A Text Information Retrieval Technique for Big Data Using Map Reduce

M.M. Kodabagi, Deepa Sarashetti and Vilas Naik

Abstract--- *Big data is a heterogeneous collection of both structured and unstructured data of larger volumes. The volume and the heterogeneity of data and the speed at which it is generated, makes it difficult for the present computing infrastructure to manage Big Data .The traditional data management, warehousing and analysis systems fall short of tools for analyzing enormous data.*

The proposed methodology is the information retrieval mechanism to improve text information retrieval using map reduce technique. This reorganization produces significant gains in performance by reducing the number of accesses made to the data file. It is examined the impact of other factors on text retrieval is also experiment in this work namely Heterogeneity, Scale, Data Representation and Complexity. A major motivation for reorganizing the structured data retrieval is to allow the application of iteration aware perfecting.

Keyword--- *Big Data Analysis, Big Data Management, Text Retrieval, Map Reduce, HDFS.*

I. INTRODUCTION

WITH increased processing power and huge storage space available at an affordable price, the size of scientific data sets has grown to the terabyte and peta byte scale. Efficiently storing and retrieving large data volumes is an important goal for the scientific data community. The overall size of the data often necessitates dividing the data across multiple disks on a single machine.

Big data refers to datasets whose size is beyond the ability of typical database management tools to capture, store, manage, and analyze. The traditional database system and data mining tools are not able to handle gigantic data at a time. Big data helps to handle this issue. Big Data illustrates an information analysis policy that consists and mixes various styles of data and data administration. Map Reduce is the heartbeat of Hadoop framework. It is a programming paradigm that allows for huge scalability across. It is a programming model that is associated with the implementation of processing and generating large data sets. The term Map Reduce basically refers to two separate and distinct tasks that Hadoop programs perform. The main purpose of this work is to develop a technique for a fast and efficient way of searching data using Map Reduce paradigm of the Hadoop Distributed File System. Hadoop is a framework which is specifically designed to process and handle vast amounts of data. Fig 1 shows the components of big data.

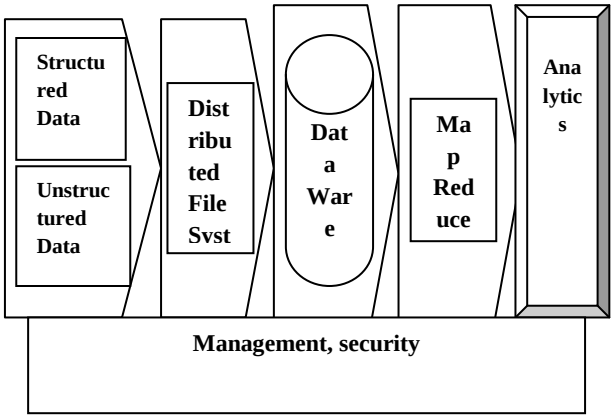

Figure 1: Big Data Components

Big Data is and normally laying on four essential magnitudes.

1. Volume: The quantity of data which is increasing immensely each day. Now days it is peta bytes but in prospect it will become zetta byte.
2. Velocity: how fast data is accessing or we are able to declare the speed of data exchange on or after different ways. Velocity means the speed and performance of Big data.
3. Variety: structured, unstructured and semi-structured data
4. Veracity: Data that we have from a variety of resources must be accurate.
5. Complexity: tremendous quantity of data we include in a variety of layouts such as structured, amorphous and semi structured.

The methodology proposed in this work is a mechanism for information retrieval technique from big data using map reduce technique. The performance evaluation of given approach is calculated using different factors. Organization of reaming part of the paper is organized into following four sections. Literature review section describe the proposed techniques for information retrieval from big data also the issues and challenges of big data with analytics tools. The next section is talks about the proposed methodology for given problem. After that describe about the experimentation and results of given solution .the last part give the conclusion with future scope.

Dr.M.M. Kodabagi, Professor, School of Computer Science and Information Technology, Reva university, Banglore.

