# ENSEM IMP DATA SCIENCE AND BIG DATA ANALYTICS

## UNIT – 3

**Q.1] What is driving data deluge? Explain with one example.**

**ANS: The data deluge is being driven by several factors, but here are some key points explained in simple terms:**

1. **Increased Connectivity: More devices are connected to the internet than ever before. This includes smartphones, tablets, computers, smartwatches, IoT devices, etc. These devices generate vast amounts of data as they communicate and interact with each other.**
2. **Emergence of IoT: The Internet of Things (IoT) refers to the network of physical objects embedded with sensors, software, and other technologies to connect and exchange data with other devices and systems over the internet. For example, smart thermostats, fitness trackers, and industrial sensors all contribute to the data deluge by constantly collecting and transmitting data.**
3. **Social Media Usage: People are increasingly sharing their lives online through social media platforms like Facebook, Instagram, Twitter, and Snapchat. Every post, photo, like, comment, and share generates data that adds to the deluge.**
4. **Advancements in Technology: Technological advancements have made it easier and cheaper to collect, store, and analyze data. Cloud computing, big data analytics, and machine learning algorithms are just a few examples of technologies that enable organizations to process and derive insights from massive datasets.**
5. **Digital Transformation: Businesses and organizations are undergoing digital transformation to stay competitive in today's fast-paced world. This involves digitizing processes, products, and services, which in turn generates large volumes of data.**
6. **Scientific Research: Scientific research, particularly in fields like genomics, astronomy, and climate science, relies heavily on collecting and analyzing massive datasets. For example, the Human Genome Project generated vast amounts of genetic data that continues to fuel research in genetics and personalized medicine.**

**Example: Imagine a scenario where you have a smart home equipped with various IoT devices such as smart lights, thermostats, security cameras, and voice assistants. Throughout the day, these devices collect data on your energy usage, temperature preferences, security alerts, and voice commands. Additionally, you use your smartphone to check social media, browse the internet, and send messages, generating even more data. All of this data is continuously being transmitted over the internet to cloud servers for storage and analysis. As a result, your smart home setup contributes to the data deluge by generating a constant stream of data points from various sources.**

**Q.2] What is is data science? Differentiate between Business Intelligence and Data Science.**

**ANS: here's a simple breakdown of data science and the difference between data science and business intelligence:**

**Data Science:**

1. **Definition: Data science is a multidisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data.**
2. **Purpose: The primary goal of data science is to uncover hidden patterns, correlations, and trends in data to aid in decision-making, predictive modeling, and optimization.**
3. **Skills Required: Data science requires a blend of skills from various domains, including statistics, mathematics, programming, machine learning, and domain expertise.**
4. **Techniques Used: Data scientists employ a wide range of techniques such as data mining, machine learning, predictive analytics, and natural language processing to extract insights from data.**
5. **Applications: Data science is applied across various industries and domains, including finance, healthcare, e-commerce, marketing, and cybersecurity, to solve complex problems and drive innovation.**

**Business Intelligence (BI):**

1. **Definition: Business intelligence involves the use of software and tools to analyze raw data and transform it into actionable insights for making informed business decisions.**
2. **Purpose: The main objective of business intelligence is to provide historical, current, and predictive views of business operations to support decision-making at all levels of an organization.**
3. **Skills Required: BI professionals typically require skills in data analysis, reporting, data visualization, and domain knowledge related to the business processes they are analyzing.**
4. **Techniques Used: Business intelligence relies on techniques such as data warehousing, online analytical processing (OLAP), dashboards, and ad-hoc reporting to analyze and present data.**
5. **Applications: BI is commonly used for functions such as financial reporting, sales forecasting, market analysis, customer segmentation, and operational optimization within organizations.**

**Difference between Data Science and Business Intelligence:**

1. **Focus: Data science focuses on extracting insights from data to solve complex problems and drive innovation, while business intelligence primarily focuses on providing insights to support operational and strategic decision-making within organizations.**
2. **Approach: Data science involves exploratory data analysis, predictive modeling, and machine learning techniques to uncover patterns and trends in data. In contrast, business intelligence typically involves descriptive analytics and reporting to provide a snapshot of past and current business performance.**
3. **Scope: Data science has a broader scope and can handle both structured and unstructured data, whereas business intelligence primarily deals with structured data stored in databases and data warehouses.**
4. **Skills Required: Data science requires advanced skills in statistics, machine learning, and programming, while business intelligence typically requires proficiency in data analysis, reporting, and data visualization tools.**
5. **Applications: Data science is often used for strategic decision-making and developing innovative products and services, while business intelligence is commonly used for operational reporting, monitoring key performance indicators, and optimizing business processes.**

