1. INTRODUCTION TO BIG DATA

Big data is a collective term for the non-traditional strategies and technologies required to collect, organize, process, and extract insights from large amounts of data. The problem of working with data that exceeds the processing power and storage capacity of a single computer is nothing new, but the prevalence, scope and value of this type of computing have increased dramatically in recent years.

1.1 Getting started with Data and Big Data

Data: The amounts, characters, or symbols that a computer performs operations on; these can be recorded on magnetic, optical, or mechanical recording media and saved and transferred as electrical signals.

Big Data: Big Data is an accumulation that contains data that is enormous in volume and is always expanding exponentially. No usual data management systems can effectively store or process this data because of its magnitude and complexity. Big data is a type of data that is very large in size.

Example of Big Data:

Stock Exchange

With a daily production of one terabyte of fresh trading data, the New York Stock Exchange is an example of big data.

Social Media

According to the estimate, Facebook's databases get more than 500 terabytes of new data each day. This information is primarily produced by the uploading of images and videos, messaging, leaving comments, etc.

Here are some more examples of how big data is used by organizations:

- Big data is used by utilities to monitor electrical grids and by oil and gas corporations to locate possible drilling sites and follow pipeline activity in the energy sector.
- Big data platforms are used by financial services companies for risk management and in-the-moment market data analysis.
- Big data is used by manufacturers and transportation firms to manage their supply networks and improve delivery routes.
- Emergency response, crime prevention, and smart city programs are further government uses.

1.2 Evolution of Big Data

The term "big data" does not just refer to the vast amount of data available today, but the entire process of collecting, storing and analyzing this data. It's important to use this process to make the world a better place. Big data as we know it is so new that there isn't much of a past to look into, but what it does show us how big data has evolved and improved in a short period of time, giving us clues for future changes. Importantly, big data is no longer just a buzzword understood by a select few. It's becoming more mainstream, and those who actually implement big data are having great success.

1940s to 1989 – Data Warehousing and Delicate Desktop

The origins of electronic storage can be traced back to the development of the world's first programmable computer, the Electronic Numerical Integrator and Computer (ENIAC). It was developed by the United States Army during World War II to solve numerical problems such as calculating the range of artillery fire. Then, in the early 1960s, International Business Machines (IBM) released the first transistor computer called TRADIC. This helped data centres move away from military and serve broader commercial purposes.

The first personal desktop computer with a graphical user interface (GUI) was the Lisa, released by Apple Computers in 1983. In the 1980s, companies such as Apple, Microsoft, and

IBM introduced a variety of personal desktop computers, and more and more people bought their own personal computers and used them in their homes for the first time. Electronic storage has finally become available to the general public.

1989 to 1999 – Establishment of the World Wide Web

Between 1989 and 1993, British computer scientist Sir Tim Berners-Lee developed the basic technology needed to run what is known today as the World Wide Web. These his web technologies were Hypertext Markup Language (HTML), Uniform Resource Identifier (URI), and he Hypertext Transfer Protocol (HTTP). And he decided in April 1993 to make the underlying code of these web technologies freely available forever.

As a result, individuals, businesses, and organizations that can afford to pay for Internet services can now go online and exchange data with other Internet-enabled computers. As more devices have access to the Internet, there has been an explosion in the amount of information that people can access and share at the same time.

2000s to 2010s – Controlling Data Volume, Social Media and Cloud Computing

In the early 2000s, companies such as Amazon, eBay, and Google contributed to generating massive amounts of web traffic and a mix of structured and unstructured data. In his 2002 he also released a beta version of AWS (Amazon Web Services), giving all developers access to the Amazon.com platform. By 2004, over 100 applications had been created for this purpose.

Then in 2006, AWS rebooted to offer a wide range of cloud infrastructure services such as Simple Storage Service (S3) and Elastic Compute Cloud (EC2). AWS' general availability attracted a wide range of customers including Dropbox, Netflix, and Reddit. These customers wanted to be cloud-ready, so by 2010 they wanted to partner with AWS.

Social media platforms such as MySpace, Facebook, and Twitter have also contributed to the increased proliferation of unstructured data. This includes sharing images and audio files, animated GIFs, videos, status posts and direct messages.

