IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Big data management in smart grids: technologies and challenges

Ameema Zainab[1,2], Member, IEEE, Ali Ghrayeb[2], Fellow, IEEE, Dabeeruddin Syed[1,2], Member, IEEE, Haitham Abu-Rub[2], Fellow, IEEE, Shady S. Refaat[2], Senior Member, IEEE, Othmane Bouhali[2], Member, IEEE

[1] Department of Electrical and Computer Engineering, Texas A and M University, TX, USA
[2] Department of Electrical and Computer Engineering, Texas A and M University at Qatar, Qatar

Corresponding author: Ameema Zainab (e-mail: azain@tamu.edu).

This publication was made possible by the NPRP grant [NPRP10-0101-170082] from the Qatar National Research Fund (a member of Qatar Foundation), an internal seed grant from Texas A&M University at Qatar, and the co-funding by IBERDROLA QSTPLLC. The findings achieved herein are solely the responsibility of the author[s]. The Open Access Funding was provided by Qatar National Library.

ABSTRACT Smart grids are re-engineering the electricity transmission and distribution system throughout the world. It is an amalgam of increased digital information with the electrical power grids. Managing the data generated from the grid efficiently is the key to successful knowledge extraction from the smart grid big data. Most of the scientific advancements are becoming data-driven and becoming an interesting area of research for data scientists. It is challenging the world computationally enough to develop new storage methods and data processing technologies. Managing big data involves data cleaning, integration of varied data sources, and decision-making applications. This paper focuses on the study of big data management and proposes a management process to help manage the data in the grid. Data management tools and techniques have been leveraged in understanding the sources and data types in the grid. The paper emphasizes the limitations of the existing solutions inclined towards applications of the smart grid big data.

INDEX TERMS Apache Spark, Big data, Data mining, Hadoop, Indexing, Management process, Smart grid, Stream Mining

I. INTRODUCTION

In the past decade, electricity consumption has evolved in practice and, the power generation modes have changed with the development of renewable energy sources and the transformation of electrical systems is a must to properly balance electrical consumption and electricity production. Smart grids provide a safe and reliable integration of different renewable sources into the generation mix and guarantee the safe operation of the electrical systems.

A smart grid can be viewed as an amalgamation of information and electrical power. It forces a cross-fertilization of electrical systems with different fields such as statistics, applied mathematics, and optimization methods. The massive amounts of high dimensional data produced from the grid brings several new challenges and opportunities to the table. The solution to these challenges would lead to a substantial contribution to the research area of big data management for smart grids. Managing the data generated in the grid, turning the data into useful information, and making decisions are a few of the most important steps in managing the smart grids. Big data management in the smart grid plays a vital role in the applications and extracting information from smart grids' data.

Big data offers potential insights and is crucial for the efficient functioning of the smart grid [1]. Information from big data being valuable, many energy companies have invested in handling the data to perceive innovative and actionable insights. It is estimated in a preliminary assessment by a utility that the amount of data required to process transactions of its customers would reach about 25 gigabytes of data points per day [2]. This set of large data to be managed is a challenge. Energy companies like ENEL are moving towards new strategies and plan to be data-driven companies exploiting huge amounts of data obtained from the grid architecture, customers, etc. [3]. It is estimated that more than 80% of the companies will be evaluating to migration of their data from data center to the cloud to estimate associated savings [4]. ENEL plans to focus on a platform model rather than a pipeline model involving data-driven networks. It is very crucial to manage the smart grid data as it would help the utilities to understand the demand and perform a dynamic balance of demand and supply. This requires deep analysis of demand data concerning different conditions of weather, different days of the month, different months of the year, weekends, and holidays for residential, commercial, and industrial customers. Also, it involves huge volumes of granular data to be processed and the correlations between different data features to be identified [5]. Also, it is very crucial to identify different analytic and data management strategies to be utilized for different applications and usage. The authors in [6] have discussed the use of block chain technology in applications such as metering, energy trading, demand response, microgrids, virtual power plants, load forecasting which are all highly dependent on effective data management in the smart grids. The biggest players in the energy market have utilized big data technologies to manage the grid. National Grid, DTE Energy, and Ausgrid are some of the largest utilities which have used the International Business Machines (IBM) insights foundation for energy to help improve their decision-making for monitoring of asset health and maintenance [7].

NextEra, one of the top renewable energy-producing companies, claims to perform ($3 to $4 per MWh) better than any other company in the United States by dynamically operating its wind turbine using machine learning techniques [8]. A SAS data management and predictive analytics platform were implemented by Électricité deFrance (EDF Energy) to perform churn modeling for the evaluation of the propensity of electrical users who change their utility provider [9]. The company has built SAS models to process immense volumes of data easily and accurately for training and then the models test or predict variables against all the data features. Incorporating a broad range of modeling techniques such as logistic regression etc. the company predicted that the top 25% of the customers are more inclined to opt dual fuels which resulted in customers being less liable to churn. This helped the company save an average of £300 million a year. The company has also utilized Hadoop to store time-series data and to perform analytics [10]. Romeo project is a five-year and €16 million project led by Iberdrola Renowables Energia [11]. This project focuses on managing the data from the wind farms using predictive models and physical fault models to lower the operation and maintenance costs of the wind farms [12]. The importance and uses of managing big data from the grid are endless. The following are the areas that get impacted the most by smart grids [13]:

- Peak Demand and Energy usage- includes subareas of advanced metering infrastructure, pricing policies, customer end sensors, etc.
- Energy efficiency in the distributed systems- includes subareas of line losses, voltage, and frequency optimization, etc.
- Operations and maintenance savings from advanced metering infrastructure- includes subareas of smart meter reading, service changes, outage management, etc.
- Operations and maintenance savings from distribution automation- includes subareas of automated and remote operations, operational efficiency and optimization, etc.
- Distribution system reliability- includes subareas of feeder switching, asset monitoring, and health sensors, etc.
- Transmission system operations and reliability- includes sub-areas of synchro-phasor technology for wide-area monitoring, visualization, operations control, etc.

The data sizes reaching petabytes is currently a challenge for the databases to process the data. To overcome this problem current utility companies are extracting processed data rather than raw data. The smart grid big data is booming, and it has become very critical to extract meaningful information from it. With the advent of smart grid systems, the energy data that is generated in huge volumes and at high velocity can be recorded and communicated for further processing. To perform the analytics on the data for required applications and visualization, relevant software technologies must be in place [14]. Few of the developed bigdata platforms in different fields summarized include TVA's Hadoop-based smart grid management system, Kyushu's cloud computing-based fast data processing platform, and others are discussed in [15][16]. There is no unique platform or tool that can serve as a solution to all the big data challenges, as each technology has its own merits and demerits in addressing the challenges and each provides a perfect specific solution to one challenge of big data or the other challenge that is being dealt with. Nevertheless, data management calls for time and energy to be invested in the development of better solutions to manage the grid. Several platforms have been proposed to manage the data from the grid ranging from cloud-based platforms to the real world implemented platforms with a primary focus on smart grid data [17][18][19][20][15][21].

The main contribution of this paper is to review big data management processes that can handle the data from the grids using the latest data handling techniques. The paper focuses on dealing with both archived and real-time energy data and intends to manage the data with the help of servers/data centers. The data is cleaned and made available to perform analytics also supporting machine learning to help make decisions related to the grid.

This paper, aiming to apply big data technologies to the smart grid, proposes an architecture for big data management with detailed discussion on data acquisition, data pre-processing, and data communication. The proposed architecture is a distributed management system that takes care of acquisition, data monitoring as a platform, and pre-processing in the form of data mining, data identification, and data sharing techniques. The distributed file system that can be used for data storage is also discussed. The paper gives an overview of the challenges of big data management in smart grids.

The rest of the paper is organized as follows: Section II details the related work in the area of big data management in the smart grid. Section III outlines the big data technologies in smart grids. Section 4 summarizes the software technologies, the various data types present in the smart grid's data, and the mathematical terminology involved in handling the smart grid data. Section 5 discusses the challenges faced and Section 6 proposes a big data management platform that will overcome the challenges discussed in Section 5. Section VII concludes the paper.

II. RELATED WORK

A. Cloud-Based Big Data Platforms

Many frameworks have been proposed in the literature to understand the data flow, analyze the data, and manage the data in smart grids. Previous works include the proposal and implementation of big data frameworks in the smart grids to take decisions on many aspects such as balancing demand, load forecasting, grid infrastructure optimization, asset management, consumer behavior analysis, state estimation, and service quality analytics, etc.

In [17], Mayilvaganan et al. proposed a cloud-based smart grid management architecture that analyses the big data for balancing the demand and supply to meet customer needs. The analysis helps in efficiently dealing with power generation and distribution. The advantages of this architecture involve the use of cloud computing and big data analytics to perform various functionalities in the smart grid, i.e.

