cases, the hybridization of metaheuristic methods for large-scale high-dimensional problems is suggested, which shows superior performance. Some scholars have also suggested combining metaheuristics with fuzzy logic for searchability enhancement [2].

V. Deployment of big data for various applications in power system operation

Some big data techniques, such as time series techniques for load and price forecasting, have been widely used in power systems since the last decades. The advent of smart grids has converted a traditional electricity network into a data-driven industry. In a power system, various data must be acquired from physical devices and infrastructures so that the complexity, speed, and size of collected data is important.

During the recent decade, the volume of generating data in a power grid has been dramatically increased, which are used mainly for power market analyzing and power system state estimation through collecting data from phasor measurement units (PMUs, also known as time-synchronized phasor). By harnessing insights from these rich sets of data, the generation companies (GENCOs) and other market participants are able to improve their performance and increase their profitability. In a restructured market, a variety of market indicators must be forecasted, such as hourly spot market price, near-real-time customers' demand, and the production of RESs. These factors are salient decision-making inputs for electric utilities and GENCOs to establish the most effective strategy for boosting their own benefit. In addition, the data collected from PMUs' remote feedback signals should be optimized in order to be used in wide-area measurement systems (WAMS) for better system monitoring and detection of intra-area oscillations. Then the oscillations could be damped through special controllers. Big data analytics can be used for real-time self-healing schemes and distribution network reconfiguration. In the following parts, some of the main applications of big data analytics in the current and future power systems are elucidated.

A. Big data in smart grids

In order to efficiently manage power systems, particularly distribution systems, the smart grid idea is suggested. This concept has a wide range of requirements that should be fulfilled. One of the key prerequisites for the launch of a smart grid is the implementation of advanced information and communication technologies (ICT) and modern telecommunication infrastructures such as 5G wireless communication networks. In addition, the materialization of smart grids requires the utilization of advanced metering infrastructure (AMI). The use of an extensive sensor network for a wide variety of connected devices will produce a vast information flow inside the network. These data are heterogenous and with different frequencies. The main concern about such a large volume of data is how to classify, quantify, analyze, and take profit from them in a certain period with the best possible accuracy. The grid's planners and operators must analyze and discover these data to find out the insights from them in order to operate the network more efficiently. However, dealing with such a high amount of data brings up a tough challenge for the experts. The urgency of data processing increases as the volume of data grows, which is a matter of primary concern. Fig. 2 shows the big data usages in the direct domain of power systems, as well as some off-domain usages in electricity grids.

In future power systems, a myriad of elements of the grid is equipped with intelligent, robust, and reliable IoT-oriented devices, which report the latest state of grid in near real-time to the monitoring center, and the local or central controlling center will issue the proper controlling command. The IoT-enabled devices will increase the level of autonomy and smartness of the grid [21,22]. This matter is specifically important for electricity networks due to the dynamic nature of elements. The existence of a secure and high-performance communication network is indispensable for the interconnection between IoT-based devices and is the prerequisite for a smart electricity network. The main challenge is that how to manage and operate the physical network by a cyber system and taking advantage of the large volume of generated data as well as boosting the intelligence and autonomy of subsystems and components by manufacturing controlling equipment powered by well-thought-out controlling schemes and well-designed analytical programs. This matter alludes to the importance of big data techniques in future networks. At present, many ongoing studies are conducted to mature these methods in terms of speed and accuracy. Some controlling measures require real-time data from IoT-enabled devices, while some measures need to be supported by historical records. Thus, a storage server is needed, which can be a local storage center or a cloud space. In addition, the computation for making a decision by a controlling scheme can be processed through cloud computation. The security of such a data-driven network is an extraordinary facet of these communication networks, which has assumed particular attention. For instance, the deployment of blockchain and cryptocurrencies are suggested by some researchers [23].

The cyber-physical smart grids have two layers of physical component and the cyber layer, which are connected with IoT-enabled devices together with excellent communication infrastructure. This communication platform must have specific attributes in terms of latency, bandwidth, interoperability, flexibility, data throughput, and cybersecurity. Fig. 3 demonstrates the paradigm of big data analytics in a cyber-physical power system.

