Performance Comparison of Statistical Techniques with Big Data Analysis

Stuti Mehla
Research Scholar
MMU, Mullana, Ambala

Saurabh Upadhyay, PhD
Associate Professor
MMU, Mullana, Ambala

ABSTRACT

In present scenario computers are involved in every field of daily life, leads to increment in volume ,variety and velocity of data. It makes difficult for Conventional Statistical techniques to handle these huge datasets and results in emergence of Big Data and different tools such as Hadoop, Hive for analyzing it. This paper describes the generic statistical techniques such as classification, regression and tools for analyzing Big Data. Analytical Comparison of these tools on different aspects are also explained.

General Terms

Database, Big Data, volume, variety, velocity

Keywords

Big Data, IOT, Hadoop, HDFS, MapReduce analytics, Hive.

1. INTRODUCTION

In present era data is exploding and in every field there is advancement whether it is health sector where trillions of information about patients, pharmaceuticals, clinical environment has to be kept [6], industrial sector which organizations are capturing millions of information about their products, consumers and suppliers or any another sector like economy [9]. Technology is booming day by day and rise of social media, e-commerce, multimedia, IOT and digital advancement in every sector [5] have created a large increment in data. In Mckinsey report it is estimated that every day 2.5 Exabyte (2.5×10^{18}) is created. It makes difficult for typical database software and tools to handle, store and analyze this large Data. Large data is termed as Big Data having huge volume, large variety and highest velocity. From different reviews Big Data is fully understandable by 6V's volume, velocity, variety, variability, veracity and value.

Volume attribute represents data which has come from different sources and it is increasing exponentially from Terabyte to Petabyte and Petabyte to Exabyte. [1]

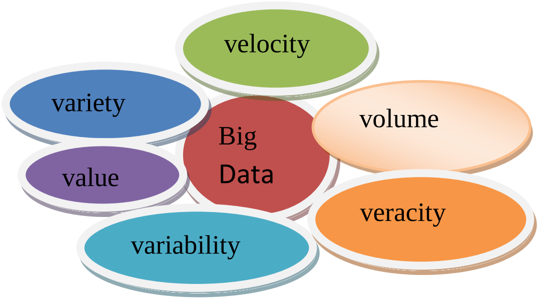

Fig 1:6v's of Big Data

Digital data is increasing in every sector. It is estimated that companies of US have more than 1petabytes data. It is concluded that manufacturing, media, healthcare and communications are major sectors for this large volume of data.

Variety [2] refers to data that is explained with multiple attributes results in structured, semi structured and unstructured format. It also refers to the data which is collected from the different type of sources i.e. social networks, IOT, companies, sensors, multiple data types e.g. audio, video, text, image and data logs, etc.To handle so large variable data is difficult for conventional database tools i.e .SQL, DB2 and this is overcome by NOSQL (not only SQL).

Velocity [3] refers to the rate at which data is generated. Data is coming in the form of streams and in recent searches it is estimated that the growing of data is speeding up at an exponential rate. It makes difficult for researchers to find out the useful information.

Variability[7]refers to the inconsistency with the flow of data and thus hampers to handle and manage the data effectively. Now it becomes a problem for researchers to interpret such a variable data.

Veracity[3] refers to data accuracy. The quality of the data that is captured greatly varies because it comes from the different origins. Accuracy of analysis depends on the veracity of the source data. How much analysis is accurate it depends upon veracity of the source data.

Value refers to find out the hidden values from large databases so that decision making based on mining and answering the queries will become easy. [3,5]

These features make Big Data so difficult for traditional software and tools to analyze. In this research paper first statistical analysis is done and then analysis of Big Data is done with the help of Hadoop and Hadoop with Hive.

Technologies which are used for Big Data having different characteristics and purpose of use is also different. According to purpose on technology Big Data is divided into two parts: operational Big Data and Analytical Big Data.

Operational Big Data[11,12]is that which provide interactive data and this type of data is generally in unstructured or semistructured format and here NOSQL come into existence. NOSQL makes easy to manage and implement the Big Data.

