Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.789 – 793

A Survey on Comparative Analysis of Big Data Tools

Mr. Piyush Bhardwaj[1], Abhishek Gupta[2], Mahima Sharma[3], Megha Gupta[4], Surabhi Singhal[5]

[1,2,3,4,5]Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, Delhi, India

Abstract— Big Data concerns large- volume, complex, data sets with multiple, autonomous sources which are growing rapidly. Big Data is rapidly expanding in all science and engineering domain due to fast development of networking, data collection capacity and its storage. The use of Big Data underpins critical activities in all sectors of our society. Apache Hadoop and IBM InfoSphere are used to analyze such vast amount of data. In this paper, we compare these two tools.

Keywords— Big Data, Hadoop, Volume, Veracity, structured data, unstructured data

I. INTRODUCTION

Big Data refers to large chunks of data and hence it becomes very difficult to analyze and derive knowledge from it. This data is extremely fast, large and difficult to process using conventional tools of data processing. It is made up of structured and unstructured information. Being able to process every item of data in a reasonable time would remove time overheads and may even generate unexpected discoveries. Such large amount of data is the blueprint of our digital life and we can put our blueprints to use to gather meaningful insights. The big data is generated mostly from the information technology industrial work, social networking sites, emails, magazines and newspapers, and blogs covering the entire World Wide Web. In the past, human genome decryption process takes approximately 10 years, now not more than a week. Multimedia data have big weight on internet backbone traffic and is expected to increase 70% by 2013. Only Google has got more than one million servers around the worlds. There have been 6 billion mobile subscriptions in the world and every day 10 billion text messages are sent. By the year 2020, 50 billion devices will be connected to networks and the internet[1]. In this paper, section 2 consists of problem statement, section 3 defines 5V's of big data, section 4 describes data mining tools namely, IBM Infosphere and Apache Hadoop and consists of the comparison of these two tools. Then, in section 5, we conclude the results.

II. PROBLEM DESCRIPTION

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years[2].

Due to Big Data, new bets are being formed by Netflix. Netflix analyses 30 million "plays" a day, including when it is paused, rewind and fast forward, four million ratings by Netflix subscribers, three million searches as well as the time of day when shows are watched and on what devices. Netflix stores three copies of every movie, more than ten years of user ratings and then extensive user account information. It includes log files, subtitles, and audio files as well. The show 'House of Cards' was a result of analysis of this data and generated a profit of millions of dollars with the lead and direction as Kevin Spacey and David Fincher respectively, eventually resulted a big hit[8]. Hence, the data has a lot of power to lead us to knowledge discovery. Conversion of unstructured data to structured data can lead us to unexpected discoveries. The main problem that we face today is how to convert unstructured data to structured data. The concept of structured data plays a vital role in order to extract some important information from the pool of data.

The term Unstructured Data refers to the information which is scattered and is not organized in a proper format. Such kind of data can't be fitted in the database relational tables. Unstructured data is human 'information' like emails, videos, tweets, Facebook posts, call center conversations, closed circuit TV footage, mobile phone calls, website clicks. Unstructured data is considered as "loosely structured data" because the data sources possesses a structure but all the data within a dataset do not have a particular predefined structure[3]. Unstructured data is a data that does not have a pre- defined data model, so it can't be analyzed.

Everyday millions of likes, comments, posts are generated on Facebook, millions of tweets on twitter, likes on Instagram and a lot of people surf Amazon, Flipkart, eBay, Myntra and other sites. All this consists of unstructured data. If we try to analyze such kind of data, a lot of inconsistencies are generated. It would be beneficial if we convert such data to structured data and analyze that data to gain some knowledge from it. Thus, these online social networking sites and companies uses this technology to convert unstructured data to structured data in order to evaluate their performance and profits.

III. 5V'S OF BIG DATA

VOLUME: - Volume refers to the enormous amount of data generated every second. Volume constitutes of myriad emails, video clips, multimedia data, Facebook and twitters messages which are produced and shared every second by people. Volume of data now is larger than terabytes and petabytes.

VELOCITY: - The term velocity refers to the speed with which data is produced and processed to meet demand. On New Year's Eve, the speed at which data is generated increases asymptotically jamming the whole network. This depicts the alarming velocity at which the new data is generated.

VERACITY:-Veracity refers to the messiness or trustworthiness of the data. Due to many forms of big data, quality and accuracy can't be controlled much, for example Twitter posts with hash tags, abbreviations, typos and colloquial speech. Big data and analytics technology allows us to work with veracious data. The volumes often make up for the lack of quality or accuracy.

