

Big Data Analytics for Dynamic Energy Management in Smart Grids

Panagiotis D. Diamantoulakis^{b,*}, Vasileios M. Kapinas^b, George K. Karagiannidis^{a,b}

^aDepartment of Electrical and Computer Engineering, Khalifa University, PO Box 127788, Abu Dhabi, United Arab Emirates

^bDepartment of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece

Abstract

The smart electricity grid enables a two-way flow of power and data between suppliers and consumers in order to facilitate the power flow optimization in terms of economic efficiency, reliability and sustainability. This infrastructure permits the consumers and the micro-energy producers to take a more active role in the electricity market and the dynamic energy management (DEM). The most important challenge in a smart grid (SG) is how to take advantage of the users' participation in order to reduce the cost of power. However, effective DEM depends critically on load and renewable production forecasting. This calls for intelligent methods and solutions for the real-time exploitation of the large volumes of data generated by a vast amount of smart meters. Hence, robust data analytics, high performance computing, efficient data network management, and cloud computing techniques are critical towards the optimized operation of SGs. This research aims to highlight the big data issues and challenges faced by the DEM employed in SG networks. It also provides a brief description of the most commonly used data processing methods in the literature, and proposes a promising direction for future research in the field.

Keywords: Big data, Smart grids, Dynamic energy management, Predictive analytics, Artificial intelligence, high performance computing.

1. Introduction

A smart grid (SG) is the next-generation power system able to manage electricity demand in a sustainable, reliable and economic manner, by employing advanced digital information and communication technologies. This new platform aims to achieve steady availability of power, energy sustainability, environmental protection, prevention of large-scale failures, as well as optimized operational expenses (OPEX) of power production and distribution, and reduced future capital expenses (CAPEX) for thermal generators and transmission networks [1]. The upcoming technology in the framework of SG facilitates the development and efficient interactive utilization of millions of alternative distributed energy resources (DER) and electric vehicles [1–3]. To this end, each consumer location has to be equipped with a smart meter for monitoring and measuring the bi-directional flow of power and data, while supervisory control and data acquisition (SCADA) systems are needed to control the grid operation.

While dynamic energy management (DEM) in conventional electricity grids is a well-investigated topic, this is not the case for SGs. This is due to its much more complicated nature, since complex decision-making processes

are required by the control centers [4, 5]. Energy management systems (EMSs) in SGs include i) real-time wide-area situational awareness (WASA) of grid status through advanced metering and monitoring systems, ii) consumers' participation through home EMSs (HEMS), demand response (DR) algorithms, and vehicle-to-grid (V2G) technology, and iii) supervisory control through computer-based systems [6]. A typical overview of the SG and the included systems and technologies is given in Fig. 1. The quality and reliability of the data collected is a key factor for the optimized operation of the SG, thus rendering data mining and predictive analytics tools essential for the effective management and utilization of the available sensor data [7]. This is because effective DEM relies dramatically on short-term power supply and consumption forecasting, which handles prediction horizons from one hour up to one week [8]. Additionally, the sensor data contains important correlations, trends, and patterns that need to be exploited for the optimization of the energy consumption and the DR, among others [4]. Most of the research related to data mining in SGs deal with predictive analytics and load classification (LC), which are necessary for the load forecasting, bad data correction, determination of the optimal energy resources scheduling, and setting of the power prices [9, 10]. The efficient processing of the produced vast amount of data requires increased data storage and computing resources, which imply the need for high performance computing (HPC) techniques.

This work differs from other related surveys in the lit-

*Corresponding author

Email addresses: padiaman@auth.gr (Panagiotis D. Diamantoulakis), kapinas@auth.gr (Vasileios M. Kapinas), geokarag@auth.gr (George K. Karagiannidis)

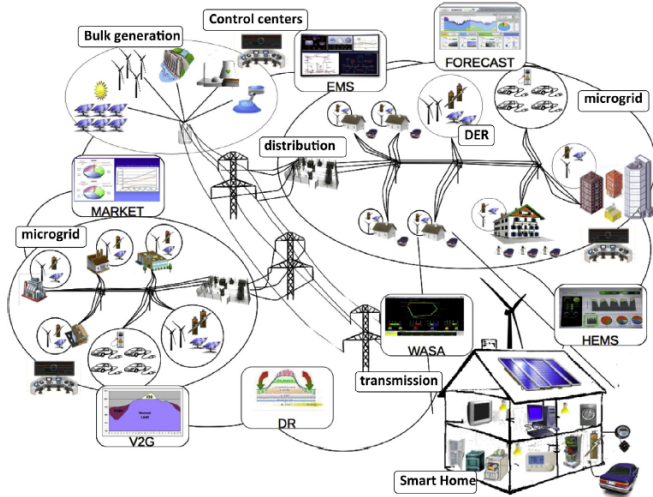


Figure 1: Smart Grid overview [6].

erature, such as [11–16], in being the first meta-analytic review on efficient SG data processing with focus on DEM. Besides, it gives useful insights into technologies and methods from the area of big data analytics (BDA) that have to be further explored into the framework of DEM, demand forecasting, and dynamic pricing. Our investigations reveal that there is free space for research in the following topics:

- Design and development of algorithms that can accurately extract the load patterns from large-scale datasets.
- Design of machine learning (ML)-based algorithms with improved forecasting performance, low memory requirements, and scalable architecture.
- Development of novel data-aware resource management systems that can provide powerful data processing in distributed computing systems and clusters for real-time processing.

It is also highlighted that scalability and flexibility, achieved through the construction of robust algorithms and fast provisioning of HPC resources, can enable the efficient processing of the large data volumes involved in DEM and short-term power demand/supply forecasting.

The rest of the paper is organized as follows. Section 2 gives insight on the reasons why conventional data processing techniques are not appropriate for DEM in SGs. Section 3 focuses on smart meter data stream mining and presents the most commonly used methods. Section 4 is dedicated to the appropriate HPC techniques. Section 5 provides promising future research directions, while Section 6 concludes the paper.

