# E-COMMERCE REVIEWS ANALYSIS

**A Project Report Submitted**
**In Partial Fulfillment of the Requirements**
**for the Degree of**

## MASTER OF COMPUTER APPLICATIONS

**by**

**Aishwarya Yadav**
**(1900290149004)**

**Under the Supervision of**

**Dr. Ajay Kumar Shrivastava**
**KIET Group of Institutions, Ghaziabad**

**to the**

**Faculty of Master of Computer Applications**

**DR. APJ ABDUL KALAM TECHNICAL UNIVERSITY**
**(Formerly Uttar Pradesh Technical University) LUCKNOW**

**July,** 2021

# DECLARATION

I hereby declare that the work presented in this report entitled “E-COMMERCE REVIEWS ANALYSIS", 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 : Aishwarya Yadav

Roll. No. : 1900290149020

Branch : Master of Computer Applications

**(Candidate Signature)**

# CERTIFICATE

Certified that **Aishwarya Yadav(1900290149004)** has carried out the project work presented in this report entitled **"E-Commerce Reviews Analysis"** for the award of **Master of Computer Application** from Dr. A.P.J. Abdul Kalam Technical University, Lucknow under my supervision. The report embodies result of original work, and studies are carried out by the student himself and the contents of the 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.

**Dr. Ajay Kr. Shrivastava** **External Examiner**

Professor & Head

KIET Group of Institutions, Ghaziabad

Date:

KIET
GROUP OF INSTITUTIONS
(A Technical Campus approved by AICTE)
Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow
An ISO-9001 : 2008 Certified Institute

# CERTIFICATE

This is to certify that the project report entitled "***E-Commerce Reviews Analysis***" submitted to KIET Group of Institutions in partial fulfillment of the requirement for the award of the degree of **MASTER OF COMPUTER APPLICATION**, is an original work carried out by **Ms. Aishwarya Yadav,** University Roll No: **1900290149004** under the guidance of Dr. Ajay Kumar Shrivastava.

The matter embodied in this project is a genuine work done by the student and has not been submitted whether to this University or to any other University / Institute for the fulfillment of the requirement of any course of study.

Dr. Ajay Kr. Shrivastava
(Head – CA)
KIET, Ghaziabad

# ABSTRACT

Nowadays people are more inclined towards ecommerce websites. This is into trend as the people these days like to spend a lot less when it comes to time. E-commerce websites have provided people with an option to sit at home and shop spending a lot less time in shopping as compared to visiting stores. Online shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser or a mobile app. People can visit these sites and can go through n number of products available on these sites. We have different E-Commerce websites like Flipkart, Myntra, Amazon, Ajio, etc. These sites not just help the users to buy the products but the users can post reviews of the products that they have purchased. Users are free to post whatever they feel for the particular purchase. Users are given the option to rate the products as well as to write down the reviews. Users can write both positive and negative reviews based on the quality, appearance, description, fit and various other parameters. In this project I am analysing reviews of the products of E-Commerce website to determine which category of the product (apparels, electronics, décor, etc) have more positive reviews. The dataset is taken from website named Kaggle. The analysis on the dataset is done using Big Data Analytics tools like Apache Hadoop. In the end the result shows the user's count of positive and negative reviews based on the words available in the dictionary file.

# ACKNOWLEDGEMENT

Success in life is never attained single handedly. My deepest gratitude goes to my Project supervisor, **Dr. Ajay Kumar Shrivastava, Professor and Head, Department of Computer Applications** for his guidance, help and encouragement throughout my research work. Their enlightening ideas, comments, and suggestions. Words are not enough to express my gratitude 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.

Aishwarya Yadav
1900290149004

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## 1.1 PROJECT DESCRIPTION

Internet has become the soul of today's world. As we are moving forward with time more and more of our daily activities are being transferred from the traditional way to online. From talking to our people to buying clothes, electronics and even our groceries, everything can be done online providing more ease and comfort and saving a lot of time too. People are migrating more towards online shopping. With this much of expansion and availability of data, the measure of web-based life information being created is increasing very fast. These all data is unstructured and big in size. This data is different from structured data (which is stored in relational database systems) in terms of five parameters –variety, volume, value, veracity and velocity (5V's). Big Data Analytics is the way of processing the large amount of data. Hadoop is a popular open-source software which is very useful in analysing the larger data. Hadoop provides several tools for this purpose like Hive, Pig, HBase, Cassandra etc. In this Project, I have used Hadoop framework. For the analysis of E-commerce Reviews dataset, Hive tool is used with Hadoop framework. The aim is to analyse the user's product reviews and classify them in either positive or negative. The reviews are classified into these categories using a dictionary text file which contains English words with its score. At last, a graph is plotted showing total number of positive and negative reviews of the product. Visualization of the sentiment counts will give better understanding of the dataset

## 1.2 PROJECT PURPOSE

In today's time internet is the soul of the world. As we are moving forward with time, we are migrating more towards doing things via internet. With time our day-to-day activities are being transferred online. Be it talking to our people or even buying groceries for our homes everything is done online by majority of the population. In countries like USA almost all the population shops online as it saves their time and they get a pretty good discount on things and there are also lot more options are available if we compare it from physical shop. Due to this migration of people from online to offline mode there is generation of a huge amount of data daily (In TB's and PB's). Therefore, it is difficult to store and manage that unstructured data hence now it has become a tedious task to do. We are going to use Hadoop technology to maintain this huge amount of data. This data will be stored in HDFS (Hadoop Distributed File System) format. Hadoop is a platform, which is used to store distributed and computational data. The main purpose of this project is to analyse large dataset of customer's reviews and perform analysis using different Big Data Analytics tools like Hadoop, Hive, MapReduce. This review analysis will result in classification of tweets in two categories – positive and negative so that it can help other buyers to identify which category is best for the purchase so that there minimal chances of taking the overhead of returning or replacing the item purchased. This project also aims towards deep learning of Big Data Analytics.

## 1.3 BIG DATA AND ITS CHARACTERSTICS

Big Data is a collection of data that is huge in volume yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

In recent years, Big Data was defined by the "*3Vs*" but now there is "*5Vs*" of Big Data which are also termed as the characteristics of Big Data as follows:

**1. Volume:**

- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.
- *Example:* In the year 2016, the estimated global mobile traffic was 6.2 Exabytes (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exabytes of data.

**2. Velocity:**

- Velocity refers to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like 'velocity'.
- *Example:* There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

**3. Variety:**

- It refers to nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.
- **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
- **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.
- **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.

