International Journal of Recent Trends in Engineering & Research
ISSN (ONLINE) : 2455 – 1457
IMPACT FACTOR : 4.101

Big Data: A Challenging Technology

Kumari Seema Rani[1], Sonia Kumari[2], Manvendra Yadav[3]
[1,2] Shyama Prasad Mukherjee College (for Women), University Of Delhi, India
[3]Department of Computer Science, Atma Ram Sanatan Dharma College, University Of Delhi, India

Abstract- Big data is an emerging trend of information technology, large amount of data are growing fast these days. To handle this huge amount of data big data is used. Based on the development of big data in this paper we describe development trend, characteristics, processing of data, structure of data, and technology which is used. Big data handles large amount of data in each and every sector of today's scenario, which generates large amount of data, like communication, media, entertainment, healthcare providers, natural resources, manufacturing, education, governance, insurance, transportation , energy and utilities etc. This technology is merged with several emerging technology of this era like relationship between cloud computing and big data, relationship between IoT (Internet of things) and big data, relationship between hadoop and big data etc. At last, we sum up the concept of big data with their applications.
Keywords- Big Data**,** Cloud Computing**,** IoT, Hadoop, MapReduce

I. INTRODUCTION

Big data is a booming technology of IT world. The term Big Data is a combination of two words big and data, which means a very large data set in size. Basically big data is same as small date only difference in size. For different size of data, it requires different types of tools, techniques and mechanism. In traditional computing technique it was not possible to generate large volume of digital data sets and not possible to analysis but after the invention of most demanding word big data, it is in reality. Different researchers have defined big data in different way. Wikipedia mentioned that big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time [2]. As per Forrester, "Big data is a frontier of a firm's ability to store, process and access (SPA) all the data it needs to operate effectively, make decisions, and reduce risks and server customers." As per Gartner, "Big data in general it is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." According to O'Reilly, "Big data is data that exceeds the processing capability of conventional databases. The data is too big, moves too fast, or doesn't fit in strictures of your database architectures. To gain value from this data, you must choose an alternative way to processes it." Whereas IBM defined it in a significant manner, "Big data is the data characterized by three attributes: volume, variety and velocity." After that Oracle defined the term in a better way, "Big data is the data characterized by four attributes: volume, variety, velocity and value." IDC (International Data Corporation) defined big data as the combination of 4V (Volume, Velocity, Value, Variety) index. Nowadays people using internet rapidly and living online, they express their ideas, views, logic, opinion etc. Big data is not about the size of data, it's about the value within the data. The value of data refers to our attitude, like, dislikes, opinion and perspective, ideas, views etc. We are generating huge amount of data in social media, in traditional model only few companies are generating data and all other companies are consuming the data but in new model, all of us are generating data, and all of us are consuming the data. These data contains a lot of information with a lot of noise. To extract the signal from the noisy data is the key of big data. Organization like Facebook, Google, Linkden, ebay were built around big data from the beginning. The rest of the

paper is organized as follows, section 2 discuss about history of big data, section 3 describes about characteristics, section 4 considers development trend, section 5 deals with processing of big data, section 6 covers about structure of big data, section 7 analyze relationship between some existing different technology with big data, section 8 discuss about application of big data, section 9 covers conclusion.

II. HISTORY

The basic idea behind the big data is from the world of computer science and econometrics [1]. The company who introduced big data firstly is Mc Kinsey & Company. In June 2011, a report was generated by Mc Kinsey & company which describe the tools, techniques, keys, impact, and application on big data. Mc Kinsey affirms that big data is one data set whose size exceeds the typical database software acquisition and storage, management and analysis. The concept of database machine (a technology specially used for storing and analyzing data) is emerged in 1970s .In 1980s a share nothing (a parallel database system to meet the demand of increasing data volume) technique was proposed. This system architecture is based on the use of cluster and each machine has their processor, storage and disk. The first successful commercial parallel database was teradata system, On June 1986, teradata became first parallel database system with the storage capacity of 1TB. However in late 1990s, the features of parallel database were widely famous in database field. With the development of internet services, indexes and query contents were growing very fast, therefore search engine company faced a lot of problems to handling these too much data. Google invented GFS and MapReduce programming model to handle these challenges by data management and analysis at the internet scale [8,9]. In January 2007 Jim Gray, a pathfinder of database software, thought the only way to cope with such challenges is to develop a new generation of computing tools to manage, visualize, and analyze huge data. In June 2011, EMC/IDC published a research report titled Extracting Values from Chaos, which introduced the theme of big data first time. Over past few years back all top rated companies like IBM, Facebook, Amazon, Oracle, Microsoft, Google etc have started their own big data projects.