2. Dynamic Energy Management in SGs: A Big Data Issue

DEM requires power flow optimization, system monitoring, real-time operation, and production planning [17]. In more detail, DEM in a SG is a complicated, multi-variable procedure, since the latter enables an interconnected power distribution network by allowing a two-way flow of both power and data. This is in contrast to the traditional power grid, in which the electricity is generated at a central source and then distributed to consumers. Thanks to the bi-directional flow of information and power between suppliers and consumers, the grids become more adaptive to the increased penetration of DER, encouraging also users' participation in energy savings and cooperation through the DR mechanism [10, 18, 19].

DR can be applied to both residential (e.g., cooling, heating, electric vehicles (EVs) charging, etc.) and industrial loads and includes three different concepts; i) energy consumption reduction, ii) energy consumption (or production) shifting to periods of low (or high) demand, and iii) efficient utilization of storage systems [20]. It should be noticed here that plug-in EVs can be considered as storage devices, while the careful scheduling of their charging and discharging can benefit both their owners and the utilities. Obviously, this further increases the parameters that the DEM algorithms have to take into account, such as the EVs charging profiles. Consequently, the associated complexity is also increased, creating at the same time storage capacity prediction problems [21]. Thus, a crucial issue in SGs is how to manage DR in order to reduce peak electricity load, utilizing at the same time renewable energies and storage systems more efficiently. Finally, effectiveness of DR algorithms depends critically on demand, price, load, and renewable energy forecasting, which highlights the need for sophisticated signal processing techniques [22].

The electricity demand and renewable production in the SG environment is affected by several factors, including weather conditions, micro-climatic variations, time of day, random disturbances, electricity prices, DR, renewable energy sources, storage cells, micro-grids, and the development of EVs [23–26]. High forecasting accuracy accommodates the generation and transmission planning, i.e., deciding which power plants to operate and how much power should be generated by them at a specific time-period, with the aim to reduce the operating cost and increase the reliability [27]. It also enables the utilities to successively estimate the electricity cost and correctly set the electricity prices, capturing the interdependency between the energy demand and the prices [28]. A typical example of this interdependency is load-synchronization, where a large portion of load is shifted from hours of high prices to hours of low prices, without significantly reducing the peak-to-average ratio [23]. Moreover, insufficient monitoring and control of the power flow can increase the possibility of failure (e.g., due to load synchronization, overloading, con-

gestion, etc.). The power grid, which is consisted of multiple components such as relays, switches, transformers, and substations, must be carefully monitored. Therefore, the SG requires intelligent real-time monitoring techniques in order to be capable of detecting abnormal events, finding their location and causes, and most importantly predicting and eliminating faults before they happen. This self-healing behaviour, renders the power grid a real “immune system”, which is one of the most important characteristics of a SG framework, targeting uninterrupted power supply [29, 30].

In order to deal with the high level of uncertainties in DEM, the extreme size of data, and the need for real-time learning/decision making, the SG demands advanced data analytic techniques, big data management, and powerful monitoring techniques [31, 32]. Since BDA is one of the major driving forces behind a SG, various techniques such as artificial intelligence, distributed and HPC, simulation and modeling, data network management, database management, data warehousing, and data analytics are to be used to guarantee smooth running of SGs. The main challenges of Big Data approaches in SGs is the selection, deployment, monitoring, and analysis of aggregated data in real-time [33]. The BD’s role in SG collective awareness, the self-organization capability of SGs, and the service interruptions limitation are thoroughly discussed in [29]. Specifically, it is explained that the reliability of the electricity grid could be enhanced if the users were aware of the effects of their personal energy use on the total consumption and overloading.

Considering the above, BDA can provide efficient solutions in specific problems related to data processing in SGs as described in the sequel and briefly summarized in Fig. 2, which illustrates the roadmap to the DEM in SGs.

3. Data Mining and Predictive Analytics in SGs

Data mining is the standard process to harvest useful information from a stream of data, such as users’ consumption, injection of renewable power, and EVs’ state of battery, and transform it into an understandable structure for further use. The data mining process involves the utilization of algorithms for discovering patterns among the data following a similarity criterion [42]. Efficient data mining is crucial towards the optimized operation of the SG, since it strongly affects the decision-making of power producers and consumers, and the reliability of the grid.

3.1. Dimensionality Reduction

Smart meters generate large volume of data, and thus acquiring and processing all of them is inefficient -if not prohibitive- in terms of communication cost, computing complexity, and data storage resources utilization. For this purpose, dimensionality reduction has been applied in [34], in order to provide a reduced version (sketch) of meters’ original data via random projection (RP). It is

shown that processing the produced summarized version of data instead of the original stream of data leads to an acceptable relative error. The main advantage of RP is scalability, complexity reduction, and execution speed increase.

Dimensionality reduction has only been sufficiently explored in the area of synchrophasor data. Particularly, online dimensionality reduction has been proposed in [43], in order to extract correlations between synchrophasor measurements, such as voltage, current, frequency etc. The proposed method can be used a preprocessing method in data analysis and storage, when a only an approximation of the initial data is required. Online dimensionality reduction has also been successfully used in for early event detection in [44], where an early event detection algorithm is proposed.

3.2. Load Classification

In classification problems a set of pre-classified data points are given and the classification algorithm tries to discover a rule, which describes as closely as possible the observed classification [10]. LC is based on clustering process, which is used to discover groups and identify distributions in the provided data.

For the successive LC in SGs, the most widely used models are Artificial Neural Networks (ANNs). ANNs are computational models consisting of a large number of simple interconnected processors, which can be used to estimate approximate functions that depend on a large number of inputs when there is not an accurate mathematical model to describe the phenomenon [45]. This can be achieved by weighting and transforming the input values by a suitable function with the aid of sequential sets of neurons (until an output neuron is activated). In [46], ANNs have been used for the successful classification of consumer load curves in order to create patterns of consumption and facilitate the selection of the appropriate DSM technique. Self-organizing mapping, also known as Kohonen neural network, which is an unsupervised neural networks method, has also been widely used for LC [9, 47].