B. Big data in electricity markets

Currently, the data mining methods are applied by power system operators for security assessment, load forecasting, price forecasting, power system control, and fault detection. There are still big data issues corresponded with power system modeling in the areas of scheduling, unit commitment, and electricity market analysis.

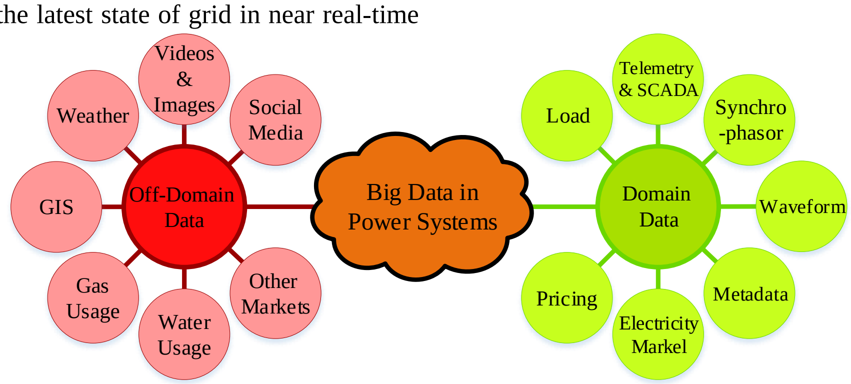

Figure 2. Big data usages in the power systems; background layer (off-domain) and direct usages in power systems.

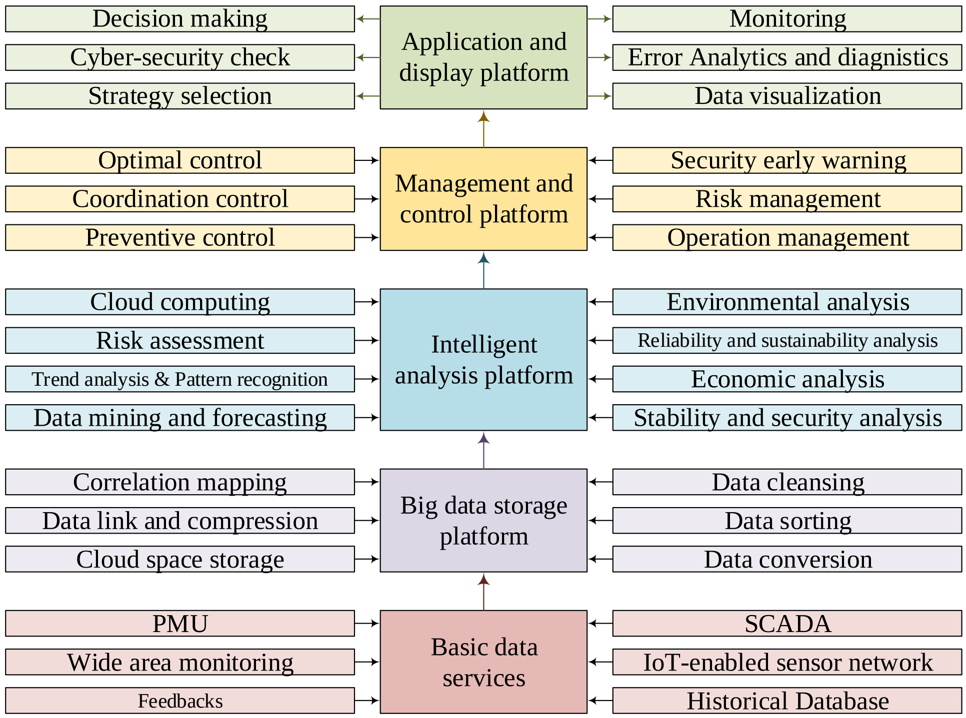

Figure 3. The structure of big data analytics in power systems.

