Big Data

Unit 1

Types of digital data/ Types of Big Data

DIGITAL DATA

Digital data is information stored on a computer system as a series of 0's and 1's in a binary language. Digital data jumps from one value to the next in a step by step sequence. Example: Whenever we send an email, read a social media post, or take pictures with our digital camera, we are working with digital data.

Digital data can be classified into three forms:

a. **Unstructured Data:** The data which does not conform to a data model or is not in a formthat can be used easily by a computer program is categorized as unstructured data. About 80—90% data of an organization is in this format.

Example: Memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, the body of an email, etc.

b. **Semi-Structured Data:** The data which does not conform to a data model but has some structure is categorized as semi-structured data. However, it is not in a form that can be usedeasily by a computer program.

Example: Emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

c. **Structured Data:** The data which is in an organized form (ie. in rows and columns) andcan be easily used by a computer program is categorized as semi-structured data. Relationships exist between entities of data, such as classes and their objects.

Example: Data stored in databases- College database, Banking database, Hospital database etc

HISTORY OF BIG DATA

The 21 st century is characterized by the rapid advancement in the field of information technology. IT has become an integral part of daily life as well as various other industries

like: health, education, entertainment, science and technology, genetics, or business operations and these industries generate a lot of data, this can be called Big Data.

The ancient history of Big Data

300 BC

The ancient Egyptians around 300 BC already tried to capture all existing ‘data’ in the library of Alexandria. Moreover, the Roman Empire used to carefully analyze statistics

of their military to determine the optimal distribution for their armies.

1884

Herman Hollerith invents the punch card tabulating machine.

Big Data in 20th century

1937

IBM got the contract to develop punch card-reading machine for this massive bookkeeping project.

1943

The first data-processing machine named **Colossus** that appeared in 1943 and was developed by the British to decipher Nazi codes during World War II. This device, named Colossus, searched for patterns in intercepted messages at a rate of 5,000 characters per second, reducing the length of time the task took from weeks to merely hours.

The internet age and the dawn of Big Data

Between 1989 and 1990

The World Wide Web and developed HTML, URLs and HTTP, all while working for CERN. The internet age with widespread and easy access to data had begun

1996

Digital data storage had become more cost-effective than storing information on paper

1998

The domain google.com was launched

NoSQL, an open-source relational database was developed that provided a way to store and retrieve data modelled

The information age

Since the early 2000s, With the expansion of web traffic and online stores, companies such as Yahoo, Amazon and eBay started to analyze customer behavior by looking at click-rates, IP-specific location data and search logs.

2005

Big Data was labelled by Roger Mougalas as he referred to a large set of data.

Hadoop, which could handle Big Data, was created by Doug Cutting and Mike Caferalla.

Big Data revolutionized entire industries and changed human culture and behavior.

For example, Big Data is being used in healthcare to map disease outbreaks and test alternative treatments. NASA uses Big Data to explore the universe. The music industry replaces intuition with Big Data studies. Utilities use Big Data to study customer behavior and avoid blackouts

Introduction to Big Data platform

A big data platform is a type of IT solution that combines the features and capabilities of several big data applications and utilities within a single solution, this is then used further for managing as well as analyzing Big Data.

It focuses on providing its users with efficient analytics tools for massive datasets.

The users of such platforms can custom build applications according to their use case like to calculate customer loyalty (E-Commerce user case), and so on.

Goal: The main goal of a Big Data Platform is to achieve: Scalability, Availability, Performance, and Security.

Example: Some of the most commonly used Big Data Platforms are :

- Hadoop Delta Lake Migration Platform
- Data Catalog Platform
- Data Ingestion Platform
- IoT Analytics Platform
- ETL Transform Platform

Hadoop - Delta Lake Migration Platform

It is an open-source software platform managed by Apache Software Foundation. It is used to manage and store large data sets at a low cost and with great efficiency.

IoT Analytics Platform

It provides a wide range of tool to work upon it; this functionality of it comes handy while using it over the IoT case.

