IMPLEMENTATION OF BIG DATA ANALYZING TECHNIQUES ON GOVERNMENT DATABASE

By

MD Solyman Ali

153-15-6665

AND

MD Saddam Hossain

153-15-6614

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Md. Sazzadur Ahamed

Senior Lecturer

Department of Computer Science and Engineering

Daffodil International University

Co Supervised By

Ms. Nasrin Akter

Lecturer

Department of Computer Science and Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2019

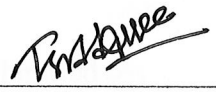
APPROVAL

This Project/internship titled "**Implementation of Big Data Analyzing Techniques on Government Database**", submitted by **Md. Solyman Ali, ID 153-15-6665 and Md. Saddam Hossain, ID 153-15-6614** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13th September, 2019

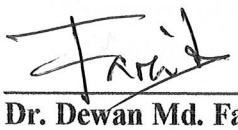
BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Shah Md. Tanvir Siddiquee — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

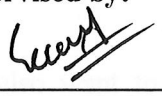


Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

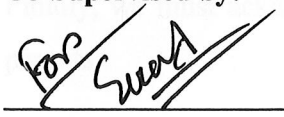
We hereby proclaim that, this project has been done by under the supervision of **Mr. Md. Sazzadur Ahamed, Senior Lecturer, Department of Computer Science and Engineering**, Daffodil International University. We also proclaim that no one this project not any part of this project has been submitted elsewhere for award of any degree or diploma

Supervised by:



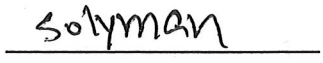
Mr. Md. Sazzadur Ahamed
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

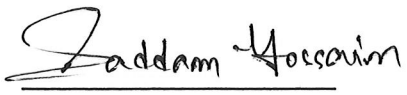


Ms. Nasrin Akter
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



Md. Solyman Ali
ID: 153-15-6665
Department of Computer Science and Engineering
Daffodil International University



Md. Saddam Hossain
ID: 153-15-6614
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First, we want to thanks our almighty Allah to help us to develop this project successfully.

We also hardly thanks to our advisor Mr. **Md. Sazzadur Ahamed, Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Our Supervisor has a great knowledge about big data and data processing. Without his effort and patience it was difficult to complete us in time.

We also want to thanks to our Head of the CSE Department Prof. Dr. Syed Akhter Hossain, for helping us to finish our project and also, we want to thanks our other faculty members for helping us to finish the project

We also thanks to our other faculty members to help us about this project.
Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Data is increasing day by day. Nowadays data flow is increasing rapidly. For this reason, we cannot control the massive data with a normal database management system. We need a system that can handle this large amount of data. We can manage a small amount of data with a normal dataset with the normal system. But if the data size is too high we can’t control it. We need a system that can handle a massive amount of data, also the data security is important. That's why the concept of big data came. Big data management system is designed to extract the information from the dataset. Also, big data can handle complex data system. We have structured, unstructured and semi-structured data set. Big data system can find out the exact information from the data set. Also, big data helps find information easily. We don't need some information rather we need the statistical info. In this process big data is helpful.

TABLE OF CONTENT

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Data is very sensitive and important in modern era. Nowdays after data is increased from 2010 to 2019 twice. Data analysers like to store data from different different resources. Data can be genareted from machine ,our search engine , web site cookie .We cannot ignore those data. They are useful to us and extract meaningful information from those resource. those data can be useful in the our future project. We can use those data to find out the total incress or decress rate for our population , health ,Technology ,business. Market busket analysis is also an inportant fact to find out the clients requrement and product recomendation system. Almost more than 8.47 crore people are connected through mobile phone. this rate has increased rocket dramatically . we cannot handle this huge amount of data with the traditional method .We need a system where we can handle this huge collection of data and manage the system efficiently. Data lose also a big factor. It is the mejor factor in the system. We want to access our system from a system.Visiting multiple system not only difficult but also not a idle to handle the huge collection of data. Hadoop is a modern solution to handle the huge collection of data from a master pc. Nowadays using hadoop has become a popular framework in modern time. Its useful to handle all collection of data from a master system, also it give the data succarity and efficient system . Because of data fratuation rate has dramatically increased , hadoop can be a best solution in modern time. Its well capable of handle this huge collection of data without any system fault.also we can easily manage the system with efficient way. The main plus point to recover the data . If one system is lose another system will take recover the system. client do not need to think about manage data system. most important part it's give the rack awareness in the system , with this user can easily handle the huge collection of data without any problem. so if the client has a multiple system in his total sytem . he don't need to think about it. hadoop give the best data managing data manageing system. we can handle big data collection. Hadoop working process is easy. we can manage the huge collection of data without any problem.

1.2 Motivation

Data is the most important tools in modern era. We need data to extract the meaningful information. The digital data processing system is not easy to handle a huge collection of data. Managing big data is challenge for us. Hadoop can be a best solution for us. We can manage system, extract meaningful information from the framework. Data managing system is not easy. We need a system to manage the data from one system. Has given the solution to handle the system.

1.3 Research Questions

- Is it possible to manage a huge collection of data?
- How to manage different system from one system?
- How much data we can manage?

1.4 Expected Outcome

- Manage huge collection of data.
- Collect data without any system problem
- Take recover from system

1.5 Layout of the Report:

- In the first chapter we have discussed about introduction to the project, motivation, research questions, and it's expected outcome.
- In the second chapter we have discussed about Hadoop and Big data.
- In the third chapter we have discussed about "Background", related works, research summary, and challenges.
- In the fourth chapter we have discussed about Research Methodology.
- In the fifth chapter we have discussed about Summary and Conclusion.

CHAPTER 2

Big Data and Hadoop

2.1 Big Data

Big data is a large scale of data. If we consider 1 GB file, it is a big size. Even 10Gg will not give some any problem. But if we talk about any large scale of data like 10TB or 100TB then it is not easy to handle. Even if we go different system to find all the data, it is not idle for data. That's why big data concept have come. We should have some system to control it. Also we need to extract valuable information from large scale of data. We cannot do it without big data. Visualize also a major factor in data management .we can easily manage big data for visualization and analysis.

2.2 Characteristics of big data

There is three characteristics of big data

- Volume
- Velocity
- Variety

Volume means the size of data. In big data terms data size will be very large. That's why is difficult to handle the data size and management. Velocity is speed of data generating .Data is increasing too fast speed. Even if we see before 2010 the data generating speed become twice then now. Variety means different types of data. We can consider three types of data:

- Structured Data
- Unstructured Data
- Semi-structured Data

Structured data means which types of data has organised way. For example data are structured data. There are row and column are organised in a structured way. Unstructured means why types of data we cannot define which types of for data has organised. For example a.txt file. There are lots of data. We cannot consider which types

of data is stored in this files. Word documents, audio, video, image all are unstructured types of data.

Semi structured means we cannot define which kind of data it is. For example an xml file. We cannot consider which kind of data it is. The data look like structured, also cannot consider unstructured. That's why this data types called semi structured type of data

2.3 Big data resources

Data is generating from different resources. For example every day we use Facebook twitter, Instagram, etc. Every day we share thousands of photos, like, message, comment. That's why data is generating day by day. We need to use bank to store money. Every day transaction come from different resources as well as instalments. Website is also producing huge collection of data. We need information to our goal. So different people are making different types of website. The most commonly use google search engine. Everyday millions of people are searching on google to get the information. Stock market also a big issue data generating for big amount of data as well as machine is producing data every day**.**

2.4 Use of big data

Every day we are using big data to get information. When we search on google or YouTube we can see different same types of products are available there. Search engines using recommendation engines to use big data. Sim Company operator everyday recording our voice call to analysis the data and make wonderful system for our future. We see different defend ford are available in the social media, internet. Big data is a great for detecting fraud. Market basket analysis is also a big plus point of big data. Market analyser using big data to analysis the market product organization. Sentimental analysis also a big part in big data. When we tweet or like ,comment any post social media operators use those data to analysis what do we need , like , dislike etc.

2.5 Hadoop

Hadoop is an open source data management system which can help to organized different data source to one system. If we have different data in different system it is very difficult to manage, and handle. Hadoop makes it easier to distribute different system into one system. It's also a very powerful data management tools. Using this data management tools we can handle a massive amount of data.

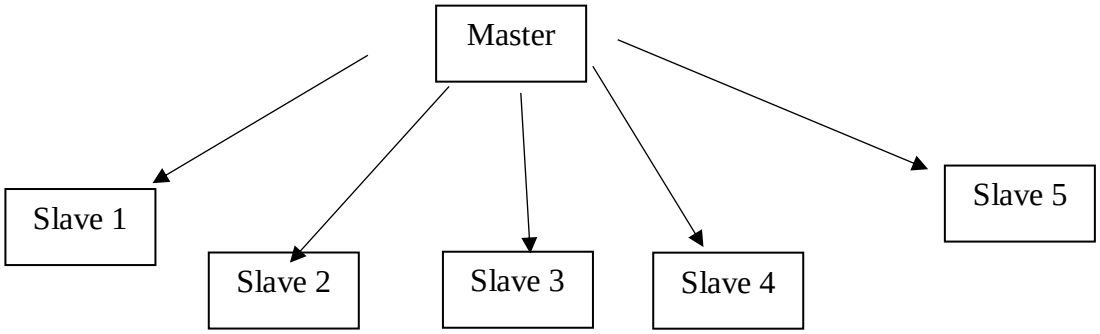


Fig 2.1 : Hadoop Distrubuted File system

Hadoop Components

There are two version of Hadoop. they are Hadoop versoin 1 and Hadoop version 2. Hadoop version 1 is older and Hadoop 2 is newer version. theough two version , the file system is almost similer. both can ran on windows and linux platform. Hadoop 2 gives better performances than Hadoop 1. also Hadoop 2 supports more than 4000 data node.

