International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Overview of Big Data

Nivedita Manohar

Assistant Professor, Dept. of CS/IT, ACED, Alliance University, Bengaluru, Karnataka, India

Abstract: ***Presently Big Data is one of the major and booming technology of this century which impact all most all streams and fields of this era. The significant research and development is improved as time went with all stream or fields. In this paper, an outline of the applications of big data are introduced with broad classification along with its tools. The research centers of India for Big Data are listed which can help to explore more knowledge along with the challenges and opportunities of the technology are listed.***

Keywords: Big Data, Hadoop, BigData Process

1. Introduction

Everyone things about Big Data as a voluminous data. But, Big data also gives meaning even though its size is small. For large amount of data Relational Database Management System technologies can provide solution for the handling of these data. So, the term big data is misnomer. As we know RDBMS follows structured data where as the emerging data from variety of data comes in different structure becomes part of so called *Big Data*. Big data not has definite definition but it can also be defined as any data set in some cases which cannot be stored using the resources of a single machine. If it can store and it may not be able to meet the mandatory service level agreements (SLAs). In other way, the definition is with the crucial action ie any scale of data can be processed virtually on a single machine even though that data cannot be stored on a single machine. In such cases data from different machine brought into one machine with help of shared storage technology such as Network Attached Storage(NAS). This process takes large amount of time with respect to the available time for that data.

In our daily life we are habituated to use many digital gadgets and also trying to adapt high end gadgets with sensors to control the situation. These gadgets are ranging from various sensors. For example, a car with many sensors are for sending the messages throughout the travel, GPS attached car produce huge amount of data at every second, which helps to manage traffic with collected data at regular intervals. The data collected includes the data of traffic commands, number of vehicles, road condition ,public movement and much more information which may be in visual form, audio or text form. The data to be collected is also depends on size of the city or location but, the budding data could be out of the blue large to make a judgment and symbolizing regular traffic situation to regularize the travelers.

Internet of things (IoT) is another new promising technology of the present world. Smart home is its application where widget swap information among themselves for getting home in order such as sensors in a refrigerator. The sensors of the refrigerator on scanning furnish the available amount of different chattels list. This list will help to prepare a list of items to be purchased and also same list is promoted to a nearby super market of the choice. In the similar way, Smart cities can be made intellectual by handing out the data of interest composed at different city points. For instance, regulating city traffic in pick time such that pollution levels at city squares do not cross a marked threshold. Such applications need processing of a huge data that emerge at instant of time.In this way Big data are applicable in all most all streams such as environment, agriculture, governance, house holdings, health, finance, security, meteorological etc., in which wide variety of data are stored in different forms. The process of Big Data involves the stages or steps as Data acquisition, Staging and Analytics as well as Visualization as shown in Figure 1. These data need to be processed to get perfect conclusion. But variety of data collection, data storage as well as analytics and visualization of these data are challenging tasks. In all these tasks Big data has to look for the techniques for all its processes and these process is difficult than its normal characteristics such as velocity, volume, variety and value. Comparing to RDBMS these characteristics are differ in nature. In RDBMS has properties such as ACID (Atomicity, Consistency, Isolation, and Durability) These are comprised in big data with only three properties such as CAP (Consistency, Availability and Partition tolerance).In any big data system, any two of these properties are achievable, not all three. The common features of Big data velocity, represents the emerging further on along the timeline and its appearance is rapid so velocity of data generation is of principal concern. Variety of big data is due to its causes of data generation that includes smart phones, sensors or social networks, etc. The types of data originate from these sources may include any form such as image, video, audio, text and data logs in either structured or unstructured format. The procedure of the extraction of information and hidden information from the originating data is passed on by value. The process of extracting hidden information from emerging data is considered as value of big data. The size of the data ambiguity in sizes so the term voluminous data is misnomer for the big data. Here data are from different resources and also expand. The collection of such data explore the hidden information and patterns through many analysis. In traditional, Three-Tier Architecture programming model such as in J2EE, the applications are centralized in a centralized application tier and data is brought into this application over the network and then processed. But in case of big data architecture, the overheads cannot be handled by the network and handling will leads to saturation of the network, inefficiency as the data size in terms of terabytes or more so the applications with independent libraries need to move to data which is distributed across the nodes. So, the big data architecture is designed to deploy the code centrally

whereas applications are allowed to move these processing nodes before execution. With this configuration of big data the programming models to be used are Massively parallel processing(MPP) database system, MapReduce systems, In-memory database systems and Bulk synchronous parallel(BSP). Among these four MapReduce systems which are more generic purpose and it includes Hadoop[1,2,3].

