

**APPLYING DATA MINING AND VISUALIZATION
TECHNIQUES ON PINGER DATA**



By

Aqsa Hameed

2014-ag-2087

A Thesis Submitted in Partial Fulfillment of
Requirements for the Degree of

MS (Computer Science)

Supervisor: Dr. Saqib Ali

**Department of Computer Science
Faculty of Sciences
University of Agriculture, Faisalabad, Pakistan**

2016

ACKNOWLEDGEMENT

I feel much pleasure in expressing my heartiest gratitude to my supervisor, Dr. Saqib Ali, Assistant Professor, Department of Computer Science, University of Agriculture, Faisalabad, for his guidance. Further, I would like to thank Prof. Dr. Les Cottrell Rodger, Prof. Dr. Bebo White and SLAC team members who guide me during the whole period of this research. Finally, my appreciation to my parents for their permanent encouragement.

ABSTRACT

Data mining is a diagnostic procedure used to investigate substantial measure of information. As databases are becoming popular quickly in business and in all different fields of life the immense measure of information is produced from various heterogeneous sources. Also, now databases are moving towards information warehouses. In Data Warehouses, there is multidimensional information in different organizations and complex in nature called Big Data. It is utilized for examination and reporting purposes in associations. It is important to mine this data to get valuable information. Current data mining systems are not pertinent on Big datasets. A project is working to measure Internet (end-to-end) performance named as PingER. It is led by SLAC since 1998. This project has been generated huge amount of data which can reveal interesting information about power cuts, network bottlenecks and packet loss, etc. It is important to analyze PingER data to look at trends on internet connections, but it is not possible currently because loading such huge amount of data is not possible and this data is also not available for user access. In this research PingER data are analyzed by loading into Big Data platform (Data Warehouse OR HDFS) and processed by using data mining MR framework. Impala OLAP queries are applied to mine the results and get information from DWH. This resulted information is in complex format so this information is further converted in graphical form by applying visualization techniques as Bar chart and Line chart. This process makes the information access easy and understandable for users and provides better enhanced storage architecture to store big data.

CONTENTS

CHAPTER NO	TITLE	PAGE NO
Chapter 1	INTRODUCTION	1-24
Chapter 2	REVIEW OF LITERATURE	25-44
Chapter 3	MATERIALS AND METHODS	45-82
Chapter 4	RESULTS AND DISCUSSION	83-94
	SUMMARY	95-96
	LITERATURE CITED	97-103

TABLE OF CONTENTS

SR. NO	TITLE	PAGE NO
1	INTRODUCTION	1
1.1	Introduction to Data Warehouse	1
1.1.1	DWH Definition	1
1.1.2	OLAP Vs DWH	2
1.1.3	DWH Architecture	3
1.1.4	OLAP Systems	6
1.2	Data Mining	6
1.2.1	Data Mining Process	6
1.2.2	Data Mining Techniques	8
1.2.3	Semantic Web Mining	10
1.3	Big Data	12
1.3.1	Dimension of Big Data	12
1.3.2	Structure of Big Data	14
1.3.3	Big Data Framework	14
1.4	Visualization	16
1.4.1	MD Data Visualization	16
1.4.2	Visualization with Big Data	19
1.5	PingER	20
1.5.1	Mechanism	20
1.5.2	Measurement Method	21
1.5.3	Data Gathering Architecture	21
1.5.4	Data Format	22
1.6	Problem Statement	23
1.7	Research Questions	23
1.8	Objectives	23
1.9	Solution	24
1.10	Significance of the Study	24
1.11	Thesis Organization	24

2	REVIEW OF LITERATURE	25
3	MATERIALS AND METHODS	45
3.1	Research Framework	46
3.1.1	Selection and Cleaning	46
3.1.2	Transformation Process	47
3.1.3	Defining Hadoop Cluster	50
3.1.4	Loading Process	53
3.1.5	Querying Data	55
3.1.6	Visualization	57
3.2	Tools to be used	57
3.2.1	Workflow Management System	57
3.2.2	Vmware Player	57
3.2.3	Cloudera	57
3.2.4	Visualization Tools (Google Charts)	59
3.2.5	XAMPP Server	59
3.3	System Development	60
3.3.1	Transforming Data	60
3.3.2	Loading CSV files on HDFS	62
3.3.3	Querying Data in Impala	74
3.3.4	Visualizing the query results	78
4	RESULTS AND DISCUSSION	83
4.1	Testing and evaluation	83
4.1.1	Unit Testing	83
4.1.2	Integration Testing	84
4.1.3	Component Testing	84
4.1.4	Comparison	85
4.1.5	Testing Usecases	85
4.2	Results	86
4.2.1	Transformation	86
4.2.2	Loading	86

4.2.3	Querying Data	86
4.2.4	Visualization	89
4.3	Discussions	93
	SUMMARY	95
	LITERATURE CITED	97

LIST OF TABLES

TABLE NO	TABLE DESCRIPTION	PAGE NO
Table 1.1	OLTP Vs DWH	3
Table 1.2	Online Occurrences in 1 second	11
Table 3.1	Comparison of GFS and HDFS	54
Table 4.1	Comparison between Testing Techniques	85

LIST OF FIGURES

FIGURE NO	FIGURE DESCRIPTION	PAGE NO
Figure 1.1	Data Warehouse Architecture	4
Figure 1.2	Data Mining Process	7
Figure 1.3	Data Mining Techniques	8
Figure 1.4	Semantic Web Infrastructure	12
Figure 1.5	Dimensions of Big Data	13
Figure 1.6	Big Data Architecture	15
Figure 1.7	Line Graph	17
Figure 1.8	Bar Chart	17
Figure 1.9	Pie Chart	18
Figure 1.10	Scatter Plot Diagram	18
Figure 1.11	Bubble Plot Diagram	19
Figure 1.12	Summary Table of PingER RTT Values	20
Figure 1.13	Data Gathering Architecture of PingER	22
Figure 3.1	PingER Dataflow Architecture	45
Figure 3.2	Proposed Research Framework	46
Figure 3.3	Star Schema of PingER Data Warehouse	48
Figure 3.4	MapReduce Architecture	49
Figure 3.5	Hadoop Cluster Architecture	50
Figure 3.6	Hadoop Cluster Components	52
Figure 3.7	Internal Working of 2 nd Component Masters	52
Figure 3.8	Internal Working of 3 rd Component Slaves	53
Figure 3.9	HDFS Architecture	55
Figure 3.10	Architecture of Impala	56
Figure 3.11	Cloudera CDH Core Components	58
Figure 3.12	Setting up system environment variables	61
Figure 3.13	Creating new system variable	61
Figure 3.14	Testing java installation	62

Figure 3.15	Opening the cloudera VM	63
Figure 3.16	Editing the VM settings	64
Figure 3.17	Changing processor and memory settings	64
Figure 3.18	Changing OS setting	65
Figure 3.19	Cloudera Desktop	65
Figure 3.20	Cloudera startup page	66
Figure 3.21	CM express launch screen	67
Figure 3.22	Running CM	68
Figure 3.23	Starting up CM services	68
Figure 3.24	Cloudera home document directory	69
Figure 3.25	Running Hue	70
Figure 3.26	Hue configuration check	70
Figure 3.27	Hue configuration check error message window	71
Figure 3.28	HDFS directories window	71
Figure 3.29	Creating new HDFS directory	72
Figure 3.30	PingER HDFS directory created	72
Figure 3.31	Changing permissions of HDFS directory	73
Figure 3.32	Uploading files to HDFS	73
Figure 3.33	Restart cloudera VM	74
Figure 3.34	Load data statement error message window	74
Figure 3.35	Introducing Impala Interface	75
Figure 3.36	Creating external database tables	75
Figure 3.37	Data loaded in external tables	76
Figure 3.38	Results of query	76
Figure 3.39	Saving query results as CSV	77
Figure 3.40	Locating CSV in downloads folder	77
Figure 3.41	Installing XAMPP	78
Figure 3.42	Starting XAMPP Apache services	78
Figure 3.43	HTML page code 1	79
Figure 3.44	HTML page code 2	80
Figure 3.45	HTML page code 3	81

Figure 3.46	Column chart of PingER data	81
Figure 4.1	Testing Techniques	84
Figure 4.2	Testing Usecase Diagram	85
Figure 4.3	Results of Impala Query 1	88
Figure 4.4	Results of Impala Query 2	89
Figure 4.5	Results of Impala Query 3	89
Figure 4.6	Visualization Bar chart of Query 1	90
Figure 4.7	Visualization Bar chart of Query 2	90
Figure 4.8	Visualization Bar chart of Query 3	91
Figure 4.9	Visualization Line chart of Query 1	91
Figure 4.10	Visualization Line chart of Query 2	92
Figure 4.11	Visualization Line chart of Query 3	92

LIST OF ABBREVIATIONS

DWH	-	Data Warehouse
BI	-	Business Intelligence
OLTP	-	Online Transaction Processing
RDBMS	-	Relational Database Management System
ODS	-	Operation Data Store
OLAP	-	Online Analytical Processing
AI	-	Artificial Intelligence
KDD	-	Knowledge Discovery
DMT	-	Data Mining Techniques
ML	-	Machine Language
NLP	-	Natural Language Processing
IRS	-	Information Retrieval System
SLAC	-	Stanford Linear Accelerator Center
PingER	-	Ping (End-to-End) Reporting
RTT	-	Round Trip Time
ICMP	-	Internet Control Message Protocol
MA	-	Measurement Agent
CSV	-	Comma Separated Values
LOD	-	Linked Open Data
RDF	-	Resource Description Framework
HDFS	-	Hadoop Distributed File System
MR	-	MapReduce
GFS	-	Google File System
SDM	-	Semantic Data Mining
PCA	-	Principal Component Analysis
MI	-	Mutual Information
RMSE	-	Root Mean Square Error
ROC	-	Receiver Operating Charectertistics
AUC	-	Area Under Curve

PDA	-	Personal Device Assistant
VC	-	Virtual Cluster
VM	-	Virtual Machine
IaaS	-	Infrastructure as a Service
IDH	-	Intel Distributed Hadoop
RDD	-	Resilient Distributed Data
CAV	-	Context Adaptive Visualization
SNA	-	Social Network Analysis
UDF	-	Unified Data Framework
ADMIRE	-	Advance Data Mining and Research for Europe
API	-	Application Programming Interface
IPC	-	Indian Penal Code
ETL	-	Extraction Transformation Loading
MPP	-	Massively Parallel Processing
CDH	-	Cloudera Distributed Hadoop
JRE	-	Java Runtime Environment
CM	-	Cloudera Manager

Chapter 1

INTRODUCTION

1.1. Introduction to Data Warehouse

The idea of a Data Warehouse (DWH) was introduced to store large amount of data that is collected from multiple sources and needs to be processed for reporting and analysis use. In the era of 1970's many organizations starting investing on new computer systems which automates the business process. This term arise from the application of Business Intelligence (BI). The new systems are able to support in decision making and effective business process. Organizations are focusing on enhance decision making by use of operational data sources (Connolly and Begg, 2005).

1.1.1. DWH Definition

As data is growing in the main challenge for organization is to utilize the storage data archives by turning into useful information and knowledge. The original concept of DWH was devised from IBM as the information warehouse and presented as a solution to access non-relational data. In data warehouse there is multi-dimensional data which are in an unstructured format. The latest work on DWH is done by Bill Inmon which developed the term DWH in the field of BI in 1993. The basic definition of DWH is.

“DWH is an integrated, non-volatile, subject oriented and time-variant storage of data used to support the business management in the decision making process.”

Data warehouses are Subject-oriented in nature as they are systematized about the main focusing interests of business organizations such as products, sales and customers instead of the major application domains such as stock control, product sales and customer invoicing. Multi-dimensional and heterogeneous data coming from multiple sources is combined as a one source of information in DWH. Because the data in DWH used for analysis purpose, therefore it contains the data according to a time interval and accurate for a specific period. The data in DWH are updated on a daily basis in real time mode the data is updated from operational data sources in a change request mode this is known as non-volatility of the data.

The Web is a massive storage of behavioral information as people connect through their Web programs with remote Web locales. The information created by this conduct is called clickstream. Utilizing an information distribution center on the Web to outfit clickstream information has prompted the advancement of Data Web-houses. The concept of web-house is introduced because semantic web is an application of data mining, which will be introduced later in this research. To discuss the semantic web it is necessary to introduce the data and web.

1.1.2. Online Transaction Processing (OLTP) Vs. DWH

A Database Management System (DBMS) worked for OLTP is by and large viewed as inadmissible for information warehousing on the grounds that every framework is outlined in light of a contrasting arrangement of prerequisites. For instance, OLTP frameworks are intended to boost the exchange handling limit, while information distribution centers are intended to bolster specially appointed question preparing. An association will regularly have various diverse OLTP frameworks for business procedures, for example, stock control, client invoicing, and purpose of-offer. Conversely, an association will ordinarily have a solitary information stockroom, which holds information that is authentic, detailed, and outlined to a different levels and subject to change rarely. A comparison between OLTP and Data warehousing system is given in Table.1.1.

Table 1.1: OLTP Vs. DWH (Connolly and Begg, 2005).

OLTP	DWH
Contains conventional data	contains historical data
Archives particular and detailed information	Archives detailed and summarized data
Data is changing continuously	Data is not changing continuously
Has duplication in processing	Has undefined, unformatted, Ad hoc and heuristic processing
Level of transactional output is high	Level of transactional output is medium to low
Usage designs are predictable	Usage designs are unpredictable
Transactions are carried out	Analysis is carried out
Based around applications	Based on subjects of users interest
Supports daily decisions	Supports planning decisions
Provide services to a large number of operational and clerical users	Provide services to low number of managerial users

1.1.3. DWH Architecture

The architecture and major components of a data warehouse are presented by (Summer and Ali, 1996). The processes, tools, and technologies associated with data warehousing and more detail is given below. The typical architecture of a data warehouse is shown in Figure 1.1.

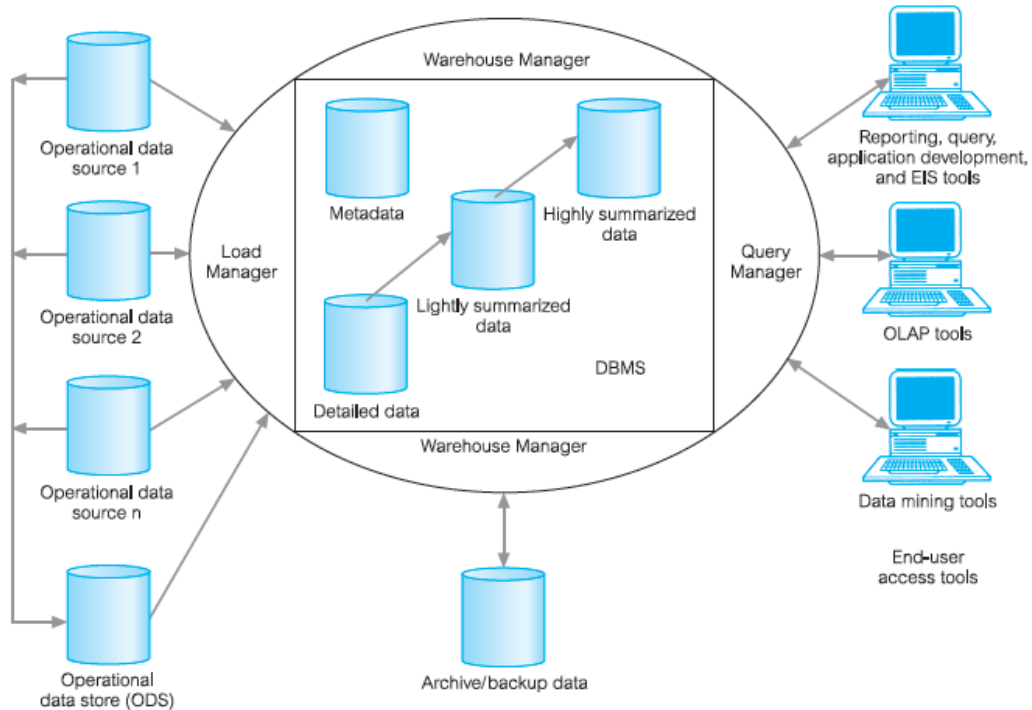


Figure 1.1: Data Warehouse Architecture (Summer and Ali, 1996)

Operational Data Sources

An Operational Data Store (ODS) is a storehouse of present and coordinated operational information utilized for investigation. It is regularly organized and supplied with information, similarly as the information warehouse yet might be going about as an arranging territory for information to be moved into the distribution center. The ODS is frequently made when legacy operational frameworks are observed to be unequipped for accomplishing reporting necessities. The ODS gives clients the convenience of a social database while staying far off from the choice bolster elements of the information stockroom.

Load Manager

The load manager performs every operation connected with the extraction and stacking of information into the stockroom. The information might be extricated straightforwardly from the information sources or all the more ordinarily from the operational information store. The operations performed by the heap administrator may incorporate basic changes of the information to set up the information for passage into the stockroom. The size and intricacy of this part will change between information stockrooms and might be developed utilizing a mix of seller information stacking devices

and custom-assembled programs.

Warehouse Manager

The warehouse manager carried out all the jobs associated with the management of the data in the warehouse. This segment is developed utilizing merchant information administration devices and custom-constructed programs. Now and again, the distribution center supervisor additionally produces question profiles to figure out which records and conglomerations are fitting. An inquiry profile can be created for every client gathering of clients or the information stockroom and depends on data that depicts the attributes of the questions, for example, recurrence, target tables, and size of result sets.

Query Manager

The query manager performs every one of the operations connected with the administration of client queries. This segment is commonly developed utilizing seller end-client information access instruments, information warehouse checking devices, database offices, and custom-manufactured projects. The multifaceted nature of the question chief is dictated by the officers gave by the end-client access apparatuses and the database. The operations performed by this segment incorporate guiding questions to the proper tables and planning the execution of queries.

Archive/Backup Data

This territory of the warehouse stores the summarized and detailed information for the motivations behind filing and reinforcement. Indeed, even albeit rundown information is produced from a point by point information, it might be important for reinforcement online synopsis information on the off chance that this information is kept past the maintenance time frames for detailed information. The information is exchanged to capacity chronicles, for example, magnetic tape or optical disk.

End-user Access Tools

The key motivation behind information warehousing is to give data to business clients to vital basic leadership. These clients cooperate with the distribution center utilizing end-client access apparatuses. The information distribution center should proficiently bolster impromptu and routine examination. High performance is accomplished by pre-arranging the prerequisites for joins, summations, and intermittent reports by end-clients.

1.1.4. OLAP Systems

OLAP stands for Analytical Processing System. It points out the advancements in technology which permit clients to effectively recover information from the information distribution center. The qualities of an OLAP framework are very not quite the same as those of value-based database frameworks, called as On-line Transaction Processing (OLTP) frameworks. To encourage complex investigations and perception, the information in a stockroom is commonly demonstrated multidimensionality.

OLAP operations incorporate rollup expanding the level of collection, drill-down diminishing the level of accumulation alongside one or more measurement chains of command, slice_and_dice determination and projection of information, and pivot re-orienting the multidimensional perspective of information (Chaudhuri and Dayal, 1997).

Decision support system generally require combining information from numerous heterogeneous sources these might incorporate outer sources, for example, securities exchange bolsters, notwithstanding a few operational databases. OLAP requires extraordinary information association, access techniques, and usage strategies, not for the most part gave by business DBMS

1.2. Data Mining

To find knowledge from information inside the stockrooms and handling a term is utilized which is called data mining. Data mining is a stage in finding learning. It is a non-unimportant procedure of distinguishing substantial, novel and beforehand obscure potential helpful data from information. Data mining is generally utilized as a part of AI (Artificial Intelligence) since 1960 furthermore applying in numerous different spaces like data frameworks, neural systems, business insight, machine learning, wellbeing sciences, natural sciences and conveyed stages like GRID and distributed computing (Wang *et al.*, 2009).

1.2.1. Data Mining Process

The term Knowledge Discovery (KDD) in databases refers to the broad process of finding knowledge from data, and focuses on the high-level application of particular data mining methods. Some people don't differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery (Kantardzic, 2011). The steps involved in DM process are discussed below and depicted in Figure 1.2.

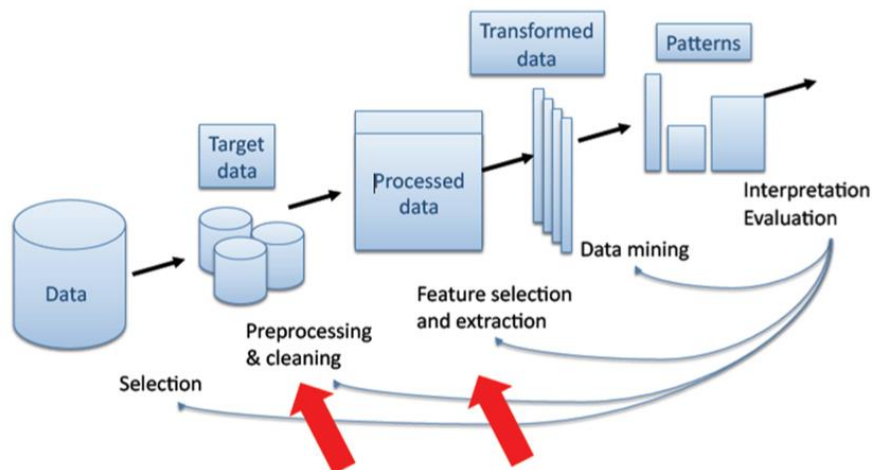


Figure 1.2: Data Mining Process (Indarto, 2013)

Selection and Cleaning

The data located in warehouses are in rough and unstructured format. Further Data warehouses also contain multidimensional data. In this step the data were selected from warehouse according to user interest. For example, user wants to analyze the specific dimensional data of sales, purchase and its prize. Then the user can select data according to these dimensions. Selection varies on the basis of user's interest.

Cleaning is the process of simplifying data. Data collection methods are loosely controlled and resulted in out of range values. Generally garbage out term is used to clean the data. It consists of performing many operations like removal of noisy data and outliers, finding missing values, collecting the necessary information to model the data, etc. The output of this step is our target data which are analyzed.

Transformation

In this step target data is transformed into a suitable structure or schema is defined for the data. Necessary summary, data aggregation, data reduction and projection operations are performed.

Data Mining

In this step appropriate data mining techniques and algorithms are applied to process the data. This is the basic step in KDD. Data mining techniques include classification, regression, association rules mining, pattern matching, etc. These techniques are discussed in detail in Section 1.2.2.

Interpretation and Evaluation

In the last step the results of data mining steps are analyzed and evaluated. This is the processed form of data turned into useful knowledge. The results are integrated with the data in warehouses and easily understandable to humans.

1.2.2. Data Mining Techniques

Different data mining techniques are introduced to process data. DMT are divided into predictive or descriptive categories. Predictive techniques are used to predict the future from the data and descriptive techniques are used to derive information from data where data involves relationships and effecting factors upon results (Elmasri and Navathe, 2011). Basic DMT are given in Figure 1.3.

Classification

Classification is a method of taking in a model which delineates unmistakable groups of data. The groups are predestined. This kind of development is moreover called directed learning. Once the model is manufactured, it can be used to portray new data. The underlying stride taking in the model is mastered by using an arrangement set of data that has starting now been gathered. Each record in the arrangement data contains a quality, called the class mark, which demonstrates which class the record has a spot with. The model that is made is as a rule as a decision tree or a game plan of precepts. A decision tree is basically a graphical representation of the delineation of each class or, by the day's end, a representation of the gathering rules.

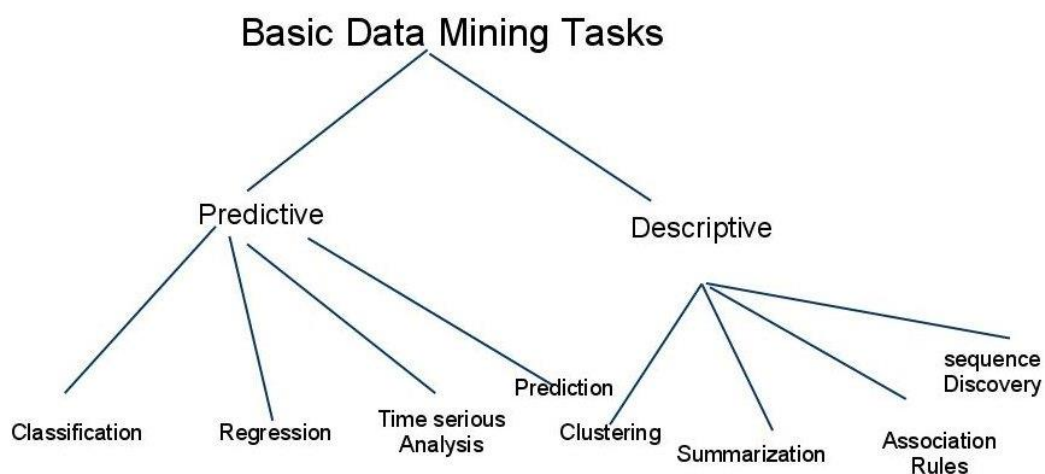


Figure 1.3: Data Mining Techniques (Bigdatanerd, 2011)

Regression

Regression is a unique use of the characterization guideline. On the off chance that an older standard is viewed as a capacity of the variables which guide these variables into an objective class variable, the tenet is known as a relapse principle. A general use of relapse happens when, rather than mapping a tuple of information from a connection to a particular class, the estimation of a variable is anticipated to take into account that tuple. Relapse investigation is an extremely regular instrument for examination of information in numerous exploration spaces. The disclosure of the capacity to anticipate the objective variable is equal to an information mining operation.

Time Series Data Analysis

Time series are groupings of occasions every occasion might be a given altered kind of an exchange. For instance, the end cost of a stock or an asset is an occasion that happens each weekday for every stock and reserve. The arrangement of these qualities per stock or reserve constitutes a period arrangement. For a period arrangement, one may search for an assortment of examples by examining groupings and subsequences. Investigation and mining of time arrangement are an amplified usefulness of worldly information administration.

Prediction

Information mining can demonstrate how certain properties inside the information will carry on later on. Case of prescient information mining incorporates the examination of purchasing exchanges to foresee what shoppers will purchase under certain rebates, the amount of offers volume a store will create in a given period, and whether erasing a product offering will return more benefits. In such applications, business rationale is utilized combining with information mining. In an exploratory setting, certain seismic wave examples may anticipate a tremor with high likelihood.

Clustering

The past information mining errand of order manages parceling information in light of utilizing a pre-arranged preparing the test. Nonetheless, it is regularly helpful to parcel information without having a preparation test in grouping. This is otherwise called

unsupervised learning. The objective of bunching is to place records into gatherings, such that records in a gathering are like each other and not at all like records in different gatherings. The gatherings are generally disjoint. A critical element of grouping is the similitude work that is utilized. At the point when the information is numeric, a similitude capacity in view of separation is commonly utilized. K-Means bunching calculation is surely understood the calculation in grouping information mining.

Association Rules

Association rules associate the nearness of an arrangement of things with another scope of qualities for another arrangement of variables. For instance, When a female retail customer purchases a satchel, she is liable to purchase shoes or when a X-beam picture containing attributes an and b is liable to likewise display trademark c.

Sequence Discovery

Sequential patterns depend on the idea of an arrangement of itemsets. The rules which represent exchange are accepted, for example, the general store wicker bin exchanges are requested from time of procurement. That requesting yields a succession of itemsets. For instance, {milk, bread, juice}, {bread, eggs}, {cookies, milk, coffee} might be such an arrangement of itemsets taking into account three visits by the same client to the store. The backing for a succession S of itemsets is the rate of the given set U of arrangements of which S is a subsequence.

1.2.3. Semantic Web Mining

From the most recent couple of years with the expansion in the utilization of web numerous association and business begin depending on the web. Numerous assets and data is accessible on the web with the goal that the web has created the enormous measure of information. Information is accessible on the web, however every one of the information is in harsh structure and it is futile until it is not composed in a great way. The client required less information from this gigantic measure of information which is just identified with its required information, remaining information is futile for him and the client has no worry with it. Web Data mining permits an organization to utilize the mass measures of information that it is collected and composed to help business and to bolster the basic leadership. To analyze the internet usage a survey is conducted

(Adamov, 2014). The result of online occurrence in one second is given below in Table 1.2.

Table 1.2: Online Occurrence in 1 second (Adamov, 2014)

Online Occurrence	Number
Instagram photos uploaded	5000
SPAM emails sent	1800000
New websites created	9
Tweets tweeted	4000
Google searches made	45000
Skype calls made	1500
YouTube videos viewed	92000
Dropbox files uploaded	10000
Global Internet transfer	25 TeraBytes
Facebook likes made	55000

Semantic web, allows us a keen administration in which coincide and mastermind of the information can be on the web in order way. Because of gigantic exploration on web mining, numerous methods and applications have been presented. Web mining has presented numerous new research fields like Machine Language (ML), Natural Language Processing (NLP), measurements and Information Retrieval (IR). Semantic web mining utilizes ontologies to mine information. Ontologies are the past space learning when cosmology coordination in web mining is determined and coordination of the past key terms are used in information to determine data. Ontologies also help us to look at the previous patterns of data. Infrastructure of semantic web mining is given below in Figure 1.4.

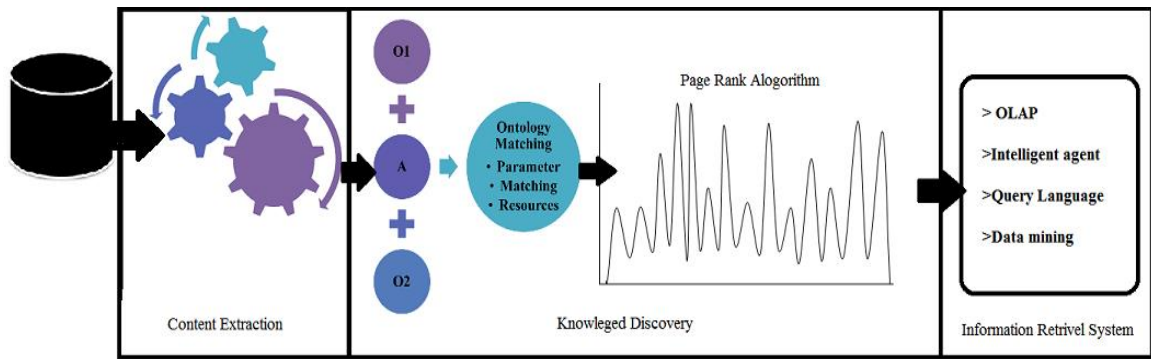


Figure 1.4: Semantic Web Infrastructure

Ontology matching is a process in knowledge discovery. Page rank algorithm measures the ranking of web sites that how frequently a web sites was accessed and after knowledge discovery the typical data retrieval tools are applied called Information Retrieval Systems (IRS).

1.3. Big Data

Big Data point out towards the information sets which are vast, complex, includes variety and self-developing. This term was presented in 1998. The applications, similar to Google, cellular telephones, Facebook, Flickr and other online networking offer ascent to huge information. 5 billion mobiles are being used in 2010. Facebook stores 15 PB information and loads more than 60 PB information day by day. Google has ordered 1 billion pages in 2000 and surpasses from 1 trillion in 2008. Glint shared 1.8 million photographs for every day this requires 3.6 TB information storage room every day. As these applications raise the information size to a degree that is unmanageable and not able to handle with customary procedures.

1.3.1. Dimensions of Big Data

Big Data dimensions are classified as 3V's (Volume, Velocity, Variety). In some literature, there is more 2V's (Veracity, Value) defined by (Burbank, 2016). All V's are discussed in this section. The dimensions of big data are given in Figure 1.5.

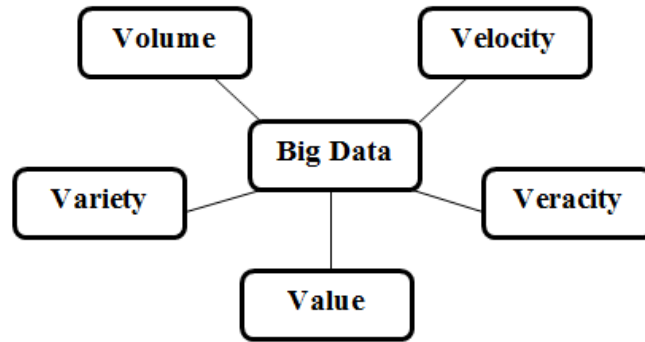


Figure 1.5: Dimension of Big Data

- i. **Volume:** It refers to the size of data. The size of Big Data is self-growing in nature. The data size of Big Data can be analyzed from previous examples that how much the size has grown in big data applications
- ii. **Velocity:** It refers to the speed of data that how quickly the data is moving. In real time systems, data comes in the form of continuous data streams and the interest is to obtain information from it. High frequency stock trading, machine to machine process exchanging data, massive logs generated from sensors is the example of velocity in big data.
- iii. **Variety:** It refers to the structure and type of data. Big data not only include text, numbers and strings, but it also includes geospatial data and 3D data, multimedia and hypermedia.
- iv. **Veracity:** It presents the context in which user wants to study the data and refers to the correctness and accuracy of data. Big Data needs to be integrated with entire information.
- v. **Value:** It includes data values for business organizations which helps in making decisions and future predictions. Except this big data technologies, discover new insights from data that derive significant business value. Customer sentiment analysis, customer usage patterns from data, marketing patterns and search patterns are examples of valuable data from different organizations.