Other commonly used algorithms are K-means, which is based on the Euclidean distance between objects, Fuzzy c-means, which is a local search fuzzy clustering method, and hierarchical clustering method, which is a model that can be viewed as a dendrogram [9]. Due to the ever-growing nature of the SG, a scalable approach is needed for the effective data harvesting and utilization. For this purpose, an effective online clustering has been proposed in [48], based on unsupervised learning techniques, improving for this case the eXtended Classifier System for clustering (XCSc). This XCSc-based method fits well the dynamic nature of SGs, while it outperforms the offline strategies in terms of the storage system performance [49].

3.3. Short-term Forecasting

Short-term load forecasting (STLF) has been widely used over the last 30 years and several models have been

Concern	Solution
DR is huge challenge in practise since each user has a different reaction to the real-time prices and stochastic parameters such as weather, etc. [10]. Also, many inserted factors are interdependent [28].	Before optimizing the power flow and setting the prices, LC can be used in order to categorize various load patterns into a specific number of groups and each group can be characterized by each characteristic load pattern [10]. Predictive analytics and ML techniques can facilitate the real-time decision making [32].
The available communication resources, e.g., bandwidth are not enough for the data acquisition in a centralized manner [34].	Distributed data mining and dimensionality reduction reduce the required resources [4, 34, 35].
Traditional computer techniques cannot handle fast data processing, which is required for real-time monitoring, dynamic energy management and power flow optimization [36].	Efficient high performance computing techniques, such as virtualization and in-memory computing, can importantly reduce the computing time [36, 37].
Increased data storage and computing resources are required in order to meet the needs of processing the huge amount of data involved in DEM, which could possibly lead to increased OPEX and CAPEX [37].	Cloud computing: pay-per-use [38].
Security [39, 40]	Data anonymization via data aggregation, encryption, etc. [41].

Figure 2: A roadmap of big data analytics in DEM

developed, which can be summarized as follows: i) regression models, ii) linear time-series-based, iii) state-space models, and iv) nonlinear time-series modeling [50]. Moreover, very little progress has been made in the field of the very-short-term (VST) and ultra-short-term (UST) load forecasting in SGs, which, among others, are necessary for the successful self-healing of SG, since they are appropriate for a few minutes load forecasting [50–52]. Among the appropriate methods for VST and UST, only basic ANNs have been widely tested. Generally, most of the research has been focused on large aggregated load data, where most individual variations are averaged out by the effect of the law of large numbers. These methods were seldom tried on individual meters or at meter aggregate levels, such as distribution feeders and substation [8]. Additionally, as shown in [8], their performance degrades when the number of the considered meters goes down. For this purpose, a short-term load forecasting based on Empirical Mode Decomposition, Extended Kalman Filter and Extreme Learning with Kernel is proposed in [53], which is more appropriate for load forecasting in micro-grids. Kernel methods has attracted the research interest in the area of STLF [54, 55], since it increases the computing energy-efficiency [56–59]. Most works in the existing literature ignore the interdependency between the demand and the electricity prices, while assuming that pricing setting follows the load forecasting. To this end, the authors in [28]

propose a multi-input multi-output forecasting engine for joint price and demand prediction using data association mining algorithms. For the effective price forecasting, the users’ comfort has also to be taken into account, which can be achieved via supervised machine algorithms [60]. For very short term wind power generation forecast, there are myriad of approaches, which can be divided into two big categories: i) forecasting the wind and direction for a specific windmill farm, and ii) forecasting the generated power in a single step [61]. More interesting details on this issue can be found in [61]. Finally, short-term photovoltaic power prediction is mainly based on the past power output [62].

3.4. Distributed Data Mining

The traditional centralized frameworks for acquiring, analyzing and processing data, require huge exchange of information among the remote sensors (e.g., the smart meters and the centralized processor), which is inefficient in terms of telecommunication resources management and economic cost. To this end, the authors in [4] present several distributed data analysis techniques that can be successively used for energy demand prediction. The provided analysis emphasizes on the problem of multivariate regression and rank ordering in a distributed scenario, based on polynomially bounded computations per node. Decentralized data mining algorithms have the advantage of scala-

bility, while they are less affected by peer failures and they need little computing and communication resources [35].

4. High Performance Computing

Real-time monitoring, DEM, and power flow optimization are all based on fast data processing and BDA, which need high computing power. Efficient data mining algorithms based on task parallelism, using multi-core, cluster, and grid computing, can reduce the computational time [36]. However, covering the increased data storage and computing resources needs is still a big economic challenge, mainly for the operators of the electricity grids. Therefore, distributed computing seems to be a promising perspective [63].

4.1. Dedicated Computational Grid

In order to enhance the existing computational capabilities and increase efficiency, a dedicated grid computing based framework is proposed in [37]. In more detail, an architecture of three layers is proposed, namely i) the resource layer, which consists of the hardware part of the computing grid, ii) the grid middleware, which provides access of grid resources to the grid services, and iii) the application layer, which consists of the services. It is shown that this computational grid can provide HPC by combining the processing power, memory and storage of the available computers.

4.2. Cloud Data

The cloud computing (CC) model meets the requirements of data and computing intensive SG applications [38, 64]. The main advantages of CC over traditional models are energy saving, cost saving, agility, scalability, and flexibility, since computational resources are used on demand [65]. Many approaches have been developed so far to further increase the energy efficiency of HPC data centers, such as energy conscious scheduling in [66], the cooperation with the SG in [67] and thermal-aware task scheduling in [53]. In [38], a model for SG data management is presented, taking advantage of the main characteristics of CC computing, such as distributed data management, parallelization, fast retrieval of information, accessibility, interoperability and extensibility. Most of smart grid applications, such as advanced metering infrastructure, SCADA, and energy management, can be facilitated by the available cloud service models, namely software as a service, platform as a service, and infrastructure as a service [68].

