BDA Unit - 1 - notes

Big Data Analytics (Bharath Institute of Higher Education and Research)

UNIT - I

Big Data – Definition, Characteristic Features – Big Data Applications - Big Data vs Traditional Data - Risks of Big Data - Structure of Big Data - Challenges of Conventional Systems– Evolution of Analytic Scalability - Evolution of Analytic Processes, Tools and methods - Analysis vs Reporting - Modern Data Analytic Tools.

Introduction to Big Data

In today's digital world, a huge amount of data is generated every second from various sources such as social media, mobile devices, sensors, online transactions, and cloud applications. This data is very large in size, complex in nature, and continuously growing. Traditional data processing tools and database systems are unable to handle such massive data efficiently. This led to the emergence of **Big Data** technologies. Big Data helps organizations store, process, and analyze large datasets to extract useful information and support decision-making.

Definition of Big Data

Big Data can be defined as a collection of large and complex datasets that cannot be processed using traditional database management systems. It involves data that is high in **volume**, generated at high **velocity**, and comes in a wide **variety** of formats. Advanced tools and techniques are required to store, manage, and analyze Big Data effectively.

What is Big Data?

According to Gartner, the definition of Big Data – "Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." This definition clearly answers the "What is Big Data?" question – Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

Characteristics of Big Data (5 V's)

Volume

Volume refers to the enormous amount of data generated every day. Data is produced from sources like social media platforms, sensors, emails, videos, and transaction records. The size of data ranges from terabytes to petabytes and even zettabytes. Handling such huge volumes of data is one of the major challenges of Big Data.

Velocity

Velocity indicates the speed at which data is generated, collected, and processed. In many applications, data must be processed in real time or near real time. Examples include stock market data, online banking transactions, and live social media updates.

Variety

Variety refers to different forms of data. Data can be structured, semi-structured, or unstructured. Structured data is stored in tables, semi-structured data includes XML and JSON, and unstructured data includes text documents, images, videos, and audio files.

Veracity

Veracity deals with the quality and reliability of data. Big Data may contain noise, inconsistency, and missing values. Ensuring accurate and trustworthy data is essential for effective analysis and correct decision-making.

Value

Value refers to the useful insights and benefits obtained from Big Data analysis. The main purpose of Big Data is to extract meaningful information that adds value to businesses and society.

Structure of Big Data

The structure of Big Data refers to the way data is organized and stored based on its format and nature. Unlike traditional data, Big Data is not limited to structured formats. It includes a wide range of data types collected from different sources such as databases, social media, sensors, mobile devices, and multimedia systems. Based on its structure, Big Data is classified into three main categories: **structured data**, **semi-structured data**, and **unstructured data**.

Structure of Big Data

- **Structured Data**: Structured data is data that is organized in a specific format, such as numbers and text in a spreadsheet. It is easy to store and analyze, and it can be used to generate reports and charts. Examples of structured data include customer data in a CRM system, financial data in an accounting system, and inventory data in an ERP system.

- **Unstructured Data**: Unstructured data is data that is not organized in a specific format. It can include images, videos, audio files, and text. Unstructured data can be difficult to store and analyze, but it can provide valuable insights when analyzed correctly. Examples of unstructured data include social media posts, customer reviews, and sensor data from IoT devices.

Unstructured Data

- **Semi-Structured Data:** Semi-Structured data is data that has some structure but not as much as structured data. It includes XML, JSON, and other similar formats. Semi-structured data can be easily analyzed and can provide valuable insights when combined with structured data. Examples of semi-structured data include email data and log data.

Semi-structured Data

Applications of Big Data

Big Data is widely used in various fields. In healthcare, it helps in disease prediction, patient monitoring, and medical research. In banking and finance, Big Data is used for fraud detection, credit analysis, and risk management. In retail and e-commerce, it supports recommendation systems and customer behavior analysis. In education, Big Data enables learning analytics and student performance tracking. Social media platforms use Big Data for sentiment analysis and trend detection. Smart cities use Big Data for traffic control, energy management, and public safety.

Big Data has become an integral part of many industries due to its ability to process and analyze large volumes of data efficiently. By extracting meaningful insights from massive datasets, Big Data helps organizations improve performance, reduce costs, and make better decisions. Some of the major applications of Big Data are explained below.