1) unit commitment

In order to better recognize the enormity of big data issues in power system operation, the following extreme condition in a security-constrained unit commitment problem can be discussed. Such a problem usually is solved by a two-stage decomposition technique, which accounts for a master problem and a subproblem to arrange a day-ahead schedule. To capture the decision, the problem must seek the solution space by solving the problem constraints many times through internal loops and with respect to optimization cuts. The calculation of any loop requires a specific computation time (usually lower than 1 second). To reach a predefined duality gap (or relative gap), the procedure must be continued to a certain number of iterations to resolve all violations. Now suppose a system with a large number of buses and a list of thousands of *N*-1 or *N*-2 contingencies for 24 consecutive intervals. Such a problem definitely involves the solution of millions of contingency cases. To achieve all post-contingency results using AC power flow will definitely take many hours for such a system using a powerful computer. The consideration of operation strategies and stochastic conditions also exacerbates the solution and extends the overall computation time. Even though AC results are more accurate, but the operators tend or are obliged to use DC-power flow along with scenario reduction techniques to shorten the computation time and obtain the results faster with an essential level of accuracy. In this case, the operators use additional tools to analyze the constraints of voltage and reactive power flow. Thus, the improvement in solving such an optimization problem with a large body of data and numerous constraints is still a challenge for the experts who are dealing with mathematical modeling and optimization in the area of power system operation. Dantzig–Wolfe decomposition method and Benders decomposition method are two decomposition techniques to tackle such large-scale mixed-integer multistage problems which are broadly used in power system analysis. The former builds a master problem consists of fewer rows compared with the original problem but a large number of columns. The looping process between a master problem and subproblems will be operated until the program reaches an optimal solution. The Benders decomposition method takes also the advantage of modularity, flexibility, and robustness. This method is particularly used for security-constrained economic dispatch, simultaneous generation, and transmission expansion planning, optimal power flow, hydrothermal coordination. However, since the Benders decomposition is a cutting-plane approach, it may demonstrate instabilities that are translated into convergence delay. Since the master problems in the scope of power system operation are often formulated as mixed-integer non-linear problems (MINLP), it entails a high computational burden that significantly affects the convergence time. Therefore, the researchers are attempting to improve the big data performance for such a problem, particularly for stochastic problems. Thus, the progressive hedging method and dual approximation dynamic programming approach are suggested. The first method decomposes the whole problem into several scenarios while defining some of the constraints as a dual problem. The second method includes dynamic subsystems as a parallel dual problem, which increases the complexity but results in a more accurate approximation.

2) Financial transmission rights

In addition, financial transmission rights (FTRs) are a vital element in the structure of any electricity market. The reason is that FTRs enable the participants to hedge against volatile market prices through uncertainty mitigation. Besides, FTRs facilitates competitive open access to the transmission network for all customers.

The inclusion of FTRs into the market clearing mechanism makes the solution more sophisticated, especially when the contingencies corresponded with the line outages are contemplated. Big data analytics could be employed to overcome the complexity of such a problem.

3) Time-constrained economic load dispatch

The short-term generation schedules usually are maintained at different intervals, particularly in weekly and day-ahead periods. The classical techniques employed for this schedule are static methods in the sense that the program solves a snapshot problem rather than a dynamic problem. In order to come up with the prevailing changes in power systems, the idea of time-constrained economic dispatch is suggested for the improvement of the grid's performance, in which the instantaneous condition of the grid topology as well as demand variations and components updates are maintained and sent to the control center to arrange a day-ahead or two-hour-ahead redispatch. This matter facilitates a more secure and more economic real-time operation and avoids unwanted load shedding and curtailments. Such a problem has interdependent time-oriented constraints between consecutive hours, which means that the time horizon of calculation is more than a single hour. Such a problem is regarded as a big data problem in practice which must be solved with big data analytics techniques.