Hadoop 1 components:

- HDFS [Hadoop distrubuted file system]
- Mapreduce

Hadoop 2 components:

- HDFS [Hadoop distrubuted file system]
- YARN/MRV2

Fig 2.2 : Hadoop Demond

Hadoop Demonds:

Hadoop 1 Demonds:

- NameNode
- Datanode
- Secondary NameNode
- job tracker
- task tracker

Fig 2.3: Hadoop 1 demond

Hadoop 2 Demonds:

- NameNode
- Datanode
- Secondary NameNode
- Resource Manager
- Node Manager

Fig 2.4: Hadoop 2 demond

Hadoop Cluster:

Hadoop cluster is a special types of storing computing system. It can store, manage data, read–write data. There are two types of cluster in Hadoop, one is master and another one is slave cluster. Master system runs Hadoop usage the node idea. There are two types of node. One is master and another one is slave. Master system run on the master node and slave demonds runs on the slave demond. Master cluster store NameNode and resource manager and slave cluster keep DataNode and node manager. All the data information are stored in node Manager and the system is stored in DataNode

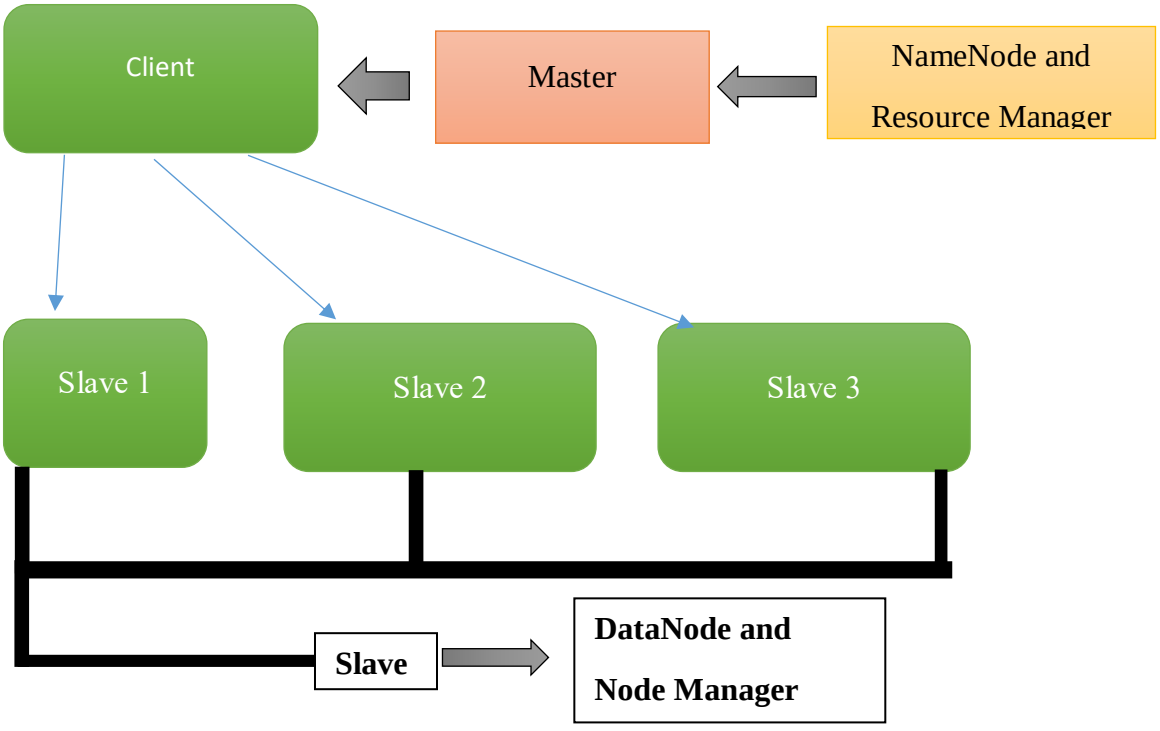


Fig 2.5: Hadoop Cluster

Secondary NameNode:

Secondary NameNode is not a backup for NameNode. It only take after a time. So if NameNode will crash secondary NameNode will not take recover .Secondary NameNode only take a backup after one or certain time if there is any restart required secondary NameNode take stand. In Hadoop 1 secondary NameNode is important. But in Hadoop 2 secondary NameNode less important

Modes of operation in Hadoop system:

There are three types of system in Hadoop.

- Stand alone
- Pseudo distributed
- Fully distributed

  Standalone mode is not for business purpose. This is only for educational purpose. Pseudo mode is for learning and implement purpose. Most of the people use this system because they can run master and slave architecture in this mode. But the device should be well configured. Fully distributed mode is for business or organization, because it need separate operation system and device. This system work well and give better performance.

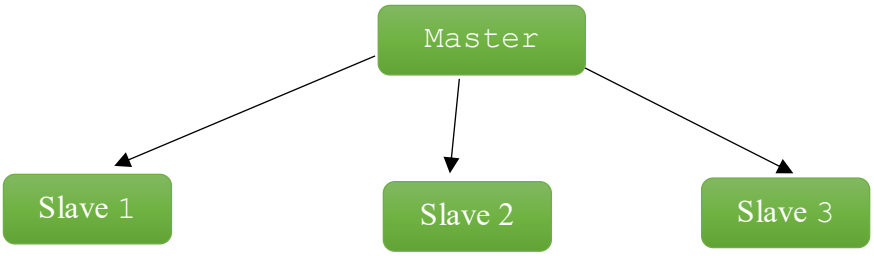


Fig 2.6: Fully distributed Mode

Hadoop 1 vs Hadoop 2:

We all know Hadoop has two basic components. In Hadoop 1 first one is hdfs and another one is MapReduce program. In Hadoop two HDFS and yarn or MRV2 are the two components. The working process of Hadoop 1 and Hadoop are almost similar.

In Hadoop 1 system there are two part MapReduce and HDFS. MapReduce also has two major components. Recourses manager and data processing. So Map reduce has two comports and that's why data the overall performance is decrees in Hadoop 1. On the other hand Hadoop 2 three parts. MapReduce manage the programming and other type of jobs. Yarn manage the resources management and HDFS handle the data management.

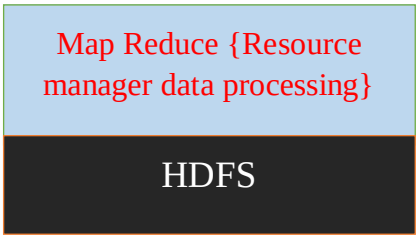


Fig 2.7: Hadoop 1 working process

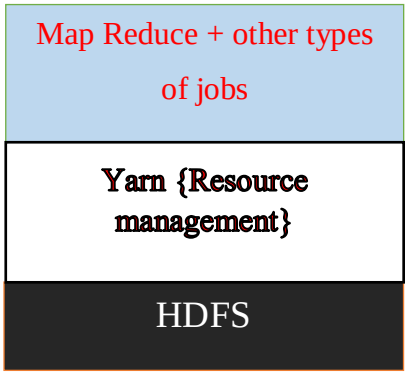


Fig 2.8: Hadoop 1 working process

Limitation of Hadoop 1:

- Only one NameNode can be configured. So if there is any failure in the NameNode it is difficult to extract the information from the DataNodes.
- Hadoop 1 cannot utilize a huge amount resources
- Less scalability when compared to Hadoop 2
- Hadoop 1 is not suitable for real time data processing
- Only max 4200 can be used in Hadoop 1. If try to use more node, the system will crash and stop working.
- Organization need suitable system. But if system fails, organization will fail huge time and money and it is difficult to extract information.

