Machine Learning for Mining Big Data: A Review

Masroor Fattah Bin Hossain
(ID: 011133055)
Abdur Rahman Mamun
(ID: 011131137)
Monika Akter Mishu
(ID: 011131180)
Department of Computer Science and Engineering
United International University

A thesis submitted for the degree of

BSc in Computer Science & Engineering

November 2018

Abstract

Development of Big Data is virtually transforming our lifestyle. It is also accelerating industrial growth through process optimization, insight discovery and improved decision making. The massive scale of big data exceeds the processing and analytic capacity of conventional database systems within an acceptable time frame. Researchers rely on the ability to extract values from such massive data through new data analytics principle; machine learning is at its core because of its ability to learn from data and provide data driven insights, decisions, and predictions. In this study, a review investigation is undertaken for exploring application of machine learning techniques in Big Data analytics of various sectors. We have reviewed 45 papers in the area of machine learning and Big Data analytics involving various sectors such as transportation, healthcare, energy, education, supply chain management, etc. Characteristics and difficulties of Big Data management are reviewed with focus on relevant solutions in order to develop overview for future researchers. We have explored the benefits of machine learning and different machine learning models. We have discussed Big Data systems and programming that is used to process and mine Big Data, and we have also given an overview of how Big Data can be manipulated to generate knowledge. Finally, the conclusion along with future recommendations is provided as a research direction underscoring the importance of inventing more machine learning algorithms, tools and techniques to be prepared for processing and mining Big Data. This review study will provide a head start for future researchers in scanning the appropriate machine learning tools and techniques for mining Big Data.

Acknowledgements

At first, our gratitude goes to our thesis supervisor, Dr. Dewan Md. Farid, Associate professor, Department of Computer Science and Engineering, United International University, for inspiring us to investigate new ideas. Throughout our thesis period, he continuously guided us that helped us to undertake this research in an orderly time efficient manner. I would like to thank Dr. Swakkhar Swatabda, Associate professor, Department of Computer Science and Engineering, United International University, and Dr. khandoker Abdullah Al Mamun, Professor, Department of Computer Science and Engineering, United International University for giving us sufficient inkling about the concept of mining Big Data employing machine learning in order to write the thesis successfully. Moreover, Dr. Salekul Islam, Professor and Head of Computer Science and Engineering department of United International University, discussed with us vividly about the relationship between Big Data and cloud computing that helped us fathom what cloud computing is. So, we are really grateful to Prof. Dr. Salekul Islam.

We would also like to thank our parents for giving us encouragement and support for completing this thesis successfully. Without their encouragement, it would have been a difficult task to complete the thesis timely.

Contents

Chapter 1

Introduction

1.1 Big Data

It has been a practice among the researchers to extract knowledge from data. Knowledge acquisition is important to derive a prognostication for the future. Data can be generated from several sources. When multifarious sources generate data continuously at a time, the datasets tend to grow in size and change frequently. This type of data falls into the term "Big Data". Big Data has become a very burning issue nowadays that interests the computational researchers. At present, it is being applied in many real time applications. Some apparently more illustrious definitions of Big Data are presented below:

- Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [1].

- Big Data is the amount of data beyond the ability of technology to store, manage and process efficiently [2].

- Big Data refers to collection of complex and large data sets whose size is beyond ability of traditional data processing applications and other relational database management tools to process, manage and capture the whole data within the desired span of time and it is basically a general term which is used to describe structured, semi structured and unstructured data [3].

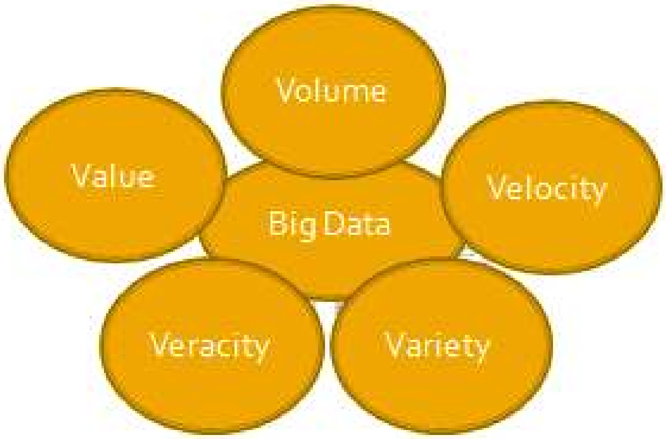

Figure 1.1: 5V's of Big Data.

- Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis [4].

1.2 Characteristics of Big Data

Big Data can be characterized into 5V's as shown in Figure 1.1 [4] [5] [6] [7] [8].

Volume: The size of the data set.

Velocity: The rate at which Big Data moves is used to represent the velocity. Big Data moves at a very high velocity.

Variety: It basically represents the diversity of Big Data. The diversity occurs due to generation of data from numerous sources separately. Numerous sources can be external or internal. There is a minimal chance of external data to be structured.

Veracity: The accuracy of data is used to represent veracity. In order to ensure accuracy, security should be provided, and data should be provided only to authentic people.

Value: It is basically the conversion of data into knowledge for prediction.

1.3 Big Data Systems and Programming

Our literature survey has led us to find that there are several tools that act the role of Big Data systems and programming [9]. They are:

Hadoop: Hadoop uses simple programming models to allow distributed processing of big datasets across clusters of networked computers. It is used to manage content and it is used to detect fraudulency. Hadoop is only suitable for Big Data that uses Java programming language to process Big Data. On top of that, Hadoop is not costly to implement. Thus, it is used extensively over Big Data.

MapReduce: MapReduce is a programming model and framework used by hadoop. It processes huge amount of data in parallel on large clusters of computer nodes. MapReduce can be used for Big Data analysis, and it can be written in any programming language. Moreover, it has a platform Hadoop that enables data to be scaled easily. However, MapReduce is not suitable for real time processing making it difficult to implement.

Weka: Weka is a java based tool for processing large amount of data. It has a vast selection of algorithms that can be used in mining data [10]. The application of Weka can be found in Big Data applied in different areas. It uses Java programming language to process Big Data.

MongoDB: MongoDB is a cross platform document oriented database management system. It uses JSON like documents instead of a table based architecture. MongoDB can be applied in Big Data. It provides options of any programming languages.

Orange: Orange is a tool for processing and mining big data that runs only on python. It has an easy to use interface with drag drop functionalities with variety of add-ons.

1.4 Thesis Contributions

Our thesis is basically a review investigation. We have reviewed 45 journal and conference proceeding papers, and have shared our findings. Moreover, we have provided research directions in order to guide the researchers who are willing to do research on Big Data in the long run.