1.3.2. Structure of Big Data

The structure of data can be further classified into 3 categories structured, semi-structured and unstructured data.

- i. Structured data: includes text, numbers, strings, dates and Relational data (row-column schema).
- ii. Semi-structured data: may be irregular or incomplete. It allows data from multiple sources with similar properties. It generally has some structure but does not conform to a fix schema like in structured data. XML documents are example of semi-structured data.
- iii. Unstructured data: can be textual and non-textual form. Textual data include data from the web like emails, blogs, power point presentation etc. and non-textual data includes audio, video files JPEG and MP3 files.

1.3.3. Big Data Framework

In the period of Big Data, it is obviously clear that associations need to utilize information driven basic leadership to increase upper hand. Preparing, incorporating and interfacing with more information ought to improve it information, giving both more all-encompassing and more granular perspectives to help vital basic leadership. This is made conceivable by means of Big Data abusing moderate and usable Computational and Storage Resources (Tekiner and Keane, 2013). Big Data framework is defined in three stages given in Figure 1.6.

Layer 1 deals with procurement and sifting of information by applying right metadata and procedures. Various information sources are incorporated and changed to add intending to the information. This procedure is the significant wellspring of increased the value of information and permits associations to increase upper hand.

Layer 2 utilizes the data readied as a part of Stage 1 to apply examination and prescient models to discover connections and examples that were not at first known. The level of insight connected relies on upon the computational capacities and ability set accessible together with the business prerequisites. Enormous Data utilizes interior and outer datasets from an assortment of sources to give data to help vital basic leadership to increase upper hand. It permits concentrate on the current and the future as opposed to customary authentic reality.

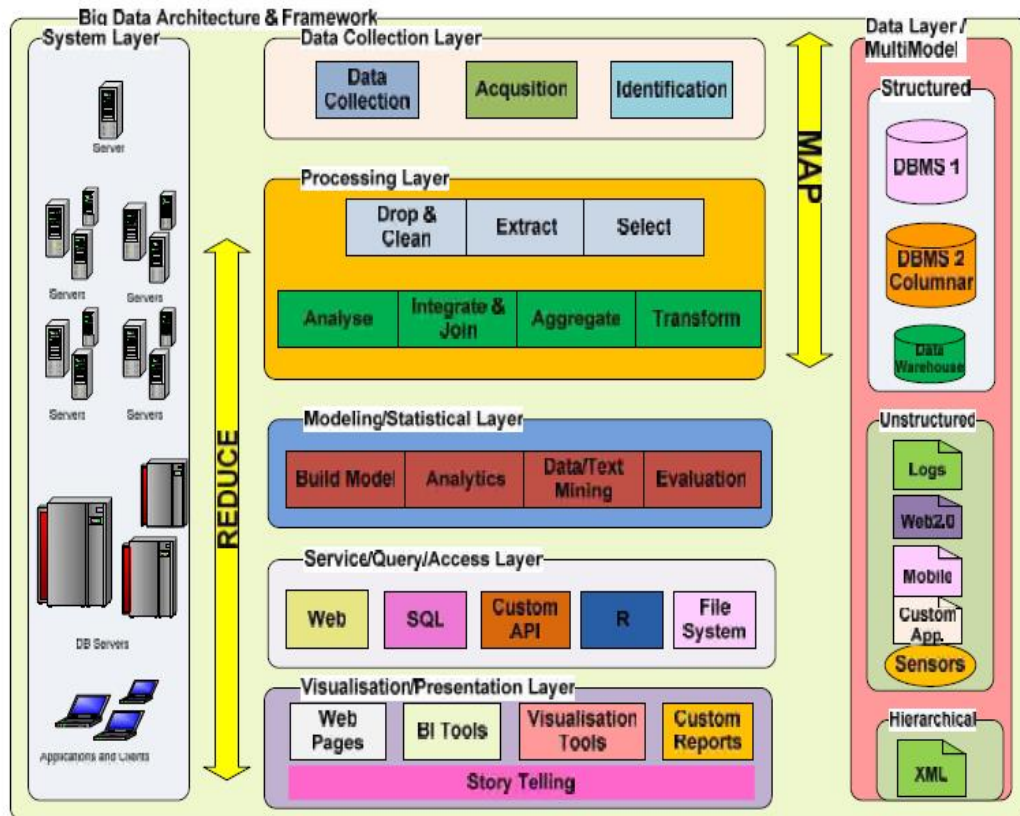


Figure 1.6: Big Data Architecture (Tekiner and Keane, 2013)

Layer 3 manages to display the source data, mapping the information to the objective model and deciphering the importance of the newfound data. The social information model does not actually suitable for the unstructured and heterogeneous information sources that are required to be accessible in Big Data applications.

Layer 4 gives techniques for getting to the information. There has been countless that attention on giving access to these information sources by means of NoSQL without utilizing SQL. They endeavor to make indexing plans like RDMBS and give fast access to information dwelling in the Hadoop document framework.

Layer 5 gives presentation and representation of information which is a critical assignment. The NoSQL choice changes the elements regarding getting to and exhibiting the information. The expanding information to be examined and handled, along these lines, yield needs to address both clarity and exactness of presentation. Moreover, elucidation of results is a noteworthy test that requires exceptionally talented staff.

Layer 6 The Processing stages portrayed guide onto the 7 layers of the structure. Every application may concentrate on various layers and may not utilize all parts of it. A Big Data application, then turns into a noteworthy coordinating exertion whereby an expansive number of moving parts should be formed to work flawlessly to accomplish results that empower upper hand.

1.4. Visualization

Visualization refers to present information in pictorial and graphical format. Mostly visualization is used to present information in a more understandable way for humans. The concept of visualization arises from data mining where it is required to process large data sets to derive information and perform data analysis. Although the algorithm of data mining is very complex and process is more abstract, but still the results of extraction are difficult to understand. A picture is worth of thousand words and close to human understanding. Therefore, visualization techniques are applied to data to make the data simpler or represent in graphical format.

Different data mining visualization techniques are used to deal with two, three, multidimensional and hierarchical data. In two dimensional data cartogram and dot distribution map can be used. In 3D timeline, timeseries and stream graph can be used. In multidimensional data line chart, pie charts, histogram, bar charts, unordered bubble charts can be used. In a hierarchical data radial tree, hyperbolic tree and tree-mapping can be used (Zoss, 2015). As this research based on multidimensional datasets so here only MD techniques are discussed in detail and these are as follows.

1.4.1. MD Data Visualization

In this section visualization techniques for multi-dimensional data are discussed. The more popular and widely used methods are Line chart, Bar Chart, Pie Chart, Bubble Chart and Scatter Plot (Choy *et al.*, 2012).

Line Chart

Line outline demonstrates the relationship of one variable to another. They are frequently used to track changes or patterns after some time. Line graphs are likewise helpful when looking at various things over the same time frame. See Figure 1.7.

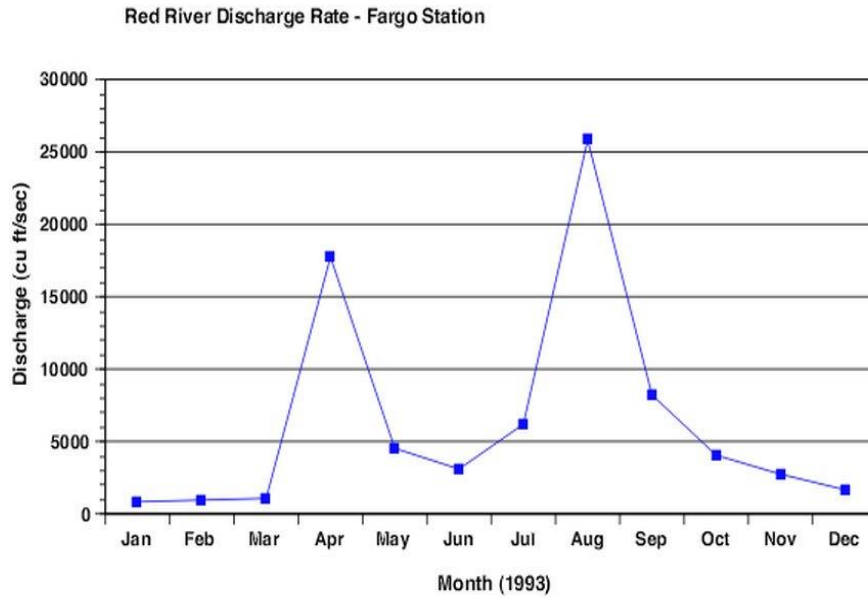


Figure 1.7: Line Graph

Bar Chart

Bar outlines are most generally utilized for looking at the amounts of various classes or gatherings. The estimations of a class are spoken to utilizing the bars, and they can be arranged with either vertical or flat bars with the length or stature of every bar speaking to the worth. Figure 1.8. Demonstrate the example of a bar chart.

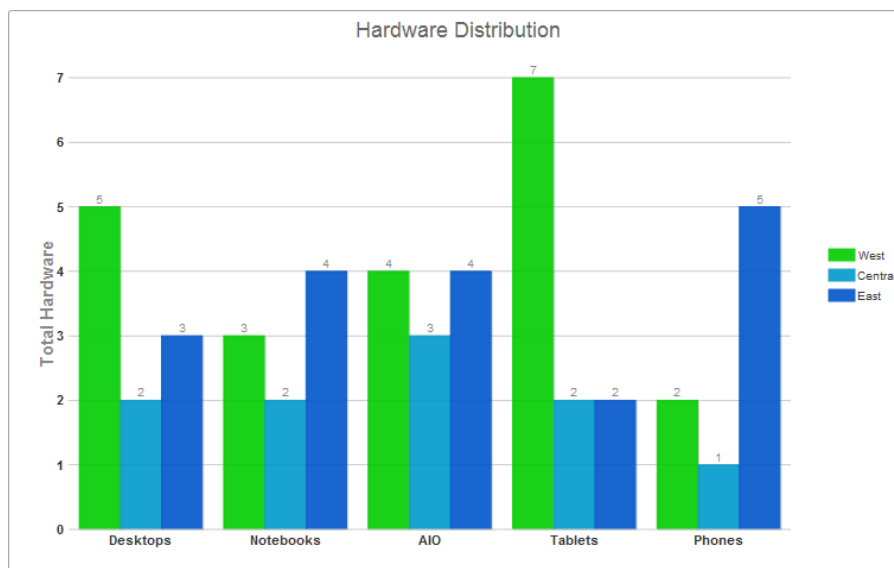


Figure 1.8: Bar Chart

Pie Chart

Pie diagrams are utilized to think about the parts of an entirety. Pie outlines are best when there are constrained parts and when and rates are incorporated to depict the substance. A Pie chart is depicted in Figure 1.9.

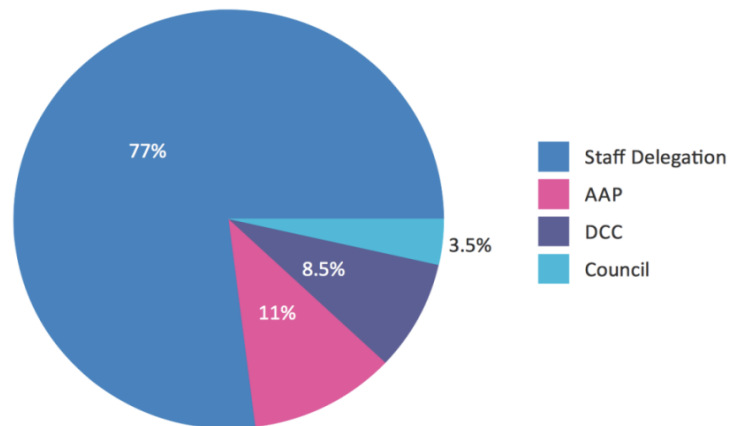


Figure 1.9: Pie Chart

Scatter Plot

A scatter plot or X-Y plot is a two-dimensional plot that demonstrates the joint variety of two informative things. In a dissipate plot, every marker (images, for example, dabs, squares and in addition to signs) speaks to a perception. The marker position shows the worth for every perception. A scatter plot is shown in Figure 1.10.

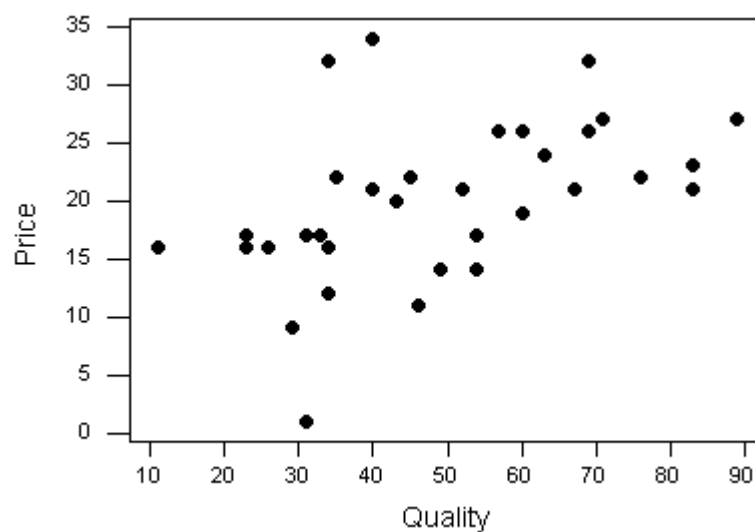


Figure 1.10: Scatter Plot Diagram

Bubble Plot

A bubble plot is a variation of a disperse plot in which the markers are supplanted with bubbles. In a bubble plot, every bubble speaks to a perception. The area of the bubble speaks to the worth for two measured tomahawks and the span of the bubble speaks to the quality for a third measure. A bubble plot is helpful for information sets with handfuls to many qualities or when the qualities contrast by a few requests of greatness. A bubble plot is given in Figure 1.11.

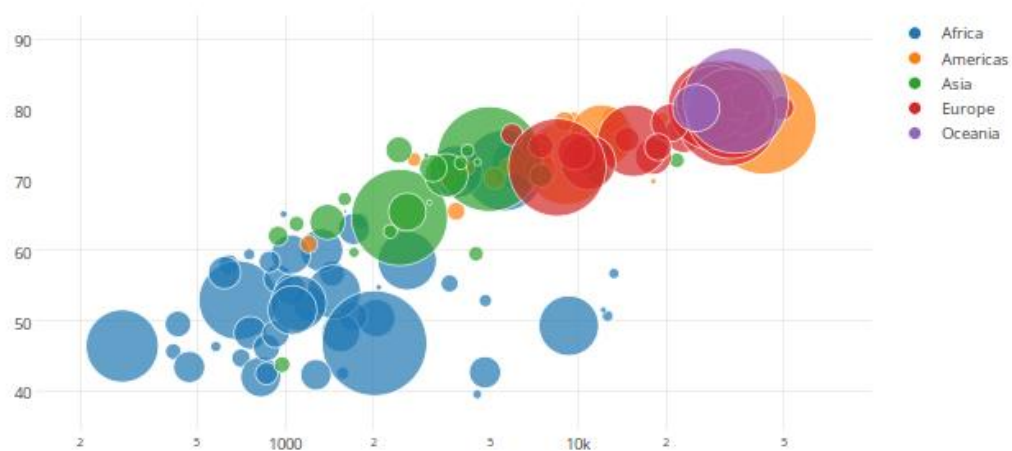


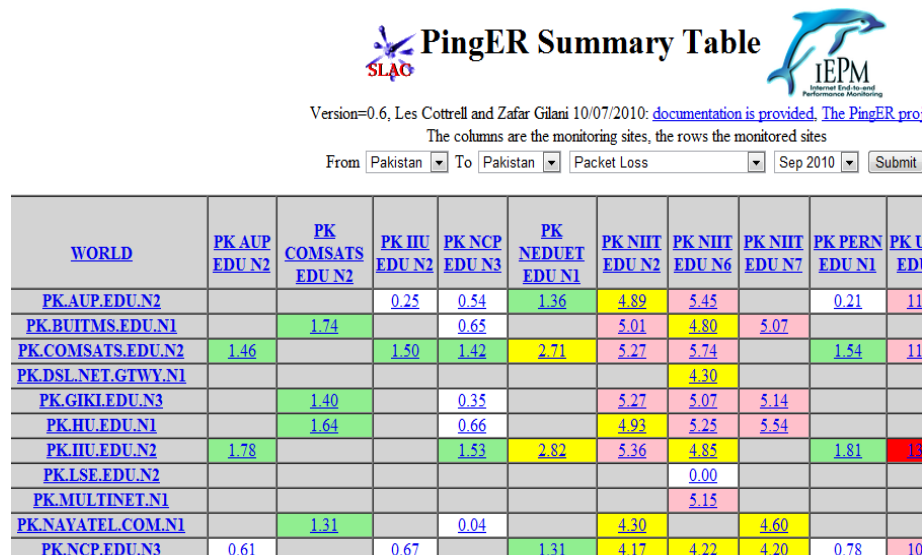
Figure 1.11: Bubble Plot Diagram

1.4.2. Visualization with Big Data

As it has been discussed above that the Big Data refers to large volume, heterogeneous and complex data sets. The data in Big Data involves various structures, mostly it is unstructured data. When visualization with Big Data sets are discussed this task becomes more complex. Further visualization in Big Data also involves a critical factor reduction of dataset and its dimensions. Because in Big Data there is multidimensional data therefore it is necessary to apply some data reduction techniques on it. The most commonly used data dimension reduction methods are Missing Value, High Correlation Filter, Backward Feature Elimination, PCA and Random Forests (Silipo *et al.*, 2014).

1.5. PingER

The operations of the internet have led to a significant importance, so it is important to measure the performance of internet connection. For this purpose PingER project is working to measure internet end-to-end performance. The IEPM group at the Stanford Linear Accelerator Center (SLAC) maintains the PingER (Ping End-to-end Reporting) project. It monitors the end-to-end performance of Internet links worldwide. The project preserves a vast data repository of network performance measurements from and to sites all around the world. The repository contains data since 1998 and several associated applications have been developed and experimented in collaboration with universities and laboratories in South America, Europe, Pakistan and Malaysia. PingER summary table of packet loss from Pakistan is shown in Figure 1.12.



Version=0.6, Les Cottrell and Zafar Gilani 10/07/2010: [documentation is provided, The PingER pro:](#)
The columns are the monitoring sites, the rows the monitored sites

From To Packet Loss Sep 2010

WORLD	PK AUP EDU N2	PK COMSATS EDU N2	PK IIU EDU N2	PK NCP EDU N3	PK NEDUET EDU N1	PK NIIT EDU N2	PK NIIT EDU N6	PK NIIT EDU N7	PK PERN EDU N1	PK U EDU N1
PK.AUP.EDU.N2			0.25	0.54	1.36	4.89	5.45		0.21	11
PK.BUITMS.EDU.N1		1.74		0.65		5.01	4.80	5.07		
PK.COMSATS.EDU.N2	1.46		1.50	1.42	2.71	5.27	5.74		1.54	11
PK.DSL.NET.GTWY.N1							4.30			
PK.GIKI.EDU.N3		1.40		0.35		5.27	5.07	5.14		
PK.HU.EDU.N1		1.64		0.66		4.93	5.25	5.54		
PK.IIU.EDU.N2	1.78			1.53	2.82	5.36	4.85		1.81	11
PK.ISE.EDU.N2							0.00			
PK.MULTINET.N1							5.15			
PK.NAYATEL.COM.N1		1.31		0.04		4.30		4.60		
PK.NCP.EDU.N3	0.61		0.67		1.31	4.17	4.22	4.20	0.78	10

Figure 1.12: Summary Table of PingER RTT values (Cottrell and Gilani, 2010)

It uses ping command to measure the performance of internet connection. This project has generated a huge amount of data and now it is needed to process this data to provide access to this information.

1.5.1. Mechanism

The fundamental component which is utilized is Internet Control Message Protocol (ICMP) reverberation system otherwise called ping strategy. It permits sending a parcel of a client chose length to a remote hub and have it resounded back. These days,

it's more often than not coming pre-introduced at all stages, so there is nothing to introduce on the customers. The server keeps running at a high need (e.g. In the piece on Unix) as will probably give a decent measure of system execution than a client application. It is extremely unobtrusive in its system data transfer capacity necessities 100 bits for every second per checking remote-host-pair.

1.5.2. Measurement Method

This anticipates ping the arrangement of remote locales from a checking node called Measurement Agent (MA). MA pings the remote locales after at regular intervals and ping order is executed 11 times sequentially 100 bytes of each with the distinction of 1 second. To begin with ping summon is discarded in light of the fact that it is ventured to be moderate. The base normal RTT time for 10 pings is recorded. This rehash the and 1000 bytes. The utilization of two sizes of information parcels makes a great appraisal of ping rates furthermore vary amongst little and huge bundles. The ping reaction time is plotted for every half hour for every node.

1.5.3. Data Gathering Architecture

The architecture of data gathering consists of 3 components, Remote Monitoring Sites, Monitoring Sites and Archives. The architecture is given below in Figure 1.13.

- i. **Remote Monitoring Sites:** These are the host remote websites whose performance are to be measured. PingER is monitoring more than 700 websites all over the world.
- ii. **Monitoring Sites:** These are the PingER monitoring servers and these sites need to install and configure set of ping tools. These are the servers which provide data on demand via the HTTP request and ping the remote sites. On these servers PingER tools are installed which provide short term data and reports.
- iii. **Archives:** These sites are may be located on a single site or on a single host. These sites contain the actual database repository of PingER. PingER currently has two such sites one is located in NUST (Islamabad, Pakistan) and second is located at SLAC (US).

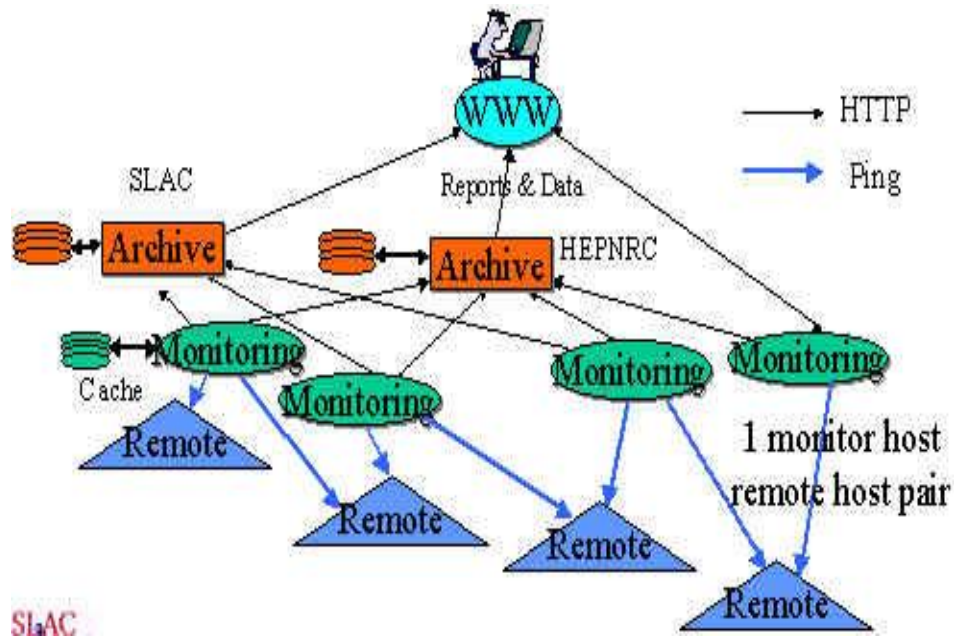


Figure 1.13: Data Gathering Architecture of PingER (Cottrell *et al.*, 2014)

1.5.4. Data Format

Initially the project started storing data in the format of flat CSV files in a form Linked Open Data (LOD). But this method is not suitable for storing and retrieving data due to lack of satisfactory access to PingER data (Cottrell *et al.*, 2010). After this data was transformed into a relational schema (Nabi *et al.*, 2011). Except this semantic web format based on RDF and SPARQL approach for ping data storage was proposed by (Souza, 2013). Although storing historical data in flat files seemed to be a good solution at the beginning of the project, after seventeen years of hourly data gathering, the manipulation of such an amount of big data becomes a very challenging task when one needs to perform fine-grained data analyses to explore the entire dataset. Therefore only more efficient computational approaches for data management are not enough, there is also need of more flexible structures for data analyses.

The above mentioned real-world scenario is quite close to the ones investigated by Data Warehousing technologies (Kimball *et al.*, 1998). It is advocated that similar solutions can be applied to the PingER project, providing a better understanding of analytical possibilities for its managers and users and providing a fair support for data aggregation and big data exploration.

1.6. Problem Statement

The operations of the internet has led to a significant importance. It is important to analyze the internet performance. Therefore a project is working to measure internet (end-to-end) performance named as PingER. It is led by SLAC since 1998. This project has been generated huge amount of data. This data can reveal interesting information about power cuts, network bottlenecks and packet loss, etc., which can help us in future forecasting and making effective business decisions. It is important to analyze this data to look at trends on internet connections, but it is not possible currently because loading such huge amount of data is not possible and users are unable to view and make an analysis of this data.

1.7. Research Questions

- i. How to develop a data warehouse on PingER data to analyze and process the PingER data?
- ii. How to provide a large scale and manageable storage architecture for storing PingER data?
- iii. How to visualize the complex finger information in a more understandable format by using effective visualization tools like Google API or D3js libraries?

1.8. Objectives

- i. To develop a data warehouse on PingER data to analyze and process the PingER data.
- ii. To provide a large scale and manageable storage architecture for storing PingER data.
- iii. To visualize the complex finger information in a more understandable format by using effective visualization tools like Google API or D3js libraries.

1.9. Proposed Solution

The proposed solution is to analyze the PingER data by loading it into Big Data platform (like Data Warehouse) and process it by using data mining techniques which supports large scale datasets (like Big Data processing techniques MR framework). And provide an effective storage architecture (like HDFS). To retrieve data from DWH execute interactive OLAP queries by using Impala and data is visualized by using Google APIs to simplify and pictorize the results.

1.10. Significance of the Study

PingER project is working to monitor Internet (End-to-End) performance of websites since 1998. The initial solution was to store PingER data by using flat files after this RDBMS was used to store data. But as PingER started monitoring more sites and expanding his working projects the data size grows to an extent which is difficult to store and manage. As PingER project is now monitoring more than 700 websites so it is necessary to have efficient and scalable storage architecture and fast processing framework. To address this problem this research provides the mechanism to store, process and visualize the PingER data.

1.11. Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 includes a literature review of 45 different research articles which represents the research work related to this research approach. Chapter 3 describes the proposed Research Framework, tools to be used and experimental setup of the system. Chapter 4 describes the results of the experiment and discussions. At the end summary of complete thesis is given.

Chapter 2

REVIEW OF LITERATURE

Adamu *et al.*, (2015) proposed an idea to increase the response of Impala queries by using indexing strategies. The idea of Big Data indexing is to fragment the datasets according to criteria that will be used frequently in query. Complex data are collected with metadata that describes their contents. Such datasets can be queried using the metadata of the contents. Instead of searching the whole database (which can be time consuming), a more efficient approach is to search the suitable groups relating to the query. To make query response quick it is necessary to choose appropriate indexing strategies. Indexing strategies are divided in two categories Artificial and Non-Artificial indexing strategies. Artificial indexing strategies include Latent Semantic Indexing and Hidden Markov Model. Non-Artificial indexing strategies includes Tree-Based indexing, Hash Indexing, Custom Indexing and Inverted Indexing.

Barbosa *et al.*, (2015) presented the overall idea of storing PingER data in HDFS and Process it by using MR and accessed by Impala queries. A two-step approach is used in this paper which requires the utilization of distributed systems. Each step uses an open-source Big Data system. In the first step, SciCumulus was used, a parallel Workflow Management System, to execute a Map-Reduce workflow to extract PingER legacy data and to transform them following a multidimensional data structure. The mapper activity maps the PingER data into star schema and reducer activity combines the data into a large CSV files. Secondly, the transformed data was loaded into HDFS and Impala queries are applied which is an analytical database, a big Data Warehouse (DWH) system that runs on top of the HDFS offering flexible and scalable support to multi-user analytical ad hoc data queries. At the end external tables was created in impala and data was loaded in impala.

Chopade *et al.*, (2015) analyzed substantial scale information sets and proposed a representation engineering X-SimVis. Existing information examination and representation environment needs utilitarian elements while working over wide range constant element frameworks. Better representation procedures will help for clear and successful comprehension in any ongoing element frameworks. The proposed X-SimViz system permits clients for intelligent continuous element information investigation and

representation. This model depends on Web Accessed Visualization Expert Systems. This model breaks into four segments Interface, Visualizer, Data and Display. The Interface gives the client access to the Visualizer. At an abnormal state, the interface can likewise incorporate a "clever operator" which can furnish extra help with the interface alongside client particular independent activities. The Visualizer gives the ability to change over the client's crude information into a visual structure. Information part is the crude information that the client needs to imagine. The Display is basically the segment which gives the visual correspondence between the client and the Interface and presentations the results of the Visualizer.

Dou *et al.*, (2015) proposed a mechanism of ontology based approaches used to mine web data that is in unstructured format. Ontologies are the domain knowledge which helps in data mining. Association Rule Mining, Classification, Clustering, Information Extraction, Recommendation Systems and Link Prediction are discussed with the effect of ontologies and their related problems and challenges. Ontologies in association rule mining provides the pruning constraints and abstraction constraints for users which permits a generalization concept of the ontology. In SDM one fundamental utilization of philosophy is to clarify the arrangement names with the arrangement of relations. Metaphysics based grouping incorporates utilizing ontologies as a part of bunching assignment for information pre-handling. Semantic data mining is the application of data mining used to mine data from the web. The survey shows that by using ontology knowledge the gaps between data, applications and algorithms can be fill out.

Kung (2015) discussed a measurable data reduction method PCA which uses an orthogonal change to move the first n directions of information set into another arrangement of n directions called primary segments. As a consequence of the change, the primary chief segment has the greatest possible contrast (that is, records for however a significant part of the variability in the data as could sensibly be normal) each succeeding part has the more astounding consumable fluctuation under the compulsory condition which is orthogonal to the first segments. The key parts are orthogonal on the grounds that they are the eigenvectors of the covariance network, which is symmetric. The motivation behind applying PCA to an information set is eventually to decrease its dimensionality, by finding another littler arrangement of m variables, for example, $m < n$, holding the vast majority of the information data.

Perrot *et al.*, (2015) discussed the accessibility problem of bigger and bigger datasets raises the need to make the customary systems (scatter plot, geographic heat maps) scale as quick as the information develops. Outlining point set representation techniques that fit into that new worldview is accordingly a basic test. To take care of this issue the center thought is to tally the quantity of focuses $c_i; j$ that are drawn at the same area $i; j$ on the screen. At that point one point $p_i; j$ is rendered for every extraordinary area $i; j$ and $c_i; j$ is utilized to decide the shade of $p_i; j$. The term heatmap is frequently used to portray that strategy. More intricate heatmap perceptions build a persistent capacity by applying a convolution on $c_i; j$. This consistent capacity is known as a density function.

Qi *et al.*, (2015) discussed the visualization techniques for different media information alongside the improvement of media data. Strategies for representation and miniaturized scale blog information perception is exhibited and procedure of geographic data information was acquainted which is capable with be more inside and out information examination and show the way of the law of the things by representation. perception covers the gathering of data, information preprocessing, learning representation, visual presentation and communication. Perception prepare chiefly incorporates framing chart and questioning spatial data. Framing geographic picture incorporates a two-dimensional picture and three-dimensional guide information and in addition information examination and assessment of visual articulation of plots, histograms and so on questioning of information incorporates snappy inquiry which can access to spatial substances and spatial data depicted in the geographic data framework. It chooses spatial substances to meet client prerequisites.

