Pervasive study on big Data Analytics

J.Ponrenga[1] M.Sembalai Sumathi[2]
Assistant Professor
Department of Computer Science
Sri Sarada College for Women
Tirunelveli. -11

Abstract:

In the world of science, Data plays a decisive role in all the fields. Data is the building block upon which any organization thrives. But the growth of data is tremendous. If the data storage is not possible, the Organization will lose its potentiality to extract valuable information and knowledge, perform detail analysis as well as provide new opportunities and advantages. Statesmen also need to gain valuable insights from such wide-ranging swiftly fluctuating data. To tackle the increasing challenges arising in the various fields, the big data analytics takes the influential place in these times. The big data is the most prominent paradigm now-a-days. Besides, the big data overrides the technological war and rule the kingdom of technology since 2009. It has changed people's way of living in some area due to its great clout on many business fields, industry, healthcare etc. This paper covers methodologies, dimensions in big data analytics and different techniques which can be applied in big data science activities.

Keywords:- Big Data, Big data analytics, collation.

Introduction:-

The world is moving towards a more connected future. It is literally drowning in enormous amounts of data, Structured and unstructured. Data is being generated at an exponentially growing rate as the world becomes digitized in every facet of human activity. Reality is essentially recorded on an ever increasingly detailed manner on a daily basis, through sensors, cameras, credit cards, phone locations, internet activity, cable TV habits, hospital records, prescriptions, motel records, real estate tax records, fitness recreation etc. Big data is an absolute technological requirement to process the enormous data sets of today.The term 'big data' has been around since the mid-1990s, but only over the past decade has it become common place in our vernacular.The world has recognized the importance of Big Data. In August 2010, President Barrack Obama announced the "Transparency and Open Government" in the "Memorandum for the Heads of Executive Departments and Agencies" proclaiming that "Big Data is a national challenge and priority along with healthcare and national security"[1]. This paper is organized as follows.We begin the paper by defining big data , Other characteristics important in defining big data and the tools needed for big data analytics.

Defining Big data:-

Clearly, size is the first characteristic that comes to mind considering the question "what is big data?" However, other characteristics of big data have emerged recently. For instance, Laney (2001)[2] suggested that Volume, Variety, and Velocity (or the Three V's) are the three dimensions of challenges in data management. The Three V's have emerged as a common framework to describe big data (Chen, Chiang, &Storey, 2012; Kwon, Lee, & Shin, 2014). For example, Gartner, Inc. defines big data in similar terms:

"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." ("Gartner IT Glossary, nd.")

Similarly, TechAmerica Foundation defines big data as follows:

"Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." (TechAmerica Foundation's Federal Big Data Commission, 2012)

Hence, "big data analysis" is the term used to describe a new generation of practices (Kempenaar et al., 2016; Sonka, 2016), designed so that farmers and related organizations can extract economic value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis (Waga and Rabah, 2014; Lokers et al., 2016). Big data analysis is successfully being used in various industries, such as banking, insurance, online user behavior understanding and personalization, as well as in environmental studies (Waga and Rabah, 2014; Cooper et al., 2013). As (Kim et al., 2014) show, governmental organizations use big data analysis to enhance their ability to serve their citizens addressing national challenges related to economy, health care, job creation, natural disasters and terrorism.

Characteristics of 'V's in Big data

Size is the first characteristic that comes to mind considering the question "what is big data?" However, other characteristics of big data have emerged recently. The term 'big data' is used for large datasets that are complex in nature, and this term is defined by nine V's: Volume, Variety, Velocity, Veracity, Validity, Visualization, Value, Vendee, Vase.(Gandomi and Haider, 2015).

- **Volume (V1)**: The volume is a huge set of data to be stored and processed [2]. The volume grows exponentially and it does not have any bound. The size of data collected for analysis.
- **Velocity (V2)**: It is the rate at which this information is generated and collected. The time window in which data is useful and relevant. The velocity is always defined with respect to volume in Big Data. The velocity in Big Data concerns mainly two things, namely, speed of growth and speed of transfer. For example, some data should be analyzed in a reasonable time to achieve a given task, e.g. to identify pests (PEAT UG, 2016) and animal diseases (Chedad et al., 2001).
- **Variety (V3)**: Multi-source (e.g. images, videos, remote and field based sensing data), multi-temporal (e.g. collected on different dates/times), and multi-resolution (e.g. different spatial resolution images) as well as data having different formats, from various sources and disciplines, and from several application domains.
- **Veracity (V4)**: The veracity is accuracy, truthfulness, and meaningfulness [3], [4], [7].The quality, reliability and potential of the data, as well as its accuracy, reliability and overall confidence.
- **Validity (V5)**: Though the process can perform a task accurately, but the data may not be valid [3]. For example, the very old data in e-commerce becomes obsolete and it can be truncated. The validity may differ from time to time. The validity refers to the data those have worthiness. The correct data may not be valid for certain processing.
- **Visualization (V6)**: On the other hand, visualization [3] is the process to show the hidden data of Big Data. This term is more precise to describe Big Data, because visualization is making visible. The Visualization is the most key process to enhance the performance of the data and business processes/decisions. Presenting complex data structures and rich information in an easy-to-understand way (Hashem et al., 2015; Karmas et al., 2016).
- **Value (V7)**: It describes the hidden value obtained from big data, such as insights about online trends and behaviors (Ekbia et al., 2015). However, the Big Data is concerned mainly on extracting value from

