Demonetization-Twitter Data Analysis using Big Data & Hadoop

Malvika Goyal[1], Anuranjana[2]

[1,2]Department of Information Technology, Amity University Uttar Pradesh
[1]goelmalvika15@gmail.com, [2]aranjana@amity.edu

Abstract: In today's fast track online and globalised world analysing data and managing it is a major consent. Majority of people use online tools to share their data and views through twitter, Facebook. It has become extremely important to analyse and detect the positive and negative response a particular topic or issue in daily lives. To make data analysis easier and more conceptual Big Data and Hadoop along with various other tools like Hive, Pig, Sqoop have been used to analyse the review of people regarding Demonetization on Twitter platform. These excel at describing data analysis problems as data flows.

Keywords: Big Data, Hadoop, Hive, Pig, Sqoop, Demonetization

I. INTRODUCTION

The World Wide Web as inferred from the word itself has reflected a major change in communication between individuals across the globe. Use of internet to share data or information has increased widely.[3] Social media has evolved and expanded beyond anyone's expectation and imagination. People today extensively share their achievements, opinions and reactions on social platforms like Facebook, Google+, Twitter. The investigation depends on social media platform for posting comments through short statuses on any cause, situation on going in the real and world. A huge number of tweets are gotten each year that could be subjected to notion investigation. As indicated by present and progressing advances the present instruments and models accessible are not adequate to oversee such measure of enormous information. Where Big Data is an issue Hadoop is its answer. HDFS in Hadoop is an appropriated filesystem that stores records over the majority of the hubs in a Hadoop group. It handles breaking the documents into vast squares and conveying them crosswise over various machines, including making different duplicates of each square so that if any one machine flops no information is lost. The primary spotlight in my venture is on the investigation on the constructive and contrary reaction of individuals amid Demonetization in our nation.

II. LITERATURE REVIEW

Paper[1] the author in his paper, on the basis of various metrics related to Indian Premier League held in 2015 concluded its popularity. On the basis of analysis done it was found that the Indian Premier League is not only famous in India but has also gained immense popularity throughout the globe. The time interval in which cricket fans tweeted the most. The most talked about players are the various metrics that have also been analyzed.

In paper [8] the author did similar work in her paper by using HADOOP for analyzing the twitter data which is also known as a big data.

In paper [9] the author used Analytics tools and models used in it are not sufficient to manage big data. Therefore, there is a requirement of using cloud storage for such type of applications. The author has utilized Hadoop for intelligent analysis and storage of big data. In this paper, the author proposed a method for sentiment analysis on tweets in the Cloud environment.

Paper [10] concluded that with the help of big data and its tools a number of enterprises were able to improve customer retention, which further helped in gaining speed and complexity. Often E-commerce companies study traffic on web sites or navigation patterns to determine probable views, interests of an individual or a group as a whole depending on the previous purchases. Taking these factors in count they compared the results obtained from various data analytic tools.

2.0 Big Data – As a Problem The presence of new headways, contraptions, infers that [1] long range relational correspondence goals, the proportion of data made by mankind is growing rapidly and more reliably than previously. Enormous Data is a gathering of substantial datasets that can't be handled utilizing conventional figuring procedures or advances. It includes different instruments, procedures and systems. As the name itself reflects, Big information speaks to the data by high volume, speed and assortment to require particular innovation and scientific techniques for its change into some esteem. Huge Data can be broke down based on the accompanying three qualities:-

1. Volume : It implies the sum data ie set away and created. The proportion of the data is settled and its potential and regard is brought into thought dependent on which decision is made that comparable data is Big Data or not.

2. Variety : This implies the sort and nature of the data. Comparative helps people who separate the data to sufficiently use it as the consequent comprehension.

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

Immense data drawn from substance, pictures, sound, video. Missing spots are filled by data blend.

3. Velocity : This recognizes the speed at which the data is made and taken care of to address the solicitations and challenges that lie in the method for advancement and change. Tremendous data is oftentimes available logically.

