Study of Big Data Analysis with Hadoop Eco-system: A Case Study of Dhaka Stock Exchange

Syed Mehedi Ashraf
ID:2015-3-50-015

Mohammad Rasel
ID:2015-3-50-020

Md.Belayet Hossain
ID:2015-3-50-023

A project is presented for the degree of
Bachelor of Science in Informations & Communications Engineering

Department Of
Electronics & Communications Engineering
East West University
25 August 2019

ABSTRACT

The age of “Big Data” is going on and the use of this amount of data is increasing day by day. We all are using it in a various way with or without our consciousness. By data firmly and with the ability given by Big Data Technologies to efficiently store and analyze those information, we can find solution to these questions and work to optimize every aspect of our behavior. For instance, Big E-commerce sites can know which products we have brought or viewed by analyzing big data gathered from thousands of websites over many years. Online and offline activities are always being tracked, aggregated, and analyzed at amazing rates. Here, we are about to deal with the stock market and analyze the fluctuation of share value over a year. Ultimately, we can say that Big Data technology is used to improve basic leadership and to give more noteworthy bits of knowledge quicker when required yet with the drawback of loss of data privacy.

APPROVAL

This is to certify that the project entitled "Big Data Analysis with Hadoop Eco-system", submitted to the respected matter of the faculty of Engineering for partial fulfillment of the requirement for the degree of Bachelor of Information and Communication Engineering (ICE) under complete supervision of the undersigned.

Submitted by
Syed Mehedi Ashraf
(2015-3-50-015)
&
Mohammad Rasel
(2015-3-50-020)
&
Md. Belayet Hossain
(2015-3-50-023)

. . . . . .
(**Dr. Anup Kumar Paul)**
Assistant Professor
Department of
Electronics and Communications Engineering
East West University

. . . . . .
(**Dr. Mohammed Moseeur Rahman)**
Assistant Professor & Chairperson
Department of
Electronics and Communications Engineering
East West University

CERTIFIED

This certifies the project entitled "Big Data Analysis with Hadoop Eco-system", being submitted by Md. Belayet Hossain, Mohammad Rasel & Syed Mehedi Ashraf to department of Electronics and Communication Engineering, East West University, Dhaka for the award of the degree of Bachelor of Science in Information and Communication Engineering, is a record of a major project carried out by them. They have worked under my supervision and guidance and have fulfilled the requirements, which, to my knowledge, have reached the requisite standard for the submission of this dissertation.

.....

(Dr. Anup Kumar Paul)
(Supervisor)
Assistant Professor
Department of
Electronics and Communications Engineering
East West University

DECLARATION

We hereby declare that we carried out the work reported in this thesis in the Department of Electronics and Communication Engineering, East West University under the supervision of Dr. Anup Kumar Paul. We declare that to the best of our knowledge, no part of this report has been submitted elsewhere for the award of any degree. All source of knowledge used in the report has been duly acknowledged.

.....

Syed Mehedi Ashraf
2015-3-50-015

.....

Mohammad Rasel
2015-3-50-020

.....

Md. Belayet Hossain
2015-3-50-023

Contents

List of Figures **7**

1 INTRODUCTION **8**
1.1 Background 8
1.2 Aim 8
1.3 Objectives 9
1.4 Challenge 9

2 CONCEPT OF BIG DATA **10**
2.1 Five V's in Big Data 10
2.2 Importance of Big Data 12
2.3 Types of Analysis 12
2.4 Source of Big Data 15

3 BIG DATA TOOLS AND TECHNIQUES **20**
3.1 Understanding Big Data Storage 20
3.2 Architecture of Big Data Analysis 21
3.3 Tools 22
3.3.1 Apache Hadoop 22
3.3.2 Hadoop Distributed File System(HDFS) 24
3.3.3 MapReduce 26
3.3.4 Yet Another Resource Negotiator (YARN) . . . 27
3.3.5 Apache Spark 29
3.3.6 Pig scripting language 30
3.3.7 Hive 31
3.3.8 Zookeeper 32
3.3.9 HBase 32
3.3.10 Sqoop 33
3.3.11 Flume 33
3.4 Hadoop Setup 33
3.5 Visualization of Big Data 33

3.6 Considerations 34

4 PROPOSED METHODOLOGY 35
4.1 Problem 35
4.2 Algorithm 36
4.3 Data Processing 36

5 RESULT & DISCUSSION 39
5.1 Result 39
5.2 Discussion 49

6 FUTURE WORK & CONCLUSION 50
6.1 Future Work 50
6.2 Conclusion 50

List of Figures

1.1 Internet users by world region since 1990 9

2.1 Importance of Big data 12

3.1 Architecture of Big data Analysis 21
3.2 Hadoop Architecture 23
3.3 Apache Hadoop 2.0 and YARN 24
3.4 Map stage 26
3.5 Yarn vs MaPReduce 28
3.6 Apache Spark Architecture 30

Chapter 1

INTRODUCTION

1.1 Background

A large amount of data has been generated in cloud in 20th century approximate 85 % of it. The word "Big data" was introduced in 2005, when the Company O'Reilly Media launched it in 2005. Whatever, the efficient usage of Big data and the importance to understand regular data which we are facing in our daily life. The person Roger Mougalas who is CEO of the company O'Reilly Media introduced the word Big Data for the very first time in 2005. After creating Web 2.0, the company was working to develop Big Data Analysis. They take only one year to introduce it. They deal with huge amount data whose are tend to impossible to manage and process at the same time using traditional intelligence tools.

Simultaneously, in 2005,Yahoo builds Hadoop. It was built on top of Google's MapReduce. The goal was to index the entire World Wide Web and nowadays the open-source Hadoop is used by lot organizations to crunch through huge amounts of data.[4]

1.2 Aim

The aim of this project is to analyze the data of Dhaka stock exchange (DSE) and to predict the share value of any company as well as to identify those companies whose may be manipulated externally. Actually, The analyzed report will be analyzed from last one year of DSE share market values. The format type of data is CSV, which is very popular among data scientists and comparatively easy to analyze.

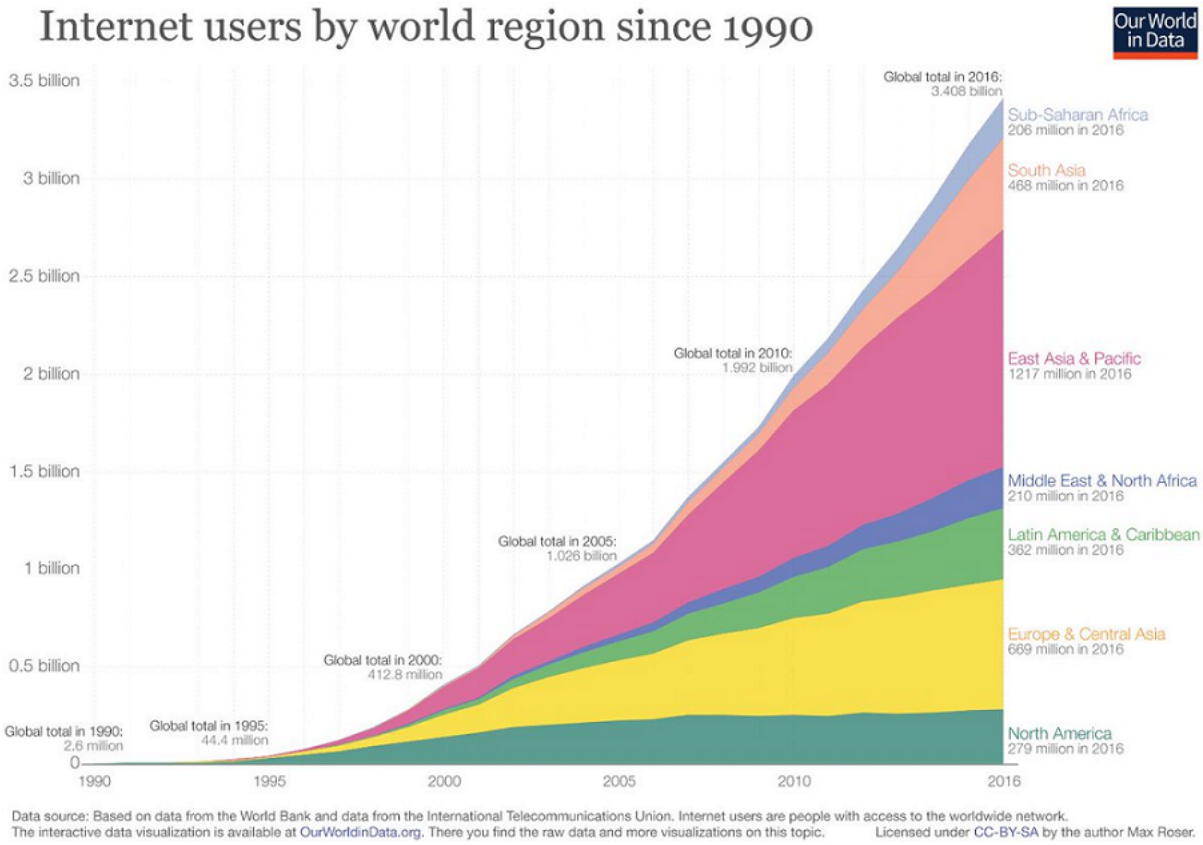

Figure 1.1: Internet users by world region since 1990

1.3 Objectives

The major objective about Big Data analysis with Hadoop Eco-system are-

1. The concept of Big Data analysis
2. Major challenges about analytics
3. What Hadoop is and How it addresses Big Data challenges
4. Description of Hadoop Eco-system
5. Finalize the analytical report

1.4 Challenge

The most critical challenge is to deal with a large number of data. It requires storage of those data as well as to process those data with suitable system so that a perfect outcome can be achieved. It also requires inserting different types of structured or unstructured data in a single system. And incoming new data should be acknowledged by the system.