Analytical Big Data[11,12] implement MapReduce which complements the SQL. In MapReduce,data is processed in parallel by making clusters. Client-server architecture is followed and server will have NameNode [13] and client will have DataNode [13] and then clients do the work and transfers it to name node.

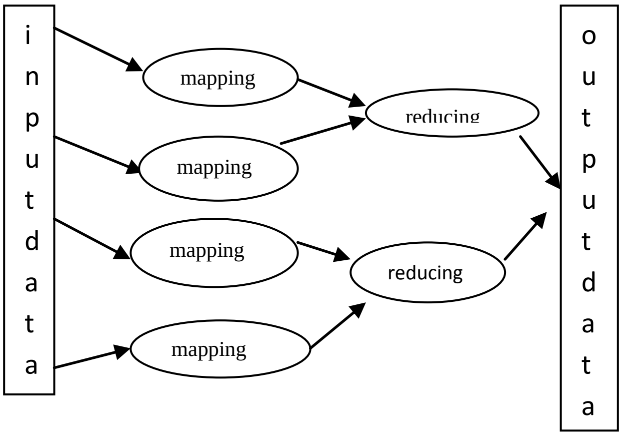

Fig2: mapreduce in bigdata

Another important distribution of tools or platforms on the basis of scaling. Scaling is defined as that feature which describes how much a system is adaptable to new conditions. It is also further divided into two categories: horizontal scaling, vertical scaling.

Horizontal scaling [5,10]is also known as scale out scaling. In this total work is distributed on multiple machines and it is not difficult to scale up for eg Hadoop, Spark.

In Vertical scaling[5,10]multiple processors are added to speedup the scaling of machine. In this scaling technique data is partitioned.

Hadoop and Hadoop with Hive is used for analysis in research work.

According to Hadoop tutorials Hadoop is an Apache open source framework which allows distributed processing of large datasets. It is designed to scale up from single server to thousands of machines. Hadoop contain four modules[14].Its architecture is described

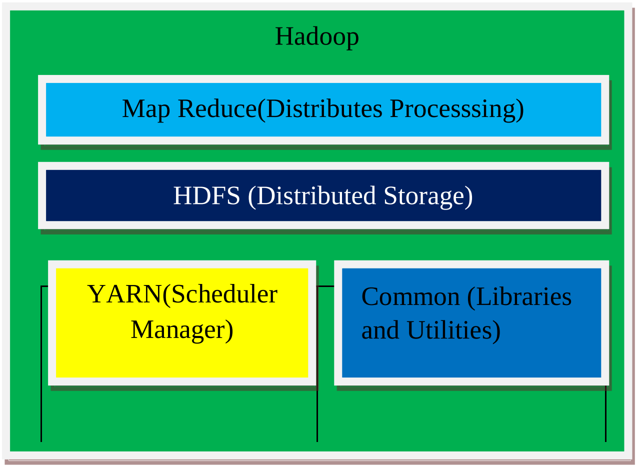
Hadoop
Map Reduce(Distributes Processsing)
HDFS (Distributed Storage)
YARN(Scheduler Manager)
Common (Libraries and Utilities)

Fig 3:architecture of hadoop

Mapreduce works on distributed processing. It first maps the input data to client nodes and then reduce the result to server nodes.

HDFS uses distributed storage ie client-server architecture is followed. Client nodes are known as datanode and server contain namenode. Hadoop YARN act as scheduler and manager.It schedules the job and manage the cluster nodes. Hadoop common include libraries and utilities which are used during processing.