Sources of Big Data :

Black Box Data: Its a piece of helicopter, planes, planes which gets voice of flight gathering, recording of enhancers or helpers in trading information about the carrier to the different aeronautics expert. Today, inside the field of substances mining and farsighted exhibiting programming, there are new revelation providers who need to offer the most minimum depictions of their computations.

Web based life Data: Social media handles like Facebook, Twitter and other social stages hold information and points of view posted by people from all around the globe[5]. Reliably on Facebook every day around 510,000 remarks are posted, 293,000 statuses are refreshed, and 136,000 photographs are transferred. Instagram clients transfer 46,740 million posts each moment. All the Facebook clients today tap the like catch on more then 4 million posts in a single day. This is not the end .

Stock Exchange Data: This type of information holds data about the 'purchase and offer' choices made on an offer of various organizations made by the customers.

Power Grid Data: It holds data about a specific hub as for a specific base station. As indicated by the investigation made in 2015 there will be there are 3111 power relations in Germany .Transport Data: It comprises of limit, separation, accessibility or essential data about a specific method of transport

Web index Data: Uncountable number of databases associated with web indexes create a lot of information. Based on investigation made by the Internet Live Stata there are around 40,000 pursuit inquiries that Google forms each second by and large, more than 3.5 billion quests for every day, and 1.2 trillion ventures for each year around the world. There are three types of data;

Structured Data:Its that data which has a defined length and format for a particular dataset. For example – Relational data including dates, numbers etc.

Unstructured Data:Its that data or information which doesn't have apredefined data model. It consists of data in the form of Media logs, Text or word.

Semi Structured Data: The most appropriate example of this kind of data is XML

III. HADOOP – A SOLUTION FOR BIG DATA

[7]Doug Cutting and his team together built an open source venture called Hadoop. It runs applications utilizing the Map Reduce Algorithm.Hadoop is utilized to create applications that performs factual investigation on gigantic measures of data. Hadoop runs applications utilizing the MapReduce calculation, where the information is prepared in parallel to various CPU hubs. Hadoop system is sufficiently able to create applications equipped for running on a number of PCs simultaneously. The six administrations of Hadoop are as per the following: JPS (Java Profile Services) Data Node is where genuine information is stored. Name Node is where Meta information is there and alter logs are seen. Optional Name Node is the place checking happens. (Hadoop Distributed File System) Resource Manager, Node Manager (YARN – Yet Another

Resource Negotiator.

3.1 Map Reduce Algorithm

[8] MapReduce is an essential yet an extreme parallel data dealing with perspective. Every movement in MapReduce includes three essential stages: In the guide organize the application has the opportunity to deal with each record in the information freely. Various maps are started immediately so that while the data may be gigabytes or terabytes in size, given enough machines, the guide stage can as a general rule be done in under one minute. In the revise organize which happens after the guide arrange, data is assembled by the key the customer has picked and spread to different machines for the diminishing stage. Each record for a given key will go to a comparative reducer. In the reduce arrange the application is shown each key, together with most of the records containing that key.

3.2 METHODOLOGY

[6]Hadoop requires the Java Virtual Machine to run. Subsequently utilizing Ubuntu working framework it is required to introduce Java before introducing Hadoop. At the point when Hadoop has been effectively introduced one needs to begin Hadoop by beginning its six administrations. This is done in the Pseudo Distributed Mode. After that the Hadoop Ecosystem was installed which includes installation of Tools like Pig, Hive, Flume, Json validator Sqoop etc. The Map reducer algorithm plays an important role in analysing data. Also Pig tool is used to perform the word count ie number of times a word is repeating in a file. Using the following steps word count was conducted.