2010s to now – Optimization Techniques, Mobile Devices and IoT

With the rise of mobile and IoT devices, new kinds of data are being collected, organized, and analyzed. Some examples are:

- Sensor data (data collected from web-enabled sensors to provide valuable real-time insight into the inner workings of your machine)

- Social data (social media data published by platforms such as Facebook and Twitter)
- Transaction data (data from online web shops, receipts, inventory records, repeat purchases, etc.)
- Health-related data (heart rate monitor, medical records, medical history).

1.3 Types of Big Data

Big data also encompasses a wide variety of data types, including the following:

- **Structured**
- **Unstructured**
- **Semi-structured**

Structured Data

Any data that can be stored, accessed and processed in a fixed format is called "structured" data. Over time, computer science talent has had great success in developing techniques for manipulating such data (if its format is known in advance) and extracting value from it. Today, however, problems arise when the size of such data increases significantly. Common sizes are in the range of a few zettabytes.

Examples of Structured Data

An ‘Employee’ table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Ramesh	Male	Finance	650000
3398	Joshi	Female	Admin	650000
7465	Roy	Male	Admin	500000
7500	Anto	Male	Finance	500000
7699	Sekar	Female	Finance	550000

Unstructured Data

Data of unknown format or structure are classified as unstructured data. Unstructured data is not only huge, it poses many challenges as to how to process it to extract value from it. A typical example of unstructured data is heterogeneous data sources that contain combinations of simple text files, images, videos, etc. Today's businesses have a wealth of data, but unfortunately, in raw or unstructured form, they don't know what to do with it.

Examples of Un-structured Data

The media file, IoT data, Document files, Analytics data etc

Semi-structured

Semi-structured data can contain both forms of data. Semi-structured data can be thought of as structured data, but is really undefined. B. A table definition in a relational DBMS. An example of semi-structured data is data represented in an XML file.

Examples of Semi-structured Data

Personal data stored in an XML file-

```
<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>
```

1.4 Characteristic of Big Data

Big data can be described by the following characteristics:

- Volume
- Variety
- Velocity
- Variability
- Value

Volume

The name big data itself refers to its huge size. Data scope plays a very important role in determining the value of data. Whether certain data is really considered big data also depends on the amount of data. Therefore, when dealing with big data solutions, "volume" must be considered. Data is ever-growing day by day of every type ever Kilobyte, Megabyte, Gigabyte, Terabyte, Petabyte, Exabyte, Zettabyte, Yottabyte of information. Within the Social Media space for example, Volume refers to the quantity of data generated through websites, portals and on-line applications.

Variety (data type)

Variety in big data refers to all or any of the structured and unstructured data that has the possibility of obtaining generated either by humans or by machines. Variety refers to heterogeneous sources and types of data, both structured and unstructured. Previously, spreadsheets and databases were the only data sources considered for most applications. Data in the form of emails, photos, videos, surveillance devices, PDFs, audio, etc. are now also considered in analytics applications. This large amount of unstructured data poses certain problems when it comes to storing, mining, and analyzing data.

Velocity (streaming data)

The term "velocity’" refers to the speed of data generation. How quickly the data is generated and processed to meet the requirements determines the actual potential of the data.

Big Data Velocity addresses the speed at which data flows in from sources such as business processes, application logs, network and social media sites, sensors, and mobile devices. The flow of data is massive and continuous.

Variability

This means that the data sometimes exhibits inconsistencies, complicating the process of effectively handling and managing it. Variability refers to the deviation of data from its mean and is often used in statistics and finance. In finance is often applied to instability in returns, and investors prefer investments with lower explosiveness and higher returns.

Value

A critical “V” from a business perspective: The value of big data typically comes from insights and pattern recognition discoveries that lead to more effective operations, stronger customer relationships, and other clear and quantifiable business benefits.

1.5 How does big data work?

Big data involves collecting, processing, and analyzing vast amounts of data from multiple sources to uncover patterns, relationships, and insights that can inform decision-making. The process involves several steps:

Data Collection: Big data is collected from various sources such as social media, sensors, transactional systems, and customer reviews.

Data Storage: Collected data should be stored in such a way that it can be easily retrieved and analyzed later. This often requires special storage technologies capable of handling large amounts of data.

