International Journal of Computer Applications Technology and Research
Volume 9–Issue 03, 98-101, 2020, ISSN : -2319-8656

A Study on Big Data Analytics: Technologies & Tools.

Mr.Chandrakant R. Gujar
Assistant Professor, MCA
W.K.B.S. Mandal's
Dr. Suryakanta R. Ajmera
MCA College for
Women,Deopur Dhule, MS
India

Abstract: Data is a very valuable asset in the world today. The economics of data are based on the idea that data value can be extracted through the use of analytics. Through Big data and analytics are still in their initial growth stage, their importance cannot be undervalued. As big data starts to expand and grow, Importance of big data analytics will continue to grow in everyday lives, both personal and business. In addition, the size and volume of data is increasing every single day, making it important to address the manner in which big data is addressed every day. A huge repository of terabytes of data is generated every day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data technologies and challenges associated with it. This paper provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.

Keywords: Big Data, Big Data technologies, Big Data Challenges, Hadoop, Big Data analytic.

Introduction

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.[1]

What is Big Data?

Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.[1]

Examples Of Big Data

Following are some the examples [1] of Big Data-

1) The New York Stock Exchange generates about *one terabyte* of new trade data per day.
2) Social Media: -The statistic shows that *500+terabytes* of new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
3) A single Jet engine can generate *10+terabytes* of data in *30 minutes* of flight time. With many thousand flights per day, generation of data reaches up to many *Petabytes.*

What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data [2].

- **Black Box Data** − It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** − Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** − The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data** − The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** − Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data** − Search engines retrieve lots of data from different databases.

Types of Big Data

- **Structured data** − Relational data.
- **Semi Structured data** − XML data.
- **Unstructured data** − Word, PDF, Text, Media Logs.

Benefits of Big Data

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

Big Data Technologies

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology −

a) Operational Big Data

This includes systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.

NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

Some **NoSQL** systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

b) Analytical Big Data

These includes systems like Massively Parallel Processing (MPP) database systems and **MapReduce** that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.

MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on Map Reduce that can be scaled up from single servers to thousands of high and low end machines.

Big Data Challenges

The major challenges associated with big data are Capturing data, Curation, Storage, Searching ,Sharing, Transfer, Analysis & Presentation, To fulfill the above challenges, organizations normally take the help of enterprise servers.

Traditional Approach

In this approach, an enterprise will have a computer to store and process big data. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc. In this approach, the user interacts with the application, which in turn handles the part of data storage and analysis.

Google's Solution

Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset

Hadoop

Apache Hadoop is an open source software framework used to develop data processing applications which are executed in a distributed computing environment.

Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers. Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at low cost.

Similar to data residing in a local file system of a personal computer system, in Hadoop, data resides in a distributed file system which is called as a Hadoop Distributed File system. The processing model is based on 'Data Locality' concept wherein computational logic is sent to cluster nodes (server) containing data. This computational logic is nothing, but a compiled version of a program written in a high-level language such as Java. Such a program, processes data stored in Hadoop HDFS [3]

What is Big Data Analytics?

Big data analytics refers to the strategy of analyzing large volumes of data, or big data. This big data is gathered from a wide variety of sources, including social networks, videos, digital images, sensors, and sales transaction records. The aim in analyzing all this data is to uncover patterns and connections that might otherwise be invisible, and that might provide valuable insights about the users who created it. Through this insight, businesses may be able to gain an edge over their rivals and make superior business decisions.

Big Data Analytics Tools

Big Data Analytics software is widely used in providing meaningful analysis of a large set of data. This software helps in finding current market trends, customer preferences, and other information.[3], There are following latest tools available in market year 2020

Azure HDInsight is a Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads.

Skytree is a big data analytics tool that empowers data scientists to build more accurate models faster. It offers accurate predictive machine learning models that are easy to use.

Talend is a big data tool that simplifies and automates big data integration. Its graphical wizard generates native code. It also allows big data integration, master data management and checks data quality.

Splice Machine is a big data analytic tool. Their architecture is portable across public clouds such as AWS, Azure, and Google.

Apache Spark is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

Plotly is an analytics tool that lets users create charts and dashboards to share online

Apache SAMOA is a big data analytics tool. It enables development of new ML algorithms. It provides a collection of distributed algorithms for common data mining and machine learning tasks

Lumify is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

Elasticsearch is a JSON-based Big data search and analytics engine. It is a distributed, RESTful search and analytics engine for solving numbers of use cases. It offers horizontal scalability, maximum reliability, and easy management.

R is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests

ACKNOWLEDGMENTS

We are thankful to our President West Khandesh Bhagini Seva Mandal’s Mr. Atulji R. Ajmera , Mrs.MeghaV.Kamerkar OS ,Director Dr. Rajeev B.Kharat, and all teaching staff /office staff of Dr. Suryakanta R. Ajmera MCA College for Women Deopur Dhule MS, India for motivation and encouragement which helped us to complete this research paper.

REFERENCES

[1] https://www.guru99.com/what-is-big-data.html#1

[2] https://www.tutorialspoint.com

[3] https://www.guru99.com/big-data-analytics-tools.html