International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT12172120

An Extensive Analysis on Big Data

M. Lakshmi Priya
Department of Computer Science, Rabiammal Ahamed Maideen College for Women, Thiruvarur, Tamilnadu, India

Article Info
Volume 7, Issue 2
Page Number : 591-595
Publication Issue :
March-April-2021

Article History
Accepted : 25 April 2021
Published : 30 April 2021

ABSTRACT

In recent years, data has rapidly developed because of the growth of the internet, the internet of things, cloud computing, and various technologies. The size of data processed and transmitted over the internet is drastically increasing. Big data refers to a database that handles huge data in real-time yet growing exponentially with time. Big data analytics uses advanced techniques on large heterogeneous datasets that are collected from different sources, and in various sizes. Big data can manage and process the data beyond the ability of a relational database.

Keywords : Data, Big Data, Analytics, Tools

I. INTRODUCTION

Assume a world without data storage, every detail about a person or organization, every action performed, or every way that can be documented is lost directly after use. It loses the ability to extract needed information, perform any analysis. Data is the unprocessed facts, after processing the facts, we get the information. Common types of data include video, audio, text, and images.

Big data refers to a database that handles huge data in real-time yet growing exponentially with time. Big data analytics uses advanced techniques on large heterogeneous datasets that are collected from different sources and in various sizes. Big data can manage and process the data beyond the ability of a relational database. Usually, we work on data of size megabytes for word documents and Excel or maximum gigabytes for Movies, but for big data the size of the data is zettabytes. These data arrived from social media sites that generate massive data on a daily basis. E-commerce Sites like Amazon generate numerous logs from which users buying things. Stock exchange in share markets across the world generates a huge amount of data through its daily transaction.

Big data not only deals with structured data but also unstructured and semi-structured data. Big data not only deals with structured data but also unstructured and semi-structured data from various sources and also in different sizes from terabytes to zettabytes Big data integrate data from distinct sources and applications, then manages the storage and examines the data set to make new discoveries. When integrating, processing and analyzing big data, it is categorized as either operational or analytical data and stored accordingly. Operational data is the data that is produced by day-to-day operations whereas analytical data is for planning functions and making decisions

II. TYPES OF BIG DATA

Big data is categorized in three ways:

A. Structured Data

Any data that can be stored and processed in fixed format is termed as structured data. Structured data are stored in a relational database that is organized in rows and columns.

B. Unstructured Data

Audios, videos, log files, images are included in unstructured data. It refers to unorganized data. It is difficult to derive the value from unstructured data.

C. Semi-Structured

Semi-Structured data can contain both forms of structured and unstructured data. The data is stored in a relational database but the schema is not well-defined.

Fig.1. Types of big data

III. CHARACTERISTICS

Big data can be represented by the following characteristics.

A. Volume

Volume relates to size. it generates volumes of data from different sources every second. Millions of new posts in social media are uploaded every day in various formats like images, audios, videos that are stored in petabytes and zettabytes. Volume is an important feature that needs to be observed when dealing with big data.

B. Variety

A variety of big data implies the data collected from heterogeneous sources that can be semi-structured, unstructured, and structured data. Most of the data is unstructured. In the early stages, the data were collected in the form of spreadsheets and databases but nowadays, data is in the form of images, audio, video, and text. Unstructured data create some issues for storage, mining, and analyzing data.

C. Velocity

Velocity refers to the speed of generating the data. For every 60 seconds, a millions of photos are uploaded on Face book, millions of tweets are posted on Twitter, a million hours of videos are uploaded on YouTube and a billion searches are performed on Google. It measures how fast the data is generated and it should be continuous.

D. Veracity

Veracity deals with the quality, accuracy, and trustworthiness of data. Most of the data is unstructured and irrelevant. If source data is incorrect, then analyses will be ineffective. It helps to keep your data clean without duplication, inconsistency and volatility.

E. Value

It defines the worthiness of data. Storing a large amount of irrelevant data has no value. If a data has no value, we cannot extract any useful information from the dataset.

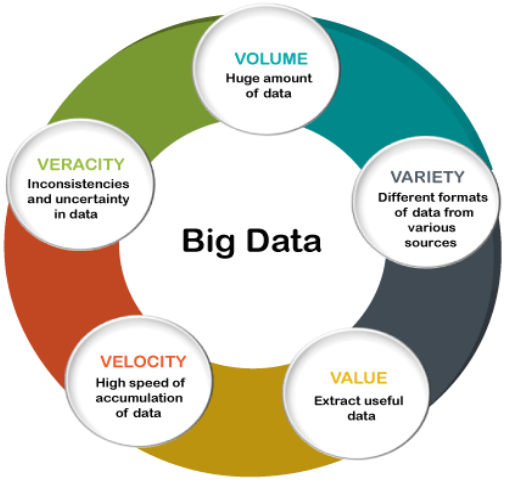

Fig.2 Characteristics

IV. CLASSIFICATION OF BIG DATA ANALYTICS

Big data analytics is the process used to extract meaningful information. It works on massive data to uncover patterns and relationships.

A. Descriptive Analytics

It summarizes the unprocessed data and converts it into a human-understandable form. It describes the detail of an event that has happened in the past. It answers the question "what happened?" It deals with the discovery of the reasons behind the positive and negative results in the past event.

B. Diagnostic Analytics

It searches the descriptive analytics in depth. It determines "why happened?" It takes a deeper look at data to find the cause of the events. It uses the techniques drill down to take from a general to a specific view of data.

C. Predictive Analytics

Predictive Analytics is used to forecast the events that happen in the future. It helps to predict the trends based on the events. It answers the question "what is going to happen?" It predicts the probability of an event happening in the future or estimates the time.