Ecosystem of Hadoop:

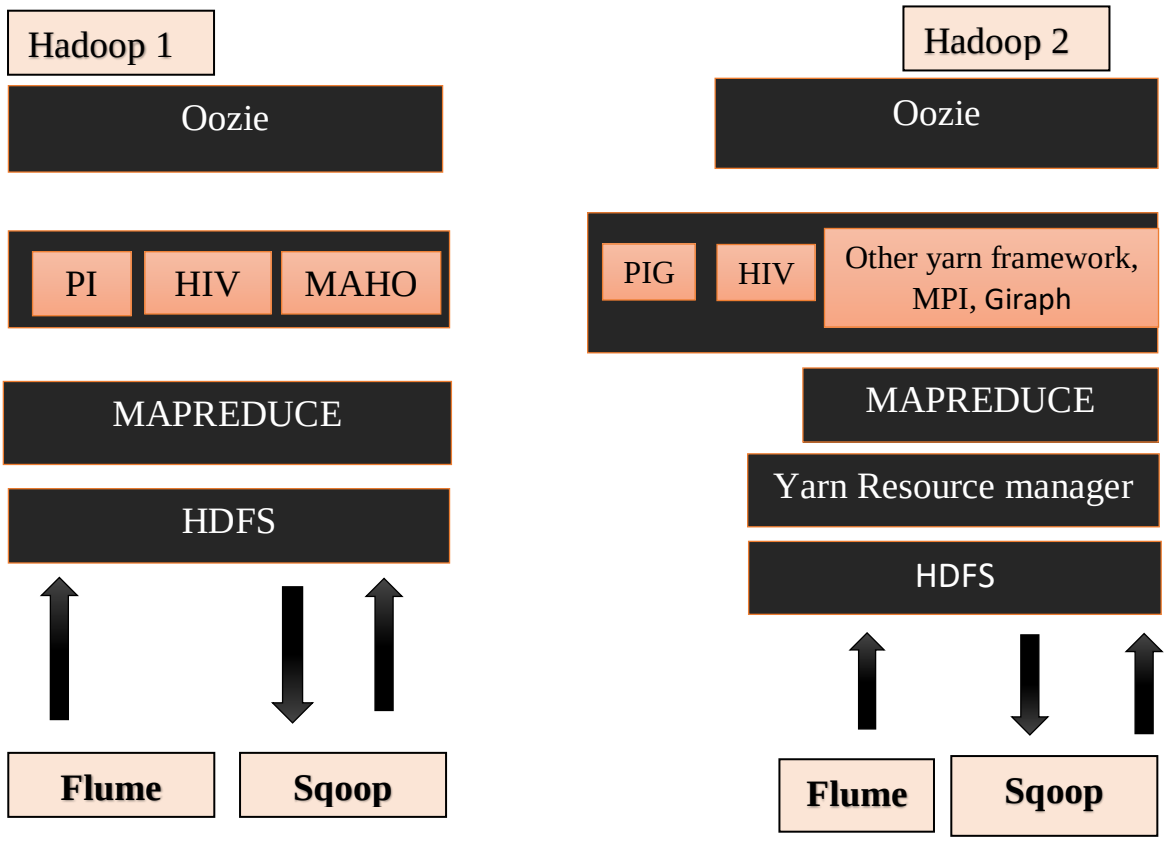


Fig 2.9: Hadoop Ecosystem

This Ecosystem is a part of Hadoop. Hadoop 1 and Hadoop 2 ecosystem almost similar. On the top of the ecosystem Oozie maintain the schedule of all task. That means which task should be execute and which work should not execute. So schedule is the main park of Oozie. PIG, HIVE, MAHOUT are the framework work below the Oozie.In Hadoop 2 MPI, graph have added. So they are almost same in the Hadoop 1 and Hadoop 2.Then MapReduce is responsible for programming in Hadoop. Which DataNode include which, MapReduce decide. In Hadoop 2 the system is almost similar. Hadoop 1 has no other process. The last stage is HDFS. But in Hadoop 2 another process have added to make smoother performance YARN. Yarn manage the resource manager.

2.6 HDFS file architecture

There are couple of things need to know in HDFS file architecture.

- DFS
- HDFS
- File block and replication
- The concept of rack

DFS

DFS means data file system. If someone have some files we cannot say this is distributed. Distributed means equally files are organised. For example if we have have a 2GB file and we want to store it equally in 5 pc. So every pc will get 400MB. This is called distributed file system. Again we have to remember we have to restore the file from the 5 pc. Because the file is cutter down.

HDFS

HDFS means Hadoop distributed file system. Hadoop use master pc and slave pc to store data. The master pc contain the information about the slave pc and slave pc know it data path location. We can access all the data from the master pc as well as we can copy a file from a local computer or copy a file from HDFS to our local Hadoop machine.

File Block and Replication:

When we add some file into our slave pc this file divided into some parts. The default file size is 128MB. We can configure the size in the Hadoop system to hdfs.xml file. If we have a 1GB of data and we give the default size of 100KB then it will divided into a lots a parts. So we have a standard size 128MB. So the file will divided into 8 part. It will increases our computer performance .again we want to copy a file to our system and our hard disk have a low write speed 100MB/S. Thought we have a 5 pc in our cluster it will give 500MB/S. This is a great advantage of Hadoop framework. Data is important, more valuable than hardware. So every company wants the data safety. So they keep the data safe. If we have a big file and we have set a value of our replication.

The default replication size is 3 .that means the divided system will be stored in our different data nodes three time. If any node fail or not working another data node will take over. Because the data blocks stored in different and replicate three time.

The Concept of Rack

Rack is a major concept in HDFS. For example a university have 5 different department. CSE, CST, Economics, BBA, EEE. So the administrator want to store data in an organised way. To keep the data structured he can give every department every rack. So rack means a network which contains a network.

2.7 MapReduce

MapReduce is a programming or logical unit how to customize data. This is divided into two sector. One is map and another one is reduce. Map is responsible for store management and reduce works for processing handling.in traditional data approach the big data were divided into multiple sector. The all the data grep from the match. After matching the divided from all split, the meaningful information can be extracted. For example if the data size is 200 GB and there are 10 system. So the all 200GB is divided into 10 sector. That means the system can store 20GB in a system. Of if someone wants to extract the information from the big data system this is divided into different sector and the person must be coded to find meaningful information.so pleating the data was not easy. So if the data size is huge it was difficult to split all the data.

MapReduce has change the data splitting system. There is key value in map reduce program. The input and output both were stored in key value pair. So in MapReduce everything is in key value pair. So in mapper program, program need to input. The format may be different. For example .txt, .html, .doc etc. Input class help to input the data into map function. The value stored in the key value pair or K/V. So the mapper output goes into the developer mode. It also stored in the key value pair. The mapper value goes into reducer. Mapper input send the data into key value pair. Then the reducer output goes into Developer. The developer keep the value into key value. Everything in MapReduce is key value pair. So all the input and output all the data are into the key value pair.

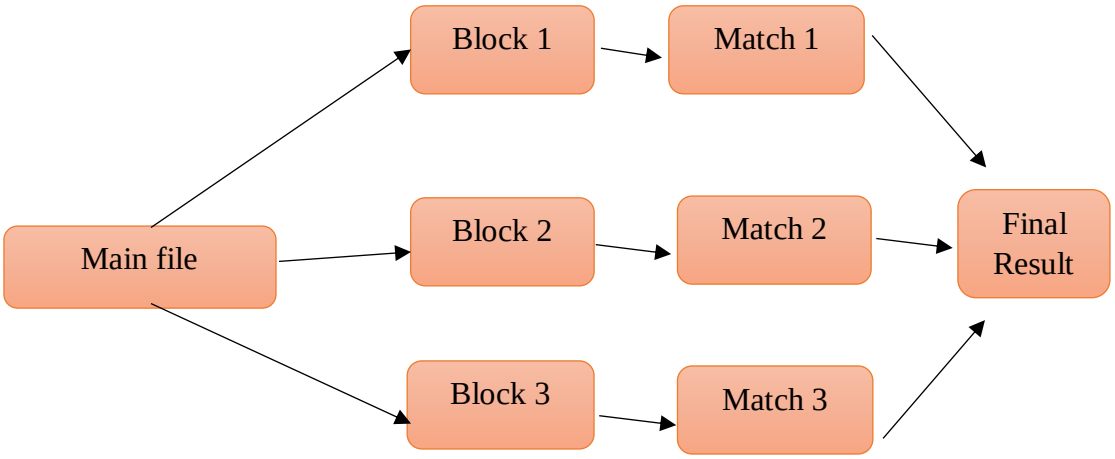


Fig 2.10: Data processing Traditional way

Mapper Work

Mapper works according to the system of file. For example if someone have a 1.5 GB of text or document type file. If the block size is 128MB by default that means the files will be divided into almost more than 10 block. The user has 20 slave pc and one master pc. The mapper theory is the number of mapper and the number block are same.so if the user has 200 block that means the mapper will be run 200 mappers. So if user has coded the mapper program into mapper and the file has to run for execution the system will run 200 mapper or no matter how many block are there. It will run automatically into the system. Is an automatic process

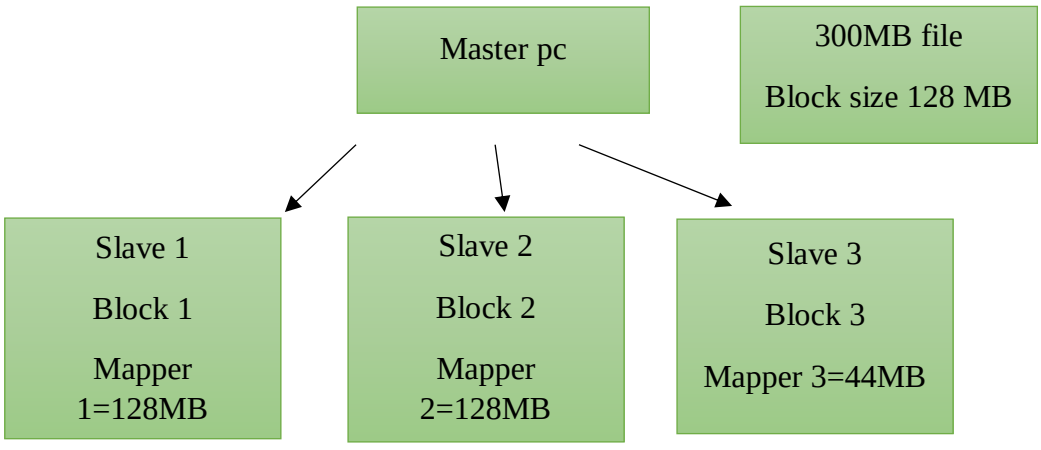


Fig 2.11: mapper working process

The mapper give the data into reducer its task to summarization the operation from the mapper. We already know that mapper works to handle the stored sector. The reducer works into mapper. By default there is only one reduce. But the user can change the reducer value of the reducer.so map take an input file from the source. It makes or it applies some business logic into the cluster and help to manage and find all the data from the source. And finally it gives the output from map. Reducer get input from the mapper, it also write some business logic. But logic is to perform into mapper. Finally it shows the output of the reducer.

