Improving Efficiency and Time Complexity of Big Data Mining using Apache Hadoop with HBase storage model .

Amita Shah[1] and Prof. Krunal Panchal[2]

1Computer Engineering Department, LJIET, Ahmedabad, Gujarat, India
2 PG Co-coordinator, Computer Engineering Department,LJIET, Ahmedabad,India

ABSTRACT

Data Mining is the science of mining the knowledge from the raw data and applying to improvement of the industrial rules. Now for the mining of " big data " we required new approach new algorithm and new techniques and analytics to mining the knowledge from it. Day by day a huge amount of data is generated and the usage is expanding .The term BIGDATA is a popular term which used to describe the data which is in zeta byte. Bigdata is large amount of data .The internet social sites and many industries also government are generating large data and it requires healthy storage too, so it is required that we mine knowledge from it and store it for future use. To address this, I aimed at training the MapReduce programming model of Hadoop framework and Hbase, since the MapReduce programming model has the ability to rapidly process large amount of data in parallel. MapReduce works in tandem with Hadoop Distributed File System (HDFS). Hbase built on top of Hadoop/HDFS and the Data stored in Hbase can be manipulated using Hadoop's MapReduce capabilities with In-Built features such as scalability, Versioning, Compression and garbage Collection.

Keywords :- *BigData, MapReduce,Hadoop/HDFS,Hbase*

1. INTRODUCTION

Data Mining is the science of mining the knowledge from the raw data and applying to improvement of the industrial rules. Now for the mining of " big data " we required new approach new algorithm and new techniques and analytics to mining the knowledge from it.

Day by day a huge amount of data is generated and the usage is expanding .The term BIGDATA is a popular term which used to describe the data which is in zeta byte[1]. Bigdata is large amount of data .The internet social sites and many industries also government are generating large data and it requires healthy storage too, so it is required that we mine knowledge from it and store it for future use.

For that we have found that Apache has developed new technology named Hadoop, we can describe the characteristics of big data in three category Volume,Velosity,Veriety [13][14]. Hadoop is well-known technology part of apache , it contains the software and library packages for the distributed environment which is used for data mining the distributed architecture supports processing of the large datasets by creating clusters of computers using simple programming tools , It is combination of HDFS (as known as NameNode) and Algorithm MapReduce.

NOSQL[3][7] means "Not Only Sql" It is an alternative for relational sql data sets . traditionalsql data sets are not enough for bigdata so they provide solution for that . example :HBase,Cassandra,MongoDb,Voldmort etc..

Apache-Hadoop -HBase have a diverse and growing user community because scale out approach using commodity hardware , Advantage of this architecture is reduction of cost , when the load increases per number of user. NOSQL data bases are scalable fault tolerance ,reliable. distributed data bases operates on a cluster commodity model .

In our world day by day data is expanding in drastic manner do it is necessary to process such data and generate knowledge from it ,to process large amount of data there are many systems and approaches available , large data cannot be processed in single machine ,our relational data bases schemes are not efficient and scalable for processing large data , also there are chances to data loss and security , In this I will use hadoop framework (HDFS)

and NOSQL database (Hbasedb) with Map reduce programming model of Hadoop framework I will try to minimize query processing time with sorted or double sorted data to process large scale datasets from different data shards.

2. ORGANIZATION OF PAPER

The organization of the paper further is as follows. The Comparative study is presented in Section III, Proposed Method in Section IV, then Result Analysis in Section V and Conclusions and Future Scope discussed in Section VI.

3. COMPARATIVE STUDY

No	Name	**HBase**	**MongoDB**
1	Description	Wide-column store based on Apache Hadoop and on concepts of BigTable	One of the most popular document stores
2	Database model	Wide column store	Document store
3	Developer	Apache Software Foundation	MongoDB, Inc
4	Initial release	2008	2009
5	Current release	1.1.4, March 2016	3.2.6, March 2016
6	License	Open Source	Open Source
7	Database as a Service (DBaaS)	No	No
8	Implementation language	Java	C++
9	Server operating systems	Linux Unix Windows	Linux OS X Solaris Windows
10	Partitioning methods	Sharding	Sharding
11	Replication methods	selectable replication factor	Master-slave replication
12	MapReduce	Yes	yes
13	User concepts	Access Control Lists (ACL)	Access rights for users and roles