Purwar and Singh (2014) researched on issues in data mining like missing value imputation, outlier detection, feature selection and cluster analysis. The discussion focus on the directions to solve the issues by defining research gaps and every issue open a broad research area. Various algorithms MI (Mutual Information), RMSE (Root Mean Square Error) for missing value imputation and ROC (Receiver Operating Characteristics), AUC (Area Under Curve), Precision and confusion matrix for imbalanced data sets are also has been presented. MI measure is utilized to figure how one variable is identified with other variable. It is likewise used to discover the relationship amongst needy and autonomous variable in MV attributions. RMSE measure firstly registers the normal of the squared contrasts amongst genuine and watched estimations of a given element. RMSE is the square base of the mean squared mistake

ascertained. ROC is diagram between the affectability and specificity of a classifier that portrays the execution of a classifier. AUC is zone under ROC bend that decreases ROC execution to a basic expected execution.

Ben Ayed *et al.*, (2014) provide a survey on clustering methods to mine big data which focuses on to choose appropriate clustering method according to nature of data. Clustering strategies like surely understood k-Mean algorithm, Gaussian Mixture Modals and their variations, the established various leveled bunching techniques like the agglomerative calculation, the fluffy grouping techniques and Big Data bunching strategies. A correlation between techniques is likewise displayed by the creators. K-Mean algorithm is unsupervised grouping and $N*k*d$ difficult algorithm. K-Means algorithm utilizes Euclidean separation and its goal is to minimize separation inside the same group and amplifies separation between bunches by minimizing the goal capacity. The progressive strategies don't bunch information straightforwardly like parceling techniques, however utilize various leveled gathering or division to bit by bit collect/dismantle information focuses in clusters.

Liu *et al.*, (2014) presented an approach to reduce data size by identifying critical feature dimensions of data. Heuristic and Empirical method has been adopted to identify critical feature dimensions and results. This paper addresses the multifaceted nature of both issues in one general hypothetical setting and demonstrates that they have the same intricacy and are profoundly immovable. Next, an observational technique is connected trying to locate the surmised basic element measurement of datasets. Three extensive datasets are utilized as a part of the examination; each is isolated into 60% for preparing and 40% for testing. The model is retrained by changing the parameters to diminish the mistake rate. Six distinctive models are assembled and retrained to get the best precision. The model that accomplishes the best preparing exactness is utilized to locate the basic measurement.

Wu *et al.*, (2014) represented the characteristics of big data by a HACE (Heterogeneous Autonomous Complex Evolving relationships) theorem. Consistently, 2.5 quintillion bytes of information are made and 90 percent of the information on the planet today were created inside the previous two years. The most crucial test for Big Data applications is to investigate the huge volumes of information and concentrate helpful data or learning for future activities. Enormous Data handling system, which incorporates

three levels from back to front with considerations on information getting to and figuring (Tier I), information security and space learning (Tier II), and Big Data mining calculations (Tier III). The big data architecture has been presented in 3 tiers (Mining Platform, Semantic and Application Knowledge, Mining Algorithms).

Vijayakumari *et al.*, (2014) performed a comparison between GFS and HDFS architecture to measure performance on the basis of some common tasks. Google introduces GFS (Google File System) and Apache introduces HDFS (Hadoop Distributed File System). Both are data storage architectures and uses MR (MapReduce) as a processing framework, but HDFS is open source software. The architecture of GFS and HDFS has been also discussed. Comparison was performed on the basis of scalability, protection, security, processes and implementation. Google have their own particular record framework called GFS. With GFS, records are part up and put away in various pieces on numerous machines. While HDFS actualizes an authorization model for records and indexes that shares a great part of the POSIX model.

Kim *et al.*, (2014) proposed Latent Semantic Analysis approach to mine big data. This plan extricates converse and proportionality connections in huge information sets. PC reenactment uncovers that it fundamentally builds believability, support, preparing time, decrease rate of the tenets and reduction rate of the thing, contrasted with the current plans. The new calculation of diminishment affiliation tenets is exhibited which has inventively upgraded and additionally speed and lessening rate, likewise validity and backing. The essential objective is discovering taking in data from tremendous measures of data. By altering the proposed arrangement, quick and strong divulgence taking care of is available. Also, the proposed arrangement finds the associations, for instance, opposite and proportionality between a game plan of things, without encountering any additional system, so more reduction is open and thing release in the rule is moreover possible.

Gohil *et al.*, (2014) implemented MR on cloud based Big Data applications. Results demonstrate the execution of enormous information cloud applications Wordcount, Pi, Terasort and Grep. Wordcount is a basic system which tallies the quantity of events of every word in a given record. The yield is another record that comprises of lines, comprising of word alongside the check. The estimation of Pi is assessed utilizing a measurable (semi Monte Carlo) technique in Pi program. Terasort utilizes three arrangements of MapReduce projects TeraGen, TeraSort and TeraValidate. Grep gets an

information a general expression, checks a group of info records, finds any coordinating strings while numbering the events of every match lastly yields the outcome into a yield document. In trial these applications are connected to Amazon EC2 cloud and execution is measured by execution time and number of hubs.

Gupta and Siddiqui (2014) proposed the possibility of representation on Big Data. As ordinary information mining and factual examination are not material to enormous information so the new instruments were presented. Perception is a successful apparatus to get experiences into the huge information. Another undertaking was composed called Nanocube which fits in a tablet or PDA memory. Pseudo code to register and inquiry a Nanocube is additionally displayed in this paper. This anticipate is driven by AT&T. Nanocube are one of a kind it could be said that perception information does not require stores of nearby assets. It can be scanned and connected with information on a standard portable workstation or tablet. A blend of home developed AT&T programming innovation, distributed computing and uncovered APIs are what permits Nanocube to run easily in a web program. Attributes of Big Data and Hadoop stage are likewise examined in subtle element. Key advances to Big Data like HDFS and MR were displayed and at the Big Data investigation is performed.

Harter *et al.*, (2014) presented a multilayer investigation of the Facebook Messages stack which depends on HBase and HDFS. The information from HDFS follows was gathered and investigated to distinguish potential upgrades assessed by means of reproduction. Messages speak to another HDFS workload where HDFS was worked to store extensive records and gets for the most part consecutive I/O. examination results demonstrated that 90% of documents are littler than 15MB and I/O is very arbitrary. Results presumed that hot information is too substantial to effortlessly fit in RAM and chilly information is too expansive to effectively fit in glimmer however cost reenactments demonstrated that including a little blaze level can enhance execution more than proportionate spending on RAM or circles. HBase's layered configuration offered effortlessness, yet at the expense of execution reenactments demonstrate that system I/O can be divided if compaction sidesteps the replication layer. Clients of FM collaborate with a web layer, which is upheld by an application group, which thusly stores information in a different HBase bunch. The application bunch executes FM-particular rationale and stores HBase columns while HBase itself is in charge of persevering information. HDFS triply reproduces information keeping in mind the end goal to give

accessibility and endure disappointments.

Conejero *et al.*, (2014) also worked on deploying Apache Hadoop on multi user IaaS (Infrastructure as a Service) cloud, four deployment strategies (Horizontal, Vertical, Master-Apart, Complete-Spread) are used to deploy virtual cluster. The work introduced in this paper concentrates on the examination of execution, force utilization and asset use by Hadoop applications while conveying Hadoop on Virtual Clusters (VCs) inside a private IaaS Cloud. All the more definitely, the effect of various VM situation techniques on Hadoop-based application execution, power utilization and asset use is measured. The outcomes received from the Sentiment Analysis Tool demonstrates that the Horizontal technique has the most exceedingly bad effect on Hadoop execution, since there is an execution change when utilizing the other arrangement methodologies as a part of all cases. The Master-Apart VM situation technique has appeared to enhance the execution in 3.38% contrasted and the Horizontal procedure.

Bal *et al.*, (2014) conducted a review on the advantages of topographical groups and also advantages and difficulties of virtual bunches. The creators work a modern virtual bunch in a UK utilizing an electronic entrance and remark on its operation and impediments. A survey was arranged and appropriated to 31 firms in Pakistan to evaluate the obstructions they confront working with the UK group. Through investigation of the information from the poll, an arrangement of proposals to enhance business interchanges in virtual groups is recommended. A Geographical Cluster (GC) is a framework in which firms acting freely to augment their own particular benefits meet up in a restricted land space and are administered by an arrangement of guidelines and standards. Each land group has a society and an arrangement of establishments of its own, which associate with the operators of the bunch and help the bunch develop and adjust to changes. Advantages of GC are Better Access to Employees and Suppliers, Innovation focal points, Trust among on-screen characters and Access to particular data and learning.

Gu *et al.*, (2014) analyzed the employment and assignment execution component of MR structure to enhance execution in hadoop bunches. The examination demonstrates the two impediments of this component. At that point two enhancements were connected to MR employment and assignment execution component which covers the impediments and enhance the execution. Firstly, enhancement was connected to the setup and cleanup undertakings of a MapReduce occupation to diminish the time cost amid the introduction

and end phases of the employment. Besides, rather than receiving the free pulse based correspondence component to transmit all messages between the JobTracker and TaskTrackers, a texting correspondence instrument for quickening execution touchy errand booking and execution was presented. This enhanced design was called SHadoop. Trial results demonstrate that contrasted with the standard Hadoop, SHadoop can accomplish stable execution change by around 25% by and large for thorough benchmarks without losing adaptability and speedup. The upgraded work has breezed through a generation level test in Intel and has been incorporated into the Intel Distributed Hadoop (IDH).

Pal *et al.*, (2014) conducted an investigation on Hadoop by applying various records as contribution to the framework and examined the execution of the Hadoop framework. The execution is measured by running a MR wordcount application. WordCount is a straightforward MapReduce assignment. It checks the quantity of times a word is rehashed into a document of an information set. The Map strategy prepared one line at once. The information can be strings or a document. It parts the contribution to tokens, for detachment it consumes room as a tokenizer and utilizations the class String Tokenizer. The Reducer for these maps is executed by the Reduce strategy. It takes contribution from all the maps and like in union sort it joins the yield of all the guide and deliver yield. A Name hub needs to record the metadata for every document independently in its document framework. As the quantity of document build metadata about the records put away will be expansion. Number of documents was multiplied in every operation test began with 499 records and afterward multiplied this number every time till 7984 records. The outcomes demonstrated that utilization of the MapReduce and HDFS for putting away BigData (where information originates from heterogeneous sources) is proficient. Bytes were effectively composed and read by the MapReduce programs. It can be utilized for examining the yield of the documents, for example, the records created by the climate sensors which are dispersed in the earth, log documents of the diverse frameworks.

Ciubancan *et al.*, (2013) discussed the challenges of data mining in distributed environment like in GRID and shows that processing of data has become more complicated. Since GRID stages permits the use of assets in a dispersed and geologically circulated situations and a proficient use to characterize a model that can deal with the heterogeneity of the included assets like PCs, information, sensors or programming

instruments. The study of Data Mining under these factors becomes more complex. Different Data Mining Techniques are used to extract patterns from GRID. The motivation behind this work is spoken to by the investigation of the effect of GRID innovation for putting away and preparing a lot of data and learning. To see what the information "mining" process comprise of, the accompanying strides: development and approval of the model and utilization of the model to new information are talked about.

Fan and Bifet (2013) explained current market challenges and future scope of big data so the other researchers can clearly identify the problematic areas and work on it. Big Data mining and its applications were introduced and importance of open-source software tools is described. Big data mining is applied in Business, Technology, Health mining and as a future scope issues in compression, visualization, distributed mining, and hidden big data need to be resolved. The challenges of Big Data are Early warning, Real-time awareness and Real-time feedback. Four contributed articles were discussed including Scaling Big Data Mining Infrastructure, Mining Heterogeneous Information Networks, and Big Graph Mining: Algorithms and discoveries and Mining Large Streams of User Data.

Gu and Li (2013) performed a comparison between Spark and Hadoop to measure memory consumption and time execution on the basis of same type of tasks. The broad investigations for iterative operations to analyze the execution in both time and memory cost amongst Hadoop and Spark was conducted. To run the experiment framework uses big data applications Hadoop and MR. Hadoop gives an open source Java execution of MapReduce. It is made out of two layers an information stockpiling layer called Hadoop appropriated document framework (HDFS) and an information preparing layer called MapReduce Framework. HDFS is a piece organized document framework oversaw by a solitary expert hub. A preparing work in Hadoop is separated to the same number of Map errands as information squares and one or more Reduce undertakings. An iterative calculation can be communicated as various Hadoop MapReduce occupations. To lead near investigation five genuine chart datasets were chosen and five engineered diagram datasets are produced. The results of experiment shows that Spark is faster than Hadoop as it uses RDD (Resilient Distributed Data) which provide fast read operations.

Wang *et al.*, (2013) presented the outline and usage of G-Hadoop, a MapReduce structure that expects to empower vast scale disseminated processing over different groups. The MapReduce worldview has developed as a profoundly effective programming model for extensive scale information escalated registering applications. Notwithstanding, current MapReduce executions are produced to work on single group situations and can't be utilized for huge scale conveyed information handling over numerous bunches. Work process frameworks are utilized for conveyed information handling crosswise over server farms. The work process worldview has a few impediments for disseminated information handling, for example, unwavering quality and effectiveness. G-Hadoop gives a parallel handling environment to enormous information sets crosswise over circulated bunches with the broadly acknowledged MapReduce worldview. Contrasted and information escalated work process frameworks, it executes a fine-grained information preparing parallelism and accomplishes high throughput information handling execution. Besides, by copying delineate diminish errands G-Hadoop can give adaptation to non-critical failure to extensive scale enormous information preparing.

Bai *et al.*, (2013) implemented a vertical model taking into account a multi-level article situated framework engineering which is intended to bolster the CAV (Context Adaptive Visualization) system. BI frameworks need to incorporate perception subsystems or be utilized together with isolated representation frameworks that offer adaptable backing for making, controlling and changing representation arrangements. The relevant elements that influence the representation results are likewise examined. CAV model was actualized by utilizing ADO.NET element system and SQL Server. In any case, a comparable framework could be produced by using proportional advancements from different merchants, for example, Oracle and IBM.

Liao *et al.*, (2012) presented a survey on data mining techniques and applications. Analysis on yearly basis shows that the usage of data mining applications has been increased. This paper studies and orders DMT, utilizing nine classes: Neural systems, Algorithm engineering, dynamic forecast based, Analysis of frameworks design, Intelligence specialist frameworks, Modeling, learning based frameworks, System advancement and Information frameworks, together with their applications in various examination and down to earth areas and multiple algorithms are also discussed with respect to each application.

Fu *et al.*, (2013) presented information handling technique for visual showing on huge information. It gives an information lessening strategy to bolster visual rundowns for huge information, and check the outline space is vast plot of multidimensional information. Visualization design for Binned Plots was introduced these are common binning schemes database data types: number, geography, sequence and time. A limitation is absence of the above four data binning methods is composite and the absence of relative information between various areas. The idea map information piece is connected to the framework, for example, Google Maps and Hot guide. Notwithstanding, the information square is diverse in two essential perspectives. To begin with, they give dynamic information representation, as opposed to pre-rendered pictures. Second, they contain multidimensional information inquiry to choose tiles, which can powerfully figure out the volume of projection information questioning and rendering.

Hlosta *et al.*, (2013) presented way to deal with perception of advancing affiliation principle models. This methodology depends on diagram perception where hubs of the chart speak to thing sets, and edges speak to affiliation rules. The work clarified how information created by information mining calculations can be stored and how information can be sifted and pictured. Two methods for chart based models are utilized force based layout algorithms and local and global layout algorithms. Three regular strategies for representation are Rule Table, Two Dimensional Matrix and Directed Graphs. A substantial info stream is isolated into mining windows in view of non-covering time interims. The model comprises of three parts. The Mining Algorithm for mining affiliation principles was connected to every Mining Window for the information mining errand. The consequences of the Mining Algorithm are spared into the capacity part called Knowledge Base. At last, affiliation standard models are reestablished from the Knowledge Base, and they are displayed to an examiner in a segment called the Presenter in a type of Evolving Models. The Presenter gives stacking of the predefined models utilizing the client indicated limitations containing a recognizable proof of the undertaking, time interim settings and separating in light of things in affiliation rules.

Bu *et al.*, (2013) presented a job scheduling methodology to alleviate impedance and in the interim saving assignment information territory for MapReduce applications. The technique incorporates an obstruction mind full booking arrangement, in light of an assignment execution expectation model, and a versatile postponement planning calculation for information area change. The impedance and region mindful (ILA)

planning technique in a virtual MapReduce system has been executed. After this viability and proficiency on a 72-hub Xen-based virtual Cluster has been assessed. In a virtual MapReduce bunch, the obstruction between virtual machines (VMs) causes execution corruption of guide and diminish assignments and renders existing information territory mindful undertaking booking strategy, similar to postpone planning. The proposed approach gives a viable answer for issue. Trial results demonstrated that with 10 delegate CPU and IO-serious applications can accomplish a speedup of 1.5 to 6.5 times for individual employments and yield a change of up to 1.9 times in framework throughput in correlation with four other MapReduce schedulers.

Chang and Sun (2013) proposed real time intelligent representation framework which incorporates a constant breaking down and perception of information. A client is permitted to interface with the framework by utilizing Kinect camera and a cell phone and toward the end intra/bury clients online networking information is pictured. The information was gathered from the cloud and the online networking are sent into the server for ongoing examination and representation from the server and presentation to a client. What's more, the continuous rendering for the 3D visual impact is likewise actualized at the server. A Kinect camera is used to track a client in the proposed framework. The body, skeletons, and joints of a client in a 3D space can be followed utilizing Kinect programming improvement pack (SDK) continuously. A client is permitted to login our framework to uncover the character of a client. In the framework, to meld the signs from the sensor of the quickening agent on a cell phone and the trigged sign of the hand joint from the Kinect camera for individuals distinguishing proof in the framework was proposed. For intra-client online networking information representation Facebook API was utilized to creep the information and for between clients Facebook API gives diagram of interpersonal organization as per NodeXL.

Zhang and Huang (2013) presented a 5w's model to visualize big data. 5w's represent five dimensions of Big Data these are What is the data content, Why the data occurred, Where the data came from, When the data arrive and Who receive the data. This structure groups BigData characteristics and examples, as well as sets up thickness designs that give more diagnostic elements. Visual grouping strategies was utilized to send and get densities which depict Big Data Patterns. The model was tested by utilizing the system security ISCX2012 dataset. The test results demonstrates this new model with grouped perception can be proficiently utilized for BigData investigation and

representation. Also, this model gives a conveyed framework that stores information on the PC hubs, with high total data transfer capacity over the group. SAS Visual Analytics were utilized to imagine and dissect the enormous dataset to discover visual examples. Heat maps and tree maps were utilized as a part of the representation.

Sachin and Vijay (2012) defined a survey on educational data. Data mining is also applicable in education it is important to mine educational data to recognize interesting learning patterns and their impact on students. The author explains the data mining process and DMT used in education mining. Prediction is used to develop a model which can infer a single aspect of data from some combination of other data. This approach is used to analyze the student performance. Clustering is a process of grouping objects into similar classes. This approach is used to investigate differences and similarities between students. Relationship mining is used for finding student's mistakes, learner behavior patterns, diagnose student's learning problems and offer student advice. Other EDM methods are outlier detections, Text Mining and Social Network Analysis (SNA) is also presented in this paper.

Ngo *et al.*, (2012) presented a Unified Data Framework (UDF) that permits the collection of appeal information sources into a solitary valuable examination asset that is identified with exploration in advanced education. The UDF bolsters the conglomeration of existing and new information sets and gives the alternative of associating and overhauling information from the first sources. The UDF give investigative apparatuses to information mining and perception of joined and complex information hotspots for analysts of advanced education. Brought together Data Framework includes 5 layers these are presentation, Data Access, Metadata, Data and Ingest layers. Presentation layer gives a coordinated perspective of information to end-clients. Information access layer shroud the many-sided quality of lower layers from clients and permit to get to the information from presentation layer. Metadata and Data layers combine various information sources which are helpful in examination. Ingest layer join the ETL procedure used to extricate, change and load information from numerous information sources.

Cheng *et al.*, (2012) designed and implemented ERMS a flexible copy administration framework for HDFS. ERMS presents a dynamic/standby stockpiling model, exploits an elite complex occasion preparing (CEP) motor to recognize the constant information sorts and gets a versatile replication arrangement for the distinctive

sorts of information. ERMS utilizes Condor to expand the replication number for hot information in standby hubs and to evacuate the additional imitations after the information chilling off. Taking into account information access designs the information in HDFS is arranged into hot, frosty and ordinary information. ERMS was executed in a private group with one Namenode and fifteen Datanodes ten dynamic hubs and five standby hubs of ware PC. ERMS is additionally actualized on Hadoop-20 which is Facebook's ongoing circulated Hadoop. The occupation are keep running from the SWIM, which gives one mouth work follow and replay scripts of a Facebook 3000-machine creation bunch follow. Information region and normal perusing throughput of these occupations were assessed under various edges ($\tau_{M1} > \tau_{M2} > \tau_{M3}$). Information territory and perusing throughput are two basic measurements for execution of HDFS.

Zaharia *et al.*, (2012) researched on another framework known as Apache Spark. An experiment is conducted and results proved that spark can perform 40X faster in read operations. The best way to share information between parallel operations in MapReduce is to compose it as a circulated filesystem which adds sub-stancial overhead because of information replication and circle I/O. Sparkle conquers this issue by giving another capacity primitive called strong conveyed datasets (RDDs). RDDs let clients store information in memory crosswise over que-ries, and give adaptation to internal failure without requiring replication, by following how to recompute lost information beginning from base information on plate. This gives RDDs a chance to be perused and reviewed to 40× quicker than ordinary appropriated filesystems, which makes an interpretation of specifically into speedier applications.

Janciak *et al.*, (2011) proposed a visualization framework which is implemented in the European project ADMIRE (Advanced Data Mining and Integration Research for Europe). The framework is consisting of two components Gateway and Benchmark. Gateway receives requests for process enactment and conducts a sequence of steps arranging that enactment or rejecting it. These requests are encoded in the DISPEL language, which is processed by a parser to generate data-flow graphs in the form of executable OGSA-DAI workflows. The execution of the workflow is fully controlled by the user and monitored by a specialized monitoring Web based application integrated to the ADMIRE Portal. Benchmark is a graphical tool called Model Visualization Application, which can process data mining results produced as outputs of the processing elements. The tool is implemented as a part of the ADMIRE Workbench which represents

the client side of the distributed environment.

Bresnahan *et al.*, (2011) presented Cumulus a capacity cloud framework that adjusts existing capacity usage to give proficient transfer/download interfaces good with S3, the true business standard. While this similarity empowers clients to effortlessly move amongst scholastic and business mists. The most imperative component of Cumulus is its all around verbalized back-end extensibility module. Cumulus engineering and execution was characterized. Cumulus gives two capacities. To begin with, it permits clients to amass and oversee information, transfer information to the cloud, screen its status and download it from the capacity cloud as required. Second, this information can be accessible specifically VM pictures. Cumulus likewise gives a picture store to Nimbus process mists. Cumulus engineering comprises of 6 layers. These are Cumulus interfaces, redirection, APIs, administration usage, storage API and execution.

Bronson *et al.*, (2011) measured the performance of MapReduce for large scale datasets to apply visualization techniques. Visualization algorithms are applied on large scale datasets are Rendering, Isosurface Extraction and Mesh Simplification. Rendering algorithm exploits the inborn properties of the Hadoop structure and permits the reterization of cross sections made out of gigabytes in size and pictures with billions of pixels. MapReduce based calculation for isosurface extraction depends on the Marching Cubes calculation which is the true standard for isosurface extraction because of its productivity and strength. Network disentanglement utilizes two MR employments to execute the calculation since it requires two sorting stages. The main Map stage containers every vertex into a general lattice to guarantee that all triangles contributing vertices to a specific canister touch base on the same hub in the Reduce stage. The primary Reduce stage gets the same key-esteem pair from the Map stage, yet sorted and gathered by key. It peruses every novel key and uses the quadric measures of all triangles falling into that receptacle to figure the delegate vertex.

Tsumoto *et al.*, (2011) proposed a fleeting information mining process which comprises of choice tree, grouping, MDS and three-dimensional directions mining. The outcomes demonstrate that the reuse of put away information will give an intense instrument to describe transient patterns of clinical activities in divisions of a healing center. OLAP questions are connected to put away and reused information. A system was proposed on advancement of clinic administrations which depends on information

mining. The imperative components like how healing facility data framework functions, information readiness and mining handle, the consequences of representation and grouping based investigation of likenesses between divisions are clarified in point of interest. Information arrangement process includes on planning of DWH. To start with, the information was ordered from heterogeneous information sets with a given center as the principal DWH. At that point DWH was part into two auxiliary DWHs: substance and histories. The outcomes demonstrated that patient records and Nursing considerations are the significant piece of requests.

Gaur (2011) proposed the advancement procedure of LIRFSS. LIRFSS (Legal Information Retrieval and Focused Semantic Search) is a framework that performs IE(Information Extraction) and IR(Information Retrieval) on Indian Supreme Court choices, particularly on criminal cases identified with homicide and give centered semantic list items, offering a high capability of help for judges, legal counselors and residents in accessing law. As examining the limitless volumes of information turns out to be progressively troublesome, so data representation is likewise done. The upside of visual information investigation is that it is instinctive and requires no comprehension of complex numerical or factual calculations or parameters. Therefore, visual information investigation as a rule permits a speedier information investigation and regularly gives better results. IE is done on the archive corpus containing criminal homicide case judgments and data, for example, date, area of event and IPC (Indian Penal code) areas were controlled by making explanations on the reports.

Chen (2011) discussed perception strategies to apply on information mining aftereffects of hoodlums record information. Information Mining comes about typically appeared to us with complex information relationship, which is outside our ability to understand. The Visualization of Data Mining result demonstrates the intricate information principles and connections in the instinctive diagram communicating way. It can help information examiner to have a superior comprehend with the Mining comes about and to frame a right judgment. The strategies for Visualization of Data Mining Results is not particular, it for the most part have a great deal more association with the calculation embraced in the Data Mining. For instance, The Decision Tree calculation embraces tree chart, bunching calculation is received conglobation graph. To make wrongdoings hazard expectation this model uses choice tree to break down the extensive records. The model has three principle parts information pretreatment, property

investigation and representation for order model of choice tree. Basic information pretreatment techniques are cleaning, joining, and change. After pretreatment the objective traits are broke down. In this progression insignificant qualities are decreased and system of credits are shown to more profound comprehend the inward hub. In last stride for perception tree diagram for choice tree and pie outline to show inside hub was utilized.

Thusoo *et al.*, (2010) discussed the challenges and capabilities of Facebook platform. Data Flow and Storage architecture, Data discovery, Data operations and resource sharing of Facebook has been explained. There are two wellsprings of information combined MySQL level that contains all the Facebook webpage related information and the web level that creates all the log information. The information from the web servers is pushed to an arrangement of Scribe-Hadoop groups. These groups include Scribe servers running on Hadoop bunches. The Scribe servers totals the logs originating from various web servers and think of them out as HDFS records in the related Hadoop bunch. With vast approaching information rates and the way that more verifiable information should be held in the group so as to backing chronicled examination, space utilization is a steady limitation for the impromptu Hive Hadoop bunch. The generation bunch for the most part needs to hold one and only month of information as the occasional creation occupations from time to time take a gander at information past that time span. At Facebook questioning and investigation of information is done transcendently through Hive. The information sets are distributed in Hive as tables with every day or hourly segments.

Thusoo *et al.*, (2010) explained data models, query language and system type of Hive. Hive is a tool used for applying OLAP queries in data warehouses. Currently Hive is used by Facebook in Hadoop Big Data environment. Hive stores information in tables where every table comprises of various lines, and every line comprises of a predetermined number of segments. Every section has a related sort. The sort is either a primitive sort or a mind boggling sort. The Hive inquiry dialect HiveQL contains a subset of SQL and a few augmentations that has been discovered valuable in our surroundings. Hive at present does not support embedding into a current table or information parcel and all additions overwrite the current information. Hive framework design comprise of 7 parts. These are Metastore, Driver, Query Compiler, Execution Engine, Hive Server, Client Components (like CLI or GUI) and Extensibility Interfaces.

Zaharia *et al.*, (2010) focused on the applications that can be reused as working data set of nodes to perform parallel operations. A new framework is proposed called Spark which supports such applications while retaining the scalability and fault tolerance of MapReduce. To achieve these goals Spark uses resilient distributed datasets (RDD). Spark can outperform Hadoop by 10x in iterative machine learning jobs, and can be used to interactively query a 39 GB dataset with sub-second response time. To utilize Spark, engineers compose a driver program that executes the abnormal state control stream of their application and dispatches different operations in parallel. Sparkle gives two primary reflections to parallel programming RDD and parallel operations on these datasets. The parallel operation that can be run on RDD are reduce, collect and for each.

Xie *et al.*, (2010) proposed the solution for the issue of how to place information crosswise over hubs in a way that every hub has an adjusted information handling load. The current Hadoop usage accept that registering hubs in a bunch are homogeneous in nature. Given information escalated application running on a Hadoop MapReduce group, proposed information situation plot adaptively adjusts the measure of information put away in every hub to accomplish enhanced information handling execution. Trial results demonstrated our information situation methodology can simply enhance the MapReduce execution by rebalancing information crosswise over hubs before performing an information escalated application in a heterogeneous Hadoop group. The MapReduce structure can rearrange the many-sided quality of running circulated information preparing capacities over different hubs in a bunch, in light of the fact that MapReduce permits a software engineer with no particular learning of conveyed programming to make MapReduce capacities running in parallel over numerous hubs in the group. MapReduce naturally handles the social occasion of results over the different hubs and return a solitary result or set. All the more vitally, the MapReduce stage can offer adaptation to non-critical failure that is completely straightforward to software engineers.

Leverich and Kozyrakis (2010) presented alteration of Hadoop to permit downsize of operational groups. The outcomes demonstrated that running Hadoop bunches in fragmentary setups can spare somewhere around 9% and half of vitality utilization, and that there is a tradeoff between execution vitality utilization. Further an examination is likewise led on the vitality productivity of these structures. The vitality effectiveness of a group can be enhanced in two routes: by coordinating the quantity of dynamic hubs to the present needs of the workload, setting the remaining hubs in low-control standby modes.

Enhancing Hadoop vitality effectiveness involves three phases: Data Layout Overview, Replication Invariant of vitality and Evaluation. Changes to Hadoop are assessed through tests on a 36-hub group combined with a vitality model construct straightly in light of CPU usage.

Qiang *et al.*, (2010) applied visualization techniques in spatial data mining and discuss the current research directions and methods of visualization. The methods of visualization are Filter, Mapping and Draw. Filter extricates the information of specialist's enthusiasm from the exceptionally unique reproduction test information and through the crude information preparing. The outcomes got by mapping the filtering values into the geometric primitives that can be plotted. Draw step makes an interpretation of geometric primitives into the picture. The progressions of perception procedure are Data operations and Data representation. Information mining and representation are two unique spaces, however perception can be connected to present data from huge datasets.

Pavlo *et al.*, (2009) gave the idea of using parallel DBMS instead MR as a processing framework. Various tasks like data loading, data aggregation, node configuration, and execution and data operations are executed in both systems. All the tasks are carried out on three systems Hadoop, DBMS-X and Vertica. The result shows that the parallel DBMS is better than MR but only for 100 node of cluster. As the number of nodes will be increase the performance of parallel DBMS will decrease. Despite the fact that the procedure to load information into and tune the execution of parallel DBMSs took any longer than the MR framework however the watched execution of these DBMSs was strikingly better on 100 nodes.

Golab and Ozsu (2003) discussed some issues of continuous information streams and proposed another design for nonstop information stream frameworks. For quite a while databases are utilized as a part of utilizations where diligent information stockpiling and complex inquiry execution is required. Be that as it may, when the situations like where constant information stream arrives as in system movement investigation, sensor systems, budgetary tickers and value-based log examination. Another design is required for this sort of frameworks. Three questioning ideal models for gushing information have been proposed: connection based, object-based, and procedural. The proposed connection based dialects are CQL, StreaQuel and AQuery each of which has SQL-like grammar and upgraded support for windows and requesting. In based displaying COUGAR sensor

database. Every sort of sensor is displayed by an ADT. In the procedural dialect Aurora framework was utilized as a part of which clients build question arranges through a graphical interface by masterminding boxes (relating to inquiry administrators) and going along with them with guided circular segments to determine information stream.

