ISSN:2320-0790

An International Journal of Advanced Computer Technology

CONCEPTUALIZING BIG DATA ANALYTICS THROUGH HADOOP

Manish Kumar Singh[1], Professor G S Baluja[2], Dr. Dinesh Prasad Sahu[3]

[1]MCA (MDU) and B.Sc. (APS) Computer Science (University of Delhi)
[2]Dept. of Computer Application, DTTE, Govt. of NCT Delhi
[3]Ph.D, M.Tech. and MCA (Jawaharlal Nehru University)

Abstract: Big Data Analytics (BDA) is the matter of great concern in the research and analysis field of data science today. Each and every day, data in petabytes are processed and analyzed from different sources on the Internet. Big organizations and Internet giants leave no stone to use the hidden information of Big Data. Hence, they make use of one algorithm or another algorithm of data science to do the analytics of Big Data. The most successful algorithm that has been developed so far to do effective Big Data Analytics is MapReduce – an algorithm developed and successfully implemented by Google. The data scientists of data science field have been trying hard to build such an architecture that can implement the philosophy of MapReduce to achieve efficient, effective, and the most economical way of doing BDA. In this effort, Apache Hadoop appears to be the most promising technology build so far. This article is an effort to bring the underlying details of Big Data Analytics done through Hadoop.

Keywords: Big Data, Big Data Analytics (BDA), 3V Model of Big Data, MapReduce Framework, Hadoop framework, Hadoop Distributed File System (HDFS),

I. INTRODUCTION

The world has experienced the production of a huge amount of data in last decade after the arrival of myriad sophisticated and advanced technologies, methodologies, devices and communication media such as social networking sites. The data collected so far has been estimated to be in petabytes and the amount of data is still growing. All this has led to the development of a well-known concept known as Big Data. What exactly a Big Data is – the answer lies in the name of Big Data itself i.e. Data which is very big. And the complexity of such data can be assumed from the fact that no traditional techniques can be applicable to process them. Today, Big Data is not restricted within a concept only rather it has evolved as a subject under research and constant practice. The idea behind research and development of Big Data is to derive useful hidden information from them through usage or devising of variety of tools, technologies, methodologies and frameworks.

Big Data evolves from various kinds of data generated from myriad data sources through the usage of different tools and techniques. On the basis of tools & techniques applied in different fields, Big Data can be categorized chiefly into following types:

***A.* Search Engine Data**: data collected from the search engines that are a huge collection of data present on the Internet.

***B.* Social Media Data:** data collected through various social media sites like Facebook, Twitter, Linkedin, Google+, YouTube, Tumblr, etc. These data are in the form of views, posts, suggestions and comments that are posted on these sites by the people of different parts of the worlds.

***C.* Power Grid Data:** data collected from the power grid that has information related to a particular node with respect to a given base station.

***D.* Transport Data:** data collected from the transport sector that involves distance and route covered by an available vehicle with some model.

***E.* Stock Exchange Data:** data collected from the various stock exchange of the world that includes data related to buy and sell based decisions with respect to customers' share in different companies of the world.

***F.* Black Box Data:** data collected through black boxes of airplanes, helicopters or jets that involve data such as

voices of flights' crew, earphones & microphones recordings, etc.

The need of the present day is to utilize Big Data in effective and efficient manner because it has great potential to contribute to good decision-making in the field of business, research, and development. That's why the field of Big Data Analytics (BDA) has evolved in the last decade and many big organizations across the world have been benefitted from the information obtained through BDA. Today, we have myriad of technologies available to do BDA but they are quite expensive, difficult to use and are very time-consuming. To resolve all such issue associated with BDA, Hadoop was introduced in 2005. And since then, most of the data scientists from data science field have accepted the power and benefits of Hadoop technology. This article puts light on the concept of Big Data Analytics done through Hadoop technology.

The section 2 of this article covers the understanding of Big Data concept while section3 discusses the most frequent technologies available today to do Big Data Analytics. The section 4 of this article discusses the evolution of Hadoop technology while section 5 discusses its architecture and the details of how Hadoop can be used to do Big Data Analytics. The section 6 of this article discusses the benefits of Hadoop technology to do BDA. And finally, section 7 wraps up the article with a conclusion.

II. Understanding Big Data Concept

The understanding for Big Data can't be possible without the knowledge of 3V model that lays the foundation for the research and analysis of Big Data. These 3Vs for Big Data include – huge volume (means Big Data have "a large amount of data"), high velocity (means Big Data have "high speed of data in and out"), and extensible variety (means Big Data have "large range of data types and sources"). On the basis of this understanding of Big Data, they can be distinguished into three types: Unstructured Data (related to Word, Text, PDF and Media Logs), Structured Data (related to relational data) and Semi-Structured Data (related to XML data).

