An Comprehensive Study of Big Data Environment and its Challenges.

Saravanan.C
Assistant Professor, Department of Master of Computer Applications, R.V.College of Engineering, Bangalore

ABSTRACT

Big Data is a data analysis methodology enabled by recent advances in technologies and Architecture. Big data is a massive volume of both structured and unstructured data, which is so large that it's difficult to process with traditional database and software techniques. This paper provides insight to Big data and discusses its nature, definition that include such features as Volume, Velocity, and Variety .This paper also provides insight to source of big data generation, tools available for processing large volume of variety of data, applications of big data and challenges involved in handling big data .

Keywords: *Big data, Challenges, Evolution, Structured data, unstructured data, Traditional database.*

I. Introduction

Big data is a term that refers to data sets or combinations of data sets whose size, complexity and rate of growth make them difficult to be captured, managed, and processed by conventional technologies and tools, within the time necessary to make them useful [1]. Big Data are high-volume, high-velocity, and high-variety information assets that require new forms of processing to enable enhanced decision making and process optimization [2]. There is no clear definition for 'Big Data'. It is defined based on some of its characteristics. There are three basic characteristics that can be used for defining big data namely volume, variety, and velocity [3] (figure 3).

Volume: It refers to size of the data such as Terabytes (TB), Petabytes (PB), Zettabytes (ZB), etc.
Variety : It refers to different types of data and sources of data. The different mediums that will produce big data are sensors, devices, social networks, the web, mobile phones, etc.
Velocity: It refers to the frequency of data generation. and how data is generated frequently For example, every millisecond, second, minute, hour, day, week, month, year. Processing frequency may also differ from the user requirements.
When big data is effectively and efficiently captured, processed, and analyzed, companies are able to gain more complete understanding of their business, customers, products, competitors which can lead to efficiency improvements, increased sales, lower costs, improved customer service, and improved products and services.

II. Big Data Generation.

In Earlier generations, fairly small volume of data generated was available through limited number of channels. Today an huge amount of data is frequently being produced and flowing from various sources, through different channels, every minute in today's digital age [4].
Traditionally few Companies were generating data and all others were consuming data. Now the scenario has been changed. All of us are generating data and everyone is consuming data. Big Data is generated by many channels such as Social Media and Networks like Facebook, twitter, YouTube Flickr ,Scientific instruments
Mobile devices and Sensor Technology and networks
Figure 1. Shows the different channels that produce big data.

Figure 1. Different Medium generating Big data.

The data generated is off different format namely Structured data ,Semi structure data and Unstructured data. Data that resides in fixed fields is termed as Structured data. Data that does not exist in fixed fields is unstructured data. Data that do not adapt to fixed fields but contain tags and other markers to separate data elements is semi structured data. According to Gartner Research, it is predicted that enterprise data will grow by 800 percent in five years, with 80 percent of data to be unstructured.[5]. According to recent survey of Gartner, September ,2015 It presents that more than 75 Percent of Companies are Investing or Planning to Invest in Big Data in the Next Two Years. Figure 2 shows the percentage of structured, unstructured and semi structured digital data.

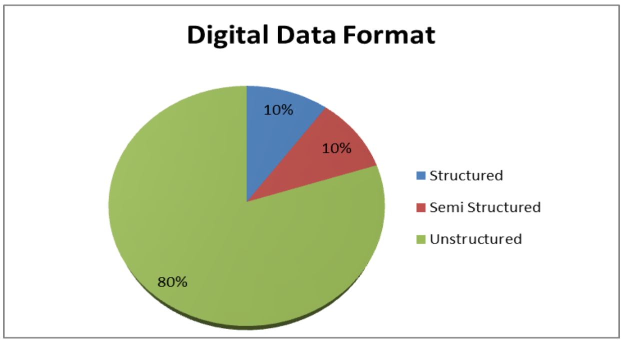

Figure 2. Percentage of Structured, Semi structure and unstructured data to be generated in next 5 years

Table 1 describes the different data format and their sources from which the data is being generated.

