BIG DATA AND CLOUD COMPUTING

UNIT I

INTRODUCTION TO BIG DATA

Big Data

Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time. These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

The amount and availability of data is growing rapidly, spurred on by digital technology advancements, such as connectivity, mobility, the Internet of Things (IoT), and artificial intelligence (AI). As data continues to expand and proliferate, new big data tools are emerging to help companies collect, process, and analyze data at the speed needed to gain the most value from it.

Big data describes large and diverse datasets that are huge in volume and also rapidly grow in size over time. Big data is used in machine learning, predictive modeling, and other advanced analytics to solve business problems and make informed decisions.

Big data examples

Data can be a company's most valuable asset. Using big data to reveal insights can help you understand the areas that affect your business—from market conditions and customer purchasing behaviors to your business processes.

Here are some big data examples that are helping transform organizations across every industry:

- Tracking consumer behavior and shopping habits to deliver hyper-personalized retail product recommendations tailored to individual customers
- Monitoring payment patterns and analyzing them against historical customer activity to detect fraud in real time
- Combining data and information from every stage of an order's shipment journey with hyperlocal traffic insights to help fleet operators optimize last-mile delivery
- Using AI-powered technologies like natural language processing to analyze unstructured medical data (such as research reports, clinical notes, and lab results) to gain new insights for improved treatment development and enhanced patient care
- Using image data from cameras and sensors, as well as GPS data, to detect potholes and improve road maintenance in cities

- Analyzing public datasets of satellite imagery and geospatial datasets to visualize, monitor, measure, and predict the social and environmental impacts of supply chain operations

CHARACTERISTICS OR V'S OF BIG DATA

1. Volume:

- The name 'Big Data' itself is related to a size which is enormous.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large, then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.
- Example: In the year 2016, the estimated global mobile traffic was 6.2 Exabytes (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exabytes of data.

2. Velocity:

- Velocity refers to the high speed of accumulation of data.
- In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
- There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Sampling data can help in dealing with the issue like 'velocity'.
- Example: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22%(Approx.) year by year.

3. Variety:

- It refers to nature of data that is structured, semi-structured and unstructured data.
- It also refers to heterogeneous sources.
- Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.
- Structured data: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
- Semi- Structured data: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.

- Unstructured data: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.

4. Veracity:

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- Example: Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

Value:

- After having the 4 V's into account there comes one more V which stands for Value! The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 6V's.

Variability:

- How fast or available data that extent is the structure of your data is changing?
- How often does the meaning or shape of your data change?
- Example: if you are eating same ice-cream daily and the taste just keep changing.

Big-Data Analytics

- Big Data Analytics is all about crunching massive amounts of information to uncover hidden trends, patterns, and relationships. It's like sifting through a giant mountain of data to find the gold nuggets of insight.
- Here's a breakdown of what it involves:
- Collecting Data: Such data is coming from various sources such as social media, web traffic, sensors and customer reviews.
- Cleaning the Data: Imagine having to assess a pile of rocks that included some gold pieces in it. You would have to clean the dirt and the debris first. When data is being cleaned, mistakes must be fixed, duplicates must be removed and the data must be formatted properly.

- Analyzing the Data: It is here that the wizardry takes place. Data analysts employ powerful tools and techniques to discover patterns and trends. It is the same thing as looking for a specific pattern in all those rocks that you sorted through.

Working of big data analytics

Big Data Analytics is a powerful tool which helps to find the potential of large and complex datasets. To get better understanding, let's break it down into key steps:

Data Collection: Data is the core of Big Data Analytics. It is the gathering of data from different sources such as the customers' comments, surveys, sensors, social media, and so on. The primary aim of data collection is to compile as much accurate data as possible. The more data, the more insights.

Data Cleaning (Data Preprocessing): The next step is to process this information. It often requires some cleaning. This entails the replacement of missing data, the correction of inaccuracies, and the removal of duplicates. It is like sifting through a treasure trove, separating the rocks and debris and leaving only the valuable gems behind.

Data Processing: After that we will be working on the data processing. This process contains such important stages as writing, structuring, and formatting of data in a way it will be usable for the analysis. It is like a chef who is gathering the ingredients before cooking. Data processing turns the data into a format suited for analytics tools to process.

Data Analysis: Data analysis is being done by means of statistical, mathematical, and machine learning methods to get out the most important findings from the processed data. For example, it can uncover customer preferences, market trends, or patterns in healthcare data.

Data Visualization: Data analysis usually is presented in visual form, for illustration – charts, graphs and interactive dashboards. The visualizations provided a way to simplify the large amounts of data and allowed for decision makers to quickly detect patterns and trends.

Data Storage and Management: The stored and managed analyzed data is of utmost importance. It is like digital scrapbooking. May be you would want to go back to those lessons in the long run, therefore, how you store them has great importance. Moreover, data protection and adherence to regulations are the key issues to be addressed during this crucial stage.

Continuous Learning and Improvement: Big data analytics is a continuous process of collecting, cleaning, and analyzing data to uncover hidden insights. It helps businesses make better decisions and gain a competitive edge.

Types of Big Data Analytics

Big Data Analytics comes in many different types, each serving a different purpose:

Descriptive Analytics: This type helps us understand past events. In social media, it shows performance metrics, like the number of likes on a post.

Diagnostic Analytics: In Diagnostic analytics delves deeper to uncover the reasons behind past events. In healthcare, it identifies the causes of high patient re-admissions.

Predictive Analytics: Predictive analytics forecasts future events based on past data. Weather forecasting, for example, predicts tomorrow's weather by analyzing historical patterns.

Prescriptive Analytics: However, this category not only predicts results but also offers recommendations for action to achieve the best results. In e-commerce, it may suggest the best price for a product to achieve the highest possible profit.

Real-time Analytics: The key function of real-time analytics is data processing in real time. It swiftly allows traders to make decisions based on real-time market events.

Spatial Analytics: Spatial analytics is about the location data. In urban management, it optimizes traffic flow from the data unde the sensors and cameras to minimize the traffic jam.

Text Analytics: Text analytics delves into the unstructured data of text. In the hotel business, it can use the guest reviews to enhance services and guest satisfaction.

Big Data Analytics Technologies and Tools

Big Data Analytics relies on various technologies and tools that might sound complex, let's simplify them:

Hadoop: Imagine Hadoop as an enormous digital warehouse. It's used by companies like Amazon to store tons of data efficiently. For instance, when

Amazon suggests products you might like, it's because Hadoop helps manage your shopping history.

Spark: Think of Spark as the super-fast data chef. Netflix uses it to quickly analyze what you watch and recommend your next binge-worthy show.

NoSQL Databases: NoSQL databases, like MongoDB, are like digital filing cabinets that Airbnb uses to store your booking details and user data. These databases are famous because of their quick and flexible, so the platform can provide you with the right information when you need it.

Tableau: Tableau is like an artist that turns data into beautiful pictures. The World Bank uses it to create interactive charts and graphs that help people understand complex economic data.

Python and R: Python and R are like magic tools for data scientists. They use these languages to solve tricky problems. For example, Kaggle uses them to predict things like house prices based on past data.

Machine Learning Frameworks (e.g., TensorFlow): In Machine learning frameworks are the tools who make predictions. Airbnb uses TensorFlow to predict which properties are most likely to be booked in certain areas. It helps hosts make smart decisions about pricing and availability.

Challenges of Big data analytics

While Big Data Analytics offers incredible benefits, it also comes with its set of challenges:

Data Overload: Consider Twitter, where approximately 6,000 tweets are posted every second. The challenge is sifting through this avalanche of data to find valuable insights.

Data Quality: If the input data is inaccurate or incomplete, the insights generated by Big Data Analytics can be flawed. For example, incorrect sensor readings could lead to wrong conclusions in weather forecasting.

Privacy Concerns: With the vast amount of personal data used, like in Facebook's ad targeting, there's a fine line between providing personalized experiences and infringing on privacy.

Security Risks: With cyber threats increasing, safeguarding sensitive data becomes crucial. For instance, banks use Big Data Analytics to detect fraudulent activities, but they must also protect this information from breaches.

Costs: Implementing and maintaining Big Data Analytics systems can be expensive. Airlines like Delta use analytics to optimize flight schedules, but they need to ensure that the benefits outweigh the costs.

Usage of Big Data Analytics

Big Data Analytics has a significant impact in various sectors:

Healthcare: It aids in precise diagnoses and disease prediction, elevating patient care.

Retail: Amazon's use of Big Data Analytics offers personalized product recommendations based on your shopping history, creating a more tailored and enjoyable shopping experience.

Finance: Credit card companies such as Visa rely on Big Data Analytics to swiftly identify and prevent fraudulent transactions, ensuring the safety of your financial assets.

