

# “Google Playstore Application Analysis and Prediction”

## A PROJECT

*Submitted in partial fulfillment of the requirements for the award of the degree of*

## BACHELOR OF TECHNOLOGY

IN

## COMPUTER SCIENCE AND ENGINEERING

Under the supervision of

**Dr. Hemraj Saini**

(Associate Professor)

By

***Shubham Ruhela (151364)***

to



**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY**

**WAKNAGHAT, SOLAN – 173234**

**HIMACHAL PRADESH, INDIA**

## CERTIFICATE

I hereby declare that the work presented in this report entitled “**Google Playstore Application Analysis and Prediction**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2018 to May 2019 under the supervision of **Dr. Hemraj Saini**, Associate Professor, department of Computer Science & Engineering.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Shubham Ruhela (151364).....

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

Dr. Hemraj Saini

(Associate Professor)

Department of Computer Science & Engineering

Dated:

## ACKNOWLEDGEMENT

We take upon this opportunity endowed upon us by grace of the almighty, to thank all those who have been part of this endeavor.

First & Foremost, we want to thank our supervisor '**Dr. Hemraj Saini**' for giving us the correct heading and legitimate direction in regards to the subject. Without his dynamic association and the correct direction this would not have been conceivable.

Last however not the minimum, we generously welcome each one of those individuals who have helped us straight forwardly or in a roundabout way in making this project a win. In this unique situation, We might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated our undertaking.

Shubham Ruhela (151364)

Dated:

## TABLE OF CONTENT

<b>Certificate.....</b>	<b>(i)</b>
<b>Acknowledgement.....</b>	<b>(ii)</b>
<b>Table of Content.....</b>	<b>(iii)</b>
<b>List of figures.....</b>	<b>(v)</b>
<b>List of Tables.....</b>	<b>(vi)</b>
<b>List of Graphs.....</b>	<b>(vii)</b>
<b>Abstract.....</b>	<b>(viii)</b>
<b>1. Introduction.....</b>	<b>1</b>
<b>1.1 Introduction.....</b>	<b>1</b>
1.1.1 Analysis and Prediction of Google Playstore Apps.....	3
1.1.2 Google Playstore Data.....	4
1.1.3 Hive.....	4
1.1.4 Hadoop Distributed File System.....	6
1.1.5 Deep Learning.....	6
1.1.6 Neural Networks.....	7
1.1.7 What is the difference between Machine Learning and Deep Learning?	8
1.1.8 Why Deep Learning?.....	9
<b>1.2 Problem Statement.....</b>	<b>10</b>
<b>1.3 Objective.....</b>	<b>10</b>
<b>1.4 Methodology.....</b>	<b>10</b>
1.4.1 Analysis.....	10
1.4.2 Prediction.....	12
<b>2. Literature Survey.....</b>	<b>15</b>
2.1 Big Data Techniques for efficient storage and processing of weather.....	15
2.2 Commercial Product Analysis Using Hadoop Map Reduce.....	16
2.3 Review paper on Hadoop and Map Reduce.....	17
2.4 Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction.....	17
2.5 Python – The Fastest Growing Programming Language.....	18
2.6 Machine Learning Algorithms: A Review.....	19
2.7 An Overview of Deep Learning.....	19
2.8 The Power of Mobile Applications.....	19
2.9 Mobile App Analytics.....	22

<b>3. System Development.....</b>	<b>24</b>
<b>3.1 Designing.....</b>	<b>24</b>
<b>3.2 Architecture.....</b>	<b>25</b>
3.2.1 Hive Architecture.....	25
3.2.2 Hdfs Architecture.....	26
<b>3.3 Data Flow.....</b>	<b>27</b>
3.3.1 Data Flow in Hive.....	27
3.3.2 Data Flow in Deep Neural Network.....	27
<b>3.4 Requirements.....</b>	<b>28</b>
3.4.1 Hardware requirements.....	28
3.4.2 Software requirements.....	29
3.4.3 Advantages of VMware Workstation.....	29
<b>3.5 Test Plan.....</b>	<b>29</b>
3.5.1 Dataset.....	29
<b>3.6 Algorithms.....</b>	<b>32</b>
3.6.1 Algorithm used for Analysis part.....	32
3.6.1.1 Map Reduce Algorithm.....	32
3.6.2 Algorithm used for Prediction part.....	34
3.6.2.1 Backpropagation.....	34
3.6.2.2 Feed forward.....	35
<b>4. Results and Performance Analysis.....</b>	<b>37</b>
<b>4.1 Analysis.....</b>	<b>37</b>
4.1.1 Using Hive.....	37
<b>4.2 Prediction.....</b>	<b>39</b>
<b>4.3 Deep Learning over Hive.....</b>	<b>42</b>
<b>5. Conclusion .....</b>	<b>43</b>
<b>References.....</b>	<b>44</b>

## LIST OF FIGURES

<b>Figure 1:</b> 5V's of BIG DATA.....	2
<b>Figure 2:</b> Value in 5V's of BIG DATA.....	3
<b>Figure 3:</b> Steps for Rating Analysis.....	4
<b>Figure 4:</b> Neural Network.....	8
<b>Figure 5:</b> The building block of Deep Neural Network.....	9
<b>Figure 6:</b> Performance of Deep Learning.....	9
<b>Figure 7:</b> Reading the dataset.....	13
<b>Figure 8:</b> Most common words in positive review.....	14
<b>Figure 9:</b> pos-to-neg ratio.....	14
<b>Figure 10:</b> Apache Hive Architecture.....	26
<b>Figure 11:</b> HDFS Architecture.....	27
<b>Figure 12:</b> Data flow diagram.....	28
<b>Figure 13:</b> Data flow in Deep Neural Network.....	28
<b>Figure 14:</b> (i) Sample report of data.....	32
<b>Figure 15:</b> (ii) Sample report of dataset.....	32
<b>Figure 16:</b> Sample report of dataset.....	33
<b>Figure 17:</b> How Map Reduce works.....	33
<b>Figure 18:</b> Display of learning rate=0.01.....	36
<b>Figure 19:</b> Top free Applications.....	38
<b>Figure 20:</b> Top paid Applications.....	39
<b>Figure 21:</b> Top reviewed Applications.....	39
<b>Figure 22:</b> Editor's choice Applications.....	40
<b>Figure 23:</b> (i) Review is positive.....	41
<b>Figure 24:</b> (ii) Review is positive.....	41
<b>Figure 25:</b> (i) Review is negative.....	41
<b>Figure 26:</b> (ii) Review is negative.....	41
<b>Figure 27:</b> Prediction for negative review.....	42
<b>Figure 28:</b> Total accuracy without training.....	42
<b>Figure 29:</b> Total accuracy with training.....	42

## LIST OF GRAPHS

<b>Graph 1:</b> Hadoop Map Reduce v/s Spark Cassandra Benchmarking.....	25
<b>Graph 2:</b> Mobile Use Grows Year over Year.....	29
<b>Graph 3:</b> Time spent per day with Digital media.....	29
<b>Graph 4:</b> Number of Apps downloaded.....	30
<b>Graph 5:</b> Number of Total Apps in mobile markets.....	31
<b>Graph 6:</b> Sigmoid function.....	43
<b>Graph 7:</b> Accuracy level.....	51

## LIST OF TABLES

<b>Table 1:</b> Dataset columns and its specifications.....	40
<b>Table 2:</b> Dataset columns and its specifications.....	41



## **ABSTRACT**

Application distribution platform, for example, Google play store gets overwhelmed with a few thousands of new applications regularly with a lot progressively a huge number of designers working freely or on the other hand in a group to make them successful. With the enormous challenge from everywhere throughout the globe, it is basic for a developer to know whether he is continuing the correct way. Dissimilar to making movies where the nearness of famous heroes raise the likelihood of accomplishment even before the movies are coming into the picture, it isn't the situation with creating applications. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. So in this project, I have tried to perform analysis and prediction into the Google Play store application dataset that I have collected from kaggle.com. Using Big Data techniques such as Hive I have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application. And using Deep Learning I have tried to make a prediction about the user reviews that which review is positive or negative.

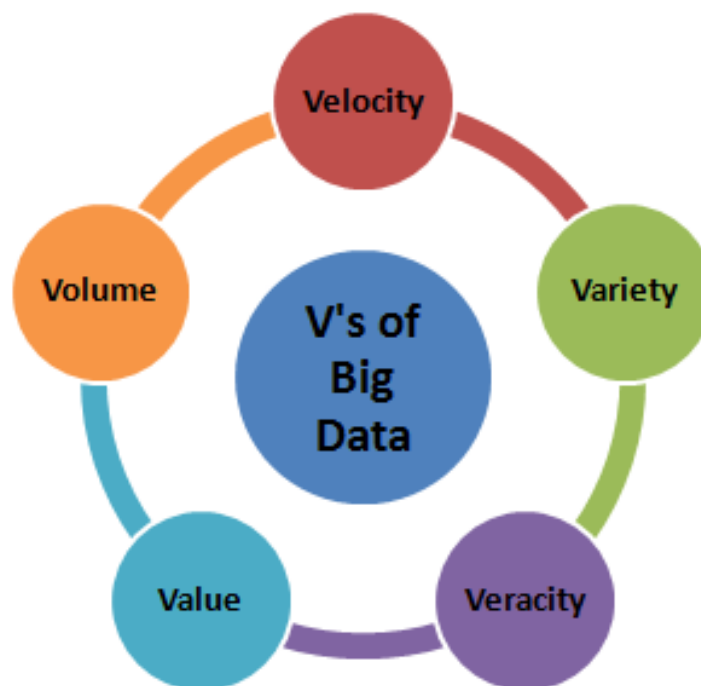
# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

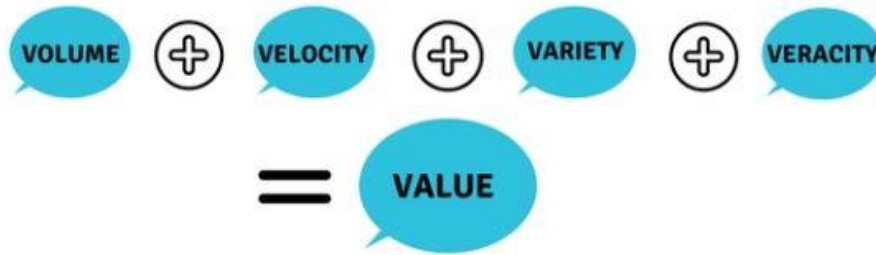
Big Data is likewise information yet with a gigantic size. Big Data is a term used to portray a gathering of information that is big in size but then developing exponentially with time. In short such information is so substantial and complex that none of the customary information the board devices can store it or procedure it proficiently.