Deepa Sarashetti, M.tech(QIP)student, Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot. E-mail: dsarashetti@gmail.com

Vilas Naik, Associated professor, Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot. E-mail: vilasnaik_h@rediffmail.com

DOI:10.9756/BIJSESC.8236

II. LITERATURE REVIEW

There are quite a few techniques have been proposed for data retrieval from big data among them are discussed here.

Table 1: Literature Review

Sl no	Author and Title	Survey
1.	Sonal Kasare, Anup Bongale (2013)" Efficiently Searching Nearest Neighbor In Documents Using Keywords"	Discusses new access method called the spatial inverted index that extends the conventional inverted index to cope with multidimensional data, and comes with algorithms that can answer nearest neighbor queries with keywords in real time.
2.	Silvia Bolohan, Sebastian Ciobanu(2013)" From Big Data to Meaningful Information with SAS® High-Performance Analytics"	Described Big data technologies – such as grid computing, in-database analytics and in memory analytics.
3.	C.Vidhya,(2014) " An analysis of Big Data usage in Big Organization"	Analysis done with researching the importance of Big Data usage in Business analytics for big organization.
4.	Tharini.N , kiruthika.K, (2015) "Towards the significant discovery of large valued Datasets Moving forward"	Highlights over the transition done from traditional RDBMS to big data. And examine big data issues and challenges.
5.	Takshak Desai, Udit Deshmukh , Prof. Kiran Bhowmick,(2015) " Machine Learning for Classification of Imbalanced Big Data"	Discussed about the existing algorithms used for classification, analyses of big data and observes the practicality of algorithms.
6.	Jaseena K.U. and Julie M. David,(2014) " Issues, challenges, and solutions: big data mining"	Describes the data analysis techniques such as Map Reduce above Hadoop with HDFS.
7.	Puneet Singh Duggal, Sanchita Paul , (2013) "Big Data Analysis: Challenges and Solutions"	Described Reduce framework above Hadoop Distributed File System (HDFS)
8.	Nikita.V.Shahane, Rutuja Pande , S.R. Vispute, (2015)" Survey of Big Data Analysis Using Predictive Analytics Algorithms and Its Use Cases"	Described K-Means clustering algorithm based on Map Reduce.
9. 5555	Shital Suryawanshi, Prof. V.S.Wadne , (2013)" Big Data Mining using Map Reduce: A Survey Paper"	Describe the Map Reduce structure used for efficient analysis of huge data
10.	Chanchal Yadav, Shuliang Wang, Manoj Kumar , (2013) "Algorithm and approaches to handle large Data-A Survey"	Described distributed technologies with various algorithms exercised to hold such bulky information.

A. *Issues and Challenges*

There are the some issues that are connected to information retrieval using big data and these are

Heterogeneity: The problems of big data analysis arise from its large scale as well as the presence of mixed data based on different types of collected and stored data analysis is a major challenge in big data mining.

Scale: Managing large and rapidly increasing volumes of data is a challenging issue. Traditional Software tools are not enough for managing the increasing volumes of data.

Complexity: Normally, there is a large number of data in raw datasets. Data organization, retrieval and modeling are also challenges due to scalability and complexity of data

Data Representation: A lot of records are diverse in type, organization, meaning, subject wise, as well as ease of understanding. Skilled data arrangement should be planned the arrangement, pecking order, and variety of information.

B. *Big Data Analysis*

Big data analysis explains the easy steps for bulky quantity information examination .The process employed toward find out data from various resource, transform this for investigative requirements, and load it in information store house meant for subsequent investigation, process recognized as "Extract, Transform & Load" (ETL). Fig 2:shows Working of ETL process.