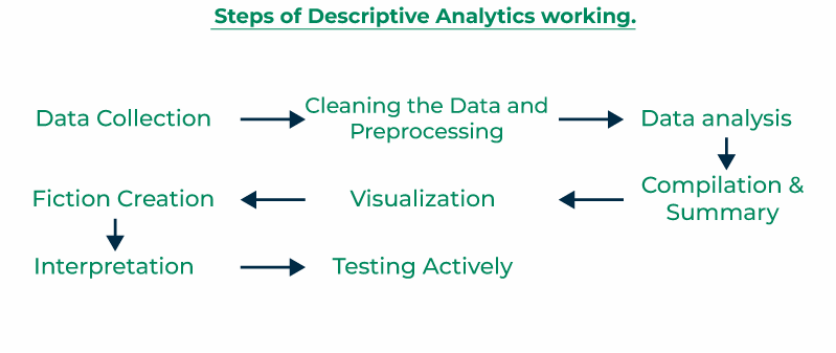
Transportation: Companies like Uber use Big Data Analytics to optimize drivers' routes and predict demand, reducing wait times and improving overall transportation experiences.

Agriculture: Farmers make informed decisions, boosting crop yields while conserving resources.

Types of Big Data Analytics

Big Data Analytics comes in many different types, each serving a different purpose:

1. **Descriptive Analytics:** This type helps us understand past events. In social media, it shows performance metrics, like the number of likes on a post.



Steps for Descriptive Analytics Work:

1. **Data Collection:** Collecting useful information is the initial stage in the descriptive analytics process. By using multiple resources such as databases, spreadsheets, and other data repositories. All of these provide this data. Since they directly affect how accurate the descriptive analytics is, the accuracy and standard of the data are extremely important.
1. **Cleaning the Data and Preprocessing:** The obtained data usually needs to be cleaned and preprocessed before analysis can start. This includes converting data into a uniform structure, standardizing formats, and handling missing or incorrect values. Clean and well-preprocessed data ensures that the subsequent analytics is reliable.
1. **Data analysis:** It provides an understanding of the structure and features of the dataset. Here EDA (**exploratory data analysis**) methods helps to find the patterns, trends, and possible outliers in the data. These methods include making histograms, scatter plots, and summary statistics.
1. **Compilation and Summary:** The goal of descriptive analytics is to offer an overview of the data at a high level. To get important metrics and statistics, such as mean, median, mode, range, and standard deviation, this frequently requires combining the data.
1. **Visualization:** In descriptive analytics, visualizations are extremely useful tools. It helps us to communicate complex information with a variety of charts, graphs, and other visual representations are employed. Data patterns and trends can be highlighted with the use of visualization, which also makes it easier to convey insights to a wide range of audiences.
1. **Fiction Creation:** Descriptive analytics can include the creation of descriptions that offer a logical and contextualized explanation of the data, in addition to visuals. When communicating findings to those in the audience who might not be familiar with the complexities of the data, this can be especially helpful.
1. **Interpretation:** To obtain significant knowledge, analysts interpret the outcomes of descriptive analytics. This involves knowing the effects of the trends and patterns seen in the data. While interpretation provides the foundation for more in-depth analyses that investigate "why" and "what might happen in the future," descriptive analytics concentrates on the "what happened" topic.
1. **Testing Actively:** The process of descriptive analytics is not one-time. Organizations continually repeat the descriptive analytics when new data becomes available in order to keep informed about the latest developments and patterns. This way, people making decisions get the newest information.

Advantages of Descriptive Analytics

- **Data-driven decision making:** It provides well-informed decision-making based on facts rather than gut instincts by evaluating and simplifying data.
- **Presents data clearly:** Descriptive analytics simplifies complex data, making it easy to understand through reports and visualizations like charts and graphs.

- **Convenient to Realize:** Data that has been summarized and graphically represented is easier to clarify and evaluate for a larger audience.
- **Identifies Relevant Data Points:** It offers straightforward metrics that give an accurate estimation of important data points.
- **Simple and cost-effective:** Descriptive analytics is simple to use and just requires basic arithmetic knowledge for execution.
- **Efficient with tools:** With the aid of tools like Python or MS Excel, which make things fast and easy.

Disadvantages of Descriptive Analytics

- **Inability of Cause Analysis:** The main goal of descriptive analytics is to explain historical events. It doesn't explore the root causes or reasons for the patterns that are seen.
- **Analysis Simplicity:** The reach of descriptive analytics is restricted to basic analyses that look at the relationships between a small number of variables.
- **Doesn't Explain Why:** History offers lessons for future generations, by offering facts, but causes and predictions are not provided to the readers.
- **Inappropriate for Making Decisions in Real Time:** Normally, descriptive analytics involves getting summary information at intervals intervals and this might not be the best option for decision- making when the time matter. In many situations, fast responsiveness is vital, therefore, sometimes only relying on the descriptive analytics might drag you behind.
- **Lack of ability to handle unstructured data:** Structured and well-organized datasets are better suited for descriptive analytics. while analyzing semi-structured or unstructured data, such as text, photos, or multimedia, it could make challenging to offer insightful analysis.

Applications of Descriptive Analytics

- **Financial Performance Evaluation:** For instance, in the past; descriptive analytics was often used to appraise and assess a specific firm's previous performances. Lots of organizations can detect trends, patterns and possibilities for a change by tracking key performance indicators (KPI's) at different periods of time. This awareness helps in the construction and building of business operations with all the required strategic planning.
- **Marketing and Analysis of Customer Behavior:** However, Companies should analyze and understand the customers' behavior. Firms need descriptive analytics to weight historical data on consumer interactions, purchasing patterns, and preferences.
- **Friction Analysis in Business Processes:** Descriptive analytics is applied descriptive approaches in business learning and development, and to detect and reduce friction in business processes. All the blockades or imparing of efficiency restraining processes from moving will be called friction. Organizations can easily pinpoint the bottlenecks of their business processes

by looking at historical data over workflow delays using of resources and process's time.

- **Social Networking Analytics:** In order to analyze user involvement, content performance, and audience demographics, descriptive analytics is used in social media. It assists businesses in customizing their social media plans according on past performance.
- **Crime and Fraud Detection:** Pattern in previous crime data is investigated by law enforcement and security agencies in order to do descriptive analysis which is one of the types of analytics. It is applied by financial organizations to make discoveries of market fluctuations and anomalies that can prevent or can be used to fight them.
- **Crypto Market Analysis:** Cryptocurrency markets are a great source of information for investors, as historical price data, market volumes aggregates, and market trends can be used to analyze the behavior of Bitcoin traders. These algorithms, mood patterns in the market, and possible factors may affect the price fluctuation of Bitcoin can all been fancy with the help of a descriptive analytics.
- **Human Resources Management:** HR uses descriptive analytics to analyze their staff. It aids businesses in the analysis of previous information on worker performance, turnover rates, training effectiveness, and other HR indicators.
- **Risk Assessment and Management:** To identify and analyze historical risk factors, descriptive analytics is used in risk assessment. Organizations need to know this information. This information is really important for companies in areas like banking and insurance to create plans that help reduce and handle risks better.

Diagnostic Analytics:

In Diagnostic analytics delves deeper to uncover the reasons behind past events. In healthcare, it identifies the causes of high patient re-admissions.

he primary purpose of diagnostic analytics is to uncover the root causes behind trends, anomalies, or issues identified through data analysis. It goes beyond simply describing what's happening (descriptive analytics) to understanding why it's happening.

- **Identify root causes:** Diagnostic analytics exactly comes to plight here by finding the exact factors that either lead to the desired effect or the one that needs to be changed.
- **Solve problems:** The route cause understanding will allow businesses to take proper actions on target to solve problems and get an opportunity to improve the performance.
- **Improve processes:** Inputs from diagnostic analytics can assist in exposing shortcomings or traffic jam in workflow, with the primary aim of process optimization.

- **Inform future decisions:** To be well-off with reasons behind things helps to make decisions more based on evidence encountered than assumptions made.

Steps in Diagnostic Analytics

1. **Identify the Anomaly:** The first step is pinpointing the deviation from the norm. Is it a sudden drop in sales, a spike in customer complaints, or an unexpected equipment failure?
1. **Data Collection:** Begin by gathering relevant data from various sources, ensuring a comprehensive dataset. Compile relevant data from various sources – website logs, customer surveys, sensor readings, financial records – to build a comprehensive picture.
1. **Data Exploration:** Dive into the data to uncover hidden insights and anomalies through statistical analysis and visualization.
1. **Pattern Identification:** Employ advanced algorithms to identify patterns and trends in the data. Analyze the data using techniques like drill-down, data mining, and correlation analysis. This involves sifting through layers of information, identifying patterns, and spotting hidden relationships.
1. **Root Cause Analysis:** Drill down into the data to understand the underlying factors contributing to specific outcomes. Based on your analysis, narrow down the possible causes. Was it a competitor's new campaign? A faulty software update? A change in supplier quality?
1. **Testing and Confirmation:** Design and implement tests to validate your hypothesis. Did the website redesign cause the traffic dip? Did the new marketing campaign trigger customer churn?