Data Ingestion Platform

This layer is the first step for the data coming from variable sources to start its journey. This means the data here is prioritized and categorized, making data flow smoothly in further layers in this process flow.

Data Catalog Platform

It provides a single self-service environment to the users, helping them find, understand, and trust the data source.

ETL (Extract Transfer Load) Data Transformation Platform

This Platform can be used to build pipelines and even schedule the running of the same for data transformation

Drivers for Big Data

Big Data has quickly risen to become one of the most desired topics in the industry.

The main business drivers for such rising demand for Big Data Analytics are :

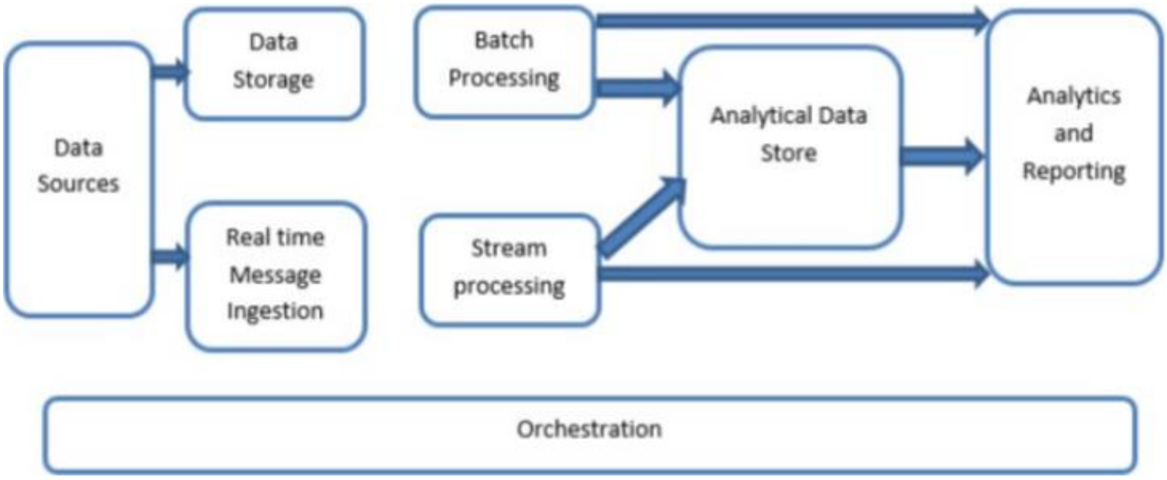
1. The digitization of society
2. The drop in technology costs
3. Connectivity through cloud computing
4. Increased knowledge about data science
5. Social media applications
6. The rise of Internet-of-Things(IoT)

Example: A number of companies that have Big Data at the core of their strategy like :

Apple, Amazon, Facebook and Netflix have become very successful at the beginning of the 21st century.

Big Data Architecture:

Big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems.



The big data architectures include the following components:

Data sources: All big data solutions start with one or more data sources.

Example,

- Application data stores, such as relational databases.
- Static files produced by applications, such as web server log files.
- Real-time data sources, such as IoT devices.

Data storage: Data for batch processing operations is stored in a distributed file store that can hold high volumes of large files in various formats (also called data lake).

Example,

Azure Data Lake Store or blob containers in Azure Storage.

Batch processing: Since the data sets are so large, therefore a big data solution must process data files using long-running batch jobs to filter, aggregate, and prepare the data for analysis.

Real-time message ingestion: If a solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing.