We can define a data is a big data with the help of these 5V's:



**Figure 1: 5Vs of BIG DATA**

- Volume: Volume means a huge amount of data that is generated in every second from any social media, cars, bank, from flights etc.
- Velocity: It means speed at which the data is generated and collected, also analyzed.
- Variety: It means different types of data we are having and using it.
- Veracity: It refers to the quality and security of the data.

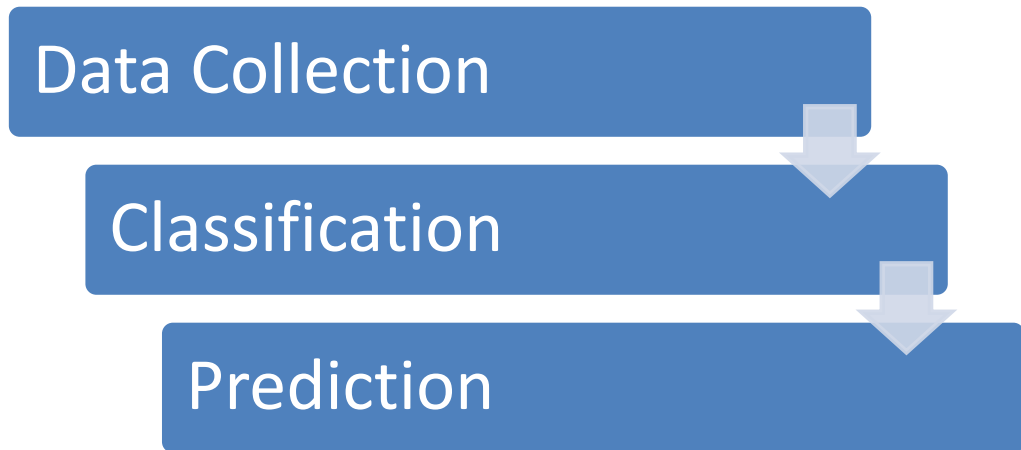


**Figure 2: Value in 5Vs of BIG DATA**

- Value: In this we refer to the worth of the data that is going to be extracted.

We can perform any type of Analysis using Big Data with the following way:

- Collection of Data
- Classification of Data
- Identification of Pattern
- Finally the Prediction
- Visualization



**Figure 3: Steps for Rating Analysis**

### **1.1.1 Analysis and Prediction of Google Playstore Apps**

In today's scenario we can see that mobile apps playing an important role in any individual's life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that in spite of the fact that it creates more than two fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination

extend. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and rating to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on generally speaking evaluations alone and would profit by understanding the a huge number of printed remarks.

### **1.1.2 Google Playstore Data**

The dataset taken is of Google play store application and is taken from Kaggle , which is the world's largest community for data scientists to explore ,analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

### **1.1.3 Hive**

Apache Hive is defined as a data warehouse system built on top of Apache Hadoop for analysis and applying query on a large dataset. It changes over SQL-like queries into MapReduce for simple execution and preparing of amazingly substantial volumes of information.

The three significant functionalities for which Hive is conveyed are: data summarization, data analysis and data query. The query language that is solely upheld by Hive is the HiveQL. This language interprets the SQL-like queries into Map Reduce jobs for conveying it on Hadoop. HiveQL likewise underpins Map Reduce contents that can be

connected to the hive queries. Hive builds the pattern structure adaptability and furthermore data serialization and deserialization.

The Hive query language is known as the **HiveQL** however isn't actually an organized query language. However, HiveQL offers different augmentations that are not part of the SQL. You can make multi-table supplements, make table as select yet it has just a fundamental help for files. In any case, HiveQL does not offer help for Online Transaction Processing and view emergence. It just offers sub-query support. Presently it is conceivable to have full ACID properties alongside update, embed and erase functionalities.

### **Why Hive?**

The Apache Hive is mostly utilized for information querying, analysis and summarization. It improves the designer profitability which comes at the expense of expanding inactivity, and diminishing proficiency. Hive is a variation of SQL and an awesome one for sure. Hive stands tall when contrasted with SQL frameworks executed in databases. Hive has numerous User Defined Functions that makes it simple to add to the UDFs. It extraordinarily helps the engineer network when working for complex explanatory preparing and information organizes that are testing.

'Data Warehouse' alludes to a framework utilized for revealing and data analysis. This means assessing, cleaning, changing, and displaying information with the objective of finding helpful data and proposing ends. Data analysis has various angles and methodologies, incorporating assorted systems under an assortment of names, in various business, science, and sociology areas.

Hive enables clients to all the while get to the information and expands the reaction time, for example the time a framework or useful unit takes to respond to a given information. Indeed Hive regularly has an a lot quicker reaction time than most different sorts of inquiries on a similar kind of colossal datasets. Hive is additionally very adaptable as more items can undoubtedly be included reaction to all the more including of group of information with no drop in execution.

### **1.1.4 Hadoop Distributed File System**

Hadoop has a storage file system known as HDFS (Hadoop Distributed File System), HDFS is a specially designed file system used to store huge volume of data with access streaming pattern on multimode cluster with commodity hardware. The framework is known for its high adaptation to internal failure and is executed on low upkeep hardware so the principle spotlight could be put on the details.. Output of this is very accurate. In this way, for managing the gigantic information, this turns out to be the best. It is much equivalent to an expert slave plan which has a singular name hub which controls the record framework get to. Hdfs uses Master-slave architecture for storage and dividing resources between node, Master nodes are Name node, Secondary Name node, Job Tracker and Slave nodes are Data node , Task Tracker.

Data replication enhanced the circumstance achieving adjustment to non-basic disappointment. The broad data gather is secured as a progression of pieces. Piece assures and the replication factor manually handled. Replication factor is 3 as a matter of course that clients there will be 3 copies of the every datum square will be there at a moment of time in the Hadoop bunch.

### **1.1.5 Deep Learning**

Deep learning is a part of machine learning methods which works on the basis of layers used in artificial neural networks. Learning has different forms: - supervised learning, semi-supervised, unsupervised.

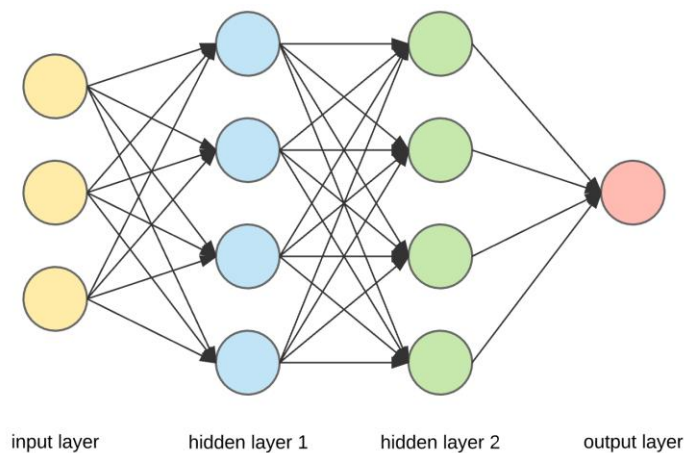
- **Supervised Learning:** It is defined as a learning in which we train a machine as per our dataset or input. From that point forward, the machine is furnished with another arrangement of examples (data) so supervised learning analyses the provided data (set of preparing models) and creates a right result from given input.
- **Unsupervised Learning:** In the Unsupervised learning we do not train our machine according to the present data or input. It means there is not any supervisor as a teacher in this learning. In this we allow algorithm to work on their own without any training or guidance. Here the main working of the machine is that it works on some definite patterns, similarities in the given dataset without any training or proper guidance. Therefore machine is restricted to find out the structure which is hidden in the given dataset.

- **Semi-supervised Learning:** This type of learning lies between the above two learning methods.

In the deep learning we mainly use neural network architectures, that is why all the models used in deep learning often referred as deep neural networks. By the term ‘deep’ in the deep learning we meant the no of hidden layers in particular neural networks. It can be 150 in number or above. In this we trained all these models using the large dataset/input which contain a large no of columns or rows.

### 1.1.6 Neural Networks

We can divide neural network into different forms such as artificial neural network, deep neural network, recurrent neural network, convolutional deep neural networks. Each form has its own importance and its own features. In neural network we have input layer, no of hidden layers, and output layer.



**Figure 4: Neural Network**

In this project I will make use of Deep Neural Network.

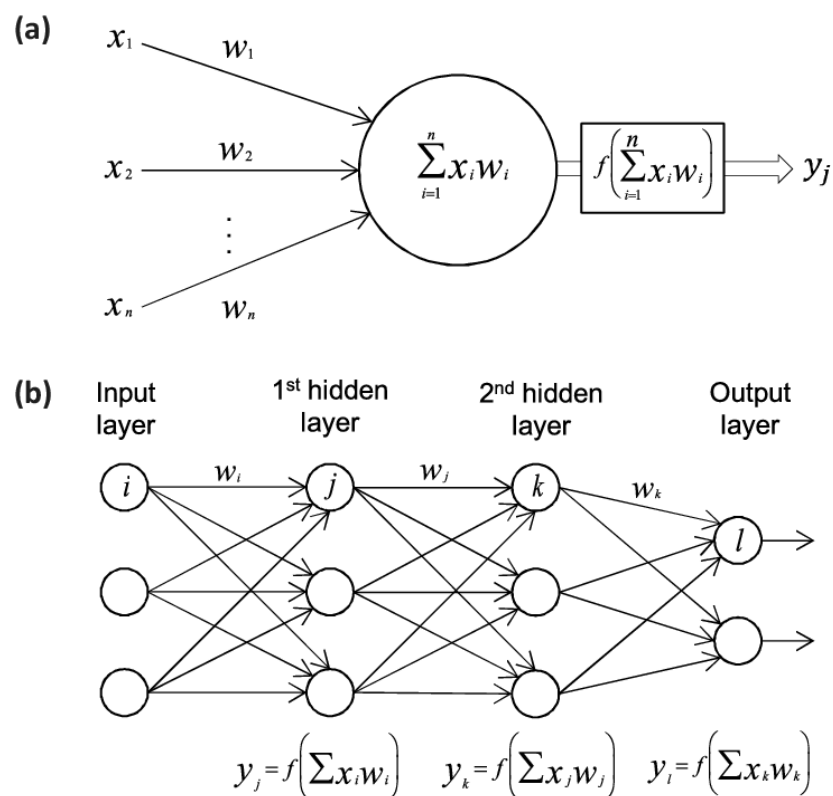
### Deep Neural Network

A deep neural network is defined as a neural network which contains certain level of complexity, like a neural network which contains more than two layers. In the deep neural network we use some mathematical model to solve any model in a proper way using all the complexities.



A neural system, when all is said in done, is an innovation worked to reproduce the action of the human brain – explicitly, design acknowledgment and the section of contribution through different layers of deep neural associations.