**Q.3] What are the sources of Big Data. Explain model building phase with example.**
**ANS: here's a simple breakdown of the sources of big data and the model-building phase in data science:**

**Sources of Big Data:**

1. **Social Media: Platforms like Facebook, Twitter, Instagram, and LinkedIn generate vast amounts of data through user interactions, posts, likes, comments, and shares.**
2. **Internet of Things (IoT) Devices: IoT devices such as sensors, wearables, smart appliances, and industrial equipment continuously generate data as they collect information about their environment and communicate with other devices.**
3. **E-commerce Transactions: Online shopping platforms generate large volumes of data through customer transactions, browsing history, product reviews, and recommendations.**
4. **Mobile Devices: Smartphones and tablets produce a significant amount of data through apps, location tracking, call logs, text messages, and app usage.**
5. **Traditional Enterprise Systems: Data from traditional enterprise systems such as customer relationship management (CRM), enterprise resource planning (ERP), and supply chain management (SCM) systems contribute to big data.**
6. **Web and Server Logs: Data from web servers, application servers, and network devices provide insights into user behavior, website traffic, and system performance.**
7. **Genomics and Healthcare: Genomic sequencing, electronic health records (EHRs), medical imaging, and clinical trials generate large datasets in the healthcare industry.**

**Model Building Phase in Data Science:**

1. **Problem Definition: The first step in the model-building phase is to clearly define the problem statement and the objectives of the project. For example, a company may want to build a predictive model to forecast sales for the upcoming year.**
2. **Data Collection and Preparation: Once the problem is defined, the next step is to gather relevant data from various sources. This data may be structured or unstructured and may require cleaning, preprocessing, and transformation to make it suitable for analysis.**
3. **Feature Engineering: Feature engineering involves selecting, extracting, and transforming the most relevant features from the dataset to use as inputs for the model. This may involve techniques such as dimensionality reduction, normalization, and encoding categorical variables.**
4. **Model Selection: In this step, data scientists choose the most appropriate machine learning algorithm or model for the problem at hand. The selection may depend on factors such as the type of data, the complexity of the problem, and the desired output.**

5. **Model Training: Once the model is selected, it is trained using historical data. The data is split into training and testing sets, and the model is trained on the training set to learn the underlying patterns and relationships in the data.**
6. **Model Evaluation: After training, the model is evaluated using the testing set to assess its performance. This may involve metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).**
7. **Hyperparameter Tuning: Hyperparameters are parameters that are set before the learning process begins. In this step, data scientists fine-tune the hyperparameters of the model to improve its performance.**
8. **Model Deployment: Once the model is trained and evaluated, it is deployed into production to make predictions on new, unseen data. This may involve integrating the model into existing systems or applications.**
9. **Monitoring and Maintenance: After deployment, the model is monitored to ensure that it continues to perform accurately over time. It may require periodic updates and retraining to adapt to changes in the data or the business environment.**

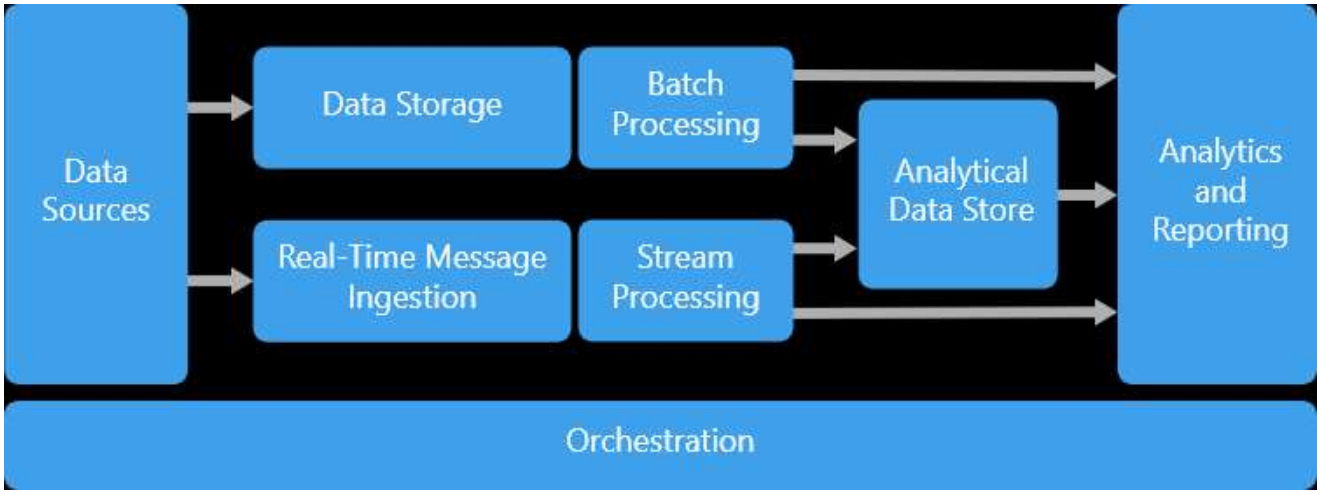
**Q.4] Explain big data analytics architecture with diagram. What is data discovery phase. Explain with example.**

