Indian Journal of Science and Technology, Vol 9(31), DOI: 10.17485/ijst/2016/v9i31/83451, August 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Implementation and Analysis of MapReduce on Biomedical Big Data

Praveen Kumar Rajendran[1*], A. Asbern[2], K. Manoj Kumar[3], M. Rajesh[4] and R. Abhilash[5]

[1]Cognizant Technology Solutions, Chennai - 600028, Tamil Nadu India; praveenkumar558@gmail.com
[2]System Design Associate, Global Knowledge Network India Private Limited, Chennai; asbern2008@gmail.com
[3]Department of Computer Science, Sri Venkateswara College of Engineering, Tirupati – 517502, Andhra Pradesh, India, kandalamanojkumar@gmail.com
[4]Sathyabama University, Chennai – 600119, Tamil Nadu, India; rajesh.manoharan89@gmail.com
[5]Amazon Development Center, Chennai, Tamil Nadu, India; mailabhi46@gmail.com

Abstract

Organizing and maintaining the big data are the two major concerns which have led to many challenges for the organization. The main objective of this research work is to give an overall idea about organizing Big data with High performance. MapReduce is one of the commonly used techniques which is used to analyze a large volume of data in an efficient manner. A common overview of Big data and the implementation of the MapReducing technique on Biomedical Big Data has been discussed in this paper with an algorithm. Discussion on performance analysis of MapReducing technique being will open the doors for further research activities in Big Data Analytics and MapReducing technique. Highlight of this research work is the data which has been selected and the output of the research work has been openly discussed to help the beginners of Big data. The proposed research work will give an insight about the implementation of Hadoop Distributed File System for small and medium sized business.

Keywords: Big Data, Big Data analytics, Biomedical Data Analysis, MapReduce, Performance

1. Introduction

IBM is one the major player in Big Data Analytics, it also states that 90% of the data in the world have created in the last two years[1]. A report from IDCS states that the data volume would grow by a factor of 300. That is from 130 Exabyte to 40,000 Exabyte between 2005 and 2020[2]. The major reasons for such tremendous growth of data are growth of Internet and Mobile devices. Transformation of static website to dynamic website can be considered as the major reason for the increased volume of data. Since mobile devices are accessible and affordable to everyone with lots of features, it has become easy for the user to share information i.e., create data in an exponential manner.

For any Business organization or Health care organization or Government organization or Educational Institution, data that are being collected from the user is the biggest asset. Analysis on the data that is being collected would have a large impact on the organization. As the data grow exponentially, it is a great challenge for an organization to analyze the ever growing data. Whatever may be the size of data, the organization has to preserve and maintain the data created by the user. Now the major goal of an organization is to maintain the large volume of increasing data efficiently. Big Data analytics are the solution for the challenges faced by the organization. Since the traditional tools and technique cannot handle large volume of data, Big data analytics has received a huge welcome from various organizations.

The following evidence for the above statement:

- Government of The United States of America has allocated of 200 million USD for big data research[3].
- The Massachusetts Institute of Technology has hosted a separate Intel Science and Technology Centre for Big Data Research[4].

*Author for correspondence

- Oracle has released big data products and solutions for enterprise to experience real business value from big data[5].
- Center for Development for Development of Advanced Computing under the department of Electronics and Information Technology, Government of India uses Big Data Analytics for the research on data that is being collected by the Government of India[6]. Experts from CDAC reach Students via Workshop and Guest lecture to provide awareness on Big Data Analytics.

In this paper a brief overview of Data, Data Structure, Big Data, and Big Data Analytics has been discussed. MapReduce concept has been implemented on large volume Biomedical Data and the result of the implementation has been narrated in this paper. The performance of the implemented MapReduce technique on biomedical data has been discussed, which would enable would help the future research activities.

2. Big Data Overview

In general the term "Big data" indicates the large volume of data. The Size of the "term" large depends upon the particular individual who is handling that data. In[7] has defined big data as large volume of heterogeneous data, from autonomous sources with distributed and decentralized control from which complex and evolving relationship are sought[7]. In[7] have defined it as HACE Theorem for Big Data. From our analysis from the HACE theorem, we could define Big data as "Large volume of heterogenous, unstructured data which cannot be processed using traditional database management system". In[8] has defined big data as a "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze". From this definition we can infer that, the traditional tools and methods will not be sufficient to manage and analyze the ever growing data. Process or manipulation or calculation, being carried using some special tool over large volume of heterogeneous (big data) is coined as "Big Data Analytics". Migrating or moving towards a new technology will arise, when there is a lot of difference between the existing system and migrating system. Here the difference is on "Data". Table 1 gives a comparison between the big data and traditional data.