In the **healthcare sector**, Big Data is used to analyze patient records, medical images, and clinical data. It helps in early disease detection, personalized treatment, and monitoring patient health. Big Data analytics also supports medical research and drug discovery by analyzing large datasets related to genetics and clinical trials.

In **banking and finance**, Big Data plays a crucial role in fraud detection and risk management. Financial institutions analyze transaction data in real time to detect suspicious activities. Big Data is also used for credit scoring, customer profiling, and investment analysis, helping banks make informed financial decisions.

In **retail and e-commerce**, Big Data is widely used to study customer behavior and preferences. Online retailers use Big Data to provide personalized product recommendations, manage inventory, and optimize pricing strategies. It also helps in analyzing customer feedback and improving customer satisfaction.

In the field of **education**, Big Data supports learning analytics by tracking student performance, attendance, and learning patterns. Educational institutions use Big Data to improve teaching methods, predict student outcomes, and provide personalized learning experiences.

Social media platforms generate massive amounts of data every second. Big Data is used to perform sentiment analysis, identify trends, and understand user behavior. Companies use this data for targeted advertising and brand monitoring.

In **smart cities**, Big Data is used to manage traffic, energy consumption, and public services. Data collected from sensors and surveillance systems helps improve urban planning, reduce congestion, and enhance public safety.

In the **manufacturing industry**, Big Data is used for predictive maintenance, quality control, and supply chain optimization. Analyzing machine and sensor data helps reduce downtime and improve production efficiency.

Big Data vs Traditional Data

Traditional data systems handle structured and limited data, whereas Big Data systems handle massive and diverse datasets. Traditional systems use centralized storage and sequential processing, while Big Data uses distributed storage and parallel processing. Big Data systems are highly scalable compared to traditional database systems.

Traditional data refers to structured data that is stored and processed using conventional database management systems such as relational databases. This type of data is usually limited in size and is generated at a relatively slow rate. Traditional data processing systems use centralized storage and sequential processing techniques. These systems work efficiently when the volume of data is small and well structured, but they fail to perform effectively when the data size increases rapidly.

Big Data, on the other hand, refers to extremely large, complex, and diverse datasets that are generated at very high speed. Big Data includes structured, semi-structured, and unstructured data such as text, images, videos, sensor data, and social media content. Due to its large size and complexity, Big Data cannot be processed using traditional database systems. It requires distributed storage and parallel processing techniques to handle the data efficiently.

In traditional data systems, scalability is limited because increasing data volume requires upgrading hardware, which is costly and time-consuming. Big Data systems are highly scalable and use distributed computing frameworks where storage and processing are spread across multiple machines. This allows the system to handle massive amounts of data efficiently.

Traditional data analytics focuses mainly on reporting and historical analysis. In contrast, Big Data analytics supports real-time processing, advanced analytics, and predictive analysis using machine learning and artificial intelligence techniques. Big Data systems also provide better fault tolerance compared to traditional systems.

Thus, Big Data differs from traditional data in terms of size, structure, processing method, scalability, and analytical capability, making it more suitable for modern data-driven applications.

Differences between Big Data and Traditional Data Processing

Parameters	Big Data	Data Processing
Data Volume	Massive, often terabytes to petabytes or more	Moderate to large, typically in gigabytes
Data Variety	Diverse, including structured, unstructured, and semi-structured data from various sources such as social media, sensors, etc.	Mainly structured data from traditional sources like databases and spreadsheets
Data Velocity	High velocity, often generated and processed in real-time or near real-time	Lower velocity, data is processed in batch mode
Data Structure	Often lacks a predefined structure, may require schema-on-read approach	Structured, with well-defined schemas
Storage Infrastructure	Requires distributed storage systems like Hadoop Distributed File System (HDFS)	Relational databases or file systems
Processing Framework	Utilizes parallel processing frameworks like Apache Spark, Hadoop MapReduce	Traditional databases or data warehouses
Scalability	Highly scalable, can easily scale out to handle increasing data loads	Limited scalability, often requires upgrading hardware or software
Analytics	Enables advanced analytics like predictive modeling, machine learning, and AI	Limited to basic analytics and reporting
Cost	Can be cost-effective due to the use of commodity hardware and open-source software	Often involves significant upfront costs for hardware, software, and licensing
Flexibility	Offers flexibility in handling various data formats and types	Limited flexibility, primarily designed for specific data formats and types
Fault Tolerance	Built-in fault tolerance mechanisms ensure resilience to hardware failures	Relies on redundancy and backup systems for fault tolerance
Real-time Processing	Capable of real-time data processing and analysis	Generally not optimized for real-time processing

Risks of Big Data

Here are the five biggest risks that big data presents for digital enterprises.