**ANS: here's a simple explanation of big data analytics architecture, followed by an explanation of the data discovery phase:**

**Big Data Analytics Architecture:**

1. **Data Sources: Data originates from various sources such as social media, IoT devices, sensors, e-commerce transactions, and traditional enterprise systems.**
2. **Data Ingestion: Data is ingested into the system from different sources. This may involve extracting data from databases, streaming platforms, APIs, or file systems.**
3. **Data Storage: In the storage layer, data is stored in distributed file systems or databases such as Hadoop Distributed File System (HDFS), NoSQL databases, or data warehouses.**
4. **Data Processing: The processing layer involves transforming and analyzing the data to derive insights. This may include batch processing with tools like Apache Spark or Hadoop MapReduce, as well as real-time processing with technologies like Apache Kafka or Apache Flink.**
5. **Data Analytics: In this layer, advanced analytics techniques such as machine learning, statistical analysis, and predictive modeling are applied to the processed data to extract valuable insights and patterns.**
6. **Data Visualization: The insights obtained from analytics are visualized using dashboards, reports, charts, and graphs to make them easily understandable and actionable for decision-makers.**
7. **Data Exploration and Discovery: Data exploration tools allow users to interactively explore and analyze data to discover hidden patterns, trends, and relationships that may not be immediately apparent.**
8. **Data Governance and Security: Data governance policies and security measures are implemented to ensure the privacy, integrity, and compliance of the data throughout the analytics process.**
9. **Scalability and Flexibility: The architecture is designed to be scalable and flexible, allowing it to handle large volumes of data and adapt to changing business requirements over time.**