4.2.1. Security

Confidentiality and privacy are big challenges towards the application of CC on SG data processing [39, 40]. To this end, the designed data architectures must be multi-tenant, following one of the three different approaches for

such architectures, namely the separate databases, separate schemas, or shared schemas [38]. Also, privacy of end users and data anonymization can be guaranteed by data aggregation, which is used in most SG architectures. However, from the electrical companies' perspective, security is still a challenging problem, since the hackers of systems located in the cloud cannot be easily traced. Authentication, encryption, trust management, and intrusion detection are important security mechanisms that can prevent, detect and mitigate such network attacks [6, 41]. Finally, a related major issue is the recovery of data in the case of a possible failure of the cloud service [68].

5. Future Research Directions & Discussion

The load data in SG environment is massive, dynamic, high-dimensional, and heterogeneous [9]. Thus, in order to build an accurate real-time monitoring and forecasting system, two novel concepts have to be taken into account in the system design. First, as shown in Fig. 3, all available information from different sources, such as individual smart meters, energy consumption schedulers, aggregators, solar radiation sensors, wind-speed meters and relays has to be integrated, while a communication point has to be designed where multiple artificial experts can interact and make decisions on data. The appropriate forecasting system should rely on effective data sampling [69], improved categorization of the information and successful recognition of the different patterns. Second, suitable adaptive algorithms and profiles for effective dynamic, autonomous, distributed, self-organized and fast multi-node decision-making have to be designed. It has been shown that the performance of multi-node load forecasting is clearly better than that of single-node forecasting [70]. The designed algorithms should be based on realistic consensus functions or scoring/voting models [71], where the large computations can be parallelized. The algorithmic results are the state estimation, the estimated production and consumption, and the STLF in SGs. For the most efficient pattern-recognition and state estimation in the SGs environment, the following methodologies and technologies can be used.

5.1. SG Feature Selection and Extraction

The factors that affect the load forecasting can be separated in two categories: a) the traditional factors and b) the SG factors [27]. The traditional factors include the weather conditions, time of the day, season of the year, and random events and disturbances. On the other hand, the smart grid factors include the electricity prices, demand response, distributed energy sources, storage cells and electric vehicles. As shown in Fig. 3, large volumes of data from the sensors installed around SG are collected, and the features extracted in this phase also need to be refined as there is noise and redundancy in the features. If the input features contain redundant information (e.g., highly

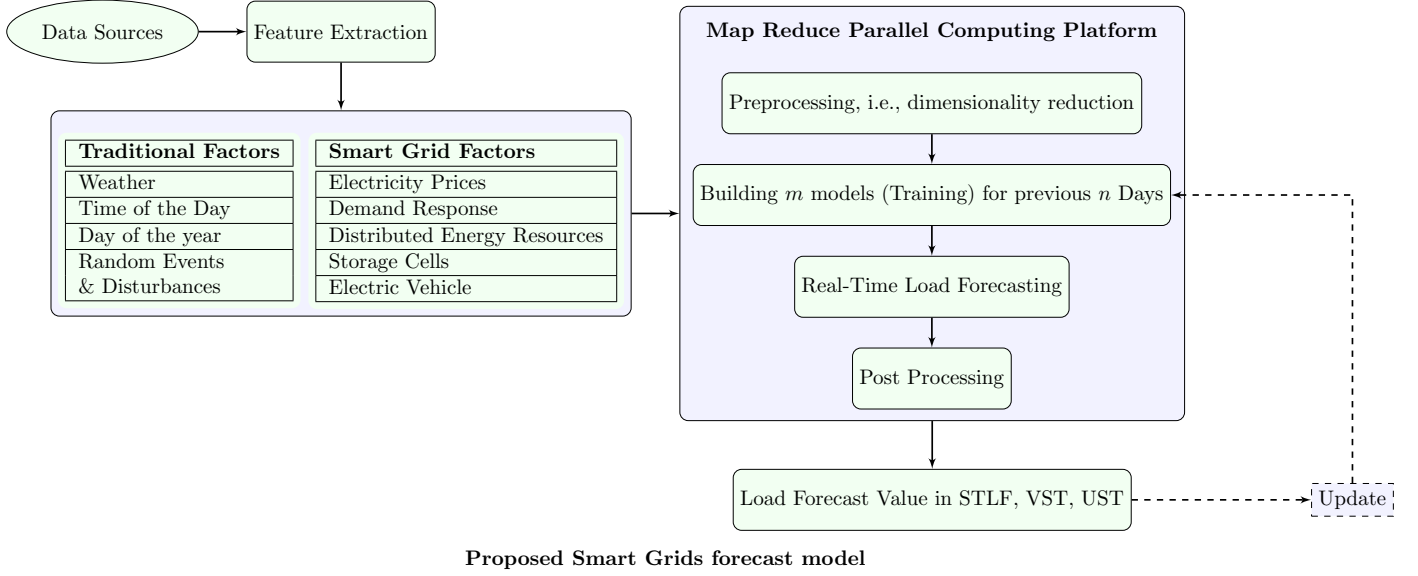


Figure 3: Proposed SG forecast model.

correlated features), ML algorithms in general perform poorly because of numerical instabilities. Some regularization techniques can be imposed to solve such problems. The techniques that can be used to build an optimal subset for the load predicting problem in SG are greedy hill climbing [72], minimum-redundancy-maximum-relevance [73], regularized trees [74], random multinomial logit [75], etc. In addition to the image preprocessing methods described above, there are many other techniques that draw on disciplines such as artificial intelligence (AI) and ML that can be used to analyze such selected features.

