Basics and Applications of Big Data

Ashwin Ravindra Khandelwal[1], Lovely Mutneja[2], Prachi Thakar[3], Priyanka Patil[4]

[1]B.E. (CSE) IV Year, Prof. Ram Meghe College of Engineering and Management, Amravati, Maharashtra, India
[2,3,4]Assistant Professor, Prof. Ram Meghe College of Engineering and Management, Amravati, Maharashtra, India

***

Abstract - *We are living in an era where a single person is generating ample amount of data per day. To apply some logic on that data and to extract information or some pattern from that data is known as big data. The paper focuses on the analytical part of the core concepts of Big Data, as analysis is the natural part of our life because it is much essential to make any sort of decisions over certain facts. Data now-a-days is increasing so far so as to touch the sky. Thus, it became necessary to analyze the data storage techniques one of which is the big data constituting of IoT i.e. Internet of Things & Cloud Computing. This article helps researchers and developers to explore and abstract big data at multiple stages. Moreover, it provides a latest platform for researchers to develop the solution, based on the challenges and open research issues as discussed in this paper.*

***Key Words*: Big Data, IoT, Cloud Computing, Big Data Analysis, RFIDs**

1. INTRODUCTION

Due to excessive digitalization of almost complete world, we are completely flooded with a large amount of data & information today. The information is generated from various sources at an uncommon scale. The ancient tools & certain techniques were unable to cope with such a large & complex data and thus the evolution of Big Data took place. For the purpose of checking the effects of Big Data, a survey was held by IBM Corporation in visual study pattern; the data that is uploaded daily is not less than 100 Tb or Terabytes and it can be said that 35 Zettabytes of data will be covered till 2020. The concept of big data refers to the grouping of extremely large and complex datasets from databases that becomes difficult or even sometimes impossible by using the traditional means of managing the database or data processing tools for data processing of such data. The term Big Data was coined by John Mashey who in the period of 1990's and was initially used by Roger Magoulas from O'Reilly media in the early 2005The term was made popular because the extremely large amount of information that was available and are made available by information administration was unable to handle due to its size and nature that is miscellaneous. The only thing that is capable of handling data storage, capturing data, data analysis, enhanced search, sharing, transfer, querying, visualization, updating and information privacy is Big Data. The term Cloud computing refers to anything that can be made available on the internet no matter whether it is data or resource or any other services that is required regardless of time and place. It depends on sharing the available resources or to make it easily available and accessible. Cloud computing is a broad term which uses the 'pay-as-you-go' prototype that enables the user to pay only for that much which is being consumed by them. Besides this, the idea that connects all the real world things with the internet is internet of things, commonly abbreviated as IoT. IoT is nothing but the combination of various technologies which is accredited next to RFID, Internet communication and technologies, smart sensors, etc. RFID is an acronym for "radio-frequency identification". RFID is often used to identify the things automatically with no human intervention and themselves fix those things into the computer systems. The Machine-to-Machine technologies also referred to as M2M technologies and the advancements of smart phones and Internet communication gave rise to IoT.

2. DEFINITION

Initially no one was aware with the concept of Big Data. And it was known to industries or companies in varying courses by different discussions and gatherings. Different types of definitions were put forth by different researchers as: The first documented use of the term *big data* appeared as: data sets are generally quite large, taxing the main memory capacities, capacity of the local disk, and even remote disk. This can be referred as the Big Data problem [15]. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources[16]. Also it can be defined as Enormous Data alludes to datasets whose size are past the capacity of regular database programming apparatuses to catch, store, oversee and break down. Moreover Big Data is Enormous Data will be information that surpasses the handling limit of ordinary database frameworks. The information is present in a tremendous amount, but can move too quickly, or does not fit the structures of existing database designs. To pick up quality from these information, there must be an option approach to process it [4].One more definition defines big data as a holistic approach to manage, process and analyze

5V's in order to create actionable insights for sustained value delivery, measuring performance and establishing competitive advantages [5].

3. CHARACTERISTICS

Big Data has four main characteristics generally referred to as “5V”: Volume, Velocity, Variety, Variability and Veracity.

- **Volume:** Big data instead of sampling just observes and tracks what happens with the data. Whether the information or data can be thought of as Big data or not is decided by the size of data and the potential of that data.
- **Velocity:** Big data is often available in real-time; the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
- **Variety:** Big data can draw the data from text, images, audio, video or from any other forms. Along with this, by the means of ‘Data Fusion’, big data can fill the missing information. Data fusion integrates data from various sources for more accuracy and completeness.
- **Variability:** Managing the data and to handle such an ample amount of data is a difficult task. Variability can be due to inconsistency or multidimensional nature of available data or information.
- **Veracity:** Accuracy of data can be greatly affected by its quality and format.