**DIAGRAM :**

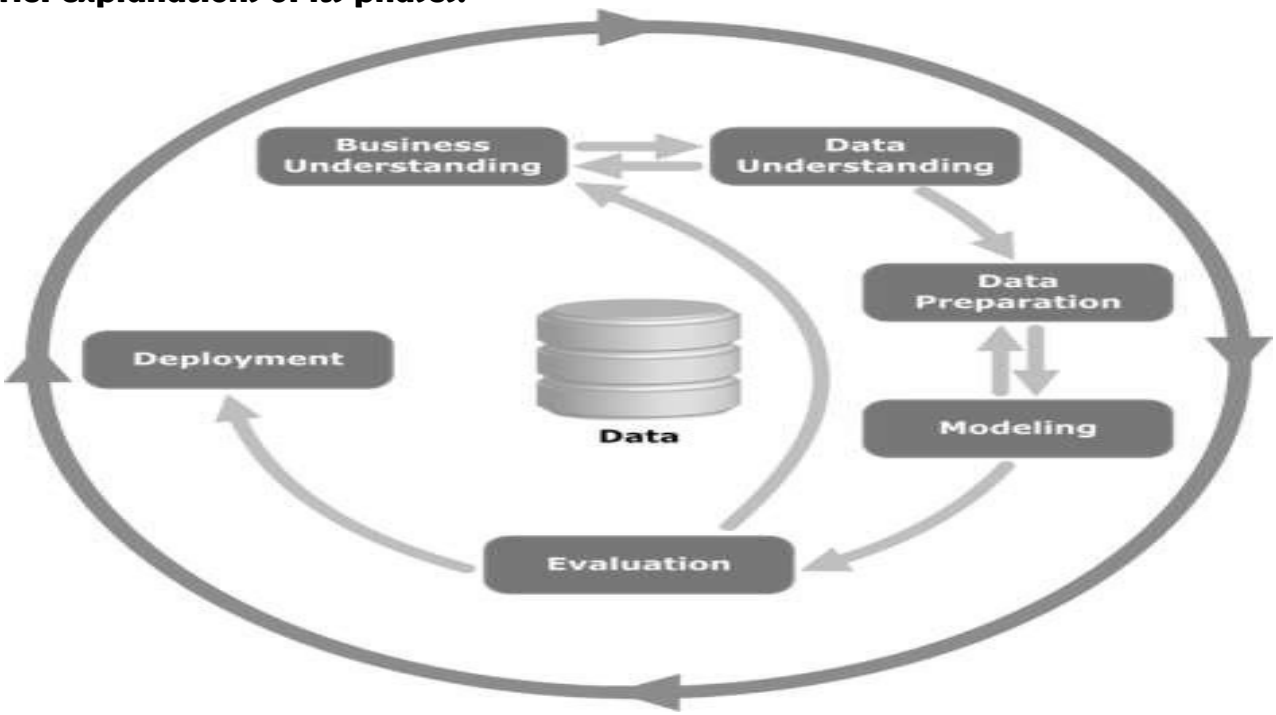


**Data Discovery Phase:**

1. **Definition: The data discovery phase is the initial stage of a data analytics project where data scientists and analysts explore and understand the available data before formal analysis begins.**
2. **Data Exploration: During this phase, data is explored to understand its structure, quality, and characteristics. This may involve examining data distributions, summary statistics, and data profiling.**
3. **Data Profiling: Data profiling techniques are used to analyze the content, structure, and relationships within the data. This helps identify missing values, outliers, duplicates, and other data quality issues.**
4. **Data Visualization: Data visualization tools are used to create visualizations such as histograms, scatter plots, and heatmaps to gain insights into the data and identify patterns or trends.**
5. **Hypothesis Generation: Based on the initial exploration and analysis of the data, hypotheses or initial insights may be generated to guide further analysis and investigation.**
6. **Example: Suppose a retail company wants to analyze its sales data to identify factors influencing sales performance. During the data discovery phase, the data scientists may explore the sales data to understand the distribution of sales across different products, regions, and time periods. They may visualize the data using charts and graphs to identify seasonal trends, popular products, and sales patterns. Data profiling techniques may be used to identify discrepancies or inconsistencies in the data, such as missing values or outliers. Based on their initial exploration, the data scientists may generate hypotheses about factors influencing sales, such as marketing promotions, pricing strategies, or customer demographics, which can be further investigated in subsequent analysis phases.**

**Q.5] Draw the diagram of data analytics life cycle in big data and briefly explain its phases.**

**ANS: Here's a simple diagram of the data analytics lifecycle in big data, along with brief explanations of its phases:**




**Explanation of Phases:**

1. **Data Collection:**
   - **In this phase, data is gathered from various sources such as databases, IoT devices, social media, and sensors.**
   - **The collected data may include structured, semi-structured, and unstructured data.**
2. **Data Preprocessing:**
   - **Data preprocessing involves cleaning, transforming, and organizing the collected data to make it suitable for analysis.**
   - **Tasks may include removing duplicates, handling missing values, and converting data into a standard format.**
3. **Data Analysis:**
   - **This phase involves applying statistical and machine learning techniques to analyze the preprocessed data.**
   - **The goal is to extract meaningful insights, patterns, and trends from the data.**
4. **Data Exploration:**
   - **Data exploration involves visualizing the analyzed data using charts, graphs, and other visualization techniques.**
   - **The purpose is to gain a deeper understanding of the data and identify interesting patterns or anomalies.**
5. **Data Interpretation:**
   - **In this phase, the insights obtained from data analysis and exploration are interpreted to derive actionable conclusions.**
   - **Data interpretation helps stakeholders make informed decisions based on the analytics results.**