**4. Veracity:**

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

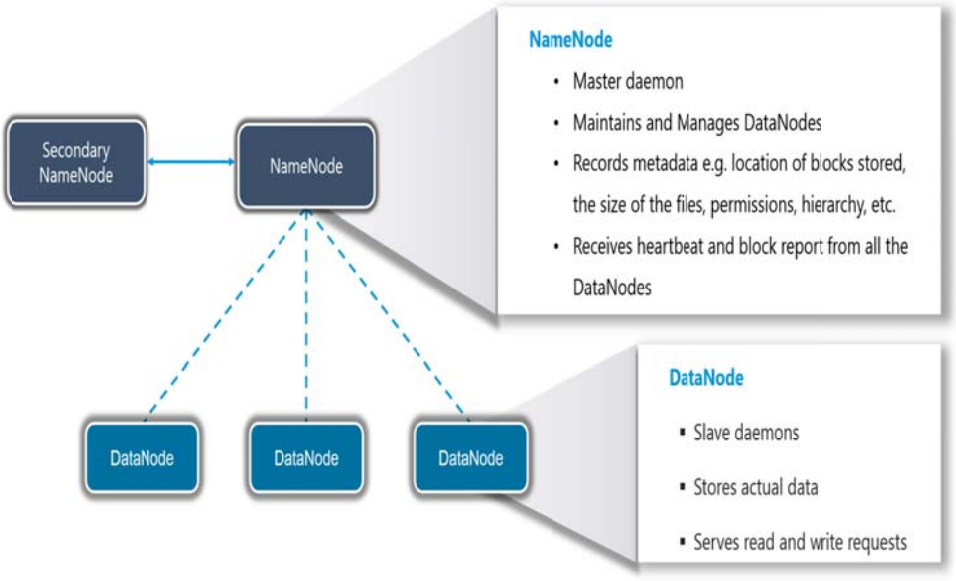
**5. Value:**

- After having the 4 V's into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V's.

## 1.3 BIG DATA ANALYSIS USING HADOOP

Hadoop is a framework that allows you to first store Big Data in a distributed environment, so that, you can process it parallelly. There are basically two components in Hadoop. The first one is ***HDFS*** for storage (Hadoop distributed File System), that allows you to store data of various formats across a cluster. The second one is ***YARN***, for resource management in Hadoop. It allows parallel processing over the data, i.e., stored across HDFS.

- **HDFS:** HDFS creates an abstraction, let me simplify it for you. Similar as virtualization, you can see HDFS logically as a single unit for storing Big Data, but actually you are storing your data across multiple nodes in a distributed fashion. HDFS follows master-slave architecture. In HDFS, Namenode is the master node and Datanodes are the slaves. Namenode contains the metadata about the data stored in Data nodes, such as which data block is stored in which data node, where are the replications of the data block kept etc. The actual data is stored in Data Nodes. I also want to add, we actually replicate the data blocks present in Data Nodes, and the default replication factor is 3. Since we are using commodity hardware and we know the failure rate of these hardwares are pretty high, so if one of the DataNodes fails, HDFS will still have the copy of those lost data blocks.




**1Fig. 1.1 HADOOP-HDFS**

- **YARN**: **YARN** performs all your processing activities by allocating resources and scheduling tasks. It has two major components, i.e., ResourceManager and NodeManager.

  ResourceManager is again a master node. It receives the processing requests and then passes the parts of requests to corresponding NodeManagers accordingly, where the actual processing takes place. NodeManagers are installed on every DataNode. It is responsible for the execution of the task on every single DataNode.

- **MapReduce** — **MapReduce** is both a programming model and big data processing engine used for the parallel processing of large data sets. Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

  A MapReduce *job* usually splits the input data-set into independent chunks which are processed by the *map tasks* in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the *reduce tasks*. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

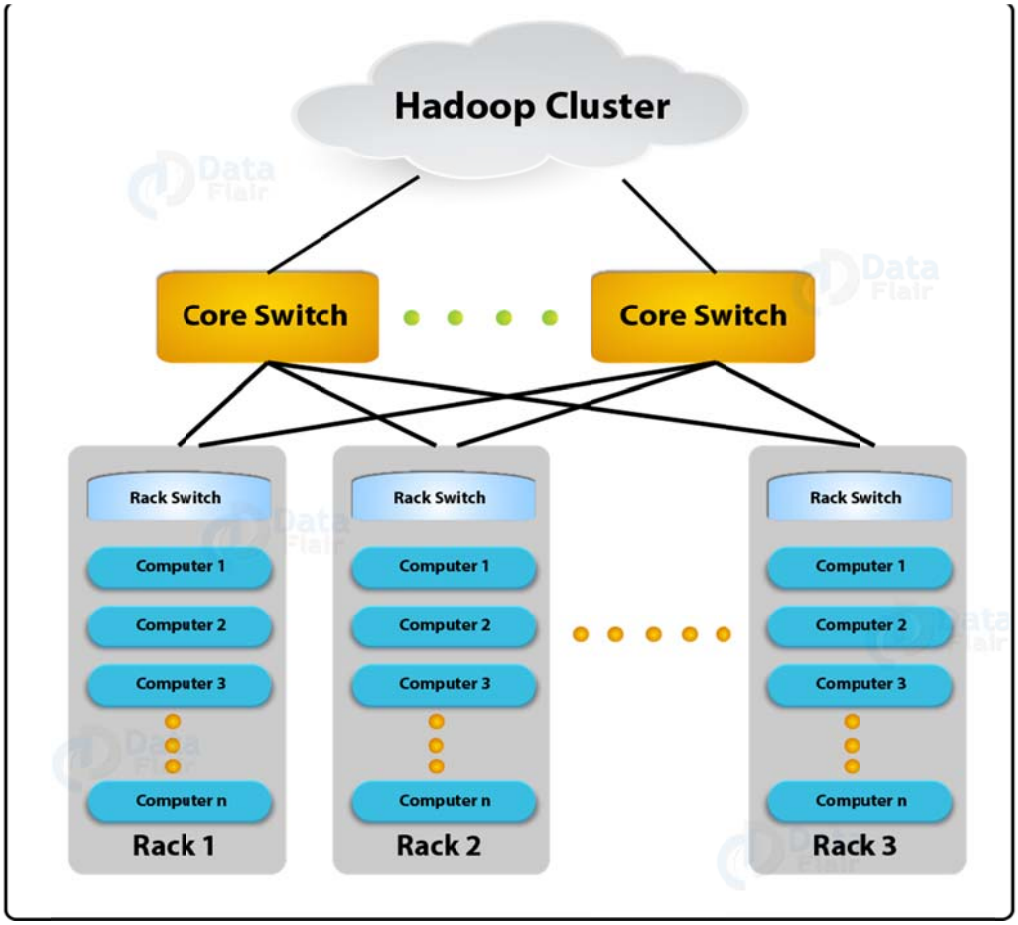
- **Hadoop Common** — Hadoop Common provides a set of services across libraries and utilities to support the other Hadoop modules.

As big data grows exponentially, parallel processing capabilities of a Hadoop cluster help in increasing the speed of analysis process. However, the processing power of a hadoop cluster might become inadequate with increasing volume of data. In such scenarios, hadoop clusters can scaled out easily to keep up with speed of analysis by adding extra cluster nodes without having to make modifications to the application logic.

A hadoop cluster architecture consists of a data centre, rack and the node that actually executes the jobs. Data centre consists of the racks and racks consists of nodes. A medium to large cluster consists of a two or three level hadoop cluster architecture that is built with rack mounted servers. Every rack of servers is interconnected through 1 gigabyte of Ethernet (1 GigE). Each rack level switch in a hadoop cluster is connected to a cluster level switch which are in turn connected to other cluster level switches or they uplink to other switching infrastructure.

In a single node hadoop cluster, all the daemons i.e., DataNode, NameNode, TaskTracker and JobTracker run on the same machine/host. In a single node hadoop cluster setup everything runs on a single JVM instance. The hadoop user need not make any configuration settings except for setting the JAVA_HOME variable. For any single node hadoop cluster setup the default replication factor is one.

