RESEARCH ARTICLE OPEN ACCESS

A Comparative Analysis of Big Data Tools

Anjali Mehta
Assistant Professor
Department of Computer Science
Aggarwal College, Ballabgarh
Faridabad – India

ABSTRACT
Big data is the term for any cluster of datasets so giant and complex that it becomes challenging to practice using traditional data mining tools. Big data is a set of techniques and technologies that require new methods of integration to uncover large unknown knowledge from large datasets that are various, complex, and of massive scale. Big Data is speedily growing in all science and engineering fields due to fast expansion of networking, data collection capability and its storage. Analysing Big Data is a tough task as it contains huge distributed file systems which should be fault tolerant, flexible and scalable. The technologies used by big data application to handle the massive data are Apache Hadoop, Map Reduce, Apache Hive, HPCC, IBM Info sphere. In this paper we discuss about these tools one by one.

Keywords:- Big Data, Apache Hadoop, Fault Tolerant, Scalable, HPCC

I. INTRODUCTION

Big data is an evolving term that describes a huge volume of structured, semi-structured and unstructured data that has the potential to be mined for information and used in machine learning projects and other advanced analytics applications.

Big Data refers to large amounts of data and hence it becomes very problematic to analyse and arise knowledge from it. This data is exceptionally fast, large and difficult to process using traditional tools of data mining. It is combined form of structured and unstructured data. Being able to process every piece of data in a realistic time would remove time overheads and may even generate unexpected discoveries. Such large volume of data is the outline of our digital life and we can put our blueprints to use to collect meaningful knowledge. The big data is generated mostly from IT, social networking sites, emails, magazine articles and newspapers, and blogs covering the entire World Wide Web. Multimedia data is the backbone of the traffic over the internet and is expected to increase 70% by 2020. Only Google has got more than 1 million servers around the worlds. There have been 6 billion mobile users in the world and 10 billion text messages are sent daily. By the year 2020, 50 billion devices will be linked to networks and the World Wide Web. In this paper, section 2 defines 6V"s of big data; section 3 describes big data mining tools namely IBM Info sphere, Apache Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. Then, in section 4, we conclude the results.

II. CHARACTERISTICS OF BIG DATA

Big data is often characterised by the 3Vs: the acute volume of data, the extensive variety of data types and the velocity at which the data must be handled. Those characteristics were initially known by Gartner analyst Doug Laney in a report printed in 2001. More recently, several other V's have been added to descriptions of big data, including veracity, value and variability. Although big data does not equate to any specific volume of data, the term is usually used to describe terabytes, petabytes and even Exabyte of data captured over time. Some people ascribe even more V's to big data -- data scientists and consultants have created various lists with between seven and 10 V's.

Fig. 1

A. Velocity

First let us talk about velocity. Clearly, velocity refers to the speed at which massive amounts of data are being produced, composed and investigated. Daily the number of emails, twitter messages, photos, videos, audios etc. increases at very high speeds around the world. Every second of every day data is increasing. Not only must it be analysed, but the speed of transmission, and access to the data must also remain rapid to allow for real-time access to website, credit card verification and instant messaging. Big data technology permits us now to analyse the data while it is being produced, without ever putting it into databases.

B. *Volume*

Volume refers to the absurd amounts of data generated each second from social media, mobiles, cars, credit cards, M2M sensors, images, videos, etc. The huge amounts of data have become so large in fact that we can no longer store and analyse data using old data mining techniques. We now use distributed systems, where parts of the data is stored in different-different locations and organized by software. With Facebook alone there are 10 billion messages, 4.5 billion times that the "like" button is pressed, and over 350 million new pictures are uploaded every day. Gathering and examining this data to convert into valuable knowledge is clearly an engineering challenge of immensely vast proportions.

C. Value

When we talk about value, we are mentioning the importance of the data being extracted. Having limitless amounts of data is one thing, but unless it can be turned into value it is of no use. While there is a clear connection between data and insights, this does not always mean there is importance in Big Data. The most important part of embarking on a big data initiative is to understand the costs and benefits of collecting and analysing the data to ensure that ultimately the data that is obtained can be monetized.

