IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 16, 2020, accepted November 24, 2020, date of publication November 27, 2020, date of current version April 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3041178

Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications

DABEERUDDIN SYED[1,2], **(Graduate Student Member, IEEE), AMEEMA ZAINAB**[1], **(Graduate Student Member, IEEE), ALI GHRAYEB**[2], **(Fellow, IEEE), SHADY S. REFAAT**[2], **(Senior Member, IEEE), HAITHAM ABU-RUB**[2], **(Fellow, IEEE), AND OTHMANE BOUHALI**[3,4], **(Member, IEEE)**

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2]Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha 23874, Qatar
[3]Research Computing, Texas A&M University at Qatar, Doha 23874, Qatar
[4]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha 5825, Qatar

Corresponding author: Dabeeruddin Syed (dsyed@tamu.edu)

This work was supported by the Qatar National Research Fund (a member of Qatar Foundation) through National Priorities Research Program (NPRP) Grant NPRP10-0101-170082. The Open Access publication of this article was funded by the Qatar National Library.

ABSTRACT Smart grids have been gradually replacing the traditional power grids since the last decade. Such transformation is linked to adding a large number of smart meters and other sources of information extraction units. This provides various opportunities associated with the collected big data. Hence, the triumph of the smart grid energy paradigm depends on the factor of big data analytics. This includes the effective acquisition, transmission, processing, visualization, interpretation, and utilization of big data. The paper provides deep insights into various big data technologies and discusses big data analytics in the context of the smart grid. The paper also presents the challenges and opportunities brought by the advent of machine learning and big data from smart grids.

INDEX TERMS Big data, data analytics, smart grid, big data management, machine learning.

I. INTRODUCTION

The electrical power system has recently witnessed massive developments. Technical developments have been witnessed not only in the power generation side but also in the transmission and distribution sides. Furthermore, the new technology is expected to revolutionize the end-user side by adopting various demand management programs and techniques. The renewable energy sources, such as solar and wind sources, are not just added to the generation side by the utility companies, but also by the end consumers and microgrids. Also, vehicle to grid technology has provided opportunities for power flow management and the management of its flow from vehicles to the grid.

Once the electricity from renewable sources is increased to a large quantity, it would bring variability in the electrical system. This variability requires that innovative flexibility measures are considered to balance the demand and supply all the time. Novel approaches are required to improve the flexibility of the energy system ranging from supply to demand side. The concept of a smart grid and the use of big data analytics will help to manage the power systems better and also to increase resilience.

There has been and will be more massive installation of smart meters at the customer premises. These meters monitor the near real-time usage of energy while also collecting and communicating the data to electric utilities. The emergence of the power system deregulation on the delivery side and the moving away from the business model of vertically integrated utility have contributed to the development of the smart grids. The principle of smart grids solves the power demand problems by providing two-way power and information flow between consumers and utility [1]. Smart meters have been installed all across the world in the past years along with the transformation of the traditional power grids into the smart grid. The development of a smart grid is fully associated with the big data flow. There are various prospective applications of big data analytics on smart grid data such as real-time and automatic processing of the electrical consumers' energy consumption, automatic billing, intelligent energy planning

The associate editor coordinating the review of this manuscript and approving it for publication was Dezhong Peng.

and pricing analysis, detection of outages due to faults and anomalies, load and generation forecast under high unpredictability, load management with demand response, and asset management [2]. A high volume of data obtained from various smart grid sources satisfies the characteristics of big data. The grid data not only displays the Volume, Velocity, and Variety characteristics but also the 'V' characteristics such as Veracity, Visibility, and Value [3]. These characteristics are the challenges when dealing with big data analytics along with other major concerns such as security, privacy, etc. [4].

The smart grid allows for the two-way energy and information flow between consumers and utilities [5]. However, managing the real-time data for making business-valued decisions is still a persisting challenge [6], [7]. Currently, many utilities are working on installing a large number of smart meters and utilizing this data for effective demand response and resource management. For example, a company named Iberdrola has installed more than 11 million smart meters in Spain, generating 240 million registers every day [8]. Big data techniques, over an estimated volume of 90 billion registers per year, are being used to improve revenue collection and to optimize energy use. In the smart grid, the number of smart sensors is much higher, and the generated data are significantly much larger.

More data utilization helps in improving grid reliability and performance and ensures better decisions by the utility provider, thereby allowing for effective demand-side management and demand response [9]. However, the high volume of raw data is not directly comprehensible or useful without a dependable and consistent ability to process, analyze, and understand the information contained within such a huge amount of data. Therefore, the data should be transformed into useful information before action can be taken based on the data. Such transformation is a complicated process as beneficial information is not obvious from the data. The factors that contribute to the complications are visualization, high data dimensionality, and the application. A part of the information needs to be used by the automated systems while other information needs to be visualized and presented to users. Also, the time scales for various applications are different, ranging from milliseconds to days. The challenges involved with the use of smart grid data for analytics can be categorized as 1) decisions on mapping the data collection infrastructure to the desired applications, 2) application of new architecture and tools to manage grid data as streams in real-time, 3) transforming processes throughout the utilities to support the big data infrastructure, 4) managing the humongous amounts of data to make decisions that allow the benefits from the information obtained from smart grids' data.

A. CONTRIBUTION AND ORGANIZATION

This paper presents the big data technologies that can be utilized with smart grids' data. It also proposes a viable platform for the real-time stream processing of big data in smart grids. The contributions of this review are the following:

1) conducting an extensive literature review of the big data sources and types of data in the smart grid.
2) presenting the big data technologies that can be employed on smart grids' big data, and reviewing the latest developments, especially in the last decade.
3) presenting the overview of the big data process, the stages, and various techniques available for big data analytics.
4) providing the well-designed commercial solutions of big data for smart grids implemented by different corporations and utilities across the world.
5) discussing the potential application areas in smart grids that can benefit from the untapped potential of big data.

This paper is structured as follows. Section II details the data flow in the smart grid while Section III provides a comprehensive review of the big data analytics process. In Section IV, the technologies used for the big data within the setting of the smart grid in the literature are discussed. Section IV presents the proposed potential technologies that aim to overcome the challenges involved in the processing of smart grid data. Applied solutions used for big data analytics in smart grids are detailed in section V. Finally, in Section VI, the potential application areas which can avail from the big data analytics in a smart grid are reviewed.

II. SMART GRID DATA FLOW

A smart grid is formed by the integration of information and communications technology, electrical networks, and automation. The smart grid as an enabling engine is depicted in Figure 1. The electrical networks in the smart grid require the deployment of smart meters, sensors, devices, and control strategies. These have evolved due to the integration of renewable energy sources that are normally considered variable and unreliable sources of energy and are completely clean to the environment [10]. The smart grid aims to incorporate all the energy sources to match not only the baseline load but also the intermediate and peak loads.

In the smart grid, there is a lot of scope with big data analytics apart from creating intelligence and obtaining information from the raw data [11]. The scope of big data analytics has been illustrated in Figure 2. It is required that the big data architecture provides the potential to perform different types of analytics on the voluminous data to interpret it and derive business-valued applications. The different application areas for big data analytics in a smart grid will be discussed in Section VI.

A. BIG DATA SOURCES

Data from the smart grids are generated in real-time at a very high rate and volume [12]. The extraction of information from smart grid data is required for grid applications and calls for deep insight into the data sources. The data in smart grids can be classified into consumer, distribution, transmission, and generation data. These data are acquired from the sensors, smart meters, grid devices, detectors, Supervisory Control And Data Acquisition (SCADA), etc. The collected signals

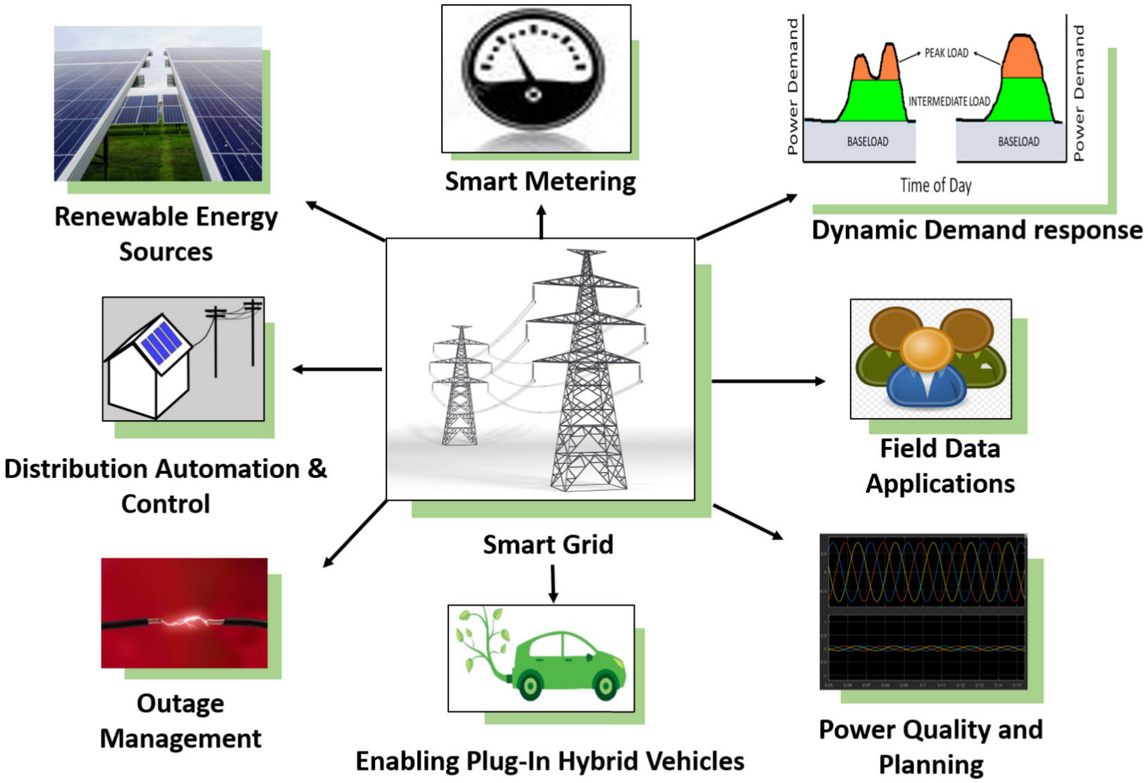


FIGURE 1. Smart grid as enabling engine - depiction of opportunities.

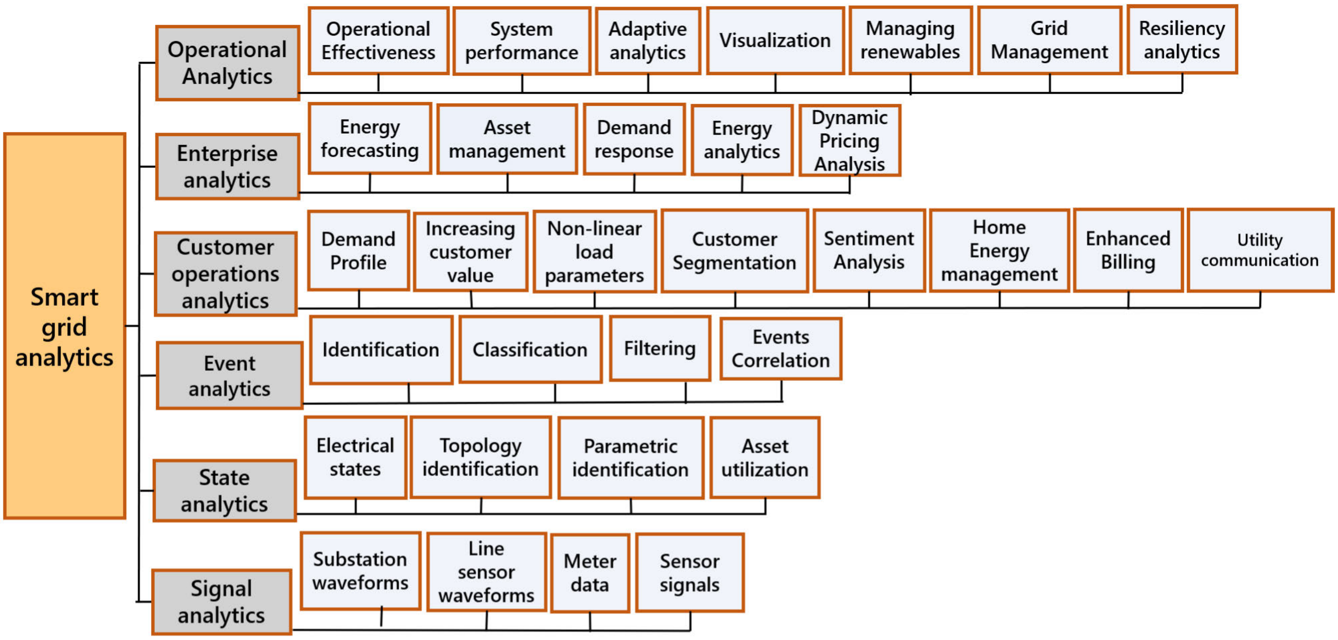


FIGURE 2. Scope of big data analytics in smart grid [11].

relate to power utilization habits of consumers, phasor measurement, energy consumption, energy pricing and bidding, operation or financials for running the utility, etc. Types of sensors, information obtained from those sensors, and other sources of information in a smart grid are described in Table 1 [13], [14].

TABLE 1. Sources of data or information in smart grids.