III. CHARACTERISTICS

As per Victor Meyer-Schonberg, the meaning of big data is using the method of all the data but not random analysis (sampling). There are four types of characteristics of big data they are Volume, Velocity, Value, and Variety [3]. Volume stands for large data set, with the huge amount of data the storage capacity expands from TB (Tera byte) to YB (Yotta Byte). In 2000, a normal personal computer might have had 10 gigabytes of storage but nowadays facebook consumes 500 terabytes of new data every day [2,3,8]. Velocity refers to streaming or mobility of data from million of events per seconds to respond the process. Some real systems support millions of concurrent users, each producing multiple instructions per second. Variety of data deals with different types of data generated by relational or non-relational database. It is not only contains number, text, strings but also included geospatial data, 3D data, audio, video, xml, multimedia etc[12,14].. With the enhancement of technology in real time the data can be structured, unstructured, and log files of social media etc. Traditional database systems were designed to smaller volume of structured data while big data analytic includes different types of data.Value or Veracity refers the significance of big data application which is uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations[10,13]. Big data refers to large amount of data sets but also refers to large amount of data processing techniques. After data collection, storage, analysis the extracted value data is achieved. Walmart handles more than one million customer transaction every hour whereas facebook handles 40 billion photos from its user base. Data models like key value, graph, document, column family Hadoop distributed file systems, Hbase, Hive etc are used for big data stores. We can choose correct data store based on data characteristics [1]. Every day we create 2.5 quintillion bytes of data. As a survey of IBM, 90% data of the world today has been created alone

in last two years. Facebook generate more than 10TB data daily while twitter generate more than 7TB data daily. At 2015, 4.4 million IT jobs in big data.

IV. DEVELOPMENT TREND

In traditional databases (data warehousing) data are organized in row and column and employ data – cleansing method on data, and it works only smaller volume of structured data with predictable updates whereas big data covers a diverse format with both batch and streaming format in wide areas like geospatial data, 3D data, audio and video, structured data, unstructured text including log files, sensor data, and social media[7,9]. The application platforms for big data is big data analytics, text mining, video analytics, web log mining, scientific data exploration, intrinsic information extraction, graph analytics, social networking, in-memory analytics, statistical analytics and predictive analytics. Although for traditional database query language like SQL is used but for data-intensive computing languages for batch processing and stream computing NoSQL, MapReduce, Hadoop etc programming languages are used. As a survey, every minute, we send 204 million emails; generate 1.8 million facebook likes, send 278 thousand tweets, and upload 200 thousand photos to facebook. The challenges of big data are analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating, and information privacy [1, 13].

V. PROCESSING OF BIG DATA

The phases of big data in organization are basically deals with Exploration & Research, Assessment & Definition, Execution, and Operation & Enhancement. Big data processing techniques consider big data sets on a very large scale like terabyte, petabyte etc. The three solutions of big data processing is batch processing, NoSQL, Real-time processing. Batch data processing is an efficient way of processing high volumes of data where a group of transactions are carried out for a period of time[2,3]. It works on scalable data and large amount of static data and supports parallelism of distributed nature of processing. Data is collected, entered, processed and then the batch results are produced. Hadoop is concerned with batch data processing. Batch processing considers separate programs for input, process and output. Example of batch processing is payroll and billing systems. Real time data processing contains a continual input, process and output of data, data must be processed in a small time period. It contains low latency with continuous unbounded streams of data it also supports parallelism of distributed nature of processing. Example of real time data processing is Radar systems, customer services and bank ATMs etc. Hybrid computation is the combination of both batch and real time (volume and velocity) data processing. It contains massive data as well as streaming data. The processing pipeline contain data acquisition, data storage, data analysis and results [2,12].

VI. STRUCTURE OF BIG DATA

The structure of big data has three types: they are structured, semi-structured and un-structured. Structures data contains a defied data type format means most traditional data sources, example of structure data is like transaction and OLAP (On-line Analytical Processing). Semi-structured data contains textual data files with different pattern and it have many sources of big data like XML data file that is self describing and defined by an XML schema, unstructured data has no inherent structure and it is usually stored in different file formats in short we can say it stores data like audio data, text, document, PDF, images and video data. As a survey 10% of data of big data is structured data, 10% semi-structured and 80% data is unstructured data. The structure of big data team is also divided into four types; they are Data Scientist, Business Analyst, Data Integration Specialist, and Application Developer. The data scientist has the knowledge of computer science; he/she has also aware about high performance applications, statistics, economics, mathematics, ad probabilistic analysis skills. Business Analyst has the capability to effectively translate business expectations into specific data analysis results. Data Integration Specialist has the experience in data extraction, transformation, loading and data transformation, for cleansing and delivery to target

system[12,13,14]. Application Developer is the technical resource person which has skills for programming and testing and testing parallel and distributed system.

VII. TECHNOLOGIES

In this section we discuss about several technologies which is closely related to big data. It includes IoT, data centers, hadoop, cloud computing.