4) Market clearing price forecasting

The forecasting methods utterly depend on the historical records of the purported variable. To forecast solar or wind power generation in a plant, the historical weather records are analyzed by the forecasting tools to predict the short-term forecasts. In a similar way, the load forecasting can be maintained for the scheduling with respect to the historical records of the loads' consumption. That is why the load forecasts usually have relatively low inaccuracy. However, price forecasting is a little different. Price forecast is influenced by demand forecast, historical records of price, as well as some other correlating factors, which makes the forecast more complicated. Hence, at present, the common using methods of price forecasting have still a relatively high level of error. The current methods usually simplify the solution in order to reach a sensible answer for price signals. These problems will be more acute when the proposed methods are employed for real cases with big historical data sets of records for a large-scale system. The market clearance mechanism is a complex paradigm per se, especially when there is congestion in the power flow analysis. This mechanism must also be taken into consideration for better forecasting. The classical methods are inappropriate and unable to deal with such a hard and complex problem. So far, some augmented and combinatorial types of neural networks are suggested to achieve better solutions, but the challenge is still open for researchers for further improvement and innovation.

Besides, in a deregulated and restructured power system, generating companies are private equities that would like to maximize their own benefit. One of the most useful tools for this purpose is price-based unit commitment. This tool helps the company owners to bid in a competitive electricity market. Generation companies employ this tool in addition to the game theory methods while receiving price signals from the operator. The calculation of such a complicated problem needs a powerful method. This matter necessitates the deployment of big data analytics in a restructured environment.

5) Collaborative operation of grid-scale energy storage facilities with RESs

There is ever-growing technological development in large-scale energy storage facilities. These resources play a vital role in the balance of instantaneous energy in the grid. An energy storage unit can be utilized for the purposes of bulk energy time-shifting, frequency regulation in small-scale, frequency stability in large-scale, and power reliability (as reserve capacity) [24]. These features boost the flexibility and dispatchability of the grid's operation. In a restructured environment, wind and solar plants usually would like to have a collaborative operation with a storage facility to offset their imbalance due to uncertainties. This matter helps them to mitigate financial detriments and improve their profitability due to curtailments. The storage plants are also prone to have a joint operation because they can increase their profit. The storage units protect the renewable units against high prices of the spot market and redress the negative and positive imbalances. This matter mitigates the risk of bidding in the day-ahead market for uncertain renewable sources. It is noticeable that a joint operation scheme needs a smart infrastructure so that the instantaneous imbalances must be managed by a collaborative operation controlling scheme. The storage unit must dedicate part of its capacity to the renewable unit's imbalance, and the power exchange between them is out of the day-ahead schedule. The storage unit absorbs the excess generation of RES when there is excess renewable generation. The storage plant has to inject power instead of RES when this unit is unable to fulfill its pledged commitment. Such a storage management scheme needs an elaborated scheme as well as modern communicational infrastructure for better coordination. The computation should be performed by big data analytics for intra-day and intra-hour forecasts in order to manage real-time power exchange [25].

6) Uncertain demand response modeling

Demand response resources (DRRs) are regarded as demand-side virtual power plants. In a broader sense, it is also referred to as demand-side management (DSM). There are some types of demand response programs categorized as time-based rate programs and incentive-based programs. The latter can be a voluntary or mandatory program. Some of these programs provide a definitive source of power (e.g., direct load control (DLC)) for scheduling while some others have uncertain nature. Even though some non-linear models are proposed for the model of such resources, these models suffer from scant accuracy and incompatibility. This matter underlies the employment of big data analytics for better modeling of such volatile and stochastic sources.