Data Processing: After storing the data, it must be processed before it can be analyzed. This includes cleaning and organizing the data to eliminate errors and discrepancies and transforming it into a format suitable for analysis.

Data Analysis: After the data is processed, tools such as statistical models and machine learning algorithms are used to analyze the data to identify patterns, relationships and trends.4

Data Visualization: Insights derived from data analysis are presented in visual formats such as graphs, charts, and dashboards, making them easier for decision makers to understand and act on.

1.6 Big Data tools

- ***Apache Hadoop*** is an open source big data tool for storage and handling large amounts of data across multiple servers. Hadoop includes a distributed file system (HDFS) and a MapReduce processing engine.
- ***Apache Spark*** is a fast and versatile cluster computing system that supports in-memory processing for accelerating iterative algorithms. Spark can be used for batch processing, real-time stream processing, machine learning, chart processing, and SQL queries.
- ***Apache Cassandra*** is a scattered NoSQL database management system deliberate to process large amounts of data on general persistence servers with high availability and fault tolerance.
- ***Apache Flink*** is an open source streaming data handling framework that supports batch processing, real-time stream processing, and event-driven applications. Flink offers low latency, high throughput data processing, fault tolerance, and scalability.

- ***Apache Kafka*** is a distributed streaming platform that enables real-time publishing and subscription to streams of datasets. Kafka is used to build concurrent data pipelines and streaming applications.
- ***Splunk*** is a software platform for searching, monitoring, and analyzing machine-generated big data in real time. Splunk collects and indexes data from various sources to provide insight into operational and business information.
- ***Talend*** is an open source data integration platform that enables enterprises to extract, transform, and load (ETL) data from various sources. Talend provisions big data skills such as Hadoop, Spark, Hive, Pig and HBase.
- ***Tableau*** is a data visualization and business intelligence tool that enables users to analyze and share data using interactive dashboards, reports and charts. Tableau cares big data stages and databases such as Hadoop, Amazon Redshift, and Google BigQuery.
- ***Apache NiFi*** is a data flow management tool for automating the movement of data between systems. NiFi supports big data technologies such as Hadoop, Spark, and Kafka to provide real-time data processing and analytics.
- **QlikView** is a commercial intelligence and data visualization tool that allows users to analyze and segment data using interactive dashboards, reports and charts. QlikView supports big data platforms such as Hadoop to provide real-time data processing and analysis.

1.7 Use Cases

Big data helps companies make better and faster decisions because they have more information to solve problems and more data to test hypotheses.

Customer Experience is a key area that has been revolutionized by the advent of big data. Industries are collecting more data than ever before about their customers and their preferences. This data is actively used by providing our customers with personalized recommendations and offers, and companies are happy to permit the collection of this data in exchange for personalized services. Netflix and Amazon/Flipkart recommendations are a gift from big data.

Machine Learning is another area that has benefited greatly from the growing popularity of big data. More data means you have a larger dataset to train your ML model on, and a better trained model will (generally) perform better as well. Thanks to big data, it is now possible to automate tasks that were previously done manually with the help of machine learning.

Demand forecasting is a more data collected about customer purchases, the more accurate it becomes. This helps companies create predictive models that help them forecast future demand and scale production accordingly. This helps companies, especially in the manufacturing sector, reduce the cost of storing unsold inventory in warehouses.

Agriculture according to United Nations estimates, the world population will reach 9.8 billion by 2050. Agriculture must be transformed to meet the food needs of such a large population. However, climate change has not only made most farmland unsuitable for agriculture, it has also affected rainfall patterns and dried up many water sources.

Automotive whether in R&D or marketing planning, big data analysis has enormous fields of application in the automotive industry, which combines several individual sectors. As a major infrastructure segment supporting many important public and private ecosystems, the automotive sector generates massive amounts of data every day.

Chapter Summary

In this chapter, discussed about the topic Big Data, structured and unstructured data, some real-world applications of Big Data, and how they can store and process Big Data using cloud platforms and Hadoop. Big data is data on a very large scale. Big data is a term that describes a huge but exponentially growing collection of data over time. Examples of big data analytics include stock exchanges, social media pages, and airplane engines.