Benefit of Diagnostic Analytics

- **Deeper Insights:** With detection of reasoning behind "what" to "why," diagnostics analytics provide a profound information about data. It serves to expose the reasons, patterns, and problems behind the trends, deviations, and phenomena, hence, you can make sound decisions based on data which backs a case, not an assumption.
- **Improved Problem-Solving:** Diagnostics analytics gives you the capacity of pinning down the real reasons behind what you perceive as a problem. Thus, you have a possibility for directed actions aimed at preventing the root of problems, what, in turn, brings about the sensible solutions and sustainable improvements.
- **Process Optimization:** Diagnostic analysis-driven insights may shed light on the presence of bottlenecks and efficiencies in work flows. The knowledge on how operations fail can help you in adopting improved techniques, boosting productivity and lowering waste.
- **Enhanced Decision-Making:** Diagnostic analytics provides you with a predictive mode of causation and exposition from your data. It enables evidence-based

decision making so managers will have tactical strategic choices and we will see a performance improvement.

- **Risk Reduction:** Via detecting issues while they are still small, the diagnostic analytics avoids the development of a crisis and helps in the further mitigation of risks. What it does is increase efficiency by cutting down the time, money and resources your organization spends on unnecessary things.
- **Increased Customer Satisfaction:** Also, retail industries can take advantage of such analytics to uncover the peculiarities about buyer habits and preferences. This in turn enables the businesses to engage in accurate and tailored marketing campaigns which ultimately culminate to personalized experiences and better product offerings, again increased customer satisfaction.
- **Competitive Advantage:** Through the application of diagnostic analytics in performance improvement, coming up with new opportunities and statistically optimizing decisions, you can achieve a great victory over your aging competitors who only depend on intuition and bare-bone data procedures.

Predictive Analytics:

Predictive analytics forecasts future events based on past data. Weather forecasting, for example, predicts tomorrow's weather by analyzing historical patterns.

Predictive analytics is a branch of data science that leverages statistical techniques, machine learning algorithms, and historical data to make data-driven predictions about future outcomes.

- **Informed Decision-Making:** By anticipating future trends and outcomes, businesses and organizations can make more strategic decisions. Imagine being able to predict customer churn (when a customer stops using your service) or equipment failure before it happens. This allows for proactive measures to retain customers or prevent costly downtime.
- **Risk Management:** Predictive analytics helps identify and mitigate potential risks. For example, financial institutions can use it to detect fraudulent transactions, while healthcare providers can predict the spread of diseases.
- **Optimization and Efficiency:** Predictive models can optimize processes and resource allocation. Businesses can forecast demand and optimize inventory levels, or predict equipment maintenance needs to avoid disruptions.
- **Personalized Experiences:** Predictive analytics allows for personalization and customization. Retailers can use it to recommend products to customers based on their past purchases and browsing behavior.
- **Innovation and Competitive Advantage:** Predictive analytics empowers organizations to identify new opportunities and develop innovative products and services. By understanding customer needs and market trends, businesses can stay ahead of the competition.

1. Define a Problem:

- Firstly data scientists or data analysts define the problem.
- Defining the problem means clearly expressing the challenge that the organization aims to focus using data analysis.
- A well- defined problem statement helps determine the appropriate predictive analytics approach to employ.

2. Gather and Organize Data:

- Once you define a problem statement it is important to acquire and organize data properly.
- Acquiring data for predictive analytics means collecting and preparing relevant information and data from various sources like databases, data warehouses, external data providers, APIs, logs, surveys, and more that can be used to build and train predictive models.

3. Pre-process Data:

- Now after collecting and organizing the data, we need to pre-process data.
- Raw data collected from different sources is rarely in an ideal state for analysis. So, before developing a predictive models, data need to be pre-processed properly.
- Pre-processing involves cleaning the data to remove any kind of anomalies, handling missing data points and addressing outliers that could be caused by errors or input or transforming the data , which can be used for further analysis.
- Pre-processing ensures that data is of high quality and now the data is ready for model development.

4. Develop Predictive Models:

- Data scientists or data analysts leverage a range of tools or techniques to develop a predictive models based on the problem statement and the nature of the datasets.
- Now techniques like machine learning algorithms, regression models , decisions trees, neural networks are much among the common techniques for this.
- These models are trained on the prepared data to identify correlations and patterns that can be used for making predictions.

5. Validate and Deploy Results:

- After building the predictive model, validation is the critical steps to assess the accuracy and reliability of predictions.
- Data scientists rigorously evaluate the model's performance against known outcomes or test datasets.

- If required, modifications are implemented to improve the accuracy of the model.
- Once the model achieve satisfactory outcomes it can be deployed to deliver predictions to stakeholders.
- This can be done through applications, websites or data dashboards, making the insights easily accessible to decision makers or stakeholders.

Predictive Analytics Techniques:

Predictive analytical models leverage historical data to anticipate future events or outcomes, employing several distinct types:

- **Classification Models**: These predict categorical outcomes or categorize data into predefined groups. Examples include Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine.
- **Regression Models**: Used to forecast continuous outcome variables based on one or more independent variables. Examples include Linear Regression, Multiple Regression, and Polynomial Regression.
- **Clustering Models**: These group similar data points together based on shared characteristics or patterns. Examples comprise K-Means Clustering and Hierarchical Clustering.
- **Time Series Models**: Designed to predict future values by analyzing patterns in historical time-dependent data. Examples include Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing Models.
- **Neural Networks Models**: Advanced predictive models capable of discerning complex data patterns and relationships. Examples encompass Feed Forward Neural Networks, Recurrent Neural Networks, and Convolutional Neural Networks.

Prescriptive Analytics Approach

Step 1 Data Collection: Gather data for a customer's locations, their requirement, company warehouses, and transportation

Step 2 Mathematical Modeling: We will create mathematical models that will handle supply chain data like customer location, time, warehouse location, and routes, we will also finalize an optimization function that will minimize company cost and delivery time

Step 3 Optimization: We will use an optimization approach like linear programming or differential calculus to solve mathematical models and find optimal locations.

Step 4 Scenario Analysis: We will perform a scenario analysis for our assumptions variables about the models.

Step 5 Decision Support: Based on our data modeling and business knowledge that we got from the raw data we will create dashboards and visualization graphs that will stakeholders in taking decisions.

Step 5 Implementation: The Final and most important part after doing all the five steps is to implement it with changes that maximizes the company's revenues

Descriptive Analytics Vs Predictive Analytics Vs Prescriptive Analytics

Descriptive analytics works over the statistical data to give us details related to the past. It helps the business to get all relatable details regarding their performance from past stats. **For Example,** Analyzation of past purchasing details of consumers/customers to decide the best time for launching a new product or any sales scheme in the market.

Predictive analytics uses a machine learning model consisting of all the relatable key trends and particular scalable patterns with the help of historical data and feeds. This model is then used in business to predict what will happen next applying the latest information. **For Example,** Statistics models are used by enterprises to through previous data whether how much consumers are using the services and which services are most popular among them so a relatable model to check in-demand services among users.

Prescriptive analytics is used to make next-level and advanced usage of predicted data. Business enterprises use the predicted possibilities to develop and provide better services to their customers/consumers. **For Example,** For a successful and cost-effective delivery system transportation enterprises used algorithms and predictive models to decide the best route with minimum energy usage for saving time and increasing profits.

Advantages of Prescriptive Analytics

- Effortlessly map Business analysis to declare out steps necessary to avoid failure and achieve success.
- An accurate and Comprehensive form of data aggregation and analysis also reduces human error and bias.
- Helping in decision-making threads related to problems rather than jumping to unreliable conclusions based on instincts.
- Removing immediate uncertainties helps in the prevention of fraud, limits risk, increases efficiency, and creates logical customers.

Real-time Analytics:

The key function of real-time analytics is data processing in real time. It swiftly allows traders to make decisions based on real-time market events.

Real-time analytics involves continuously processing and analyzing data immediately as it is collected, enabling instant decision-making and responses. businesses to get awareness and take action on data immediately or soon after the data enters their system. Real-time app analytics respond to queries within seconds. They grasp a large amount of data with high velocity and low reaction time. For example, real-time big data analytics uses data in financial databases to notify trading decisions. Analytics can be on-demand or uninterrupted. On-demand notifies results when the user requests it. Continuous renovation users as events happen and can be programmed to answer automatically to certain events. For example, real-time web analytics might refurbish an administrator if the page load presentation goes out of the present boundary.

Real-time data analytics lets users see, examine and recognize data as it enters a system. Logic and mathematics are put into the data, so it can give users a perception for making real-time decisions.

Examples

Examples of real-time customer analytics include the following.

1. Viewing orders as they happen for better tracing and to identify fashion.
1. Continually modernize customer activity like page views and shopping cart use to understand user etiquette.
1. Choose customers with advancement as they shop for items in a store, affecting real-time decisions.

Real-Time Analytics – working

Real-time analytics involves a comprehensive and intricate process that encompasses several critical components and steps. Here's a more detailed breakdown of how it operates:

Data Ingestion

- **Continuous Data Collection**: Real-time analytics systems continuously collect data from various sources, such as sensors, IoT devices, social media feeds, transaction logs, and application databases. This data can come in various formats, including structured, semi-structured, and unstructured data.
- **Stream Processing**: Data is ingested as streams, meaning it is captured and processed in real-time as it arrives. Technologies like Apache Kafka, RabbitMQ, and Amazon Kinesis are commonly used for data ingestion due to their ability to handle high-throughput data streams reliably.

