A Survey on Data mining and Analysis in Hadoop and HBase

Mrs. Amita V. Shah[1], Mr. Krunal Panchal[2]
[1, 2] L.J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India.

Abstract- *Data Mining is a process to generate pattern and rules from various types of data marts and data warehouses, in this process there are several steps which contains data cleaning data anomaly detection then clean data is mined with various approaches. In this research we have discussed data mining on large datasets (Big Data) with this large data set major issues are scalability and security ,Hadoop is the tool to mine the data and Mongo db provides input for it, which is a key-value paradigm for parsing the data ,Other approaches are discussed with this report and their capability for data storage ,Map reduce is method which can be used to reduce the data set to reduce query processing time and improve system throughput, In the Proposed system we are going to mine the big data this Hadoop and Mongo db and we will try to mine the data with sorted or double sorted key value pair ,for and analyze the outcome of system.*

Keywords- Data Mining, Hadoop, MapReduce, HDFS, HBase.

I. INTRODUCTION

"Big Data" is data whose scale, diversity, and complexity require new architecture, Techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

Amount of data generated every day is expanding in drastic manner. Big data is a popular term used to describe the data which is in zeta byte.[1] . Big Data is large amount of data. This vast amount of data is generated by social media and networks, scientific instruments, mobile devices, sensor technology and networks. Ability to manage, analyze, summarize, visualize, and discover knowledge from the collected unstructured data in a timely manner and in a scalable fashion is very difficult task using traditional data mining tools. To analyze the data Apache introduce a new technology called Hadoop. We can describe the characteristics of big data using three Vs Volume, Variety and Velocity. [13][14]

Hadoop is the part of Apache projects, Hadoop software library is a framework that supports distributed processing of large data across clusters of computers using simple programming models.

NOSQL [3][7] is the term related to " Not Only Sql " Sql is a relational database language but for big data analysis these techniques are not enough so alternative solutions are NoSql databases like Mongodb, Cassandra,Voldmort etc.

II. LITERATURE SURVEY

2.1 Applications

This proposed will provide a new approach analysis big data mining with hadoop and mongodb which is based on MapReduce Paradigm. This new approach will try to improve the computational time, more fault tolerance of system and will handle or deal with Bigdata analysis.

2.2 Related Work

This chapter will provide information about the work done in big data mining and various approaches use and method proposed

In [1] author has discussed the meaning and importance of big data analysis programming tool use for big data mining and important of big data , with the example of facebook we can understood that today it is required to process large number of data sets ,our traditional data sets are not enough for that ,for example instead of taking large MySql tables we can use caching approach from memcached for n tier elements as Mysql has very good performance in read but they are lagging in write ,which leads us to very high reliability but low partition tolerance in our CAP model ,another example author has given is Yelp which uses *AWS and Hadoop* for data analysis which uses *Amazon S3* server to store large datasets which is *RAID* service.

The author proposed such data analysis using Apache Hadoop and *JSON* and data stored. From Amazon web services using their web services and analyze the data, the analysis showed that this method can analyze the large data from different sources with minimum utilization of resources

In [2] In this paper author has utilized Nosql database Mongo db to implement the big data analysis as it is advantageous over rigid sql tables which is not useful in

today's large scale data for web logs generated every day .more over author has compared performance between Mongo db and HDFS frame work using inbuilt map reduce method with mongo db , author has not defined the modern data store technology and integration available with hadoop like Hbase , and HIVE for that experiments and results are shown for large amount of data sets , this is the motive why we choose mongo db data store for Large data sets.

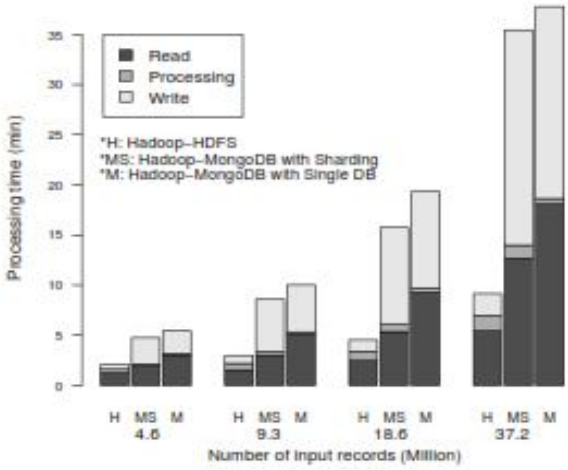

Figure 2.1 [HDFS –Mongo Db comparison]