Sources	Quantity being measured	Information extracted and applications
Advanced Metering Infrastructure (AMI)	E, Cumulative energy usage, peak load, load curve, phase, failure counts & logs, P.F., tamper factor, last interval demand	Market pricing, real-time on demand, remote meter configuration, demand-side management, electric usage, power quality monitoring, and local control
Distributed Generation Sensors	V, P.F.	Load balancing
Digital Fault Recorder (DFR)	Power swing, load variation, transient phase angle changes, frequency fluctuations, also records power system events like time of fault, power disturbance	Faults classification
Electrical Measurement Sensors (EMS)	V, I, E, V_{sag}, P.F., Q_{reac}, electric & magnetic fields	Revenue
Fibre Bragg Grating sensor (FBG)	wavelength shift under changes in strain & temperature	Prediction of overheating, sag, vibration, galloping
Geographical Information System (GIS)	GIS data	Asset management & map the location of outages
Hall Effect sensor	V and magnetic field	Current sensing, proximity switching, positioning, speed detection
High Voltage Line Temperature and Weather Condition Sensors	T, record weather conditions	Preventive maintenance
Intelligent electronic device (IED)	Records status changes in substation and outgoing feeders	Relay protection
Line Fault detectors	V, I, P, harmonics, phase angle	Transmission or Distribution faults
Magnetoresistive sensors	Current, power, total energy, frequency, modulation	Transient Magnetic Field, EMI in substation
Phasor Measurement Unit (PMU)	V, I, P, harmonics, phase angle	Time synchronized measurements with phase angles, electrical waves measurement of power grid
Remote Terminal Unit (RTU)	Transmits telemetry data and controllable by micro-processor	system operation status
Smart Capacitor control	V, I, VAR and harmonic monitoring	Monitoring & control of capacitor banks remotely
Sagometer	T	Line Sagging
Smart Sensors for Outage Detection	T, I	Outage detection
SCADA	V, I, E, P.F.	Automatic control, protection, system monitoring, event processing and alarm
Smart Sensors for Transformer Monitoring	V, I, T, load tap changer values, partial discharge, dissolve gas data	Preventive maintenance
Smart Voltage Sensors	V	Voltage Regulation
Wide area monitoring system (WAMS)	Deals with incoming data from PMUs	Dynamic stability of the grid
V = Voltage, I = Current, P = Power,	E = Energy, V_{sag} = Voltage sag, P.F. = Power factor,	T = temperature

Also, large datasets, not directly related to the grid, such as weather data, GIS data, etc. should be used for situational awareness and decision making. Owing to security and privacy concerns, the electric utilities do not share the smart meter data publicly and this poses a challenge to the research community. There are several benchmark and publicly-accessible data that have been anonymized or semi-anonymized and that the researchers can use to validate the performance of their proposed modeling and data analytics methodologies. The summarization of the list of public data sources is given in Table 2.

B. DATA STRUCTURES

Contrary to traditional data analysis, big data analysis deals with semi-structured, quasi-structured, and unstructured data in addition to structured data [25], [26].

- **Structured data**: Structured data is the data that comprises clearly defined data types, structure, and format whose patterns make the data easily searchable. Few examples include data that can be stored in spreadsheets, Comma-separated Values (CSV) file, a traditional Relational Database Management System (RDBMS), data cubes in Online Analytical Processing (OLAP), relational tables containing customer information, electrical consumption data in numbers and strings, etc. Meters' data, distribution management data, equipment parameters, load control data, marketing system data in relational format, etc. are examples of structured data in smart grids.
- **Semi-structured data**: Semi-structured data is textual data that contains perceptible data patterns and enables parsing. For example, the XML and JSON data files are

TABLE 2. Publicly-accessible data sources.

Data Source Name	Data Description
Ausgrid network [15]	Load profile data at the substation level.
Commission for Energy Regulation (CER) smart metering project [16]	Smart meter data from Ireland.
Cornell campus smart grid [17]	Smart meter data.
The École polytechnique fédérale de Lausanne (EPFL) smart grid data (Switzerland) [18]	PMU data.
Electric Reliability Council of Texas (ERCOT) data [19]	Market data.
North American SynchroPhasor Initiative (NASPI) data [20]	PMU data.
Pecan Street project [21]	Smart meter data.
Pennsylvania-New Jersey-Maryland (PJM) market data [22]	Market data.
Residential or commercial data [23]	Consumption, electric vehicles, power quality, PV generation, reliability, weather, wind-based generation, and general energy data.
University of California (UC) Berkeley campus smart grid [24]	Smart meter and building consumption data.

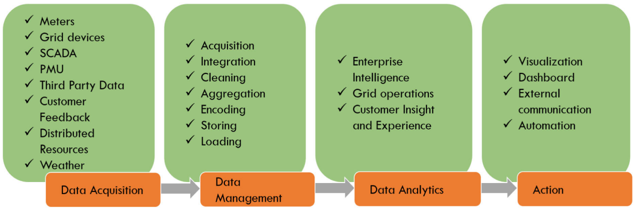


FIGURE 3. High-level view of the flow of data into the utility [27].

self-describing and defined by its schema. Web service data, load monitoring, power quality data, etc. are examples of semi-structured data in smart grids.

- **Quasi-structured data**: Quasi-structure data is textual data that contains erratic data formats but can be properly formatted with tools after time and effort. The only difference between the semi-structured and quasi-structured data is that semi-structured data has metadata associated with them and the metadata can be easily used to structure or format the data. Whereas the quasi-structured data requires intelligence-aware approaches to structure or format them. For example, web click-stream that contains erratic formats and data values, web scrapping data, search engine results, etc. are quasi-structured data.
- **Unstructured data**: Unstructured data is data that has no pre-defined models or schema. Examples include publicly collected census and text, social media streams and tweets, audio, video, photographs, etc. Meteorological information, customer service data, economy data of distribution regions, etc. are examples of unstructured data in smart grids.

The high-level view of the data flow into the utility is illustrated in Figure 3 [27]. The first step is the data collection in which the major classes of data are collected from various sources, eg. the customer data is collected using smart meters, grid data is measured on distribution and transmission lines using PMUs and synchrophasors, etc. Other important data, that are collected, include SCADA data, market data, weather data, and customer feedback in the form of tweets, text, videos, audio, and pictures. The complex and heterogeneous data from multiple sources are then transmitted through various communication networks and stored in the relational database, data warehouse, file servers, application servers, Hadoop clusters, etc. This comes under the phase of data management where the data undergoes extraction, cleaning, aggregation, and encoding. Finally, the data are loaded into any in-memory distributed databases for further analytics. The third phase is analytics where the actual information stored in data is extracted to represent business value. The data analytics is performed using approaches such as time series analysis, feature selection, feature extraction, machine learning modeling, deep learning modeling, clustering, incremental learning, adaptive learning, reinforcement learning, etc. with an aim to enhance applications for enterprise intelligence, grid operations, and customer insight. The applications may include the following but are not limited to: load profiling, load forecasting, demand response, program marketing, outage management, bad data detection, etc [28]. Finally, the information should enable action in the form of automation, external communication, and monitoring through visualization and dashboards.

III. BIG DATA ANALYTICS PROCESS

Big Data analytics requires pre-defined strategies because of the high volume of data. Also, the velocity and variety of data pose challenges in the data analytics process. It is very crucial that the data from the smart grid are processed in real-time because significant patterns can be recognized from the data to make better decisions. Data analytics deals with the extraction of actionable knowledge and patterns from the available data [29]. The big data analytics process is illustrated in Figure 4.

There are four major types of big data analytics [30]. These are described as follows:

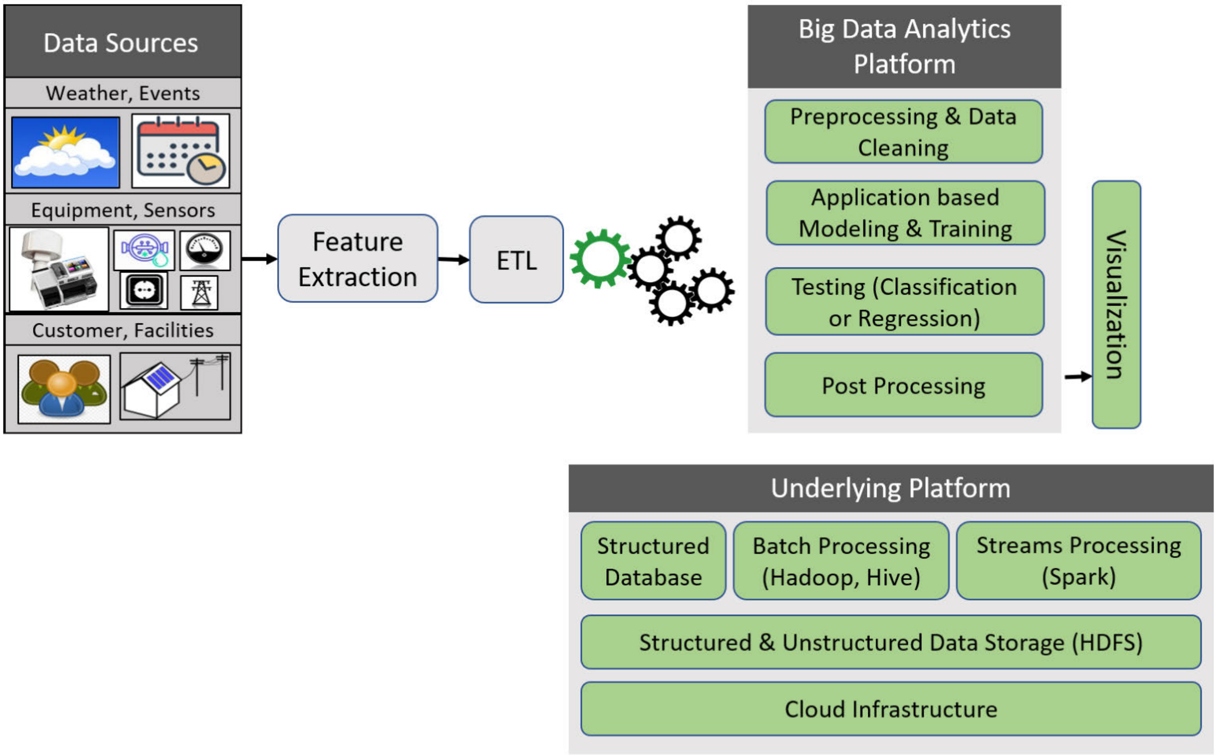

Data Sources
Weather, Events
Equipment, Sensors
Customer, Facilities
Feature Extraction
ETL
Big Data Analytics Platform
Preprocessing & Data Cleaning
Application based Modeling & Training
Testing (Classification or Regression)
Post Processing
Visualization
Underlying Platform
Structured Database
Batch Processing (Hadoop, Hive)
Streams Processing (Spark)
Structured & Unstructured Data Storage (HDFS)
Cloud Infrastructure

FIGURE 4. Big data analytics process.

a: DESCRIPTIVE ANALYTICS

Descriptive analytics illustrates what happened in the past using the historical data available and shows the data in an easily understandable form or visualization. In general, the data is illustrated using graphs, bar diagrams, pie diagrams, maps, scatter plots, etc. In short, descriptive analytics is performed to understand or illustrate the patterns in the data.

b: PREDICTIVE ANALYTICS

It extrapolates from the data available to predict what can happen in the future. The tools that are used for predictive analytics are time-series analysis using statistical methods and other data mining algorithms. Predictive analytics is usually performed to predict which events can happen in the future.

c: EXPLORATORY ANALYTICS

It finds hidden correlations or relationships between features in the data. This helps us to estimate values for a dependent feature when information is available for the independent features. Exploratory analytics is basically performed to determine the cause behind the events that have happened in the past.

d: PRESCRIPTIVE ANALYTICS

It is used to discover the best outcome of past events when the features of the data and operating parameters of a system are given. It helps to develop strategies for future events under similar conditions. The techniques involve simulation tools and these simulate the operating conditions or features to finally come up with the best outcome. The simulation techniques strategize how to plan for similar events in the future. Prescriptive analytics is basically performed to know how preferable events can be made to happen in the future. Example: power flow analysis, etc.

Data analytics starts with the acquisition of data following which the data is processed to reveal information.

A. DATA ACQUISITION

The first step in any of the data analytics process is the collection of data. The data in the smart grid are collected from various sources as mentioned in the earlier section. With the data collection already in place, the other sub-tasks in data acquisition are data communication and data preprocessing. The raw data need to be transmitted either to a real-time stream processing system or to a storage system from where the data can be sent to the offline batch processing system for further analysis. Since the data have been collected from diverse and multiple sources, the data aggregation and cleaning are the foremost and crucial steps. Data aggregation services should be in place to integrate the data from varied sources and furnish a unified view of the available data. In data pre-processing, the inconsistent and missing data are to be filled or one among the records, and the features are to be removed to improve the data quality [31]. It is crucial to refine the features in the extracted data as there are noise and redundancy in the collected raw data. Refining the features involve either feature selection or feature extraction. If the data contain highly correlated features, then the machine learning algorithms, in general, perform poorly. Regularization techniques are used to overcome the issues of overfitting whereas underfitting would require the acquisition of more data and that is not an issue in the case of big data [32].

B. DATA PROCESSING

The data collected and transmitted should be stored in storage infrastructure for further processing. The stage, at which the data is processed, classifies data processing into the following types:

1) BATCH PROCESSING

Batch analytics is fundamentally the analysis of data in batches. It involves the workflow on offline data where all the data are available, pre-extracted, and ingested using scripts and a huge group of data is analyzed in a single execution. Distributed file systems (DFS) provides for the fault-tolerant scalable storage of data across commodity hardware where the storage nodes do not share memory but are connected virtually through networking [33]. MapReduce and Hadoop framework provides such a DFS framework. In MapReduce, a huge amount of data is processed by dividing the job into a set of sub-jobs and each sub-job handles a small portion of data and all the sub-jobs operate in parallel to obtain the intermediate outcomes. The final result is then obtained by the aggregation of the intermediate outcomes. The advantage of the MapReduce paradigm with respect to batch processing is the data locality principle. In this principle, the algorithm or the user code is moved close to data rather than moving the data to the algorithm. This requires the movement of computational resources to where the data is located and thus, prevents overhead from the data transmission. The disadvantage of batch processing is that it cannot provide analytics results in real-time. An example of batch processing in smart grids includes the training of data-driven models using offline data for applications of topology identification, predictive maintenance, energy forecasting, etc. These models would require re-training if new data become available and need to be included in the modeling performance. There is no specific time interval defined to term processing as batch analytics. However, it is usually considered that if the processing is scheduled to happen with an interval equal to or greater than 20 minutes, then it is batch processing.