Benefits and Advantages of using Real-Time Analytics

- **Faster Decision-Making**: Real-time analytics provides instant access to data insights, allowing businesses to make informed decisions quickly.
- **Proactive Problem Solving**: By continuously monitoring data streams, organizations can identify and address issues as they arise, preventing potential problems from escalating.

Enhanced Operational Efficiency

- **Optimized Processes**: By analyzing data as it is generated real-time analytics, businesses can streamline processes, reduce waste, and improve overall productivity.
- **Resource Allocation**: Organizations can optimize the allocation of resources, such as labor, inventory, and energy, based on real-time demand and usage patterns.

Improved Customer Experience

- **Personalized Interactions**: Real-time analytics enables businesses to tailor their interactions with customers based on current data.
- **Responsive Service**: By analyzing customer behavior and feedback in real-time, businesses can quickly address issues and adapt their services to meet customer expectations.

Competitive Advantage

- **Market Responsiveness**: Businesses that leverage real-time analytics can quickly adapt to changing market conditions and emerging trends. This allows them to stay ahead of competitors and capitalize on new opportunities.
- **Innovation**: Real-time data insights can drive innovation by highlighting emerging trends and customer preferences. Businesses can use these insights to develop new products and services that meet market demands.

Enhanced Risk Management

- **Fraud Detection**: Real-time analytics is critical for identifying and preventing fraudulent activities. By continuously monitoring transactions and behavior patterns, businesses can detect anomalies and take immediate action to mitigate risks.
- **Operational Risk Management**: Real-time monitoring of operations allows businesses to identify and address potential risks before they cause significant disruptions.

Improved Financial Performance

- **Revenue Optimization**: Real-time analytics can help businesses identify and capitalize on revenue opportunities. For example, dynamic pricing models can adjust prices based on current demand and market conditions, maximizing revenue.
- **Cost Reduction**: By optimizing operations and resource allocation, real-time analytics can lead to significant cost savings. Businesses can reduce waste, improve efficiency, and lower operational expenses.

Enhanced Collaboration and Communication

- **Data-Driven Culture**: Real-time analytics helps employees across different departments can access and analyze real-time data, leading to more informed decisions and better collaboration.
- **Transparent Operations**: Real-time data visualization tools, such as dashboards and reports, provide a clear and transparent view of operations.

Regulatory Compliance

- **Real-Time Monitoring**: For industries with stringent regulatory requirements, real-time analytics provides continuous monitoring and reporting capabilities.
- **Audit Trails**: Real-time analytics systems can maintain detailed audit trails of data access and modifications, aiding in compliance and accountability.

Challenges of Real-Time Analytics in Big Data

Data Latency

- **Minimizing Delay**: Ensuring minimal delay in data processing and analysis can be challenging due to the sheer volume and velocity of Big Data. Achieving true real-time processing requires robust infrastructure and optimized algorithms.

Scalability

- **Handling Massive Data**: As data volumes grow, maintaining the scalability of real-time analytics systems becomes difficult. Organizations need to invest in scalable infrastructure and distributed processing frameworks to manage the increasing load.

Data Quality

- **Ensuring Accuracy**: Maintaining high data quality and accuracy in real-time environments is critical. Inconsistent or erroneous data can lead to incorrect insights and decisions, impacting business outcomes negatively.

Integration Complexity

- **Seamless Integration**: Integrating real-time analytics with existing systems and processes can be complex. Organizations need to ensure seamless data flow between various sources, analytics platforms, and applications.

Resource Intensive

- **High Computational Demands**: Real-time analytics requires significant computational resources for processing and storing data. This can lead to increased costs and the need for advanced hardware and software solutions.

Data Security

- **Protecting Data**: Ensuring the security and privacy of data in real-time analytics is crucial. Real-time systems are often more vulnerable to cyber-attacks due to continuous data transmission and processing.

Technical Expertise

- **Skilled Professionals**: Implementing and maintaining real-time analytics systems require skilled professionals with expertise in Big Data technologies, data science, and system integration. Finding and retaining such talent can be difficult.

Cost Implications

- **Financial Investment**: The infrastructure, tools, and human resources needed for real-time analytics can be expensive. Organizations must weigh the benefits against the costs to justify the investment.

Addressing these challenges is essential for successfully implementing real-time analytics in Big Data environments, enabling organizations to leverage timely insights for better decision-making and competitive advantage

Applications of Real-Time Analytics

Real-time analytics is a powerful tool that finds applications across various industries. Here are some key applications:

Predictive Maintenance: Manufacturing, utilities, and transportation sectors utilize real-time analytics to monitor equipment health and predict failures before they occur.

Fraud Detection: Financial services, e-commerce platforms, and insurance companies continuously monitor transactions and user behavior, to identify anomalies and take immediate action to mitigate fraud risks.

Customer Experience Management: Retailers, hospitality providers, and online services analyze customer interactions and feedback in real-time, businesses can personalize services, optimize marketing campaigns, and promptly address customer issues, leading to higher satisfaction and loyalty.
Smart Cities: Urban planners and city administrations employ real-time analytics in traffic management, public transportation optimization, and real-time monitoring of public safety and environmental conditions.
Healthcare: Healthcare providers use real-time analytics to monitor patient vitals, manage hospital resources, and provide timely interventions. For instance, real-time analysis of patient data can alert medical staff to potential emergencies, improving patient outcomes and operational efficiency.
Financial Trading: Financial institutions and traders rely on real-time analytics to make quick, informed trading decisions. By analyzing market data as it happens, traders can identify trends, detect anomalies, and execute trades at the optimal moment to maximize profits.
Supply Chain Management: Logistics and supply chain companies use real-time analytics to track shipments, manage inventory, and optimize delivery routes. This ensures timely deliveries, reduces costs, and improves overall supply chain efficiency.
Telecommunications: Telecom operators use real-time analytics to monitor network performance, detect outages, and manage bandwidth. This helps in maintaining service quality, reducing downtime, and enhancing customer satisfaction.
Energy Management: Utility companies and large enterprises employ real-time analytics for energy consumption monitoring and optimization. By analyzing real-time data from smart meters and sensors, businesses can optimize energy usage, reduce costs, and support sustainability initiatives.
Marketing and Advertising: Marketers and advertisers use real-time analytics to measure the effectiveness of campaigns and adjust strategies on the fly. Real-time insights into customer behavior and engagement help in creating targeted and impactful marketing efforts.
Retail and E-commerce: Retailers and e-commerce platforms leverage real-time analytics to manage inventory, optimize pricing strategies, and enhance the shopping experience. Analyzing real-time sales data and customer interactions helps in making informed decisions that drive sales and improve customer satisfaction.

Spatial Analytics:

Spatial analytics is about the location data. In urban management, it optimizes traffic flow from the data unde the sensors and cameras to minimize the traffic jam.

The world is overflowing with data, but this data only becomes valuable when we can derive meaningful insights from it. **Spatial analysis** is the process of

using **analytical tools** to study and **represent data, uncovering relationships** and patterns within geospatial data. This method transforms raw data into actionable information by analyzing **geographic features collected through satellites, maps, and other sources.** It employs a range of analytical techniques, **algorithms, and computational models** to draw connections between data points and apply them to targeted systems such as environmental management, urban planning, and more.

What is Spatial Data?

Spatial data also called geospatial data contains information that has a geographic component. **Spatial data is broadly classified into two categories, vector and raster.** Let's take a look at each one of them.

1. Vector Data

Vector data represents spatial features using points, lines, and polygons. In GIS, vector data is used to represent addresses and points of interest with points; rivers, railways, roads using lines and lakes, and buildings with polygons.

- **Point** - A point is depicted by a single dot on the layer. It is the simplest type of vector data and can be accessed using a single pair of coordinates i.e. x and y coordinates of that point. A point has zero dimension. Examples of points include the position of cities, landmarks, schools, etc. on a map.
- **Line** - Lines can be depicted as a sequence of connected points depicting the shape and location of a linear feature. A line is **one-dimensiona**l vector data. Examples of lines include rivers, roads, and power lines on a map.
- **Polygon** - A polygon is formed by connecting points in such a way that it forms a closed loop. It can also be seen as a line with the same start and end point, hence a closed loop. Each polygon can be differentiated from the others by assigning different colors to each polygon. Examples where polygons are used where we need to depict a defined area or boundary like buildings, closed water bodies, etc.

2. Raster Data

Raster data in contrast to vector data is a grid of cells where each cell represents a specific value. Examples of raster data include aerial photographs, imagery from satellites, digital pictures, and scanned maps. In raster data, each cell of the grid holds a single value representing various attributes like elevation, depth, etc.

- **Digital Elevation Models(DEM)** - This kind of raster data depicts the topography of the surface in terms of elevation or depth.
- **Satellite Imagery** - This kind of raster data depicts the aerial photographs taken by satellites where each cell in the grid takes up a color to imitate the image taken by the satellite.
- **Temperature map**s - This kind of raster data stores the temperature at each location in the cells of the grid.