In DNN, data flows forward it means from input layer to output layer without having any loopholes. At first, the DNN makes a guide of virtual neurons and allots irregular numerical qualities, or "loads", to create link between them. The loads and data sources are increased and return a yield somewhere in the range of 0 and 1. On the off chance that the system didn't precisely perceive a specific example, a calculation would alter the loads. That way the calculation can make certain parameters progressively compelling, until it decides the right scientific control to completely process the information.



**Figure 5: The building block of deep neural network**

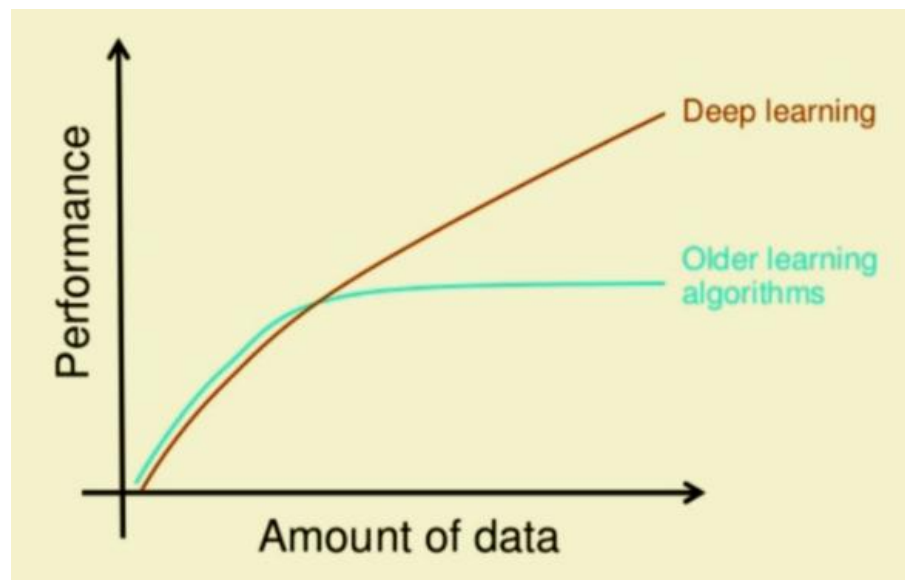
**1.1.7 What is the Difference Between Machine Learning and Deep Learning?**

- Deep learning is a particular type of machine learning. A machine learning works process begins with significant features/attributes being physically removed from pictures. The attributes/features are then used to make a model that classifies the items in the picture. With a deep learning work process, applicable

features/attributes are consequently separated from pictures. What's more, deep learning realizing performs "end to end learning" – where a system is given crude information and an undertaking to perform, for example, arrangement, and it figures out how to do this automatically.

- Another key contrast is deep learning calculations scale with information, while machine learning converges. Machine learning strategies that level at a specific dimension of execution when you include more precedents and training a data to the system. From this we can conclude that deep can easily work with large complex data with full scalability as compare to machine learning.
- The main advantage of using deep learning is that they continue to improve as the size of dataset continues to increase.

### **1.1.8 Why Deep learning?**



**Figure 6: Performance of Deep Learning**

Deep Learning requires top of the line machines in opposition to Machine Learning calculations. GPU has turned into an essential part currently to execute any Deep Learning calculation.

In conventional Machine learning systems, the greater part of the connected highlights should be recognized by an area master so as to diminish the unpredictability of the information and make designs increasingly unmistakable to learning calculations to work. The greatest preferred standpoint Deep Learning calculations as talked about before are

that they attempt to take in abnormal state highlights from information in a steady way. This dispenses with the need of special expertise for feature extraction.

## **1.2 Problem Statement**

In this project we have focused on our 2 objectives. We have taken the dataset and observed it nicely and as per our need we have taken various attributes to analyze and further display the result. By doing this, we can clearly and easily observe the dataset. Moreover, Firstly, I will analyze different attributes given in dataset. Secondly, I will do prediction of those different attributes like predict whether the user review is positive or negative.

## **1.3 Objective**

- The main goal of this project is to analyze different attributes of given application like application name, category, rating, reviews, size, installs, type, price, content rating, genres, last updated, current version, android version. And to find out the most rated and most reviewed apps and also to distinguish between the apps which are either free or paid using Hive technology.
- The second Objective is to predict whether the user review for different application positive or negative using Deep Neural Networks in Deep Learning.

## **1.4 Methodology**

The Implementation of this Project is divided into 2 parts:

### **1.4.1 Analysis**

In this various attributes like application name, category, rating, reviews, size, installs, type, price, content rating, genres, last updated, current version, android version are analyzed using HiveQL.

### Steps for implementing the analysis part:

- First I have selected the dataset.
- Then that information is send to HDFS.
- Then in hive I have created a database.
- In the above created database tables are created.

To create the table named playstore:

```
create table playstore (app string, category string, rating float, reviews int, size string, installs string, type string, price int, content string, genres string, lastup string, currentver string, androidver string) row format delimited fields terminated by ',' stored as textfile;
```

To load the data in the table created by using command, we will have to use the following command:

```
Load data local inpath 'googleplaystore.csv' into table playstore;
```

Below mentioned are the various queries that a user can use to retrieve information:

```
//Names of all free apps
```

```
select playstore.app,playstore.type from google.playstore where type='Free';
```

```
//Names of all paid apps
```

```
select playstore.app,playstore.type from google.playstore where type='Paid';
```

```
//Names of top free apps
```

```
select playstore.app,playstore.type,playstore.rating
```

```
//Names of top free apps
```

```
select playstore.app,playstore.type,playstore.rating from google.playstore where type='Free' and rating>=4;
```

```
//Names of top paid apps
select playstore.app,playstore.type,playstore.rating from google.playstore where
type='Paid' and rating>=4;

//Names of all distinct categories
select distinct playstore.category from google.playstore;

//Names of top reviewed apps
select playstore.app,playstore.reviews from google.playstore where
reviews>=20000000;

//Editors' choice
select playstore.app,playstore.rating,playstore.reviews from google.playstore where
rating>=4 and reviews>=20000000;
```

### **1.4.2 Prediction**

- Initially I have read the given dataset which is in the text form and labeled it accordingly as shown in figure:

```
g = open('reviews.txt','r') # What we know!
reviews = list(map(lambda x:x[:-1],g.readlines()))
g.close()

g = open('labels.txt','r') # What we WANT to know!
labels = list(map(lambda x:x[:-1].upper(),g.readlines()))
g.close()
```

**Figure 7: Reading the dataset**

- I have broken the dataset into training and testing data. The training input data and testing input data is one hot encoded by seeing the words in a review. For example if there is a review like “this is a good application”. So the positioning of each word will be setup in an array and the position of that word in an array is set to one and remaining will be zero. Let us assume this has a position of 3 in an array, good has a position of 1, has a position of zero, application has a

position of 4 and is has a position of 2. Total words are 10 let's assume. So the array will look like

```
import numpy as np
arr=np.zeros(10)
arr[0]=1
arr[1]=1
arr[2]=1
arr[3]=1
arr[4]=1
```

So this will be the input to our neural network and output will be either positive or negative.

Positive output will be set to -->1

Negative output will be set to -->0

So in this case “This is a good movie” our input will look like

[1, 1, 1, 1, 1, 0, 0, 0, 0, 0] and output will be 1 because this is a positive review

Similarly other reviews will be stored in this fashion.

- From the above array of data I have tried to find out the most common words used in positive reviews as can be seen from the figure below:

```
# Examine the counts of the most common words in positive reviews
positive_counts.most_common()

[('.', 550468),
 ('the', 173324),
 ('.', 159654),
 ('and', 89722),
 ('a', 83688),
 ('of', 76855),
 ('to', 66746),
 ('is', 57245),
 ('in', 50215),
 ('br', 49235),
 ('it', 48025),
 ('i', 40743),
 ('that', 35630),
 ('this', 35080),
 ('s', 33815),
 ('as', 26308),
 ('with', 23247),
 ('for', 22416),
 ('was', 21917),
 ('f', 20927)]
```

**Figure 8: most common words in positive review**

- After finding the most common words in positive or negative review I have calculated the positive to negative ratio as shown in figure:

```
print("Pos-to-neg ratio for 'the' = {}".format(pos_neg_ratios["the"]))
print("Pos-to-neg ratio for 'amazing' = {}".format(pos_neg_ratios["amazing"]))
print("Pos-to-neg ratio for 'terrible' = {}".format(pos_neg_ratios["terrible"]))
```

```
Pos-to-neg ratio for 'the' = 0.05902269426102881
Pos-to-neg ratio for 'amazing' = 1.3919815802404802
Pos-to-neg ratio for 'terrible' = -1.723488697472832
```

**Figure 9: pos-to-neg ratio**

- Since the English alphabetical data is converted to mathematical value. And now will be passed through the neural network for testing purposes.
- The data is passed through the neural network with some randomized weights on each link using feed forward process.
- Based on the weight model predict the value and based on these predictions the value is compared with the actual output.
- Error functions are calculated based on these predictions and actual values.
- Backpropagation technique is used to adjust these weights by minimizing the error function.
- These error is minimized using stochastic gradient descent and after that new weight is adjusted.
- This step is performed several times keeping in mind that model doesn't over fit the given training data.
- After training process testing is done and based on this dataset our accuracy factor is calculated.
- In this model we set the probabilistic values i.e. if we get the actual output  $> 0.5$  then review would be positive and if the output  $< 0.5$  then the review would be negative otherwise neutral.

## **CHAPTER 2**

### **LITERATURE SURVEY**

In the recent years, enormous work is carried out in the domain of Weather forecasting. Weather forecasting is one of the applications to predict state of climate in future at a given location.

#### **2.1 "Big Data Techniques for efficient storage and processing of weather" [1]**

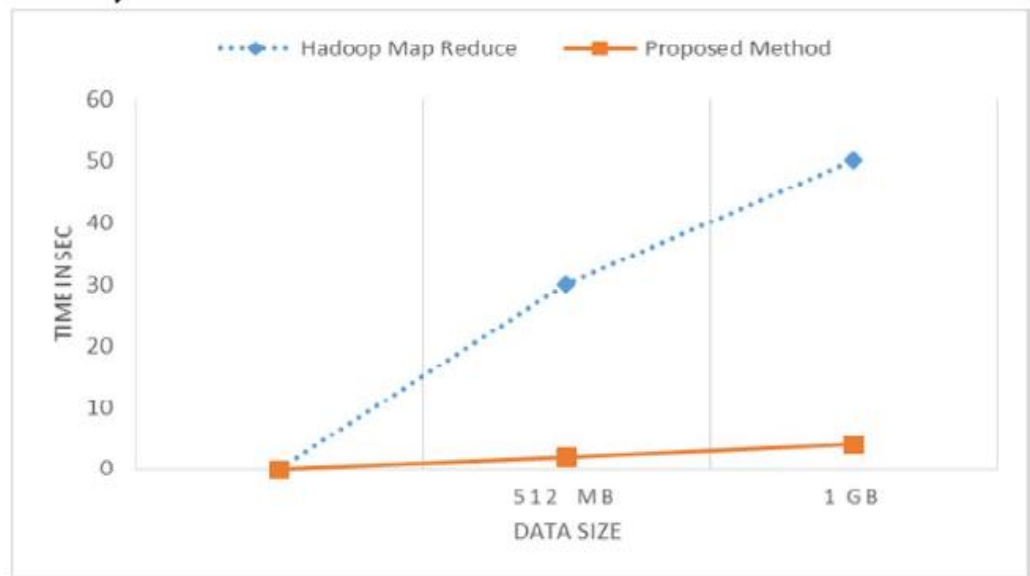
This Research paper proposes an efficient Big Data technique for storage and processing of weather data. In general Apache Hadoop framework is most popularly use for storing and processing of enormous dataset. During this study, Apache Spark and Cassandra integration is experimented to judge the time taken to efficiently store any datasets and process it and therefore the result is evaluated with Hadoop Map Reduce

Weather datasets is collected from National Climatic Data Centre (NCDC). In weather forecasting the raw information is received through satellite delivered over to the various weather stations and this data stored in cluster. Traditional Database like SQL are not best to handle unstructured data or weather data. Input datasets contain field like location, date, temperature, humidity, pressure, rain, wind etc.