With this framework proposed by author the output comparison shows that Figure shows the effect of the split size on performance using mongo-hadoop. The number of input records is _ 9:3 million, or 4GB of input data. With the default split size of 8MB, Hadoop schedules over 500 mappers; by increasing the split size, we are able to reduce this number to around 40 and achieve a considerable performance improvement. The curve levels o_ between 128MB and 256MB, so we decided to use 128MB as the split size for the rest of our tests both for native Hadoop-HDFS and mongo-hadoop.

In , MS At el.[3] has discussed various security issues and threats available with Big data as data is in zeta byte size it also contains some sensitive and confidential information it is necessary to prevent unauthorized use of data so apart from storage retrieve and processing security is also an important concern for data mining , data application from social web ,consumer oriented work has large impact on big data security according to author vast use of smart phones has increased photo uploading and other sensitive information on web it is an issue for that author has proposed metadata analysis in big data which creates an index of each images uploaded on social web and we can identify from link which gives confidentiality over social media, so each images can be scanned from big data bases of social media and can be apply for future security policies .

In **[4]** After consideringsecurity in analysis we again come with our problem of analysis the big data with this paper integration of NOSQL with big data analysis author proposed model of unity architecture for analysis of data as shown in figure 3.2

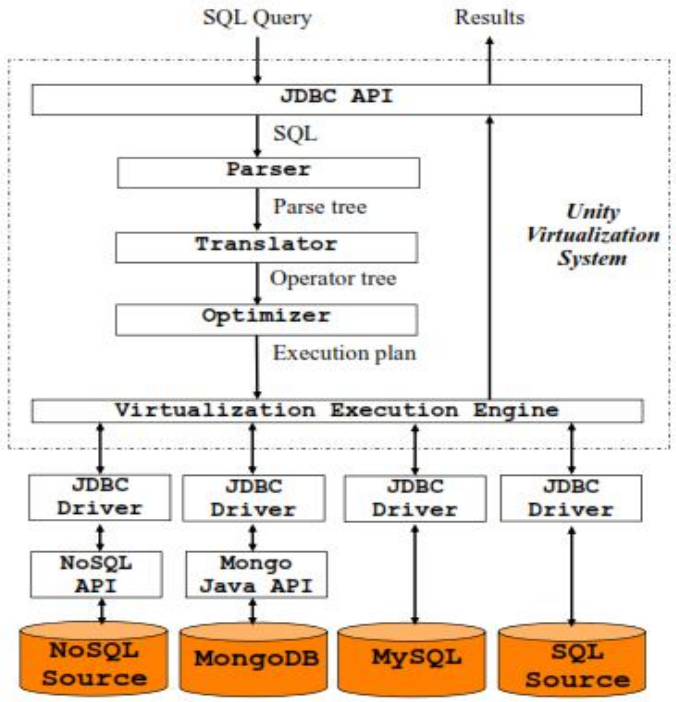

Figure 2.2 Unity Architecture

The objectives of this architecture is as follow

- SQL is a declarative language that allows descriptive queries while hiding implementation and query execution details.
- SQL is a standardized language allowing portability between systems and leveraging a massive existing knowledge base of database developers.
- Supporting SQL allows a NoSQL system to seamlessly interact with other enterprise systems that use SQL and JDBC/ODBC without requiring changes.

This system provides combination of both Relational data base system and Nosql system for this interaction we can translate one schema to another schema by JDBC API and Mongo dbconnector.

In [6] Mongo db and Oracle databases are compared by their storage method ,syntax and their retrieval methods also various experiments conducted with different query processing time and number of processing the results here we are discussing few results achieved with this research .

