Module – 4

WHAT IS BIG DATA

- As Gartner defines it – "Big Data are high volume, high velocity, or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization."
- The term 'big data' is self-explanatory − a collection of huge data sets that normal computing techniques cannot process.
- The term not only refers to the data, but also to the various frameworks, tools, and techniques involved.
- Technological advancement and the advent of new channels of communication (like social networking) and new, stronger devices have presented a challenge to industry players in the sense that they have to find other ways to handle the data.
- Big data is an all-inclusive term, representing the enormous volume of complex data sets that companies and governments generate in the present-day digital environment.
- Big data, typically measured in petabytes or terabytes, materializes from three major sources— transactional data, machine data, and social data.

The Difference Between Traditional Data and Big Data

Traditional Data	Big Data
Traditional data is generated in enterprise level.	Big data is generated outside the enterprise level.
Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
Traditional database system deals with structured data.	Big data system deals with structured, semi-structured,database, and unstructured data.
Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.
Data integration is very easy.	Data integration is very difficult.
Normal system configuration is capable to process traditional data.	High system configuration is required to process big data.
The size of the data is very small.	The size is more than the traditional data size.
Traditional data base tools are required to perform any data base operation.	Special kind of data base tools are required to perform any database schema based operation.
Normal functions can manipulate data.	Special kind of functions can manipulate data.
Its data model is strict schema based and it is static.	Its data model is a flat schema based and it is dynamic.

Traditional Data	Big Data
Traditional data is stable and inter relationship.	Big data is not stable and unknown relationship.
Traditional data is in manageable volume.	Big data is in huge volume which becomes unmanageable.
It is easy to manage and manipulate the data.	It is difficult to manage and manipulate the data.
Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc.	Its data sources includes social media, device data, sensor data, video, images, audio etc.

TYPES OF BIG-DATA

Big Data is generally categorized into three different varieties. They are as shown below:

- **Structured Data**
- **Semi-Structured Data**
- **Unstructured Data**

TYPES OF BIG-DATA

Structured Data owns a dedicated data model, it also has a well-defined structure, it follows a consistent order and it is designed in such a way that it can be **easily accessed** and used by a person or a computer. Structured data is usually stored in well-defined columns and also Databases.

Example: Database Management Systems (DBMS)

Semi-Structured Data can be considered as another form of Structured Data. It inherits a few properties of Structured Data, but the major part of this kind of data fails to have a

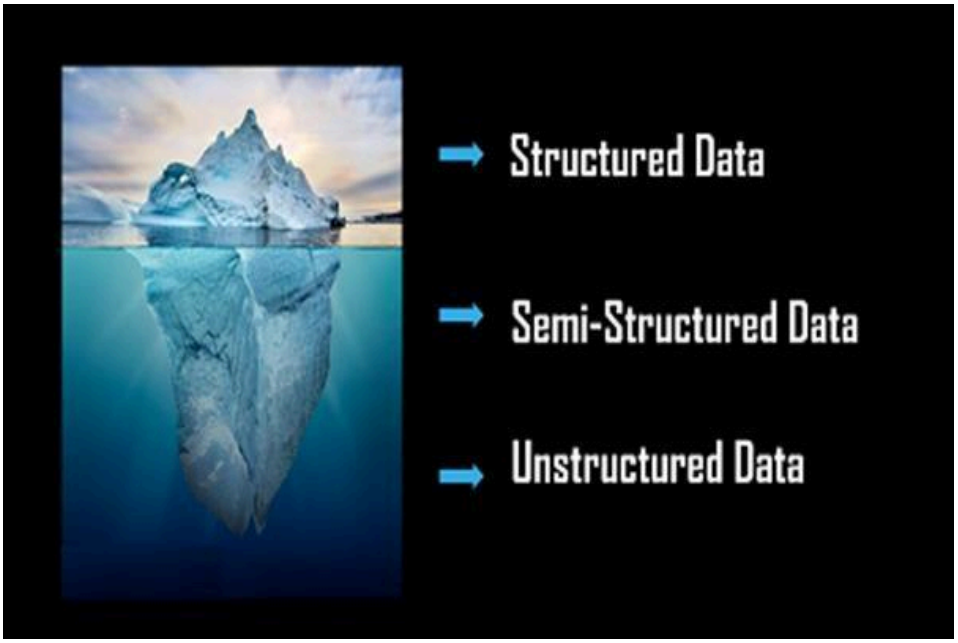
definite structure and also, it does not obey the formal structure of data models such as an RDBMS.