Hadoop is that framework which helps in processing large data sets.In hadoop there is mapreduce and HDFS where mapreduce works on parallel processing and HDFS works on distributed file processing.[15]

According to Jerome Serranoto enhance the workability of Hadoop, it is used with Hive. Basically Hive is a datawarehouse which is built on top of mapreduce and HDFS. It converts QUERIES into mapreduce jobs and run them into cluster

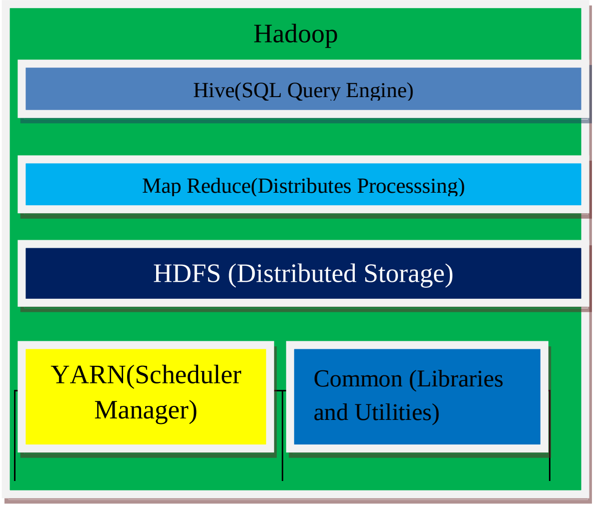
Hadoop
Hive(SQL Query Engine)
Map Reduce(Distributes Processsing)
HDFS (Distributed Storage)
YARN(Scheduler Manager)
Common (Libraries and Utilities)

Fig4: position of Hive with Hadoop

2. MATHEMATICAL FORMULATION AND FLOWCHART

(i) Suppose BG is a matrix of Bigdata containing BigData BG=(d1,d2……..dn)where n€N

(ii) Statistical technique Correlation,regression followed by Hadoop and Hadoop with Hive are termed as(t1,t2,t3,t4) €T.

(iii) Every instance of set T is applied on set BG.

(iv) Results are compared in terms of time and throughput.

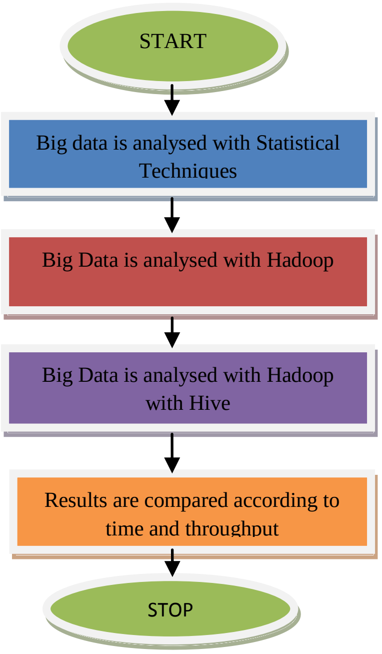
START
Big data is analysed with Statistical Techniques
Big Data is analysed with Hadoop
Big Data is analysed with Hadoop with Hive
Results are compared according to time and throughput
STOP

Fig 5:flowchart

3. METHODOLOGY

During analysis of Big Data using Hadoop we have followed the model which is explained below. During analysis we have taken tweet data which is mixed data and then preprocessing of that is done. After preprocessing text data is tokenized and then TF-IDF is calculated and it is converted into an variable array. This variable array is converted into positive and negative tweets. These tweets then fed to Naive Bayes Model which is classifier model and output comes according to accuracy, precision and recall features. This proposed model is applied on Hadoop and Hadoop with Hive which concludes that Hadoop with Hive is not only better in terms of time, throughput but also in accuracy, precision and recall.