#### **Methodology used in this Research Paper:-**

- Hadoop Map Reduce implementation:-Hadoop Framework works on parallel processing distributed system which are conventionally based on map reduce jobs. The Input data is splitted into split. These splits are passed to mapper and the result output is given as input to reducer. Hence in this research paper spark used to process weather data compared to Hadoop Map Reduce.
- Spark implementation: - Due to its in-memory computation it can perform 10x better than Map Reduce. Core concept in Apache Spark is RDD which act as a table in database and can hold various type of data and store on different partition.
- NoSQL Database Cassandra:-NoSQL Database provides a provision for storage and retrieval of unstructured data unlike the traditional database which use tabular relations. NoSQL Database are being efficiently used for real time web application and high speed online transactional data.





**Graph 1: Hadoop Map Reduce v/s Spark Cassandra Benchmarking [1]**

## **2.2 “Commercial Product Analysis Using Hadoop Map Reduce” [2]**

It examines how an association can find certified open entryways in solidifying disengaged and online data to give cleverness on how consolidating separated and online data can be helpful. Associations use proposition estimations which have the above favorable circumstances. Proposition computations are best seen for their use on online business Web destinations. Here they use customer's interests as a commitment to make a record of endorsed things.

The first one is called content based sifting. Content based separating can moreover be called as intellectual sifting, which endorses things dependent on an examination between the substance of the things and a customer profile.

Additionally, the next one is community oriented sifting. It relies on not just the attributes of the things yet rather how person's for example various customers respond to comparable articles. Affiliations need to get all of the data characteristics, detached and on the web, into a lone database, which would be moreover refined by front line examination procedures, and use the solidified data for exactness concentrating on.

### **2.3 “Review paper on Hadoop and Map Reduce” [3]**

Gigantic Data is a data whose scale, superior to average collection, and multifaceted nature require new planning, procedures, figuring, and examination to guide it and center respect and masked picking up from it.

Hadoop is the center stage for dealing with Big Data, and manages the issue of making it huge for examination purposes. Hadoop is an open source programming experience that draws in the appropriated treatment of huge enlightening accumulations crosswise over packs of thing servers. It is proposed to scale up from a singular server to a large number machines, with an anomalous condition of acclimation to inside disillusionment.

This paper totally inspects why hadoop is better in all edges. Seeking after focuses base on the upsides of hadoop.

Flexibility surpasses desires at dealing with data of complex nature and its open-source nature makes it much surely understood. In this paper the structure of hadoop and guide diminish is cleared up. Guide Reduce has been appeared as a free stage as preference layer legitimate for various need by cloud suppliers. It in addition empowers clients to value the data dealing with and investigating.

### **2.4 "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" [4]**

The paper examined a securities exchange expectation probability dependent on gathering of information originating from the required tweets from Twitter small scale blogging stage.

Tweets just in dialect English were utilized in this undertaking work. Retweeted posts were viewed as excess for arrangement, so were evacuated. After information pre-handling, each tweet was spared as a model of pack of-words, a standard procedure for disentangling the spoke to data utilized in data recovery.

The structure of the framework comprised of four noteworthy parts:

- Retrieving, pre-processing and sparing the twitter information to our database
- Stock showcase information recovery

Extremity examination is that part of sentiment investigation, in which the information is gathered either as positive or negative. Customized estimation identification of tweets was pulled off by utilizing Senti Word Net. Future stock costs expectation is performed in this paper by joining the consequences of grouped assumption tweets and stock prices from some past interim.

Viewing huge amounts of information as sorted and the reality they are a composed content, the Naïve Bayes calculation was chosen for its quick procedure of preparing even with huge amount of preparing information and the way that it very well may be expanded. Considering huge amounts of information additionally result choice to execute the guide diminish rendition of Naïve Bayes calculation.

## **2.5 “Python – The Fastest Growing Programming Language” [5]**

This paper started with introduction of python as a high state programming.

Why python is growing so rapidly is discussed next because of its attributes like mobility, simple to learn, open source, etc. Further its drawbacks are discussed like its moderate nature and besides it is hard to keep up.

Various activities have been recorded in python and it has been used in Irobot, Google, and Intel, etc.

## **2.6 “Machine Learning Algorithms: A Review” [6]**

In this investigation paper, distinctive AI estimations have been discussed. This paper gives quick and dirty illumination of the classes of the computations. In the coordinated learning arrangement, three counts have been inspected for example gullible bayes, bolster vector machine and choice tree.

In the unsupervised learning, two algorithms have been discussed for example K-implies grouping and vital part examination.

## **2.7 “An Overview of Deep Learning” [7]**

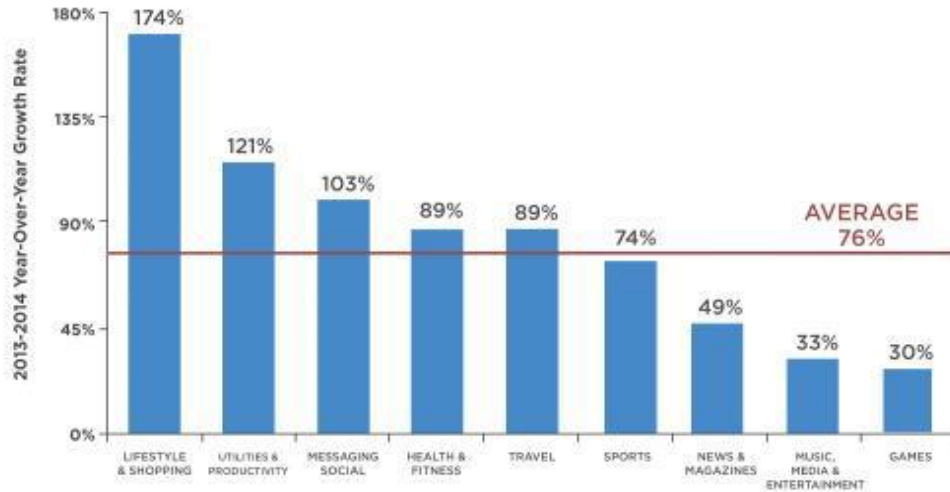
Deep learning approaches are essential for us to take care of numerous issues. In this paper, we present deep learning models and structures in detail. Deep learning various types of models and structures, and it has had numerous applications in numerous perspectives. From these, we can see that deep learning has an incredible advancement potential.

In future, it is predictable that deep learning could set up ideal speculations to clarify its exhibitions. In the mean time, its capacities of unsupervised learning will be improved since there are the great many information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize this points of interest to achieve more assignments.

## **2.8 “The Power of Mobile Applications” [8]**

The importance of cell phones and applications has been developing massively for past years as far as use measurements, download checks and number of applications on the business sectors. As indicated by Flurry Analytics (2015), portable application utilization developed by %76 in 2014.

## Mobile Use Grows 76% Year-Over-Year (Sessions)



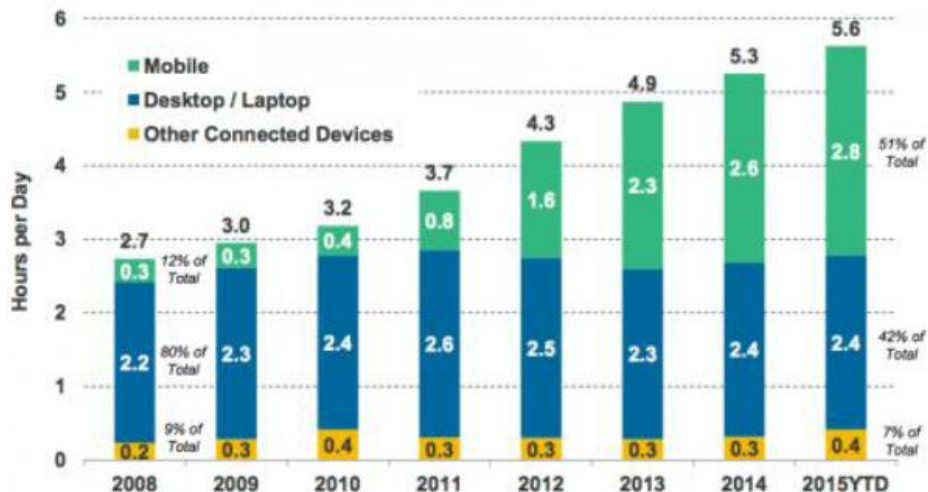
FLURRY

Source: Flurry Analytics

**Graph 2: Mobile Use Grows Year-Over-Year [8]**

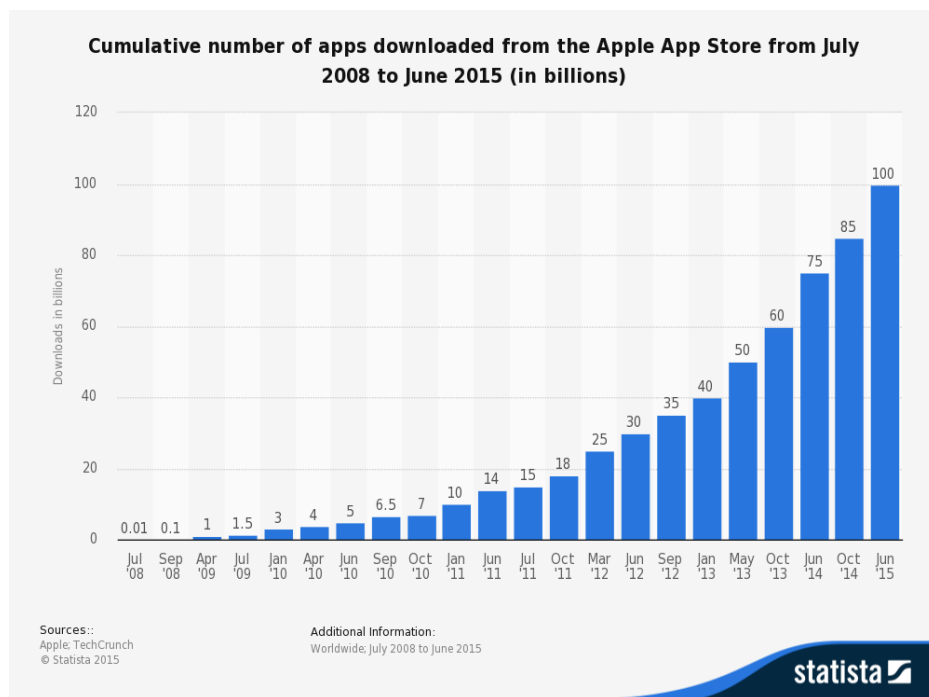
In addition, late insights uncovers that versatile media use overwhelms most of absolute media utilization. Kleiner Perkins Caufield and Byers' report (2015) demonstrates that time spent by clients on versatile advanced media is the most astounding contrasted with others