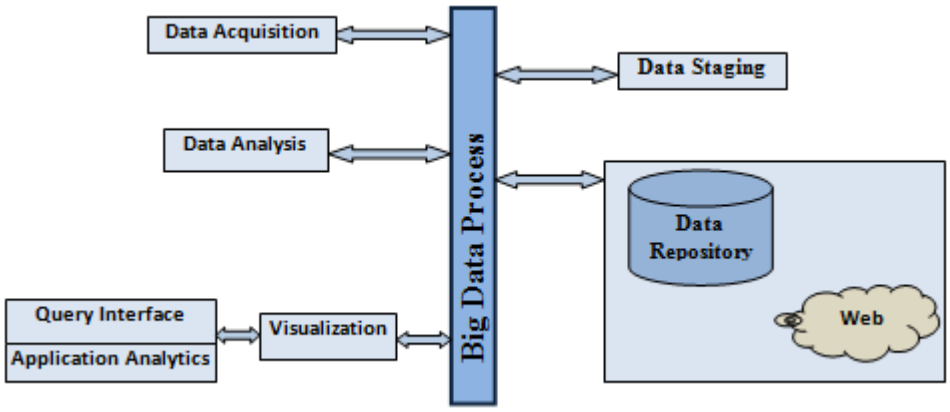

Figure 1: Overview of Big Data Process

2. Big Data Classification

As we mentioned Big data are collection of different data set. These data can be classified with aspect to the following: Content format, Data stores, Data Staging, Data processing and Data sources. All these classification is based on characteristics as well as convolution of each class and shown in Figure 2.

- **Content Format:** This class includes the structured, unstructured as well as semi structured data. Structured data use SQL or programming language in RDBMS. Unstructured data includes text message, audio, video and social media data. These data have not specific format and these data keep on increasing with use of smart phones etc., Semi structured data not follow conventional database systems rules. But, semi structured data may be structured also in some cases.
- **Data stores:** Document-oriented, Column-oriented, Key-value, Graph database are part of the data stores. The standard formats like JSON, pdf, xml are supported by these document oriented data stores. To store in column oriented data column must make use of tables with columns and these representation is differ from data base management systems. The graph model with nodes, edges, properties of with other in a table is mentioned. With key-value which is an alternative relational database system to store and access data of large size.
- **Data Staging:** This class includes Cleaning, Transform and Normalization. Cleaning process is removal of incomplete as well as unreasonable data. The cleaned data has to be prepared for the analysis in a required format in transform process, and the process of structuring database schema to minimize redundancy which is termed as normalization.
- **Data processing:** It process in Batch ie Map Reduced based system or Real time based systems. In many organization, for long-running job or tasks, MapReduce based systems are adopted. This arrangement helps to scale the large cluster of thousands of nodes' machine. Real time based systems are powerful and scalable streaming system as S4. S4 is pluggable platform of distributed computing systems to allow programmers to conveniently develop applications for processing continuous unbounded streams of data with partially fault tolerant.
- **Data sources:** Data are collected from sources such as Social media, from sensing devices, Machine-generated data, Transactional data, Uniquely identifiable Internet objects which are called IoT data [2,3,4].

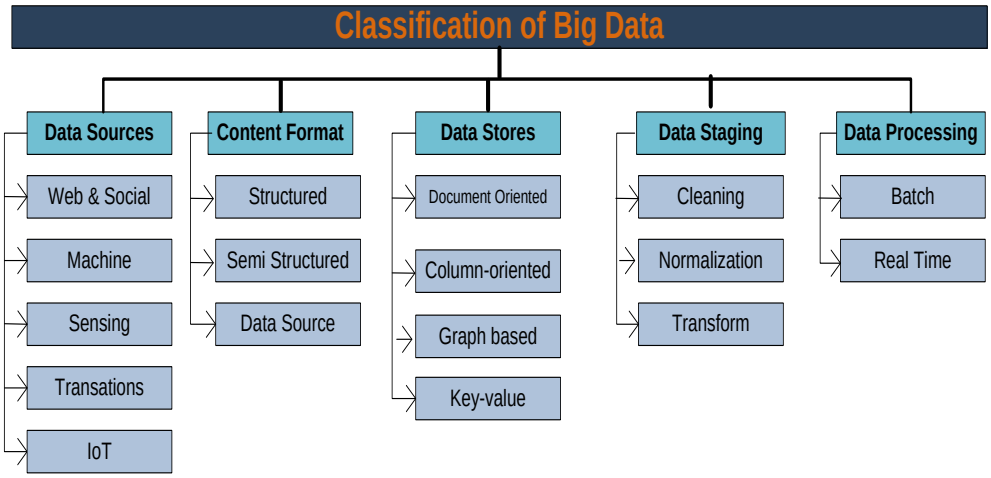

Figure 2: Big Data Classification

3. Tools of Big Data

The tools of big data are listed in Table 1[6].