1.5 Organization of the Thesis

This thesis is written in five chapters. The purpose of each chapter is described below:

In chapter 2, we have focused on modeling Big Data and difficulties in managing Big Data. Moreover, we have provided brief description of Big data tools along with the delineation of Big Data analytics using spark and Hadoop.

In chapter 3, we have focused on what machine learning is and its potential to extract knowledge from Big Data. Different types of machine learning models have been discussed. Moreover, the techniques to evaluate machine learning models have been discussed elaborately.

In chapter 4, we have showcased the application of machine learning with Big Data by presenting some of the papers based on different areas.

In chapter 5, we have provided overall review of Big Data and the challenges of machine learning applications with Big Data. We have discussed our findings. Moreover, the future recommendations for mining and processing Big Data employing machine learning are provided so that higher number of researchers can carry out research with Big Data using machine learning in a more efficient way.

Chapter 2

Big Data

2.1 Big Data Modeling

Data modeling is an act of organization of data so that it can be used for various purposes. Initially, Data must be logically designed in order to associate the related data for various purposes. In order to model the data, SQL is always used to convert logical design into a physical design because it impeccably associates the dataset keys and data types. However, it has always been a challenging task to model Big Data, as it cannot be modeled using SQL due to its humongous size and its changing nature. It is necessary to model Big Data in spite of difficulty in modeling because it facilitates the provision of valuable perspicacity and crucial decision for the future. In order to model Big Data, only data that is related to your research interest should be modelled. Not only that, but also, Big should be created in a system for modeling. Furthermore, researchers should quest for data tools that support Big Data and Hadoop framework.

2.2 Designing a Big Data Management System

Organizing and maintaining data to serve for specific purposes is called data management. The aim of data management is to quest for the infrastructure that supports data. This is like checking out whether statistical computation should be applied on data to obtain result. If we can find the operational requirements for once, the apposite system can be chosen that will allow us to perform the operations. However, it has always been a challenging task to design a Big Data Management system due to its huge size and changing nature. Although the task of managing Big Data is challenging,

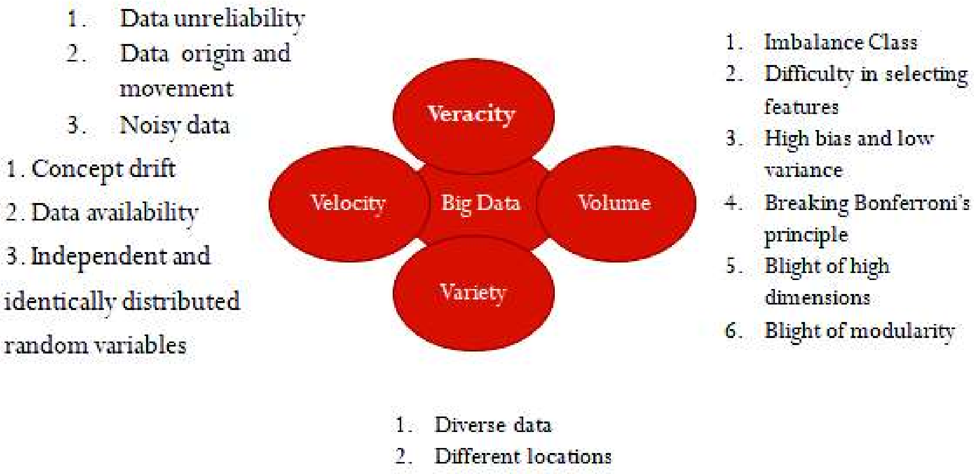
1. Data unreliability
2. Data origin and movement
3. Noisy data
1. Concept drift
2. Data availability
3. Independent and identically distributed random variables
Veracity
Velocity
Big Data
Volume
Variety
1. Imbalance Class
2. Difficulty in selecting features
3. High bias and low variance
4. Breaking Bonferroni's principle
5. Blight of high dimensions
6. Blight of modularity
1. Diverse data
2. Different locations

Figure 2.1: Challenges of 4v's of Big Data.

there are some solutions that can minimize the challenges to some extent. In subsection 2.2.1, the challenges of managing Big Data are presented below with respect to 4Vs as shown in Figure 2.1, and in subsection 2.2.2 , the solutions to minimize the challenges of managing Big Data to some extent are presented.

2.2.1 Challenges of managing Big Data with respect to 4V's

Four dimensions of Big Data provide many challenges as shown in figure 2.1 [11]. Some of the challenges are presented below :

Challenges of volume of Big Data

Imbalance class: Big Data is not usually uniformly distributed across all the classes.

Difficulty in processing data: The humongous size of Big Data makes it difficult for Big Data to get processed .

Non linearity problem: Big Data tends to be non linear forming a gigantic cloud due to its many points.

Difficulty in selecting features: Selection of features has always been difficult due to its high dimensionality.

High Bias and low variance: Variance and Bias error in machine learning should be optimal if we want to get a corrected desired output. But, in case of Big Data, the bias error tends to increase.

Breaking Bonferronis principle: The principle declared by Bonferroni tells us that one can always predict the event from data. But in case of Big Data, it is not always true because Big Data is usually generated in different formats from various sources at a high velocity. The formats can be structured or unstructured. So, the assumption of Boneferronis principle does not hold in case of Big Data.

Blight of high dimensions: Big Data tends to be high dimensional. High dimensional means it contains numerous features.

Blight of modularity: Big Data, even if it is processed, it cannot be kept entirely in memory.

Challenges of Velocity of Big Data

Concept drift: Big Data is mobile i.e it is moving continuously. The entire dataset cannot be obtained so easily before processing. Thus, it is difficult to understand the similarity of current data distribution with future data distribution.

Data availability: Big Data cannot be acquired easily due to its continuous nature of travelling at high velocity.

Independent and identically distributed random variables: The generation of Big Data from different sources are not random, independent and identically distributed.

Challenges of veracity of Big Data

Data unreliability: The ways of accumulating Big Data from different sources can add unreliability in the dataset.

Data origin and movement: It is difficult to detect the source and movement of Big data between locations due to its huge size and complexity.

Noisy data: Big Data tends to be noisy due to its generation from different sources.

Challenges of Variety of Big Data

Diverse data: Big Data can have differences in formats, types and features due to its generation from various sources. It is difficult to process data containing different formats, types and features.

Different locations: Big Data is usually located in different files containing different formats. Thus, it is too difficult to bring entire datasets into memory.