2.8 Apache Hive

Hive is a software developed by apache . It works above the mapreduce. hadoop works according the the hadoop query of the the system. it gives the flaver of sql. the code is almost similer of mysql . Hive works without any complexity of the system. Working with mapreduc is difficult because the user need to write code a lot of java programming . So who do not have any idea about java its difficult to manage and coding with java programming. Normally hive convert the code into mapreduce and then execute the code.hive is above the mapreduce in the hadoop ecosystem . So if a user make code in hive it will convert into mapreduce. hive support a large collection of data .the working process of hive is very simple. User have to write some query according to the hive rules and hive convert the code into their logic.it support different different type of file formet like .text , RCfile ,Hbase etc.because it convert it into MapReduce ,that's why it is really easy to code.

CHAPTER 3
BACKGROUND STUDY

3.1 Introduction

In this section, we are going to describe the related works which is based on Hadoop framework. We will review others research paper, their work, their methodology used in paper and their success that related to our paper. In the research summary section we will summaries the related works. In the challenge section we will talk how we succeed in the implementation. We want to work with Big Data. We are motivated from Facebook and some Companies who are manage their data system using big data. And this technology did not widely spread yet. So this is the opportunity for us to make something unique by using this technology. We will apply Big Data on Government websites like Educational Board site and others website where lots of data are stored. In educational board website there are millions of students' result published after board exam and this data are stored using MySQL query language to save the data into database. But students face the problem when thousands of people visit the website to see their result at a same time and that time the server can't take the load and the site crash. There are various type websites in various sector of Government. So sometimes it's difficult to find that website or information that people needs. So our plan is we want to make a single website where all the Government sector information will be included.

3.2 Related Works

B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan have proposed some steps to prevent security issues emerge in Hadoop base layer called Hadoop Distributed File System (HDFS). First approach is "Kerberos Mechanism" based in HDFS which is used to block data correctly and only uses by the authentic user. The next approach is "Bull Eye Algorithm Approach" which is used to secure sensitive data where all the data are stored in without any risk and provide security in 3600 from mode to node. The last approach is based on one "NameNode Approach" which is used for security by achieved replicating and reduce server crashes [1].

Raissa uskenbayeva, Abu kuandykov, Young Im Cho, Tolganay Temirbolatova, Saule Amanzholova, Dinara Kozhamzharova have proposed some method for integrating of data using the Hadoop and R. The approaches are R and Hadoop, R and Streaming, Rhipe and RHadoop where R and Hadoop used for processing the large scale of data, RHadoop and Rhipe give permission to users to make and call their own map. RHadoop is dependent on R packages. Streaming is used for easy jobs where the solution is limited input data files but for the complex jobs, Rhipe and RHadoop are used [2].

To classify Health Information Exchange (HIE) Wang Lijun, Huang Yongfeng, Chen Ji, Zhou Ke, Li Chunhua have proposed a medical information platform which is based on Hadoop and after named Medoop. Medoop builds a platform for health data storage and exchange in the Hadoop ecosystem. For this project, they proposed some approaches which are HDFS based CDA file storage, HBased indexing for query and MapReduce algorithm for CDA information and by they may store the data, organize and may analyze the health data smoothly using the advantage of Hadoop [3].

In this project, authors Said Jai-Andaloussi, Abdeljalil Elabdouli, Abdelmajid Chaffai, Nabil Madrane, Abderrahim Sekkaki have discussed medical content image retrieval by using the Hadoop framework. Nowadays medical content is getting larger and digitized, and these data are stored in a medical image database. So collecting the desired image from the database it's quite complicated. To prevent this problem they took the challenge and applying two models which are MapReduce distributed computing model and the other is the HDFS storage model. They have also used two methods called BEMD-GGD and BEMD-HHT, and used a database named DDSM image database. For implementing the project they collect 2,500 patient files with 10,000 image where each image is 2000by 5000 pixels. After implementing they made an experiment on mean precision at 20 [4].

In this project, LI Jing-min and HE Guo-hui have discussed distributed database management system based on Hadoop. In modern technology, data are increasing day by day. For a huge amount of data it’s quite impossible to manage the data because there is various kind of data like images, audio, video, etc. So for this big amount of data storage, data distribution is complex. So the authors have taken some model which are Hadoop core, HBase and Hive, and also used MapReduce data processing. HBase is used to store the data and Hive is used for data query and analysis. By using these model large-scale

cluster transform in a multi-node cluster and reduce the cost of inputs and provide a suitable method for the cloud computing system [5].

In this project, Mehul Nalin Vora has discussed large scale data which using Hadoop HBase . there are 3 important components of HBase architecture which is HMaster, Region Server and ZooKeeper. For experimenting the project they have user Intel-Itanium(2) processor with 4-core master pc with 16 GBs of RAM and the slave pc was dual core with 2GB RAM. They tested for 2 million image files and the average size of these files are about 50 KB to 100GBs [6].

Yonggang Wang and Sheng Wang has discussed about research and implementation spatial data storage based on Hadoop. A second major area relates to structural data the process of discovering interesting and possibly useful arrangement in structural databases. Finally, as structurally referenced data sets become more extensive in scientific applications, the ambition for full assumption and authentic assessment of model unevenly has become increasingly important [7].

In this research, Ashwani Kumar Kushwaha and Sweta Bhattachrya tried to predict the quality of crop yields based on soil, climate and different diseases of crops. They used the widely used agricultural algorithm 'Agro algorithm'. They used the big data Hadoop platform to deal with three large categorical datasets namely climate data, crop disease related data and soil data. They preprocessed these agricultural dataset and implemented Agro algorithm to predict quality crop yields. [8]

In this research, Ahmed Slama Ismail and Haytham Al-Feel tried to make a recommender system named DLRS which will search research papers in big data environment Based on Hadoop and Hive-Ql as a query engine using MapReduce parallel programming framework. They used both single and three cluster machine to reduce time in search from this large dataset. In both approach the retrieval time was almost similar but the cluster machine approach was slight better then single machine. [9]

3.3 Research Summary

From the above all research paper we can say that all of those authors who research different type of topic was appreciable. From the papers we see all the researches are

most important in this modern technology, and all the authors used Hadoop platform in their project and they got a good result from it. Some authors tried to predict the quality of crop yields based on soil, climate and different diseases of crops using agricultural algorithm 'Agro algorithm'. Our research is on government database and we also used Hadoop platform. Our topic is different from these paper. In our government database there are lots of data and these data are increasing so we are working to manage government data like educational and others using Hadoop. We are working on Hadoop system with can run on windows platform. This system also can run into Linux operating system like Ubuntu, kubuntu, Jubuntu, Linux mint etc. But we have work in windows 10 format. Because windows is the most popular operating system and available to everywhere. With this file system we can handle a huge collection of data. Almost more than 4000 nodes. If one data nodes contains 50 terabyte, then 4000 DataNodes will contain 200000 terabyte. That's a huge collection. This system is easy to use. Data security also a big issue on Hadoop filesystem. We can use multiple master pc. That will save our NameNode. If one data node will destroy we can recover our data with other data nodes.

3.4 Challenges

We are facing lots of challenges in our project and the main problem of our project is data collection because we need lots of data. Government data is sensitive data so we are collecting some dummy data to implement the project and we hope we can solve all the problem that we are facing. To implement this project. We have several problem. First we make sure this system is working correctly we have to use high configuration pc. The slave node or slave demond should minimum 8 GB of ram and master pc should have at least 32 GB ram. Also we need lots of pc, vlan switch, route, cat6 cable. This is too expensive. That's why we have used single pc with single server. This system is for practice. But if we want to see the real result we have to see those device. But we don't have any option. That's why we are using medium range pc. But this system is highly recommended to use. If we use big collection of data the pc will create problem and we cannot run our program. At the start we did not have any knowledge about big data. We had faced different problem to solve big resources. Also Data are not properly available in the internet. We it was a big challenge for us.