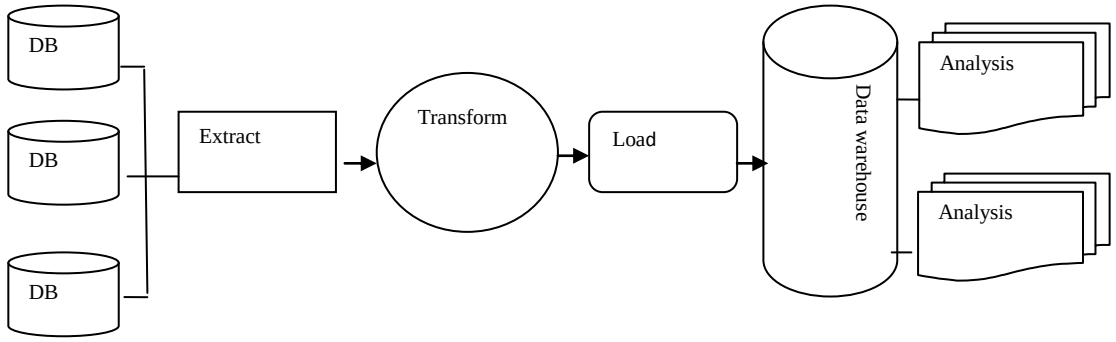

Figure 2: Working of ETL

Big data analytics perform the analysis in following areas

1. Text analytics: Text analytics (text mining) is a technique that extracts information from textual data. Text analytics involve statistical analysis.
2. Audio diagnostics: Audio analytics evaluate also extract information from unstructured audio data. Audio diagnostics is as well denoted to while speech diagnostics is applied to human verbal communication.
3. Video diagnostics: it occupies a multiplicity of methods to observe, examine, and fetch significant data from video flows.
4. Predictive analytics: Predictive analytics encompass a variety of techniques that predict future outcomes based on historical and current data.

The proposed methodology is related to text analytics area where given text is searched in big data using following methodology.

III. PROPOSED METHODOLOGY

This work is related to text information retrieval from big data using map reduce technique .following figure 3 describe the solution for given work.

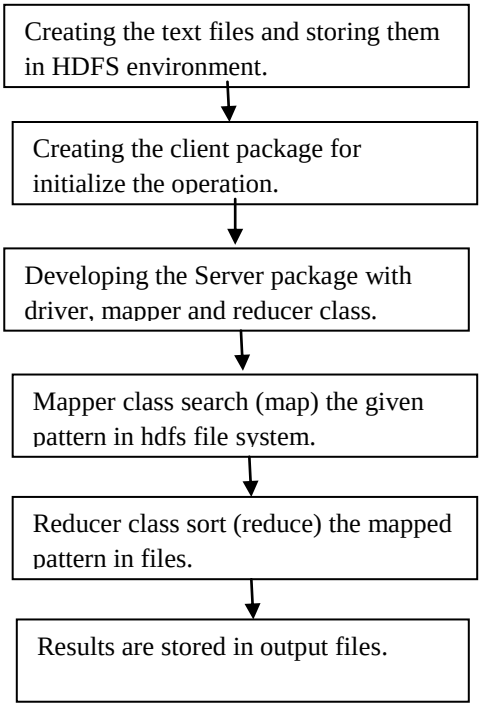

Figure 3:Working Solution

The work focused towards on retrieval of data from bigger size collection data using big data analysis tools. The major issue of big data is scale and complexity of huge information.

In proposed method text data is collected processed and retrieved using the techniques of map reduce to draw meaning full insights of the large data sets. Project takes the pattern to be searched and the pattern of text is examined for the occurrences in different text files and finish with the pattern occurrence count to visualized for superior understanding.

This framework works with client server principle.

1. Server section is implementing three different classes under one package. First is driver class which initiate the mapper function and reducer function which is written in the two different classes as Mapper class and Reducer Class respectively.
2. In this phase extract the useful information from dataset and calculate average occurrence of each tuple.
3. At this step Computation time for next job will be less time consuming as it has to deal with smaller data and time to extract data from tuple is reduced.
4. In last phase implemented the client class to initiate the server functions with the help of driver class.
5. Hadoop (High Availability Distributed Object Oriented Platform)Hadoop Architecture is divided into 2 core layers. One is a framework written in java to allow the system to store the various forms of data HDFS. the other is the programming engine of Hadoop which gives control to the user to access the data and perform analysis on it, which is treated as Map Reduce .

MAP REDUCE **:** Map Reduce is a programming model. The main idea Of the Map Reduce model is to hide details of parallel execution and allow users to focus only on data processing strategies. The Map Reduce model consists of two primitive functions: Map and Reduce. The input for Map Reduce is a list of (key1, value1) pairs and Map () is applied to each pair to compute intermediate key-value pairs, (key2, value2). The intermediate key-value pairs are then grouped together on the key equality basis, i.e. (key2, list (value2)).

