International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-3, Issue-7, July 2015

Effective Execution Time for Processing Large Date

Shikha Malik, Rajiv K Nath

***Abstract*— This paper enumerates methods to look around for the alternate to process Big Data in minimum amount of time using point in time analysis. Now a days with available databased SQL Server, Oracel, DB2 etc. it is quite complex to manage and process larger amount of data for any health care industries. There are lot many hurdles a database engineer face in capturing, storing, maintaining relational data with help of tables and schemas, searching, sharing, presenting to higher management for taking critical decisions related to any org to grow. This paper starts with what is Big Data, why we are moving towards this evolution in IT industry. Following the advantages and challenges in opting Big Data and what all requisites we need to move on to Big Data. The focus on my paper will be to discuss various possible solutions for the issues in managing Big Data using Hadoop (YRAN, HDFS and MapReduce etc.).Cloud Computing plays a subsequent role in managing data using big data tools. Cloud Computing together with Big Data can help any industries to grow much faster pace with promising results.**

***Index Terms*— Big Data, Hadoop, YRAN, HDFS, Map Reduce.**

I. INTRODUCTION

In order to maintain very large and complex data many industries, research organizations, various institutes, schools etc. some or the another kind of relational databases which directly or indirectly involves hardware maintenance to setup servers, network system to build connectivity with all servers, man power cost to build up a team who is dedicatedly working on maintaining, saving, executing, decomposing very large and complex data and also cost spend on data security. MapReduce Framework introduced by Google to process data which is very large and complex. Map Reduce framework works on programming architecture which is used to process data with certain type of parallel and distributed type of algorithms. This framework is used to increase scalability and to sustain fault lines as minimum by optimizing engine used to execute the data processing. Apache's distributed file system known as HDFS(Hadoop Distributed File System) is emerging as software integrated with MapReduce framework ,enable to work on cloud computing[1].

II. BIG DATA

First, Big Data refers to large and very complex cluster of data which is structured and unstructured lying on distributed servers and it is very difficult to process the data with defined fault-tolerance and execute the data within best optimized time period with existing soft wares and databases available. On broader scale we can proclaim big data based on following properties: Volume, Variety and Velocity, Variability and Complexity.

- Volume: Many factors devoting towards increasing volume streaming data and data collected from very sensitive devices etc.,
- Variety: Data can be of any format:emails,datasheets,video, audio, transactions scripts, csv files, encrypted data etc.,
- Velocity: Execution time for data producing and processing.
- Variability: With respect to time amount of data coming in and going out.
- Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources.

The data must be associated, equated, clarified and transformed into required formats before actual processing [2].

III. HADOOP

A. About Hadoop in Big Data

Big data is encircled with datasets that are very large to be handled by traditional available databases soft wares. To overcome this hurdle business executive's release a need to remain competitive business executives need to ratify the new technologies and techniques emerging due to big data which can solve their problems to deal with cumbersome data. Big Data which is large dataset can have all types of data which can be structures, unstructured and semi structured. Traditional databases that are large enough to analyse, execute, process and save hence a new term "Big Data" came into existence. Big Data serves the attempt to quantify the growth rate of data in terms of volume. Volume of Big Data depends on sectors like we are using data in health care companies or for business analysis purpose to set up or grow the business in particular segment. Its volume may vary from terabytes to petabytes and it is keeping on increasing in present IT world.Big data is a term not any tool/software or framework which works on data, to work on Big Data we have some available tools/softwares in merchandise like Hadoop. Hadoop,which is an open source freely available provided by Apache is used to process big data which can be structured, unsturcted or semistructured based on data storing logic.Hadoop uses the Mapreduce model to trace and locate all admissible data over a distributed network following to select the data which is a result of the query.There are different available Databases that can be used to process structured big data like NoSQL,MongoDB,HBase, TerraStore etc[3].

Shikha Malik received the M.Sc degree in Information Technology from Sikkim Manipal University in 2012. During 2007-2008, she worked as a trainee in LG Electronics Greater Noida(UP).

B. Big Data Storage Technologies

Need to go for Big Data arises from the fact to explore the potentiality to capture large amounts of data which is used by business executives and analytics.Means of storing big data are clustered network-attached storage(NAS) and object-based storage systems.As without changing the mean to store big data technology, executives/analytics will not be to collect meaning information from big data.Scale-out NAS is based on an existing earlier NAS system.NAS is like a storage device in which computer in used as storage device so no hardware peripherals devices like keyboard,mouse are required.In NAS system computers are connected with each other via network and each NAS system serves a mean to act like a server. This system is used to reinforcing the demands of big data.In another mean to store Big Data that is Object-based storage systems, users deal with sets of objects instead of files and the objects are distributed across several servers/systems/devices. This storage system provide high efficiency to store Big Data with high capacity and throughput which in turns increase the reliability, scalability and hence minimize the processing time ,which is the ultimate target of Big Data Storage. Both storage systems scale-out NAS and Object-based storage systems are built for increasing scaling storage. The difference maker is their respective metadata characteristics. They are not mutually exclusive[1].