- Prediction of energy production by historical data analysis.
- Prediction of demand in advance by consumer behavior analysis.
- The decision of high or low priority demands

In [18], Yogesh et al. have proposed 'Floe's, a continuous data flow engine that utilizes a private cloud infrastructure. The proposed cloud-based D2R (Dynamic Demand Response) platform performs intelligent dynamic demand response management relieving the load peaks in the power grid. The platform has been validated on a microgrid and it is adaptable to ingest dynamic data flow. Demand forecasting has been performed by training massive datasets with scalable machine learning models. In [19], Baek et al. proposed 'Smart-Frame' as a secure cloud-computing-based big data platform to analyze a voluminous amount of data acquired from power assets, smart meters, and distinct types of front-end devices in the grid. Along with the structural framework to form hierarchical cloud computing services for big data analysis and information management, and identity-based encryption security solution has also been presented. A popular cloud computing opensource platform called eucalyptus has been utilized for the prototype implementation [20]. It is in sync with the Amazon Web Services (AWS) industry standards and cloud APIs which also support virtualization technologies like VSphere, Xen, VMWare, and KVM. The platform can also be developed and implemented on major operating system distributions like Ubuntu, Debian, etc. The platform is built such that the following cloud computing services can be accessed:

- Infrastructure-as-a-Service (IaaS)- This layer stands as a backbone of the system with the main tasks involved such as information gathering, storing, and processing. This layer serves all their sources demanded by the services and applications deployed in the smart grid system.
- Software-as-a-Service (SaaS)- At the top of the system, all the services (smart grid) will be set up in the SaaS layer. The SaaS applications will have a user-friendly interface. For example, Google Power Meter tracks almost real-time electricity usage statistics and helps customers optimize or save energy.
- Platform-as-a-Service (PaaS) - In this layer, applications and services are developed based on cloud computing with the help of tools and libraries provided, e.g. Salesforce. Platform as a service in the field of smart grids will help in the implementation of customized applications. It will make data management easier and quicker to some extent, as the service already integrates the special security requirements and lawful interceptions needed. Cloud characteristics will be inherited from the applications developed.
- Data-as-a-Service (DaaS) - This layer provides useful information for statistical use from the extremely large smart grid data files. This layer can be used not only by the customers but also by electricity providers. This service is provided only as read-only, and the data provided cannot be downloaded.

The proposed framework provides security to the system by enabling hierarchical identity-based cryptography. The cryptography makes the framework secure in addition to being scalable and flexible.

B. REAL WORLD IMPLEMENTED BIG DATA PLATFORMS

Big data has made its presence in numerous industries such as finance, smart buildings, commerce, etc. A few of the developed bigdata platforms in different fields have been summarized below:

1) THE SMART GRID: HADOOP AT THE TENNESSEE VALLEY AUTHORITY (TVA)

TVA was selected by NERC (North American Electric Reliability Corporation) in 2009 as the repository for PMU data nationwide. America's power grids at the TVA producing hundreds of terabytes of data have been handled with the help of apache Hadoop [15]. The platform of Hadoop has enabled TVA to perform a deeper analysis of the data at very lower costs compared to the existing solutions. Hadoop aids data management with its distributed file system called H.D.F.S. to store huge amounts of PMU data while making the platform available and reliable at all times. Hadoop's aggressive replication scheme has helped the organization to have an operational file system even in cases of losing whole physical machines. The data flow from the measurement device to TVA is described below:

- The measurement device of the substation timestamps various data samples with the help of a GPS clock and sends the samples to a central location over optical fiber,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080433, IEEE Access

coaxial cables, twisted-pair cables, or any other suitable lines.

- A VPN tunnel over a LAN to LAN connects the TVA and the participant companies. Several partners also use the Multi-Protocol Label Switching (MPLS) connection in the case of more remote regions.
- Data concentrator called the Super Phasor Concentrator (SPDC) was developed by the TVA and it receives the data after a few network hops. The PMU input is ordered in a time-aligned sequence compensating any delayed or missing data introduced by the network delay, latency, or congestion.
- 19 companies, 10 different PMU device manufacturers, and 103PMUs comprise the entire stream. 16 measured values at a rate of 30 samples a second are passed to the servers. Archive files of the PMU moved via an FTP interface into the Hadoop cluster. Real-time data is continuously streamed, processed, and fed to client visualization tools.

Cloudera along with TVA was successfully able to store the PMU data and make the data available for analysis. Utilizing Hadoop for the process has made the platform very cost-efficient. This is one of the best examples to show how big data management from the smart grid took place in a real case scenario. The next steps include crunching the greater amounts of data to be stored or analyzed in real-time for the multi-sensor data from the smart grids and not only PMU data.

2) CIDAP - BIG DATA PLATFORM FOR SMART CITIES

In [21], architecture named ‘SmartSantander’ which is a live city flexible big data platform has been introduced. A practical system has been built in the testbed city of Smart Santander to evaluate the platform. The work provides insight into the future smart city platforms that can address various issues that can be encountered at the time of building the system. The main emphasis of the proposed system architecture is to take the helm of both the historical and real-time data. It also emphasizes handling different scales or types of data Figure 1 explains the platform overview and the workflow is described below:

- Data from various sources is collected via the Internet of Things (IoT) broker and stored in the big data repository [22].
- A set of pre-defined tasks processes the collected data. The processing is done at different levels, depending on the complexity of the process.
- Basic processing such as format transformations, creating new structured views for data indexing, etc. are performed at the big data repository level. While the complex processing such as mining the data with advanced analytics is performed on the separate computation resource supported by a spark cluster [23] which comprises a huge number of compute nodes.
- A web-based data management portal is designed to monitor and operate the entire big data platform.

3) OTHER BIG DATA PLATFORMS

In [24], ’SCOPE’ was presented as a smart city Cloud-based Open platform and ecosystem by Boston University. It is a platform that is open to innovators to develop smart-city services, with a focus on being an open platform to collectively innovate and monetize the big data assets. It acts as a template to help break the technological silos involving deep citizens’ involvement in the wide-spread adoption of smart city services. It makes use of sensor-based information to develop services such as transportation, energy, health care, commerce, business, and social applications amongst others. City Pulse [25] is proposed by Osborne Clarke, a smart city consulting firm from Europe [26]. City Pulse is a large-scale data analytics framework for smart cities. The framework combines and operates large-scale streaming data of the cities in an extensible and flexible manner. Application Programming Interface (API)s that are exposed by CityPulse components facilitate the application creation and services.

FIWARE [27] is a smart energy platform for the development of intelligent applications in the future internet. It serves as an energy platform capable of supporting various

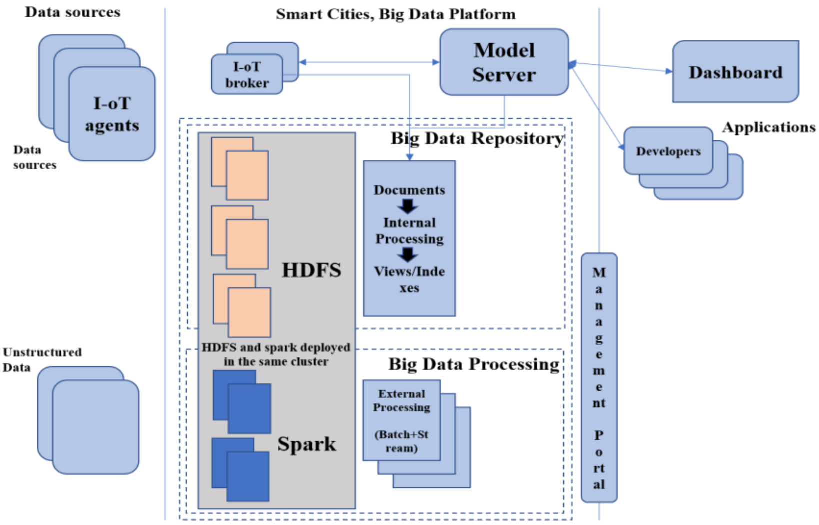


FIGURE 1. CiDAP platform architecture overview [18]

business models for different smart energy industries. It is desired to work as a testbed for facilitating new services. Wang et al. proposed wireless computing architecture for the processing and analysis of smart grid data [28]. The proposed inner and outer optimization method has improved the storage planning scheme resulting in better energy scheduling and lower costs to the customers. Zhou et al. presented data mining and visualization techniques for smart grid data and achieved real-time monitoring of power consumption [29]. Their work also highlights the need for distributed technologies for increased computation and scalability to accomplish unified data management.

Apart from smart grid data management some of the other big data platforms developed include: In [30], 'SealedGRID' a highly trusted and interoperable smart grid security platform has been presented which abides the blockchain concept with the web of trust. An anamoly detection framework has also been proposed based on big data and machine learning using the blockcahin technology [31]. In [32] 'UlTraMan a unified platform for big data management and analytics for trajectory data is proposed. It offers a customized pipeline extension of modules offering enhanced computing. ASTROIDE a unified big data processing engine over spark for astronomical data. It introduces efficient query execution, by data partitioning with Hierarchical Equal Area isoLatitude Pixelization **(**HEALPix) on Spark [33].