No. of records	Oracle Database	MongoDB
10	31	800
100	47	4
1000	1563	40
10000	8750	681
100000	83287	4350
1000000	882078	57871

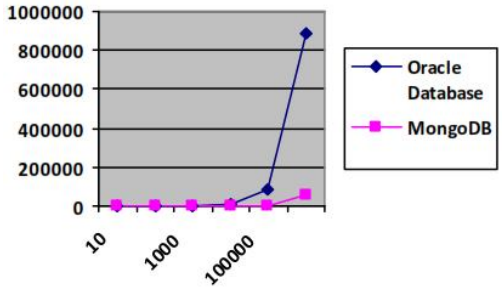

Figure 2.3. Insert query with Mongo db and Oracle[6]

As we can see for small records inserting oracle databases are faster then mongo db but as the size increases for records the mongo db is impressively ahead then Oracle database Same results are achieved with update query comparison

No. of records	Oracle Database	MongoDB
10	453	1
100	47	1
1000	47	1
10000	94	1
100000	1343	2
1000000	27782	3

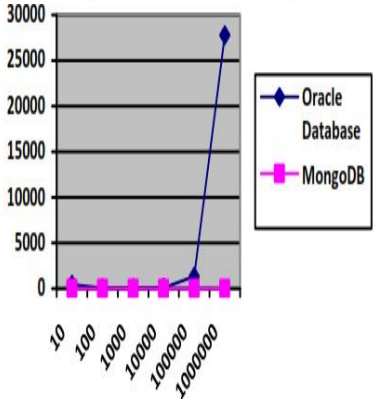

Figure 2.4 Update query on records and time comparison Mongo dbvs Oracle

From this we can conclude that mongodb is flexible and scalable for large data sets which provides batter integration for data storage and retrieval .

In [5] In this paper author has discussed about some very important parameters of mongo db focusing on *CAP* model and compared various types of data store available with no Nosql and tested them among various business intelligent system provided , and concluded that Nosql data stores provides huge opportunity where Sql data bases are not useful basic advantages are their scalability and cross node operation .the intersection algorithm for mongo db states the effectiveness of mongo db data store for key value approach to modelize the data.

In [7] [10] and [11] some practical approaches are shown to interact no sql data stores with various systems such as distributed architecture[7] ,Hashdoop[11] and evolution in hadoop[10] are proposed in distributed system data bases are handled by structure system but it fails when data items increase so unstructured data stores are useful for such problems some major industries are capable to develop their own unstructured data stores for ex. Google's Big table, Yahoo's PNUTS ,Hadoop'sHbase and many more but what about small industries ,author stated that there are many open source products are available to handle such data the comparison between them is shown in below figure.

	Indexing	Caching	SQL-Like	Joins	Aggregations
Hive	✓	X	✓	✓	✓
Pig	X	X	X	✓	✓
Redis	✓	✓	✓	X	✓
Hypertable	X*	✓	X	X	X
Project Voldmart	✓	✓	X	X	X
Risk Core+ Search	✓	✓	X	X*	✓
HSearch	✓	✓	X	X	X
HBase	X*	✓	X	X	X
Lucene	✓	X	X	X	✓
Cassandra 0.75	✓	✓	X	X	X
Cassandra 1.1.6	✓	✓	✓	✓	✓
MongoDB	✓	✓	✓	X*	✓

Figure 2.5 Comparison of Difference Data stores [7]

Among all this data stores mongo db is better replacement for MySql as it is semi structured, and provides batter joins contains laser time for searching and in performing other queries.

In paper [10] multiple Nosql data stores are compared and we can see that mongo db provides consistency, partition tolerance and crash handling over any other data stores But in this paper author has limited computation power this system can be improve by adding some more computation power over large datasets by cloud computing or distributor

approach .[11] is an example of hadoop hash function for anomaly detection using map reduce programming model .the hashdoop framework splits the traffic using has functions and the detector detects the anomaly from carious hadoop clusters then the traffic has been divided in less traffic lines however author has not applied to store the data back to original data sets which will be lost vice versa.