Keim (2002) proposed a classification of information visualization and visual information mining systems which depends on the information sort to be imagined, the perception procedure and the collaboration and mutilation strategy. The fundamental thought of visual information investigation is to exhibit the information in some visual structure which permits the clients to get understanding into the information, make determinations and straightforwardly communicate with the information. The information sorts that were envisioned are 1Dimensional information, 2Dimensional information, Multi-Dimensional information, Text and hypertext and Hierarchies and charts. The representation strategies are characterized into Standard 2D/3D shows, geometrically changed presentations, Icon-based showcases, Dense pixel shows and Stacked showcases. The third measurement of the grouping is the cooperation and twisting procedure utilized. Communication and twisting procedures permit clients to specifically interface with the representations. Collaboration and mutilation systems are further characterize into various classes.

Chapter 3

MATERIALS AND METHODS

This research is purely based on experimental setup. An experiment is conducted to apply visualization on PingER data (Qiang *et al.*, 2010). A data warehouse is created by using big data platform Hadoop. PingER data was loaded in to HDFS and processed by using MR framework. The data is collected from PingER files that are stored on PingER servers and transformed into star schema to create DWH. After that, data is accessed by applying OLAP queries by using Impala (Barbosa *et al.*, 2015). To visualize this processed data or information data visualization techniques are used which makes this data visible to the general public.

In Section 1.5, the PingER project was introduced, explaining which Internet performance data are collected, and how these data are stored. It was also has been discussed that storing a large amount of data in multiple files is not efficient for big data analyses. In this section, the proposed solution that facilitates fine-grained data exploration, presenting the designed dimensional data model and the ETL MapReduce dataflow, highlighting the loading process into Impala is depicted. The architecture of PingER dataflow is given in Figure 3.1.

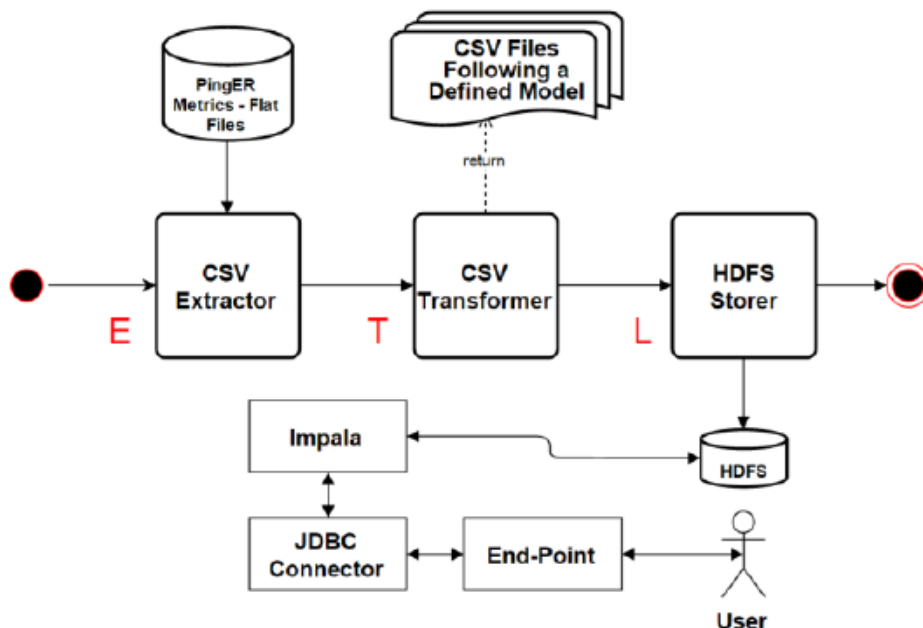


Figure 3.14: PingER Dataflow Architecture

3.1. Research Framework

The flow of this research work has been explained in Figure 3.2. The output from one step will be used as input for the second step.

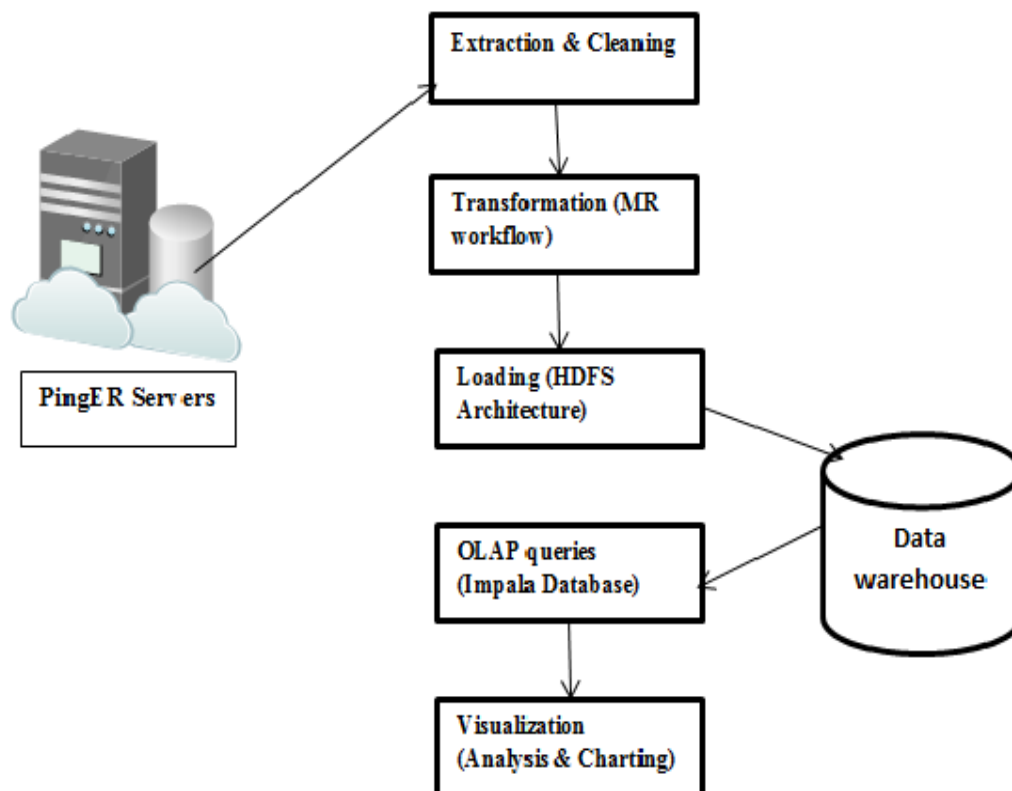


Figure 3.15: Research Framework

3.1.1. Selection and Cleaning:

In this step the data from PingER servers was downloaded. As explained above in introduction the PingER data save RTT values in text flat files. PingER uses small and large data packets of 100 bytes and 1000 bytes. For this research the data from 1994 to 2014 was collected by PingER Measurement Agents. The dataset comprises of 100,000 files summing over than 60GB. Further PingER performs data analysis of RTT values by applying 16 different metrics like jitter, MOS, unpredictability, unreachability etc. Different information is derived from RTT values are stored as a separate file for all years. These files called aggregated small files and contain average values of all days for a year. The aggregated data files used in research are Average RTT, Minimum Average RTT, Maximum Average RTT, Alpha, MOS, Conditional Loss Probability, Duplicate Packets, IPDV, IQR, Minimum Packet Loss, Out-of-Order Packets, Packet Loss, Throughput, Unpredictability, Unreachability and Zero Packet Loss Frequency.

All these files are collected for 100 and 1000 bytes data size. The total size of these aggregated data files is 20GB. Because this data is already in simplified form, transformed from raw and rough files to fine granularity aggregated files so there is no need to perform cleaning process on this data.

3.1.2. Transformation Process:

This process is divided into two steps. In first step the dimensional model of data was designed to transform data in multidimensional model of DW. In second step PingER data was transformed into clean CSV Files. By using parallel workflow management system a MapReduce program was executed. Map activity maps the data values according to defined star schema and reduce activity combines the data of each year from a separate flat file into a large CSV file. As a result there are 17 large CSV files to load into HDFS.

Dimension Modeling

For dimension modeling star blueprint was utilized. In data warehousing and business Intelligence (BI), a star outline is the minimum complex kind of a dimensional model, in which data is sorted out into convictions and estimations. An actuality is an event that is tallied or measured, for instance, an arrangement or login. A measurement holds reference information about the sureness, for instance, date, thing, or customer. A star blueprint is charted by enveloping each truth with its related measurements. This consequent chart results in star formed.

Star schemas are enhanced for questioning huge information sets and are utilized as a part of information distribution centers and information stores to bolster OLAP 3D shapes, business knowledge and systematic applications, and impromptu inquiries. Inside the information stockroom or information store, a measurement table is connected with an actuality table by utilizing a remote key relationship. The measurement table has a solitary essential key that interestingly distinguishes every part record. The reality table contains the essential key of each related measurement table as an outside key. Consolidated, these outside keys structure a multi-part composite essential key that extraordinarily recognizes every part record in the actuality table. The actuality table additionally contains one or more numeric measures. Star schema designed for PingER is given in Figure 3.3.

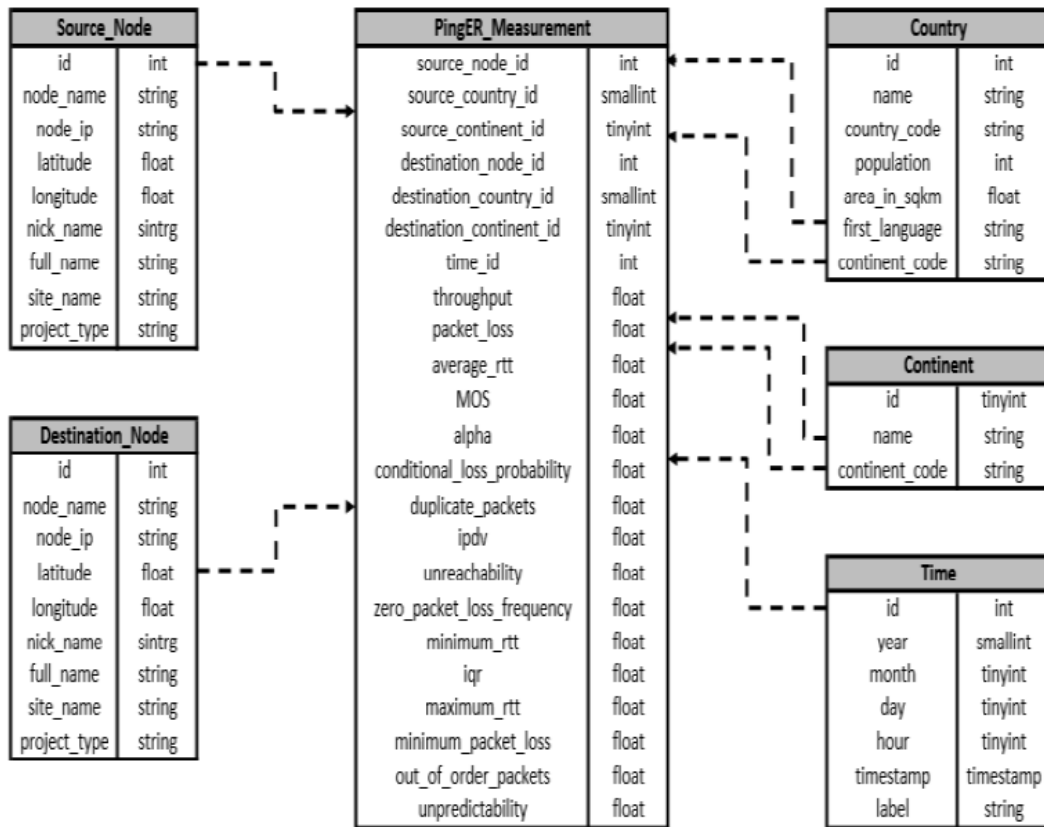


Figure 3.16: Star Schema of PingER Data Warehouse

To define a dimensional data model, network measurements and their perspectives of analysis were identified. As previously explained a ping measurement is defined by a combination of parameters (source, destination, timestamp) and associated with specific network metrics. This combination defines one measurement occurrence or, as it is called DW fact table. Additionally, each parameter corresponds to a DW dimension table in a dimensional data model. Besides, as each source or destination node is physically located in a country and continent these are called DW dimensions, to enable aggregations by regions (countries or continents) directly relating them to the PingER measurement fact table.

MR Workflow System

Map-Reduce was acquainted by Google all together with procedure and store vast datasets on ware equipment. It gives a programming worldview which permits useable and sensible circulation of numerous computationally escalated undertakings. Accordingly, numerous programming dialects now have Map-Reduce executions which augment its uptake. Then again, Hadoop is a very famous Map-Reduce usage by the Apache Foundation. Its establishments depend on standards of parallel and disseminated

preparing with no database reliance. The adaptability of MapReduce lies in the capacity to handle dispersed calculations on a lot of information on bunches of product servers, with basic errand based models for administration (Krishnan, 2013).

The MapReduce structure comprises of a library of various interfaces. The significant interfaces utilized by engineers are Mapper, Reducer, JobConf, JobClient, Partitioner, OutputCollector, Reporter, InputFormat, OutputFormat, and OutputCommitter. The basic architecture of MR is given in Figure 3.4.

The Map function composed by the client will get an info pair of keys and values, and after the calculation cycles, will create an arrangement of halfway key-esteem sets. Library works then are utilized to amass together all middle qualities connected with a moderate key 1 and passes them to the Reduce capacity.

The Reduce function composed by the client will acknowledge a middle key 1, and the arrangement of qualities for the key. It will consolidate together these qualities to frame a perhaps littler arrangement of qualities. Reducer yields are only zero or one yield esteem for every conjuring. The middle of the road qualities are supplied to the Reduce capacity by means of an iterator.

In this system implementation Map function reads PingER legacy data files and map the data according to designed dimensional model given in Figure 3.3. Each mapper invocation reads a text file containing measurement data for a given matrix and day of the year and transform this file into a resulting file that contains the same information but follows a dimensional structure called CSV file.

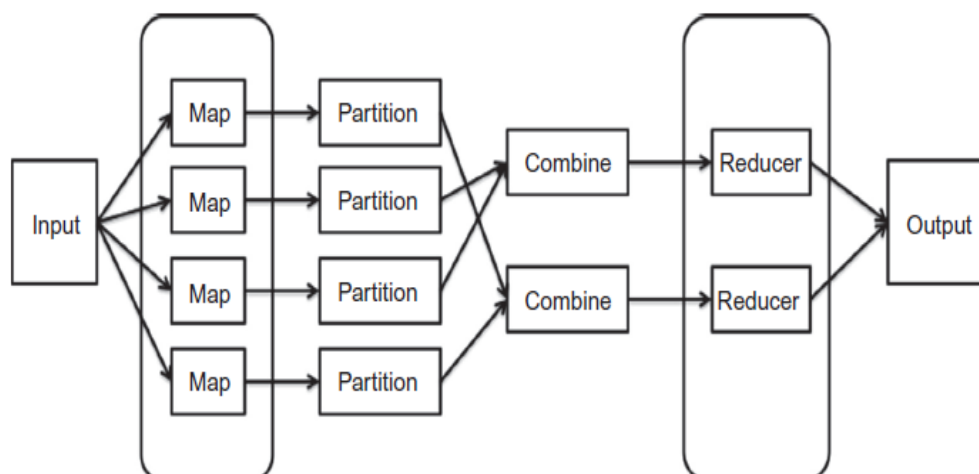


Figure 3.17: MR Architecture (Karishnan, 2013)

In this system implementation Reduce function simply combines all files for a given year into a single yearly big file. Each yearly file size corresponds to the sum of all file sizes for a particular year. Finally these files can be easily loaded into HDFS. Before loading data into HDFS it is necessary to define a Hadoop Cluster.

3.1.3. Defining Hadoop Cluster

A Hadoop cluster is an uncommon sort of computational group planned particularly to store and dissecting gigantic measures of unstructured information in an appropriated registering environment. Such clusters run Hadoop's open source circulated preparing programming on ease product PCs. Commonly one machine in the group is assigned as the NameNode and another machine the as JobTracker; these are the experts. Whatever remains of the machines in the bunch go about as both DataNode and TaskTracker these are the slaves. Hadoop clusters are frequently alluded to as "shared nothing" frameworks on the grounds that the main thing that is shared between hubs is the system that associates them (Hedlund, 2011).

Hadoop clusters are known for boosting the velocity of information examination applications. They likewise are exceptionally adaptable. On the off chance that a group's preparing force is over whelmed by developing volumes of information, extra bunch hubs can be added to expand throughput. Hadoop cluster additionally are exceedingly impervious to disappointment in light of the fact that every bit of information is duplicated onto other bunch hubs, which guarantees that the information is not lost on the off chance that one hub falls flat. Facebook was perceived as having the biggest Hadoop group on the planet. Other example illustrations are Google, Yahoo and IBM.

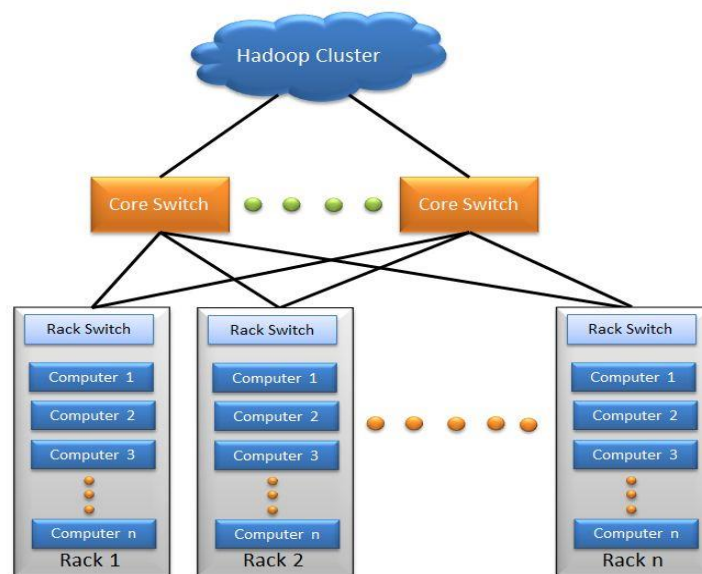


Figure 3.18: Hadoop Cluster Architecture (Hedlund, 2011)

Hadoop Cluster Architecture

The Hadoop cluster architecture is explained in Figure 3.5. The Hadoop Cluster can comprise of 110 distinctive racks every rack can have around 40 slave machines. At the highest point of every rack there is a rack switch every slave machine (rack server in a rack) has links turning out it from both the closures. Links are associated with rack switch at the top which implies that top rack switch will have around 80 ports. There are worldwide 8 center switches.

The rack switch has uplinks associated with center switches and thus interfacing all different racks with uniform data transmission framing the bunch. In the group you have few machines to go about as Name hub and as JobTracker. They are alluded as Master machines. These expert machines can have diverse setup supporting more DRAM and CPU and less neighborhood stockpiling. Most of the machines goes about as DataNode and Task Trackers and is alluded as Slaves. These slave hubs have heaps of neighborhood circle stockpiling and direct measures of CPU and DRAM.

Cluster Components

Hadoop cluster has 3 components Client, Master and Slave. The role of each component is shown in the Figure 3.6. Clients play the role of loading the data into clusters and submit MapReduce jobs describing how the data should be processed and then retrieve the data to see the response after job completion.

The Masters comprises of 3 parts NameNode, Secondary NameNode and JobTracker. NameNode does not store the records it stores just the metadata documents. NameNode administers the health of DataNode and directions access to the information put away in DataNode. NameNode monitors all the document framework related data, for example, Which segment of record is spared in which part of the bunch, Last get to time for the records and client authorizations like which client have entry to the document. JobTracker arranges the parallel handling of information utilizing MapReduce. The job of Masters is given in Figure.3.7.

Hadoop Server Roles

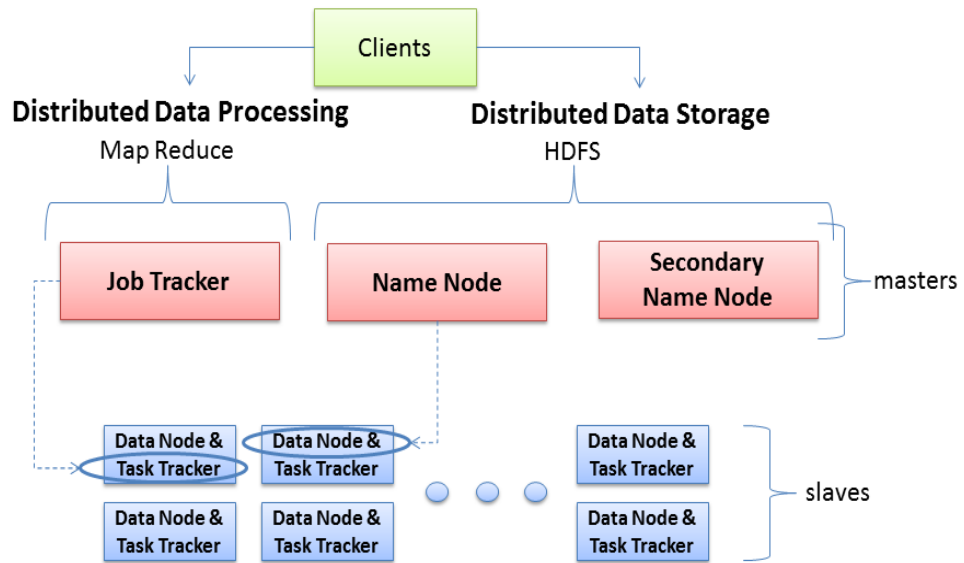


Figure 3.19: Hadoop Cluster Components (Hedlund, 2011)

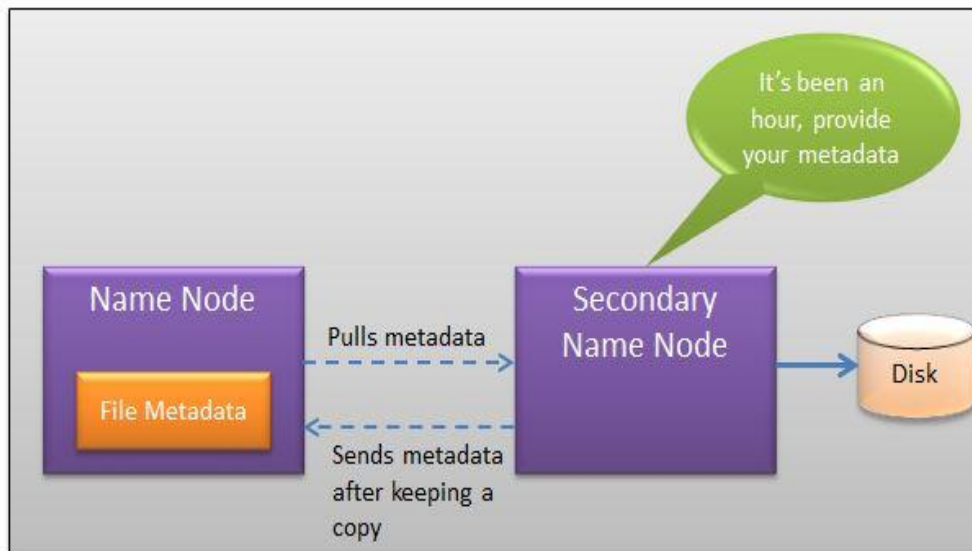


Figure 3.20: Internal working of 2nd Component Masters (Hedlund, 2011)

The job of Secondary Node is to contact NameNode in an occasional way after certain time interim (of course 60 minutes). NameNode which keeps all filesystem metadata in RAM has no ability to process that metadata on to plate. So if NameNode crashes, you lose everything in RAM itself and you don't have any reinforcement of document framework. What auxiliary hub does is it contacts NameNode in a hour and hauls duplicate of metadata data out of NameNode. It rearrange and combine this data

into clean document envelope and sent to back again to NameNode, while keeping a duplicate for itself.

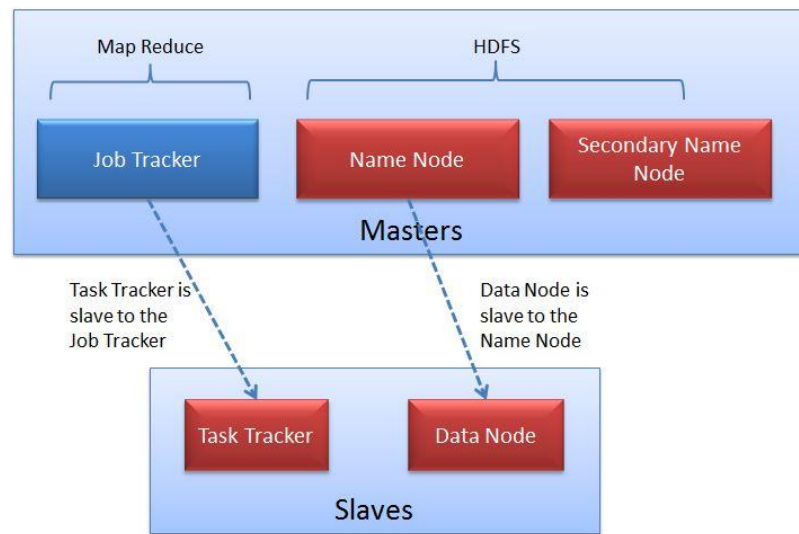


Figure 3.21: Internal working of 3rd component Slaves (Hedlund, 2011)

Figure 3.8 explains the working of slave nodes. Slave nodes is the greater part of machine in Hadoop Cluster and are capable to Store the information Process the calculation. Every slave runs both a DataNode and Task Tracker daemon which imparts to their lords. The Task Tracker daemon is a slave to the JobTracker and the DataNode daemon a slave to the NameNode.

3.1.4. Loading process:

In this process Hadoop HDFS storage architecture was used. CSV files are loaded on HDFS. Data directories for loading CSV files were created in HDFS for dimensional model. One directory was created for fact table and other for dimension tables and files were loaded.

When the term large scale data analysis or Big Data analysis occurs the keywords like MR, GFS and HDFS are used to explain key Big Data technologies. Google File System is an exclusive disseminated record framework created by Google and uniquely intended to give effective, solid access to information utilizing substantial bunches of product servers. Documents are isolated into pieces of 64 megabytes, and are normally annexed to or read and just to a great degree once in a while overwritten or contracted. Contrasted and customary record frameworks, GFS is planned and advanced to keep

running on server farms to give amazingly high information throughputs, low idleness and survive singular server disappointments.

Then again the open source Hadoop Distributed File System (HDFS) stores expansive documents over different machines. It accomplishes unwavering quality by duplicating the information over various servers. Thus to GFS, information is put away on various geo-differing hubs. The record framework is worked from a bunch of information hubs, each of which serves squares of information over the system utilizing a piece convention particular to HDFS. Keeping in mind the end goal to play out the specific operations in GFS and HDFS a programming model is required which is MR (Vijayakumari *et al.*, 2014). Table 3.1. explains comparison between GFS and HDFS.

Table 3.3: Comparison of GFS and HDFS

	GFS	HDFS
File Structure	Divided into 64MB chunks Chunks replicated	Divided into 128MB block Namenode holds block replica
Communication	TCP connections are used for communication. Pipelining is used for data transfer over TCP connections.	RPC based protocol on top of TCP/IP
Data Balancing	Placing new replicas on chunk servers with below average disk space utilization	Avoiding disk space utilization on write (prevents bottle-neck situation on a small subset of DataNodes).
Database Files	Bigtable is the database used by GFS. Bigtable is a proprietary distributed database of Google Inc.	HBase provides Bigtable like capabilities on top of Hadoop Core.

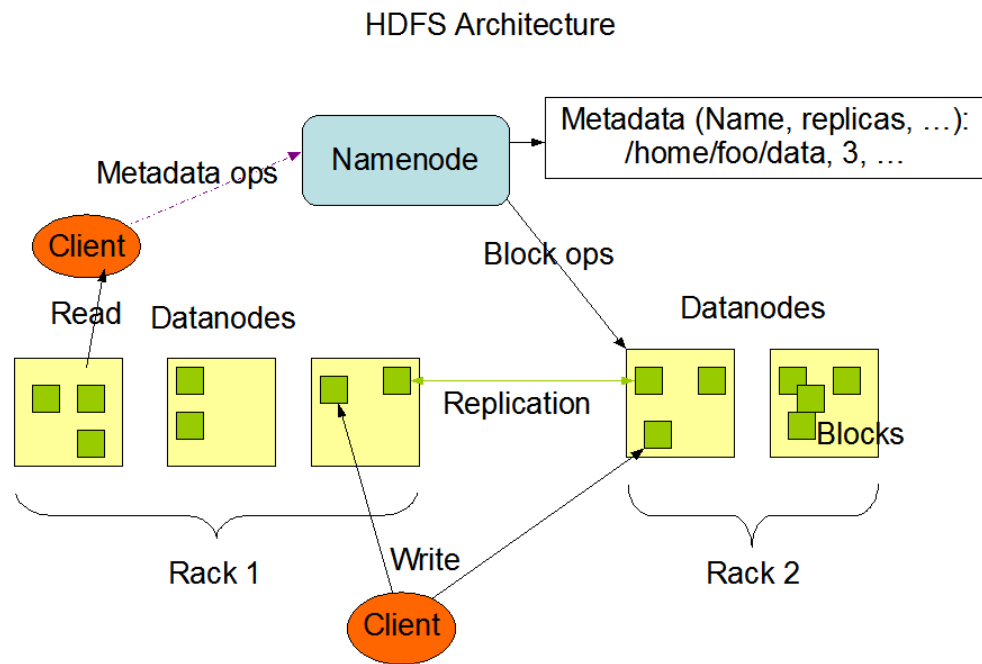


Figure 3.22: HDFS Architecture (Borthakur, 2008)

HDFS has master/slave engineering. A HDFS bunch comprises of a solitary NameNode, an expert server that deals with the document framework namespace and manages access to records by customers. Furthermore, there are various DataNodes, typically one for every hub in the group, which oversee capacity connected to the hubs that they keep running on. HDFS uncovered a document framework namespace and permits client information to be put away in records. Inside, a record is part into one or more squares and these pieces are put away in an arrangement of DataNodes. For better understanding of HDFS storage architecture is given in Figure 3.9.

The NameNode executes record framework namespace operations like opening, shutting, and renaming documents and registries. It likewise decides the mapping of pieces to DataNodes. The DataNodes are in charge of serving read and compose demands from the record framework's customers. The DataNodes likewise perform square creation, erasure, and replication upon direction from the NameNode (Borthakur, 2008).

3.1.5. Querying Data:

Once the files have been loaded into HDFS these files can be used in impala to query the data files. In impala the external tables were created as define above in star schema and load data from CSV files into external tables. Once the data has been loaded into external tables Impala queries can be applied and the output of query are displayed. Cloudera Impala is a distributed, massively parallel processing (MPP) database engine on Hadoop. Impala doesn't build on MapReduce. It is its own execution engine that accepts

queries in SQL and is capable of querying data from HDFS and HBase (Cloudera, 2015).

Impala daemon Installed on N occasions of a N hub group Impala daemons are the one that do "genuine work". They frame the center of the Impala execution motor and are the ones perusing information from HDFS/HBase and conglomerating it. StateStored introduced on 1 example of a N hub bunch. StateStore daemon is a name administration. It monitors which Impala daemons are up and running, and transfers this data to all the Impala daemons in the bunch so they know about this data when disseminating undertakings to other impala daemons. Catalog daemon introduced on 1 occasion of a N hub group. It conveys metadata (table names, section names, sorts, and so forth.) to Impala daemons through the statestored. The architecture of Impala is given in Figure 3.10.

An impala daemon can be separated into 3 sections. Planner is in charge of parsing out the question and making a Directed Acyclic Graph of administrators from it. The arranging happens in 2 sections. Initial, a solitary hub arrangement is made, as though every one of the information in the bunch dwelled on only one hub, and besides, this single hub arrangement is changed over to a disseminated arrangement in view of the area of different information sources in the group. Coordinator is in charge of organizing the execution of the whole question. Execution Engine is in charge of perusing the information from HDFS/HBase and/or doing aggregations on information (Grover, 2014).

