Indian Education Statistics

Submitted in partial fulfilment of the Requirements for the Degree of

Master of Computer Applications

A PROJECT REPORT
Submitted by

APOORVA SRIVASTAVA
(University Roll No 1900290140008)

SURUCHI SINHA
(University Roll No 1900290140040)

Batch:2019-2022

Under the Supervision of
VIDUSHI
(Assistant Professor)

DEPARTMENT OF COMPUTER APPLICATIONS
DR.APJ ABDUL KALAM TECHNICAL UNIVERSITYLUCKNOW
(Formerly Uttar Pradesh Technical University, Lucknow)

DECLARATION

I hereby declare that the work presented in this report entitled “Indian Education Statistics", was carried out by me. I have not submitted the matter embodied in this report for the award of any other degree or diploma of any other University or Institute. I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programs, experiments, results, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources. I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable.

Name :

Roll. No.

Branch :

CERTIFICATE

Certified that **Apoorva Srivastava (University Roll No 1900290140008), Suruchi Sinha (University Roll No 1900290140040)**, have carried out the project work having "Indian Education Statistics" for Master of Computer Applications from Dr. A.P.J. Abdul Kalam Technical University (AKTU) (formerly UPTU),Technical University, Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself/herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:

Apoorva Srivastava
University Roll No. 1900290140008
Suruchi Sinha
University Roll No. 1900290140040

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Vidushi Mishra
(Assistant Professor)
Department of Computer Applications
KIET Group of Institutions, Ghaziabad

Signature of External Examiner **Signature of Internal Examiner**

ACKNOWLEDGEMENT

Success in life is never attained single handed. My deepest gratitude goes to my thesis supervisor, **Vidushi** for his guidance, help and encouragement throughout my research work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to **Dr. Ajay Kumar Shrivastava, Professor and Head, Department of Computer Applications**, for his insightful comments and administrative help at various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

Apoorva Srivastava

Suruchi Sinha

Table of Contents

CHAPTER 1

INTRODUCTION

Today the data is growing in a very high speed. These data can be produced by any sources like industries, social media, cell-phones, scientific source etc. We can refer this large data as Big Data. Till now there is no particular measure is defined on size of the Big Data. In the beginning, Big Data was adopted by Facebook, LinkedIn, Google etc. The reason might be the rapid change of the data. The three characteristics of Big Data are volume, velocity and variety. Volume refers to the size of the data. It is growing bigger day by day. According to the expert's analysis, few years later the data can cross 25 Zettabytes. Velocity refers to how fast the data is being processed. For this we can consider the examples like posting comment or image on Facebook and watching video on YouTube etc. Variety is referring to different types of data and different sources which produce these data. The data can be structured, semi-structured or unstructured. It can be in any formats like image, csv files, text files, audio, video etc. In addition to these characteristics, two more have been defined: veracity and value. Veracity refers to the trustworthiness of the data and value refers to extracting meaningful information from datasets.

As the data are produced, we need to store and analyse it. The traditional data processing techniques are failed to analyse the Big Data. Today about 80% of data are unstructured and these unstructured data are impossible to analyse using relational database systems. So Big Data analytics is introduced for processing these large amount of structured, unstructured and semi-structured data with the help of existing software tools and with very less amount of time. Big data analytics makes use of Hadoop framework for analyzing the larger datasets. Hadoop is an open-source, reliable, scalable and shared computing framework developed especially to process the huge amount of data. It was initially developed by Google in 2004 and at present it is maintained by Apache. The main two components Hadoop are Hadoop Distributed File System (HDFS) and MapReduce. HDFS stores the data in the form of blocks. MapReduce is used for processing, sharing and clustering.

Hadoop ecosystem provides several tools for Big Data analytics. Some of the tools are Hive, Pig, Hbase, Cassandra, Mahout, Flume, Avro etc.

Pig is an analysis tool of Hadoop ecosystem. Pig has its own programming language called Pig Latin. To covert the Pig Latin scripts into MapReduce job Pig

Runtime is used. Hive is called as data warehousing software. It has its own query language called HiveQL. It is used for processing the larger datasets stored indatawarehouse. Hbase is used to store the large amount of data especially structured data. It stores data in the form of tables and Pig or Hive queries can be used to analyse these data. Like Hbase, Cassandra is also a database which stores the structured data. In Casandra, the data is replicated in several nodes so even if the one node fails it doesn’t make any difference. Its data will be available in other nodes. ThereforeCassandra is called as fault-tolerant system. Hbase and Cassndra are NoSQL databases and they are column oriented. Mahout is a framework developed by Apache. The main purpose of this is to implement the data mining algorithms. Flume is another data analyzing tool provided by Hadoop eco-system. It is used especially to analyse the log data. Its architecture is very simple and it is robust. Avro is schema-dependent. It is mainly used to serialise the data so that it can analysed easily. It declares different data structures with the help of JSON format. Java, C, C++, Python, C# and Ruby are the languages supported by Avro.

In this project we have used Hive with Hadoop framework for analysing Indian Education dataset. Hive is built on the top of Hadoop and it has its own query language HiveQL which is similar to SQL. Hive will internally convert the queries into MapReduce job. The reason for why Hive is better than MySQL is that, Hive is most suitable for larger datasets and MySQL is suitable for smaller datasets.

CHAPTER 2

LITERATURE REVIEW

Ammar Fuad et al. proposed a method to analyze the performance MySQL cluster, Hive and Pig [7]. For the experiment three different sizes of movie lens datasets are considered. The result showed that MySQL cluster processing time increases as the data size increases. Because of step by step execution nature of the Pig, its processing time exceeded the processing time of the Hive.

Karan Sachdeva et al. compared the performance of MapReduce, Hive and Pig by considering unstructured, semi structured and structured dataset [8]. From the result it's been proved that to process structured data Hive is the efficient tool. For processing semi-structured and unstructured data Pig and Map-Reduce respectively are efficient.

Analysis of Meteorological and Oceanographic data is difficult using MySQL because it consists of several number of small files and each file might contain 20 to 300 columns. Ali Usman Abdullahi et al. have analyzed the Meteorological and Oceanographic data for indexed and non-indexed table using Hive [9]. There are three different types of queries have been executed on the tables. From the result it is shown that type 1 and type 3 queries have shown better response time for indexed table. The type 2 showed different processing time for indexed and non-indexed table depending upon the size of the data. It is possible to reduce the number of mappers so that response time can be increased.

Aditya Bhardwaj et al. have analyzed Twitter data using Hive [10]. Analysis is performed to predict the Map-Reduce time and total job completion time for different cluster size. From the result it is proved that as the cluster size increases the Map-Reduce time also increases and total job completion time decreases.

It is possible to increase the performance of Hadoop/Hive with help of Multi Query Optimization technique and distributed Hive. Varun Garg [11] considered 11 queries from TPC-H and different sizes of datasets are used. From the experiments the author has shown that performance of distributed Hive is greater than the conventional Hive.

Xiaoyu Wang et al. [12] performed analysis on Internet traffic data using Hadoop and Hive based traffic analysis system. The libpcap files are pre-processed with less amount of time. They have shown that the system proposed by them is error free and increases execution speed.

Taoying Liu et al. [13] presented the implementation Standard Science DBMS which is a benchmark of distributed scientific data on Hive. Hive queries are compared with SciDB queries. The amount of time taken to load the data by both SciDB and Hive is same. For the smaller input data, the performance of Hive is slower when compared toSciDB.

S K Pushpa et al. [14] analyzed airport data using Hive and found out that it is more efficient and faster when compared to traditional approach. Dharaben Patel et al. have analyzed the huge amount of network traffic data using Hive [15]. Hive queries are written to find-out different types of security attacks. To visualize the result Apache Zeppelin tool is used. As the part of future work using this method more number of security attacks can be found.

Hive performance time can be predicted by determining the Map-Reduce job execution time [16]. Amit Sangroya et al. have proposed a linear regression model. Hive processing time decreases with increase in the size of data. The proposed method helps in predicting the Hive performance time with reduced error rate.

The Connected Vehicles can exchange information about location and security. Large amount of data are produced by Connected Vehicle. Weija Xu et al. [17] analyzed these data using Hive and compared result with PostgreSQL. The experiment conducted showed that Hive query performance time is lesser than PostgreSQL.

The Earth science data are always in NetCDF format. This format is not supported in HDFS. Therefore, we cannot analyze these data using Hadoop tools. Shujia Zhou et al. [18] proposed a system that will convert the NetCDF format to CSV format making it easy to visualize and to analyze by Hadoop Tools like Spark, Hive.