Table1: The overall comparison analysis of MongoDB and Hbase

4. PROPOSED METHOD

As we discussed earlier to analyze big data we required Hadoop HDFS system ,a database i.e. Hbasedb

The following steps we will consider to

(1) Import data from Hbase to hadoop
First we will create columnar database stored in a key value pair and the data will be fetch into system It is convenience to fetch the data from hbase because it stores data in column format and can be easily parse to the hadoop using JSON script. We can also create a java script to maintain the autonomous batch system for input.

(2) Processing Data in hadoop
Map Reduce algorithm provides us two steps to process the data (map), and the answer is reduced data sets following functions and algorithms are used in hadoop Mapper and reduce[1]

(3) Store the data back to data base
The output will be in key value format , and we have to store it back to avoid data loss the key and value will considered the answer the key and value pair will be extracted from database via hadoopHBase connecter and stored again in database also multiple jobs can be run by creating multiple node on hadoop.

(4) analyze the different pattern
Finally we will analyze the different pattern generated by hadoop for bigdata and reduced data can be use for generate new rule for organization.

PREPROCESS: Create Data base into Hbase create Master Node and ZooKeeper.

INPUT: input Key value Data set
from data connector of HBase
Function : Mapper() Reducer()
MAPPER()
I : start hadoop node
II:StartNamenode and HDFS
III :impor t data from database
document
IV :call Mapper()
REDUCE()
I: trace the mapper output
II: aggregate the output received
from different mapper
III: generate log
OUTPUT:
Save the Data in HBase, Mine the results Efficiency Scalability And Pattern Evolution

This is An improvement on the basic algorithm(4.1)as shown in Algorithm 4.2 (the mapper is modified but the reducer remains the same as and therefore is not repeated). An associative array (i.e.,Key value) is introduced inside the mapper to tally up term counts within a single document instead of emitting a key-value pair for each term in the document.

Problem: Map reduce uses key value pair for analyzing multiple values from data base but for large database the data associated with key might be change we will try to sort the data using Hadoop and analyze time for map reduce algorithm ,also if primary sorting is not enough we will perform secondary sorting which guarantees the sorted order,

Hadoop does not have this facility in built and for that if we try to analyze large data the reducer will run out of memory to keep buffer the map values there is only one solution for these problem that we can convert the map data into key value format, for example (a,b)→(r1001),(c,d)→(r1002), here HBasedb comes in a picture.