Table 1. Comparison of big data vs traditional data

Sl. No.	Traditional Data	Big Data
1.	Here the data is "Structured" data	Here the data is "Unstructured or Semi structured" data
2.	The size of the data is very small	The size is more than the traditional data size
3.	Here the data is Centralized	Here the data are distributed
4.	It is easy to work or manipulate	It is difficult to handle the data
5.	Normal system configuration is sufficient to process	High system configuration is required to process the data
6.	A traditional database tools is enough	Special kind of tools are required
7.	Normal functions are enough to manipulate the data	Requires special kind of functions to manipulate the data

3. Research Motivation

From the above Table 1, certain important facts about Big data are quite essential for moving towards a new kind of methodology, thereby giving a clear view for the research scholars to explore on analyzing big data. As the data gets increased rapidly, it is very essential to manage the ever growing data. Rapid growth in data from Megabyte to Gigabyte, from Gigabyte to Terabyte, from Terabyte to Petabyte, from Petabyte to Exabyte[9] has made the research scholars to think beyond the traditional tools. The two major research challenges in migrating from traditional data to big data would be the performance and processing of data. Processing of data indicates the processing of large volumes of unstructured data. Unstructured data indicate data of different data types (Structured data indicate the same kind of data type). In the real time scenario, large volume of unstructured data type is being produced by the users. It is not feasible to process and manage the unstructured data using the traditional database management tools.

Another major challenge for the research scholar with big data would be "Cost". In order to manage large volume of data, it would be mandatory to invest much on infrastructure. The ultimate goal of Big data research can be defined as 3V's- Volume, Variety and Velocity[10]. In[8] has coined "Value" as another parameter. The term "Value" indicates the output of the data process. So 3V's of goal can be expanded as 4V's of Big data. The 4V's of big data can be defined based upon the goal as "Large *Volume* of *Variety* of data has to be reduced in high *Velocity* and

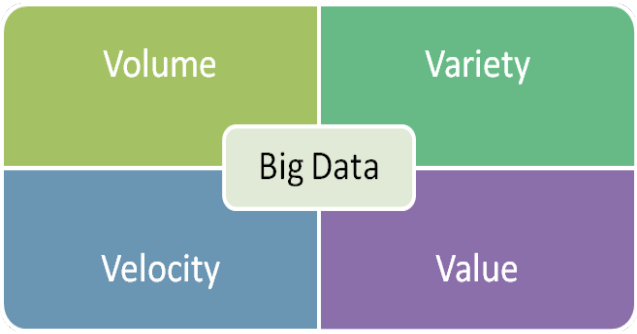

Figure 1. Overview of big data.

arrive a *Value* or solution of the process". Figure 1 will give a pictorial representation overview of Big data.

Predicting the future by analyzing the previously processed data is the main reason that motivates "Big Data Analytics" research. Analyzing the data that has been collected in a certain period of time can be processed and using the result of the process, the future can be predicted. The prediction can be a correct one, or it may be end up in a different manner. This kind of activities would be much useful for business and real time scenario, where the data that is being collected are unstructured and large in volume. These are some of the main reasons that motivates research on Big data analytics. Section 4 of this paper would give a clear idea on "MapReduce" concept of Big data analytics. Implementation of MapReduce concept has been discussed the rest of the sections.

4. MapReduce

MapReduce is one of special kind of Big data analytics, which is commonly used in large volume of unstructured data[11]. The two major operations of MapReduce concept are, Mapping and Reducing. Mapping is the process of filtering and sorting of large volume of Big Data and Reducing is the process of counting the data that has been filtered and sorted. Map-reducing concept can also be utilized with the data which is present in the cloud based environment, where the services are hosted to the client over Internet[12,13]. The three major advantages of MapReduce programming are.

- Simple and Powerful, where large number of task can be performed in an easy manner.
- Scalability, the framework can be expanded and reduced based upon the requirement.
- Low Fault tolerance, failure of a particular node will not affect the entire process.