D. Variety

Variety is defined as the different types of data we have. Data today looks very different than the data from the past. We no longer just have structured data (name, phone number, address, financials, etc.) that fits easily and neatly into a data base. Today's data is unstructured. In fact, 80% of the entire world's data fits into this category, including images, video clips, social media updates, etc. New and advanced Big Data technology is now allowing structured and unstructured data to be collected, stored, and used all together.

E. Veracity

Veracity is the quality or consistency of the data. Just how accurate is all this data? For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc., and the trustworthiness and accuracy of all that content. Lots of data is of no use if the quality or reliability is not accurate.

F. Variability

This also often applies to sets of Big Data, which are less dependable than conventional transaction data and may have numerous meanings or be organized in different ways from one data source to another -- things that further complicate efforts to process and analyse the data.

III. TOOLS OF BIG DATA

A. APACHE HADOOP:

Apache Hadoop is written in java and an open-source software framework for distributed storage and distributed processing of huge data sets on computer clusters built from hardware. The core of Apache Hadoop consists two parts-

1. *Storage part*- Storage part is also known as Hadoop Distributed File System (HDFS).
2. *Processing part*- This part is known as MapReduce.

Hadoop divides files into big blocks. It then allocates these blocks to the nodes in the cluster. HDFS is the last part of Hadoop software stack which is a distributed file system. In HDFS, each file appears contiguous sequence of bytes. Hadoop map reduce system which is the processing part forms the centre layer of the stack and it applies map operations to the data in partitions of an HDFS file, sorts and redistributes the results supported key values within the output data and then performs reduce operations on the collections of output data 0items with identical keys from the map segment of the job.

Hadoop is:

Reliable: The package is fault tolerant; it expects and handles all the hardware and software failures

Scalable: Designed for enormous scale of processors, memory, and connected storage

Distributed: Handles replication. Offers massively parallel programming model, Map Reduce.

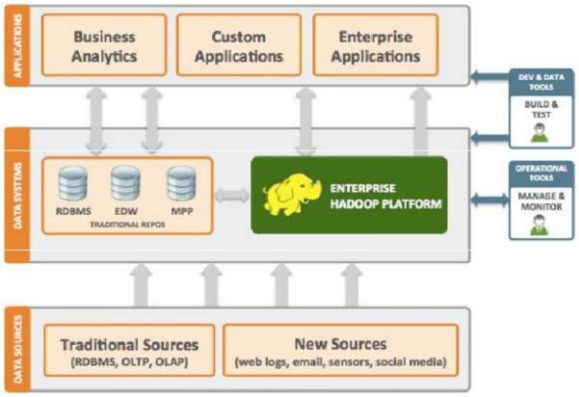

Fig. 2 Architecture of Hadoop System

B. *MAPREDUCE*:

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. To functions are present in this model-

The map() function- This produces filtering and sorting

The reduce() function- performs a summary operation.

By enhancing the execution engine just the once, map reduce programming model have the characteristics of scalability and fault-tolerance. MapReduce libraries are coded in different programming languages, with required optimization. The name MapReduce was initially referred to the technology which was proposed by Google. By 2014, Google were no longer using MapReduce as a big data processing model, and development on Apache Mahout had moved on to more capable and less disk- oriented mechanisms that incorporated full map and reduce capabilities.

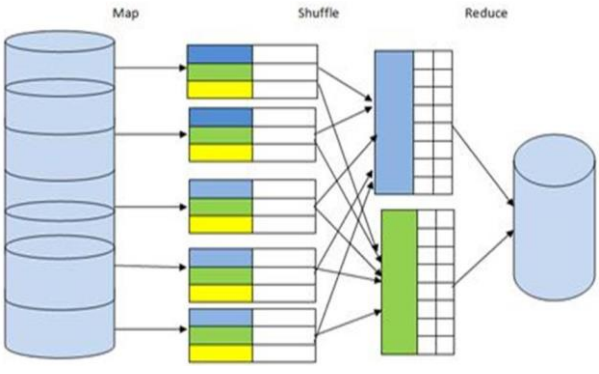

Fig. 3 Architecture of Map Reduce

Advantage- In map reduce large variety of problems are easily expressible as Map reduce computations and group of machines handle thousands of nodes and their fault-tolerance.
Disadvantage- In map reduce we need Real-time processing which is not always very easy to do.