D. Prescriptive Analytics

It allows users to define possible solutions. It answers the question "what is the best course of action?" It works with both descriptive and predictive analytics.

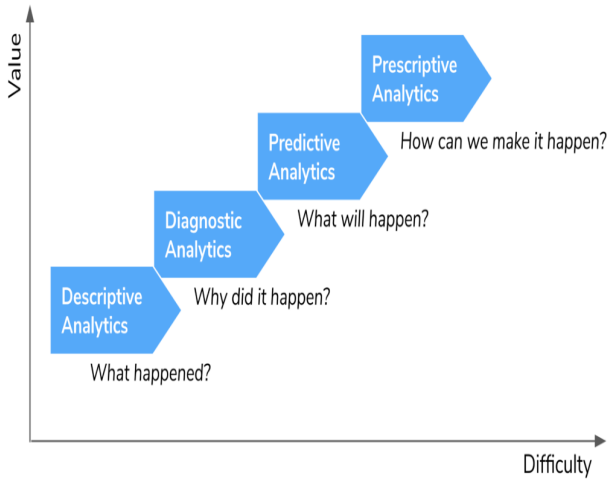

Fig.3 Classifications of analytics

V. CHALLENGES OF BIG DATA

In the digitalized world, we are producing a massive amount of data every second. It is challenging to store, process, and analyse the data.

A. Handling Huge Data

Data is growing with every passing day. It specifies that organizations need to deal with large amounts of data on a daily basis. It is hardly amazing to get voluminous data into the big data.

B. Collecting Real-Time Data

Data keeps updating every second, it is important to keep ourselves updated with this data. This will helps to get accurate data and enhance a deep understanding of data.

C. Security

Big data collect structured and unstructured data from multiple sources. Maintaining security, integrity, and privacy is the major challenge in big data.

D. Shortage of skilled people

There is severe inadequacy of professionals who understand big data. With the exponential growth of data, a huge demand of big data scientists and analysts. It is necessary to hive data scientists having skills on big data.

E. Synchronization

The data collected from heterogeneous sources that differ in various formats and sizes. There is a big challenge to combine them into a single dataset. If this is overlooked, it leads to incorrect message and create gaps.

VI. TOOLS

Nowadays, there is a number of big data tools that are available. Hadoop, Map Reduce, Hive, spark are some of them are commonly used.

A. Hadoop

Hadoop offers storage for all kinds of data. there is a storage system is known as the Hadoop distributed file system which helps to divide a large amount of data and replicate it to nodes in the cluster. Name node in HDFS acts as a master node that contains all information about the data node. Data nodes act as slaves used to store data.

B. Map Reduce

Map reduce is a software framework is used for processing a large amount of data. It consists of two functions: Map and Reduce. Map function deals with splitting the input data and distribute it to the slave. Reduce function aggregates the data from the shuffling phase then produces a single output.

C. Spark

Spark is the open-source big data analytical tool by apache. Spark is a general compute engine for large-scale data processing. It supports real-time processing which involves continuous input and output of data.

D. MongoDB

MongoDB is a database that is written in C++ and belongs to NoSQL. It is used for storing and retrieving information.

E. Cassandra

Cassandra is a distributed database management system by Apache that handles a large amount of data from the various data centers. Cassandra query language is used to access the database.

F. Hive

Hive is the open-source data warehourse software for managing large data sets, It is used to analyze only the structured data. It can analyse large data set stored in hadoop distributed file system or in other storage systems.

VII. CONCLUSION

This paper presents the essential concepts of big data. A large volume of data is collected in different formats such as audio, video, text, images, etc. It is hard to handle the massive amount of data. It is the process of converting a huge amount of unstructured raw data from heterogeneous sources to a data product. It is the complex process of examining large and varied data sets. It deals with the set of data to find patterns, and extract information. Representing

the unstructured data in a visual format can be a challenging task. Various types of graphs and tables can be used to represent the data. By applying big data, valuable information can be extracted to enhance decision-making. Data warehouses are used to gather and store a large amount of unstructured data. The problem arises when we combine unstructured data and inconsistent data from various sources, it leads to error, missing, and duplicate data results in data quality.

VIII. REFERENCES

[1]. Chahal, Dr & Gulia, Preeti. (2016). Big Data Analytics, Research Journal of Computer and Information Technology Sciences E-ISSN 2320 – 6527 Vol. 4(2), 1- 4, February (2016)

[2]. Zakir, Jasmine, Big Data Analytics Issues in Information Systems Volume 16, Issue II, pp. 81-90, 2015.

[3]. R.V. Gandhi, Ch. Rathan Kumar, P. Vamshi Krishna, Big Data: Issues And Challenges, I Journals: International Journal of Software & Hardware Research in Engineering ISSN-2347-4890 Volume 5 Issue 7 July, 2017

[4]. Acharjya, D. P., & Ahmed, K. (2016). A survey on big data analytics: challenges, open research issues and tools. International Journal of Advanced Computer Science and Applications, 7(2), 511-518.

[5]. Mukherjee, S., & Shaw, R. (2016). Big data–concepts, applications, challenges and future scope. International Journal of Advanced Research in Computer and Communication Engineering, 5(2), 66-74.

[6]. B.Thillaieswari., “Comparative Study on Tools and Techniques of Big Data Analysis” International Journal of Advanced Networking & Applications (IJANA) Volume: 08, Issue: 05 Pages: 61-66 (2017) Special Issue

Cite this article as :

M. Lakshmi Priya, "An Extensive Analysis on Big Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 591-595, March-April 2021. Available at

doi : https://doi.org/10.32628/CSEIT12172120

Journal URL : https://ijsrcseit.com/CSEIT12172120