An Overview on Big Data Analysis

Jeevitha I
Assistant Professor
Department of BCA & M.Sc. SS
Sri Krishna Arts & Science College, Tamilnadu, India

Giri Pai U
Student
Department of BCA & M.Sc. SS
Sri Krishna Arts & Science College, Tamilnadu, India

Mohana Prasath
Student
Department of BCA & M.Sc. SS
Sri Krishna Arts & Science College, Tamilnadu, India

Abstract

Now the world is moving digitalized. To bring enhancement in modern world, we are ongoing on in a new concept known as the big data. Almost eighty to ninety percent of businesses that are running today seek a new and better approach to remain competitive and profitable. To do this, big data leads them in a path that stays ahead of the curves. Thus big data is an approach that helps people to make their life more comfortable, profitable and compatible one. Big data plays a major role in planning important strategic and operational plans and implement them. Apart from business people also use big data for many other reasons. Many people browse, collaborate, and shop for goods and services online, big data gives them hand to perform all these tasks. Not only consumer's even business to business transaction takes place in this platform. This paper totally discusses about big data analysis, its history, various definitions, background and finally its applications. The platform which helps people in almost all means of their livelihood and makes their life comfortable is considered to be the best one. In such consideration big data is one of the best one in all means.

Keywords: Big Data, Internet of Things, Massive Parallel Processing, Business Intelligence

I. INTRODUCTION

Mobiles are the masters of man update. Everyone in this generation are looking at their mobile for something or the other. People are online, browsing information, shopping goods and services and perform many other tasks like these. Not only customers even the business men are profited by the means of online transaction. In this type of fast moving mobile economy, it is important for the business men to retain their customers and run business in an efficient way. To do this big data helps them in all means. The concept of digitalized data along with big data concept helps the user to enhance the means of his life. The data is woven into sectors and functions. Other essential factors of production (i.e., hard assets and human capita) much of modern economic activity could not take placed with the woven data sectors. The use of big data which are viewed as large pool of data are bought together and analyzed to make better decisions. This forms as the basis for the growth of industrial firm. The data sets that are too large in size or complex which cannot be processed by any traditional data processing application software is known as big data. Big data faces n-number of encounters include seizing data, data storage, visualization, querying, data analysis, search, sharing, transfer, updating and information privacy.

II. WHAT IS BIG DATA?

Big data refers to predicative analytics or any other advanced data analytics method that generally extract value from data and seldom to a particular size of data set. Big data now has bought enhancement in the fields of science, business, medicine, advertisements and government tasks. Before big data concept all these fields found it difficult to handle large data sets in areas such as internet, search, fintech, urban, informatics and business informatics. Always scientists found it difficult to move with e-science as there were huge numbers of limitations. Now data sets are growing rapidly in part because they are increasing gathered by IOT (Internet of Things) devices. Thus big data are large parts of data that are used for analysis and processed for better decisions.

III. DEFINITION

The term was coined and popularized by John Mashey around 1990's. Generally big data is defined as data sets with a size which are beyond the ability of common software tool to manage and process within the elapsed time. The big data consists of instructed, semi structured and structured data when the main focus is on unstructured data. The size of big data has moved from few terabytes to many pent byte of data during the year 2012. Thus big data requires returned and fine techniques and technologies to reveal insights from datasets that are diverse, complex and of a massive scale.

Definition according to Gartner: it is high volume, high velocity, and high variety information assets that demand cost-effective, innovative from of information processing for enhanced insight and decision making" [1].

A. Explanation over Gartner's Definition:

Now the definition is to be split into segments for explanation. There are typically three constraints to describe big data.

1) "High Volume, High Velocity & High Variety:

These are known as the 3V's of big data

1) Volume: Generally we create massive amount of data every day. The amount of data created in past two years is estimated to be 85% of the entire set of data available. And it is expected to reach 40,000 Exabyte around 2020. This indicates the volume of big data.
2) Variety: data is collected from various disparate sources. Almost 80% of these data are constructed which do not suit for today's corporate database. This shows the variety of data in big data
3) Velocity: all this data are generated and collected in fraction of second. It works with the speed of lightning. This indicated the speed at which the big data is being responding to data.