2.2.2 Solutions to minimize the challenges of managing Big Data

There are steps that can be undertaken to minimize the challenges of managing Big Data to some extent. The two ways are:

Imitating non-Big Data: Transform Big Data in such a way so that it emulates non-big data. This can be done in two ways. They are:

-Dimensionality reduction: This transforms high dimensional data into low dimensional data. Dimensionality reduction can be acquired using principal component analysis and auto encoders.

-Instance selection: This deals with the selection of some datasets that indicate the whole dataset. Instance selection can be acquired by cluster sampling, random selection from domain knowledge and genetic algorithm-based selection.

Remove noise from data: There can be several missing values in Big Data due to its generation from any data sources. The presence of missing values adds to unreliability in the dataset. Thus, the missing values should be removed in order to ensure certainty in the dataset. This can be accomplished using autoencoders.

Vertical Scaling: Vertical scaling is used to abate the difficulty of processing Big Data. Vertical scaling is the embodiment of adding more power to CPUs.

Horizontal scaling: Horizontal scaling is another technique that is used to abate the difficulty of processing Big Data. Horizontal scaling is the embodiment of adding more CPUs to process Big Data. Horizontal scaling is of two types. The two types are:

-Batch oriented system: It basically emphasizes on throughput than latency. It processes data in batches.

-Stream-oriented system: It basically functions on one data element or a small set of recent data in real-time or near real-time.

Modification of Algorithms: Algorithms are needed to be modified to manage Big Data so that they perform well on Big Data. For example, optimized version of SVM was propounded by Pegasos for scaling large text [12].

2.3 Big Data Tools

There are several Big Data tools that exist to process and mine Big Data. In section 2.3.1, we have delineated Big Data analytics using Spark, and in section 2.3.2, we have delineated Big Data analytics using Hadoop.

2.3.1 Big Data analytics using Spark

Big Data analytics becomes easier with Spark, as Spark divides the dataset into several clusters to perform Big Data analytics.Spark makes data analysis easier, and it provides result quickly. Not only, it facilitates the provision of build in libraries to perform analytics, but also, it provides an option for many programming languages. Moreover, Spark provides an opportunity to implement diverse Hadoop applications that are used for processing and storing Big Data. Big Data analytics using Spark can be used to detect fraudulency.

2.3.2 Big Data analytics using Hadoop

Hadoop is commonly used nowadays to perform Big Data analytics. It utilizes simple programming models to allow distributed processing of big datasets across clusters of networked computers. Hadoop distributed file system and MapReduce are two components of Hadoop that are primarily responsible for performing analytics over Big Data. Hadoop distributed file system is used for storing huge amount of data. It splits the data into several blocks, and then allocate the nodes into disparate groups. In addition to that, the fault tolerance concept of Hadoop distributed file system allows replica of different nodes fallen into divergent groups in order to preclude data from getting lost. MapReduce is used for processing huge amount. MapReduce performs two operations in a sequestered fashion sequentially. First, it does the operation of mapping. During Mapping, the required data is fetched from a cluster.After mapping, it does the operation of reducing. During reducing, all data points are fetched, and merged to provide predictions from Big Data.

Chapter 3

Machine Learning

It has been a practice among the computational researchers to accumulate, store and interrelate huge amount of data, and acquiesce knowledge from it. But, when data spirals out of control falling into a term “Big Data”, it becomes difficult to derive perspicacity from data. Very recently, the researchers have begun to realize the power of machine learning in deriving knowledge from Big Data. So,they have started applying machine learning with Big Data. In section 3.1, we have introduced the term “machine learning”, and discussed the types of machine learning that are used for mining Big Data. Moreover, we have discussed different evaluation methods to evaluate the machine learning models in section 3.2.

3.1 Machine Learning

Machine learning manages and analyzes data. Moreover, it helps the companies in revealing hidden patterns from intricate data. The concept of machine learning is being applied in numerous areas like healthcare, bioinformatics, transportation, smart city, etc. Machine Learning is primarily of two types. They are:

3.1.1 Supervised Learning

Supervised learning is basically a synonym of classification. The supervision in the learning comes from the labeled examples in the training data set [13]. There are two types of supervised learning. They are:

-Classification: Classification is a type of machine learning models that falls into the term “supervised learning”. It basically deals with labelled nominal data to predict data for the future. KNN, naive bayes classifier, decision tree, neural network and random forest are some of the common classification algorithms that are used to derive knowledge from data.

-Regression: Regression is a type of machine learning models that falls into the term “supervised learning”. It basically deals with labelled numeric data to predict data for the future.

3.1.2 Unsupervised Learning

Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled [13]. It basically deals with non-labelled data to assign synonymous data into similar groups and different data into different groups. K-means clustering and hierarchical clustering are used extensively for clustering data.

3.2 Evaluation of Machine Learning Models

It is important for machine learning models to be evaluated once it is trained to ensure whether it is working correctly or not. If the models are not evaluated, there is a high probable chance that the models can be over fitted or under fitted. The following below is the techniques that are used to evaluate the machine learning models. The techniques are:

Classification accuracy(ACC): The number of correct predictions by total number of input samples is called classification accuracy. Classification accuracy can also termed as accuracy. The formula of classification accuracy can be written as:

$$Acc = NCP(Numberofaccuratepredictions)/TN(Totalnumberofsamples)) \tag{3.1}$$

Confusion matrix(CAC): It facilities the provision of output matrix along with the delineation of entire execution of the model. The formula of confusion matrix can be written as:

$$CAC(Confusionmatrixaccuracy) = TP{+}FN/TNS(Totalnumberofsamples) \tag{3.2}$$

Confusion matrix allows us to get familiarized with four crucial terms. The four terms are:

-True negative(TN): The situations when our prognostication is no and the real output is yes.

-True positive(TP): The situations when our prognostication is yes and the real output is also yes.

-False negative(FN): The situations when our prognostication is no and the real output is also no.

-False positive(FP): The situations when our prognostication is yes and the real output is also no.

Area under the curve(AUC): It is the representation of area under the curve of False Positive Rate(FPR) versus True Positive Rate(TPR) at disparate data points, where the formula of TPR and FPR can be computed as:

$$TPR = TP(Truepositive)/FN(FalseNegative) + TP(Truepositive) \tag{3.3}$$

$$FPR = FP(Falsepositive)/TN(TrueNegative) + FP(Falsepositive) \tag{3.4}$$

Train/test/validation split: In this evaluation technique, the dataset is divided into training and testing. The training data is then divided into training and validation. After the division, random instances are selected and are trained using a classification algorithm.

F1 score: Balanced mean between precision(prec) and recall(rec) is used to define F1 score. The formula of F1 score can be written as:

$$F1score = 2 * 1/1/prec + 1/rec \tag{3.5}$$

Here, prec is called precision and rec is called recall.