Big Data has a wide range of applications. Some of the major applications of Big Data include:

A. The task of searching web content on the Internet through various search engines becomes easy and efficient due to the myriad algorithms of data science exploited by the search engines.

B. The information collected from the social media sites and community sites like suggestions, comments, views and queries, various online retailers do the eye catchy digital advertising to draw the attention of the customers to buy their products.

C. One can easily make a search for information through images. This can be done by uploading an image in the "search by image" feature of Google that uses the concept of Big Data to recognize the image for doing effective and fast image-based searching online.

D. Similar to image recognition, various speech recognition software techniques are available online to search for given information after recognizing the speech-based instructions given online by a user.

E. Various financial institutions around the world make use of Big Data and do analytics of them to help themselves to predict frauds, defaults, and risks that might take place while granting a loan to a person or a company by them.

F. Various Gaming application building companies make use of Big Data to enable a player to interact with a game in real time mode. It further helps one to track one's performance in a game through interactive mode and thus one can make choice of the best move out of all possible moves to play the game effectively.

III. Technologies to Harness Big Data

There are various technologies that are devised to use Big Data effectively, especially in decision-making. This benefits business houses and Internet giants to achieve the task of data processing and meaningful information derivation in an efficient manner with reduced cost as well as reduced risks. There are many honchos of Internet industry that uses such technologies for utilizing Big Data. Some of them include IBM, Amazon, Microsoft, Flipkart, Snapdeal, etc. These technologies are significant for the reason that they are helpful to do accurate analysis from Big Data to benefit the companies to derive information that may have some sense and utmost use. But to do so, the companies and organizations need to develop a robust, flexible and interoperable infrastructure to process and manage large data volumes of the unstructured, structured or semi-structured type in real time by ensuring data security and their privacy at the same time.

Big Data can be made operational in real time with interactive workloads with the help of technologies such as MongoDB that primary captures and stores data. Further, there are other technologies like NoSQL that engage themselves with cloud computing architectures to make the carrying out of hefty computations more cheaper and in more efficient manner. A NoSQL DB is a non-SQL or non-relational" database where data is modeled through means other than tabular relation forms as observed in relational databases. The benefit of NoSQL-based systems is that they are helpful to get deep insights of trends and patterns which are based on data of the real-time type that requires minimal coding and no need for any additional infrastructure or any data scientist.

Big Data need to be analyzed in order to get their hidden patterns, useful information and uncovered patterns derived. For Big Data analytics, wide ranges of technologies are under operation. The most frequently used

technologies among them include MapReduce and MPP (acronyms for Massively Parallel Processing). They can be considered as complimentary of SQL that are scalable from single servers to hundreds of thousands of low as well high machines. The most amazing fact is that MapReduce and MPP can easily be configured together to obtain best results from Big Data.

IV. Evolution of Hadoop to do Big Data Analytics

The traditional approach to harness Big Data and do their analytics was quite complex and cumbersome task. The approach can be understood in three simple steps:

A. Data are stored in the database of RDBMS (acronyms for Relational Data Base Management System) such as Oracle, MS SQL Server etc.

B. The costlier and heavier software are programmed to do interactivity with the Big Data stored in the database.

C. The tedious task of processing Big Data is carried out and are finally presented for analysis to users.

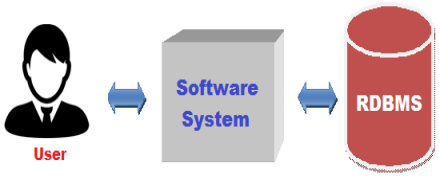

Fig 4.1: Traditional approach to handle Big Data

Although this approach is well enough for dealing data with small volume, the approach becomes unattractive and useless when it has to deal data with data with large volume.

The challenges that are posed by the traditional approach for doing Big Data Analytics is resolved through an algorithm developed and implemented by Google in the form of MapReduce. MapReduce works by dividing the Big Data Analytics task into smaller subtasks which are later put into different computers that are connected together through a robust and interoperable network. Finally, the results generated by all computer systems are collected together to form the final dataset of results that are delivered to users for their further usage.

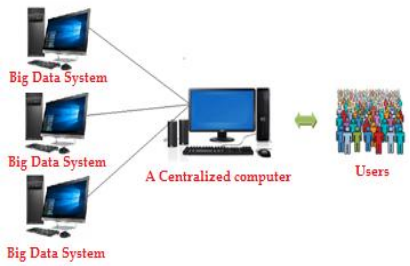

Fig 4.2: MapReduce: Google's solution to BDA

On the basis of the solution offered by Google in terms of MapReduce, a team of Mike Cafarella and Doug Cutting in 2005 developed HADOOP – an Open Source Project under Apache Software Foundation's registered trademark. It is an amazing fact that Doug named the project HADOOP following the name of a toy elephant of his little son.

