UNIT I

INTRODUCTION TO BIG DATA

Big Data Platforms are a complete system that helps organizations work with large and complex datasets. It gives them the tools and technology to turn raw data into valuable information.

This, in turn, helps them make decisions based on data and come up with innovative solutions.

IMPORTANCE OF BIG DATA PLATFORMS

Big Data has transformed the way businesses operate, and it has become a valuable resource that, when harnessed effectively, can drive innovation, enhance decision-making, and create competitive advantages. This is where Big Data Platforms play a pivotal role.

a) Data-driven decision-making: In the past, decisions relied on intuition, but today, data-backed decisions are crucial. Big Data Platforms enable real-time collection, processing, and analysis of vast datasets, empowering businesses to make informed decisions, identify trends, and predict outcomes more accurately.

b) Improved customer experiences: Understanding customer behaviour is vital for personalised experiences. Big Data Platforms gather and analyse data from various touchpoints, like websites and social media. This allows companies to tailor products, services, and marketing to individual needs, boosting customer satisfaction and loyalty.

c) Enhanced operational efficiency: Big Data Platforms streamline operations by optimising processes and reducing waste. For instance, in manufacturing, real-time data analysis identifies bottlenecks and maintenance needs, saving costs. In logistics, it optimises routes and reduces fuel consumption, improving overall efficiency.

d) Innovation and product development: Big Data Platforms drive innovation by revealing market trends and consumer behaviour. Analysing large datasets helps companies identify gaps and develop products that meet demand, driving revenue and maintaining a competitive edge.

e) Fraud detection and security: In an era of cyber threats, Big Data Platforms swiftly detect and mitigate risks by analysing real-time patterns and anomalies, bolstering security with robust access controls and encryption.

f) Healthcare advancements: Big Data Platforms revolutionise healthcare by analysing patient data and genomic information, leading to advances in disease detection, drug development, and personalised medicine.

g) Competitive advantage: Firms leveraging Big Data Platforms adapt swiftly to market changes, capitalise on opportunities, and deliver superior products and services, gaining a competitive edge.

h) Scientific and research advancements: Beyond business, Big Data Platforms accelerate scientific research by analysing vast datasets, facilitating breakthroughs in fields like climate science and genomics.

i) Government and social impact: Public organisations utilise data analytics from Big Data Platforms to enhance services, allocate resources optimally, and make informed decisions, improving citizens' quality of life.

j) Economic growth and job creation: The adoption of Big Data Platforms fuels economic growth by creating job opportunities in Data Analytics and Data Science, contributing to economic development.

Components of Big Data Platforms

Big Data Platforms are complex systems designed to handle vast volumes of data, process it efficiently, and turn it into valuable insights. These platforms consist of several essential components, each playing a critical role in overall functionality.

Data ingestion and collection

This is the first step in the Big Data journey. Data can come from various sources, including sensors, applications, social media, and databases. The data ingestion component is responsible for gathering this diverse data and making it ready for processing. It involves data connectors, adapters, and protocols to ensure data from different sources can be efficiently brought into the platform.