In a multi-node hadoop cluster, all the essential daemons are up and run-on different machines/hosts. A multi-node hadoop cluster setup has a master slave architecture where in one machine acts as a master that runs the NameNode daemon while the other machines act as slave or worker nodes to run other hadoop daemons. Usually in a multi-node hadoop cluster there are cheaper machines (commodity computers) that run the TaskTracker and DataNode daemons while other services are run on powerful servers. For a multi-node hadoop cluster, machines or computers can be present in any location irrespective of the location of the physical server.




**2Fig. 1.2 HADOOP CLUSTER**

# CHAPTER 2

## LITERATURE REVIEW

### 2.1 BIG DATA

Big data analysis requires examining very large amounts of data. This is to discover the hidden patterns, and give the insight to make decisions on the correct business. Companies follow it, because it uses the power of data and technology be a more objective data-driven. The main purpose is to describe the big data analysis method to improve company performance and customer's experience.[1]

Big Data in simple terms is a dataset with 3 V's that are Volume, Variety and Velocity. Because this it is difficult to store and process using traditional database management systems. Big Data Analytics is the mean of processing this such a large amount of data. Hadoop is a popular open-source software which is very useful in analyzing the larger data. Hadoop provides several tools for this purpose like Hive, Pig, HBase, Cassandra etc. [15]

Many research works deal with big data platforms looking forward to data science and analytics. These are complex and usually distributed environments, composed of several systems and tools [2].

### 2.2 E-COMMERCE REVIEWS ANALYSIS

Online reviews on e-commerce websites play quite a decisive role in customers' purchase decisions on these platforms. The current ranking patterns of these platforms are mostly based on the ratio of helpful and total votes. Thus if a recent review which is also helpful cannot get a good ranking due to this ranking pattern. With an increasing number of reviews on these e-commerce websites, even an average product can get enough reviews making it a difficult task for consumers to go for each review. Formerly researchers have mostly worked only on the helpfulness of review but very few

have worked on the extent of review helpfulness. The textual content of reviews on e-commerce websites is examined with different helpfulness votes to further classify a new review and give the recently submitted review a proper rank or place in the current set of reviews i.e. our model can not only decide helpfulness of review but can also decide up to what extent a particular review is helpful. Thus an e-commerce website consumer could be highly benefited from this as the consumer doesn't have to go through all the reviews before purchasing. [13].

Although statistical learning methods have achieved success in e-commerce platform product review sentiment classification, two problems have limited its practical application:1) The computational efficiency to process large-scale reviews;2) the ability to continuously learn from increasing reviews and multiple domains.

Sentiment classification originated from the analysis and mining of the product review text of e-commerce platforms (Das, Chen, 2001, Turney, 2002). The reviews in e-commerce platforms have natural user-labeled ratings. [14]

Studies of gender-related differences in the perception of ecommerce websites dependent on the websites' aesthetics, usefulness, ease of use, and purchase intentions, give contradictory results. men and women do not significantly differ in their expressed evaluations of ecommerce websites. However, the neural results indicate that gender-related differences in the perception of ecommerce websites are influenced by unconscious effects, which might explain the inconsistent gender-specific research findings. Men tend to require greater neural activity when using ecommerce websites.[7]

The emergence of ecommerce has revolutionized the manner in which firms could conduct business with customers by eliminating spatial and temporal barriers. However, the personal information that customers are often required to disclose poses a threat to individuals' information privacy. While information privacy threatens the adoption of ecommerce, different countries with different cultural values can further inhibit ecommerce adoption, while increasing individuals' concern for the privacy of their personal information. the level of Internet safety perceptions, ecommerce acceptance, privacy concerns and personal interest are salient factors in informing individuals' intention to transact online.[8]

Online reviews have a significant impact on the decisions of consumers, providing valuable information which must be managed from two different perspectives: that of the user who

reads the review and the people who gave those opinions. [9]

Consumers must divulge personal data in order to utilize services or interact with websites. Ecommerce requires consumers to provide information necessary for fulfilling and completing an online purchase (i.e., address, phone number, credit card information). Further, consumers may disclose information in exchange for a more personalized shopping experience or for product recommendation. As the frequency with which individuals provide private information over the Internet increases, protecting personal information has become critically important. The lack of comprehensive policies in the United States aimed at protecting consumer privacy and controlling access to consumer information has created a sense of urgency around issues of consumer privacy.[10]

In modern ecommerce platforms, product content information may have two origins: one is tree-structured taxonomy attributes, and the other is free-form folksonomy tags.[11]

Electronic commerce (E-commerce) websites have grown significantly over the years. However, due to security and usability issues, only 29% of users convert their online search into a purchase. There is a lack of one comprehensive model that is able to measure all the usability components together with the security components. There is a need to design an evaluation model that will be able to evaluate usability and security together for e-commerce website in order to improvise the e-commerce website.[12]

## 2.3 HADOOP

Apache Hadoop has recently been one of the most popular systems for distributed storage and parallel processing of big data. Hadoop is designed for efficiently dealing with large-scale computing on clusters of hardware with a scalable and fault-tolerant framework. Apache Hadoop (2006) with MapReduce is a distributed architecture on clusters for dealing with large-scale datasets. In the environment of MapReduce, the developers can utilize the distributed computing program for solving large-scale problems with the map and reduce functions. Also, the MapReduce is also a simple programming model, where a scalable, flexible, and fault-tolerant algorithm for computations can be easily established.

In the master–slave model there is a master node to manage all sub-populations and distribute individuals among slave nodes. Subsequently, the fitness values of individuals are calculated in the corresponding slave nodes.[3].

Hadoop is Java based programming framework that is used for distributed storage and processing of large data sets on commodity hardware. Apache Software Foundation developed it as an open-source framework. Hadoop consist of two main components. First is the Hadoop Distributed File System (HDFS) for distributed storage of large data sets and second part is MapReduce for distributed processing of data sets. [16]

Apache Hadoop is considered as one of the most popular systems in the cloud community. Hadoop utilizes the MapReduce programming model to decompose individual jobs into multiple smaller tasks and distributes them to different computing nodes to proceed their execution in parallel . By doing so, larger jobs could fasten their execution accordingly.[5]

## 2.4 HDFS

Hadoop's File System (HDFS) uses a master-slave architecture to store and access data. The entire HDFS is managed by a single server called the NameNode (NN) as the master and files content stored on DataNodes (DNs) as slaves. By default, each HDFS data block size is 128 MB but configurable according to the desired I/O performance. To store a big file, Hadoop splits it into many data blocks and stores them in different DNs. With Hadoop's built-in replication system, each data block is replicated on several DNs (3 by default) to avoid data loss in case of a DN failure. HDFS is very efficient when storing and processing large data files. HDFS is a file system which builds on the existing file system. MapReduce is a programming model which is used for processing and generating large data sets with a parallel, distributed algorithm on a cluster.[4]

## 2.5 MAP REDUCE

MapReduce is a software framework for distributed computing.[3]

MapReduce (MR) is a programming model for distributed computing environment that is designed to handle big datasets efficiently. It follows the master-slave architecture: the master node configures the job and distributes the computing tasks over the cluster workers. MapReduce consists of two important phases, namely Map and Reduce. In the Map phase, a transformation is carried out on the data and the Reduce phase aggregates the outputs and produces the summary results. MR solves the problems in a divide-and-conquer fashion with parallel algorithms. It means that the input data are split into smaller partitions distributed over the worker nodes. Each map task processes the data in its related partition by applying a user-defined function. The output of each map task is a set of < *key*,

*value* > pairs which are considered as the input data for the Reduce task. In other words, the Reducer takes the output from a map as an input and aggregates the data into a smaller set of tuples. MapReduce is easy to scale the processing over multiple cluster nodes.

