ISSN:0975-9646

Prediction, Sentimental Analysis and Visualization of Static and Dynamic football data using Hadoop in a Multi-Node system.

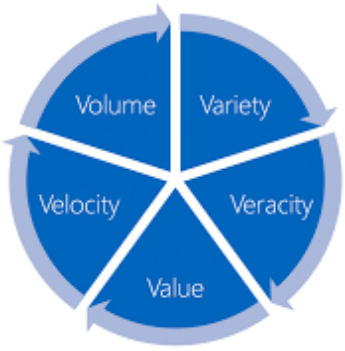
Shalom Mathews , Rohan Naik, Keziah Elsa John
B.E Computer Engineering
St.John College of Mumbai, India.

Abstract: In today's world huge amounts of data is present in the form of Datasets (Structured) and Twitter Dumps, logs (unstructured). These datasets are petabytes and zettabytes in size. Querying these datasets using traditional database techniques in inefficient and slow. Hadoop provides us with a framework which allows us to process this data using multi-cluster networks for parallel processing.
Hive is a tool in Hadoop provides us the capability to query huge datasets and extract required information from it. Flume which is another tool is used to stream live unstructured data into Hadoop. MapReduce is a programming model used to design programs which can execute on multimode Hadoop clusters
In this paper we are going to concentrate on sports related applications of Big data. We have used Hive is to query huge football related datasets spanning hundreds of years and extract current as well as historic information. Naive Bayes Algorithm is used to predict the future match results based on previous data. Flume is used to stream football related tweets into Hadoop. We have created a MapReduce program to perform sentimental analysis on this unstructured twitter data to find the fans sentiments about the manager and club's players. Popularity analysis is done based on the number of tweets in a minute.

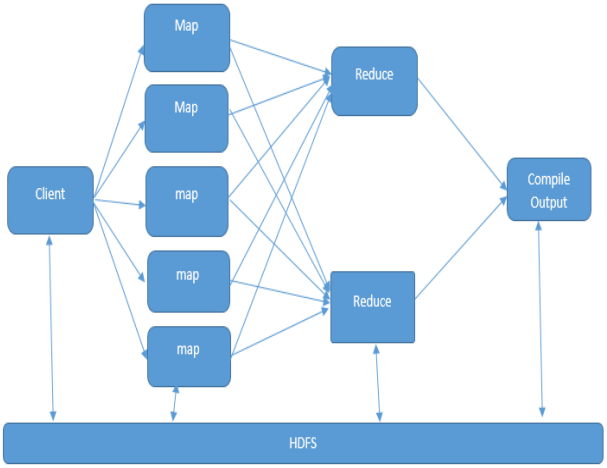
Keywords: Hadoop, Flume, Hive, Naïve Bayes Algorithm, Sentimental Analysis, MapReduce,

I. INTRODUCTION:

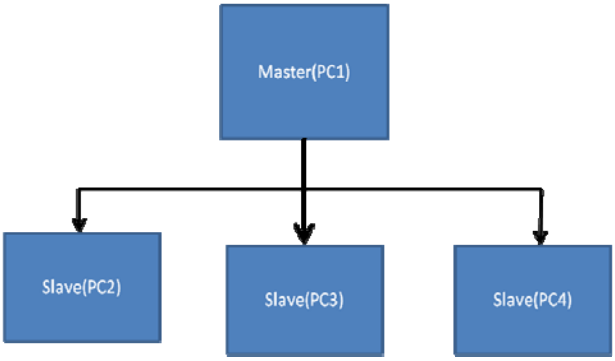
A there is a rapid explosion of Internet users across the globe. People use different media to express their view, ideas, emotions through varied online applications like Facebook, Twitter, WordPress and etc. For a clear illustration of data generation let's take Twitter into consideration. Twitter nearly generates Zettabytes of data per year to be more precise, it generates approximately 1 TB of data per week which are in the form of tweets.
It is very difficult for a relational database to store and process such large amount of data. Such large amount of data is coined as Big Data. Big data is vivid and complex in form. The 5V's of Big Data are: 1. Volume 2. Velocity 3.Variety 4.Veracity 5.Value
Volume: It refers to the vast amount of data generated every second. In Twitter as mentioned above millions of tweets are tweeted per day. It includes the amount of data generated from where, its depth and etc..
Velocity: It refers to the speed at which new data is generated and the speed at which data moves around. Within seconds the tweets get posted and viral also. Big Data is a collection of massive datasets obtained from different data sources that can be used for revealing business insights and for optimized decision making. Big data technology allows us now to analyze the data while it is being generated.
Variety: It refers to the different types of data we can now use. There are mainly three types of data: structured, unstructured and semi-structured data. Structured data is something that can be easily used by relational databases which can be in the form of tables or numeric or alphabetic. Unstructured data consists of images, videos, audios etc. Unstructured data is a combination of both structured as well as unstructured, example: chat box, email.
Veracity: It refers to the messiness, trustworthiness or availability of the data. With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations) but big data analytics helps to work with such types of data.
Value: It refers to the ability to turn the data into value which can be statistical, events or hypothetical.