Example: Comma Separated Value (CSV) Files

Unstructured Data is completely a different type of which neither has a structure nor obeys to follow the formal structural rules of data models. It does not even have a consistent format and it found to be varying all the time. But, rarely it may have information related to data and time.

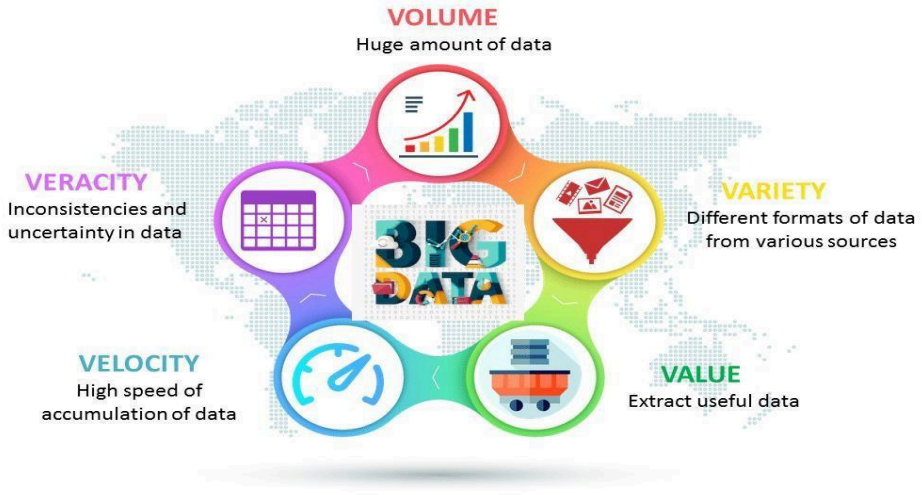
Example: Audio Files, Images etc

TYPES OF BIG-DATA



THE CHARACTERISTICS OF BIG DATA

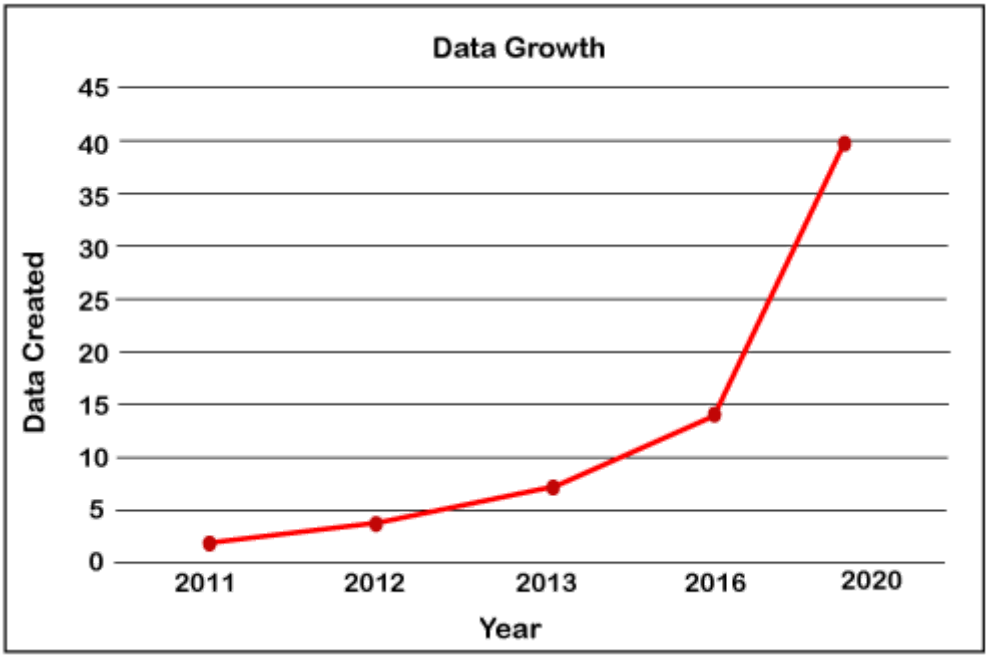
- Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit.
- It is used by many **multinational companies** to **process** the data and business of many **organizations**.
- The data flow would exceed **150 exabytes** per day before replication.



