

# **“Amazon Rating Analysis and Prediction”**

Project report submitted in partial fulfillment of the requirement for  
the degree of Bachelor of Technology in

**Computer Science and Engineering/Information Technology**

By

Divya Ameta (151376)

Rahul Verma (151384)

Under the supervision of

Dr. Hari Singh

to



Department of Computer Science & Engineering and Information  
Technology

## Candidate's Declaration

I hereby declare that the work presented in this report entitled “ **Amazon Rating Analysis and Prediction** ” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** 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. Hari Singh (Assistant Professor Sr. Grade Dept. of CSE)**.

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

**Divya Ameta (151376)**

**Rahul Verma (151384)**

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

**Dr. Hari Singh**

**(Assistant Professor Sr. Grade)**

**Dept. of CSE**

Dated: 08/05/2019

## Acknowledgement

It is our privilege to express our sincerest regards to our project supervisor **Dr. Hari Singh (Assistant Professor Sr. Grade)**, for their valuable inputs, able guidance, encouragement, whole-hearted cooperation and direction throughout the duration of our project.

We deeply express our sincere thanks to our Head of Department **Prof. Dr. Satya Prakash Ghreera** for encouraging and allowing us to present the project on the topic “**AMAZON RATING ANALYSIS AND PREDICTION**” at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree.

At the end we would like to express our sincere thanks to all my friends and others who helped me directly or indirectly during this project work.

Date: 08-05-2019

Divya Ameta (151376)

Rahul Verma (151384)

## Table of Content

Sr.No.	Topics	Page No.
<b>1</b>	<b>Introduction</b>	<b>1-10</b>
<b>1.1</b>	Introduction	<b>1</b>
<b>1.2</b>	Problem Statement	<b>7</b>
<b>1.3</b>	Objective	<b>7</b>
<b>1.4</b>	Methodology	<b>7</b>
<b>1.4.1</b>	Using Hive	<b>7</b>
<b>1.4.2</b>	Using Machine Learning	<b>10</b>
<b>2</b>	<b>Literature Survey</b>	<b>11-18</b>
<b>2.1</b>	Hive A Warehouse Solution Over A Mapreduce Framework	<b>11</b>
<b>2.2</b>	Pattern Finding In Log Data Using Hive on Hadoop	<b>11</b>
<b>2.3</b>	Scalability Study Of Hadoop Mapreduce And Hive In Big Data Analytics	<b>12</b>
<b>2.4</b>	Commercial Product Analysis Using Hadoop MapReduce	<b>12</b>
<b>2.5</b>	Hive – A Petabyte Scale Data Warehouse using Hadoop	<b>13</b>
<b>2.6</b>	Storage and Processing Speed For Knowledge from Enhanced Cloud Framework	<b>13</b>
<b>2.7</b>	An Overview of the Hadoop/Mapreduce/Hbase framework and its current applications in bioinformatics	<b>14</b>
<b>2.8</b>	Migration of Hadoop to Android platform using ‘Chroot’	<b>14</b>
<b>2.9</b>	Review Paper on Hadoop And Map Reduce	<b>15</b>
<b>2.10</b>	Machine Learning Algorithms: A Review	<b>15</b>
<b>2.11</b>	A Survey on Machine Learning: Concept, Algorithms and Applications	<b>16</b>
<b>2.12</b>	Python – The Fastest Growing Programming Language	<b>16</b>
<b>2.13</b>	Machine Learning for Computer Security	<b>16</b>
<b>2.14</b>	Sentiment Analysis for Amazon Reviews	<b>17</b>
<b>2.15</b>	Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning	<b>17</b>
<b>2.16</b>	Sentimental analysis of Amazon reviews using naïve bayes on laptop products with MongoDB and R	<b>17</b>
<b>2.17</b>	Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining	<b>18</b>

## Table of Content (contd.)

<b>Sr.No.</b>	<b>Topics</b>	<b>Page No.</b>
<b>3</b>	<b>System Development</b>	<b>19-41</b>
<b>3.1</b>	Designing	<b>19</b>
<b>3.1.1</b>	Apache Hive Data Models	<b>19</b>
<b>3.1.2</b>	Machine Learning Model	<b>20</b>
<b>3.2</b>	Experimental Setup	<b>21</b>
<b>3.2.1</b>	For Hive	<b>21</b>
<b>3.2.2</b>	For Machine Learning	<b>23</b>
<b>3.3</b>	Analytical Model	<b>24</b>
<b>3.3.1</b>	Traditional Model- “Schema on Write”	<b>24</b>
<b>3.3.2</b>	Big Data Model- “Schema on Read”	<b>25</b>
<b>3.3.3</b>	Data Modeling in the Big Data Ecosystem	<b>26</b>
<b>3.4</b>	Analysis	<b>27</b>
<b>3.4.1</b>	Data Analysis using Hive	<b>27</b>
<b>3.4.2</b>	Data Analysis using Machine Learning	<b>28</b>
<b>3.5</b>	Algorithms	<b>29</b>
<b>3.5.1</b>	MapReduce Algorithm	<b>29</b>
<b>3.5.2</b>	Machine Learning Algorithms	<b>32</b>
<b>3.6</b>	Test Plan	<b>38</b>
<b>3.6.1</b>	Dataset	<b>38</b>
<b>3.6.2</b>	Split Dataset	<b>40</b>
<b>4</b>	<b>Results And Performance Analysis</b>	<b>42-49</b>
<b>4.1</b>	Using Hive	<b>42</b>
<b>4.1.1</b>	Analysis	<b>42</b>
<b>4.1.2</b>	Prediction	<b>44</b>
<b>4.1.3</b>	Query with and without optimization	<b>44</b>
<b>4.2</b>	Using Machine Learning	<b>47</b>
<b>7</b>	<b>Conclusion</b>	<b>50</b>
<b>8</b>	<b>References</b>	<b>51</b>

## List of Figures

<b>Figure No.</b>	<b>Figure Name</b>	<b>Page No.</b>
<b>1</b>	<b>5Vs of BIG DATA</b>	<b>1</b>
<b>2</b>	<b>Value in 5Vs of BIG DATA</b>	<b>2</b>
<b>3</b>	<b>Steps for Rating Analysis</b>	<b>5</b>
<b>4</b>	<b>Machine learning</b>	<b>6</b>
<b>5</b>	<b>Cross validation</b>	<b>10</b>
<b>6</b>	<b>Design of a Machine Learning Model</b>	<b>20</b>
<b>7</b>	<b>Data Models for converting data into tabular format</b>	<b>25</b>
<b>8</b>	<b>Process for uploading data and creating table using the HDFS</b>	<b>25</b>
<b>9</b>	<b>Data Modeling for handling structured, unstructured and semi structured data</b>	<b>26</b>
<b>10</b>	<b>Flowchart of data analysis using Hive</b>	<b>27</b>
<b>11</b>	<b>Flowchart of data analysis using Machine Learning</b>	<b>28</b>
<b>12</b>	<b>Steps for Mapper &amp; Reducer class</b>	<b>29</b>
<b>13</b>	<b>Combiner stage</b>	<b>30</b>
<b>14</b>	<b>Reducer stage</b>	<b>31</b>

## List of Figures (contd.)

<b>Figure No.</b>	<b>Figure Name</b>	<b>Page No.</b>
15	Sample record of the Amazon dataset	39
16	Sample record of the Amazon dataset(after pre-processing)	40
17	Training set and Test set	41
18	Display of Brand-wise Data	42
19	Display of Manufacturer-wise Data	43
20	Display of Rating-wise Data	43
21	Display of predicted rating-value	44
22	Display of query-run on Hive (without optimization)	45
23	Display of query-run on Hive (with optimization)	46
24	Comparison between query without optimization and query with optimization	46
25	Display of accuracy score and time taken of the algorithms	47
26	Display of accuracy score of the algorithms	49
27	Display of time taken of the algorithms	49

## List of Graphs

<b>Graph No.</b>	<b>Name</b>	<b>Page No.</b>
<b>1</b>	<b>Distribution of ratings for 1,239,581 electronics reviews on Amazon</b>	<b>3</b>
<b>2</b>	<b>Monthly Average Ratings of new Amazon electronics reviews</b>	<b>3</b>
<b>3</b>	<b>Distribution of helpfulness on 167,514 electronics reviews on Amazon</b>	<b>4</b>
<b>4</b>	<b>Distribution of average ratings for reviews written by 11,676 Amazon users</b>	<b>4</b>



## List of Tables

<b>Table No.</b>	<b>Name</b>	<b>Page No.</b>
<b>1</b>	<b>Dataset columns and its specifications</b>	<b>38</b>
<b>2</b>	<b>Comparison between query without optimization and query with optimization</b>	<b>46</b>
<b>3</b>	<b>Five times execution of algorithms</b>	<b>48</b>
<b>4</b>	<b>Final result of Accuracy and Time taken of algorithms</b>	<b>48</b>

## Abstract

Big data is the large amount of data that is generated every second. Such large data becomes very difficult to process using manual checking and old methods. Now, we have a number of tools and techniques to handle such data. With continually increasing customer-base on online platforms, it is necessarily required to know customer's likes and interests. It is really necessary as through this the preferences of the customers can be well known. The best way to handle this problem is by applying Big Data Analytics.

This is a project report on “**Amazon Rating Analysis and Prediction**”. During the exploration of this project, we develop new ideas and functionalities while using the Hive and Machine Learning technologies. This project is based on the Amazon rating details of different products from various categories purchased by a number of customers within a time-span.

The project report covers the implementation of the project on Hive and Machine Learning technologies and the reason why there is a shift from Hive to Machine Learning with a concluding result at the end.

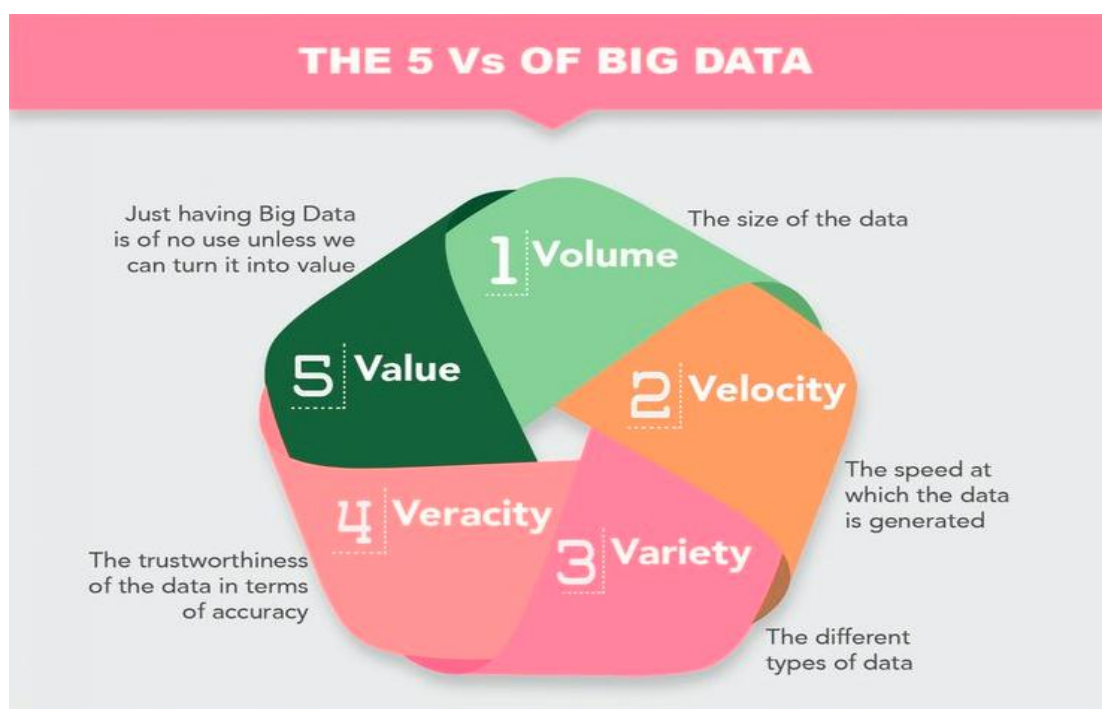
# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

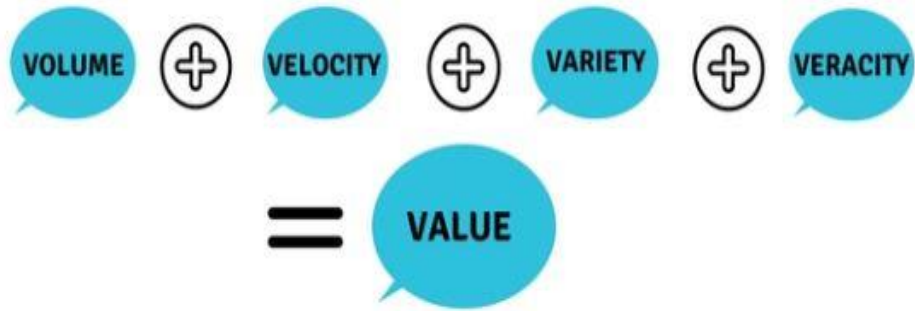
Lately, an extraordinary multiplication has been seen in the measure of information created. This rate is developing as information is consistently developing. Outfitting such tremendous measure of information is an intense activity and in this way it needs some out of the container thinking as it can't be handled with conventional apparatuses and techniques. This insufficiency prompted the introduction of the term Big Data and alongside it the difficulties, for example, storage, processing and privacy. Huge Data isn't just about being enormous in size. The definition is widened utilizing five attributes or "V's".

5 Vs of Big Data :



**Fig. 1 5Vs of BIG DATA**

- Volume: This trademark implies colossal voluminous information; it is in requests of terabytes and even petabytes.
- Velocity: This trademark means the high speed with which the information is created.
- Variety: This trademark alludes to the enormous assortment in the huge information.
- Veracity: This trademark alludes to vulnerabilities in huge information, for example, missing, copy and fragmented sections.