enormous stored data [5], [4], [7], [3]. The Big Data extract values from data, reveal hidden truth from the data, uncover useful message from the data,
Creates value of data.

• **Vendee (V8)**: The one more new V in Big Data is "Vendee" to define the client size associated with the Big Data to deal with. The Vendee is the most significant component to define the Big Data, where the 9V's are made only for clients to conform as their requirements.

• **Vase (V9)**: According to Merriam-Webster dictionary, the vase is a container that is used for holding flowers or for decoration. The Big Data is high volume, and stored in the datacenter. The datacenter requires farmhouse, land, huge electric power, thousands of hardware, thousands of manpower, varieties of hardware, and many more small products to enable Big Data. In the Big Data paradigm, the flower refers to Big Data and the vase refers to underlying requirements
to enable Big Data.

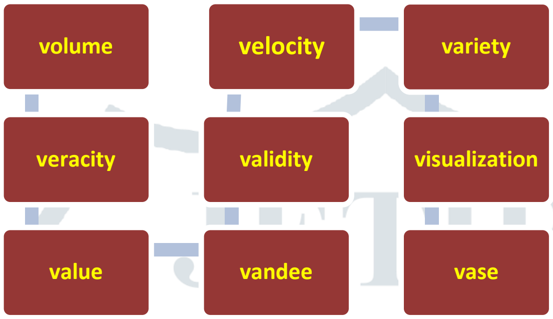

Fig. 1. Big Data Dimensions

Big Data Analytics

Big data analytic is a process of discovering patterns and trends from a large amounts of data to extract its value and correlations Based on the report done, the stages that are needed to be done in data retrieval for big data are as followed:

1) Data acquisition: Data acquired from the medium where data generation is growing at an exponentially rate. Although, the data that is being produced continuously mostly made up of unprocessed data that are useless and due to its unstructured form, selecting and discarding unneeded data can be quite challenging.

2) Data extraction: Majority of the acquired raw data are not useful. Hence, deciding which data are needed to be kept and which one should be discarded is a difficult task to perform as well as there is an abundance of it.

3) Data collation: Most of the time, utilizing data from one sample is inadequate to be used in an analysis or prediction. So, data is retrieved from different sources and would be combined or superimposed so that a bigger and more detailed picture is formed. From this point, the data can be analyzed more properly.

4) Data structuring: It is important for the analyzed data to be organized in a structured form. This enables the retrieval of information to be easier.

5) Data visualization: Usually, case studies would concentrate on certain part of an area or region. The data that is being pulled from these areas are then analyzed and converted into a more specific and visual format.

6) Data Interpretation: This is where valuable information will be extracted. There are two types of information that can be acquired: Retrospective Analysis and Prospective Analysis. Retrospective involves gaining insights from the past events and actions. Prospective Analysis is distinguishing patterns and discovering trends for future based off the data that was recorded.