Unorganized data

Big data is highly versatile. It comes from number of sources and in number of forms. There's structured data, there's unstructured data. There's data coming from online and offline sources. And all this data keeps piling up each day, each minute. It's overwhelming for enterprises to tackle such unorganized and siloed data sets effectively. A well planned governance strategy can bring you out of your dark data and help you make sense of it.

Data storage and retention

This is one of the most obvious risks associated with big data. When data gets accumulated at such a rapid pace and in such huge volumes, the first concern is its storage. Traditional data storage methods and technology are just not enough to store big data and retain it well. Enterprises today need a shift to cloud based data storage solutions to store, archive and access big data effectively.

Cost management

The process of storing, archiving, analyzing, reporting and managing big data involves costs. Many small and medium enterprises think that big data is only for big businesses, and they cannot afford it. However, with careful budgeting and planning of resources, big data costs can be mitigated well. Once the initial set up, migration and overhauling costs are taken care of, big data acts as an incredible revenue generator for digital enterprises.

Incompetent analytics

Without proper analytics, big data is just a pile of trash lying unnecessarily in your organization. Analytics is what makes data meaningful, giving management valuable insights to make business decisions and plan strategies for growth. With data growing at such an alarming rate, there's obviously a lack of skilled professionals and technology to analyze big data efficiently. It exposes enterprises to the risk of misinterpretation of data, and wrong decision making. Hiring the right talent and applying the right tools is crucial to make relevant decisions from a big data project.

Data privacy

With big data, comes the biggest risk of data privacy. Enterprises worldwide make use of sensitive data, personal customer information and strategic documents. When there's so much confidential data lying around, the last thing you want is a data breach at your enterprise. A security incident can not only affect critical data and bring down your reputation; it also leads to

legal actions and heavy penalties. Taking measures for data privacy is not just a good initiative anymore, it’s a compliance necessity.

Challenges of Conventional Systems

Big data has revolutionized the way businesses operate, but it has also presented a number of challenges for conventional systems. Here are some of the challenges faced by conventional systems in handling big data:

Big data is a term used to describe the large amount of data that can be stored and analyzed by computers. Big data is often used in business, science and government. Big Data has been around for several years now, but it's only recently that people have started realizing how important it is for businesses to use this technology in order to improve their operations and provide better services to customers. A lot of companies have already started using big data analytics tools because they realize how much potential there is in utilizing these systems effectively.

However, while there are many benefits associated with using such systems - including faster processing times as well as increased accuracy - there are also some challenges involved with implementing them correctly.

Challenges of Conventional System in big data
- Scalability
- Speed
- Storage
- Data Integration
- Security

Scalability
A common problem with conventional systems is that they can't scale. As the amount of data increases, so does the time it takes to process and store it. This can cause bottlenecks and system crashes, which are not ideal for businesses looking to make quick decisions based on their data. Conventional systems also lack flexibility in terms of how they handle new types of information--for example, if you want to add another column (columns are like fields) or row (rows are like records) without having to rewrite all your code from scratch.

Speed
Speed is a critical component of any data processing system. Speed is important because it allows you to process and analyze your data faster, which means you can make better-informed decisions about how to proceed with your business. Make more accurate predictions about future events based on past performance.

Storage
The amount of data being created and stored is growing exponentially, with estimates that it will reach 44 zettabytes by 2020. That's a lot of storage space! The problem with conventional

systems is that they don't scale well as you add more data. This leads to huge amounts of wasted storage space and lost information due to corruption or security breaches.

Data Integration

The challenges of conventional systems in big data are numerous. Data integration is one of the biggest challenges, as it requires a lot of time and effort to combine different sources into a single database. This is especially true when you're trying to integrate data from multiple sources with different schemas and formats. Another challenge is errors and inaccuracies in analysis due to lack of understanding of what exactly happened during an event or transaction. For example, if there was an error while transferring money from one bank account to another, then there would be no way for us to know what actually happened unless someone tells us about it later on (which may not happen).