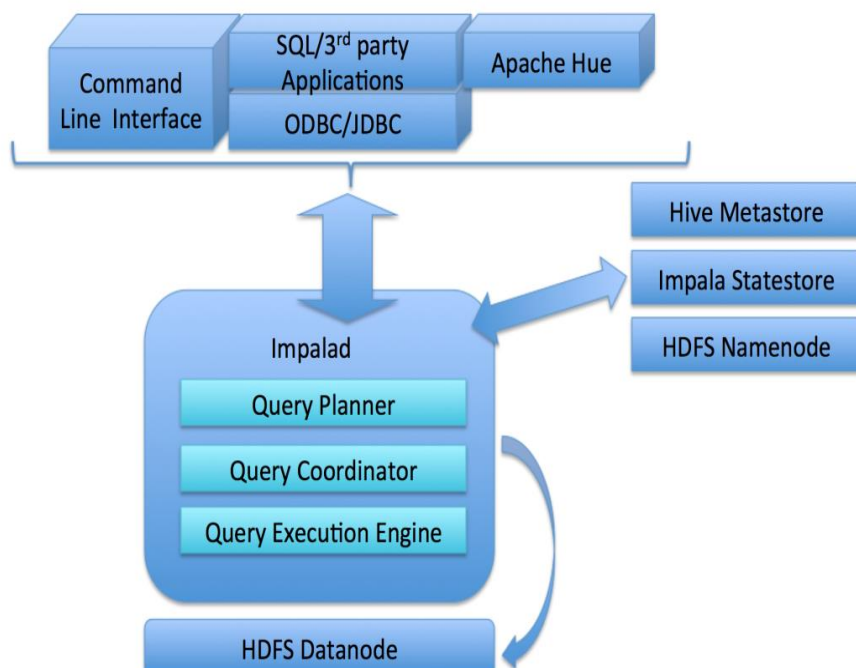


Figure 3.23: Architecture of Impala (Grover, 2014)

3.1.6. Visualization:

After the data has been transformed and processed it can be used to visualize by using Google Chart API. The result of data mining is processed information but complex enough to understand for a general user. Therefore it is necessary to apply visualization techniques. The information was displayed in graphical form on a web page.

3.2. Tools to be Used

In this section, all basic tools used in this research and their features are described and minimum hardware/software requirements for each tool are discussed. And how these tools help us in the research process has been also explained in this section.

3.2.1. Workflow management system

SciCumulus/C² (SCC) is a parallel Scientific Workflow Management System (SWMS), which allows scientific workflow modeling, execution and analysis. The SCC was developed for supporting the parallel execution on High Performance Computing (HPC) environments. Specifically, SCC was developed to execute computer simulations on cluster and cloud environments, as an integration of Chiron and SciCumulus (SciCumulus, 2016).

The parallel processing of SCC follows the MapReduce (e.g. Hadoop) style. Although, the SCC implements his own algebraic approach that can perform optimization, dynamic scheduling and workflow steering. In this system implementation this tool was used to design dimensional model and MR workflow in Transformation Process.

3.2.2. Vmware player

VMware Player is a streamlined desktop virtualization application that runs one or all the more working frameworks on the same PC without rebooting. With its basic client interface, unmatched working framework backing and compactness, it's presently less demanding than at any other time for IT experts to get their clients up and running with a corporate desktop (Vmware Products, 2016). Vmware player is used to install cloudera quickstart VM (Virtual Machine). It is pre-requisite for installing cloudera.

3.2.3. Cloudera

Cloudera gives an adaptable, adaptable, incorporated stage that makes it simple to oversee quickly expanding volumes and assortments of information in your venture. Cloudera items and arrangements empower you to convey and oversee Apache Hadoop and related ventures, control and dissect your information, and keep that information

secure and ensured.

Cloudera CDH

Cloudera Distribution of Hadoop (CDH) is the most finish, tried, and well known conveyance of Apache Hadoop and related undertakings. CDH conveys the center components of Hadoop adaptable stockpiling and appropriated registering alongside a Web-based client interface and indispensable endeavor capacities. CDH is Apache-authorized open source and is the main Hadoop answer for offer bound together clump preparing, intelligent SQL and intuitive hunt, and part based access controls (Cloudera, 2015). Cloudera provides all services used for big data analysis at single place. All available core components are described in Figure 3.11.

Cloudera HDFS

HDFS is a deficiency tolerant and self-mending disseminated document framework intended to transform a bunch of industry-standard servers into an enormously versatile pool of capacity. Grown particularly for expansive scale information preparing workloads where adaptability, adaptability, and throughput are basic, HDFS acknowledges information in any organization paying little respect to blueprint, streamlines for high-data transmission gushing, and scales to demonstrated arrangements of 100PB and past.

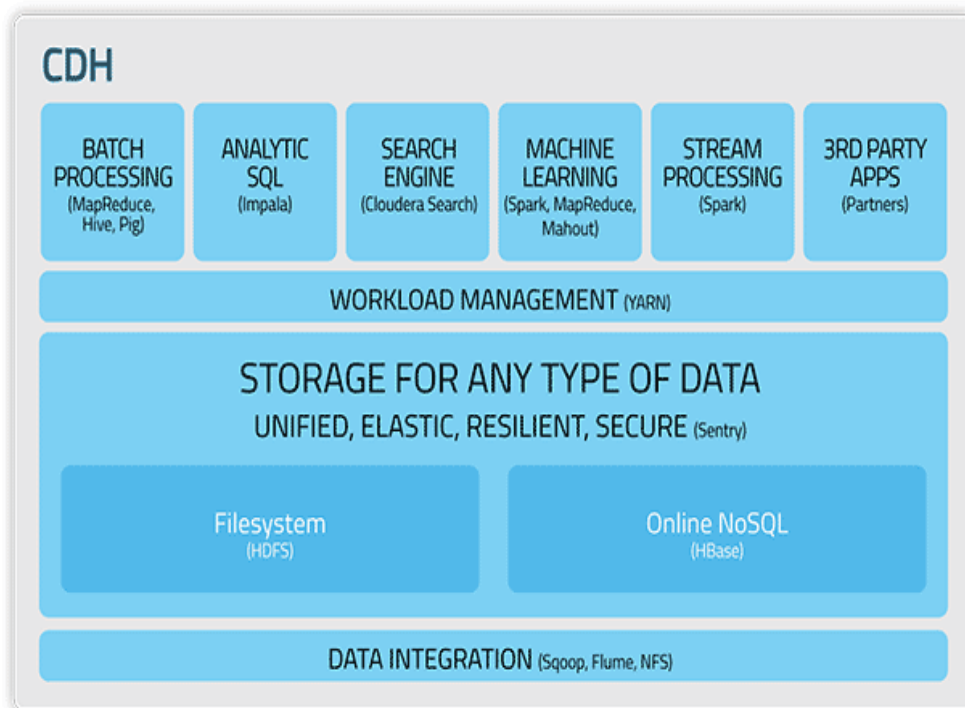


Figure 3.24: Cloudera CDH core components (Cloudera, 2016)

Cloudera MR

MapReduce keeps on being a famous cluster preparing apparatus; Apache Spark's adaptability and in-memory execution make it an a great deal all the more capable clump execution motor. Cloudera has been working with the group to bring the systems presently running on MapReduce onto Spark for quicker, more strong preparing. MapReduce is intended to handle boundless measures of information of any sort that is put away in HDFS by isolating workloads into numerous undertakings crosswise over servers that are keep running in parallel.

Cloudera Impala

Impala can support high simultaneousness workloads to give expansive access to business examiners over the whole business, for the speediest time-to-bits of knowledge. For multi-user queries, Impala is on average 16.4x faster than Hive and 7.6x faster than Spark SQL, with an average response time of 12.8s compared to over 1.6 minutes or more. All 99 TPC-DS-derived queries in the benchmark run on Impala, with similar results in Impala's favor. Impala uses Parquet file format which is a columnar stockpiling group accessible to any undertaking in the Hadoop biological community, paying little respect to the decision of information handling structure, information model or programming dialect.

3.2.4. Visualization Tools (Google Charts)

The Google Chart API is a tool that lets individuals effortlessly make an outline from a few information and insert it in a site page. Google makes a PNG picture of a diagram from information and arranging parameters in a HTTP ask. Numerous sorts of graphs are upheld, and by making the solicitation into a picture label, individuals can basically incorporate the diagram in a website page. Initially it was an inner apparatus to bolster quick inserting of outlines inside Google's own applications (Google, 2016).

Google charts provide a large range of charts and easily customizable. Here are the names of few Area chart, Line chart, Pie chart, Bubble chart, Scatter plot, Bar chart, Gantt chart, Histograms etc.

3.2.5. XAMPP Server

XAMPP is the most famous PHP advancement environment. XAMPP is a totally free, simple to introduce Apache circulation containing MariaDB, PHP, and Perl. The XAMPP open source bundle has been set up to be inconceivably simple to introduce and

to utilize. XAMPP helps you make and build up your own applications utilizing Web server advancements. Introducing an apache web server is not a simple errand the XAMPP server gives your own particular localhost server to test and run the website pages.

3.3. System Development

In this section, details about how to configure and use the above tools and how these tools help us to achieve our required output and goals is given. In this section, the overall system development, experimental setup, basic requirements to startup and assumptions to develop this system has been explained.

Software Requirements

- i. 64-bit Operating System (OS) Windows 8 or higher versions

Hardware Requirements

A laptop or PC with:

- i. Core i5 processor with visualization and VTx technology support
- ii. Minimum 12GB RAM
- iii. 250 GB Hard Disk

Implementation

As it has been discussed above in first step the data was downloaded from PingER website which is collected by PingER Measurement Agents (MA). This data stores in flat files and some aggregated operations are applied on this data after this there is cleaned data which is used in next steps.

3.3.1. Transforming Data

In transformation process the data is transforms into a DWH dimensional model by using MR framework. For this purpose install SciCumulus and performs the following steps. To install the SciCumulus it is needed to install three components. Firstly Java Runtime Environment (JRE) it provides basic java libraries to write and execute MR program written in java. In this system JRE version 1.8.0_91 for windows 64-bit is used. Download this program setup from the link given in literature cited section. After download the installation process is quiet simple just run the application and follow the

instruction on installation wizard. After installation it is necessary to setup the environment variables of system as JAVA_HOME. To setup goto system properties advanced tab click on environment variables option and add new system variable as given in Figure.3.12.

Clicking on new button opens a new window define the variable name and Value which is the path of folder where the JRE is installed as shown in Figure.3.13. To check the JRE is installed properly run the `java -version` command on windows command prompt it shows the version of JRE as given in Figure.3.14.

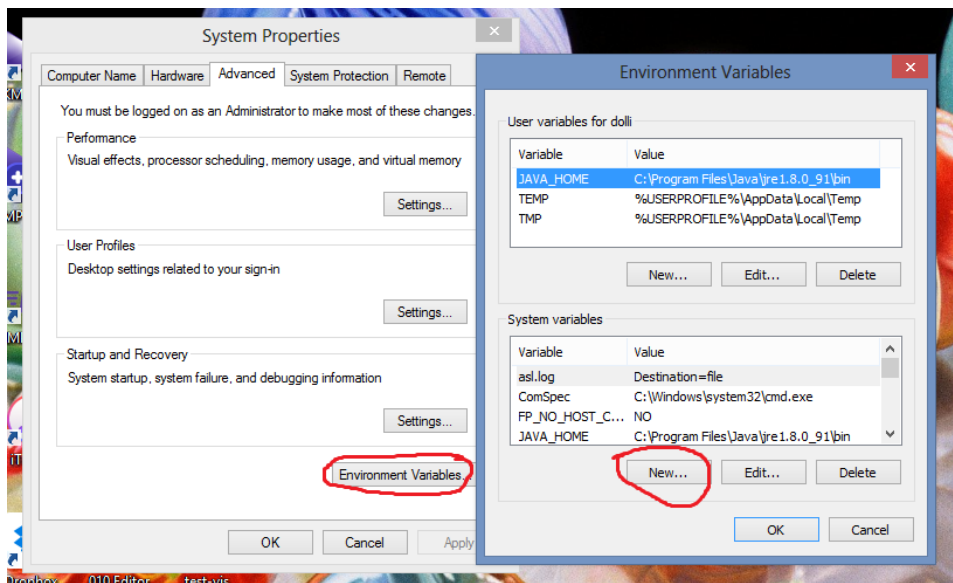


Figure 3.25: Setting up system environment variables

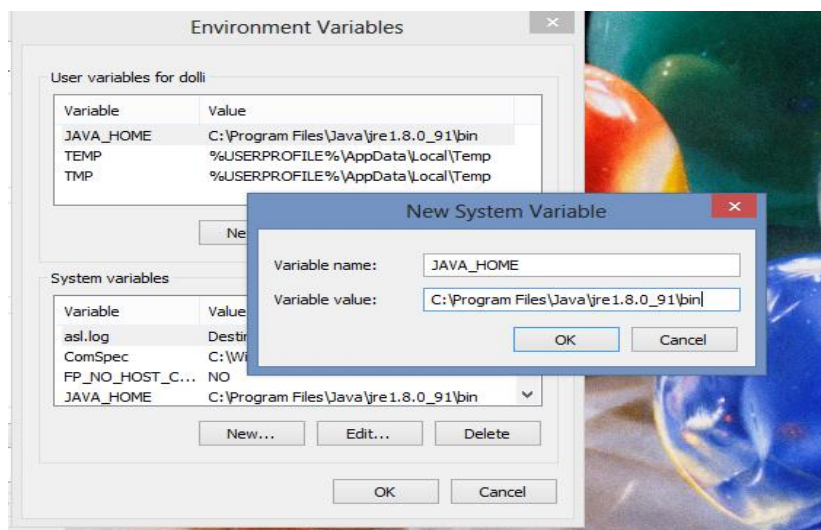
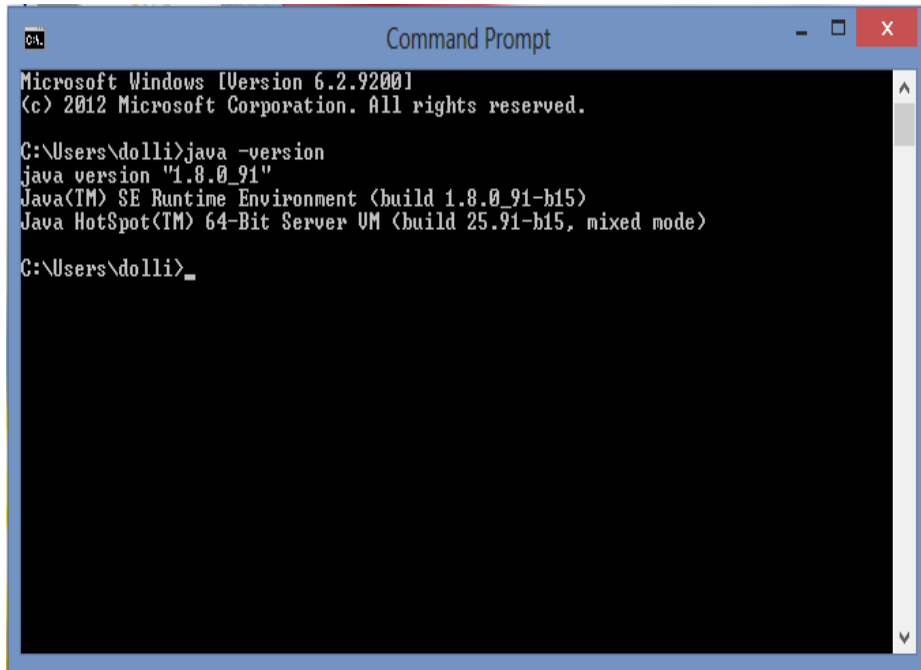


Figure 3.13: Creating new system variable

A screenshot of a Windows Command Prompt window. The title bar reads "Command Prompt". The window content shows the following text:

```
Microsoft Windows [Version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\dolli>java -version
java version "1.8.0_91"
Java(TM) SE Runtime Environment (build 1.8.0_91-b15)
Java HotSpot(TM) 64-Bit Server VM (build 25.91-b15, mixed mode)

C:\Users\dolli>_
```

Figure 3.14: Testing Java Installation

The second step is to install the postgresQL database but in this system development it is not necessary to install this database so it is skipped optionally. The third step is to install the MPJ which is a message passing library and provide execution of java programs in parallel and distributed environments like Hadoop cluster. In this research MPJ Express version 0.44 was used download and unzip the file. It is also necessary to setup MPJ system variables as above it is done for JRE. MPJ_HOME and value is the path *C:/mpj*. MR java program is written in MPJ express as *PingER.java* and compile by using the command in command prompt *javac -cp .;%MPJ_HOME%/lib/mpj.jar PingER.java* and to execute *mpjrun.bat -np 4 PingER.java*.

MR program was written in java language which includes two functions map function maps the *PingER* legacy data into dimensional model of DWH and reduce function reduces the files according to each year. The java program is executed by using MPJ express named as *PingER.java*. *PingER.java* take input of flat text *PingER* files and outputs multiple transformed files in CSV format which contains the same data as in flat files but the data holds a dimensional structure.

3.3.2. Loading CSV files on HDFS

To load the CSV files on HDFS it is required to setup a Hadoop cluster. As HDFS is a storage architecture provided by Apache Hadoop so to setup Hadoop cluster in this implementation Apache Cloudera is used which provides a single node Hadoop

cluster. More nodes can be added by using CDH provided by cloudera or it can be setup manually.

To install cloudera a single node cluster download cloudera quickstart VM version 5.4.2-0 and vmware player. VMware provide an interface on which multiple VMs can be install and run. After downloading the vmware player install it simply by using installation wizard. After installation it is necessary to open cloudera quickstart VM in vmware as shown in Figure.3.15 (a). When cloudera quickstart VM is opened in vmware it displays in the left pane of window Figure.3.15 (b). After this click on the cloudera quickstart VM in left pane and there is a new window appear which displays the option of edit VM settings. As shown in Figure3.16.

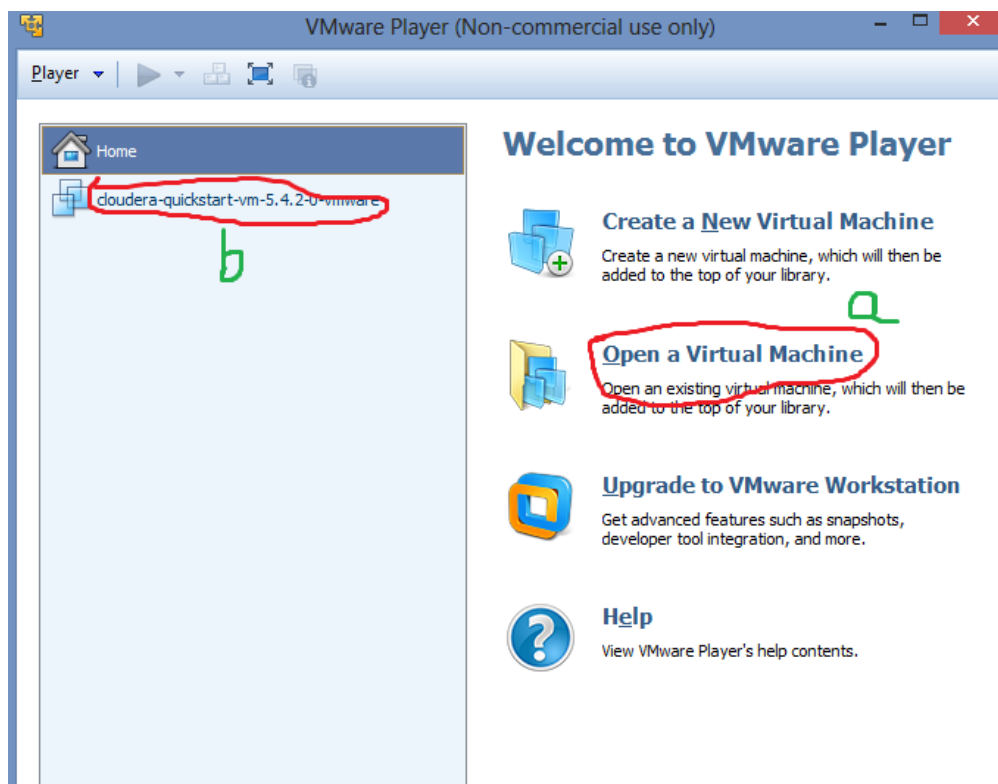


Figure 3.15: Opening the Cloudera VM



Figure 3.16: Editing the VM settings

To use cloudera manager which provides all basic services of Big Data it is necessary to edit the virtual machine settings. After click on edit VM settings option a new window appears change the following settings given in Figure3.17. Select the memory maximum 8GB as in Figure.3.17 (a) and change the processor cores to 2 as in Figure. 3.17 (b) because to run cloudera manager it is required the 8GB RAM and 2 processors. After this click on the options tab as given in Figure. 3.17 (c) to change the OS settings

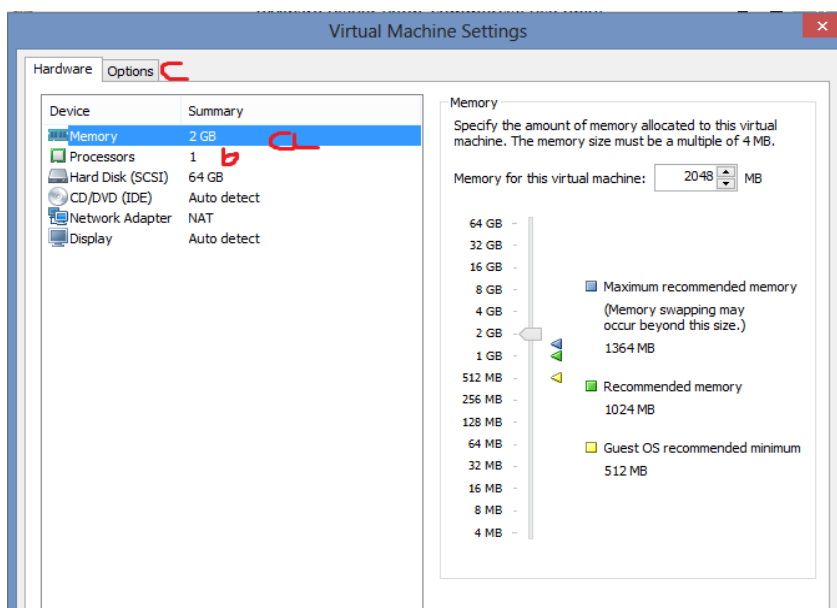


Figure 3.17: Changing processor and memory settings

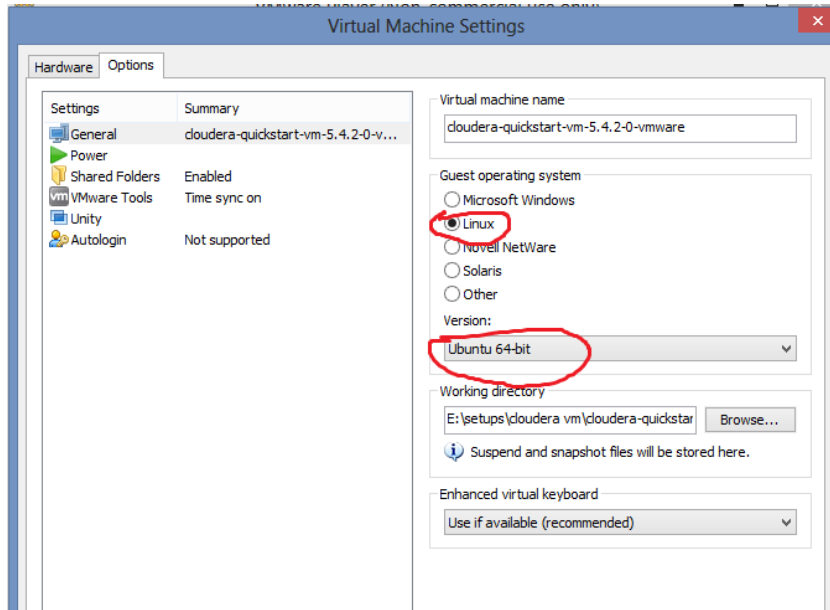


Figure 3.18: Changing OS settings

Cloudera only supports linux 64 bit OS. The VM settings are set to Ubuntu 64 bit as given in Figure.3.18. After changing the settings close the edit VM setting window and click on the play VM. This cause the starting up of VM. Start up process boot the SentOS which is linux core kernel component. After booting the centOS cloudera quickstart desktop appears as given in Figure.3.19.



Figure 3.19: Cloudera Desktop

Figure.3.19 (a) is a application of Cloudera Manager (CM) it is used to manage all services of cloudera the user can start, stop and monitor the performance of services. CM help to manage Hadoop cluster, adding Datanodes and monitoring the host services. Figure. 3.19 (b) explains the default browser which is mozilla firefox. It is used to open all pages like CM, Hue, Impala and HDFS. All these services are run on a specific port numbers so it can be accessed by default browser. Figure.3.19 (c) is the terminal which provides the interface for running all Hadoop commands. As in windows OS there is a command prompt to run commands same like in cloudera VM there is terminal. In MAC OS the command prompts are also called terminal. Figure.3.19 (d) is the document library for Cloudera VM you can view and save file in this folder.

When the Cloudera VM will be started the default browser Mozilla Firefox starts automatically and shows the startup page as given in Figure.3.20. The page shows the cluster information and default IP addresses of worker and manager node. On the top there is bookmark bar which displays all service pages bookmarked. It helps to open all services on one click instead of typing the IP address and port number.

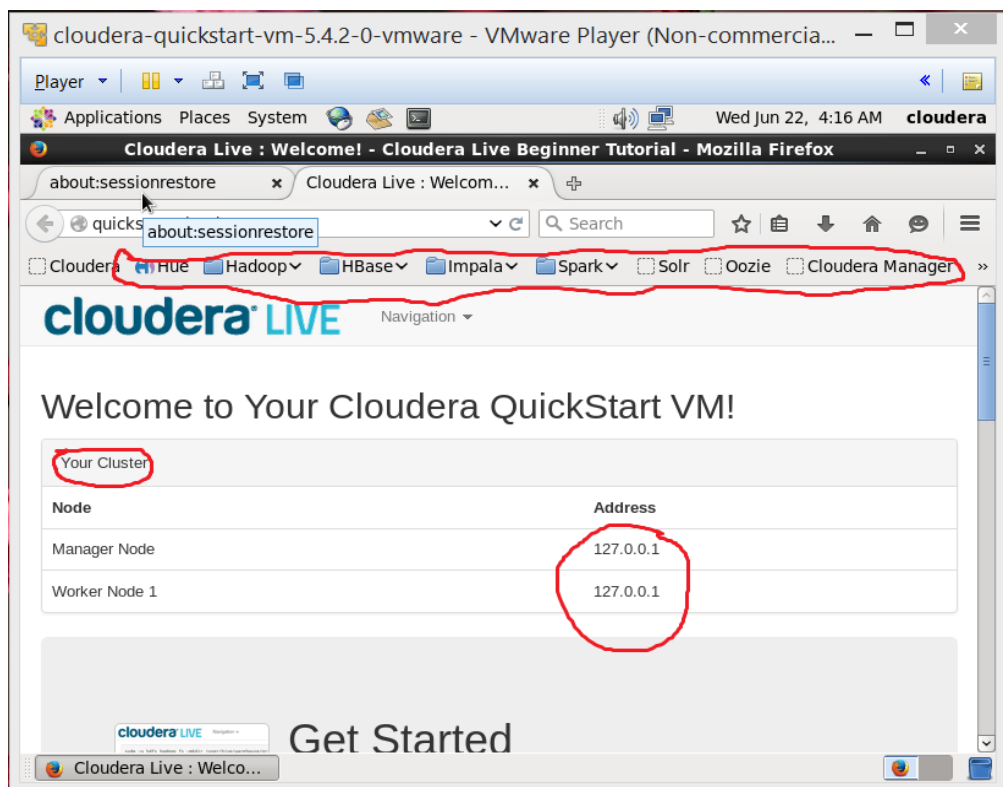


Figure 3.20: Cloudera startup page

First of all it is necessary to start up all basic services and servers of cloudera manager which provides services like HDFS, MR and Impala. To start the services click on cm express application as given in Figure.3.19 (a). When the CM express application was run a window appears with the message as given in Figure.3.21. This application script required to be executed only once there is no need to run this application each time the Cloudera VM start ups. Because this application has been executed before therefore the message is different but when this application was executed first time the different message will be appear. After executing this application just close the application. Because CM will be accessed by default browser.

Click on the default browser and click on the cloudera manager. It requires a login the default username and password is cloudera. CM is installed on port 7180 but it is not necessary to remember the port numbers for all services, bookmarked links can be use simply to access the services as given in Figure.3.22.

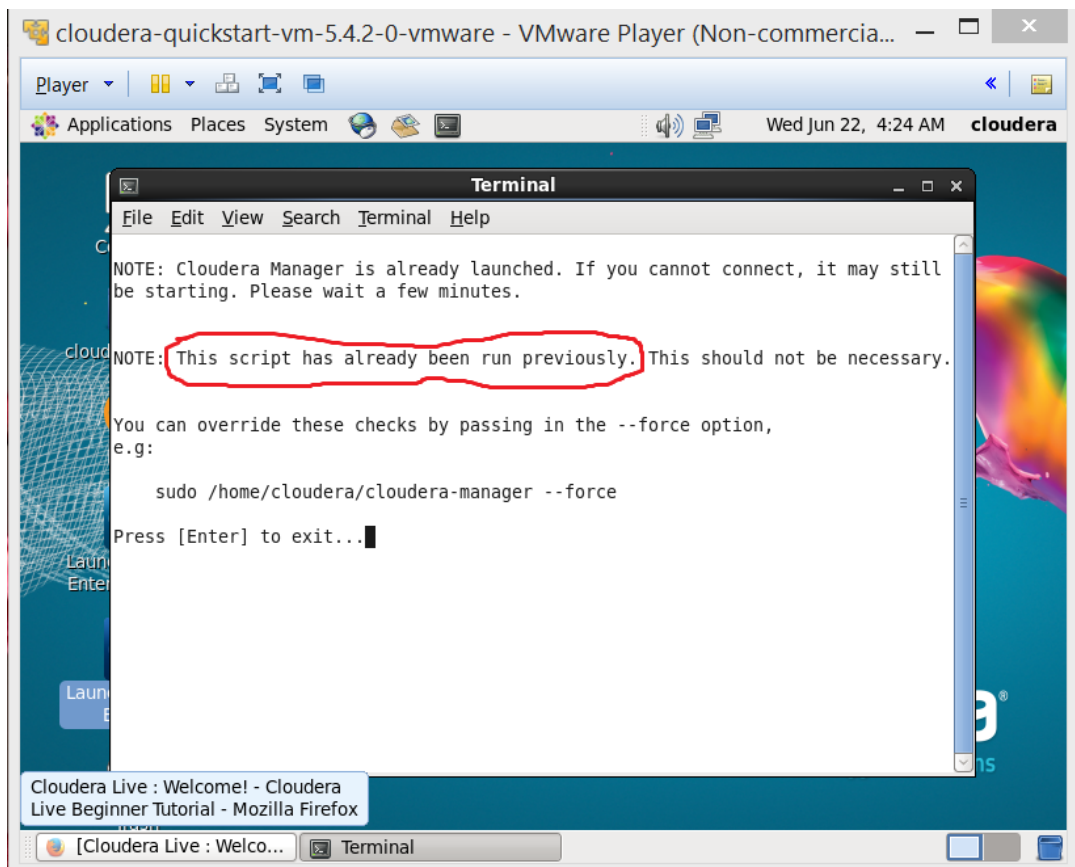


Figure 3.21: CM express launch screen

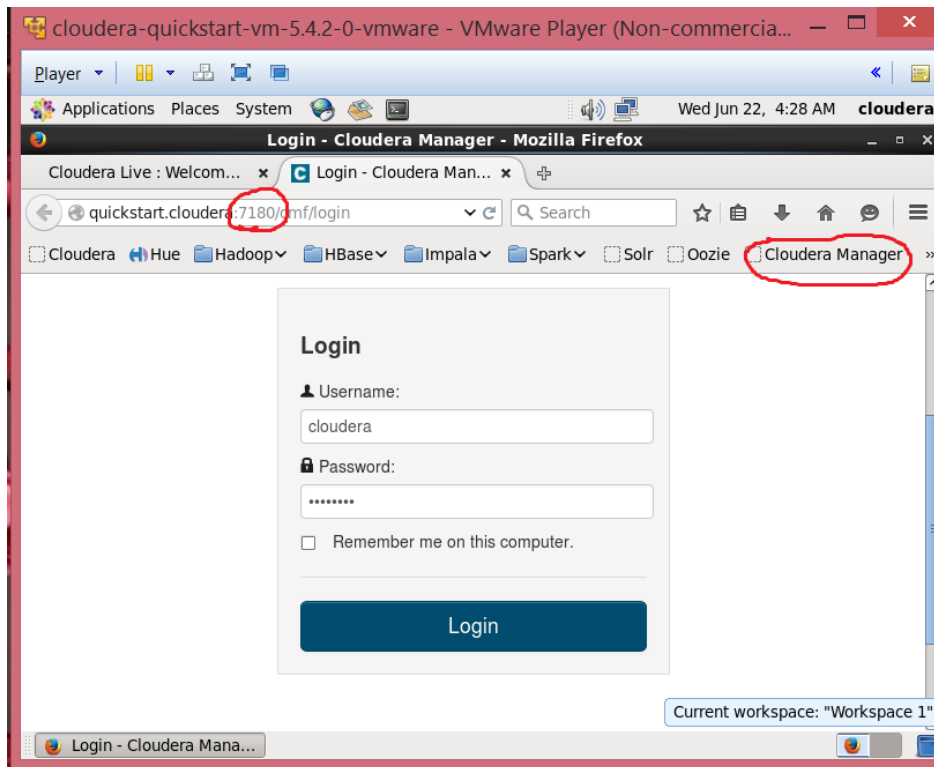


Figure 3.22: Running CM

After logging in the following screen appears just click on the arrow in front of every service and click on start as given in Figure.3.23. Start all services one by one. Each service takes some configuration time to start up the servers.