Data storage

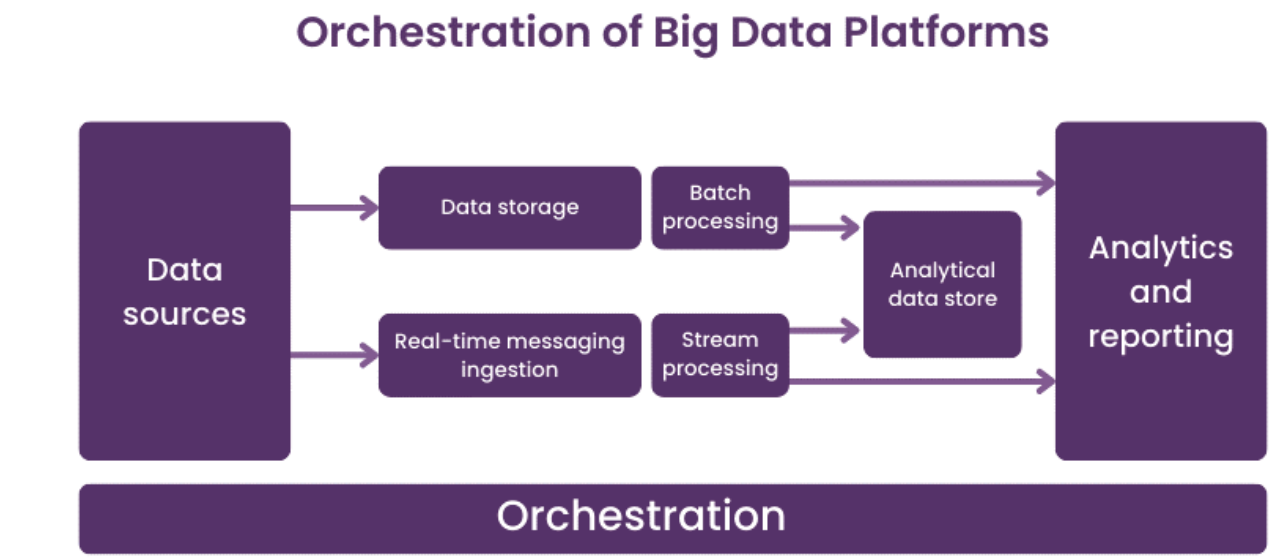
Once data is ingested, it needs a place to reside. Big Data Platforms employ a variety of storage solutions designed to handle large datasets. Common storage systems include distributed file systems (e.g., Hadoop HDFS, Amazon S3) and NoSQL databases (e.g., Apache Cassandra, MongoDB). These storage systems are optimised for scalability, fault tolerance, and high availability, ensuring data remains accessible and reliable even as it grows.

Data processing and analysis

This is the heart of Big Data Platforms, where data is transformed, processed, and analysed to extract meaningful insights. Processing engines and frameworks like Apache Spark, Apache Flink, and Hadoop MapReduce play a vital role in this component. They distribute and parallelise computations across clusters of machines, enabling the platform to handle massive workloads efficiently.

Data management and orchestration

Managing and orchestrating data processing tasks across a distributed infrastructure is a complex task. The management layer includes components for resource allocation, job scheduling, and workflow orchestration. This layer ensures that data processing tasks run smoothly and efficiently, optimising resource utilisation.



Data visualisation and reporting

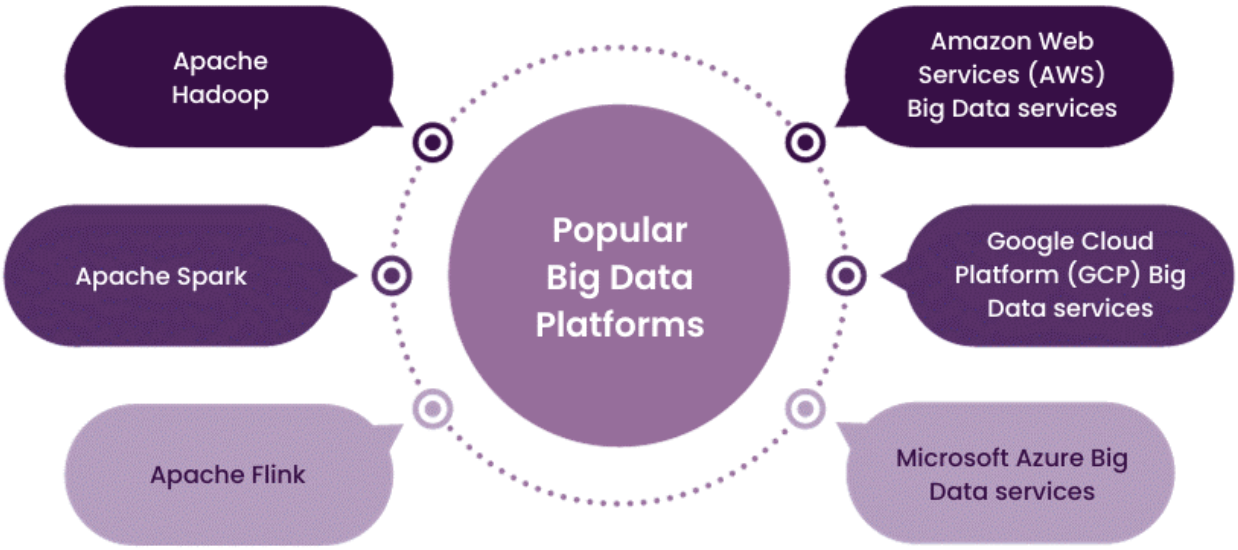
The insights derived from Big Data analysis are only valuable if they can be understood and acted upon. This includes tools and technologies for data visualisation and reporting. This component allows users to create interactive dashboards, generate reports, and visualise trends and patterns in the data.

