A Detailed Study of Big Data

-Ankit Gautam

Abstract

Billions of people spend their time daily on social media exchanging messages, posting comments, uploading videos, and doing other activities. These all things lead to the production of large amounts of data. The size of the data will be in petabytes within an hour and this data is known as big data. Big Data is simply data, but huge in volume, yet growing exponentially with time. Only social media isn't the source of big data, stock exchange, telecom companies and jet engines are also sources of it. The main purpose of this article was to explore the types and applications of big data in real life. This article's interpretation and conclusion give a brief explanation of its types, 3Vs, processing software, and applications.

1.Introduction

Larger, more complicated data sets that can be structured, unstructured, or partially structured are referred to as big data. In another term, big data are data that contain greater variety, arriving in increasing volumes and with more velocity. This is also known as 3Vs.

a. Volume

Volume refers to the amount of data that exists. It is like the base of big data, as it is the initial size and the amount of data that is collected. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.

b. Velocity

Velocity refers to how quickly data are generated and how quick data move.The data are received at an unprecedented speed and are acted upon in a timely manner. It also requires real-time evaluation and action in the case of Internet of Things applications.

c. Variety

Variety refers to different formats of data. It may be structured, unstructured or semi-structured. The data can be audio, video, text or email. In this additional processing is required to derive the meaning of data and also to support the metadata.

These data sets are so voluminous that traditional data processing software can't store and process them efficiently. So, new technologies like Hadoop are used to process them. The examples of big data are New York Stock Exchange, Social Media, Jet engines and many more. The New York Stock Exchange generates about one terabyte of new trade data every day and a Jet engine generates around 10+ terabytes of data in 30 minutes flight time.

2.Types of Big Data

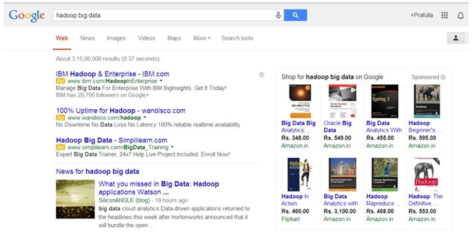
2.1. Structured Big Data

Any big data which can be accessed, processed and manipulated in the fixed form is called structured big data. It has predefined nature due to which each field is discrete and can be accessed separately or jointly along with data from other fields. This makes structured data extremely valuable, making it possible to collect data from various locations in the database quickly. The data related to RDBMS is its type.

Employee_ID	Employee_Name	Gender
2365	Rajesh Kulkarni	Male
3398	Pratibha Joshi	Female
7465	Shushil Roy	Male
7500	Shubhojit Das	Male
7699	Priya Sane	Female

2.2. Unstructured Big Data

Any big data which have no fixed form to access, process and manipulate is called unstructured big data. This data can't be easily interpreted or analyzed by the standard database or data model. Unstructured data accounts for the majority of big data and comprises information such as dates, numbers, and facts. It is a heterogeneous combination of data. Video and audio files, mobile activity, satellite imagery, and No-SQL databases are its examples. Photos we upload on Facebook or Instagram and videos that we watch on YouTube or any other platform contribute to the growing pile of unstructured data.



2.3. Semi Structured Big Data

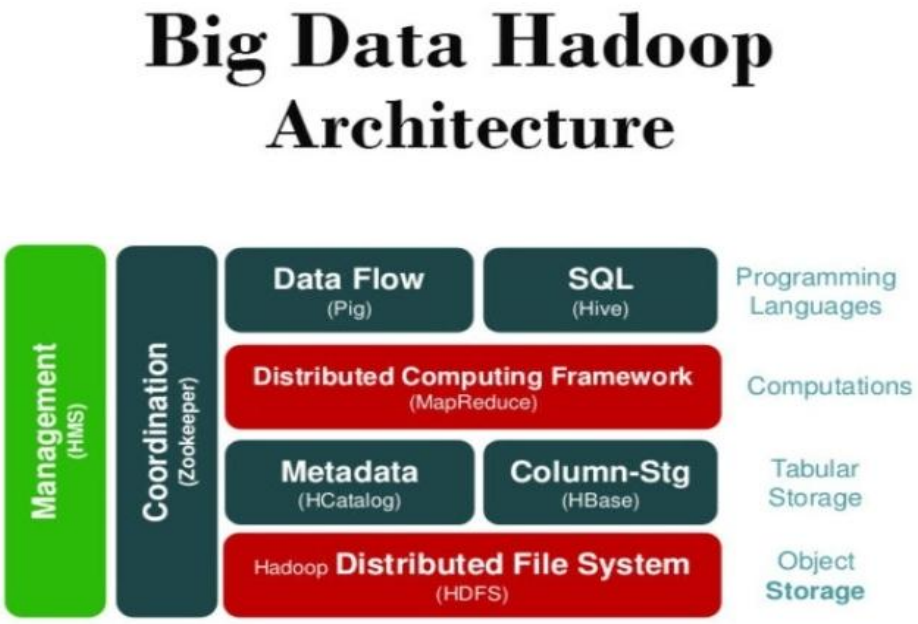
From the name only we can understand that semi structured big data is data which has partial properties of both structured and unstructured data. This means it inherits the few characteristics of structured data, but nonetheless contains information that fails to have a definite structure and does not conform with relational databases or formal structures of data models. Data represented in an XML file is an example of semi structured big data.

```
<rec><name>Prashant Rao</name>
<sex>Male</sex><age>35</age>
</rec>
<rec><name>Seema R.</name>
<sex>Female</sex><age>41</age>
</rec>
<rec><name>Satish Mane</name>
<sex>Male</sex><age>29</age>
</rec>
<rec><name>Subrato Roy</name>
<sex>Male</sex><age>26</age>
</rec>
<rec><name>Jeremiah J.</name>
<sex>Male</sex><age>35</age>
</rec>
```