## 2.6 APACHE HIVE

SQL-on-Hadoop engines such as Hive provide a declarative interface for processing large-scale data over computing frameworks such as Hadoop. The increasing need to process analytical queries over large-scale semi-structured data has led to the development of SQL-on-Hadoop engines. These systems evaluate SQL-like queries over data stored in distributed file systems such as the Hadoop Distributed File System (HDFS) . Hive was the first SQL-on-Hadoop system to provide an SQL-like query language, namely HiveQL, and can use MapReduce or Tez as its underlying framework for executing queries. [6]

# REFRENCES


1. Xu Yang **(2021).** Business big data analysis based on microprocessor system and mathematical modeling.

2. Rogério Luís de ,C.Costa ,JoséMoreira , Paulo Pintor, Veronicados Santos ,Sérgio Lifschitz(2021)A survey on Data Driven performance Tuning for Big Data Analytics Platform

3. Hao-ChunLu, F.J.Hwang, Yao-HueiHuang(2020) Parallel and Distributed architecture of generic alogrithm on Apache Hadoop and Spark.

4. Yanlong Zhai, JudeTchaye-Kondi, Kwei-JayLin, Liehuang Zhu, Wenjun Tao, Xiaojiang Du, Mohsen Guizani (2021) Hadoop perfect file: A fast and memory efficient meta data access archive file to face small problem in HDFS.

5. Tsozen Yeh, Hsinyi Huang (2019) Realising integrated prioritized service in Hadoop cloud system.

6. Chunlin Li, Jing Zhang, Yi Chen, Youlong Luo (2019) Data prefetching and file synchronising for performance optimization in Hadoop based hybrid cloud.

7. Anika Nissen, Caspar Krampe (2021) Why he buys it and she doesn't-exploring self-reported and neutral gender difference of e-commerce website.

8. Zareef A.Mohammed, Gurvirender P.Tejay(2017) Examining privacy concern of ecommerce adoption in developing countries: The impact of culture in shaping individual's perception towards technology.

9. JesusSerrano-Guerrero, Jose A.Olivas, Francisco P.Romero(2020) A T1OWA and aspect-based model for customising recommendation on eCommerce.

10. CoryRobinson (2016) Disclosure of personal data in ecommerce: A cross-national comparison of Estonia and the United States.

11. Mingsong Mao, Sihua Chen, Fuguo Zhang, Jialin Han, Quan Xiao(2021) Hybrid ecommerce recommendation model incorporating product taxonomy and folksonomy.

12. Nur Azimah bt Mohd, Zarul Fitri Zaaba(2020) A Review of Usability and Security Evaluation Model of Ecommerce Website.

13. ParitoshTripathi, Sonu Singh, Pragya Chhajer, Munesh Chandra Trivedi, Vineet K.Singh, (2020) Analysis and prediction of extent of helpfulness of reviews on E-commerce websites.

14. Feng Xu, Zhenchun Pan, Rui Xia (2020) E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework.

15. Lucas Filho, E. R., de Almeida, E. C., Scherzinger, S., & Herodotou, H. (2021). Investigating Automatic Parameter Tuning for SQL-on-Hadoop Systems. *Big Data Research*, *25*, 100204.

16. Becken, S., Alaei, A. R., & Wang, Y. (2019). Benefits and pitfalls of using tweets to assess destination sentiment. *Journal of Hospitality and Tourism Technology*.
17. Hamidreza Kadkhodaei, Amir Masoud Eftekhari Moghadam, Mehdi Dehghan(2021) Big data classification using heterogeneous ensemble classifiers in Apache Spark based on MapReduce paradigm.

# CHAPTER 3

## REQUIREMENT SPECIFICATIONS

## 3.1 HARDWARE REQUIREMENTS

- RAM: 8GB
- Operating system: Linux (32bit or 64 bit)

## 3.2 SOFTWARE REQUIREMENTS

- **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.


**3Fig. 3.1 ORACLE VIRTUALBOX**

- **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.


**4Fig. 3.2 CENTOS**

- **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

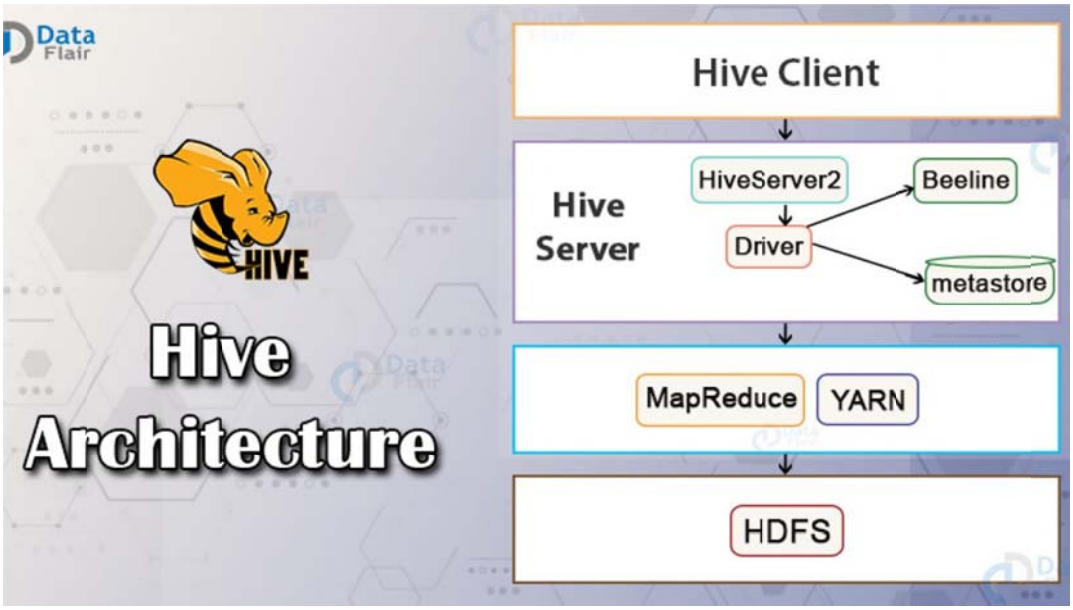
  (e) Secure

- **Apache Hadoop**

  Apache Hadoop is an open source, Java-based software platform that manages data processing and storage for big data applications. Hadoop works by distributing large data sets and analytics jobs across nodes in a computing cluster, breaking them down into smaller workloads that can be run in parallel. Hadoop can process structured and unstructured data and scale up reliably from a single server to thousands of machines. Hadoop clusters are gaining popularity for enhancing the speed of data analysis applications. Hadoop clusters are extremely scalable and highly efficient as they are resistant to failures.


