Analyzing Big Data using Hadoop

Tanuja A
Department of Computer Science & Engineering
VTU Belgaum, India

Swetha Ramana D
Department of Computer Science & Engineering
VTU Belgaum, India

Abstract

The major challenge of Big Data is extracting the useful information from the terabytes of raw data and analyzing the extracted information which is essential for the decision making. The above factor can be found in the proposed system which is simplified in three units. The first, Data Acquisition Unit (DAU) filters the collected data. The second, Data Processing Unit(DPU) process the data by collecting the useful information and the third, Analysis and Decision Unit(ADU), analyses the reduced data and supports decision making.

Keywords: Big data, Map Reduce, ADU, DPU, DAU, Hadoop

I. INTRODUCTION

The data is tremendously increasing day today leads to a Bigdata. The advancement in Big Data sensing and computer technology revolutionizes the way remote data collected, processed, analyzed, and managed in effective manner[8]. Big Data are normally generated by online transaction, video/audio, email, number of clicks, logs, posts, social network data, scientific data, remote access sensory data, mobile phones, and their applications [12], [13]. Storing and analysing the data gathered from remote in a traditional database is not possible as it cannot handle unstructured and streaming data. So it leads to a transition from traditional to Bigdata tool, which can analyse the data in well manner.

Hadoop is a tool used by many organizations for managing and analysing the data. Hadoop uses parallel execution of data using large clusters of tiny machines or nodes which results in faster execution. And even data is distributed among the nodes so the node failure can be easily handled.

MapReduce is a programming style, for Distributed processing on Hadoop. It contains the two functions; Map function will take the input as key/value pair and splits the data on several nodes for processing. Reduce function combines the results from Map function. The architecture and algorithm are implemented using Hadoop.

This paper is organized as follows. In section II, we give review of authors. In section III and IV provides existing system and proposed system, section V presents architecture of system and in section VI, conclusion and future enhancement.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before improving the tools it is compulsory to decide the economy strength, time factor. Once the programmer‘s create the structure tools as programmer require a lot of external support, this type of support can be done by senior programmers, from websites or from books.

A. Reference Paper [1]:

- Paper Title: Transitioning from Relational Databases to Big Data
- Originators: Sangeeta Bansal, Dr. Ajay Rana
- Techniques & Procedures: Hadoop & Relational Database Management System.
- Description: This paper gives a clear differentiation between relational database management system and Hadoop which led me to choose Hadoop tool. Hadoop can perform on larger data with less cost has it runs on tiny nodes and makes execution faster with its parallel processing nature which is not found in traditional one.

Hadoop can process text data along with multimedia also like audio, video, images etc but in relational database management system we can store and process only text data.

B. Reference Paper [2]:

- Paper Title: MapReduce Simplified Data Processing on Large Clusters
- Originators: Jeffrey Dean and Sanjay Ghemawat
- Techniques & Procedures: MapReduce.
- Description: In this paper, author are performed series of operation on larger datasets. Operation contains map and reduce function. Map function splits input data, which will be in the form of key/value pair into intermediate key / value pair and reduce function, combines the results of each intermediate key/value pairs. MapReduce functions will run on larger clusters containing small machines. These clusters run simultaneously, but this parallelization is hided to the user making the environment easy. Moreover optimization is done locally, where the intermediate data is read and written in host machine

without duplication, reflecting in reduced bandwidth of the network used. Data is split and stored in multiple nodes, even if one node fails to execute, the lost data is recovered from the rest the nodes.

C. *Reference Paper [3]:*

Excerpts from white paper on big data analytics

- Description: Big data not only deals with the large volumes, but also with the variety and the complexity of the data. Even small organizations end up with such complex data, but are not affordable to set the big data environment. In such a case, small organizations opt for the services for analyzing their data. This paper introduces big data analytics as a service.

This is performed in following ways

- Data is equally distributed among the nodes increasing in high availability of data and can perform simultaneous executions.
- Data transfer is avoided by bringing the service where the data is stored.
- The capabilities of software as a service analytics is integrated with the Hadoop framework.