Along with the following two main characteristics:

- **Machine learning:** Big data never asks any questions related to “Why’s” instead it simply detects patterns.
- **Digital footprint:** The footprint or byproduct of digital interaction is Big data which too is free-of-cost.

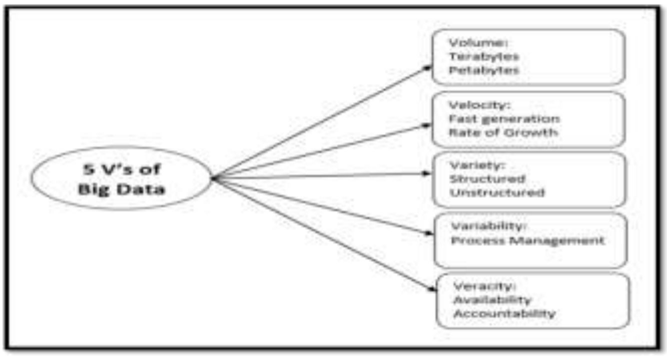
5 V's of Big Data
Volume: Terabytes Petabytes
Velocity: Fast generation Rate of Growth
Variety: Structured Unstructured
Variability: Process Management
Veracity: Availability Accountability

Fig -1: Characteristics of Big Data

4. APPLICATIONS

Thus being such a wonderful and useful technology there has been a certain increase in the demand of Big data for information management specialists so much, such that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year about twice as fast as the software business as a whole.

There are many different Applications of Big Data as follows:

- **Government**

Big Data allows the efficiencies in terms of cost, productivity, and innovative ideas at different governmental processes. Besides these there are certain flaws or defects which cannot be overcome normally. Thus big data came up with Data Analysis. Data analysis is collaborated with the existing services of various governmental bodies to create new and more innovative processes to deliver the required outcome.

For ex. In India Big data analysis was tried out for the BJP to win the Indian General Election 2014. The Indian government used large number of techniques to ensure how the Indian electorate is responding to government action, as well as ideas for policy augmentation.

- **International development**

There is more need for research and development through big data on the effective usage of communication and information sources of technologies, but at is seems to be an open international challenges.

The increasing advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management.

Longstanding challenges for developing regions such as inadequate technological infrastructure and economic and human resource scarcity sharpen the existing concerns with big data such as privacy, imperfect methodology, and interoperability issues.

- **Finance**

The Financial Market Data uses the concept of Technical Analysis for the purpose of analyzation.
Use of non-finance data for market estimation is sometimes called alternative data.

- **Manufacturing**

The greatest benefit of big data for manufacturing is improvements in supply planning and product quality in TCS 2013. Research and further study proofs that Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unzip doubts such as inconsistent component performance and availability. Historical data along with huge amount of sensory data leads to the enrollment of Big data in the field of manufacturing. The big data that is generated by such methods acts as the input for the predictive tools and technologies and preventive methodologies. For ex. Prognostics and Health Management (PHM).

- **Healthcare**

By providing diagnosed medicine and prescriptive analytics, big data analytics has helped healthcare improvement within the short span of time. Some areas of improvement are more characterized than they are actually implemented. The level of data generated within healthcare systems is not insignificant. Big Data has placed in market the concept of mHealth, eHealth and wearable technologies with am aim that the volume of data will continue to increase. This includes recorded data from electronic healthcare services, imaging data, data generated by patients, sensor data, and other forms of difficult to process data.

- **Media**

It is a big question that how the media utilizes big data? It is mandatory requirements for the media process to have some content. The industrial sectors are going far away from the traditional methods which use environments such as newspapers, magazines, or television shows and instead taps into consumers with technologies that reaches to the targeted people in a limited amount of time and cost. The ultimate aim is to serve or convey a message or content that is statistically spoken in line with the consumer's mindset.

- **Targeting of consumers (for advertising by marketers)**

Data-capture And Data journalism: Publishers and journalists use big data tools to provide unique and innovative insights and infographics.

The British public-service television broadcaster from Channel-4, plays a leading in the field of big data and data analysis.

- **Technology**

The prominently used shopping site Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. Linux-based is the core technology that keeps Amazon running. From a survey of 2005, they had the world's three largest Linux databases, which has a capacities of 7.8 TB, 18.5 TB, and 24.7 TB.

The Facebook application which is used in our daily life has a capacity of handling 50 billion photos from its user database.

From a survey of Google, it is observsed that roughly 100 billion searches were made per month.

Most commonly used Oracle and NoSQL Database is tested to past the 1M ops/sec mark with 8 shards and is expected to hit 1.2M ops/sec with 10 shards.