Security

Security is a major challenge for enterprises that depend on conventional systems to process and store their data. Traditional databases are designed to be accessed by trusted users within an organization, but this makes it difficult to ensure that only authorized people have access to sensitive information. Security measures such as firewalls, passwords and encryption help protect against unauthorized access and attacks by hackers who want to steal data or disrupt operations. But these security measures have limitations: They're expensive; they require constant monitoring and maintenance; they can slow down performance if implemented too extensively; and they often don't prevent breaches altogether because there's always some way around them (such as through phishing emails).

Conventional systems are not equipped for big data. They were designed for a different era, when the volume of information was much smaller and more manageable. Now that we're dealing with huge amounts of data, conventional systems are struggling to keep up. Conventional systems are also expensive and time - consuming to maintain; they require constant maintenance and upgrades in order to meet new demands from users who want faster access speeds and more features than ever before.

1.7 Evolution of analytics scalability

In analytic scalability, we have to pull the data together in a separate analytics environment and then start performing analysis.

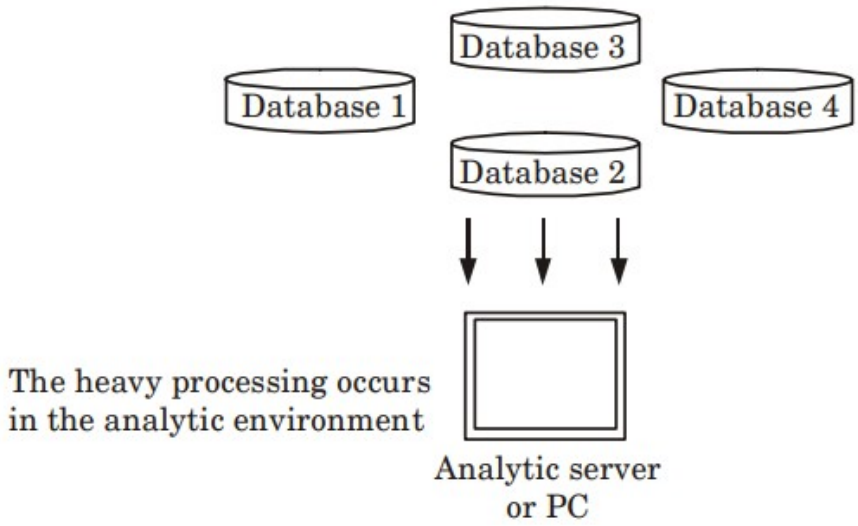
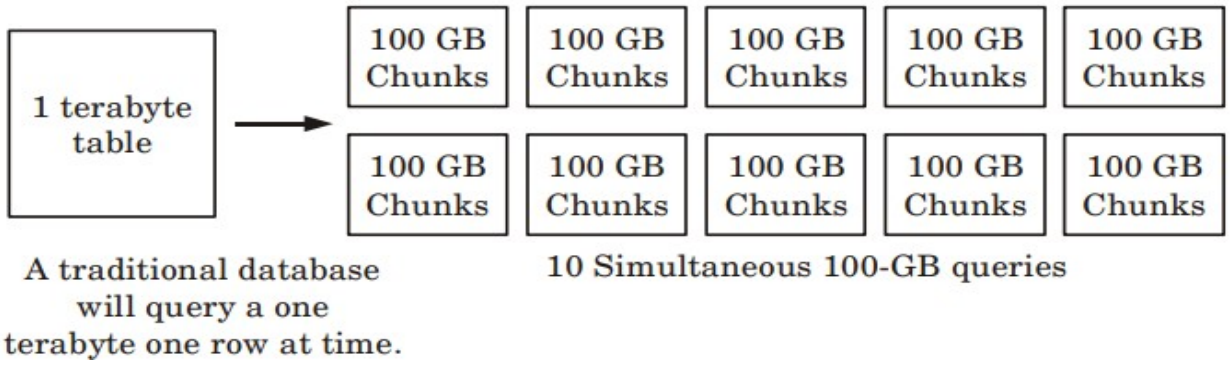


Fig 1.8 Evolution of analytics scalability

Analysts do the merge operation on the data sets which contain rows and columns. The columns represent information about the customers such as name, spending level, or status. In merge or join, two or more data sets are combined together. They are typically merged/joined so that specific rows of one data set or table are combined with specific rows of another.

Analysts also do data preparation. Data preparation is made up of joins, aggregations, derivations, and transformations. In this process, they pull data from various sources and merge it all together to create the variables required for analysis. The massively Parallel Processing (MPP) system is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.