In traditional systems, the demand was treated as inelastic loads, which could not be interrupted or deferred. Unlike, the modernized power systems use digital information for better energy saving by changing the electric usage of end-users, particularly in critical conditions. DRRs are also integrated for the balancing services such as frequency regulation, valley filling, and peak load leveling. Together with the rise in the implementation of DR programs at high-resolution level, load serving entities (LSEs) have encountered new challenges and difficulties such as collecting, storing, and processing of such a large volume of data, which generally encompasses power consumption of different appliances and a high quantity of residential end-users, in addition to the data gathered from industrial and commercial consumers. LSEs must take the forecasted (expected) consumption values of the loads and expected generation values of available GENCOs, together with the varying electricity prices, in the schedule decision-making process, and resulting load reductions. This matter further complicates the solution of a data management problem. Hence LSEs are inclined to employ big data management techniques to handle such enormous data sets. Big data analytics applications in demand response can be summarized into three categories of consumption pattern assessment and demand forecasting, electric load classification, and dynamic pricing. In order to improve the cyber-security at various layers of intelligent demand response execution, some methods such as encryption and anonymization are used in order to achieve services such as remote access control and authentication [20].

C. Large scale wind and solar power exploitation

Various types of neural networks and clustering methods have been widely used and developed by researchers and experts to forecast and estimate the generated power of RES. In these methods, the models of wind turbines or solar panels are trained by meteorological data, such as wind speed, humidity, solar irradiance, temperature, etc., while respecting to wind turbine parameters and solar panel characteristics. These methods are used in order to maintain long-term, mid-term, and short-term predictions. The mentioned method highly relies on a massive amount of heterogeneous data, which reflect always a high level of non-smoothness, non-convexity, and non-linearity. The extraction of the correct information and precise results is still a big challenge in this field, and many ongoing research works are conducting to improve the methods that have been using in practice.

D. Stability and security control by PMUs

Synchrophasors or time-synchronized vectors are the measured values by PMUs which report the instantaneous phase angle and magnitude of the sine waves of current or voltage at the point of installation in the grid. These high-speed high-precision PMUs are approximately 100 times faster than traditional SCADA system, and provide a high level of accuracy via GPS for the monitoring system to identify instabilities, power swing conditions, and transient phenomena. Besides, PMUs provide data of the grid for off-line analysis as well as real-time operation. This technology is typically adopted for wide-area monitoring and wide-area protection subject to improve efficiency and reliability and to mitigate operation cost.

One of the vital signs of an electric grid is the frequency. A PMU can report the instantaneous frequency along with the speed of change of frequency according to the IEEE C37.118 communication protocol. The aggregated data of PMUs, topological data of the grid, and the data of configuration of equipment must be integrated into a controlling center in a wide-area monitoring facility in order to be analyzed for occurred of possible blackouts in a large-scale network.

E. Grid operation with large penetration of the hybrid electric vehicle

With respect to the recent advances in battery technology resulting in a considerable decrease in charging time along with remarkable improvement in the efficiency, the pervasiveness of electric vehicles (EV) has had an accelerated ever-growing trend. It is estimated that in the near future, the penetration of millions of electric vehicles in a power system can affect the operation of grids. It should be noted that the model of electric vehicles in a power system is very complex because they are regarded as mobile distributed energy storage sources that have highly stochastic, intermittent, and volatile consumption and generation nature. Hence, the integration of massive amounts of these vehicles will pose a big challenge in future power systems. They can demand power from any point in the network and can also inject power to the grid at any point. This challenge can only be managed by taking advantage of IoT-enabled devices, cloud space accessibility, and big data processing techniques. As a solution, each electric must be equipped with IoT devices. It must be defined for the controlling center that the electric car has vehicle-to-grid (V2G) capability or not. In addition, the instantaneous charging status of EVs' batteries must be reported. The number of EVs in the vicinity of a specific parking lot must be instantaneously reported. With respect to these items, along with some other factors, the highly uncertain behavior of EVs can be estimated. The EVs should be classified and clustered into different categories, and the historical records pertaining to each individual EV can specify the most probable and routine power exchange pattern of an owner. Such a massive generation and storage of data needs a cloud space all over the power system. In addition, enjoying the cloud computing technology, the big data techniques must be employed to process and deal with such a huge volume of produced data. In essence, the use of big data techniques for such a complex problem is inevitable.