5.2. Online Learning

Online learning is a powerful way of dealing with load monitoring and prediction in SGs. An online learning algorithm observes a stream of examples and makes a prediction for each element in the stream [76]. The algorithm receives immediate feedback about each prediction and uses this feedback to improve its accuracy on subsequent predictions. In contrast to statistical ML, online learning algorithms don't make stochastic assumptions about the observed data, and even handle situations where the data is generated by a malicious adversary. There has been a recent surge of scientific research on online learning algorithms, largely due to their broad applicability to web-scale forecasting problems. In load forecasting in SGs, the statistical ML properties of the target variable, which the model is trying to forecast, change over time in unforeseen ways. This causes the so-called "concept drift" problem because the predictions become less accurate as time passes. In [77], the employed online learning set up mitigates such problem. In online learning, soon after the forecasting is made, the true label of the instance is discovered. This information can then be used to refine the forecasting hypothesis used by the algorithm. The goal of

the algorithm is to make forecasting that is close to the true labels.

5.3. Randomized Model Averaging

ML concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data. A major focus of ML research is to automatically learn to recognize complex patterns such as the features of SGs, and make intelligent decisions based on data. Using multiple predictive ML models (each developed using statistics and/or ML) to obtain better predictive performance than could be obtained from any of the constituent models [78, 79]. In addition, the proposed scheme should be utilizing the model averaging technique [80] in order to improve the stability and accuracy of ML algorithms via reducing variance [56, 81, 82].

5.4. MapReduce Parallel Processing

Load forecasting problems, involving intelligent ML solutions pertaining to large volume of data source generation, are a perfect fit for MapReduce deployments [76]. MapReduce is a programming model for processing large datasets with a parallel, distributed algorithm on a computing cluster of low cost commodity computers. A MapReduce application typically consists of two phases (or operations); "map" and "reduce" with many tasks in each phase [83]. A map/reduce task deals with a chunk of data independently and thus, tasks in a given phase can be easily parallelized and effectively processed in a large-scale computing environment (i.e., a cloud platform). However, the existing resource management models pay little attention to the data and/or utilize a simple technique, called MapReduces rack-aware task placement. Such parallelization model could be reconstructed by explicitly taking the application characteristics and network topology into

account [84, 85]. For instance, in a tree network topology, two sub-trees (racks) adjacent to each other will be a better combination than two of dispersed sub-trees. The presented framework can enable the minimization of data movement and in turn reduce the occurrence of network contention, and this will eventually enhance the cloud systems efficiency. Two main mechanisms that have to be taken into account explicitly are a) the data locality-aware scheduling algorithm, and b) the application-specific resource allocation mechanisms. Specifically, tasks requiring common datasets are dispatched to computers (compute nodes) with close proximity to those data sets. For most data processing applications, storage capacity, such as disk and memory, is more important than computing power (i.e., CPUs).

5.5. Available Testbeds and Platforms

The majority of the available power grid systems focus on modeling of traditional network components, i.e., the generation systems, loads, and transmission network. A different approach is followed in [86], where a distribution grid testbed has been proposed, which can be used to test the designs of integrated information management systems. The purpose of this testbed is to successfully represent the correlation and interdependency among data sets, aiming to efficiently monitor the status of the SG and detect abnormalities. Interestingly enough, extensive sets of SG's detailed trial data, which can be used in order to test the designed schemes, can be easily acquired, thus facilitating the research in this area [87–91]. In [87], registered users can also access the public model, including key functions, assumptions, and analytical tools.

Storing and processing the huge amount of data generated by the smart meters, requires improved platforms, appropriate for big data analytics, such as Hadoop, Cassandra, and Hive [92]. Hadoop is a promising platform for the distributed processing of large SG's data sets. It is a collection of open source tools and includes the concept of MapReduce. Cassandra database, which supports the cloud infrastructure, can be used in order to store the large data sets that are needed for the effective DEM. Moreover, Hive data warehouse software, which uses a simple SQL-like language, can be used to query datasets that are stored in a distributed environment.

6. Conclusions

In this paper, we have summarized the state-of-the-art in the exploitation of big data tools for dynamic energy management in smart grid platforms. We have first highlighted that, in order to deal with the extreme size of data, the smart grid requires the adoption of advanced data analytics, big data management, and powerful monitoring techniques. Next, we elaborated on the utilization of the most commonly used smart grid data mining and predictive analytics methods, focusing on the smart meter data that are necessary for the accurate and efficient

power consumption/supply forecasting. We proceeded with a brief survey on the works dealing with high performance computing, insisting on cost efficiency and security issues in the context of SG control. Finally, we discussed several interesting techniques and methods that have to be further explored into the framework of a real-time monitoring and forecasting system, and we provided promising research directions for future research in the field.

Acknowledgement

This work was supported by the project of “Scalable Real-Time Load Forecasting in Smart Grid” under a grant from the Khalifa University Internal Research Fund (KUIRF-Level 2; Fund # 210063).

References

References

- [1] S. F. Bush, S. Goel, G. Simard, IEEE vision for smart grid communications: 2030 and beyond roadmap, IEEE Std. Association (2013) 1–19.
- [2] P. Goncalves Da Silva, D. Ilic, S. Karnouskos, The impact of smart grid prosumer grouping on forecasting accuracy and its benefits for local electricity market trading, IEEE Trans. Smart Grid 5 (1) (2014) 402–410.
- [3] M. A. Lopez, S. de la Torre, S. Martin, J. A. Aguado, Demand-side management in smart grid operation considering electric vehicles load shifting and vehicle-to-grid support, Int. J. Electr. Power Energy Syst. 64 (0) (2015) 689–698.
- [4] R. Mallik, N. Sarda, H. Kargupta, S. Bandyopadhyay, Distributed data mining for sustainable smart grids, in: Proc. of ACM SustKDD'11, 2011, pp. 1–6.
- [5] N. Balac, “Green machine” intelligence: Greening and sustaining smart grids, IEEE Intell. Syst. 28 (5) (2013) 50–55.
- [6] E. Ancillotti, R. Bruno, M. Conti, The role of communication systems in smart grids: Architectures, technical solutions and research challenges, Comput. Commun. 36 (17–18) (2013) 1665–1697.
- [7] Z. Fan, Q. Chen, G. Kalogridis, S. Tan, D. Kaleshi, The power of data: Data analytics for M2M and smart grid, in: Proc. 3rd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe), 2012, pp. 1–8.
- [8] P. Mirowski, S. Chen, T. K. Ho, C.-N. Yu, Demand forecasting in smart grids, Bell Labs Tech. J. 18 (4) (2014) 135–158.
- [9] K. le Zhou, S. lin Yang, C. Shen, A review of electric load classification in smart grid environment, Renew. Sustai. Energy Rev. 24 (0) (2013) 103–110.
- [10] Z. Vale, H. Morais, S. Ramos, J. Soares, P. Faria, Using data mining techniques to support DR programs definition in smart grids, in: Proc. IEEE Power and Energy Society General Meeting, 2011, pp. 1–8.
- [11] A. Ukil, R. Zivanovic, Automated analysis of power systems disturbance records: Smart grid big data perspective, in: Proc. IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2014, 2014, pp. 126–131.
- [12] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, V. Prasanna, Cloud-based software platform for big data analytics in smart grids, Comp. Sci. Eng. 15 (4) (2013) 38–47.
- [13] D. J. Leeds, The soft grid 2013-2020: Big data & utility analytics for smart grid, <http://www.greentechmedia.com/research/report/the-soft-grid-2013>, Accessed: 2014.12.06 (2012).
- [14] C. L. Stimmel, Big Data Analytics Strategies for the Smart Grid, CRC Press, 2014.