Users can define the Map () and Reduce () functions however they want the Map Reduce framework works. Map Reduce utilizes the Google File System (GFS) as an underlying storage layer to read input and store output.GFS is a chunk-based distributed file system that supports fault-tolerance by data partitioning and replication. In Hadoop is broken down to as many Map tasks as input data blocks and one or more Reducer tasks. A single Map Reduce (MR) job is performed in two phases: Map and Reduce stages.

1. Mapper Stage: Before starting the Map task, an input file is loaded on the distributed file system. At loading, the file is partitioned into multiple data block have the same size, typically 64MB.Each block is then assigned to a mapper, Map () to each record in the data block. The intermediate outputs produced by the mappers are then sorted locally for grouping key-value pairs. Combine () is optionally applied to performer-aggregation on the grouped key-value pairs outputs to reducers is minimized.
2. The mapped outputs are stored in local disks of the mappers, partitioned into R.
3. Reducer stage: A reducer reads the intermediate results and merges them by the intermediate keys. This grouping is done by external merge-sort. The output of reducers is stored and triplicate in HDFS.
4. Proposed Approach: creating HDFS file system. Implementing mapper class. Implementing reducer

class. Implementing driver class. Implementing client class.

A. Mapper

1. Reads each transaction of input file and generates the data set of the items:
2. (<V1>, <V2>, …,<Vn>)
3. 2.Sort all data set <Vn> and generates sorted data set
4. <Un>: (<U1>, <U2>, …,<Un>)
5. Loop While <Un> has the next element.
6. End Loop While
7. Data set is created as input of Reducer: (key, <value>) = (P, <number of occurrences>).

B. Reducer

1. 1Study (P, <count of presence>) records from workstations
2. The reducer is collect the amount of values per key.
3. note: Vn= (key,value) pair , Un=sorted(Vn),p=pattern

C. Execution Outline

To build to problem taken up following steps. Collecting the data files in Varity of formats like text, pdf, ppt etc and storing in HDFS. The data is to be processed and retrieved using techniques of map reduce programming tool. Calculating performance analysis of retrieval system.

a) User Program

1. Begins with the user program.
2. Map Reduce libraries that are imported into the program are used in splitting operations.
3. X number of tasks and Y reduce operations to perform.

b) Map Workers

1. Map task takes the split input data and generates the key/value pair for each segment of input data.
2. Invokes the user-defined Map function.
3. Resultant values of the Map function are buffered in the memory.

c) Reduce Workers

1. Remote procedure calls to access from the Map workers.
2. Reduce read all the intermediate data, it groups together all the data of the same intermediate key.
3. different keys map to the same task
4. The Output of the reduce function is written to an output usually to a distributed file system.

d) Return to the User Program

After all Map and Reduce functions have been run. The Master sends control back to the user side.

IV. EXPRIMENTS AND PERFORMANCE

Performance of the system is estimated .For evaluating performance given solution can be considered the different criteria's. But check analysis as per the size of pattern and different size of inputs.

1. Pattern size: pattern is searching term which is differing from user to user. This is one topic which is considered for evaluating performance.
2. File size in hdfs: amount of content is differing from one file to another file. This is one topic which is considered for evaluating performance.

As the system considers text types file only so third observation is not considered. The performance as per the size of pattern is considered. For the performance checking here given different size inputs. To find evaluate performance as per the time. Size of inputs is very as one word, two words, three words; one sentence. Following chart gives performance analysis through graph. As per the time verses analysis is done .observe that it takes more time as per the pattern size. Following figure 4 and figure 5 depicts the performance. With reference to given observation value in following table.

pattern size	*time*
one word	0.7
two word	0.9
three word	1.2
sentence	1.6
two sentence	1.95