2) STREAM PROCESSING

Stream processing is primarily the processing of each new data instance as soon as it is available instead of waiting for batches of offline data. The idea behind the stream processing is that the potential worth of information from data relies on the freshness of data [34]. Hence, it is crucial that the stream processing model processes the data as soon as the data instance is available to obtain approximation results. If the data are continuously available in huge streams, a portion of the data can be stored in memory until it is processed.

In the subsequent sections, we will focus on the technologies that possess the capability of processing big data in real-time. They provide a huge advantage of handling data with high-velocity requirements. In our platform, the Hadoop File system is used as a storage system and spark streaming provides the real-time processing solution along with tools such as Spark Structured Query Language (SQL), Spark Machine Learning Library (MLlib), GraphX, etc. Examples of stream processing in a smart grid include stateless conversion, stateless filtering, aggregation, pre-processing, wavelet transformations, etc. of the data. The time interval for data processing to be termed as stream processing is typically seconds or milliseconds.

3) ITERATIVE PROCESSING

There are a few big data problems that require the processing of data iteratively and demand more number of read and write operations than batch processing and stream processing. These involve a high number of Input-Output transfers and are time-consuming.

For big data analytics on smart grids' data, we focus on the batch and stream processing and the comparison of these is given in Table 3.

TABLE 3. Batch v/s stream processing.

	Batch Processing	Stream Processing
Input form	Chunks of data	Streams of data
Input data size	Known & finite	Unknown & infinite
Is data stored?	Yes	Data is not stored (or) small streams stored in memory
Hardware used	Multiple Central Processing Units (CPU) & memory	Restricted memory
Processing	Multiple rounds	Single round processing
Time	Longer time	Seconds or milliseconds
Applications	Widely adopted	Sensor networks, web mining, etc.

C. DATA ANALYTICS TECHNIQUES

Multiple machine learning algorithms are used as data analytics techniques. These techniques are used to map the relationship between the features in the data and the prediction label usually. If the labels exist, the techniques employed are named as supervised techniques. Whereas the data may not explicitly consist of labels and it is up to the algorithm to recognize the patterns in the data. These techniques that work on data without labels are termed as unsupervised techniques. The summarization of the different classes of machine learning techniques, that have been previously applied in smart grids, is presented in Table 4, Table 5, and Table 6.

IV. TECHNOLOGIES FOR BIG DATA ANALYTICS

In this section, we present the hierarchical architecture of state-of-the-art core components of big data analysis for

TABLE 4. Dimensionality reduction algorithms.

Algorithm	Description
Principle Component Analysis (PCA) [35]	Most widely used unsupervised technique; heuristic approach to extract variance structure from high-dimensional data; involves 1) feature scaling & mean normalization, 2) calculation of covariance matrix and 3) sorting the eigenvectors that represent components.
Linear Discriminant Analysis (LDA) [36]	Supervised technique; projection of data from higher dimensional space to lower one so that it maximizes between-class & minimizes within-class distances.
Kernel Discriminant Analysis (KDA) [37]	Obtains linear separation by non-linear mapping of input space to high-dimensional feature space.
t-distributed stochastic neighbor embedding (t-SNE) [38]	Converts high-dimensional data into a matrix of pair-wise similarities using conditional probabilities, and variation of stochastic neighbor embedding.

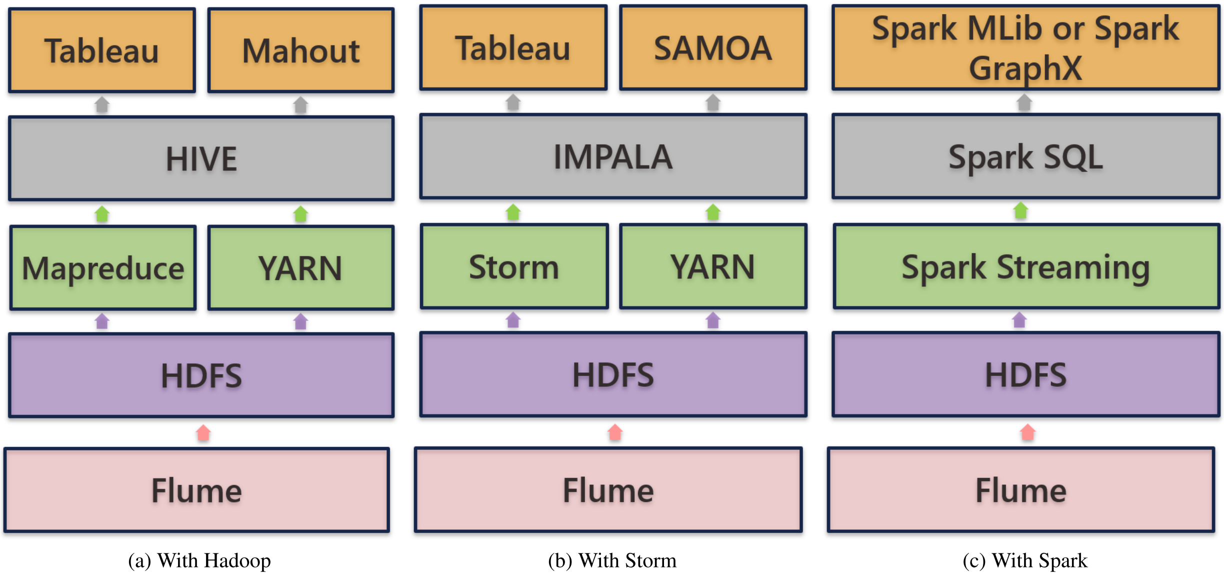
TABLE 5. Supervised algorithms.

Algorithm	Description
Linear Regression (LR) [39]	Curve fitting regression technique for linear functions; the hypothesis function is linear.
Polynomial Regression [40]	Curve fitting regression technique for non-linear functions; the hypothesis is a linear model of basis functions (linear, polynomial, Gaussian Radial Basis Function (RBF), sigmoid, etc.)
Logistic Regression [41]	Classification technique to identify decision boundary; the hypothesis function is sigmoid.
Neural Networks (NN) [42], [43]	Performs classification & regression; capable of modeling highly non-linear relationships with large feature space; parametric model; can represent complex logic operations & comprises input, hidden & output layers with activation functions (threshold, logistic, arctan, gaussian & relu); and types: convolutional, and recurrent.
Support Vector Machines (SVM) [44]	Large margin classifier; classifies non-linear data by introducing slack variables; SVM is found by minimization formulation under constraints that are overcome by the use of a Lagrangian multiplier. types: linear, and kernel.
Naive Bayes [45]	The parametric approach for likelihood estimation assumes that the data features are independent.
k-Nearest Neighbor (kNN) [46]	Non-parametric approach for likelihood estimation; classifies a data point to the majority class among k Neighbors;
Decision Tree (DT) [47]	Recursive, partition-based tree model that predicts a class based on split points; the algorithm takes leaf size and purity threshold as inputs; the process stops when leaf size or purity threshold is reached.
Random Forest (RF) [48]	Collection of low-bias, high-variance trees; and outputs mode of the classes or mean prediction.

TABLE 6. Unsupervised algorithms.

Algorithm	Description
K-Means Clustering [49]	The representative-based technique includes steps of initializing cluster centroids, grouping data points to nearest centroids, updating centroids, and uses euclidean distances & variables are to be quantitative.
Expectation Maximization Clustering [50]	The representative-based technique includes steps of initializing cluster mean, calculating posterior probability, and re-estimating means, covariance & priors.
Gaussian Mixture Clustering [51]	Fits k-Gaussians to cluster the data. The result is the weighted average of K-gaussian distributions.
Hierarchical Clustering [49]	Involves creating a sequence of nested partitions that can be visualized by a tree or hierarchy of clusters.
Density-based Spatial Clustering of Applications with Noise (DBSCAN) [52]	Density-based clustering that computes neighborhood to classify data points into core, border & noise points while also using a threshold called minimum points (minpts).
Association Rules [53]	Usually applied in market basket analysis, text mining, web usage mining, linguistics mining, etc. to determine the co-occurrence relationships or associations between all items in the database.
Collaborative Filtering [54]	Generally employed in recommender systems where preferences of a target user are predicted based on the user searches where users are similar to the target & mining on their preferences.

smart grids using Hadoop as shown in Figure 5a and using Storm as shown in Figure 5b. We also discuss the proposed platform and technologies for big data analytics for smart grids using Spark (shown in Figure 5c) [55], [56]. The major components perform the collection, storage, processing, visualization, and querying of data. There are a variety



(a) With Hadoop (b) With Storm (c) With Spark

FIGURE 5. Architecture for big data analytics platform.

of workloads present in the scenario of massive-scale data analytics. A combination of these workloads will present a potentially effective solution for the business goals in the scenario of smart grids.

A. EVOLUTION OF BIG DATA TECHNOLOGIES

When dealing with massive-scaled data, the framework was initially developed for the processing of offline large datasets. Apache Hadoop and MapReduce models provide open-source software frameworks for the distributed processing of offline data spread across data nodes or clusters using simple programming paradigms of the map and reduce functions. **MapReduce** abstracts from distributed programming but it still requires programming to a certain level. Moreover, MapReduce is efficient for batch tasks and not for ad-hoc queries or iterative processing. If the offline analysis or background task of indexing websites is required, then MapReduce is a suitable option. Hence, the combination of a distributed file system and MapReduce is suitable for write once and read many, or sequential data access, but not for random read or write access applications [57]. Yet, random read/access is required for the online analysis of data or the ad-hoc querying.

As a solution to the ad-hoc querying issue, **Not only SQL (NoSQL) databases** can be used. NoSQL Databases are of two types [58]. These are mentioned in the following:

- Column databases: A column-oriented database is a database that stores data in columns rather than rows. Furthermore, it is very effortless to add columns and these columns can be added row by row as well. The databases offer great flexibility, performance, and efficiency. Also, the performance of the column databases can be significantly enhanced by compression, late materialization, and batch processing.
  Examples of column databases include BigTable, HBase in Amazon Dynamo, Google Bigtable, Apache HBase, etc.
- Key-value stores: These are distributed data structures that provide key-based access to data and are also called Distributed Hash Tables.
  An example is Apache Cassandra.

NoSQL Databases are very efficient when dealing with massive-scale data even if the data type is unstructured or semi-structured. However, the only disadvantage is that these do not offer SQL-like querying. To make querying SQL-like, many NoSQL databases have been evolved with the SQL-like interface (Contextual Query Language (CQL) of Cassandra, Hive, Pig, etc.). There are developments in the form of SQL interfaces that can directly connect to the NoSQL databases (such as PrestoDB, etc.). The SQL-like interfaced NoSQL databases are termed as **NewSQL** and these possess the inherent capability of organizing massive-scaled data and sorting to enable efficient offline analyses (H-Store, Google Spanner, etc.) [59].

There has been a massive growth in the availability of digital data and the data are available in continuous streams. Therefore, NoSQL databases have been evolved to cater to the **stream-processing** solution with the fault-tolerant distributed data ingest systems such as Apache Kafka, Flume, etc. [60]. Examples of stream processing solutions are Apache Storm and Samsa. Also, there are standalone stream processing frameworks that are faster. Additionally, there have been solutions developed to employ OLAP-like processing in the big data landscape. Built on top of data structures, there are now libraries available for machine learning and big data analytics for real-time analytics processing. For example, there is an Apache Spark framework that contains

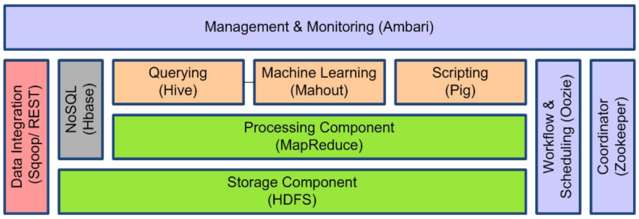

Management & Monitoring (Ambari)
Data Integration (Sqoop/ REST)
NoSQL (Hbase)
Querying (Hive)
Machine Learning (Mahout)
Scripting (Pig)
Processing Component (MapReduce)
Storage Component (HDFS)
Workflow & Scheduling (Oozie)
Coordinator (Zookeeper)

FIGURE 6. Apache Hadoop ecosystem [63].

machine learning libraries and can be used for massive-scale data analytics.

B. APACHE HADOOP AND MapReduce

1) HADOOP FRAMEWORK

Heterogeneity, volume, performance, scaling, cost, and security concerns of big data hinder the process of data analytics at every stage [61]. Apache Hadoop is an open-source framework that provides the distributed storage and processing of big data. It consists of the core (for storage part) called the Hadoop Distributed File System (HDFS), the processing component that is the MapReduce programming model and resource scheduler called Hadoop YARN (Yet Another Resource Negotiator) [62].