Table 1. Sources of Data format

Sl.No	Data Format	Sources
1	Structured Data	Access, SQL, Spread sheet , Online Transaction processing (OLTP).
2	Semi Structured Data	Blog, XML, Emails, Electronic data Interchange [EDI]
3	Unstructured Data	Social media data, chat messages, Location stream data, Streamed Video, Audio/Music files, Images, Sensor data

Global data volumes—flowing from social Web sites, sensors, smartphones are increasing faster than every two years[6].The Amount of data stored in a country differs from one country to another. Fig 3 shows amount of new data stored across the globe[7].

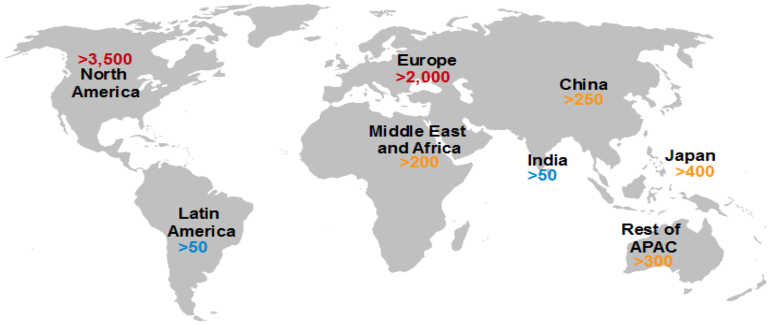

Figure 3:Amount of new data stored across globe
SOURCE:IDC storage reports; McKinsey Global Institute analysis

III. Tools Available for Big Data

There are various tools available for handling big data

3.1 Hadoop.: This is open source software platform helpful in storing and managing vast amounts of data cheaply and efficiently. It has two main parts –

i) Data processing framework
ii) Distributed file system for data storage.

The distributed file system is array of storage clusters i,e the Hadoop component that holds the actual data

The data processing framework is the tool used to work with the data itself. By default, this is the Java-based system known as Map Reduce. It is a general programming model and an associated implementation for generating and processing large data sets. [8]

3.2 NOSQL

NoSQL databases are fast becoming an essential tool for managing big data while databases based on the relational model guarantee certain properties which at first sight might seem more important or even necessary nowadays it is impossible to handle certain volumes without relaxing some of them. It is precisely out of this relaxing and out of the need to provide other properties that a new data management paradigm has arisen: NoSQL. There are four different forms of NOSQL technologies namely Key value , document store, wide column stores and graph databases.

Key-values Stores : This database uses hash table where there is a unique key and a pointer to a particular item of data. The Key/value model is the simplest and easiest to implement. Tokyo, Redis, Voldemort, Oracle BDB, Amazon Simple DB, Riak are few examples of Key value Stores databases.

3.3 Wide Column store

This type of database were created to store, process very large amounts of data distributed over many machines. There are still keys but they point to multiple columns. Cassandra which is open source Nosql database [9] and Hbase are examples of wide column store databases.

3.4 Document Databases

These databases are similar to key-value stores. The model is basically versioned documents that are collections of other key-value collections. The semi-structured documents are stored in formats like JSON. CouchDB[12] and MongoDb which is open source Nosql database[9] are examples for this type of databases.

3.5 Graph Databases

This type of database are built with nodes, relationships between notes and the properties of nodes. Instead of tables of rows and columns and the rigid structure of SQL, a flexible graph model is used which can scale across many machines. Neo4J, InfoGrid, Infinite Graph are examples of Graph databases.

IV .Applications

Big data technology has become part and parcel of our life. It can be applied in every fields namely Optimization of IT Infrastructure ,Social Network Analysis, Optimization of Traffic flow in a city ,Web app Optimization, Natural resource exploration, Weather Forecasting ,Health care outcomes, Fraud detection, Life science research, Advertising analysis, Equipment monitoring, smart meter monitoring and so on.