Security and governance

Data security and governance are paramount in Big Data Platforms, especially when dealing with sensitive information. This layer includes components for authentication, authorisation, encryption, and auditing. It ensures that data is protected from unauthorised access and maintains compliance with regulatory requirements.

Popular Big Data Platforms

Big Data Platforms are like the superheroes of the digital world, capable of handling massive amounts of data and turning it into valuable information. Here, we'll introduce you to a list of Big Data Platforms:



a) Apache Hadoop: Apache Hadoop is a platform that's excellent at storing and processing large volumes of data. It's like a robust storage and data processing system that companies use to handle and manage massive datasets.

b) Apache Spark: Apache Spark is known for its speed and efficiency in analysing data. It's like a powerful tool that helps organisations quickly make sense of their data and extract valuable insights from it.

c) Apache Flink: Apache Flink is another data processing platform, similar to Spark, that specialises in real-time data analysis. It's used for tasks where speed and low latency are critical, like monitoring online activities or financial transactions.

d) Amazon Web Services (AWS) Big Data services: AWS offers a suite of Big Data services that run in the cloud. These services make it easier for companies to store, process, and analyse data without the need for extensive infrastructure management.

e) Google Cloud Platform (GCP) Big Data services: Similar to AWS, Google Cloud Platform provides a range of Big Data services in the cloud. These services help organisations leverage Google's computing power and data analytics capabilities.