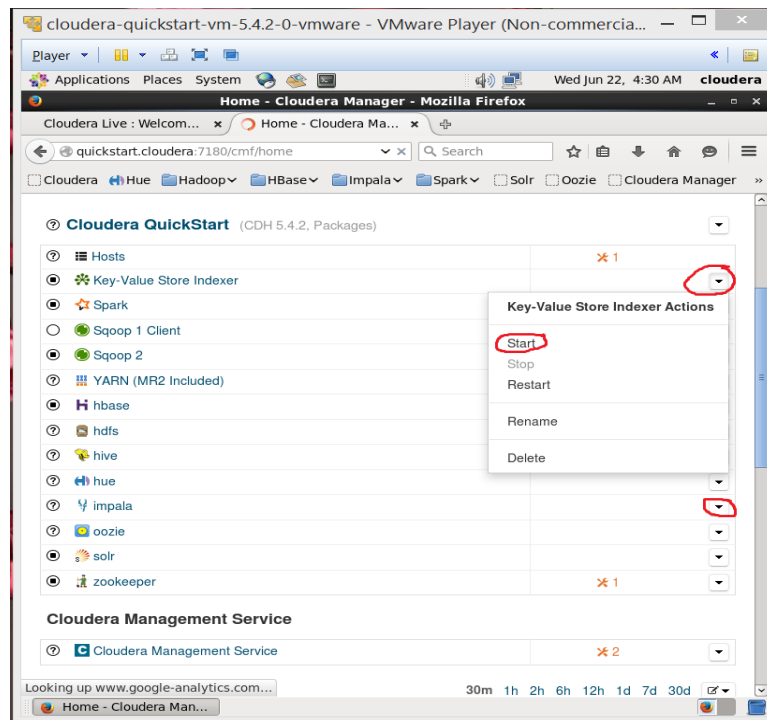


Figure 3.23: Starting up CM services

Remember that if the HDFS service is not running the files cannot be uploaded on HDFS. There are two ways of uploading files on HDFS one is by using command prompt and second is by using Hue. Hue provides graphical interface to upload the files it located files from cloudera home document libraries but if terminal was used then the file can be uploaded directly from windows desktop. In this system development Hue was used so it is required to paste the CSV files in cloudera document libraries. Open the cloudera home as in Figure.3.19 (d). and just drag drop the CSV files from windows desktop to cloudera home directory as shown in Figure.3.24. Now open the Hue webpage which is bookmarked on the default browser of cloudera as given in Figure.3.25. and provide login/password which is cloudera/cloudera.

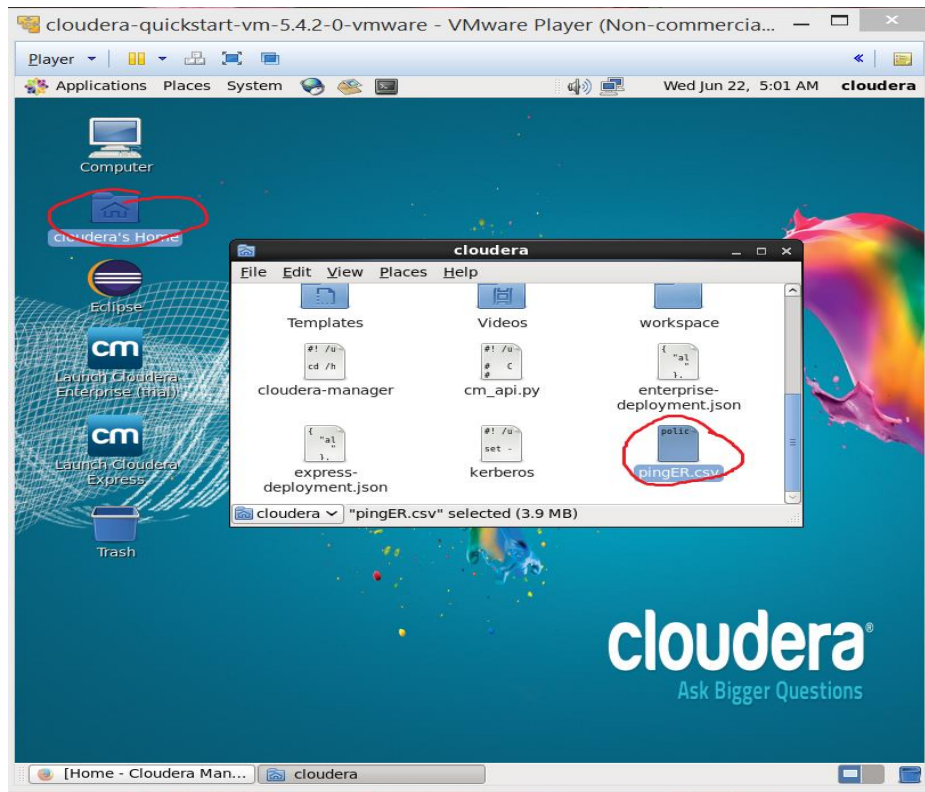


Figure 3.24: Cloudera home document directory

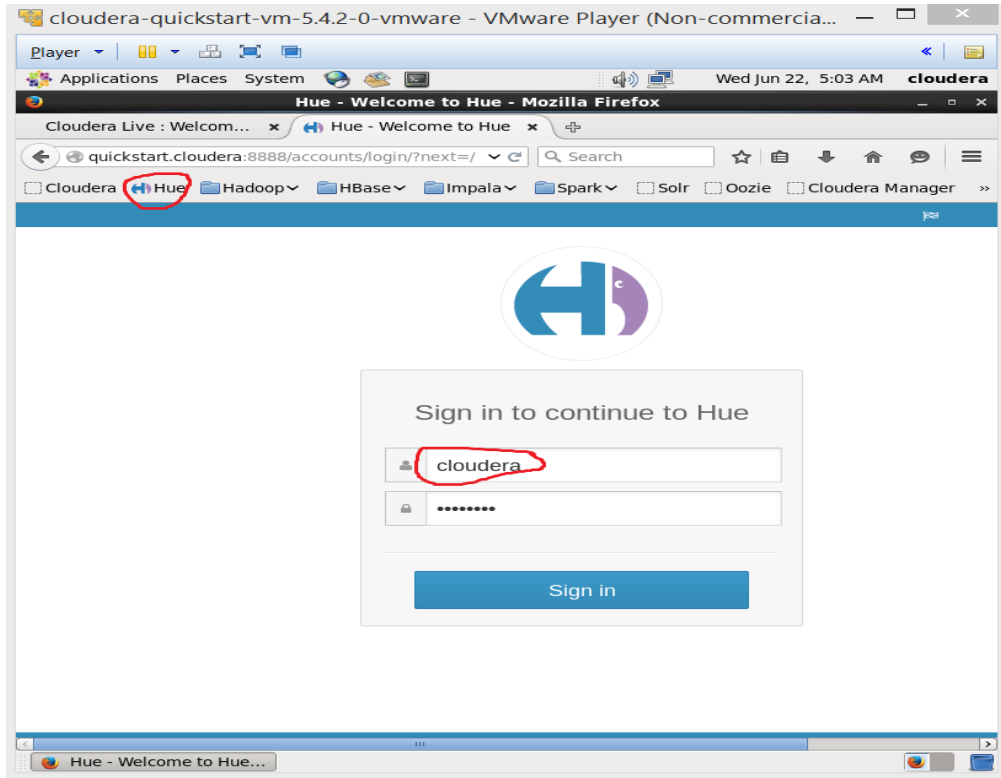


Figure 3.25: Running Hue

Hue requires all services in running and in up condition. If following screen appears as in Figure.3.26. means all services are running. But if any error message screen appears as in Figure.3.27. its mean any of your service is not configure properly and not running in this case Impala is not running and error message for Impala demons are not running. You can stop this service and restart this service by using cloudera manager.

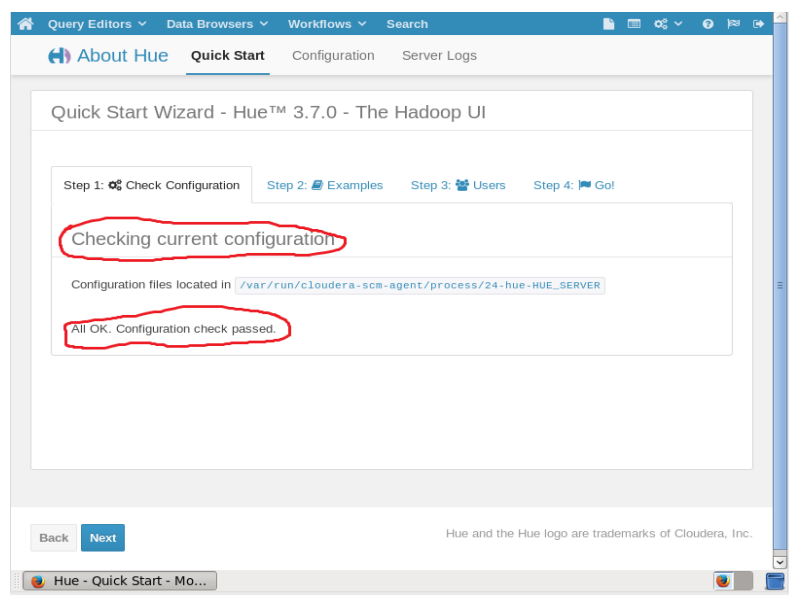


Figure 3.26: Hue configuration check

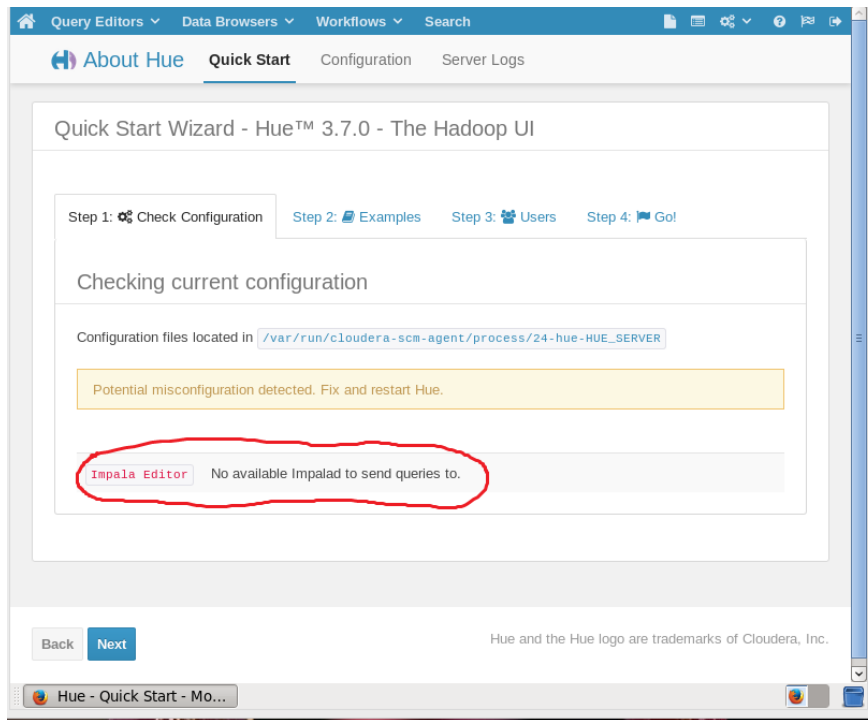


Figure 3.27: Hue configuration check error message screen

After this click on the File browser on the blue bar of hue services and the HDFS directories will appear as given in Figure.3.28. Now to upload the files there should exist the PingER file directory on HDFS. To create HDFS directories run the command `hdfs dfs -mkdir PingER` on terminal as given in Figure.3.29. The file directory has been created as given in Figure.3.30. Similarly the file directories for each csv has been created inside PingER directory.

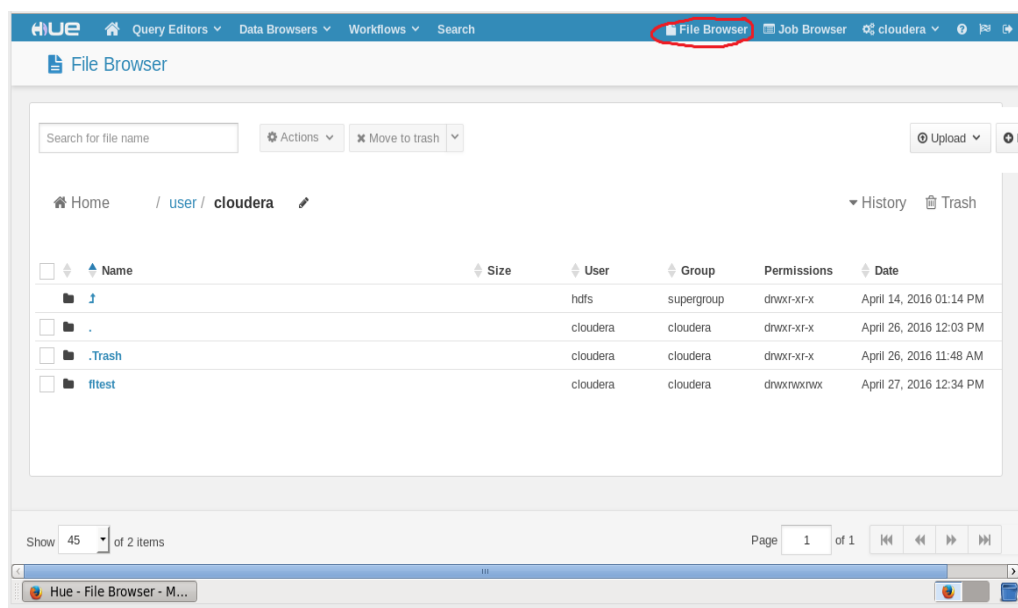


Figure 3.28: HDFS directories window

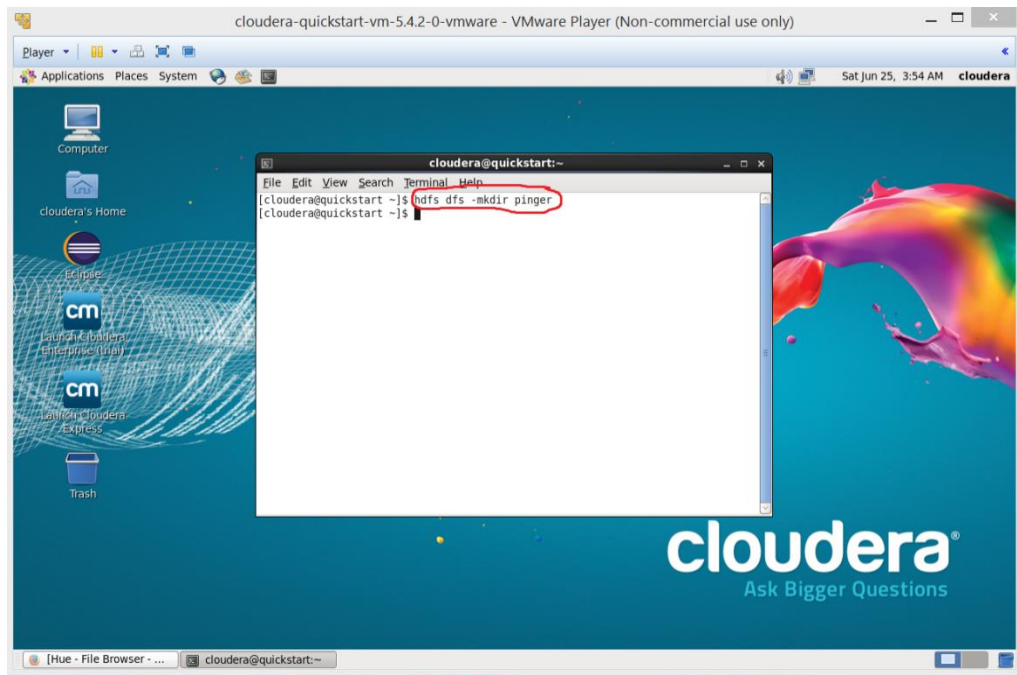


Figure 3.29: Creating new HDFS directory

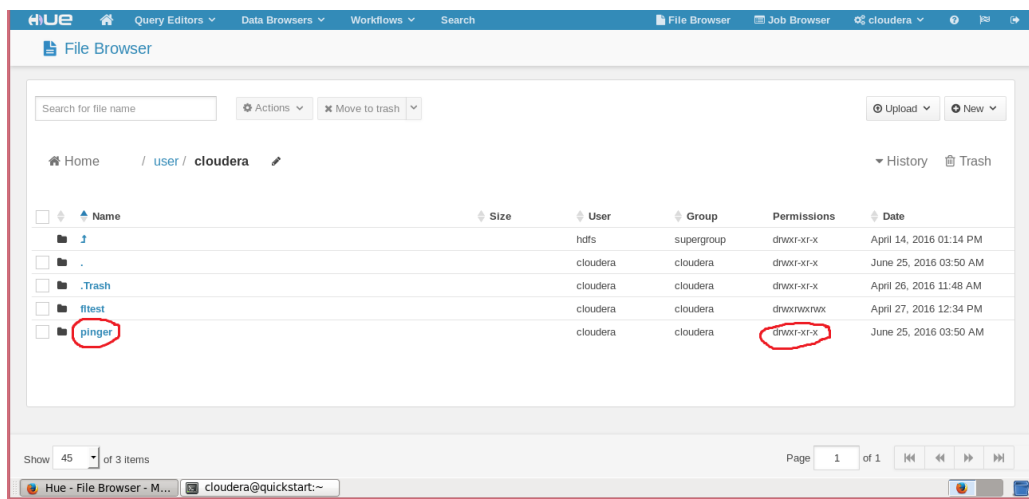


Figure 3.30: PingER HDFS directory created

Note the access permissions of file directory given in Figure.3.30. which are *drwxr-xr-x* to access this directory on Impala it is necessary to change the permissions of file directory because Impala does not have read and write access by default to this directory. To change the permissions right click on the directory and select the change permissions the given window will appear check all uncheck options except sticky and recursive options and click on submit as given in Figure.3.31. Note the permission after changing it should be *drwxrwxrwx*.

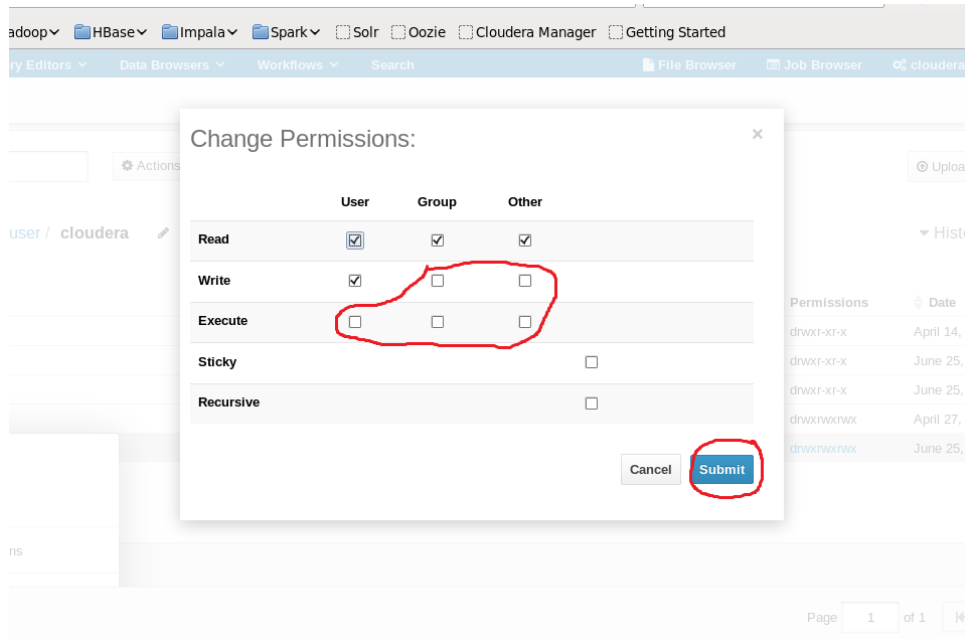


Figure 3.31: Changing permissions of HDFS directory

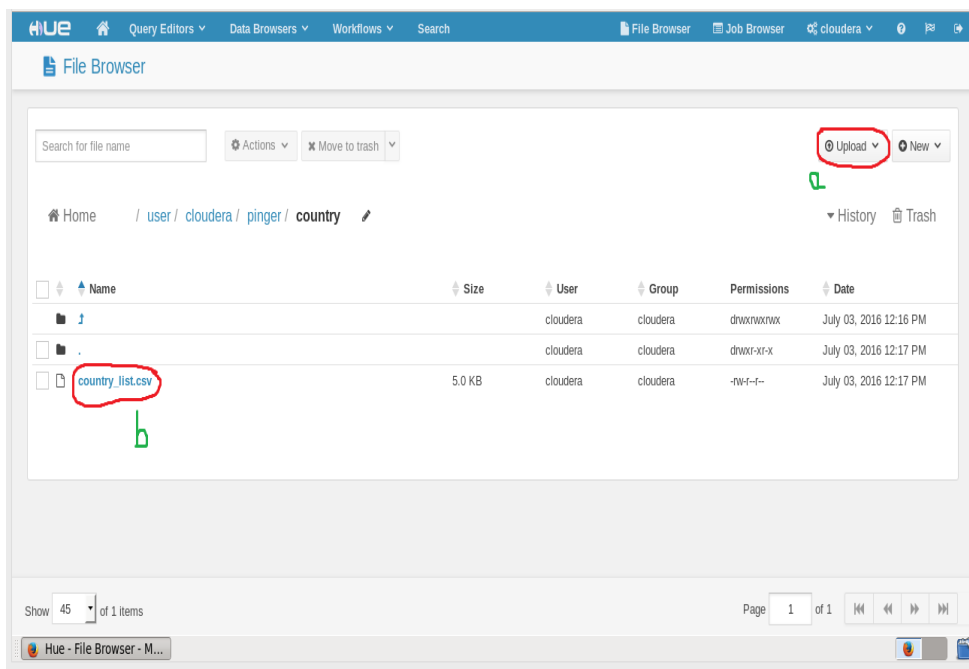


Figure 3.32: Uploading files to HDFS

To upload file click on the PingER directory, click on country directory and click on the upload button as given in Figure.3.32 (a). Locate the file from cloudera home document library and upload. Country_list.csv file has been uploaded on HDFS as shown in the Figure.3.32 (b). Similarly all other 17 CSV files for different matrices are uploaded.

3.3.3. Querying Data in Impala

Before starting and launching Impala the system needs to be restart properly so the all configurations and settings saved properly. Click on the system option and click on shut down as given in Figure.3.33. and play the Cloudera VM again from vmware player. If the VM is not restarted then the permissions for the HDFS directory may not be saved and you will see the error in Impala while loading the data into table as given in Figure.3.34

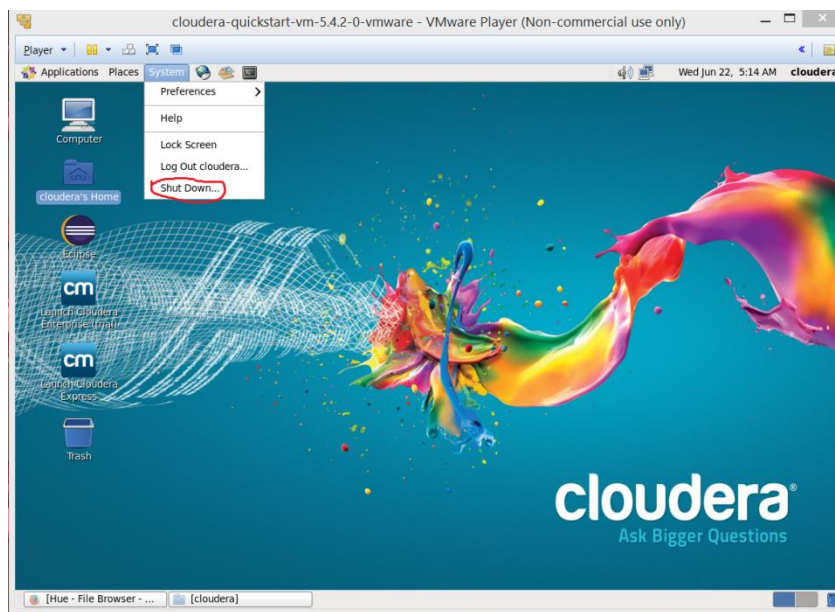


Figure 3.33: Restart Cloudera VM



Figure 3.34: Load Data statement error message window

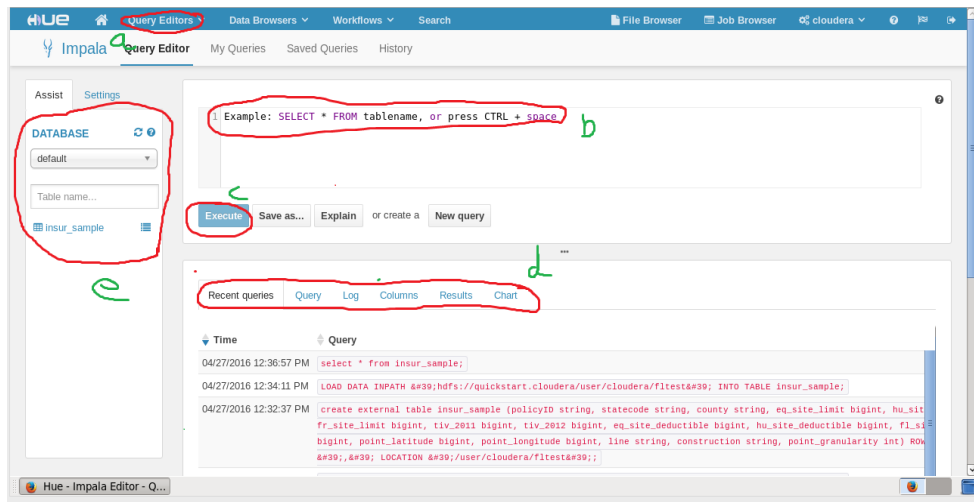


Figure 3.35: Introducing Impala Interface

To launch impala load the hue page in browser and click on the query editors option as given in Figure.3.35 (a). The Figure.3.35 shows the environment of impala. Figure.3.35 (b) shows the query editor where the impala queries can be edited and impala queries can be run by using the execute button as in Figure.3.35 (c). Figure.3.35 (d) shows the advanced tabs all recent queries or query results and logs can be viewed by using these tabs. Figure.3.35 (e) shows default database of cloudera and list of all tables can be viewed in left pane.

To load the data from CSV files it is necessary to create an external table. External table represent the basic schema of CSV file and external table which has been created is the empty table. To create the external table type following query in query editor and execute as given in Figure.3.36. Note that the names of fields and their types should be the same as in CSV file and the sequence of fields also should be in same order.



Figure 3.36: Creating external database table

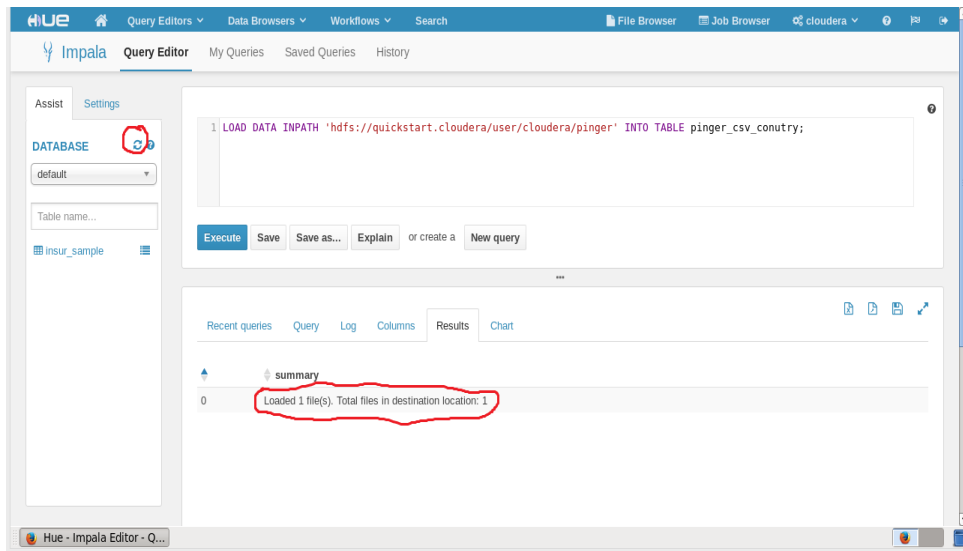


Figure 3.37: Data loaded in external table

After creating the external table it is necessary to load the data from CSV file. To load the data run the *LOAD DATA INPATH* statement as given in Figure.3.37. After executing the statement the message appeared in the lower pane of impala window. On the left pane the button highlighted used to refresh the database click the button and the new created table displays in the list as given in Figure.3.38.

After creating the table and loading data any query can be executed to retrieve the data according to user's requirement in this case just a simple *select * from PingER_csv_country* statement was executed as given in Figure.3.38.

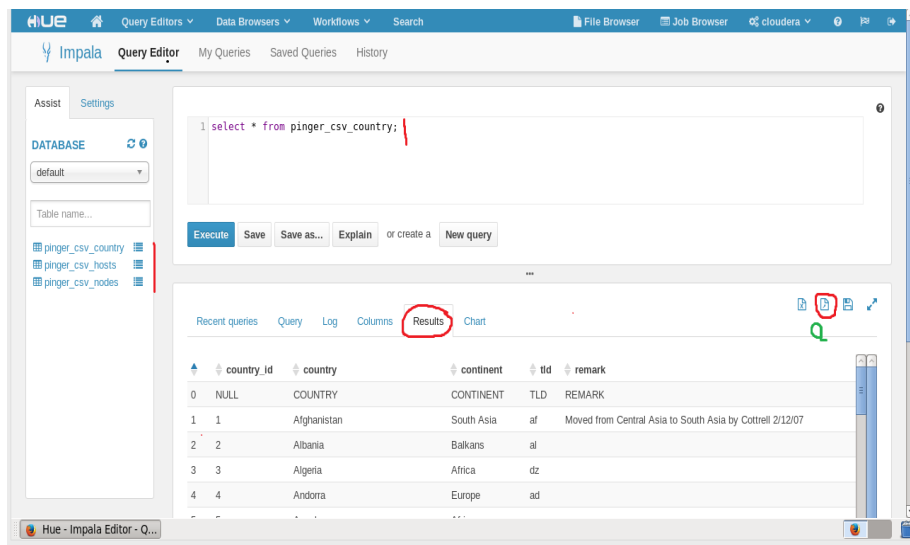


Figure 3.38: Results of query

Now for the next step visualizing the query results it must be in CSV files because Google API only can be integrated with CSV files. It is necessary to export the query results as a CSV file. To export the results click on the button given in Figure.3.38 (a). and a window will appear to save file as given in Figure.3.39. click on save file option and click ok button. The file starts downloading in browser and ater downloading you can explore this file in cloudera home document directories in downloads folder as shown in Figure.3.40. When the file has been located in the downloads you can just again use drag/drop to move the file on windws desktop.