-Precision(Prec): Number of positive accurate results by number of predicted positive results from classifier. The formula of precision can be written as:

$$Prec = TP(Truepositive)/TP(Truepositive) + FP(Falsepositive) \tag{3.6}$$

-Recall(Rec): Number of positive accurate results by number of related samples(Samples must be positive). The formula of recall can be written as:

$$Rec = TP(Truepositive)/TP(Truepositive) + FN(Falsenegative) \tag{3.7}$$

Mean absolute error(MAE): Summation of difference between actual value(yactual) and predicted value(ypredicted) by number of training examples(NT) represents the mean absolute error. The formula of mean absolute error can be written as:

$$MAE = \sum(yactual - ypredicted)/NT \tag{3.8}$$

Mean square error(MSE): Summation of squared difference between actual value(yactual) and predicted value(ypredicted) by number of training examples(NS) represents the mean square error. The formula of mean square error can be written as:

$$MSE = \sum((yactual - ypredicted) * (yactual - ypredicted))/NS \tag{3.9}$$

Chapter 4

Machine Learning with Big Data

We have reviewed 45 papers garnered from various journals and conferences related to machine learning with Big Data. Review of the papers are presented and grouped according to the research application areas. The following below is a review of some of the papers pertaining to transportation, health care, smart city, education, smart grid, electricity, company and customer gender analysis, data science, Big Data security, and energy awareness.

4.1 Transportation

Intelligent or smart transportation system is crucial for sustainable development of a city. If not, we cannot expect our lives to ameliorate. However, with the emergence of Big Data, we can expect our city to have better transportation. The following below is the review of some of the papers pertaining to transportation Big Data. They are described below:

The applications of Big Data on transportation system have become prevalent nowadays to provide smart transportation system. This system facilitates in diminishing traffic gridlock. Not only that, but also this system subsides carbon emission. There are numerous models that exist to predict the flow of traffic,but nevertheless, the models aren't still robust enough to provide more accurate prediction. In 2015, a deep learning approach was propounded to predict the flow of traffic [14]. This deep learning method embodies the utilization of eminent model called stacked auto encoder(SAE). This model comprises of a stack of autoencoders which is basically a deep network formed from considering the outputs of autoencoder, a neural network as present input layers. The model uses logistic regression to foretell the flow of traffic.

The applications of Big Data in social transportation facilitate in diminishing the shortcomings of transportation system. In 3rd March, 2016, Big Data for social transportation was explored deeply [15]. Social transportation Big Data can be based on anything pertaining to transportation. Each data is liable for performing a specific task as for example, taxi GPS data can also fall into a term ”Social transportation Big Data” that can be utilized to measure travel time on roads in real time. There are three main benefits of performing data analytics over social transportation data. They are:

- If social transportation data is analyzed, there is a very high probable chance of intelligent transportation system to ameliorate i.e the model can be able to interpret people’s emotion and then model people’s behavior.

- The data analytics over social transportation data can be able to act as a medium for informing irregular schedules, waiting time and alarm.

- In addition to that, the analysis of social transportation data can be able to help the government fathom people’s problems, and by that, he/she can solve transportation problems.

Sundry models pertaining to machine learning are accessible to foretell the flow of traffic, but none of them provide prognostication on time. In 2017, a DeepTFP model was propounded to provide on time prognostication of traffic data [16]. The proposed method performs two tasks. Initially, it applies deep learning to mine generate attributes from data, and then, it applies time series function to foretell the flow of traffic from sequential data at every time interval.

Intelligent transportation system based on Big Data helps in detecting the number of people moving from one place to another. Not only that, but also, it can also identify and predict whether traffic congestion has taken place nearby shopping centres, airport, etc. In order to facilitate these services, ShenZhen Transportation system was propounded in 2015 [17]. The system runs on Hadoop, and it is based on authentic data of ShenZhen pertaining to transportation.

Smart vehicles with capability of connecting and exchanging information with their surrounding vehicles leads to rapid development of automative telematics. Resulting efficient and intelligent future transport system needs to expand the network scale and conduct dynamic continuous information processing emerging to the concept of internet of vehicles. Wenchao et al. [18] very recently has investigated the introduction of Big Data concepts in the areas of IOV(Internet of Vehicles) focusing on data transmission, storage, computation, IOV characterization, performance evaluation and communication protocol development in vehicular environment. For IOV Big Data sourcing, the authors have mentioned the possibilities of satellite networks, high altitude platform, near ground unmanned aerial vehicles, terrestrial networks, vehicular social networks and vehicular sensor networks. The study describes the Big Data generation sources from on-board and on-road sensing with probable range of respective bandwidths. The on-board sensed data items includes vehicle position, data from global positioning system(GPS), driving state data such as (velocity, tire pressure, etc,), engine controlled data such as (RPM and Coolen temperature),accident event data(accident, collision and brake). The on-road sensing technology includes rudder, sonar, leader and camera. Using the accumulated Big data, the study explores IOV applications in terms of vehicle management through ubiquitous connections, in-vehicle high definition video streaming through large bandwidth, autonomous driving through faster both way data communication and high definition map with high data volume. The study has also suggested wide range of MAC protocols to minimize the access time and transmission interference for ameliorating the throughput, robustness and scalability. Furthermore, in order to mitigate the challenges of making optimal routing strategy for Big Data transmissions, the study has suggested two routing protocols namely:topology-based routing protocols and position based routing protocols. The study concluded that with some emerging issues and desired directions of potential IOV and Big Data synchronization.

4.2 Healthcare

Smart health care is needed to cure diseases, and make people conscientious about their health. If not, people won't get proper treatment, and our health care sector will aggravate. However, with the revolution of Big Data, we can expect for a better healthcare. The following below is the review of some of our papers pertaining to healthcare that how Big Data can improve our healthcare.

The increased reliance on technology has led to widespread utilization of Big Data and Internet of things in health care. This modern technique has aided the doctors in promulgating accurately the disease of patients. In 2018, the prime issues of Internet of things germane to smart sensors implemented in health care were mentioned [19]. The Internet of things related to smart sensors embodies Wearable and body sensors, digitalized healthcare systems, and Big Data analytics principle i.e machine learning to apprise the sensor devices.

In 2013, the numerous applications of Big Data to healthcare were explored [20]. In doing so, it was found that there are many benefits of applying Big Data to healthcare. Some of the benefits are :

- Big Data facilitates in finding the diseases of targeted patients, as it can interconnect conventional medical data of patients with private data of patients.

- The provision of knowledge to doctors how treatment can be provided to patients through Big Data.