CHAPTER 4

RESEARCH METHODOLOGY

4.1 Introduction:

In this chapter we will discuss about our research methodology. To implement this system we have used Hadoop 2.7. We know Hadoop is a very powerful framework to manage big data system. Nowadays data is increasing day by day. It is impossible to manage this data system with normal data management system. Hadoop is widely used all over the world. In this chapter we will discuss how we have used Hadoop and the data system. Hadoop is a very simple data processing system to handle big amount of data. If we have a huge amount of data we can easily handle of data. We can use huge collection of node in Hadoop. That's why data managing is easy in this system. In this system we have used spark, Scala and Hadoop.

4.2 Tools idle for Hadoop:

- Hadoop 2.7.2
- JDK-8u211
- Scala 2.11.12
- Spark 2.4.3
- Windows 10/ Windows 7
- Core i7, minimum 6 core
- Ram 16 GB
- 30GB hard disk space
- Intel gaming PC

Our system configuration

Table 4.1: System Configuration for our usage PC

PC NO	**Motherboard**	**Ram**	**Hard Disk**	**Processor**	**Processor Core**
1	Gigabyte GA Z170-M D3H DDR3	8GB	2TB	Core i5-Intel 6500	8
2	Gigabyte GA-AX370-Gaming DDR4 AM4	8GB	1TB	Core i5-Intel 6500k	8
3	MSI B450 GAMING PRO Carbon AC DDR4	12GB	1TB	Intel Coffee Lake Core i7 8700K	12

4.3 Directory Setup:

To make this system first we have to set the Hadoop environment .We have to download all the component and make a directory in the c drive. The folder name should not include any error because that can cause error in our system. Than we have to install JDK and Scala to our PC. We have to copy the java and Scala to our program directory to work directory because the program folder contain space that can make error while running our program. Than we have to extract Hadoop and spark to our work directory. Should be remembered that there should not be any space in Hadoop or other folder otherwise it will not rum. We have to create directory in the c drive with the name tmp and the tmp folder there should be another folder called hive. Also we have to create folder in the Hadoop with the name of data and inside the data directory there should be another two folder named NameNode and DataNode.

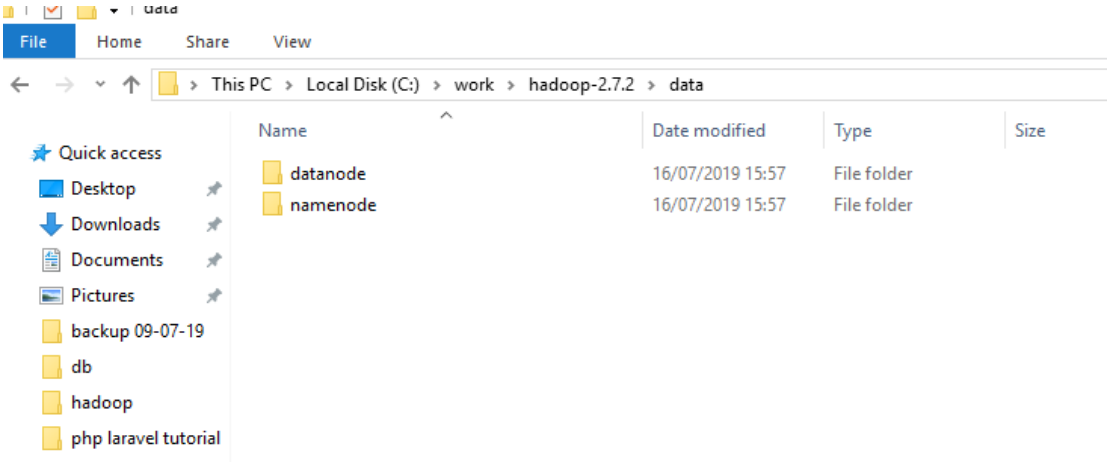


Fig 4.1: Folder inside the Hadoop

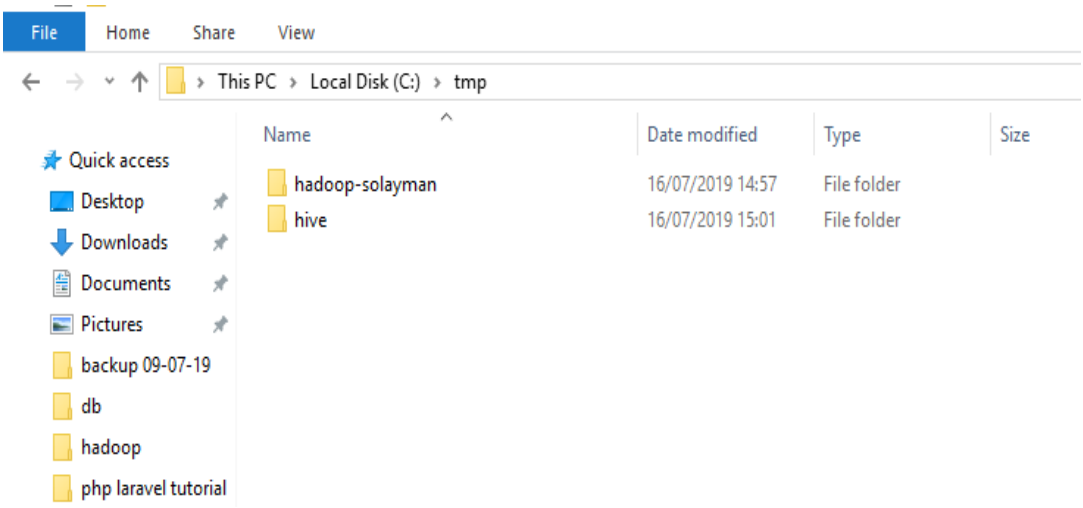


Fig 4.2: Folder inside the tmp

Environment variable setup:

This will be our folder setup. Now we have to setup our variable link to the system to get proper execution to the system. Without making this the framework will not properly work. We have to copy the Hadoop, spark, Scala and java directory link and have to add to our user environment variable and system variable. Also inside our Hadoop and spark folder there are two directory. One is bin and another is sbin. Are also important. We have to copy those as well as we have to set those link in the system variable .linking the variable is the most important part in our system. Hadoop run through java system. So without proper link of java file it will not work properly give exception.

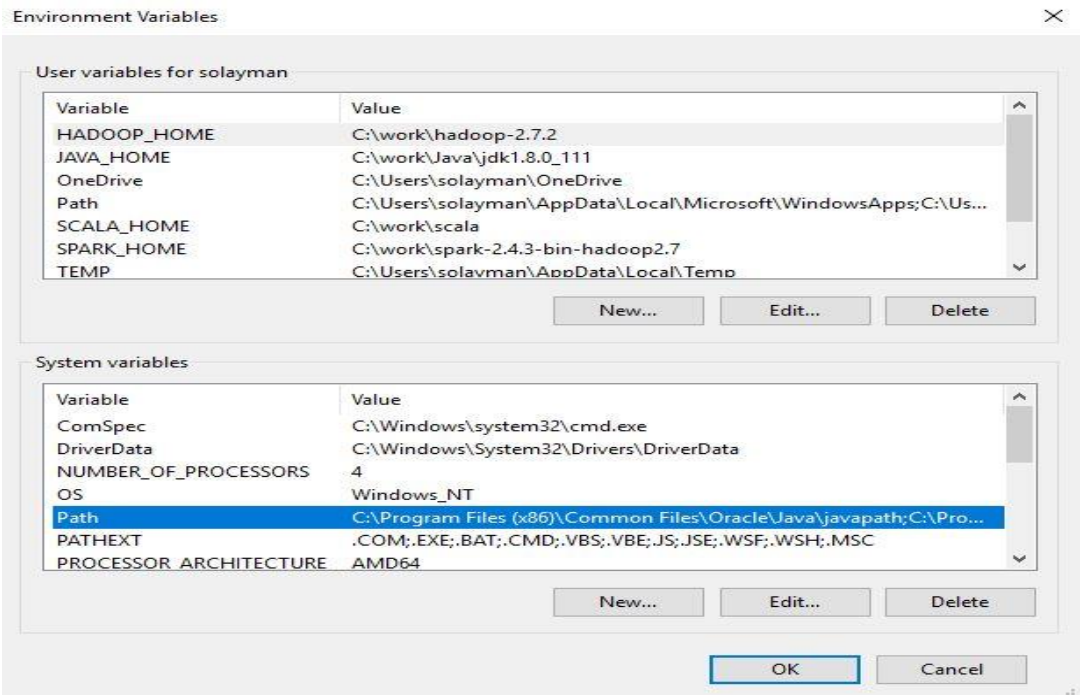


Fig 4.3: Setup Environment Variable

4.4 Hadoop configure file setup:

Now we have to set configure our Hadoop etc. folders files and bin files. We need to change the core-site.xml and set the property name to fs. Default FS and property value to hdfs://localhost:9000. It will be configure our etc file. After the completing our core-site file we have to change Hadoop-env.cmd file and set need to change the path of java location. Otherwise it will not work. The loalhost is our local server name. With this server address we have to access our system.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
   Licensed under the Apache License, Version 2.0 (the "License");
   you may not use this file except in compliance with the License.
   You may obtain a copy of the License at

     http://www.apache.org/licenses/LICENSE-2.0

   Unless required by applicable law or agreed to in writing, software
   distributed under the License is distributed on an "AS IS" BASIS,
   WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
   See the License for the specific language governing permissions and
   limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