3. Big Data Hadoop

As mentioned above, big data can't be processed by the traditional database software so new technologies are used and one of them is Hadoop. Hadoop is an open source framework that is used to efficiently store and process large datasets. Hadoop provides storage and processing power for big data. Instead of using a single computer, Hadoop allows clustering multiple computers to analyze massive datasets in parallel.

HDFS, MapReduce and YARN are the components of Hadoop architecture. HDFS (Hadoop Distributed File Systems) is used to store large amounts of data and mainly designed for working on commodity hardware devices. Data in HDFS is always stored in terms of blocks. So, the single block of data is divided into multiple blocks of size 128MB which is default and you can also change it manually. MapReduce is an algorithm which is used to process the data in a distributed manner across multiple machines. YARN (Yet Another Resource Negotiator) is used for data processing resources like CPU, RAM, and memory. Resource Manager and Node Manager are the elements of YARN. These two elements work as master and slave.



Hadoop has a great importance in today's world. It can easily store huge chunks of data from social media and IoT (Internet of Things) and process them fast. It also provides security from malware and hardware failure. You can store data according to your needs and the system can be grown easily just by adding nodes. Everything in the world has its own bad side so there are few demerits of Hadoop. It is not suitable for iteration and interaction tasks. It is suitable for problems which can be divided into units. Tools for data quality and standardization are missing.

4. Application of Big Data

4.1 Education

Big data are used by educational institutions to understand the need and interest of the students and build unique, customized curriculums for students. This process leads to higher achievements as students have access to more information in a format that is designed for their requirements. Data can also be used to monitor the students' problem, their weak and strong areas of academics, achievements to guide and help them to improve. Big data not only helps students it also helps teachers. Big data help educators evaluate their own course content. It provides unbiased feedback on the structure and design of their course and also helps them understand how efficient their teaching methods are. For examples: The university of Tasmania has deployed a Learning and Management System that tracks when a student logs onto the system,

how much time is spent on different pages and overall progress of the students. The Office of Educational Technology in the U. S. Department of Education is using Big Data to develop analytics to help correct course students who are going astray while using online Big Data certification courses. Click patterns are also being used to detect boredom.

4.2 Health

Big data are used in the healthcare sectors for clinical data analysis, pattern analysis, medicine supply, drug discovery and various other such analytics. These analytics have supported in a major way to curb the cost of rising healthcare and also improved the health care system. For example: Beth Israel Hospital has collected data from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital. Free public health data and Google Maps have been used by the University of Florida to create visual data that allow for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

4.3 Government

Big data allows the government to access large amounts of data that are essential for day to day operation. With real time access, the government can identify areas that require attention, make better and more timely judgements about how to proceed, and make necessary changes. It is used to improve transparency and efficiency in public management, and fraud detection. The Food and Drug Administration (FDA) is using Big data to detect and study patterns of food-related illnesses. This allows for a faster response, which has led to more rapid treatment and less death. Big data are being used in the analysis of large amounts of social disability claims made to the Social Security Administration (SSA) that arrive in the form of unstructured data. The analytics are used to process them and detect the suspicious claim.

4.4 Banking and Security

The Banking Industry uses big data to track the activities of customers for personalized product offering. They also warn the customers if they find any suspicious activities with their bank accounts. They also depend on big data for risk analytics, including: anti-money laundering, fraud mitigation, high speed trading and many more. The Securities Exchange Commission (SEC) is using Big Data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity. Big Data providers specific to this industry include 1010 data, Panopticon Software, Streambase Systems, and Nice Actimize.

Conclusion

Big data is a huge chunk of structured, unstructured and semi structured data whose size is in petabytes and contains useful information. Utilization of all three types of big data with the help of data analytics leads to the success of the company in any sector as it can offer better experience to customers. A slight change in the efficiency can lead to a huge profit. This is why most organizations are moving towards big data. Big data is a game changer.

Reference

https://www.guru99.com/what-is-big-data.html
https://www.oracle.com/in/big-data/what-is-big-data/
https://www.simplilearn.com/tutorials/big-data-tutorial/big-data-applications
https://techsparks.co.in/thesis-topics-in-big-data-and-hadoop/