INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

A Review on Big Data, Hadoop and its Impact on Business

Kankana Kashyap
School of Computing Sciences, Kaziranga University, Jorhat, Assam, India
Champak Deka
School of Computing Sciences, Kaziranga University, Jorhat, Assam, India
Sandip Rakshit
School of Computing Sciences, Kaziranga University, Jorhat, Assam, India

Abstract:
"We are drowning in data, but starving for knowledge!" John Naisbett
Big data is defined as a large amount of data which continues to grow so much that it is difficult to manage and requires new technologies to store and analyse these data. Due to such enormous explosion of data it becomes difficult to manage data with the help of traditional techniques. This is a review paper which gives the basic introduction of big data, it properties, traditional data analysing technique like relational database management system (RDBMS) and Hadoop as the latest technology for analysing big data. The paper also includes the impact that Hadoop has upon business analytics.

Keywords: *Big data, RDBMS, Hadoop*

1. Introduction

We are living in a world where data is like a vast never ending ocean; Data goes on increasing day by day. The invention of new technologies has lead to usage of a large amount of data. As the amount of data is going on increasing, there is a need to conserve, process, and extract these data. This is what leads to the term called the "BIG DATA". From the word itself we know that it refers to an amount of data which is *"BIG"*. But the term big data is not only confined to this, it has a more vast meaning to it. Big data can be described as *much bigger, faster and much harder* [2]. It is defined as a large amount of data, which continues to grow so much that it is difficult to manage and require new technologies and architectures to store, and analyse this data. Big data is a challenge that can be overcome by 5 V's. Data *volume*, it refers to the increasing amount of data and the data that is available. Data *velocity* refers to the speed of data i.e. data is not only big but also needs to be processed quickly. Data *variety* refers to the different types of data and their representation. Data *value* measures the usefulness of making decisions for the data and data *veracity* refers to the messiness of data.
A decade before, 'big data' was just a term which was not known to many. Back then, data was not created in such a spine-chilling rate as it is now. Data created in a span of five or more years is created in a mere 5 minutes today. For storing small data sets, traditional techniques like RDBMS (SQL, oracle) were just about enough. But, at the rate in which data volume is growing RDBMS is no match for it. This database is not enough to hold such enormous volumes of data. Big data is not only about getting a big database but also about constructing new methods for constructing and analysing such data. The creation of new tools and techniques to overcome the dilemma of rising data has led to the use of a new database management framework such as Hadoop [4].

2. Traditional Technique of Analysing Data

There is a tremendous difference in the amount of data that is created today and before. There are two types of data structured data and unstructured data. *Structured data* are those types of data that are pre-processed and organized so that it becomes easier to use it for queries in a database system. *Unstructured data*, as the name suggest does not have any structure to it, it is not in an organized manner i.e. all the data acquired is in a perplexed manner like plain text, video files, image file ,web logs ,social media, GPS files. [6]
Database management system is a traditional technology to store and analyse data. A DBMS is a software system which provides a database to define, create and store smaller sets of structured data. A RDBMS is a database management system where we find a relation between the tables. The difference between both is that DBMS stores file in form of flat-files and in RDBMS a single file can stored in several tables. Practically, both are almost considered to be same. [13].The RDBMS was created to that it makes applications, development and maintenance of business. RDBMS is used for the analysis of structured data; this is a reason because of which it is not much preferable.

3. New Technology for Analysis of Big Data

3.1. Hadoop

Hadoop is a program that was developed by Apache. Hadoop was developed because the traditional data warehouses were not enough to store the massive amount of data. It is open-source software, which helps in the processing and managing of very huge data sets. Hadoop is a framework which consists of different parts i.e. clusters; these have their own individual work. The main task of Hadoop is to store big data and it helps to increase the rate of processing of data [5].

Hadoop basically has two core components:

3.1.1. HDFS: Hadoop Distributed File System

Hadoop distributed file system is actually like municipal dustbin. All the garbage is dumped in the same place from where the garbage is divided into degradable and non-degradable by collecting and managing and then sent for recycling. In a very similar way, all the structured and unstructured data is dumped in the HDFS. The data waits in the HDFS until it is analysed. Anything can be done with the data till it is in the HDFS like storing, collecting or analysis of the data.