2) "Demand Cost Effective, Innovative Forms of Information Processing":

Any of the traditional computing environments, database or tools is not capable of handling big data sets. Cost effective cloud storage and search tool is the only way to organize and build big data solutions. For this we need the help of advanced technologies such as Hadoop or NO SQL etc.

3) "For Enhanced Insight & Decision Making':

It represents the value of big data. This enables us to become more innovative and competitive by gaining insights into our businesses and find relationships between data sets not available previously.

B. Other Definitions:

The growing maturity of the concept more starkly delineates the difference between big data and business intelligence.

1) Business Intelligence:

It uses descriptive statistics. It implements this statistics to data which are highly informative to measure things, detect trends etc

2) Inductive statistics:

Big data makes use of inductive statistics and nonlinear system identification. This is implemented to infer laws from large set of data which has low information density.

IV. HISTORY

Big data has a very long old and brief history. Earlier computers were common place. It was very easy to forget. Gradual evaluation began on which the ability to store and analyze information has been increased. It was further speeded up with invention of digital storage and the internet. With the invention of big data, the efficiency was much more enhanced.

A. Ancient History of Big Data:

1) Year 1999:

The term big data appeared in "Visually Exploring Gigabyte Data sets in Real time" which was published by computation Machinery. The concept of "Middle man" which deals with the communication of devices online with each other with the help of human was first described in the term IoT (Internet of Things). IoT was first used by big data.

2) Year 2000:

Peter Lyman and Hal Varian tried to quantify and rate the growth of digital data in the world. They came up with a conclusion that it would require at least 1.5 billion gigabyte to store all forms of data in the world

3) Year 2001:

The definition of Gartner came into existence. It was accepted as characteristics of big data. The term "Software as a service" came into existence. This concept aimed on providing basis for cloud based applications. It was released in the article "Strategic Backgrounder: Software as a Service" by software and information Industry.

4) Year 2005:

Commentators encountered the birth of "Web 2.0". This consists of contents created by the users rather than the service providers. It was built with the help of traditional HTML style web page with the backend of SQL 5. This year also created a system called Hadoop. It is an open source framework which was created to store and analyze big data sets. It was helpful in managing unstructured data.

5) Year 2007:

Wired, bought the concept of big data in the article "the end of Theory: The data Deluge makes the scientific model obsolete.

6) Year 2008:

The worlds serve process 9.57 zettabytes of information which was equivalent to 12 gigabyte of information. It was estimated that 14.7 Exabyte of new information was produced during that year in the article "International Production and bissemination of information".

7) Year 2009:
It was found that an average US company with about 1000 employees stored 200 terabytes of data. The big data came into existence here.
8) Year 2010:
Eric Schmidt estimated that almost 80% of the data is created shortly of the whole data from human evolution
9) Year 2011:
According to Mc Kinsey Prediction it is expected that US will face a short fall for professional DATA SCIENTISTS ranging between 1,40,000 TO 1,90,000 IN 2020 TO resolve the issue of privacy, security and intellectual property that world occur in Big Data. It must be resolved before the value of big data is released.
10) Year 2014:
Most of the people started using mobile devices to access digital data other than the office or home computers. Almost 88% of the business executes surveyed with Ge working with Accenture report that big data analytics is a top priority for their business.
11) Final Thought:
Big data is not a new or an isolated phenomenon. It is a part of long evolution of capturing and using data. Big data provides a better way in running business and society by efficiently storing the data. It also lays a foundation for many other evolutions to build.

V. CHARACTERISTICS

Big data has the following characteristics:

A. Volume:

The amount of data generated and stored data determines the value and potential understanding and confirms whether to consider it as big data or not.

B. Variety:

The type and nature of data will help to analyze if effectively and use the resulting insight.

C. Velocity:

The rate at which the data is produced and processed to meet the difficulties as well as the challenges that lies on the path is considered.

D. Variability:

Inconsistency of the data set can hamper process to handle and manage it.

E. Veracity:

The quality of the captured data varies greatly the affects the accurate analysis.

F. Factory Work & Cyber:

Physical system may have 6c system:
- Connection
- Cloud
- Cyber
- Content/contest
- Community
- Customization

To reveal meaningful information, a data must be processed with advanced tool. Also the information generation algorithms must be able to detect issues such as machine degradation, component wear etc.

VI. ARCHITECTURE

There are many forms of existence of big data repositories update. All these are built with specific need by the corporations. Commercial vendors offered parallel database management system for big data in 1990's.