S. Karimian-Aliabadi et al. continuous struggle of data scientists with increasing size of data to be analyzed, led to handful of practical tools and methods. In 2008, Dean and Ghemawat proposed MapReduce (MR) paradigm to process large amount of data on multiple node cluster to increase parallelism and therefor improved performance. [1]

The MR paradigm was not globally used until useful Hadoop framework developed in 2011 by Apache. The Hadoop Distributed File System (HDFS) is a primary layer of the Hadoop ecosystem but not the only one. In 2013, Vavilapalli et al., introduced YARN layer to the Hadoop cluster in order to specialize the resource management and make it dynamic rather than Hadoop's earlier static allocation scheme. With more complex dataflow in MR applications there was a need to cut down the complexity into multiple stages and thus Directed Acyclic Graphs (DAG) was chosen by Tez developers to demonstrate the dataflow between stages of a complex application. Taking advantage of the memory’s high speed and the Resilient Distribute (RDD) concept, Spark was created and became popular due to high speed and the ease of application development**.**[2]

Tuning the framework and cluster parameters in order to reduce the execution time of a BigData application was a challenge from the earliest steps and a main part of this optimization process is to predict the execution time for a given set of parameters. But with each step in development of a more advanced framework for processing BigData, new set of parameters and complexity is created and execution time prediction made more and more challenging. A lot of work have been done in literature to simulate, model, or learn the process, but their accuracy and scalability is only enough for simple runs with a single job running by one or more users and not for more complex applications with multiple multi-stage jobs running by number of users. [3]

In today’s world there is a huge growth in data. This data is generated from variety of sources like social media, industry, transaction records, cell phone, GPS signals etc. It is difficult and challenging to store such a huge amount data in traditional data warehouse. Big Data is the dataset with 3 V’s that are Volume, Variety and Velocity and difficult to store and process using traditional database management systems. Big Data Analytics is the way of processing the large amount of data. Hadoop is a popular opensource software which is very useful in analyzing the larger data. Hadoop provides several tools for this purpose like Hive, Pig, HBase, Cassandra etc. In this paper, we have used Hadoop framework. For the analysis of movie dataset Hive tool is used with Hadoop framework. We have got significant improvement in processing time for analyzing dataset compared to traditional system.[6]

HDFS needs to work with massive amounts of data stored in very large files. When dealing with large HDFS files, MapReduce splits the files into multiple pieces at record boundaries, so it can read data from the large file simultaneously by starting multiple mapper processes. A splitable data format lets a file be correctly split into pieces at the record boundaries. Hadoop environments prefer to use binary formats rather than text formats when dealing with HDFS, because binary formats prevent incomplete records being written to files, by catching and ignoring incorrect records that may be created due to data corruption or incompleteness. This type of issue can occur, for example, when a cluster accidentally runs out of space during a write. Data compression capability is also a key requirement for a good HDFS file format. A popular binary format used by many is the Avro container file format, which is splitable and can also be compressed. Another common HDFS data format is a Sequence File, which is a splitable file format represented as a list of keys and values. Users can also customize the data format by using serializers, which let them write data in any format they choose.

In paper, predictive models for the box office performance of the movies was represented by factors derived from social media and IMDb. According to our models, we have identified the following patterns: the popularity of leading actress is crucial to the success of a movie, the combination of past successful genre and a sequel movie is another pattern for success, a new movie in the not popular genre and an actor with low popularity could be a pattern for a Flop. It is surprising that sentiment score and view and comment counts were not identified as relevant in our experiments. Author believe it is related to how weights are assigned to each attribute. Further studies to determine different weighting methods will be beneficial. In addition, our prediction is for movies yet to be released. The preliminary result of tracking 13 of the movies shows a good prediction performance from our model. A follow-up study on the final performance of our models will be validated and presented once all of the movies are released. Future work to improve our models will include further refinement of the Neutral class and characterization of movie box office performance in terms of net profits and profit ratios.

Apache Hive provides a SQL interface that enables you to use HDFS data without having to write programs using MapReduce. It's important to understand that unlike Apache HBase, Hive is not a database. It simply provides a mechanism to project a database structure on data you store in HDFS and lets you query that data using HiveQL, a SQL-like language. Hive uses a type of SQL that lets you query HDFS data in ways that are similar to how you query data stored in a relational database. While HiveQL doesn't have the full range of features available in SQL, it offers more than enough SQL capabilities for you to efficiently work with HDFS data. When you use a Hive query, Hive parses the SQL query and generates a MapReduce job to process the data to get you the query results. The main rationale for Hive is to reduce effort by doing away with developing MapReduce programs. It also provides a data warehouse capability when handling large amounts of data, is analyst friendly and is ideal for making use of HDFS data for business intelligence (BI) analysis.[5]

Two mainstream approaches for large-scale data analysis are parallel database systems and MapReduce-based systems. Both approaches share certain common design elements: they both employ a shared-nothing architecture, and deployed on a cluster of independent nodes via a high-speed interconnecting network; both achieve parallelism by partitioning the data and processing the query in parallel on each partition. However, parallel database approach has major limitations on managing and querying spatial data at massive scale. Parallel database management systems (DBMSs) tend to reduce the I/O bottleneck through partitioning of data on multiple parallel disks and are not optimized for computational-intensive operations such as spatial and geometric computations. Partitioned parallel DBMS architecture often lacks effective spatial partitioning to balance data and task loads across database partitions. While it is possible to induce a spatial partitioning, fixed grid tiling, for example, and map such partitioning to one dimensional attribute distribution key, such an approach fails to handle boundary objects for accurate query processing. Scaling out spatial queries through a parallel database infrastructure is possible while being costly, and such approach is explored.[4]

CHAPTER 3

REQUIREMENT GATHERING AND ANALYSIS

3.1 Hardware Requirements

- RAM 4GB – 8GB
- Operating system (32bit or 64 bit)
- 2 GHz or Higher Processor
- 20 GB Hard Disk

3.2 Software Requirements

1. Oracle VM VirtualBox

Oracle VM VirtualBox is cross-platform virtualization software that allows users to extend their existing computer to run multiple operating systems at the same time. Designed for IT professionals and developers, Oracle VM VirtualBox runs on Microsoft Windows, Mac OS X, Linux, and Oracle Solaris systems and is ideal for testing, developing, demonstrating, and deploying solutions across multiple platforms on one machine.

Oracle VM VirtualBox has been designed to take advantage of the innovations introduced in the x86 hardware platform, and it is lightweight and easy to install and use. Yet under the simple exterior lies an extremely fast and powerful virtualization engine. With a well-earned reputation for speed and agility, Oracle VM VirtualBox contains innovative features to deliver tangible business benefits: excellent performance; a powerful virtualization system; and a wide range of supported guest operating system platforms.VirtualBox is used to setup the virtual hadoop servers.

2. CentOS 7

CentOS is a community-driven free software effort that provides two Linux distribution (CentOS Linux and CentOS Stream) and a variety of Special Interest Groups releasing packages to run on those distributions. CentOS Linux provides a free, community-supported computing platform functionally compatible with its upstream source, Red Hat Enterprise Linux (RHEL).CentOS Stream is a continuously delivered distribution that tracks just ahead of RHEL and acts as an upstream for RHEL development.

3. Apache NiFi version-1.9.0

Apache NiFi supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic. Some of the high-level capabilities and objectives of Apache NiFi include:

a) Web-based user interface
    - Seamless experience between design, control, feedback, and monitoring

b) Highly configurable
    - Loss tolerant vs guaranteed delivery
    - Low latency vs high throughput
    - Dynamic prioritization
    - Flow can be modified at runtime
    - Back pressure

c) Data Provenance
    - Track dataflow from beginning to end

d) Designed for extension
    - Build your own processors and more
    - Enables rapid development and effective testing

e) Secure

4. Apache Hadoop

- Hadoop is an open source framework utilized for processing humungous datasets and also used for distributed storage.
- A particular special type of computational cluster is built in order to store and analyze large volumes of unstructured data is known as a Hadoop cluster.
- Hadoop clusters are gaining popularity for enhancing the speed of data analysis applications. Hadoop clusters are extremely scalable.
- Hadoop clusters are highly efficient as they are resistant to failures.

5. Apache Hive

- Hive is a data warehouse system for Hadoop.
- It allows querying, data analysis utilizing HiveQLetc.
- Hive enables users to potray structure on huge unstructured data.
- Hive has the ability to understand organized and unorganized data which may include text files where fields are circumscribed by specific characters.

6. Tableau

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data in a very easily understandable format. Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards.

The best features of Tableau software are-

- Data Blending
- Real time analysis
- Collaboration of data

3.3 Data Requirements

The dataset requirement for our project is fulfilled through kaggle repository. From here we downloaded the dataset and used as an input. Kaggle.com is a website that provides dataset for free for its users. Thus we got dataset for free of cost.