VARIETY: - Variety refers to the different types of data we can now use. In the past we focused on structured data that neatly fits into tables or relational databases such as financial data (for example, sales by product or region). In fact, 80 percent of the world's data is now unstructured and therefore can't easily be put into tables or relational databases—think of photos, video sequences or social media updates. With Big Data technology we can now harness differed types of data including messages, social media conversations, photos, sensor data, and video or voice recordings and bring them together with more traditional, structured data.

VALUE: - Value refers to our ability turn our data into value. Cost is one major factor that all organizations need to look into when it comes to big data implementation, or for that matter any software package or framework implementation. The initiative of entering into Big Data is a very critical and the Value needs to be consciously deliberated in the value to price perspective. A thorough understanding of Big Data is a must and the usefulness via Value of big data initiative to the organization needs to be clearly put down on paper.

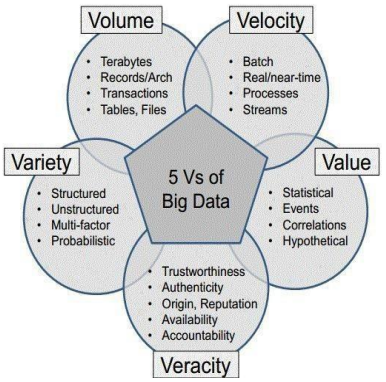

Fig 1. 5v's of Big Data

IV. DATA MINING TOOLS

IBM INFOSPHERE: - The InfoSphere Platform is used to manage the data which is very important now a days in any business and this platform includes various functional modules such as data integration, data warehousing, master data management, big data and information governance[6]. The platform provides an enterprise-class foundation for different big data projects, providing the excellent performance, scalability, reliability and accuracy needed to beat various challenges and deliver useful and trusted information to your enterprise faster.

IBM InfoSphere DataStage is an ETL tool. It is part of the IBM InfoSphere and IBM Information Platforms Solutions suite. It is a very useful platform which is used to build various data integration solutions. It has various versions which include the Server Edition, the Enterprise Edition, and the MVS Edition [5].

FIG 2.IBM INFOSPHERE

APACHE HADOOP: - Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop splits files into large blocks. It then distributes these blocks amongst the nodes in the cluster. HDFS is at the bottom of Hadoop software stack which is a distributed file system. In HDFS, each file appears adjacent sequence of bytes. Hadoop map reduce system forms the middle layer of the stack and it applies map operations to the data in partitions of an HDFS file, sorts and redistributes the results based on key values in the output data and then performs reduce operations on the groups of output data items with matching keys from the map phase of the job .

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[7].

The map() and reduce() are the two functions of this model. The map() function produces filtering and sorting and reduce() function performs a summary operation. By optimizing the execution engine once, map reduce programming model supports features namely, scalability and fault-tolerance. MapReduce libraries have been written in different programming languages, with necessary optimization. The name MapReduce was initially referred to the technology which was proposed by Google, but has since been genericized. 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 [8].

V. COMPARISON OF BIG DATA TOOLS

	Apache Hadoop	IBM InfoSphere
Mode of software	Open source and free source	Commercial
Type of Data	Unstructured data, time series, textual data	Unstructured data, semi structured data and structured data
Data Sources	Files, the network scripted Output	IBM Warehouse
Database Support	HBASE, Sybase, SAP	Mongo DB, DB2, Oracle
Operating System	Windows, Linux	Windows

The above table demonstrates the comparative aspects of the two chosen tools in big data by us based on compatible data sources and its operating system. The main objective of this comparison is not to criticize which is the best tool in big data, but to demonstrate its usage and to create alertness in various fields.

VI. CONCLUSION

In this paper, we have discussed about the big data, significance of big data, the problem of unstructured data which is generated in big data and 5 V's problems of big data. Also we have done a comparative study of different tools on which we can convert unstructured data to structured data. The main objective of this comparison is not to criticize which is the best tool in big data, but to demonstrate its usage and to create alertness in various fields. Apache Hadoop is used to process the big data and other related projects of Hadoop. Map reduce programming model has been successfully used at Google for many different purposes. Success of map reduce is based on various reasons. First, the model is easy to use. For example, the programmers without any experience with parallel and distributed systems can use it with ease as it hides the details of parallelization, load balancing, locality optimization, and fault tolerance. Second, a large set of problems with veracious nature are easily expressible as MapReduce computations. Third, MapReduce can implement on large clusters of commodity hardware. Also The InfoSphere Platform provides all the foundational building structure of trusted information, including 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 system.

REFERENCES

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

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

[7] https://en.wikipedia.org/wiki/MapReduce.
[8]http://www.bigdatanews.com/profiles/blogs/getting-real-about-big-data-from-roi-to-insights-what-can-it.