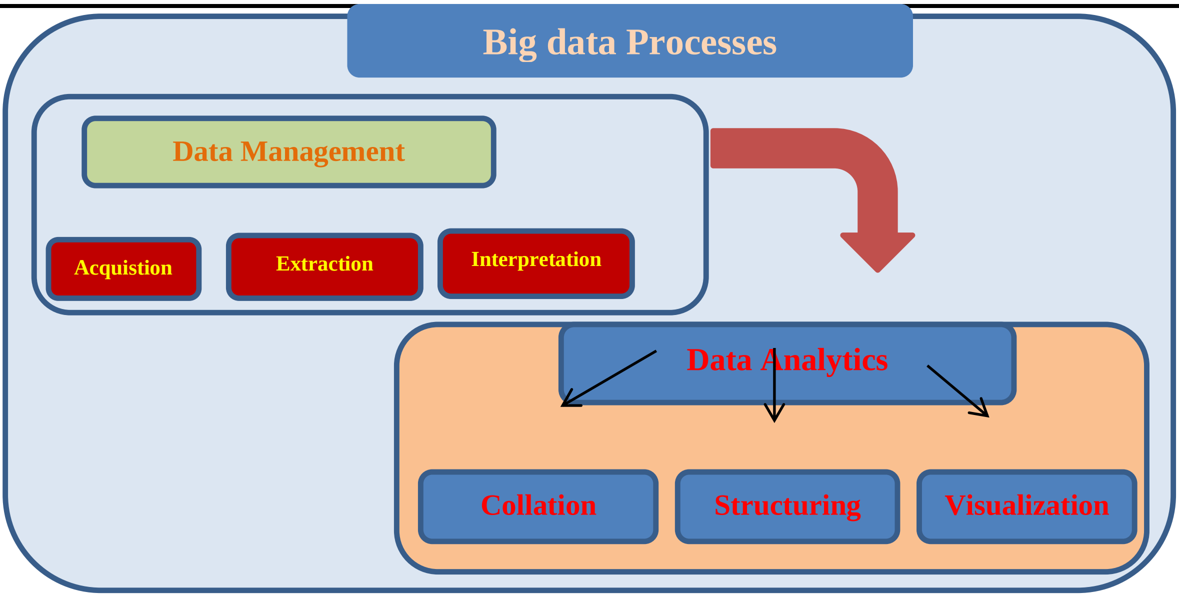

Fig. 1. Processes for extracting insights from bigdata

Big Data Tools

1. Hadoop

Apache Hadoop is the most prominent and used tool in big data industry with its enormous capability of large-scale processing data. This is 100% open source framework and runs on commodity hardware in an existing data center. Furthermore, it can run on a cloud infrastructure. Hadoop consists of four parts:

- Hadoop Distributed File System: Commonly known as HDFS, it is a distributed file system compatible with very high scale bandwidth.
- MapReduce: A programming model for processing big data.
- YARN: It is a platform used for managing and scheduling Hadoop's resources in Hadoop infrastructure.
- Libraries: To help other modules to work with Hadoop.

2. NoSQL:-

A NoSQL is originally referring to ”non SQL”or”non relational” database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases**.**

3. Presto:-

Facebook has developed and recently open- sourced its query engine (SQL-on-Hadoop) named Presto which is built to handle petabytes of data.Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.

4. Sqoop:-

This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.

5.PolyBase:-

This works on top of SQL Server 2012 Parallel Data Warehouse(PDW) and is used to access data stored in PDW.PDW is a data warehousing appliance built for processing any volume of relational data and provides an integration with Hadoop allowing us to access non-relational data as well.

Conclusion:-

Business organizations are in need to consider big data to make more accurate analysis, leading to a better decision making. Big data can be characterized by nine Vs; Volume, Variety,Velocity,Veracity,validity,visualization,value,vendee,vase.Bigdataunlocks the value of data y quickly and seamlessly integrating new insights into every aspect of production,across all departments, available to all employees in languages they can understand. and tools to its data analytics' complexity, data size and type. Big data analytic does process multiple sources of data to present any patterns, trends, or customers' behavior. By then. Big data can be used to detect future transactions based on customers' behavior.

References:-

[1] American Institute of Physics (AIP). 2010. College Park, MD,(http://www.aip.org/fyi/2010/)
[2] L. Douglas."3D Data Management: Controlling Data Volume,Velocity and Variety" (PDF). Gartner. Retrieved 6 February 20.
[3] Eileen McNulty, "Understanding Big Data: The Seven Vs",[online] Retrieved 10, June, 2016 from http://dataconomy.com/seven-vs-big-data/
[4] landmark.solutions, "The 7 pillars of Big Data", A White Paper of Landmark Solutions, Retrieved on 10, June, 2016 from https://www.landmark.solutions/Portals/0/LMSDocs/ Whitepapers/The 7 pillars of Big Data Whitepaper.pdf
[5] M Ali-ud-din Khan, M F Uddin, and N Gupta, "Seven V s of Big Data: Understanding Big Data to extract Value", In 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1), pages 3-5, April, 2014, DOI://http://dx.doi.org/10.1109/ASEEZone1.2014.6820689.
[6] https://en.wikipedia.org/wiki/Big_data.
[7] Yuri Demchenko, Cees de Laat, and Peter Membrey, "Defining Architecture Components of the Big Data Ecosystem", In 2014 International Conference on Collaboration Technologies and Systems (CTS), pages 104 - 112, 2014, DOI: http://dx.doi.org/10.1109/CTS.2014.6867550