**Q.6] Explain in detail how the model building phase is built by team in data analytics life cycle?**

**ANS:**

1. **Problem Definition:**
   - **The team begins by clearly defining the problem they aim to solve or the objective they want to achieve through data analytics. This involves understanding the business requirements and goals.**
2. **Data Preparation:**
   - **The team collects relevant data from various sources and prepares it for analysis. This includes data cleaning, integration, transformation, and feature engineering to make the data suitable for modeling.**
3. **Exploratory Data Analysis (EDA):**
   - **Before building models, the team conducts exploratory data analysis to understand the characteristics and relationships within the data. This involves visualizing data distributions, identifying correlations, and detecting outliers.**
4. **Model Selection:**
   - **Based on the problem definition and data characteristics, the team selects appropriate machine learning algorithms or models to build. This decision depends on factors such as the type of data, the nature of the problem (classification, regression, clustering, etc.), and the desired outcomes.**
5. **Feature Selection:**
   - **The team identifies the most relevant features or variables from the dataset that are likely to have the most significant impact on the model's performance. Feature selection techniques such as correlation analysis, feature importance ranking, and domain expertise are used.**
6. **Model Training:**
   - **In this phase, the selected model is trained on the prepared dataset. The dataset is typically split into training and testing sets to evaluate the model's performance. The model learns from the training data to identify patterns and relationships.**
7. **Hyperparameter Tuning:**
   - **The team fine-tunes the model's hyperparameters to optimize its performance. This involves experimenting with different parameter values and evaluating the model's performance using techniques like cross-validation.**
8. **Model Evaluation:**
   - **After training and tuning the model, the team evaluates its performance using appropriate evaluation metrics. Common metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), depending on the nature of the problem.**
9. **Validation and Deployment:**
   - **Once the model meets the desired performance criteria, it undergoes validation to ensure its effectiveness and reliability. If validated successfully, the model is deployed into production for real-world use. This involves integrating the model into existing systems or applications to make predictions or recommendations.**

**Q.7] List and explain the steps in data preparation phase of data analytics life cycle.**

**ANS: here are the steps in the data preparation phase of the data analytics life cycle explained in simple points:**

1. **Data Collection:**
   - **Gather relevant data from various sources such as databases, files, APIs, or sensors.**
   - **Ensure the data collected is comprehensive and represents the problem domain adequately.**
2. **Data Cleaning:**
   - **Identify and handle missing or null values in the dataset.**
   - **Remove duplicates to avoid skewing analysis results.**
   - **Standardize formats and correct errors in data entries.**
3. **Data Transformation:**
   - **Convert data into a suitable format for analysis, such as numerical or categorical.**
   - **Normalize or scale numerical data to bring them within a similar range.**
   - **Encode categorical variables into numerical representations using techniques like one-hot encoding.**
4. **Feature Engineering:**
   - **Create new features from existing ones to enhance the predictive power of the model.**
   - **Select relevant features that contribute most to the analysis and discard irrelevant ones to reduce dimensionality.**
5. **Data Integration:**
   - **Combine data from multiple sources into a single dataset for comprehensive analysis.**
   - **Ensure consistency and compatibility of data formats and structures during integration.**
6. **Data Reduction:**
   - **Reduce the volume of data while preserving its informational content.**
   - **Techniques such as dimensionality reduction or sampling may be used to achieve this.**
7. **Data Formatting:**
   - **Format the data according to the requirements of the analytics tools or algorithms being used.**
   - **Ensure that the data is in a format that can be easily processed and analyzed further.**
8. **Data Splitting:**
   - **Split the dataset into training, validation, and testing sets to evaluate the performance of the analytics model.**
   - **Typically, the data is divided into subsets such as 70% for training, 15% for validation, and 15% for testing.**

**Q.8] Write short note on the following:**
**i) ETL**
**ii) Common tools for the model building.**
**iii) Model selection for data analytics.**
**ANS: Here are short notes on each of the given topics:**

**i) ETL (Extract, Transform, Load):**

1. **Extract: Involves retrieving data from various sources such as databases, files, or APIs.**
2. **Transform: Data undergoes cleaning, integration, and transformation processes to ensure consistency and usability.**
3. **Load: The processed data is loaded into a target database or data warehouse for analysis and reporting.**
4. **ETL plays a crucial role in data warehousing and business intelligence by facilitating the movement of data from disparate sources to a centralized repository.**

**ii) Common Tools for Model Building:**

1. **Python Libraries: Popular libraries like scikit-learn, TensorFlow, and Keras offer extensive functionalities for building machine learning and deep learning models.**
2. **R Programming: R provides a wide range of packages such as caret, glmnet, and randomForest for statistical modeling and machine learning.**
3. **KNIME: A visual data analytics platform that allows users to build and execute machine learning workflows without writing code.**
4. **Apache Spark: Distributed computing framework with MLlib library for scalable machine learning tasks.**
5. **Microsoft Azure ML: Cloud-based platform offering tools and services for building, training, and deploying machine learning models.**