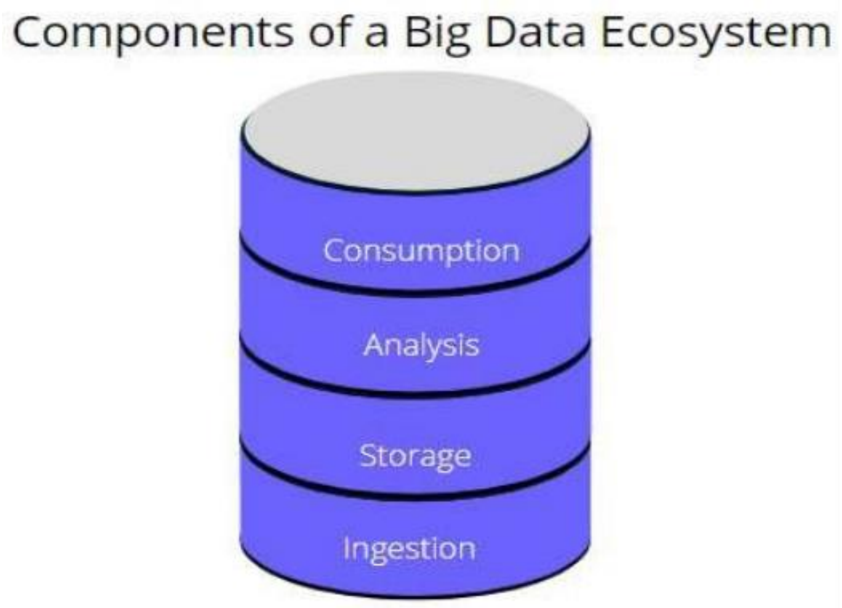
Stream processing: After capturing real-time messages, the solution must process them by filtering, aggregating, and preparing the data for analysis. The processed stream written to an output sink. We can use open-source Apache streaming technologies like Stormand Spark Streaming for this.

Analytical data store: Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. Example: Azure Synapse Analytics provides a managed service for large-scale, cloud-based data warehousing.

Analysis and reporting: The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower users to analyze the data, the architecture may include a data modelling layer. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts.

Orchestration: Most big data solutions consist of repeated data processing operations, that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report. To automate these workflows, we can use an orchestration technology such as Azure Data Factory.

Big Data Technology Components (Layers of Big Data) :



1. Ingestion :

The ingestion layer is the very first step of pulling in raw data.

It comes from internal sources, relational databases, non-relational databases, social media, emails, phone calls etc.

There are two kinds of ingestions :

Batch, in which large groups of data are gathered and delivered together.

Streaming, which is a continuous flow of data. This is necessary for real-time data analytics.

2. Storage :

Storage is where the converted data is stored in a data lake or warehouse and eventually processed.

The data lake/warehouse is the most essential component of a big data ecosystem.
It needs to contain only thorough, relevant data to make insights as valuable as possible.
It must be efficient with as little redundancy as possible to allow for quicker processing.

3. Analysis :
In the analysis layer, data gets passed through several tools, shaping it into actionable insights.
There are four types of analytics on big data :

- **Diagnostic:** Explains why a problem is happening.
- **Descriptive:** Describes the current state of a business through historical data.
- **Predictive:** Projects future results based on historical data.
- **Prescriptive:** Takes predictive analytics a step further by projecting best future efforts.

4. Consumption :
The final big data component is presenting the information in a format digestible to the end-user.

This can be in the forms of tables, advanced visualizations and even single numbers if requested.
The most important thing in this layer is making sure the intent and meaning of the output is understandable.

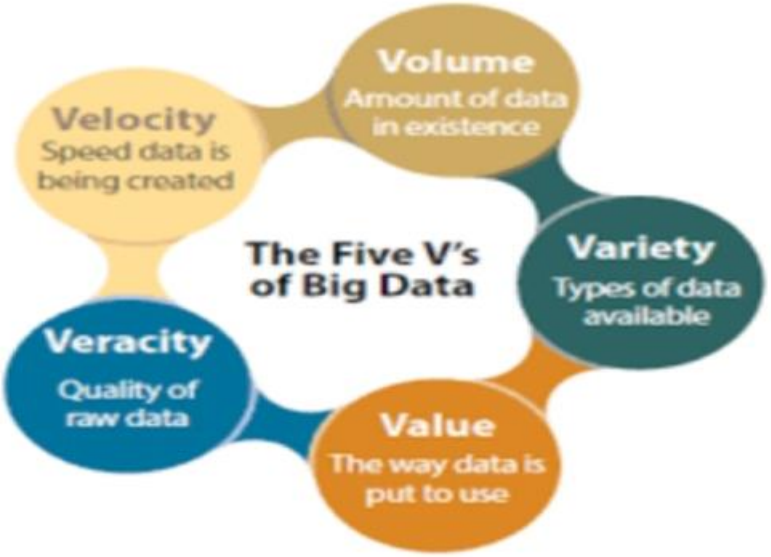
Big Data Characteristics/ Dimensions/ 5 V's :