**Fig. 2 Value in 5Vs of BIG DATA**

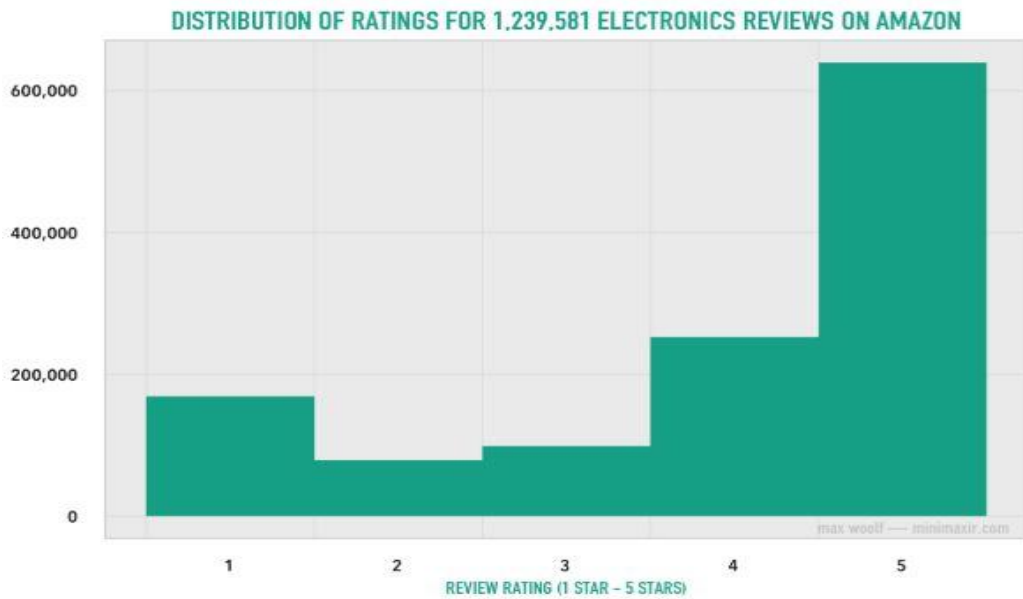
- Value: This trademark alludes to the inherent esteem contained in huge information.

Today, with a basic snap, it ends up less demanding than at any other time to influence examinations, to get more experiences and pick between comparable brand contributions dependent on online surveys. The advantages for composing surveys and give evaluations is it helps the merchant (if it's a great rating) and in the event that it is a legitimate audit the greatest advantage is to customers who will think about the item later on.

When purchasing the most recent items on Amazon, perusing reviews is an essential piece of the acquiring procedure. Client surveys/evaluations from clients who have really obtained and utilized the item being referred to can give you more setting to the item itself. Every commentator rate the item from 1 to 5 star-rating, and then gives a content outline of the encounters and feelings about the item. The evaluations for every item are arrived at the midpoint of together with the end goal to get a general item appraising.

Stanford analysts Jure Leskovec and Julian Mc Auley and gathered all the important Amazon related reviews from the administration's online presentation in 1995 to 2013. Dissecting the dataset of around 1.2 million Amazon surveys of items in the Electronic's segment, they discovered some intriguing factual patterns; some being instinctive and self-evident, however rest offer knowledge to how Amazon review framework really functions.

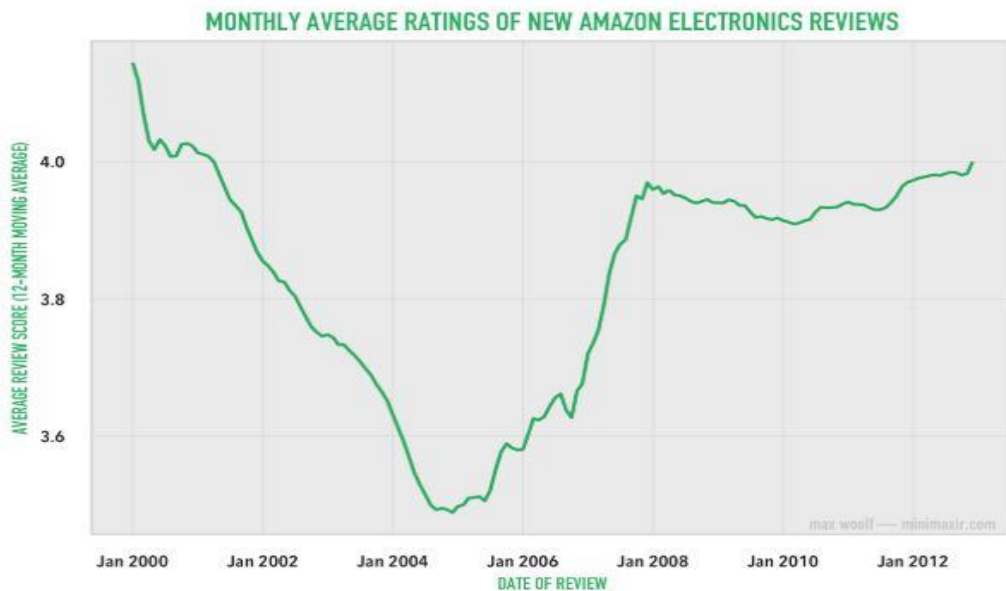
To begin with, how about we perceive how the ratings of the user are disseminated among the reviews.



**Graph. 1 Distribution of ratings for 1,239,581 electronics reviews on Amazon**

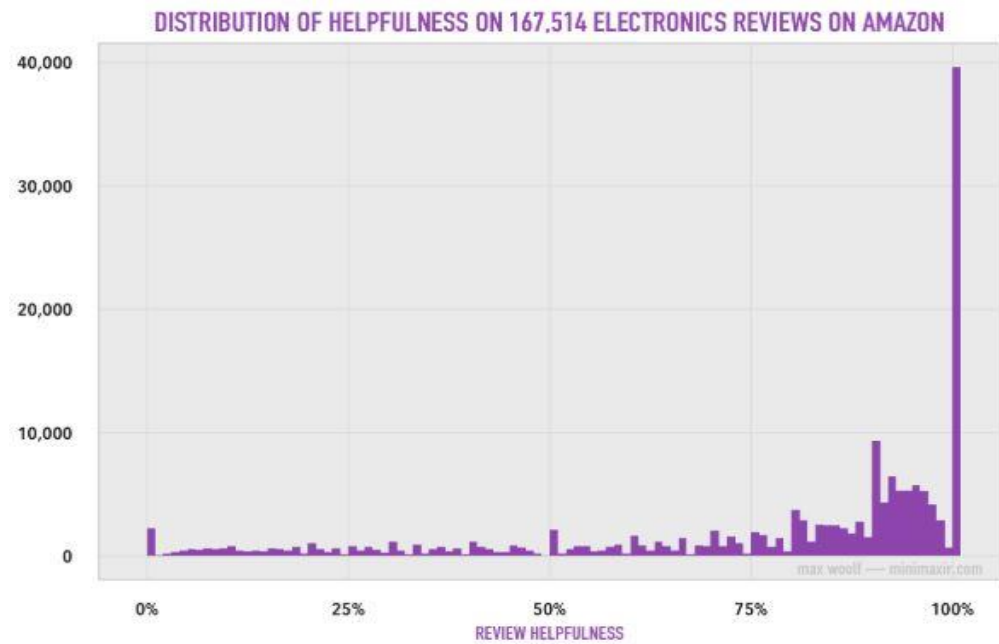
What we got was the 5 star rating in the greater portion of the reviews. Beside impeccable surveys, most users give 4 star or 1 star evaluations, with not very many giving 2 star or 3 star generally.

As resulted, the factual normal for all the review evaluations is on high part of the scale at about 3.90. Actually, normal review rating for recently composed ratings has different from 3.4 to 4.2 after some time-span.



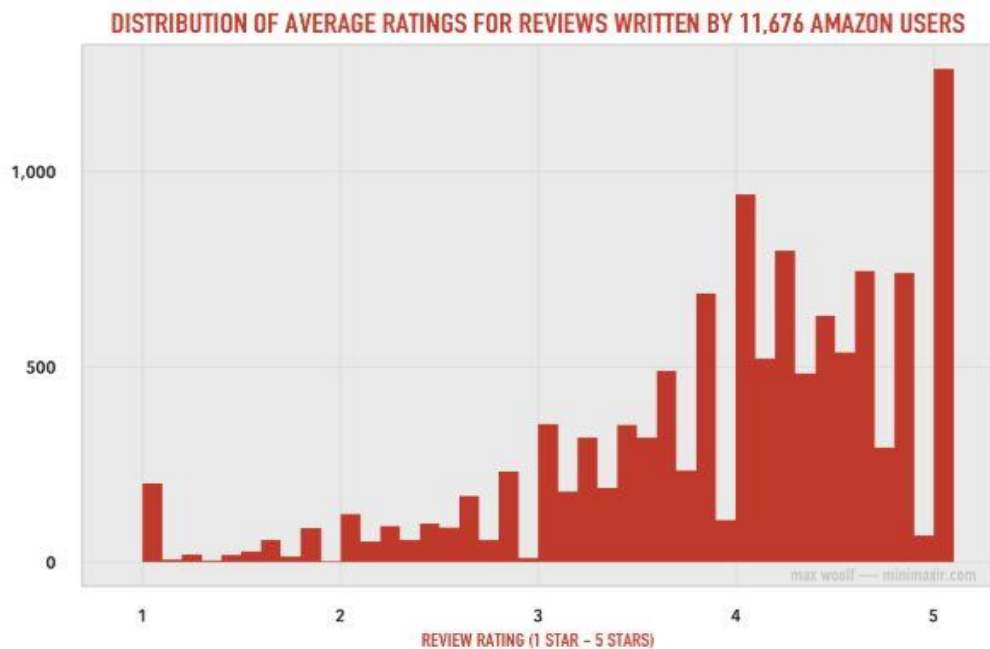
**Graph. 2 Monthly Average Ratings of new Amazon electronics reviews**

Another measurement being used to gauge review is rating supportiveness. Some other Amazon analysts rate a specific survey - "supportive" / "not supportive." This gives a sign of review quality to an imminent purchaser.



**Graph. 3 Distribution of helpfulness on 167,514 electronics reviews on Amazon**

That would bode well; in case you are composing any review (particularly a 5 star-rating), you are composing having the aim to encourage the other planned buyers.



**Graph. 4 Distribution of average ratings for reviews written by 11,676 Amazon users**

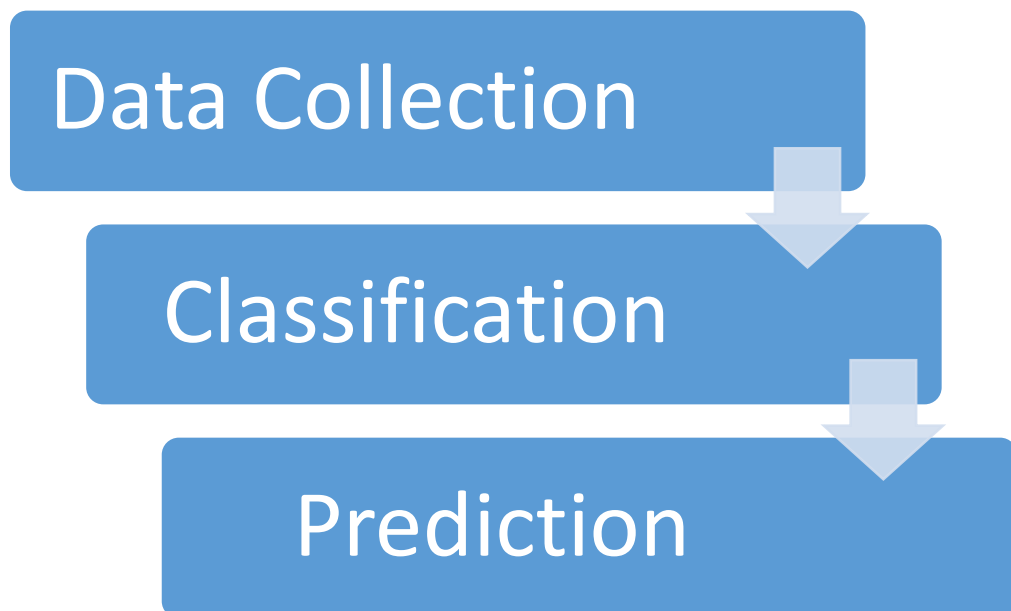
Dispersion of review evaluations when found the middle value of crosswise over is like alternate conveyances of survey appraisals. Be that as it may, this appropriation is less skewed toward 5 star and is more uniform b/w 4 star and 5 star.

Ratings done on Electronics items from Amazon often rate the product 4 or 5 star, and such reviews are found to be quite often viewed as supportive. 1-stars given are utilized to imply objection, and 2 star and 3 stars by any means have no noteworthy effect.

With a 5 star framework, one can enable the forthcoming user to make more educated examination b/w two items a: a user might be bound to buy an item that evaluates 4.2 star than an item that is appraised 3.8 star, which is a nuance that can't without much of a stretch be copied with a like/hate framework.