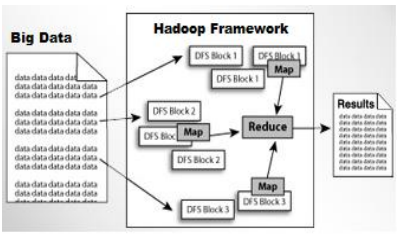

Fig 4.3: Working of Hadoop Framework

Hadoop follows MapReduce algorithm to build applications through Big Data that are processed from a large number of systems running in parallel. The results obtained are prepared under full statistical analysis in order to serve them to the users.

V. Understanding Hadoop Architecture

In order to understand the architecture of Hadoop, it is important to understand the underlying details of the two components lying at the core of Hadoop. These two components are:

A. **A Hadoop Distributed File System**: abbreviated as HDFS, this component is responsible for splitting and storing large sized files into hundreds or thousands of computer nodes. HDFS nodes store small data chunks that are termed as blocks which are later fed into MapReduce framework.

B. **A MapReduce Framework**: this component is responsible for processing each block. The processing of block is basically performed in two phases – first is to map blocks and then reduce them effectively. During map phase of data blocks, they are filtered and sorted whereas during reduce phase, they are aggregated for better results.

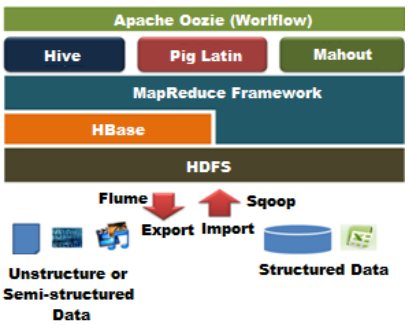

Fig 5.1: Block Diagram of Hadoop Architecture

The data blocks are enabled to import/export from/to HDFS through sophisticated tools called Sqoop and Flume.

Sqoop is responsible to import/export data from/to RDBMS whereas Flume is simply used to export data from unstructured file sources like logs.

HBase: it is considered as Hadoop database which is basically a very large data source that is highly scalable and widely distributed. This is very similar to Google's Big Table.

Hive: it is considered as one of the most important components of Hadoop's architecture for the reason that it plays the role of Hadoop's warehouse system. It is responsible for performing summarization of data in an easy manner, doing quick ad-hoc queries online and for carrying out large dataset analysis for the datasets stored in HDFS.

Apache Pig: it is responsible for doing large data sets analysis. The datasets basically use server-side scripting languages for explaining the programs related to data analysis.

Apache Mahout: this component of Hadoop architecture is considered as the machine learning library that helps to carry out the task of machine learning process to do data set analysis. It further helps to develop large scalable machine learning libraries.

Apache Oozie: this component is considered as the workflow (or sometimes coordination) system that is responsible for job management tasks of Apache Hadoop.

Working of Hadoop MapReduce

MapReduce of Hadoop architecture is responsible for easy application development that involves writing applications that can process Big Data on large clusters (sometimes hundreds to thousands of clusters) in parallel on Big Data system (sometimes referred to as commodity hardware). The Hadoop MapReduce perform this task in fault-tolerant and reliable manner.

MapReduce has two of the tasks associated with it:

A. Map: this task involves taking of input data HDFS that are then converted into datasets. These datasets are then broken into small chunks known as data blocks or sometimes tuples containing a key-value pair.

B. Reduce: this output produced from the map task is given as input to reduce task that converts the received datasets into much smaller tuples.

HDFS contains all the input and output of Hadoop. The scheduling tasks and the monitoring of them are the responsibility of MapReduce framework. Even, when the tasks get failed, the framework has the responsibility to re-executes them. Most importantly, the framework is configured as master-slave JobTracker for each cluster node. While the responsibility of master JobTracker is to do management of resources, the slave JobTrackers are fully responsible for tracking down the resources' availability and their consumption along with performing the task of job scheduling. Whatever tasks performed by a slave JobTracker are the tasks assigned to it by a master JobTracker. This entire mechanism of MapReducer has one major limitation that the failure of JobTracker may lead to the halting of all currently executing jobs.

How does Hadoop do Big Data Analytics?

Hadoop has fine and robust architecture to do Big Data Analytics (BDA). The BDA through Hadoop can be performed in three steps:

A. Hadoop receives a job request from a user or any application. The job request made for Hadoop consists of three items including the input-output files' location in HDFS, java jar file containing classes for achieving map and reduce tasks, and the job configuration related information that is obtained as distinct parameters for executing the job.

B. The job client of Hadoop then places the job request along with all necessary items and configuration to JobTracker. The JobTracker, on receiving the job request, simply initialize the master JobTracker to distribute the items and configuration of the job request to slave JobTracker so that it can perform the task of scheduling and monitoring job request. After this, the slave JobTracker simply returns the diagnostic and status related information to the client which had made job request earlier to Hadoop.

