International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

Big data – A Review

Dipti Shikha Singh[1], Garima Singh[2]

[1] *Student, Computer Science Department, Babu Banarasi Das University, Lucknow, U.P, India*
[2] *Assistant Professor, Computer Science Department, Babu Banarasi Das University, Lucknow, U.P, India*

Abstract - *The use of the Internet and various technologies worldwide, whether for social, personal or professional use, give rise to Big Data with an incredible speed. The Big data analysis has emerged as an important activity for many organizations. There is still a debate about the tools and traditional management frameworks are ineffective with Big Data. This document sheds light on many of these documents that help us with the idea of Big Data and new technologies that help Big Data. Also, we discuss the challenges that increase the use of large data while trying to get the right approach to get valuable information from large data stack.*

Key Words: Hadoop, Big Data, 5V's, Hive, Pig, etc.

1. INTRODUCTION

Traditional data sources such as business data, sensor data generated automatically, social data and data from billions of devices such as mobile phones, smartphones, laptops, cameras, and pictures are a wealth of information to create. A few years ago the data were measurable in megabytes and gigabytes, while today data are measured in terabytes and petabytes. With this growing momentum to come more in the future. The current data rate is estimated with approximately 1,000,000 terabytes [1] that is 2.5 exabytes of data per day. The sources of these data vary from a variety of data sources, including sensors that transmit meteorological data, generated data from social networking sites like Facebook and Twitter, and digital content sites such as YouTube [1]. Gone are the days when the data were generated by the people and usually recorded in tabular form. Now the challenge is how to transform these unstructured data into information. Various challenges arise when using Big Data deal with an application requiring unstructured data for management and provide near real-time analysis, along with fault tolerance. In addition, you must have high storage and processing capacities. The great variety and large data set sizes are becoming impractical for tools and applications of traditional data management. Therefore, Big Data requires a new set of applications, tools, and frameworks for themselves.

2. BIG DATA: DEFINITION AND CHARACTERISTICS

The word Big Data seems to rule the data on the basis of the size to define is not limited to a certain extent, but it is a solution to analyze the data in order to make sense and its value for valuable information to use. The massive size of these data goes beyond petabytes and exabytes of data, even if the volume, speed, and various data on the storage capacity of an organization are calculated. Doug Laney defined the 3V model in 2001, characterizing Big Data with respect to the three V. The three basic characteristics of the data are large-volume, variety and velocity. Many organizations and professionals have expanded this model 3V to 4V model with a new "value" of "V". While the extension of the model 4V to 5V is by the concept of veracity [1, 3].

• Volume: Refers to the size of the data. Along with the growth of social media, the data volume is also growing very fast. Large amount of data generated by machines and surpasses the man-made data. Therefore, the occurrence of data size is known as the large data volume.

• Velocity: Refers to the speed at which data is generated. In today's competitive world, decision makers want information to provide important data in a fraction of a second in real-time. Twitter Tweets, status updates / likes / shares in Facebook, etc.

• Variety: refers to the different formats in which data is generated. 70% of the data generated today is in an unstructured manner. Earlier the development of Big Data, the industry did not have powerful management tools to manage unstructured data. The competition between the organizations was not only due to semi-structured data but also unstructured data like the traditional tables, flat files, relational databases and unstructured data stored as images, audio, web logs, sensor data, etc.

• Value: refers to the ability of companies to analyze data and to provide a better understanding of the various key areas that include customer behavior, provide personalized services, and provide information about problems that do not access previously. Therefore, the value can be viewed as the monetary value in a company or an organization that includes a data technology.

• Veracity: Refers to the accuracy or truth of the data. Uncertainty in the data can be caused for various reasons in the data, which may be legal questions, privacy issues, duplication, etc.

3. RECENT STATISTICS RELATED TO BIG DATA

Every month around 1.86 million users use Facebook, 317 million users use twitter, 467 million users are on LinkedIn, 1 million users use YouTube, while YouTube is watched by 4 billion a day. Therefore, large data come mainly from large companies such as Facebook, Instagram, Netflix, Paytm, Uber and much more. To get the hidden patterns and many other useful information from the sea of Big Data is a need for large data analysis. It is important to use, integrate, and adapt Big Data analytical techniques to new trends that evolve in the Big Data paradigm [4].

4. DIFFERENCE BETWEEN BIG DATA AND TRADITIONAL DATA

There are various differences between the traditional data and big data that can be seen in the table 1 below.

Table -1: Difference between traditional and big data

FEATURES	RELATIONAL DATABASE	BIG DATA
Database architecture	Centralized	Distributed
Data types	Structured	Semi-structured, unstructured
Data volume range	Gigabytes to terabytes	Terabytes, petabytes and beyond
Data schema	Fixed or static	Dynamic
Hardware / software cost	Higher	Lower

5. HADOOP: THE BIG DATA NEED

One of the fundamental problems with Big Data is storage. There are several approaches to managing this problem, including Hadoop and Hive, Pig, Hbase, Zookepear, Flume, Oozie, Sqoop, and so on. Several vendors show great interest in the integration of these tools with their own product because it is an open source and has become a standard for many companies.

Apache Hadoop is a Java open source implementation of MapReduce. Hadoop consists of two layers: a data storage layer Hadoop Distributed file system called (HDFS) and a layer called MapReduce for data processing. HDFS is the memory area, while MapReduce is the edit area. It is designed to run parallel processing of a large number of records distributed across clusters of computers with simple programming models that can scale to hardware products from individual nodes to thousands of computers. Thus cluster, in which each node in the cluster provides local computing and memory area. Rather than offering high availability on hardware, it is possible to identify its own framework and to calculate the single point of failure in the application layer, which provides high availability to the end user. It has a number of commercially supported distributions from companies like Cloudera, Hortonworks, etc.