THE CHARACTERISTICS OF BIG DATA

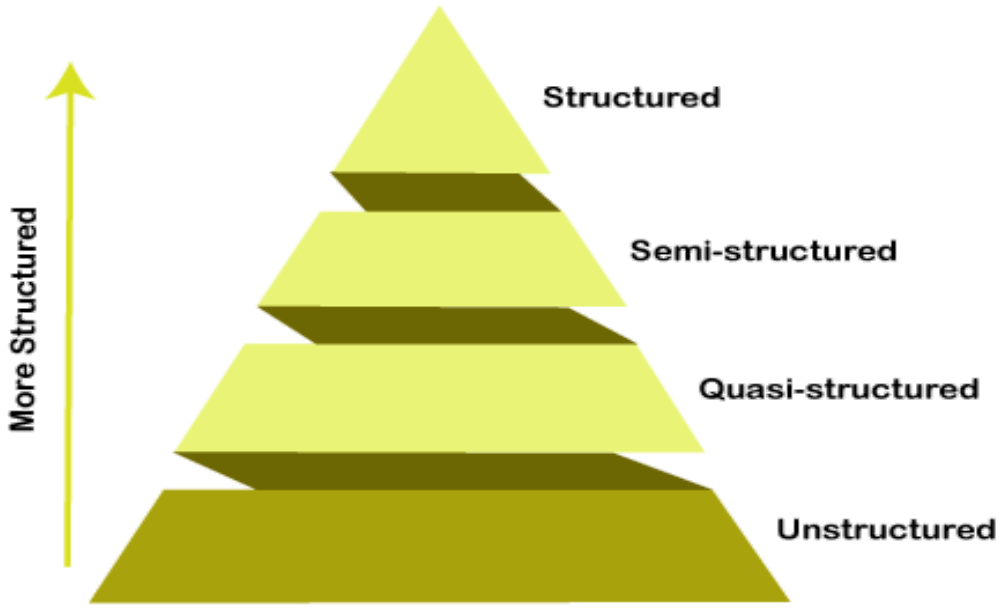
Volume

- Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using **distributed systems**, to store data in several locations and brought together by a software Framework like **Hadoop**.
- Facebook alone can generate about **billion** messages, **4.5 billion** times that the "like" button is recorded, and over **350 million** new posts are uploaded **each day.** Such a huge amount of data can only be handled by Big Data Technologies



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.



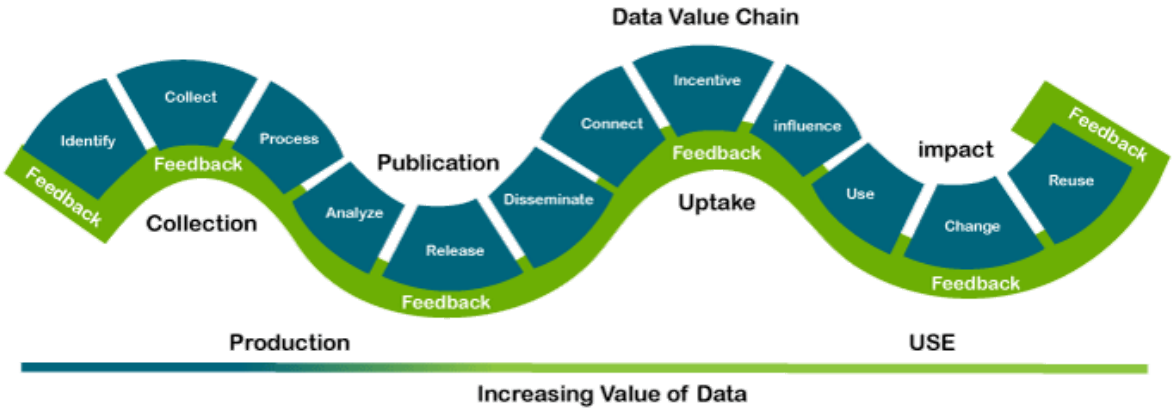
- **Structured data:** It is in a tabular form. Structured Data is stored in the relational database management system.
- **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.
- **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data.
- **Quasi-structured Data:** The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