Rating analysis is done through the following steps :

- Data Collection
- Classification
- Pattern Identification
- Prediction
- Visualization



**Fig. 3 Steps for Rating Analysis**

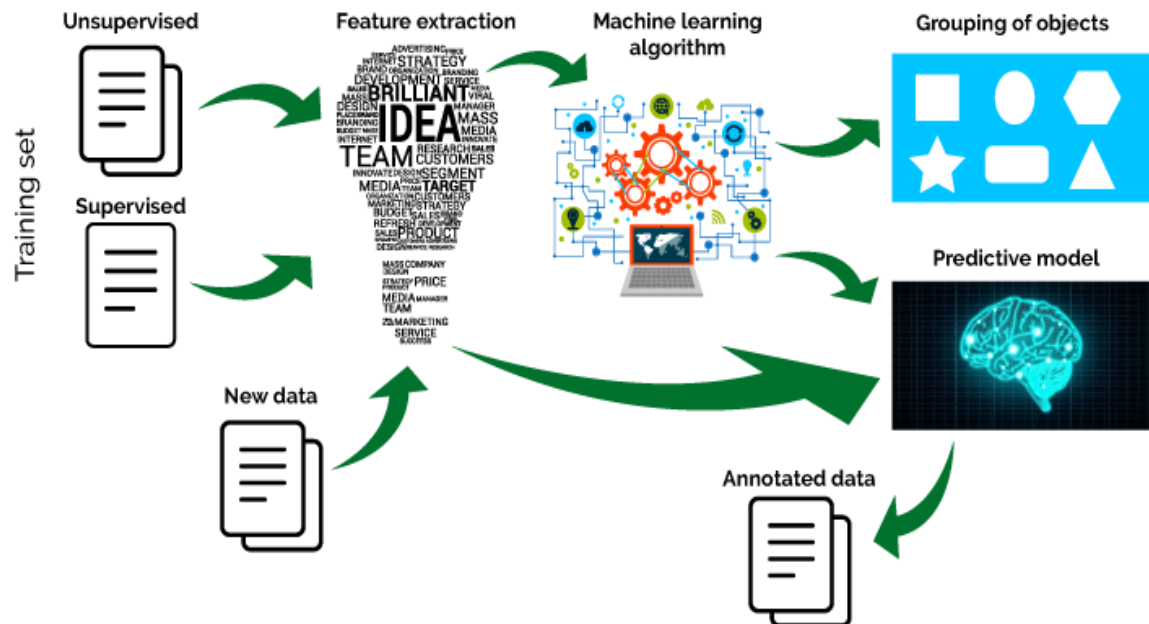
## Machine Learning:

Machine Learning being a computer science subdomain is utilized so as to investigate the information, which computerizes the development of systematic models. The motivation behind algorithms of machine learning is to gain from the current information "*without being expressly customized*".

A significant viewpoint with respect to machine learning is that, when the models are connected on new datasets these attributes originates from the iterative element of machine learning. These models are gaining from preceding computations for creating certain and replicable choices and results.

As the innovation propels, the old machine learning strategies don't fit well with the substantial large data. Machine learning being a field which is raised out of AI. Applying AI, building better intelligent machines is what we wanted. In any case, with the exception of couple of insignificant undertakings, for example, finding the most brief way between point An and B, we were unfit to program increasingly intricate and continually developing difficulties. There was an acknowledgment that to let machine learn from itself is the best way to have the capacity to accomplish this errand. This sounds like self-learning.

Thus, ML was created as another ability for PCs. Furthermore, presently ML is available in such huge numbers of sections of innovation, that we don't have any realization while utilizing it.



**Fig. 4 Machine learning**



## **1.2 Problem Statement**

Our main motive for the development of this project is regarding the two objectives.

Firstly, the given dataset is observed and according to the requirement, various attributes are taken and displayed together. The same dataset is run on two different platform i.e. Hadoop and Python. Moreover, the data can be displayed both in tabular format as well as in graphical form.

Secondly, we will forecast/predict what sort of product purchased by the customer is to be getting higher or lower rating.

Finally, we will examine the above-said problem through Machine Learning algorithms due to the limitations in the Hadoop-Hive prediction. The reason of the switch from Hive to machine learning is lack of limited subquery support. Also, hive cannot handle real time queries. The main Concern for using Hive is that is does not support OLTP(Online transaction processing).

## **1.3 Objective**

The objective is:

- To develop the project of Big Data Analytics using Hadoop on Hive
  - Display the different attributes of the dataset in tabular and graphical form
  - Predict the rating of a particular product.
- To analyze the limitations of hive and switching to machine learning to get better prediction results using some adequate machine learning algorithms and hence comparing Hadoop and python platforms.

## **1.4 Methodology**

This part covers the implementation of the described problem statement using Hive and Machine Learning.

### **1.4.1 Using Hive**

The implementation part is divided into 2 parts :

- **Analysis**

Analytics of Big-Data defines the procedure of dissecting expansive volume of information, or huge information. The huge information is accumulated from a vast assortment of sources, also including interpersonal organizations, recordings, advanced pictures, sensors, and exchange records.

The point to break down this information is to reveal eg. and associations that may for some way or another be imperceptible, and it may give significant point of knowledge about the users who made that. Through this type of understanding, organizations have the capacity to take over an edge over the adversaries and settle by unrivaled business choices.

In this, product\_id, name of product, category, review\_added, review\_recommend, review\_title, review\_text, and review\_username are analyzed individually & also collectively.

- **Prediction**

Predictive analysis is one of cutting edge procedure that utilizes information to gauge conduct, action and patterns. This includes applying the measurable investigation systems and explanatory questions to indexes of information to make such models that put a numerical esteem.

In this, review\_rating is predicted using regression technique for any given upcoming product\_id among the categories .

The project is implemented via the GUI which comprises :

- main JFrame form which gives options to select among the two, i.e. , Analysis & Prediction.
- In Analysis form , we can select from the different fields and then can analyze the respected data record in tabular form in an individual form.
- In Prediction form , we can enter a product\_id & then therefore see the predicted rating value for that particular product.
- Also , we can display the results from analysis & prediction form in the form of line graph for better visual comparison.

Steps for implementing the project :

- Selecting the dataset
- Data is stored in HDFS
- Database in created in Hive
- Tables are created in that database
- To load the data in the table created by above command we will have to use the following command:

**Load data local in 'path of the file' into table tablename;**

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

a) List the various products id:

```
select rating.id from data.rating
```

- b) List the various names of products:  
`select rating.name from data.rating`
- c) List the various brands:  
`select rating.brand from data.rating`
- d) List the various colors of the product:  
`select rating.colors from data.rating`
- e) List the various manufacturers:  
`select rating.manunfacturer from data.rating`
- f) List the various manufacturer\_number:  
`select rating.manufacturerNumber from data.rating`
- g) List the recommendation of the reviews:  
`select rating.doRecommendtitle from data.rating`
- h) List the rating value:  
`select rating.rating title from data.rating`
- i) List the username of reviewed user:  
`select rating.username from data.rating`
- j) Predict the rating of a product having product id 5:  
`select((sum(rating.id*rating.id)*sum(rating.rating)-  
sum(rating.id)*sum(rating.id*rating.rating))/  
(sum(rating.id*rating.id)*count(*)-  
sum(rating.id)*sum(rating.id)))+  
((sum(rating.id*rating.rating)*count(*)-  
sum(rating.id)*sum(rating.rating))/  
(sum(rating.id*rating.id)*count(*)-  
sum(rating.id)*sum(rating.id))) *5from data.rating;`
- k) Predict the rating of a product having product id 5:  
`Select a+5*b from reg;`

Optimization of the query is done as follows :

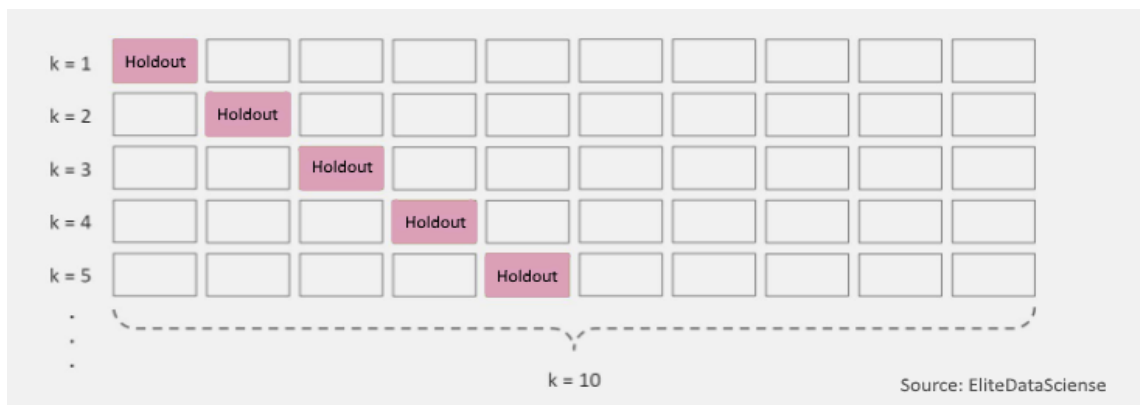
- A new table is created in which the variables of linear regression are stored after it is made to run using select query.
- Then the variables are directly fetched from the new table, as using Map-Reduce program we had already got the values of the variables.

## 1.4.2 Using Machine Learning

The big-data project of Amazon Rating Analysis & Prediction is implemented using Machine learning algorithms on Python.

- Initially, the dataset (amazon.csv) is taken from Kaggle, then according to the type of dataset and the described objective, various machine learning algorithms are selected.
- Then, data pre-processing takes place where data cleaning takes place to remove incomplete and useless data objects.
- Data is splitted into 2: training set and test set.
- Then model training takes place where the algorithms are trained using training data. Then algorithm will process the data and output a model which is able to find a target-value in the new data.
- Then model evaluation and testing takes place. The goal of such a step is developing a simplest model which is able to calculate a target-value well and fast enough.

One of the more efficient methods for model evaluation and tuning is cross-validation.



**Fig. 5 Cross validation**

### **Cross-validation:**

Cross-validation is the most commonly used tuning technique. It involves splitting of training-dataset into ten a balance of (folds). A given model is trained on just nine overlays and afterward tested on the tenth one (the one left out earlier). Training proceeds until each overlap is left aside and utilized for testing. Because of model execution measure, one figures out a cross-validated score for every set of hyper-parameters. A data researcher trains models with various sets of hyper-parameters to characterize which model has the most accuracy of prediction.

The cross-validated score demonstrates normal model execution crosswise over ten hold-out folds. At that point, one tests models with a lot of hyper-parameter values that got the best cross-validated score. There are different error-metrics for ML undertakings.

## **CHAPTER 2**

### **LITERATURE SURVEY**

#### **2.1 “Hive A Warehouse Solution Over A Mapreduce Framework” [1]**

In this paper, we present the Hive as open-source information warehousing arrangement based over Hadoop. Hive likewise incorporates a framework list, Hive-Metastore, containing constructions and insights, which is valuable in information investigation and question streamlining. Hive gives a SQL-like inquiry dialect called HiveQL which bolsters select, project, join, aggregate, union all and sub-inquiries in the from proviso.

In this paper how hive functions in the field of warehousing is appeared:

1. Usefulness — It shows that HiveQL gets developed by means of the StatusMeme application .
2. Tuning — It exhibit the inquiry plan watcher which demonstrates how HiveQL questions are converted into physical plans of guide lessen employments.
3. UI — It demonstrates the graphical UI which enables clients to investigate a Hive database, creator HiveQL inquiries, and screen inquiry execution.
4. Versatility — It delineates the adaptability of the framework by expanding the sizes of the info information and the multifaceted nature of the questions.

#### **2.2 “Pattern Finding In Log Data Using Hive on Hadoop” [2]**

In this paper the device utilized is Apache Hive to deal with the log documents. Log records are by and large documents which get routinely produced and kept up by a web server. For changing over the unstructured and complex log document into structure forbidden arrangement, we can utilize Regex SerDe properties into hive which can change the unstructured information into organized configuration.

In this we can likewise improve the hive inquiry execution, we can perform serialization process at the beginning table and store the resultant table into new table and after that apply all the question In this paper a test was directed in which the host or IP address which has most extreme recurrence or hit checks and the time taken by hive question takes 47.099 seconds to complete the execution.

At long last we saw that time taken by Hive is lesser than time taken by Pig in all viewpoints.

### **2.3 “Scalability Study Of Hadoop Mapreduce And Hive In Big Data Analytics” [3]**

First inquiry emerges that how would you move a current information framework to Hadoop, when that foundation depends on conventional social databases and the Structured Query Language (SQL)?

This is the place Hive comes in. Facebook created Hive which depends on natural ideas of tables, sections and segments, giving an abnormal state inquiry apparatus for getting to information from their current Hadoop distribution center.

In this paper this is a correlation between Hive versus Mapreduce :

Guide decrease programming model is low dimension and expects engineers to compose custom projects which are difficult to keep up and reuse though hive plays out the mapreduce work inside.

An examination was directed on word check in which hive played out the activity inside 35 seconds and customary guide lessen took around 1 min 10 seconds.

Along these lines this examination paper presumes that hive is unmistakably more superior to ordinary mapreduce and outperforms mapreduce execution.

### **2.4 “Commercial Product Analysis Using Hadoop MapReduce” [4]**

This paper discusses how an organization can discover genuine open doors in consolidating disconnected and online information to give astuteness on how combining disconnected and online information can be useful. Organizations utilizes proposal calculations which have the above advantages. Proposal calculations are best perceived for their utilization on online business Web sites. Here they utilize client's interests as a contribution to create an index of prescribed things.

These calculations are partitioned into 2 types :

The initial ones are called content based filtering. Content based filtering can likewise be called as cognitive filtering, which prescribes items based on an examination between the substance of the items and a client profile.

Also, the second one is collaborative filtering. It depends upon not simply the characteristics of the items but rather how individuals i.e different clients react to similar articles. Associations need to get every one of the information traits, disconnected and on the web, into a solitary database, which would be additionally refined by cutting edge examination strategies, and utilize the consolidated information for accuracy focusing on.

## **2.5 “Hive – A Petabyte Scale Data Warehouse using Hadoop” [5]**

The span of informational indexes being gathered and examined in the business for business insight is developing quickly, making customary warehousing arrangements restrictively costly. Hadoop is a prevalent open-source outline usage which is being utilized in organizations like Yahoo, Facebook and so forth to store and process to a great degree substantial informational indexes on ware equipment. Notwithstanding, the guide diminish programming model is low dimension and expects designers to compose custom projects which are difficult to keep up and reuse.