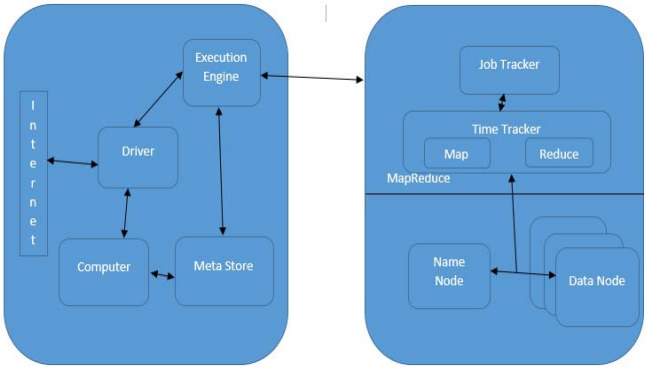
Hadoop: Then Hadoop comes in the picture. Apache Hadoop is an open -source software that is used for processing and analyzing large amount of data. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Than to depend on hardware to deliver high-availability of the data, the Hadoop is designed to detect and handle failures, so as to deliver a highly-available service on top of a cluster of computers, each of which may be prone to failures.



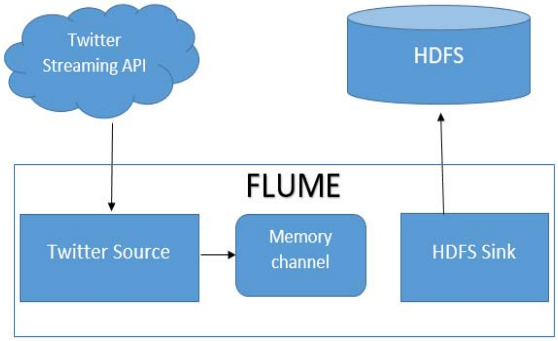
A multi-node set up:



Hive: Hive is tool designed which helps data engineers to query huge amount of structured data. Hive uses queries similar to SQL to query the data. The advantage with hive is that the queries are familiar and the MapReduce program code is generated by hive itself.



Flume: It is mainly used for online streaming of live and dynamic data.
Hadoop can handle any type of data like structured, unstructured, semi-structured or graphs.



Naïve Bayes: Naive Bayes is a simple classification algorithm.
The formula is given by the formula.
Subsequent= (Prior x Likelihood) ÷Evidence
Previous is called as base rates. It is information about the number of tuples classified in the classes.
Evidence contains the probability of attributes used to classify the tuples.
Likelihood: It is the probability of the tuple belonging to a certain class with certain attributes used to classify it.

II. Related Work:

The proposed system deals with future predictions and sentimental analysis. In the earliest system Hadoop and its tools has been used for election predictions, football match predictions, prediction of box office revenue, analyzing the sentiments of football fans.

3.1 Sentiment Analysis and Summarization of Twitter Data

Studied and analyzed the online text. Mostly prefer to write their experiences about the product that is in textual form. By using different algorithms Sentimental analysis over tweets and provides a summarized view over useful functionalities and features of such textual data in the form of tweets is performed

3.2 Market Sentiment Analysis for Popularity of Flipkart

FlipKart is a well-known ecommerce website. Sentimental Analysis is performed based on the Twitter tweets updated by the FlipKart user. It includes their reviews about the website and about the product which they bought. Reviews contain their sentiments which can be good or bad.