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

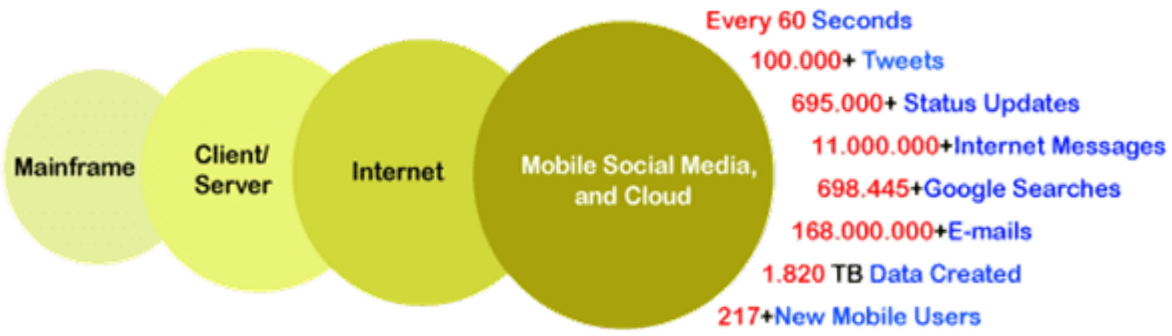
- It is not the data that we process or store.
  It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.



Velocity

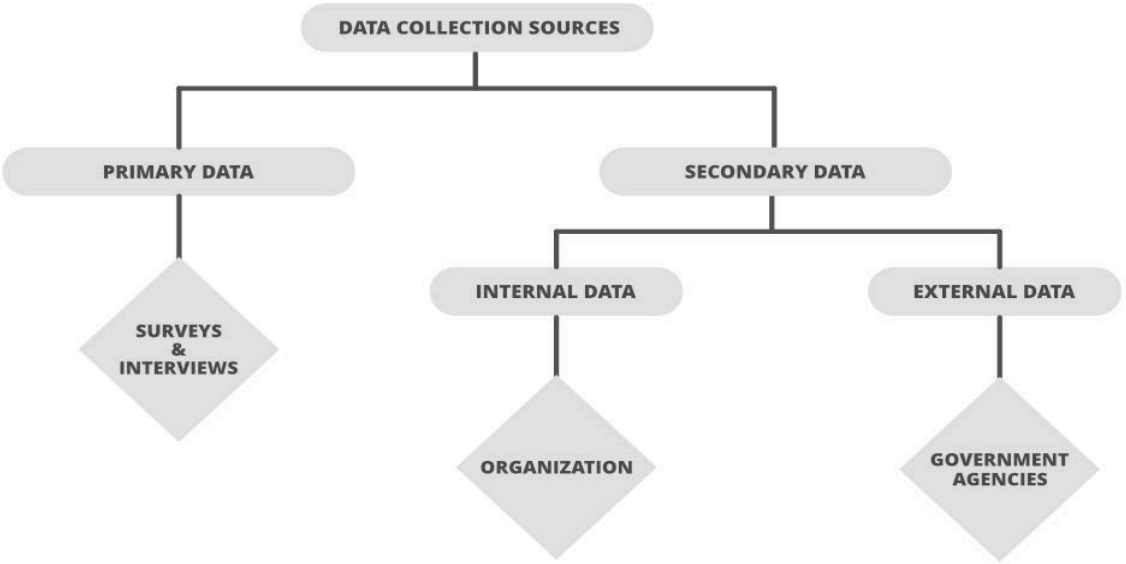
- Velocity creates the speed by which the data is created in **real-time**.

- 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.



DIFFERENT SOURCES OF DATA GENERATION

- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.
- In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data.
- The data which is to be analyzed must be collected from different valid sources.



Primary data:

- The data which is Raw, original, and extracted directly from the official sources is known as primary data.
- This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys.
- The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

Interview method:

- The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee.
- Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing.
- These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

Survey method:

- The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video.
- The survey method can be obtained in both online and offline mode like through website forms and email.

- Then that survey answers are stored for analyzing data.
- Examples are online surveys or surveys through social media polls.

Observation method:

- In this method, the data is collected directly by posting a few questions on the participants.
- For example, observing a group of customers and their behavior towards the products.
- The data obtained will be sent for processing.

Experimental method:

- The experimental method is the process of collecting data through performing experiments, research, and investigation.

Secondary data:

- Secondary data is the data which has already been collected and reused again for some valid purpose.
- This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source:

- These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc.
- The cost and time consumption is less in obtaining internal sources.

External source:

- The data which can't be found at internal organizations and can be gained through external third party resources is external source data.
- The cost and time consumption is more because this contains a huge amount of data.
- Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

Other sources:

- **Sensors data:** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.

- **Web traffic:** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis.

UNDERSTANDING RDBMS AND WHY IT IS FAILING TO STORE BIG DATA

RDBMS:

- **RDBMS** stands for **Relational Database Management Systems.**
- A database is an organized collection of data stored in a computer system and usually controlled by a database management system (DBMS).
- The data in common databases is modeled in tables, making querying and processing efficient.