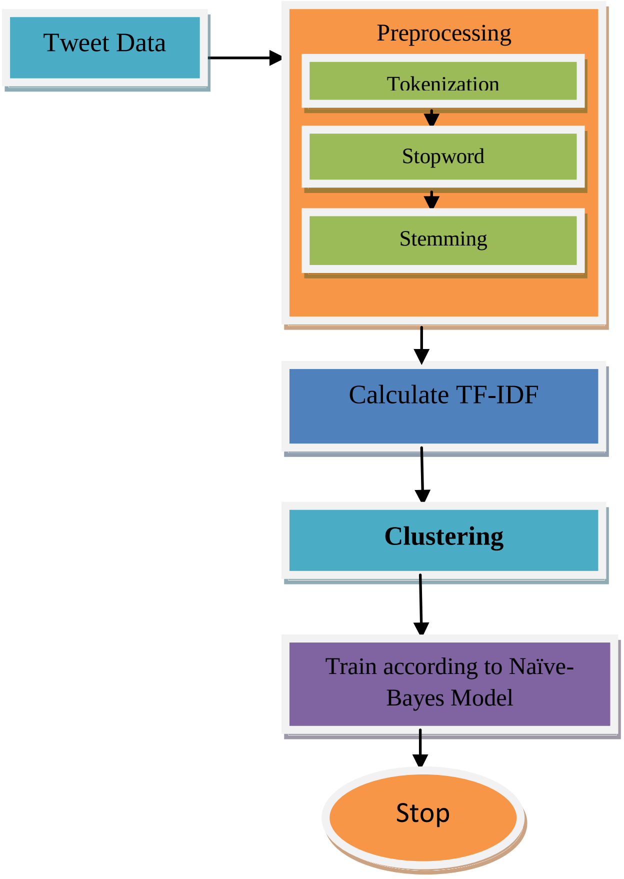

Fig 6: Block diagram of processing of Big Data

4. RESULTS

This graph shows the comparison of time on y-axis and number of tweets on horizontal axis between regression, correlation, hadoop and hadoop with hive.

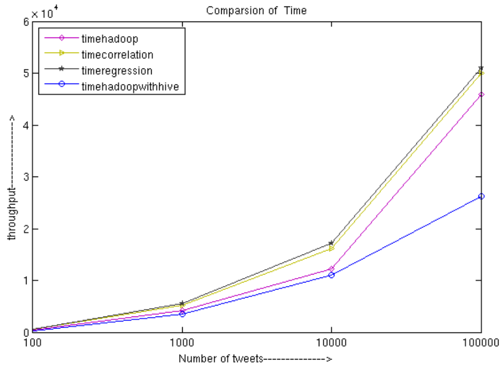

Fig 7: comparison between time and number of tweets

The graph shows Hadoop with Hive has done processing fast.

Throughput refers to no of jobs done in a particular time. This graph shows the comparison of throughput among statistical techniques, Hadoop and Hadoop with Hive.

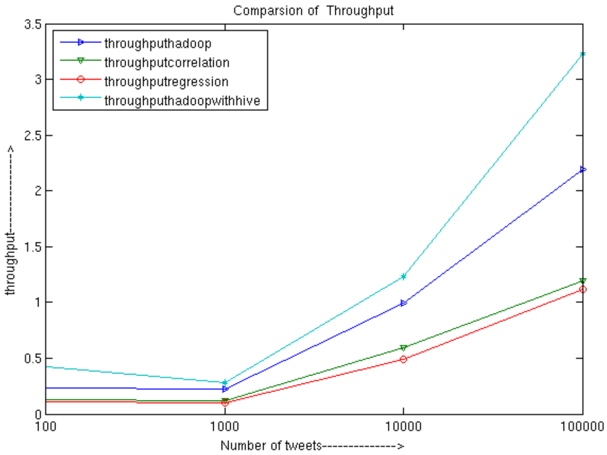

Fig 8: Comparison between throughput and number of tweets

This graph shows that hadoop with hive has the best throughput.

This graph shows the comparison between Hadoop and Hadoop with Hive according to accuracy, precision and recall parameters.