**5Fig. 3.3 APACHE HADOOP**

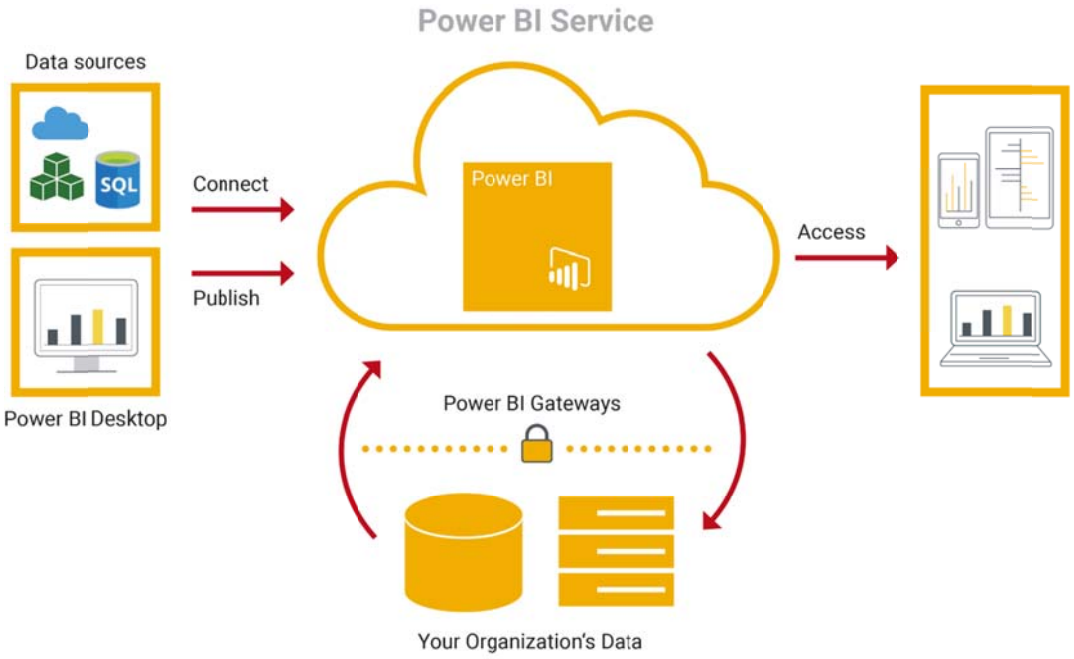
- **Apache Hive**
  - Hive is a data warehouse system for Hadoop.
  - It allows querying, data analysis utilizing HiveQL etc.
  - Hive enables users to portray structure on huge unstructured data.
  - Hive can understand organized and unorganized data which may include text files where fields are circumscribed by specific characters.




**6Fig. 3.4 HIVE ARCHITECHTURE**

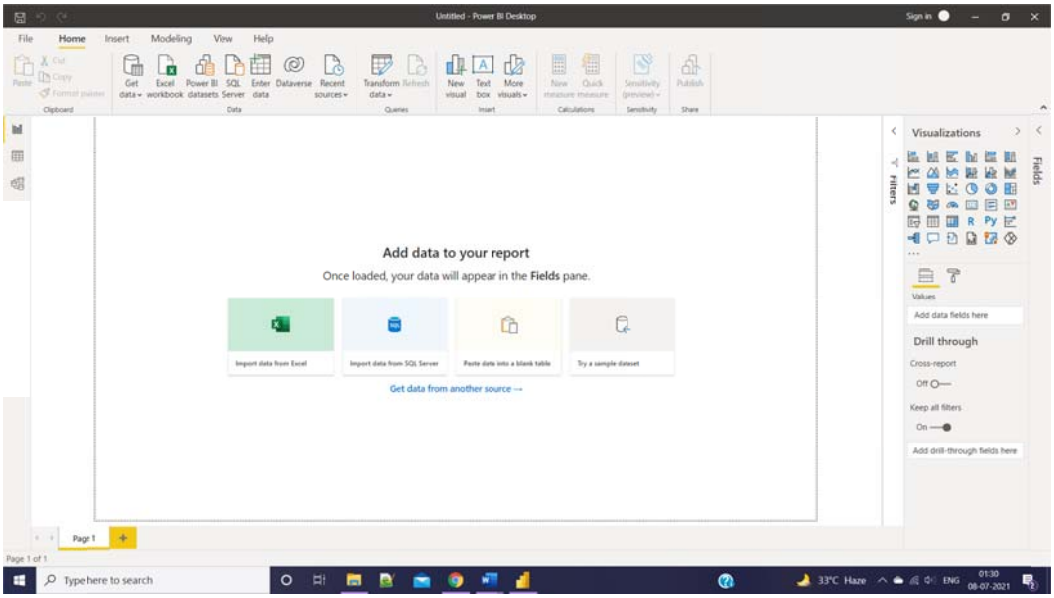
- **Putty**
  - PuTTY is a free implementation of SSH and Telnet for Windows and Unix platforms, along with an xterm terminal emulator.

- **Microsoft Power BI**
  - Microsoft Power BI is a suite that is a collection of business intelligence tools such as software services, apps and data connectors. It is a cloud-based platform used to consolidate data from varied sources into a single data set. These data sets are used for data visualization, evaluation, and analysis by making sharable reports, dashboards, and apps. **Power BI** is **Microsoft's** interactive data visualization and analytics tool for **business intelligence** (**BI**). It is used to visualize the final table of hive having count of each of positive, negative and neutral tweets.




**7Fig. 3.5 POWER BI SERVICE**


**8Fig. 3.6 POWER BI FEATURES**




**9Fig. 3.7POWER BI DESKTOP**

## 3.3 DATA REQUIREMENTS

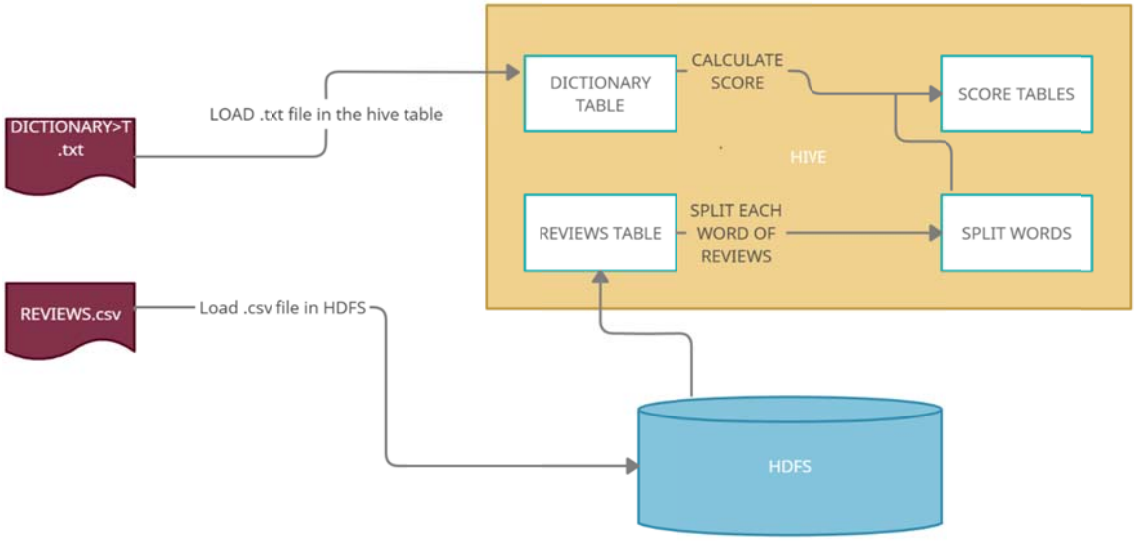
To analyize the data we require a really large amount of data to obtain some trend or pattern. The e-commerce reviews data set is downloaded from Kaggle. It is a website where we can get the data sets in large amounts and it is free of cost for the users. All the entries and Reviews text in data set are related to e-commerce shopping websites.