Following is the list of modules in the Apache Hadoop Ecosystem (as shown in Figure 6 [63]):

1) **Hadoop core**: Hadoop core contains a pre-defined collection of utilities and libraries that can be used by other modules within the Hadoop ecosystem [63]. For instance, if the data access module such as HBase, Hive, etc. needs to access the file storage system in Hadoop, then these are required to build Java Archive (JAR) files stored in the Hadoop core.
2) **HDFS**: The default distributed storage system in Apache Hadoop is the HDFS. The huge datasets are dumped in the HDFS and when required, access to the data is provided to other Hadoop modules using utilities [64]. HDFS component provides reliable and quick access to the data by creating several copies of the data block and these copies are distributed across multiple clusters. HDFS works on the master-slave architecture model and comprises three components namely NameNode, DataNode, and Secondary NameNode [65].
3) **Hadoop YARN**: YARN is the dynamic resource management component that lets the user run multiple Hadoop applications without having to worry about the aggravating workloads. YARN provides for improved cluster utilization. Key components of YARN are Resource Manager, Application Master, Node Manager, and containers.
4) **Hadoop MapReduce**: This is a framework for parallel computations of massive data sets.

2) MapReduce PROGRAMMING MODEL

MapReduce model is employed for the parallel computation and interpretation of massive-scale data and has three stages: map, shuffle, and reduce [66]. All the jobs are written in a functional programming style to create map and reduce tasks. Dynamic systems for the MapReduce model are commonly clusters that perform tasks such as data partitioning, scheduling of jobs, and communication between the cluster nodes and hence, are more suitable when dealing with massive-scale data. In the map phase, the data are read from the DFS and partitioned into clustered systems where the input is processed to compute the intermediary results which are then stored on the local node of the cluster where the map phase has run and waits for all the map functions to generate output in key-value pairs. The output in key-value pairs is then given as input to the reduce function to generate the final result. The advantage of the MapReduce model is that it takes processing to where the data resides and hence, decreases the transmission of data and improves efficiency. Therefore, the MapReduce model is more apt for the distributed computing of massive-scale data. The summary of the Hadoop module is illustrated in Table 7 [67].

TABLE 7. Summary of Hadoop module.

Stage	Software	Function
Data Acquisition	Flume	Data acquisition from varied sources to a centralized location
	Sqoop	Data Import & Export between centralized location & Hadoop
Data Storage	HDFS	Distributed File System
	HBase	Non-relational key-value based columnar data store
Computation	MapReduce	A parallel computation programming model
Querying	Pig	Procedural Data Flow platform
	Hive	SQL like language for querying
Analysis	Mahout	Machine Learning Library
Process Management	Zookeeper	Centralized service to maintain configuration information & synchronization.
	Chukwa	System Monitoring

Hadoop has provided for storing and analyzing data at massive scales. However, data analytics technology cannot be applied to real-time systems [68]. The advent of the Internet-of-Things, smart meters, and devices has led to the possibility of real-time analysis of data for the benefits of business and many other advantages such as smart grid stability, and management. The real-time handling of data falls under one of the categories: Stream processing or Iterative processing. The stream processing framework would work efficiently for big data analytics in the smart grid for real-time decisions about generation, control, etc.

C. APACHE STORM

It is a scalable and distributed framework for reliable computation and processing of streams of real-time data with processing latencies in the order of milliseconds. Apache Storm can ingest the data from multiple sources using Kafka or Kinesis. A storm cluster is very alike to the data cluster in Apache Hadoop [69]. In Hadoop, MapReduce jobs are executed while topologies are executed in Apache Storm. Topologies are very similar to jobs, but topologies process messages or data forever until these are killed.

In a Storm cluster, there are two types of nodes, namely master node and worker nodes [70]. A background process called Nimbus runs on the master node and this is analogous to Hadoop's job-tracker. Nimbus process distributes the code in the cluster i.e. assigns tasks to the machines and monitors for any failures. On the machines other than the master node, the process called Supervisor runs and it listens for the work assigned to its machine by the Nimbus daemon. It starts and stops the worker node process depending upon the task assigned to the machine. Every worker process runs a subset of topologies. That means the execution of topology requires multiple worker processes that are assigned to different machines across the cluster. It requires coordination between Nimbus daemon and Supervisor processes and this is taken care of by Zookeeper which is the coordinating service in the distributed environment [71]. Zookeeper takes care of naming, configuration, synchronization, etc. The important point to note is that all daemons in the Apache Storm are stateless and fail-fast and these come back up even if these are killed by issuing manual commands. This provides for the stable and reliable real-time analysis of big data.

D. APACHE SPARK

Apache Spark is an open-source cluster computing framework for analyzing massive-scaled data. It was originally developed by Matei Zaharia at UC Berkeley AMPLab [72]. Spark has the capability for stream processing of big data and has many advantages over Hadoop MapReduce and Storm. In Apache Spark, data analytics is more stream processing than batch processing and hence, it avoids the reprocessing of the data [73]. This provides the stream processing model of Apache Spark to be dynamic and it becomes more crucial during the real-time processing of huge volumes of data collected from different sources. Even for iterative processing, the leading framework currently is Apache Spark as it possesses the capability of processing and holding the data in the memory nodes across the cluster.

1) CHARACTERISTICS OF APACHE SPARK

- **Speed**: Spark extends the MapReduce model to execute computations of stream processing and interactive querying. In literature, it is proven to be 10 times faster than the Hadoop MapReduce model.
- **Ease of Use**: Applications written in any language such as Java, python, scala, etc. are compatible with Apache Spark.
- **Advanced Analytics**: Spark supports the MapReduce model of Hadoop, SQL-like querying, streaming data, machine learning algorithms, and graph algorithms as well.
- **Iterative and Interactive Applications**: Spark is designed to execute both in-memory and on-disk. It holds the intermediary results in memory rather than writing to disk to avoid reprocessing the data if required again. Spark operators perform external operations on the data if it does not fit into memory.
- **In-memory Computation**: The data is stored in memory rather than written on disk. Hence, Spark reduces the response time to a great extent when the data is queried.
- **Directed Acyclic Graph (DAG)**: DAG in Apache Spark is a set of vertices and edges where the vertices are the representations of the Resilient Distributed Datasets (RDDs) while edges represent the operations to be performed on the RDDs. DAGs in Spark can contain any number of stages. Even the MapReduce model of Hadoop is a DAG of two stages - Map and Reduce. This allows for simple jobs to be completed in one stage and more complex jobs to be completed in one run of many stages unlike multiple jobs in the MapReduce model. Thus, jobs in Spark execute faster than they would in the MapReduce paradigm.

2) SPARK FRAMEWORK

Other than core Spark, there are multiple components in the Spark ecosystem. These components as shown in Figure 7 [74].

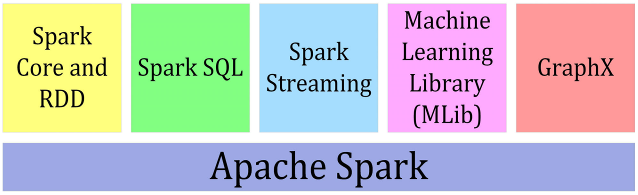


FIGURE 7. Apache spark [74].

Spark Core is the base of all the Spark projects and it allows basic input/output operations, distributed task dispatching, and scheduling through an Application Programming Interface (API) centered on RDD abstraction. RDD is a read-only collection of objects partitioned across a set of machines and it can be rebuilt if any of the partitions are lost [75]. RDDs are fault-tolerant, can be cached in-memory across machines, and can be reused in MapReduce-like simultaneous computations.

Spark SQL: Spark SQL is the Apache Spark module that is commonly worked with structured data. It lies on top of Spark core and is used to execute SQL queries. It introduces the schema RDD which can be manipulated. Users can interact with the SQL interface using the command line or over Open Database Connectivity (ODBC), Java Database Connectivity (JDBC) server, etc.

Spark Streaming: It is the component of the spark that enables the processing of live streams of data. Spark streaming gives a programming interface for processing data

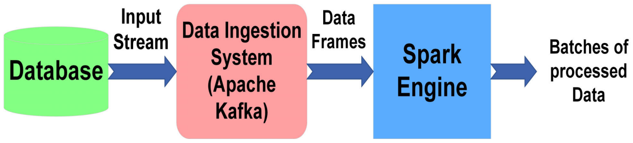

Database
Input Stream
Data Ingestion System (Apache Kafka)
Data Frames
Spark Engine
Batches of processed Data

FIGURE 8. Stream processing using Apache spark.

streams. It resembles the Spark core's RDD API, pushes data in small chunks, and does RDD transformations on the batches of data.

MLib: Apache Spark comprises a library with common machine learning functionality and this library is called MLib. It processes data faster when compared to Hadoop's disk-based machine learning library called Mahout.

GraphX: The GraphX API provides for users to view data in graphical format and to view RDDs without data movement or duplication. It uses the fundamental operators such as subgraphs, joinVertices, aggregateMessages, etc.

The summary of the proposed module of Spark on top of Hadoop is illustrated in Table 8.

TABLE 8. Summary of Apache spark module.

Stage	Software	Function
Data Acquisition	Flume	Data collection from sources to a centralized location
	Sqoop	Data import and export between centralized location & HDFS
Data Storage	HDFS	Storage of data across nodes with high bandwidth across the cluster
	HBase	Column-oriented key-value data store to store spark data sets
Processing	Spark Streaming	Computation framework
Querying	SQL	SQL-like language for querying
Analysis	MLib	Machine learning libraries
Visualization	GraphX	Visualizations

E. APACHE DRILL

It is an open-source software framework that provides for data-driven distributed applications requiring interactive processing of massive-scaled data. Apache Drill is the first and only distributed SQL engine that does not require schemas. Drill automatically understands the data when data are provided. This saves a lot of time and effort in defining schemas, transforming data, and maintaining those schemas. It is designed to handle Petabytes (PBs) of data spread across thousands of clusters and it responds to ad-hoc queries with high performance and low latency.

It is a query layer that functions even when multiple data sources are present. It primarily scans the full tables instead of maintaining indices. The workers in Apache Drill are named Drillbits and run on each of the data nodes in the cluster. The coordination between the drillbits, optimization, scheduling, and execution is performed in a distributed way.

The architecture of Apache Drill contains the following components:

User interface: It provides an interface for the user or application-driven interaction. For example, interface through a command line, Representational state transfer (REST), JDBC, ODBC, etc.

Processing layer: It comprises SQL Parses, Optimizer, Execution Engine, and Storage Engine.

Data Sources: The data in the pluggable data sources may be spread across thousands of nodes (in-cluster) or they can be local.

The comparison of the different frameworks can be summed up as shown in Table 9.

V. APPLIED SOLUTIONS FOR BIG DATA ANALYTICS IN SMART GRIDS

As mentioned before, there are few works that have been reported in the literature for big data analytics specifically in the smart grids. In particular, there are only a few commercial solutions available in the market. One of the earlier practical works on big data analytics was based on the Naive Bayes classification method using the MapReduce paradigm for novel transient power quality assessment [76]. In [77], the authors proposed a cloud-based architecture using Hadoop, Cassandra, and Hive for big data analytics in a smart grid using the data on power usage patterns of customers, historical weather data, supply and demand data.

In [78], Munshi *et al.* presented an implementation of cloud-based Lambda architecture for smart grid big data analytics using Hadoop data lake. The Lambda architecture is aimed to provide a trade-off between latency throughput and fault tolerance while providing the batch and stream processing capabilities for parallel computation of arbitrary functions on distributed data. The Lambda architecture is based on three layers aptly named as a batch layer, speed layer, and serving layer [79]. The batch layer is required to perform two tasks including the storage of data in a distributed manner and the computation of batch views for the distributed data for low latency. The speed layer utilizes an online technique to store and update the real-time views of the recent data which have not been considered by the batch layer. The serving layer is a specialized distributed database that integrates the data views provided by the batch and speed layers with an aim for real-time and online big data analytics in smart grids. The authors have integrated the capabilities of tools such as Hadoop, Spark SQL, Hive, Impala, etc., and depicted generalized, low latent, scalable, and robust results for smart grid big data analytics.

In [80], several challenges faced at each stage of performing big data analytics are presented. These challenges can be classified into three categories: data acquisition and handling, data processing, and system issues [81], [82].

In data acquisition and handling, the challenges are related to the competent presentation of heterogeneous data to reflect

TABLE 9. Comparison between different frameworks for big data analytics.

Features	Hadoop	Storm	Spark	Drill
Source Code	Open	Open	Open	Open
Complexity	Simple	Simple	Simple	Complex
Type of Processing	Batch Processing	Real-time Stream Processing	Real-time Stream Processing	Interactive Ad-hoc Querying
Latency	High	Low	Low	Low

the diversity, hierarchy, and granularity of data. Also, the raw datasets often contain redundancy that needs to be reduced along with data compression without deteriorating the information in the data. Data life cycle management is of utmost importance because of the availability of huge amounts of data and the current storage systems cannot store the massive data available at an unprecedented rate. Therefore, there needs to be a practical system where the data is analyzed on the go and for that, the stream processing framework using HDFS and Apache Spark has been proposed in this paper. The system challenges for analytics are faced with massive storage and high-speed processing. Furthermore, there are concerns about privacy and security since the data might contain personal information.

In data analytics, the challenges posed are that of huge data and the requirement of real-time processing. One of the solutions to these challenges could be approximate analytics providing approximate but real-time results. Mining on social media and customer feedback could present challenges as the data is generally unstructured.

Solving these challenges requires the use of large-scale parallel systems that further brings additional challenges such as energy management, scalability, and real-time collaboration. The energy usage of the large-scale parallel systems has been alarming due to massive data volume and analytics demand. Hence, system-wide energy management techniques should be utilized in big data system solutions.

In the smart grid discipline, a cloud-based platform project has been presented in [12] where the University of South California microgrid was deployed as a testbed to transform the electrical utility into a smart grid in the future.