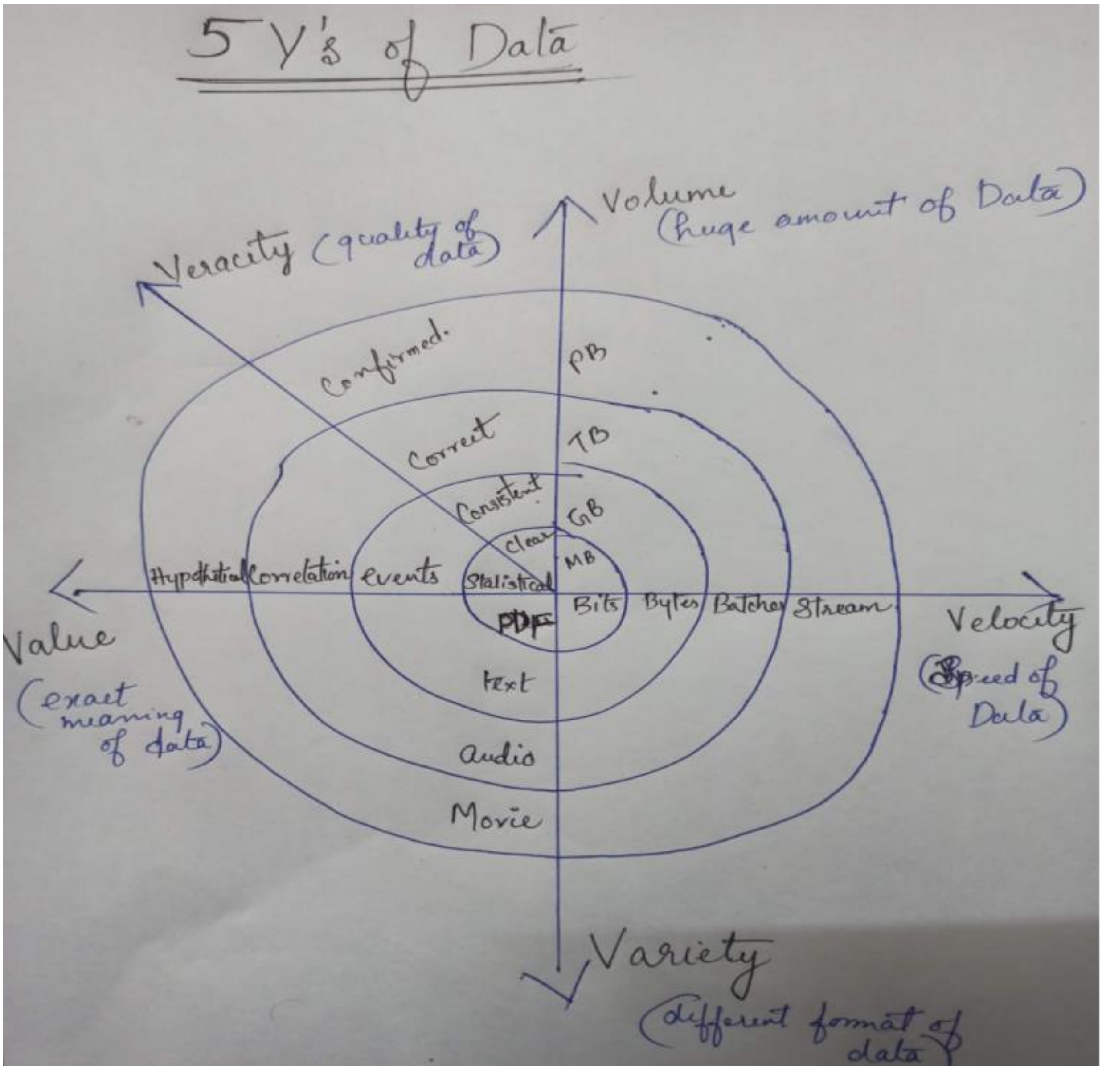
Big data can be described by the following characteristics:

- Volume
- Variety
- Velocity
- Veracity
- Value

5 Vs of Big Data, Big Data technology components

5 Vs of Big Data :


The Five V's of Big Data
Volume
Amount of data in existence
Variety
Types of data available
Value
The way data is put to use
Veracity
Quality of raw data
Velocity
Speed data is being created


5 V's of Data
Volume (huge amount of Data)
Veracity (quality of data)
Value (exact meaning of data)
Velocity (Speed of Data)
Variety (different format of data)
Confirmed.
Correct
Consistent
Clean
PB
TB
GB
MB
Hypothetical
Correlation
events
Statistical
Bits
Bytes
Batches
Stream
PDF
text
audio
Movie

1. Volume : (Huge amount of data)

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 so on. Example: Facebook generates approximately a billion messages, 4.5 billion times 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.

2 Variety : (Different types of data)

Big Data can be structured, unstructured, and semi-structured that are being collected from different sources.

Data were only collected from databases and sheets in the past, But these days the data will come in an array of forms ie.- PDFs, Emails, audios, Social Media posts, photos, videos, etc.

3 Velocity : (High speed of data)

Velocity refers to the speed with which data is generated in real-time.
Velocity plays an important role compared to others.

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.

Example of data that is generated with high velocity - Twitter messages or Facebook posts.

4 Veracity : (quality of data)

Veracity refers to the quality of the data that is being analyzed.
It is the process of being able to handle and manage data efficiently.
Example: Facebook posts with hashtags.

5 Value : (Exact meaning of data)

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 analyse.

Big Data importance and applications

Big Data Importance :

Big Data importance doesn’t revolve around the amount of data a company has but lies in the fact that how the company utilizes the gathered data.

Every company uses its collected data in its own way. More effectively the company uses its data, more rapidly it grows.

By analysing the big data pools effectively the companies can get answers to :

Cost Savings :

- Some tools of Big Data like Hadoop can bring cost advantages to business when largeamounts of data are to be stored.
- These tools help in identifying more efficient ways of doing business.

Time Reductions :

- The high speed of tools like Hadoop and in-memory analytics can easily identify newsources of data which helps businesses analyzing data immediately.
- This helps us to make quick decisions based on the learnings.

Understand the market conditions :

- By analyzing big data we can get a better understanding of current market conditions.
- For example: By analyzing customers’ purchasing behaviours, a company can find out theproducts that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

Control online reputation :

- Big data tools can do sentiment analysis.Therefore, you can get feedback about who is saying what about your company.
- If you want to monitor and improve the online presence of your business, then big datatools can help in all this.

Using Big Data Analytics to Boost Customer Acquisition(purchase) and Retention :

- The customer is the most important asset any business depends on.
- No single business can claim success without first having to establish a solid customer base.
- If a business is slow to learn what customers are looking for, then it is very likely to deliverpoor quality products.

- The use of big data allows businesses to observe various customer-related patterns andtrends.

Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights :

- Big data analytics can help change all business operations.
- Like the ability to match customer expectations, changing
company’s product line, etc.
- And ensuring that the marketing campaigns are powerful.

Big Data Applications :

In today’s world big data have several applications, some of them are listed below :

Tracking Customer Spending Habit, Shopping Behavior :

In big retails stores, the management team has to keep data of customer’s spending habits, shopping behaviour, most liked product, which product is being searched/sold most, based on that data, the production/collection rate of that product gets fixed.