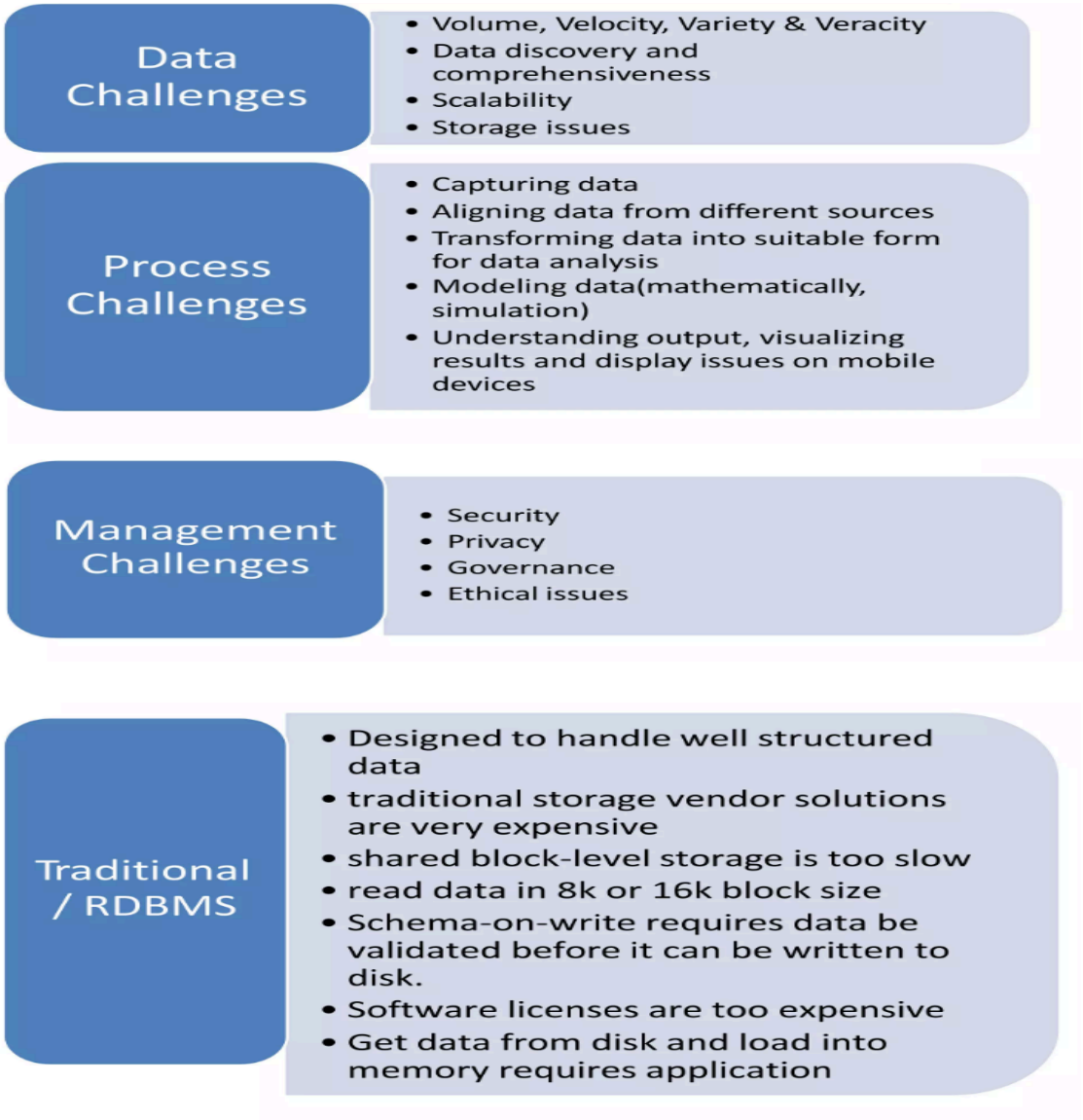
f) Microsoft Azure Big Data services: Microsoft Azure offers various Big Data services, including data storage, processing, and analytics tools. These services are designed to help businesses work with their data efficiently and effectively.

CHALLENGES OF CONVENTIONAL SYSTEMS

'Analytics' has been used in the business intelligence world to provide tools and intelligence to gain insight into the data Data mining is used in enterprises to keep pace with the critical monitoring and analysis of mountains of data

Common Challenges

It cannot work on unstructured data efficiently It is built on top of the relational data model. It is batch oriented and we need to wait for nightly ETL (extract, transform and load) and transformation jobs to complete before the required insight is obtained. Parallelism in a traditional analytics system is achieved through costly hardware like MPP (Massively Parallel Processing) systems. I Inadequate support of aggregated summaries of data.



INTELLIGENT DATA ANALYSIS DEFINITION

Intelligent Data Analysis (IDA) is an interdisciplinary study that is concerned with the extraction of useful knowledge from data, drawing techniques from a variety of fields, such as artificial intelligence, high-performance computing, pattern recognition, and statistics.

Data intelligence platforms and data intelligence solutions are available from data intelligence companies such as Data Visualization Intelligence, Strategic Data Intelligence, Global Data Intelligence.

What is Intelligent Data Analysis?

Intelligent data analysis refers to the use of analysis, classification, conversion, extraction organization, and reasoning methods to extract useful knowledge from data. This data analytics intelligence process generally consists of the data preparation stage, the data mining stage, and the result validation and explanation stage.