3.3 Predicting the Future with Social Media

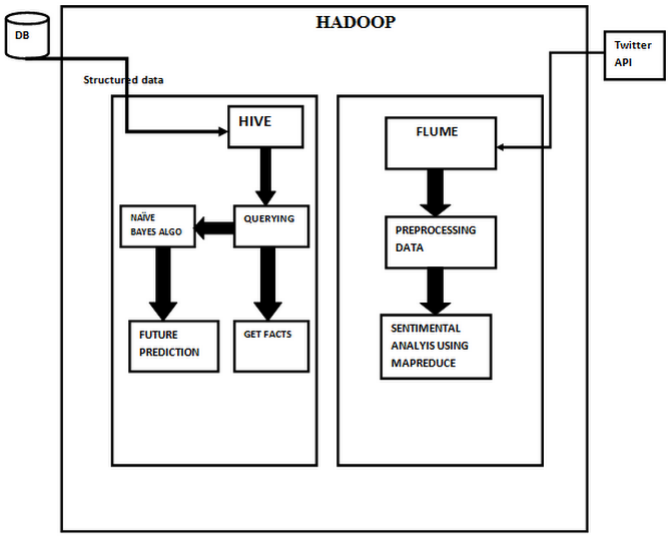
Demonstrated how social media can foresee the office revenues for movies depending on the Box-office hit and fail. It is based on the sentiments of the reviewers who tweeted about the movies.

III. Existing System:

In the current systems, RDBMS was the key medium to store data. Due to the data explosion and the growth of Internet and usage of certain applications like Facebook, Twitter etc. large chunks of data are generated on a daily basis following from Terabytes or Petabytes of data. Scalability, Fault Tolerance, Reliability were the certain issues faced by RDBMS. Processing and analyzing the

data were also some hurdles that RDBMS couldn't perform well. RDBMS couldn't process live or dynamic data. RDBMS has to ETL principles or make different schemas. As different types of data are generated like structured, unstructured or semi-structured, maintain and processing such variety of data is not possible by RDBMS. In today's high-tech world analyzing and processing the dynamic data plays a very vital role.

IV. PROPOSED SYSTEM:



By using Hadoop, we have overcome certain drawbacks of RDBMS like content management, scalability, handling real-time data etc. In the proposed system certain of useful tools belonging to the Hadoop ecosystem are used.
Flume.
Hive.

1. **Hive:** Data warehouse platform built on top of Hadoop query and is used for managing datasets across distributed storage. It queries the data on the basis of SQL and divides the data in tables, partitions or buckets.
A Hive Query to find out how many matches a team (Manchester United) won at their home ground in a season.

```
hive> select COUNT(*) from foot WHERE home='Man United' AND  home>away;
Query ID = root_20160421001612_32b7ff64-c959-4143-a509-5e2fcc87c0fa
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 mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2016-04-21 00:16:13,686 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local836694064_0005
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 193180 HDFS Write: 634 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
10
Time taken: 1.401 seconds, Fetched: 1 row(s)
```

2. **Flume:** It is mainly used for online streaming of live and dynamic data.
Hadoop can handle any type of data like structured, unstructured, semi-structured or graphs. In our project we have used static data that are mostly in structured form and dynamic data that are unstructured in nature. We have used the static and dynamic Twitter data of Manchester United's fans tweets.

```
1 RT @VintageMUFC: How we see Fellaini and how Van Gaal sees Fellaini
  https://t.co/9nucruoCXg
2 RT @RedDevilTimes: The good thing about this international break is
  that I won't get my weekend ruined by van Gaal.
3
4 #MUFC https://t.co/qL…
5 RT @strictvangaal: Van Gaal manages Mike Smalling
6
7 Smalling plays with Vardy for England
8
9 Vardy scores wonder goal for England after 5 minu…
10 RT @EPL_NewsID: Martial Hanya Tertawa Jika Dimarahi Van Gaal https://
  t.co/h1n96cdFhO
11 Louis Van Gaal https://t.co/G6ZhnF2qmn
12 RT @SkyFootball: Manchester United's Anthony Martial says he has
  learned to laugh when Louis van Gaal yells  https://t.co/obx5UKB5ie
  https:…
13 @JoeKarizma at least he'll show passion and determination unlike van
  Gaal
14 Van Gaal fancies Carrick and Fellaini instead. The mind boggles.
```

Naïve Bayes:
Naive Bayes is used to predict the results of future of football matches. The classifier classifies the unknown result into 3 different classes (Win, Loss, Draw) based on different attributes. Consider a sample football dataset (Rooney is a player).