## Time Spent per Adult User per Day with Digital Media, USA, 2008 – 2015YTD



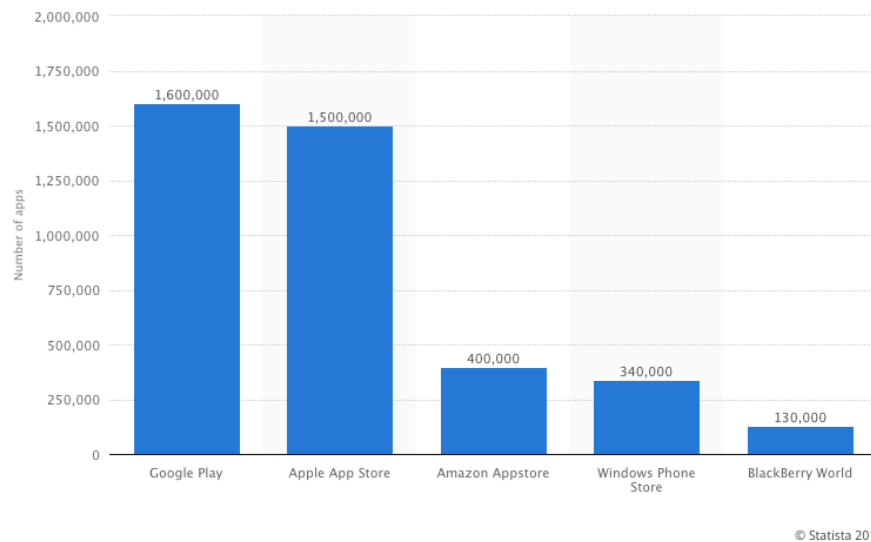
**Graph 3: Time Spent per Day with Digital Media [8]**

Apple as of late reported that the App Store passed 100 billion application downloads since it is opened in 2008 (CNET, 2015). Apple's CEO Tim Cook made the declaration at Apple's Worldwide Developers Conference (WWDC) while he noticed that it has been paid \$30 billion dollars to application designers by Apple, the realistic by Statista (2015) demonstrates the development in number of downloads in App Store from 2008 to 2015.



**Graph 4: Number of Apps Downloaded [8]**

Increment in the fame of versatile condition makes the field focused for individual engineers, organizations and merchants (for example Apple, Google, Windows, Amazon and so forth.). Since engineers need to get required to portable segment, the quantity of applications in stores increments and make the challenge increasingly extreme. The quantity of accessible applications for July, 2015 is appeared in the Figure 4 which uncovers that there are in excess of 3 million applications in Google Play and Apple App Store (Statista, 2015).



**Graph 5: Number of Total Apps in Mobile Markets [8]**

This circumstance causes issues for the two designers and buyers. For customer's case, the principle issue is essentially to pick right application for the correct reason in a great many applications. Then again, engineers ought to keep up solid input procedure to improve their applications, make new features and defeat the shortcomings (Chen et al., 2014) so as to draw in more clients. Regardless of other programming circulation channels, versatile stores offer clients to capacity to rate applications and make remarks.

## **2.9 “Mobile App Analytics” [9]**

Bitterer (2011) portrayed Mobile Business Intelligence (Mobile BI) as one of the new innovations which have the capability of disturbing the Business Intelligence (BI) advertises. Thinking about the intensity of versatile promotions (Snider, 2012), the nature of the portable environment, which offers capacity to gather customized and area explicit substance, has given significant open doors for Business Intelligence and Analytics (Chen et. al, 2012).

One of the significant information which versatile application investigation decides to process is "Star Ratings" of utilizations. Star evaluations which exhibit the normal voted rating of the applications can impact incomes and point of confinement the rate of development (Vasa et. al., 2012). So as to exist in this exceedingly aggressive market, engineers ought to demonstrate the nature of their applications with high application evaluations (Khalid, 2013).

As it is expressed by Chevalier and Myzalin (2006), he said that review of the customer has a great impact on sales of particular application. He also said that review contain some important information including functionality, about failure of apps, weakness of user interface, bugs at the time of update etc. The main motive of the developers is how to respond those feedbacks so that all those vulnerabilities can be removed easily from the particular app. Hence, there is a requirement for mechanized arrangements dissecting surveys and changing to instructive information.



## CHAPTER 3

### SYSTEM DEVELOPMENT

#### 3.1 Designing

##### (i) Hive Data Models

Hive is an open source platform that can store information as tables and then we can use its features i.e. data summarization, data analysis, data query.

##### **Tables**

Hive containing tables resemble social database tables. Tables of Hive contain data and their plan is depicted with the help of related metadata. Channel, join and association exercises can be performed on these tables.

Table containing in Hive:

For managed table, following syntax is used:

**Create table tablename(Var String, Var Int) row format delimited fields terminated by ‘;’;**

**Eg;** create table rating(id String, name String, categories String, dateAdded String, doRecommend String, rating in , text String , title String , username String) row format delimited fields terminated by ‘;’;

Data is loaded by:

**Load data local inpath ‘<pathof the file>’ into table tablename;**

**E.g.;** load data local inpath ‘googleplaystore.csv’ into table playstore;

## (ii) Deep Neural Network models

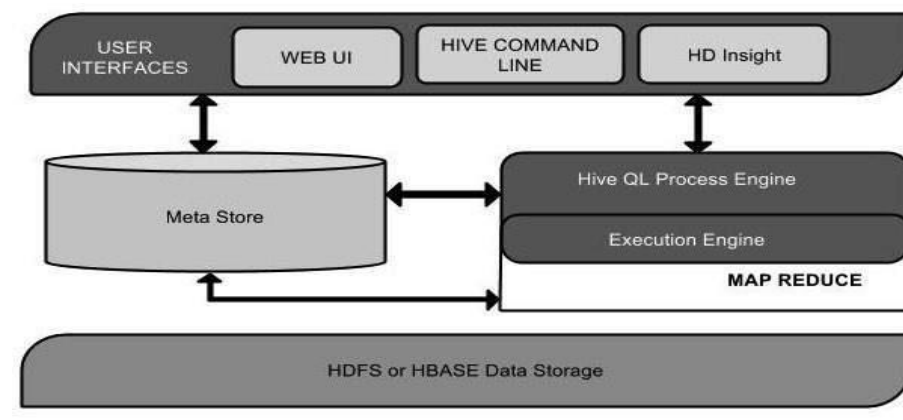
**Training from Scratch:** To prepare a deep learning model from the very initial stage, we assemble a substantial named informational collection and design a related network structure that will get familiar with all the attributes/features and model. This is useful for new applications, or applications that will have an expansive number of yield classifications. This is a less normal methodology in light of the fact that with the substantial measure of information and rate of learning, these systems ordinarily take days or weeks to prepare.

Designing of DNN model for my project as:

- Set the position of each word into an array
- Make the different layers i.e. one input layer, one hidden layer, one output for neural network.
- Send the randomized weight to each layer.
- Add sigmoid function to hidden layer and output layer.
- Use the Feed forward to calculate and comparing it with actual output and then calculate the error. Based on this error weights are adjusted using Back propagation.

## 3.2 Architecture

### 3.2.1 Hive Architecture



**Figure 10: Apache Hive Architecture**

**User Interface:** Hive is an information warehouse framework programming that can make cooperation among client and HDFS. The UIs that Hive supports are Hive Web UI, Hive order line, and Hive HD Insight (In Windows server).

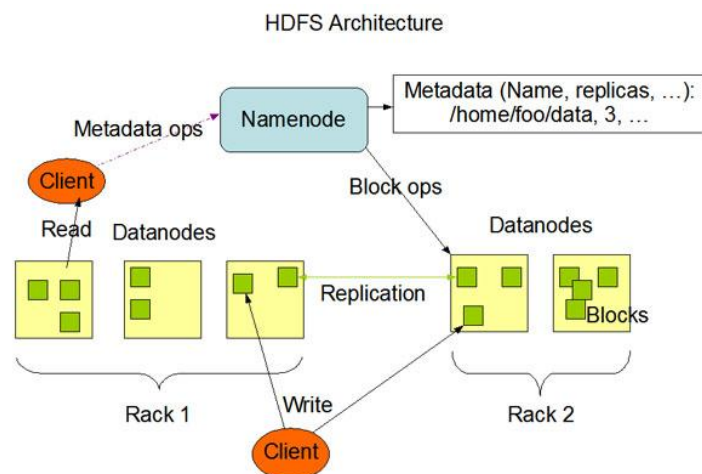
**Meta Store:** Hive picks individual database servers to store the composition or Metadata of tables, databases, sections in a table, their information types, and HDFS mapping.

**HiveQL Process Engine:** HiveQL is like SQL for querying on pattern data on the Meta store. It is one of the substitutes of conventional methodology for MapReduce program. Rather than composing MapReduce program in Java, we can compose an inquiry for MapReduce work and analyze it.

**Execution Engine:** The combination part of HiveQL process Engine and MapReduce is Hive Execution Engine. Execution engine procedures all queries and creates results as same as MapReduce results. It utilizes the kind of MapReduce.