Data preparation involves the integration of required data into a dataset that will be used for data mining; data mining involves examining large databases in order to generate new information; result validation involves the verification of patterns produced by data mining algorithms; and result explanation involves the intuitive communication of results.

DATA

Data can be defined as a representation of facts, concepts, or instructions in a formalized manner.

Data can be defined as a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing by human or electronic machines.

Definition of Information

Information is organized or classified data, which has some meaningful values for the receiver. Information is the processed data on which decisions and actions are based.

Characteristics of Data

The following are six key characteristics of data which discussed below:

1. Accuracy
2. Validity
3. Reliability
4. Timeliness
5. Relevance
6. Completeness

Accuracy

Data should be sufficiently accurate for the intended use and should be captured only once, although it may have multiple uses. Data should be captured at the point of activity.

Validity

Data should be recorded and used in compliance with relevant requirements, including the correct application of any rules or definitions. This will ensure consistency between periods and with similar organizations, measuring what is intended to be measured.

Reliability

Data should reflect stable and consistent data collection processes across collection points and over time. Progress toward performance targets should reflect real changes rather than variations in data collection approaches or methods. Source data is clearly identified and readily available from manual, automated, or other systems and records.

Timeliness

Data should be captured as quickly as possible after the event or activity and must be available for the intended use within a reasonable time period. Data must be available quickly and frequently enough to support information needs and to influence service or management decisions.

Relevance

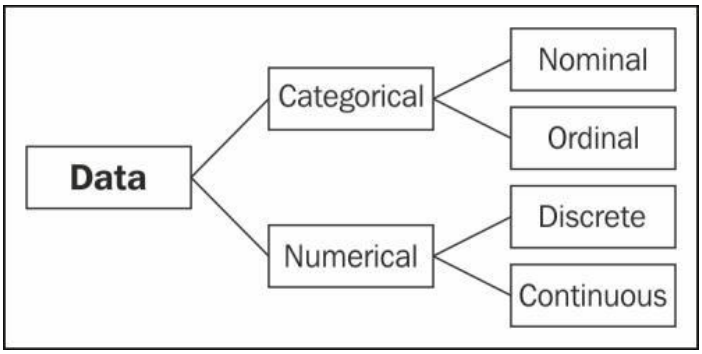
Data captured should be relevant to the purposes for which it is to be used. This will require a periodic review of requirements to reflect changing needs.

Completeness

Data requirements should be clearly specified based on the information needs of the organization and data collection processes matched to these requirements.

NATURE OF DATA

Data is the plural of datum, so it is always treated as plural. We can find data in all the situations of the world around us, in all the structured or unstructured, in continuous or discrete conditions, in weather records, stock market logs, in photo albums, music playlists,or in our Twitter accounts. In fact, data can be seen as the essential raw material of any kind of human activity.



Categorical data are values or observations that can be sorted into groups or categories. There are two types of categorical values, nominal and ordinal.

A nominal variable has no intrinsic ordering to its categories. For example, housing is a categorical variable having two categories (own and rent).

An ordinal variable has an established ordering. For example, age as a variable with three orderly categories (young, adult, and elder).

Numerical data are values or observations that can be measured. There are two kinds of numerical values, discrete and continuous.

Discrete data are values or observations that can be counted and are distinct and separate. For example, number of lines in a code.

Continuous data are values or observations that may take on any value within a finite or infinite interval. For example, an economic time series such as historic gold prices.

- E-mails (unstructured, discrete)
- Digital images (unstructured, discrete)
- Stock market logs (structured, continuous)
- Historic gold prices (structured, continuous)
- Credit approval records (structured, discrete)
- Social media friends and relationships (unstructured, discrete)
- Tweets and trending topics (unstructured, continuous)
- Sales records (structured, continuous)

EVOLUTION OF BIG DATA

If we see the last few decades, we can analyze that Big Data technology has gained so much growth. There are a lot of milestones in the evolution of Big Data which are described below:

1. ***Data Warehousing:***
   In the 1990s, data warehousing emerged as a solution to store and analyze large volumes of structured data.
2. ***Hadoop:***
   Hadoop was introduced in 2006 by Doug Cutting and Mike Cafarella. Distributed storage medium and large data processing are provided by Hadoop, and it is an open-source framework.
3. ***NoSQL Databases:***
   In 2009, NoSQL databases were introduced, which provide a flexible way to store and retrieve unstructured data.
4. ***Cloud Computing:***
   Cloud Computing technology helps companies to store their important data in data centers that are remote, and it saves their infrastructure cost and maintenance costs.