**iii) Model Selection for Data Analytics:**

1. **Define Objectives: Clearly define the goals and objectives of the analysis to guide the model selection process.**
2. **Explore Data: Understand the characteristics of the dataset and identify the problem type (e.g., classification, regression, clustering).**
3. **Select Algorithms: Choose appropriate algorithms based on the problem type, data characteristics, and computational resources available.**
4. **Evaluate Performance: Use metrics such as accuracy, precision, recall, or F1 score to evaluate the performance of different models.**
5. **Iterate and Refine: Experiment with different algorithms and hyperparameters, and refine the model based on performance feedback.**
6. **Consider Trade-offs: Balance between model complexity and interpretability, as well as computational resources and performance.**
7. **Validation: Validate the selected model using cross-validation or holdout validation to ensure its generalization to unseen data.**
8. **Deployment: Deploy the final model into production environment for real-world application and decision-making.**

**Q.9] What is Model Building elaborate this phase of data analytics with the help of suitable example.**

**ANS: Here's an explanation of the model building phase of data analytics with the help of suitable examples, presented in easy and simple points:**

1. **Definition of Goals:**
   - **Before starting the model building phase, it's crucial to define clear objectives and goals that the model aims to achieve. For example, in a retail business, the goal might be to predict customer churn to improve customer retention strategies.**
2. **Data Preparation:**
   - **This phase involves collecting, cleaning, and transforming data to make it suitable for model building. For instance, in our retail example, data about customer demographics, purchase history, and interactions might be collected and preprocessed.**
3. **Feature Selection/Engineering:**
   - **Identify and select relevant features or variables from the dataset that can significantly impact the model's performance. In our retail example, features like frequency of purchases, average purchase amount, and customer loyalty status could be important predictors of churn.**
4. **Algorithm Selection:**
   - **Choose appropriate algorithms based on the problem type and dataset characteristics. For instance, for predicting customer churn, classification algorithms like logistic regression, decision trees, or random forests might be suitable.**
5. **Model Training:**
   - **Use the prepared data to train the selected algorithms. This involves feeding the algorithm with input data and corresponding output labels (in supervised learning) to learn patterns and relationships. For example, the model would learn from historical data which customer characteristics are indicative of churn.**
6. **Hyperparameter Tuning:**
   - **Fine-tune the model's hyperparameters to optimize its performance. Hyperparameters are parameters that are set before the learning process begins, such as the learning rate or the number of trees in a random forest. This process involves experimentation to find the best combination of hyperparameters that maximizes the model's accuracy or other performance metrics.**
7. **Model Evaluation:**
   - **Assess the performance of the trained model using evaluation metrics such as accuracy, precision, recall, or F1 score. This step helps determine how well the model generalizes to new, unseen data. For instance, in our retail example, the model's accuracy in predicting customer churn can be evaluated using a holdout dataset.**
8. **Validation and Testing:**
   - **Validate the model's performance on independent datasets to ensure its reliability and generalizability. This typically involves splitting the data into training and testing sets or using techniques like cross-validation.**

9. **Deployment:**
    - **Once the model has been trained, evaluated, and validated, it can be deployed into production for real-world use. In our retail example, the churn prediction model could be integrated into the customer relationship management (CRM) system to identify at-risk customers and implement targeted retention strategies.**

**Q.10] Explain any three sources of Big Data. Differentiate BI versus Data science.**

**ANS: Three Sources of Big Data:**

1. **Social Media Data:**
   - **Data generated from social media platforms such as Facebook, Twitter, and Instagram.**
   - **Includes text, images, videos, likes, shares, comments, and other interactions.**
   - **Used for sentiment analysis, customer behavior analysis, and targeted marketing.**
2. **Sensor Data:**
   - **Data collected from various sensors embedded in devices, machinery, or IoT (Internet of Things) devices.**
   - **Examples include temperature sensors, GPS trackers, accelerometers, and RFID tags.**
   - **Used for monitoring and optimizing processes in industries like manufacturing, healthcare, and transportation.**
3. **Transactional Data:**
   - **Data generated from transactions occurring in business operations, such as sales transactions, financial transactions, or online purchases.**
   - **Includes information about products, customers, prices, quantities, and timestamps.**
   - **Used for business analytics, forecasting, fraud detection, and customer segmentation.**