D. *Reference Paper [4]:*

- Paper Title: MAD Skills: New Analysis Practices for Big Data
- Originators: Jeffrey Cohen, Greenplum, Brian Dolan, Fox Audience Network, Mark Dunlap, Evergreen Technologies,Joseph M. Hellerstein, U.C. Berkeley,Caleb Welton, Greenplum
- Techniques & Procedures: Magnetic, Agile, Deep analytics, Greenplum.
- Description: Here the analysts use agile working methods along with the deep analytics and magnetic, focusing on the volume of the data gathered from one of the widest network advertisers. These dataset are stored in the Greenplum which is the parallel database system. Finally, here they use agile development on database with SQL and MapReduce algorithm as an interface for storage of data.

E. *Reference Paper [5]:*

- Paper Title: Big Data and Cloud Computing: Current State and Future Opportunities
- Originators: Divyakant Agrawal, Sudipto Das, Amr El Abbadi, Department of Computer Science,University of California, Santa Barbara
- Techniques & Procedures: DBMS, Cloud, deep analytics, Ad-Hoc analytics
- Description: This paper focuses on the frequently updating applications as well as decision support for deep analytics. The above two factors ensures absolute departure from traditional to the modern cloud infrastructure. It analyzes the applications which have larger database creating abound between cloud and DBMS.

F. *Reference Paper [6]:*

- Paper Title : A Big Data Architecture for Large Scale Security Monitoring
- Originators: Samuel Marchal, SnT - University of Luxembourg, Luxembourg, Xiuyan Jiangz, Faculty - University of Luxembourg, Radu State, University of Luxembourg, Luxembourg, Thomas Engel, University of Luxembourg, Luxembourg.
- Techniques & Procedures: Hadoop, Spark, Shark, Pig, Hive.
- Description: Huge data is collected in network flows and traffic which lead to a big data , the data can be collected from HP, domain name system, hyper text transfer protocol traffic for intrusion recognition and forensic study. Here author has taken 5 tools for analysis of traffic data from various sources of network and measured the performance of tools. The 5 tools taken here are Hadoop, spark, pig, hive and shark.

According to the analysis made by others, the results states spark and hadoop will analyze faster than other tools.

G. *Reference Paper [7]:*

- Paper Tile: A Hadoop-based Distributed framework for efficient managing and processing big remote sensing images
- Originators: C. Wang, Hainan Geomatic Center, China,F. Hu, George Mason University, USA. X. Hu, Hainan Geomatic Center, China,S. Zhao, Mapping and Geoinformation of China, China. W. Wen, Hainan Geomatic Center, China. C. Yang, George Mason University, USA.
- Techniques & Procedures: HDFS, MapReduce, Orfeo toolbox, Parallel computing
- Description: Satellites around the earth are generating various variety of data every second leading to big data. Images are one among those data, which need to be processed efficiently. In this paper, Orfeo toolbox, used to process images in large volumes, is combined with MapReduce and HDFS.

The paper concludes that with the integration of the above techniques, operations on images can be processed side by side with a efficient results and reduced execution time.

III. EXISTING SYSTEM

Present sensors used in the various fields streams the data continuously. In case of data on the images, major implementations have been done on the images generated through remote sensory satellites.

A. *What lags behind?*

- Transformation of remotely sensed data to scientific calculation and analysis is a major challenge
- Analysis needs a standard format for processing, but the data retrieved from the satellites possess undesirable format.

IV. PROPOSED SYSTEM

- Here, Big Data refers to the tremendously streaming data, both in terms of speed and volume.
- This paper deals with the architecture of high volume data analytics. This architecture shows the process of analysis of both real time and offline data.
- When it comes to offline data, it is first moved to the data storage devices, and processed when necessary.
- But real time data directly steps into the filtration process. Here it is passed through the algorithms, where the unrelated data is taken off.
- Meanwhile load balance algorithms are applied for equal distribution of the data to the servers.
- Along with the filtration and load balancing, the overall efficiency of the system is enhanced.

Fig(1) gives the flow of the project, where the data will be stored in Hadoop DFS, then it split and reduce the size by Mapreduce technique, later the reduced data is fed as a input for analysis, which can lead to a better solution and even the analyzed data can put to dashboards.