Chapter 2

CONCEPT OF BIG DATA

2.1 Five V's in Big Data

There are five "V" to represent Big Data. May be, the initial guess is Big Data is a concept of storing a large number of data. But It is not. The five V is representing-

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value

1. Volume:

   The main characteristic that makes data "big" is the large amount of volume. It is not going to make any sense to focus on tiny storage units because the total amount of information is growing exponentially every year. In 2020, Thomson Reuters estimated in its annual report that it believed the world was "awash with over more than thousand Exabyte's of data and growing." For that equivalent year, EMC, an equipment organization that makes information stockpiling gadgets, thought it was more like 900 Exabyte's and would develop by 50 percent consistently. Nobody truly realizes how much new information is being produced; however the measure of data being gathered is enormous.

2. Velocity:

   Velocity is the speed of approaching information that should be processed. Considering present world that our gadgets are brimming with SMS messages, Facebook status, or MasterCard swipes

are being sent on a specific telecom bearer each moment of consistently, and we'll have a decent energy about speed. A spilling application like Amazon Web Services, Kinesis is a case of an application that handles the speed of information.

3. Variety:

   Variety is calm intriguing advancements with regards to innovation as increasingly more data is digitized. Conventional information types, for example, organized information incorporate things on a bank statement like date, amount, and time. These are things that fit flawlessly in a relational database. Structured data is expanded by unstructured information, which is the place things like Twitter channels, sound documents, MRI pictures, pages, web logs are put — anything that can be caught and put away however doesn't have a meta model (a lot of standards to outline an idea or thought — it characterizes a class of data and how to express it) that flawlessly characterizes it. Unstructured data is an essential and recognize idea in huge information. The ideal method to comprehend unstructured information is by contrasting it with organized information. Consider organized information as information that is all around characterized in a lot of standards. For instance, cash will consistently be numbers and have in any event two decimal focuses; names are communicated as content; and dates pursue a particular example.

4. Veracity:

   The Veracity in Big data refers to the biases, noise and abnormality in data. It is the data that is being stored, and minimized like meaningful to the problem being analyzed. In that case, veracity in data analysis is the biggest challenge when compares to things like volume and velocity. In scoping out big data strategy it is need to have a team and partners work to help keep the data clean and processes to keep 'dirty data' from accumulating in your systems.

5. Value:

   Last but not least, the V for value sits at the top of the big data pyramid. This refers to the ability to transform a tsunami of data into business. An example that is the source of company pride at MetLife: "We now know within a two-month period when it

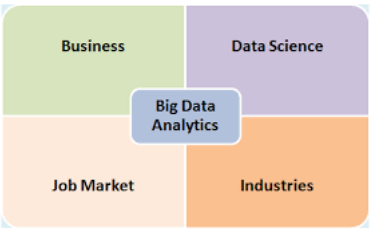

Figure 2.1: Importance of Big data

is highly likely that a customer will cancel his or her policy or purchase a new one."[9]

2.2 Importance of Big Data

The Big Data analytics is obviously a revolution in the field of Information Technology. The use of Big Data analytics by the companies is enhancing every year. The initiative focus of the companies is on customers. The field is flourishing in Business to Consumer (B2C) applications. We divide the analytics into different types as per needed of application. We have three divisions of Big Data analytics: Prescriptive Analytics, Predictive Analytics, and Descriptive Analytics. This field represents immense potential, and here, we will discuss four perspectives to explain why big data analytics is so important nowadays.[7]

1. Data Science Perspective
2. Business Perspective
3. Real-time Usability Perspective
4. Job Market Perspective

2.3 Types of Analysis

There are five types of Big Data Analysis. Those are -

1. Prescriptive Analytics
2. Diagnostic Analytics
3. Descriptive Analytics
4. Predictive Analytics
5. Outcome Analytics

1. Prescriptive Analytics:

   This type of analysis is greatly used to analysis to get perfect solution from a set of decision. Most of the company is going for this type of analytics because a lot of companies going through confusion like which decision can give them the best result. This technique actually focuses on to answer a particular question as well as helps to visualize the current data fact. It can also suggest few tasks to improve decision-making policy with Client and customer. So that Improvement of relation can occur. This is the technique, which is used most among other analysis for its performance.The facts behind prescriptive analytics are[8] -

   - Onward looking
   - Focused on optimal decisions for future situations
   - Simple rules to complex models that are applied on an automated or programmatic basis
   - Discrete prediction of individual data set members based on similarities and differences
   - Optimization and decision rules for future events

2. Diagnostic Analytics

   These types of analysis are doing when we are looking for some facts whose are responsible to occur a particular situation, in short to get reason behind a incident. It is very much helpful and efficient when determining about something like loyal customer preferences and his/her habit as well as attraction. Examples of diagnostic analytics can be, a company is looking for the habit of customer so that the company can serve the best services to them. This type of analysis is doing through diagnostic.[8]

   - Reverse looking
   - Deals with traditional connections and trade.
   - Create sorting through variable relation and the priority of data.
   - Try to make connection with dependent variable.
   - Deals with frequents and Bayesian causal inferential analyses

3. Descriptive Analytics

   This procedure is the most time-serious and more often than not delivers the least esteem; in any case,it is helpful for revealing examples inside a specific fragment of bunch. Elucidating investigation give the profound view into what has happened truly and will furnish you with patterns to dive into in more detail. Instances of engaging examination incorporate rundown insights, bunching and making new guidelines utilized in market bin investigation.[8]

   - Reverse looking
   - Focused on portrayals and examinations
   - Pattern identification and depictions
   - MECE (fundamentally unrelated and by and large thorough) classification
   - Category advancement dependent on likenesses and contrasts (division)

4. Predictive Analytics

   The most normally utilized and famous procedure, which is generally used to foresee some future development of any issue. Prescient examination use models to estimate what may occur in explicit situations. Instances of prescient investigation incorporate next best offers; agitate hazard and recharging hazard examination. [8]

   - Onward looking
   - Focused on non-discrete forecasts of future states, relationship, and examples
   - Description of forecast result set likelihood dispersions and probabilities
   - Model application
   - Non-discrete estimating (figures conveyed in likelihood dispersions)

5. Outcome Analytics

   Result examination alluded to as utilization investigation; this procedure gives profound view into the conduct or example of information that drives explicit results. This investigation is utilized to give better benefits and figure out how they are interfacing

with the framework items and administrations. Key purposes of Outcome examination are[8]-

- Reverse ward looking, Real-time and Onward looking
- Focused on utilization designs and related business results
- Description of use limits
- Create a model

2.4 Source of Big Data

Big Data is frequently come down to a couple of assortments including social information, machine information, and value-based information. Online networking information is giving amazing bits of knowledge to organizations on buyer conduct and slant that can be incorporated with CRM information for investigation, with 230 million tweets posted on Twitter every day, 2.7 billion Likes and remarks added to Facebook consistently, and 60 hours of video transferred to YouTube consistently (this is the thing that we mean by speed of information).

Machine information comprises of data created from mechanical gear, constant information from sensors that track parts and screen apparatus (regularly additionally called the Internet of Things), and even web logs that track client conduct on the web. At arcplan customer CERN, the biggest molecule material science research focus on the planet, the Large Hadron Collider (LHC) creates 40 terabytes of information consistently during examinations. Concerning information, huge retailers and even B2B organizations can create huge numbers of information all the time thinking about that their exchanges comprise of one or numerous things, item IDs, costs, installment data, producer and wholesaler information, and significantly more. Real retailers like Amazon.com, which posted $10B in deals in Q3 2011, and cafés like US pizza chain Domino's, which serves more than 1 million clients for every day, are creating petabytes of value-based huge data. The thing to note is that enormous information can take after conventional organized information or unstructured, high recurrence data. In characterizing huge information, it's additionally essential to comprehend the blend of unstructured and multi-organized information that contains the volume of data.[1]

In short, we can collect data from-

1. Social Media Data
2. Black Box Data
3. Stock Exchange Data
4. Transport Data
5. Power Grid Data
6. Search Engine Data
7. Power Grid Data
8. Action records from electronic devices
9. Business transactions
10. Electronic Files
11. Broadcasting

- Social Media Data: The best and easiest way to collect data is to collect data from Social Media. Because people spend a large number of time in social media and posted every single details of their daily life as well as their personal information also which is even not suppose to do. Is information created by human collaborations through a system, similar to Internet? The most widely recognized is the information created in interpersonal organizations. This sort of information suggests subjective and quantitative viewpoints, which are of some enthusiasm to be estimated. Quantitative viewpoints are simpler to gauge tan subjective angles, initial ones infers tallying number of perceptions gathered by geological or transient attributes, while the nature of the subsequent ones for the most part depends on the precision of the calculations connected to extricate the importance of the substance which are usually found as unstructured content written in regular language, instances of investigation that are produced using this information are notion examination, pattern themes examination, and so on.

- Black Box Data:

  It is a component of helicopter, airplane and jets e.t.c. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.

- Stock Exchange Data: The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.

- Transport Data: Transport data includes model, capacity, distance and availability of a vehicle.

- Search Engine Data: Search engines retrieve lots of data from different database.