MPP systems build in redundancy to make recovery easy.

MPP systems have resource management tools:

- Manage the CPU and disk space

- Query optimizer

Evolution of analytic process

- With increased level of scalability, it needs to update analytic processes to take advantage of it.
- This can be achieved with the use of analytical sandboxes to provide analytic professionals with a scalable environment to build advanced analytics processes.
- One of the uses of MPP database system is to facilitate the building and deployment of advanced analytic processes.
- An analytic sandbox is the mechanism to utilize an enterprise data warehouse.
- If used appropriately, an analytic sandbox can be one of the primary drivers of value in the world of big data.

Analytical sandbox

- An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions.
- An analytic sandbox is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping.
- Once things progress into ongoing, user-managed processes or production processes, then the sandbox should not be involved.
- A sandbox is going to be leveraged by a fairly small set of users.
- There will be data created within the sandbox that is segregated from the production database.
- Sandbox users will also be allowed to load data of their own for brief time periods as part of a project, even if that data is not part of the official enterprise data model.

1.8 Tools and methods

TOOLS

- Big Data Analytics is the process of collecting large chunks of structured/unstructured data, segregating and analyzing it and discovering the patterns and other useful business insights from it.
- These days, organizations are realizing the value they get out of big data analytics and hence
- They are deploying big data tools and processes to bring more efficiency in their work environment.
- Many big data tools and processes are being utilized by companies these days in the processes of discovering insights and supporting decision making.

- Big data processing is a set of techniques or programming models to access large- scale data to extract useful information for supporting and providing decisions.

There are hundreds of **data analytics tools** out there in the market today but the selection of the right tool will depend upon your business **NEED, GOALS, and VARIETY** to get business in the right direction.

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. It can process both structured and unstructured data from one server to multiple computers. Hadoop also offers cross-platform support for its users. Today, it is the best big data analytic tool and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

Features of Apache Hadoop:

- Free to use and offers an efficient storage solution for businesses.
- Offers quick access via HDFS (Hadoop Distributed File System).
- Highly flexible and can be easily implemented with MySQL, and JSON.
- Highly scalable as it can distribute a large amount of data in small segments.
- It works on small commodity hardware like JBOD or a bunch of disks.

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. It can deliver thousands of operations every second and can handle petabytes of resources with almost zero downtime. It was created by Facebook back in 2008 and was published publicly.

Features of APACHE Cassandra

- Data Storage Flexibility: It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
- Data Distribution System: Easy to distribute data with the help of replicating data on multiple data centers.
- Fast Processing: Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
- Fault-tolerance: The moment, if any node fails, it will be replaced without any delay.

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. It is capable of configuring multi-cloud services such as AWS, Azure, and Google Cloud. Besides, it also helps in lowering the cost of cloud computing by 50%.

Features of Qubole:

- Supports ETL process: It allows companies to **migrate data from multiple sources in one place**.
- Real-time Insight: It monitors user's systems and allows them to view real-time insights
- Predictive Analysis: Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.
- Advanced Security System: To protect users' data in the cloud, Qubole uses an advanced security system and also ensures to protect any future breaches. Besides, it also allows encrypting cloud data from any potential threat.

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. The best part of using Xplenty is its low investment in hardware & software and its offers support via **email, chat, telephonic and virtual meetings**. Xplenty is a platform to process data for analytics over the cloud and segregates all the data together.

Features of Xplenty

- Rest API: A user can possibly do anything by implementing Rest API
- Flexibility: Data can be sent, and pulled to databases, warehouses, and salesforce.
- Data Security: It offers SSL/TSL encryption and the platform is capable of verifying algorithms and certificates regularly.
- Deployment: It offers integration apps for both cloud & in-house and supports deployment to integrate apps over the cloud.

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. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is **capable of handling multi-petabytes of data** as well. Recently, Spark made a record of processing **100 terabytes of data in just 23 minutes** which

broke the previous world record of **Hadoop (71 minutes)**. This is the reason why big tech giants are moving towards spark now and is highly suitable for ML and AI today.

Features of APACHE Spark

- Ease of use: It allows users to run in their preferred language. (JAVA, Python, etc.)
- Real-time Processing: Spark can handle real-time streaming via Spark Streaming
- Flexible: It can run on, Mesos, Kubernetes, or the cloud.