III. BIG DATA TECHNOLOGIES

Because the data is both complex and has different formats, handling the data is not straightforward. Big data technologies offer scalability, persistence, and are computationally efficient. Various technologies offer services that help in dealing with big data complexities. A comprehensive review of the storage and processing structures, database management systems, software technologies, architectures, systems benchmarking, and data indexing.

A. STORAGE AND PROCESSING

1) HADOOP

A unified and centralized storage platform to manage various types of data. Hadoop augments itself by providing a repository where structured, semi-structured, and unstructured data may be processed together easily [34]. Along with being an open-source software, Hadoop is fault-tolerant and has a very reliable storage system. Having a programmable storage system, it is flexible for users to analyze the data directly attached to the disk where it resides. However, Hadoop has limitations i.e. it supports only batch processing and is not efficient with real-time, iterative, and stream processing. The data collected from dispatched sources in the grid is stored in huge datasets. This data needs to be accessible by multiple users on multiple machines for analytics. The Hadoop framework helps in parallelizing the processing in cloud computing environments and permits users to attain a local copy of the stored data. The Hadoop distributed file system is also well known for efficient storage of data as it provides fault-tolerance, high availability, and scalability. However, for applications such as smart meter analytics, load forecasting, and scheduling which require stream processing, Hadoop is not very efficient as it cannot produce output in real-time with low latency. The Hadoop ecosystem is built of two components, MapReduce and Hadoop distributed file system (HDFS) and these are discussed below:

2) MAPREDUCE

It is a parallel data processing system of Hadoop. It is the programming model used within Hadoop and it is efficient at processing huge volumes of data. MapReduce works on the concept of job scheduler which assigns multiple tasks in parallel to Data Nodes in a single cluster or shared clusters and results are collated, filtered, sorted, and then passed out as an output. If the task assigned to a node is overloaded or failed in a cluster, then the task is executed by another server in the cluster as shown in Figure 2. MapReduce can execute in a potential number of high-level languages such as C, C++, and scripting programming languages i.e. Python, Perl, and PHP. It can also be noted that as MapReduce processes large datasets, it requires a large amount of time and might result in increased latency. Running on various clusters results in increased time and lesser processing speeds. This limitation can be overcome by the in-memory computation capability of the Hadoop spark. MapReduce does not have an interactive mode. However, this can be overcome by adding Hive Hadoop [35] or Pig Hadoop [36] and this enables users to have an interface to deal with MapReduce paradigm without having to code complex java MapReduce programs.

3) HDFS AND HOPSFS

The file storage system in Hadoop is called Hadoop distributed file system. Because of its write once and read many models, it is best suited for data integrity when a read operation is performed. Many grid centers utilize Hadoop with HDFS file storage to collect various types of data from the grid such as phasor measurement units (PMU). HDFS however doesn't support random reading of small file sizes. It is designed in a way to support a small number of large datasets rather than a large number of small datasets. This can be overcome by merging the small files into one and then copying the bigger files to HDFS.

HOPSFS is an open-source file system and it is an alternative to HDFS [37]. It uses the active and standby name nodes and thereby overcome the deficiencies of HDFS. The name nodes in HopsFS can process the metadata not just locally in memory but also the metadata stored in the

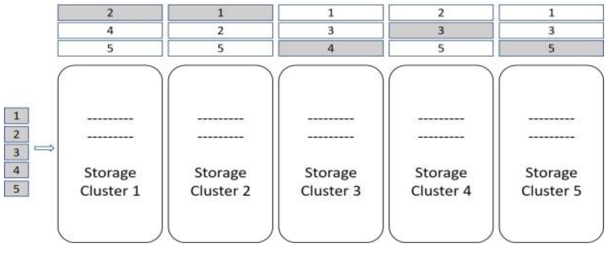


FIGURE 2. **Software Framework – MapReduce**

database. HopsFS works with different varieties of NewSQL databases even if the databases have different licenses. It is since HopsFS uses Data Access Layer (DAL) as encapsulation to the database operations.

5) APACHE SPARK

It is a lightning-fast framework that processes data that exists in data storage systems such as HDFS, Amazon S3 [38], MapR FileSystem [39], Cassandra [40], etc. The data processing also utilizes a cluster manager such as spark cluster, Apache Mesos, HadoopYARN, etc. [41]. Spark can process the data as it comes, even millions of events per second as it uses Resilient Distributed Datasets (RDDs) which reside in memory. The flexibility, speed, and scalable features of spark address the challenges of big data in smart grids. Spark also supports user-friendly APIs such as Python, Scala, Java, etc. and this makes developers easily use spark for machine learning libraries [42]. The very nature of data from smart grids (for example, the data from SCADA) is dynamic and anomalies in electrical systems tend to occur in milliseconds. Apache spark supports the real-time processing of the data and it can capture real-time information from the grid. Memory management in spark is crucial and involves various levels such as memory only, memory and disk, memory only serialization, and memory and disk serialization. Based on the size of the data and the memory allocation is altered.

6) RESOURCE SCHEDULER

A key to efficient utilization of a large asset is the choice of a suitable resource scheduler. Both supercomputers and big data systems use schedulers to allocate computing resources for the execution of submitted processes. The authors in [43] analyze 15 schedulers in both supercomputing and big data architectures. In [44], the authors utilized upto 32 processors with the help of Slurm resource scheduler. Four of the most popular schedulers include Slurm, apache YARN, Apache Mesos, and Kubernetes are open-sourced.

B. DATABASE MANAGEMENT SYSTEMS

Picking a relational (SQL) or a non-relational (NoSQL) database is one of the crucial decisions in choosing a database system. Both types of databases are suitable options, however, non-relational databases are constantly replacing relational databases as non-relational databases are efficient for big data applications. The cost of scaling of relational databases is very high and the volume of data is ever-increasing in big data. Moreover, the ACID properties (Atomicity, Consistency, Isolation, and Durability) set unrequired constraints and hindrances to applications and these pose a challenge [45]. Therefore, relational databases are best avoided in big data applications.

NoSQL data storage has more ability to perform better adaptability, scaling, and performance when compared to relational databases. Although it must be noted that NoSQL does not have a universal query language that fits with all data models. Instead, it allows for RESTful coherence to the data and the query APIs. A comprehensive study explains the uses and performance comparisons between relational and non-relational databases [46]. Some of the non-relational databases include Redis, MemCached, Dynamo, Cassandra, PNUTS, MongoDB, CouchDB, Neo4j, HyperGraph DB, etc. The comparison between the relational and NoSQL databases is discussed in Table 1**.**

There are many other databases in the market that provide support to the requirements of huge data size, different data types, and high speed. The big databases include in-Memory or main memory databases, object-oriented databases, time-series databases, and spatial and GIS (Geographical Information systems) databases. Even though in-memory databases are quite fast they are not durable, and it might be subject to data loss. The spatial databases are useful when data has geospatial attributes, but at the same time, it is hard to query upon [47]. Also, it requires good visualization to interpret the data patterns. Streaming data from SCADA and oscillography data are usually stored in time-series databases.

C. SOFTWARE TECHNOLOGIES

The evolution of big data technologies started way early in the 1990s. A boost to big data technology started with Hadoop in 2011 and it has been an open-source platform. Big data technologies have evolved in the past decade performing batch processing at one stage to real-time processing later. In [48], Sebnemet al. has explained the evolution of big data technologies starting with Google File system performing batch processing (2003) to Google Data Flow and apache spark (2003) performing real-time analytical streams processing. Different software applications were released in the market and many were open-source, and these handled the high data volumes and high speeds while decreasing the latency of processing. One of the most widely used state-of-the-art lambda architecture has been discussed in the section below along with the system requirements to handle the software technologies:

D. ARCHITECTURES

1) LAMBDA ARCHITECTURE

The advantages of data systems built with the assistance of lambda architecture go beyond just scaling and supporting real-time and batch processing on the distributed data. In support, the architecture will not just be capable of handling the data only but will also be able to accumulate more data to interpret information from it. Increasing the number of data types and volumes stored will result in further opportunities to mine the data including, predicting performance, avoiding more than one version of a schema to be operative at the same time, and building new applications. Lambda architecture (Figure 3), a unique software design, is adopted to overcome the need to process two different systems considering batch processing and stream processing [49].