The challenges and solutions to handle big data from smart grid units have been researched in academics and industrial centers. Solutions have also been implemented at the commercial level by a few utility companies. These utilities always strive to meet the goals of moving to a smarter grid to support distributed generation, distribution automation devices, providing new products and services, improving operational efficiency, and finally enhancing the system reliability. Some of the prominent industry efforts are described in the following:

A. ACCENTURE SOLUTION

Accenture proposed a system that uses grid observability to drive performance (Fig. 9) and to govern five distinct smart grid data classes such as operational data, non-operational data, meter usage data, event message data, and metadata [83]. All the classes of data should be treated and managed differently owing to their inherent characteristics and different sources. The architecture was aimed to overcome the challenges of corresponding the data collection infrastructure to the desired outcome, application of tools to manage massive-scaled data, and analysis of master data to benefit from smart grid potential. The commercial solution is proposed to discover the information through the components as shown in Figure 9.

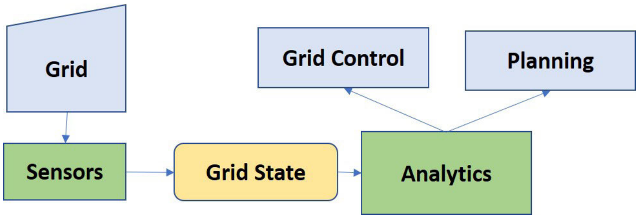


FIGURE 9. Using observability to ensure performance.

The provided solution explains the analytical aspects of the proposed architecture, however, it does not provide detailed information on the data treatment, management, and storage processes.

The Accenture architecture named Intelligent Network Data Enterprise (INDE) has the following components:

- The software layer in the architecture acts as a layer between the grid data sources and the current utility enterprise IT platforms. It aims to integrate the data from various sources to enhance the utility business operations and customer operations.
- The integration layer is prevalent to provide a unifying platform to the smart grid ecosystem products such as smart meters, communication lines, sensors, and other electrical network components.
- The visualization layer is provided to observe and monitor the different components in the grid. It also aims to recognize patterns in the raw data to correlate with different events and metadata.

The implemented solutions by Accenture at their clients' sites indicate their emphasis on the following five major application areas for smart meter data [84]:

- **Enhancing outage management:** The main goal of smart meter analytics has always been to enhance outage management. Outage management can be enhanced if the disturbances in the electric network are accurately predicted, localized, and restored by integrating the outage notifications, sectionalizing, and reclosing systems.

- **Power quality assessment:** The smart meter data can be used to monitor the quality of power at every point in the electrical distribution network. The fluctuations in the frequency and voltage can cause damage or failure to the electric equipment. The remote assessment of power quality can help utilities to investigate legitimate claims of customers saving field effort and time.
- **Protect customers and detect losses:** The system should protect the interests of all customers by detecting different losses including technical and non-technical losses in the electrical network. The non-technical losses occur when the customers tamper with electrical meter readings to reduce their bills. The integration of data from feeder meters and smart meters into the work management system will help utilities to identify electricity theft and investigate claims easily.
- **Renewable energy forecasting:** The generation of renewable energy is increasing in the grids. This calls for the proper management on part of the grid operator. Renewable energy is less predictable. However, accurate forecasting should be in place by integrating the data from smart meters and weather stations. This would help in the operational and investment decisions of utilities. With accurate forecasting, the grid operators can stabilize the supply and quality of power throughout the electrical distribution network.
- **Future market developments:** Long-term planning is required for balancing generation and load demand, flexible energy tariff planning, etc.

B. IBERDROLA

Big Data techniques are used to yield knowledge management solutions to control high turnover environments and to minimize the impact on call centers. Iberdrola has a part of its ambitious Digital Transformation Program [85] in the use of big data techniques. The company group targets to invest 4.8 billion euros in the digital transformation between 2019 and 2022 to boost the performance and conservation of its assets using data analytics and artificial intelligence. Digital analytics provides for creating an analytical environment to inspire knowledge that aids to maintain the three lines of business: Networks, Renewables, and Customers. Some examples of these applications are:

- Detection of non-technical losses and design of optimal time-of-use tariffs with the use of customer load curves to improve energy utilization [8]. The company has installed more than 11 million smart meters in Spain, generating 240 million registers every day. Big data techniques over an estimated volume of 90 billion registers per year are being used to improve revenue collection and to optimize energy use.
- Improvement of the operation and maintenance of the utility's assets expanding the availability of its generation plants. For example, in the U.S., Iberdrola is saving $3 million monthly by feeding wind turbine power generation data across multiple wind farms to develop curtailment optimization plans [86]. Iberdrola is leading a five-year project called Romeo, €16 million EU Horizon 2020 project, aiming at the reduction of the preservation cost of wind turbines using predictive machine learning algorithms, artificial intelligence, and cloud computing [87]. Utility's relationship with the customers can be transformed by the development of applications such as managing electricity consumption from mobile phones or scheduling electric vehicle charging [85]. Big Data techniques are also used to provide knowledge management solutions to command high turnover environments and prune the impact on call centers.

C. ITRON-TERADATA SOLUTION

Itron-Teradata architecture is established on active smart grid analytics (ASA) as depicted in Figure 10. As per the solution, the data warehouse actively provides strategies for the parallel ingestion of massive-scale data from varied sources and executes complex analytics for applications such as energy diversion detection, power quality, demand response, transformer load management, load forecast, customer profiling, etc. The data arrive triggering actions and activating workflows [88]–[90].

ASA is based upon the comprehensive Utility Logical Data Model (ULDM) of smart grids' data. The ASA solution helps the customers through self-service with insight on how to convert their usage to green energy, to make savings in energy and billing, etc. The solution assists the utilities to develop communication channels for customer-utility interaction and to invest in assets that boost customer experience. Also, the regulatory agencies benefit from the ASA solution with insights on the efficiency standards of operations, the percentage of energy from alternative sources, the fair pricing of energy, etc.

D. INTERNATIONAL BUSINESS MACHINES (IBM) SOLUTION FOR E.ON

Since 2013, IBM has worked on the smart metering infrastructure on the private cloud for E.ON with an aim to enhance the deployment and management of smart meters and to help incorporate renewable energy sources easily into the current grid [92]. The platform addresses the challenges of high data storage, low speed of report generation and analytics, etc. With the platform, customers have better control of the energy usage with information on their usage profile, on electricity tariff for the time of use, and on changes in consumption patterns when compared to their historical data. IBM intends the platform to be scalable with low startup and operational costs in order to provide for future growth. The platform has a high emphasis on ensuring the privacy of sensitive customer data, however, the data would be retained for a longer time to help with the emerging regulatory requirements in the future.

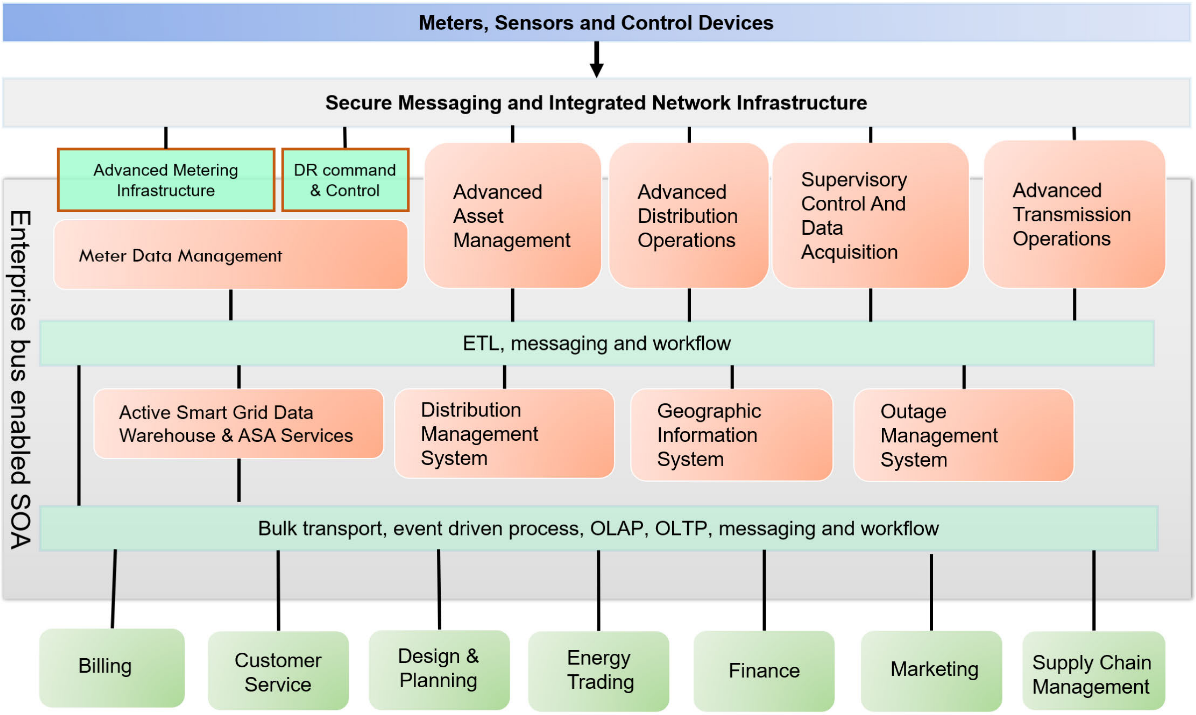

Meters, Sensors and Control Devices
Secure Messaging and Integrated Network Infrastructure
Enterprise bus enabled SOA
Advanced Metering Infrastructure
DR command & Control
Meter Data Management
Advanced Asset Management
Advanced Distribution Operations
Supervisory Control And Data Acquisition
Advanced Transmission Operations
ETL, messaging and workflow
Active Smart Grid Data Warehouse & ASA Services
Distribution Management System
Geographic Information System
Outage Management System
Bulk transport, event driven process, OLAP, OLTP, messaging and workflow
Billing
Customer Service
Design & Planning
Energy Trading
Finance
Marketing
Supply Chain Management

FIGURE 10. Service-oriented architecture of Itron-Teradata solution [91].

E. USA EXELON

Since 2014, Baltimore Gas and Electric (BGE) and Exelon have been working on a project by employing C3's cloud-based data processing platform to control the working of millions of smart meters installed in the regions of Chicago and Philadelphia Electric Company (PECO) utilities [93]. They have been successfully tapping the data from the smart meters with an aim to locate and avert energy theft. They employ machine learning algorithms to encode every rule of meter tampering and unbilled power delivery as these change over time. The algorithms also integrate various types of data from systems in place for the management of data from meters, outage prevention, user profiling, billing, and asset management. These applications led to the program of Business Intelligence Data Analytics (BIDA) and the solution of Data Analytics Platform (DAP). The solution supports the domains of business support, customer service, smart energy services, grid management, and AMI with a vision to assist future utilities, energy regulators, and customers.

F. KOREA ELECTRICAL POWER CORPORATION (KEPCO)

KEPCO launched two projects to use big data analytics on smart grids' data to improve demand management, and load forecasting and has been achieving considerable success in its goals ever since [94]. The first project helps customers to save electricity by comparing similar customers energy consumption data and allows KEPCO to prevent brownouts and manage load demand. The second project involves analyzing the business risks of blackouts, user complaints, weather changes, climate change statistics with the aid of social networking data, internet data, and complaints.

The companies do not explicitly describe their commercial solutions and do not release the information of the components of data management architecture in detail. However, noticing the potential of big data analytics to manage the demand-side response and user service, the utilities have now and again been cooperating with IT companies to tap the potential. This paper has also presented our proposed architecture aiming for the streams data processing to provide real-time information and visual analytics.

VI. APPLICATIONS OF BIG DATA ANALYTICS IN SMART GRID

This section discusses a few of the potential application areas which would avail from the big data analytics in the smart grid. It also details the previous application-based works and their proposed methodologies.

A. FAULT CLASSIFICATION AND IDENTIFICATION

The invention of the smart grid was driven by the need for clean and alternative forms of energy. The utilization of distributed energy sources in distribution grids brings the integration of renewable energy sources to reality. The microgrids allow for energy generation closer to load and hence, assist the improvement of power delivery and reduction in the power transmission losses. Furthermore, the microgrids can be used in islanded mode, and consequently, the loads can be protected from the damages resulting due to fluctuations in voltage and frequency [95].

The fluctuations of the energy produced by renewable energy sources bring uncertainty in the energy generation from distribution grids. Usually, Inverted Integrated Distribution Grids (IIDG) are used to improve the power quality in microgrids. However, these IIDGs have low inertia and hence if the faults caused in microgrids are not detected and cleared in short times, this is a huge threat to the microgrids. The classical approaches to fault identification and clearing [96] are based on the measurement of overcurrent and negative sequences of current. These approaches are not suited to microgrids due to their low current capacity. The statistical features are extracted using the wavelet transforms on the current measurements in the branches sampled by protective relays. The deep learning model is developed with the training data available on the statistical features to detect faults, classify them, and localize the faults in [97].

B. PREVENTIVE MAINTENANCE

The pieces of equipment of the power grid are vulnerable to failures and a robust plan for preventive maintenance of equipment, devices, etc. in the power grid can play a crucial part in reducing the probability of occurrence of failures in the power grid. Preventive maintenance can signal for and provide maintenance for equipment before these fail and hence, will avert major events and disruption of power supply for long periods. The integration of renewable energy sources at the distribution level of grids through microgrids supply clean energy. Nevertheless, the uncertainty of supply and fluctuations of frequency and voltage increase vulnerability to failure. It is required that the occurrence of failures is detected before failure and the clearance time is averted using preventive maintenance. Preventive maintenance is categorized into two types - time-based and condition-based. In time-based maintenance, the components are subjected to maintenance at periodic intervals of time irrespective of their condition. This approach does not utilize the service life of the components efficiently. Condition-based maintenance monitors the health of the components and draws a correlation between the current status and future faults of the components so that the future maintenance plans are scheduled [98]. One of the approaches to prognostic maintenance is the design of a proposed integrated fault detection system developed after analysis of the data from SCADA and Pole Mounted Auto Reclosers (PMARs) [99]. PMAR is a breaker that trips for intermittent fault currents and closes automatically to supply the power after a short duration of time nonetheless, it stays open for a permanent fault.