Thus, rater data is used to store continuous data whereas vector data is used to store data with well-defined boundaries.

Spatial Analysis Work

Spatial analysis is the process of using analytical tools to analyze and represent data, relationships, and patterns among various geospatial data. This task of analyzing and recognizing patterns is discussed as follows.

1. Data Collection

Data collection is the foundation of spatial analysis. It involves gathering information from various sources, including remote sensing devices like **LiDAR (Light Detection and Ranging)** and **airborne systems**. This data, often **high-resolution images or photographs from satellites or aerial systems**, is used to create maps that depict the geographic distribution of entities. For example, maps showing temperature variations across different regions are created using this data.

2. Data Analysis

Once collected, the data undergoes spatial analysis using **artificial intelligence (AI) and machine learning (ML)** solutions to extract meaningful insights. ML models can be trained to detect and identify objects or structures within vast datasets, such as millions of images. These objects can include **schools, playgrounds, traffic zones, and residential areas**. In spatial analysis there are visualization tools further enhance this process by highlighting different objects with **distinct colors, shapes, or annotations, making it easier to identify** and analyze these objects within large datasets.

3. Data Presentation

In spatial analysis presenting analyzed data is crucial and can be time-consuming, as it involves emphasizing key findings. Data visualization tools, including **tables, charts, and graphs,** simplify this task by effectively projecting relevant data and facilitating communication with stakeholders. Additionally, **3D visualization tools enhance 2D data** by adding depth and perspective, optimizing planning and implementation strategies for better problem-solving outcomes.

Critical Capabilities of Spatial Analysis Workflows

- **Geographic Search**
    - Spatial analysis enables visualization of specific data on maps through user-friendly interfaces.
    - Users can search for geographic data using elements such as city names, country names, zip codes, etc.
    - This search functionality helps identify points of interest, such as schools in a specific area.
- **Clustering of Datasets**
    - Spatial analysis allows for the clustering of data points to understand demographic patterns.
    - Authorities can analyze the density of data points to determine the proximity of amenities like schools.
    - This helps identify areas with easy or limited access to facilities.
- **Comprehensive Data View**

    - Using various colors, shapes, and annotations provides a detailed overview of an area.
    - Different entities, such as hospitals, colleges, and repair shops, can be distinctly marked on maps for better visualization.
- **Visual Mapping**
    - Users can represent data sets on maps using layers, similar to heatmaps or bubble charts.
    - For instance, weather data can be displayed in layers to facilitate visual interpretation.
- **Highlighting Target Entities**
    - Different types of data can be combined and displayed on simple graphs.
    - For example, combining population data with the locations of nearby clinics helps determine if there are sufficient health centers for a given population.

Application of Spatial Analysis

1. Urban Development

- **Create Resilient Urban Cities** - Climate change has a great impact on urban life. Thus, policymakers are working on ways to minimize the effect of climate by analyzing deforestation patterns, sea level analysis due to **increasing global warming**, and **emission analysis** and **strategy to shift to efficient energy resources.**
- **Monitor and Reduce Urban Heat Island (UHI) effect** - The **UHI effect is the phenomenon where natural vegetation is replaced with buildings.** This leads to more heat retention. Spatial analysis techniques like thermal remote sensing, satellite imagery, and field observations can be used to collect relevant data and understand spatial patterns.
- **Determine Quality of Life - Spatial data can be used to determine the socioeconomic quality of life.** For example, areas with distributed hospital services have a better quality of life. Areas near industrial areas have a poor quality of life due to emissions.
- **Traffic Analysis** - Spatial imagery can be used to recognize congestion and high-traffic routes. This identification of busy routes can help improve public transportation infrastructure.

2. Public Health Sector

- **Mapping Spreading of Disease** - Satellite data can be used to **monitor the spread of disease in an area that helps policymakers come up with prevention plans.** The disease data can be integrated with climatic attributes and nearby water bodies' presence to analyze how various factors combine to increase the spread.
- **Sanitation and Health Facilities Analysis** - Spatial data can be used to identify areas with low sanitation and health facilities. Recognizing these areas can help the authorities to come up with a better healthcare management system.

- **Vaccination Statistics** - GIS technologies and spatial data can be used by authorities to come up with vaccination strategies and track even distribution of vaccines among the population.

- **Crop Monitoring** - Remote sensing can be used to collect data related to climate, soil nutrients, and sunlight which play a major role in crop productivity.
- **Crop Yield Prediction** - Satellite imagery can be used to provide insights about climate, weather conditions, and soil nutrients. Using this information, farmers can make better decisions for the best crop yield.
- **Farm Animals Monitoring** - Spatial analysis can be used to monitor freely roaming livestock which play a major role in methane production and soil and water contamination.
- **Soil Analysis** - Spatial analysis can help soil specialists retrieve important information about soil like **pH level, nitrogen levels**, **moisture content**, etc. which play an important role in a better crop.

Text Analytics:

Text analytics delves into the unstructured data of text. In the hotel business, it can use the guest reviews to enhance services and guest satisfaction.

Text Analytics is a process of analyzing and understanding written or spoken language. It employs computer algorithms and techniques to extract valuable information, patterns, and insights from extensive textual data. In simpler terms, text analytics empowers computers to understand and interpret human language. In simpler terms, text analytics helps computers understand and interpret human language. **Here's a real-world example to illustrate text analytics:** Let's say a company receives customer reviews for its products online. These reviews can be a goldmine of information, but it's not feasible for humans to read and analyze thousands of reviews manually. This is where text analytics comes in. The text analytics system can automatically analyze the reviews, looking for patterns and sentiments. It can identify common words or phrases that customers use to express satisfaction or dissatisfaction

Text Analytics Importance

Text analytics has become a crucial tool in today's information age for two main reasons: the **massive growth of text data** and its unique ability to **extract valuable insights** hidden within that data.

Steps of Text Analytics Process

Language Identification

- **Objective:** Determine the language in which the text is written.

- **How it works:** Algorithms analyze patterns within the text to identify the language. This is essential for subsequent processing steps, as different languages may have different rules and structures.

Tokenization

- **Objective:** Divide the text into individual units, often words or sub-word units (tokens).
- **How it works:** Tokenization breaks down the text into meaningful units, making it easier to analyze and process. It involves identifying word boundaries and handling punctuation.

Sentence Breaking

- **Objective:** Identify and separate individual sentences in the text.
- **How it works:** Algorithms analyze the text to determine where one sentence ends and another begins. This is crucial for tasks that require understanding the context of sentences.

Part of Speech Tagging

- **Objective:** Assign a grammatical category (part of speech) to each token in a sentence.
- **How it works:** Machine learning models or rule-based systems analyze the context and relationships between words to assign appropriate part-of-speech tags (e.g., noun, verb, adjective) to each token.

Chunking

- **Objective:** Identify and group related words (tokens) together, often based on the part-of-speech tags.
- **How it works:** Chunking helps in identifying phrases or meaningful chunks within a sentence. This step is useful for extracting information about specific entities or relationships between words.

Syntax Parsing

- **Objective:** Analyze the grammatical structure of sentences to understand relationships between words.
- **How it works:** Syntax parsing involves creating a syntactic tree that represents the grammatical structure of a sentence. This tree helps in understanding the syntactic relationships and dependencies between words.

Sentence Chaining

- **Objective:** Connect and understand the relationships between multiple sentences.
- **How it works:** Algorithms analyze the content and context of different sentences to establish connections or dependencies between them. This step is crucial for tasks that require a broader understanding of the text, such as summarization or document-level sentiment analysis.

Overall, text analytics involves a combination of linguistic rules, machine learning models, and statistical techniques to extract valuable information from text data.

The specific techniques and tools used may vary depending on the application and the complexity of the text analysis task.

Various Text Analytics Techniques

There are numerous applications of text analytics across various industries. Here are some notable examples:

1. **Sentiment Analysis:** Analyzing social media comments, customer reviews, or survey responses to understand and evaluate the sentiment towards a product, brand, or service.
1. **Customer Feedback Analysis:** Extracting valuable insights from customer feedback to identify areas of improvement, track customer satisfaction, and enhance product or service offerings.
1. **Social Media Monitoring:** Monitoring and analyzing social media content to gain insights into public opinions, trends, and reactions related to a particular topic, brand, or event.
1. **Market Research:** Analyzing large volumes of textual data to identify market trends, consumer preferences, and competitive intelligence.
1. **Email Filtering and Classification:** Automatically categorizing and filtering emails based on content, helping in prioritizing and organizing incoming messages.
1. **Content Summarization:** Summarizing lengthy documents, articles, or reports to provide concise and informative summaries for quick understanding.
1. **Chatbot Development:** Implementing natural language processing to develop intelligent chatbots that can understand and respond to user queries in a human-like manner.