[4]In Hive's use of separating, data inside a table is part over various designations. Each bundle analyzes to a particular value(s) of portion column(s) and is secured as a sub-file inside the table's vault on HDFS. Right when the table is addressed, where material simply the required packages of the

table are addressed, in this way diminishing the I/O and time required by the request. Portions are level record of data which grants considerable datasets to get disconnected into more sensible pieces. In Hive, allotting is maintained for both managed dataset in coordinators and for outside tables also. Table distributing secluding table data into a couple of areas subject to the estimations of particular fragments like date or country, separate the information records into different reports/indexes reliant on date or country. Partitioning ought to be conceivable reliant on more than section which will constrain multidimensional structure on vault storing. Each one of the inventories or reports isolated in Tweets of people be it positive or negative is fetched from the above mentioned Twitter Handle . Json Validator was used to convert the tweets into human readable form. After fetching the data time zone wise analysis and statistical conclusion was made i.e how many people from around the world were satisfied and not satisfied with Demonetization on the to January 2017, satisfied and not satisfied with Demonetization – a major movement This graph the statistical analysis on demonetization for our country which affected all citizens in some or the way. Basis of data fetched for three months. Pig tool is used to load tweets in the Hadoop terminal from the extracted file in the directory of the system Different commands are run in the terminal to conduct time zone wise analysis using Hive. Graph was made using MS Excel.

Fig. 3.2.2. This table shows if a user has retweeted, if yes then there tweet count.

IV. RESULT

```
hive> select * from customer;
OK
Jaideep 1234    21      Rohini,Delhi
Simran  2345    23      Jaipur,Rajasthan
Karan   3456    24      Ajmer,Rajasthan
Ashish  4567    43      Pitampura,Delhi
Ashutosh        5678    54      Faridabad,Haryana
Chandi  6789    22      Noida,UP
Kannu   7890    20      Lajpat Nagar,Assam
Abhishek        3214    66      Gandhi Nagar,Gujarat
Riya    7654    43      Bikaner,Rajasthan
Rahul   6532    30      Juhu,Mumbai
Time taken: 0.143 seconds, Fetched: 10 row(s)
hive>
```

Fig. 4.1. This figure shows the timezone of different places from where users have tweeted.

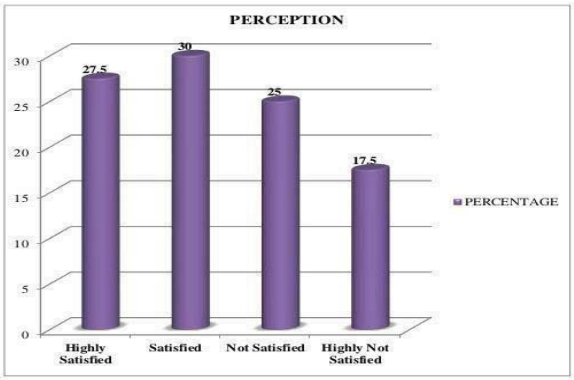

Fig. 4.2. The graph shows the overall analysis of the people from the months of November 2016 to January 2017, satisfied and not satisfied with Demonetization.

V. CONCLUSION

The openness of Big Data, and new information organization and illustrative programming have made an extensive variation in the verifiable background of data examination. The gathering of these examples infers that we have the capacities required to analyze astonishing data. They address a genuine bounce forward and a sensible opportunity to recognize huge gains similar to viability, proficiency, wage, and advantage. In view of all the examination done on demonetization it was found that the amount of people agreeing and contrasting have fluctuating extents.

REFERENCES

[1] A CaseStudy with analysing Twitter Data using Apache Hive by Ankit Kumar, AdityaBhardwaj (2015)

[2] Big Data as a service solution for building graphs and social networks.By Sihan Yousifi Dalila Chiadmi (2016)

[3] Data analysis techniques in qualitative research paper by Barbara B.K, State university of West Georgia.

[4] Book On Big Data Analysis by Andy Konwinski, Holden Karau, Matei Zaharia, and Patrick Wen. [2016]

[5] Book on Sqoop on Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database

[6] Evaluation datasets for twitter sentiment analysis: A survey and a new dataset. By Saif, H., Fernandaz, M., & Alani, H. (2013)

[7] Classifying the political sentiment of tweets.By Johnson, C., Shukla, P., & Shukla, S. (2012). s

[8] Sentimental Analysis of Applications using Twitter data with help of Hadoop framework by Divya Sehgal (2016)

[9] Analysing Twitter Sentiments through Big Data by Monu Kumar

[10] Big Data Analytics : Hadoop and Tools by Ms Shikha Pandey(2016)