In this paper, Hive, an open-source information warehousing arrangement based over Hadoop is used. Hive bolsters questions communicated in a SQL-like decisive dialect - HiveQL, which are gathered into mapreduce occupations that are executed utilizing Hadoop. Also, HiveQL empowers clients to connect custom guide decrease contents into inquiries.

The dialect incorporates a sort framework with help for tables containing crude sorts, accumulations like clusters and maps, and settled sytheses of the equivalent. The basic IO libraries can be stretched out to question information in custom arrangements. Hive additionally incorporates a framework list - Metastore – that contains mappings and insights, which are helpful in information investigation, question streamlining and inquiry assemblage.

In Facebook, the Hive stockroom contains a huge number of tables and stores over 700TB of information and is being utilized widely for both announcing and specially appointed examinations by in excess of 200 clients for every month.

## **2.6 “Storage and Processing Speed For Knowledge from Enhanced Cloud Framework” [6]**

Cloud being the pool of servers, every one of the servers are inter connected through web, The principle issue in cloud is recovering of information (learning) and process that assortment of information and here other issue is security for that information, Generally now a day’s distinctive kinds of, means assortment of information (Structured, semi-organized and Unstructured information) is existed in the diverse social applications (confront book).

So, and another issue with verifiable information recovering. These kinds of issues are settled with help of hadoop outline work and Sqoop and flume devices. Sqoop is stack the information from database to Hadoop (HDFS), and flume stacks the information from server documents to hadoop appropriated record framework. Limit issue is settling with help of squares in hadoop dispersed record system and taking care of is settling with help of guide diminishing and pig and hive and begin, etc.

This paper here condenses the capacity and handling speed in the upgraded cloud with hadoop system.

## **2.7 “An Overview of the Hadoop/Mapreduce/Hbase framework and its current applications in bioinformatics” [7]**

Hadoop is basically a product structure that is to be introduced on a ware Linux group to allow substantial scale appropriated information examination. The Hadoop Distributed File System (HDFS) is a strong file system given by hadoop and in addition a Java-based API that permits parallel handling over the hubs of the group. Projects utilize a Guide/Reduce execution motor which works as a blame tolerant appropriated processing framework over extensive informational indexes - a strategy promoted by use at Google.

There are discrete Map and Reduce steps, where each of the progression are done in parallel, working one by one on sets of key-esteem sets. Preparing can be parallelized more than thousands of hubs chipping away at terabyte or bigger measured informational indexes. The Hadoop structure consequently plans outline near the information on which they will work, with "close" which means a similar hub or, at any rate, the same rack. Hub disappointments are additionally taken care of consequently. Notwithstanding Hadoop itself, which is a best dimension Apache venture, there are subprojects expand over Hadoop.

For example, Hive, an information stockroom system utilized for specially appointed questioning (with a SQL type inquiry dialect) and utilized for more unpredictable investigation; and Pig , an abnormal state information stream dialect and execution system whose compiler produces successions of Map/Reduce programs for execution inside Hadoop. Preparing of high throughput sequencing information (for instance, mapping to a great degree expansive quantities of short peruses onto a reference genome) is a territory where Hadoop-based programming is having an effect. bioinformatics work not exclusively is the versatility allowed by Hadoop and HBase essential, yet additionally of outcome is the simplicity of coordinating and breaking down different dissimilar information sources into one information distribution center under Hadoop, in moderately few HBase tables.

## **2.8 “Migration of Hadoop to Android platform using ‘Chroot’ ” [8]**

This paper is expected to enhance the current hadoop usage on low vitality expending android gadgets. Presently it gives ventures on the best way to utilize AndroHadoop stage :

1. Build an interface to introduce Linux.
2. Build or utilize existing VNC application to connect with GUI.
3. Install and arrange hadoop for running pseudo-dispersed mode.
4. Run example test program Hadoop we use for vast information handling.

What is chroot? Chroot'changes root condition.

There are 2 manners by which we can introduce chroot :



1. Manual: Using terminal emulator application and busybox from Google play store, and performing ordinary chroot on android
2. Automatic: Using an application called Linux send from Google play store Update Linux movement.

## **2.9 “Review Paper on Hadoop And Map Reduce” [9]**

Enormous Data is an information whose scale, better than average assortment, and multifaceted nature require new designing, strategies, figurings, and examination to direct it and focus regard and disguised gaining from it.

Hadoop being the center-stage for dealing with the Big Data, and manages the main issue of making it noteworthy for the examination purposes. Hadoop being an open-source programming adventure which engages appropriated treatment of immense informational collection across over-packs of thing servers. It is proposed to scale up from a solitary server to a great many machines, with an abnormal state of adjustment to interior disappointment.

This paper altogether discusses why hadoop is better in all angles. Following focuses center around the benefits of hadoop:

Adaptability , exceeds expectations at handling information of complex nature and its open-source nature make it much well known. In this paper the design of hadoop and mapreduce is clarified. MapReduce has been shown as a free stage as the advantage layer proper for different need by cloud providers.

It moreover enables customers to appreciate the information handling and exploring.

## **2.10 “Machine Learning Algorithms: A Review” [10]**

In this exploration paper, different machine learning calculations have been talked about. This paper gives nitty gritty clarification of the classes of the calculations.

In the directed learning classification, three calculations have been examined i.e. naïve bayes, support vector machine and decision tree.

In the unsupervised learning, two algorithms have been talked about i.e. K- means clustering and principal component analysis.

### **2.11 “A Survey on Machine Learning: Concept, Algorithms and Applications ” [11]**

In this examination paper, the paper begins with the utilization of machine learning in measurements , human learning, and artificial intelligence . This exploration gives a point by point clarification about the present research situation.

One such precedent given in the exploration paper is the means by which administered calculations can be utilized for unlabelled information. Besides, another example is to interface diverse kinds of machine learning calculations and so on. From that point different machine calculations have been talked about.

In conclusion the models naïve bayes, support vector machine and decision tree has been thought about regarding exactness .forecast time and preparing time. As per the paper, naïve bayes turns out to be the best.

Additionally it has likewise been expressed that ML calculations perform better as indicated by the dataset.

### **2.12. “Python – The Fastest Growing Programming Language” [12]**

This paper began with presentation of python as an high state programming.

Why python is developing so quickly is talked about next due to its characteristics like movability, easy to learn, open source and so on. Further its downsides are talked about like its moderate nature and furthermore it is difficult to keep up.

Numerous projects have been recorded in python and it has been utilized in Irobot, Google, Intel and so on.

### **2.13. “Machine Learning for Computer Security” [13]**

As traditional strategies for removing dangers requires a great deal of human exertion, this prompted the improvement of machine learning calculations to distinguish dangers and attacks.DMC and PPM models have been made for email messages.

DMC takes in a Markov demonstrate steadily by means of a cloning procedure to present new states in the model. PPM takes in a table of settings and the recurrence of the image following the specific circumstances.

To discover spam in the email ,MCE and MDL is utilized. MCE computes the quantity of bits to encode a message. MDL measures the extra number of bits expected to encode a message subsequent to adding it to the contending models.

### **2.14 “Sentiment Analysis for Amazon Reviews” [14]**

An expanding measure of research endeavors extended in understanding opinion in printed resources. In this paper, creator has utilized the LSTM with 128 concealed units and afterward utilized a dense net with soft max being the activation function to anticipate these 5 classes. The information has been prepared for 20 epoch sin tests utilizing LSTM. Adam enhancer has been utilized to enhance the parameters the learning rate is 0.01 and the batch-size being 32. To forestall over fitting, a dropout rate of 0.2 was set in the LSTM layer.

Be that as it may, the test accuracy comes out only 65.6 %, which implies this model has over fitted on the resampled information, since there are many rehashed precedents.

### **2.15 “Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning” [15]**

Amazon.com has developed quickly and been a microcosm for client provided surveys. Before long, Amazon open alters audits to purchasers, and in the end enabled any client to post a survey for any of the a huge number of items on the site. With this expansion in unknown client produced content, endeavors must be made to comprehend the data in the right setting, and create techniques to decide the aim of the creator. The implementation rely upon many sub-portions of their executions, yet the exhibitions of both the Decision List and Naïve Bayes additionally rely upon the numbered features being considered.

Moreover, the number of principles that are connected in the Decision List before a tag is made will influence its execution. These parameters do likewise rely upon the range of the test and training sets, however these tests are just intended to give a gauge with the goal that the frameworks can be tried better in different ways.

Naive Bayes had the most noteworthy exactness when it was utilized with 800 highlights.

### **2.16 “Sentimental analysis of Amazon reviews using naïve bayes on laptop products with MongoDB and R” [16]**

The essential methodology should be possible securing the information in this manner as time was being propelled along these lines the through the headway. The innovation was likewise being at fast change from years to years.

In this manner the client are being from past century are being becoming accustomed to pay high add up to the merchant in light of extra expense.

The conclusion is a characteristic procedure of passing on the type of assessment by the client for the specific item that is accessible in the ecommerce site as we are worry about the assumption approach of the electronic devices that is PC which is being considered to one piece of real work areas in the Indian w-trade site there are a few suggestion and methodologies proposed by the client with respect to the services and the nature of the workstation usefulness and it include and the sort of nature of item to client get the services from the web based business industry through the web based business site.

Naive Bayes characterization has autonomy among its highlights while Bayesian systems can be said that it has dependency for all highlights. It very well may be utilized as non-cyclic chart and highlights as nodes and has different connections between them. Exactness for the probabilistic model is around 73%.

### **2.17 “Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining” [17]**

The auto identifying of helpful-reviews isn't as simple as it might appear, on the grounds that the survey content must be semantically examined. A starter trial contemplate led on a vast freely dataset (117,000 Amazon surveys) confirms this could be extremely possible and a productive research bearing. The objective in executing the analysis is to foresee whether are see is viewed as helpful, by considering either its textual substance just, or, moreover, additionally the argumentation mining information originating from MARGOT.

At the end of the day, the creator has worked in binary classification situation. In this paper, we have seen a first exploratory investigation that intends to demonstrate how includes originating from an off-the-rack argumentation mining framework can help in anticipating whether a given survey is valuable.

## **CHAPTER 3**

### **SYSTEM DEVELOPMENT**

#### **3.1 Designing**

This part covers the designing of the Hive and Machine Learning models.

##### **3.1.1 Apache Hive Data Models**

Hive, which is an open source information stockroom and based on the top of Hadoop can break down and store even substantial datasets, put away in Hadoop documents. Apache Hive can store information as tables.

- **Tables**

Apache Hive tables are like social database tables. Tables of Hive comprise of information and are their design is portrayed with the assistance of related metadata. Channel ,join and association activities can be performed on these tables.

Typically, in Hadoop, the information are put away in HDFS however Hive stores metadata in a social database as opposed to HDFS.

Table types containing in Hive :

- **Managed or Internal Tables**
  - Managed Tables of Hive are additionally called internal tables and are the default tables.
  - All managed tables are made or put away in HDFS and the information of the tables are made or put away in the/client/hive/distribution center catalog of HDFS. On the off chance that one erases the table, both the table information and metadata are erased from HDFS.

- 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 Int ,text String , title String , username String) row format delimited fields terminated by ‘;’;

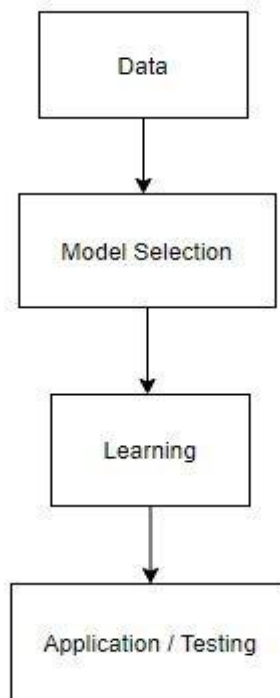
- By the use of above command, managed table is created.

Data is loaded by :

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

**Eg;** load data local inpath ‘amazon.csv’ into table rating;

### 3.1.2 Machine Learning Model



**Fig. 6 Design of a Machine Learning Model**

Machine Learning model consists of various stages :

1. **Data**
2. **Model Selection**
3. **Learning**
4. **Application / Testing**

These are explained as following :

### 1. Data

Data may need a lot of:

- **Cleaning** : It involves getting rid of errors and noise from the data. It also deals in removal of redundancies in the data.
- **Preprocessing** : In this relabeling is done which is converting categorical values to numbers. And, moreover rescaling is done where continuous values are transformed to some range[1,-1].

### 2. Model Selection

- This involves the selection of best model for the desired result
- Selection will be much easier if there is a prior knowledge in terms of which algorithm fits best according to the problem. And, also knowing initial data analysis and visualization.
- We can make a good guess about the form of the distribution

### 3. Learning

- Learning is in equivalence to optimizing problem
- Optimization problem may be sometimes hard to makeout, but correct selection of a model gives us the the desired output with better accuracy.

### 4. Application / Testing

In this, predictive performance of the learning model is evaluated and hence compared according to different parameters.

## 3.2 Experimental setup

This part lists the hardware and software used while implementing the given problem statement using Hive and Machine Learning.

### 3.2.1 For Hive

For Single Node , hardware requirements for Data Analytics System are described by the following details :

We have developed the whole system on the following hardware :

<b>Type of hardware</b>	<b>Hardware used</b>
Hardware 4 GB RAM	- Quad core Intel i3 compatible processor with  - multiprocessor-based computer with a 2.00 Ghz processor - 64-bit operating system ( minimum 256 MB of RAM required)
Disk space	5 GB free disk space (minimum). Requirements increase as data is gathered and stored in HDFS.
Memory	4 GB (2.4 GB on virtual machine)

The following details describes the basic software requirements we have used :

<b>Type of software</b>	<b>Software versions</b>
Operating System	Linux (Ubuntu 12.04 LTS)
VMware workstation	14.1.2
Hadoop	2.7.3
Hive	1.2.0
Netbeans IDE	7.0.1
Web browser	Mozilla Firefox 28.0
Text Editor	gedit



**Benefits of VMware Workstation include :**

- Cloning ability of virtual machines.
- Software development under various working frameworks(OS).
- Turning of physical PC into virtual machine.
- Import the machine and modify virtual machines.
- Test programming under different working frameworks.