With the above advantages, the process of Map-Reducing can be implemented using Apache Hadoop, which is a popular open source tool for Big Data Analytics. Map-Reducing concept can also be utilized for association rule mining process in data mining[14]. In[15] has defined the process of Map-Reducing as the process which consists of map phase, shuffle phase, sort phase and reduce phase. The process involved in these phases has been discussed in the following section, with an illustrative implementation on Biomedical Big Data.

5. MapReduce on Biomedical Big Data

5.1 Overview

In general, the map reducing process is carried out in Apache Hadoop framework using Hadoop Distributed File System (HDFS)[16]. It divides the data into name node and data node. Name node can also be termed as Index node, which has the index of the data which has been spitted into data node. The data has been spitted into data node in a redundant manner, i.e., the same data will be stored in the multiple name nodes. When the map-reduce program initiated, the data will be fetched from the nearest name node based, which has been redirected by the name node.

When the map-reduce program has been initiated, the data which has been preloaded into the HDFS framework will be divided across the nodes and mapping process will be carried out. The output of the mapping process will be again stored in the HDFS framework. From the mapped data which has been stored in the data node, reducing process will be carried out.

In our experiment, biomedical data from the PubMed database has been fetched into the HDFS framework. MapReduce program has been initiated to identify the behavior of the HDFS framework at various time stamps. MapReduce program has been initiated for the 40,000+ data which has obtained from the PubMed database, using keyword EDTA (Ethylenediaminetetraacetic acid). Figure 2 gives the information about the data which has been taken for the experiment from the PubMed Database. The steps

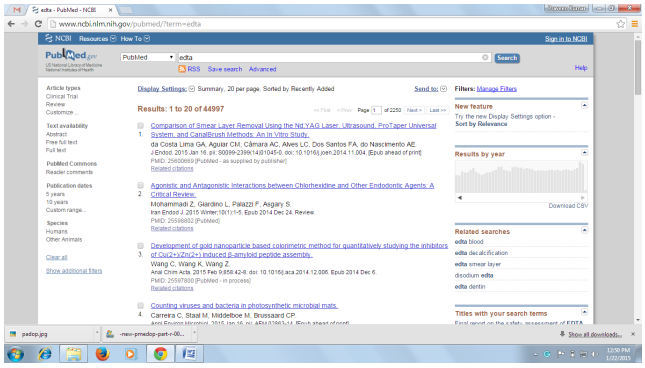

Figure 2. Data Fetched in HDFS Framework.

Map-reduce algorithm.

1. Start
2. {
3. Download the data for which the Map-Reduce program has to be implemented.
4. Download and install the Hadoop from https://hadoop.apache.org/
5. Initialize the hardware required to run the Map-Reduce program.
6. }
7. {
8. Initialize the data set for which the Map-Reduce program has to be implemented into the Hadoop Distributed File System (HDFS).
9. Code the Map-Reduce program on the loaded data.
10. {
11. Mapping of Data takes place on the data which has been loaded.
12. Reducing of Data takes place on the data which has been mapped.
13. Extraction of process takes place on the mapped and reduced data.
14. }
15. Result of the program has been derived from the extraction step.
16. Step is repeated 'N' number of times to analyze the performance of the program.
17. End.

involved in the experiment have been narrated in the following section as an algorithm.

5.2 Algorithm for Map-Reduce Program on Biomedical Data

The following Table 2 gives step by step procedure to be carried out for the implementation of Map-Reduce program on Biomedical Big Data.

5.3 Implementation

The above proposed algorithm has been implemented using the Hadoop Framework. When the data has been loaded into the Hadoop Distributed File System, using the Map-Reduce program the entire data has been mapped and reduced. Figure 3 shows the output, when the Map-Reduce program has been initiated.

The map-reduce program has been initiated on unstructured data, which consist of Integer data type and Character data type. When the large volume of unstructured data (variety) of data has been loaded into the HDSF, the volume of the data loaded has been reduced. The proof for such reduced volume can be identified from the Figure 4. Number of times each and every word and numbers, repeated in the data has been obtained as the result of Map-Reduce program. Inside the Figure 4 we can find the output of the Map-Reduce program, which shows the number of times each and every element of the data has been repeated in the side.