C. Figures and Tables

Data in IT, Healthcare and institutions are keep on increasing. These data potential will go high exponentially in coming future. If we try to plot data vs availability we can have plot curve as shown.

Figure 1: Big Data Flow

D. Big Data Analytics

On sitting on the front end storing data is the only part we can see from the picture. In actual it requires special techniques, algorithms to analyze and store such a big data that it can be processed correctly and in minimum amount of time. For this analyst and executives need to do lot of researches on big data and they then try to get familiar with all methodologies applied in Big Data, after doing lot of research and development they come to end to adopt the technology which get fitted with their business and ensure that engineers or employees of the company get hands on the skill after certain trainings and learnings. Data storage technologies are different based on the fact the data are structured or unstructured. Unstructured data can be analyzed by available software like Hadoop. And for structured data available softwares are NoSQL, MongoDB etc. Hadoop uses the framework called MapReduce. MapReduce is the core construct of Hadoop this name (MapReduce) came from its operations that Hadoop program perform using key-value pairs which we try to access Big Data with help of a query. In MapReduce model mapping task is given to a piece of data known as a key to search on and find all relevant values based on the key and then it converts the key and its values into subset of a query which again work based on key-value pair.To work on structured data software available is NoSQL which works on BASE rather than traditional ACID.BASE works on Basically available, Soft state and eventually consistent. Basically available refer to the perceived availability of the data. If a single node fails, part of the particular data will not be available but the query executes the results for all active nodes and remains operational. Soft state refers the state of a system which may change over time, even without giving input because of the eventual consistency model. Eventual consistency refers the system will become steady over the time, given that the system doesn't obtained any incoming input during that time[5].

E. Importance of Big Data

Some major contributions big data can make to businesses:1) Transparency in data 2) Improvement in performance 3)Decision making support 4)Minimizes the analysis time on Big Data. Big data can provide data more accurately and detailed. Big data can help in business to get decisions based on data facts that are more detailed, accurate and are processed and executed within minimum amount of time. Several companies spend quality amount of time in analyzing data pattern and then do strategies based on data patterns. Using Big Data we can get this data in minimum possible time which is helping any business to take decisions faster and grow[4].

IV. FUTURE WORK

To successfully identify and implement big data solutions and benefit from the value that big data can put forward, all IT Organizations need to give some time and resources to visioning and outlining. This will provide the foundation needed for effective processing. Without this result, organizations will not realize the envisioned benefits of big data and will risk being left behind championships. There are huge number of available technology solutions for dealing with big data. All of these solutions tend to have low time to value and maintenance but relatively high total cost of ownership. Cloud-hosted software as a service (SaaS) solutions can help reduce the barriers of participating in the big data field. Google and Amazon appliance MapReduce-based solutions to process huge datasets using a large number of computers — e.g., terabytes of data on thousands of computers. MapReduce algorithms take large problems and divide them into a set of discrete tasks that can then be distributed to a large number of computers for processing and the results combined into a problem solution. We can fuse Big Data with cloud that can be more mature

outcome of all existing technologies and will help to process Big Data with security protocols.

ACKNOWLEDGMENT

I would like to express my sincere gratitude and appreciation to everyone who made this thesis possible. Most of all, I would like to thank my supervisor, honorable Mr. Rajiv K Nath Asst. Prof., Department of Computer Science and Engineering, for his valuable guidance and providing all sorts of assistance throughout this dissertation work and for encouraging and inspiring me to carry out the project in the department. Finally, I would like to thank my family, IJETR and classmates for the support, productive discussions and their help to finalize this content of paper work.

REFERENCES

[1] Coronel, C., Morris, S., & Rob, P. (2013). Database Systems: Design, Implementation, and Management, (10th Ed.). Boston: Cengage Learning. Eaton, Deroos, Deutsch, Lapis, & Zikopoulos. (2012). "Understanding big data: Analytics for enterprise class Hadoop and streaming data". New York: McGraw-Hill.

[2] N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens:2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011

[3] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing referencearchitecture: Cloud service management perspective.". Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.

[4] Eaton, Deroos, Deutsch, Lapis, & Zikopoulos. (2012). "Understanding big data: Analytics for enterprise class Hadoop and streaming data". New York: McGraw-Hill.

Shikha Malik received the M.Sc degree in Information Technology from Sikkim Manipal University in 2012. During 2007-2008, she worked as a trainee in LG Electronics Greater Noida(UP).