Fig 4.4: Setup Core-site file

```
rem Extra Java CLASSPATH elements.  Automatically insert capacity-scheduler.
if exist %HADOOP_HOME%\contrib\capacity-scheduler (
  if not defined HADOOP_CLASSPATH (
    set HADOOP_CLASSPATH=%HADOOP_HOME%\contrib\capacity-scheduler\*.jar
  ) else (
    set HADOOP_CLASSPATH=%HADOOP_CLASSPATH%;%HADOOP_HOME%\contrib\capacity-scheduler\*.jar
  )
```

Fig 4.5: Setup Java Path file

Another important file is our hdfs-site.xml file. We have to set our NameNode and DataNode path inside the file system. Inside the property the topper part value is our name and bottom part is our destination path. There are two path in our Hadoop system. One is NameNode another part is data node. NameNode store all the information about data node. Name node store in the master pc and data node store in the slave pc. Data node is responsible for storing all the necessary data in the system and if data node is lost another data node have the same backup but if the name node will lost all data will be lost.so for this purpose another 2 or three master pc have to add in the system to safe the system .If one master fail another master will take position and help to recover and reconnect the system again.

```
<configuration>
 <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>C:\work\hadoop-2.7.2\data\namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>C:\work\hadoop-2.7.2\data\datanode</value>
    </property>
</configuration>
```

Fig 4.6: Setup hdfs-site.xml file

4.5 Installation Hadoop On system:

To install Hadoop on the system first we have to run our command prompt. First we have to check if the system has properly installed java, spark Scala and spark. If those are properly installed the display will show a proper message. Otherwise it will give exception on the system. If there is any error we have to every single element inside the environment variable, java configuration and Hadoop configuration.

If there is no error there must be file problem. If it happened, we have to use different version of file system. Otherwise we can continue.

If everything is right then we have to format your Hadoop system. To format our Hadoop and install on our system we have to use this command

- hdfs.cmd NameNode –format
- start-dfs.cmd && start-yarn.cmd

If we get any windows permission popup we have to approve this popup with position click. Otherwise it will not work on the system.

Now to check our system is properly work or not we have to give this command.

- Jps
- 10864 NameNode
- 14992 Resource Manager
- 19048 DataNode
- 3016 Jps
- 11036 Node Manager

If we get those command view then our system is properly. Also there will be four popup command.

Now to check our machine we have to open our browser and inside the url box we have to type:

http://localhost:8088

There will be a web page of Hadoop. This is our main local system where we can see every file structure of Hadoop system. There are all the files of Hadoop are organised way. Also we can see the files available on different directory, files permission and file size block size. The default block size is 128MB.

We can change the block size in Hadoop configure.

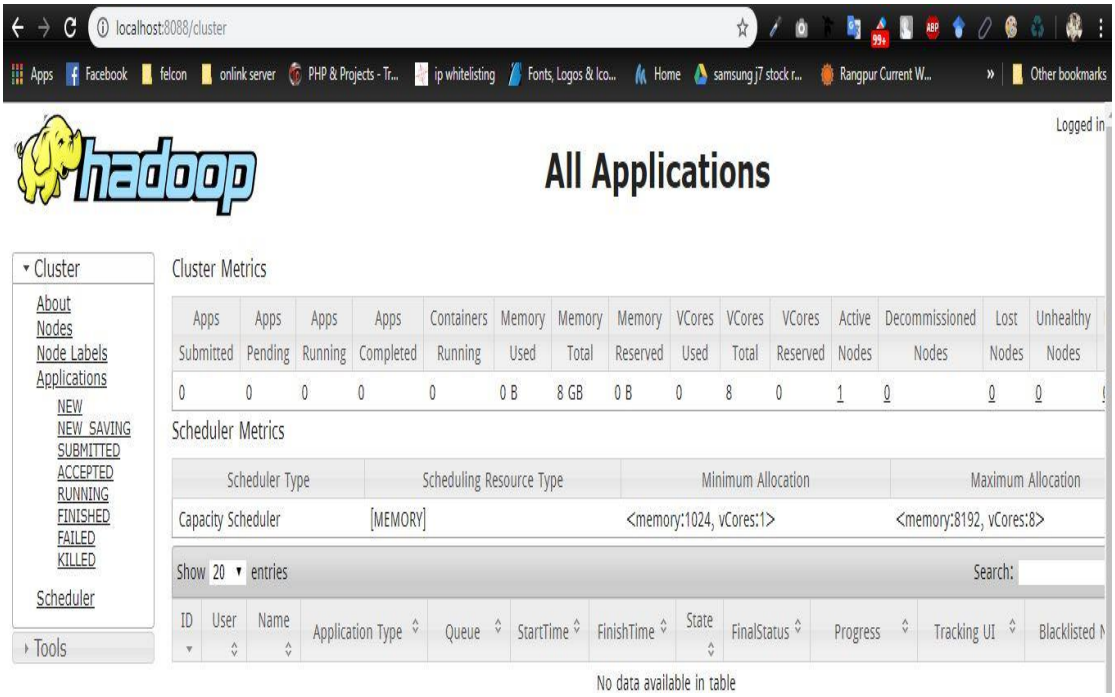


Fig 4.7: Hadoop Cluster Metrics

We have to see another local server to check the machine is properly working or not. This is our main local server. Inside this server we can see our file system inside this local server.

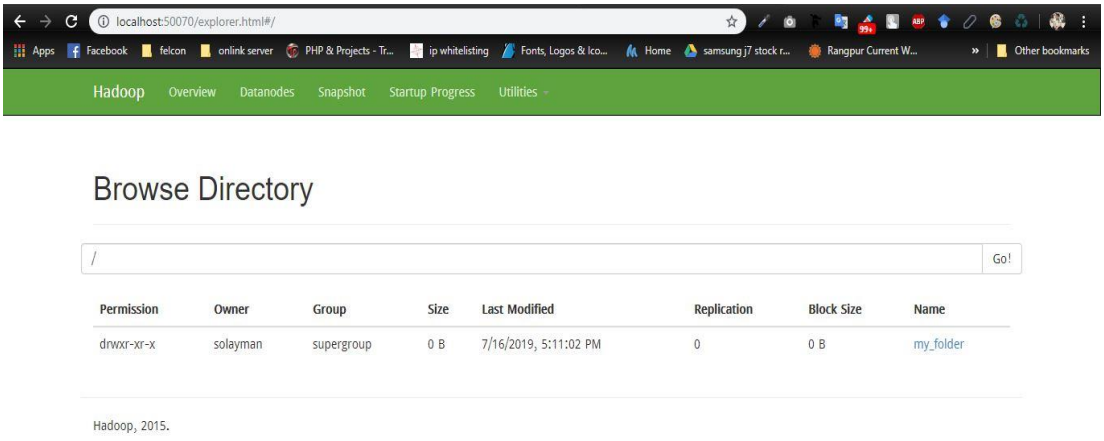


Fig 4.8: Hadoop Browse Directory

Hadoop Useful Command:

We can use some useful command to manage data system.

- Hadoop dfs -mkdir /user/data/ [to create a directory we have to use this command. mkdir means make directory and /user/data/ is the path. That means inside the root folder there will be a path named user and inside the user data directory will be created.
- Hadoop dfs–rmr /myName [to remove a directory need to use this command, rmr means remove and /myName is directory path.
- Hadoop fs –ls /Myfile [too see inside directory file list this command is used
- hdfs dfs -put /home/dataflair/Desktop/sample /user/dataflair/dir1 [to copy a file from a local system and put into hdfs system this command is used/home/dataflair/Desktop/sample /user/dataflair/ is the local directory and /dir1 is the Hadoop directory.
- There are also lots of command available on google

CHAPTER 5

RESULT

5.1 Introduction

In this section we have worked with dataset and showed the result after collecting the data. The dataset was not pre-processed. So first we have preprocessed the data then apply hive on these dataset and got the exact result that we wanted.

5.2 Result Analysis

Hive works above the map reduce. The query is almost similar to sql. So if there is anyone who comfortable with sql query with sql hive can works with hive. A programmer have write huge amount of code in java programming. But in hive the programmer have to run some basic query. Hive can manage a huge collection of dataset. The database design is almost similar to sql. But hive have some extra advantage in working process.

We have work with different different dataset. And the result was great with our device. It can manage a large collection of data. The database stored in data wirehouse. Inside the database the table stored.

First we have inserted some dummy data inside the table. It works fine with those files. But the files size was too low. We have created 3 dummy data and inserted into our database table. We were able to extract meaningful information with those files. We had counted number of column, one row information with unique id. The column types are integer, string.

But if the dataset is big collection Hadoop system will scan will the whole table. That's why Hadoop has given a better system to manage the dataset. The dataset can be divided into multiple partition with specific category. If we need to find the exact information it will only scan the specific category files. It will improve our working performance. For our working purpose we have used Dhaka stock exchange data from 2008 to 2017. Each file contains minimum 650 column of info. Total number of file 10. We had inserted the files into our table.

The count, show full information, row information worked fine. We can find the whole information or a category type information. This types of partition is static partition.