Type	Home	Not Home	Rooney Playing	Rooney Injured	Rainy	Not Rainy	Total
WIN	600	200	400	400	450	350	800
Looss	0	300	200	100	150	150	300
Draw	200	200	150	250	50	350	400
Total	800	700	750	750	650	850	1500

We can pre-compute a lot of things about our Football Match collection.

Prior
P(WIN) = 0.53 (800/1500)
P(Loose) = 0.2
P(Draw) = 0.26
Evidence
p(Home) = 0.53
P(Rooney Playing) = 0.5
P(Rainy) = 0.43
Likelihood
P(Home|WIN) = .75
P(Home|Loss) = 0
....
P(Rainy|Draw) = 50/400 = 0.12
P(Not Rainy|Draw) = 0.87

Given a Football Match, how to classify it?

P(WIN|Home, Rooney Playing and Rainy) = (P(Home|WIN) .P(Rooney Playing|WIN).P(Rainy|WIN) x P(WIN))/ (P(Home) x P(Rooney Playing) x P(Rainy))/(P(evidence))

P(Draw|Home, Rooney Playing and Rainy) =(P(Home|Draw) x P(Rooney Playing|Draw) x P(Rainy/Draw) x P(Draw))/(P(evidence))

MapReduce

The Hadoop system partitions the given input data and schedules the program execution over multiple clusters or workstations. This helps in solving various computational problems. MapReduce follows the divide and conquer fashion.

Huge amount of data is split into small chunks of 64MB that are processed by mappers initially in parallel. After the map phase the data is given to the shuffle or sort phase and it creates intermediate results. These results are given to the reducers as an input. At the reducer phase, the final result is obtained with proper sorted result. As in Sentimental Analysis it maps the different emotions of the users through their tweets. Shuffles them on the basis of positive and negative comments and is given to the reducer phase to form the final result.

Formula used for Sentimental Analysis:

Sentiment = (positive - negative) / (positive + negative)

Map phase:

```
Map phase:
map(word,value):
String word = WORD_BOUNDARY.split(line);

  // Count instances of each (non-skipped) word.
  currentWord = new Text(word);
  context.write(currentWord,one);

  // Filter and count "good" words.
  if goodWords in word:
  value:context.getCounter(Gauge.POSITIVE).increment(1);
      emit(goodWords,value)


      else
  // Filter and count "bad" words.
  if badWords in word:
  value:context.getCounter(Gauge.NEGATIVE).increment(1);
      emit(badwords,value)
```

Reduce phase:

```
Reduce phase:
reduce(word,values):
//Get the counters from the Map class.
                Counters counters = job.getCounters();
  float good = counters.findCounter("org.myorg.MapSGauge", "POSITIVE").getValue();
  float bad = counters.findCounter("org.myorg.MapSGauge","NEGATIVE").getValue();
for value in values:

        // Calculate the basic sentiment score by dividing the difference of good and bad
words by their sum.

float sentiment = ((good - bad) / (good + bad));

// Calculate the positivity score by dividing good results by the sum of good and bad results.
Multiply by 100 and round off to get a percentage.
// Results 50% and above are more positive, overall

float positivity = (good / (good + bad))*100;
int positivityScore = Math.round(positivity);

emit(word,sentiment)
        else
emit(word,positivity)
```

V. METHODOLOGY:

1. Querying the static data using Hive and creating buckets:

The static football datasets are queried using multi-node clustered Hive network. The queried data are stored in the form of partitions or tables which makes the data more precise for analytics.

2. Future Predictions on the queried data using Naïve Bayes Algorithm:

Naïve Bayes algorithm is used for data classification and prediction of football results.

3. Creating Twitter Application:

A sample application on twitter was created and the consumer session key has been copied. Those keys are used to stream tweets into Hadoop.

4. Getting dynamic data using Flume:

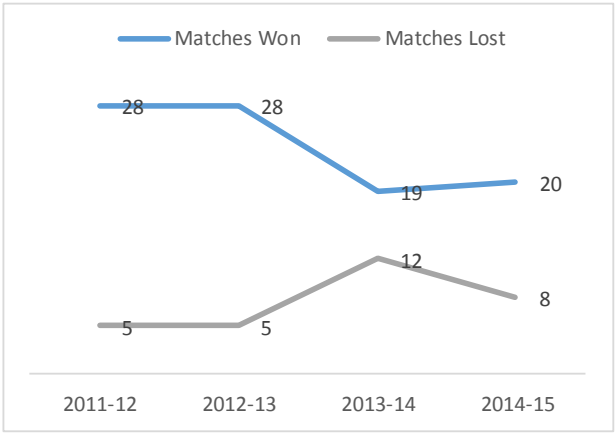
A flush is created using Flume. The data is flushed or in literal given to Hadoop system.