5. ***Machine Learning:***
   Machine Learning algorithms are those algorithms that work on large data, and analysis is done on a huge amount of data to get meaningful insights from it. This has led to the development of artificial intelligence (AI) applications.
6. ***Data Streaming:***
   Data Streaming technology has emerged as a solution to process large volumes of data in real time.
7. ***Edge Computing:***
   dge Computing is a kind of distributed computing paradigm that allows data processing to be done at the edge or the corner of the network, closer to the source of the data.

Overall, big data technology has come a long way since the early days of data warehousing. The introduction of Hadoop, NoSQL databases, cloud computing, machine learning, data streaming, and edge computing has revolutionized how we store, process, and analyze large volumes of data. As technology evolves, we can expect Big Data to play a very important role in various industries.

BIG DATA

Big data refers to huge amount of data.

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data.

Big Data *is a collection of data that is huge in volume, yet growing exponentially with time.* It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

CHALLENGES OF BIG DATA

Storage

With vast amounts of data generated daily, the greatest challenge is storage (especially when the data is in different formats) within legacy systems. Unstructured data cannot be stored in traditional databases.

Processing

Processing big data refers to the reading, transforming, extraction, and formatting of useful information from raw information. The input and output of information in unified formats continue to present difficulties.

Security

Security is a big concern for organizations. Non-encrypted information is at risk of theft or damage by cyber-criminals. Therefore, data security professionals must balance access to data against maintaining strict security protocols.

Finding and Fixing Data Quality Issues

Many of you are probably dealing with challenges related to poor data quality, but solutions are available. The following are four approaches to fixing data problems:

- Correct information in the original database.
- Repairing the original data source is necessary to resolve any data inaccuracies.
- You must use highly accurate methods of determining who someone is.

Scaling Big Data Systems

Database sharding, memory caching, moving to the cloud and separating read-only and write-active databases are all effective scaling methods. While each one of those approaches is fantastic on its own, combining them will lead you to the next level.

Evaluating and Selecting Big Data Technologies

Companies are spending millions on new big data technologies, and the market for such tools is expanding rapidly. In recent years, however, the IT industry has caught on to big data and analytics potential. The trending technologies include the following:

- Hadoop Ecosystem
- Apache Spark
- NoSQL Databases
- R Software
- Predictive Analytics
- Prescriptive Analytics

Big Data Environments

In an extensive data set, data is constantly being ingested from various sources, making it more dynamic than a data warehouse. The people in charge of the big data environment will fast forget where and what each data collection came from.

Real-Time Insights

The term "real-time analytics" describes the practice of performing analyses on data as a system is collecting it. Decisions may be made more efficiently and with more accurate information thanks to real-time analytics tools, which use logic and mathematics to deliver insights on this data quickly.

Data Validation

Before using data in a business process, its integrity, accuracy, and structure must be validated. The output of a data validation procedure can be used for further analysis, BI, or even to train a machine learning model.

Healthcare Challenges

Electronic health records (EHRs), genomic sequencing, medical research, wearables, and medical imaging are just a few examples of the many sources of health-related big data.

Barriers to Effective Use Of Big Data in Healthcare

- The price of implementation
- Compiling and polishing data
- Security
- Disconnect in communication

BIG DATA CHARACTERISTICS

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many **multinational companies** to **process** the data and business of many **organizations**. The data flow would exceed **150 exabytes** per day before replication.