Table 1: Tools and Feature

Tool	*Feature*
Hadoop	• Caliber to process tremendously large data of both structure as well as unstructured formats, dependably reproducing portions of data to nodes in the cluster and preparing it for the local machine and is with a new approach with objective to minimize the data movement to store and process the complex data.
MapReduce	• It is a programming model and support for writing applications and works to quickly process massive amounts of data in comparable on large clusters of compute nodes.
GridGain	• It is in-memory processing for fast analysis of real-time big data with java based programming.
HPCC	• The abbreviation for HPCC *is "High Performance Computing Cluster".* It privileges superior performance than Hadoop and delivers on single platform, a single programming language and a single architecture.
Storm	• Storm is not like batch processing systems and is extremely scalable, fast and works with all most language.
Cassandra	• It is a highly scalable NoSQL database to monitor massive data across multiple data centers and the cloud.
HBase	• It stores non-relational data for Hadoop and is g a column-oriented database management system, linear and modular scalability, strictly consistent reads and writes, automatic failover support and much more also suits for sparse data sets. Java is used to write applications such as Avro, REST and Thrift. Its features include.
MongoDB	• The documented oriented storage for NoSQL written in C++, full index support, high availability and replication. It supports to scale horizontally without effecting compromise functionality.
Neo4j	• It boasts performance improvements of up to 1000x or more when in comparison with relational databases. Stores data structured in graphs instead of tables and is a disk-based, fully transactional Java engine. Organizations can purchase advanced and enterprise versions from Neo Technology Developed by Neo Technologies, which is the world's leading graph database.
CouchDB	• Using query with Java Scripts or web it allows to access the stored JSON documents and also offers fault tolerant storage with distributed scaling. The tools comfortable for real-time change notification, web administration and document transformation.
Splice Machine	• This tools is to derive real time actionable approaches for faster development with SQL with of support of hardware and it supports for the languages such as .NET, Java and Python, JavaScript/AngularJS.
MarkLogic	• Suitable for the heavy data load and these data can be accessed though real time updates, it also provides geographical data with content and location relevance along with data filtering tools and supports flexible API's such as Client API, Node.js, NoSQL .It also help to implement a architecture for reference.
Google Charts	• Tool capable for visualizing data from hierarchical tree maps, simple charts. With JavaScript, data can be stored, modified as well as filtered with connection to website or outside website.

4. Big Data Challenges

The main challenges of big data are many. The most important are listed here. The scarcity of human resources with statics and mathematical background with statistical computing and statistical learning concepts which are research areas of assured results. Machine learning algorithms of both supervised as well as unsupervised algorithms leads to series of problems while processing big data so, these algorithms still need to be improved.

Big data is also facing problem of data integrity which can cause due to transmission of huge data over the network which requires large bandwidth. These problems can be solved by cloud technology to some extent. Another hardware challenge is regarding hard disk drive. For the high speed input and output transaction of data it requires high technology such as solid-storage device (SSD) as well as phase change memory(PCM) and these are promising techniques for big data. In this way there are many challenges and research issues. To drive the big demands it is very essential to overcome these challenges to develop fortified big system.

The central Govt. of India has launched soil health card which is an initiative with big data technology. Similarly we can implement for harvesting for rainfall water, forest monitoring, whether forecasting etc. The research labs for big data are Xerox Research lab, IBM Lab.

5. Programming modules of Big Data

There are four programming modules of big data such as

- In-memory database systems : Examples are Oracle Exanlytics and SAP HANA.
- Massively parallel processing (MPP) database system examples are IBM's Netezza as well as EMC's Greenplum.
- MapReduce systems which includes Hadoop, which is the most general-purpose of all the big data systems.
- Bulk synchronous parallel (BSP) systems, Apache HAMA and Apache Giraphe are the examples.

6. Conclusion

The processing of big data into many streams is increasing in geometric progression, the development is not noticeable in manufacturing field where we need to cover with data mining also as a part of big data. This research is most promising and also impact on the productivity of the manufacturing industries. This article has tried to bring some

aspects of the big data in to one place to help the beginners of the field.

References

[1] Nathan Marz , James Warren, "Big Data A Principles and best practices of scalable real-time data system", Manning Publications Co. 2015.
[2] Rajendra Arkekar, "Big Data Computing" , CRC Press, 2014.
[3] Keon Myung Lee, Seung-Jong Park, Jee-Hyong Lee, "Soft Computing in Big Data Processing", Springer International Publishing Switzerland 2014.
[4] Hrushikesha Mohanty, Prachet Bhuyan, Deepak Chenthati, "Big Data A Premier", Springer India 2015.
[5] Madhukar Dayal, Sachin Garg and Rubaina Shrivastava, "Big Data: Road Ahead for India", IMJ, Volume 6 Issue 2, July - December 2014.
[6] http://www.cbronline.com/news/big-data/analytics/10-of-the-most-popular-big-data-tools-for-developers-4570483.