What is RDBMS?

- RDBMS stands for **Relational Database Management Systems**.
- It is a program that allows us to create, delete, and update a relational database.
- A Relational Database is a database system that stores and retrieves data in a tabular format organized in the form of rows and columns.
- It is a smaller subset of DBMS which was designed by E.F Codd in the 1970s.
- The major DBMSs like SQL, My-SQL, and ORACLE are all based on the principles of relational DBMS.
- Relational DBMS owes its foundation to the fact that the values of each table are related to others.
- It has the capability to handle larger magnitudes of data and simulate queries easily.

Why Traditional DBMS fails to supports Big Data?

1. **Scalability Limitations** – RDBMS designed to handle structured data only.

2. **Schema Rigidity** – RDBMS requires predefined schema, Inflexible for handling semi-structured or unstructured data.

3. **Data Size Limitations** – RDBMS has limitations on data size

4. **Performance Issues** – RDBMS can experience performance degradation as data volume increase.

Big data requires flexible schema, high scalability and high performance, making NoSQL database like Hbase, Cassandra, and MongoDB more suitable for big data storage.

Three Big Data Challenges

- Big data has many qualities—it's unstructured, dynamic, and complex.
- Big data is big. Humans and IoT sensors are producing trillions of gigabytes of data each year.

- It is modern data, in an increasingly diverse range of formats and from an ever-broader variety of sources.

1. Big Data Is Too Big for Traditional Storage

- Perhaps the most obvious of the big data challenges is its enormous scale. We typically measure it in petabytes (so that's 1,024 terabytes or 1,048,576 gigabytes).
- To give you an idea of how big data can get, here's an example: Facebook users upload at least 14.58 million photos per hour. Each photo garners interactions stored along with it, such as likes and comments. Users have "liked" at least a trillion posts, comments, and other data points.
- But it's not just tech giants like Facebook that are storing and analyzing huge quantities of data.
- Even a small business taking a slice of social media information—for example, to see what people are saying about its brand—requires high-capacity data storage architecture.

2. Big Data Is Too Complex for Traditional Storage

- A relational database—the type of database that stores traditional data—consists of records containing clearly defined fields.
- You can access this type of database using a relational database management system (RDBMS) such as MySQL, Oracle DB, or SQL Server.
- A relational database can be relatively large and complex: It may consist of thousands of rows and columns.
- But crucially, with a relational database, you can access a piece of data by reference to its relation to another piece of data.
- Big data doesn't always fit neatly into the relational rows and columns of a traditional data storage system.
- It's largely unstructured, consisting of myriad file types and often including images, videos, audio, and social media content.

3. Big Data Is Too Fast for Traditional Storage

- Traditional data storage systems are for steady data retention. You can add more data regularly and then perform analysis on the new data set. But big data grows almost instantaneously, and analysis often needs to occur in real time.
- Take sensor data, for example. **Internet of things (IoT)** devices need to process large amounts of sensor data with minimal latency. Sensors transmit data from the "real world" at a near-constant rate. Traditional storage systems struggle to store and analyze data arriving at such a velocity.
- Example: **cybersecurity.** IT departments must inspect each packet of data arriving through a company's firewall to check whether it contains suspicious code. Many gigabytes might be passing through the network each day. To avoid falling victim to cybercrime, analysis must occur instantaneously—storing all the data in a table until the end of the day is not an option.

USE CASES OF BIG-DATA ANALYTICS

Real-Time Big Data Use Cases Across Industries

- Retailers analyze big data to understand customer preferences and buying patterns, enabling targeted marketing campaigns and personalized recommendations.
- Healthcare organizations leverage big data to improve patient outcomes by identifying trends, predicting disease outbreaks, and optimizing treatment plans based on large-scale data analysis.
- Financial institutions utilize big data to detect fraudulent activities, manage risk, and make data-driven investment decisions.
- Manufacturing firms employ big data to optimize production processes, reduce downtime, and predict maintenance needs, resulting in increased productivity and reduced costs.
- Government agencies utilize big data for policy-making, urban planning, and resource allocation, enabling evidence-based decision-making and improving public services.
- Energy companies leverage big data to optimize energy generation and distribution, identify consumption patterns, and promote energy efficiency.

Big Data Use Cases in Healthcare

- **Predictive Analytics**
- Big data analytics is used to analyze vast amounts of patient data, including electronic health records (EHRs), genomic data, and real-time monitoring data, to predict disease outcomes and identify patients at high risk of developing certain health conditions.
- This enables healthcare providers to take early actions and offer personalized healthcare plans, leading to better patient treatment outcomes.
- For instance, analyzing data from wearable devices to predict health issues, such as heart attacks or failures, allows for timely interventions.