F. Active distribution networks

Unlike tradition distribution networks, an active distribution network is comprised of many demand-side generation resources, which are also called virtual power plants (VPPs), and the operator has overall real-time monitoring and control over all feeders and intrinsic microgrids. The utilization of distributed generation, mostly small-scale RES, can boost the flexibility of the operation and enhance operational efficiency and economy in low-voltage grids. Modern distribution networks must be equipped with advanced meters and two-way communicational tools in order to facilitate the execution of demand-side management tasks in order to improve power quality and voltage instabilities and mitigating losses. The use of AMIs enabled with IoT technology will produce a big data set that can be analyzed for enhancement of observability and controllability of the grid. Hence, a higher level of autonomy will facilitate the control of these systems. So far, various methods of artificial intelligence, machine learning, statistical techniques, and pattern recognition methods are suggested to deal with data analytics in active distribution networks [26]. Big data analytics techniques can be applied in active distribution networks for the purposes of grid visualization, high impedance fault detection, loss detection, demand response execution, integration of RES in demand side, equipment diagnosis, asset management, etc. To tackle the abovementioned problems, many methods are used, such as decision tree, regression, clustering, support vector machine, neural networks, Bayesian estimation, classification, and sequence mining. The important point is what kind of data must be collected and how often these data must be produced. These data must be processed by big data techniques to extract targeted benefits. Some scholars have also suggested big data techniques such as greedy algorithms and clustering methods for topology learning and state estimation in radial distribution networks. Some researchers have also employed unsupervised learning techniques and data visualization methods for voltage unbalance analysis in distribution grids.

G. Microgrids and local-scale energy Hubs

Microgrids are proliferating throughout the world day by day. The idea is to satisfy part of demand in demand-side using various decentralized small-scale sources. The priority of generation is by renewable clean sources, but the existence of storage facilities and fossil-fuel burning generators is yet inevitable. The microgrids are usually connected to an upstream distribution grid as a back-up source. So far, many controlling schemes are introduced for the operation of microgrids as well as some schemes for the control and protection of interconnected microgrids. The implementation of such schemes requires dealing with a large body of data for data-driven optimization to maximize the exploitation of internal resources and meeting the security of loads. Moreover, some innovative schemes have suggested multi-carrier and multi-agent energy schemes for the microgrids so that the optimization of electricity, heat, and gas consumption is addressed at the same time. These systems are sometimes referred to as multiform energy systems. In these models, an energy hub is defined for the microgrid, and the controlling scheme has the responsibility to optimize the cost and emission simultaneously by receiving updated data of the system continuously [27]. In these systems, combined cooling heating and power (CCHP) facilities, micro-turbines usually are used for flexible and reliable operation. The execution of such a controlling scheme requires the employment of big data analytics to provide a high level of autonomy and intelligence. On a larger scale, district heat and electricity supply schemes and centralized energy management systems are also suggested [28]. Fig. 4 shows the hierarchical paradigm of a smear management scheme using big data analytics.

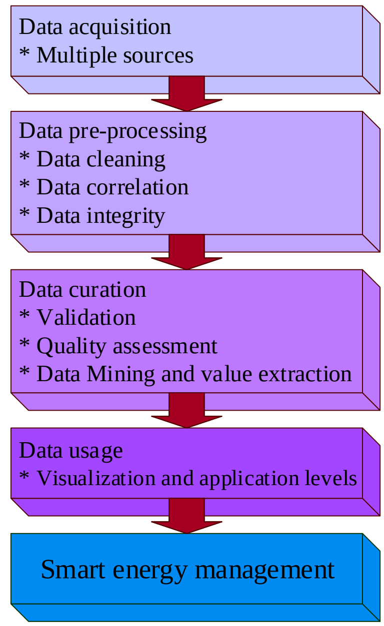

Figure 4. Hierarchical paradigm of smart energy management scheme.

