International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Digital Library Application Using Hadoop

Ms.Chandan Kaveri A[1]., Ms.Nagre Jayshree G[2].,Ms.Sainkar Varsha B[3].,Ms.Shinde Sandhya S[4].Prof.Khivsara B. A[5]
[1]*Department of Computer Engineering, S.N.J.B's KBJ COE, Chandwad,(kaverichandan@gmail.com)*
[2]*Department of Computer Engineering, S.N.J.B's KBJ COE, Chandwad, (jayshreenagare9158@gmail.com)*
[3]*Department of Computer Engineering, S.N.J.B's KBJ COE, Chandwad,(vsainkar@gmail.com)*
[4]*Department of Computer Engineering, S.N.J.B's KBJ COE, Chandwad, (sandhyashinde669@gmail.com)*
[5]*Department of Computer Engineering, S.N.J.B's KBJ COE, Chandwad, (bhavana.khivsara@gmail.com)*

Abstract -Due to rapid uses of social or network services such as Facebook, Gmail, Google, Twitter, etc, data is continuously growing. So traditional databases like DBMS, RDBMS etc are not sufficient to store Data which is continuously growing. Such type of Data is Big Data. To handle such type of Big data Hadoop is use. Applications that are Digital Library there are different kind of operations like Searching, analysis, issue, renew, registration, etc. To perform these different kinds of operations we use Hadoop Map reduce Technology.

Keyword- Hadoop, Map reduce, HDFS, Big Data.

I. INTRODUCTION

In the Past few year, data increasing rapidly such as like Facebook, twitter, Google, yahoo, web crawler, etc. Usage of data exceeds beyond Terabyte so to maintain such type of data the traditional databases are not sufficient to store data. So how to improve and maintain such type of data has become a major challenge. So Hadoop is use, the aim of Hadoop is used to store structured and unstructured large amount of data. Digital Library Application searching Data from large dataset, according to keyword like Author name, Book title, Content given by User.

1.1 Big Data

Big data is an all-Enwinding term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.[3] It is difficult to work with using most relational database management systems and desktop statistics, requiring instead massy parallel software running on thousands of servers. Big data varies depending on the capabilities of the organization, and on the capabilities of the applications that used to process and analyze the data set in its domain. It takes hundreds ofterabytes before data size becomes a significant consideration. Examples of Big Data are Google, Facebook, and Twitter etc.

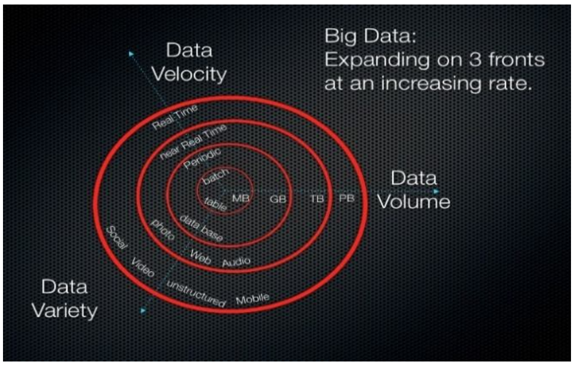

Fig.1: Big data[9]

Volume: - Many factors accommodate to the increase in data volume. Unstructured data streaming from social media. Data is collected from increasing amount of sensor and machines[8]
Velocity: - Data is streaming in at unusual speed and must be with in a timely. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time[8] **Variety :-** Today data comes in all types of formats. Structured, emails, numeric data, audios, unstructured data, videos, text document[8].

1.2 Hadoop

Hadoop is an open source framework. It allows to store an process big data efficiently. It also uses in a distributed environment[1]. Hadoop is designed to increase the performance from single server to thousands of server.
There are two major layers:
1)HDFS 2) Mapreduce
HDFS that is Hadoop Distributed File System. It is storage purpose. It has components namely

- Datanode(Slave)
- Namenode(Master)
- Secondary Namenode

Namenode consist the name of data. Datanode consist the Data. Secondary namenode is use for backup purpose whenever the namenode get fails then it retrieve data from secondary namenode.
Mapreduce is used to analysis large Dataset in map reduce data is written ones and read many times.It have two functions Mapper and Reducer. It has component namely

- Jobtracker(Master)
- Tasktracker(Slave)

Jobtracker keep the record of each job and assign job to the tasktracker. It also monitoring the jobs. Tasktracker run task or job which is assign by jobtracker and produce the output.
Mapper map the input coming from master and given as a output to the reducer. Reducers reduce that output which is coming from mapper and produce the final output.