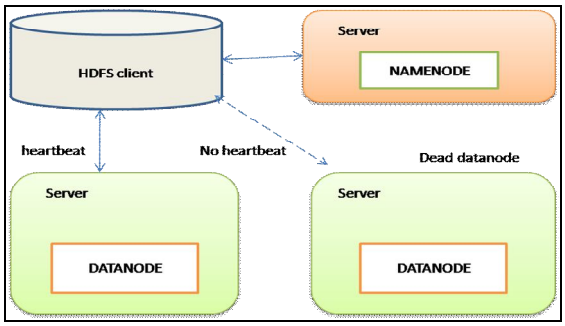

Figure 1: Two functions of HDFS

HDFS is basically a master/slave structure. The master part being the name node and slave being the data nodes.

The main work of the name node is that it stores data over data; the location of the files, and also different attributes of the documents. The data nodes store the content of the many blocks of files stored in different nodes. When the hdfs is implemented then the data node has to send a message to the name node to specify its startup and unique ID. There is only one name node and various data nodes in the hdfs design. The name node doesn't connect directly with the data nodes but instead it connects with the servers the data nodes are sitting in. The name node helps in accessing the files or the queries that are sent by the clients. The name node has a file system and its properties are recorded in the name node. There are a number of copies made of the file and is replicated in all the data nodes present. This duplication of files is managed by name node [14]. The message to be sent by data node during its startup is called heartbeat. If a data node doesn't send a heartbeat then it is said to be lost and blocked by the name node; if the name node has a failure in it then all the files of the file system is lost [15].

3.1.2. Map Reduce

Google introduced a java-based programming model to generate large data from several nodes, which is known as Map reduces.

Map Reduce is two different parts. The "map" function and the "reduce" function with each of the parts having their own different importance.

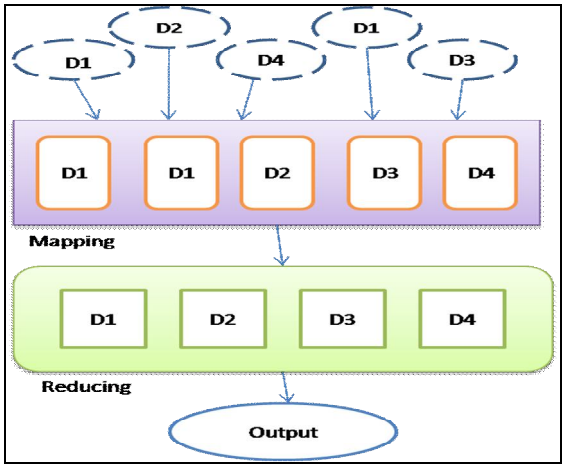

Figure 2: Two functions of map reduce

The Hadoop Map Reduce is strongly dependent on the Hadoop distributed file system (HDFS) to carry its input-output functions. The Hadoop Map Reduce is divided into different jobs, the job tracker and task tracker. Since, Hadoop Map Reduce is entirely based on java; it makes all precise tasks easier to perform on it. It is not like SQL, where you need to have knowledge about databases but instead it only requires users to know a little bit about java. This is one reason for which Hadoop Map Reduce is hugely used by developers today [7].

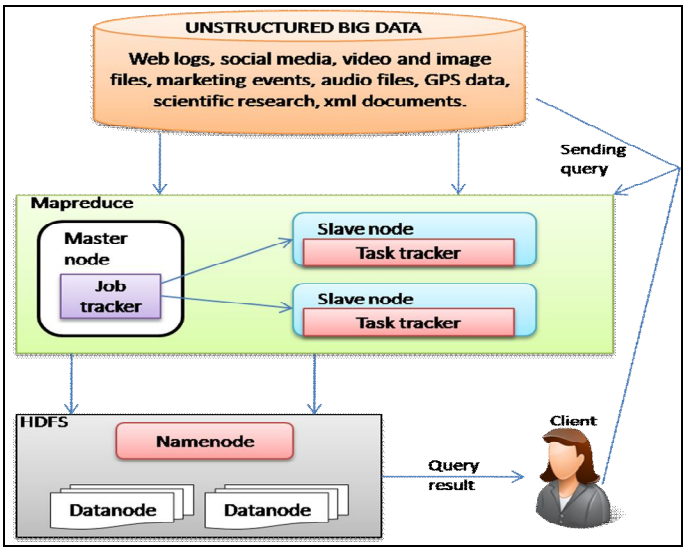

Figure 3: The Hadoop working system [12]