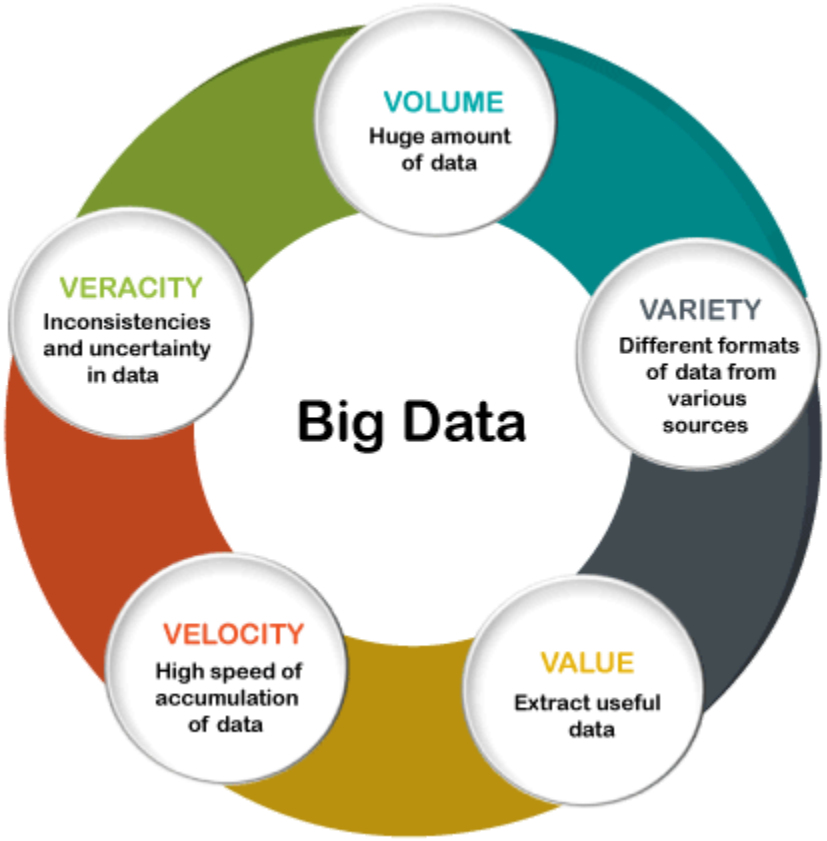
There are five v's of Big Data that explains the characteristics.