Recommendation :

By tracking customer spending habits, shopping behaviour, big retail stores provide recommendations to the customers.

Smart Traffic System :

Data about the condition of the traffic of different roads, collected through cameras, GPS devices placed in the vehicle.

All such data are analyzed and jam-free or less jam way, less time taking ways are recommended.

One more profit is fuel consumption can be reduced.

Secure Air Traffic System :

At various places of flight, sensors are present.

These sensors capture data like the speed of flight, moisture, temperature, and other environmental conditions.

Based on such data analysis, an environmental parameter within flight is set up and varied. By analyzing flight’s machine-generated data, it can be estimated how long the machine can operate flawlessly and when it can be replaced/repaired.

Auto Driving Car :

In the various spots of the car camera, a sensor is placed that gathers data like the size of the surrounding car, obstacle, distance from those, etc.

These data are being analyzed, then various calculations are carried out.
These calculations help to take action automatically.

Virtual Personal Assistant Tool :

Big data analysis helps virtual personal assistant tools like Siri, Cortana and Google Assistant to provide the answer to the various questions asked by users.

This tool tracks the location of the user, their local time, season, other data related to questions asked, etc.
Analyzing all such data provides an answer.
Example: Suppose one user asks "Do I need to take Umbrella?"The tool collects data like location of the user, season and weather condition at that location, then analyzes these data to conclude if there is a chance of raining, then provides the answer.

IoT :

Manufacturing companies install IOT sensors into machines to collect operational data. Analyzing such data, it can be predicted how long a machine will work without any problem when it requires repair.

Thus, the cost to replace the whole machine can be saved.

Education Sector Energy Sector :

Online educational courses conducting organization utilize big data to search candidates interested in that course.
If someone searches for a YouTube tutorial video on a subject, then an online or offline course provider organization on that subject sends an ad online to that person about their course.

Media and Entertainment Sector :

Media and entertainment service providing company like Netflix, Amazon Prime, Spotify do analysis on data collected from their users. Data like what type of video, music users are watching, listening to most,

how long users are spending on site, etc are collected and analyzed to set the next business strategy.

Big Data Features –security, compliance, auditing and protection

Features is the collective term which is used to guard the huge amount of data.

BIG DATA SECURITY :

Big data security is the collective term for all the measures and tools used to guard both the data and analytics **processes from attacks, theft, or other malicious activities that could harm or negatively affect them**

For companies that operate on the cloud, big data security challenges are multi-faceted.

When customers give their personal information to companies, they trust them with personal data which can be used against them if it falls into the wrong hands.

Best practices for strengthening Big Data security.

1. Encryption
2. User Access Control
3. Cloud Security Monitoring
4. Insider Threat Detection
5. User behavior analysis

BIG DATA COMPLIANCE :

Data compliance is the practice of ensuring that sensitive data is organized and managed in such a way as to enable organizations to meet enterprise business rules along with legal and governmental regulations.

Organizations that don’t implement these regulations can be fined up to tens of millions of dollars and even receive a 20-year penalty.

AUDITING AND PROTECTION:

- Reviewing & tracking data usage, access & changes over a period to ensure with policies & regulations.
- Auditors can use big data to expand the scope of their projects and draw comparisons overlarger populations of data.
- Role-based access audits
- login attempts & authentication audits.
- Detailed log – any change made to data
- Version history - record of all modifications made to data, documents, or files over time. It helps track changes, allowing users to view previous versions, compare updates, restore an older version if needed, and maintain accountability for edits.
- Big data also helps financial auditors to streamline the reporting process and detect fraud.
- These professionals can identify business risks in time and conduct more relevant andaccurate audits.

BIG DATA PROTECTION :

- Data Protection Regulation, if adopted, could improve the level of data protection for individuals in the context of big data analytics, in that it aims to increase the transparency of the processing, enhance the rights of data subjects and introduce a requirement for privacy by design and privacy impact assessments.
- Data protection is also important as organizations that don't implement these regulations can be fined up to tens of millions of dollars and even receive a 20-year penality.