**Hive 1.2.0 is being used as :**

- this version is being supported only on **Hadoop** 2 clusters.

**3.2.2 For Machine Learning**

Following details describes the hardware requirements for the Data Analytics System.

We have developed the whole system on the following hardware :

<b>Type of hardware</b>	<b>Hardware used</b>
Hardware 4 GB RAM	- Quad core Intel i3 compatible processor with  - multiprocessor-based computer with a 2.00 Ghz processor - 64-bit operating system
Memory	- 4 GB

The following details describes the basic software requirements we have used :

<b>Type of software</b>	<b>Software versions</b>
Operating System	Windows 10
Python	3.7
Anaconda	5.3.1
Jupyter notebook	5.7.4
Web browser	Chrome
Text editor	Notepad++

### **3.3 Analytical Model**

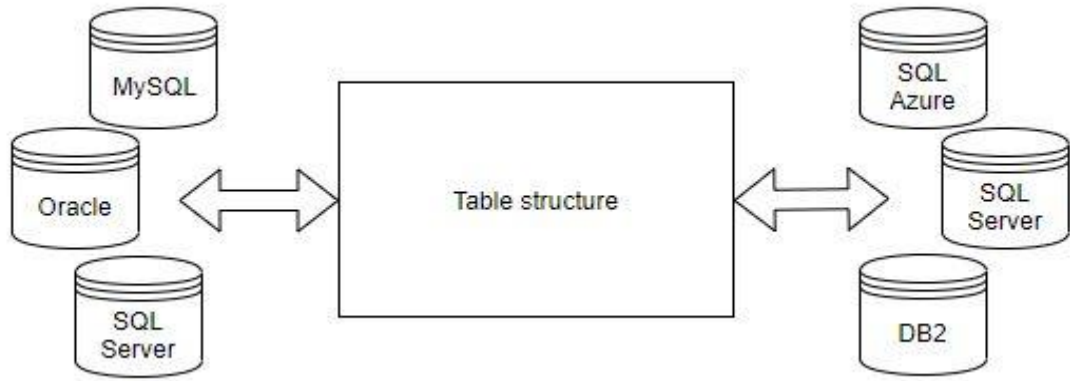
This part describes the analytical models in the field of Big Data Analytics.

#### **3.3.1 Traditional Model – “Schema on Write”**

- **In traditional**

In customary databases, the table's pattern is forced amid the information stack time, on the off chance that the information being stacked does not adjust to the construction, the information stack is rejected, this procedure is know as Schema-on-Write. Blueprint on-Write helps in quicker execution of the inquiry, as the information is as of now stacked in a specific arrangement and it is anything but difficult to find the section file or pack the information.

The fundamental points of interest of mapping on compose are exactness and question speed. In customary way(RDBMS).So assume we make a Schema comprising of 10 Columns and we attempt to stack information which could fulfill just 9 segments that information would be rejected as a result of Schema on compose, here the information is perused against blueprint before it is kept in touch with the database.



**Fig. 7 Data Models for converting data into tabular format**

### 3.3.2 Big Data Model – “Schema on Read”

In HIVE, the information composition isn't checked amid the heap time, rather it is confirmed while handling the question. Henceforth this procedure in HIVE called Schema-on-Read. Construction on-Read helps in quick starting information stack, since the information does not need to pursue any inward schema(internal database design) to peruse or parse or serialize, as it is only a duplicate/move of a record.Organized is connected to the information just when it's perused, this enables unstructured information to be put away in the database. Since it's not important to characterize the pattern before putting away the information it makes it less demanding to get new information sources.

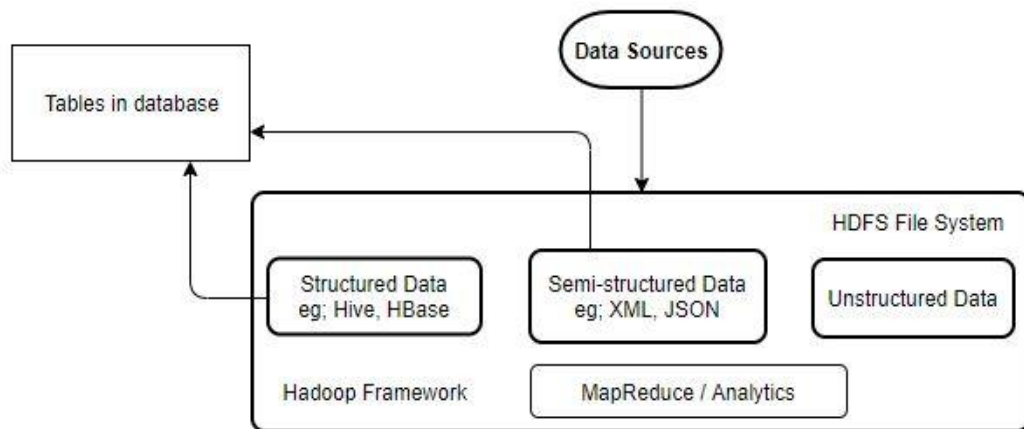
With the Big Data and NoSQL worldview , "Schema-on-Read" implies we don't have to know -how we will utilize our information when we are putting away it. We do need to know how we will utilize our information when we are utilizing it and model in like manner.

Model: We may initially put the information on HDFS in records , then apply a table structure in Hive.

<b>HDFS</b>	<b>File System</b>
<b>Exploration</b>	<b>File System</b> Analyze and understand the data.
<b>Hive</b>	<b>File System</b> HDFS:

**Fig. 8 Process for uploading data and creating table using the HDFS**

### 3.3.3 Data Modeling in the Big Data Ecosystem



**Fig. 9 Data Modeling for handling structured, unstructured and semi structured data**

In the huge information biological system , there are three kinds of data i.e. structured ,unstructured and semi structured . There are a few manners by which we could execute these kinds of information.

For example , let us consider the instance of organized information for that we have to utilize HIVE with HQL.

For unstructured data we have to initially stack the record into HDFS document framework and after that convert it into an unthinkable organization utilizing individual mapreduce strategies.

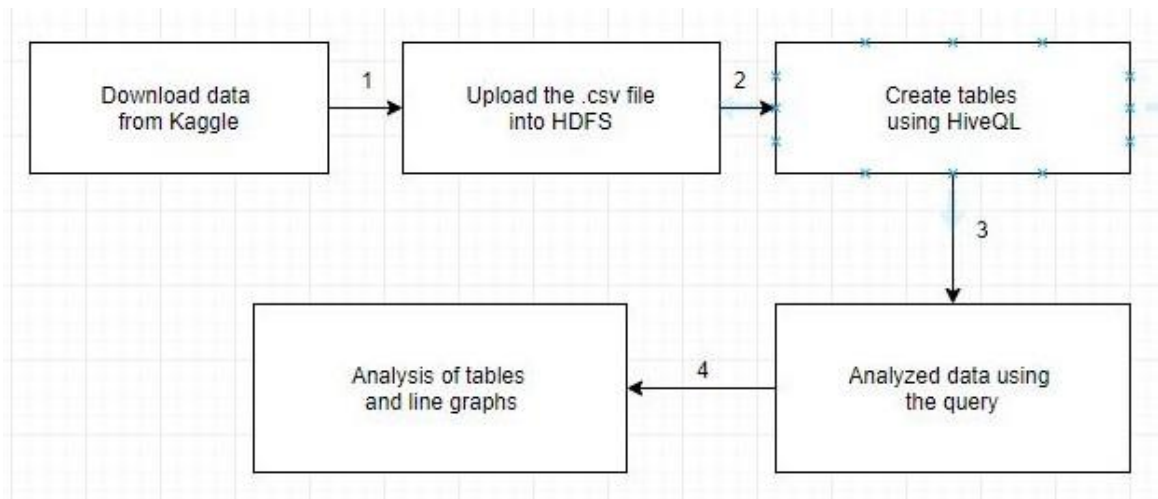
For semi organized data we have to utilize JSON/XML records to change over it into an unthinkable configuration.

On the whole, every one of these kinds of information chip away at HADOOP system.

### 3.4 Analysis

This section describes the data analysis according to the problem statement using Hive and Machine Learning.

#### 3.4.1 Data analysis using Hive

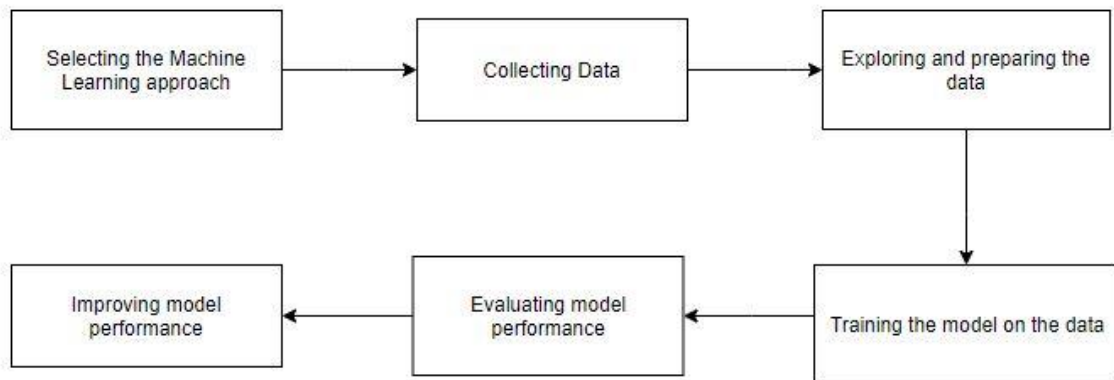


**Fig. 10** Flowchart of data analysis using Hive

In the Data Analysis, following processing takes place:

1. The downloaded .csv file (amazon.csv) is added into the HDFS using:  
**load data local inpath 'amazon.csv' into table rating;**
2. After the file is added to the HDFS, database (name - amazon) is created.
3. Queries are written to analyze the data.
4. Data is displayed using tables and line graphs.

### 3.4.2 Data analysis using Machine Learning



**Fig. 11 Flowchart of data analysis using Machine Learning**

The steps involved in data analysis using Machine Learning:

- 1. Selecting the Machine Learning approach:** Before beginning any means, the ML issue should be communicated. What would we like to discover? Would we like to order our information, to anticipate new values, to bunch our data dependent on certain criteria? After we choose what kind of ML task we might want to perform, we select our model.
- 2. Collecting data:** Data can be composed on paper, recorded content documents and spreadsheets or put away in a SQL database. Information should be assembled in an electronic organization reasonable for analysis.
- 3. Exploring and preparing the data:** The nature of any ML project depends on the nature of the data it employs. It is proposed that 80-% of the exertion in ML is dedicated to preparation of data. This progression requires a lot of human intercession.
- 4. Training the model on the data:** The particular ML undertaking will illuminate the choice regarding a suitable algorithm. We then "feed" the data into the model amid this stage and then we will get a learner. A learner being a ML algo that has been prepared on certain data and changed in accordance with fit the data as most ideal as.
- 5. Evaluating model performance:** Because every learner results in a one-sided arrangement, it is essential to assess how well the algorithm gained from its experience. Contingent upon the model utilized, we may most likely assess the precision of the learner utilizing a test dataset.
- 6. Improving model performance:** If better execution is required, it ends up important to use further developed strategies to improve the execution of the model, or change to an alternate model, supplement with extra data and play out extra preparation-work on the data (stage 3).

### 3.5 Algorithms

This section describes the algorithms used in the context of the problem statement of the project.

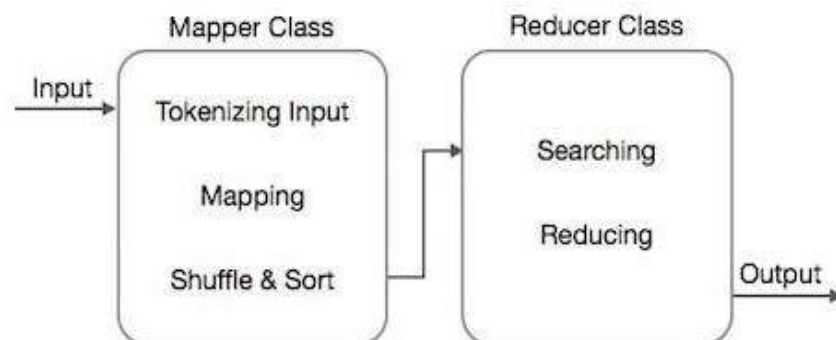
#### 3.5.1 MapReduce Algorithm

The MapReduce calculation contains two imperative undertakings, in particular Map and Reduce.

The mapping is finished by methods for Mapper Class

The reducing is finished by methods for Reducer Class.

Mapper class takes the information, tokenizes it, maps and sorts it. The yield of Mapper class is utilized as contribution by Reducer class, which thus seeks coordinating sets and lessens them.



**Fig. 12 Steps for Mapper & Reducer class**

MapReduce actualizes different numerical calculations to separate an errand into little parts and appoint them to various frameworks. In specialized terms, MapReduce calculation helps in sending the Map and Reduce undertakings to suitable servers in a bunch.

These above-said algorithms may incorporate the accompanying :

- Sorting
- Searching

- **Sorting**

Sorting is one of the fundamental MapReduce calculations to process and break down information. MapReduce executes arranging calculation to naturally sort the yield key-value sets from the mapper by their keys.