Hadoop operates on the master and slave architecture in a name or master node, and with multiple data node or slaves. HDFS is a data processing system with high fault tolerance, reliable and cost-effective, designed to run on low-cost hardware and storage terabytes (TB) and petabytes (PB) distributed by unstructured data without any problems. MapReduce, acts as software large amounts of data with two main functions map and reduce processing. Map the data in <key, value> pair and generate the intermediate value, while reduce the final output completion of intermediate value generated by the map function. The workflow map and reduce is made up of mapping, sharing and transforming. Hadoop few restrictions include the processing time of the CPU, power consumption, hoping that all the map jobs are completed (or not or skip) only then you can start to work on reduce, etc.

6. SUB-PROJECT SUPPORTING HADOOP: A LITERATURE REVIEW

Some implementations of the latest popular software, which are often used to develop MapReduce-based systems and applications, are Apache Pig and Apache Hive. It aims to enable declarative query languages for the MapReduce Framework to support the independence of consultations, the reuse of queries, and automatic query optimization. Facebook came up with the solution known as Hive, the same year, Yahoo developed Pig. The intention of the Hive and the Pig was to bring simplicity to the complex MapReduce code. Both Hive and Pig is an open source solution built on Hadoop. Hive is a data warehouse that supports SQL queries such as HiveQL or HQL, which are compiled in the map reduction work and are run with Hadoop. Pig is a data flow language generated in Pig Latin. Unlike Pig, in Hive schema is required.

Pig is an open source project designed to support the ad-hoc analysis of very large data, a scripting language for Google MapReduce. Pig Latin supports nested data models and a set of predefined UDFs that can be customized accordingly. The pig translation frame first generates a logical query plan, then creates the logical plan in a series of MapReduce jobs. It is built on the framework of Hadoop and Hadoop need not be changed.

Hive is an open source project that aims to provide data warehouse solutions on the Hadoop and supports ad hoc queries. Hive creates an HQL query in a directed acyclic graph (DAG) of MapReduce Jobs. HQL includes the data definition language (DDL) for managing data integrity and system catalog, which contains schema information and statistics as DBMS engines. Currently, Hive only offers a simple and naive based rules optimizer.

YSmart aims to create a generic framework to translate into optimized SQL queries to create and run

MapReduce jobs on a large scale efficiently distributed cluster systems. YSmart can be combined into Hive for better performance, and can also be a SQL-to-MapReduce translator.

In some cases, several similar queries, common tables, and combination tasks come simultaneously, many opportunities that are sharing computing work together. Performing common tasks can only significantly reduce the overall execution time of a query batch. Therefore, SharedHive is a framework to optimize multiple queries that works to improve the overall performance of Hadoop. SharedHive converts a series of HQL queries into a new set correlated queries from generated output within a shorter implementation time.

7. CONCLUSIONS

It is now a phase for Big Data development. This article focuses on the concept of Big Data with 3V to present the boundaries and challenges in big-data processing. To get out of Big Data limitations these challenges have to be met. The document describes various advantages and disadvantages of Hadoop as a tool for large data management. Although Hadoop with its ecosystem is a powerful solution to deliver big data but does not yet sound good for frequent data changes. In other words, we can currently say that there is no transactional support in Hadoop. Hadoop is only used for OLAP. The machine learning algorithm for Big Data needs to be more robust and easier to use. Therefore, a further improvement is required in the Big Data solution.

REFERENCES

[1] Gema Bello-Orgaza, Jason J.Jungb, David Camachoa, “Social big data: Recent achievements and new challenges” ,in press.

[2] Tansel Dokeroglu, Serkan Ozal, Murat Ali Bayir, Muhammet Serkan Cinar and Ahmet Cosar, “Improving the performance of Hadoop Hive by sharing scan and computation tasks”, Journal of Cloud Computing Advances, Systems and Applications 2014.

[3] Dr. Narasimha Rao Vajjhala, Dr. Ervin Ramollari, “Big Data using Cloud Computing – Opportunities for Small and Medium-sized Enterprises”, European Journal of Economics and Business Studies, Jan-Apr 2016 Vol.4 Nr. 1.

[4] “(February 2017) How Many People Use the Top Social Media, Apps & Services?” Craig Smith, Accessed: http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/

[5] “Difference between traditional data and big data”, Deepali Aggrawal, Accessed: http://www.projectguru.in/publications/difference-traditional-data-big-data/

[6] Rubao Lee, Tian Luo, Yin Huai, Fusheng Wang, Yongqiang He, Xiaodong Zhang, “YSmart:Yet Another SQL-to-MapReduce Translator”, Proceedings of 31st International Conference on Distributed Computing Systems (ICDCS 2011), Minneapolis, Minnesota, June 20-24, 2011.

[7] Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, Bongki Moon, “Parallel Data Processing with MapReduce: A Survey”, SIGMOD Record, December 2011 (Vol. 40, No. 4).

[8] Hadoop Wiki, "Apache Hadoop", Accessed: http://wiki.apache.org/

[9] Palak Gupta, Nidhi Tyagi, “An Approach towards Big Data-A Review”, International Conference on Computing, Communication and Automation (ICCCA2015), ISBN: 978-1-4799-8890-7.