Relationship between Cloud Computing and big data- Big data is basically extracting Value out of variety, velocity and volume from the information assets available, whereas cloud contains on-demand, scalability, pay-per-use, elasticity etc. Cloud Computing is closely related to big data. By using on-demand feature of cloud computing we get enough computing power and distributed storage to handle the 3V data problem of big data without investing in infrastructure assets, and cloud provides the elastic on-demand computer required for the same. Cloud is used "As-a-service" model by hiding the complexity of building a scalable elastic self service, and this is the requirement for big data processing. In a similar way hadoop hides the complexity of large scale distributed processing from the end user. Another feature of cloud is pay-per-use, which is used for delivering value to enterprise by lowering the cost of ownership. If data sources are spread around the world, we can use public cloud to allow those sources for faster access to storage. For more storage a cloud platform can dynamically expand to accommodate the storage needs, if data don't need the storage anymore cloud can shrink it and no need to pay more. NoSQL is a set of data retrieval and storage tools, cloud computing providers host have their own version of NoSQL tools like Amazon and Hadoop. Big data represents content and cloud computing is infrastructure [11,12].

Relationship between IoT and big data- The huge amount of data produced by IoT would be useless without the analytic power of big data. Without the IoT, big data wouldn't have the raw materials from which to fashion solutions that are expected of it. IoT generates big data which has different characteristics as compared to general big data due to different types of data collected; most classical characteristics include heterogeneity, variety, unstructured feature, noise, and high redundancy. As per HP, the current IoT data is not the dominant part of big data, by 2030, the quantity of sensors will reach one trillion and then the IoT data will be the prominent part of big data[15,16]. A report by Intel mentioned that big data in IoT has three features that conform to the big data paradigm: abundant terminals generating masses of data; data Generated by IoT is usually semi-structured or unstructured; data of IoT is useful only when it is analyzed. Big data is about of data but IoT is about data, device and connectivity. The motive of IoT is creating smarter products, delivering intelligent insights and providing business outcomes. As millions of devices get connected IoT activate inflow of big data. In short we can say that "IoT is the sense, big data is the fuel, and artificial intelligence is the brain to realize the future of a smart connected world."

Relationship between hadoop and big data – Big data is a concept for handling large amount of data sets where as hadoop is just a single framework out of many tools used for handling big data. It is used for batch processing. Hadoop is open source software (free), java-based programming framework that supports the processing of large data sets in distributed computing environment. Hadoop is part of Apache project sponsored by the Apache software foundation. It is based on Google's MapReduce programming model for large scale data processing, MapReduce is a software framework in which an application is broken down into a large number of small parts. The current hadoop system contains Hadoop Kernal, MapReduce, the Hadoop distributed file (HDFS). Companies like Google, Yahoo, and IBM use hadoop framework for applications involving search engine and advertising[12,14].

Big Data Acquisition Engine- Security is a major part for any technology. The rule engine and finite automaton merged together, for verify the security and correctness of big data acquisition flow

[4,5]. When a new collection node is added, the rule engine automatically makes the whole system more flexible and scalable.

MFA (Mean Field Analysis) - The MFA is effective and reliable in evaluating the performance of very large big data. It is able to model performance of big data architectures indices in a bounded time. It can set up and evaluate in a lesser time, because it does not depend on the number of instances. MFA technology is very effective for assessing the performance of big data.

VIII. APPLICATIONS

Communication, Media and entertainment- The challenges in the communication, media and entertainment industry include collecting, analyzing and utilizing customer perspective, it is effectively used for social media and mobile content, it works on understanding patterns of real-time, media content usage. The industry organization basically analyze customer data along with behavioral data to create description of customer profiles and it can be used for create content for different target perspectives, analyze measure content performance, and recommend content on demand[15,12].

Healthcare Providers- The healthcare sector has access huge amount of data. To create visual data for faster identification and efficient analysis of healthcare information free public health data and Google Maps is used by the University of Florida. It is used in tracking the spread of chronic disease.
Education- Big data is effectively used in higher education, An Australian University, University of Tasmania with over 26000 students, have developed a learning and management system which measure the overall progress of a student over time, that tracks among things, when a student logs onto the system, how much time is spent on different pages in the system. It is also used to measure teacher's effectiveness to analyze a good experience between both students and teachers[14].

Manufacturing and Natural Resources- Big data is used to solve manufacturing challenges as well as in natural resource industry, big data can help to reduce defects and control cost during automated production. Every details of each part of the product can be track from manufacture to installation for the better solutions. To help suppliers the performance is also evaluated on the basis of monitoring defect rates and on time delivery.

Government- In public organization, big data has a very wide range of applications including: financial market analysis, fraud detection, energy exploration, health related research and environmental protection. The Food and Drug Administration (FDA) is using big data to detect and study patterns of food related illness and diseases. It is used for faster response which has led to faster treatment and less death. Big data is also used in security concern for several different use cases.