- Sorting strategies are actualized in the mapper class itself.
- The arrangement of middle key-value sets for a given Reducer is naturally arranged by Hadoop to frame key-values (K2, {V2, V2, ...}) before they are exhibited to the Reducer.

- **Searching**

Searching plays an imperative job in MapReduce calculation. It helps in the combiner stage (discretionary) and in the Reducer stage.

- The Map stage forms each information document and gives the information in key-value sets (<k, v> : <id, rating>).
- The combiner stage will acknowledge the contribution from the Map stage as a key-value match with item id and rating. Utilizing searching method, the combiner will check all the item id's to locate a specific rating an incentive in each record.

```
<k: id, v: rating>
R= rating of a particular product-id

if(v(product).rating == R){
    R = v(rating);
}
else{
    Continue checking;
}
```

**Fig. 13 Combiner stage**



- Reducer stage – From each document, we will discover the item id's of that specific rating. To maintain a strategic distance from repetition, all the  $\langle k, v \rangle$  sets are checked and copy sections are disposed of, assuming any.

$\langle 10, 5 \rangle$

**Fig. 14 Reducer stage**

### 3.5.2 Machine Learning Algorithms

There are three kinds of machine learning algorithms :

#### 1) **SUPERVISED MACHINE LEARNING** (eg; Linear regression & Random forest):

Most of the practical ML utilizes supervised learning.

Supervised learning is the place we have input factors (x) and a yield variable (Y) and we utilize a algorithm to take in the mapping capacity from the i/p to the o/p.

$$Y = f(X)$$

The objective is to estimated the mapping function so well that when we have new i/p data (x) that we can foresee the o/p factors (Y) for that data.

It is called supervised learning in light of the fact that the procedure of a algo learning from the training-dataset can be thought of as an educator regulating the learning procedure. We know the right answers, the algo iteratively have predictions on the training-data and after that is rectified further. Learning stops when the algorithm accomplishes a worthy dimension of execution.

Supervised learning issues can be additionally gathered into regression and classification problems:

- a) **Classification:** It is the point at which the o/p variable is a class, for example, "red" or "blue" or "illness" and "no ailment".
- b) **Regression:** It is the point at which the o/p variable is a genuine value, for example, "dollars" or "weight".

#### 2) **UNSUPERVISED MACHINE LEARNING** (eg; k-means):

Unsupervised learning is the place where we just have input information (X) and no relating o/p factors.

The objective for unsupervised learning is to display the fundamental structure or appropriation in the data so as to study the data.

These are called unsupervised learning in light of the fact that dissimilar to supervised learning above there is no right answers and there is no educator. Algorithms are left to its own devises to find and present the intriguing structure with regards to the data.

Unsupervised learning issues can be additionally gathered into clustering and association problems.

- a) **Clustering:** It is the place we need to find the inherent groupings in the information, for example, gathering clients by buying behavior.
- b) **Association:** It is the place we need to find rules that depict huge parts of your data, for example, individuals that purchase X additionally will in general purchase Y.

### 3) SEMI-SUPERVISED MACHINE LEARNING :

Issues where we have a lot of i/p data (X) and just a portion of the data is marked (Y) are called semi-supervised learning problems.

These issues sit in the middle of both supervised and unsupervised learning.

A genuine example is a photograph file where just a portion of the pictures are named, (for example person, cat, dog) and the greater part are unlabeled.

Numerous genuine ML issues fall into this region. This is on the grounds that it very well may be costly or tedious to label information as it might expect access to domain specialists. Though unlabeled data being cheap and simple to gather and store.

We can utilize unsupervised learning methods to find and become familiar with the structure in the i/p factors. We can likewise utilize supervised learning procedures to make best prediction expectations for the unlabeled information, feed that information again into the supervised learning algo as training data and utilize the model to make forecasts on new concealed data.

Machine learning algorithms used in context of the problem statement :

- 1) Linear Regression
- 2) Logistic Regression
- 3) KNN
- 4) Decision Tree

5) SVC

6) Random forest

7) Naïve Bayes

- The above algorithms are explained as following:

### **1) Linear Regression:**

In this calculation we attempt to discover a connection between an autonomous and a needy variable utilizing best fit line. The best fit line is spoken to by  $Y = a * X + b$  where

Y – represents the “Dependent Variable”

a – gives the Slope

X – indicates the Independent variable

b – is the Intercept

These coefficients a and b are determined dependent on limiting the aggregate of squared contrast of separation between information focuses and relapse line.

Linear Regression is of two kinds:

- Simple linear regression(one free factor)
- Multiple Linear Regression (in the excess of one autonomous variable).

## 2) Logistic Regression:

In **Logistic Regression**, we wish to display a dependent variable(Y) in means of at least one independent variables(X). It is a technique for classification. This calculation is utilized for the dependent variable that is Categorical. Y is demonstrated utilizing a function that gives o/p somewhere in the range of 0 and 1 for all estimations of X. In Logistic Regression, the Sigmoid Function is utilized.

## 3) KNN:

**K-Nearest Neighbors (KNN)** is one of the most straightforward calculations utilized in Machine Learning for classification and regression problems. KNN calculations utilize a data and group new data focuses dependent on a likeness measures (for example distance func). Classification is finished by a greater part vote to its neighbors. The information is relegated to the class which has the most closest neighbors.

## 4) Decision Tree:

**Decision Trees** are a kind of Supervised Machine Learning (that is we clarify what the i/p is and what the comparing o/p is in the training data) where the information is ceaselessly split as per a specific parameter. The tree can be clarified by two substances, in particular decision nodes and leaves. The leaves are the decisions or the ultimate results. Furthermore, the decision nodes are the place the information is split.

There are two principle kinds of Decision Trees:

- Classification trees (Yes/No sorts)
- Regression trees (Continuous information types)

## 5) SVC:

“**Support Vector Machine**” (SVM) is a directed ML algo which can be utilized for both classification and regression difficulties. Nonetheless, it is for the most part utilized in classification problems. In this algo, we plot every data thing as a point in n-dimensional space (where n is number of features we are having) with the estimation of each feature being the estimation of a specific coordinate. At that point, we perform characterization by finding the hyper-plane that separate the two classes great.

## 6) Random Forest:

**Random forest** is a kind of managed ML algo dependent on ensemble learning. Ensemble learning being a kind of learning where we join various sorts of algorithms or same algo on numerous times to frame an all the more dominant prediction model. The random forest algo consolidates numerous algorithms of a similar kind for example different decision trees, bringing about a forest of trees, thus the name "Random Forest". The random forest algo can be utilized for both classification and regression tasks.

## 7) Naïve Bayes:

It is a classification method dependent on Bayes' Theorem with a presumption of independence among indicators. In straightforward terms, a Naive Bayes classifier accept that the existence of a specific component in a class is inconsequential to the existence of some other element.

Naive Bayes model is easier to construct and especially valuable for exceptionally substantial datasets. Alongside simplicity, Naive Bayes is known to beat even very complex classification methods.

The major Naive Bayes supposition which is that each component makes an:

- independent
- equal

commitment to the result.

With connection to our dataset, this idea can be comprehended as:

- We expect that no pair of features being dependent. For instance, the temperature being 'Hot' has nothing to do with the stickiness or the viewpoint being 'Stormy' has no impact on the breezes. Thus, the featured are thought to be **independent**.
- Secondly, each feature is given the equivalent weight(or significance). For instance, knowing just temperature and dampness alone can't anticipate the result accurately. None of the properties is insignificant and thought to contribute **equally** to the result.

In **Gaussian Naive Bayes**, continuous values related with each feature are thought to be conveyed by a Gaussian distribution. It is additionally called Normal distribution.

- In ML, there is something many refer to as the "No free Lunch" hypothesis. More or less, it expresses that nobody algorithm works best for each problem and it's particularly significant for supervised learning. Thus, we should attempt various algorithms for our concern to assess execution and afterward select the algorithm as it needs to be.

According to the results and the parameters considered, we found Naïve Bayes algorithm to be much better than other algorithms in terms of accuracy (to find out the desired result), and in terms of time taken by algorithms to execute.

**The following are the reasons for choosing Naïve Bayes over other algorithms:**

**Naïve bayes** is the best algorithm for prediction in case we have moderate or large training dataset. It performs very well when the i/p values are categorical. It requires less training data than the other algorithms.

Thinking about the instance of **logistic regression**, it utilizes a logistic function to outline just binary o/p model. It can't be utilized for classification problems that are non-linear. Co-linearity and exceptions alters the exactness of the LR model. It is additionally not effective for the categorical values.

If there should arise an occurrence of **K nearest neighbor**, k ought to be carefully chosen as the prediction relies upon the estimation of k. Slight variety in k results in a lot of error. It is a moderate algorithm contrasted with other ML algorithms. Likewise, the computational expense amid runtime is high if there should be an occurrence of KNN. Remembering, the project to be financially savvy and achievable we have disposed of the utilization of KNN. Moreover, KNN is non parametric algorithm.

If there should be an occurrence of **decision tree**, as it gets confounded while training datasets. It is extremely inclined to anomalies. Odds of blunder are extremely high on the off chance if we continue constructing the tree to accomplish higher purity. It is likewise serves non parametric method. It is a discriminative model and prompts lesser accuracy.

If there should arise an occurrence of **random forest algorithm**, it is vey complex to utilize which builds the odds of blunder in our execution. It isn't feasible for small to moderate datasets and requires large dataset for calculation. In our task we have utilized a moderate dataset, so we have disposed of our decision to utilize random forest algorithm. Moreover, using this algorithm for prediction is a very tedious to build.

In the event of **support vector machine**, picking a good kernel method being never simple. It requires longer investment of time for datasets. It is difficult to picture the effect of the hyper parameters and difficult to tune them which makes it an unpredictable and complex algorithm. Having a go at picking the most simple and effective algorithm for our concern, that is the reason we have disposed of the decision to utilize SVC.

### 3.6 Test Plan

This section describes the test plan for the handling of the data.

#### 3.6.1 Dataset

A dataset is an accumulation of information. Most normally an informational collection relates to the substance of a solitary database table, where each segment of the table addresses an explicit variable, and each line looks at to a given individual from the educational file being alluded to.

The dataset taken is of Amazon rating and is taken from Kaggle (source – [www.kaggle.com](http://www.kaggle.com) ), which is the world's biggest network for information researchers to investigate ,break down and share information.

The dataset contains 9 columns and is about the review ratings of various products of different categories.

In Table 1, the field id stands for the id of the product which is a unique number to describe a product. The field brands stand for various brands which are available in the market. Color field defines the color of the product. Manufacturer field represent the manufacturer's name for a given product. ManufacturerNumber represents a unique number of the manufacturer. Name stands for the name of the product. Reviews.doRecommend stands whether the reviewed user recommend the product to various customers or not. Review.rating stands for the rating given to any product between 1 to 5. Review.username tells the username of the user who has reviewed the product.

Field Name	Description
<b>id</b>	Id of the product
<b>brand</b>	Brand of the product
<b>colors</b>	Color of the product
<b>manufacturer</b>	Manufacturer name
<b>manufacturerNumber</b>	Manufacturer number
<b>name</b>	Name of the product
<b>reviews.doRecommend</b>	Recommendation of the product (TRUE/FALSE)
<b>reviews.rating</b>	Rating of the product
<b>reviews.username</b>	Username of the reviewed user

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



One sample record from the offline-dataset (amazon.csv) is shown in the following figure:

id	brand	colors	manufacturer	manufacturerNumber	name	reviews.doRecommend	reviews.rating	reviews.username
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	5	JNH1
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	4	Appa
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	4	Kman
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	5	UpstateNY
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	5	Glickster
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	5	gjohnsonxc
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	4	nakulrk
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	FALSE	3	Angie
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft Surface	TRUE	4	papabear

**Fig. 15 Sample record of the Amazon dataset**

On this pattern , the actual dataset that is analyzed contains 6140 rows.

### **Description of the dataset**

- This dataset is for the Amazon rating analysis and prediction.
- This is an offline dataset which can be used by a user to analyze the ratings of various products of different categories (like Electronics data).
- With the help of this, user can predict whether any given product is going to get lower or higher rating level.
- This dataset can be further used for future references for the recommendation of any product.
- Moreover, the offline dataset is selected so as to determine the prediction accurately as online data gets updated very frequently.
- Also, to keep a track of its old customers to understand their preferences better.

This dataset is taken to perform the following tasks:

- to analyze the data records of a particular column field.  
Eg; To display all the names of the products available.

- to analyze the data records of various column fields collectively.  
Eg; To display the name of a product having rating of 5 given by username Droi.
- to analyze the data records of “reviews.rating” column of various products of different columns and therefore predict the ratings of a particular product from a fixed category.  
Eg; To predict the rating of All-new Fire HD 8 Tablet from Electronics category.

### 3.6.2 Split Dataset

For the implementation of the algorithms over the dataset (amazon.csv) according to the problem statement defined, **data preprocessing** is done.

The motivation behind preprocessing is to change over raw data into a structure that fits ML. Clean and structured data enable us to get increasingly precise results from an applied machine learning model. The technique includes data cleaning.

**Data Cleaning:** This set of procedures allows for removing noise and fixing inconsistencies in data. In case of any erroneous data, it is deleted or corrected if possible. This stage also includes removing incomplete and useless data objects.