Big Data containing details of sick ones increases in size and changes continuously. Therefore, doctors find challenging to know the health conditions of sick ones. The issues to solve the challenges of Big Data have become a burning issue nowadays. In 2016, a framework was propounded that can minimize the challenges of health care Big Data [21]. The framework is based on machine learning principles containing different parts that are responsible for performing different tasks pertaining to applications of cloud computing, Big Data analytics, sensing technologies and internet of things. In this way, doctors can cure the sick ones as the proposed technique helps in managing and dissecting data of sick ones.

The materialization of big data in health care has subsided the extortionate health care prices. Owing to that, very recently, USA has started implementing big data in health care to minimize the health care costs in USA. After that the remarkable amelioration of health care system can be seen in USA. The six advantages of minimization of health prices employing big data were illustrated in 2014 [22]. On top of that, the acquisition of kinds of perspicacity from mining big data related to sick ones were mentioned. Furthermore, the models needed to mine big data related to sick ones were vividly described in the paper. At present, the procedures that are used to store, process and mine healthcare Big Data for diagnosis are not up to the mark. The challenges of mining Big Data still exist today. In 2014, Big Data analytics framework for healthcare was propounded [23]. The model stores, processes and mines medical Big Data in five stages. Initially, the model covers issues of collection and storage of health care Big Data. Secondly, it maintains the stored data that is processed. Thirdly, it converts data into a specific format for dissecting data. Fourthly, it manipulates algorithms over medical data. Finally, it extracts knowledge from data.

Reviewing the prospect of Big Data applications in health sector, Tao Huang et.al. claimed that The augmentation in number of smart devices and cloud computing has led to a significant increase in accumulation of health data. They explored the contemporary usages of Big Data in healthcare such as digital epidemic monitoring,application of sensors in health monitoring, inspection of food safety, bioinformatics, inspection of air quality, etc [24]. Moreover, they explored the contemporary techniques of Big Data accumulation, Big Data retention, Big Data transfer process, and application of smart analytical procedures. The study highlighted the importance of fusion between software and hardware referring to various sensors which create numerous useful applications for inspecting health condition and food safety. At the end of the study, they highlighted the future prospect of personalized health data generated from millions of personal smart phones in terms of personal health management, accumulated Big Data analysis and data privacy. Since the study, we can see the manifestation of the authors forecast in the areas of bioinformatics and personal health management developments centering around Big Data generation.

4.3 Smart City

Smart city is very important for sustainable development of a country. With smart city, we can expect our education, health care, transportation and economy to ameliorate. Smart city can be obtained using Big Data. The following below is a review of our some of our papers related to applications of Big Data in Smart City. They are:

With the expansion of digital technology, the applications of Big Data to obtain smart city has increased. In 2015, the challenges and opportunities of applying Big Data to smart cities were mentioned. Moreover, the requirements are needed to be fulfilled for applying Big Data to smart cities were delineated [25]. The following below is the description of challenges and opportunities of applying Big Data to smart cities along with requirements needed to be fulfilled for applying Big Data to smart cities.

Requirements for applying Big Data to smart cities: Appropriate tools are necessary for the utilization of Big Data in smart cities as Big Data applied in smart cities tend to be gigantic with different formats. The conventional algorithms designed for mining small datasets should be adapted to perform well on large datasets like Big Data. Correct and safe utilization of ICT solutions for smart city should not be made oblivious to the citizens . The provision of information pertaining to various events from the citizens confrontation with applications of smart city can act as an aid for improving the data quality that was accumulated. Hence, garnered Big Data can provide better decisions to amplify various components of smart city. Smart networks are required for associating different parts of smart city. The network must have the ability to send information between the components. Furthermore, effective and reliable interaction must be set by governing entities of smart cities so that Big Data can be transferred between various components safely without breaching peoples privacy.

Challenges of applying Big Data smart cities: Data distribution among various city departments can breach peoples privacy because every city department has its own data warehouse containing detailed description about its own people. So, it is challenging task to maintain the synchronization of accumulating and utilizing Big Data to smart cities and checking peoples privacy. Moreover, the comprehension of data semantics can also be difficult because incomplete data can be updated quickly from various distributed data sources that are responsible for generating Big Data. It is difficult to assess quality of data in Big Data because data is gathered by various people at different times and it is difficult to converse with different entities of smart cities data policy and make citizens understand every time the data analysis and its corresponding outputs. Furthermore, high levels of security policies are required to ensure security of Big Data because Big Data can contain secret information pertaining to people, and people have rights to maintain their privacy.

Opportunities of applying Big Data to smart cities: Applications of Big Data to smart cities can introduce smart traffic light that can control huge traffic gridlock. Smart grid can be acquired by applying Big Data to smart cities. Smart grid is a revamped electrical grid system that utilizes information and communication technology from the data. It leads to augmentation in effectiveness and reliability of electric power.

It is difficult to make businesses related to smart city prosperous without the applications of Big Data. Thus, in 2013, it was propounded that Big Data can be manipulated to provide smart city using the concept of API [26]. The conditions needed to obtain smart city were also illustrated .

The increased volume of number of people makes it difficult to meet the needs of the citizens. Smart technologies that run on internet must be provided to interact with the needy ones. In order to facilitate that service, in 2015, IV-tier model was propounded [27]. The model uses Big Data analytics principles i.e machine learning to conduct planning of making smart city. The model consists of four tiers.The first tiers embodies the issues of sources of data, data creation and data accumulation. The second tier embodies the issues of interaction between different sensors. The third tier embodies the issues of managing and processing Big Data employing Hadoop framwork. Finally, the last tier embodies the issues of dissecting and predicting result from data.

Energy consumption is a concerning issue for global communication network among millions of Internet enabled devices popularly known as Internet of Things(IOT). Murad et al. [28] addressed this issue through suggesting an energy-aware communication network for IOT environment. The research propounded for sensor network for identification, quantification and relevant data generation from IOT devices manifested by recently evolved concept of smart home, office and city. The proposed model uses Hadoop system for data controlling and storing job of the data management. The author used adaptive job scheduling mechanism to load the MapReduce system in real time which divides specific task into sub-tasks to assign among the Hadoop cluster system. In order to stabilize the High end low performance nodes into Hadoop ecosystem. The task scheduler assigns tasks according to present workload of each node. The propounded framework is experimented in a smart home condition comprising of twenty six sensors connected home devices. The experiment revealed that proposed energy-aware scheduling mechanism can mitigate excess energy consumption of the smart home appliances. The experimental data are then analyzed in Hadoop system, and the result shows that propounded scheme works as hypothesized. This paper shows an encouraging result in terms of energy-aware and energy-efficient sensor based IOT environment which opens up the possibility of further applied research on IOT environment oriented smart city, offices and homes.