III. BACKGROUND STUDY

3.1 Big Data

3.1.1 Architecture

Big data is a distributed architecture for storing large amount of data ,According to a research recently the online data has increased in size CERN research says that data without operating online. For example, "will produce roughly 15 peta bytes (15 million gigabytes) of data annually – enough to fill more than 1.7 million dual-layer DVDs a year!" [11]

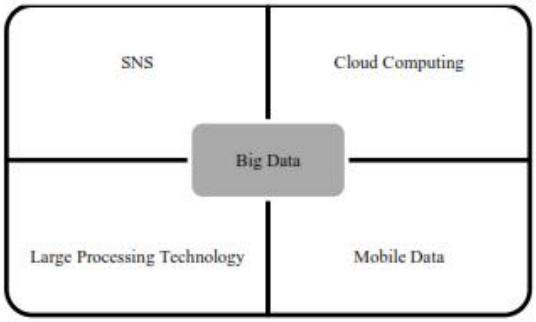

Figure 3.1.1. Bigdata[9]

Bigdata architecture consists following three segments

- Storage System
- Processing
- Analysis

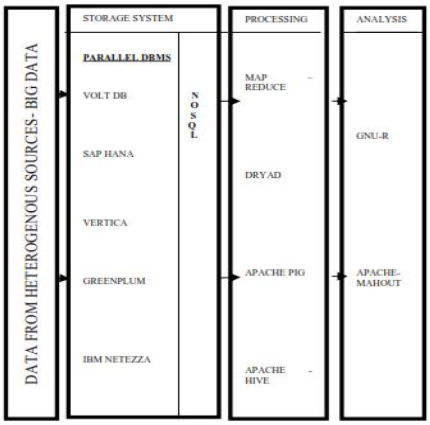

Figure 3.1.2BigData system [9]

3.2 What is NOSQL?

No SQL means an alternative oftraditional database system the term was generated ny a scientist named Eric Evans in Sanfransisco. The nosql databases have variety of different database systems and they provides the data manipulation as well as low time in reading and writing

Many large organizations have their own NoSql Databases as BIG Table in Google which has much effect on the no sql The whole point is that they provides alternatives to the traditional databases product. For Example the many nosql products are available in the market and they are widely used by many companies.

3.2.1HBase

Hbase is very different because it is developed by and for hadoop HDFS archirecture , so any other products developed for hadoop can easily implemented by Hbase , Hbase has following three types :

(1) Standalone Installation: Basic installation for client to connect with clusters.
(2) Pseudo Installation: basic distributed installation single server multiple workstations
(3) Fully Distributed installation which is used for all application based on hadoop multiple deployment environment.

3.2.1.1:Data Model

Hbase is distributed file system that scale to hundreds or thousands of node. HDFS is good for the batch processing, not good for record lookup and not good for updates or small batches, Hbase is designed for Fast Record Lookup, support for record-level insertion and support for updates.

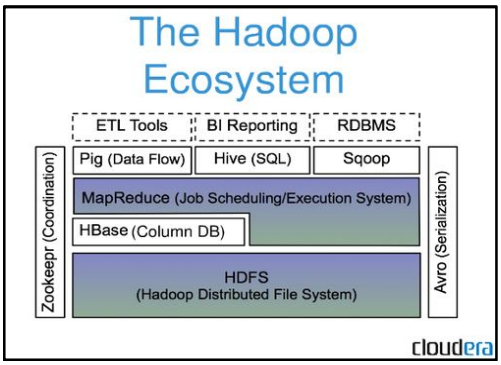

Fig:2.3 HBase System based on google'sBigTable[14]

3.3Apache Hadoop

Apache Hadoop is java based programming framework which is used for processing large data sets in

distributed computer environment. Hadoop is used in system where multiple nodes are present which can process terabytes of data hadoop uses its own file system HDFS which facilitates fast transfer of data which can sustain node failure and avoid system failure as whole.[1]

3.3.1 Architecutre

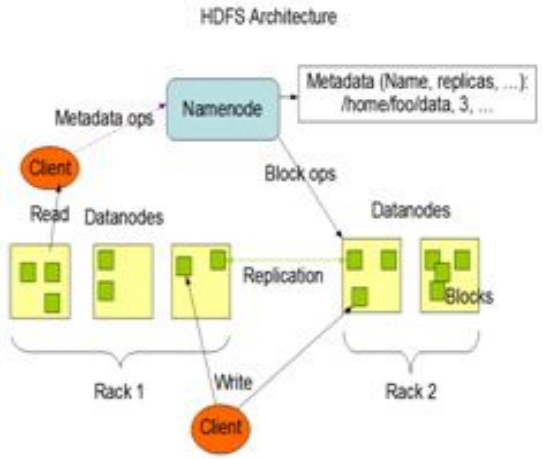

Figure 3.3.1 HDFS ARCHITECTURE [14]

3.4 Map Reduce Algorithm and Approaches

Map/Reduce is a programming paradigm that was made popular by Google where a task is divided into small portions and distributed to a large number of nodes for processing (map), and the results are then summarized into the final answer (reduce). Hadoop also uses Map/Reduce for data processing. Hence different functions for the processing are written in the form of Hadoop job. A Hadoop job consists of mapper and reducer functions framework i.e. STS is used as the integrated development environment (IDE) [1].