4. Comparison between RDBMS and Hadoop

- Hadoop is a latest technology and RDBMS is a traditional technique.
- Hadoop is a framework by RDBMS is a database system.
- Hadoop blends well with both structured and unstructured data whereas RDBMS is only about processing structured data.
- It is very difficult and time consuming to add data in a traditional database system like RDBMS but extra node file anything can be added to Hadoop.
- RDBMS is all about query statements but Hadoop is about java codes.
- Hadoop is evolving, fast and flexible; RDBMS is not too fast and is confined to some constraints.
- Hadoop is about parallel processing and running different jobs which RDBMS lacks [16].

5. Impact of Big Data Technology in Business Analytics

The use of big data has a huge impact in the world of business and its analytics. Big data has led to the development of mammoth of an opportunity for large enterprises as well as small businesses. Though the traditional database system is still in use by some enterprises it is seen that with the emergence of big data solutions these traditional systems are falling out of place. Hadoop is one technology which is a near perfect solution for analysing big data. In the world of IT, every technology has its set of pros and cons. Advantages of using Hadoop for business analytics:

- **Scalability**
  Traditional databases were unable to measure and process large amounts of data. The Hadoop framework helps to fork out the gigantic amounts of data and store it in several thousands of computers consisting of data in the range of millions of gigabytes.
- **Inexpensive**
  The main disadvantage of traditional database management system is that it lacks the property of processing data in voluminous amounts due to which it is extremely expensive to store big data. It was seen that lately to reduce the cost of traditional database management system companies had to decrease the amount of data in the process of which many valuable raw data was lost. This is where Hadoop comes in; It is very cost-effective in comparison and designed in a way to store data without having to tamper with the data.
- **Swift, versatile and no failure**
  Earlier, the traditional database management systems were only able to store and manage structured data. So, it had become the need of the industries to have a framework which would be able to help them process, manage and store all the structured and unstructured data. Hadoop helps in generating value from both the types of data which is why it is used in business for variety of purposes. Hadoop having a distributed file system and mapreduce as its core components is divided into a large number of nodes which makes the processing large volume of data done in a matter of a few minutes. Since, Hadoop is

actually about a large number of servers. When data in sent to a server, it is replicated to all the other servers. This helps as there is always a copy of the data in case of failure of a node [10, 11].

Disadvantages of Hadoop for business analytics:

Calculation of data is a little time consuming in Hadoop due to the versatility of the data.

Although Hadoop is capable of processing very large amounts of data but still, since Hadoop is a new concept for the business industry. It is very difficult in collecting and processing data from different parts of the enterprise. [9]

5.1. Hadoop and its Use Cases

- Advertisements: The most popular thing in the internet is the advertisements. These needs to be calculated, their relevancy should be checked and the profit for the company. Companies auction their ads which causes the generation of huge amounts of data in petabytes. Hadoop helps companies to process these data to increase the speed and reliability of their ads according to the need of the user.
- Financial services: The financial services background have been very clear about solving their issues like fraud detection, reducing risks, identifying miscellaneous traders and marketing mostly by using Hadoop. Hadoop services like map R is used to keep a check on the interaction with customers, transactions, funds for industries and 24 hour service to the customers which results to huge amounts of data and fraud which can be overcome by processing data using Hadoop.
- Health care: The healthcare industry is one of the biggest industries in terms of data realisation. Hospital administrators, pharmaceutical providers, drug developers, researchers have to make which was a huge task with traditional databases as they did not have data transparency. By using Hadoop to analyze this data the healthcare industry is able to take care of patients more effectively since it helps in earlier detection of diseases, keeping information about certain diseases, development of drugs and diving patients according to their diseases and requirements.
- Gaming: Technology has led to a world of gaming where gamers are passionate about gaming battlefields, achieving new rewards and updating of their games. Hadoop is used to analyze the behaviour of the players so that gaming experience can be enhanced. It also processes the advertisements, updation criteria required according to the behaviour of the player.
- Web: The most important resource of data is the web. It consists of online services like social networking websites, online gaming, online travelling, ads. Every sector and company has its own websites adding to creation of large data. There are lot of possibilities of fraud which can be overcome by the use of Hadoop. Facebook is one such website which generates large data in form of videos, images, texts and other files and by using Hadoop all the data can be analyzed without the fear of data being lost.[17].