# CHAPTER 4

## DESIGN

### 4.1 0-LEVEL DATA FLOW DIAGRAM


**Fig. 4.1 0-LEVEL DFD**




4.2 1-LEVEL DATA FLOW DIAGRAM

## 4.3 DATA DICTIONARY

### 4.3.1 Tables Used

- ReviewsTab Table stores the actual dataset and it contains all the fields that has been specified in .csv file.

| ReviewsTab | |
|---|---|
| user_name | STRING |
| user_location | STRING |
| user_contact | INT |
| user_authenticity | STRING |
| user_created | STRING |
| user_purchase | INT |
| user_whishlist | INT |
| user_address | STRING |
| user_payment_details | STRING |
| dates | STRING |
| text | STRING |

**Table 1 Table 4.1 ReviewsTab**

- Dictionary Table contains 2 fields word and score same as in dictionary.txt file. Dictionary.txt file has scores associated with each word. The scores are given in such way that, the more the word is positively expressed the higher will be its value and for negative words, more the word is negative more negative value is specified in the score. If the score of any word is 0 then it means that word is Neutral in nature.

| Dictionary | |
|---|---|
| word | STRING |
| score | INT |

**Table 2 Table 4.2 Dictionary**

- Splitwords Table contains user_name field and each splitted word of user's reviews. Score of each word can be determined independently using these split words.

| Splitwords | |
|---|---|
| user_name | STRING |
| word | STRING |

**Table 3 Table 4.3 Splitwords**

- ScoreTable is generated to analyze each user's reviews. It contains the score of each word in review corresponding to the word's value in dictionary table.

| ScoreTable | |
|---|---|
| t.user_name | STRING |
| t.word | STRING |
| d.score | STRING |

**Table 4 Table 4.4 scoretable**

- ReviewsScore table is the final table having 2 columns – sentiment and count.

| ReviewsScore | |
|---|---|
| sentiment | STRING |
| counts | INT |

**Table 5 Table 4.5 ReviewsScore**

# CHAPTER 5

# TESTING

## 5.1 BIG DATA TESTING STRATEGY

Big Data Testing is a testing process of a big data application to ensure that all the functionalities of a big data application work as expected. The goal of big data testing is to make sure that the big data system runs smoothly and error-free while maintaining the performance and security. Since big data is a collection of large datasets that cannot be processed using traditional computing techniques, traditional data testing methods do not apply to big data. This means your big data testing strategy should include big data testing techniques, big data testing methods and big data automation tools, such as Apache's Hadoop. There are several areas in Big Data where big data testing strategy is required. There are various types of testing in Big Data projects such as Database testing, Infrastructure, and Performance Testing, and Functional testing. Big Data defined as a large volume of data structured or unstructured. Data may exist in any format like flat files, images, videos, etc. The primary Big Data characteristics are three V's - Volume, Velocity, and Variety where volume represents the size of the data collected from various sources like sensors, transactions, velocity described as the speed (handle and process rates) and variety represents the formats of data. Some primary examples of Big Data are Social Networking sites like Twitter and Facebook and E-commerce sites such as Amazon, Flipkart, Snapdeal and any other E-commerce site which have millions of visitors and products.

Big Data Testing plays a vital role in Big Data Systems. If Big Data systems not appropriately tested, then it will affect business, and it will also become tough to understand the error, cause of the failure and where it occurs. Due to which finding the solution for the problem also becomes difficult. If Big Data Testing performed correctly, then it will prevent the wastage of resources in the future.

## 5.2 HOW BIG DATA TESTING STRATEGY WORKS

- **Data Ingestion Testing:** In this, data collected from multiple sources such as CSV, sensors, logs, social media, etc. and further, store it into HDFS. In this testing, the primary motive is to verify that the data adequately extracted and correctly loaded into HDFS or not. Tester must ensure that the data properly ingests according to the defined schema and also have to verify that there is no data corruption. The tester validates the correctness of data by taking some little sample source data, and after ingestion, compares both source data and ingested data with each other. And further, data loaded into HDFS into desired locations.

- **Data Processing Testing:** In this type of testing, the primary focus is on aggregated data. Whenever the ingested data processes, validate whether the business logic is implemented correctly or not. And further, validate it by comparing the output files with input files. **Tools -** Hadoop, Hive, Pig, Oozie

- **Data Storage Testing:** The output stored in HDFS or any other warehouse. The tester verifies the output data correctly loaded into the warehouse by comparing the output data with the warehouse data. **Tools -** HDFS, HBase

- **Data Migration Testing:** Majorly, the need for Data Migration is only when an application moved to a different server or if there is any technology change. So basically, data migration is a process where the entire data of the user migrated from the old system to the new system. Data Migration testing is a process of migration from the old system to the new system with minimal downtime, with no data loss. For smooth migration (elimination defects), it is essential to carry out Data Migration testing.

## 5.3 HOW TO TEST HADOOP APPLICATIONS

Big Data Testing or Hadoop Testing can be broadly divided into three steps.

**Step 1 : Data Staging Validation**

The first step in this big data testing tutorial is referred as pre-Hadoop stage involves process validation.

- Data from various source like RDBMS, weblogs, social media, etc. should be validated to make sure that correct data is pulled into the system
- Comparing source data with the data pushed into the Hadoop system to make sure they match
- Verify the right data is extracted and loaded into the correct HDFS location

**Step 2 : "MapReduce" Validation**

The second step is a validation of "MapReduce". In this stage, the Big Data tester verifies the business logic validation on every node and then validating them after running against multiple nodes, ensuring that the

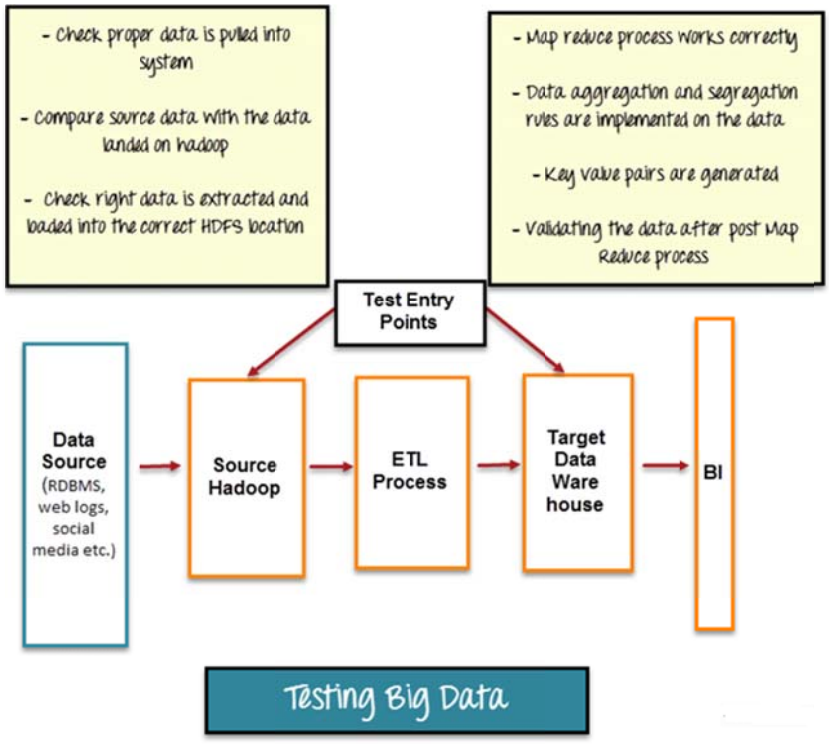
- Map Reduce process works correctly
- Data aggregation or segregation rules are implemented on the data
- Key value pairs are generated
- Validating the data after the Map-Reduce process

**Step 3 : Output Validation Phase**

The final or third stage of Hadoop testing is the output validation process. The output data files are generated and ready to be moved to an EDW (Enterprise Data Warehouse) or any other system based on the requirement.