- [15] T. Nguyen, V. Nunavath, A. Prinz, Big Data Metadata Management in Smart Grids, in: *Big Data and Internet of Things: A Roadmap for Smart Environments*, Vol. 546 of *Studies in Computational Intelligence*, Springer International Publishing, 2014, pp. 189–214.
- [16] I. S. Group, Managing big data for smart grids and smart meters, IBM Corporation, whitepaper (May 2012).
- [17] M. Manfren, Multi-commodity network flow models for dynamic energy management-mathematical formulation, *Energy Procedia* 14 (0) (2012) 1380–1385.
- [18] P. Samadi, H. Mohsenian-Rad, V. W. S. Wong, R. Schober, Real-time pricing for demand response based on stochastic approximation, *IEEE Trans. Smart Grid* 5 (2) (2014) 789–798.
- [19] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, A. Leon-Garcia, Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid, *IEEE Trans. Smart Grid* 1 (3) (2010) 320–331.
- [20] P. Siano, Demand response and smart grids A survey, *Renew. Sustain. Energy Rev.* 30 (0) (2014) 461–478.
- [21] W. Xiaojun, T. Wenqi, H. Jinghan, H. Mei, J. Jiuchun, H. Haiying, The application of electric vehicles as mobile distributed energy storage units in smart grid, in: *Proc. Power and Energy Engineering Conference (APPEEC'11)*, 2011 Asia-Pacific, 2011, pp. 1–5.
- [22] S. C. Chan, K. M. Tsui, H. C. Wu, Y. Hou, Y.-C. Wu, F. F. Wu, Load/price forecasting and managing demand response for smart grids: Methodologies and challenges, *IEEE Signal Process. Mag.* 29 (5) (2012) 68–85.
- [23] A.-H. Mohsenian-Rad, A. Leon-Garcia, Optimal residential load control with price prediction in real-time electricity pricing environments, *IEEE Trans. Smart Grid* 1 (2) (2010) 120–133.
- [24] G. Carpinelli, G. Celli, S. Mocci, F. Mottola, F. Pilo, D. Proto, Optimal integration of distributed energy storage devices in smart grids, *IEEE Trans. Smart Grid* 4 (2) (2013) 985–995.
- [25] A. Y. Saber, G. K. Venayagamoorthy, Resource scheduling under uncertainty in a smart grid with renewables and plug-in vehicles, *IEEE Syst. J.* 6 (1) (2012) 103–109.
- [26] F. Avila, D. Saez, G. Jimenez-Estevéz, L. Reyes, A. Nunez, Fuzzy demand forecasting in a predictive control strategy for a renewable-energy based microgrid, in: *Proc. European Control Conference (ECC)*, 2013, pp. 2020–2025.
- [27] S. Balantrapu, Load forecasting in smart grid, <http://www.energycentral.com/enduse/demandresponse/articles/2760>, Accessed: 2014.12.06.
- [28] A. Motamedi, H. Zareipour, W. D. Rosehart, Electricity price and demand forecasting in smart grids, *IEEE Trans. Smart Grid* 3 (2) (2012) 664–674.
- [29] J. Pitt, A. Bourazeri, A. Nowak, M. Roszczynska-Kurasinska, A. Rychwalska, I. Santiago, M. Sanchez, M. Florea, M. Sanduleac, Transforming big data into collective awareness, *Computer* 46 (6) (2013) 40–45.
- [30] J. DongLi, M. Xiaoli, S. Xiaohui, Study on technology system of self-healing control in smart distribution grid, in: *Proc. International Conference on Advanced Power System Automation and Protection (APAP)*, Vol. 1, 2011, pp. 26–30.
- [31] Z. Aung, M. Toukhy, J. Williams, A. Sanchez, S. Herrero, Towards accurate electricity load forecasting in smart grids, in: *Proc. 4th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA)*, 2012, pp. 51–57.
- [32] P. Mack, Chapter 35 - big data, data mining, and predictive analytics and high performance computing, in: L. E. Jones (Ed.), *Renewable Energy Integration*, Academic Press, Boston, 2014, pp. 439 – 454.
- [33] J. Baek, Q. Vu, J. Liu, X. Huang, Y. Xiang, A secure cloud computing based framework for big data information management of smart grid, *IEEE Trans. Cloud Comput.* PP (99) (2014) 1–1.
- [34] A. Dieb Martins, E. Gurjao, Processing of smart meters data based on random projections, in: *Proc. IEEE PES Conference On Innovative Smart Grid Technologies Latin America (ISGT LA)*, 2013, 2013, pp. 1–4.