CHAPTER 4

Design

4.1 Data Flow Diagram

- **0-level DFD**

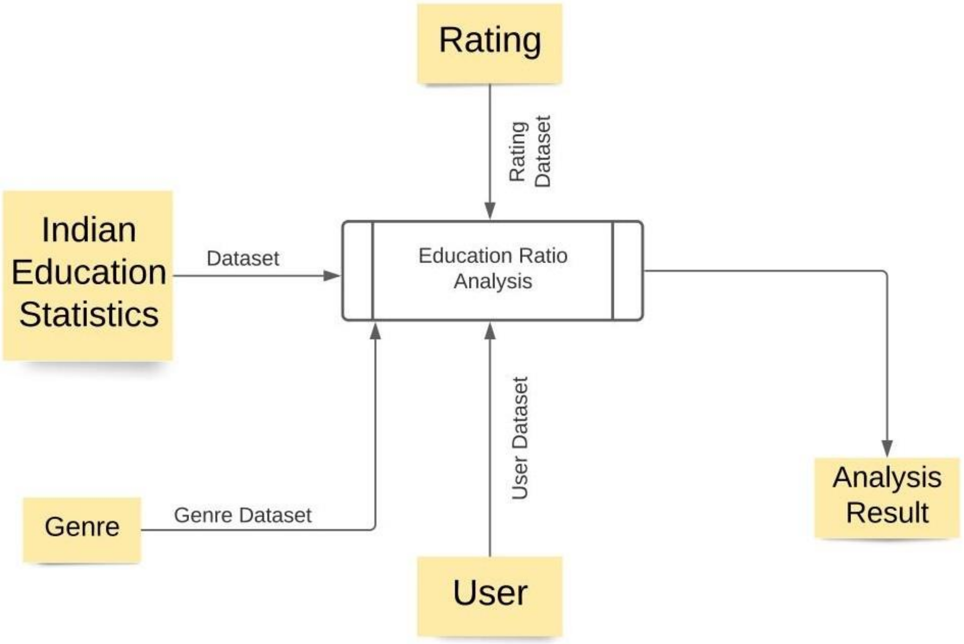


Figure: 4.1.1 0-level DFD

- **<u>1-level DFD</u>**

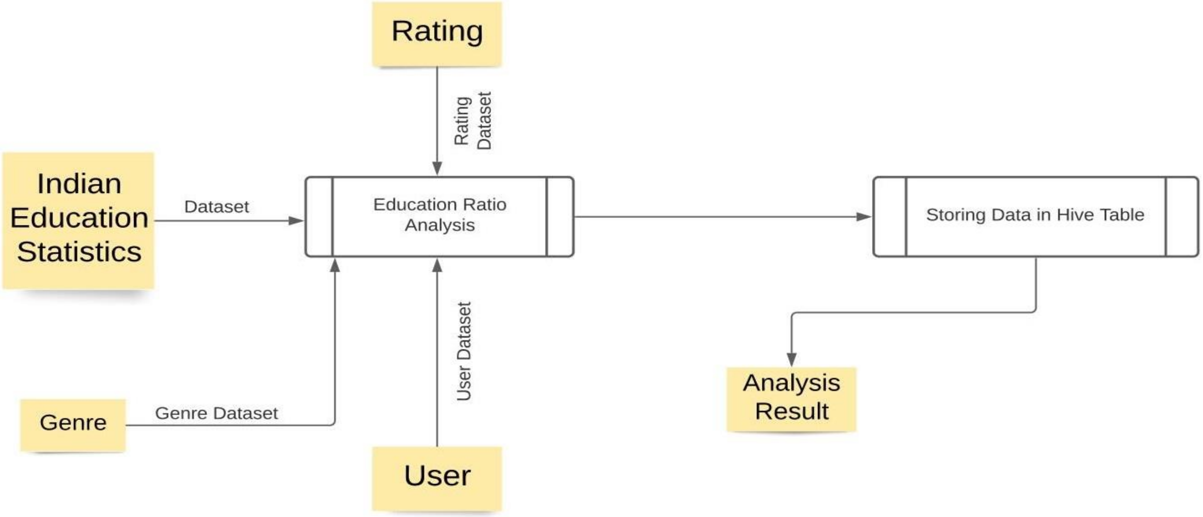


Figure: 4.1.2 1-level DFD

4.2 Data Dictionary

The datasets are taken from kaggle. The Movie dataset consists of information about the year of release, title, language, imdb ratings, FB likes, genre etc. In this report we have considered four datastes for analysis: Movie dataset, Genre dataset, Rating dataset, User dataset. Movie dataset consists of 10650 rows, Genre dataset consists of 5044 rows, Rating dataset consists of 6-7 lakh rows and User dataset consists of 6040 rows. The description about the datasets is as follows.

STATE CODE	Code of the State
DISTRICT CODE	Code of the District
COUNTRY	Country to which the state belongs
AGE GROUP	Age of the Person
ANALYSIS	Literate or Illiterate

Figure: 4.2.1 Indian Education dataset

STATE	State to which user belongs.
ABILITY	Ability to read or write or both.
RATINGS	The rating given by the user based on abilities.

Figure: 4.2.2 Rating dataset

GENDER	Describes whether user is male or female.
AGE	Age of the person.
AGE GROUP	Describes the age group to which a person belongs.

Figure: 4.2.3 Age dataset

USER ID	Unique ID for a particular user.
GENDER	Describes whether user is male or female.

Figure: 4.2.4 User dataset

CHAPTER 5

Project Workflow

5.1 Setup Linux Machine using Oracle VirtualBox and CentOS

- **Virtual Box**

VirtualBox is a powerful x86 and AMD64/Intel64 virtualisation product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of guest operating systems including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7, Windows 8, Windows 10), DOS/Windows 3.x, Linux (2.4, 2.6, 3.x and 4.x), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever-growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

Figure 5.1.1 Oracle VM VirtualBox

- **CentOS**

The CentOS Project invites you to be a part of the community as a contributor. There are many ways to contribute to the project, from documentation, QA, and testing to coding changes for SIGs, providing mirroring or hosting, and helping other users.

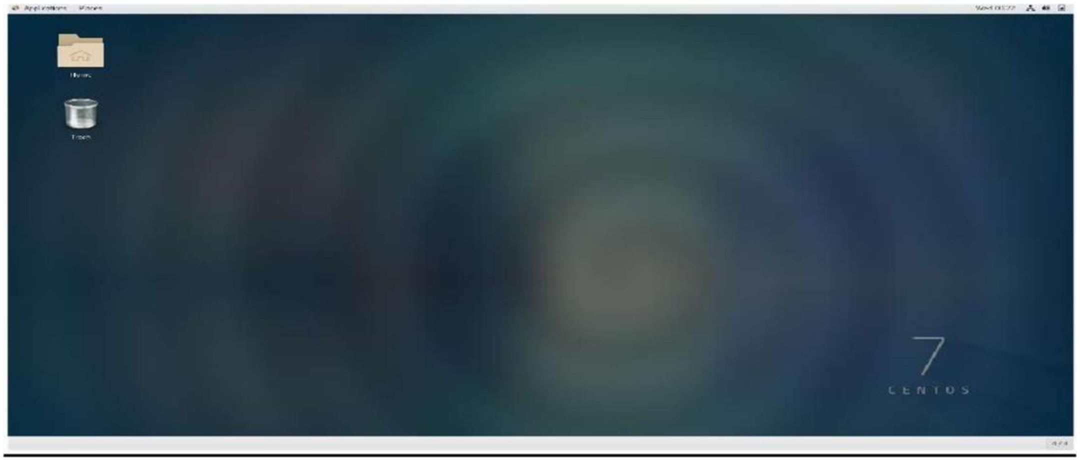


Figure 5.1.2 CentOS

- **Installation:**

Download the appropriate version of both the packages. The 64-bit architecture has been used for this demonstration, so download the software accordingly.

Install VirtualBox and open it. Click on the **new** to set up the new VM.

⬇

The ISO image of the downloaded CentOS has to be linked to the newly created virtual machine.

⬇

Select “**Install CentOS Linux 7**” and proceed.

⬇

After finishing the initial setup, you need to execute some additional steps. First, you need to accept the OS EULA.