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. It is so popular among developers due to its availability for multi-programming languages such as Python, Jscript, and Ruby.

Features of Mongo DB

- Written in C++: It's a schema-less DB and can hold varieties of documents inside.
- Simplifies Stack: With the help of mongo, a user can easily store files without any disturbance in the stack.
- Master-Slave Replication: It can write/read data from the master and can be called back for backup.

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. When we talk about real-time data processing, Storm leads the chart because of its distributed real-time big data processing system, due to which today many tech giants are using APACHE Storm in their system. Some of the most notable names are Twitter, Zendesk, NaviSite, etc.

Features of Storm:

- Data Processing: Storm process the data even if the node gets disconnected
- Highly Scalable: It keeps the momentum of performance even if the load increases
- Fast: The speed of APACHE Storm is impeccable and can process up to 1 million messages of 100 bytes on a single node.

8. SAS

Today it is one of the best tools for creating statistical modeling used by data analysts. By using SAS, a data scientist can mine, manage, extract or update data in different variants from different

sources. Statistical Analytical System or SAS allows a user to access the data in any format (SAS tables or Excel worksheets). Besides that it also offers a cloud platform for business analytics called **SAS Viya** and also to get a strong grip on AI & ML, they have introduced new tools and products.

Features of SAS:

- Flexible Programming Language: It offers easy-to-learn syntax and has also vast libraries which make it suitable for non-programmers
- Vast Data Format: It provides support for many programming languages which also include SQL and carries the ability to read data from any format.
- Encryption: It provides end-to-end security with a feature called **SAS/SECURE**.

9. Data Pine

Datapine is an analytical used for BI and was founded back in 2012 (Berlin, Germany). In a short period of time, it has gained much popularity in a number of countries and it's mainly used for data extraction (for small-medium companies fetching data for close monitoring).

Features of Datapine:

- Automation: To cut down the manual chase, datapine offers a wide array of AI assistant and BI tools.
- Predictive Tool: datapine provides forecasting/predictive analytics by using historical and current data, it derives the future outcome.
- Add on: It also offers intuitive **widgets, visual analytics & discovery, ad hoc reporting**, etc.

10. Rapid Miner

It's a fully automated visual workflow design tool used for data analytics. It's a no-code platform and users aren't required to code for segregating data. Today, it is being heavily used in many industries such as ed-tech, training, research, etc. Though it's an open-source platform but has a limitation of adding **10000 data rows and a single logical processor**. With the help of Rapid Miner, one can easily deploy their ML models to the web or mobile (only when the user interface is ready to collect real-time figures).

Features of Rapid Miner

- Accessibility: It allows users to access 40+ types of files (SAS, ARFF, etc.) via URL
- Storage: Users can access cloud storage facilities such as AWS and dropbox
- Data validation: Rapid miner enables the visual display of multiple results in history for better evaluation.

Methods of Big Data Analytics

1. A/B Testing

A/B testing is comparing a control group to a test group to figure out what variable changes will create the best outcomes. For example, this type of data analysis method is often used in email marketing to test different subject lines or images. Once the system figures out which one is generating the greatest level of engagement, it becomes the chosen selection.

2. Data Mining

Data mining finds patterns from large data sets by utilising statistical methods and machine learning. A great use of data mining is to figure out customer behaviour in order to offer the most fitting products to a specific segment of your entire customer database.

3. Regression Analysis

As a statistical method, regression analysis aims to determine the effect of an independent variable on a dependent variable. It figures out how a dependent variable changes with fluctuations in the independent variable. It can be used to decipher how customer satisfaction rates affect customer loyalty, for example.

4. Natural Language Processing (NLP)

Under the umbrella of computer science and artificial intelligence, natural language processing (NLP) uses algorithms to process human language. This can be used in a variety of settings, such as text-to-talk when phone assistants like Apple's Siri are able to transcribe what humans say out loud and then respond with a fitting answer.

5. Cluster Analysis

This type of analysis is used when looking to extract patterns from large data. It works by grouping data elements that are like one another in a sense. It can also be used to add context to trends.

With a large customer base, it's hard to understand each person's behaviour one-by-one. Cluster analysis is a way to combine records together based on specific elements, be it demographics, purchasing behaviours, or something else.

6. Time Series Analysis

Time series analysis analyses data within a defined period of time. It helps to see how different variables affect outcomes at different points in time. It's very effective for understanding seasonality effects of customer decisions.