**HDFS or HBASE:** These are the tools to store all the data into the file.

### 3.2.2 HDFS Architecture



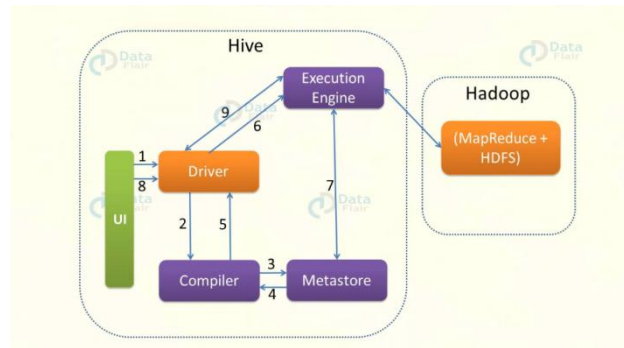
**Figure 11: HDFS Architecture**

The Hadoop Distributed File System pursues master/slave information design. Each group contains a single Name node that goes about as the master server so as to deal with all the document framework namespace and give the correct access to clients. The following phrasing in the HDFS bunch is the Data node that is normally one for each node in a HDFS cluster. The Data node is relegated with a task of dealing with the capacity connected to the node that it keeps running on. HDFS likewise incorporates a document framework namespace that is being executed by the Name node for general

activities like record open, close, rename and even registries. The Name node additionally maps the blocks to the Data node.

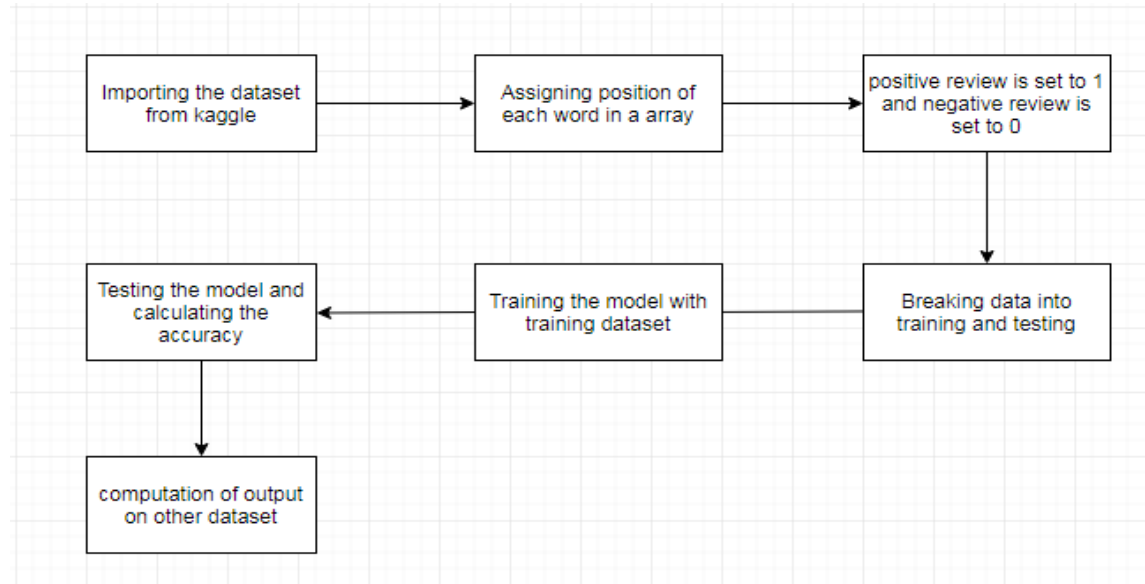
### 3.3 Data Flow

#### 3.3.1 Data Flow in Hive



**Figure 12: Data Flow Diagram**

#### 3.3.2 Data Flow in Deep Neural Network



**Figure 13: Data Flow in Deep Neural Network**

### **Steps in flowchart:**

1. Firstly, the dataset is imported from the web (kaggle.com) and divided into training and testing.
2. Assigning the position of each word from the dataset in an array because our dataset is in text form so we cannot directly pass this dataset into our model.
3. Now, set the positive review to 1 and negative review to 0.
4. Break the data into two forms i.e. training and testing for accuracy purpose.
5. Train the model according to the features/attributes of training dataset.
6. Pass the testing dataset into testing model and calculate the accuracy.
7. Now, pass your own dataset for calculating the output and check the accuracy.

### **3.4 Requirements**

#### **3.4.1 Hardware Requirements**

<b>Type of hardware</b>	<b>Hardware used</b>
Processor	-multiprocessor-based with a 2.00 GHz 64-bit Operating system
Hardware	-Quad core Intel i3 has 4 GB RAM
Memory	-4 GB (2.4 GB on virtual machine)
Disk space	-8 GB free disk space. Requirements increase as data is gathered and stored in HDFS.

### **3.4.2 Software Requirements:**

<b>Type of software</b>	<b>versions</b>
Operating System	Linux (Ubuntu )/Windows
VMware workstation	14.1.2
Horton works sandbox	HDP 1.2.4
Hadoop	2.7.3
Netbeans IDE	7.0.1
Eclipse IDE	4.8
Web browser	Mozilla Firefox 28.
Python Jupyter	4.3.0

### **3.4.3 Advantages of VMware Workstation:**

- Ability to clone virtual machines frameworks.
- It can develop various software under many frameworks.
- Turn your own machine into a virtual machine framework.
- It can Import and modify virtual machines.

## **3.5 Test Plan**

### **3.5.1 Dataset**

Most regularly a dataset relates to the matter of the single database table, or the single factual information framework, where each segment of the table speaks to a specific variable, and each column compares to a given individual from the informational collection being referred to.

This information is of Google play store application and is taken from **Kaggle** , which is the world's largest community for data scientists to explore, analyze and share data. I have two datasets one is googleplaystore.csv and other is googleplaystore\_user\_reviews.csv. With the help of these two dataset I will perform both the analysis and prediction part respectively.

### (1) googleplaystore.csv

This dataset is for Web scrapped information of 10k Play Store applications for examining the Android advertises. This is the offline dataset which can be utilized by a client to have the Android market of different utilization of various classes music, camera etc. With the assistance of this, client can foresee whether any given application will get lower or higher rating level, which app is free or which is paid, about the reviews, about the latest version of the particular app and many more things to analyze. This dataset can be additionally utilized for future references for the suggestion of any application. In addition, this offline dataset is chosen in order to decide the forecast precisely as online information gets refreshed all around much of the time. Likewise, to monitor its old clients to comprehend their inclinations better.

This dataset have 13 columns of various categories of the application. In this project I have analyzed all these various columns of the dataset.

The 13 columns of the dataset are as follows:

App	Application Name
Category	Category the app belongs to
Rating	Overall rating of the app
Reviews	Number of user reviews for the app
Size	Size of the app
Installs	No of user downloads/installs the app
Type	Paid or free
Price	Price of the app
Content rating	Age group the app is targeted at-children/Mature/Adult
Genres	An app can belongs to multiple genres For eg-a musical family, game.
Last updated	Date when the app was last updated on google playstore
Current ver	Current version of the app
Android ver	Min required android version

**Table 1: Dataset columns and its specifications**

sample record from the offline-dataset (googleplaystore.csv ) is shown in the following table :

▲ App	▲ Category	# Rating	# Reviews	▲ Size	▲ Installs	▲ Type
Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free
Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free
U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free

**Figure 14: (i) Sample record of dataset**

▲ Price	▲ Conten...	▲ Genres	📅 Last Up...	▲ Current...	▲ Androi...
0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

**Figure 15: (ii) Sample record of dataset**

## (2) googleplaystore\_user\_reviews.csv

This dataset contains the principal 'most significant' 100 surveys for each application. Each survey content/remark has been pre-handled and credited with 3 new highlights - Sentiment, Sentiment Polarity and Sentiment Subjectivity. Through this dataset I am going to predict which review is positive and which is negative.

This dataset contains 5 columns of the user review for the particular app. The 5 columns of the dataset are as follows:

App	Name of the app
Translated review	User review
Sentiment	Positive/negative/neutral
Sentiment_polarity	Sentiment polarity score
Sentiment_subjectivity	Sentiment subjectivity score

**Table 2: Dataset columns and its specifications**



Record from the downloaded-dataset ( googleplaystore\_user\_reviews.csv ) is shown in the following table:

App	Transla...	Sentim...	# Sentim...	# Sentim...
10 Best Foods for You	I like eat delicious food. That's I'm cooking food myself. case "10 Best Foods" helps lot. also "Best Before (Shelf Life)"	Positive	1.0	0.5333333333333333
10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.28846153846153844

Figure 16: Sample record of dataset

### 3.6 Algorithms

#### 3.6.1 Algorithm used for Analysis part

**3.6.1.1 Map Reduce Algorithm:** This algorithm is a Distributed Data Processing algorithm, presented by Google. This algorithm is used for the most part valuable to process colossal measure of information in parallel, solid and productive path in bunch conditions

MapReduce Algorithm uses the following three main steps:

1. Map Function
2. Shuffle Function
3. Reduce Function



Figure 17: How Map Reduce works

#### **1. Map Function: -**

- This function works on the basis of Key & value pairs.
- It takes input undertakings and partitions them into littler sub-errands and after that perform required calculation on each sub-task in parallel.
- Those key & value pairs should be in the form of byte offset values.
- Then this should be pass to the mapper function map().
- Then the output of this map() function should be in form (K,V) pairs.

**2. Shuffle Function:-** This function is also known as combine function.

- Output of the mapper function will be taken as input to this function.
- In this mainly we works on the grouping of data collecting from different nodes which is based on the keys.
- In this we have two sub-steps:-
  - Merging- it is used to combines all key-value that has same keys.
  - Sorting- input of this step is the output of the above step and sort them.

**3. Reduce Function:-**

- This takes the input as the output of the Shuffle function and performs its own operation.
- Reducer consolidates every one of these qualities together and gives single yield an incentive to the particular key.
- Reduce step in this function <Key, Value> pairs are totally different from mapping step <Key, Value> pairs.

After function of the Reduce Phase, the group gathers the information to frame a suitable outcome and sends it back to the Hadoop server.

### 3.6.2 Algorithm used for Prediction part

**3.6.2.1 Backpropagation:** Backpropagation is a strategy utilized in artificial neural systems to figure a slope that is required in the estimation of the loads to be utilized in the network.

Backpropagation is shorthand for "the backward propagation of error," since an error is processed at the yield and circulated in reverse all through the network's layers. It is usually used to prepare deep neural networks.

Backpropagation is a speculation of the delta principle to multi-layered feedforward systems, made conceivable by utilizing the chain rule to iteratively compute gradient for each layer. It is firmly identified with the Gauss– Newton calculation and is a piece of proceeding with research in neural backpropagation.

Backpropagation is an extraordinary instance of a progressively broad procedure called programmed separation. With regards to learning, backpropagation is normally utilized by the gradient descent optimization algorithm to change the heaviness of neurons by ascertaining the slope of the loss function.