A reinforcement learning-based framework is proposed in [100]. The framework monitors the health of the equipment, models the degradation, and computes the remaining useful life of the grid components. The framework tested on a case study on the power grid performs with good approximation capability by using an ANN ensemble model. All of the data or subset of data from grid operations data, weather information, diagnostics data of the relay protection systems, galloping of power lines, fault tolerance current, and voltage signals have been used for the design of data-driven models for preventive maintenance in the power grids. Different machine learning models such as SVM [101], extreme learning machines [102], Long Short Term Memory (LSTM) [103], hybrid ensemble models [104], etc. are used to build data-driven models. The correlation between the actual faults that have occurred in the past and the features extracted from the data has been studied. These analyses models and studies are required to have high learning without iterative computations to converge faster, predict with higher accuracy and earliness. This would be an ideal solution for big data analysis for predictive maintenance.

C. TRANSIENT STABILITY ANALYSIS

Transient stability analysis (TSA) is performed to study the safe operation of the power grid. However, the challenges to the TSA these days are the integration of intermittent renewable energy sources at the distribution level, fluctuating demand of load, and deregulated energy market. The efficient approaches that extract information and patterns in the highly redundant records of big data are required for TSA. The techniques for TSA can be classified into automatic learning approaches, direct techniques, and time-domain techniques. Automatic learning approaches have edge over direct and time-domain techniques for real-world applications. The direct techniques [105] have demerits in the construction of energy functions for large-scale power systems whereas time-domain techniques are computationally inefficient for real-time applications [106].

Steady-state variables are used as features for TSA in [107] thus avoiding the use of time-domain simulation. The approach takes into account the size of the electrical network, the topology, the location of a fault, and operating status.

In [97], Yu *et al.* used time-series synchrophasor measurement data under different simulation contingency models to train the deep learning model of LSTM for online-assessment of transient stability status post-contingencies. Although the training of the TSA LSTM model was computationally expensive and time-consuming, the time adaptive nature and self-learning of temporal dependency by the LSTM model achieve better test accuracy and highly responsive time. Moreover, to reduce the training time, simpler models such as Extreme Learning Machines (ELMs) that are single-layer neural networks are used [108]. To address the uneven class distribution of power systems' data with a higher number of data points representing stability and a lower number of data points representing contingency condition, Baltas *et al.* proposed a response-based ensemble model of diverse ELM [109].

Rahmatian *et al.* worked on the implementation of transient stability assessment in real-time using characteristic features of voltage and current phasors from PMU data, Classification and Regression Trees (CART), and Multiregression Adaptive Regression Splines (MARS) models [110]. The models predict if a situation is stable or unstable using CART and applies MARS along with online TSA to indicate the level of severity of a contingency and instability of the system with high accuracy.

D. HEALTH MONITORING

Failure in crucial components of the power grid such as transformers, etc. will lead to brownouts or blackouts in the electrical grid network. It is crucial that the health of the electrical components in the grid is monitored. Classically, the monitoring system is based on a threshold mechanism that monitors different parameters and readings for different grid components.

The uncertainty and intermittent nature of renewable energy sources at the distribution level bring uncertainty in the life estimation of crucial components such as power transformers, etc. In [111], Aizpurua *et al.* proposed a probabilistic health monitoring framework for power transformers by using a probabilistic forecasting approach along with Monte Carlo-based Kalman-filtering techniques. The lifetime estimation of transformers in these models is adaptive as the dynamics of smart grids is propagated to the power transformers to determine the probabilistic thermal model and lifetime model.

There are different artificial intelligence-based approaches used for health monitoring using big data in smart grids. These include artificial neural networks [112], deep learning models [113], expert systems [114], fuzzy logic [115], [116], genetic algorithm [117], etc.

Mileta *et al.* analyzed the Mamdani model and Sugeno-model in the fuzzy expert system to compute the probability of occurrence of faults in the future and to determine the urgency of intervention or maintenance on the transformers based on their current condition [114]. The models utilized the online and offline data on historical and current conditions of transformers' age, lower oil level, frequency response analysis, oil temperature, insulation temperature, insulation degradation, polarization index, etc.

Hybrid models are utilized to overcome the shortcomings of single models. For instance, a health monitoring system was developed by Allen *et al.* for the health diagnostics of building automation systems and variable air valve units using a fuzzy logic model [118]. The fuzzy logic model detected anomalies in the operating conditions and generated fault signatures. The neural network-based model was used to classify the fault signatures into different faults. The monitoring of the health of the components at lower granularity ensures that the energy consumption observed at higher levels is reduced and finally helps for energy savings, and efficient monitoring.

E. POWER QUALITY MONITORING

When the frequency, magnitude, and waveforms of current and voltage are steady and within the prescribed limits, it is defined as power with high quality. Power quality also defines the performance and health of the smart grid components and the accuracy of utility metering. With the integration of non-linear sources of energy and power electronics-based devices, the harmonics appear in the voltage or current waves and it is essential that the power quality of the supply is maintained for the health of the devices, sensors, and appliances connected to the electric network. The power quality issues are currently addressed using dynamic voltage regulator, inverter, power quality monitoring, static synchronous compensator, unified power quality conditioner, etc.

Power quality monitoring is performed using conventional approaches through the integration of SCADA, AMI, etc., or by using artificial intelligence-based approaches. Multiple machine learning modeling such as Support Vector Machines [119], decision trees [120], Bayesian networks [121], k-Nearest neighbors [122], etc. have been employed for monitoring power quality disturbances.

Wang *et al.* employed deep learning in each of the stages of power quality classification i.e. signal analysis, feature selection, and classification [123]. They used a deep convolutional neural network consisting of the 1-D convolutional layer along with pooling and batch normalization layers for the automatic extraction of features from disturbance samples. They presented evidence in terms of accuracy and training time cost that deep Convolutional Neural Networks (CNNs) performs better for applications of automatic power quality classification when compared to other deep learning models such as gated recurrent networks, long short-term memory, ResNet50, and stacked auto-encoders. To overcome the non-distributed computing and feature extraction-based power quality classification, Chen *et al.* presented an integrated solution based on deep belief networks for real-time and distributed power quality disturbance analysis [124]. The developed models proved to have higher accuracy and more robustness on distributed platforms, however, the training time is also very high.

F. TOPOLOGY IDENTIFICATION

The topology identification problem in the smart grid includes the identification of the structure of power distribution network, identification of customer phase connectivity, and associating a customer with a transformer at the distribution level. The identification of phase connectivity is crucial to the analysis of distribution system including distributed network estimation, power flow analysis, optimal power flow, distribution network reconfiguration and restoration, load balancing, etc. Topology identification could be possible using specialized sensors such as micro-synchrophasors, phase meters, etc. However, using a special sensor for each customer is impractical and expensive. There are many approaches developed to identify topology using the data made available by the current infrastructure such as AMI, SCADA, GIS, Outage Management System (OMS), and besides machine learning approaches have been developed using training data on field validated phase connectivity.

Voltage time series data have been utilized to extract feature vectors after the application of principal component analysis and the authors have suggested that the voltage data are predictive of the phase connectivity [125]. Afterward, the k-means clustering approach has been applied to cluster the different customers into the three different phases for phase connectivity identification. The innovative model was tested on a real distribution feeder and the test accuracy was about 90%.

G. ENERGY THEFT

Energy theft is defined as the act of changing the electricity consumption reading in order to reduce the bill through physical approaches such as bypassing the smart meter, tampering with meters, cyber approaches such as hacking into a smart meter to change the energy consumption values, etc. Data-driven approaches are currently applied to identify energy theft and these approaches are classified into different types depending on the type of available data. When the smart meter data were not available, machine learning models such as fuzzy clustering [126], SVM [127], etc. have been applied to the annual energy consumption, and credit scores were determined to identify theft detection. When the smart meter data and theft cases data are available, then supervised machine learning models such as neural networks [128], deep learning models [129], etc. can be applied. Usually, energy theft cases are not available or disclosed for research. In such cases, the energy theft identification can be performed using smart meter data, and network topology information can be determined using state-estimation based approaches [130], and other anomaly detection techniques [131].

H. RENEWABLE ENERGY FORECASTING

The renewable energy (RE) sources are environment-friendly, clean, and unlimited replenishable sources of energy. Nevertheless, the uncertain and intermittent behavior of the supply poses many challenges in the generation of power using renewable energy sources. The reliable and accurate RE forecasting helps in the grid operations, load management, planning of capacity, scheduling of generation, regulation of energy, etc. Multiple approaches including physical models, statistical models, machine learning approach, hybrid models, etc. have been used to date for renewable energy forecasting.

Physical models include the simulation of geographic characteristics of an area. These models utilize weather forecasting, geographical information, meteorological information, etc. Physical methods require huge computational resources, are less accurate and also, are not suitable for short-term forecasting. Statistical models apply mathematical modeling to recognize the patterns in time-series data of renewable energy sources. The methods such as Auto-Regressive Moving Average [132], Kalman Filters [133], Markov models [134], etc. have been applied previously. With the widespread popularity of machine learning models, these have been applied reliably on renewable energy forecasting. The machine learning algorithms include models such as linear regression [135], decision trees regression [136], multi-layer perceptrons [137], support vector machines [138], etc. Owing to the inherent intermittent and non-linear nature of renewable energy supply, deep learning models have been found to be extremely efficient and effective [139]–[141]. Deep learning models such as deep belief networks, autoencoders, convolutional neural networks, long short-term memory, deep learning ensemble models, etc. have been applied to predict renewable energy from sources. The patterns of temporal changes in renewable energy are captured in the parameters of the deep layers. The high accuracy of renewable energy forecast will help in the planning, and development of reliable, and resilient integration of the sources in the distribution grids through microgrids.

VII. CONCLUSION

The review paper presents a comprehensive study of technologies and techniques for big data processing. These technologies are also applicable when dealing with data from smart grids.

Apache Hadoop is the most suitable platform for big data analytics when time is not a crucial consideration and when batch processing of large datasets of offline data is required. Apache Storm is best suited for real-time stream processing of real-time data. Apache Spark is suitable for both batch processing and stream processing. Whereas Apache Drill is suited for data-intensive applications requiring interactive processing of massive-scaled data. It can be concluded that there is a need for a big data analytics platform that utilizes different types of technologies for real-time solutions in smart grids and also a middle-ware software is required to integrate all of the technologies with reliability and stability. Enterprises dealing with big data are required to address the challenges of security, privacy, data handling, etc. Before any big data techniques are employed in the smart grid, it is always necessary to consider steps such as data acquisition, data management, analytics, and visualization along with the requirement of the real-time processing of data. This paper has suggested the big data analytics technologies for the smart grid to make the real-time processing of data a reality.

The paper presents that Apache Spark is more suitable for both batch and real-time processing in smart grids. However, Apache Spark does not provide its own distributed storage system and hence, it requires a storage system provided by a third party. Hence, the paper advises the installation of Apache Spark on top of Hadoop so that the advanced analytic applications provided by Spark can be used along with the parallel distributed storage system of HDFS. The big data analytics on the smart grids' data will achieve proper management of renewable energy sources in the generation and the distribution side.

REFERENCES

[1] Y. Zhang, R. Yu, M. Nekovee, Y. Liu, S. Xie, and S. Gjessing, "Cognitive machine-to-machine communications: Visions and potentials for the smart grid," *IEEE Netw.*, vol. 26, no. 3, pp. 6–13, May 2012.

[2] T.-H. Dang-Ha, R. Olsson, and H. Wang, "The role of big data on smart grid transition," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 33–39.

[3] M. K. Saggi and S. Jain, "A survey towards an integration of big data analytics to big insights for value-creation," *Inf. Process. Manage.*, vol. 54, no. 5, pp. 758–790, Sep. 2018.

[4] D. B. Rawat and C. Bajracharya, "Cyber security for smart grid systems: Status, challenges and perspectives," in *Proc. SoutheastCon*, Apr. 2015, pp. 1–6.

[5] J. N. Bharothu, M. Sridhar, and R. S. Rao, "A literature survey report on smart grid technologies," in *Proc. Int. Conf. Smart Electr. Grid (ISEG)*, Sep. 2014, pp. 1–8.

[6] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 5–20, 1st Quart., 2013.

[7] J. Joy, D. E. Jasmin, and V. R. John, "Challenges of smart grid," *Int. J. Adv. Res. Electr., Electron. Instrum. Eng.*, vol. 2, no. 3, pp. 976–981, 2013.

[8] Expansión. *Iberdrola Bets on 'Big Data' to Manage Its Electricity Networks*. Accessed: Mar. 7, 2018. [Online]. Available: http://www.expansion.com/empresas/energia/2015/04/23/5538cc0e22601de3448b458e.html

[9] P. Siano, "Demand response and smart grids—A survey," *Renew. Sustain. Energy Rev.*, vol. 30, pp. 461–478, Feb. 2014.

[10] F. Mwasilu, J. J. Justo, E.-K. Kim, T. D. Do, and J.-W. Jung, "Electric vehicles and smart grid interaction: A review on vehicle to grid and renewable energy sources integration," *Renew. Sustain. Energy Rev.*, vol. 34, pp. 501–516, Jun. 2014.

[11] Accenture. *Unlocking the Value of Analytics*. Accessed: Jul. 20, 2020. [Online]. Available: https://www.accenture.com/t20171213T064437Z__w__/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_9/Accenture-Grid-Analytics-Report-Digitally-Enabled-Grid.pdf

[12] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna, "Cloud-based software platform for big data analytics in smart grids," *Comput. Sci. Eng.*, vol. 15, no. 4, pp. 38–47, Jul. 2013.