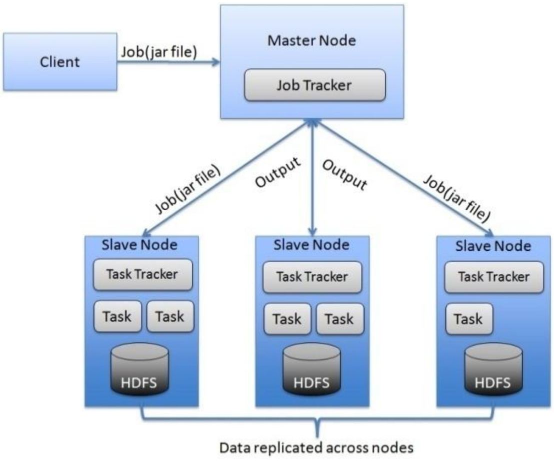

Fig .2: Hadoop Architecture [10]

II. LITERATURE SURVEY

2.1 DBMS

Database Management System(DBMS) is a software system designed is used to allow the definition, creation, updation of a databases. DBMS store only structural data.

Advantages	Disadvantages
• Improved data sharing. • Better data integration. • Improved data access.	• Disadvantages of DBMS are difficult ,complex to understand and time-consuming to design. • Information theft, loss, and storage.

2.2 RDBMS

Relational Database Management System (RDBMS). RDBMS data is structured in database .It include tables, fields and records.

Advantages	Disadvantages
• Data is only stored once. • Complex queries can be carried out. • Better security • Ease of use • Flexibility	• Performance • Physical Storage Consumption • Slow extraction of meaning from data • Data Complexity • Broken Keys and Records

2.3 Google Books

Google Books is a service provided from Google. These search the full text of books and magazines. Using optical character recognition Google convert scanned into text, and store in digital database. Google Books store all book according to its contents such as name of book, author of book, price, id .according to constraints it give result of search.[5] A click on a result from Google Books opens an interface in which the user may view pages from the book. For books where permission for a "preview" has been refused, only permission for two to three lines of text may be permission, but the limited basis book is searchable according to text. Neither a "full view" nor "preview" for other Books, the text is not search at all, and Google Books provides *no* identification of content transed of the title of book[6]. For those the site provides links to the website of the publisher and book sellers. Google books also search which keyword is present how many times and on which page it is available so that we can access it fast.

III. SYSTEM ARCHITECTURE

3.1 Digital Library:

Digital libraries provide a search interface which allows resources to be found. This is a concept of searching the book and the text file according to Keyword. Distributed searching typically involves a client sending multiple search requests in parallel to a number of servers.The results are collect, duplications are removed and/or clustered, and the remaining items are sort again and presented back to the client. Searching over previously harvested metadata involves searching a locally stored index of information that has previously been collected from the libraries. This application searching according to the Books from given Keyword. Keywords like Book_Title, Author_Name, Book_Id, Word .Whenever user enter the input in the form of keyword then this input send to the master node for processing. Then name node in master node save the input and create copy on secondary node for backup. Then Job tracker in Master node assign that keyword to slave node for additional processing. Then data node which is present in slave node save input keyword.. Task tracker in slave node use the MapReducer to perform the task which is assign by Master Node.