One sample record from the dataset after data preprocessing is shown in the following table:

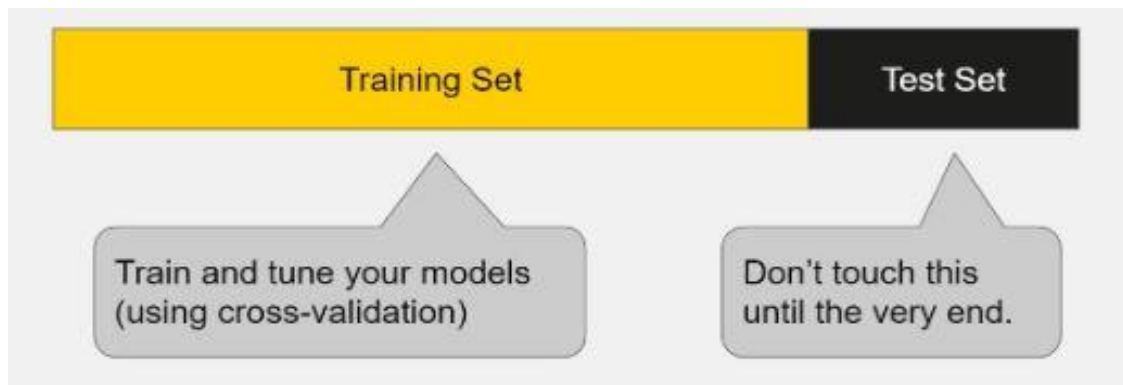
id	reviews.r	reviews.rating
1	TRUE	5
1	TRUE	4
1	TRUE	4
1	TRUE	5
1	TRUE	5
1	TRUE	5
1	TRUE	4
1	FALSE	3
1	TRUE	4

**Fig. 16** Sample record of the Amazon dataset(after pre-processing)

The above dataset is then splitted into 2 parts:

- **Training set**  
This is the actual dataset that is used to train the model. The model *sees* and thus *learns* from this data.
- **Test set**  
This being sample of data that is used to provide an unbiased evaluation for a final model fit over the training dataset.

After splitting,  
Training set contains **4605 rows**  
Test set contains **1535 rows**.



**Fig. 17 Training set and Test set**

- One should always split data before doing any steps further.
- This being the best method to get reliable estimates of the performance of the model.

If evaluating of the model takes place on the same data we used to train it, the model could be very outfit and then we wouldn't even know ! A model should be judged on the ability to predict new and unseen data.

## CHAPTER 4

### RESULTS AND PERFORMANCE ANALYSIS

#### 4.1 Using Hive

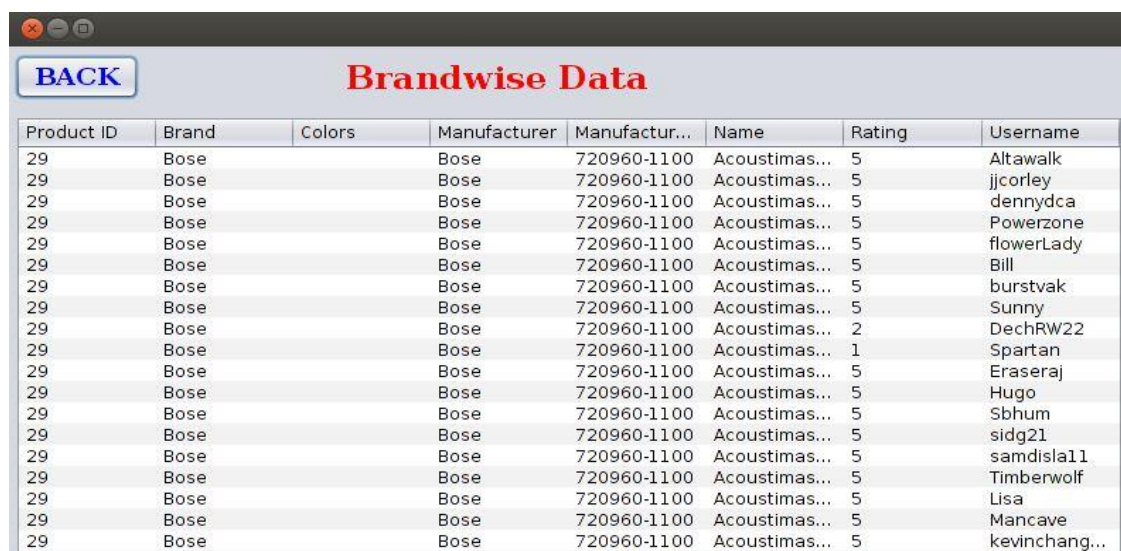
We implemented the project using Big-Data Analytics using Hadoop on Hive , and using the same following results have been obtained :

##### 4.1.1 Analysis

Analysis/categories of rating is done on the basis of brand-wise, manufacturer-wise and rating-value data.

As the user select one of the data from the combo-box from main.java, MapReduce program starts on the back-end and searches for the details of that particular brand, manufacturer and rating-value respectively and finally combine all the details and display that in a table using query (c), (e) and (h) respectively as described in Section 1.4.1.

- Analysis of rating done on the basis of brand-wise data



Product ID	Brand	Colors	Manufacturer	Manufactur...	Name	Rating	Username
29	Bose		Bose	720960-1100	Acoustimas...	5	Altawalk
29	Bose		Bose	720960-1100	Acoustimas...	5	jjcorley
29	Bose		Bose	720960-1100	Acoustimas...	5	dennydca
29	Bose		Bose	720960-1100	Acoustimas...	5	Powerzone
29	Bose		Bose	720960-1100	Acoustimas...	5	flowerLady
29	Bose		Bose	720960-1100	Acoustimas...	5	Bill
29	Bose		Bose	720960-1100	Acoustimas...	5	burstvak
29	Bose		Bose	720960-1100	Acoustimas...	5	Sunny
29	Bose		Bose	720960-1100	Acoustimas...	2	DechRW22
29	Bose		Bose	720960-1100	Acoustimas...	1	Spartan
29	Bose		Bose	720960-1100	Acoustimas...	5	Eraseraj
29	Bose		Bose	720960-1100	Acoustimas...	5	Hugo
29	Bose		Bose	720960-1100	Acoustimas...	5	Sbhum
29	Bose		Bose	720960-1100	Acoustimas...	5	sidg21
29	Bose		Bose	720960-1100	Acoustimas...	5	samdsla11
29	Bose		Bose	720960-1100	Acoustimas...	5	Timberwolf
29	Bose		Bose	720960-1100	Acoustimas...	5	Lisa
29	Bose		Bose	720960-1100	Acoustimas...	5	Mancave
29	Bose		Bose	720960-1100	Acoustimas...	5	kevinchang...

**Fig. 18 Display of Brand-wise Data**

- Analysis of rating done on the basis of manufacturer-wise data

The screenshot shows a web application window titled "Manufacturer-wise Data". It features a "BACK" button and a table with the following columns: Product ID, Brand, Colors, Manufacturer, Manufactur..., Name, Rating, and Username. The table contains 20 rows of data, all for Product ID 40, Brand JBL, and Colors Black.White. The manufacturers listed are JBL, and the product names are Everest Elit... The ratings range from 1 to 5, and the usernames are various user identifiers.

Product ID	Brand	Colors	Manufacturer	Manufactur...	Name	Rating	Username
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	thedishman
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	2	JPeterson
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	Rokinthere...
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	4	usasf
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	4	markyson
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	BigAI
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	hotb0x
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	4	PerAmadeus
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	4	MarkSta1969
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	2	MAHNYC
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	ryanthec
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	chenowitz
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	Audiofile2015
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	ArcticGoofball
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	1	Danny
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	1	Kamigawa
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	5	MDB983
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	2	NYITpro
40	JBL	Black.White	JBL	V700NXTWHT	Everest Elit...	3	21342156

**Fig. 19 Display of Manufacturer-wise Data**

- Analysis of rating done according to a particular rating-value (eg; 4)

The screenshot shows a web application window titled "Rating-wise Data". It features a "BACK" button and a table with the following columns: Product ID, Brand, Colors, Manufactur..., Manufactur..., Name, Rating, and Username. The table contains 20 rows of data, all for Product ID 1, Brand Microsoft, and Colors Black. The manufacturer is listed as Microsoft, and the product names are Microsoft S... The rating for all entries is 4, and the usernames are various user identifiers.

Product ID	Brand	Colors	Manufactur...	Manufactur...	Name	Rating	Username
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Appa
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Kman
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	nakulrk
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	papabear
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	lundi3
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	mevans
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Looloo
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	kooah
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Pod3000
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Wcgiv
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Brian
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	snobbycube
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	MyBestBuy
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Techfreak
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Pete
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	iceice
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Steve
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	HunterCub
1	Microsoft	Black	Microsoft	RH7-00001	Microsoft S...	4	Kooah

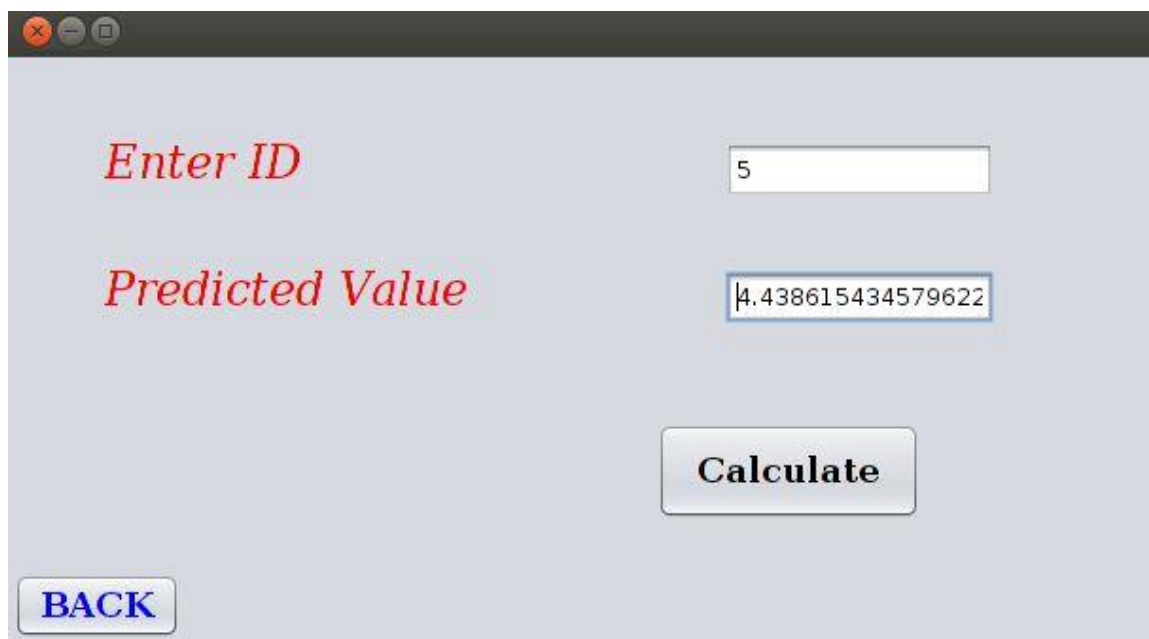
**Fig. 20 Display of Rating-wise Data**

### 4.1.2 Prediction

Prediction of rating of a particular product having product-id is done.

As the user enter the desired product-id, MapReduce program starts on the back-end and calculate the coefficients a and b of the equation ( $Y=a*X+b$ ) in the Linear Regression method, & finally the predicted rating-value ranging between 1-5 is calculated and displayed using query (j) described in Section 1.4.1.

- Prediction of rating of a particular product having product-id (eg; 5)



The screenshot shows a web application window with a light blue background. At the top left, there are standard window control buttons (close, minimize, maximize). The main content area contains the following elements:

- The text *Enter ID* in red, italicized font, positioned to the left of a text input field containing the number "5".
- The text *Predicted Value* in red, italicized font, positioned to the left of a text input field containing the value "4.438615434579622".
- A "Calculate" button with a grey gradient and rounded corners, centered below the input fields.
- A "BACK" button with a blue gradient and rounded corners, located in the bottom left corner of the window.

**Fig. 21** Display of predicted rating-value

### 4.1.3 Query with and without optimization

- Query for evaluating variables for linear regression is made to run, and using the variables predicted rating-value is calculated using query (j), as described in the Section 1.4.1.

When the query (j) for predicting rating of the product using Linear Regression was made to run on Hive(without optimization), then it takes 107.702 seconds as shown in Fig. 22.



```

hive> select ((sum(rating.id*rating.id)*sum(rating.rating) - sum(rating.id)*sum(
rating.id*rating.rating))/ (sum(rating.id*rating.id)*count(*) - sum(rating.id)*s
um(rating.id))) +
((sum(rating.id*rating.rating)*count(*) - sum(rating.id)*sum(rating.rating))/ (s
um(rating.id*rating.id)*count(*) -sum(rating.id)*sum(rating.id))) *5
from data.rating;
Query ID = hduser_20181129021616_202c6b44-35fb-45ec-87d8-a2beedd54afd
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201811290213_0001, Tracking URL = http://localhost:50030/jobd
etails.jsp?jobid=job_201811290213_0001
Kill Command = /usr/local/hadoop/libexec/./bin/hadoop job -kill job_2018112902
13_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-29 02:17:23,187 Stage-1 map = 0%,  reduce = 0%
2018-11-29 02:17:46,749 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.6 sec
2018-11-29 02:18:09,943 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 7.6 se
c
2018-11-29 02:18:11,963 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.99
sec
MapReduce Total cumulative CPU time: 10 seconds 990 msec
Ended Job = job_201811290213_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.99 sec   HDFS Read: 934268
HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 990 msec
OK
4.4386154345796225
Time taken: 107.702 seconds, Fetched: 1 row(s)

```

**Fig. 22** Display of query-run on Hive (without optimization)

- Query for evaluating variables for linear regression is made to run and variables after evaluation are then stored in a table, and using simple select query the predicted rating-value is fetched. It is accomplished using query (k) described in Section 1.4.1.