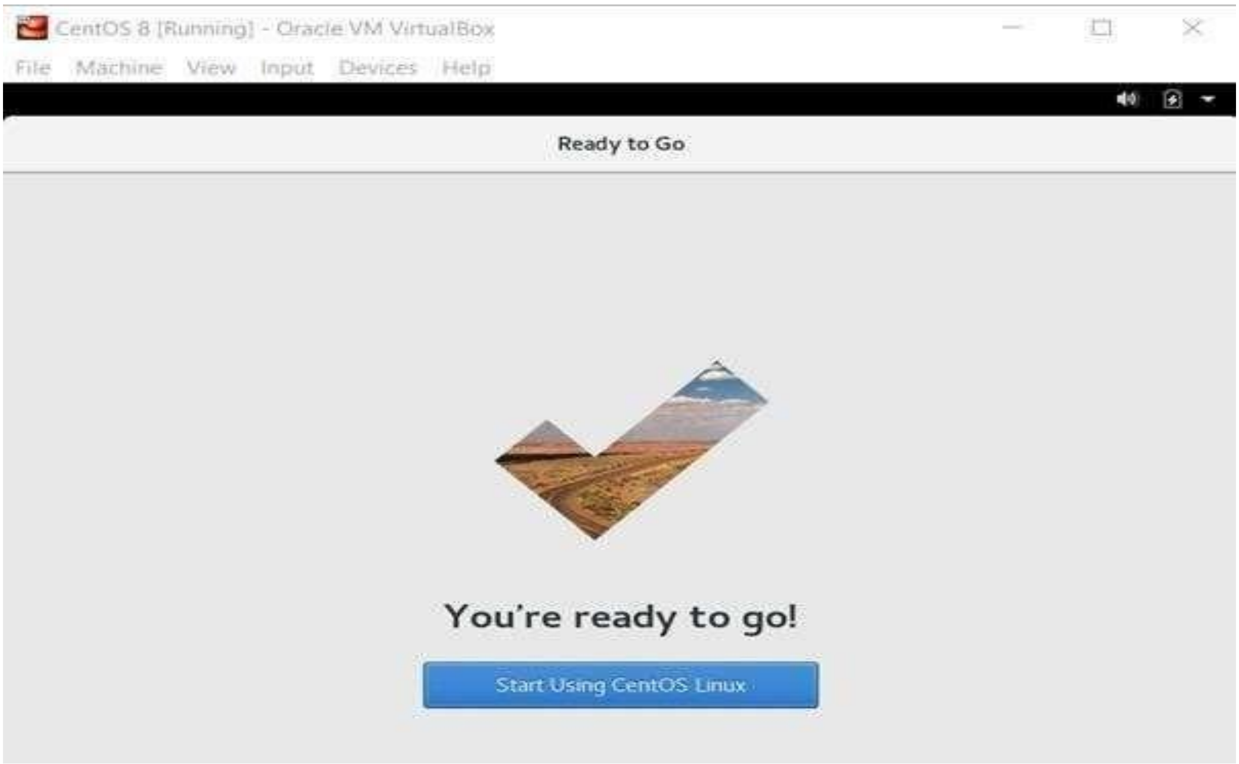

CentOS 8 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Ready to Go
You're ready to go!
Start Using CentOS Linux

Figure 5.1.3 Installed CentOS

5.2 Hadoop Installation and Single Node Cluster Setup

- **Hadoop**

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

- **Hadoop Installation**

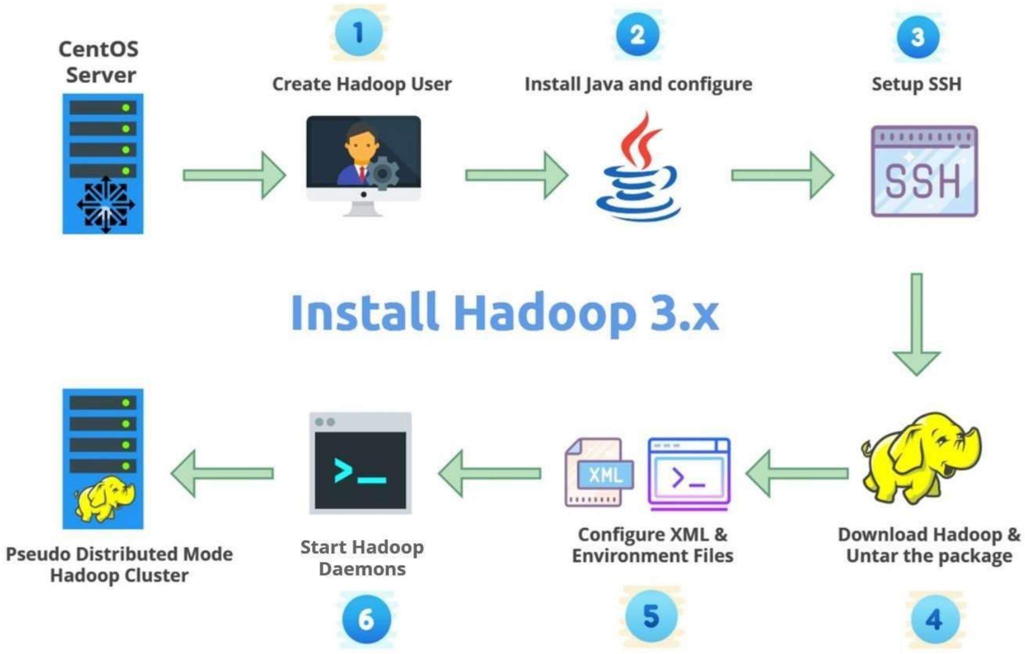

CentOS Server
1
Create Hadoop User
2
Install Java and configure
3
Setup SSH
SSH
Install Hadoop 3.x
Download Hadoop & Untar the package
4
XML
Configure XML & Environment Files
5
Start Hadoop Daemons
6
Pseudo Distributed Mode Hadoop Cluster

Figure 5.2.1 Hadoop Installation

- **<u>Hadoop Architecture</u>**

Hadoop has a Master-Slave Architecture for data storage and distributed data processing using MapReduce and HDFS methods. The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System). The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.

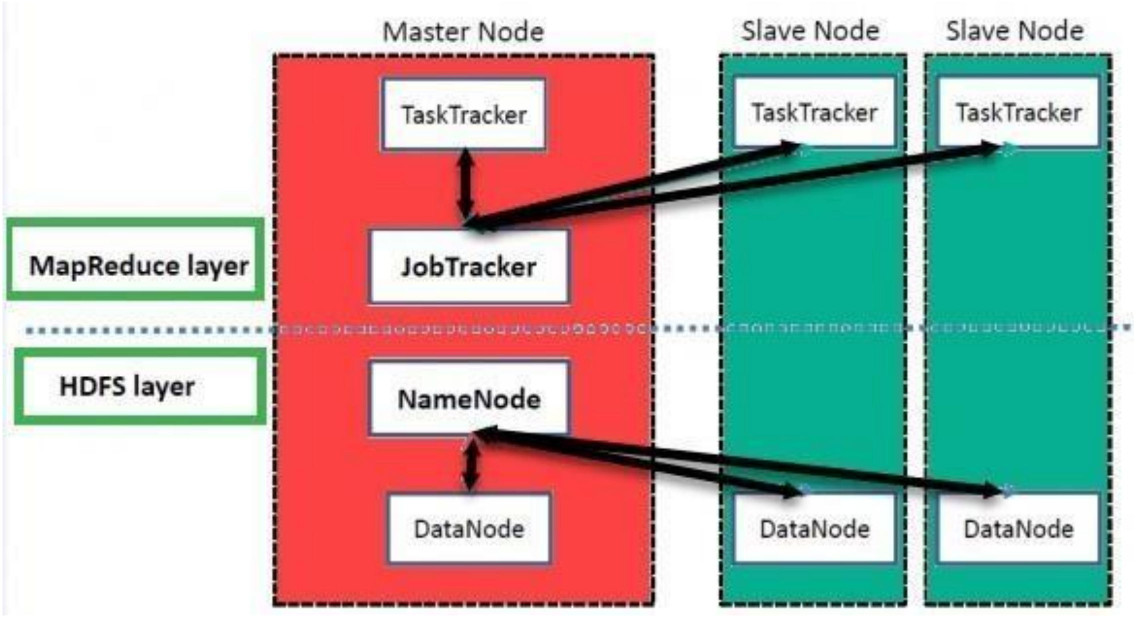


Figure 5.2.2 Hadoop Architecture

- **Single Node Cluster**

Single node cluster means only one DataNode running and setting up all the NameNode, DataNode, ResourceManager and NodeManager on a single machine.
It can easily and efficiently test the sequential workflow in a smaller environment as compared to large environments which contains terabytes of data distributed across hundreds of machines.