DBC 1012 system was a parallel processing system that was marketed by Teradata Corporation in 1984. In 1992 Teradata Corporation was the first to store 1 terabyte of data. The definition of big data continuously evolved according to Kryder's Law in 1991 when the hard disks were only 25 GB pent byte class RDBMS system was installed by Teradata in the year 2007. There is few dozen of RDB installed update. The largest one exceeds 50PB. Till 2008 there existed only structured relational data.

In 2000 C++ based distributed file sharing frame work for data storage query was developed. This stored structured, semi structured and unstructured data from many servers. Generally in this system user built queries in C++ dialect which is known as

ECL. To infer the structure of stored data, ECL uses "apply schema on read" method. Seisint Inc and choice point Inc were acquired in 2004 and 2008 respectively. Both the systems were merged together in HPCC (High Performance Computing Cluster) system in 2011. HPCS was an open source under Apache V2.0 license.

A paper on "MapReduce" was published by Google which also had the same architecture. This model provided parallel processing and associated implementation was related to process huge amount of data. With the help of this system the queries are distributed to different nodes and processed parellely. The result is then gathered together and then is displayed. This project caves successfully. Thus Apache open source project replicated the algorithm and named it Hadoop.

Mike 2.0 is an open approach for information management. This methodology addresses handling big data in terms of permutation of data sources, complexity in interrelationships and difficulty in deleting individual records.

To resolve the issues in big data, one of the efficient means was established in 2012 which was the multilayer architecture. This system distributes the data among the servers and is parallel executed. This eventually increases the data processing speed. This architecture inspects data into DBMS which implements the usage of MapReduce framework. This framework tries to process the data in front end with the help of application server.

Big data analytics for manufacturing application involves 5C architecture. They are:-

- Connection
- Conversion
- Cyber
- Cognition
- Configuration

Data always ties to shift organizational control to centralized control in order to share models and respond to the changing dynamics of information management. This reduces the overhead time by quick segregation of data into the data lake.

VII. TECHNOLOGIES

The main characteristics, main components and ecosystem of big data was established by McKinsey Global Institute report in 2011. They are as follows:

Techniques for analyzing data such as MB testing, Machine Language and netural language processing.

Big data technologies, like business intelligence, cloud computing and databases.

Visualization such as charts, graphs and other displays of the data are available now.

Tenson based computing is effectively handled by multidimensional big data that is represented as tensous. Any kind of additional technologies when added to big data enhances to MPP (Massive Parallel Processing) databases. Even though there are many advancement made it is still difficult for big data to move to the field of machine

A. Learning:

The practitioners of big data analytics processes are generally hostile to slower shared storage, preferring direct attached storage (DAS) in its various forms from solid state Drive (SSD) to high capacity SATA disk buried inside parallel processing modes. The perception of shared storage architectures storage area network (SAN) and Network Attached Storage (NAS) is that they are relatively slow, complex and expensive. These qualities are not suitable for big data system that emphasizes on performance, commodity infrastructure and low cost.

One of the defining characteristics of big data is their real or near real time information delivery. Thus this helps in avoiding the latency whenever and where ever possible. Always the data at the memory is good but the data at the spinning dist at the end of FCSAN is comparatively not. The SAN is very costly when compared to other storage techniques.

There are both advantage and disadvantage in sharing the sorted data in big data analytics, but as per 2011 practitioners it does not favors it.

VIII. APPLICATIONS

Big data seeks it importance in almost all the fields of science and technology. It has a great demand of information management specialists who have spent $15 billion on software from specializing in data management and analytics. It has an increase percent of almost 10 percent a year which is above twice as fast as any other software business as a whole.

A. Government:

Big data is adapted in government sectors that slow efficiencies in terms of cost, productivity and innovation.

B. International Development:

The development in the field of information and technology leads to the development of big data which eventually leads to the international development. The improvement in big data analysis provides cost effective decision making in crucial departments such as health case, security and natural disaster etc.

C. Manufacturing:

Big data provides an infrastructure for transparency in manufacturing industry. Generally a conceptual framework for predictive manufacturing starts with data acquisition with the help of sensory data. Vast amount of physical data is combined with old data to form big data. The generated big data acts as the input into predictive tools.