**Difference between Business Intelligence (BI) and Data Science:**

1. **Scope and Purpose:**
   - **BI focuses on analyzing past and current data to provide insights for making informed business decisions.**
   - **Data Science encompasses a broader scope, including predictive analytics, machine learning, and statistical analysis, to uncover hidden patterns and insights in data and make future predictions.**
2. **Tools and Techniques:**
   - **BI primarily uses reporting tools, dashboards, and OLAP (Online Analytical Processing) for data visualization and analysis.**
   - **Data Science employs advanced statistical and machine learning algorithms, programming languages like Python and R, and big data technologies such as Hadoop and Spark for analyzing large datasets and building predictive models.**
3. **Data Types and Sources:**
   - **BI mainly deals with structured data from traditional databases and data warehouses.**
   - **Data Science handles both structured and unstructured data from diverse sources such as social media, sensors, and web logs.**
4. **Time Horizon:**
   - **BI provides insights into historical and current data to support operational and strategic decision-making.**
   - **Data Science focuses on analyzing historical data to make predictions about future trends and outcomes.**

**Q.11] What are the three characteristic of Big Data and what are the main consideration in processing Big Data.**

**ANS: Here are the three characteristics of Big Data and the main considerations in processing Big Data, presented in easy and simple points:**

**Three Characteristics of Big Data:**

1. **Volume:**
   - **Refers to the vast amount of data generated and collected from various sources.**
   - **Traditional data storage and processing systems may struggle to handle the sheer volume of data.**
   - **Examples include sensor data, social media posts, transaction records, and multimedia content.**
2. **Velocity:**
   - **Describes the speed at which data is generated, collected, and processed.**
   - **Big Data sources often produce data at high velocity, requiring real-time or near-real-time processing.**
   - **Examples include streaming data from IoT devices, social media feeds, and financial transactions.**
3. **Variety:**
   - **Indicates the diversity of data types and formats that Big Data encompasses.**
   - **Data can be structured, semi-structured, or unstructured, and may include text, images, videos, audio, and sensor readings.**
   - **Managing and analyzing such diverse data types pose significant challenges for traditional data processing systems.**

**Main Considerations in Processing Big Data:**

1. **Scalability:**
   - **Processing Big Data requires scalable systems capable of handling increasing volumes of data without compromising performance.**
   - **Distributed computing frameworks like Hadoop and Apache Spark enable parallel processing of data across multiple nodes in a cluster, ensuring scalability.**
2. **Performance:**
   - **Efficient processing of Big Data necessitates high-performance computing resources and optimized algorithms.**
   - **Techniques such as data partitioning, parallel processing, and in-memory computing help improve processing speed and performance.**
3. **Data Quality and Cleansing:**
   - **Ensuring data quality is crucial in Big Data processing to avoid inaccuracies and biases in analysis results.**
   - **Data cleansing techniques, such as deduplication, outlier detection, and error correction, are essential for enhancing data quality and reliability.**
4. **Security and Privacy:**
   - **Big Data processing involves sensitive and confidential information, making security and privacy paramount concerns.**
   - **Implementing robust security measures, access controls, encryption, and anonymization techniques help protect data from unauthorized access and breaches.**

5. **Cost-Efficiency:**
    - **Processing and storing Big Data can be costly, especially in cloud computing environments.**
    - **Optimizing resource utilization, adopting cost-effective storage solutions, and implementing efficient data processing pipelines are essential for minimizing costs while maximizing performance.**

**Q.12] Explain Descriptive, Diagnostic, Predictive analytics**
**ANS: Here's an explanation of descriptive, diagnostic, and predictive analytics, presented in easy and simple points:**

1. **Descriptive Analytics:**
    - **Descriptive analytics focuses on summarizing historical data to understand past events and trends.**
    - **It answers the question: "What happened?"**
    - **Descriptive analytics techniques include data aggregation, visualization, and statistical analysis.**
    - **Example: Analyzing sales data to identify the best-selling products, sales trends over time, or geographical distribution of customers.**

2. **Diagnostic Analytics:**
    - **Diagnostic analytics aims to uncover the reasons behind past events or trends by analyzing historical data.**
    - **It answers the question: "Why did it happen?"**
    - **Diagnostic analytics involves deeper analysis and exploration of data to identify patterns, correlations, and causal relationships.**
    - **Example: Investigating a decrease in website traffic by analyzing user behavior data, traffic sources, and website performance metrics to determine the root cause of the decline.**