- Power Grid Data: The power grid data holds information consumed by a particular node with respect to a base station.

  Sensors/meters and action records from electronic devices: These sort of data is created on constant, the number and periodicity of perceptions of the perceptions will be variable, in some cases it will depend of a lap of time, on others of the event of some occasion (per model a vehicle going by the vision edge of a camera) and in others will depend of manual control (from an exacting perspective it will be a similar that the event of an occasion). Nature of this sort of source depends for the most part of the limit of the sensor to take exact estimations in the manner it is normal.

- Business transactions: Data created because of business exercises can be recorded in organized or unstructured databases. At the point when recorded on organized information bases the most well-known issue to examine that data and get measurable markers is the enormous volume of data and the periodicity of its generation in light of the fact that occasionally these information is created at a quick pace, a great many records can be delivered in a second when huge organizations like grocery store chains are recording their deals. In any case, these sort of information isn't constantly created in organizations that can be straightforwardly put away in social databases, an electronic receipt is a case of this instance of source, it has pretty much a structure yet on the off chance that we have to put the information that it contains in a social database, we should apply some procedure to disperse that information on various tables (so as to standardize the information in like manner

with the social database hypothesis), and perhaps isn't in plain message (could be an image, a PDF, Excel record, and so on.), one issue that we could have here is that the procedure needs time and as recently stated, information possibly is being delivered excessively quick, so we would need various methodologies to utilize the information, preparing it all things considered without putting it on a social database, disposing of certain perceptions (which criteria?), utilizing parallel handling, and so on. Nature of data delivered from business exchanges is firmly identified with the ability to get delegate perceptions and to process them;

- Electronic Files: These allude to unstructured records, statically or powerfully delivered which are put away or distributed as electronic documents, similar to Internet pages, recordings, sounds, PDF documents, and so forth. They can have substance of extraordinary intrigue yet are hard to remove, various systems could be utilized, similar to content mining, design acknowledgment, etc. Nature of our estimations will for the most part depend on the ability to concentrate and effectively decipher all the agent data from those reports;

- Broadcasting: Mainly alluded to video and sound created on ongoing, getting factual information from the substance of this sort of electronic information at this point is excessively mind boggling and suggests enormous computational and interchanges control, when tackled the issues of changing over "advanced simple" substance to "computerized information" substance we will have comparative inconveniences to process it like the ones that we can discover on social communications.

  The employments of Big Data are nearly as differed, as they are huge. Unmistakable models we're most likely effectively acquainted with including Internet based life system breaking down their individuals' information to get familiar with them and interface them with substance and promoting important to their interests, or web search tools taking a gander at the connection among inquiries and results to offer better responses to clients' inquiries.

Yet, the potential uses go a lot further! Two of the biggest wellsprings of information in enormous amounts are value-based information, including everything from stock costs to bank information to singular vendors' buy accounts; and sensor information, quite a bit of it originating based on what is normally alluded to as the Internet of Things (IoT). This sensor information may be anything from estimations taken from robots on the assembling line of an automobile producer, to area information on a mobile phone arrange, to momentary electrical use in homes and organizations, to traveler boarding data taken on a travel framework.[1]

Chapter 3

BIG DATA TOOLS AND TECHNIQUES

3.1 Understanding Big Data Storage

Indeed, even through we are utilizing elite gadgets, if not every huge data applications accomplish their exhibition and versatility through sending on an accumulation of capacity and registering assets bound together inside a run time situation. Fundamentally, the capacity to configuration, create, and actualize a major information application is legitimately subject to a consciousness of the design of the hidden processing stage, both from an equipment and all the more significantly from a product point of view. One shared trait among the various machines and systems is the adjustment of devices to use the blend of accumulations of four key figuring assets:

1. Capability of processing: Often alluded to as a CPU, processor, or hub. Fundamentally, present day handling hubs regularly join different centers that are singular CPUs that offer the memory of hubs and are overseen and booked together, enabling various errands to be kept running at once; this is known as multi threading.

2. Memory: Which holds the information that the preparing node is as of now chipping away at? Most single node machines have a point of confinement to the measure of memory.

3. Storage: Providing steadiness of information;where datasets are stacked, and from which the information or data is stacked into memory to be handled.

4. Network:Which gives the "funnels" through which datasets are traded between various handling and capacity hubs? Since leaf-

hub PCs are restricted in their ability, they cannot effectively oblige monstrous measures of information. That is the reason the superior stages are made out of accumulations of PCs in which the gigantic measures of information and prerequisites for preparing can be dispersed among a pool of assets.

3.2 Architecture of Big Data Analysis

Associating various nodes together by means of an assortment of system topologies makes most superior stages. Particularly apparatuses may change in the points of interest of the designs, as do programming machines.Whatever, the general architecture creates a fine line between the management of computing resources (and corresponding allocation of tasks) and the management of the data across the network of storage nodes. A client level architecture of Big Data

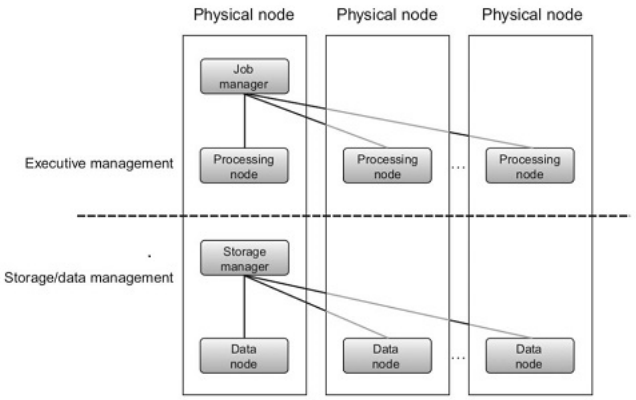

Figure 3.1: Architecture of Big data Analysis

analysis consists of NoSQL databases, distributed file systems and a distributed processing framework. NoSQL is commonly interpreted as "Not Only SQL." NoSQL databases are non-relational, non-SQL based, and store data in key-value pairs, which work well with unrelated data. NoSQL databases give services like distribution, highly scalable data storage for Big Data. It presented a NoSQL taxonomy that includes key value cache, key value store, and eventually consistent key value store, ordered key value store, data structures server, tuple-store, object database, document store, and wide columnar store. NoSQL listed 150 examples of NoSQL databases by different categories. A famous example of NoSQL database is Apache Hbase. According to Apache, Apache Hbase is an open-source, distributed, versioned, column-oriented store that provides random, real-time read

and write access to Big Data. That is why, It is important to select a perfect architecture.

3.3 Tools

There are a lot of tools whose are used to analyze and process the data. The Big data tools are categorized on the basis of storage and processing. Following are the top tools used in Big Data Analysis.

1. Apache Hadoop
2. MapReduce
3. YARN
4. Pig
5. Apache Spark
6. TEZ
7. HIVE
8. Zookeeper
9. Hbase
10. Sqoop
11. Flume
12. Mongo DB
13. R Programming

Even there are various other Big Data Tools to manage data, which is so large that it even exceeds terabytes in size.

3.3.1 Apache Hadoop

If we talk about Big Data, then it is closely impossible to neglect Apache Hadoop. It is an open-source framework for processing and managing big data across clusters of computers using simple programming models. Hadoop file system (HFS) is used for distributed processing and storage of data as well as high aggregate bandwidth across the cluster. Traditional data storage and analytics systems were not built like that it needs of big data. And it is no longer easily and cost-effectively supports the large amount of data sets. That is why;

we need Hadoop framework so that we can deal with a large number of unstructured or unorganized data. There are several advantage of using Apache Hadoop. For instance, Fault tolerant, built in redundancy, Flexibility and scalability, High computing power and most importantly low cost. Hadoop is an enormous scale, huge clump data

handling framework, which uses MapReduce for calculation and HDFS for information stockpiling. HDFS is the most dependable dispersed document framework with a configurable replication system intended to be conveyed on ease item equipment. HDFS breaks documents into pieces of at least 64 MB squares, where each square is imitated multiple times. The rehashed factor can be arranged, and it must be flawlessly adjusted; which is relying on the information. The accompanying chart portrays a common two hub Hadoop group set up on two exposed metal machines, in spite of the fact that you can utilize a virtual machine also.[2]