Personalized Medicine

- Big data enables personalized medicine, which includes personalizing medical treatments based on an individual's unique genetic profile, lifestyle, and other factors.
- By analyzing large datasets of genomic data, clinical data, and other relevant information, big data is helping healthcare providers to identify targeted treatments for patients with complex medical conditions, such as cancer, cardiovascular diseases, rare genetic disorders, etc.
- For instance, medical care facilities can use genomic data to identify targeted treatment alternatives for cancer patients based on their genetic mutations.

Telemedicine and Remote Patient Monitoring

- Big data facilitates telemedicine and remote patient monitoring, allowing healthcare providers to monitor patients' health conditions and collect real-time data remotely.
- Big data analytics can be used to analyze this and other patient data to find patterns and trends, allowing the early identification of possible health risks and timely treatment.
- For instance, hospitals may offer virtual consultations and follow-up treatment for patients with chronic diseases, reducing hospital visits and enhancing patient outcomes.
- Hospitals can also employ telemedicine to provide mental health treatments in far-off places, enhancing underprivileged people's access to healthcare.

Health Data Analytics

- Big data analytics is helping healthcare organizations analyze large volumes of data to gain valuable business insights into population health patterns, disease prevalence, and treatment efficiency.
- Healthcare centers can use this data to create evidence-based treatment guidelines, allocate resources more effectively, and assist public health activities like disease surveillance and outbreak control.
- For instance, medical centers can analyze population health data to identify trends and patterns, enabling healthcare officials to develop targeted interventions to prevent disease outbreaks.

Big Data Use Cases in Retail

- The retail sector has increasingly used big data analytics to obtain valuable business insights and improve business processes, including customer experiences, inventory management, pricing strategies, and supply chain management.
- For instance, Amazon, the biggest online retailer in the world, utilizes big data to analyze customer information and behavior, including browsing and purchase history, to tailor the shopping experience for each customer.
- Amazon also uses big data to optimize its supply chain management, accurately forecasting demand and optimizing inventory levels to reduce costs and ensure timely deliveries.
- By leveraging big data, retailers like Amazon can gain a competitive edge and deliver a better customer experience.

BIG DATA FOR DECISION MAKING IN MARKETING:

Big data analytics can help improve marketing decision-making in a number of ways, including:

- **Better customer understanding**

Big data analytics can help marketers understand their customers' preferences, behaviors, and habits. This can help marketers personalize messages and offers to increase customer engagement and loyalty.

- **Identifying target audiences**

Big data analytics can help marketers identify their ideal target audiences through in-depth consumer analysis.

- **Optimizing marketing spend**

Big data analytics can help marketers optimize their spending budget by targeting only valuable customers.

- **Reducing customer churn**

Big data analytics can help marketers identify customers at risk of leaving and target them with personalized offers and messages.

- **Improving customer experience**

Big data analytics can help marketers identify areas for improvement related to customer experience and make changes based on this information.

- **Real-time campaign performance**

Big data analytics can help marketers see what campaigns are driving brand awareness, qualified leads, and customer conversion. This can help marketers optimize campaigns to improve metrics and drive measurable business results.

- **Identifying trends**

Big data analytics can help marketers detect new trends, customer preferences, and competitive threats.

BIG DATA IN REAL WORLD

Big data can be used in many real-world applications to help businesses make better decisions, improve efficiency, and understand their customers:

- **Predictive maintenance**

In manufacturing, big data can help predict equipment failure and remaining life of systems and components. This can help manufacturers maximize uptime and deploy maintenance more cost effectively.

- **Product recommendations**

Companies like Amazon use big data to learn what their customers want and like, and then recommend products to them.

- **Sentiment analysis**

Big data can be used to provide real-time sentiment analysis on tennis matches for TV, mobile, and web users.

- **Music recommendations**

Spotify uses big data analytics to collect data from its users and then use that data to make music recommendations.

- **Education**

Schools, colleges, and technology providers are using big data to enhance the educational experience.

- **Healthcare**

Big data is used in healthcare for a number of purposes, including predicting epidemic outbreaks, early symptom detection, and prediction and prevention of serious medical conditions.

- **Cost savings**

Big data can help businesses pinpoint ways to enhance operational efficiency, such as analyzing energy use.