Table 5.1: Stock Exchange Dataset

Table Patition	Sample Data
stock exchange 2008	30/12/2008 00:00,1STBSRS,823,840,819,825,826.25,825.75,86,4.1475,5000
stock exchange 2009	30/12/2009 00:00,1STBSRS,1250,1290,1236,1280,1253,1283.25,67,5.4105,4250
stock exchange 2010	30/12/2010 00:00,1JANATAMF,12.9,13.9,12.3,12.3,13,12,4708,124.868,9623500
stock exchange 2011	29/12/2011 00:00,1JANATAMF,9,9,8.6,8.6,8.9,8.8,305,12.2408,1364500
stock exchange 2012	30/12/2012 00:00,1JANATAMF,7,7.2,7,7.2,7,7.1,59,0.5651,80000
stock exchange 2013	30/12/2013 00:00,1JANATAMF,6.1,6.1,5.9,5.9,6,5.9,68,0.8164,135500
stock exchange 2014	30/12/2014 00:00,1JANATAMF,5.1,5.2,5,5,5.1,5.1,107,0.295,58281
stock exchange 2015	31/12/2015 00:00,1JANATAMF,4.2,4.3,4.2,4.3,4.3,4.3,16,0.377,87878
stock exchange 2016	29/12/2016 00:00,1JANATAMF,5.8,6.2,5.8,6.2,5.9,6.2,443,26.409,4427760
stock exchange 2017	28/12/2017 00:00,1JANATAMF,6.4,6.5,6.4,6.4,6.4,6.5,79,1.888,294720

Those data maintain a sequential structure. If we load the data inside expected table we will get the result for showing the data. We can also count, sum, average with those result.

Table 5.2: Stock Exchange Result

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col 11
30/12/2008 00:00	1STBSRS	823	840	819	825	826.25	825.75	86	4.1475	5000
28/12/2008 00:00	1STBSRS	817	829	802	802	825.75	790	59	2.7838	3400
24/12/2008 00:00	1STBSRS	789	809.75	786	790	790	785.75	74	3.527	4450
23/12/2008 00:00	1STBSRS	785	798.5	785	785	785.75	782.5	38	2.0541	2600
22/12/2008 00:00	1STBSRS	778.25	794.5	775.5	787	782.5	797.75	19	0.7825	1000
21/12/2008 00:00	1STBSRS	790.5	810	790	798	797.75	779	64	2.999	3750
18/12/2008 00:00	1STBSRS	786	786	765	784.75	779	764	37	1.9409	2500
17/12/2008 00:00	1STBSRS	769	782	752.75	760.25	764	761.5	17	0.8023	1050

Table 5.3: Temperature and Rain Status Dataset

temperature	Month	Year	rain
16.976	1	1901	18.5356
19.9026	2	1901	16.2548
24.3158	3	1901	70.7981
28.1834	4	1901	66.1616
27.8892	5	1901	267.215
28.8925	6	1901	341.042
28.3327	7	1901	540.907
28.3327	8	1901	493.21

This is the situation we know the category type with the specific data. We can dynamically select the data type and the file will be automatically partition by the hive. If we have a file contains 50 different category .Hive will convert the data type into different sector. So if we want to search a table with specific table we will get the result.

Those are the sample unorganized data .we had created different category according to the month wise data. So there was 12 dynamic category partition. So every year has one month divided category. But if there will be any timestamp we will not divided it because there will be a huge collection of table any that can be unmanageable. So our file was divided into 12 different sector. And each category has a month. And inside the month

there was year. So if we want to search any particular table the table will not scan the whole sector. It will only scan the particular sector.

Table 5.4: Temperature and Rain Status Result after Partition

tem	Month	Year	rain
16.976	1	1901	18.5356
18.5455	1	1902	1.29152
17.7023	1	1903	2.4001
17.7866	1	1904	0.3257
17	1	1905	7.07224
17.5423	1	1906	13.9546
19.0642	1	1907	10.7854
17.2889	1	1908	7.99686

Hive has given another option to divide the data into multiple sector. If there is no way to divide the data into multiple sector or category we can divided the dataset into some limited sector. For example we want to divide the dataset into 200 sector. So the dataset of the file will be divided into 200 sector. We can define the dataset into multiple sector.

We have a dataset of more than 1 million column .so we divided the dataset into 100 sector. Each bucket or cluster contains more than 10000 column. So if we want to make any operation inside the bucket, the query will execute the particular operation.

1487927612,4880850486,I20120726000331,HALL,ESTHER,F,LIFE CHIROPRACTIC COLLEGE - WEST,2010,CHIROPRACTIC,PLACERVILLE,CA,956673933,5306228041
1235146762,2365435336,I20040406000367,WHITE,BARBARA,F,OTHER,1992,CLINICAL SOCIAL WORKER,ENGLEWOOD,NJ,76312530,2014102812
1285727842,5799762829,I20040707000479,CASHA,MARK,M,OTHER,1981,CHIROPRACTIC,FAIR OAKS,CA,956283541,9169677436
1295821098,840459889,I20120309000459,MORANZ,JANICE,F,OHIO STATE UNIVERSITY COLLEGE OF MEDICINE,1984,DERMATOLOGY,ALBUQUERQUE,NM,871101437,5052563648
1063514289,1951391671,I20040514000778,MCGREGOR,VICTOR,M,OTHER,1997,NURSE PRACTITIONER,SAUGERTIES,NY,124771804,8455322493
1346282258,5395768527,I20060113000139,DAVIDSON,JOHN,M,OTHER,1999,CLINICAL SOCIAL WORKER,BINGHAMTON,NY,139015821,6072456259
1932283124,5193762862,I20050415000143,CAGEN,STEVEN,M,SHERMAN COLLEGE OF STRAIGHT CHIROPRACTIC,1997,CHIROPRACTIC,BREVARD,NC,287129524,8288857100
1902950462,7416123666,I20120110000522,ESPY,LEISHA,F,LIFE CHIROPRACTIC COLLEGE,1985,CHIROPRACTIC,ROSSVILLE,GA,307411348,7068667557

Fig 5.1: Physician Compare National Sample Dataset

we can divided the dataset into multiple partition than again the dataset can be divided into multiple bucket. It can be use a big collection of dataset.so if the data size is big it will not check the whole dataset. Rather first it will check the partition then it will go to the particular data requirement.

Table 5.5: Physician Compare National Sample Result

NPI	PAC ID	Professional Enrollment ID	First Name	Last Name	Gender
1487927612	4880850486	I20120726000331	ESTHER	HALL	F
1235146762	2365435336	I20040406000367	BARBARA	WHITE	F
1285727842	5799762829	I20040707000479	MARK	CASHA	M
1295821098	840459889	I20120309000459	JANICE	MORANZ	F
1063514289	1951391671	I20040514000778	VICTOR	MCGREGOR	M
1346282258	5395768527	I20060113000139	JOHN	DAVIDSON	M
1932283124	5193762862	I20050415000143	STEVEN	CAGEN	M
1902950462	7416123666	I20120110000522	LEISHA	ESPY	F
1518981026	7719166586	I20110125001223	DAN	PETROSKY	M

We can retrieve different information from the input dataset. Output will show the specific information without scan the whole table. This will improve our working performance.

```
hive> select * from stock_market limit 10;
OK
 30/12/2012 00:00  1JANATAMF 7.0 7.2 7.0 7.2 7.0 7.1 59 0.5651 80000
 27/12/2012 00:00  1JANATAMF 7.1 7.2 7.0 7.2 7.1 7.1 73 0.9062 127500
 26/12/2012 00:00  1JANATAMF 7.1 7.3 6.9 7.0 7.1 7.0 128 1.9881 277000
 24/12/2012 00:00  1JANATAMF 7.0 7.1 6.8 6.9 7.0 6.8 186 3.064 441500
 23/12/2012 00:00  1JANATAMF 6.9 7.0 6.8 6.9 6.8 7.0 144 1.9875 287500
 20/12/2012 00:00  1JANATAMF 7.1 7.2 7.0 7.2 7.0 7.1 128 2.4237 344500
 19/12/2012 00:00  1JANATAMF 7.2 7.2 7.0 7.0 7.1 7.1 104 1.7669 248500
 18/12/2012 00:00  1JANATAMF 7.2 7.2 7.1 7.2 7.1 7.2 94 0.8079 113500
 17/12/2012 00:00  1JANATAMF 7.2 7.3 7.2 7.2 7.2 7.1 77 1.2892 177500
 13/12/2012 00:00  1JANATAMF 7.2 7.2 7.0 7.1 7.1 7.2 77 0.5173 72500
10 rows selected (0.12 seconds)
```

Fig 5.2: Data load inside the table

But it will scan the whole table. If the data size is too many it can reduce the performance of the data table. But in some case, if the data size is small there is no complexity, we can use this data system. The time complexity will be too high in this static table system. But it's a great way to manage the data table from the dataset.

We can also partition the data files inside the tables, this is called static partition. If we have different different data for category we can partition those data into multiple partition. We have used Dhaka stock exchange data from 2008 to 2017. Those data are inserted into the table with partition. So if we want to scan the file we do not need to scan the whole table document. We have to give the condition and hive will automatically find the information without scan the whole table. It improve the quality of time and performance.