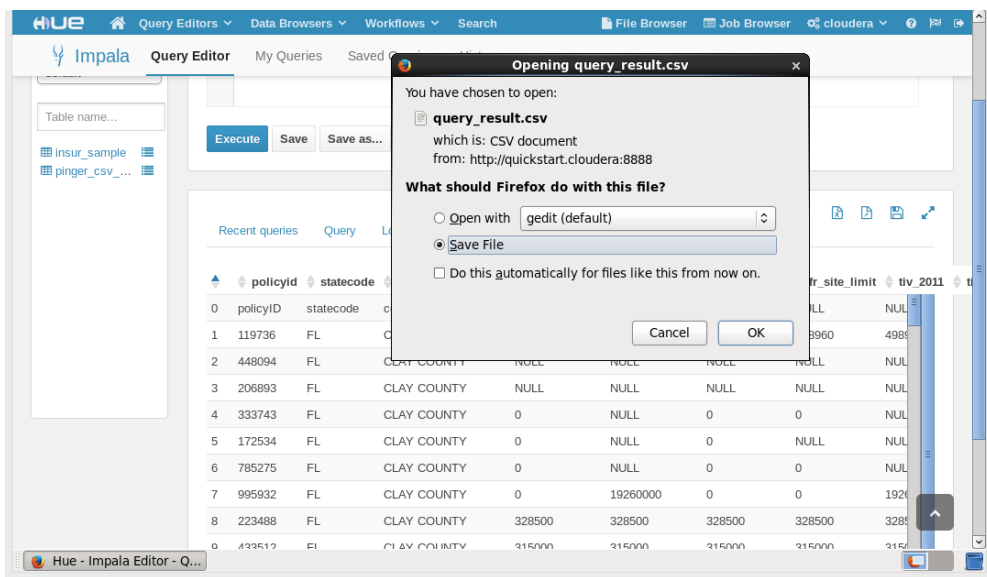


Figure 3.39: Saving query results as CSV

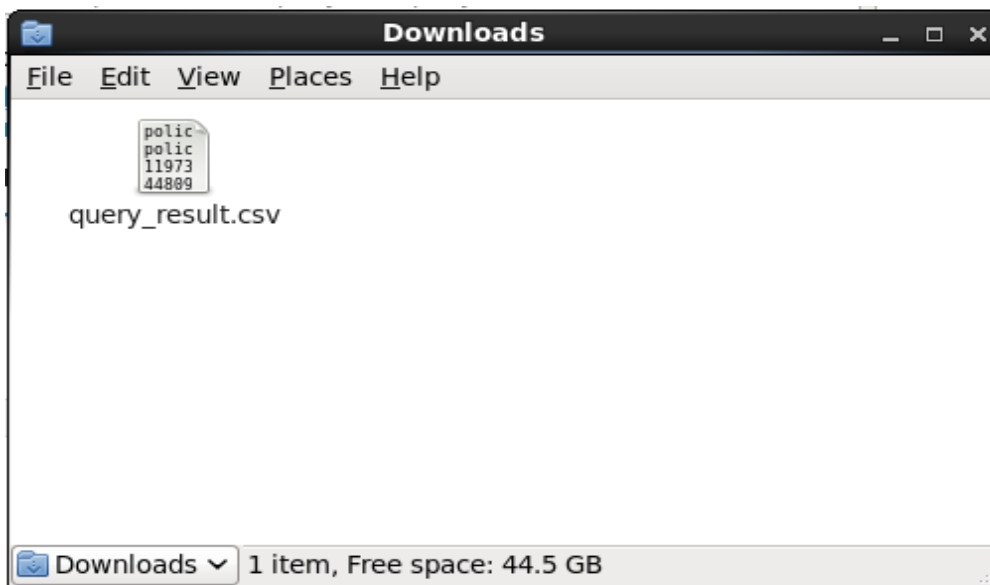


Figure 3.40: Locating CSV in downloads folder

3.3.4. Visualizing the query results

The process of visualization data is quiet simple. To start with visualization it is required to install the Xampp server version 2.5. download the setup file and complete the installation by just following the installation wizard. It is given in Figure.3.41. During installation specify the directory path in C drive or system drive of your windows.

After installation run the xampp control panal application from the xampp folder located in *C:\xampp* and start the Apache and SQL services as shown in Figure.3.42 (a) and after starting the status of services are as running and start button shows the stop option to stop the services as given in Figure.3.42 (b).

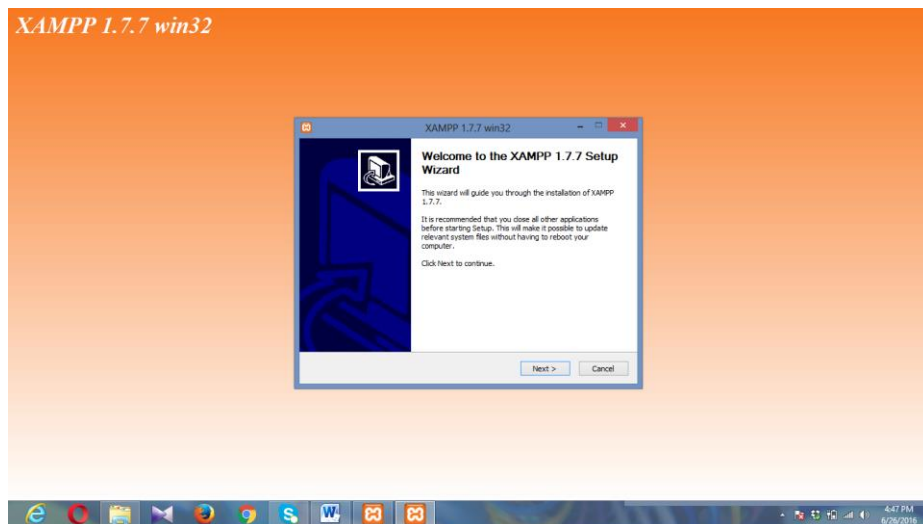


Figure 3.41: Installing xampp

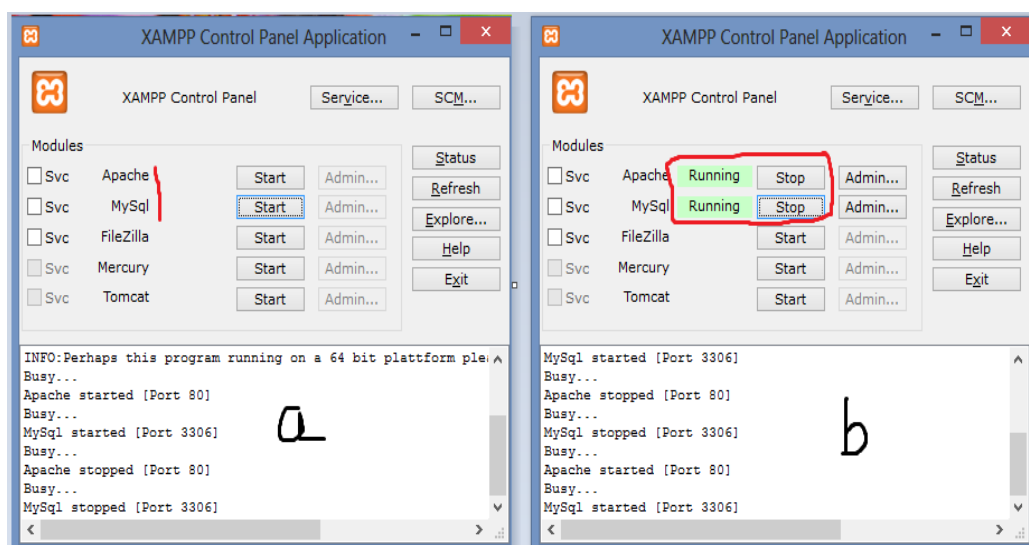


Figure 3.42: Starting xampp apache services

Now create a new folder in htdocs folder with the name PingER-visualization htdocs folder is the root document library of xampp. In this folder paste the CSV file which was exported from impala containing the impala query results and to display the charts it is necessary to create a basic html or php page. To create the basic html page you can use any text editor and save the file with .html or .php extension. If the simple windows text editor was used it is necessary to add the file extension manually while saving or if notepad++ text editor was used then it provides you basic options code the html and save the html page. In this research Notepad++ editor was used. To create the html page open the Notepad++ editor paste the following code and save as index.html page in the PingER-visualization folder.

It is not necessary to understand the whole code but just few lines of code are important to understand. Now there are two files in the PingER-visualization folder the CSV and html page. First it is necessary to look at the code let just review the basic sections. Figure.3.43 shows the 3 lines of code which is used to include the Google API. To use google API it is necessary to include these javascript libraries. There are three .js libraries JQuery.min.js, JQuery.csv.min.js and Jsapi as given in Figure.3.43 (a). These files can be download from Google or can be linked directly to web by using http:// link. Figure.3.43 (b). shows the GET function which is used to get the CSV file mention the name of CSV file in this line. GET function transforms the data of CSV file into array to represent in the html page.

```

1 <!DOCTYPE html>
2 <head>
3 <meta charset="utf-8">
4 <meta http-equiv="X-UA-Compatible" content="IE=edge">
5 <meta name="viewport" content="width=device-width, initial-scale=1">
6 <!-- The above 3 meta tags *must* come first in the head; any other head content must come *after* these -->
7 <!-- http://t.co/dKFP3o1e -->
8 <meta name="HandheldFriendly" content="True">
9 <meta name="MobileOptimized" content="320">
10
11 <title>Google Graph and CSV</title>
12 <meta name="description" content="test">
13
14 <script src="http://ajax.googleapis.com/ajax/libs/jquery/2.1.4/jquery.min.js"></script>
15 <script src="jquery.csv.min.js"></script>
16 <script type="text/javascript" src="http://www.google.com/jsapi"></script>
17
18 <script type="text/javascript"> // load the visualisation API
19 google.load('visualization', '1', { packages: ['corechart', 'controls'] });
20 </script>
21
22 <script type="text/javascript">
23 function drawVisualization() {
24 //alert("function executes");
25 $.get("query_result.csv", function(csvString) {
26 // transform the CSV string into a 2-dimensional array
27 //alert("csv has been received");
28 var arrayData = $.csv.toArrays(csvString, {onParseValue: $.csv.hooks.castToScalar});
29

```

Figure 3.43: HTML page code 1

Figure.3.44 (a). shows the google.load function which is used to load all basic controls and packages of Google library to implement the visualization. Figure.3.44 (b) shows the drawVisualization function. It is user defined function in which the CSV file data transforms into a javascript array and then this data-array populated in data variable as given in Figure.3.44 (c). data variable populate Google's data table which is used to draw charts.

Google API provide a large variety of charts like column, bar, pie, line ,bubble and many others according to data type. In tis research the column chart is drawn to visualize the data. If it is required to change the chart type it can be mention in the code line as given in Figure.3.45 (a). Figure.3.45 (b) code line is just simply a function call to drawVisualization function. At the end there is simple html code which forms the page body and contains a simple div tag as shown in Figure.3.45 (c). The Google chart is drawn in the div area.

```

17
18 <script type="text/javascript"> // load the visualisation API
19 google.load('visualization', '1', { packages: ['corechart', 'controls'] });
20 </script>
21
22 <script type="text/javascript">
23 function drawVisualization() {
24 //alert("function executes");
25 $.get("query_result.csv", function(csvString) {
26 // transform the CSV string into a 2-dimensional array
27 //alert("csv has been received");
28 var arrayData = $.csv.toArrays(csvString, {onParseValue: $.csv.hooks.castToScalar});
29
30
31 // this new DataTable object holds all the data
32 var data = new google.visualization.arrayToDataTable(arrayData);
33 // CAPACITY - En-route ATFM delay - YY - CHART
34 var crt_ertdlyYY = new google.visualization.ChartWrapper({
35 chartType: 'ColumnChart',
36 containerId: 'crt_ertdlyYY',
37 dataTable: data,
38 options: {
39 width: 450, height: 160,
40 title: 'EU-wide en-route ATFM delays (year to date)',
41 titleTextStyle: {color: 'grey', fontSize: 11},
42 }
43 });
44 crt_ertdlyYY.draw();

```

Figure 3.44: HTML page code 2

```

32 var data = new google.visualization.arrayToDataTable(arrayData);
33 // CAPACITY - En-route ATFM delay - YY - CHART
34 var crt_ertdlyYY = new google.visualization.ChartWrapper({
35   chartType: 'ColumnChart',
36   containerId: 'crt_ertdlyYY',
37   dataTable: data,
38   options: {
39     width: 450, height: 160,
40     title: 'EU-wide en-route ATFM delays (year to date)',
41     titleTextStyle: {color: 'grey', fontSize: 11},
42   }
43 });
44 crt_ertdlyYY.draw();
45 });
46 }
47 google.setOnLoadCallback(drawVisualization);
48
49
50
51 </script>
52 <html>
53 <body>
54
55 <div id="crt_ertdlyYY"></div>
56
57 </body>
58 </html>

```

Figure 3.45: HTML page code 3

Now to load this page and view the chart it is necessary to load this page from a server. For this purpose xampp server has been used. Open the default browser of your windows and type the following URL *http://localhost/PingER-visualization/index.html* and the column chart is appeared as given in Figure.3.46.

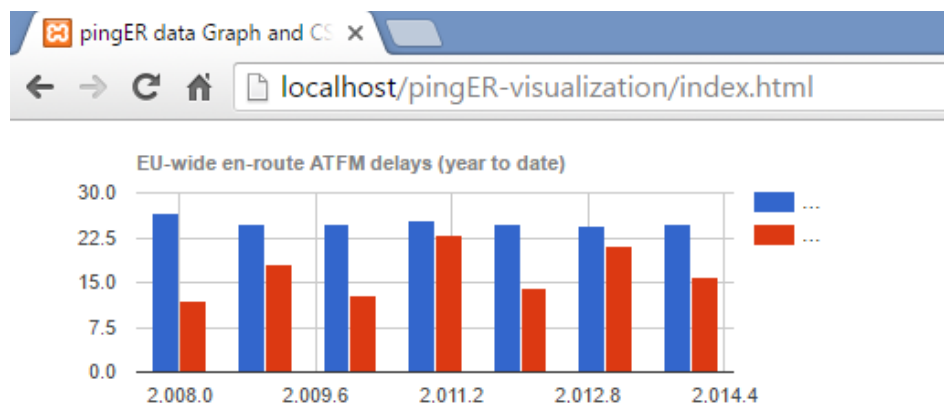


Figure 3.46: Column chart of PingER data

The overall system implementation and tools has been discussed in this chapter. First half portion of this chapter explains the working of MR and HDFS and defines Hadoop cluster. MR is a java program which includes two functions map function maps the data set according to given schema and reduce function reduces the data files. HDFS provides a storage architecture and store files as distributed blocks. Impala provides an interactive interface to run Impala queries. The jobs and roles of Hadoop cluster has been also explained in Section 3.1.3.

Section 3.3 covers overall installation process. Firstly, it is necessary to install Scicumulus MR workflow. Because MR is a java program the installation of Scicumulus requires JRE and MPJ installation. After Scicumulus cloudera was installed and CM was used to start all services of cloudera. To upload files on HDFS first data directories was created by using terminal and then files are uploaded by using Hue. At the end Hue was used to type and execute Impala queries. The last step is to install Xampp server and using Google APIs to display information in pictorial form.

The installation process is quiet lengthy but simple enough to understand. After the system implementation completed once the users can run different queries according to their data requirements. DWH contains huge amount of data and many analytical queries can be performed through Impala. Results of queries are displayed in cloudera professionals can use these results and general users can access these results on a simple web page after visualization process.

Chapter 4

RESULTS AND DISCUSSION

In this chapter overall implementation of system and results are discussed. As it has been discussed above that in our research process output from one step is used in as input to next step. But here the combine results and effects of this research study are discussed. Further the testing techniques are applied to test the performance of system and execution time of queries was recorded.

4.1. Testing and Evaluation

Test and Evaluation is the method by which a componet or segments are checked to match basic requirements and details by applying testing. The results are analyzed to survey the factors like execution, enhanced outline, supportability etc. Formative test and assessment is a designed tool used to decrease all errors which are occurring in a development lifecycle. Functional testing and assessment is the actual or propagate business, by regular clients, of a developed component under applicable functional conditions.

Testing is an procedure to assure the quality of an item, developed model, or ability (e.g., right product, produced right). To make right product, testing cannot happen only just toward the end of an advancement process. It must be occur persistently in all phases of the whole life cycle. Test and Evaluation includes assessing an item from the segment level, to remain solitary framework, coordinated framework, and, if proper, arrangement of-framework and venture (Stevens *et al.*, 2016). The evaluation levels and how the levels are work along with government DT, OT, and accreditation and certification testing are given below in Figure 4.1.

4.1.1. Unit Testing

The basic purpose of unit testing is to take the small piece of testable programming in the application, separate it from whatever is left of the code, and make sense of in the event that it demonstrations absolutely as you foresee. Each unit is attempted autonomously before planning them into modules to test the interfaces between modules. Unit testing has shown its value in that an immense rate of disfigurements are perceived in the midst of its usage. The most broadly perceived approach to manage unit testing requires drivers and stubs to be made.

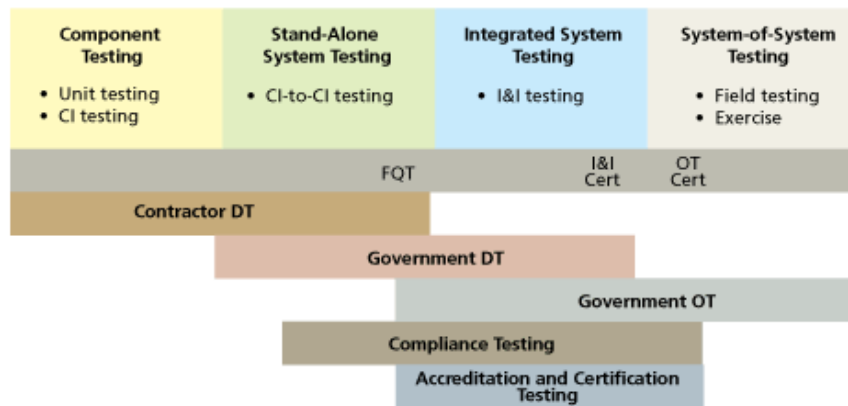


Figure 4.26: Testing Techniques (Stevens, 2016)

4.1.2. Integration Testing

Integration testing is a method to test how parts of the framework cooperate. Incorporation tests are like unit tests, however there's one major contrast while unit tests are detached from different parts, joining tests are most certainly not. For instance, a unit test for database access code would not converse with a genuine database, but rather a joining test would. Coordination testing is principally valuable for circumstances where unit testing is insufficient. Some of the time you need tests to confirm that two separate frameworks like a database and your application how they cooperate, and that requires a coordination test.

4.1.3. Component Testing

Component testing is a strategy where testing of every part in an application is done independently. Assume, in an application there are 5 segments. Testing of every 5 parts independently and proficiently is called as component testing. Component testing is otherwise called module and project testing. It finds the imperfections in the module and checks the working of programming. Part testing is finished by the analyzer. Segment testing might be done in confinement from rest of the framework relying upon the advancement life cycle model decided for that specific application.

4.1.4. Comparison

A comparison between testing techniques is given in Table 4.1.

Table 4.4: Comparison between testing techniques

Validation Strategy	Explanation
Unit Testing	Testing of single program
Integration Testing	Testing of related programs
System Testing	Testing of entire system
Performance Testing	Testing for performance of the application

4.1.5. Testing Usecases

As above the testing techniques has been discussed in this research unit testing is carried out as the data mining process divided in multiple steps and the output from step is injected as input in second step so it is necessary to test the performance of each step. In this testing usecases are defined which has been tested one by one in results section and output results has been explained. Testing usecases are depicted in Figure.4.2.

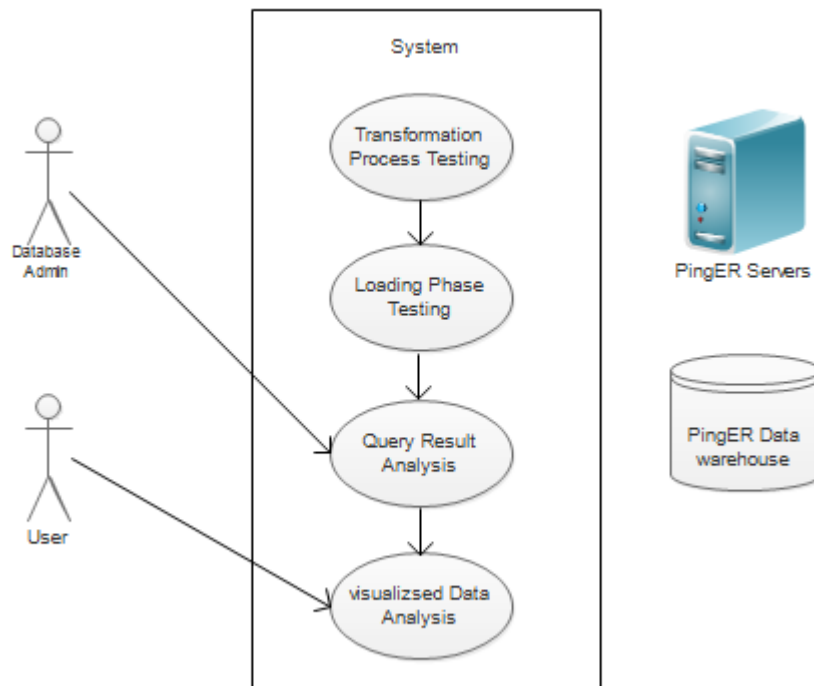


Figure 4.27: Testing Usecase Diagram

4.2. Results

The results section is divided according to testing usecases. As the research process has been explained above and depicted in Figure.3.2. And testing usecases are depicted in Figure.4.2. The results of each testing phase is given below.

First phase selection and cleaning will not be discussed here as any cleaning procedure on data was not applied. The refined granularity data from PingER servers is downloaded and processed by passing it to next phase.

4.2.1. Transformation

For transformation process our required output is the Comma Separated Value (CSV) files. The performance of MR program is also need to be tested and the data in the CSV file that data is according to defined dimensional model or not. For transformation process MR data flow is carried out. The mapper activity transformed over 100,000 flat files into the same number of CSV files, but each of which followed the designed dimensional data model. The reducer activity combined those transformed files into 17 large text files, summing 45 Gigabytes of transformed files. The entire transformation process took 6h 12 min to run (De Oliveira *et al.*, 2010).

4.2.2. Loading

After that, the CSV files were loaded into HDFS (Cloudera HDFS, 2016). The data directories were created in the HDFS using the Hadoop commands via Linux command line. The fact table was partitioned in 16 files, one file for each year. One directory was created for the DW dimensions (one subdirectory for each dimension) and one directory was created for the facts (one subdirectory for each partition). After that, the generated data was also loaded into the file system using the following commands:

```
hdfs dfs -mkdir pinger
hdfs dfs -mkdir pinger/country
hdfs dfs -mkdir pinger/hosts
hdfs dfs -mkdir pinger/nodes
hdfs dfs -mkdir pinger/rtdtdata
```

Each file took 32 seconds on average to be uploaded to HDFS.

4.2.3. Querying Data

First it is required to create external tables in impala then the data from CSV files is loaded into impala tables. An external table uses arbitrary HDFS directories, where the data files are typically shared between different Hadoop components. It works as a link

between Impala and HDFS. The external tables took only 0.28 seconds on average to be created on Impala using statements, such as:

```
create external table pinger_csv_country (country_ID int,  
    country string, continent string, TLD string, REMARK string)  
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
    LOCATION '/user/cloudera/pinger/country';  
create external table pinger_csv_hosts (host_id int,  
    host_source string, host_destination string)  
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
    LOCATION '/user/cloudera/pinger/hosts';  
create external table pinger_csv_nodes (node_id int,  
    nodename string, ipaddress string, sitename string,  
    nickname string, fullname string, location string,  
    country string, country_ID int, latandlong string,  
    projecttype string, pingserver string, traceserver string,  
    dataserver string, url string, gmt string, comments string,  
    appuser string, contacts string, ping_size string, misc string)  
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
    LOCATION '/user/cloudera/pinger/nodes';  
create external table pinger_csv_rttdata(source string,  
    destination string, h01 float, h02 float, h03 float, h04 float,  
    h05 float, h06 float, h07 float, h08 float, h09 float, h10 float,  
    h11 float, h12 float, h13 float, h14 float, h15 float, h16 float,  
    h17 float, h18 float, h19 float, h20 float, h21 float, h22 float,  
    h23 float, h24 float, source_2 string, destination_2 string)  
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
    LOCATION '/user/cloudera/pinger/rttdata';
```

After the external tables has been created, the Impala tables were created using Parquet format. Parquet is as a column-oriented binary file format that was created to be highly efficient for large-scale queries (Cloudera Impala, 2015). This fits exactly with what is required in our analytical environment. After this the data from CSV files to external tables is loaded by using Load Data statement such as:

```
LOAD DATA INPATH 'hdfs://quickstart.cloudera/user/cloudera/pinger/country'  
    INTO TABLE pinger_csv_country;
```

```

LOAD DATA INPATH 'hdfs://quickstart.cloudera/user/cloudera/pinger/country'
    INTO TABLE pinger_csv_hosts;
LOAD DATA INPATH 'hdfs://quickstart.cloudera/user/cloudera/pinger/country'
    INTO TABLE pinger_csv_nodes;
LOAD DATA INPATH 'hdfs://quickstart.cloudera/user/cloudera/pinger/country'
    INTO TABLE pinger_csv_rttdata;

```

Inserting data into impala tables took 6.75 seconds on average. After inserting data finally the queries are applied over the data. The results of three queries has been shown here moreover the user can run many other queries according to his data requirements.

Query 1.

```

SELECT destination,h01 FROM pinger_csv_rttdata WHERE source =
'pinger.stanford.edu';

```

The results of query 1 are given in Figure 4.3.

Query 2.

```

SELECT destination,h22 FROM pinger_csv_rttdata WHERE source =
'pinger.sprace.org.br';

```

The results of query 2 are given in Figure 4.4.

Query 3.

```

SELECT destination,h22 FROM pinger_csv_rttdata WHERE source =
'ping.riken.jp';

```

The results of query 3 are given in Figure 4.5.

	destination	h01
0	ping.miki.kfki.hu	215.9949951171875
1	www.afe.mr	252.4680023193594
2	brunsvigia.tenet.ac.za	391.1820068359375
3	elitca1.epfl.ch	195.9219970703125
4	kadri.ut.ee	202.32499694824219
5	www.gov.bw	339.62600708007812
6	www-05.nexus.ao	352.364990234375
7	pinger.unesp.br	222.51699829101562
8	www.kazrena.kz	281.86300659179688
9	pingermtu.pern.edu.pk	298.12296583984375
10	ninnes.unimas.my	756.79000954497188

Figure 4.28: Results of Impala Query 1

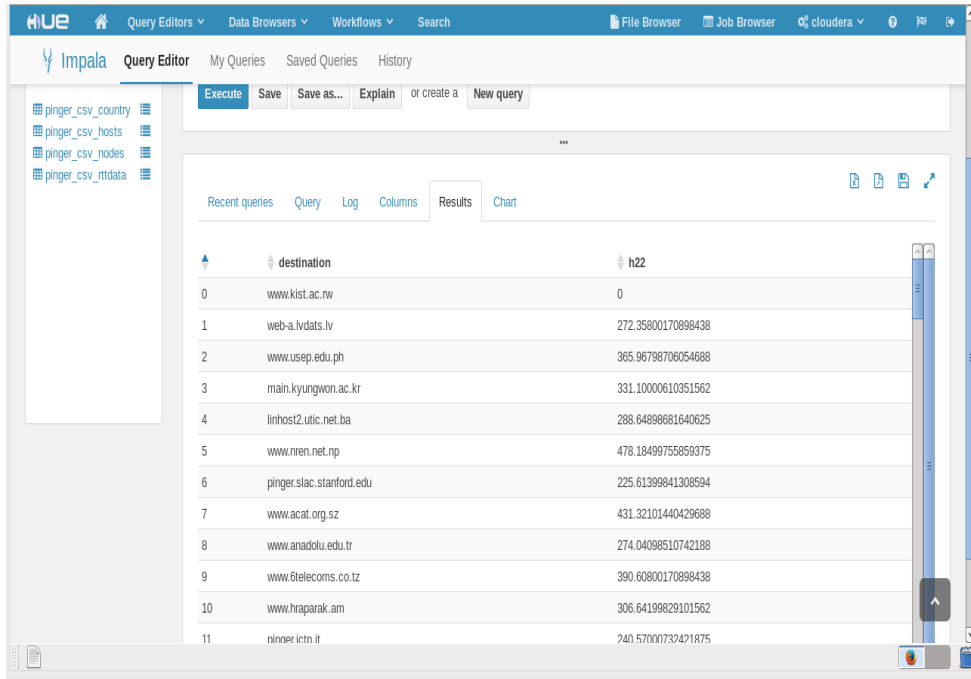


Figure 4.29: Results of Impala Query 2

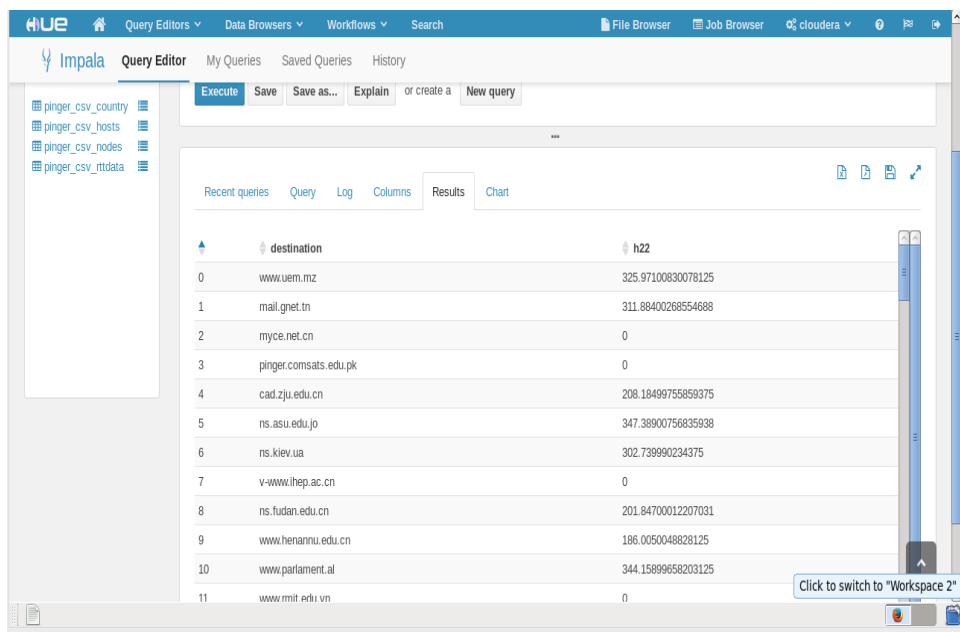


Figure 4.30: Results of Impala Query 3

4.2.4. Visualization

The results after querying data are not publically accessible. To make it publically accessible and display the information to users, visualization process was applied by using Google charts. The results of 3 queries are exported into CSV files and pass to the Google library (Google Charts, 2016). Figure 4.6, 4.7, 4.8 shows the visualization Bar charts of three queries respectively.