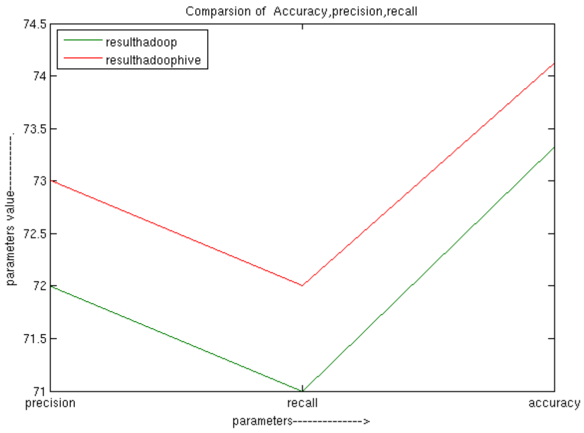

Fig 9: comparison between hadoop and hadoop with hive

This result shows Hadoop with Hive has better accuracy, precision and recall.

5. CONCLUSION

Big Data is now an emerging research field. Different tools have overcome traditional statistical tools for analysis. In future tools like hadoop can be enhanced by changing the features of clustering. Big Data is the most challenging area for researchers because Big Data is related to every field whether it is social sites,e-commerce, government, healthcare, industrial or any other sector which is hitech. In this research paper it is shown that hadoop is better than correlation and regression (standard analysis techniques) and hadoop is also enhanced using Hive with it. Results also show that when Hive is used with Hadoop it has better accuracy, precision, recall features.

6. REFERENCES

[1] ScaDiPaSi: An Effective Scalable and Distributable MapReduce - Based Method to Find Patient Similarity on Huge Healthcare Networks. Mohammad hossein Barkhordari[1,], Mahdi Niamanesh[1,] Copyright © 2015 Elsevier Inc.

[2] Promises and Challenges of Big Data Computing in Health Sciences Tao Huang[b, 1,] ,Liang Lan[c, 1,] , Xuexian Fang[a,] , Peng An[a, d,] , Junxia Min[d,] , Fudi Wang[a, ,] Copyright © 2015 Elsevier Inc.

[3] Significance and Challenges of Big Data Research Xiaolong Jin[a, ,] , Benjamin W. Wah[a, b], Xueqi Cheng[a], Yuanzhuo Wang[a] Copyright © 2015 Elsevier Inc.

[4] Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix Shaokun Fan[a, ,] , Raymond Y.K. Lau[b,] , J. Leon Zhao[b,] Copyright © 2015 Elsevier Inc.

[5] The rise of "big data" on cloud computing: Review and open research issues Ibrahim Abaker Targio Hashem[a, ,] , Ibrar Yaqoob[a,] , Nor Badrul Anuar[a,] , Salimah Mokhtar[a,] , Abdullah Gani[a,] , Samee UllahKhan[b,] Copyright © 2014 Elsevier Ltd.

[6] Big data analytics in healthcare: promise and potential Wullianallur Raghupathi and Viju Raghupathi Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 http://www.hissjournal.com/content/2/1/3

[7] Mining Big Data: Current Status, and Forecast to the Future Wei Fan, Albert Bifet SIGKDD Exploration Volume 14,Issue 2

[8] CRITICAL QUESTIONS FOR BIG DATA danahboyd & Kate Crawford Information, Communication & Society Vol. 15, No. 5, June 2012, pp. 662–679 ISSN 1369-118X print/ISSN 1468-4462 online # 2012 Microsoft.

[9] Big data: The next frontier for innovation, competition, and productivity McKinsey Global Institute June 2011

[10] The Parable of Google Flu: Traps in Big Data Analysis BIG DATA David Lazer, 1, 2 Ryan Kennedy, 1, 3, 4 Gary King, 3 Alessandro Vespignani 3,5,6

[11] Operational Vs Analytical Big Data https://www.mongodb.com/scale/operational-vs-analytical-big-data

[12] Analytical Master Data Management https://blogs.oracle.com/mdm/entry/operational_vs_analytical_master_data_management

[13] Hadoop Tutorial http://www.tutorialspoint.com/hadoop/

[14] Hadoop Tutorial http://www.tutorialspoint.com/hadoop/ ord.

[15] What is difference between Hadoop, Hive, and AWS RedShift? https://www.quora.com/What-is-difference-between-Hadoop-Hive-and-AWS-RedShift