Hadoop discussed in section 3.1 can handle the data at rest with the help of Hadoop's MapReduce functionality. The data received would be pulled into HDFS and MapReduce jobs are executed using Pig, Hive etc. As all the data would

TABLE I
SOFTWARE FRAMEWORK – MAPREDUCE

Characteristic	Relational Databases	NoSQL Databases
Data representation	Predefined schemas. Schema represents a logical view in which the data is organized & the relations are displayed.	Dynamic schema for unstructured data
Data Structure	Structured	Unstructured or lenient structure
Scaling	Vertically scalable. The amount of data stored depends on the physical memory of the system. Relational databases are scaled by increasing the hardware resources like CPU, RAM, SSD etc. on a single server.	Horizontally scalable. No limit on data storage. NoSQL databases are scaled by increasing database servers.
Examples	MySQL, Oracle, SQLite, Postgres, MS-SQL, etc.	MongoDB, Bigtable, Redis, RavenDB, Cassandra, HBase, CouchDB, Graph databases like Neo4j, OrientDB, InfiniteGraph, AllegroGraph, etc.
Types	Table based databases	Column DB, Graph DB, Key-value pair DB, Document DB, etc.
Properties	ACID (Atomicity, Consistency, Isolation, Durability)	CAP (Consistency, Availability, Partition tolerance)
Language	Structured Query Language for data definition & manipulation	Unstructured Query Language
Development	Mix of open source (PostgreSQL) & closed (Oracle)	Open source
Model	source databases	
Complex Querying	Suitable for complex querying	does not have standards to perform complex queries.
Complexity	If records do not fit in the pre-defined schema tables, then the design of the database table becomes complex.	Schema is easily changed here as it is dynamic.
Community	Widely supported from vendors	Only community support
Normalization	Necessary	No constraint of normalization
Maintenance	High maintenance	Low maintenance with features like automatic repair, easier distribution of data & simpler data models is available. So, administration is easy & so is tuning requirement.
Consumer friendly	GUI mode tools available.	GUI mode tools not available.

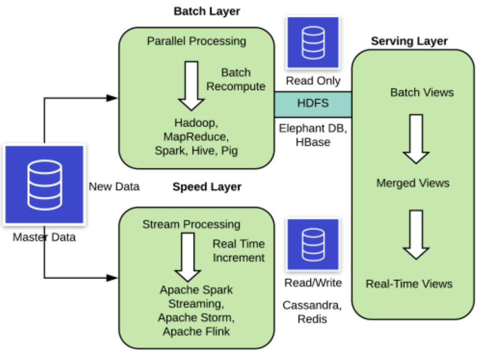


FIGURE 3. Lambda Architecture [49]

be in HDFS, there will be a full view of the data available to process it. Streaming analytics engines such as spark and flink will assist to perform processing and analytics on incomplete data or when data is being updated [50]. These engines process ables the data as it comes in and does it a lot faster. These help in processing the data even before the data is transferred to HDFS. A portion of the data that is collected is analyzed instantaneously as and when the data is generated, and the rest of the data is stored for batch processing. Table II refers to some current systems in the field of stream analytics. Analyzing the data as it is available from the source to the memory of a distributed platform needs stream mining systems. If working with stream only frameworks is desired, then apache storm [50] is one of the best-suited frameworks as it offers a great range of language support, but at the same time, it cannot guarantee to order in its default configuration. The best fit always relies on the data being analyzed, the required latency, and the application required. The three layers of Lambda Architecture are:

- Batch layer: stores all the data as 'master data', manages it, and precomputes batch views.
- Speed Layer: processes the incoming streaming data as per user-defined requirements and increments the real-time views.
- Serving Layer: a linearly scalable data management system on top of the batch layer and speed layer exposing queried views by the user.

E. SYSTEMS BENCHMARKING

TABLE II
STREAM MINING SYSTEMS

Current Systems	Year
R's stream package (clustering only) [71]	2017
streamDM (github) [72]	2016
Moa.cs.woikato.ac.nz (Massive Online Analysis) [73]	2014
Samoa-project.net [74]	2014
lambda-architecture.net [49]	2013
Spark.apache.org/streaming [75]	2012
Rapid Miner stream plugin [76]	2012
Apache Samza [77]	2012
Apache Storm [78]	2011

Big data in the smart grid sector involves not only data at rest but also real-time data. Owing to the data being real-time and continuous, additional resources and high computational speeds are required.

As discussed earlier, the use of cloud computing helps electrical companies to reduce cost and power requirements. Table III shows the minimum requirements needed to install the platforms Hadoop, Strom, Spark, and Flink and work with the big data frameworks [51]. A minimum of 8 GB RAM is required to have any of the mentioned software technologies to be installed. A supercomputer will help in the processes to run faster.

F. DATA INDEXING

Indexing plays an important role when it comes to big data management. The speed of data retrieval from a database system is vital for efficient data access. Time-series data is one of the massive types of smart girds. An index format is chosen based on the type of storage system. Table IV shows a summary of advanced data indexing techniques that exhibit comprehensive distributed functionality. As the paper suggests the utilization of a distributed framework, the section focuses on distributed data indexing techniques.

TABLE III
BIG DATA FRAMEWORK HARDWARE REQUIREMENTS [53]

Framework	Hadoop	Storm	Spark	Flink
RAM(Min)	64 GB	64 GB	64 GB	64 GB
CPU (at least)	2	8	8	8
Hard Disk (for each 1TB at least)-Disks per node	12-24	6	4-8	12-24
Operating Sytems	64 bit:SUSELinux Enterprise Server	CentOS, Red HatEnterprise Linux, Windows	Windows XP/7/8, Windows (Cygwin), Linux, MacOSX, CentOS, Linux	Linux

TABLE IV
DISTRIBUTED DATA INDEXING TECHNIQUES

Indexing	Year	Property	Underlying storage system
FITing-Tree	2019	A data-aware index structure that captures data trends and fits an index to a dataset with the help of piecewise linear functions.	-
Parallel B+ trees [79]	2019	Tree-based: maximizing terminal nodes and minimizing height of a B+ tree	Hadoop
FastPM [80]	2018	Extends k-d tree indexing to a distributed framework	
IndexedHBase [81	2014	Historical and streaming dat scalable indexing	HBase
E3 [82]	2013	Avoiding irrelevant data splits accesses	Hadoop
HIndex [83]	2013	Secondary Index (server side)	HBase
HAIL [84]	2012	Less index creation cost	Hadoop
MD-HBase [85]	2011	Quad-Tree and K-d based multi-dimensional index	HBase
Trojan Index [86]	2010	Created at data load time an at query time no penalty	Hadoop

IV. SMART GRID DATA

An automated big data management pipeline for a smart grid must have the following qualities:

- The platform should be able to support the acquisition of dynamic data at variable rates and high volumes.
- The platform should be adaptive to the operational needs of current data sources.

The data sources in the smart grid fall under four categories i.e. Historical (archived), real-time, multimedia, and time series [1]. Data sources from SCADA, PMUs, Automated Metering Infrastructure (AMIs), smart meter, Digital Fault Recorders (DFRs), Digital Protective Relays (DPRs), Intelligent Electronic Device (IEDs), Asset management, operational and weather are real-time data sources. The real-time data flows in high volumes and the data is either collected at once or streamed in chunks continually. For instance, standard SCADA polls every 4 seconds. PMU, weather or lightning, and GIS are mostly historically based. The data is usually available in bursts from devices in the grid or as files stored in any of the storage devices and this data can be captured when there is a triggered event. On-demand, this data is transferred by the utility for different kinds of analyses. Data in the form of text, voice, and video (e.g., video surveillance cameras) are multimedia and PMU data are time series. Most often event messages are generated in response to any unusual physical events. These responses might be in the form of commands communicated to the grid devices by grid-control systems, e.g., an asynchronous business process such as meter ping [52].

A. DATA ANALYSIS APPROACHES

Big data management deals with finding the hidden patterns in the data to get meaningful information as an output. As the data grows in volume, variety, and velocity, it tends to be multi-dimensional. To handle big data with multiple dimensions, Random Matrix Theory (RMT) is particularly useful [53]. The most fundamental concepts of dealing with big data account for the representation and modeling of big data.

The random matrices are natural building blocks in modeling big data [1]. The non-asymptotic theory is a unified treatment to a lot of big data problems, which was proposed to model the datasets as large random matrices in 2010 [54]. A single dataset can be expressed as an m×n matrix given by

$$X = U \wedge V \tag{1}$$

$U(m \times n)$ - Orthonormal rows matrix

$\wedge(n \times n)$ - Diagonal matrix with real and non-negative entries

$V(n \times n)$ - Unitary matrix

Where XX^H and $X^H X$ are Hermitian matrices with diagonal entries of $\wedge 2$ correspondings to the eigenvalues. U and V correspond to the eigenvectors. When it comes to large random matrices $m \to \infty$, $n \to \infty$, both Hermitian and Non-Hermitian are utilized in various applications based on the variety of data [55]. Some of the differences between Hermitian and non-Hermitian matrices have been stated in Table V.

In a high dimensional setting, it is often desired to cut down the dimension of the matrix by working on a low-rank matrix approximation and often require solving for eigenvalues. The most prevalent methods are Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). PCA is one of the most widely used dimensionality reduction techniques [56].

It is used to reduce the number of features in the data. It selects the features which have the most variance in the data and neglects the features that have the least information in the data. We can explicitly specify the number of principal components or features that we wish to consider. The reduction in the features decreases the training and the testing time to a great extent and this knowledge can help in the reduction of data that is to be managed.