- [35] K. Bhaduri, H. Kargupta, An efficient local algorithm for distributed multivariate regression in peer-to-peer networks, in: *SIAM Conference on Data Mining (SDM)*, 2008, pp. 153–164.
- [36] R. C. Green, L. Wang, M. Alam, Applications and trends of high performance computing for electric power systems: Focusing on smart grid, *IEEE Trans. Smart Grid*, 4 (2) (2013) 922–931.
- [37] M. Ali, Z. Y. Dong, X. Li, P. Zhang, RSA-grid: A grid computing based framework for power system reliability and security analysis, in: *IEEE Power Engineering Society General Meeting*, 2006, pp. 1–7.
- [38] S. Rusitschka, K. Eger, C. Gerdes, Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain, in: *Proc. IEEE International Conference on Smart Grid Communications (SmartGridComm'10)*, 2010, pp. 483–488.
- [39] M. Yigit, V. C. Gungor, S. Baktir, Cloud computing for smart grid applications, *Computer Networks* 70 (0) (2014) 312–329.
- [40] A. Kasper, Legal aspects of cybersecurity in emerging technologies: Smart grids and big data, in: T. Kerikme (Ed.), *Regulating eTechnologies in the European Union*, Springer International Publishing, 2014, pp. 189–216.
- [41] G. N. Ericsson, Cyber security and power system communication-essential parts of a smart grid infrastructure, *IEEE Trans. Power Del.* 25 (3) (2010) 1501–1507.
- [42] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, Customer characterization options for improving the tariff offer, *IEEE Trans. Power Syst.* 18 (1) (2003) 381–387.
- [43] N. Dahal, R. L. King, V. Madani, Online dimension reduction of synchrophasor data, in: *Proc. IEEE PES Transmission and Distribution Conference and Exposition (T&D)*, 2012, pp. 1–7.
- [44] L. Xie, Y. Chen, P. R. Kumar, Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis, *IEEE Trans. Power Syst.* 29 (6) (2014) 2784–2794.
- [45] J. Taheri, A. Y. Zomaya, Artificial neural networks, in: *Handbook of Nature-Inspired and Innovative Computing*, Springer US, 2006, pp. 147–185.
- [46] M. N. Q. Macedo, J. J. M. Galo, L. A. L. de Almeida, A. C. de C. Lima, Demand side management using artificial neural networks in a smart grid environment, *Renewable Sustain. Energy Rev.* 41 (0) (2015) 128–133.
- [47] S. V. Verdu, M. O. Garcia, F. Franco, N. Encinas, A. G. Marin, A. Molina, E. G. Lazaro, Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters, in: *IEEE Power Systems Conference and Exposition (PES)*, Vol. 2, 2004, pp. 899–906.
- [48] A. Monti, F. Ponci, Power grids of the future: Why smart means complex, in: *Proc. IEEE Complexity in Engineering (COM-PENG'10)*, 2010, pp. 7–11.
- [49] A. Sancho-Asensio, J. Navarro, I. Arrieta-Salinas, J. E. Armendáriz-Íñigo, V. Jiménez-Ruano, A. Zaballos, E. Golobardes, Improving data partition schemes in smart grids via clustering data streams, *Expert Syst. Appl.* 41 (13) (2014) 5832 – 5842.
- [50] E. Kyriakides, M. Polycarpou, Short term electric load forecasting: A tutorial, in: K. Chen, L. Wang (Eds.), *Trends in Neural Computation*, Vol. 35 of *Studies in Computational Intelligence*, Springer Berlin Heidelberg, 2007, pp. 391–418.
- [51] C. Guan, P. B. Luh, L. D. Michel, Y. Wang, P. B. Friedland, Very short-term load forecasting: Wavelet neural networks with data pre-filtering, *IEEE Trans. Power Syst.* 28 (1) (2013) 30–41.
- [52] X. S. Han, L. Han, H. B. Gooi, Z. Y. Pan, Ultra-short-term multi-node load forecasting-a composite approach, *IET Gener., Transm. Distrib.* 6 (5) (2012) 436–444.
- [53] Q. Tang, S. K. S. Gupta, G. Varsamopoulos, Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach, *IEEE Trans. Parallel and Distrib. Syst.* 19 (11) (2008) 1458–1472.
- [54] H. Mori, E. Kurata, An efficient kernel machine technique for short-term load forecasting under smart grid environment, in: *Proc. IEEE Power and Energy Society General Meeting*, 2012, pp. 1–4.
- [55] O. Kramer, F. Gieseke, Analysis of wind energy time series with kernel methods and neural networks, in: *Proc. 7th International*