[13] R. Mahmud, R. Vallakati, A. Mukherjee, P. Ranganathan, and A. Nejadpak, "A survey on smart grid metering infrastructures: Threats and solutions," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, May 2015, pp. 386–391.

[14] Y.-J. Kim, M. Thottan, V. Kolesnikov, and W. Lee, "A secure decentralized data-centric information infrastructure for smart grid," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 58–65, Nov. 2010.

[15] Ausgrid. *Ausgrid Average Electricity Use*. Accessed: Sep. 26, 2020. [Online]. Available: https://data.gov.au/data/organization/ausgrid

[16] Irish Social Science Data Archive. *Electricity Smart Meter Data*. Accessed: Sep. 26, 2020. [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[17] Cornell University. *Smart Meter Data*. Accessed: Jul. 26, 2019. [Online]. Available: http://buildingdashboard.net/cornell/#/cornell

[18] École Polytechnique Fédérale de Lausanne (Switzerland). *Smart Grid Data*. Accessed: Sep. 26, 2020. [Online]. Available: http://nanotera-stg2.epfl.ch/data/

[19] Electric Reliability Council of Texas. *Grid Information Load Data*. Accessed: Sep. 26, 2020. [Online]. Available: http://www.ercot.com/gridinfo/load

[20] North American SynchroPhasor Initiative. *PMU Data*. Accessed: Mar. 1, 2020. [Online]. Available: https://www.naspi.net/PmuRegistry/#

[21] Pecan Street Inc. *Pecan Street Dataport*. Accessed: Mar. 1, 2020. [Online]. Available: https://dataport.pecanstreet.org/

[22] Pennsylvania-New Jersey-Maryland Interconnection. *PJM Data*. Accessed: Sep. 26, 2020. [Online]. Available: https://www.pjm.com/markets-and-operations/data-dictionary.aspx

[23] IEEE. *Intelligent Data Mining and Analysis Data Sets*. Accessed: Jul. 26, 2019. [Online]. Available: https://site.ieee.org/psace-idma/data-sets/

[24] UC Berkeley Campus. *Smart Grid and Building Consumption Data*. Accessed: Mar. 1, 2020. [Online]. Available: https://us.pulseenergy.com/UniCalBerkeley/dashboard/#/overview

[25] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Bus. Horizons*, vol. 60, no. 3, pp. 293–303, May 2017.

[26] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[27] A. Mohamed, S. S. Refaat, and H. Abu-Rub, "A review on big data management and decision-making in smart grid," *Power Electron. Drives*, vol. 4, no. 1, pp. 1–13, Jun. 2019.

[28] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019.

[29] V. Rajaraman, "Big data analytics," *Resonance*, vol. 21, pp. 695–716, Sep. 2016.

[30] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: A review," *Energy Informat.*, vol. 1, no. 1, p. 8, Aug. 2018.

[31] J. Sessa and D. Syed, "Techniques to deal with missing data," in *Proc. 5th Int. Conf. Electron. Devices, Syst. Appl. (ICEDSA)*, Dec. 2016, pp. 1–4.

[32] I. Nusrat and S.-B. Jang, "A comparison of regularization techniques in deep neural networks," *Symmetry*, vol. 10, no. 11, p. 648, Nov. 2018.

[33] K. Wadhwa, "Byte: Big data roadmap and cross-disciplinary community for addressing societal externalities," in *Proc. Eur. Data Forum*, 2014, pp. 108–116.

[34] N. Tatbul, "Streaming data integration: Challenges and opportunities," in *Proc. IEEE 26th Int. Conf. Data Eng. Workshops (ICDEW)*, Long Beach, CA, USA, Mar. 2010, pp. 155–158.

[35] R. M. A. Velásquez and J. V. M. Lara, "Principal components analysis and adaptive decision system based on fuzzy logic for power transformer," *Fuzzy Inf. Eng.*, vol. 9, no. 4, pp. 493–514, Dec. 2017.

[36] J. C. Palomares-Salas, J. J. Gonzalez de la Rosa, A. Aguera-Perez, and J. M. Sierra-Fernandez, "Smart grids power quality analysis based in classification techniques and higher-order statistics: Proposal for photovoltaic systems," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2015, pp. 2955–2959.

[37] E. De Santis, L. Livi, A. Sadeghian, and A. Rizzi, "Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification," *Neurocomputing*, vol. 170, pp. 368–383, Dec. 2015.

[38] A. E. Lazzaretti, D. M. J. Tax, H. Vieira Neto, and V. H. Ferreira, "Novelty detection and multi-class classification in power distribution voltage waveforms," *Expert Syst. Appl.*, vol. 45, pp. 322–330, Mar. 2016.

[39] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, "Load forecasting, dynamic pricing and DSM in smart grid: A review," *Renew. Sustain. Energy Rev.*, vol. 54, pp. 1311–1322, Feb. 2016.

[40] Y. Weng, R. Negi, C. Faloutsos, and M. D. Ilic, "Robust data-driven state estimation for smart grid," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1956–1967, Jul. 2017.

[41] Y. Cai and M.-Y. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in smart grids," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2009, pp. 1–6.

[42] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.

[43] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.

[44] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, Oct. 2015.

[45] K. Chahine, K. E. K. Drissi, C. Pasquier, K. Kerroum, C. Faure, T. Jouannet, and M. Michou, "Electric load disaggregation in smart metering using a novel feature extraction method and supervised classification," *Energy Procedia*, vol. 6, pp. 627–632, Jan. 2011.

[46] A. I. Saleh, A. H. Rabie, and K. M. Abo-Al-Ez, "A data mining based load forecasting strategy for smart electrical grids," *Adv. Eng. Informat.*, vol. 30, no. 3, pp. 422–448, Aug. 2016.

[47] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1005–1016, Jun. 2016.

[48] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy Buildings*, vol. 147, pp. 77–89, Jul. 2017.

[49] K.-L. Zhou, S.-L. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 24, pp. 103–110, Aug. 2013.

[50] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[51] X. Yang, P. Zhao, X. Zhang, J. Lin, and W. Yu, "Toward a Gaussian-mixture model-based detection scheme against data integrity attacks in the smart grid," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 147–161, Feb. 2017.

[52] I. Khan, J. Z. Huang, and K. Ivanov, "Incremental density-based ensemble clustering over evolving data streams," *Neurocomputing*, vol. 191, pp. 34–43, May 2016.

[53] X. Yuan, "An improved apriori algorithm for mining association rules," in *Proc. AIP Conf.*, 2017, Art. no. 080005.

[54] M. K. Najafabadi, M. N. Mahrin, S. Chuprat, and H. M. Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," *Comput. Hum. Behav.*, vol. 67, pp. 113–128, Feb. 2017.

[55] A. A. Munshi and A.-R. M. Yasser, "Big data framework for analytics in smart grids," *Electr. Power Syst. Res.*, vol. 151, pp. 369–380, Oct. 2017.

[56] A. El Khaouat and L. Benhlima, "Big data based management for smart grids," in *Proc. Int. Renew. Sustain. Energy Conf. (IRSEC)*, Nov. 2016, pp. 1044–1047.

[57] D. Vohra, *Practical Hadoop Ecosystem*, 1st ed. New York, NY, USA: Apress, 2016.

[58] A. Moniruzzaman and S. A. Hossain, "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison," *Int. J. Database Theory Appl.*, vol. 6, no. 4, pp. 1–13, 2013.

[59] R. Kumar, B. B. Parashar, S. Gupta, Y. Sharma, and A. Gupta, "Apache Hadoop, NoSQL and NewSQL solutions of big data," *Int. J. Advance Found. Res. Sci. Eng.*, vol. 1, no. 6, pp. 28–36, 2014.

[60] G. Liu, W. Zhu, C. Saunders, F. Gao, and Y. Yu, "Real-time complex event processing and analytics for smart grid," *Procedia Comput. Sci.*, vol. 61, pp. 113–119, Jan. 2015.

[61] C. L. Stimmel, *Big Data Analytics Strategies for the Smart Grid*. Boca Raton, FL, USA: Auerbach, Apr. 2016.

[62] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google file system," *ACM SIGOPS Operating Syst. Rev.*, vol. 37, no. 5, pp. 29–43, Dec. 2003.

[63] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *J. Big Data*, vol. 2, no. 1, p. 24, Nov. 2015.

[64] K. Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler, "The Hadoop distributed file system," in *Proc. IEEE 26th Symp. MSST*, May 2010, pp. 1–10.

[65] D. Borthakur *et al.*, "HDFS architecture guide," *Hadoop Apache Project*, vol. 53, nos. 1–13, p. 2, 2008.

[66] A. Bahga and V. Madisetti, *Big Data Science & Analytics: A Hands-on Approach*. Johns Creek, GA, USA: VPT, 2016.

[67] G. Turkington, T. Deshpande, and S. Karanth, *Hadoop: Data Processing and Modelling*. Birmingham, U.K.: Packt, 2016.

[68] V. Kalavri and V. Vlassov, "MapReduce: Limitations, optimizations and open issues," in *Proc. 12th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Jul. 2013, pp. 1031–1038.

[69] A. S. Foundation, *Apache Storm*. Accessed: Jul. 26, 2018. [Online]. Available: http://storm.apache.org/

[70] B. Peng, M. Hosseini, Z. Hong, R. Farivar, and R. Campbell, "R-storm: Resource-aware scheduling in storm," in *Proc. 16th Annu. Middleware Conf.* New York, NY, USA: ACM, 2015, pp. 149–161.

[71] Apache Software Foundation. *Apache ZooKeeper*. Accessed: Jul. 26, 2018. [Online]. Available: http://zookeeper.apache.org/

[72] B. Chambers and M. Zaharia, *Spark: The Definitive Guide: Big Data Processing Made Simple*. Newton, MA, USA: O'Reilly Media, 2018.

[73] S. R., B. Ganesh H. B., S. Kumar S., P. Poornachandran, and K. P. Soman, "Apache spark a big data analytics platform for smart grid," *Procedia Technol.*, vol. 21, pp. 171–178, Jan. 2015.

[74] S. Sakr, "Big data processing stacks," *IT Prof.*, vol. 19, no. 1, pp. 34–41, Jan. 2017.

[75] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," *HotCloud*, vol. 10, no. 10, p. 95, Jun. 2010.

[76] H. Zhiwei, G. Tian, Z. Huaving, H. Xu, C. Junwei, H. Ziheng, Y. Senjing, and Z. Zhengguo, "Transient power quality assessment based on big data analysis," in *Proc. China Int. Conf. Electr. Distrib. (CICED)*, Sep. 2014, pp. 1308–1312.

[77] M. Mayilvaganan and M. Sabitha, "A cloud-based architecture for big-data analytics in smart grid: A proposal," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2013, pp. 1–4.

[78] A. A. Munshi and Y. A.-R.-I. Mohamed, "Data lake lambda architecture for smart grids big data analytics," *IEEE Access*, vol. 6, pp. 40463–40471, 2018.

[79] Nathan Marz. *Big Data Lambda Architecture*. Accessed: Jul. 14, 2020. [Online]. Available: http://www.databasetube.com/database/big-data-lambda-architecture

[80] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[81] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, and J. Han, "Challenges and opportunities with big data. A community white paper developed by leading researchers across the united states," Comput. Res. Assoc., Washington, DC, USA, Tech. Rep. 1, Mar. 2012.

[82] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012.

[83] D. Haak, "Achieving high performance in smart grid data management," Accenture, Dublin, Ireland, Tech. Rep. WSS141, 2010.

[84] Accenture. *Applying Smart Meter Analytics*. Accessed: Jul. 14, 2020. [Online]. Available: https://www.accenture.com/nl-en/blogs/insights/did-you-consider-these-ways-to-apply-smart-meter-analytics

[85] S. A. Iberdrola. *At the Forefront of Digital Transformation*. Accessed: Mar. 7, 2018. [Online]. Available: https://www.iberdrola.com/about-us/utility-of-the-future/digital-transf ormation

[86] Wood Mackenzie Power Renewables. *Big Data Is Boosting Power Production, Reducing Downtime Across Wind Fleets*. Accessed: Jul. 28, 2020. [Online]. Available: https://www.greentechmedia.com/articles/read/big-data-is-boosting-power-production-reducing-downtime-across-wind-fleets

[87] S. A. Iberdrola. *Romeo Project Lands in East Anglia One and Wikinger*. Accessed: Mar. 7, 2018. [Online]. Available: https://www.iberdrola.com/top-stories/iberdrola-shares-with-you/romeo-project

[88] S. Johnson, "Openway demand response: Maximizing value and efficiency in energy delivery," Itron, Liberty Lake, WA, USA, Tech. Rep. 101025WP-01, 2010.

[89] L. Hogg, "Business intelligence for enterprise energy management," Itron, Liberty Lake, WA, USA, Tech. Rep. 100710WP-02, 2007.

[90] S. Moore, "Key features of meter data management systems," Itron, Liberty Lake, WA, USA, Tech. Rep. 100910WP-01, 2008.

[91] S. Moore and S. Butler, "Active smart grid analytics: Maximizing your smart grid investment," Itron, Liberty Lake, WA, USA, Tech. Rep. 100961WP-01, 2009.