- **Information Technology**

The emergence of Big Data with the prominence took place especially since 2015, within Business Operations as a tool to help employees work more efficiently and streamline the collection and distribution of Information Technology (IT). The IT and data collection issues within the enterprises were resolved by the use of big data through IT Operations Analytics (ITOA). Big data principles were applied to the concepts of machine intelligence and deep computing so that IT departments can depict the potential issues and can get the possible solutions before the problems actually occur. Meanwhile, ITOA businesses started to play a major role in systems management by offering platforms that brought individual data silos together and generated insights from the whole of the system rather than from isolated pockets of data.

- **Education**

With a case study it is known that a McKinsey Global Institute study found a shortage of 1.5 million highly trained data professionals and managers and a number of universities which also includes the University of Tennessee and UC Berkeley that created masters programs to meet the requirements. Private boot camps have also developed programs to meet that demand, including free programs like The Data Incubator or paid programs like General Assembly so that each of them were able to learn the new theoretical as well as practical knowledge.

- **Retail**

Considering a retail example Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data which approximately equals to 167 times of the information contained in all the books in the US Library of Congress.

- **Retail banking**

With the help of security modules Card Detection System protects accounts worldwide. According to some estimates made in the banking sector, the business data volume worldwide, throughout all companies, doubles every 1.2 years.

- **Real estate**

For the home buyers to buy any estate, the factor of location accuracy needs to be met. For the accomplishment of the same, a large amount of agencies are working.

- **Science and research**

The Sloan Digital Sky Survey (SDSS) is a survey platform which started the collection of astronomical data in the year 2000, it collected such an amount of information in its first few weeks which was more than all data collected in the history of astronomy. A rate of about 200 GB per night was used by SDSS to reach information of more than 140 terabyte.

The cost of sequencing is highly divided by the DNA by 10,000 in the last ten years, which estimates approximately 100 times cheaper than the reduction in cost predicted by Moore's Law.

- **Sports**

Sports sensors can be developed by using Big Data to improve training and understanding competitors. The big data analysis can be used to predict the winners of specific sports. Similarly the future performance of the players can be estimated. Thus, player's value and salary is determined by data collected throughout the season.

The use of Big Data analysis in the field of Sports can be demonstrated by a Movie named as MoneyBall. Also undervalued players are identified with the big data only.

- **Computational social sciences**

APIs stands for Application Programming Interfaces. Everybody can use big data holders supplied APIs, such as Google and Twitter, for the purpose of research in the social and behavioral sciences. These APIs are available at free of cost.

Traditionally no algorithmic challenges were present which is made available by Big data. Thus, it became necessary to fundamentally change the processing ways.

Open research issues in Big Data Analysis:

Because of it's such a high importance, big data is included in various courses of academics and industrial placements and in industries too. Big data analysis are classified into these broad categories namely Internet of Things (IoT), cloud computing.

IoT:

IoT stands for Internet of Things. "The **Internet of things** (**IoT**) is the either the inter or intra-networking of the devices that are physically present such as vehicles, appliances often used at home and other items integrated with electronics, software, sensors, actuators, and network connectivity which enable these objects to connect and exchange data[17]. We can uniquely identify anything through its computing system using embedded systems but can be inter-operated within the existing Internet infrastructure". The concept of IoT is gaining more importance from the realistic world due to the development of mobile devices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. IoT devices are used to generate continuous streams of data and the study by the researchers develop tools to extract meaningful information from the generated data using machine learning techniques and data processing technologies. IoT is being exponentially accepted for collection of sensory data, which can be used in development of medical science and manufacturing contexts.

The term Internet of Things is defined as: "If we had computers that knew everything there was to know about things—using data they gathered without any help from us—we would be able to track and count everything, and greatly reduce waste, loss and cost. We would know when things needed replacing, repairing or recalling, and whether they were fresh or past their best"[18].

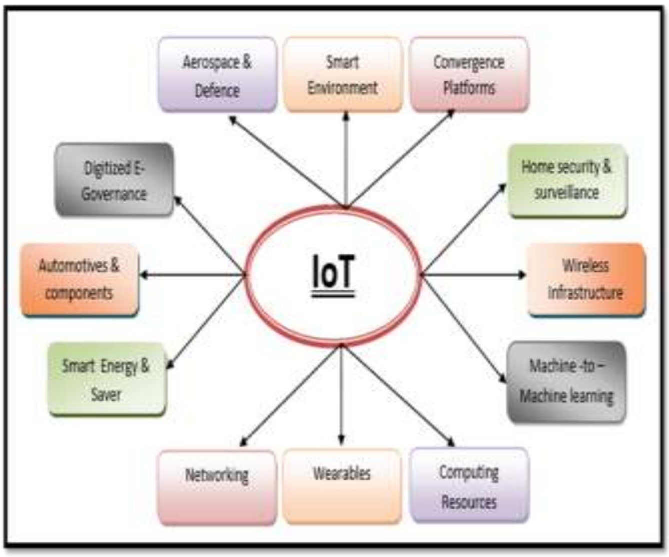

Fig -1: Applications of IoT with Big Data

Cloud Computing:

Cloud computing, often referred to as simply "the cloud", can be defined as 'the delivery of on-demand computing resources, from small scale applications to huge data centers over the internet which can be used on a pay-for-use basis' meaning that we only need to pay for the amount, that we have used . Cloud computing offers:

Elastic resources – can easily meet the user demand and is scalable.