- Conference on Natural Computation (ICNC), Vol. 4, 2011, pp. 2381–2385.
- [56] P. D. Yoo, A. Y. Zomaya, Combining analytic kernel models for energy-efficient data modeling and classification, *J. Supercomput.* 63 (3) (2013) 790–799.
 - [57] P. D. Yoo, J. W. P. Ng, A. Y. Zomaya, An energy-efficient kernel framework for large-scale data modeling and classification, in: *Proc. IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW)*, 2011, pp. 404–408.
 - [58] P. D. Yoo, B. B. Zhou, A. Y. Zomaya, A modular kernel approach for integrative analysis of protein domain boundaries, *BMC Genomics* 10 (Suppl 3) (2009) S21.
 - [59] P. D. Yoo, Y. S. Ho, B. B. Zhou, A. Y. Zomaya, Adaptive locality-effective kernel machine for protein phosphorylation site prediction, in: *Proc. IEEE International Symposium on Parallel and Distributed Processing (IPDPS)*, 2008, pp. 1–8.
 - [60] B. Li, S. Gangadhar, S. Cheng, P. K. Verma, Predicting user comfort level using machine learning for smart grid environments, in: *Proc. IEEE PES Innovative Smart Grid Technologies (ISGT)*, 2011, pp. 1–6.
 - [61] M. Couceiro, R. Ferrando, D. Manzano, L. Lafuente, Stream analytics for utilities. Predicting power supply and demand in a smart grid, in: *Proc. International Workshop on Cognitive Information Processing (CIP)*, 2012, pp. 1–6.
 - [62] Y. Murakami, Y. Takabayashi, Y. Noro, Photovoltaic power prediction and its application to smart grid, in: *Proc. IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, 2014, pp. 47–50.
 - [63] A. Y. Zomaya, Y. C. Lee, *Energy Efficient Distributed Computing Systems*, Wiley-IEEE Computer Society Press, 2012.
 - [64] S. Bera, S. Misra, J. J. P. C. Rodrigues, Cloud computing applications for smart grid: A survey, *IEEE Trans. Parallel and Distrib. Syst.* PP (99) (2014) 1–1.
 - [65] B. Hayes, *Commun. ACM*, *Cloud Computing* 51 (7) (2008) 9–11.
 - [66] Y. C. Lee, A. Zomaya, Energy conscious scheduling for distributed computing systems under different operating conditions, *IEEE Trans. Parallel and Distrib. Syst.* 22 (8) (2011) 1374–1381.
 - [67] M. Ghamkhari, A.-H. Mohsenian-Rad, Energy and performance management of green data centers: A profit maximization approach, *IEEE Trans. Smart Grid* 4 (2) (2013) 1017–1025.
 - [68] D. S. Markovic, D. Zivkovic, I. Branovic, R. Popovic, D. Cvetkovic, Smart power grid and cloud computing, *Renewable Sustain. Energy Rev.* 24 (0) (2013) 566–577.
 - [69] P. Yang, P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, A. Y. Zomaya, Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications, *IEEE Trans. Cybern.* 44 (3) (2014) 445–455.
 - [70] L. Han, X. Han, J. Hua, Y. Geng, A hybrid approach of ultra-short term multinode load forecasting, in: *Proc. International Power Engineering Conference (IPEC’07)*, 2007, pp. 1321–1326.
 - [71] A. Hajdu, L. Hajdu, A. Jonas, L. Kovacs, H. Toman, Generalizing the majority voting scheme to spatially constrained voting, *IEEE Trans. Image Process.* 22 (11) (2013) 4182–4194.
 - [72] B. N. Alajmi, K. H. Ahmed, S. J. Finney, B. W. Williams, Fuzzy-logic-control approach of a modified hill-climbing method for maximum power point in microgrid standalone photovoltaic system, *IEEE Trans. Power Electron.* 26 (4) (2011) 1022–1030.
 - [73] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8) (2005) 1226–1238.
 - [74] H. Deng, G. Runger, Feature selection via regularized trees, in: *Proc. 12th IEEE International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–8.
 - [75] A. Prinzie, D. V. den Poel, Random forests for multiclass classification: Random multiNomial logit, *Expert Systems with Applications* 34 (3) (2008) 1721–1732.
 - [76] B. Wang, S. Huang, J. Qiu, Y. Liu, G. Wang, Parallel online sequential extreme learning machine based on MapReduce, *Neurocomputing* 149, Part A (0) (2015) 224–232.
 - [77] T. Anderson, *The theory and practice of online learning*, Athabasca University Press, 2008.
 - [78] D. Lauri, J. V. Salcedo, S. Garcia-Nieto, M. Martinez, Model predictive control relevant identification: Multiple input multiple output against multiple input single output, *IET Control Theory Appl.* 4 (9) (2010) 1756–1766.
 - [79] P. D. Yoo, A. R. Sikder, B. B. Zhou, A. Y. Zomaya, Improved general regression network for protein domain boundary prediction, *BMC Bioinformatics* 9 (2008) S12.
 - [80] J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky, Bayesian model averaging: A tutorial, *Stat. Sci.* 14 (4) (1999) 382–401.
 - [81] P. D. Yoo, S. Muhaidat, K. Taha, J. Bentahar, A. Shami, Intelligent consensus modeling for proline cis-trans isomerization prediction, *IEEE/ACM Trans. Comput. Biol. Bioinf.* 11 (1) (2014) 26–32.
 - [82] P. D. Yoo, B. B. Zhou, A. Y. Zomaya, Machine learning techniques for protein secondary structure prediction: An overview and evaluation, *Current Bioinformatics* 3 (2) (2008) 74–86.
 - [83] N. B. Rizvandi, J. Taheri, R. Moraveji, A. Y. Zomaya, A study on using uncertain time series matching algorithms for MapReduce applications, *Concurr. Comput.: Pract. Exper.* 25 (12) (2013) 1699–1718.
 - [84] L. Breiman, Random forests, *Machine learning* 45 (1) (2001) 5–32.
 - [85] P. D. Yoo, A. Y. Zomaya, K. Alromaihi, S. Alshamsi, Tree-based consensus model for proline cis-trans isomerization prediction, in: *Proc. International Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW)*, 2013, pp. 454–458.
 - [86] N. Lu, P. Du, P. Paulson, F. L. Greitzer, X. Guo, M. Hadley, The development of a smart distribution grid testbed for integrated information management systems, in: *Proc. IEEE Power and Energy Society General Meeting*, 2011, pp. 1–8.
 - [87] <https://ich.smartgridsmartcity.com.au/>, Accessed: 2015.22.02.
 - [88] H. Liang, B. J. Choi, A. Abdrabou, W. Zhuang, X. Shen, Decentralized economic dispatch in microgrids via heterogeneous wireless networks, *IEEE J. Sel. Areas Commun.* 30 (6) (2012) 1061–1074.
 - [89] Canadian wind energy atlas, <http://www.windatlas.ca/>, Accessed: 2015.22.02.
 - [90] Nrel (national renewable energy laboratory), <http://www.nrel.gov/rredc/pvwatts>, Accessed: 2015.22.02.
 - [91] Waterloo north hydro, <http://www.wnhydro.com/>, Accessed: 2015.22.02.
 - [92] M. Mayilvaganan, M. Sabitha, A cloud-based architecture for big-data analytics in smart grid: A proposal, in: *Proc. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2013, pp. 1–4.