[92] International Business Machines Corporation. *E.ON and IBM Deliver Innovative Service Offerings to Customers With New Smart Energy Solutions*. Accessed: Aug. 7, 2018. [Online]. Available: https://www-03.ibm.com/press/us/en/pressrelease/41921.wss

[93] Wood Mackenzie Power Renewables. *C3's Tom Siebel Opens Up About His Secretive Firm's Smart Grid Data Analytics*. Accessed: Jul. 7, 2020. [Online]. Available: https://www.greentechmedia.com/articles/read/c3-energy-unveils-first-big-test-of-smart-grid-data-analytics#gs.IRv0C1Y

[94] Smart Energy International. *KEPCO Pilots Big Data Projects for AMI and Customer Service Systems*. Accessed: Aug. 7, 2018. [Online]. Available: https://www.smart-energy.com/regional-news/asia/kepco-pilots-big-data-projects-for-ami-and-customer-service-systems/

[95] Z. Chen, X. Pei, M. Yang, L. Peng, and P. Shi, "A novel protection scheme for inverter-interfaced microgrid (IIM) operated in islanded mode," *IEEE Trans. Power Electron.*, vol. 33, no. 9, pp. 7684–7697, Sep. 2018.

[96] A. Zainab, S. S. Refaat, D. Syed, A. Ghrayeb, and H. Abu-Rub, "Faulted line identification and localization in power system using machine learning techniques," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2975–2981.

[97] J. J. Q. Yu, Y. Hou, A. Y. S. Lam, and V. O. K. Li, "Intelligent fault detection scheme for microgrids with wavelet-based deep neural networks," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1694–1703, Mar. 2019.

[98] B. Liu, S. Wu, M. Xie, and W. Kuo, "A condition-based maintenance policy for degrading systems with age- and state-dependent operating cost," *Eur. J. Oper. Res.*, vol. 263, no. 3, pp. 879–887, Dec. 2017.

[99] X. Wang, S. M. Strachan, S. D. J. McArthur, and J. D. Kirkwood, "Automatic analysis of pole mounted auto-recloser data for fault diagnosis and prognosis," in *Proc. 18th Int. Conf. Intell. Syst. Appl. Power Syst. (ISAP)*, Sep. 2015, pp. 1–6.

[100] R. Rocchetta, L. Bellani, M. Compare, E. Zio, and E. Patelli, "A reinforcement learning framework for optimal operation and maintenance of power grids," *Appl. Energy*, vol. 241, pp. 291–301, May 2019.

[101] G. Peng, S. Tang, Z. Lin, and Y. Zhang, "Applications of fuzzy multilayer support vector machines in fault diagnosis and forecast of electric power equipment," in *Proc. IEEE 2nd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2017, pp. 457–461.

[102] M. Rafiei, T. Niknam, J. Aghaei, M. Shafie-Khah, and J. P. S. Catalao, "Probabilistic load forecasting using an improved wavelet neural network trained by generalized extreme learning machine," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6961–6971, Nov. 2018.

[103] S. Zhang, Y. Wang, M. Liu, and Z. Bao, "Data-based line trip fault prediction in power systems using LSTM networks and SVM," *IEEE Access*, vol. 6, pp. 7675–7686, 2018.

[104] C. Hu, "Ensemble feature learning-based event classification for cyber-physical security of the smart grid," M.S. thesis, Dept. Inf. Syst. Eng., Concordia Univ., Montreal, QC, Canada, 2019.

[105] D. P. Wadduwage, C. Q. Wu, and U. D. Annakkage, "Power system transient stability analysis via the concept of Lyapunov exponents," *Electr. Power Syst. Res.*, vol. 104, pp. 183–192, Nov. 2013.

[106] S. Zadkhast, J. Jatskevich, and E. Vaahedi, "A multi-decomposition approach for accelerated time-domain simulation of transient stability problems," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2301–2311, Sep. 2015.

[107] C. He, L. Guan, and W. Mo, "A method for transient stability assessment based on pattern recognition," in *Proc. Int. Conf. Smart Grid Clean Energy Technol. (ICSGCE)*, Oct. 2016, pp. 343–347.

[108] L. Zhang, X. Hu, P. Li, F. Shi, and Z. Yu, "ELM model for power system transient stability assessment," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017, pp. 5740–5744.

[109] G. N. Baltas, C. Perales-Gonzalez, P. Mazidi, F. Fernandez, and P. Rodriguez, "A novel ensemble approach for solving the transient stability classification problem," in *Proc. 7th Int. Conf. Renew. Energy Res. Appl. (ICRERA)*, Oct. 2018, pp. 1282–1286.

[110] M. Rahmatian, Y. C. Chen, A. Palizban, A. Moshref, and W. G. Dunford, "Transient stability assessment via decision trees and multivariate adaptive regression splines," *Electr. Power Syst. Res.*, vol. 142, pp. 320–328, Jan. 2017.

[111] J. I. Aizpurua, U. Garro, E. Muxika, M. Mendicute, I. P. Gilbert, B. G. Stewart, S. D. J. McArthur, and B. Lambert, "Probabilistic power transformer condition monitoring in smart grids," in *Proc. 6th Int. Adv. Res. Workshop Transformers (ARWtr)*, Oct. 2019, pp. 42–47.

[112] W. L. Woon, Z. Aung, and A. El-Hag, "Intelligent monitoring of transformer insulation using convolutional neural networks," in *Proc. Int. Workshop Data Anal. Renew. Energy Integr.* New York, NY, USA: Springer, 2018, pp. 127–136.

[113] H. F. Cindy Tan, W. Lok Woo, A. Sharma, T. Logenthiran, and D. S. Kumar, "Study of smart condition monitoring using deep neural networks with dropouts and cross-validation," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT Asia)*, May 2019, pp. 3965–3969.

[114] M. Žarković and Z. Stojković, "Analysis of artificial intelligence expert systems for power transformer condition monitoring and diagnostics," *Electr. Power Syst. Res.*, vol. 149, pp. 125–136, Aug. 2017.

[115] N. G. Chothani, M. B. Raichura, D. D. Patel, and K. D. Mistry, "Real-time monitoring & protection of power transformer to enhance smart grid reliability," in *Proc. IEEE Electr. Power Energy Conf. (EPEC)*, Oct. 2018, pp. 1–6.

[116] N. Mahmood and S. Yadav, "An enhanced MPPT technique by using fuzzy logic controller," *J. Multimedia Technol. Recent Advancements*, vol. 6, no. 2, pp. 1–10, 2019.

[117] G. Gui, H. Pan, Z. Lin, Y. Li, and Z. Yuan, "Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection," *KSCE J. Civil Eng.*, vol. 21, no. 2, pp. 523–534, Feb. 2017.

[118] W. H. Allen, A. Rubaai, and R. Chawla, "Fuzzy neural network-based health monitoring for HVAC system variable-air-volume unit," *IEEE Trans. Ind. Appl.*, vol. 52, no. 3, pp. 2513–2524, May 2016.

[119] A. A. Abdoos, P. K. Mianaei, and M. R. Ghadikolaei, "Combined VMD-SVM based feature selection method for classification of power quality events," *Appl. Soft Comput.*, vol. 38, pp. 637–646, Jan. 2016.

[120] P. D. Achlerkar, S. R. Samantaray, and M. S. Manikandan, "Variational mode decomposition and decision tree based detection and classification of power quality disturbances in grid-connected distributed generation system," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3122–3132, Jul. 2018.

[121] Y. Luo, K. Li, Y. Li, D. Cai, C. Zhao, and Q. Meng, "Three-layer Bayesian network for classification of complex power quality disturbances," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3997–4006, Sep. 2018.

[122] R. Zhu, X. Gong, S. Hu, and Y. Wang, "Power quality disturbances classification via fully-convolutional siamese network and k-Nearest neighbor," *Energies*, vol. 12, no. 24, p. 4732, Dec. 2019.

[123] S. Wang and H. Chen, "A novel deep learning method for the classification of power quality disturbances using deep convolutional neural network," *Appl. Energy*, vol. 235, pp. 1126–1140, Feb. 2019.

[124] Z. Chen, M. Li, T. Ji, and Q. Wu, "Real-time recognition of power quality disturbance-based deep belief network using embedded parallel computing platform," *IEEJ Trans. Electr. Electron. Eng.*, vol. 15, no. 4, pp. 519–526, Apr. 2020.

[125] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase identification in electric power distribution systems by clustering of smart meter data," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 259–265.

[126] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Sci. Technol.*, vol. 19, no. 2, pp. 105–120, Apr. 2014.

[127] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Support vector machine based data classification for detection of electricity theft," in *Proc. IEEE/PES Power Syst. Conf. Exposit.*, Mar. 2011, pp. 1–8.

[128] H. Huang, S. Liu, and K. Davis, "Energy theft detection via artificial neural networks," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Eur. (ISGT-Europe)*, Oct. 2018, pp. 1–6.

[129] D. Yao, M. Wen, X. Liang, Z. Fu, K. Zhang, and B. Yang, "Energy theft detection with energy privacy preservation in the smart grid," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7659–7669, Oct. 2019.

[130] M. Wen, D. Yao, B. Li, and R. Lu, "State estimation based energy theft detection scheme with privacy preservation in smart grid," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[131] S.-C. Yip, W.-N. Tan, C. Tan, M.-T. Gan, and K. Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids," *Int. J. Electr. Power Energy Syst.*, vol. 101, pp. 189–203, Oct. 2018.

[132] S. N. Singh and A. Mohapatra, "Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting," *Renew. Energy*, vol. 136, pp. 758–768, Jun. 2019.

[133] A. Samalot, M. Astitha, J. Yang, and G. Galanis, "Combined Kalman filter and universal kriging to improve storm wind speed predictions for the northeastern United States," *Weather Forecasting*, vol. 34, no. 3, pp. 587–601, Jun. 2019.

[134] F. O. Hocaoglu and F. Serttas, "A novel hybrid (Mycielski-Markov) model for hourly solar radiation forecasting," *Renew. Energy*, vol. 108, pp. 635–643, Aug. 2017.

[135] M. Abuella and B. Chowdhury, "Solar power probabilistic forecasting by using multiple linear regression analysis," in *Proc. SoutheastCon*, Apr. 2015, pp. 1–5.

[136] L. Cai, J. Gu, J. Ma, and Z. Jin, "Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees," *Energies*, vol. 12, no. 1, p. 159, Jan. 2019.

[137] R. Azimi, M. Ghayekhloo, and M. Ghofrani, "A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting," *Energy Convers. Manage.*, vol. 118, pp. 331–344, Jun. 2016.

[138] A. Zendehboudi, M. A. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: A review," *J. Cleaner Prod.*, vol. 199, pp. 272–285, Oct. 2018.

[139] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Convers. Manage.*, vol. 198, Oct. 2019, Art. no. 111799.

[140] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep LSTM-RNN," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2727–2740, Jul. 2019.

[141] D. Syed, S. S. Refaat, H. Abu-Rub, O. Bouhali, A. Zainab, and L. Xie, "Averaging ensembles model for forecasting of short-term load in smart grids," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2931–2938.

DABEERUDDIN SYED (Graduate Student Member, IEEE) received the B.E. degree in electronics and electrical engineering from University College, Osmania University, India, in 2013, and the M.Sc. degree in data science and engineering from Hamad Bin Khalifa University, Qatar, in 2018. He is currently pursuing the Ph.D. degree in electrical engineering (E.E.) with Texas A&M University (TAMU), College Station, TX, USA. He has a work experience of three years in the industry. He has worked as a Teaching Assistant in electrical circuits and learning from data. His current research interests include smart grids, big data analytics, load forecasting, and distributed computing.

AMEEMA ZAINAB (Graduate Student Member, IEEE) received the bachelor's degree in electronics and communication engineering from Osmania University, Hyderabad, India, in 2013, and the M.Sc. degree in data science and engineering from Hamad Bin Khalifa University (HBKU), Qatar. She is currently pursuing the Ph.D. degree in electrical engineering (E.E.) with Texas A&M University (TAMU), College Station, TX, USA. She has worked as an Analytics Professional supporting Audit in Data Analytics with Deloitte Touche LLP, Hyderabad, India, for a period of three years. She is also a Base SAS Certified Professional. Her research interests include data analytics, artificial intelligence, and big data management in the smart grids.

ALI GHRAYEB (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Arizona, USA. He is a Professor in the Department of Electrical and Computer Engineering (ECEN), TAMU-Q, Qatar. He has worked as a Professor in the Department of ECEN at Concordia University, Canada. His primary research interests include massive MIMO, wireless communications, physical layer security, and visible light communications. He has served on the editorial board of several IEEE and non-IEEE journals.

SHADY S. REFAAT (Senior Member, IEEE) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from Cairo University, Egypt, in 2013, 2007, and 2002, respectively. He acquired the industrial work experience of over 12 years. He worked as a Senior Electrical Engineer, a Team Leader, and a Design Engineer on several engineering projects. He currently works as an Associate Research Scientist with TAMU at Qatar (TAMU-Q). He is also serving as a Member of the Smart Grid Center Qatar (SGC-Q). He has published over 105 research articles. His research interests include big data, condition monitoring, electrical machines, energy management systems, fault detection, power systems and reliability, fault-tolerant systems, and smart grid.

HAITHAM ABU-RUB (Fellow, IEEE) received the Ph.D. degrees. He has been working with TAMU-Q for the past 16 years. He is currently a Full Professor. He attained research and teaching experience in diverse countries, including Germany, Palestine, Poland, Qatar, and USA. He has served as a Chair of the Electrical and Computer Engineering (ECEN) Program with TAMU-Q for a period of five years. He is currently working as the Managing Director (MD) of SGC-Q. He has published over 400 research articles. He has five books and six book chapters to his credit. His research interests include renewable energy, power electronic converters, smart grid, and electric drives. He was a recipient of numerous recognitions, international, and national awards. He received the German Alexander von Humboldt Fellowship and the American Fulbright Scholarship.

OTHMANE BOUHALI (Member, IEEE) is currently a Physics Research Professor with TAMU-Q. He is also the Founder and the Director of the TAMUA Advanced Scientific Computing Center. He has been involved in the Large Hadron Collider Research Program for more than 25 years. He has supervised various research projects. His research interests include large scale modeling, medical physics, and high-performance computing and detector technologies for radiation.