Activities in the third stage include

- To check the transformation rules are correctly applied
- To check the data integrity and successful data load into the target system
- To check that there is no data corruption by comparing the target data with the HDFS file system data




**10Fig. 5.1 BIG DATA TESTING**

## 5.4 BENEFITS OF BIG DATA TESTING STRATEGY

- **Data Validation Testing:** Every organization strives for accurate data for business planning, forecasting and decision-making. This data needs to be validated for its correctness in any big data application. This validation process should confirm that:
    - the data injection process is error-free
    - complete and correct data is loaded to the big data framework
    - the data process validation is working properly based on the designed logic
    - the data output in the data access tools is accurate as per the requirement
- **Improved Business Decisions:** Accurate data is the pillar for crucial business decisions. When the right data goes in the hands of genuine people, it becomes a positive feature. It helps in analysing all kinds of risks and only the data that contribute to the decision-making process comes into the picture, and ultimately becomes a great aid to make sound decisions.

- **Cost-Effective Storage:** Behind every big data application, there are multiple machines that are used to store the data injected from different servers into the big data framework. Every data requires storage-and storage doesn't come cheap. That's why it's important to thoroughly validate if the injected data is properly stored in different nodes based on the configuration, such as data replication factor and data block size.
  Keep in mind that any data that is not well structured or in bad shape requires more storage. Once that data is tested and is structured, the less storage it consumes, thus ultimately becoming more cost-effective.

- **Right Data at the Right Time:** Big data framework consists of multiple components. Any component can lead to bad performance in data loading or processing. No matter how accurate the data may be, it is of no use if it is not available at the right time. Applications that undergo load testing with different volumes and varieties of data can quickly process a large amount of data and make the information available when required.

- **Reduces Deficit and Boosts Profits:** Indigent big data becomes a major loophole for the business as it is difficult to determine the cause and location of errors. On the other hand, accurate data improves the overall business, including the decision-making process. Testing such data isolates the useful data from the unstructured or bad data, which will enhance customer services and boost business revenue.

## 5.5 BIG DATA TESTING CHALLENGES

Challenges faced when testing unstructured data are expected, especially when new to implementing tools used in big data scenarios.

- **Heterogeneity and Incompleteness of Data**

  - **Problem**: Many businesses today are storing exabytes of data in order to conduct daily business. Testers must audit this voluminous data to confirm its accuracy and relevance for the business. Manual testing this level of data, even with hundreds of QA testers, is impossible.
  - **Solution**: Automation in big data is essential to your big data testing strategy. In fact, data automation tools are designed to review the validity of this volume of data. Make sure to assign QA engineers skilled in creating and executing automated tests for big data applications.

- **High Scalability**
    - **Problem**: A significant increase in workload volume can drastically impact database accessibility, processing and networking for the big data application. Even though big data applications are designed to handle enormous amounts of data, it may not be able to handle immense workload demands.
    - **Solution**: Your data testing methods should include the following testing approaches:
        - **Clustering Techniques**: Distribute large amounts of data equally among all nodes of a cluster. These large data files can then be easily split into different chunks and stored in different nodes of a cluster. By replicating file chunks and storing within different nodes, machine dependency is reduced.
        - **Data Partitioning**: This automation in big data approach is less complex and is easier to execute. Your QA testers can conduct parallelism at the CPU level through data partitioning.

- **Test Data Management**
    - **Problem:** It is not easy to manage test data when it's not understood by your QA testers. Tools used in big data scenarios can only carry your team so far when it comes to migrating, processing and storing test data- that is, if your QA team doesn't understand the components within the big data system.
    - **Solution:** First, your QA team should coordinate with both your marketing and development teams in order to understand data extraction from different resources and data filtering as well as pre and post-processing algorithms. Provide proper training to your QA engineers designated to run test cases through your big data automation tools so that test data is always properly managed.

# CHAPTER 6

## IMPLEMENTATION & WORKFLOW

## 6.1 LINUX MACHINE SETUP USING ORACLE VIRTUALBOX AND CENTOS

- Download Oracle VM Virtualbox for windows and install.
- Download CentOS 7
- Setup Linux 64-bit Machine.

## 6.2 SETTING UP 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.

**Install Hadoop**


**11 Fig. 5.1 HADOOP SETUP**

Step 1: Download the Java 8 Package. Save this file in your home directory.

Step 2: Extract the Java Tar File.

***Command*:** tar -xvf jdk-8u101-linux-i586.tar.gz

**Step 3: Download the Hadoop 2.7.3 Package.**

***Command*:** wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz

Step 4: Extract the Hadoop tar File.

***Command***: tar -xvf hadoop-2.7.3.tar.gz

Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).

Open**. bashrc** file. Now, add Hadoop and Java Path as shown below.

***Command*****:** vi .bashrc

Then, save the bash file and close it.

For applying all these changes to the current Terminal, execute the source command.

***Command*****:** source .bashrc

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and hadoop version commands.

***Command*****:** java –version

***Command*****:** hadoop version

**Step 6:** Edit the **Hadoop Configuration files**.

***Command:*** cd hadoop-2.7.3/etc/hadoop/

***Command:*** ls

All the Hadoop configuration files are in **hadoop-2.7.3/etc/hadoop** directory.

Step 7: Open *core-site.xml* and edit the property mentioned below inside configuration tag:

*core-site.xml* informs Hadoop daemon where NameNode runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.

***Command*****:** vi core-site.xml

Step 8: Edit *hdfs-site.xml* and edit the property mentioned below inside configuration tag:

*hdfs-site.xml* contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.

***Command*****:** vi hdfs-site.xml

Step 9: Edit the *mapred-site.xml* file and edit the property mentioned below inside configuration tag:

*mapred-site.xml* contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.

***Command*:** cp mapred-site.xml.template mapred-site.xml

***Command*:** vi mapred-site.xml.

Step 10: Edit *yarn-site.xml* and edit the property mentioned below inside configuration tag:

*yarn-site.xml* contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

***Command*:** vi yarn-site.xml

**Step 11:** Edit *hadoop-env.sh* and add the Java Path as mentioned below:

*hadoop-env.sh* contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

***Command*:** vi hadoop–env.sh

Step 12: Go to Hadoop home directory and format the NameNode.

***Command*:** cd

***Command*:** cd hadoop-2.7.3

***Command*:** bin/hadoop namenode -format

This formats the HDFS via NameNode. This command is only executed for the first time. Formatting the file system means initializing the directory specified by the dfs.name.dir variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

Step 13: Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.

***Command:*** cd hadoop-2.7.3/sbin

Either you can start all daemons with a single command or do it individually.