4.4 Education

Online activities of students generate Big Data, mining which, can pave the way for smart learning management. Katrina and Lognathan [9] reviewed recent uses of Big Data techniques in education and learning analytics. They mentioned four commonly used educational data mining techniques as regression, nearest neighbor, clustering and classification. It is also claimed that significant number of open source tools are available for educational Big Data analytics mentioning MongoDB, Hadoop, MapReduce, Orange and Weka as some of the top tools. Additionally, SAP HANA was mentioned as a proprietary tool capable of utilizing parallel in memory relational query techniques, columnar stores and compression technology. Again, the study highlighted application area of Big Data learning analytics such as students performance prognostication, attrition risk detection, data visualization, intelligent feedback and course recommendation.

4.5 Data Science

The term “Data Science” has become a burning issue nowadays. Especially, the revolution of Big Data always interests the computational researchers though it is challenging to manage Big Data. The following below is the review of some of the papers related to Data Science. They are:

Big Data has always attracted the researchers ever since its evolution. With the manipulation and proper management of Big Data, the valuable perspicacity related to data can be derived easily. Thus, the challenges related to Big Data must be diminished by hooks or crooks. In order to do that, Xindong Wu et al. [1] presented a HACE theorem where Big Data is compared with an elephant and a data sources are compared with blind men. They mentioned that data can also grow from small to Big like elephant which can be generated from various sources that are independent of each other. They also asserted that every source is restrained to its own view like blind man viewing nose instead of a tail i.e one source cannot view data generated from another source. Furthermore, they explored Big Data processing model where the challenges of processing and storing Big Data were discussed. The model comprises of three tiers. The tier 1 underscores the difficulties of accessing data. The tier 2 focuses on the issues of data privacy and difficulties of designing algorithms from the domain knowledge of data. The tier 3 focuses on the challenges of mining Big Data. They claimed that high performance computing platform is required to store and manage Big Data.

The usage of current data to derive a perspicacity for the future event has always enticed the researchers. Thus, the term ” Big Data ” can be heard very frequently nowadays.In delineating Big Data, Hakan et al. [29] elaborated on the definitions of Big Data, classification of Big Data, operations of Big Data and applications of Big Data.Moreover, they also discussed the past and future directions of Big Data. The past direction showed that there was no enthusiasm in applying Big Data between 1999 to 2011, but after 2011, the interest in Big Data started increasing significantly. They also mentioned that in the coming future, India will have the highest interest in Big Data, and Singapore will have the second highest interest in Big Data. After reviewing this paper, we believe that people are not at all lackadaisical about the application of Big Data in real life scenarios even in 2018.

Manipulating Big Data to generate knowledge has always been a challenging task. In order to diminish the challenges to some extant, Alexandra L'heureux et.l [11] delineated the challenges of processing and storing Big Data with respect to 4V'S along with some possible solutions. They claimed that there should be better techniques to manipulate Big Data into following Bonferroni's principles.

Hang and Simon [30] have solved interference problem of real time online data feeds employing a new incremental decision tree algorithm named incrementally optimized very very fast decision tree(IOVFDT). They claimed that IOVFDT can perform very well in terms of accuracy and diminutive model size.

Albert Bifet [31] analyzed the present and succeeding challenges of mining Big Data in real time and the solutions precursor to mining Big Data. They showed that there are many difficulties as well as benefits for mining Big Data in real time as datasets tend to enlarge incessantly. Jinlong Wang [32] discussed the improvement of data mining over the years as well as its roles, algorithms and data mining procedures. They claimed that there should be more focus on modifications of data mining techniques to ameliorate the technology of data science. Katarina Grolinger et.l [33] discussed the challenges of MapReduce for Big Data. The challenges are categorized into four categories i.e problems of data storage, challenges of Big Data analytics, challenges of online processing, problems of security and problems of privacy.

4.6 Security of Big Data

Big Data lacks security due to its generation from different sources. It is challenging to maintain security of Big Data. The following below is some of the review papers related to security of Big Data. They are:

It is difficult to capture, process and manage Big Data. Not only that, but also, it exceeds the capacity of conventional database management system for storing and managing Big Data.Thus, Big Data provides security problems to the users as it may contain confidential information of different citizens. Owing to that, Neetu and Dr. Satyajee [34] explored the security issues of Big Data and proposed approaches that can diminish the security problems of Big Data to some extant. They claimed that there should be some changes in techniques as to ensure security of Big Data.

It is difficult to make Big Data secured with current techniques. Thus, Vijey and Aiiad [35] discussed a new method called quantum cryptography. The propounded method uses PairHand protocol to minimize the security problems of Big Data. They claimed that the security of Big Data depends on disparate factors like traffic and size of data.

Cloud computing is used extensively to store and process Big Data. It is difficult to maintain the security of Big Data in cloud computing area. The conventional method named encryption does not work well in securing Big Data. Gunasekaran et al. [36] discussed the challenges of storing Big Data in cloud area, and provided possible approaches that can diminish the challenges of security problems of Big Data in cloud areas. Furthermore, in order to address the challenges, the framework named Meta-Cloud data storage architecture was propounded to preserve Big Data in cloud area. They showed that we can effectively manage Big Data in cloud area using the proposed model.

The applications of Hadoop can be used to store and process Big Data. The security issues of Hadoop applications in Big Data still remain a big problem. Thus, to secure Big Data applications in Hadoop to some extant, B. Saraladevi et al. [37] focused on the security issues of Big Data in Hadoop. Moreover, they provided solutions to secure Big Data in Hadoop. They claimed that Big Data can be secured in Hadoop utilizing kerberos, algorithms and name nodes.

Aayush et al. [38] propounded a model i.e a mixture of role base access control and Hadoop distributed file system to provide security of Big Data. They proved that the storage system of a file modifies each time when the file is accessed. This leads to secured storage of Big Data in cloud area.

Pietro and Elena [39] in order to make aware of security problems of Big Data, delineated the research issues pertaining to model that minimizes the security problems of Big Data. They claimed that application of this model is similar to where MapReduce and Nosql are utilized.

4.7 Smart Job

The growth of Internet has enabled many web portals to automate the hiring processes of employees. This automatic process of spreading job offer and recruiting new people for the job is called smart job. Without smart job, it is difficult for the job seekers to look for the job that is appropriate for them. Sidahmed et al. [40] propounded Big Data analytics framework that utilizes time series forecasting and semantic classification for intelligent job offers. They confirmed that the model performed well in offering the jobs.

