Big Data Technology: The Backbone for Building Remarkable Business Applications

Er.Sukhwinder Kaur[1] Er.Gagandeep Singh Kahlon[2] Dr. R.C.Gangwar[3] Er.Nishant K Pathak[4]

[1,4]Assistant Professor [2,3]Head of Dept.

[1,2,3,4]Department of Computer Science & Engineering

[1]CTIEMT [2,4]CTPC [3]BCET

Abstract— This paper gives you a brief glimpse of everything about big data and how we can make it very useful for making new applications. By reading this you will see that how big data analytics is finding out hidden patterns in huge amount of data, thus making it very much useful for the managers to take decisions about where to invest and how to increase the sales in a particular region. The paper focuses on very great outcomes that we are having from big data analytics in healthcare, production, employment, businesses, security management, marketing management etc. Big data is very important and it enables you to analyze large amounts of raw data either for cost management or technological reasons.

Key words: Building Remarkable Business Applications, Big Data Technology
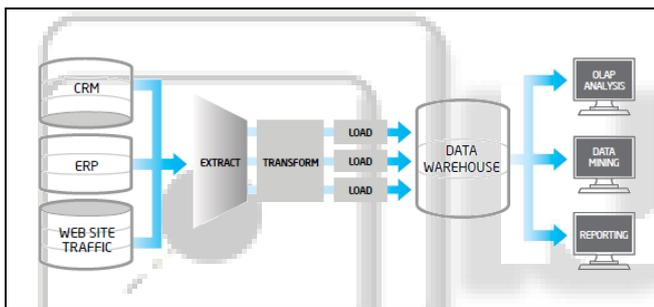
I. INTRODUCTION

Fig. 1: Big Data

Big Data will definitely change the decision support system in each and every area of business. It will change the revolution of data into revolution to produce information for making strategic decisions. Application needs to be developed to implement the above mentioned changes. Big Data will prove to be a catalyst for the development of developing countries like India, where decision support system has to be up to the mark. Big Data technology based application is not as simple to be build. As it involves data, we have to consider the privacy of the individual who owns the data or is associated with the data. We have to consider the privacy issues. The academic, public sector and private sector collaboration is also needed.

Fig. 2: The 5 Vs in big data

1) Volume: It refers to that enormous amount of data that is generated day by day. 90 percent of the data in the world today has been created in the last two years alone. Nowadays data volume is increasing from gigabytes to peta bytes. 40 Zetta bytes of data will be created by 2020 which is 300 times from 2005.
2) Velocity: Velocity is the speed at which data is developing and processed. For example social media posts.
3) Variety: Variety is one more important characteristic of big data. It refers to the type of data. Data may be in different styles such as Text, numerical, images, audio, video data. On twitter 400 million tweets are sent per day and there are 200 million active users on it.
4) Veracity: Veracity means accuracy of data. Data is uncertain due to the inconsistency and in completeness. It should also be free from ambiguities.
5) Variability: It refers to the changes in the data during its processing and lifecycle.

II. TECHNIQUES AND TECHNOLOGY FOR BIG DATA

A. Apache Hadoop for Big Data

When Yahoo, Google, Face book, and other companies extended their services to web-scale, the amount of data they collected on daily basis from user interactions online would have overwhelmed the capabilities of traditional IT architectures. In the interest of advancing the technological development of core infrastructure components rapidly, researches were made so as to overcome with these problems. Of these components, Apache Hadoop has rapidly emerged as the de facto standard for managing large volumes of unstructured data. Apache Hadoop is an open source distributed software platform for storing and processing data. Written in Java, it runs on a cluster of industry-standard servers configured with direct-attached storage. A hadoop frame worked applications work in an environment that provides distributed storage and computation across clusters of computers. Apache hadoop gives us the ability to store and process huge amounts of any kind of data quickly. Moreover the computation is very fast.

Fault tolerance in hadoop: In architecture of hadoop when a node goes down, the processing jobs are automatically redirected to other nodes to make sure that the distributing computing does not fail. Also it is having the flexibility that it manages unstructured data also.