5.2. Companies Using Hadoop

There are many companies all over the world that uses Hadoop for business analysis. Some of the large industries are:

- Amazon web services: It is one of the largest enterprises using Hadoop. It doesn't purely use Hadoop but instead uses a service called Amazon EMR. This service offers wide variety of offers and makes it easy to process vast amounts of data very quickly and also cost-effectively.
- Cloudera: It is a software company that provides Hadoop support and services to different business. cloudera provides speed and accuracy to customers.
- Hortonworks: This is an enterprise that focuses on the development of Hadoop. It is an industry that builds and distributes apache hadoop. It deals with data from various sources and formats to analyze and store data.
- IBM: IBM info sphere big insights is the service that uses Hadoop. It provides with extreme accuracy, reliability, security and support to customers.
- MapR technologies: It is a company that is not much popular among people. It is the combination of apache Hadoop with various other capabilities to make data processing easier. It has a very high-tech architecture due to which it increases its reliability, dependability and also easy to use.
- Microsoft: Microsoft is a very popular among people. It is not an industry that would be using Hadoop and increasing its use. Microsoft is a name that is not related to open-source but it has let Hadoop to run on windows which has advanced the processing of data in windows.[18,19].

6. Conclusion

In this paper we have a basic review on big data. We came to know what actually is big data and its properties and challenges. We know about traditional data and the technique that was used earlier to store small amount of data. We see through the technology to analyse big data with the help of Hadoop. We have also mentioned the comparison between traditional and latest data analysing techniques. Another main part of the paper is the impact that the big data technologies have over business and helps in business analytics.

This is a review paper; we hope that this paper proves to be a great help to the readers in the field of research and also help amateur researchers.

7. Acknowledgements
I am very thankful to my acquaintance Nasrin Hussain for helping me and guiding me throughout the framing of this review paper.

8. References

1. P. A. Goloboff, "Techniques for analyzing large data sets", Techniques in Molecular Systematic and Evolution, ISBN 978-3-0348-8125-8, pp(70-73)
2. S. Madden, MIT, "From databases to big data", IEEE Computer Society, 1089-7801/12, 2012, (pp 4)
3. R. Gupta, "Journey from Data Mining to Web Mining to Big Data", International Journal of Computer Trends and Technology (pp 18).
4. what are big data techniques and why do you need them, FCW, big data techniques, http://www.fcw.com,(pp 2)
5. Hadoop: What it is and why it matter, SAS, http://www.sas.com, (pp 1/12)
6. T. White, "Hadoop: The Definitive Guide", Third Edition, O'rielly|yahoo press, ISBN 978-1-449-31152-0.
7. J. Dittrich, Jorge-Arnulfo Quian´e-Ruiz, "Efficient Big Data Processing in Hadoop Map Reduce", (pp 2014)
8. B. Proffit , "Hadoop: What It Is and How It Works", website http://www.readwrite.com,2013.
9. Big Data & Hadoop Solutions, Attunity website, http://www.attunity.com.
10. Alteryx and hortonworks (eds): "The business analyst's guide to Hadoop."
11. M. Nemschoff," 5 major advantages of Hadoop", http://www.itproportal.com.
12. S. Prakash, "Storing and querying big data in Hadoop distributed file system (HDFS).",sri's technology blog, website http://www.ecomcanada.wordpress.com.
13. M. Whitehorn, "What is the difference between DBMS and RDBMS?" university of Dundee, website http://www.searchdatamanagement.techtarget.com.
14. D. Borthakur, "The Hadoop Distributed File System: Architecture and Design." The Apache Software Foundation, 2007.
15. S. Blazehievsky, "Introduction to Hadoop, map reduce and HDFS for big data applications".
16. S. Ozcan, Database In Action -Integrated Approach to Oracle, "Difference between Hadoop and RDBMS", website http://www.oraclesys.com.
17. D. Weldon, "Examining business cases for Hadoop", industry insider M. Nemschoff, 2014.
18. B.Butler, "Nine Hadoop companies you should know", network world Inc., 2014.
19. Top 14 hadoop technology companies, technavio's blog-website, http://www.technavio.com