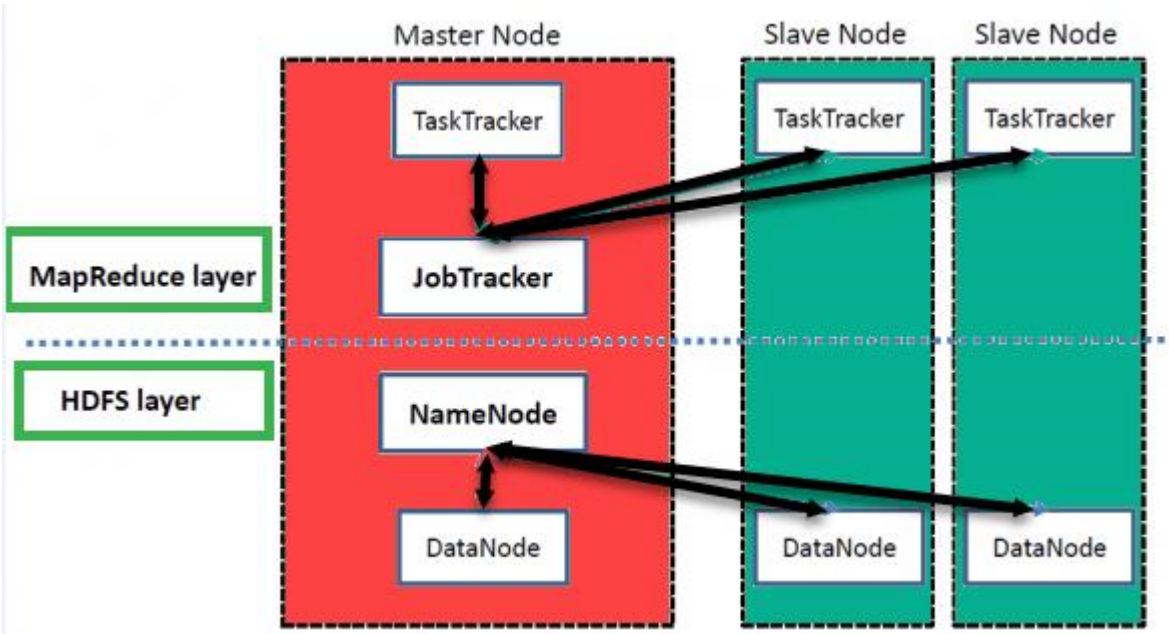

Figure 3.2: Hadoop Architecture

3.3.2 Hadoop Distributed File System(HDFS)

HDFS represents Hadoop Distributed File System. HDFS endeavors to empower the capacity of huge records, and does this by disseminating the information among a pool of information hubs. A solitary name hub (at times alluded to as Name Node) keeps running in a group, related with at least one information hubs, and gives the administration of a normal progressive record association and namespace. The name hub adequately connects with the conveyed information hubs. The making of a record in HDFS has all the earmarks of being a solitary document, despite the fact that it squares "lumps" of the document into pieces that is put away on individual information hubs.[3]

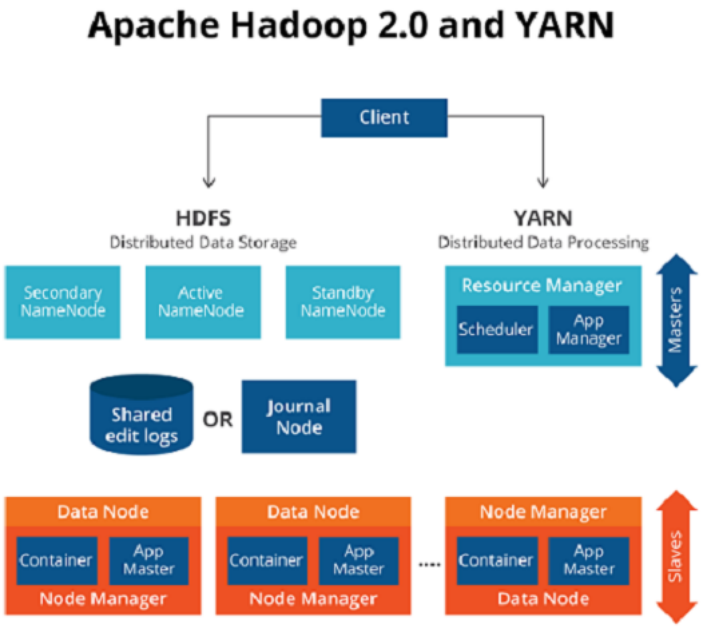

Figure 3.3: Apache Hadoop 2.0 and YARN

The name node manages metadata about specific document just as the historical backdrop of changes to record metadata. That metadata incorporates an identification of the oversaw documents, properties of the records, and the document framework, just as the mapping of squares to records at the information hubs. The information node itself does not deal with any data about the legitimate HDFS record; rather, it treats every datum hinder as a different document and offers the basic data with the name node.

The capacity to utilize HDFS exclusively as a methods for making an adaptable and expandable document framework for keeping up

quick access to huge datasets gives a sensible offer from an Information Technology viewpoint:

1. Decreased the cost of specialty large-scale storage systems;
2. Provides the ability to rely on commodity components;
3. Enabled the ability to deploy using cloud-based services;
4. Reduced system management costs.

Enabling this level of reliability should be facilitated through a number of key tasks for failure management, some of which are already deployed within HDFS while others are not currently implemented:

- Monitoring: There is a continuous "heartbeat" communication between the data nodes to the name node. If the name node does not hear a data node's heartbeat, the data node is considered to have failed and is no longer available. In this case, a replica is employed to replace the failed node, and a change is made to the replication scheme.
- Re-balancing: This is a process of automatically migrating blocks of data from one data node to another when there is free space, when there is an increased demand for the data and moving it may improve performance (such as moving from a traditional disk drive to a solid-state drive that is much faster or can accommodate increased numbers of simultaneous accesses), or an increased need to replication in reaction to more frequent node failures.
- Managing integrity: HDFS uses checksums, which are effectively "digital signatures", associated with the actual data stored in a file (often calculated as a numerical function of the values within the bits of the files) that can be used to verify that the data stored corresponds to the data shared or received. When the checksum calculated for a retrieved block does not equal the stored checksum of that block, it is considered an integrity error. In that case, the requested block will need to be retrieved from a replica instead.
- Metadata replication: The metadata files are also subject to failure, and HDFS can be configured to maintain replicas of the corresponding metadata files to protect against corruption.
- Snapshots: This is incremental copying of data to establish a point in time to which the system can be rolled back.

- Cost: Actually, in case of storage cost, Hadoop has lower cost comparing with others HDBMS. Even Hadoop provides highly saleable storages and process with fraction of the EWD cost.

3.3.3 MapReduce

MapReduce is a framework or a program model that allows performing parallel and distributed processing from a set of data in a distributed environment such as Hadoop. There are three stages to execute a process in MapReduce. The stages are-

1. Map stage
2. Shuffle stage
3. Reduce stage

1. Map stage:The executing process begins with map stage where input data is passing into mapped function line by line. The input data type can be multiple or dictionary which will be stored in HDFS. To process the mapper will use several small chunks of data.

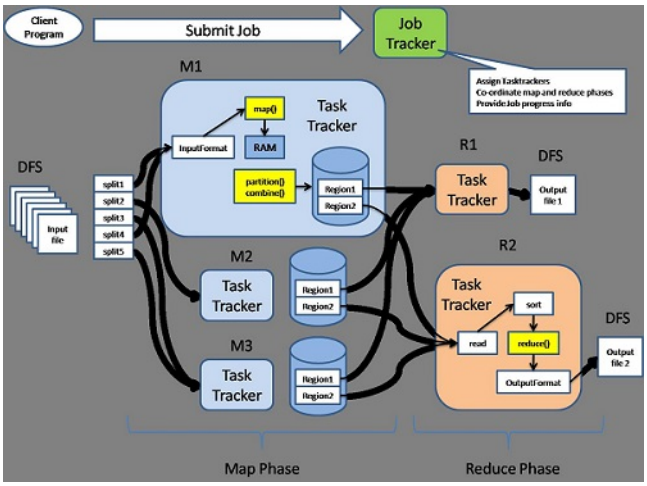

Figure 3.4: Map stage

2. Reduce stage: Basically, Reduce stage is the combination of shuffle and reduces stage. The job of the reducer stage is to collect the data from the Map stage and it processes the data and makes a new set of it. After that new data set is stored the data into HDFS again. When MapReduce performs, Hadoop sends the "Map and Reduce" tasks to proper servers in the cluster so that the task can

be completed easily and as fast as possible. In Hadoop, MapReduce really consolidated employment the board, oversight and the programming model for execution. The MapReduce execution condition utilizes an ace execution model, in which one ace node (called the Job Tracker) deals with a pool of slave processing assets (called Task Trackers) that are called upon to do the real work. The job of the Job Tracker is to deal with the assets with some particular duties, including dealing with the Task Trackers, ceaselessly observing their openness and accessibility, and the various parts of occupation the board that incorporate planning errands, following the advancement of allotted assignments, responding to recognized disappointments, and guaranteeing adaptation to non-critical failure of the execution. The job of the Task Tracker is a lot more straightforward: hang tight for an errand task, start and execute the mentioned assignment, and give status back to the Job Tracker on an occasional premise. Various customers can make demands from the Job Tracker, which turns into the sole authority for assignment of assets. There are impediments inside this current MapReduce model. To begin with, the programming worldview is pleasantly fit to applications where there is region between the preparing and the information, however applications that request information development will quickly move toward becoming impeded by system inactivity issues. Second, All applications are not effectively mapped to the MapReduce model, with the goal that applications created utilizing elective Programming strategies would even now require the MapReduce framework for occupation the executives. To wrap things up, the designation of preparing hubs inside the group is fixed through portion of specific hubs as "map spaces" versus "diminish openings." When the calculation is weighted into one of the stages, the hubs relegated to the next stage are to a great extent unused, which results in processor under utilization.[5]

3.3.4 Yet Another Resource Negotiator (YARN)

YARN is a part of Hadoop distributed framework, which is used to manage resources, and to schedule of tasks (JOB). YARN is used for allocating system resources into the various applications running in a Hadoop cluster and Job scheduling tasks to be executed on various cluster nodes. YARN is stand for "Yet Another Resource Negotia-