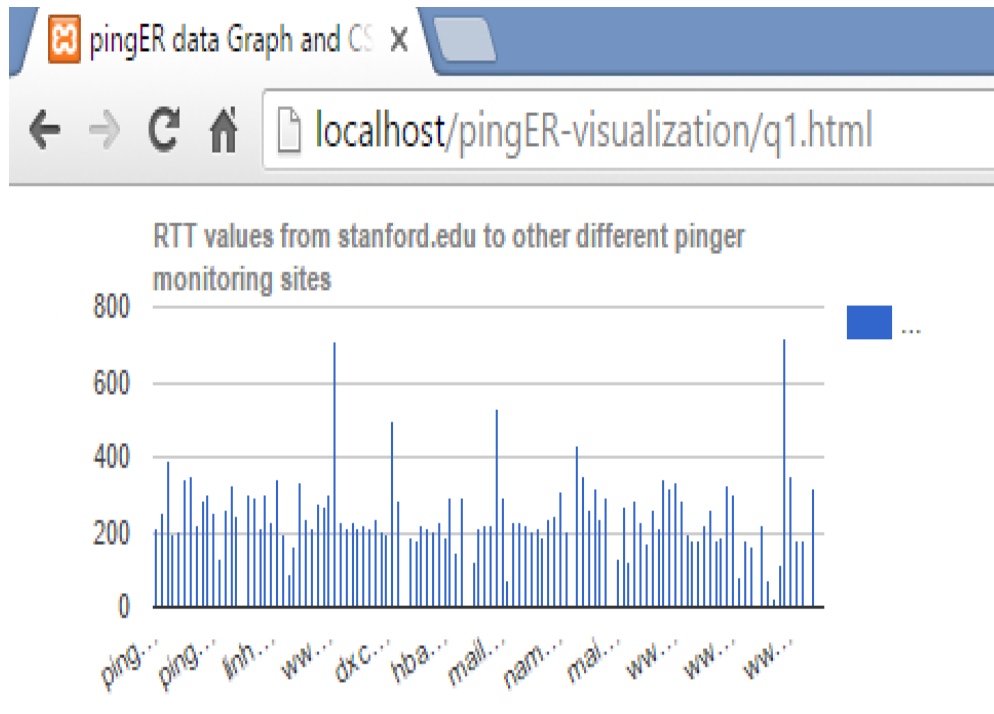


Figure 4.31: Visualization Bar Chart of Query 1

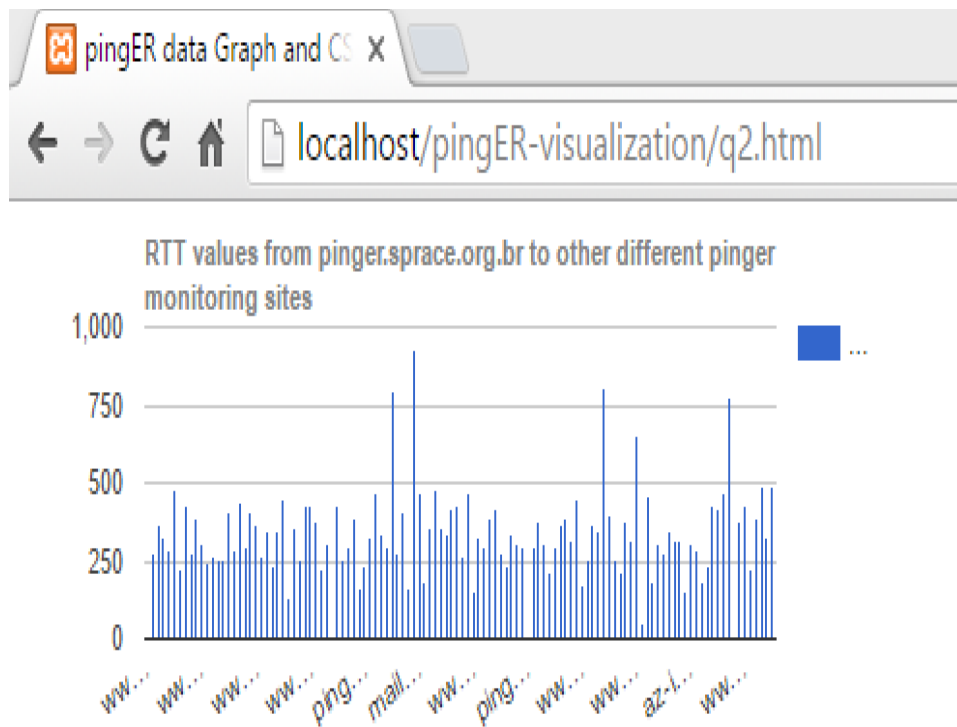


Figure 4.32: Visualization Bar Chart of Query 2

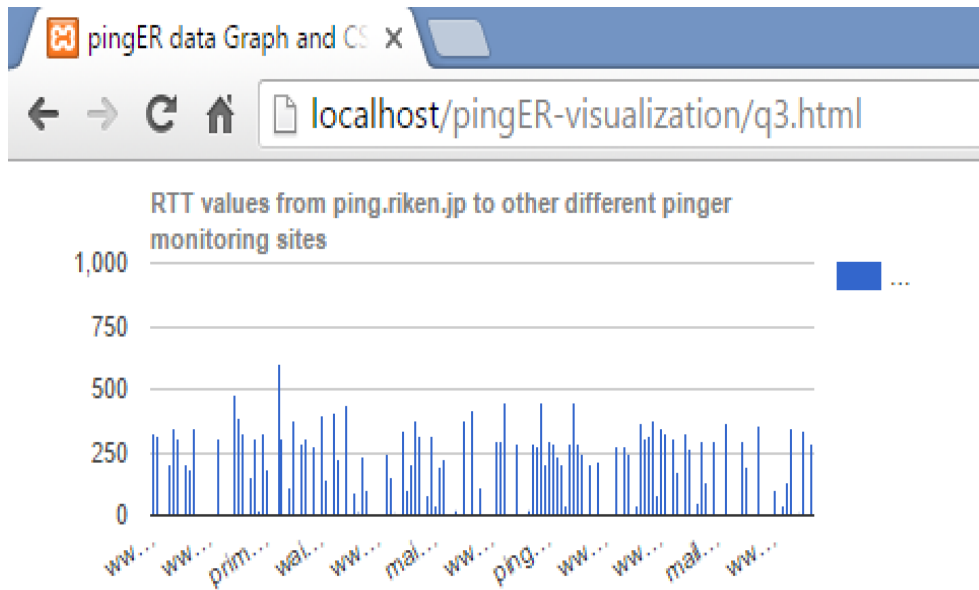


Figure 4.33: Visualization Bar Chart of Query 3

Figure 4.9, 4.10, 4.11 shows the visualization Line charts of three queries respectively.

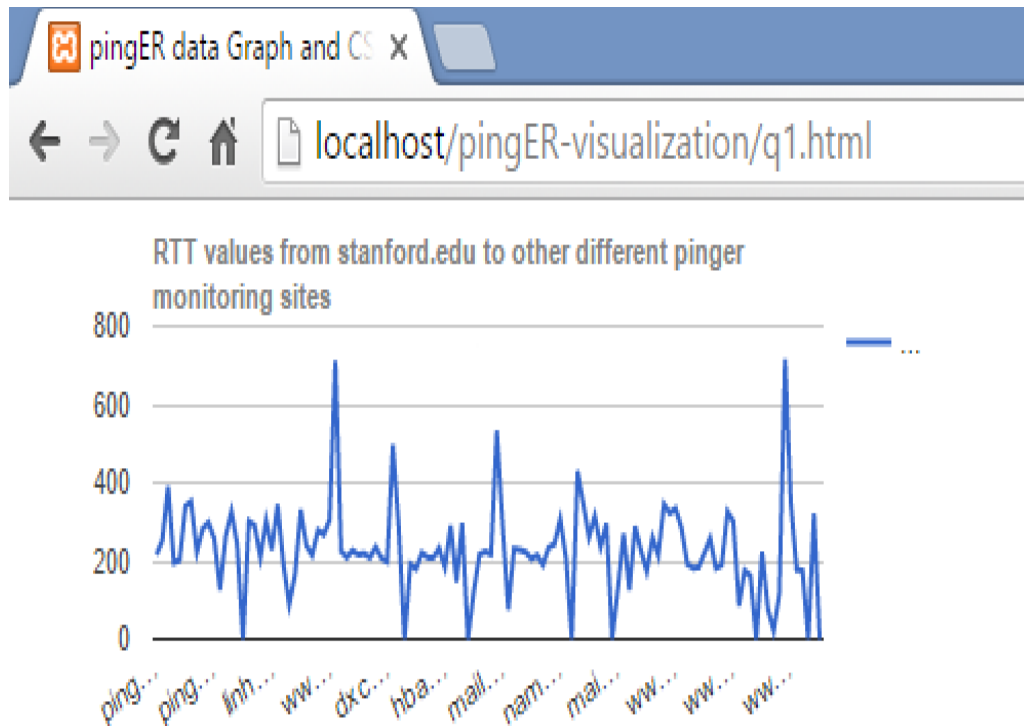


Figure 4.34: Visualization Line Chart of Query 1

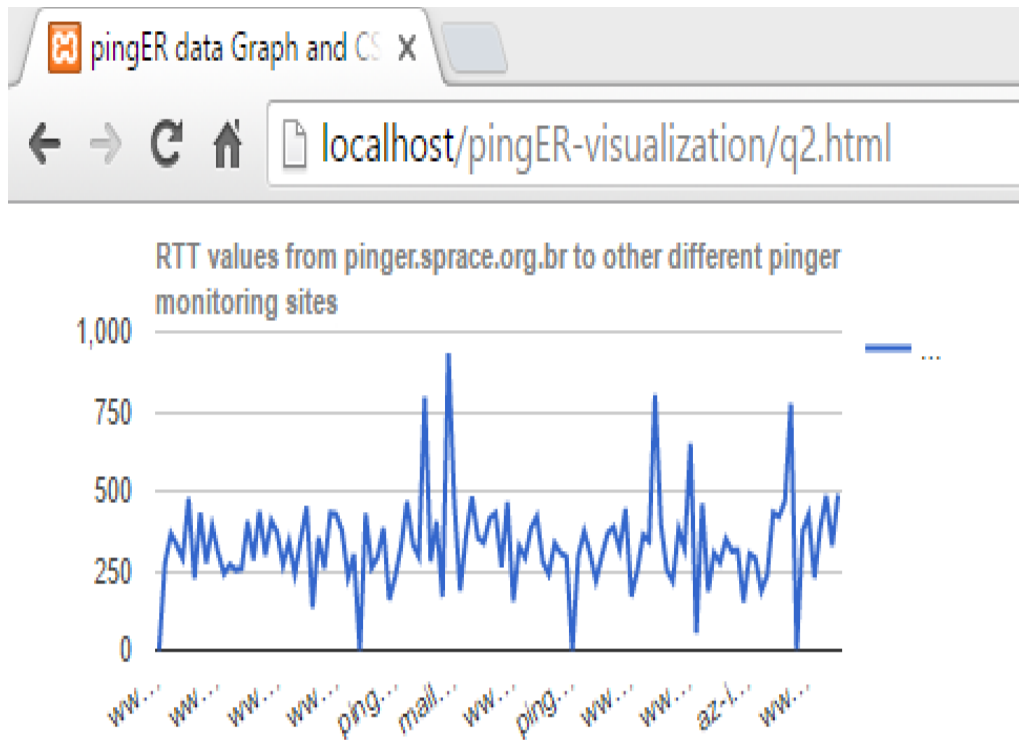


Figure 4.35: Visualization Line Chart of Query 2

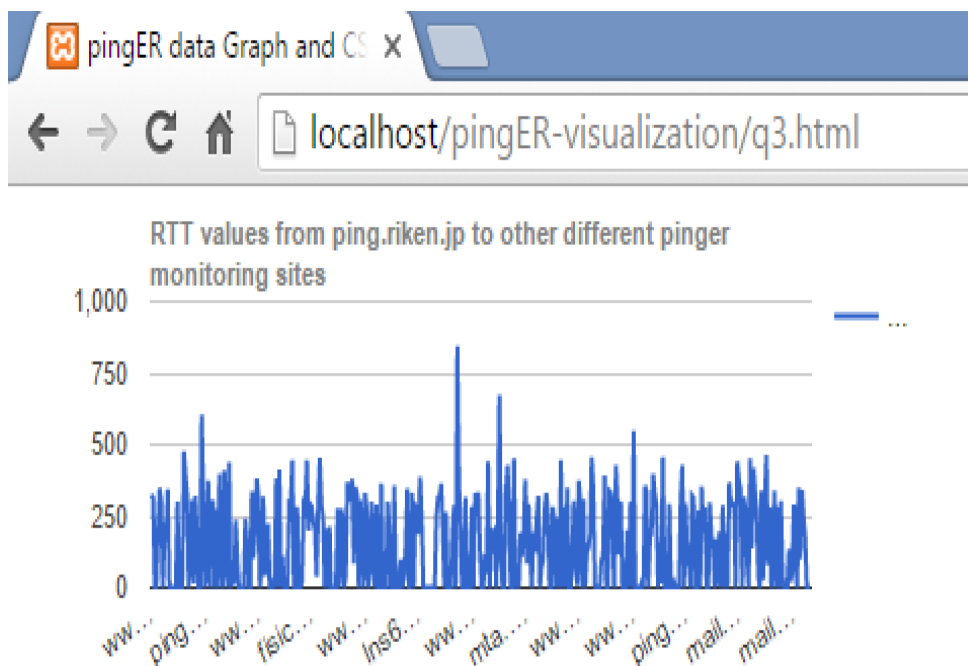


Figure 4.36: Visualization Line Chart of Query 3

4.3. Discussions

With the rapid growth of data it has become difficult to store and process this data. As organization's requirements are increasing towards the analysis of data because processing of data can reveal interesting information and helps in reporting and effective decision making which supports the business organizations. One more question arises why it is necessary to store a lot of data if it cannot provide us any useful information. For this purpose data warehouses and data mining are used widely for processing of data.

But when the era of Big Data arrived these traditional techniques are not suitable for processing and cannot provide effective results. Big Data involves large volume of data, arrive from heterogeneous sources and complex in nature. Cloud computing, Internet of Things (IoT), data centers, and Hadoop are key technologies of Big Data (Chen *et al.*, 2014).

Despite of some benefits that are achieved in this research there are some limitations of system which needs to be focused. The designing of Data Warehouse requires pre-dimensional modeling like defining a schema and performing complex operations. Modeling is a time consuming task and when the updation is required it is necessary to update the schema design with the Data warehouse.

Despite the fact that business associations are the expected clients of Data Warehouses however they will in any case need to work with the IT experts on the grounds that the displaying undertakings require an intricate demonstrating methodology and clients need to compose an incredible quantities of codes/scripts/SQL. A few operations like selecting a typical scientific model, building map measurements of model to the names of fields, tables, and perspectives of the business database and performing ETL process can't be finished without the help of a professional.

There is also a limitation under the used tool cloudera. When the results was transformed and impala queries was executed the processed output can be viewed by using cloudera. But users cannot interact with the tools directly so visualize this results Google Chart API was used. But Google chart API cannot be integrated with Impala because Impala Hadoop distributed Big Data Platform. So to visualize it is necessary to export the query results into a CSV file and then it can Integrate the data with Google Chart API.

Our basic purpose of this research was to store the PingER raw data on big data architecture and process to derive useful information. Despite of these limitations the proposed approach provides a solution to the defined problem. The work presented in this research showed to be very scalable, making the solution capable of dealing with even more data. The main advantage of the proposed approach is that it provides the facility to search the PingER database and filter on ways that the original PingER project was not capable of. The previous solution was about to store PingER data in RDBMS which lacks the scalability and efficiency of PingER data. The further PingER project is lacking to display this information for general public and make it accessible to all users. The solution to this problem is given by representing PingER data in graphical form to display the results and information to general users in a more understandable format.

Summary

In the IT world as new generations have moved towards the emerging technologies like web, social media, smart phones and cloud. Many business organizations started using these technologies to expand their business. This results in generation of lots of data. Data is an important source for every organization. It involves useful information and key factors to improve business performance by predicting and forecasting future values. And now with the growing data it has become difficult to manage and store this data with traditional databases and this data requires new technologies to process data. As user requirements for data processing has been increased the new emerging technologies are data warehouses and Big Data technologies now a days. With the increased usage of web a lot of multidimensional data and unstructured data has been generated. This arose a new concept of data mining. The basic concept of data mining arrived from data warehouses where multidimensional data stored and used for analysis and reporting purposes. This data can be processed by using many data mining techniques like classification, clustering, regression and association rules mining. But when Big Data was discussing these data mining techniques are not enough. Big Data refers to large volume, variety and velocity in data. It involves complex relationships and difficult to process. To store big data a new architecture is required, which is HDFS and processing framework MR instead of data mining. PingER is a project starting from 1996 and led by SLAC laboratories. It uses a ping command to measure internet end-to-end performance. The results of RTT values are stored in flat files and aggregated data operations are applied to these RTT values and stored in smaller data files. But as data size grows, it is difficult to store and manage data with flat files and RDBMS. With proposed RDBMS solution PingER data lack the efficiency and scalability of data. Now finger is monitoring the performance of over 700 websites, therefore PingER has generated a lot of data and requires processing of this data. This research supports processing of data by using Data Mining and Big data technologies. A data warehouse for the PingER project has been created. These data files are processed by using MR framework and store the data in HDFS architecture. To retrieve the processed data Impala queries are applied. The results from impala queries are simple enough, but it cannot be displayed to users. To present this useful information to user visualization techniques like Bar chart and Line chart is drawn on the query results. This research provides an effective approach to process PingER data and the results of this research concluded that the

processing of data can reveal interesting information for finger project. It provides information like power shortcuts, server bottlenecks and the management can use this information for making an effective decision making and forecasting future predictions. This research also provides a manageable way to store data and the PingER management can store and retrieve data easily. After retrieval of data it is also available in pictorial form so the users can understand it more easily.

LITERATURE CITED

- Adamov, A., 2014. Data mining and analysis in depth. case study of Qafqaz University HTTP server log analysis. *2014 8th IEEE International Conference on Application of Information and Communication Technologies (AICT)*, 1-4 October 2014, Astana, Kazakhstan.
- Adamu, F. B., A. Habbal, S. Hassan, R. L. Cottrell, B. White, and I. Abdullahi, 2015. A Survey On Big Data Indexing Strategies. *4th International Conference on Internet Applications, Protocols and Services (NETAPPS)*, December 2015, Putrajaya, Malaysia.
- Bai, X., D. White, and D. Sundaram, 2013. Context adaptive visualization for effective business intelligence. *2013 15th IEEE International Conference on Communication Technology*, 786–790 November 2013, Guilin, China.
- Bal, J., S. Lashari, and S. M. Nabavieh, 2014. *Making successful virtual clusters*. WMG, University of Warwick, 1-27 July 2014, Coventry, England.
- Barbosa, T. M. S., R. Souza, S. M. S. Cruz, M. L. Campos, and R. L. Cottrell, 2015. Applying Data Warehousing and Big Data Techniques to Analyze Internet Performance. *4th International Conference on Internet Applications, Protocols and Services (NETAPPS)*, 31–36 December 2015, Putrajaya, Malaysia.
- Ben Ayed, A., M. Ben Halima, and A. M. Alimi, 2014. Survey on clustering methods: Towards fuzzy clustering for big data. *6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 331–336 August 2014, Tunis, Tunisia.
- Borthakur, D., 2008. HDFS Architecture Guide [Online]. Available at https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf (June, 2008).
- Bresnahan, J., D. Labissoniere, T. Freeman, and K. Keahey, 2011. Cumulus : An Open Source Storage Cloud for Science. *2nd Workshop on Scientific Cloud Computing (ScienceCloud 2011)*, November 2011, San Jose, CA.
- Bronson, J., B. Summa, J. Freire, V. Pascucci, and C. T. Silva, 2011. Parallel Visualization on Large Clusters using Map Reduce. *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, 81–88 October 2011, Providence, Rhode Island.
- Bu, X., J. Rao, and C. Xu, 2013. Interference and locality-aware task scheduling for MapReduce applications in virtual clusters. *Proceedings of the 22nd International Symposium on High-Performance Parallel and Distributed Computing (HPDC '13)*, 227-238 June 2013, New York, USA.
- Burbank, D., The 5 V's of Big Data [Online]. Available at <http://enterprisearchitects.com/the-5v-s-of-big-data/> (May, 2016).
- Chang, Y., 2013. A realtime interactive visualization system for knowledge transfer from social media in a big data. *2013 9th International Conference on Information,*

- Communications & Signal Processing (ICICS)*, 1–5 December 2013, Tainan, Taiwan.
- Chaudhuri, S., and U. Dayal, 1997. An overview of data warehousing and OLAP technology. *Special Interest Group on Management Of Data (SIGMOD) ACM*, 26(1): 65–74.
- Chen, C., 2011. Research on the visualization of Data Mining results. *2011 6th International Conference on Computer Science & Education (ICCSE)*, 938–941 August 2011, Singapore, Malaysia.
- Chen, M., S. Mao, and Y. Liu, 2014. Big Data: A Survey. *Springer Mobile Networks and Applications*, 19(2): 171-209.
- Cheng, Z., Z. Luan, Y. Meng, Y. Xu, D. Qian, A. Roy, ... G. Guan, 2012. ERMS: An Elastic Replication Management System for HDFS. *2012 IEEE International Conference on Cluster Computing Workshops*, 32–40 September 2012, Beijing, China.
- Chopade, P., J. Zhan, K. Roy, and K. Flurchick, 2015. Real-Time Large-Scale Big Data Networks Analytics and Visualization Architecture. [*2015 12th International Conference & Expo on Emerging Technologies for a Smarter World \(CEWIT\)*](#), 1-6 October 2015, Melville, New York.
- Ciubancan, M., G. Neculoiu, O. Grigoriu, I. Halcu, V. Sandulescu, M. Marinescu, and V. Marinescu, 2013. Data mining processing using GRID technologies. *11th RoEduNet International Conference*, 1–3 January 2013, Sinaia, Romania.
- Cloudera, 2016. CDH Overview [Online]. Available at http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cdh_intro.html (20 May, 2016).
- Cloudera, 2016. Cloudera Impala [Online]. Available at <http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html> (20 May, 2016).
- Cloudera, 2016. Using the Parquet File Format with Impala Tables [Online]. Available at http://www.cloudera.com/content/cloudera/en/documentation/cloudera-impala/latest/topics/impala_parquet.html (20 May, 2016).
- Cloudera, 2016. Cloudera Data Management [Online]. Available at <http://www.cloudera.com/documentation/enterprise/latest/topics/datamgmt.html> (20 May, 2016).
- Conejero, J., B. Caminero, and C. Carri, 2014. Analysing Hadoop Performance in a Multi-user IaaS Cloud. *International Conference on High Performance Computing and Simulation (HPCS)*, 399–406 July 2014, Bologna, Italy.
- Connolly, T., and C. Begg, 2005. Data Warehousing Concepts. P.1200-1212 *In Database Systems. Part 9 (4th Ed.)* Pearson Education Limited.

- Cottrell R. L., 2015. Tutorial on Internet Monitoring & PingER at SLAC [Online]. Available at <http://pinger.seecs.edu.pk/tutorial/tutorial.html> (2015, Jul 07).
- De Oliveira, D., E. Ogasawara, F. Baião, and M. Mattoso, 2010. SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific Workflows. *2010 IEEE 3rd International Conference on Cloud Computing*, 378–385 July 2010, Miami, Florida.
- Dou, D., H. Wang, and H. Liu, 2015. Semantic data mining: A survey of ontology-based approaches. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 244–251 February 2015, Anaheim, California.
- Elmasri R., and S. B. navathe, 2011. Data Mining Concepts. P.1035-1067 *In* Fundamentals of Database Systems. Part 11 (6th Ed.) Addison Wesley.
- Fan, W., and A. Bifet, 2013. Mining Big Data : Current Status , and Forecast to the Future. *Special Interest Group on Knowledge Discovery in Data (SIGKDD)*, 14(2): 1–5.
- Fu, Q., W. Liu, T. Xue, H. Gu, S. Zhang, and C. Wang, 2014. A BIG DATA PROCESSING METHODS FOR VISUALIZATION. *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, 571-575 November 2014, Shenzhen, China.
- Gaur, D., 2011. Data mining and visualization on legal documents. *2011 International Conference on Recent Trends in Information Systems*, 132–136 December 2011, Kolkata, India.
- Gohil, P., 2014. A Performance Analysis of MapReduce Applications on Big Data in Cloud based Hadoop. *2014 International Conference on Information Communication and Embedded Systems (ICICES)*, 2–7 February 2014, Chennai, India.
- Golab, L., and M. T. Ozsu, 2003. Issues in Data Stream Management *. *Special Interest Group on Management Of Data (SIGMOD) ACM*, 32(2): 5–14.
- Google Inc, 2016. Display live data on your site: About Google chart tools [Online]. Available at <https://developers.google.com/chart/> (20 May, 2016).
- Grover, M., 2014. The architectural design and features of Cloudera Impala [Online]. Available at <https://www.quora.com/What-is-the-architectural-design-and-features-of-Cloudera-Impala> (14 February, 2014).
- Gu, L., and H. Li, 2013. Memory or Time: Performance Evaluation for Iterative Operation on Hadoop and Spark. *2013 IEEE 10th International Conference on High Performance Computing and Communications and 2013 IEEE International Conference on Embedded and Ubiquitous Computing*, 721–727 November 2013, Zhangjiajie, China.

- Gu, R., X. Yang, J. Yan, Y. Sun, B. Wang, C. Yuan, and Y. Huang, 2014. SHadoop: Improving MapReduce performance by optimizing job execution mechanism in Hadoop clusters. *Journal of Parallel and Distributed Computing*, 74(3): 2166–2179.
- Gupta, D., and S. Siddiqui, 2014. BIG DATA IMPLEMENTATION AND VISUALIZATION. [*2014 International Conference on Advances in Engineering and Technology Research \(ICAETR\)*](#), 1-10 August 2014, Unnao, India.
- Harter, T., W. Madison, D. Borthakur, S. Dong, A. Aiyer, L. Tang, ... S. Clara, 2014. Analysis of HDFS Under HBase : A Facebook Messages Case Study. *Proceedings of the 12th USENIX Conference on File and Storage Technologies (FAST 14)*, 199-212 February 2014, Santa Clara, California.
- Hedlund, B., 2011. Understanding Hadoop Clusters and the Network [Online]. Available at <http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/> (10 September, 2011).
- Hlosta, M., M. Sebek, and J. Zendulka, 2013. Approach to visualisation of evolving association rule models. *2013 Second International Conference on Informatics & Applications (ICIA)*, 47–52 September 2013, Lodz, Poland.
- Indarto, E., 2013. Data Mining [Online]. Available at <http://recommender-systems.readthedocs.io/en/latest/datamining.html> (5 July, 2013).
- Janciak, I., M. Lenart, P. Brezany, L. Novakova, and O. Habala, 2011. Visualization of the Mining Models on a Data Mining and Integration Platform. *Proceedings of the 34th International Convention MIPRO*, 215–220 May 2011, Opatija, Croatia.
- Keim, D. A. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1): 1–8.
- Kim, K. T., W. S. Seol, U. M. Kim, and H. Y. Youn, 2014. Latent Semantic Analysis for Mining Rules in Big Data Environment. *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 193–200 October 2014, Shanghai, China.
- Kimball, R., and M. Ross, 1998. Dimensional Modeling Techniques Overview. P. 37-68 *In The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses Architecture*. (3rd Ed.).
- Krishnan, K., 2013. Big Data Processing Architectures. P.29-42 *In Data Warehousing in the Age of Big Data*. Part 1 (2nd Ed.) Elsevier
- Leverich, J., and C. Kozyrakis, 2010). On the energy (in)efficiency of Hadoop clusters. *ACM SIGOPS Operating Systems Review*, 44(1): 61-65.
- Liao, S.-H., P.-H. Chu, and P.-Y. Hsiao, 2012. Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12): 11303–11311.

- Liu, Q., B. Ribeiro, A. H. Sung, and D. Suryakumar, 2014. Mining the Big Data: The Critical Feature Dimension Problem. *IIAI 3rd International Conference on Advanced Applied Informatics*, 499–504 September 2014, Kitakyushu, Japan.
- Nabi, 2011. [Implementation of Relational archive site for PingER](https://confluence.slac.stanford.edu/display/IEPM/Implementation+of+Relational+archive+site+for+PingER) [Online]. Available at <https://confluence.slac.stanford.edu/display/IEPM/Implementation+of+Relational+archive+site+for+PingER> (26 July, 2011).
- Ngo, L., V. Dantuluri, M. Stealey, S. Ahalt, and A. Apon, 2012. An Architecture for Mining and Visualization of U.S. Higher Educational Data. *2012 Ninth International Conference on Information Technology - New Generations*, 783–789 April 2012, Las Vegas, Nevada.
- Pal, A., K. Jain, P. Agrawal, and S. Agrawal, 2014. A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop. *2014 Fourth International Conference on Communication Systems and Network Technologies*, 587–591 April 2014, Bhopal, India.
- Pavlo, A., E. Paulson, A. Rasin, D. J. Abadi, S. Madden, M. I. T. Csail, D. J. Dewitt, 2009. A Comparison of Approaches to Large-Scale Data Analysis. *Special Interest Group on Management Of Data (SIGMOD) ACM*, 165-178.
- Perrot, A., R. Bourqui, N. Hanusse, F. Lalanne, and D. Auber, 2015. Large interactive visualization of density functions on big data infrastructure. *2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV)*, 99–106 October 2015, Chicago, Illinois.
- Purwar, A., and S. K. Singh, 2014. Issues in data mining: A comprehensive survey. *IEEE International Conference on Computational Intelligence and Computing Research*, 1–6 December 2014, Coimbatore, India.
- Qi, Y., and X. Yu, G. Shi, Y. Li, 2015. Visualization in Media Big Data Analysis, [2015 IEEE/ACIS 14th International Conference on Computer and Information Science \(ICIS\)](#), 571–574 July 2015, Las Vegas, Nevada.
- Qiang, X., Y. Wei, and Z. Hanfei, 2010. Application of Visualization Technology in Spatial Data Mining. *International Conference on Computing, Control and Industrial Engineering*, 153–157 June 2010, Wuhan, China.
- Silipo R., I. Adae, A. Hart, M. berthold, 2015. Seven Techniques for Dimensionality Reduction [Online]. Available at <https://www.knime.org/blog/seven-techniques-for-data-dimensionality-reduction> (12 May, 2015).
- Sachin, R. B., and M. S. Vijay, 2012. A Survey and Future Vision of Data Mining in Educational Field. *Second International Conference on Advanced Computing and Communication Technologies*, 96–100 January 2012, Rohtak, India.
- SciCumulus, 2016. [SciCumulus/C2 – Parallel Scientific Workflow Management System](#) [Online]. Available at <https://scicumulusc2.wordpress.com/starter-guide-2/> (20 May, 2016).

- Souza, 2014. [PingER Linked Open Data \(PingERLOD\) overview](https://confluence.slac.stanford.edu/display/IEPM/PingER+Linked+Open+Data+(PingERLOD)+overview) [Online]. Available at [https://confluence.slac.stanford.edu/display/IEPM/PingER+Linked+Open+Data+\(PingERLOD\)+overview](https://confluence.slac.stanford.edu/display/IEPM/PingER+Linked+Open+Data+(PingERLOD)+overview) (19 February, 2014).
- Stevens, 2007. Test and Evaluation [Online]. Available at <https://www.mitre.org/publications/systems-engineering-guide/se-lifecycle-building-blocks/test-and-evaluation> (1 september, 2007).
- Summer, E., and D. L. Ali, 1996. A practical guide for implementing data warehousing. *Computers & Industrial Engineering*, 31(2): 307–310.
- Tekiner, F., and J. A. Keane, 2013. Big data framework. *2013 IEEE International Conference on Systems Man and Cybernetics*, 1494–1499 October 2013, Manchester, England.
- Thusoo, A., J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, ... R. Murthy, 2010. Hive - a petabyte scale data warehouse using Hadoop. *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 996–1005 March 2010, Long Beach, California.
- Thusoo, A., Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sen Sarma, ... H. Liu, 2010. Data warehousing and analytics infrastructure at facebook. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 1013-1020 June 2010, Indiana, USA.
- Tsumoto, S., S. Hirano, and Y. Tsumoto, Y. 2011. Visualization of Hospital Services Using Data Mining Methods. *2011 IEEE 11th International Conference on Data Mining Workshops*, 1183–1190 December 2011, Vancouver, BC.
- Vijayakumari, R., R. Kirankumar, and K. G. Rao, 2014. Comparative analysis of Google File System and Hadoop Distributed File System. *International Journal of Advanced Trends in Computer Science and Engineering*, 3(1): 553–558.
- Vmware Player, 2016. VMware Workstation Player (formely known as Player Pro) [Online]. <https://www.vmware.com/products/player#sthash.OUv7Thx3.ZujSPdYm.dpuf> (20 May, 2016).
- Wang, L., J. Tao, R. Ranjan, H. Marten, A. Streit, J. Chen, and D. Chen, 2013. G-Hadoop: MapReduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems*, 29(3): 739–750.
- Wu, X., X. Zhu, and S. Member, 2014. Data Mining with Big Data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 26(1): 97–107.
- Xie J., S. Yin, X. Ruan, Z. Ding, Y. Tian, J. Majors, A. Manzanares, and X. Qin (2010). Improving MapReduce performance through data placement in heterogeneous Hadoop clusters. *2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)*, 1–9 April 2010, Atlanta, Georgia.

Zaharia, M., M. Chowdhury, T. Das, and A. Dave, 2012. Fast and Interactive Analytics over Hadoop Data with Spark. *USENIX ;login.*, 37, 45–51.

Zhang, J., and M. L. Huang, 2013. 5Ws Model for Big Data Analysis and Visualization. *2013 IEEE 16th International Conference on Computational Science and Engineering*, 1021–1028 December 2013, Sydney, New South Wales.

Zoss A., 2015. [Introduction to Data Visualization: Visualization Types](http://guides.library.duke.edu/datavis/vis_types) [Online]. Available at http://guides.library.duke.edu/datavis/vis_types (8 December, 2015).