4.8 Electricity

Machine Learning with Big Data can be used to predict the amount of power required to generate electricity in the city. It is known to us all that machine learning can be applied in Big Data to create powerful prognostication model. Thus, Mohammad et al. [41] propounded an electricity generation forecasting system that can predict the power required to generate electricity. The forecasting system uses Hadoop application to predict the power needed to generate electricity. The application stores the data in a hadoop cluster. The data is then distributed into several nodes using parallel programming under the MapReduce. Artificial neural network is then applied on separate datasets stored in individual nodes to generate the result. They claimed that applications of machine learning with Big Data is a good approach to predict the amount of power required in the city.

4.9 Smart grid

One of the significant application area of Big Data analytics is dynamic management of power grid leading to the concept of smart and intelligent grid management system. In smart grid system, large pool of sensors is integrated in the power grid system for collecting all sorts of grid characteristics data which results enormous data volume and analytics of this Big Data leads to automated intelligent management of grid system.

Presenting Apache Spark as a unified computing platform suitable for Big Data analytics on grid data, Shyam et al. [42] claimed that a demand responsive real time pricing of smart grid is feasible through this platform. Suggesting data integration schemes for sensor generated four different types of heterogenous data covering generation, transmission, distribution and consumption, the authors propounded three main catergories of data processing methodologies such as batch, stream and iterative processing. Thus, Apache Spark is claimed to be an integrated platform that synchronizes batch, real time and iterative data processing for delivering advanced analytics and machine learning approaches for power grid Big Data with useful applications such as real time pricing, automated demand response, pick hour load balancing, fault monitoring and online grid operation management.

4.10 Energy consumption

One of the drawbacks of using Big Data in supercomputers is that it wastes huge amount of energy. In order to diminish the rate of energy usage, there should be methodologies that preserve energy as much as possible. For that, the effective techniques such as hardware and software are needed to save energy. There has been considerable development in hardware section, but nevertheless, there isn't still sufficient amelioration of effectiveness of software. That is primarily due to devoid of tools and systems that are needed to conserve energy. In order to shorten the differences between the efficiency of hardware and software in saving energy, Ziliang et al. [43] propounded Marcher: A heterogeneous system to resolve the maximum energy required to perform Big Data analytics in supercomputers. The propounded Marcher facilitates the services of Intel Xeon CPUs, Intel Many Integrated Cores, Nvidia GPUs, power-aware memory systems and hybrid storage with Hard Disk Drives (HDDs) and Solid State Disks (SSDs). The framework can compute the capacity of CPUs, DRAMs, disks, accelerators and coprocessors. Moreover, effective energy and education can be ensured on multiple programs using the model. There are three layers of Marcher covering top layer, middle layer and bottom layer. The device allows the programmers to create and keep the program. Moreover, the model facilitates the provision of various programming languages and mainstream parallel programming models. Currently, Texas, United States is already using the model, and the authors claimed that the model works quite well.

4.11 Company and Customer gender analysis

Company: Companies usually contain details of various customers and stakeholders who are associated with the company. It is really important to know the details of every customer who is buying his/her product. Thus, it is crucial to manipulate information to convert into knowledge for the benefits of the company. In order to do that, Milan et al. [44] discussed how Big Data can be manipulated for the benefits of a company. They claimed that Big Data can provide entire picture of a customer who is buying a product from a particular company. Knowing the gender of customers plays a seminal role in finding the information about shoppers.

Customer gender analysis: Knowing the gender of customers taking online services plays a seminal role in finding the information about shoppers because this provides more description about purchasers. If sellers acquiesce more information, they will become more benefited. But, customers usually do not provide such confidential information. In order to know the gender of clients, Duc Duong et al. [45] used machine learning techniques and found the attributes to predict the gender of buyers. They confirmed that the proposed system works well in terms of finding the gender of buyers

Chapter 5

Conclusions

5.1 Conclusion

Throughout the journey of this thesis study, we have realized the potential and challenges of Big Data analytics. We have been introduced with different types of machine learning models i.e. supervised learning and unsupervised learning. The motivation of this thesis is to explore the concept and characteristics of Big Data and to explore deeply on modeling and managing Big data for mining Big Data employing machine learning. In our thesis, we have reviewed 45 papers. Our literature search led us to find various definitions, characteristics of Big Data and different application areas of Big Data. We have listed Big Data systems i.e. tools and techniques for processing and mining Big Data. Furthermore, we have discussed how Big data tools i.e Spark and Hadoop make processing and mining Big Data easier.What we have found from our thesis is that, there should be more improvement on data mining methods to process and mine Big Data. Adaptive algorithms like incrementally optimized decision tree are required to process and mine Big Data. High performance computing platform is required for performing Big Data analytics in supercomputers, and Deep learning can be used to mine Big Data.

5.2 Future Work

In order to work further on machine learning applications in Big Data, the usage of deep learning for mining Big Data can be a good option because deep learning abates the difficulty of selecting features. Moreover, researchers can check our review of 45 contemporary papers. There are many new machine learning models that have been propounded in the papers, we have reviewed, but at the same time, the study also has some limitations. Therefore, more conference and journal papers are required to be reviewed that we could not garner due to shortage of time. So, the future researchers should focus on collecting more papers that underscores better machine learning platforms for processing and mining Big Data.

Bibliography

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014. 1, 23

[2] J. Gantz and D. Reinsel, "Extracting value from chaos," *IDC iview*, vol. 1142, no. 2011, pp. 1–12, 2011. 1

[3] N. Garg, S. Singla, and D. S. Jangra, "Challenges and techniques for testing of big data," *Procedia Computer Science*, vol. 85, pp. 940–948, 2016. 1

[4] I. A. T. Hashem, I. Yaqoob, N. B. A. S. Mokhtar, A. Gani, and S. U. Khan, "The rise of big data on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015. 2

[5] A. Elragal, "Erp and big data: the inept couple," *Procedia Technology*, vol. 16, pp. 242–249, 2014. 2

[6] S. O. Fadiya, S. Saydam, and V. V. Zira, "Advancing big data for humanitarian needs," *Procedia Engineering*, vol. 78, pp. 88–95, 2014. 2

[7] C. Yang, X. Zhang, C. Zhong, C. Liu, J. Pei, K. Ramamohanarao, and J. Chen, "A spatiotemporal compression based approach for efficient big data processing on cloud," *Journal of Computer and System Sciences*, vol. 80, no. 8, pp. 1563–1583, 2014. 2

[8] V. Lopez, S. del Rio, J. M. Benitez, and F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data," *Fuzzy Sets and Systems*, vol. 258, pp. 5–38, 2015. 2

[9] K. Sin and L. Muthu, "Application of big data in education data mining and learning analytics–a literature review." vol. 5, no. 4, 2015. 3, 22