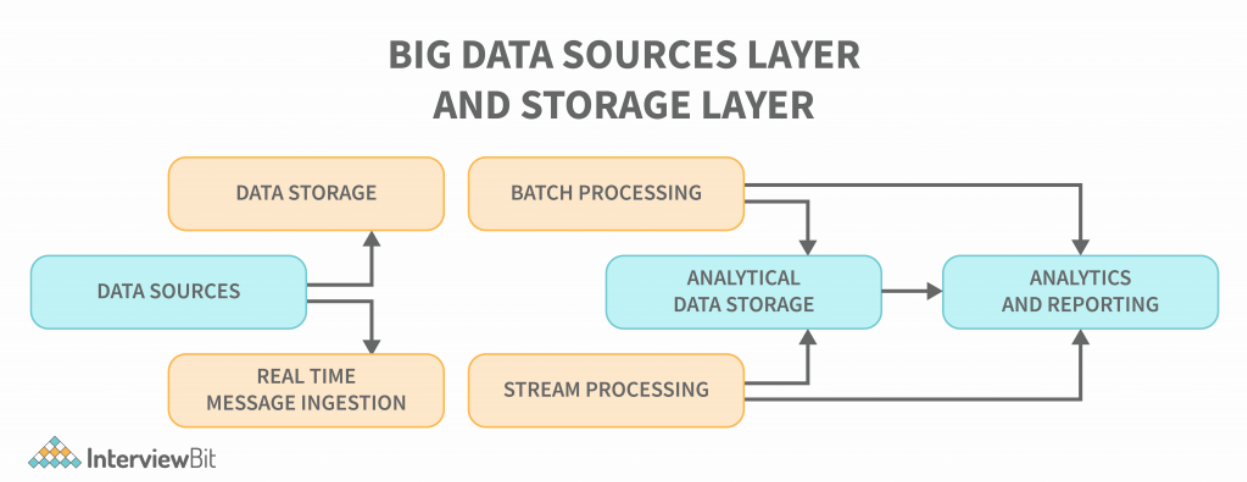
Application of Text Analysis

1. **Legal Document Analysis:** Assisting legal professionals in parsing and summarizing complex legal documents for faster comprehension.
1. **Healthcare Data Insights:** Analyzing medical records and patient feedback to derive insights for improved healthcare services and patient care.
1. **Financial Data Evaluation:** Enhancing fraud detection by scrutinizing large volumes of financial texts to identify irregularities and potential risks.
1. **Educational Content Enhancement:** Improving educational materials by analyzing student feedback and adapting content to better suit learning needs.

Big data Architecture

Big data architecture is a comprehensive solution to deal with an enormous amount of data. It details the blueprint for providing solutions and infrastructure for dealing with big data based on a company's demands. It clearly defines the components, layers, and methods of communication. The reference point is the ingestion, processing, storing, managing, accessing, and analysing of the data. A

big data architecture typically looks like the one shown below, with the following layers:

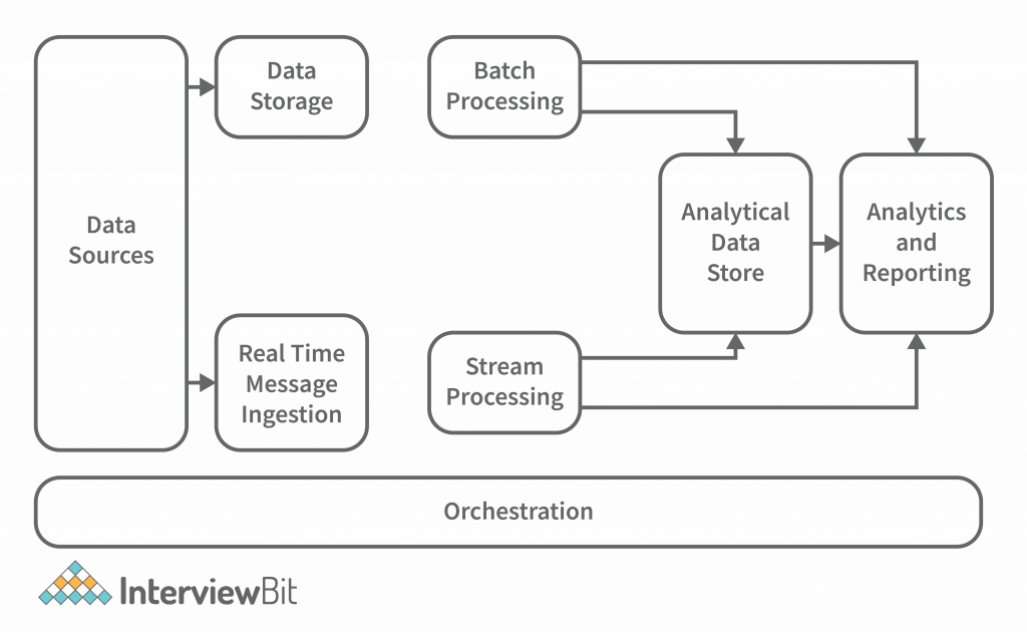


- A database management system that handles data ingestion, processing, and analysis is too complex or too large to handle a traditional architecture. A traditional architecture handles all of that in one fell swoop, while a database management system handles it in chunks.
- Some organisations have threshold values for gigabytes or terabytes, while others, even millions of gigabytes or terabytes, are not good enough.
- As an example, if you look at storage systems and commodity markets, the values and costs of storage have significantly decreased due to this occurrence. There is lots of data that requires different methods to be stored.
- A big data architecture addresses some of these problems by providing a scalable and efficient method of storage and processing data. Some of them are batch-related data that occurs at a particular time and therefore the jobs must be scheduled in the same way as batch data. Streaming class jobs require a real-time streaming pipeline to be built to meet all of their demands. This process is accomplished through big data architecture.

What is Big Data Architecture?

There is more than one workload type involved in big data systems, and they are broadly classified as follows:

1. Merely batching data where big data-based sources are at rest is a data processing situation.
2. Real-time processing of big data is achievable with motion-based processing.
3. The exploration of new interactive big data technologies and tools.
4. The use of machine learning and predictive analysis.



- **Data Sources:** All of the sources that feed into the data extraction pipeline are subject to this definition, so this is where the starting point for the big data pipeline is located. Data sources, open and third-party, play a significant role in architecture. Relational databases, data warehouses, cloud-based data warehouses, SaaS applications, real-time data from company servers and sensors such as IoT devices, third-party data providers, and also static files such as Windows logs, comprise several data sources. Both batch processing and real-time processing are possible. The data managed can be both batch processing and real-time processing.

- **Data Storage:** There is data stored in file stores that are distributed in nature and that can hold a variety of format-based big files. It is also possible to store large numbers of different format-based big files in the data lake. This consists of the data that is managed for batch built operations and is saved in the file stores. We provide HDFS, Microsoft Azure, AWS, and GCP storage, among other blob containers.

- **Batch Processing:** Each chunk of data is split into different categories using long-running jobs, which filter and aggregate and also prepare data for analysis. These jobs typically require sources, process them, and deliver the processed files to new files. Multiple approaches to batch processing are employed, including Hive jobs, U-SQL jobs, Sqoop or Pig and custom map reducer jobs written in any one of the Java or Scala or other languages such as Python.

- **Real Time-Based Message Ingestion:** A real-time streaming system that caters to the data being generated in a sequential and uniform fashion is a batch processing system. When compared to batch processing, this includes all real-time streaming systems that cater to the data being generated at the time it is received. This data mart or store, which receives all incoming messages and discards them into a folder for data processing, is usually the only one that needs to be contacted. Message-based ingestion stores such as Apache Kafka, Apache Flume, Event hubs from Azure, and others, on the other hand, must be used if message-based processing is required. The delivery process, along with other message queuing semantics, is generally more reliable.

- **Stream Processing:** Real-time message ingest and stream processing are different. The latter uses the ingested data as a publish-subscribe tool, whereas the former takes into account all of the ingested data in the first place and then utilises it as a publish-subscribe tool. Stream processing, on the other hand, handles all of that streaming data in the form of windows or streams and writes it to the sink. This includes Apache Spark, Flink, Storm, etc.

- **Analytics-Based Datastore:** In order to analyze and process already processed data, analytical tools use the data store that is based on HBase or any other NoSQL data warehouse technology. The data can be presented with the help of a hive database, which can provide metadata abstraction, or interactive use of a hive database, which can provide metadata abstraction in the data store. NoSQL databases like HBase or Spark SQL are also available.