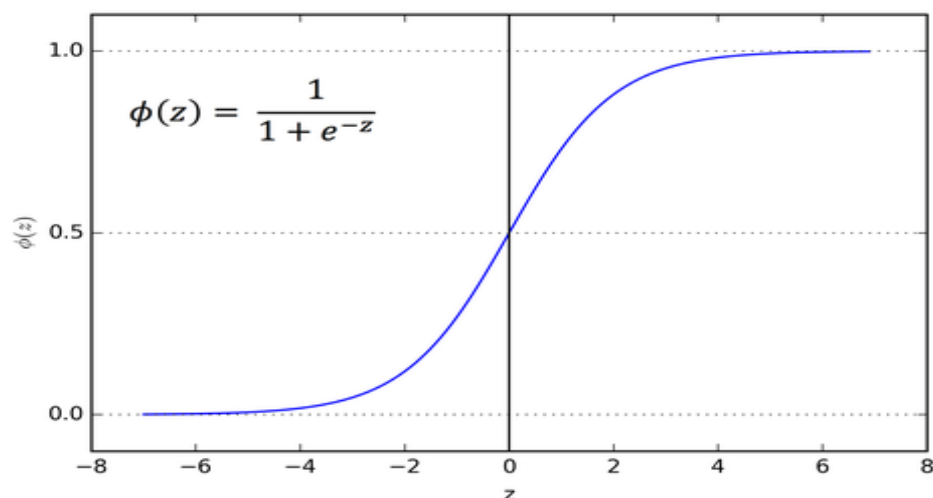
In this project we have used this algorithm as follows:

1. Initially, we will pass the converted dataset through the neural network which has hidden layer and the sigmoid activation function is used. Sigmoid activation main work is to scale the given input between zero and one.

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

$$-\text{INF} < x < \text{INF}$$

$$0 < \text{Sigmoid}(x) < 1$$



### Graph 6: Sigmoid function

2. We will set the randomized weights for each link in our neural networks. The random from distributions like random, Gaussian or normal distribution. The weights initialization is picked from an appropriate distribution because sometimes the incorrect picking of hyper parameters may lead to a huge problem.

**3.6.2.2 Feed Forward:** A feed forward neural system is an artificial neural system wherein associations between the nodes don't frame a cycle.

The feed forward neural system was the first and most straightforward sort of neural system contrived. In this system, the data moves in just a single bearing, forward, from the info nodes, through hidden hubs (assuming any) and to the yield nodes. There are no cycles or circles in the system.

In this project we have used this algorithm as follows:

1. In this step the complex multiplication will be done and output is calculated. For example let's say there is a review "This is a worst application". In this case the model predicted a value let's say 0.4. So our model will say that value should be as close to zero as possible. The error is calculated and minimized by the backpropagation techniques.

$$\text{Error} = (\text{Actual} - \text{Predicted})^2$$

This error needs to be optimized.

2. In this step error function is optimized by using various techniques. One of the technique is stochastic gradient descent. In this technique we will optimize the function on the basis of gradient.

We will take a step in the direction of negative gradient. In our case the step size is set to 0.01 i.e a small step towards minima. This is done so that we don't skip the minima by taking a large steps. This will help us to achieve good accuracy.

```
[39] mlp = SentimentNetwork(reviews[:1000], labels[:1000], min_count=20, cutoff=0.2, learning_rate=0.01)
```

**Figure 18: Display of learning rate=0.01**

D->Denotes the partial derivative

New weight = Old Weight – (D(E)/D(W))\*Learning Rate.

3. After this step we will predict the accuracy on the testing dataset and our model is now good to go. Testing data is also one hot encoded.

4. After this we will write our own review of app and the prediction is done by our neural network model.

## CHAPTER 4

### RESULTS AND PERFORMANCE ANALYSIS

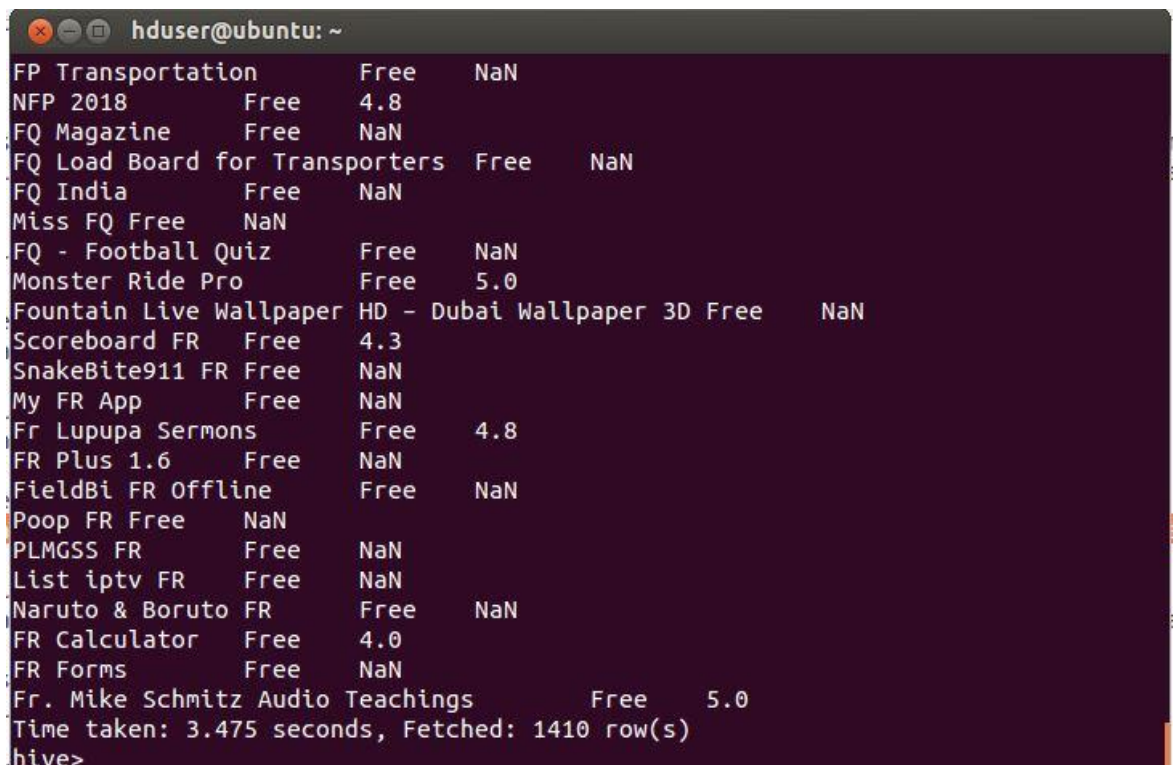
After implementing the project on Hive and Deep neural network I have observed the following results.

#### 4.1 Analysis.

##### 4.1.1 Using Hive

In this I have loaded the given dataset into hdfs and then I have created database and tables on hive and then by applying different queries to the different attributes of the given google playstore dataset I have tried to analyze those attributes such as which is free or paid apps among all the applications present in the given dataset, which app is most reviewed by the user, about the rating of the particular application.

(i) Firstly, I have tried to find out names of those applications which come on the top of the entire free app with their respective rating. This picture shows all the names of top free apps.



```
hduser@ubuntu: ~
FP Transportation      Free      NaN
NFP 2018              Free      4.8
FQ Magazine           Free      NaN
FQ Load Board for Transporters  Free      NaN
FQ India              Free      NaN
Miss FQ Free          NaN
FQ - Football Quiz    Free      NaN
Monster Ride Pro      Free      5.0
Fountain Live Wallpaper HD - Dubai Wallpaper 3D Free      NaN
Scoreboard FR        Free      4.3
SnakeBite911 FR      Free      NaN
My FR App            Free      NaN
Fr Lupupa Sermons     Free      4.8
FR Plus 1.6          Free      NaN
FieldBi FR Offline   Free      NaN
Poop FR Free         NaN
PLMGSS FR            Free      NaN
List iptv FR         Free      NaN
Naruto & Boruto FR    Free      NaN
FR Calculator         Free      4.0
FR Forms             Free      NaN
Fr. Mike Schmitz Audio Teachings      Free      5.0
Time taken: 3.475 seconds, Fetched: 1410 row(s)
hive>
```

**Figure 19: Top free applications**

(ii) In this I have tried to find out names of those applications which comes on the top of all the paid apps with their respective rating. This picture shows all the names of top paid apps among all the presented apps in given dataset.

```

hduser@ubuntu: ~
Eu Sou Rico      Paid      NaN
IF YOU TO EU PEGO      Paid      NaN
I'm Rich/Eu sou Rico/أنا غني/我很有钱      Paid      NaN
EX File Explorer File Manage Pro      Paid      NaN
An Elite Warrior Ex      Paid      4.7
Volume Control Ex      Paid      4.2
Galaxian(FC)      Paid      4.5
SCI-FI UI      Paid      4.7
FJ Toolkit      Paid      NaN
FN pistol Model 1906 explained      Paid      NaN
FN pistol model 1903 explained      Paid      NaN
"The FN "Baby" pistol explained"      Paid      NaN
FN FAL rifle explained      Paid      NaN
The FN HP pistol explained      Paid      NaN
FN model 1900 pistol explained      Paid      NaN
Pistolet FN GP35 expliqué      Paid      NaN
Pistolet FN 1906 expliqué      Paid      NaN
Circle Colors Pack-FN Theme      Paid      4.2
FO Bixby      Paid      5.0
Mu.F.O. Paid      5.0
FP VoiceBot      Paid      NaN
Word Search Tab 1 FR      Paid      NaN
Time taken: 0.146 seconds, Fetched: 239 row(s)
hive>

```

**Figure 20: Top paid applications**

(iii) In this I have tried to find out names of those applications which come on the top of all the reviewed apps. This picture shows all the names of apps which are most reviewed by the users among all the presented apps in given dataset.

```

hduser@ubuntu: ~
Candy Crush Saga      22429716
Candy Crush Saga      22430188
Subway Surfers      27725352
Clash Royale      23136735
Clash of Clans      44893888
Subway Surfers      27725352
Candy Crush Saga      22430188
Facebook      78158306
Instagram      66577313
Instagram      66577446
Instagram      66577313
YouTube      25655305
Subway Surfers      27711703
WhatsApp Messenger      69109672
Instagram      66509917
YouTube      25623548
Facebook      78128208
Clash of Clans      44881447
Clash Royale      23125280
Candy Crush Saga      22419455
Clean Master- Space Cleaner & Antivirus      42916526
Messenger - Text and Video Chat for Free      56642847
Time taken: 0.077 seconds, Fetched: 36 row(s)
hive>

```

**Figure 21: Top reviewed applications**

(iv) In this I have tried to find out names of those applications which come on the top of Editor's choice. This picture shows all the names of the apps which come under the editor's choice whose rating is above 4 and has more reviews among all the presented apps in given dataset.

```

hduser@ubuntu: ~
Candy Crush Saga      4.4      22429716
Candy Crush Saga      4.4      22430188
Subway Surfers        4.5      27725352
Clash Royale          4.6      23136735
Clash of Clans        4.6      44893888
Subway Surfers        4.5      27725352
Candy Crush Saga      4.4      22430188
Facebook              4.1      78158306
Instagram             4.5      66577313
Instagram             4.5      66577446
Instagram             4.5      66577313
YouTube 4.3          25655305
Subway Surfers        4.5      27711703
WhatsApp Messenger    4.4      69109672
Instagram             4.5      66509917
YouTube 4.3          25623548
Facebook              4.1      78128208
Clash of Clans        4.6      44881447
Clash Royale          4.6      23125280
Candy Crush Saga      4.4      22419455
Clean Master- Space Cleaner & Antivirus 4.7      42916526
Messenger - Text and Video Chat for Free 4.0      56642847
Time taken: 0.076 seconds, Fetched: 36 row(s)
hive>

```

**Figure 22: Editor's choice apps**


## **4.2 Prediction**

After analyzing all the attributes of given dataset using hive I have made the prediction on the user reviews of the particular application that which particular review is positive or negative using Deep neural network .First I have read the given dataset which is in the form of text and converted it into mathematical form then pass this dataset in the deep neural network and calculate the output compare it with actual output and then train the model and adjust the weights to minimize the error using backpropagation this process is performed several times. Finally by comparing the output after training process with the set range to find out which review is positive or negative.