V. Challenges In Handling Big Data

There are various challenges to be addressed when handling big data. Some of the inherited challenges in big data are capture, storage, search, analysis, and virtualization. The other challenges are discussed below

i. Organizational Challenge

The management of the organization often lacks the understanding of the value in big data as well as how to solve this value. Many organizations do not have the talent pool or experts to derive insights from the new technology big data .According to Business Intelligence and Information Management Survey of 541 business technology professionals, October 2012, 31% people are not sure how big data analytics will create business opportunities for their business organization[10]

ii. Lack of Analytical Skills

There is scarcity of analytical and managerial skills to make best use of big data. It is found that United States alone faces a shortage of 1,40,000 to 1,90,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings.[11].According to Business Intelligence and Information Management Survey of 541 business technology professionals, October 2012, around 38% people find scarcity in big data expertise is scarce and find big data is expensive.[10].

iii. Lack of Analytical tools in big data platform

Big data can be managed by tools like Hadoop and Nosql. Analytical tools are lacking for big data platforms.Hadoop and nosql technologies also lack management features [10].

iv. Design hurdles

The barrier is in technology i,e new architecture, algorithms, techniques are needed to handle or manage big data.

v. Data discovery

Large volumes of different data can be combined and explored freely using a visual analytic development environment. Data discovery provides a new insight to use the enriched data to make better-informed decisions. The challenge lies in finding out high quality data from the vast collections of data that is available.

vi. Data volume

The capacity to process the high volume of data at an adequate speed so that the information is available to decision makers when they are in need of it.

vii. Data integration

The capacity to integrate data which is dissimilar in nature,quickly with in short period of time from various source at reasonable cost is challenging task

VI. Conclusion

In recent times, big data has invited a lot of attention from academia, industry, public as well as government. The increasing volume and details of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future. Improvements in professional data management will result in better science and will benefit the social community. A proposed direction for the future work could be the study of big data analytics, which has become an immense tool affecting every part of the economy.

References

[1]. Mukherjee A. Datta, J. , Jorapur, R. ,Singhvi, R., Haloi, S., Akram, W. "Shared disk big data analytics with Apache Hadoop" High Performance Computing (HiPC), 2012 19th International Conference.

[2] Douglas and Laney, "The importance of 'big data': A definition," 2008.

[3] D. Laney, "Application Delivery Strategies", http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data- Management-Controlling Business Intelligence and Information Management Survey Data-Volume-Velocity-and-' Variety.pdf

[4] King, Gary. "Ensuring the Data-Rich Future of Social Science." Science Mag 331, Feb 2011, 719-721 .

[5] Win with Advanced Business Analytics.Creating Business Value from Your Data, Wiley and SAS Business Series,Oct12.

[6]. Jacques Bughin, Michael Chui, and James Manyik,"Ten IT-enabled business trends for the decade ahead", McKinsey & Company, Tech.rep June 2013.

[7]. James Manyika et al., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global
Institute, Tech. rep. May 2011.

[8] JeffreyDean,SanjayGhemawat, ” MapReduce: Simplified Data Processing on Large Clusters” USENIX Association OSDI ’
04: 6th Symposium On Operating Systems Design and Implementation

[9] Van der Veen, J.S , van der Waaij. B , Meijer, R.J. Sensor Data Storage Performance: SQL or NoSQL, Physical or Virtual.
In Proceedings of IEEE 5th International Conference on cloud Computing (CLOUD 2012), Nice,France ,22-27 July 2012;
pp. 431–438

[10]. InformationWeek 2013 Analytics, Business Intelligence and Information Management Survey of 541 business technology
professionals, October 2012

[11] Big data: The next frontier for innovation, ompetition and productivity , McKinsey Global Institute, June 2011.

[12]. “Post-relational data management for Ineractive software systems, NoSQL database Technology, www.Couchbase.com.