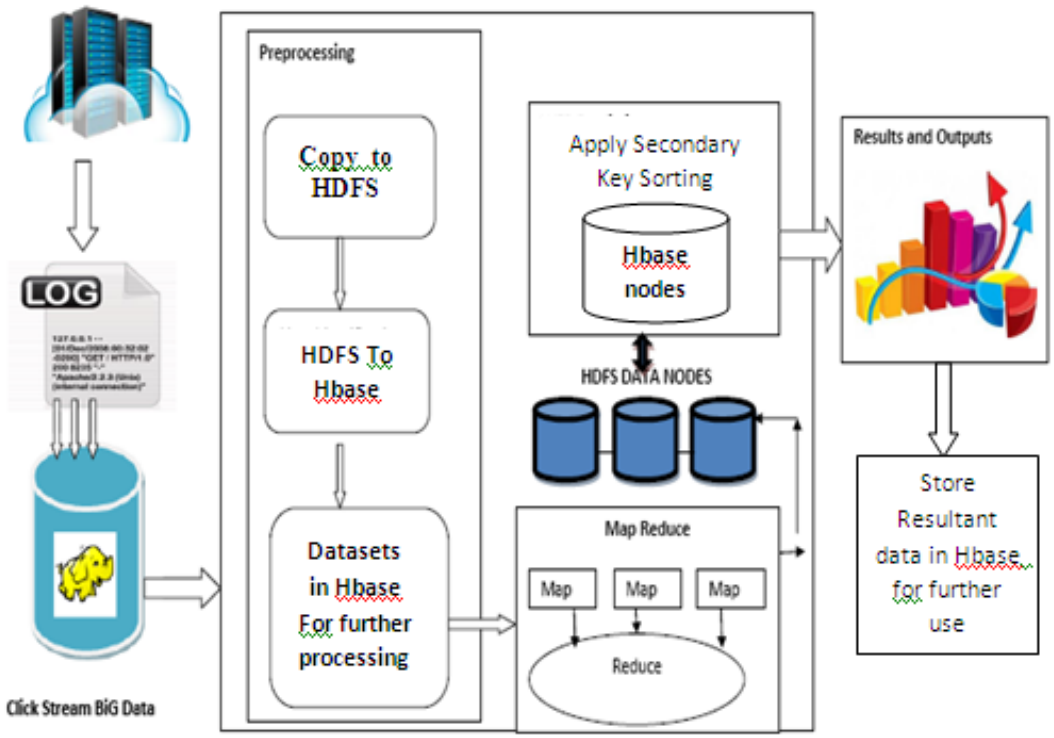


Fig -1: Flow diagram of proposed work

5. RESULT ANALYSIS

We have performed Secondary key sorting on different Number of record and we can see that for more record it will take less time to execute but it is up to some threshold value then time will be increased when number of records increases. so it will work more efficiently for large dataset.
The effect of increasing records is shown in Table 1

Number of Records	Time(in minute)
50,000	3.59
1,00,000	1.30
1,50,000	6.25

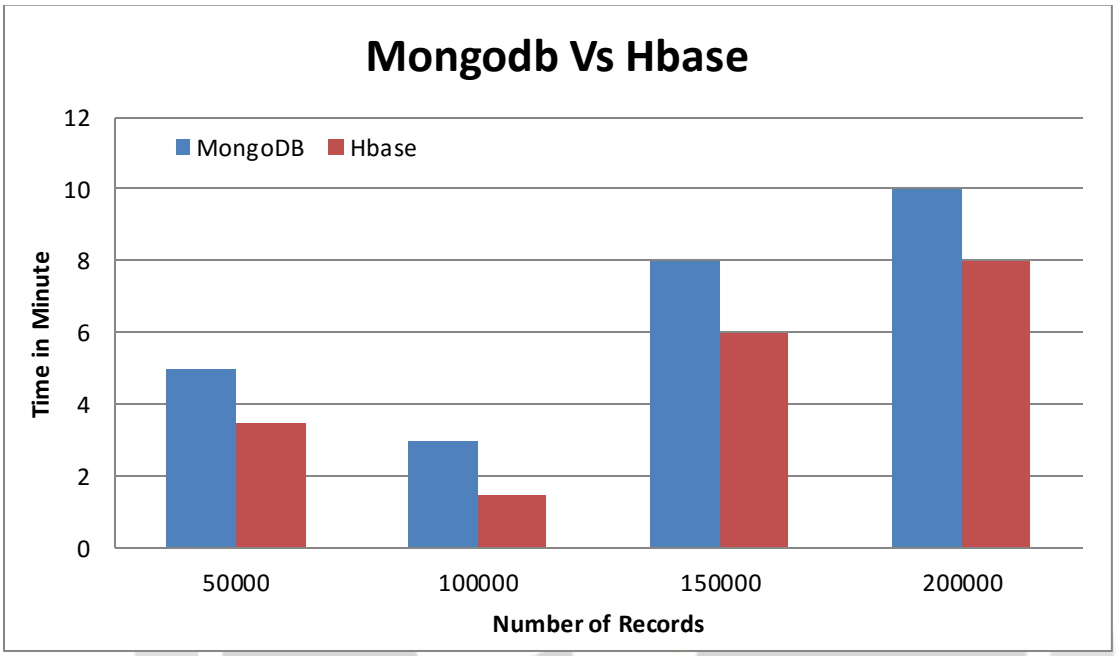
Table 2:- Elapsed Time to evaluate different number of records in Dataset

From the above results it is proved that processing of large Dataset in Hbase will take less time to execute.We can also use the MongoDb instead of Hbase but it takes more time than processing the same file in Hadoop Mapreduce.. This function is capable to do the processing in less amount of time when compared to MongoDb.