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. W. Witten, "The weka data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009. 3

[11] A. L'heureux, K. Grolinger, H. F.Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," vol. 5, pp. 7776–7797, 2017. 6, 24

[12] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," vol. 127, no. 1, pp. 3–30, 2011. 9

[13] J. Han, J. Pei, and M. Kamber, *Data mining: Concepts and Techniques.* Elsevier, 2011. 11, 12

[14] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015. 15

[15] X. Zheng, W. Chen, P. Wang, D. Shen, S. Chen, X. Wang, Q. Zhang, and L. Yang, "Big data for social transportation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 3, pp. 620–630, 2016. 16

[16] Y. Chen, L. Shu, and L. Wang, "Traffic flow prediction with big data: A deep learning based time series model," in *Computer Communications Workshops (INFOCOM WKSHPS), 2017 IEEE Conference on.* IEEE, 2017, pp. 1010–1011. 16

[17] W. Xiong, Z. Yu, L. Eeckhout, Z. Bei, F. Zhang, and C. Xu, "Szts: A novel big data transportation system benchmark suite," in *2015 44th International Conference on Parallel Processing (ICPP).* IEEE, 2015, pp. 819–828. 17

[18] W. Xu, H. Zhou, N. Cheng, F. Lyu, W. Shi, J. Chen, and X. Shen, "Internet of vehicles in big data era," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 19–35, 2018. 17

[19] F. Firouzi, A. M. Rahmani, K. Mankodiya, M. Badaroglu, G. Merrett, P. Wong, and B. Farahani, "Internet-of-things and big data for smarter healthcare: From device to architecture, applications and analytics," pp. 583–586. 18

[20] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013. 18

[21] S. Sakr and A. Elgammal, "Towards a comprehensive data analytics framework for smart healthcare services," *Big Data Research*, vol. 4, pp. 44–58, 2016. 18

[22] D. W. Bates, S. Saria, O.-M. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014. 19

[23] M.-H. Kuo, T. S. Kuo, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, "Health big data analytics: Current perspectives, challenges and potential solutions," *International Journal of Big Data Intelligence*, vol. 1, no. 1-2, pp. 114–126, 2014. 19

[24] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," *Big Data Research*, vol. 2, no. 1, pp. 2–11, 2015. 19

[25] E. A. Nuaimi, H. A. Neyadi, N. Mohamed, and J. Al-Jaroodi, "Applications of big data to smart cities," *Journal of Internet Services and Applications*, vol. 6, no. 1–15, p. 25, 2015. 20

[26] I. Vilajosana, J. Llosa, B. Martinez, M. Domingo-Prieto, A. Angles, and X. Vilajosana, "Bootstrapping smart cities through a self-sustainable model based on big data flows," *IEEE Communications magazine*, vol. 51, no. 6, pp. 128–134, 2013. 21

[27] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics," *Computer Networks*, vol. 101, pp. 63–80, 2016. 21

[28] M. Khan, M. Babar, S. H. Ahmed, S. C. Shah, and K. Han, "Smart city designing and planning based on big data analytics," *Sustainable Cities and Society*, vol. 35, pp. 271–279, 2017. 22

[29] H. Ozkose, E. Sertac, Ari, and p. y. p. Cevriye Gencer) journal=Procedia-Social and Behavioral Sciences, volume=195, "Yesterday, today and tomorrow of big data." 23

[30] H. Yang and S. Fong, "Incrementally optimized decision tree for noisy big data," in *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. ACM, 2012, pp. 36–44. 24

[31] A. Bifet, "Mining big data in real time," vol. 37, no. 1, 2013. 24

[32] J. Wang, J. Liu, R. Higgs, L. Zhou, and C. Zhou, "The application of data mining technology to big data," in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC)*, vol. 2. IEEE, 2017, pp. 284–288. 24

[33] K. Grolinger, M. Hayes, W. A. Higashino, A. L'Heureux, D. S. Allison, and M. A. Capretz, "Challenges for mapreduce in big data," in *2014 IEEE World Congress on Services (SERVICES)*. IEEE, 2014. 24

[34] N. Chaudhari and D. S. Srivastava, "Big data security issues and challenges," in *International Conference on Computing, Communication and Automation (ICCA2016)*. IEEE, 2016, pp. 60–64. 24

[35] V. Thayananthan and A. Albeshri, "Big data security issues based on quantum cryptography and privacy with authentication for mobile data center," *Procedia Computer Science*, vol. 50, pp. 149–156, 2015. 25

[36] G. Manogaran, C. Thota, and M. V. K. Kumar, "Metaclouddatastorage architecture for big data security in cloud computing," *Procedia Computer Science*, vol. 87, pp. 128–133, 2016. 25

[37] B. Saraladevia, N. Pazhanirajaa, P. V. Paula, M. S. Bashab, and P. Dhavachelvanc, "Big data and hadoop-a study in security perspective," *Procedia computer science*, vol. 50, pp. 596–601, 2015. 25

[38] A. Gupta, K. Pandhi, P. V. B. Bindu, and P. S. Thilagam, "Role and access based data segregator for security of big data," *Procedia Technology*, vol. 24, pp. 1550–1557, 2016. 25

[39] P. Colombo and E. Ferrari, "Privacy aware access control for big data: A research roadmap," vol. 2, no. 4, pp. 145–154, 2015. 25

[40] S. Benabderrahmanea, N. Melloulia, and P. P. Myriam Lamollea, "Smart4job: A big data framework for intelligent job offers broadcasting using time series forecasting and semantic classification," *Big Data Research*, vol. 7, pp. 16–30, 2017. 26

[41] M. N. Rahman, A. Esmailpour, and J. Zhao, "Machine learning with big data an efficient electricity generation forecasting system," *Big Data Research*, vol. 5, pp. 9–15, 2016. 26

[42] S. R, B. G. HB, S. K. S, P. Poornachandran, and S. K. P, "Apache spark a big data analytics platform for smart grid," *Procedia Technology*, vol. 21, pp. 171–178, 2015. 27

[43] Z. Zong, R. Ge, and Q. Gu, "Marcher: A heterogeneous system supporting energy-aware high performance computing and big data analytics," *Big Data Research*, vol. 8, pp. 27–38, 2017. 27

[44] M. Kubina, M. Varmus, and I. Kubinova, "Use of big data for competitive advantage of company," *Procedia Economics and Finance*, vol. 26, pp. 561–565, 2015. 28

[45] D. Duong, H. Tan, and S. Pham, "Customer gender prediction based on e-commerce data," in *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2016, pp. 91–95. 28