Hadoop cluster consists of three components –

- Master Node – Master node in a hadoop cluster is responsible for storing data in HDFS and executing parallel computation the stored data using MapReduce. Master Node has 3 nodes – NameNode, Secondary NameNode and JobTracker. JobTracker monitors the parallel processing of data using MapReduce while the NameNode handles the data storage function with HDFS. NameNode keeps a track of all the information on files (i.e. the metadata on files) such as the access time of the file, which user is accessing a file on current time and which file is saved in which hadoop cluster. The secondary NameNode keeps a backup of the NameNode data.

- Slave/Worker Node- This component in a hadoop cluster is responsible for storing the data and performing computations. Every slave/worker node runs both a TaskTracker and a DataNode service to communicate with the Master node in the cluster. The DataNode service is secondary to the NameNode and the TaskTracker service is secondary to the JobTracker.

- Client Nodes – Client node has hadoop installed with all the required cluster configuration settings and is responsible for loading all the data into the hadoop cluster. Client node submits mapreduce jobs describing on how data needs to be processed and then the output is retrieved by the client node once the job processing is completed.

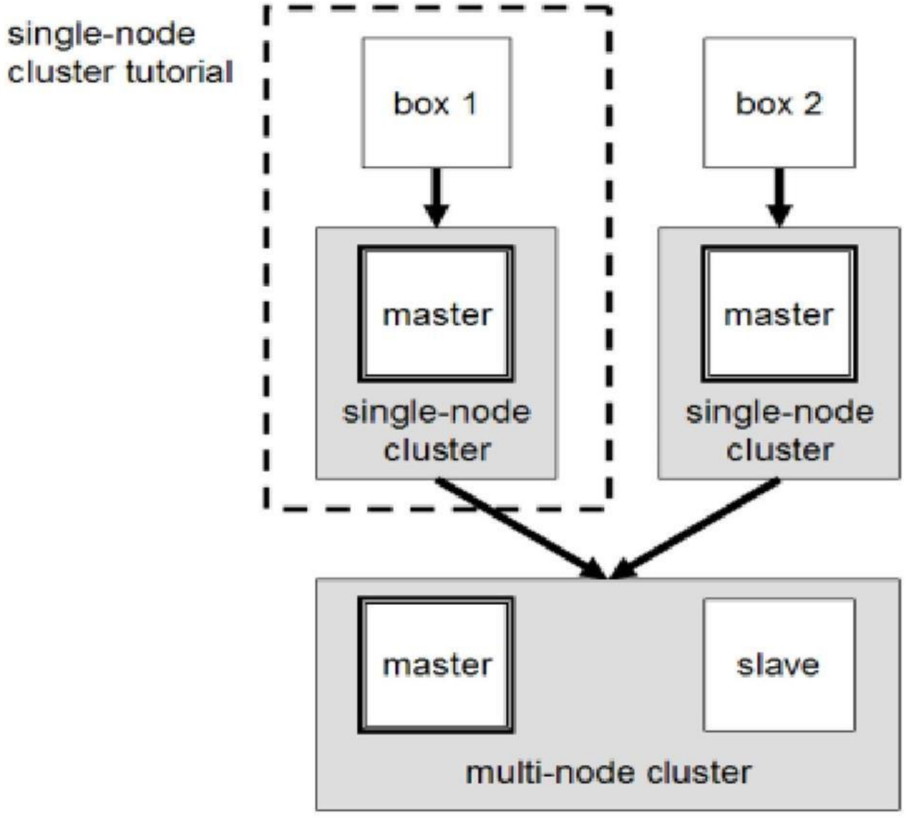
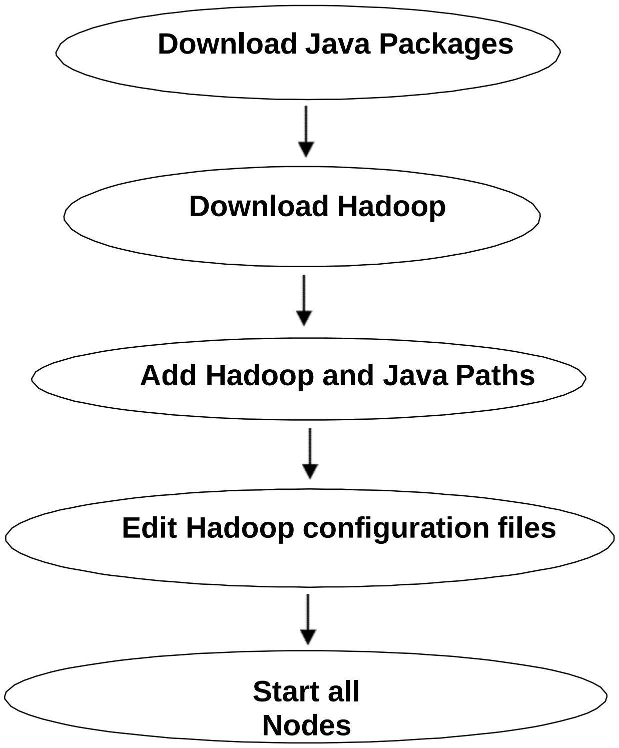
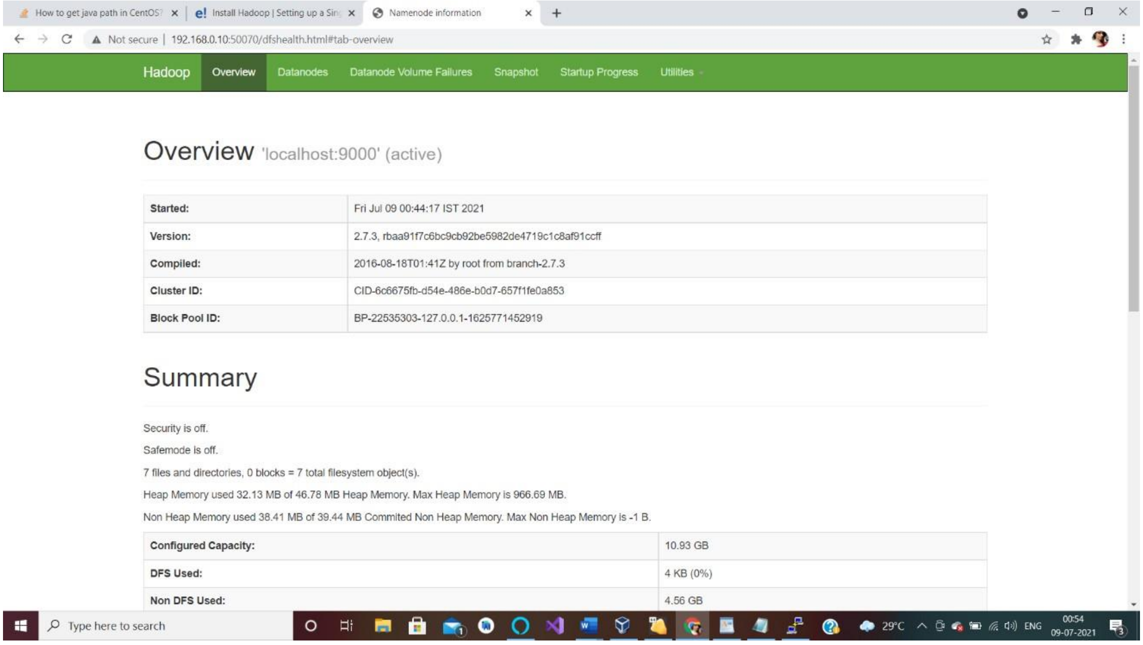


Figure 5.2.3 Single Node Cluster

- **Single Node Cluster Setup**





Overview 'localhost:9000' (active)

Started:	Fri Jul 09 00:44:17 IST 2021
Version:	2.7.3, rbaa91f7c6bc9cb92be5982de4719c1c8af91ccff
Compiled:	2016-08-18T01:41Z by root from branch-2.7.3
Cluster ID:	CID-6c6675fb-d54e-486e-b0d7-657f1fe0a853
Block Pool ID:	BP-22535303-127.0.0.1-1625771452919

Summary

Security is off.

Safemode is off.

7 files and directories, 0 blocks = 7 total filesystem object(s).

Heap Memory used 32.13 MB of 46.78 MB Heap Memory. Max Heap Memory is 966.69 MB.