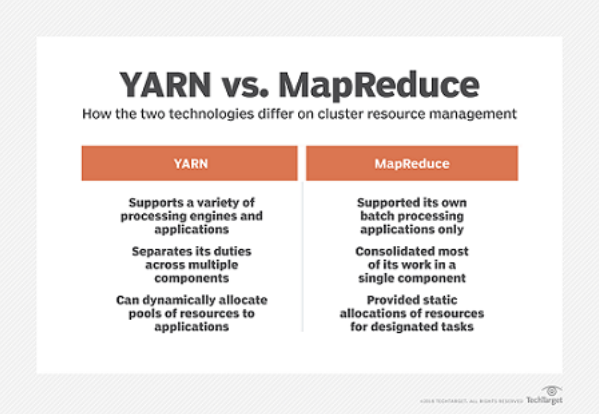

Figure 3.5: Yarn vs MaPReduce

tor", but it is commonly referred to acronym alone; this name was self-deprecating humor on the part of its developers. The technology became an Apache Hadoop sub-project within the Apache Software Foundation in 2012 and was one of the key features added in Hadoop 2.0, which was released for testing that year and became generally available in October 2013. This is utilized to state that in future forms of Hadoop through the isolation of obligations inside an amendment called YARN. In this methodology, generally speaking asset the executives has been concentrated while a neighborhood Node Manager presently performs the executives of assets at every hub. Likewise, there is the idea of an Application Master that is related with every application that legitimately consults with the focal Resource Manager for assets while assuming control over the duty regarding checking advancement and following status. Pushing this obligation to the application condition permits more prominent adaptability in the task of assets just as is progressively compelling in booking to improve hub usage. There are three important elements architecture in YARN. Those are:

1. Resource Manager
2. Application Master
3. Node Managers

- Resource Manager: The Resource Manager, or RM, which is typically one for each bunch, is the ace server. Resource Manager has the data of area of DataNode and how much assets they have. This data is alluded to as Rack Awareness. The Resource Manager runs different administrations, the most significant of which is the Resource Scheduler that chooses how to allot the assets.

- Application Master:

  The Application Master is a procedure of system that instate assets for a solitary application, that is, a solitary activity or a coordinated non-cyclic chart of occupations, which keeps running in the main compartment assigned for the reason. The general undertaking of Application Master is to demand assets from the Resource Manager and afterward works with compartments given by Node Managers.

- Node Managers:

  The Node Managers can be numerous in one bunch. They are the captives of the foundation. When it begins, it reports itself to the RM and occasionally sends a heartbeat to the RM. Every Node Manager offers assets to the bunch. The asset limit is the measure of memory and the quantity of v-centers, short for the virtual center. At run-time, the Resource Scheduler takes the choice that how to utilize this limit. A holder is a small amount of the Node Manager limit, and the customer to run a program utilizes it. Every Node Manager takes directions from the Resource Manager and reports and handles holders on a solitary node.

The Resource Manager start with the customers through an interface, which is known as the Client Service. Every customer can submit or end an application and got recognized about the booking line or bunch insights through the Client Service. Authoritative solicitations are served through a different interface, which is known as the Admin Service through which administrators can get refreshed data about the group activity.

At the same time, the Resource Tracker Service gets hub pulses from the Node Manager by which it can follow new or decommissioned hubs.

The Node chief Liveliness Monitor and Nodes List Manager keep a refreshed status of which hubs are solid with the goal that the Scheduler and the Resource Tracker Service can designate work autonomously and fittingly.

3.3.5 Apache Spark

Apache Spark is the cluster-computing framework for large-scale data processing. Spark offers a set of libraries in three languages that are Java, Scala, and Python for its unified computing engine. The definition is actually indicated on three particular things. Unified:By

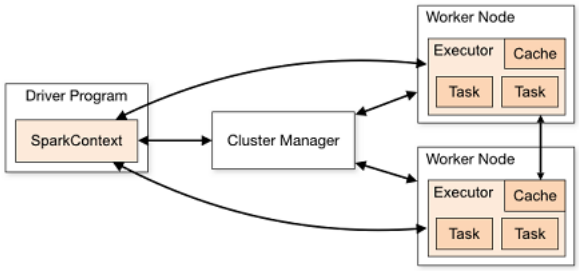

Figure 3.6: Apache Spark Architecture

using Spark, there is zero need to integrate an application out of multiple APIs or systems. Apache Spark provides us with enough built-in APIs to get the job properly. Computing Engine: Spark handles loading data from various file systems and runs computations on it, but does not store any data itself permanently. Spark do its work entirely in memory allowing unparalleled performance and speed. Libraries: Apache Spark is comprised of a large series of libraries, which is built for data science. Spark includes libraries for SQL, Machine Learning, Stream Processing (Spark Streaming and Structured Streaming), and Graph Analytics.

3.3.6 Pig scripting language

In spite of the fact that the programming model MapReduce is moderately direct, despite everything it takes some aptitude and comprehension of both parallel and appropriated programming and Java to best exploit the model. The undertaking of Pig is endeavoring to make simpler the advancement procedure of utilization by expending a portion of the subtleties away through a higher level programming language called Pig Latin. As indicated by the venture's site, Pig's abnormal state programming language enables the designer to determine how the investigation is performed. Then again, a compiler changes the Pig Latin determination into MapReduce programs.

The plan is to implant a critical arrangement of parallel administrators and capacities contained inside a control succession of mandates to be connected to datasets in a manner that is fairly like the way SQL explanations are connected to customary organized databases. A few models incorporate producing datasets, sifting through subsets, joins, parting datasets, expelling copies. For couple of essential applications, utilizing Pig serves critical simplicity of improvement, and progressively complex undertakings can be built as arrangements

of connected administrators. Last but not least, the use of a high-level language for instance Pig also allows the compiler to identify opportunities for optimization that might have been ignored by an inexperienced programmer. At the same time, the Pig environment allows developers to create new user defined functions (UDFs) that can subsequently be incorporated into developed programs.

3.3.7 Hive

One of the regularly noted issues with MapReduce is that in spite of the fact that it star vides a procedure for creating and executing applications those utilization enormous measures of information, it isn't more than that. And keeping in mind that the information can be overseen inside documents utilizing HDFS, numerous business applications expect portrayals of information in organized database tables. That was the inspiration for the improvement of Hive, which (as per the Apache Hive site) is an "information stockroom framework for Hadoop that encourages simple information rundown, questions, and the investigation of huge datasets used to put away in Hadoop circulated document frameworks." Hive is explicitly built for information distribution center questioning and announcing and isn't expected for use as inside exchange handling frameworks that require constant inquiry execution or exchange semantics for consistency at the line level.

Hive is layered over the record framework and execution system for Hadoop and empowers applications and clients to compose information in an organized information stockroom and in this way inquiry the information utilizing a question language called HiveQL that is like SQL (the standard Structured Query Language utilized for most present day social database the executives frameworks). It empowers local access to the MapReduce model, enabling software engineers to create custom Map and Reduce capacities that can be legitimately incorporated into HiveQL questions. Hive gives adaptability and extensibility to bunch style inquiries for detailing over enormous informational collections that are normally being extended while depending on the shortcoming tolerant parts of the fundamental Hadoop execution model.

3.3.8 Zookeeper

When there are any multiple tasks and jobs running in Hadoop Ecosystem, there is a need for configuration management and synchronization of various aspects of naming and coordination. The Zookeeper projects website specifies that: "Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services." Zookeeper manages library, which is naming and successfully executes a framework for dealing with the different static and transient named questions in a various leveled way, much like a record framework. Likewise, it empowers coordination for practicing command over shared assets that are affected by race conditions (in which the normal yield of a procedure is affected by varieties in timing) and stop (in which various errands competing for control of a similar asset adequately lock each other out of any undertaking's capacity to utilize the asset). Common coordination administrations like those gave in Zookeeper enable engineers to utilize these controls without creating them starting with no outside help.

3.3.9 HBase

HBase is another case of a none relational data management tool that circulates enormous datasets over the hidden Hadoop system. HBase is gotten from Google's Big Table and is a section-situated information format that, when layered over Hadoop, gives a flaw tolerant technique to putting away and controlling enormous information tables. As was talked about in Chapter 6, information put away in a columnar design is manageable to pressure, which expands the measure of information that can be spoken to while diminishing the genuine stockpiling impression. What's more, HBase underpins in-memory execution. HBase is certifiably not a relational database, and it doesn't bolster SQL inquiries. There are some fundamental activities for HBase: Get (which access a particular line in the table), Put (which stores or updates a line in the table), Scan (which emphasizes over an accumulation of columns in the table), and Delete (which expels a line from the table). Since it very well may be utilized to compose datasets, combined with the exhibition given by the parts of the columnar direction, HBase is a sensible option as a tenacious stockpiling worldview when running MapReduce applications.

3.3.10 Sqoop

It is a command line interface to create relation between a rational database and Hadoop. It actually deals with structured data. It is used to insert data into HDFS.

3.3.11 Flume

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log data, events (etc.) from various web servers to a centralized data store. It is greatly reliable, distributed, and configurable tool that is principally designed to transfer streaming data from various sources to HDFS.

3.4 Hadoop Setup

First need to setup Hortonworks Sandbox in the device. Hortonworks Sandbox is a Hadoop learning and development environment that runs as a virtual machine. It is a widely accepted and user friendly to learn Hadoop as it comes with most of latest stack of applications of Hortonworks Data Platform (HDP). We have used Hortonworks Sandbox throughout the book. At the time of this writing, the latest version of the sandbox is 2.5. Setting up the Hortonworks Sandbox: The following steps are needed to set up Hortonworks Sandbox:

1. Download the Oracle VirtualBox installer from https://www.virtualbox.org
2. Launch the installer and accept all the default options.
3. Download the Hortonworks Sandbox virtual image for VirtualBox, located at http://hortonworks.com/products/hortonworks-sandbox

3.5 Visualization of Big Data

To visualize the analyze pattern using various charts or table or graph based on Hive data, following steps should be taken[6]-

1. Evolution of data visualization and its classification
2. Data source preparation
3. Consumption of HDFS-based data through HiveQL

4. Preparation of various charts
5. Beautification of the charts using styling

3.6 Considerations

Big data applications convey an assortment of apparatuses and systems for execution. When sorting out our considerations about building up those applications, it is imperative to consider the parameters that will outline our requirements for evaluation and procurement, measuring and configuration, methods for information association (Data Organizing) , and expected calculations to be utilized or created starting with no outside help.