BIG DATA PRIVACY AND ETHICS

Big Data privacy (or information privacy or data protection) is about access, use and collection of data, and the data subject's legal right to the data.

- **Personal Data Protection**
- **Data Minimisation** – collecting only data necessary for a specific purpose.
- **Informed Consent** – individuals understand & are informed about the purpose of data collection, processing & use of their personal data (fundamental principle – GDPR).
- **Data Security Measures** – methods to protect sensitive data (e.g., encryption, access control, authentication, audits).

Ethics

- **Fair & Bias** – It should make decisions without discrimination Eg.AI-driven recruitment tool algorithms (formal & informal bias).
- **Respecting Privacy** – individuals' privacy matters.
- **Data Minimisation** – avoid collecting excessive personal data
- Data privacy protection is complex due to **socio-techno risk**, a new security concern. This risk occurs with the abuse of technology that is used to store and process data.
- For example, **taking a company universal serial bus (USB) device home** for personal convenience runs the risk of breaching a company regulation that no company property shall leave company premises without permission. That risk becomes a data risk if the USB contains confidential corporate data.
- Organizations that don't implement these regulations can be fined up to tens of millions of dollars and even receive a 20-year penalty.

Big Data Analytics:

Big data analytics is a complex process of examining big data to uncover information, such as - hidden patterns, correlations, market trends and customer preferences.

- In simple terms, Big Data analytics is **the process of collecting, organizing, processing, and analyzing** huge volumes of data (Big Data).
- Data Analytics technologies and techniques give organizations a way to analyze data sets andgather new information.

  Examples

- **Amazon & Netflix** – Recommend products/movies based on user behavior.

- **Banks** – Detect fraud by analyzing unusual transaction patterns.
- **Google Maps** – Predicts traffic and suggests the fastest route.
- **Healthcare AI** – Identifies disease risks from patient records.
- **Facebook Ads** – Shows targeted ads based on browsing history.
- **Smart Cities** – Optimize public transport using real-time data.
- **Retail Stores** – Analyze purchase history to suggest related products

Several Organizations uses Big Data Analytics Examples **to generate various reports and dashboards** based on their huge current and past data sets in the form of Structured, Semi-structured or Unstructured.

Examples are

- **Fraud Management Report**
- **Live Tracking Report**
- **Sales Report and Future target Report**
- **Live Data Report**

Big Data Analytics Tools

- R programming
- Python
- Jupyter Notebook
- Apache Spark
- Splunk
- Tableau
- RapidMiner

Importance of Big Data Analytics :

- Organizations use big data analytics systems and software to make data-driven decisions that can improve business-related outcomes.
- The benefits include more effective marketing, new revenue opportunities, customer personalization and improved operational efficiency.

- With an effective strategy, these benefits can provide competitive advantages over rivals. Big Data Analytics tools also help businesses save time and money and aid in gaining insights to inform data-driven decisions.
- Big Data Analytics enables enterprises to narrow their Big Data to the most relevant information and analyze it to inform critical business decisions.

Challenges of conventional Systems:

Conventional systems refer to traditional methods used for data storage, processing, and analysis before the advent of modern frameworks.

- Uncertainty of Data Management
- Inability to handle unstructured data (RDBMS)
- Data processing speed
- Delayed batch processing
- Limited handling of complex queries
- Distributed queries, complex joins across large datasets
- Handling large volumes
- Cost & resource management
- Lack of flexibility
- Relying on fixed solutions
- Limited scalability-Usually on vertical scaling (adding more power to a single server)
- Little number of professionals
- Little number of professionals on pressure management from top

To overcome these challenges and drawbacks **Intelligent data analysis** is used

Intelligent Data Analysis(IDA):

- IDA refer to use of **advanced data techniques such as AI, ML & statistical methods** to extract meaningful insights from complex datasets.
- Helps address these challenges by automating decision-making, detecting patterns & generating actionable insights.
- IDA **automatically extracts online information, necessary knowledge** from online data inorder to make right choices