5.4 Implementation Analysis

As the implementation satisfied the basic concepts of Big Data analytics, MapReduce program has been initiated

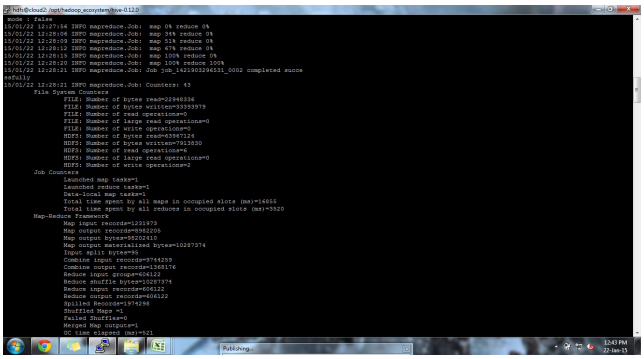

Figure 3. Working of map-reduce program.

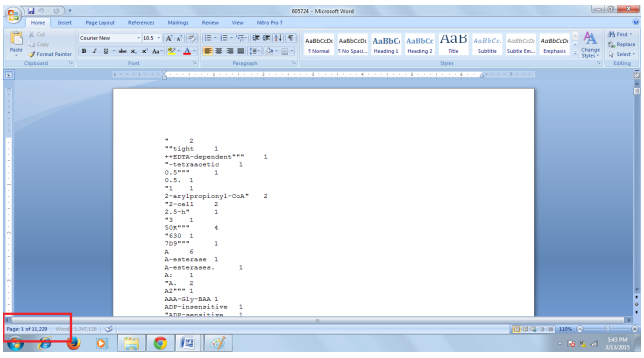

Figure 4. Output of map-reduce program.

Table 3. Performance analysis

Sl. No.	Mapping(ms)	Reducing(ms)
1.	16855	3520
2.	16560	4082
3.	16579	3560
4.	16265	3701
5.	16443	3662

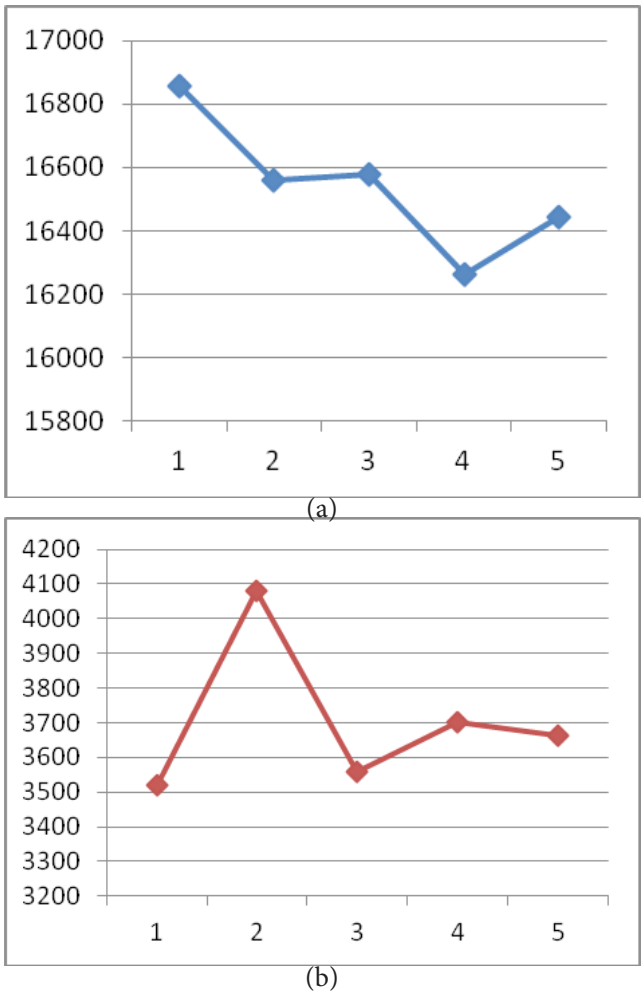

Figure 5. **(a)** Mapping. **b)** Reducing.

on same data for 5 times in a different interval of time to evaluate the performance of the implementation. Table 3, shows the result of the implementation analysis. The values obtained are in Milliseconds. From the obtained values, the following are the conclusions arrived.