V. CHALLENGES

Even though the benefits of big data management are many, it also has many challenges and requires a high level of attention. Some of the key challenges related to big data management in the smart grid are summarized below [57]:

Data recovery and capture: Sensors' data sometimes is updated and overwritten discarding the previous data. However, the discarding of data should not take place until the information from the data is extracted [58].

Data size explosion: Data is generated with a precision of seconds resulting in Terabytes of data and so the analytical value per unit of data is low [59].

Data compression: The data communication requires the high-volume of data to be compressed before flow. Special compression methods are required for the electrical data because under normal conditions, the data either has constant values or is sparse [60]. The data compression will allow tackling the issue of network congestion and bandwidth requirements.

TABLE V
DEALING WITH BIG DATA MATRICES IN SMART GRID

Operations	Hermitian matrices	Non-Hermitian matrices
Diagonalization	$XV = V\wedge$	$XX_R = X_R \wedge$ and $X_L X = \wedge X_L$ X_R right-hand eigenvectors X_L left-hand eigenvectors
Eigenvalues	Real	Real or complex conjugate pairs
Eigenvectors	Orthonormal	Not orthonormal

Data loss: Data generated at measuring points or sensors and not usually streamed or transferred to storage units for analysis. The data loss should be dealt with carefully in the pre-processing step of data analysis and there are several techniques to deal with missing data like imputation methods [61].

Data Coherence: The data sources in smart grids are numerous and, the data is collected in substations at different locations, it will always be a challenge to share the data or to have centralized data storage [62]. The data integration should address the challenges of multi-source datasets and different formats or datatypes.

Real-time processing: The real-time processing of big data is of high priority because the applications required by utilities are very critical and require faster clearing times. The cloud-based infrastructure with Hadoop or spark is an apparent solution for real-time processing [63]. However, there are still inherent challenges of latency, network congestion, complex algorithms and computational speed to be solved and this solution is required to be feasible with the electrical data. Even though the spark is considered a lightning-fast framework there are many challenges in configuring spark. It involves the choice of memory storage levels, ineffective storage levels will result in overhead [64]. While running in a supercomputing environment selection of the right job scheduler plays a challenging role.

Performance: Utilizing the data in the grids to generate applications is a challenging problem to solve because it is an amalgam of both model-based analytical methods and data-driven IT methods.

Visualization: Visualization usually helps the operators to recognize the patterns in the data, monitor the real-time changes in frequency or voltage. So, it is crucial to represent the correlation and patterns in the multi-source data through innovative and effective graphs, charts, or images using efficient data integration, management, extraction, analytic, and visualization techniques [65].

Communication security: Fast, secure, and reliable communication channel is a challenge for applications involving real-time analytics [66]. Securing the streaming pipeline is a complex and time-consuming task. The grid data needs to be kept secure from network security attacks by maintaining data integrity and confidentiality.

Listed above are some of the challenges discovered and many others in managing the big data of the smart grid are yet to be discovered.

VI. PROPOSED BIG DATA MANAGEMENT ARCHITECTURE

The proposed architecture is broadly categorized into three mainstages - data collection, storage or transfer, and, mining and analytics. The challenges discussed in section V have been addressed in these three categories. To accommodate big data in the overall data management process, a plan needs to be in place. The plan should begin with integrating data as part of an operational process and finally, should involve understanding the workflow and addressing other characteristics of the big data i.e. Validity, Veracity, and

Volatility. The real-time implemented platforms discussed in section II such as TVA, uses Hadoop to store the data, CiDAP uses spark and hdfs for batch processing and CouchDB for real-time processing, SCOPE, a cloud-based smart city platform, City pulse, an ongoing European project, and Fiware smart energy presents only some high-level design architecture and details are not open.

Smart grid data contains both data in transit and data at rest (see Figure 4). Keeping in view that the data generated from the grids may be historical or real-time and voluminous or varied, architecture has been proposed for handling the data and managing it to accommodate the desired applications. The data management process can be depicted as in Figure 5 and the three main sections of the process data collection, data storage and transfer, and data mining and analytics are discussed in detail below:

A. DATA COLLECTION

There are various data sources in the smart grid such as, SCADA, Advanced Metering Infrastructure (AMI), smart meters, sensors, PMUs, distributed generation units, weather, customers, etc. The data is collected or transferred to the servers or cloud through secure channels. By selected querying and indexing from the endless flow of data, the relevant data stream can be ingested. The big data platform can also be to connected to simulated micro-grid system software's i.e. MATLAB or Simulink, PSSE, PSCAD, Power World, etc. This will add flexibility in terms of experimenting with the platform to handle the data received from a plethora of sources to customized sources and parameters.

B. DATA STORAGE AND TRANSFER

It is the nature of databases to get slower over time because of the increased usage as it results in staggering amounts of data being stored and databases getting bigger. As the data size can be in petabytes or more, computationally high storage devices are required. This calls for database management systems to perform better as per the application requirements. To overcome the challenges, the system needs to be robust and must have a strong disaster recovery plan keeping in view the worst-case scenarios. More on data storage and on trending big data technologies has been discussed in section III. Apache Hadoop is utilized in the proposed big data architecture. If an energy company, has archived data on the cloud or stroage facilities it can be easily transferred to the hadoop storage with the help of HDFS. To maintain fault tolerance in the data, database sharding is applied which horizontally partitions data across various nodes or servers. Data is either stored on the data centers or cloud, shared nodes or clusters based on the platform selected and the requirements of the application. Secure LAN or wireless channels are used to migrate the data from the grid to mining tools. Cybersecurity is a major concern when it comes to data storage or transfer on the cloud. Kerberos aids the security concern by providing authentication technology in the Hadoop cluster [67].

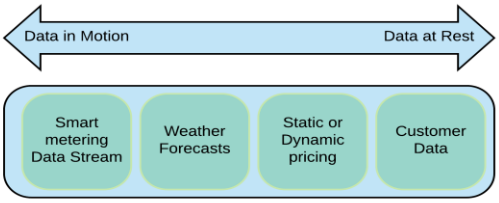


FIGURE 4. **Perception on the Methodologies**

Data streaming action over time is elaborated in Figure 6. Each of the devices can produce the data endlessly and continuously at a fixed rate or a custom pattern or in a random intermittent way. The multiple data streams from the stream