Success of map reduce is based on many reasons given below-

1. The model is very easy to use. Programmers without any skill with parallel and distributed systems can use it with easiness as it hides the facts of parallelization, load balancing, locality optimization, and fault tolerance.
2. A huge set of problems with different nature are simply expressible as MapReduce computations.
3. MapReduce can implement on large collections of commodity hardware.

C. *HIVE* :

Hive is a distributed agent platform which is a decentralized system for constructing applications by using networking local system resources. Apache Hive data warehousing component which is based on cloud Hadoop ecosystem and offers a query language called Hive QL. Hive QL is used to translate SQL-like queries into Map Reduce jobs spontaneously. SQL, Oracle, IBM DB2 are the applications of Hive.

D. *HPCC*:

HPCC is an open source platform used for computing and that delivers the facility for handling of huge big data work flow. HPCC system is the only platform having a single architecture and only one programming language which is used for the data simulation. HPCC system was designed to investigate the massive amount of data for the determination of solving difficult and complex problems of big data. HPCC system is built on enterprise control language which has very declarative and on-procedural nature.

The main components of HPCC are:

HPCC Data Refinery: Parallel ETL engine mostly used in this.

HPCC Data Delivery*:* It is hugely based on structured query engine used in it. Enterprise Control Language divides the workload between the nodes in proper and even load.

E. *IBM INFOSPHERE*:

The InfoSphere software is used to handle the data which is very important now a day in any business and this software includes several functional modules like data integration, data warehousing, and management of data, big data and at last information governance. This platform offers an enterprise-class base for different big data projects, offering the exceptional performance, scalability, consistency and precision required to beat many challenges and deliver useful and trusted information which we can say knowledge to your enterprise quicker. IBM InfoSphere is an ETL tool which is the part of the IBM InfoSphere and IBM Information Platforms Solutions suite. It is very useful software which is used to construct numerous data integration solutions. It has different versions which include the Server Edition, the Enterprise Edition, and the MVS Edition.

Fig. 4

IV.CONCLUSION

In this paper, we have discussed about the big data, 6 V"s of big data. Also we have got done a comparative study of various tools on which we will convert unstructured data to structured data. The main objective of this comparison isn't to condemn which is the best tool in big data, but to demonstrate its usage and to create alertness in numerous fields. Apache Hadoop is employed to handle the big data and other interrelated projects of Hadoop. Map reduce programming model has been with success used at Google for several different functions for many years. Also The InfoSphere Platform offers all the foundational building structure of reliable information which includes data integration, data warehousing, master data management, big data and information governance but it needs large amount of RAM to run it on a personal computer.

The main objective of this paper was to create a survey of different big data handling methods those handle a huge quantity of data from totally different sources and improves overall performance of systems.

REFRENCES

[1] Mikin K. Dagli, Brijesh B. Mehta, "Big Data and Hadoop: A Review" in IJARES, ISSN: 2347-9337(Online), Volume: 2(Issue: 2), Pg. No. 192, Feb, 2014.

[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding "Data Mining With Big Data", in IEEE Transactions on Knowledge and Data Engineering, IEEE, ISSN: 1041-4347, Volume: 26(Issue: 1) Page No. 97-107, 2014.

[3] Abdul Raheem Syed, Kumar Gillela, Dr. C. Venugopal, "The Future Revolution On Big Data", In International Journal of Advanced Research in Computer and Communication Engineering, e- ISSN: 2278-1021, p-ISSN: 2319-5940, Volume: 2 (Issue:6) , Page No. 2446- 2451,2013.

[4] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions", In RGPV, Page No.269-276, 2013.

[5] Amogh Pramod Kulkarni, Mahesh Khandewal, —Survey on Hadoop and Introduction to YARN .International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[6] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.

[7] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N —Analysis of Big Data using Apache Hadoop and Map Reduce Volume 4, Issue 5, May 2014.

[8] Suman Arora, Dr.Madhu Goel, —Survey Paper on Scheduling in Hadoop. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[9] Jimmy Lin —Map Reduce Is Good Enough? The control project, IEEE Computer 32 (2013)

[10] http://www.thefullwiki.org/IBM_InfoSphere_DataStage.

[11] http://rootshellinc.com/index.php/services/ibm-infosphere- services.

[12] https://en.wikipedia.org/wiki/MapReduce.

[13]http://www.bigdatanews.com/profiles/blogs/getting-real-about-big-data-from-roi-to-insightswhat-can-it.