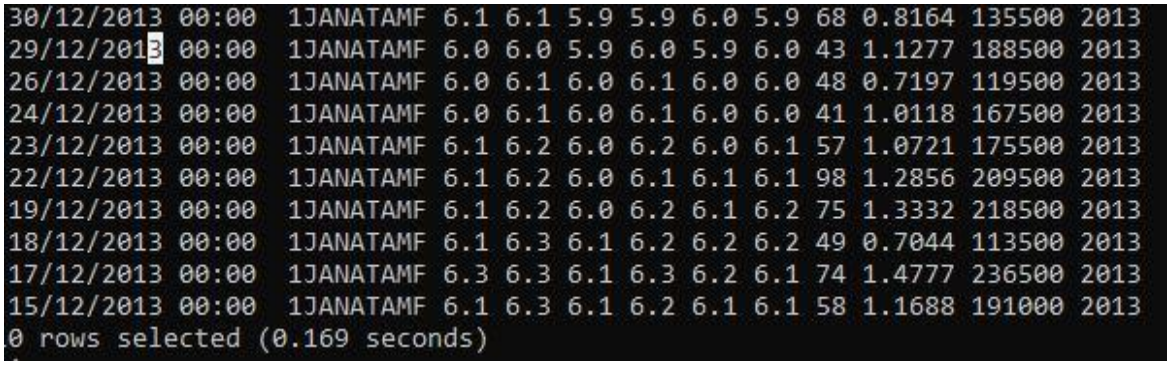


Fig 5.3: Static partition sample dataset for 2013 Dhaka Stock Exchange

But every time static partition does not work. If the size of the file is too high or there are multiple category then it is difficult to manage the partition with static way. Then we need to work with dynamic partition.

We used cluster to divided the big dataset into multiple sector .if we have a 1 million of dataset and we do not have any option to divided the dataset into multiple category. for example if you have a dataset which contains 20 years record . that menas there are 20*365=7300 days available. if we divided the dataset into multiple it will be almost unmanageable to handle the huge collection of data. the solution is clustering or bucketing.

CHAPTER 6
CONCLUSION

6.1 Conclusion

Data is generating rapidly day by day. From 2010 to 2019 data increasing rate is almost double. Every day millions of terabyte is generating day by day. Data increasing rate will not stop. But the data management is going to difficult to difficult day by day. We have to think so bigger. Hadoop can be a recent solution. Maybe after 50 or more years later Hadoop will not work. But at present this framework is a grate choice for the data management lovers. We have to accept this big data management system. And I think Hadoop is a perfect solution for everyone.

REFERENCES

[1]. Saraladevi, B., et al. “Big Data and Hadoop-a Study in Security Perspective.” Procedia Computer Science, vol. 50, Elsevier BV, 2015, pp. 596–601. Crossref, doi:10.1016/j.procs.2015.04.091.

[2]. Uskenbayeva, Raissa, et al. “Integrating of Data Using the Hadoop and R.” Procedia Computer Science, vol. 56, Elsevier BV, 2015, pp. 145–49. Crossref, doi:10.1016/j.procs.2015.07.187.

[3]. Wang Lijun, et al. “Medoop: A Medical Information Platform Based on Hadoop.” 2013 IEEE 15th International Conference on E-Health Networking, Applications and Services (Healthcom 2013), IEEE, 2013. Crossref, doi:10.1109/healthcom.2013.6720779.

[4]. Jai-Andaloussi, Said, et al. “Medical Content Based Image Retrieval by Using the Hadoop Framework.” ICT 2013, IEEE, 2013. Crossref, doi:10.1109/ictel.2013.6632112.

[5]. Li, Jing-min, and Guo-hui He. “Research of Distributed Database System Based on Hadoop.” The 2nd International Conference on Information Science and Engineering, IEEE, 2010. Crossref, doi:10.1109/icise.2010.5689141.

[6]. Mehul Nalin Vora. “Hadoop-HBase for Large-Scale Data.” Proceedings of 2011 International Conference on Computer Science and Network Technology, IEEE, 2011. Crossref, doi:10.1109/iccsnt.2011.6182030.

[7]. Yonggang Wang, and Sheng Wang. “Research and Implementation on Spatial Data Storage and Operation Based on Hadoop Platform.” 2010 Second IITA International Conference on Geoscience and Remote Sensing, IEEE, 2010. Crossref, doi:10.1109/iita-grs.2010.5603956.

[8]. Kushwaha, Ashwani Kumar. “Crop yield prediction using Agro Algorithm in Hadoop.” (2015).

[9]. Ismail, Ahmed Slama, and Haytham Al-Feel. “Digital Library Recommender System on Hadoop.” 2015 IEEE Fourth Symposium on Network Cloud Computing and Applications (NCCA), IEEE, 2015. Crossref, doi:10.1109/ncca.2015.27.

Big Data..2

ORIGINALITY REPORT

12%	5%	8%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	topics.sciencedirect.com Internet Source	1%
3	Yi Li, Fang Shi, Hengxu Zhang. "Panoramic synchronous measurement system for wide-area power system based on the cloud computing", 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018 Publication	1%
4	Uskenbayeva, Raissa, abu Kuandykov, Young Im Cho, Tolganay Temirbolatova, Saule amanzholova, and Dinara Kozhamzharova. "Integrating of Data Using the Hadoop and R", Procedia Computer Science, 2015. Publication	1%
5	Wang Lijun, Huang Yongfeng, Chen Ji, Zhou Ke, Li Chunhua. "Medoop: A medical information platform based on Hadoop", 2013	<1%

IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013), 2013
Publication

6	Submitted to National College of Ireland Student Paper	<1%
7	Syeda Sana Bukhari, JinHyuck Park, Dong Ryeol Shin. "Hadoop based Demography Big Data Management System", 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2018 Publication	<1%
8	B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. Saleem Basha, P. Dhavachelvan. "Big Data and Hadoop-a Study in Security Perspective", Procedia Computer Science, 2015 Publication	<1%
9	Jing-min Li, Guo-hui He. "Research of distributed database system based on Hadoop", The 2nd International Conference on Information Science and Engineering, 2010 Publication	<1%
10	Qin, Shihong, and Xiaolong Li. "Design of a Log Analysis System Based on Hadoop", Industrial Engineering Machine Design and Automation	<1%

(IEMDA 2014) & Computer Science and Application (CCSA 2014), 2015.
Publication

11	Shujaat Hussain, Maqbool Hussain, Muhammad Afzal, Jamil Hussain, Jaehun Bang, Hyonwoo Seung, Sungyoung Lee. "Semantic preservation of standardized healthcare documents in big data", International Journal of Medical Informatics, 2019 Publication	<1%
12	Submitted to CSU, San Jose State University Student Paper	<1%
13	Submitted to University of Westminster Student Paper	<1%
14	Submitted to University of Bedfordshire Student Paper	<1%
15	dblp.dagstuhl.de Internet Source	<1%
16	Submitted to Jamia Milia Islamia University Student Paper	<1%
17	Ahmed Slama Ismail, Haytham Al-Feel. "Digital Library Recommender System on Hadoop", 2015 IEEE Fourth Symposium on Network Cloud Computing and Applications (NCCA), 2015 Publication	<1%

18	paper Student Paper	<1%
19	Submitted to Institute of Technology, Nirma University Student Paper	<1%
20	kb.psu.ac.th Internet Source	<1%
21	Submitted to University of North Texas Student Paper	<1%
22	fruct.org Internet Source	<1%
23	Seema Maitrey, C.K. Jha. "Handling Big Data Efficiently by Using Map Reduce Technique", 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 2015 Publication	<1%
24	Submitted to University of Waikato Student Paper	<1%
25	repository.up.ac.za Internet Source	<1%
26	Madhavi Vaidya, Shrinivas Deshpande. "Critical Study of Performance Parameters on Distributed File Systems Using MapReduce", Procedia Computer Science, 2016	<1%

	Publication	
27	ijarcsse.com Internet Source	<1%
28	C Srimathi, Soo-Hyun Park, N Rajesh. "Proposed framework for underwater sensor cloud for environmental monitoring", 2013 Fifth International Conference on Ubiquitous and Future Networks (ICUFN), 2013 Publication	<1%
29	ecohs.nl Internet Source	<1%
30	www.semanticscholar.org Internet Source	<1%
31	Submitted to Yildirim Beyazit Universitesi Student Paper	<1%
32	Submitted to The University of Memphis Student Paper	<1%
33	Lecture Notes in Electrical Engineering, 2014. Publication	<1%
34	Submitted to University of Technology, Sydney Student Paper	<1%
35	K. Meena, J. Sujatha. "Reduced Time Compression in Big Data Using MapReduce Approach and Hadoop", Journal of Medical Systems, 2019	<1%

	Publication	
36	"Beyond the Internet of Things", Springer Science and Business Media LLC, 2017 Publication	<1%
37	Submitted to National Institute Of Technology, Tiruchirappalli Student Paper	<1%
38	Submitted to University of Central Lancashire Student Paper	<1%
39	www.opensourceforu.com Internet Source	<1%
40	Submitted to University of Western Sydney Student Paper	<1%
41	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
42	Madjid Khalilian, Maryam Fathi Ahmadsaraei, Lida Farajpour. "Security threats and their mitigation in big data recommender systems", Institution of Engineering and Technology (IET), 2019 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		