Insurance- Big data is used in industry to provide customer observation by analyzing and predicting customer behavior through data derived from social media, GPS enabled devices and CCTV footage, for transparent and simpler products, Predictive analytics from big data is used for faster services, fraud detection has also been enhanced.

Transportation- Big data is also used in traffic control, route planning, intelligent transportation systems, congestion management (by predicting traffic conditions), revenue management, technological enhancements, and logistics for competitive advantage; it is also used in tourism for travel arrangement in route planning to save fuel.

Energy & Utilities- Applications of big data in used in the energy and utilities industry. Smart meter readers allow data to be collected almost every 15 minutes. The granular data is used to analyze

consumption of utilities better. In Utilities Company the use of big data performs better workforce management which is useful for recognizing errors and correcting them.

Retail and Whole sale trade- Big date is used to reduce fraud and it performs timely analysis of inventory, a part from this the wholesale store and retail gathered data from unutilized data, this is derived from customer loyalty cards, POS scanners, RFID etc.

IX. CONCLUSIONS

Big data is a demanding technology of IT industry; to improve the decision making and enhance the business skill to another level the data is required. So the data is generated by sensor enabled machines, mobile devices, social media, cloud computing, satellites help different organizations to enhance their features. It is the next generation of data warehousing and business analytics. According to Susan Hauser, corporate vice president of Microsoft about big data is; "Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives," To understand Indian citizen‘s sentiments and ideas through crowd sourcing platform, Indian Prime Minister‘s office is using Big Data analytics which is available on www.mygov.in social media and through this we get a clear picture of common people‘s thought and opinion on government actions[6]. This research paper discusses the current trend, processing, characteristics, technologies and applications of big data. Big data professional is treated as the bridge between raw data and useable information. They have the ability to manipulate data on the lowest levels and know how to interpret trends, patters, outliers in different forms. Big data is a platform which brings exciting and new features of company, which elaborates how IT trend is faster by traditional technology. Big data is future; currently more researchers are giving their contribution in this field. As we know data is growing in faster rate to handle such amount of huge data there is need of such tools and technology, hadoop is the most emerging framework used by mostly organization like facebook, amazon, IBM, Yahoo, Microsoft etc.

REFERENCES

[1] Church, A.H. and Dutta, S. (2013) The Promise of Big Data for OD: Old Wine in New Bottles or the Next Generation of Data-Driven Methods for Change. OD Practitioner, 45, 23-31.
[2] Wikipedia, Big Data. http://en.wikipedia.org/wiki/Bigdata
[3] Mayer-Schönberger, V. and Cukier, K. (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston
[4] Xu, X.B., Yang, Z.Q., Xiu, J.P. and Chen, L.I.U. (2013) A Big Data Acquisition Engine Based on Rule Engine. The Journal of China Universities of Posts and Telecommunications, 45-49. http://dx.doi.org/10.1016/S1005-8885(13)60250-2
[5] American Journal of Industrial and Business Management, 2015, 5, 192-197, Research of Big Data Based on the Views of Technology and Application, http://dx.doi.org/10.4236/ajibm.2015.54021
[6] http://dataconomy.com/indian-government-using-big-data-to-revolutionise-democracy.
[7]http://datasciencedegree.wisconsin.edu/blog/big-data-headlines-of-january-2016
[8]https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big- data/&refURL=https://www.google.co.in/&referrer=https://www.google.co.in/
[9]https://datafloq.com/read/big-data-history/239
[10] http://www.dataintensity.com/characteristics-of-big-data-part-one/
[11] http://www.dataversity.net/the-top-10-big-data-analytics-technologies/
[12]Survey of Recent Research Progress and Issues in Big Data, Bo Li, boli@seas.wustl.edu (A paper written under the guidance of Prof. Raj Jain)
[13] Big Data – Concepts, Applications, Challenges and Future Scope, Samiddha Mukherjee1, Ravi Shaw.
[14]Research of Big Data Based on the Views of Technology and Application, American Journal of Industrial and Business Management, 2015, 5, 192-197 Published Online April 2015 in SciRes.
[15] Big Data computing and clouds: Trends and future directions, Marcos D. Assunção a,□, Rodrigo N. Calheiros b, Silvia Bianchi c, Marco A.S. Netto c,
Rajkumar Buyyab,□

[16] The anatomy of big data computing Raghavendra Kune1,*,†, Pramod Kumar Konugurthi1, Arun Agarwal2, Raghavendra Rao Chillarige2 and Rajkumar Buyya, SOFTWARE: PRACTICE AND EXPERIENCE Softw. Pract. Exper. 2016; 46:79–105, Published online 9 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2374.
[17] Big Data: The Structure & Value of Big Data Analytics, Hak J. Kim Hofstra University hak.j.kim@hofstra.edu