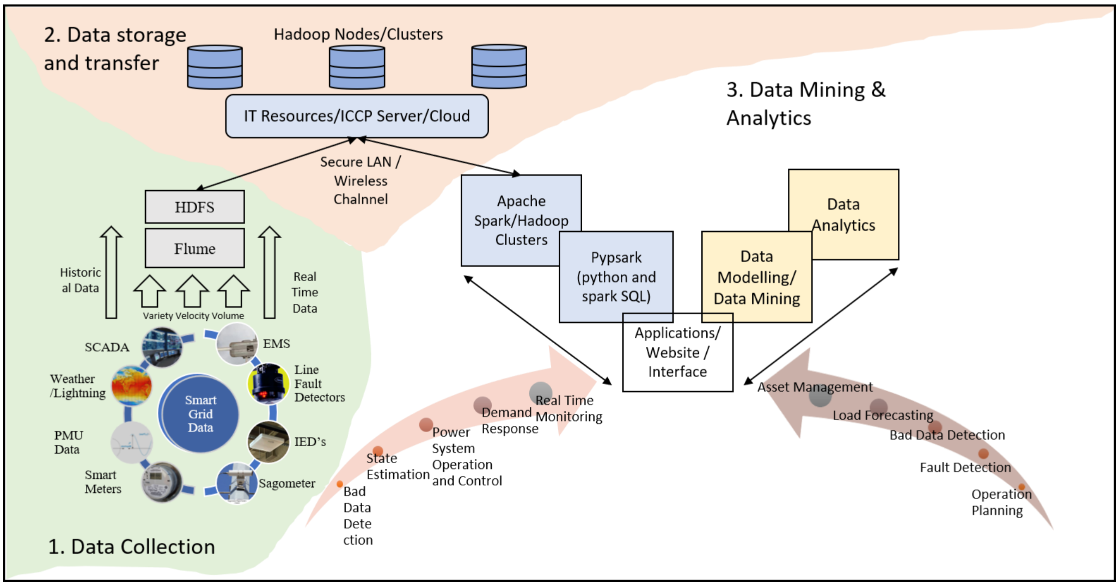


FIGURE 5. **Proposed Big Data Management architecture**

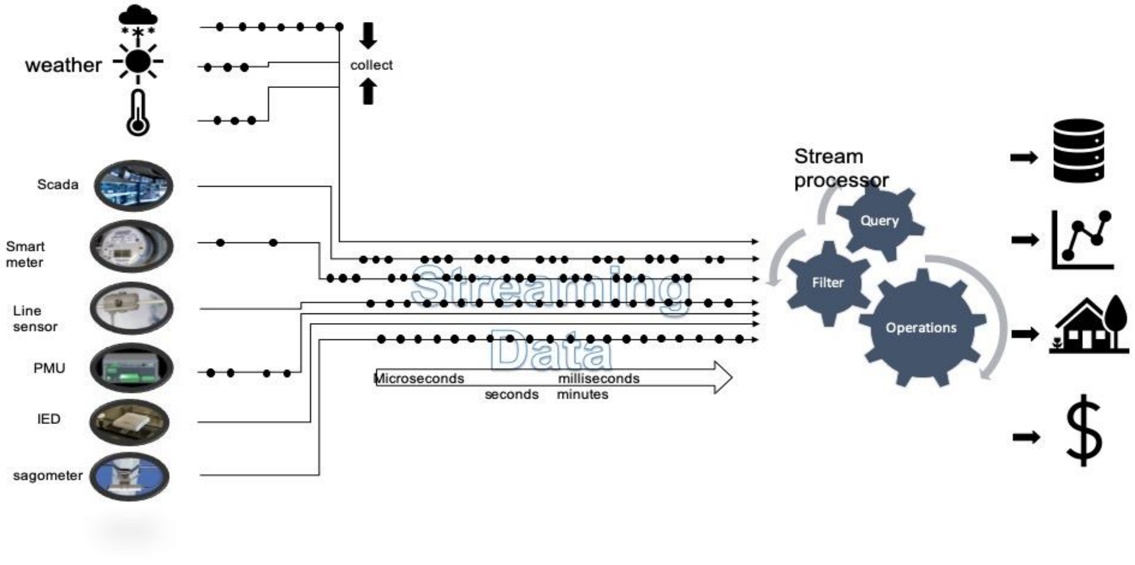


FIGURE 6. High-level view of Data Streaming

producers can be chosen and selected to be processed as a single data stream. This is possible as there is interdependency between the various data sources. The stream processing tool is responsible to perform operations such as a map, reduce, filter, etc. Spark streaming is fault-tolerant and is compatible with the HDFS file system of Hadoop and can process real-time data. The chain of operations is performed yielding the output of one operation as an input to the next operation. The result is delivered in an on-demand fashion to the customer, application, database, etc.

C. DATA MINING AND ANALYTICS

The data from the storage or transfer units is accessed by data mining tools and applications over a secured network channel. Information is extracted from the data after cleaning and mining the data and then analytics is performed. Based on the application, data from different sources is combined if needed and catered to the need of the application. Many applications such as load forecasting, bad data detection, state estimation, asset management, etc. can be performed with the help of the latest big data tools, i.e. Apache Hadoop, Apache Spark, Apache Hive, Cassandra, etc. along with programming languages such as Python or Scala or Java (discussed in section III). As the Machine Learning(ML) applications are massive, the use of ML in smart grid applications has been intense. To aid ML although spark and flink both support the python API, the spark has a mature community for ML applications and has more ML libraries than flink. Pyspark, a python API that supports both python and spark integrates both machine learning modeling and the spark platform and is an excellent framework to work with big datasets. A web interface in node js is also added to help as a communication medium between the smart grid and the operator or customer. Smart grid applications such as load forecasting, fault detection, and many others use machine learning capabilities to produce better prediction [68]. Data modeling is discussed in detail in the following section.

1) DATA MODELING

Modeling the data is an important step in mining and analytics. This is a part of the data mining process where an algorithm is selected, trained, tested, and finally deployed. The process is depicted in Figure 7. The data in hand is cleaned and validated to verify its completeness. If any of the data is missing, then different imputation methods or techniques can be used to deal with the missing data [61]. Redundant features such as a linear combination of other features that add no relevant information can be removed. Hence, feature selection methods are applied to deal with any correlated data, as having this redundant data only adds up to increased computational time. The reduced data is split into training and validation datasets and the experiment starts with the training dataset. Various models are

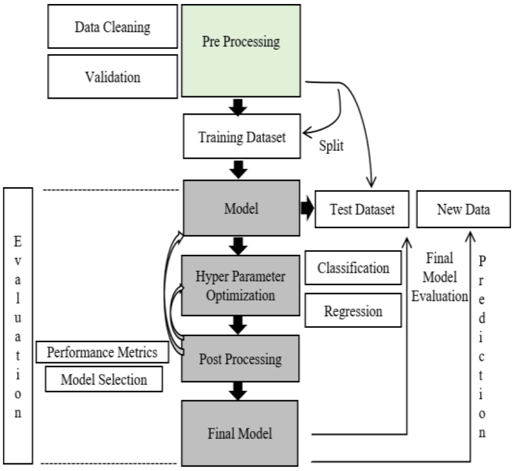


FIGURE 7. Data Mining phases/steps - process flow to find the best algorithm that fit the data

used on the training dataset to select models and algorithms that best suit the data. Once the best model is identified, it is implemented on the validation dataset to assess the model accuracy.

This process is repeated for various combinations of training and validation data splits making improvements until an acceptable level of accuracy is attained. The repetitive process at this step can make use of the parallel processing technology to spread the work across various computational nodes or cores. Cross-validation is one such method that can be used for the split of data into different combinations of training and validation datasets. The performance metrics such as accuracy, f-score, precision, recall measure, and processing time can be used to evaluate different machine learning models on the training and validation datasets and the best model with optimized hyperparameters can be selected [69]. Parallely distributed ML models are now available from spark's resource that are scalable and can be readily utilized resulting in more than 10x faster accuracy compared to Hadoop's mapreduce [70].

2) PERFORMANCE EVALUATION

Two important factors determine the robustness of the proposed data management architecture i.e., time and accuracy of models developed for the applications. The time includes the extract, transfer, and load (ETL) time and time to process the data. For example, for an application such as load forecasting, the time to access the data from HDFS, process the data, and run ML models is summed and a tradeoff between the total execution time and the accuracy of the ML model is considered.

VII. CONCLUSION

In the era of big data, where information is one of the key factors in making decisions, this paper has drawn attention to the need for data management in smart grids. Smart grid data from the electrical power utilities is very crucial to be worked upon for business valued applications and for saving energy sources. In this paper, a detailed list of big data platforms in the smart grid domain has been studied and the methodologies have been discussed. Numerous storage and real-time databases have been discussed to come up with an effective database for the proposed bigdata management process. Several computation models have been reviewed, necessary to manage the big data in the smart grid. A comprehensive review on the indexing, software technologies, storage and processing, frameworks, and architectures have been presented. Along with the review a centralized process flow has been proposed to manage the data in the smart grids. Challenges faced in data management has also been discussed.

REFERENCES

[1] R. C. Qiu and P. Antonik, *Smart Grid using Big Data Analytics*. John Wiley & Sons, 2017.

[2] D. Alahakoon and X. Yu, "Advanced analytics for harnessing the power of smart meter big data," in *2013 IEEE International Workshop on Inteligent Energy Systems (IWIES)*, 2013, pp. 40–45.

[3] M. Nisi *et al.*, "Transparently mining data from a medium-voltage distribution network: A prognostic-diagnostic analysis," in *CEUR Workshop Proceedings*, 2019, vol. 2322.

[4] "3 Journeys for Migrating a Data Center to Cloud IaaS - Smarter With Gartner." [Online]. Available: https://www.gartner.com/smarterwithgartner/3-journeys-for-migrating-a-data-center-to-cloud-iaas/. [Accessed: 01-Apr-2021].

[5] I. S. Group and others, "Managing big data for smart grids and smart meters," *IBM Corp. whitepaper (May 2012)*, 2012.

[6] A. Aderibole *et al.*, "Blockchain Technology for Smart Grids: Decentralized NIST Conceptual Model," *IEEE Access*, vol. 8, pp. 43177–43190, 2020, doi: 10.1109/ACCESS.2020.2977149.

[7] N. B. Reinprecht, G. White, and M. Peters, "Enabling European electrical transmission and distribution smart grids by standards," *IBM J. Res. Dev.*, vol. 60, no. 1, pp. 1–3, 2016, doi: 10.1147/JRD.2015.2482878.

[8] J. R. CEO, "Turning Big Data into Clean Electrons at NextEra," *From Transcr. Provid. by Seek. Remarks made 2018 Wolfe Res. Util. Energy Conf. New York, NY https//seekingalpha.com/article/4209709-nextera-energy-inc-nee-ceo-jim-robo-2018-wolfe-research-utilities-and-energy-conference?pag*.

[9] "Understand the propensity of your customers to defect, EDF Energy analyzes its customer base to build its marketing strategy." Published, 2018.

[10] M.-L. Picard, "A smart elephant for a smart-grid:(Electrical) Time-series storage and analytics within hadoop," in *TERATEC Forum*, 2013.

[11] "Romeo project lands in East Anglia ONE and Wikinger." Online: accessed, 2019.