Number of Records	MongoDB(Time in S)	Hbase(Time in S)
50,000	312	239
1,00,000	201	90
1,50,000	490	385

Table 3:-Comparative results

From Table 2 it is observed that the time to execute dataset in both the environment shows a difference of time in seconds. So we can say that we can improve Time complexity of our work by using Hbase as a NoSql database for large Dataset.



6. CONCLUSIONS & FUTURE SCOPE

In therapeutic sciences, picture handling has enabled for exact and speedy quantitative examination and 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 Bigdatamining using Hadoop and Mapreduce with NoSql Storage model(Hbase) and represent the efficiency,and Time complexity for the same , The secondary sorting algorithm will arrange the data with any key,value pair mined with hadoop . This Experiments provides important insights and improvement for the data analysis using hadoop and hbase

In the future we will try to implement this environment on multiple servers to find the time complexity of hadoop – Hbase, also we can measure failure rate of the system as it improves over a time because as we increase parallel task the map can fail but if we create a multiple node cluster, it can achieve Fault tolerance compare to single node cluster .

7. REFERENCES

1. Jyoti Nandimath , Ankur Patil , Ekata Banerjee , Pratima Kakade ,"Big Data Analysis using Apache Hadoop " In SKNCOE Pune India,2013,ISBN:978-1-4799-1050-2.
2. Mehul Nalin Vora (TCS),"Hadoop-Hbase for Large-Scale Data" In IEEE 2011 International Conference On Computer Science And Network Technology,ISBN:978-1-4577-1587-7.
3. 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
4. Matthew Smith, Christian Szongott,Benjamin Henne, Gabriele von Voigt" Big Data Privacy Issues in Public Social Media",2013,ISBN:978-1-4673-0

5. Laurent Bonne, Anne Laurent, Michel Sala,Benedicte Laurent,Nicolas Sicard:" 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,DOI 10.1109/DEXA.2011.71
6. Alexandru Boicea, Florin Radulescu, Laura Ioana Agapin" MongoDB vs Oracle - database comparison "2012 Third International Conference on Emerging Intelligent Data and Web Technologies,2012,DOI 10.1109/EIDWT.2012.32
7. Suyog S. Nyati, Shivanand Pawar ,Rajesh Ingle: " Performance Evaluation of Unstructured NoSQL data over distributed framework" 2013 International Conference on Advances in Computing, Communications andInformatics (ICACCI),2013,ISBN:978-1-4673-6217-7.
8. Lior Okman,Nurit Gal-Oz,, Yaron Gonen,, Ehud Gudes ,Jenny Abramov:" Security Issues in NoSQL Databases" 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11,2011,DOI 10.1109/TrustCom.2011.70.
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,DOI 10.1109/CSCI.2014.56.
10. Ruxandra Burtica, Eleonora Maria Mocanu, Mugurel Ionut Andreica,
Nicolae Tapus: Practical application and evaluation of no-SQL databases in Cloud Computing , ©2012 IEEE, ISBN:978-1-4673-6317-2.
11. Romain Fontugne,,Johan Mazel, Kensuke FukudaHashdoop: A MapReduce Framework for NetworkAnomaly DetectionCERN – European organization for nuclear Research 2014 IEEE INFOCOM Workshops: 2014 IEEE INFOCOM Workshop on Security and Privacy in Big Data,2014, ISBN:978-1-4600-4562-5.
12. Big Data Characteristics : http://en.wikipedia.org/wiki/Big_data#Characteristics 12/10/2015,11:10:55
13.Hadoop Architecture, Available at: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html 18/09/2015,1:11:17
14. Hbase Architecture, Available at: http://netwovenblogs.com/2013/10/10/hbase-overview-of-architecture-and-data-model/ 10/11/2015,20:08:05
15. System Properties Comparison HBase vs. MongoDB,: Available at http://db-engines.com/en/system/HBase%3BMongoDB
16. By Jiawei Han And Micheline Kamber,Data Mining Concept and Techniques, Copyright 2006, Second Edition,pp 5-9 ,119-145.