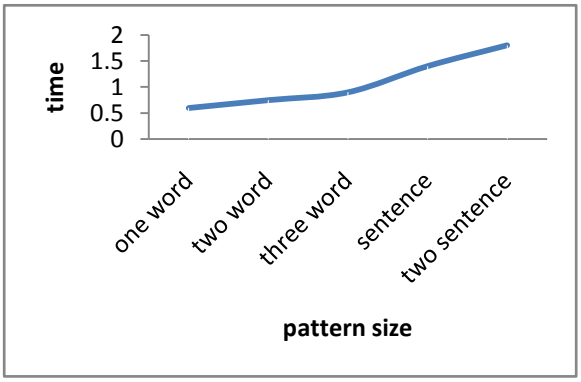

Figure 4: Performance Chart Pattern size vs. Time

Figure 5 depicts the performance. With reference to given observation value in following table.

Pattern	*Relevancy*
Freamwork	0.2
hadoop freamwork	0.34
mapreduce hadoop freamwork	0.57
hdfs and mapreduce hadoop freamwork	0.89

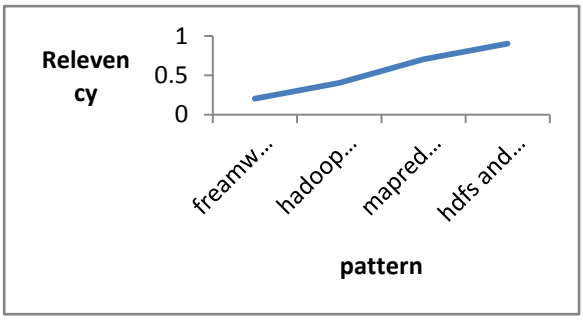

Figure 5: Performance Chart Pattern Size vs. Relevancy

V. CONCLUSION AND FUTURE SCOPE

The proposed work is a methodology for text retrieval from collection of heterogeneous text data from big data. The retrieval is accomplished through the support of map reduce

technique.In proposed system one machine is used with one job tracker .The search operation accepts one input string to search within all file systems. Because of this performance of algorithm is increased with respect to time and relevancies in search.

In future the work can be extent used in distributed system for retrieving the data from many data nodes.

REFERENCES

[1] C.C. Aggarwal, "An introduction to social network data analytics", In Social network data analytics, Pp. 1-15, 2011.

[2] G. Barbier and H. Liu, "Data mining in social media", In Social network data analytics, Pp. 327-352, 2011.

[3] H. Chen R.H. Chiang and V.C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact", MIS quarterly, Vol. 36, No. 4, Pp.1165-1188, 2012.

[4] J. Fan, F. Han and H. Liu, "Challenges of big data analysis", National science review, Vol.1, No.2, Pp.293-314, 2014.

[5] http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/

[6] P.P. Sharma and C.P. Navdeti, "Securing big data hadoop: a review of security issues", threats and solution. Int. J. Comput. Sci. Inf. Technol, Vol. 5, No. 2, Pp. 2126-2131, 2014.

[7] R. Gupta, S. Gupta and A. Singhal, "Big data: overview", arXiv preprint arXiv:1404.4136, 2014.

[8] C. Yadav, S. Wang and M. Kumar, "Algorithm and approaches to handle large Data-A Survey", arXiv preprint, Vol. 2, No. 3, Pp. 2277-5420, 2013.

[9] X. Wu, X. Zhu, G.Q. Wu and W. Ding, "Data mining with big data", IEEE transactions on knowledge and data engineering, Vol. 26, No. 1, Pp. 97-107, 2014.

[10] P.S. Duggal and S. Paul, "Big Data Analysis: Challenges and Solutions", In International Conference on Cloud, Big Data and Trust, Pp. 13-15, 2013.

[11] Apache Hadoop Project http://hadoop.apache.org/

[12] http://www.edupristine.com/courses/big-data-hadoop-program/bigdata-Hdoop-course/

[13] G. Shah, Annappa2 and K.C. Shet, "International Journal on Design An Efficient Big Data Analytic Architecture For Retrieval Of Data Based On Web Server In Cloud Environment", Vol. 2, 2014.

[14] P.P. Sharma and C.P. Navdeti, "Securing big data hadoop: a review of security issues, threats and solution", Vol. 5, No. 2, Pp. 2126-2131, 2014.