Non Heap Memory used 38.41 MB of 39.44 MB Commited Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	10.93 GB
DFS Used:	4 KB (0%)
Non DFS Used:	4.56 GB

Figure 5.2.4 Single Node Cluster Setup

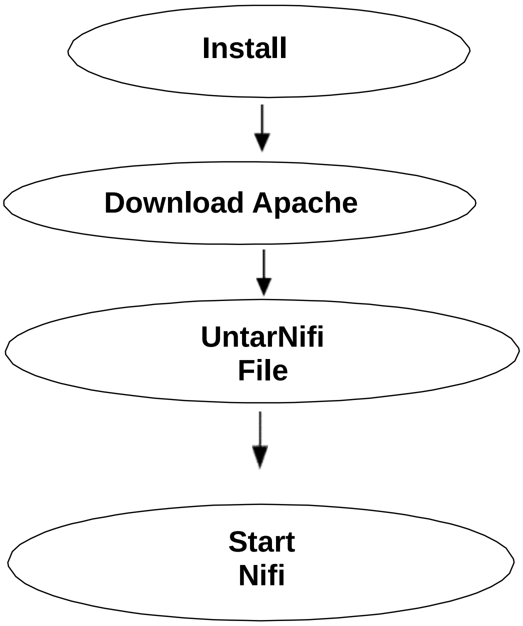
5.3 Nifi Installation

- **Apache Nifi**

Apache NiFi is an open source data ingestion platform. It was developed by NSA and is now being maintained and further development is supported by Apache foundation. It is based on Java, and runs in Jetty server. It is licensed under the Apache license version 2.0.The guide we are giving in this tutorial is intended to provide knowledge how to work with NiFi. To work with NiFi, you should have the basic knowledge of Java, Data ingestion, transformation, and ETL. You should also be familiar with the regex pattern, web server, and platform configuration.

Figure 5.3.1 Apache Nifi

- **Nifi Setup**



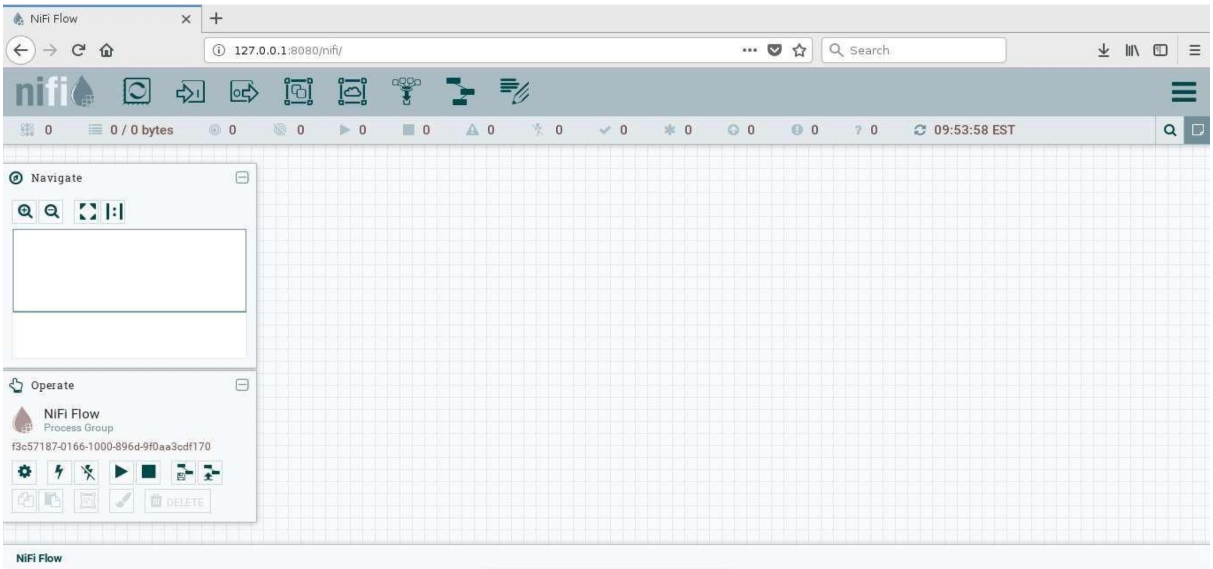


Figure 5.3.2 Nifi Setup

5.4 Hive Installation

- **Apache Hive**

Apache Hive is a data ware house system for Hadoop that runs SQL like queries called HQL (Hive query language) which gets internally converted to map reduce jobs. Hive was developed by Facebook. It supports Data definition Language, Data Manipulation Language and user defined functions. Hive abstracts the complexity of Hadoop MapReduce. Basically, it provides a mechanism to project structure onto the data and perform queries written in HQL (Hive Query Language) that are similar to SQL statements. Internally, these queries or HQL gets converted to map reduce jobs by the Hive compiler.

Figure 5.4.1 Apache Hive

- **<u>Hive Architecture</u>**

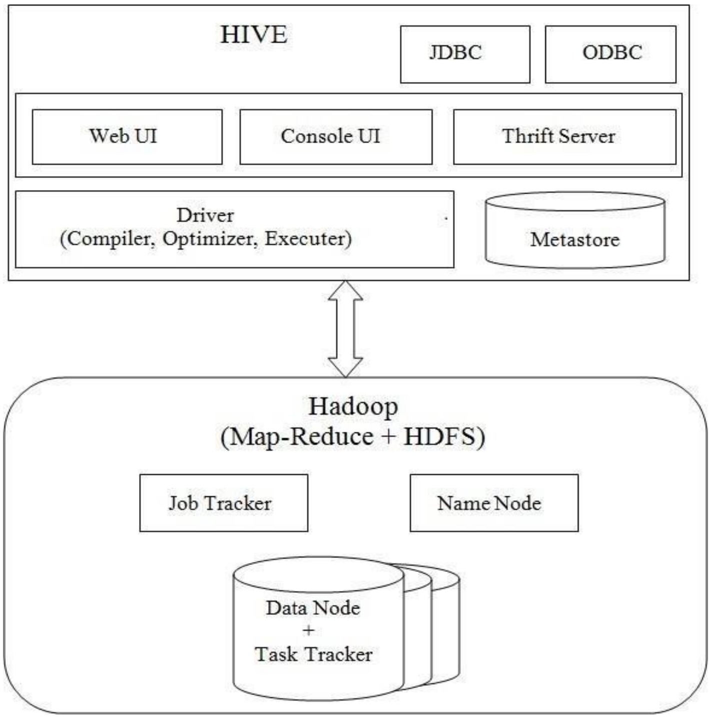
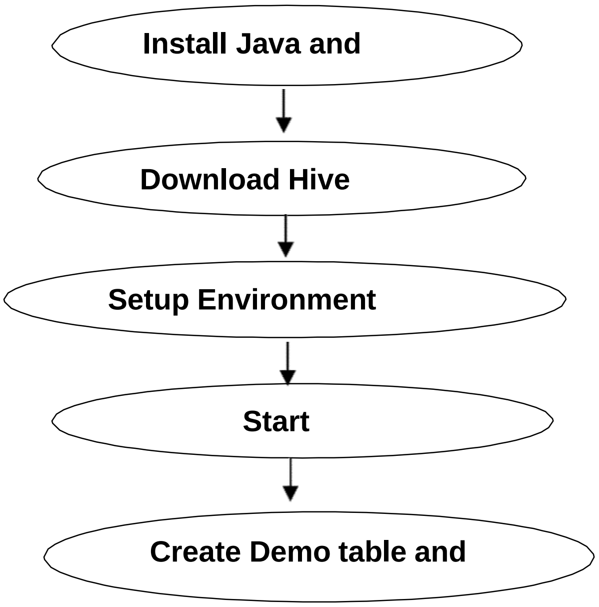
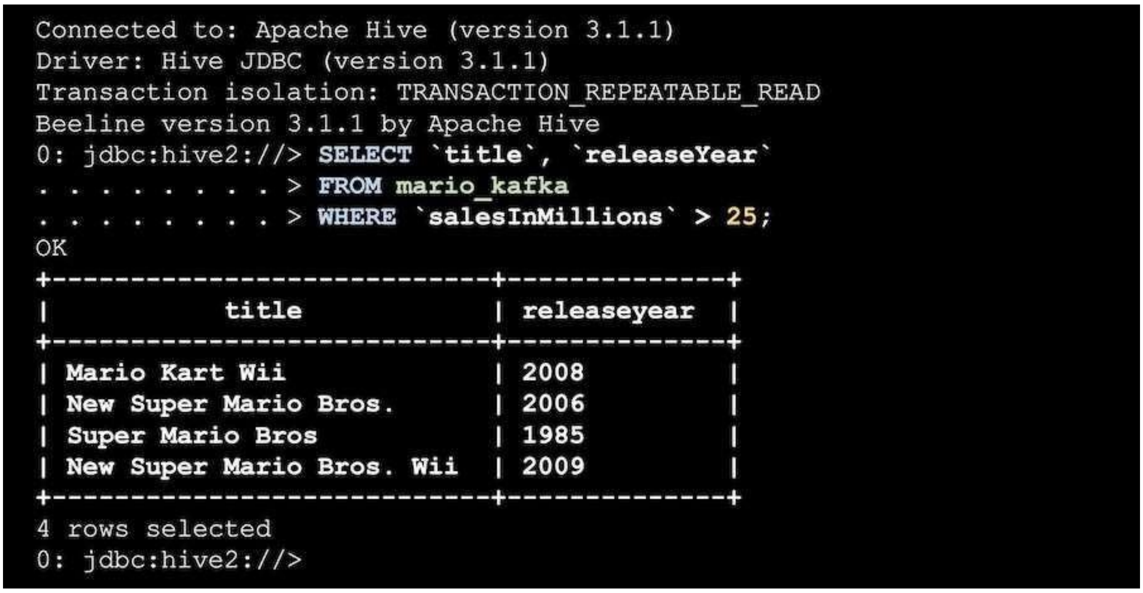