HDFS (Hadoop distributed file system): The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size.
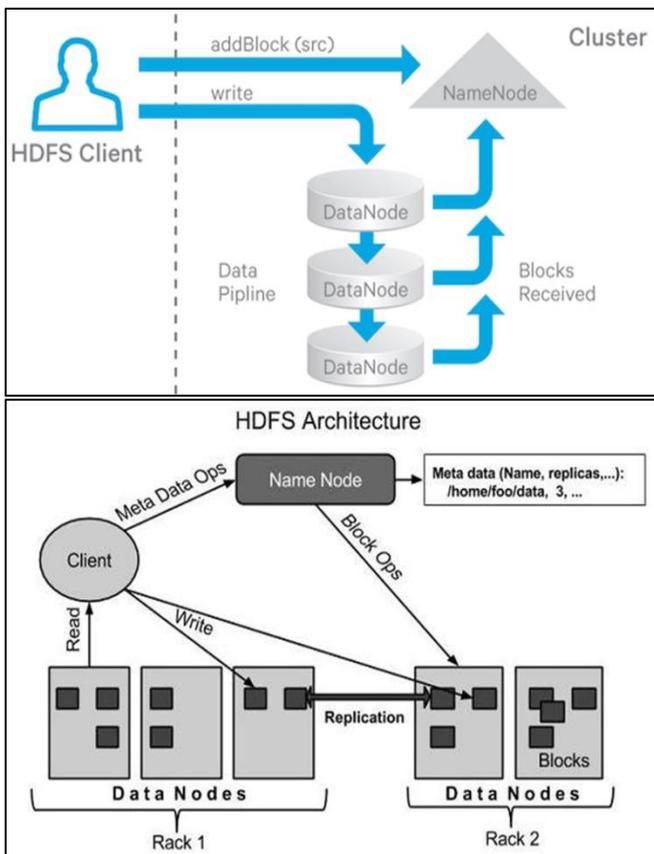
Fig. 3: HDF5 Architecture

1) Name node: The name node is the commodity hardware that contains the GNU/Linux operating system and the name node software. It is software that can be run on commodity hardware.

2) Data node: The data node is a commodity hardware having the GNU/Linux operating system and data node software. For every node (Commodity hardware/System) in a cluster, there will be a data node.

3) Block: Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block.

Data mining with big data: The emerging area of data mining with big data includes distributed and collaborative mining. By effectively mining the data available at different geographical sites will definitely change the way the decisions are made in business which helps us to make an effective decision support system.

Application needs to be developed to implement the above mentioned changes. Big Data will prove to be a great catalyst for the development of developing countries like India, where decision support system has to be up to the mark. Big Data technology based application is not as simple to be developed. As it involves data, we have to consider the security and privacy of the individual who owns the data or is linked with the data. We have to consider the privacy issues. Mainly academic, public sector and private sector collaboration is also needed. Since big data has to be stored and implemented using sophisticated hardware which is costly, government should provide

required resources to the researchers to carry on the research. The project aims to define a strategy in terms of research and innovation to guide supporting actions from the European Commission in the successful implementation of the Big Data economy. So to extract more value from enormous amount of data it will prove to be very useful and effective strategies in today's competitive world.

REFERENCES

[1] A White Paper by Bloor Research, Author: Philip Howard and published in the year April 2012.

[2] Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review"

[3] Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;(17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing.

[4] Grosso, P. ; de Laat, C. ; Membrey, P.,(20-24 May 2013)," Addressing big data issues in Scientific Data Infrastructure"

[5] Zhang, Du,(16-18 July,2013)," Inconsistencies in big data"

[6] Tien, J.M.(17-19 July,2013)," Big Data: Unleashing information"

[7] Kogge, P.M.,(20-24 May,2013), "Big data, deep data, and the effect of system architectures on performance"

[8] K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012.

[9] Suhas V. Ambade and Prof. Priya Deshpande. Hadoop Block Placement Policy for Different File Formats. International Journal of Computer Engineering and Technology, 5(12), 2014, pp. 249-256.

[10] Gandhali Upadhye and Astt. Prof. Trupti Dange. Nephele: Efficient Data Processing Using Hadoop. International Journal of Computer Engineering and Technology, 5(7), 2014, pp. 11-16.