Preceding jumping straightforwardly into downloading and installing software, center around the sorts of huge information business applications and their relating execution needs scaling, for example, those recorded in the figure.

The specialized necessities will control both the equipment and the software. This additionally enables us to adjust the improvement of the stage to the business application advancement needs.

Table 3.1 Variables to consider when framing Big Data Environment

Variable	Intent	Technical Requirements
Predisposition of parallelization	Number and type of processing node	Number or processors Types of processors
Size of data to be persistently stored	Amount and allocation of disk space for distributed file system	Size of disk drives Number of disk drivers Type of drivers(SSD versus magnetic versus optical) Bus configuration
Amount of data to be accessible in memory	Amount and allocation of core memory	Amount of RAM memory Cache memory
Need for cross-node communication	Optimize speed and bandwidth	Network/cabinet configuration Network speed Network bandwidth
Types of data organization	Data management requirements	File management organization Database requirements Data orientation Type of data structures
Developer skill set	Development tools	Types of programming tools,compilers, execution model,debuggers etc.
Types of algorithms	Analytic functionality requirements	Data warehouse/marts for OLAP data mining and predictive analytics

Table 3.1: Variables to consider when framing Big Data Environment

Chapter 4

PROPOSED METHODOLOGY

4.1 Problem

From the first week of April, Dhaka Stock Exchange (DSE) has been in the features of local dailies as its major index (DSEX) kept on declining. As per the dailies, Dhaka Stock Exchange Brokers' Association (DBA) believed absence of certainty to be the purpose for the market droop. Some disappointed little financial specialists took to the road to dissent. Be that as it may, the finance minister AHM Mustafa Kamal stated, "The stock market is fine. It is typical for the market to have changes." But the thing is, in our local share market, the external corruption occur which manipulate our share market. But most of the time, it becomes tough to identify which companies are manipulated and dealing with those companies. Security Exchange Council (SEC) is the council, which deals with that security problem. So, it is necessary to identify those companies whose are manipulated as fast as possible. The discussion on whether the market is controlled or not, regardless of whether there is any scheme (hypothesis) or not, is beyond the scope of this article. Scholastics like to learn about the thesis through which market control, or 'changes'. It is closely impossible to check out every share of thousands of company and taking steps physically to judge it. So, we are here to make the identification process much more easier.

4.2 Algorithm

1. Start
2. Downlaod the DSE data
3. Upload the data into HDFS.
4. Create the Pig script
    - Read the data from HDFS by using LOAD command
    - Select with or without schemas for a relation by using FOREACH command
    - Collect records from the given inputs by using GROUP BY command
    - Apply expression to each record and generate max closeprice of a particular company by using FOREACH command
    - Apply expression to each record and generate min closeprice of a particular company by using FOREACH command
    - join the data
    - Distinct the data
    - Store that data into HDFS by using STORE command
    - Then again read that file from the HDFS for further use
    - Describe the relation
    - Generate output by using DUMP command
5. Run pig script on Tez
6. Get the expected output
7. End

4.3 Data Processing

The solving process starts with the data of Dhaka Stock Exchange (DSE). Here, we have collected the data of last two year of Dhaka Stock Exchange (DSE). The format of data is "CSV". CSV stands for "Comma Separated value" which is very popular among data analyst. Usually, the data format CSV is comparatively gives faster results.

The sample of data is here,

00DS30,01-01-2019,1880.78,1911.37,1880.78,1908.2,53017420
00DS30,02-01-2019,1908.2,1916.06,1896.44,1912.98,69645700
00DS30,03-01-2019,1912.98,1945.22,1912.98,1941.99,92490920
00DS30,06-01-2019,1941.99,1980,1935.08,1974.1,102676760
00DS30,07-01-2019,1974.1,1983.02,1960.66,1963.89,96553530
00DS30,08-01-2019,1963.89,2002.64,1963.89,2001.19,101003250
00DS30,09-01-2019,2001.19,2013.68,1999.63,2011.12,102573570
00DS30,10-01-2019,2011.12,2021.41,2005.42,2011.74,89757530
00DS30,13-01-2019,2011.74,2034.7,2011.74,2030.43,97394350
00DS30,14-01-2019,2030.43,2042.32,2023.71,2026,114632380
00DS30,15-01-2019,2026,2033.79,2020.02,2029.07,113931890
00DS30,16-01-2019,2029.07,2039.95,2014.74,2017.27,100052550
00DS30,17-01-2019,2017.27,2020.58,2005.13,2009.49,101156870
00DS30,20-01-2019,2009.49,2032.81,2009.49,2029.9,96891260
00DS30,21-01-2019,2029.9,2038.22,2019.72,2021.7,88534150
00DS30,22-01-2019,2021.7,2030.49,2018.7,2026.11,89066340
00DS30,23-01-2019,2026.11,2050.02,2026.11,2043.03,104525550
00DS30,24-01-2019,2043.03,2051.71,2043.03,2049,103757110
00DS30,27-01-2019,2049,2061.99,2040.91,2043.52,119857260
00DS30,28-01-2019,2043.52,2057.8,2042.7,2046.77,100954840
00DS30,29-01-2019,2046.77,2054.31,2042.27,2046.29,94585990
00DS30,30-01-2019,2046.29,2048.49,2030.12,2031.73,102461480
00DS30,31-01-2019,2031.73,2035.81,2004.99,2007.96,99273790
00DSES,01-01-2019,1232.82,1247.5,1232.82,1245.8,53017420
00DSES,02-01-2019,1245.8,1253.19,1241.32,1249.81,69645700
00DSES,03-01-2019,1249.81,1273.91,1249.81,1271.37,92490920
00DSES,06-01-2019,1271.37,1299.37,1271.37,1295.35,102676760
00DSES,07-01-2019,1295.35,1302.11,1287.32,1289.55,96553530
00DSES,08-01-2019,1289.55,1311.64,1289.55,1308.06,101003250
00DSES,09-01-2019,1308.06,1321.05,1308.06,1315.99,102573570
00DSES,10-01-2019,1315.99,1324.81,1313.82,1318.66,89757530
00DSES,13-01-2019,1318.66,1331.05,1318.66,1329.87,97394350
00DSES,14-01-2019,1329.87,1338.36,1325.83,1326.58,114632380

Here, the 1st column represents the title of the share, 2nd column represents the date of market value, 3rd column represents the opening price of that share, 5th column represents the maximum price of that share on that day, 6th share represents the minimum price of the share on that day, 7th column represents the closing price of that

share on that day and lastly that is the ID of that unique share. The environment setup was done by us with “Apache Ambari”. Apache Ambari is a complete open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters. Apache Ambari takes the guesswork out of operating Hadoop. Apache Ambari, as part of the Hortonworks Data Platform, allows enterprises to plan, install and securely configure HDP making it easier to provide ongoing cluster maintenance and management, no matter the size of the cluster. We have run it on virtual box.You have used Sqoop to insert CSV data into the Hadoop Distributed File system. We have used structured data so that it requires Sqoop instead of flume.

Chapter 5

RESULT & DISCUSSION

5.1 Result

After completion of processing we have got our desire result, sample is putted below: percent_dse: companyID: chararray,percentage: double
(GLAXOSMITH,9269.412231445312)
(SEMLFBSLGF,700.0)
(SONARBAINS,347.33331203460693)
(BATBC,315.3094530105591)
(BIFC,285.714316368103)
(DUTCHBANGL,240.12250900268555)
(UNITEDINS,238.64541053771973)
(MONNOCERA,236.9696855545044)
(MONNOSTAF,225.0)
(EMERALDOIL,210.98899841308594)
(06.Food_&_Allied,210.80503463745117)
(11.Miscellaneous,201.7378807067871)
(BXSYNTH,197.14285135269165)
(CNATEX,184.99999046325684)
(GLOBALINS,182.35293626785278)
(MEGHNAPET,182.14287757873535)
(ILFSL,179.03225421905518)
(AL-HAJTEX,176.49768590927124)
(SEMLIBBLSF,173.84613752365112)
(INTECH,172.9257583618164)
(UNITEDAIR,164.2857313156128)
(TALLUSPIN,161.29034757614136)
(FAREASTFIN,160.0000023841858)
(PROVATIINS,155.8139443397522)
(16.Tannery_Industries,155.69924116134644)