[12] H. Owen, M. Avila, A. Folch, L. Cosculluela, and L. Prieto, "A high performance finite element model for wind farm modeling in forested areas," in *EGU General Assembly Conference Abstracts*, 2015, vol. 17.

[13] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact.," *MIS Q.*, vol. 36, no. 4, 2012.

[14] D. Syed, A. Zainab, S. S. Refaat, H. Abu-Rub, and O. Bouhali, "Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications," *IEEE Access*, pp. 1–1, Nov. 2020, doi: 10.1109/access.2020.3041178.

[15] C. Bisciglia, "The smart grid: Hadoop at the tennessee valley authority (tva)," *Blog Cloudera, Inc., USA*, 2009.

[16] S. Kawasoe, Y. Igarashi, K. Shibayama, Y. Nagashima, and S. Nagashima, "Examples of distributed information platforms constructed by power utilities in Japan," *44th Int. Conf. Large High Volt. Electr. Syst. 2012*, pp. 108–113, 2012.

[17] M. Mayilvaganan and M. Sabitha, "A cloud-based architecture for Big-Data analytics in smart grid: A proposal," in *2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013*, 2013, pp. 1–4, doi: 10.1109/ICCIC.2013.6724168.

[18] Y. Simmhan *et al.*, "Cloud-based software platform for big data analytics in smart grids," *Comput. Sci. Eng.*, vol. 15, no. 4, pp. 38–47, 2013, doi: 10.1109/MCSE.2013.39.

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

[20] R. Kumar and S. Gupta, "Open source infrastructure for cloud computing platform using eucalyptus," *Glob. J. Comput. Technol. Vol*, vol. 1, no. 2, pp. 44–50, 2014.

[21] B. Cheng, S. Longo, F. Cirillo, M. Bauer, and E. Kovacs, "Building a big data platform for smart cities: Experience and lessons from santander," in *2015 IEEE International Congress on Big Data*, 2015, pp. 592–599.

[22] "Aeron open source project." Online: accessed, 2018.

[23] X. Meng *et al.*, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1235–1241, 2016.

[24] A. C. C. Bestavros, L. Hutyra, and E. Terzi, "SCOPE: Smart-city Cloud Based Open Platform and Ecosystem," *Bost. Univ. Boston, MA, USA*, 2016.

[25] D. Puiu *et al.*, "Citypulse: Large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.

[26] "Osborne Clarke: the smart cities law firm." Online: accessed, 2018.

[27] T. Zahariadis *et al.*, "FIWARE lab: managing resources and services in a cloud federation supporting future internet applications," in *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, 2014, pp. 792–799.

[28] K. Wang *et al.*, "Wireless Big Data Computing in Smart Grid," *IEEE Wirel. Commun.*, vol. 24, no. 2, pp. 58–64, 2017, doi: 10.1109/MWC.2017.1600256WC.

[29] Y. Zhou, P. Li, Y. Xiao, A. Masood, Q. Yu, and B. Sheng, "Smart grid data mining and visualization," in *PIC 2016 - Proceedings of the 2016 IEEE International Conference on Progress in Informatics and Computing*, 2017, pp. 536–540, doi: 10.1109/PIC.2016.7949558.

[30] A. Farao, C. Ntantogian, C. Istrate, G. Suciu, and C. Xenakis, "SealedGRID: Scalable, trustEd, and interoperAble pLatform for sEcureD smart GRID," 2019, doi: 10.14236/ewic/icscsr19.10.

[31] M. Li, K. Zhang, J. Liu, H. Gong, and Z. Zhang, "Blockchain-based anomaly detection of electricity consumption in smart grids," *Pattern Recognit. Lett.*, vol. 138, pp. 476–482, Oct. 2020, doi: 10.1016/j.patrec.2020.07.020.

[32] X. Ding, L. Chen, Y. Gao, C. S. Jensen, and H. Bao, "Ultraman: a unified platform for big trajectory data management and analytics," *Proc. VLDB Endow.*, vol. 11, no. 7, pp. 787–799, 2018.

[33] M. Brahem, K. Zeitouni, and L. Yeh, "Astroide: A unified astronomical big data processing engine over spark," *IEEE Trans. Big Data*, 2018.

[34] M. Olson, "Hadoop: Scalable, flexible data storage and analysis," *IQT Quart*, vol. 1, no. 3, pp. 14–18, 2010.

[35] A. Thusoo *et al.*, "Hive: a warehousing solution over a map-reduce framework," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1626–1629, 2009.

[36] A. F. Gates *et al.*, "Building a high-level dataflow system on top of Map-Reduce: the Pig experience," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1414–1425, 2009.

[37] M. Ismail, S. Niazi, M. Ronstrom, S. Haridi, and J. Dowling, "Scaling HDFS to more than 1 million operations per second with HopsFS," in *Proceedings - 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017*, 2017, pp. 683–688, doi: 10.1109/CCGRID.2017.117.

[38] M. R. Palankar, A. Iamnitchi, M. Ripeanu, and S. Garfinkel, "Amazon S3 for science grids: a viable solution?," in *Proceedings of the 2008 international workshop on Data-aware distributed computing*, 2008, pp. 55–64.

[39] M. C. Srivas *et al.*, "Map-reduce ready distributed file system." Google Patents, 2015.

[40] A. Lakshman and P. Malik, "Cassandra - A decentralized structured storage system," *Oper. Syst. Rev.*, vol. 44, no. 2, pp. 35–40, 2010, doi: 10.1145/1773912.1773922.

[41] M. Zaharia *et al.*, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, Nov. 2016, doi: 10.1145/2934664.

[42] A. Zainab, S. S. Refaat, H. Abu-Rub, and O. Bouhali, "Distributed Computing for Smart Meter Data Management for Electrical Utility Applications," in *2020 Cybernetics & Informatics (K&I)*, 2020, pp. 1–6, doi: 10.1109/KI48306.2020.9039899.

[43] A. Reuther, C. Byun, W. Arcand, … D. B.-J. of P. and, and undefined 2018, "Scalable system scheduling for HPC and big data," *Elsevier*.

[44] A. Zainab *et al.*, "A Multiprocessing-based sensitivity analysis of Machine Learning algorithms for Load Forecasting of Electric Power Distribution System," *IEEE Access*, pp. 1–1, 2021, doi: 10.1109/ACCESS.2021.3059730.

[45] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, 2015.

[46] A. Makris, K. Tserpes, V. Andronikou, and D. Anagnostopoulos, "A classification of NoSQL data stores based on key design characteristics," *Procedia Comput. Sci.*, vol. 97, pp. 94–103, 2016.

[47] S. Le, Y. Dong, H. Chen, and K. Furuse, "Balanced Nearest Neighborhood Query in Spatial Database," in *2019 IEEE International Conference on Big Data and Smart Computing, BigComp 2019 - Proceedings*, 2019, pp. 1–4, doi: 10.1109/BIGCOMP.2019.8679425.

[48] H. A. Abdelhafez, "Big Data Technologies and Analytics," in *International Journal of Business Analytics*, 2014, vol. 1, no. 2, pp. 1–17, doi: 10.4018/ijban.2014040101.

[49] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable real-time data systems*. New York; Manning Publications Co., 2015.

[50] Carbone and Asterios Katsifodimos and Stephan Ewen and Volker Markl and Seif Haridi and Kostas Tzoumas, "Apache Flink {^{TM}} : Stream and Batch Processing in a Single Engine Paris," *Undefined*, vol. 36, no. 4, 2016.

[51] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big Data management in smart grid: concepts, requirements and implementation," *J. Big Data*, vol. 4, no. 1, p. 13, 2017, doi: 10.1186/s40537-017-0070-y.

[52] B. Taube, S. G. Solutions, and V. Corporation, "Leveraging big data and real-time analytics to achieve situational awareness for smart grids."

[53] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 674–686, 2017.

[54] G. W. Andersonf, A. Guionnet, and O. Zeitouni, "An introduction to random matrices, volume 118 of Cambridge Studies in Advanced Mathematics." Cambridge University Press, Cambridge New York.

[55] A. Basak and M. Rudelson, "Invertibility of sparse non-Hermitian matrices," *Adv. Math. (N. Y).*, vol. 310, pp. 426–483, 2017, doi: 10.1016/j.aim.2017.02.009.

[56] L. Xie, Y. Chen, and P. R. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2784–2794, 2014, doi: 10.1109/TPWRS.2014.2316476.

[57] S. S. Refaat, H. Abu-Rub, and A. Mohamed, "Big data, better energy management and control decisions for distribution systems in smart grid," in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 2016, pp. 3115–3120, doi: 10.1109/BigData.2016.7840966.

[58] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mob. networks Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[59] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2013, pp. 995–1004, doi: 10.1109/HICSS.2013.645.

[60] L. Wen, K. Zhou, S. Yang, and L. Li, "Compression of smart meter big data: A survey," *Renew. Sustain. Energy Rev.*, vol. 91, pp. 59–69, 2018, doi: 10.1016/j.rser.2018.03.088.