Figure 5.4.2 Hive Architecture

- **<u>Hive Setup</u>**





```
Connected to: Apache Hive (version 3.1.1)
Driver: Hive JDBC (version 3.1.1)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.1 by Apache Hive
0: jdbc:hive2://> SELECT `title`, `releaseYear`
. . . . . . . . > FROM mario_kafka
. . . . . . . . > WHERE `salesInMillions` > 25;
OK
+---------------------------+--------------+
|           title           | releaseyear  |
+---------------------------+--------------+
| Mario Kart Wii            | 2008         |
| New Super Mario Bros.     | 2006         |
| Super Mario Bros          | 1985         |
| New Super Mario Bros. Wii | 2009         |
+---------------------------+--------------+
4 rows selected
0: jdbc:hive2://>
```

Figure 5.4.3 Hive Setup

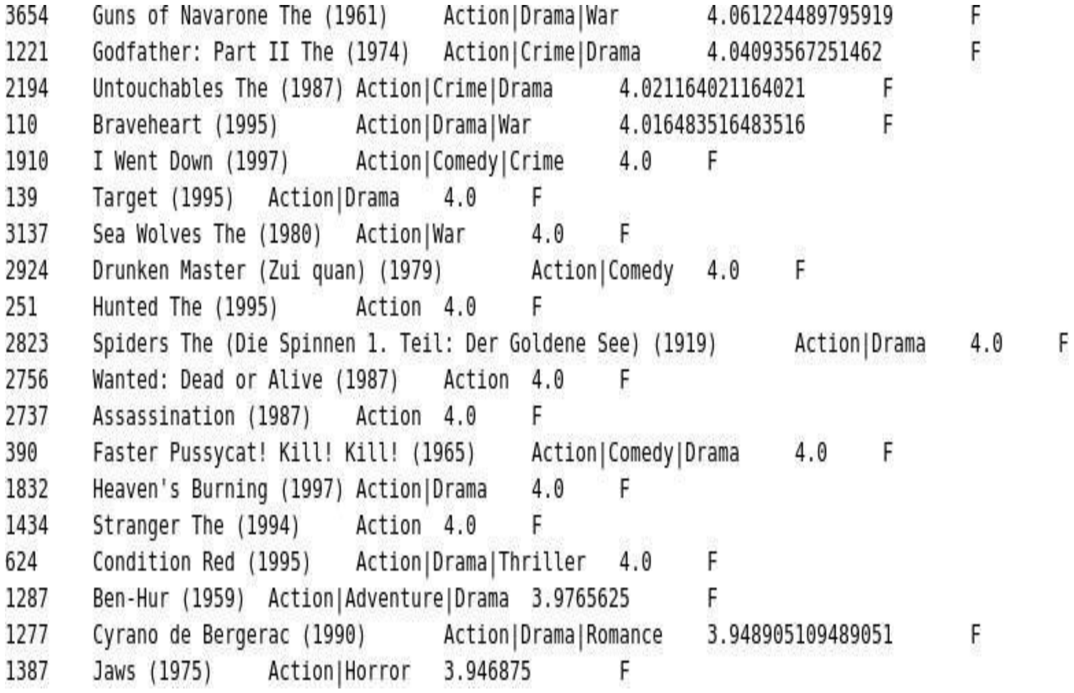
5.5 Analysis of Data

The analysis involves following steps.

- Rating of action movies on user gender.
- Rating of adventure movies on user gender.
- Highest rated movie in a year.
- Maximum FB likes in a year.

For analyzing dataset, we have created the views of each table. The purpose of creating the view is to minimize the execution time. For example if there is a table with 1000 columns and we want to access 10 columns from that. In this case a view of those 10 columns is created. Instead of scanning the entire table for 10 columns, query can be executed on the view to minimize the execution time. In our analysis these views are then joined on the condition to get the result.

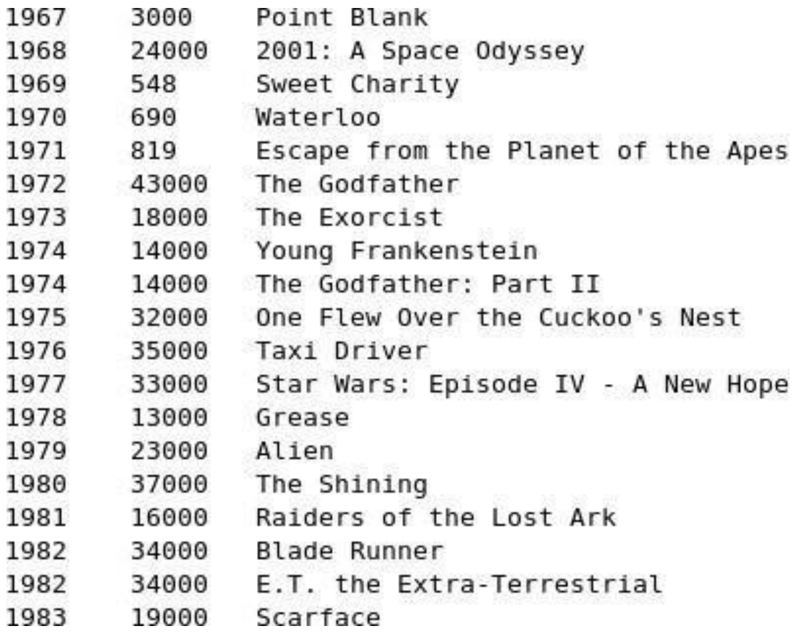
The figure 5.5.1 shows the result of average rating given by the female gender for each action category movie. We have used three views: Movie_Action which has the columns movie_id, title, genre as Action from Genre dataset. Female_view has the colums user_id and gender as F from user dataset. Rat_F has movie_id, ratings, gender.



```
3654    Guns of Navarone The (1961)     Action|Drama|War        4.061224489795919       F
1221    Godfather: Part II The (1974)   Action|Crime|Drama      4.04093567251462        F
2194    Untouchables The (1987) Action|Crime|Drama      4.021164021164021       F
110     Braveheart (1995)       Action|Drama|War        4.016483516483516       F
1910    I Went Down (1997)      Action|Comedy|Crime     4.0     F
139     Target (1995)   Action|Drama    4.0     F
3137    Sea Wolves The (1980)   Action|War      4.0     F
2924    Drunken Master (Zui quan) (1979)        Action|Comedy   4.0     F
251     Hunted The (1995)       Action  4.0     F
2823    Spiders The (Die Spinnen 1. Teil: Der Goldene See) (1919)       Action|Drama    4.0     F
2756    Wanted: Dead or Alive (1987)    Action  4.0     F
2737    Assassination (1987)    Action  4.0     F
390     Faster Pussycat! Kill! Kill! (1965)     Action|Comedy|Drama     4.0     F
1832    Heaven's Burning (1997) Action|Drama    4.0     F
1434    Stranger The (1994)     Action  4.0     F
624     Condition Red (1995)    Action|Drama|Thriller   4.0     F
1287    Ben-Hur (1959)  Action|Adventure|Drama  3.9765625       F
1277    Cyrano de Bergerac (1990)       Action|Drama|Romance    3.948905109489051       F
1387    Jaws (1975)     Action|Horror   3.946875        F
```

Figure 5.5.1 Average Rating By Female Gender

The figure 5.5.2 shows the result of which movie has got maximum Facebook likes in a particular year. The views created here are Movie_FB with year, title and fb_likes from Movie dataset. Movie_FB_Max with year and maximum fb_likes for that year.



```
1967    3000    Point Blank
1968    24000   2001: A Space Odyssey
1969    548     Sweet Charity
1970    690     Waterloo
1971    819     Escape from the Planet of the Apes
1972    43000   The Godfather
1973    18000   The Exorcist
1974    14000   Young Frankenstein
1974    14000   The Godfather: Part II
1975    32000   One Flew Over the Cuckoo's Nest
1976    35000   Taxi Driver
1977    33000   Star Wars: Episode IV - A New Hope
1978    13000   Grease
1979    23000   Alien
1980    37000   The Shining
1981    16000   Raiders of the Lost Ark
1982    34000   Blade Runner
1982    34000   E.T. the Extra-Terrestrial
1983    19000   Scarface
```

Figure 5.5.2 Maximum Facebook Like

5.6 Visualization

- **<u>Tableau</u>**

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data in a very easily understandable format. Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards. Data analysis is very fast with Tableau tool and the visualizations created are in the form of dashboards and worksheets.