***Command:*** ./start-all.sh

The above command is a combination of ***start-dfs.sh, start-yarn.sh*** & ***mr-jobhistory-daemon.sh***

Or you can run all the services individually as below:

Start NameNode:

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.

***Command:*** ./hadoop-daemon.sh start namenode

Start DataNode:

On startup, a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.

***Command:*** ./hadoop-daemon.sh start datanode

Start ResourceManager:

ResourceManager is the master that arbitrates all the available cluster resources and thus helps in managing the distributed applications running on the YARN system. Its work is to manage each NodeManagers and the each application's ApplicationMaster.

***Command:*** ./yarn-daemon.sh start resourcemanager

Start NodeManager:

The NodeManager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the ResourceManager.

***Command:*** ./yarn-daemon.sh start nodemanager

Start JobHistoryServer:

JobHistoryServer is responsible for servicing all job history related requests from client.

***Command*:** ./mr-jobhistory-daemon.sh start historyserver

Step 14: To check that all the Hadoop services are up and running, run the below command.

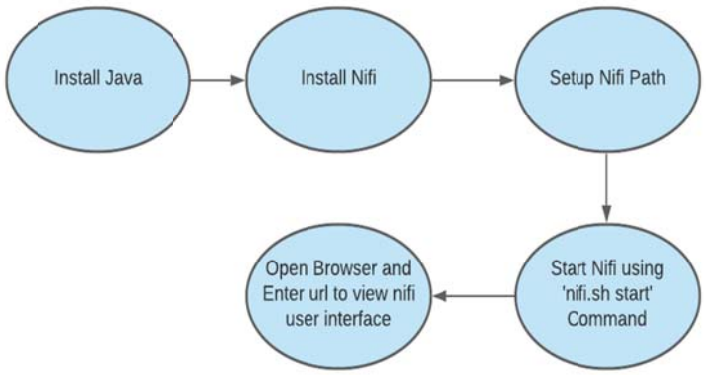
***Command:*** jps

**Step 15:** Now open the Mozilla browser and go to **localhost**:**50070/dfshealth.html** to check the NameNode interface.

## 6.3 DATA FLOW FROM LOCAL DISK TO HDFS USING APACHE NIFI

- **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.
We are using Nifi for ingesting our dataset from local disk into our HDFS. Nifi has a user friendly interface so it becomes very easy to perform these file flows.

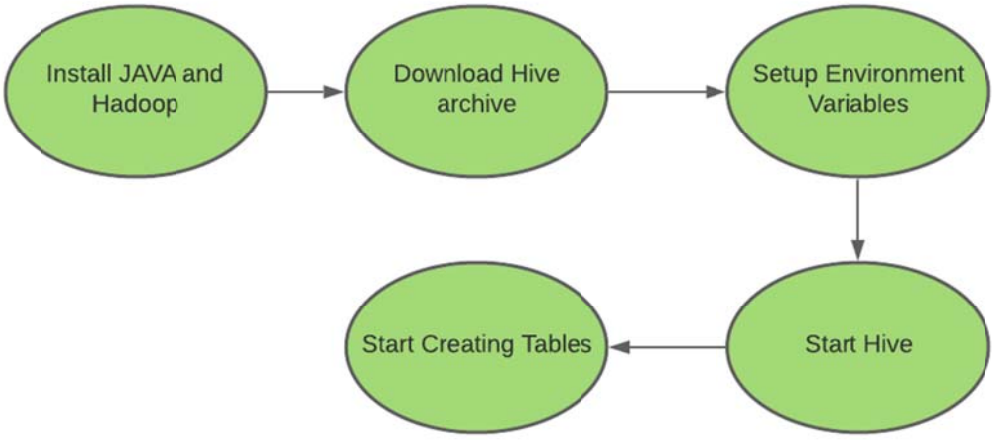



**12 Fig. 5.2 NIFI SETUP**

- After successful installation of Nifi, open nifi interface inside browser. Here create 2 Processors. First is GetFile. It will fetch our dataset from local disk. Second One is the PutHDFS processor. This processor will fetch dataset from GetFile and ingest dataset into the HDFS directory.

## 6.4 ANALYSIS OF TWEETS USING HIVE

The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using an SQL-like language called HiveQL.




**13 Fig. 5.3 HIVE SETUP**

1. Start Hadoop. I have used ./start-all.sh command for starting the Hadoop.
   Used jps command to view the running process at present.
2. We have sample Reviews data to perform a few basic analysis of the data.
   Reviews.csv.

i. This .csv file we have stored in HDFS.
   Dictionary.txt — Contains words with a rating for each word.
ii. Dictionary.txt file is in local disk.

3. To analyze the user reviews that has been posted and figure out whether they are positive or negative based on the words available in the Dictionary file.

4. Start Hive.
   Let's start creating Tables:

5. Load Dataset from HDFS and into Hive Table.

```
CREATE SCHEMA IF NOT EXISTS reviewatab;
CREATE EXTERNAL TABLE IF NOT EXISTS reviewtab
(user_name STRING,
user_location STRING,
user_contact INT,
user_authenticity STRING,
user_created STRING,
user_purchase INT,
user_wishlist STRING,
user_address STRING,
user_paymentdetails STRING,
date STRING,
```

```
text STRING,
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/reviewtab';
```

6. Now, creating a table for dictionary and load data using the given dictionary file

   CREATE TABLE dictionary(word string, score int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

7. LOAD DATA local INPATH 'Desktop/Dictionary.txt' into table dictionary;

8. Split up each tweet (called text in the twitter table) into individual words for comparison.

```
splitwords as
select user_name, word from reviewstab raw lateral view explode(split(text,
" ")) text_ex as word;
select * from splitwords;
```

9. Compare the words with the dictionary to get the “scores”.

```
Create table scoretable as
select
     t..user_name,
    t. word,
    .d.score
from splitwords t join dictionary d where t.word = d.word;
```

10. Group the scores again by date, id, user_name and this is how I have attempted to do that:

```
SELECT
      user_name,
      SUM(score),
      case
        when sum( score ) > 0 then 'positive'
```

```
  when sum( score ) < 0 then 'negative'
 end as sentiment
 FROM scoretable
GROUP BY user_name ;
```

11. Final output using hive and select query is used to show the name, the reviews whether they are positive or negative. If score is >0 it is positive.

1. Insert above results into new table.

```
INSERT into table reviewscore
select  *  from (SELECT
     FromUser,
     SUM(score),
     case
       when sum( score ) > 0 then 'positive'
       when sum( score ) < 0 then 'negative'
     end as sentiment
     FROM scoretable_new
GROUP BY  FromUser)
tmp;
```

2. Display result of sentiment counts.

```
SELECT COUNT(sentiment),sentiment FROM reviewscore
GROUP BY sentiment;
```

# CHAPTER 7

# CONCLUSION

## 7.1 CONCLUSION

This project involved working on Big Data technology using its various tools. This helped me to better understand the way in which a large dataset that is unstructured can be handled and made useful for companies as well as the customers. This dataset that is hard to handle can be handled efficiently and can be made useful for decision making by analyzing these humungous datasets. This project gave me exposure of the working of the single node cluster . I have performed analysis on our datasets and our project is completed using these.

- Nifi
- Hadoop
- Map Reduce
- HDFS
- Hive
- Microsoft Power BI