VI. Conclusions

Big data refers to a new breed of structured or unstructured data with high velocity, volume, and variety in a highly complicated structure for being processed, which cannot be analyzed via classic methods. The electricity grids are experiencing a gradual transition toward smart future power systems. The most predominant feature of this era is high-frequency generation and exchange of a large volume of data in real-time that must be analyzed to make proper decisions and taking appropriate actions as quickly as possible. The data-driven environment of future power systems has salient attributes of high penetration of renewable energy sources, proliferated use of IoT-enabled and cloud-based devices, higher integration of demand-side resources, and better management of electrical vehicle fleets, especially with V2G capabilities, higher implementation of individual and interconnected microgrids and nanogrids, decentralization of controlling and monitoring centers capable of high-frequency communication data processing, high integration of cryptocurrencies secured by blockchain techniques, etc. The focus of big data analytics is mainly on the field of artificial intelligence, machine learning, data mining, and advanced statistics. The extensive deployment of big data analytics in various interconnected intelligent mechanisms in future power systems will shift the traditional grids, which are pure physical, to cyber-physical grids with a high level of autonomy, intelligence, adaptability, flexibility, functionality, efficiency, safety, and reliability. The incorporation of big data techniques also procures economic and environmental benefits and alleviates the system vulnerability.

REFERENCES

[1]. Stimmel CL. *Big data analytics strategies for the smart grid*. Auerbach Publications; 2016 Apr 19.

[2]. Zobaa AF, Bihl TJ, editors. *Big Data Analytics in Future Power Systems*. CRC Press; 2018 Aug 14.

[3]. Yu, N., Shah, S., Johnson, R., Sherick, R., Hong, M., & Loparo, K. (2015, February). Big data analytics in power distribution systems. In *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)* (pp. 1-5). IEEE.

[4]. Wang, B., Fang, B., Wang, Y., Liu, H., & Liu, Y. (2016). Power system transient stability assessment based on big data and the core vector machine. *IEEE Transactions on Smart Grid*, *7*(5), 2561-2570.

[5]. Guan, L., Zhang, J., Zhong, L., Li, X., & Xu, Y. (2017, July). Enhancing security and resilience of bulk power systems via multisource big data learning. In *2017 IEEE Power & Energy Society General Meeting* (pp. 1-5). IEEE.

[6]. Song, D., Wen, B., Yang, X., Gu, Y., Wei, S., & Ma, S. (2015, September). A new method for processing and application of wide area measurement big data in a power system. In *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (pp. 137-140). IEEE.

[7]. Liu, Y., Liang, S., He, C., Zhou, Z., Fang, W., Li, Y., & Wang, Y. (2019, December). A Cloud-computing and big data based wide area monitoring of power grids strategy. In *IOP Conference Series: Materials Science and Engineering* (Vol. 677, No. 4, p. 042055). IOP Publishing.

[8]. He, X., Ai, Q., Qiu, R. C., Huang, W., Piao, L., & Liu, H. (2015). A big data architecture design for smart grids based on random matrix theory. *IEEE transactions on smart Grid*, *8*(2), 674-686.

[9]. Ghorbanian, M., Dolatabadi, S. H., & Siano, P. (2019). Big Data Issues in Smart Grids: A Survey. *IEEE Systems Journal*, *13*(4), 4158-4168.

[10]. Tu, C., He, X., Shuai, Z., & Jiang, F. (2017). Big data issues in smart grid–A review. *Renewable and Sustainable Energy Reviews*, *79*, 1099-1107.

[11]. Wang, C., Sun, H., & Yang, Y. B. (2019, September). Research on Demand Response Model of Electric Power Interruptible Load Based on Big Data Analysis. In *International Conference on Advanced Hybrid Information Processing* (pp. 336-345). Springer, Cham.

[12]. Torres, J. F., Troncoso, A., Koprinska, I., Wang, Z., & Martínez-Álvarez, F. (2018, June). Deep learning for big data time series forecasting applied to solar power. In *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications* (pp. 123-133). Springer, Cham.

[13]. Mujeeb, S., Alghamdi, T. A., Ullah, S., Fatima, A., Javaid, N., & Saba, T. (2019). Exploiting Deep Learning for Wind Power Forecasting Based on Big Data Analytics. *Applied Sciences*, *9*(20), 4417.