Figure 5.6.1 Tableau

- **<u>Visualization Outputs</u>**

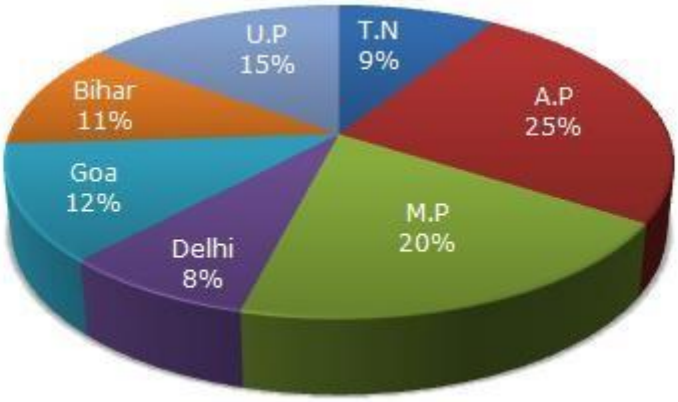


Figure 5.6.2.1 State wise Literacy

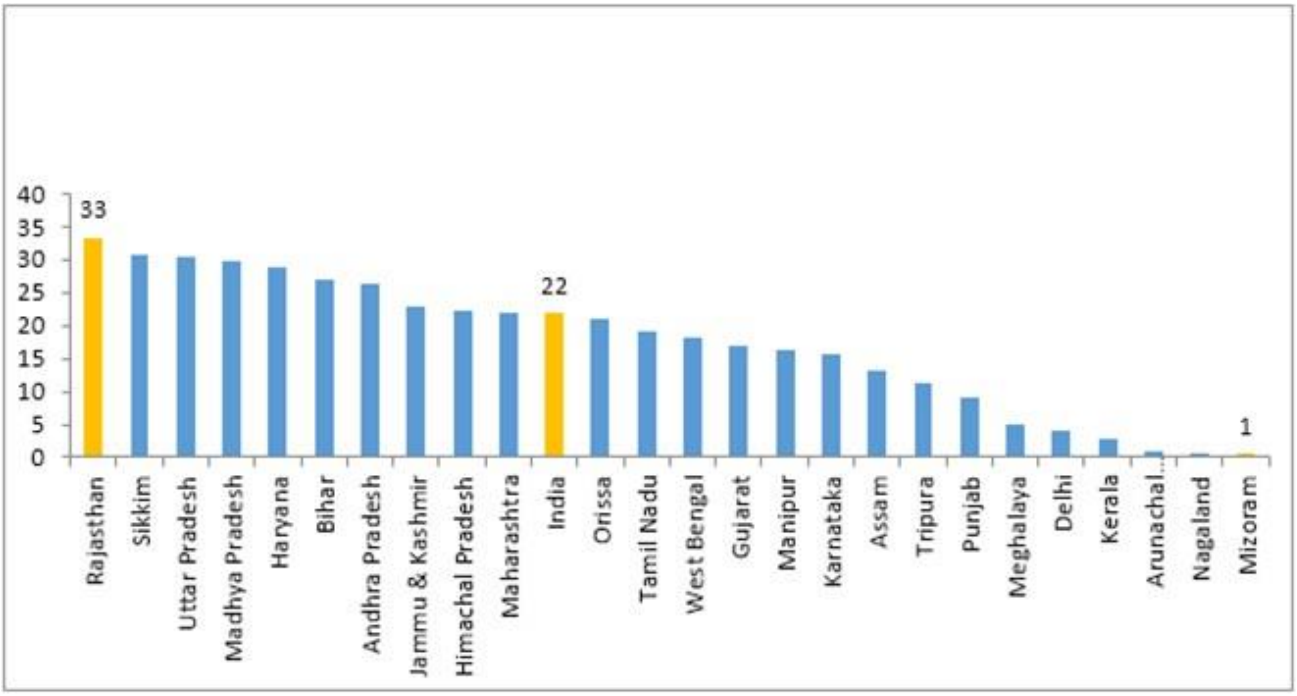


Figure 5.6.2.2 Graph Representation

CHAPTER 6

CONCLUSION AND FUTURE WORK

Hive is mainly used to process large amount of data. It is faster than SQL on low-cost machine. Hive performance is poor on smaller dataset but as the data size increases its processing time decreases. SQL is efficient and more robust for smaller data.

In order to find the relationship between state, district, country, age group, literacy, the dataset is analyzed using HiveQL and its performance is compared with SQL performance. The result states that Hive performs better than SQL for larger dataset. For the future work we can consider various countries and all other state of different countries as well.

REFERENCES

[1] Lucio Grandinetti, Seyedeh Leili Mirtaheri, Reza Shahbazian (Eds.), “High- Performance Computing and Big Data Analysis”, Second International Congress, TopHPC 2019 Tehran, Iran, April 23–25, 2019 Revised Selected Papers, Springer.

[2] Zongben Xu, Xinbo Gao, Qiguang Miao, Yunquan Zhang, Jiajun Bu (Eds.)., “Big Data”, 6th CCF Conference, Big Data 2018 Xi’an, China, October 11–13, 2018 Proceedings,Springer.

[3] Mohammed M. Alani, Hissam Tawfik, Mohammed Saeed, Obinna Anya, “Applications of Big Data Analytics”, Trends, Issues, and Challenges, Springer.

[4] FEI HU, “Big Data Sharing Storage and Security”, CRC Press Taylor & Francis Group.

[5] Peng, Xiao, Shao Liangshan, and Li Xiuran, “Improved Collaborative Filtering Algorithm in the Research and Application of Personalized Movie Recommendations”.

[6] Ashwitha T AAnisha P Rodrigues Niranjan N Chiplunkar, “Movie Dataset Analysis using Hadoop-Hive.

[7] Ammar Fuad, Alva Erwin, Henru Purnomo Ipung, “ Processing Performance on Apache Pig, Apache Hive and MySQL Cluster”, IEEE, 2014 International Conference on Information, Communication Technology and System.

[8] Karan Sachdeva et al., “Comparison of Data Processing Tools in Hadoop”, IEEE, 2016 International Conference on Electrical, Electronics,Communication, Computer and Optimization Techniques.

[9] Ali Usman Abdullahi, Rohiza Ahmad, Nordin M Zakaria, “Big Data: Performance Profiling of Meteorological and Oceanographic Data on Hive”, IEEE, 2016 3rd International Conference On Computer And Information Sciences.

[10] Aditya Bhardwaj et al., “Big Data Emerging Technologies: A CAseStudy with Analyzing Twitter Data using Apache Hive”, IEEE, 2015 RAECS UIET Panjab University Chandigarh.

[11] Varun Garg, “Optimization of Multiple Queries for Big Data with Apache Hadoop/Hive”, IEEE, 2015 International Conference on Computational Intelligence and Communication Networks.

[12] Abdeltawab M. Hendawi et al., “Hobbits: Hadoop and Hive Based Internet Traffic Analysis”, 2016 IEEE International Conference on Big Data.

[13] Taoying Liu, Jing Liu, Hong Liu, Wei Li, “A Performance Evaluation of Hive for Scientific Data Management”, 2013 IEEE International Conference on Big Data.

[14] S K Pushpa, Manjunath T N, Srividhya, “Analysis of Airport Data using Hadoop- Hive: A Case Study”, International Journal of Computer Applications (0975– 8887) National Conference on “Recent Trends in Information Technology” (NCRTIT-2016).

[15] Dharaben Patel, Xiaohong Yuan, Kaushik Roy, Aakiel Abernathy, “Analyzing Network Traffic Data Using Hive Queries”, 978-1-5386-1539-3/17/2017 IEEE.

[16] Amit Sangroya, Reha Singhal, “Performance Assurance Model for HiveQL on Large Data Volume”, 2015 IEEE 22nd International Conference on High Performance Computing Workshops.

[17] Weijia Xu et al., “Supporting Large Scale Connected Vehicle Data Analysis using Hive”, 2016 IEEE International Conference on Big Data.

[18] Shujia Zhou et al., “Visualization and Diagnosis of Earth Science Data through Hadoop and Spark”, 978- 1-4673-9005-7/16/2016 IEEE International Conference on Big Data.