Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data and its Handling Technologies-A Study

Sukhdev Singh Ghuman
Dept.of Comp.Sci., SBDSM Khalsa College,
Domeli, Kapurthala, Punjab, India

Abstract---Big Data are very large data sets which are difficult to manage with ordinary data handling techniques. This data consist of structured and unstructured data that has the potential to be mined for information. Big Data analysis is concerned with finding patterns, trends and associations hidden within. The three main characteristics of Big Data are volume, variety and velocity. Many technologies are available for handling Big Data like MapReduce, Hadoop Hive, PIG etc. All the big data technologies are closely associated with cloud.

Keywords— Big data, Hadoop, structured, analysis, volume

I. INTRODUCTION

Big data means extremely large data sets that may be analysed computationally to find patterns, trends, and associations, especially relating to human behaviour and interactions. Big data [1] is a broad term for very large and complex data sets which cannot be processed by traditional data processing applications. It includes analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reductions and reduced risk. Big data [2] is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data. Big data [3] is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Despite these problems, big data has the potential to help companies improve operations and make faster, more intelligent decisions. Big data [4] is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety.

Figure 1: Big data [9]

To extract meaningful value from big data, we need great processing power and analytics capabilities. The term big data [3], may refer to the technology (which includes tools and processes) that an organization requires to handle the large amounts of data and storage facilities. The term big data is believed to have originated with Web search companies who needed to query very large distributed aggregations of loosely-structured data. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources like Web, sales, customer contact center, social media, mobile data etc. Big Data [5] is the ocean of information we swim in every day – vast zetabytes of data flowing from our computers, mobile devices, and machine sensors. With the right solutions, organizations can dive into all that data and gain valuable insights that were previously unimaginable. The paper has been divided into the sections. The first part try to define Big data gives its brief introduction. In the second part important characteristics of Big data has been listed. The third section is detailed study about different technologies available to handle big data.

II. CHARACTERISTICS OF BIG DATA

Big data can be can be distinguished from the ordinary data with the help of following characteristics.

- **Volume** – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not [1].
- **Variety** - It means that the category to which Big Data belongs. This helps the analysts, who are closely analyzing the data.
- **Velocity** – It means the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges [1].
- **Variability** - This refers to the inconsistency which can be shown by the data from time to times, thus causing problems to handle and manage the data effectively [1].
- **Veracity** - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data [1].
- **Complexity** - Data management can become a very complex process, particularly when large volumes of data come from multiple sources. This is known as the 'complexity' of Big Data[1].

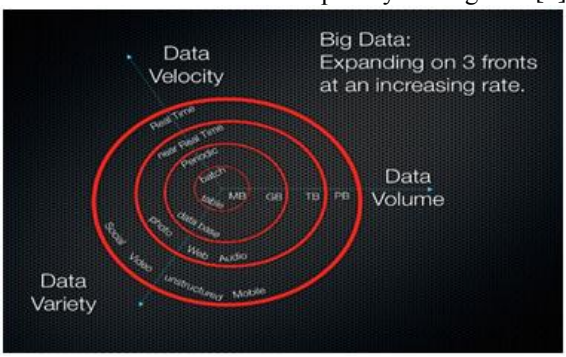

Figure 2: Big Data Characteristics [1]

III. BIG DATA TECHNOLOGIES

Big data technologies are important in providing more accurate analysis, which may lead to much better decision-making resulting in greater productivity, cost reductions, and reduced risks for the business. The infrastructure that can manage and process huge volumes of structured and unstructured data is required to tap the power of Big Data. Big data[2] takes too much time ,costs and money to load into a traditional relational database for analysis, new approaches to storing and analyzing data have emerged that rely less on data schema and data quality. Instead, raw data with extended metadata is aggregated in a data lake and machine learning and artificial intelligence (AI) programs use complex algorithms to look for repeatable patterns. When dealing with larger datasets [3], organizations face difficulties in being able to create, manipulate, and manage big data. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets. This data, when captured, formatted, manipulated, stored, and analyzed can help a company to gain useful insight to increase revenues, get or retain customers, and improve operations.

Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time requires a platform like Hadoop to store large data sets across a distributed cluster and MapReduce to coordinate, combine and process data from multiple sources. Hadoop [2] is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

New skills are needed to fully harness the power of big data. Though courses are being offered to prepare a new generation of big data experts, it will take some time to get them into the workforce. Leading organizations are developing new roles, focusing on key challenges and creating new business models to gain the most from big data [4].

The big data includes data produced by different devices. The different sources of Big data are as given below [6]

- Black Box Data
- Social Media Data
- Stock Exchange Data
- Power Grid Data
- Transport Data
- Search Engine Data

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology:

Operational Big Data

NoSQL Big Data systems such as document databases have emerged to address a broad set of applications. These technologies were developed to address the shortcomings of relational databases in the modern computing environment.

They are much faster and scale much more quickly and inexpensively than relational databases [7]. This makes operational big data workloads much easier to manage, cheaper, and faster to implement. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure [6].

Analytical Big Data

Analytical Big Data workloads are addressed by MPP database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data [6][7]. These technologies are also a reaction to the limitations of traditional relational databases and their lack of ability to scale beyond the resources of a single server. Furthermore, MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL. MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines[6].

These two classes of technology are complementary to each other and generally deployed together

Operational vs. Analytical Systems [6]

	Operational	**Analytical**
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 – 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce MPP Database

Big Data is broad and encompasses many trends and new technology developments, some of the emerging technologies that are helping users cope with and handle Big Data in a cost-effective manner are as listed below [8]:-

1. NoSQL databases

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

2. MapReduce

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Its implementation consists of two tasks [8]:

- The "Map" task, where an input dataset is converted into a different set of key pairs, or tuples;
- The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples.

3. Hadoop

Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources. It works by either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but it is mainly used for large volumes of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data [8].

4. Hive

It is a technology that allows conventional BI applications to run queries against a Hadoop cluster. It was developed by Facebook, but later it was made open source. It provides framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store.

5. PIG

PIG is another technology that tries to bring Hadoop closer to the realities of developers and users. PIG consists of a Perl like language that allows for query execution over data stored on a Hadoop cluster, instead of a SQL like language.

6. WibiData

It is a combination of web analytics with Hadoop. WibiData is built on top of HBase, which is a database layer on top of Hadoop[8]. It allows web sites to better explore and work with their user data. It provides real-time responses to user behavior and helps to give user specific content and recommendations,

7. PLATFORA

It is a platform that converts user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

8. Storage Technologies

Data volumes grow tremendously, so there is the need for efficient and effective storage techniques. The main technologies used are data compression and storage virtualization.

9. SkyTree

It is a high-performance machine learning and data analytics platform focused specifically on handling Big Data. Machine learning is an essential part of Big Data because the massive data volumes make manual exploration, or even traditional automated exploration methods unfeasible [8].

These technologies are closely associated with the cloud. Most cloud vendors are offering hosted Hadoop clusters that can be scaled on demand according to the requirement of the user. Many of the products and platforms are either entirely cloud-based or have cloud versions themselves. Big Data and cloud computing go hand-in-hand. It helps companies to get more value from their data by enabling fast analytics at a fraction of previous cost.

IV. CONCLUSIONS

Big Data is very voluminous data which can be analysed with the help of technologies discussed in this paper. There are a large number of technologies available for this but they have their own advantages and disadvantages. Hadoop is very common for processing Big Data because it is an open source platform and it has the strength to process different varieties of data.

REFERENCES

[1] http://en.wikipedia.org/wiki/Big_data
[2] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
[3] http://www.webopedia.com/TERM/B/big_data.html
[4] http://www.ibm.com/big-data/us/en/
[5] http://go.sap.com/solution/big-data.html
[6] http://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm
[7] https://www.mongodb.com/big-data-explained
[8] http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data
[9] http://www.isymmetry.com