Stages of IDA

- Problem definition
    - Identifying business needs
    - Understanding data sources
- Data collection
    - DBs, APIs, web scraping, logs, IoT, sensor data, etc.
- Data preprocessing
    - Data cleaning
    - Data integration
    - Data transformation
- Exploratory Data Analysis (EDA)
    - Helps to understand patterns, trends & anomalies
- Model selection & training
- Model evaluation & optimization
- Deployment & integration
- Monitoring & maintenance
    - Performance tracking, logging & alerts

Nature of Data

- ➢ Introduction of data (explanation from previous same topic)
- ➢ Characteristics of data (explanation from previous same topic)
- ➢ Types of Data (explanation from previous same topic)
- ➢ Data Sources
    - Social Media Data
    - Sensor Data
    - Transactional Data
    - Web & Log Data
    - Scientific & Healthcare Data

- Data Representation & Storage
    - Data warehousing & Data Lakes
    - Traditional DBs
    - File Formats (CSV, JSON)
- Data Analytics Preprocessing
    - Issues in Raw Data (Noise, Redundancy, Incompleteness)
    - Data Cleaning & Transformation
- Data Privacy & Ethics (Explanation from previous same topics)

Analytic processes and tools

- Big Data analytics is **the process of collecting, organizing, processing, and analyzing** huge volumes of data (Big Data).
- **Phases of Big Data Analytics Processing**

a) **Data Collection**

- Sources of Big Data – Social Media, Sensors, IoT, Web logs, etc.
- Data Ingestion Methods – (Batch Processing, Stream Processing)

b) **Data Storage**

- Big Data Storage – HDFS, NoSQL
- Data Warehouses & Delta Lake

c) Data Processing

- **Batch Processing** (MapReduce, Apache Spark)
- **Stream Processing** (Apache Kafka, Apache Flink, Storm)
- **ETL** (Extract, Transform, Load) pipelines

d) Data Analysis & Insights Generation

- Descriptive
- Diagnostic
- Predictive
- Prescriptive

d) Data Visualization

- **Graphical representation of data** using charts, graphs, maps & dashboards
- **Helps in transforming complex datasets** into an easily understandable format

Analysis vs reporting

Analysis	**Report**
taking the organized data & analyzing it	Process of organising data
Interpretation of data	Representation of data
Transforms data & information into insights.	Translates raw data into information
Analysis shows - why it is happening, and what we can do	Report explains -what happened
Process: • Exploring • Questioning • Interpreting	**Process:** • Organizing • Promoting
Example: A financial analyst examines company performance to predict future trends.	**Example:** A financial report represents revenue, expenses & profit over time.

Modern data analytic tools

Modern analytic tools are advanced software platforms and frameworks designed to process, analyze, and visualize large-scale data.

Tools – BD, ML, AI & Cloud computing to extract meaningful insights.

Importance

- Handle large volumes of data.
- Enable real-time & predictive analytics.
- Provide interactive visualizations for decision-making.
- Support cloud computing & distributed processing.

Examples

BD Processing Tools:

- Apache Hadoop – Distributed storage & batch processing of large datasets.
- Apache Spark – Fast in-memory data processing for large-scale analytics**.**

Cloud-Based Analytics

- **Google BigQuery** – Warehouse for SQL-based analytics
- **Amazon Redshif-** It is a fully managed, petabyte-scale data warehouse service by AWS. It is optimized for fast querying and analytics using SQL.
- **Snowflake-** Snowflake is a cloud-native data platform designed for analytics, offering a flexible and scalable architecture.

ML & AI for Analytics

- **TensorFlow & PyTorch** – Frameworks for AI-driven analytics
- **H2O.ai** – AutoML for predictive modeling
- **Databricks**

Business Intelligence & Data Visualizations

- **Tableau** – Interactive data visualization & dashboarding
- **Microsoft Power BI** – Enterprise reporting & analytics
- **Google Cloud** – Advanced data modeling & visualization

X-------------------X