C. Finally, the TaskTrackers that are located on myriad nodes start the process of execution of the scheduled job received from MapReduce framework and simply stores the result obtained in the HDFS.

VI. BENEFITS OF HADOOP FOR DOING BDA

There are many reasons to choose Hadoop to carry out the task of Big Data Analytics (BDA). These are the following benefits that are offered by Hadoop while performing the task of BDA:

***A. Fast Computation*:** Due to the distributed computing feature provided by Hadoop, it is obvious that the computation of BDA becomes fast as compared to other technologies available.

***B. Adaptability*:** Hadoop is able to adapt to any size or kind of data – be it unstructured, structured or semi-structured. Moreover, it always makes prior treatment of Big Data before their storage in HDFS. One can have **choices** to store data of any size that one wishes.

***C. Most Economical*:** Since Hadoop is an Open Source Software Framework, one can understand how economical Hadoop can be.

***D. Scalability*:** Hadoop, as distributed computing system, has one significant benefit - it is scalable i.e. the nodes can be easily added to the Hadoop framework anytime with limited administration required.

VII. CONCLUSION

Big Data Analytics (BDA) is one of the most challenging tasks in the data science industry today. Many big organizations are using one technology or another to harness information from Big Data. Some of them who apply the traditional approach for doing BDA might have achieved a little while those who use the MapReduce algorithmic approach of Google to do BDA might have been able to get the best results out of Big Data. This can be understood from the fact that fast content searching over the Internet, digital advertising, speech recognition and image recognition to search data on the Internet and many more are the instances that show the effective BDA performed by big organizations like Google, Facebook, Twitter, Amazon, etc. to achieve all that they are known for today. And somewhere the reason is same, Google's MapReduce algorithm. But, the infrastructure needed to implement MapReduce algorithm to achieve BDA is not as effective, scalable, robust, efficient and economical as Hadoop is. This has been analyzed by us in our article.

We discussed how MapReduce is used in Hadoop to do splitting of Big Data into small data blocks through Map operation and then how data blocks are later scheduled and monitored by Reduce operation to carry out analytics of Big Data to achieve better results. We discussed the various benefits offered by Hadoop in achieving BDA. So, we can conclude that Hadoop is a fine technology to work with at present scenario for doing Big Data Analytics. But, if we look for other alternatives for doing BDA like crowdsourcing or crowdcomputing, it would be good for BDA and so for data science field.

VIII. REFERENCES AND BIBLIOGRAPHY

[1] White, Tom. P. 2012.Hadoop: The definite Guide. SPD O'Reilly.

[2] Turkington, Garry. P. 2013. Hadoop Beginner's Guide. SPD.

[3] Holmes, Alex. P. 2014. Hadoop in Practice. Second Edition. Manning.

[4] Agneeswaran, Vijay. P. 2014. Big Data Analytics Beyond Hadoop. Pearson Education Inc.

[5] Introduction to Apache Hadoop. 03/28/2017. Web. http://hadoop.apache.org/.

[6] A Complete Wikipaedia for Hadoop. May 2017. Web. https://en.wikipedia.org/wiki/Apache_Hadoop.

[7] A Basics of Hadoop. May 2017. Web. https://www.ibm.com/analytics/us/en/technology/hadoop/

[8] Figure 1 and 2: 19 May, 2017. Hadoop Big Data Solutions. Tutorials Point. Web.

[9] Figure-3: 19 May, 2017. Hadoop Tutorial. codegravity.com. Web.

[10] Figure-4: 19 May, 2017. Pradhan, Manaranjan. Enabling You For Cloud and Data. blog.enablecloud.com.

ABOUT AUTHORS

Manish Kumar Singh has published two research papers in international journals on web mining – chiefly "Web Mining: Penning an Era of Information Age" and "Understanding How Crucial Hidden Value Discovery In Data Warehouse Is?" and he has also published an article in international journal on the issue of net neutrality with title "Analysing Net Neutrality with Indian Netizens' Perspective".

GS Baluja is a famous Indian author of Computer Science field who has authored numerous books so far - chiefly Data Structure Through C, Data Structures Through C++, Object Oriented Programming Using C++, Java Programming & many more. He has done B.E (Com. Sci.) from Marathwada University and M.Tech from IIT, Delhi.

Dinesh Prasad Sahu received the Master degree (Computer Science & Application) M.Tech (Computer Science & Application) from Jawaharlal Nehru University, New Delhi, India. Currently, He is doing Ph.D. (Computer Science & Engineering) under the guidance of Dr. Karan Singh, from Jawaharlal Nehru University, New Delhi, India & is working in school of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi. His primary research interests include parallel and distributed system and Grid Computing. He has published 3 papers in proceedings of peer-reviewed Conferences.