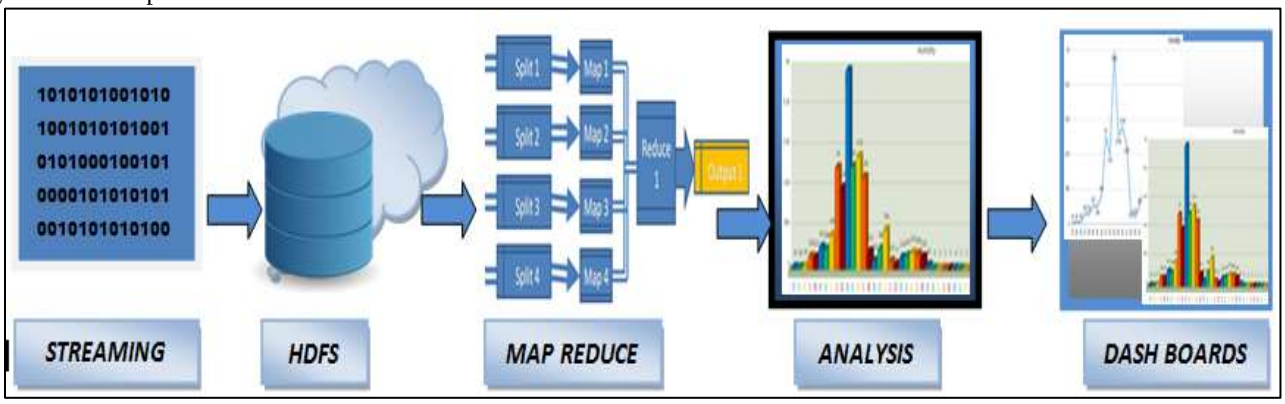

Fig. 1: Flow of Existing System

Advantages of proposed system

- Load balancer by Mapreduce is used here will distribute the work equally among all Tasktracker.
- Data acquisition used here will filter the data and even compression also will be undertaken if necessary.
- Data extraction will help to take up a useful information, leaving behind all the noise and unwanted data or information.
- Analysis used here can be placed on dashboards of company where development team can take up a decision fast.

V. SYSTEM ARCHITECTURE

The below fig (2) shown is the system architecture, where it contains 3 major units

- Data Acquisition Unit(DAU)
- Data Processing Unit(DPU)
- Analysis and Decision Unit(ADU)

In DAU, the data is collected and placed in the local file of the host machine, later it is stored in the Hadoop DFS where data is divided into many datanode and their location is stored in namenode but in project only single cluster has been used.

In DPU, the main reduction of data takes place by Mapreduce technique present again in Hadoop, where the work is divided among several workers and master will take care of workers. Once the task is done either combined or sorting is been done and placed back in Hadoop DFS by creating a particular directory so that it will be clearly known to all.

In ADU, the data stored after the Mapreduce is placed back in local host machine as we don't have analysis option in Hadoop. Later using java- script language in project it has been analyzed and plotted in the form of graph i.es line and bar graph so that everyone can take decision easily without much difficulty in understanding.