When the query for predicting rating of the product using linear regression was made to run on Hive (with optimization), then it takes 0.114 **seconds** as shown in Fig. 23.

```
hive> select a+5*b from reg;
OK
4.4386154345796225
Time taken: 0.114 seconds, Fetched: 1 row(s)
```

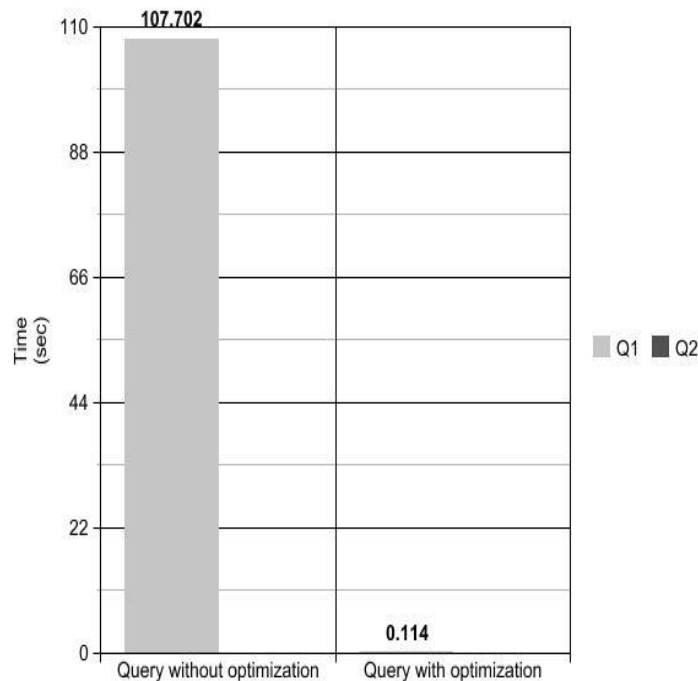
**Fig. 23** Display of query-run on Hive (with optimization)

**Table 2:** Comparison between query without optimization and query with optimization

	Query without optimization	Query with optimization
Time(sec)	107.702	0.114

So , on comparing both the scenarios it can be clearly observed that after optimization, time taken by Hive to run the query was decreased by **107.588 seconds** which is a significant number in the case of Big-Data as we have to deal with large datasets and access the data in shortest time period.

As the graph is an effective way to compare both the scenarios and allows the user to recognize pattern and trend far more easily than looking at a table of numerical data, therefore the same result of time consumed by the query without optimization and with optimization is shown with the help of bar-graph as shown in Fig. 24.



**Fig. 24:** Comparison between query without optimization and query with optimization



## 4.2 Using Machine Learning

This section discusses the results obtained for Machine learning algorithms on Python.

After training the dataset using 6 machine learning models, the models are evaluated over the test set and their accuracy scores are calculated.

Fig.25 describes the accuracy scores of the respective machine learning model after training the dataset. Accuracy score defines how much precise the trained model is to give the desired output.

And the time taken describes which algorithm takes how much time to get executed.

```
1)LR: 0.647557003257329
Time: 0.3275110721588135 sec
2)knn: 0.5322475570032573
Time: 0.027995824813842773 sec
3)Decison tree: 0.6358306188925081
Time: 0.06398582458496094 sec

4)SVC: 0.6482084690553745
Time: 1.3460557460784912 sec
5)Random forest: 0.6390879478827362
Time: 0.06678032875061035 sec
6)Naive Bayes: 0.6384364820846905
Time: 0.015622854232788086 sec
```

**Fig. 25 Display of accuracy score and time taken of the algorithms**

All the machine learning models :

Logistic Regression,  
KNN,  
Decision Tree,  
SVC,  
Random Forest, and  
Naïve Bayes

are executed 5 times and then the mean of their accuracy and time taken is calculated:

		<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>Mean</b>
<b>Logistic Regression</b>	Accuracy	64.755	64.755	64.755	64.755	64.755	64.755
	Time(sec)	0.2418	1.4018	0.2705	0.3079	0.3275	0.5099
<b>KNN</b>	Accuracy	53.224	53.224	53.224	53.224	53.224	53.224
	Time(sec)	0.0519	0.0189	0.0309	0.0156	0.0279	0.0290
<b>Decision Tree</b>	Accuracy	63.583	63.583	63.583	63.583	63.583	63.583
	Time(sec)	0.0499	0.1652	0.0440	0.0468	0.0639	0.0739
<b>SVC</b>	Accuracy	64.820	64.820	64.820	64.820	64.820	64.820
	Time(sec)	1.1833	1.0834	1.1175	1.2976	1.3460	1.2055
<b>Random Forest</b>	Accuracy	64.234	64.560	64.364	64.690	63.908	64.151
	Time(sec)	0.0559	0.3199	0.0600	0.0781	0.0667	0.1161
<b>Naïve Bayes</b>	Accuracy	63.843	63.843	63.843	63.843	63.843	63.843
	Time(sec)	0.0079	0.0209	0.0099	0.0156	0.0156	0.0138

**Table 3:** Five times execution of algorithms

**Final result is shown as following:**

	<b>Logistic Regression</b>	<b>KNN</b>	<b>Decision Tree</b>	<b>SVC</b>	<b>Random Forest</b>	<b>Naïve Bayes</b>
<b>Accuracy</b>	64.755	53.224	65.583	64.820	64.151	63.843
<b>Time(sec)</b>	0.5099	0.0290	0.0739	1.2055	0.1161	0.0138

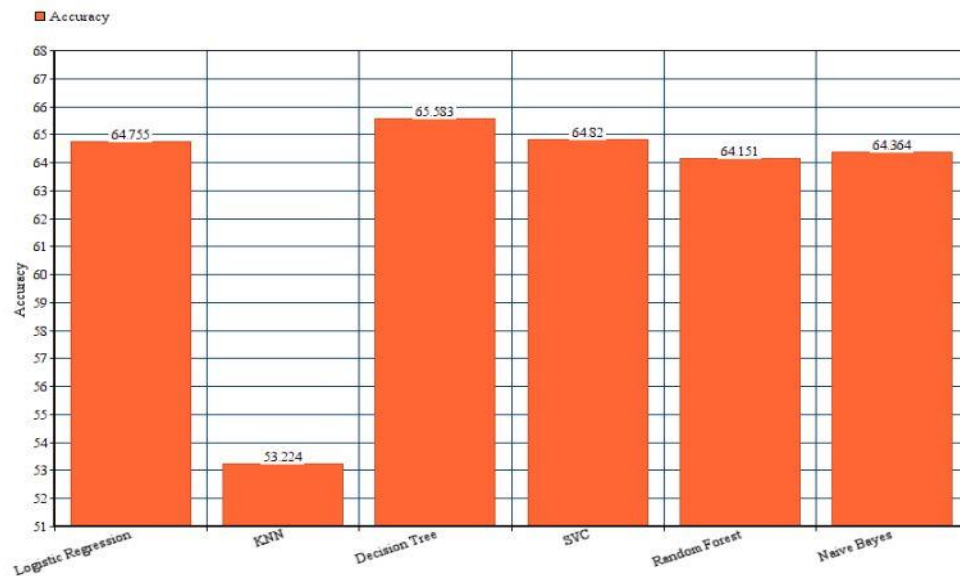
**Table 4:** Final result of Accuracy and Time taken of algorithms

So, on comparing all the scenarios it can be clearly observed that time taken for Naïve Bayes algorithm for execution is the least out of all the compared algorithms. Moreover, the accuracy score of Naïve Bayes algorithm comes out to be almost close to the highest accuracy score.

Therefore, taking both the scenarios into consideration, it is observed that for solving the desired problem statement, **Naïve Bayes algorithm** is the best out of all the other algorithms

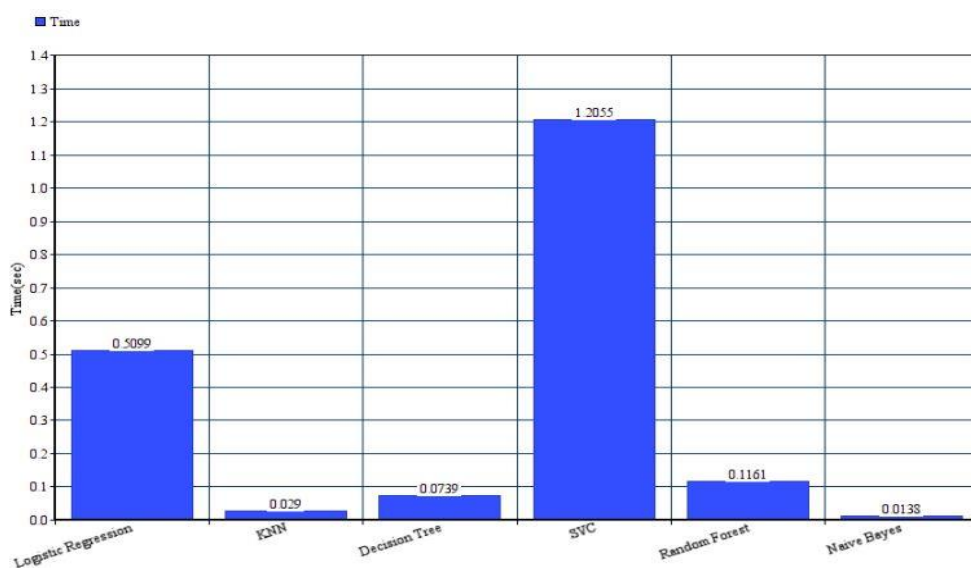
As the graph is an effective way to compare both the scenarios and allows the user to recognize pattern and trend far more easily than looking at a table of numerical data, therefore the same result of accuracy score and the time taken by algorithms is shown with the help of bar-graph as shown in Fig. 26 and Fig.27.

- The same result of accuracy score is shown with the help of bar-graph:



**Fig. 26** Display of accuracy score of the algorithms

- The same result of time taken is shown with the help of bar-graph:



**Fig. 27** Display of time taken of the algorithms

## CHAPTER 5

### Conclusion

#### a) Using Hive

As per the results achieved, we can clearly take this consideration that Hive can handle large datasets very efficiently, that includes fetching and concluding results in less time-frame.

Big-Data is being used by almost all the big companies so as to understand their business operations in a better way and also to extract some meaningful resulted data from the raw dataset being generated on a daily-basis. The results gives us a better information that after optimization we can use this system to predict the rating of a product very quickly, that is experimentally proved comes out to be 0.114 seconds, which is 107.588 seconds faster than the usual method without optimization.

- We have shifted the application of our problem statement from Hive to Machine Learning. The reason of the switch from Hive to machine learning is lack of limited subquery support. Also, hive cannot handle real time queries. The main Concern for using Hive is that is does not support OLTP(Online transaction processing).
- We are tend to be more interested in comparably small datasets where overfitting is the problem and moreover taking into consideration the concept to learn from trained data and predict future result, we have thus implemented the desired result using machine learning.

#### b) Using Machine Learning

After executing the given problem statement using six algorithms: Logistic Regression, KNN, Decision Tree, SVC, Random Forest, and Naïve Bayes; Naïve Bayes comes out to be the most appropriate as the application requires good performance and lesser time execution which is a key aspect of our project.

In our implemented work we can see that there is balance between accuracy and time taken in Naïve Bayes. In our project, we require such balance so as to not compromise with time as well as accuracy. Since, we have paid more attention to time factor therefore we select **Naïve Bayes** as the most accurate algorithm for our project.

## REFERENCES

- [1] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff and Raghotham Murthy , ' Hive - A Warehousing Solution Over a Map-Reduce Framework'
- [2] Swapna Sahu , 'Pattern Finding In Log Data Using Hive on Hadoop', IJRMPS | Volume 6, Issue 4, 2018
- [3] Scalability Study of Hadoop MapReduce and Hive in Big Data Analytics Khadija Jabeen1, Dr TSS Balaji2 1B , International Journal Of Engineering And Computer Science ISSN: 2319-7242 ,Volume 5 Issue 11 Nov. 2016, Page No. 18790-18792
- [4] Kshitij Jaju1, Vishal Nehe2,Abhishek Konduri3,' Commercial Product Analysis Using Hadoop MapReduce', International Research Journal of Engineering and Technology (IRJET, Volume: 03 Issue: 04 | April-2016 , © 2016 IJSRSET | Volume 2 | Issue 2
- [5] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy , ' Hive – A Petabyte Scale Data Warehouse Using Hadoop '
- [6] SK. Jilani Basha, P. Anil Kumar, S. Giri Babu , 'Storage and Processing Speed for Knowledge from Enhanced Cloud Computing With Hadoop Frame Work : A Survey'
- [7] Ronald Taylor , ' An Overview of the Hadoop/Mapreduce/Hbase framework and its current applications in bioinformatics'
- [8] Namrata B Bothe , 'Migration of Hadoop To Android Platform Using 'Chroot', Volume 1 | Issue 5
- [9] 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
- [10] Kajaree Das, Rabi Narayan Behera , ' A Survey on Machine Learning: Concept, Algorithms and Applications', Vol. 5, Issue 2, February 2017
- [11] Ayon Dey,' Machine Learning Algorithms: A Review', Vol. 7 (3), 2016,ISSN:0975-9646
- [12] K. R. Srinath,' Python – The Fastest Growing Programming Language' International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 12 | Dec-2017.

- [13] Philip K. Chan, Richard P. Lippmann, 'Machine Learning for Computer Security', Journal of Machine Learning Research 7 (2006) 2669-2672, Submitted 12/06, Published 12/06.
- [14] Wanliang Tan Xinyu Wang Xinyu Xu, 'Sentiment Analysis for Amazon Reviews, 2015
- [15] Callen Rain, 'Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning, 2016
- [16] Mohan Kamal Hassan, Sana Prasanth Shakthi and R Sasikala, 'Sentimental analysis of Amazon reviews using naïve bayes on laptop products with MongoDB and R', 14th ICSET-2017
- [17] Marco Passon †, Marco Lippi ‡, Giuseppe Serra †, Carlo Tasso, 'Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining, Proceedings of the 5th Workshop on Argument Mining, pages 35–39 Brussels, Belgium, November 1, 2018