3 V's of Big Data

- ***Volume***
- ***Veracity***
- ***Variety***

Other Characteristics of Big Data

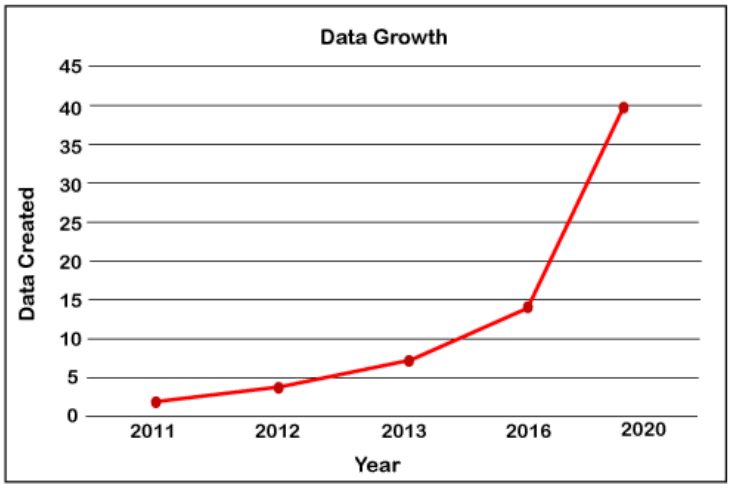
- ***Value***
- ***Velocity***



Volume

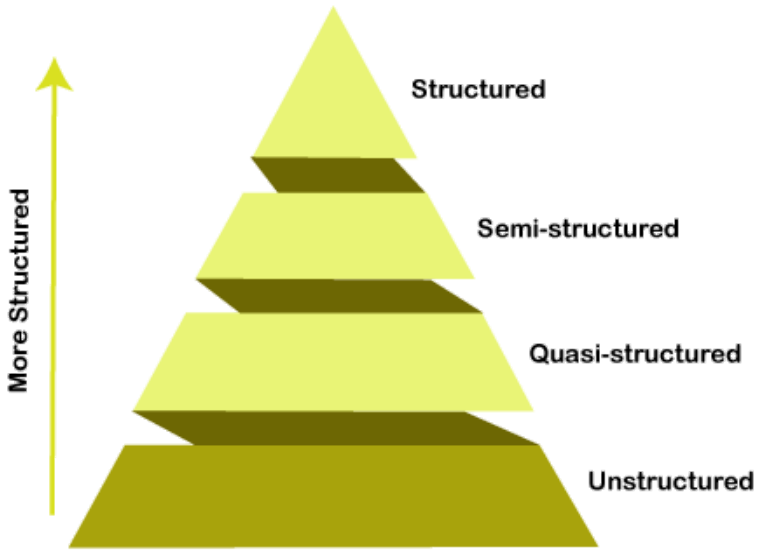
The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more.

Facebook can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.



Variety

Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.



The data is categorized as below:

a. **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

a. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.

b. **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.

c. **Quasi-structured Data:**The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of **activities**.

Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in **real-time**. It contains the linking of incoming **data sets speeds, rate of change**, and **activity bursts**. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices,** etc.

Veracity

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.

For example, **Facebook posts** with hashtags.

Value

Value is an essential characteristic of big data. It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.

NEED FOR BIG DATA

Big Data initiatives were rated as "extremely important" to 93% of companies. Leveraging a Big Data analytics solution helps organizations to unlock the strategic values and take full advantage of their assets.

It helps organizations:

- To understand Where, When and Why their customers buy
- Protect the company's client base with improved loyalty programs
- Seizing cross-selling and upselling opportunities
- Provide targeted promotional information
- Optimize Workforce planning and operations
- Improve inefficiencies in the company's supply chain
- Predict market trends
- Predict future needs
- Make companies more innovative and competitive
- It helps companies to discover new sources of revenue

ANALYTICS PROCESSES AND TOOLS:

1. APACHE Hadoop

It's a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel.

2. Cassandra

APACHE Cassandra is an open-source NoSQL distributed database that is used to fetch large amounts of data. It's one of the **most popular tools for data analytics** and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance.

3. Qubole

It's an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines

4. Xplenty

It is a data analytic tool for building a data pipeline by using minimal codes in it. It offers a wide range of solutions for sales, marketing, and support. With the help of its interactive graphical interface, it provides solutions for *ETL, ELT*, etc.

5. Spark

APACHE Spark is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools.

6. Mongo DB

Came in limelight in 2010, is a free, open-source platform and a **document-oriented (NoSQL) database** that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of Mongo DB.

7. Apache Storm

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to **handle a pool of large data in fault-tolerance and horizontally scalable methods**

ANALYTICS VS REPORTING

Analytics is the technique of examining data and reports to obtain actionable insights that can be used to comprehend and improve business performance. Business users may gain insights from data, recognize trends, and make better decisions with workforce analytics.

Analytics	Reporting
Analytics is the method of examining and analyzing summarized data to make business decisions.	Reporting is an action that includes all the needed information and data and is put together in an organized way.
Questioning the data, understanding it, investigating it, and presenting it to the end users are all part of analytics.	Identifying business events, gathering the required information, organizing, summarizing, and presenting existing data are all part of reporting.
The purpose of analytics is to draw conclusions based on data.	The purpose of reporting is to organize the data into meaningful information.
Analytics is used by data analysts, scientists, and business people to make effective decisions.	Reporting is provided to the appropriate business leaders to perform effectively and efficiently within a firm.

Analytics and reporting can be used to reach a number of different goals. Both of these can be very helpful to a business if they are used correctly.