5. Preprocessing and performing Sentimental Analysis using MapReduce:

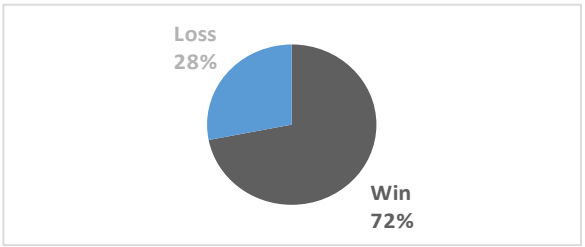
The dynamic data obtained from twitter has got many unwanted data in it. By preprocessing, the useful data are obtained by using delimiters.

MapReduce is a programming paradigm used for parallel processing of large amount of data is used to analyze the sentiments of twitter users through their tweets about the football matches.

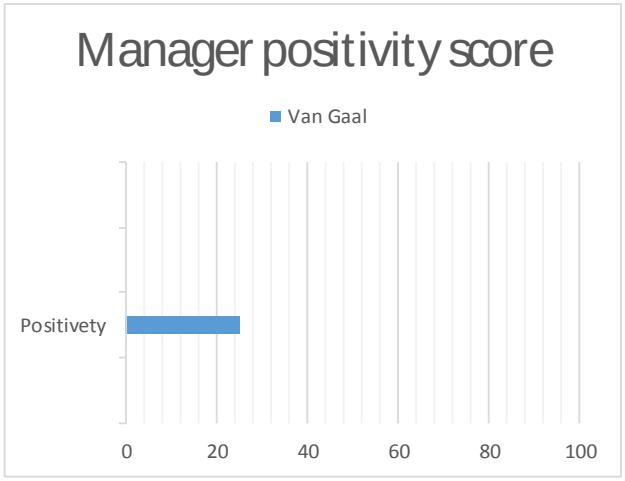
VI. EXPERIMENTAL RESULT:



Statics of the Manchester United's Performance (Lost or Won) queried using hive.



Probable Result of Manchester United's Next match at home calculated using Naïve Bayes.



Sentimental Analysis using tweets of Manchester United's Coach Van Gaal.

VII. CONCLUSION

In this paper, we have successfully queried the a static dataset and also perform future prediction using Naïve Bayes .We have also done sentimental and popularity analysis of real-time twitter data using Map Reduce.

REFERENCES

[1] witter Sentiment Analysis:The Good the Bad and the OMG! Efthymios Kouloumpis,i-sieve Technologies Athens, Greece.Theresa Wilson,HLT Center of Excellence Johns Hopkins University Baltimore, MD, USA

[2] Predicting the Future with Social Media -Sitaram Asur, Social Computing Lab, HP Labs, Palo Alto, California Bernardo A. Huberman, Social Computing Lab, HP Labs Palo Alto, California

[3] Twitter mood predicts the stock market.-Johan Bollen, Huina Mao,Xiao-Jun Zeng2.

[4] Stock Prediction Using Twitter Sentiment Analysis-Anshul Mittal, Stanford University Arpit Goel, Stanford University

[6] Big Data Processing Using Hadoop MapReduce Programming Model-Anumol Johnson, Avinash P H, Vinvce Paul, Calicut, Kerala

[7] Prediction of Movie Success using Sentiment Analysis of Tweets-Vasu Jain, Department of Computer ScienceUniversity of Southern California

[8] Integrating Predictive Analytics and Social Media Yafeng Lu, Robert Kr¨uger, Student Member, IEEE, Dennis Thom, Feng Wang,Steffen Koch, Member, IEEE, Thomas Ertl, Member, IEEE, and Ross Maciejewski, Member, IEEE

[9] Tweet Analysis: Twitter Data processing Using Apache Hadoop Manoj Kumar Danthala Dept. Computer Science Engineering Keshav Memorial Institute of Technology (KMIT