7. **Machine Learning**

Machine learning enables systems to learn from data and improve performance without explicit programming. Algorithms like decision trees, neural networks, and support vector machines are commonly used.

- **Example:** Detecting fraud in financial transactions using anomaly detection techniques.

8. Predictive Modeling

Predictive modeling uses statistical and machine learning models to forecast future outcomes. Regression analysis, time-series analysis, and supervised learning are popular techniques.

- **Example:** Predicting energy consumption patterns based on historical usage.

9. Stream Analytics

Stream analytics processes data in real time, providing immediate insights for decision-making.

- **Example:** Monitoring sensor data in manufacturing for predictive maintenance.

10. Statistical Analysis

Statistical analysis involves hypothesis testing, correlation, and regression to explore data relationships and validate findings.

- **Example:** Analyzing clinical trial data to validate the efficacy of a drug.

1.9 Analysis vs Reporting

In the context of big data, **analysis** and **reporting** are two distinct but interconnected activities that serve different purposes. Here's a breakdown of each:

1. Analysis in Big Data

Analysis refers to the process of examining large datasets to extract valuable insights, identify patterns, and uncover trends. It is often exploratory and involves the use of advanced techniques like statistical modeling, machine learning, and data mining. The goal is to understand underlying patterns, predict future outcomes, and make informed decisions.

Key characteristics of analysis

- **Exploratory**: Often performed to discover insights from raw, unstructured, or semi-structured data.
- **Complex**: Involves sophisticated methods such as regression analysis, predictive modeling, clustering, and classification.
- **Iterative**: Analysts may need to refine hypotheses or techniques to gain deeper insights.
- **Dynamic**: It adapts as new data is acquired or as models are improved.
- **Tools**: Data scientists typically use programming languages like Python or R, and frameworks like Apache Spark, Hadoop, and TensorFlow.

Types of Analysis

- **Descriptive Analysis**: Summarizes past data to understand what has happened.

- **Predictive Analysis**: Uses historical data to predict future events.
- **Prescriptive Analysis**: Recommends actions based on predictions.
- **Diagnostic Analysis**: Explores why something happened.

2. Reporting in Big Data

Reporting, on the other hand, is focused on summarizing the results of data analysis in a structured, often standardized format. The objective is to communicate insights clearly to stakeholders in an accessible way. Reports often use visualizations, charts, tables, and dashboards to present data and trends.

Key characteristics of reporting

- **Descriptive**: It tends to be more focused on summarizing what has happened or is happening, rather than explaining or predicting.
- **Static**: Reports often present data in a specific format without much interactivity or modification.
- **Predefined**: Reports are typically created following a specific template or structure, often on a scheduled basis (e.g., weekly or monthly).
- **Tools**: Business Intelligence (BI) tools such as Tableau, Power BI, and Google Data Studio are commonly used for reporting.

Types of Reporting

- **Operational Reporting**: Focuses on day-to-day business metrics (e.g., sales performance).
- **Strategic Reporting**: Involves higher-level metrics that support long-term decision-making.
- **Ad-hoc Reporting**: Custom reports generated on-demand to address specific questions or problems.

Key Differences Between Analysis and Reporting:

Aspect	Analysis	Reporting
Purpose	Discover insights, patterns, and trends.	Present summarized data to inform decisions.
Methodology	Uses advanced techniques (e.g., machine learning).	Summarizes data in a structured format.
Complexity	Can be complex and exploratory.	Generally simpler, focused on clarity.
Timeframe	Often continuous or iterative.	Periodic (e.g., daily, weekly, monthly).
Outcome	Predictive insights and actionable conclusions.	Static reports or dashboards showing historical data.
Audience	Data scientists, analysts, and decision-	Business leaders, managers, or general

Aspect	Analysis	Reporting
	makers.	stakeholders.
Tools	Advanced analytics platforms (e.g., R, Python, Spark).	BI tools (e.g., Power BI, Tableau).

- **Analysis** is typically more complex, aimed at understanding data through in-depth exploration, often leading to predictive or prescriptive outcomes.
- **Reporting** is about presenting data in a digestible form for decision-makers, often focusing on what has happened rather than why or what could happen.

Both are essential in the big data ecosystem: **analysis** helps to generate actionable insights, while **reporting** ensures those insights are communicated effectively to stakeholders for decision-making.