- The time taken for the mapping is less than the reducing process. This is obvious because, the process of sorting would take less time than arrangement.
- The time taken for the first mapping and reducing is more than the other values. Since the data has been fetched into HDFS, the time taken for the first time is more than other time.
- Forming a graph using the obtained values gives us information that there is not much deviation in the performance of the MapReduce program on the Big Data. Figure 5.1 and Figure 5.2 gives the insight of Mapping and reducing process. In Figure 5(a) and Figure 5(b), X axis indicates the instance number of the process and Y axis indicates Time taken for the process in milliseconds.
- The average time taken for the mapping process is 16540.4 milliseconds and 3705 milliseconds.

6. Future Scope and Conclusion

From the implementation and performance analysis, we can come to a conclusion that the MapReduce program can be implemented in any form of Big Data and meaningful information can be arrived as a result of the implementation. In the proposed implementation, the data has been fetched from the standalone system. In the future research work, a deep analysis can be made on the data which has been stored in Cloud. This will open to a new era of computing and a next step in Cloud computing, where Analytics as a Service can be provided as services to the customers. Implementing MapReduce program on the data which has been collected in IoT environment will take the analytics and IoT into different frontiers.

7. References

1. IBM. Big Data at the Speed of Business. Available from: http://www-01.ibm.com/software/data/bigdata/
2. Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC analyze the future. 2012 Dec; 2007:1–6.
3. Liu S. Exploring the future of computing. IT Professional. 2013:2–3.
4. Big data at CSAIL. Available from: http://bigdata.csail.mit.edu/
5. Oracle. Oracle big data for the enterprise; 2012. Available from: http://www.oracle.com/caen/technologies/big-data
6. Centre for Development of Advanced Computing. Available from: http://cdac.in/
7. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering. 2014 Jan; 26(1):97–107.

8. Manyika J, et al. Big data: The next frontier for innovation, competition, and productivity. San Francisco, CA, USA: McKinsey Global Institute; 2011. p. 1–137.
9. Hu H, et al. Towards scalable systems for big data analytics: A technology tutorial; 2014.
10. Zikopoulos P, Eaton C. Understanding big data: Analytics for enterprise class hadoop and streaming data. New York, NY, USA: McGraw-Hill; 2011.
11. Jiang D, Tung AK, Chen G. Map-join-reduce: Toward scalable and efficient data analysis on large clusters. IEEE Transactions on Knowledge and Data Engineering. 2011 Sep; 23(9):1299–311.
12. Rajendran PK, Muthukumar B, Nagarajan G. Hybrid intrusion detection system for private cloud: A systematic approach. Procedia Computer Science. 2015; 48:325–9.
13. Muthukumar B, Rajendran PK. Intelligent intrusion detection system for private cloud environment. Security in Computing and Communications. Springer International Publishing; 2015 Aug 10. p. 54–65.
14. Asbern A, Asha P. Performance evaluation of association mining in Hadoop single node cluster with Big Data. 2015 International Conference on Circuit, Power and Computing Technologies (ICCPCT); 2015 Mar 19. p. 1–5.
15. Kailasam S, Dhawalia P, Balaji SJ, Iyer G, Dharanipragada J. Extending MapReduce across Clouds with BStream. IEEE Transactions on Cloud Computing. 2014 Jul 1; 2(3):362–76.
16. Lakshmi M, Sowmya K. Sensitivity analysis for safe grain-storage using big data. Indian Journal of Science and Technology. 2015 Apr 1; 8(S7):156–64.
17. Kim BS, Kim TG, Song HS. Parallel and distributed framework for standalone monte carlo simulation using mapreduce. Indian Journal of Science and Technology. 2015 Oct; 8(25).
18. Noh K-H Lee D-S. Bigdata platform design and implementation model. Indian Journal of Science and Technology. 2015 Aug; 8(18).
19. Dhamodaran S, Sachin KR, Kumar R. Big data implementation of natural disaster monitoring and alerting system in real time social network using hadoop technology. Indian Journal of Science and Technology. 2015 Sep; 8(22).
20. Rajendran PK, Rajesh M, Abhilash R. Hybrid Intrusion Detection Algorithm for Private Cloud. Indian Journal of Science and Technology. 2015 Dec; 8:35.
21. Rajesh M, Abhilash R, Kumar RP. URL ATTACKS: Classification of URLs via Analysis and Learning. International Journal of Electrical and Computer Engineering (IJECE). 2016 Jun 1; 6(3).
22. Muthukumar B ,Praveen Kumar Rajendran, S. Murugan, G.Nagarajan. Multilevel Intrusion Detection System for Private Cloud Environment. Advances in Intelligent Systems and Computing(Accepted).