(TUNGHAI,151.85186862945557)
(1STPRIMFMF,148.74999523162842)
(AGRANINS,148.55492115020752)
(MEGCONMILK,147.44526147842407)
(03.Ceramics_Sector,147.08367586135864)
(FASFIN,144.615375995636)
(SAVAREFR,141.07142686843872)
(LEGACYFOOT,138.97435665130615)
(RUPALILIFE,137.2844696044922)
(UNIONCAP,133.33332538604736)
(DACCADYE,132.1428656578064)
(PLFSL,130.00000715255737)
(KEYACOSMET,129.03226613998413)
(JUTESPINN,126.64713859558105)
(ASIAINS,123.59551191329956)
(MIDASFIN,121.67832851409912)
(EASTERNINS,121.4545488357544)
(SSSTEEL,117.67240762710571)
(APOLOISPAT,114.28570747375488)
(RNSPIN,110.00000238418579)
(DHAKAINS,108.99999141693115)
(SONALIANSH,108.73712301254272)
(FAMILYTEX,107.1428656578064)
(UNITEDFIN,105.66039085388184)
(FIRSTFIN,104.99999523162842)
(PARAMOUNT,104.54546213150024)
(PREMIERLEA,104.4776201248169)
(IMAMBUTTON,103.08642387390137)
(SEMLLECMF,103.03030014038086)
(BDFINANCE,102.8571367263794)
(RENWICKJA,98.22485446929932)
(GENNEXT,97.4358856678009)
(STANCERAM,97.40437865257263)
(MITHUNKNIT,97.29728698730469)
(PRIMEINSUR,96.35036587715149)
(DULAMIACOT,95.68106532096863)
(DELTASPINN,95.12195587158203)
(BDAUTOCA,94.73684430122375)
(PF1STMF,94.44444179534912)

(QUASEMIND,91.58248901367188)
(STYLECRAFT,91.2701666355133)
(JMISMDL,90.98959565162659)
(GEMINISEA,90.69767594337463)
(BSCCL,90.3743326663971)
(HEIDELBCEM,90.19999504089355)
(17.Telecommunication,89.8692786693573)
(CVOPRL,88.37209343910217)
(IPDC,88.0658507347107)
(NORTHRNINS,86.66666150093079)
(BNICL,86.66666150093079)
(BEACHHATCH,83.99999737739563)
(SHYAMPSUG,83.85651111602783)
(GOLDENSON,83.82352590560913)
(CITYGENINS,82.03124403953552)
(GSPFINANCE,81.94445371627808)
(KPPL,81.74602389335632)
(KAY&QUE,81.38260841369629)
(SUNLIFEINS,80.90452551841736)
(EASTLAND,80.41237592697144)
(AIL,80.0000011920929)
(GENEXIL,79.5597493648529)
(KARNAPHULI,79.33332920074463)
(JANATAINS,79.104483127594)
(SHASHADNIM,78.90411615371704)
(BDWELDING,78.33333015441895)
(SILVAPHL,78.17257642745972)
(SAIHAMTEX,77.5456964969635)
(SINOBANGLA,77.30337381362915)
(MALEKSPIN,77.08333730697632)
(MHSML,76.84209942817688)
(PROGRESLIF,76.80491805076599)
(LANKABAFIN,76.47058963775635)
(SAMATALETH,76.14107728004456)
(RECKITTBEN,76.0986328125)
(NPOLYMAR,76.04456543922424)
(MAKSONSPIN,76.00000500679016)
(REPUBLIC,75.78948140144348)
(ZAHINTEX,75.67567229270935)

(BAYLEASING,75.18248558044434)
(10.Jute,74.53402876853943)
(MEGHNALIFE,74.48405623435974)
(CONTININS,74.17582273483276)
(SIMTEX,73.70892763137817)
(RUPALIINS,73.05388450622559)
(PREMIERBAN,72.99998998641968)
(VAMLRBBF,72.72726893424988)
(ALLTEX,72.72726893424988)
(TAKAFULINS,72.22222089767456)
(02.Cement,71.89888954162598)
(BDTHAI,71.7391312122345)
(HAKKANIPUL,71.61904573440552)
(CAPMIBBLMF,71.42857313156128)
(IFIC,71.27660512924194)
(ARAMITCEM,70.9876537322998)
(NHFIL,70.48709988594055)
(FEDERALINS,70.4081654548645)
(ECABLES,70.05624771118164)
(DSHGARME,69.87179517745972)
(PHENIXINS,69.56523060798645)
(PURABIGEN,69.2307710647583)
(OAL,68.23529601097107)
(ARAMIT,68.21818351745605)
(IFADAUTOS,68.1415855884552)
(CENTRALPHL,67.96115636825562)
(ASIAPACINS,67.5000011920929)
(DSSL,67.1052634716034)
(NATLIFEINS,66.99387431144714)
(CENTRALINS,66.66666865348816)
(AZIZPIPES,66.66666865348816)
(MIRACLEIND,65.69037437438965)
(ISLAMIINS,65.16854166984558)
(FARCHEM,64.76189494132996)
(KTL,64.49704170227051)
(NITOLINS,63.63636255264282)
(FUWANGCER,63.333332538604736)
(VAMLBDMF1,63.07693123817444)
(PHOENIXFIN,63.052207231521606)

(SANDHANINS,62.72726655006409)
(ICBAGRANI1,62.68656849861145)
(BANGAS,62.44344115257263)
(NORTHERN,61.70629262924194)
(AFTABAUTO,61.43791079521179)
(SINGERBD,61.31471395492554)
(SPCERAMICS,60.95237731933594)
(ACFL,60.92714071273804)
(SAPORTL,60.80402135848999)
(STANDARINS,60.2189838886261)
(ISLAMICFIN,59.58903431892395)
(PHARMAID,59.528905153274536)
(EASTRNLUB,59.433531761169434)
(SAFKOSPINN,59.23566818237305)
(ABBANK,58.620697259902954)
(BRACBANK,58.3629846572876)
(NFML,57.831329107284546)
(CAPMBDBLMF,57.575756311416626)
(ZEALBANGLA,57.46753811836243)
(SONARGAON,57.432425022125244)
(SEAPEARL,57.142847776412964)
(ISNLTD,56.79011940956116)
(IBP,56.756746768951416)
(BERGERPBL,56.71641826629639)
(12.Mutual_Funds,56.620997190475464)
(MERCINS,56.521737575531006)
(AMBEEPHA,56.280338764190674)
(MICEMENT,56.07985854148865)
(INTRACO,55.95855712890625)
(USMANIAGL,55.950915813446045)
(WMSHIPYARD,55.86206912994385)
(MERCANBANK,55.725181102752686)
(APEXFOODS,55.40540814399719)
(ONEBANKLTD,55.37189245223999)
(ATCSLGF,55.20832538604736)
(PEOPLESINS,54.97075915336609)
(BEXIMCO,54.913294315338135)
(YPL,54.36892509460449)
(SHEPHERD,53.96825671195984)

(PRIMETEX,53.846150636672974)
(13.Paper_&_Printing,52.49016880989075)
(KBPPWBIL,51.68539881706238)
(04.Engineering,51.30105018615723)
(PRAGATILIF,50.94339847564697)
(BENGALWTL,50.862061977386475)
(NAHEEACP,50.63830614089966)
(SAIFPOWER,50.62500238418579)
(LIBRAINFU,50.410956144332886)
(ZAHEENSPIN,50.000011920928955)
(FUWANGFOOD,50.0)
(NEWLINE,50.0)
(PRIMELIFE,50.0)
(FIRSTSBANK,50.0)
(ACI,49.99999701976776)
(BPML,49.84026551246643)
(AAMRANET,49.55357313156128)
(SPCL,49.31130111217499)
(EXIM1STMF,48.93617630004883)
(POPULARLIF,48.863643407821655)
(MARICO,48.72084558010101)
(UPGDCL,48.58456552028656)
(KPCL,48.43423068523407)
(RDFOOD,48.275864124298096)
(DOREENPWR,48.199450969696045)
(WATACHEM,47.91616201400757)
(ALIF,47.72726893424988)
(CITYBANK,47.2573846578598)
(PRAGATIINS,47.058820724487305)
(SOUTHEASTB,46.39999270439148)
(QUEENSOUTH,46.12794816493988)
(RAKCERAMIC,45.97315788269043)
(ACTIVEFINE,45.909082889556885)
(MLDYEING,45.90747058391571)
(DHAKABANK,45.80152630805969)
(NBL,45.67900896072388)
(BGIC,45.05493640899658)
(BSC,44.897955656051636)
(IDLC,44.80873942375183)

(SALVOCHEM,44.696974754333496)
(NURANI,44.525542855262756)
(ADVENT,44.366198778152466)
(PHPMF1,44.18603479862213)
(PIONEERINS,44.16058659553528)
(05.Financial_Institutions,44.1266268491745)
(STANDBANKL,43.95604133605957)
(RSRMSTEEL,43.74999701976776)
(SAIHAMCOT,43.63636672496796)
(BDLAMPS,43.342775106430054)
(METROSPIN,43.28359067440033)
(PADMALIFE,43.25580894947052)
(SALAMCRST,43.13725531101227)
(PRIME1ICBA,43.10344755649567)
(CONFIDCEM,42.88904666900635)
(EBLNRBMF,42.8571492433548)
(ICBAMCL2ND,42.85713732242584)
(EHL,42.672401666641235)
(TRUSTBANK,42.58555471897125)
(EBL1STMF,42.372873425483704)
(RUNNERAUTO,42.16560423374176)
(18.Textile,41.915297508239746)
(BEACONPHAR,41.81818068027496)
(14.Pharmaceuticals_&_Chemicals,41.7510449886322)
(BDCOM,41.70040190219879)
(ACMELAB,41.17647111415863)
(OIMEX,41.05263352394104)
(TRUSTB1MF,40.90908467769623)
(HRTEX,40.76923429965973)
(01.Bank,40.73896110057831)
(REGENTTEX,40.44117629528046)
(FBFIF,39.99999761581421)
(1JANATAMF,39.534878730773926)
(FORTUNE,39.33333158493042)
(SHURWID,39.31035101413727)
(ANLIMAYARN,39.20971751213074)
(RUPALIBANK,39.16914165019989)
(DESHBANDHU,39.13043439388275)
(POWERGRID,38.95832598209381)