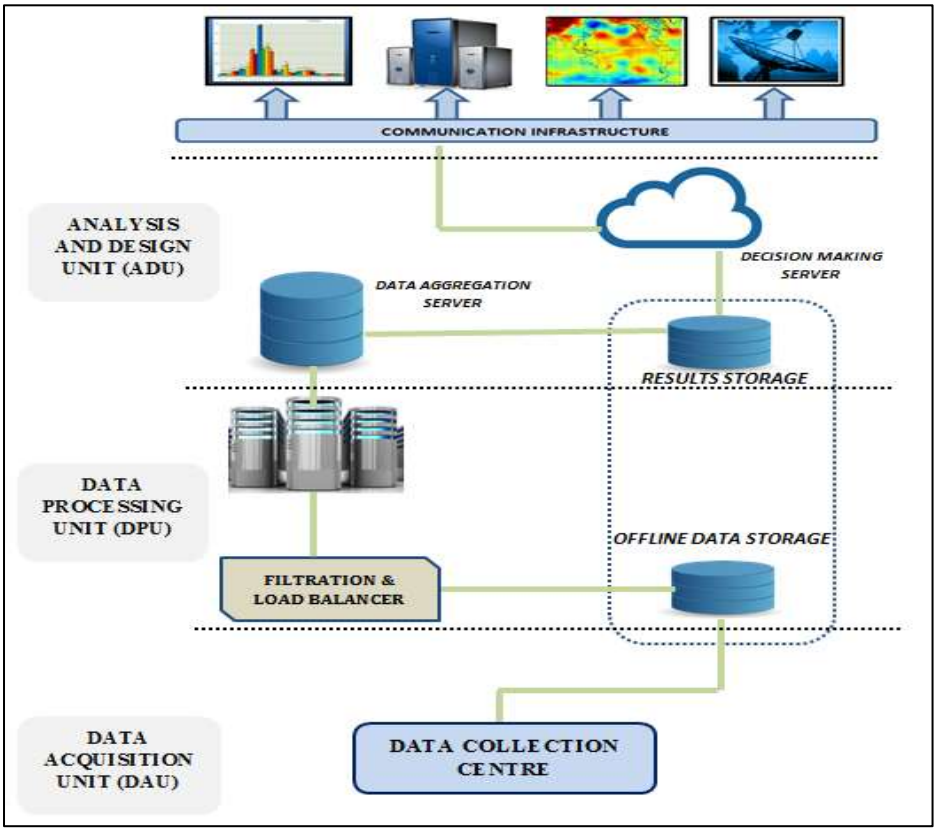

Fig. 2: System Architecture

VI. Conclusion And Future Scope

In this paper, provides the framework for reducing the file size so that can be executed faster. Terabytes of a data can be processed in a minute which was not possible in traditional database. Hadoop tool used for a data is also increases the performance comparatively other tools.

In the project input is taken as a file filtered and processed by MapReduce concept and stored the details in Hadoop DFS and the analysis is done on the output of MapReduce.

The main advantage of the design is that analysis is done very fast so that the organization can take up a decision quicker.

References

[1] Sangeeta Bansal, Dr. Ajay Rana ,Department of Computer Science & Engineering Amity University, Noida (U.P.) India
[2] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, pp. 107–113, 2008.
[3] Big Data Meets Big Data Analytics Three Key Technologies for Extracting Real-Time Business Value from the Big Data That Threatens to Overwhelm Traditional Computing Architectures
[4] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, "Mad skills: New analysis practices for Big Data," PVLDB, vol. 2, no. 2, pp. 1481–1492, 2009.
[5] D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloud computing: Current state and future opportunities," in Proc. Int. Conf. Extending Database Technol. (EDBT), 2011, pp.
[6] S. Marchal, X. Jiang, R. State, and T. Engel, "A Big Data architecture for large scale security monitoring," in Proc. IEEE Int. Congr. Big Data, 2014, pp. 56–63.
[7] A Hadoop-Based Distributed Framework For Efficient Managing And Processing Big Remote Sensing Images C. Wanga,b, F. Hub,*,X. Hua, S. Zhaoc, W. Wena, C. Yangb.
[8] Real-Time Big Data Analytical Architecture for Remote Sensing Application Muhammad Mazhar Ullah Rathore, Anand Paul, Senior Member, IEEE, Awais Ahmad, Student Member, IEEE,Bo-Wei Chen, Member, IEEE, Bormin Huang, and Wen Ji, Member, IEEE
[9] R. A. Dugane and A. B. Raut, "A survey on Big Data in real-time," Int. J.Recent Innov. Trends Comput. Commun., vol. 2, no. 4, pp. 794–797, Apr.2014.
[10] X. Yi, F. Liu, J. Liu, and H. Jin, "Building a network highway for BigData: Architecture and challenges," IEEE Netw., vol. 28, no. 4, pp. 5–13,Jul./Aug. 2014.
[11] E. Christophe, J. Michel, and J. Inglada, "Remote sensing processing:From multicore to GPU," IEEE J. Sel. Topics Appl. Earth Observ. RemoteSens., vol. 4, no. 3, pp. 643–652, Aug. 2011.
[12] Y.Wang et al., "Using a remote sensing driven model to analyze effect of land use on soil moisture in the Weihe River Basin, China," IEEE J. Sel.Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 9, pp. 38923902, Sep. 2014.
[13] "C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.
[14] R. D. Schneider, Hadoop for Dummies Special Edition. Hoboken, NJ, USA: Wiley,2012