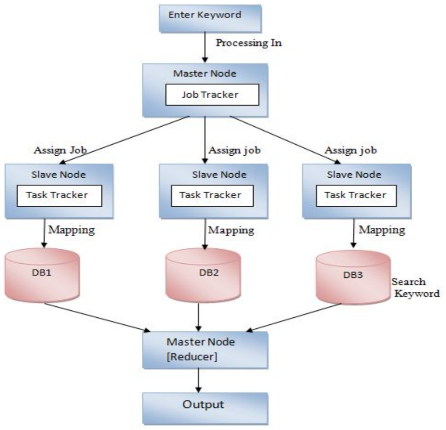

Fig .3: System Architecture

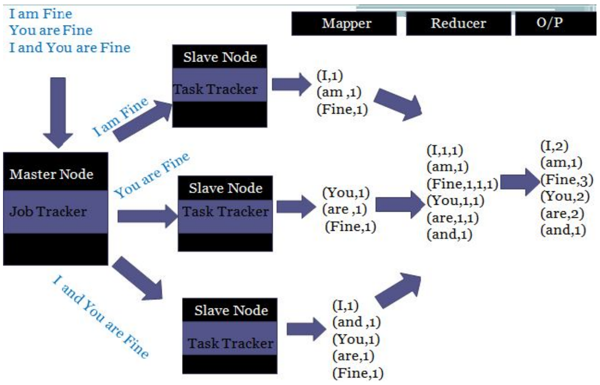

Fig.4: Ex. of Mapreduce-Word count

IV. ALGORITHM ANALYSIS

Proposed system falls under 'Divide And Conquer' method.

4.1 Problem Computational Analysis:

Proposed system falls into Deterministic. Deterministic has property that, result of every operation is uniquely defined.

4.2 Time Complexity:

O (n) =m (log n)   Where, m : number of slaves

V. MATHEMATICAL MODULE

S={I,O,P,C}

I={B,A,K}

O={S,U}
P={T,I,Ii,Se}

5.1 TF-IDF(Term frequency/Inverse Document frequency) ranking:

Let n(d) = number of terms in the document *d*
n(d, t) = number of occurrences of term *t* in the document *d*.
Relevance of a document *d* to a *term t*

$$TF(d,t) = \log\left(1 + \frac{n(d,t)}{n(d)}\right)$$

The log factor is to avoid excessive weight to frequent terms relevance of document to query*Q*

$$r(d,Q) = \sum_{t \in Q} \frac{TF(d,t)}{n(t)}$$

5.2 Indexing and Inverted Indexing:

An inverted index maps each keyword K_i to a set of documents S_i that contain the keyword. Documents identified by identifiers, Inverted index may record Keyword locations within document to allow proximity based ranking. Counts of number of occurrences of keyword to compute TFand operation: Finds **documents that contain all of K_1, K_2, ..., K_n**.Intersection $S_1 \cap S_2 \cap$..... $\cap Sn$ or operation: documents that contain at least one of K_1, K_2, …, K_n union, $S_1 \cap S_2 \cap$..... $\cap S_n$,.Each S_i is kept sorted to allow efficient intersection/union by merging "not" can also be efficiently implemented by merging of sorted lists

CONCLUSION

In the propose system to handle the Big data efficiently using Hadoop. System will help to find out result of searching in less time .The traditional databases DBMS,RDBMS have some drawbacks like complexity, difficult and time consuming to design. And also have less storage capacity. So the Digital Library Application find out the result of searching from the big data of digital library application using Hadoop.

REFERENCES

[1] Roman, Javi. "The Hadoop Ecosystem Table". github.com. Retrieved 2014-12-06.
[2] Ashlee Vance (2009-03-17). "Hadoop, a Free Software Program, Finds Uses Beyond Search". The New York Times. Archived from the original on 11 February 2010. Retrieved 2010-01-20.
[3] Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0 (Sebastopol CA: O'Reilly Media) (11).
[4] Layton, Julia. "Amazon Technology". Money.howstuffworks.com. Retrieved 2013-03-05
[5] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.
[5] Robert Darnton (February 12, 2009). The New York Review of Books.
[6] John Timmer (30 January 2010)."The sequel stinks: critics trash new Google Books settlement".Ars Technica.
[7] He, B.; Fang, W.; Luo, Q.; Govindaraju, N. K.; Wang, T. (2008). "Mars: a MapReduce framework on graphics processors". Proceedings of the 17th international conference on Parallel architectures and compilation techniques ACT '08
[8] http://www.sas.com/en_in/insights/big-data/what-is-big-data.html
[9] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
[10]http://blog.raremile.com/wp-content/uploads/2012/09/hadoop_architecture.jpg