[14]. Bhagat, M. J. (2019). *Forecasting Price Interval of Wholesale Electricity Market Using a Hybrid Extreme Learning Machine and Fuzzy System* (Doctoral dissertation, The University of Texas at San Antonio).

[15]. Bhattarai, B. P., Paudyal, S., Luo, Y., Mohanpurkar, M., Cheung, K., Tonkoski, R., ... & Manic, M. (2019). Big data analytics in smart grids: State-of-the-art, challenges, opportunities, and future directions. *IET Smart Grid*, *2*(2), 141-154.

[16]. Gutierrez, R. J., Boehmke, B. C., Bauer, K. W., Saie, C. M., & Bihl, T. J. (2017). anomalyDetection: implementation of augmented network log anomaly detection procedures. *The R Journal*, *9*(2), 354-365.

[17]. Akhavan-Hejazi, H., & Mohsenian-Rad, H. (2018). Power systems big data analytics: An assessment of paradigm shift barriers and prospects. *Energy Reports*, *4*, 91-100.

[18]. Qiu RC, Antonik P. *Smart grid using big data analytics: a random matrix theory approach*. John Wiley & Sons; 2017 Apr 17.

[19]. Erdinc O, Tascıkaraoglu A, editors. *Pathways to a Smarter Power System.* Academic Press; 2019 May 1.

[20]. Arghandeh R, Zhou Y, editors. Big *Data Application in Power Systems*. Elsevier; 2017 Nov 27.

[21]. Moradi, J., Shahinzadeh, H., Nafisi, H., Gharehpetian, G. B., & Shaneh, M. (2019, June). Blockchain, a Sustainable Solution for Cybersecurity Using Cryptocurrency for Financial Transactions in Smart Grids. In *2019 24th Electrical Power Distribution Conference (EPDC)* (pp. 47-53). IEEE.

[22]. Shahinzadeh, H., Moradi, J., Gharehpetian, G. B., Nafisi, H., & Abedi, M. (2019, January). IoT Architecture for Smart Grids. In *2019 International Conference on Protection and Automation of Power System (IPAPS)* (pp. 22-30). IEEE.

[23]. Shahinzadeh, H., Moradi, J., Gharehpetian, G. B., Nafisi, H., & Abedi, M. Internet of Energy (IoE) in Smart Power Systems. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)* (pp. 627-636). IEEE.

[24]. Arbabzadeh, M., Sioshansi, R., Johnson, J. X., & Keoleian, G. A. (2019). The role of energy storage in deep decarbonization of electricity production. *Nature communications*, *10*(1), 1-11.

[25]. Moradi, J., Shahinzadeh, H., & Khandan, A. (2017). A Cooperative Dispatch Model for the Coordination of the Wind and Pumped-storage Generating Companies in the Day-ahead Electricity Market. *International Journal of Renewable Energy Research (IJRER)*, *7*(4), 2057-2067.

[26]. Hosseini, S. J. A. D., Moradian, M., Shahinzadeh, H., & Ahmadi, S. (2018). Optimal Placement of Distributed Generators with Regard to Reliability Assessment using Virus Colony Search Algorithm. *International Journal of Renewable Energy Research (IJRER)*, *8*(2), 714-723.

[27]. Shahinzadeh, H., Moradi, J., Gharehpetian, G. B., Fathi, S. H., & Abedi, M. (2018, November). Optimal Energy Scheduling for a Microgrid Encompassing DRRs and Energy Hub Paradigm Subject to Alleviate Emission and Operational Costs. In *2018 Smart Grid Conference (SGC)* (pp. 1-10). IEEE.

[28]. Shahinzadeh, H., Moradi, J., Gharehpetian, G. B., Abedi, M., & Hosseinian, S. H. (2018, November). Multi-Objective Scheduling of CHP-Based Microgrids with Cooperation of Thermal and Electrical Storage Units in Restructured Environment. In *2018 Smart Grid Conference (SGC)* (pp. 1-10). IEEE.