[61] J. Sessa and D. Syed, "Techniques to deal with missing data," in *International Conference on Electronic Devices, Systems, and Applications, Sarawak, Malaysia*, 2017, pp. 1–4, doi: 10.1109/ICEDSA.2016.7818486.

[62] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5820–5830, 2018, doi: 10.1109/TSG.2017.2697440.

[63] N. Mohamed and J. Al-Jaroodi, "Real-time big data analytics: Applications and challenges," in *Proceedings of the 2014 International Conference on High Performance Computing and Simulation, HPCS 2014*, 2014, pp. 305–310, doi: 10.1109/HPCSim.2014.6903700.

[64] A. Zainab, A. Ghrayeb, H. Abu-Rub, S. S. Refaat, and O. Bouhali, "Distributed Tree-based Machine Learning for Short-Term Load Forecasting with Apache Spark," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3072609.

[65] Y. Song, G. Zhou, and Y. Zhu, "Present status and challenges of big data processing in smart grid," *Dianwang Jishu/Power Syst. Technol.*, vol. 37, no. 4, pp. 927–935, 2013.

[66] D. Syed, T.-H. Chang, D. Svetinovic, T. Rahwan, and Z. Aung, "Security for Complex Cyber-Physical and Industrial Control Systems: Current Trends, Limitations, and Challenges.," in *PACIS*, 2017, p. 180.

[67] Z. Dou, I. Khalil, A. Khreishah, and A. Al-Fuqaha, "Robust insider attacks countermeasure for hadoop: Design and implementation," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1874–1885, Jun. 2018, doi: 10.1109/JSYST.2017.2669908.

[68] D. Syed *et al.*, "Deep Learning-based Short-Term Load Forecasting

Approach in Smart Grid with Clustering and Consumption Pattern Recognition," *IEEE Access*, pp. 1–1, 2021, doi: 10.1109/ACCESS.2021.3071654.

[69] A. Zainab, A. Ghrayeb, M. Houchati, S. S. Refaat, and H. Abu-Rub, "Performance Evaluation of Tree-based Models for Big Data Load Forecasting using Randomized Hyperparameter Tuning," 2021, pp. 5332–5339, doi: 10.1109/bigdata50022.2020.9378423.

[70] "MLlib: Main Guide - Spark 3.0.1 Documentation." [Online]. Available: https://spark.apache.org/docs/latest/ml-guide.html. [Accessed: 26-Nov-2020].

[71] M. Hahsler, M. Bolaños, and J. Forrest, "Introduction to stream: An extensible framework for data stream clustering research with R," *J. Stat. Softw.*, vol. 76, no. 1, pp. 1–50, 2017, doi: 10.18637/jss.v076.i14.

[72] A. Bifet, S. Maniu, J. Qian, G. Tian, C. He, and W. Fan, "Streamdm: Advanced data mining in spark streaming," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 1608–1611.

[73] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive Online Analysis," *J. Mach. Learn. Res.*, vol. 11, no. May, pp. 1601–1604, 2010.

[74] G. D. F. Morales and A. Bifet, "SAMOA: scalable advanced massive online analysis.," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 149–153, 2015.

[75] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," *2nd USENIX Work. Hot Top. Cloud Comput. HotCloud 2010*, vol. 10, no. 10–10, p. 95, 2010.

[76] C. Bockermann and H. Blom, "The streams Framework," 2012.

[77] G. Hesse and M. Lorenz, "Conceptual survey on data stream processing systems," in *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, 2016, vol. 2016-Janua, pp. 797–802, doi: 10.1109/ICPADS.2015.106.

[78] "Apache Storm Project." Online: accessed, 2018.

[79] H. C. V. Ngu and J.-H. Huh, "B+-tree construction on massive data with Hadoop," *Cluster Comput.*, vol. 22, no. 1, pp. 1011–1021, 2019.

[80] D. Yang *et al.*, "Fastpm: An approach to pattern matching via distributed stream processing," *Inf. Sci. (Ny).*, vol. 453, pp. 263–280, 2018.

[81] X. Gao and J. Qiu, "Supporting queries and analyses of large-scale social media data with customizable and scalable indexing techniques over NoSQL databases," in *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2014, pp. 587–590.

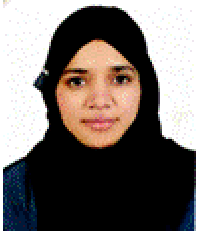
[82] M. Y. Eltabakh, F. Özcan, Y. Sismanis, P. J. Haas, H. Pirahesh, and J. Vondrak, "Eagle-eyed elephant: Split-oriented indexing in Hadoop," in *ACM International Conference Proceeding Series*, 2013, pp. 89–100, doi: 10.1145/2452376.2452388.

[83] "HIndex," 2013. .

[84] J. Dittrich, J.-A. Quiané-Ruiz, S. Richter, S. Schuh, A. Jindal, and J. Schad, "Only aggressive elephants are fast elephants," *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1591–1602, 2012.

[85] S. Nishimura, S. Das, D. Agrawal, and A. El Abbadi, "MD-HBase: A scalable multi-dimensional data infrastructure for location aware services," in *Proceedings - IEEE International Conference on Mobile Data Management*, 2011, vol. 1, pp. 7–16, doi: 10.1109/MDM.2011.41.

[86] J. Dittrich, J. A. Quiané-Ruiz, A. Jindal, Y. Kargin, V. Setty, and J. Schad, "Hadoop++: Making a yellow elephant run like a cheetah (without it even noticing)," *Proc. VLDB Endow.*, vol. 3, no. 1, pp. 518–529, 2010, doi: 10.14778/1920841.1920908.



AMEEMA ZAINAB (Member, IEEE) is a P.h.D candidate in Electrical Engineering from Texas A&M University (TAMU), College Station, TX, USA. She pursued her M.S. degree in data science and engineering from Hamad Bin Khalifa University (HBKU), Qatar. She received her bachelor's degree in electronics and communication engineering from Osmania University, Hyderabad, India, in 2013. She also has industry experience of 3 years working as a Data Analytics professional supporting Audit in Deloitte Touche LLP, Hyderabad, India. She is a base SAS certified programmer. Her research interests include Data Science, Big data machine learning, Power forecasting, and Big data management in the smart grids.

ALI GHRAYEB (Fellow, IEEE) received a Ph.D. degree in electrical engineering from the University of Arizona, Tucson, AZ, USA, in 2000. He was a Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada. He is currently a Professor with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar. His research interests include wireless and mobile communications, physical layer security, massive MIMO, and visible light communications. He served as an Instructor or Co-Instructor in technical tutorials at several major IEEE conferences. He served as the Executive Chair for the 2016 IEEE WCNC conference. He has served on the editorial board of several IEEE and non-IEEE journals.

DABEERUDDIN SYED (Member, IEEE) is a Ph.D. candidate in Electrical Engineering (E.E.) at TAMU, College Station, U.S.A. He received his M.Sc. degree in Data Science & Engineering from HBKU, Qatar in 2018 and a B.E. degree in Electronics and Electrical Engineering from University College, Osmania University (O.U.), India in 2013. He has work experience of 3 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, Artificial Intelligence, and distributed computing.

HAITHAM ABU-RUB (Fellow, IEEE) is currently a Full Professor of electrical engineering (EE) and is holding two Ph.D. s. He has been with many universities in many countries, including Poland, Palestine, the USA, Germany, and Qatar. Since 2006, he has been with Texas A&M University at Qatar (TAMU-Q). He is currently the Managing Director of Smart Grid Center - Extension in Qatar (SGC-Q). His principal research areas are smart grid, power electronic converters, renewable energy, and electric drives. He has published more than four hundred journal and conference papers, 5 books, and 5 book chapters. He has supervised many research projects on the smart grid and renewable energy systems. Dr. Abu-Rub is a recipient of many national and international awards and recognitions. He is the recipient of the American Fulbright Scholarship and the German Alexander von Humboldt Fellowship.

SHADY S. REFAAT (Senior Member, IEEE) obtained the B.A.Sc, M.A.Sc., and Ph.D. degrees in EE in 2002, 2007, and 2013, respectively, all from Cairo University, Giza, Egypt. He has worked in the industry for more than 12 years as Engineering Team Leader, Senior EE, and Electrical Design Engineer. Currently, he is an associate research scientist in the Department of ECEN, TAMU-Q. He is a member of IET and a member of the SGC-Q. He has published more than 100 journal and conference papers. His main work interest includes power systems, electrical machines, smart grid, Big Data, development of fault-tolerant systems, reliability of power grids and electric machinery, fault detection, condition monitoring, and energy management systems.

OTHMANE BOUHALI is a physics research professor at TAMU-Q. His research interests are large scale modeling, high-performance computing, and detector technologies for radiation and medical physics. He has been involved in the Large Hadron Collider research program for more than 25 years and has supervised various research projects. He is the founder and director of the TAMU-Q Advanced Scientific Computing Center.