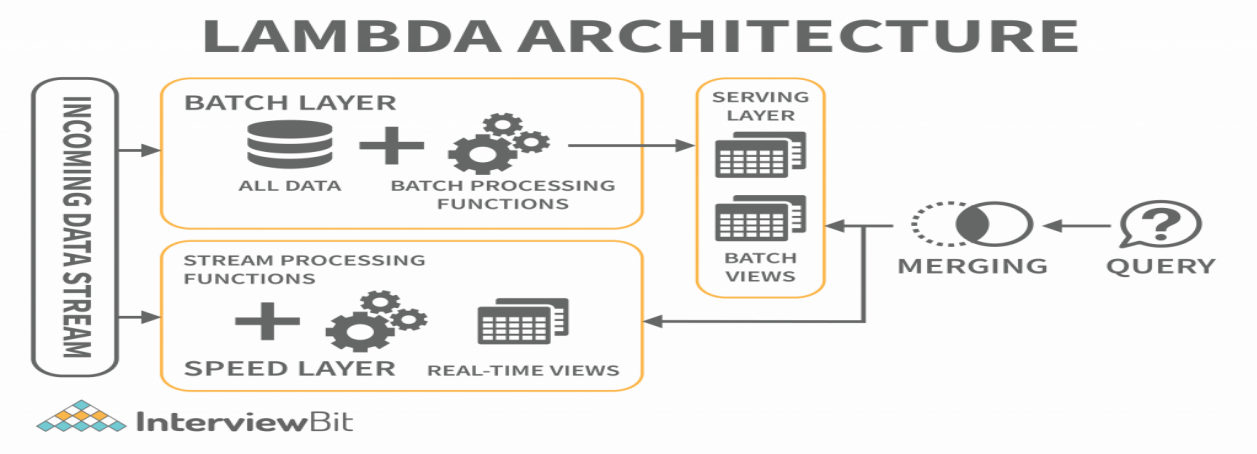
- **Reporting and Analysis:** The generated insights, on the other hand, must be processed and that is effectively accomplished by the reporting and analysis tools that utilize embedded technology and a solution to produce useful graphs, analysis, and insights that are beneficial to the businesses. For example, Cognos, Hyperion, and others.

- **Orchestration:** Data-based solutions that utilise big data are data-related tasks that are repetitive in nature, and which are also contained in workflow chains that can transform the source data and also move data across sources as well as sinks and loads in stores. Sqoop, oozie, data factory, and others are just a few examples.

Types of Big Data Architecture

Lambda Architecture

A single Lambda architecture handles both batch (static) data and real-time processing data. It is employed to solve the problem of computing arbitrary functions. In this deployment model, latency is reduced and negligible errors are preserved while retaining accuracy. The big data architecture illustrated below is similar to that described:



The lambda architecture is comprised of these layers:

- **Batch Layer:** The batch layer of the lambda architecture saves incoming data in its entirety as batch views. The batch views are used to prepare the indexes. The data is immutable, and only copies of the original data are created and preserved. The batch layer ensures consistency by making the data append-only. It is the first layer in the lambda architecture that saves incoming data in its entirety as batch views. The data cannot be

changed, and only copies of the original data are created and preserved. The data that is saved is immutable, meaning that it cannot be changed, and only copies of the original data are preserved and stored. The data that is saved is append-only, which ensures that it is prepared before it is presented. The master dataset and then pre-computing the batch views are handled this way.

- **Speed Layer:** The speed layer delivers data straight to the batch layer, which is responsible for computing incremental data. However, the speed layer itself may also be reduced in latency by reducing the number of computations. The stream layer processes the processed data from the speed layer to produce error correction.

- **Serving Layer:** The batch views and the speed outcomes traverse to the serving layer as a result of the batch layers batch views. The serving layer indexes the views and parallelizes them to ensure users' queries are fast and are exempt from delays.

Kappa Architecture

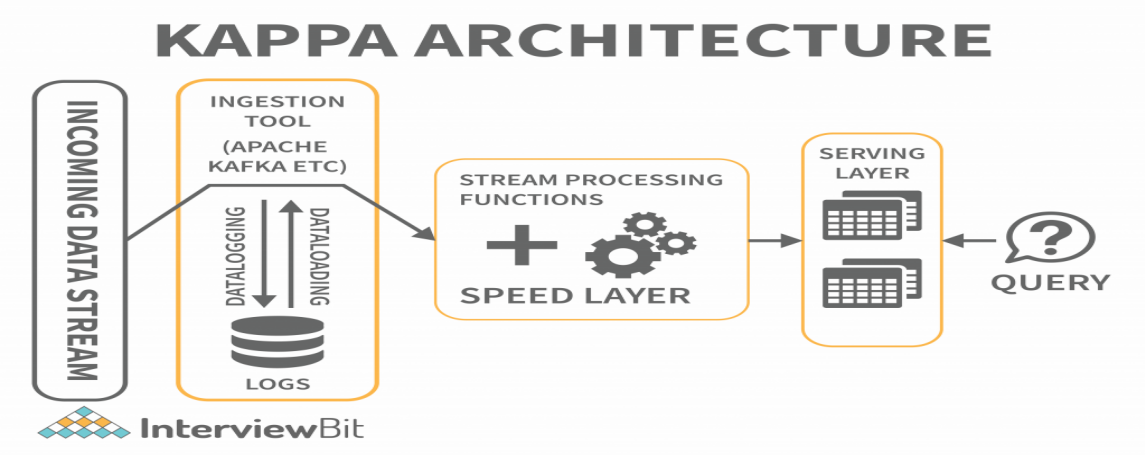
When compared to Lambda architecture, Kappa architecture is also intended to handle both real-time streaming and batch processing data. The Kappa architecture, in addition to reducing the additional cost that comes from the Lambda architecture, replaces the data sourcing medium with message queues.

The messaging engines store a sequence of data in the analytical databases, which are then read and converted into appropriate format before being saved for the end-user.

The architecture makes it easy to access real-time information by reading and transforming the message data into a format that is easily accessible to end users. It also provides additional outputs by allowing previously saved data to be taken into account.

The batch layer was eliminated in the Kappa architecture, and the speed layer was enhanced to provide reprogramming capabilities. The key difference with the Kappa architecture is that all the data is presented as a series or stream. Data

transformation is achieved through the steam engine, which is the central engine for data processing.



Big Data Tools and Techniques

A big data tool can be classified into the four buckets listed below based on its practicability.

1. Massively Parallel Processing (MPP)
2. No-SQL Databases
3. Distributed Storage and Processing Tools
4. Cloud Computing Tools

Massively Parallel Processing (MPP)

A loosely coupled or shared nothing storage system is a massively parallel processing construct with the goal of dividing up a large number of computing machines into discrete pieces and proceeding in parallel. An MPP system is also referred to as a loosely coupled or shared nothing system. Processing is accomplished by breaking a large number of computer processors into separate bits and proceeding in parallel.

Each processor works on separate tasks, has a different operating system, and does not share memory. It is also possible for up to 200 or more processors to work on applications connected to this high-speed network. In each case, the processor handles a different set of instructions and has a different operating system, which is not shared. MPP may also send messages between processes via a messaging system that allows it to send commands to the processors.

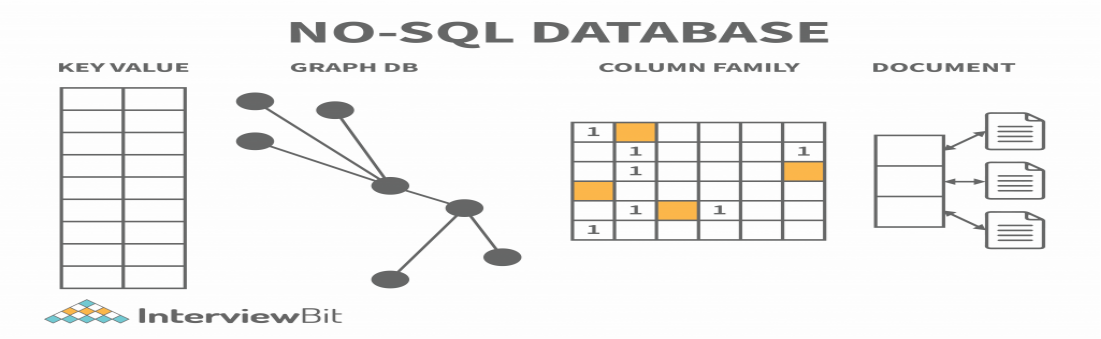
MPP-based databases are IBM Netezza, Oracle Exadata, Teradata, SAP HANA, EMC Greenplum.

No-SQL Databases

Structures are employed to help associate data with a particular domain. Data cannot be stored in a structured database unless it is first converted to one. SQL (or NoSQL) is a non-structured language used to encapsulate unstructured data and create structures for heterogeneous data in the same domain. NoSQL databases offer a vast array of configuration scalability, as well as versatility, and scalability in handling large quantities of data. There is also distributed data storage, making data available locally or remotely.

NoSQL databases include the following categories:

1. Key-value Pair Based
2. Graphs based
3. Column-oriented Graph
4. Document-oriented



Key-value model: Dictionaries, collections, and associative arrays can often use hash tables to store data, but this database stores information in a unique key-value pair. The key is required to access data and the value is used to record information. It helps store data without a schema. The key is unique and used to retrieve and update data, while the value is a string, char, JSON, or BLOB. Redis, Dynamo, and Riak are the key-value store databases.

Graph-based model: Graph databases store both entities and relationships between them, and they are multi-relational. Nodes and links and entities are stored as elements on the graph, and relationships between these elements are represented by edges (or nodes). Graph databases are employed for mapping, transportation, social networks, and spatial data applications. They may also be used to discover patterns in semi-structured and unstructured data. The Neo4J, FlockDB, and OrientDB graph databases are available.