Metered service - Amount to be paid is only for the space and time it consumes.

Self service - Provides high flexibility level by means of self-service.

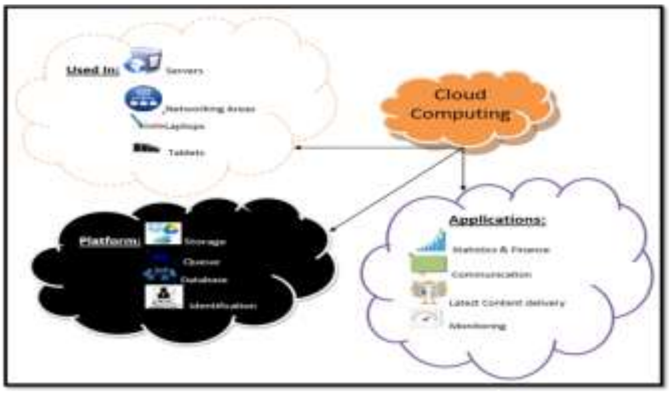

Fig -1: Cloud Computing

The concept related to the use of the virtual computers is known as cloud computing and in recent condition it has been one of the most robust big data technique. Both the technologies namely Big Data and cloud computing are developed with a view of developing a flexible, scalable and on-demand availability of resources as well as data. Data analytics and development are well supported by cloud computing, an application of big data. The cloud computing resources provides tools that allow data scientists & investigators and business analysts to collaboratively explore knowledge acquisition data for further processing and extracting fruitful results as the traditional ways were unable to solve the problems.

5. CONCLUSION

In this paper, the basic information and applications of Big Data is discussed which is the concept used worldwide since recent years dealing with the challenges and new innovative data management techniques and advancements. Various specified functionalities are being handled by Big Data theory. Also it has been minutely observed that the Big Data model will become more than thrice the upcoming years.

REFERENCES

[1]. Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". martinhilbert.net. Retrieved 7 October 2015.

[2]. Acharjya, D. P., & Kauser Ahmed, P. “A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools”. Article in International Journal of Advanced Computer Science and Applications February 2016.

[3]. Mishra, N., Lin, C. C., & Chang, H. T. “A cognitive adopted framework for IoT big-data management and knowledge discovery prospective”. International Journal of Distributed Sensor Networks, 11(10), 718390

[4]. Sharma, Sunny, and Prithvipal Singh. "A Review toward Powers of Big Data" (2016).

[5]. Feki, Mondher, Imed Boughzala, and Samuel Fosso Wamba "Big Data Analytics-enabled Supply Chain Transformation: A Literature Review". In 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 1123-1132. IEEE, 2016.

[6]. DT&SC 7-3: What is Big Data? YouTube 12 August 2015.

[7]. Lee, Jay; Wu, F.; Zhao, W.; Ghaffari, M.; Liao, L. "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications". Mechanical Systems and Signal Processing. **42** (1). January 2013.

[8]. "Tutorials". PHM Society. Retrieved 27 September 2016

[9].Online material from:
https://en.wikipedia.org/wiki/Big_data#cite_ref-67

[10]. Lampitt, Andrew. "The real story of how big data analytics helped Obama win". InfoWorld. Retrieved 31 May 2014.

[11]. Hoover, J. Nicholas. "Government's 10 Most Powerful Supercomputers”. Information Week. UBM. Retrieved 26 September 2012.

[12]. "News: Live Mint". Are Indian companies making enough sense of Big Data?. Live Mint. 23 June 2014. Retrieved 22 November 2014.

[13]. Biswas, Abdur Rahim, and Raffaele Giaffreda. "IoT and cloud convergence: Opportunities and challenges." In 2014 IEEE World Forum on Internet of Things (WF-IoT), IEEE, 2014. pp. 375-376.

[14]. Shah, Sajjad Hussain, and Ilyas Yaqoob. "A survey: Internet of Things (IOT) technologies, applications and challenges". In Smart Energy Grid Engineering (SEGE), 2016 IEEE, pp. 381-385. IEEE, 2016.

[15]. Kulshrestha, Sanatan. "Big data in military information & intelligence." (2016).

[16]. Online material available at:
https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#3327623c65a1

[17]. Online material available at:
https://petrowiki.org/Internet_of_things_(IoT)

[18]. “A Comprehensive Review: Internet of Things (IOT)” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. III (Jul.-Aug. 2017), PP 62-72 www.iosrjournals.org