3.4.1 Map reduce with Hadoop

A Many company uses Hadoop for big data analysis, Forexample facebook use HIVE with Hadoop [1]

B Yelp: uses AWS and Hadoop Yelp uses Amazon S3 to store daily logs and photos, generating around 100GB of logs per day. The company also uses Amazon Elastic Map Reduce to power approximately 20 separate batch scripts, most of those processing the logs.

Features powered by Amazon ElasticMapReduce include: [1]
1. People Who Viewed this Also Viewed
2. Review highlights
3. Auto complete as you type on search
4. Search spelling suggestions
5. Top searches

IV. CONCLUSION

BigData Mining is an emerging trend for new researcher, as well as Mining knowledge from bigdata is useful for many areas like medical, Engineering, Government, In our research we will represent the novel approach for Big data mining using Hadoop and Map reduce with NoSql Storage model and represent the efficiency, and space complexity for the same , The secondary sorting algorithm will arrange the data with any key, value pair mined with hadoop

REFERENCES

[1] JyotiNandimath ,AnkurPatil , Ekata Banerjee , PratimaKakade :*"Big Data Analysis using Apache Hadoop " In SKNCOE Pune India,2013*

[2] E. Dede, M. Govindaraju ,D. Gunter, R. Canon, L. Ramakrishnan"Performance Evaluation of a MongoDB and HadoopPlatform for Scientific Data Analysis" In Lawrence Berekely National Lab Berkeley, CA 94720

[3] Matthew Smith, Christian Szongott,BenjaminHenne, Gabriele von Voigt" *Big Data Privacy Issues in Public Social Media",2013*

[4] Ramon Lawrence :"*Integration and Virtualization of Relational SQL and NoSQL Systems including MySQL and MongoDB"At*2014 International Conference on Computational Science and Computational Intelligence,2014

[5] Laurent Bonne, Anne Laurent, Michel Sala, Benedicte Laurent,NicolasSicard:"*REDUCE, YOU SAY: What NoSQL can do for Data Aggregation and BI in Large Repositories* "In 2011 22nd International Workshop on Database and Expert Systems Applications,2011

[6] AlexandruBoicea, Florin Radulescu, Laura IoanaAgapin" Mongo DBvs Oracle – database comparison **"**2012 Third International Conference on Emerging Intelligent Data and Web Technologies,2012

[7] Suyog S. Nyati, ShivanandPawar ,Rajesh Ingle: " Performance Evaluation of Unstructured NoSQL data over distributed framework" 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI),2013

[8] LiorOkman,Nurit Gal-Oz,, YaronGonen,, Ehud Gudes ,Jenny Abramov:" Security Issues in NoSQL Databases"

2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11,2011

[9] Chanchal Yadav, Shuliang Wang , Manoj Kumar: Algorithm and approaches to handle large Data- A Survey. IJCSN International Journal of Computer Science and Network, Vol 2, Issue 3, 2013

[10] Ruxandra Burtica, Eleonora Maria Mocanu, Mugurel Ionut Andreica, Nicolae Tapus: Practical application and evaluation of no-SQL databases in Cloud Computing , ©2012 IEEE

[11] Romain Fontugne, Johan Mazel, Kensuke Fukuda Hashdoop: A Map Reduce Framework for Network Anomaly Detection CERN – European organization for nuclear Research 2014 IEEE INFOCOM Workshops: 2014 IEEE INFOCOM Workshop on Security and Privacy in Big Data,2014

WEBREFERENCES

[12] Bigdata-Wikipedia :http://en.wikipedia.org/wiki/Bigdata

[13] Big Data Characteristics :http://en.wikipedia.org/wiki/ Big_data#Characteristics

[14] Hadoop Architecture:http://hadoop.apache.org/docs/ r1.2.1/hdfs_design.html