Column-based NoSQL database: Columnar databases work on columns. Compared to relational databases, they have a set of columns rather than tables. Each column is significant, and it is viewed independently. The values in the database are stored in a contiguous manner and may not have values. Because columns are easy to assess, columnar databases are efficient at performing summarisation jobs such as SUM, COUNT, AVG, MIN, and MAX.

Column families are also known as Wide columns or Columnar columns or Column stores. These are used for data structures, business intelligence, CRM, and catalogues of library cards.

The columnar databases Cassandra, HBase, and Hypertable use NoSQL databases that use columnar storage.

Document-Oriented NoSQL database: The document-oriented database stores documents in order to make them essentially document-oriented rather than data-oriented. JSON or XML are the formats used for data, and key-value pairs and the format of JSON or XML are used for data. E-commerce applications, blogging platforms, real-time analytics, Content Management systems (CMS), are among the applications that benefit from these databases.

MongoDB, CouchDB, Amazon SimpleDB, Riak, Lotus Notes. NoSQL document databases are MongoDB, CouchDB, Amazon SimpleDB, Riak, and Lotus Notes.

Distributed Storage and Processing Tools

A distributed database is a set of data storage chunks that is distributed over a network of computers. Data centres may have their own processing units for distributed databases. The distributed databases may be physically located in the same location or dispersed over an interconnected network of computers. The distributed databases are heterogeneous (having a variety of software and hardware), homogeneous (having the same software and hardware across all instances), and different, supported by distinct hardware.

The leading big data processing and distribution platforms are Hadoop HDFS, Snowflake, Qubole, Apache Spark, Azure HDInsight, Azure Data Lake, Amazon EMR, Google BigQuery, Google Cloud Dataflow, MS SQL.

Cloud Computing Tools

Cloud Computing Tools refers to the network-based computing services that utilise the Internet's development and services. The shared pool of configurable computing resources, which are available at any time and anywhere and at any time, are shared by all network-based services. This service is available for paid-for use when required and is provided by the service provider. The platform is very useful in handling large amounts of data.

Amazon Web Services (AWS) is the most popular cloud computing tool, followed by Microsoft Azure, Google Cloud, Blob Storage, and DataBricks. Oracle, IBM, and Alibaba are also popular cloud computing tools.

Big Data Architecture Application

An important aspect of a big data architecture is using and applying big data applications, in particular, the big data architecture utilises and applies big data applications are:

- The data structure of the big data architecture allows deleting of sensitive data right at the beginning because of its data ingesting procedure and because of its data lake storage.

- A batch-or real-time-involving big data architecture ingests data both in the batch and real-time. Batch processing has a frequency and recurring schedule. The ingestion process and the job scheduling for the batch data are simplified as the data files can be partitioned. The query performance is improved by partitioning the tables. Hive, U-SQL, or SQL queries are used to partition the table data.

- Distributed batch files can be split further using parallelism and reduced job time. Another application is to disperse the workload across processing units. The static batch files are created and saved in formats that can be split further. The Hadoop Distributed File System (HDFS) can cluster hundreds of nodes and can parallelly process the files, eventually decreasing job times.

Benefits of Big Data Architecture

- **High-performance parallel computing:** Big data architectures employ parallel computing, in which multiprocessor servers perform lots of calculations at the same time to speed up the process. Large data sets can be processed quickly by parallelising them on multiprocessor servers. Part of the job can be handled simultaneously.

- **Elastic scalability:** Big Data architectures can be scaled horizontally, allowing the environment to be tuned to the size of the workloads. A big data solution is usually operated in the cloud, where you only have to pay for the storage and processing resources you actually utilise.

- **Freedom of choice:** Big Data architectures may use various platforms and solutions in the marketplace, such as Azure-managed services, MongoDB Atlas, and Apache technologies. You can pick the right combination of solutions for your specific workloads, existing systems, and IT expertise levels to achieve the best result.

- **The ability to interoperate with other systems:** You can use Big Data architecture components for IoT processing and BI as well as analytics

workflows to create integrated platforms across different types of workloads.

Big Data Architecture Challenges

- **Security:** When it comes to static big data, the data lake is the norm. Because security is required to safeguard your data from intrusion and theft, robust security is required. In addition, setting up secure access can be difficult. Other applications must also consume data in order for them to function.

- **Complexity:** The moving parts of a Big Data architecture typically consist of many interlocking elements. These components may have their own data-injection pipelines and various configuration settings to improve performance, in addition to many cross-component configuration interventions. Big Data procedures are demanding in nature and require a lot of knowledge and skill.

- **Evolving technologies:** Choosing the right solutions and components is critical to meeting Big Data business objectives. It can be challenging to determine which Big Data technologies, practices, and standards are still in the midst of a period of advancement, as many of them are relatively new and still evolving. Core Hadoop components such as Hive and Pig have reached a stable stage, but other technologies and services are still in development and are likely to change over time.

- **Expertise in a specific domain:** As Big Data APIs built on mainstream languages gradually become popular, we can see Big Data architectures and solutions using atypical, highly specialized languages and frameworks. Nevertheless, Big Data architectures and solutions do generally use unique, highly specialised languages and frameworks that impose a substantial learning curve for developers and data analysts alike.

Applications of big data

1. Tracking Customer Spending Habit, Shopping Behavior: In big retails store (like Amazon, Walmart, Big Bazar etc.) management team has to keep data of customer's spending habit (in which product customer spent, in which brand they wish to spent, how frequently they spent), shopping behavior, customer's

most liked product (so that they can keep those products in the store). Which product is being searched/sold most, based on that data, production/collection rate of that product get fixed.
Banking sector uses their customer's spending behavior-related data so that they can provide the offer to a particular customer to buy his particular liked product by using bank's credit or debit card with discount or cashback. By this way, they can send the right offer to the right person at the right time.

2. Recommendation: By tracking customer spending habit, shopping behavior, Big retails store provide a recommendation to the customer. E-commerce site like Amazon, Walmart, Flipkart does product recommendation. They track what product a customer is searching, based on that data they recommend that type of product to that customer.
As an example, suppose any customer searched bed cover on Amazon. So, Amazon got data that customer may be interested to buy bed cover. Next time when that customer will go to any google page, advertisement of various bed covers will be seen. Thus, advertisement of the right product to the right customer can be sent.

3. Smart Traffic System: Data about the condition of the traffic of different road, collected through camera kept beside the road, at entry and exit point of the city, GPS device placed in the vehicle (Ola, Uber cab, etc.). All such data are analyzed and jam-free or less jam way, less time taking ways are recommended. Such a way smart traffic system can be built in the city by Big data analysis. One more profit is fuel consumption can be reduced.
4. Secure Air Traffic System: At various places of flight (like propeller etc) sensors present. These sensors capture data like the speed of flight, moisture, temperature, other environmental condition. Based on such data analysis, an environmental parameter within flight are set up and varied.
By analyzing flight's machine-generated data, it can be estimated how long the machine can operate flawlessly when it to be replaced/repaired.

5. Auto Driving Car: Big data analysis helps drive a car without human interpretation. In the various spot of car camera, a sensor placed, that gather data like the size of the surrounding car, obstacle, distance from those, etc. These data are being analyzed, then various calculation like how many angles to rotate, what should be speed, when to stop, etc carried out. These calculations help to take action automatically.
6. Virtual Personal Assistant Tool: Big data analysis helps virtual personal assistant tool (like Siri in Apple Device, Cortana in Windows, Google Assistant in Android) to provide the answer of the various question asked by users. This tool tracks the location of the user, their local time, season, other data related to question asked, etc. Analyzing all such data, it provides an answer.

As an 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 analyze these data to conclude if there is a chance of raining, then provide the answer.

7. IoT:

- Manufacturing company install IOT sensor into machines to collect operational data. Analyzing such data, it can be predicted how long machine will work without any problem when it requires repairing so that company can take action before the situation when machine facing a lot of issues or gets totally down. Thus, the cost to replace the whole machine can be saved.
- In the Healthcare field, Big data is providing a significant contribution. Using big data tool, data regarding patient experience is collected and is used by doctors to give better treatment. IoT device can sense a symptom of probable coming disease in the human body and prevent it from giving advance treatment. IoT Sensor placed near-patient, new-born baby constantly keeps track of various health condition like heart bit rate, blood presser, etc. Whenever any parameter crosses the safe limit, an alarm sent to a doctor, so that they can take step remotely very soon.

8. Education Sector: Online educational course conducting organization utilize big data to search candidate, interested in that course. If someone searches for YouTube tutorial video on a subject, then online or offline course provider organization on that subject send ad online to that person about their course.

9. Energy Sector: Smart electric meter read consumed power every 15 minutes and sends this read data to the server, where data analyzed and it can be estimated what is the time in a day when the power load is less throughout the city. By this system manufacturing unit or housekeeper are suggested the time when they should drive their heavy machine in the night time when power load less to enjoy less electricity bill.

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