(ICBIBANK,38.88889253139496)
(FAREASTLIF,38.218921422958374)
(IFIC1STMF,38.095247745513916)
(07.Fuel_&_Power,37.85248100757599)
(BANKASIA,37.82050907611847)
(EXIMBANK,37.254905700683594)
(NCCBANK,37.20931112766266)
(SIBL,37.14286386966705)
(HFL,36.95652782917023)
(ICBSONALI1,36.764705181121826)
(POPULAR1MF,36.58536672592163)
(TOSRIFA,36.47059202194214)
(FEKDIL,36.42857372760773)
(BBS,36.3636314868927)
(JAMUNABANK,36.24999523162842)
(PENINSULA,35.89743375778198)
(ICB,35.74879169464111)
(BSRMLTD,35.43307185173035)
(RANFOUNDRY,35.176655650138855)
(PTL,35.13513505458832)
(SKTRIMS,35.09615957736969)
(RELIANCINS,35.01198589801788)
(BATASHOE,34.97343063354492)
(LHBL,34.80662405490875)
(DBH,34.78260934352875)
(GBBPOWER,34.78260636329651)
(VFSTDL,34.736841917037964)
(MEGHNACEM,34.44444537162781)
(UCB,34.21052694320679)
(PRIMEFIN,34.11764204502106)
(AFCAGRO,33.83457958698273)
(AAMRATECH,33.75527262687683)
(AMANFEED,33.67088437080383)
(DAFODILCOM,33.4894597530365)
(BBSCABLES,32.947975397109985)
(ITC,32.74999558925629)
(NAVANACNG,32.36010670661926)
(OLYMPIC,32.31793940067291)
(DELTALIFE,32.00882971286774)

(FINEFOODS,31.94444477558136)
(ABB1STMF,31.81818425655365)
(ICBEPMF1S1,31.48147761821747)
(09.IT_Sector,31.389740109443665)
(PRIMEBANK,31.25)
(08.Insurance,30.916130542755127)
(ICB3RDNRB,30.769237875938416)
(NTC,30.447760224342346)
(ORIONPHARM,30.065354704856873)
(SHAHJABANK,29.999998211860657)
(ALARABANK,29.999998211860657)
(GHAIL,29.965150356292725)
(ETL,29.906541109085083)
(GP,29.50310707092285)
(GQBALLPEN,29.014083743095398)
(ATLASBANG,28.947368264198303)
(ARGONDENIM,28.634360432624817)
(PREMIERCEM,28.52664589881897)
(RAHIMTEXT,28.485199809074402)
(GREENDELMF,27.94117033481598)
(GHCL,27.9069721698761)
(NCCBLMF1,27.692312002182007)
(15.Services_&_Real_Estate,27.44992971420288)
(APEXSPINN,27.331706881523132)
(PDL,27.272731065750122)
(UTTARAFIN,27.241384983062744)
(ESQUIRENIT,27.227720618247986)
(BSRMSTEEL,27.22008228302002)
(ENVOYTEX,27.062708139419556)
(MPETROLEUM,26.861703395843506)
(SAMORITA,26.837065815925598)
(AGNISYSL,26.775965094566345)
(KDSALTD,26.494023203849792)
(DESCO,26.433920860290527)
(RELIANCE1,26.373621821403503)
(JAMUNAOIL,26.293110847473145)
(NLI1STMF,26.050424575805664)
(ISLAMIBANK,26.008975505828857)
(BARKAPOWER,25.968995690345764)

(ANWARGALV,25.875192880630493)
(LINDEBD,25.320139527320862)
(LRGLOBMF1,24.615390598773956)
(GRAMEENS2,24.166662991046906)
(ORIONINFU,24.060148000717163)
(DBH1STMF,23.749995231628418)
(MJLBD,23.52941185235977)
(SQUARETEXT,22.249384224414825)
(GREENDELT,22.066548466682434)
(NTLTUBES,21.998080611228943)
(EBL,21.508382260799408)
(ACIFORMULA,21.40783667564392)
(GPHISPAT,20.87227702140808)
(BXPHARMA,20.026008784770966)
(UTTARABANK,20.0000062584877)
(APEXFOOT,19.878368079662323)
(DSEX,19.804322719573975)
(00DSEX,19.804322719573975)
(APEXTANRY,19.536681473255157)
(PUBALIBANK,19.502069056034088)
(MBL1STMF,19.402988255023956)
(AMCL(PRAN),19.15268898010254)
(AIBL1STIMF,18.840575218200684)
(19.Travel_&_Leisure,17.834648489952087)
(TITASGAS,17.58241057395935)
(SUMITPOWER,17.105263471603394)
(DSES,16.895928978919983)
(00DSES,16.895928978919983)
(MATINSPINN,15.427003800868988)
(SILCOPHL,15.416669845581055)
(00DS30,15.361876785755157)
(DS30,15.361876785755157)
(UNIQUEHRL,15.208332240581512)
(IFILISLMF1,14.92537409067154)
(MTB,14.920637011528015)
(HWAWELLTEX,14.917120337486267)
(PADMAOIL,14.701922237873077)
(SEBL1STMF,13.513512909412384)
(SQURPHARMA,11.505860090255737)

(KOHINOOR,10.747406631708145)
(IBBLPBOND,9.383675456047058)
(20.Bond,9.383675456047058)
(IBNSINA,9.036387503147125)
(RENATA,7.66100287437439)

Those are the samples result of our data analysis.

5.2 Discussion

The companies whose share value fluctuated more than 50% over a month, then we can consider them as manipulated companies. After that Security Council can look over those companies to physically investigate and go for proper solution of it. Without this analysis, it is quite impossible to investigate all the companies to identify the manipulation core.

Chapter 6

FUTURE WORK & CONCLUSION

6.1 Future Work

Our present study was just to learn about the Big Data Analysis. So we have studied about the environment of Hadoop and we have provided a solution of DSE related task. Now, we tend to think that we may be able to do any sort of complex task, which is related to data analysis. If any companies or organization wants any suggestion from their previous data, we can create the relation between the problem and its solution. So, for this case, future work can be to maximize the efficiency of the result.

6.2 Conclusion

Indeed Big Data analysis is growing field in modern technology. If the question is, what is the future of it? Then the answer can be, there is no future of Big Data analysis but the future itself. Because future world will be dominated by them whose are containing as much as data. But the question is: right to privacy, trust, ethics, transparency, context, and global differences. We are still do not know what we are going to be capable of by using this type of technology. So we can say that the usage of Data analysis is going to stop at all but we should use it to make an efficient world.

Bibliography

[1] *Big Data Analytics*. https://www.slideshare.net/GhulamImaduddin1/big-data-analytics-58553692?fbclid\wAR1W1eCoU3ZAEeofZVxFE\\3777Xn0Ux2vfedwpsCDkDjY5nPF_97TJONXxBQ. (Accessed on 16/05/2019).

[2] *Big Data Analytics — Big Data Explained — Big Data Tools & Trends — Big Data Training — Edureka - YouTube*. https://www.youtube.com/watch?v\k7zu3NXEiGY&list=PLusdUbcINbvOjgGRTgtbeNJ1NyQBOBH4Y&index=2&fbclid\wAR2vt1H6R2SM7Uhf6YiY_idLn6peU\\vUmJF57ggKZqMdXk5rkJRwynC9X_gM. (Accessed on 14/06/2019).

[3] *Big Data Tutorial For Beginners — What Is Big Data — Big Data Tutorial — Hadoop Training — Edureka - YouTube*. https://www.youtube.com/watch?v=zez2Tv-bcXY&list=PLusdUbcINbvOjgGRTgtbeNJ1NyQBOBH4Y&index=2&t\\=2179s&fbclid=IwAR2-Ye5qLQNoxXA7AQs2d_2oqvRTOEcinbHdf3\\ToIDlG_6qu66mJpHhrz78. (Accessed on 10/06/2019).

[4] *Big data - Wikipedia*. https://en.wikipedia.org/wiki/Big_data?\\fbclid=IwAR0KfOqTltE5\\UEylmQdZYz9ghH9csuge1VZQ7\\ErexGpM3fq6DOyCePPAKwY. (Accessed on 03/05/2019).

[5] David Loshin. *Big data analytics: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*. Elsevier, 2013.

[6] Manoj R Patil and Feris Thia. *Pentaho for big data analytics*. Packt Publishing Ltd, 2013.

[7] Rotsnarani Sethy and Mrutyunjaya Panda. "Big Data Analysis using Hadoop: A Survey". In: *International Journal of Advance Research in Computer Science and Software Engineering* 5 (Aug. 2015).

[8] *Sources of BigData – BigData*. http://www.hadoopadmin.co.in/\\sources-of-bigdata/?fbclid=IwAR2vMw6xU16OSqbKEiWztnjZCMdUOzTJT\\mE6iuymjH4eqBcVMkozMOlvpgU. (Accessed on 05/06/2019).

[9] *The V's of Big Data: Velocity, Volume, Value, Variety, and Veracity*. https://www.xsnet.com/blog/bid/205405/\\the-v-s-of-big-data-velocity-volume-value-variety-and-\\veracity?fbclid=IwAR2o_Tgl0nlLBrxhdSgDRssF-\\lYmuAUomNOI4gp8UiFmjoVYXZOU8XdnKRE. (Accessed on 14/05/2019).