Here are some results which come after performing prediction on the user reviews on the particular application in the given dataset.



(i) These pictures showing the result after providing the review on any particular application.



```
print(mlp.run("What an amazing application"))
```

POSITVE

---

**Figure 23: (i) Review is positive**



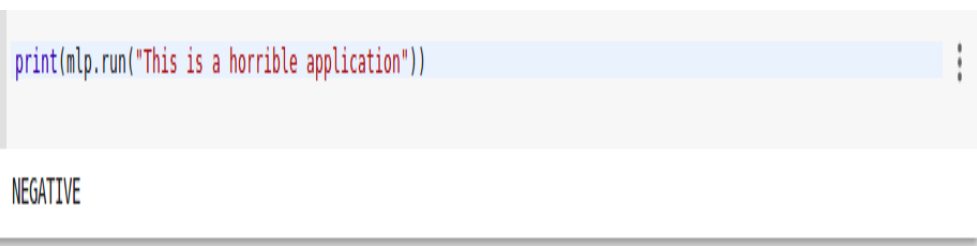
```
print(mlp.run("This is a great application"))
```

POSITIVE

---

**Figure 24: (ii) Review is positive**

(ii) These pictures showing the result after providing the review on any particular application.



```
print(mlp.run("This is a horrible application"))
```

NEGATIVE

---

**Figure 25: (i) Review is negative**



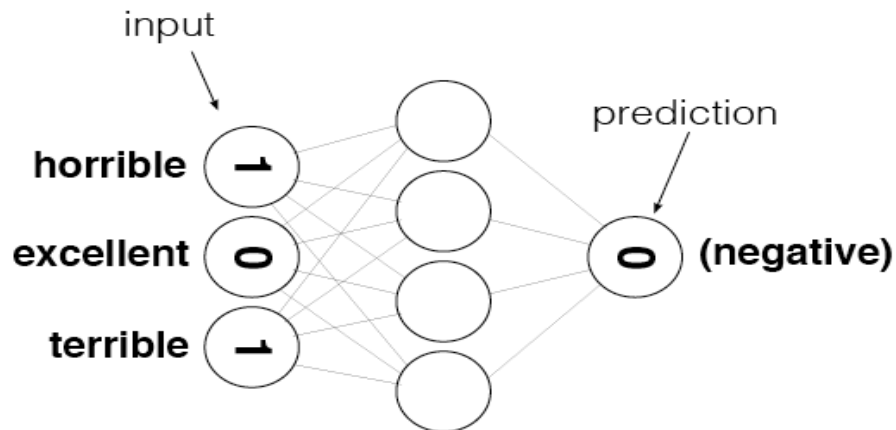
```
print(mlp.run("This is a worst application I have ever used"))
```

NEGATIVE

---

**Figure 26: (ii) Review is negative**

(iii) This picture shows how we are predicting the positive or negative user review using deep neural network.



**Figure 27: Prediction of negative review**

(iv) This picture shows the accuracy level of model without training for the given dataset.

```
m1p.test(reviews[-1000:]*2, labels[-1000:]*2)
Progress:99.9% Speed(reviews/sec):1858. #Correct:1000 #Tested:2000 Testing Accuracy:50.0%
```

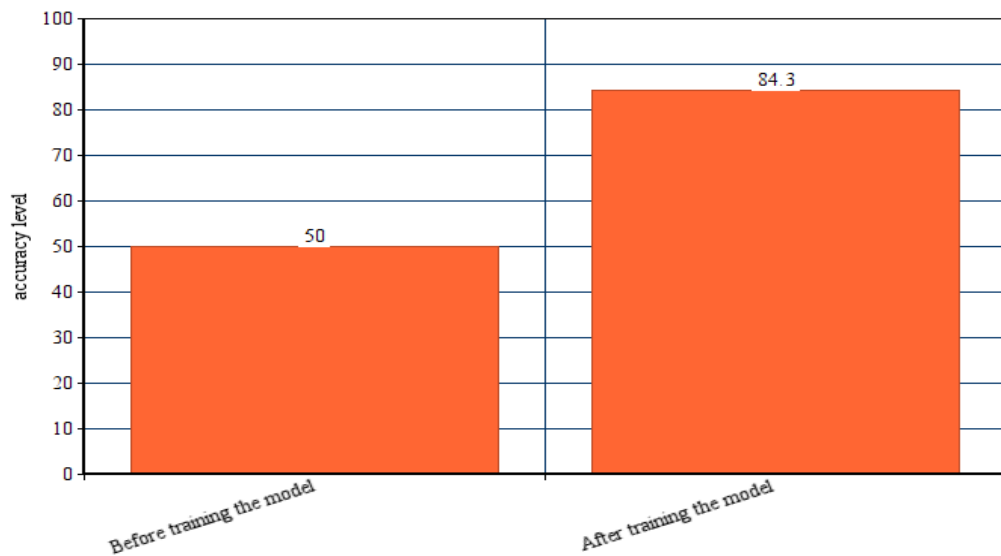
**Figure 28: Total Accuracy without training**

(v) This picture shows the accuracy level of model with training for the given dataset.

```
m1p.train(reviews[: -1000], labels[: -1000])
Progress:0.0% Speed(reviews/sec):0.0 #Correct:0 #Trained:1 Training Accuracy:0.0%
Progress:10.4% Speed(reviews/sec):2292. #Correct:1729 #Trained:2501 Training Accuracy:69.1%
Progress:20.8% Speed(reviews/sec):2133. #Correct:3777 #Trained:5001 Training Accuracy:75.5%
Progress:31.2% Speed(reviews/sec):2157. #Correct:5915 #Trained:7501 Training Accuracy:78.8%
Progress:41.6% Speed(reviews/sec):2140. #Correct:8054 #Trained:10001 Training Accuracy:80.5%
Progress:52.0% Speed(reviews/sec):2136. #Correct:10213 #Trained:12501 Training Accuracy:81.6%
Progress:62.5% Speed(reviews/sec):2146. #Correct:12363 #Trained:15001 Training Accuracy:82.4%
Progress:72.9% Speed(reviews/sec):2155. #Correct:14520 #Trained:17501 Training Accuracy:82.9%
Progress:83.3% Speed(reviews/sec):2155. #Correct:16714 #Trained:20001 Training Accuracy:83.5%
Progress:93.7% Speed(reviews/sec):2161. #Correct:18920 #Trained:22501 Training Accuracy:84.0%
Progress:99.9% Speed(reviews/sec):2159. #Correct:20246 #Trained:24000 Training Accuracy:84.3%
```

**Figure 29: Total Accuracy with training**

(vi) This graph depicts the accuracy level of model before training and after training the model for the given dataset.



**Graph 7: Accuracy level**

As we can see from the graph that accuracy changes from 50% to 84.3% after training the model for the given dataset. So using deep neural network we have predicted the user reviews as shown above with optimized error.

### **4.3 Deep Learning over Hive**

Deep Learning allows us to train the model as per our dataset means for the given dataset we can train the model according to our output and also as many times we train the model we will get more and more accuracy every time we train the model. But using Hive we are not allowed to do the above training. We cannot train the model as per our need no matter how much accuracy we get. In Deep Learning we can work on large dataset with more and more accuracy and with more processing speed as compare to Hive.

At the point when there is absence of understanding for highlight thoughtfulness, Deep Learning strategies eclipses others as you need to stress less over element designing.

## **CHAPTER 5**

### **CONCLUSION**

The Google Play Store is the biggest application advertises on this earth. It produces more than the download of the Apple App Store, yet not profits as the App Store. We scratched information from the Play Store to lead examine on it. In this project I have used Big data technique such as Hive to analyze the different attributes of the given dataset of Google playstore application such as top free apps, top paid apps ,most reviewed apps, apps under editor's choice with the help of HiveQL and displayed the results as shown above.

Moreover in the rest of my project I have tried to predict the positive or negative user review on the basis of given dataset using deep neural network. First I have read the given dataset which is in the form of text and converted it into mathematical form then pass this dataset in the deep neural network and calculate the output compare it with actual output and then train the model and adjust the weights to minimize the error using backpropagation this process is performed several times. Finally by comparing the output after training process with the set range to find out which review is positive or negative.

Accuracy of model increases to 84.3% from 50% after training the model which is far better than the previous model.

## REFERENCES

- [1] K.Anusha and K.Usha Rani, "Big Data Techniques for efficient storage and processing of weather." International Journal for Research and Applied Science & Engineering Technology (IJRASET).
- [2] Kshitij Jaju<sup>1</sup>, Vishal Nehe<sup>2</sup>,Abhishek Konduri<sup>3</sup>, ' Commercial Product Analysis Using Hadoop MapReduce', International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 04 | April-2016 , 2016 IJSRSET | Volume 2 |
- [3] Nishant Rajput , Nikhil Ganage ,and Jeet Bhavesh Thakur,"Review Paper on Hadoop and Map Reduce", IJRET: International Journal of Research in Engineering and Technology, Volume: 06 Issue: 09 | Sep-2017
- [4] Andrzej Romanowski, Michal, and Skuza, "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction", 2015 FederatediConference on, pp. 1349-1354. IEEE,i2015.
- [5] K. R. Srinath,' Python – The Fastest Growing Programming Language' International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 12 | Dec-2017.
- [6] Ayon Dey,' Machine Learning Algorithms: A Review', Vol. 7 (3), 2016, ISSN:0975-9646.
- [7] Xuedan Du, Yinghao Cai, Shuo Wang and Lejie Zhang," An Overview of Deep Learning " International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 13 | Nov-2016
- [8] Kleiner Perkins Caufield and Byers",Power of Mobile Applications" International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 23 | Sep-2016.

[9] Ece Calikus, "Mobile App Analytics ", Department of Engineering and Technology, Halmstad University, Volume: 02 Issue: 16 | Nov-2016