D. Health care:

Big data is helpful in the field of health care by providing personalized medicines and prescriptive analytics etc.

E. Education:

The concept of big data enhances the use of browsing in the field of education. Almost most of the universities are benefited with bid data in their genera

F. Media:

Media uses big data in a more effective and efficient way. People are benefited through media which seeks support from big data.

- Targeting of consumers
- Data capture
- Data Journalism
- These are the few things the, media concentrates and for which big data supports it.

G. IT (Information Technology):

Effective measures and steps have been taken in the field of IT since 2015. Big data to use in IT to resolve the challenges faced by IT sectors in all means big data a part of IT helps to efficiently achieve the goals of IT in full-fledged manner

H. Science:

It is estimated that more than 150 sensors that are delivering data which is above 40 million times per second. There exist almost 600 million collisions per second in this universe. After the process of filtering and refraining from records, about 99.9% of these streams still have 100 collisions per second. All these are estimated, analyzed and judged for decision with the help of big data.

I. Sports:

Using sports sensor one can improve training and understand competitions. With the help of big data one can predict the winner in the match. Thus big data being revolution in sports field too

Almost all the fields are covered by big data. Thus in short we can say that big data plays role in all most all means of human life.

IX. CONCLUSION

This paper is intended with big data tools, techniques, issues affiliated with big data. It also engrossed and supplied the information about how to perform big data visualization. Research tenors in big data, maneuvers of big data such as storage, search and retrieval, big data analytics and computations on big data are mentioned. Big data analytics targets on tools, algorithm, and architecture which perform proper analysis and transfer large and massive volume of data. Computing covenants with processing, transforming, handling and information storage. This paper has discussed on the basic concepts of big data and its applications.

REFERENCES

[1] Neelam Singh, Neha Garg, Varsha Mittal, Data – insights, motivation and challenges, Volume 4, Issue 12, December-2013, 2172, ISSN 2229-5518 2013.
[2] Karthik Kambatlaa, Giorgos Kollias b, Vipin Kumarc, Ananth Gramaa, Trends in big data Analytics, (2014) 74 2561–2573
[3] Francis X. "On the Origin(s) and Development of the Term \Big Data"_ Francis X., 2012
[4] Venkata narasimha inukollu1, sailaja arsi1 and srinivasa rao ravuri3 Security issues associated with big data in cloud computing Vol.6, No.3, May 2014
[5] Matzat1, Ulf-Dietrich Reips2,3 1 Eindhoven "Big Data" 2012, 7 (1), 1–5 ISSN 1662-5544
[6] Hong Kong, Park Shatin, Mining Big Data: Current Status, and Forecast to the Future
[7] Anil K. Jain Clustering Big Data, 2012
[8] Daniel Keim Big-Data Visualization.
[9] Hsinchun Chen Business Intelligence and Analytics: From Big Data to Big Impact AZ 85721, OH 45221-0211 U.S.A. Mack Robinson, GA 30302-4015.
[10] Ibrahim Abaker Targio Hashema,n, Ibrar Yaqooba, Nor Badrul Anuara, Salimah Mokhtara, Abdullah Gania, Samee Ullah Khanb, The rise of "big data" on cloud computing: Review and open research issues. 2014
[11] Edd Dumbill, Making Sense of Big Data
[12] Silva Robak , prof. Z. Szafrana, Zielona Góra Uniwersytet Zielonogórski Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management 2014 249 DOI: 10.15439/2014F472
[13] C.L. Philip Chen , Chun-Yang Zhang Data-intensive applications, challenges, techniques and technologies: A survey on Big Data 275 (2014) 314–347
[14] Chaitanya Baru,1 Milind Bhandarkar,2Raghunath Nambiar,3 Meikel Poess,4and Tilmann Rabl Survey of Recent Research Progress and Issues in Big Data 2013.
[15] Tackling the Challenges of Big Data 2014.

[16] Stephen Kaisleri_SW. Alberto Espinosa Big Data: Issues and Challenges Moving Forward Stephen Kaisleri_SW. Alberto Espinosa 013 46th Hawaii International Conference on System Sciences
[17] Challenges and Opportunities with Big Data A community white paper developed by leading researchers across the United States 1819
[18] Danyang Dua , Aihua Lia*, Survey on the Applications of Big Data in Chinese Real Estate Enterprise 1st International Conference on Data Science,2014