3. **Predictive Analytics:**
    - **Predictive analytics focuses on forecasting future outcomes or trends based on historical data and statistical models.**
    - **It answers the question: "What is likely to happen?"**
    - **Predictive analytics leverages techniques such as regression analysis, machine learning, and artificial intelligence to build predictive models.**
    - **Example: Using historical sales data, customer demographics, and market trends to predict future sales revenue, customer churn, or demand for products.**

**Q.13] and Explain the various activities involved in identifying potential data resources as a part of discovery phase in Data Analytics Life Cycle?**

**ANS: Here's a simple explanation of the various activities involved in identifying potential data resources as part of the discovery phase in the Data Analytics Life Cycle, presented in easy and simple points:**

1. **Define Objectives:**
   - **Clearly define the goals and objectives of the data analytics project to guide the identification of relevant data resources.**
2. **Identify Stakeholders:**
   - **Identify key stakeholders, including business owners, subject matter experts, and data owners, to understand their requirements and perspectives on data resources.**
3. **Gather Requirements:**
   - **Collect requirements from stakeholders regarding the types of data needed, the scope of analysis, and the expected outcomes of the project.**
4. **Understand the Problem Domain:**
   - **Gain a deep understanding of the problem domain and the context in which the data analytics project will operate. This includes understanding the industry, business processes, and relevant regulations.**
5. **Inventory Existing Data Sources:**
   - **Conduct a comprehensive inventory of existing data sources within the organization, including databases, data warehouses, spreadsheets, and other repositories.**
6. **Explore External Data Sources:**
   - **Identify potential external data sources such as public datasets, industry databases, open data repositories, and data marketplaces that may contain relevant information for analysis.**
7. **Assess Data Quality:**
   - **Evaluate the quality of potential data resources by assessing factors such as completeness, accuracy, consistency, and timeliness. This may involve conducting data profiling and quality checks.**
8. **Consider Data Accessibility:**
   - **Assess the accessibility of data resources in terms of availability, permissions, and legal constraints. Ensure that the data can be obtained and used ethically and legally.**
9. **Prioritize Data Resources:**
   - **Prioritize data resources based on their relevance to the project objectives, quality, accessibility, and potential impact on analysis outcomes.**
10. **Document Findings:**
    - **Document the findings of the data resource identification process, including a list of identified data sources, their characteristics, quality assessments, and any constraints or limitations.**

**Q.14] List and explain the key roles for successful analytics project.**
**ANS: Here are the key roles for a successful analytics project, explained in easy and simple points:**

1. **Project Manager:**
   - **Oversees the entire analytics project, including planning, execution, and delivery.**
   - **Coordinates activities among team members, manages timelines and budgets, and communicates with stakeholders.**
2. **Data Analyst:**
   - **Responsible for collecting, cleaning, analyzing, and interpreting data to extract meaningful insights.**
   - **Utilizes statistical and analytical tools to identify patterns, trends, and correlations in the data.**
3. **Data Scientist:**
   - **Applies advanced statistical and machine learning techniques to solve complex analytical problems.**
   - **Develops predictive models, algorithms, and data-driven solutions to address business challenges.**
4. **Subject Matter Expert (SME):**
   - **Provides domain-specific knowledge and expertise to guide the analytics project.**
   - **Helps define project goals, identify relevant data sources, and interpret analysis results in the context of the business.**
5. **Data Engineer:**
   - **Designs and implements data pipelines to collect, process, and transform raw data into usable formats for analysis.**
   - **Manages data infrastructure, including databases, data warehouses, and big data platforms.**
6. **Business Analyst:**
   - **Bridges the gap between technical analysis and business requirements.**
   - **Translates business needs into analytics requirements, defines KPIs, and communicates insights to business stakeholders.**
7. **Visualization Specialist:**
   - **Creates visually compelling and informative data visualizations to communicate analysis results effectively.**
   - **Uses tools like Tableau, Power BI, or Python libraries to design dashboards, charts, and reports.**
8. **Project Sponsor/Stakeholder:**
   - **Provides leadership, support, and resources for the analytics project.**
   - **Defines project objectives, priorities, and success criteria, and ensures alignment with organizational goals.**

**Q.15] Write short note on :**
**i) Common Tools for the Model Building**
**ANS:**
**i) Common Tools for Model Building:**

1. **Python Libraries: Popular libraries like scikit-learn, TensorFlow, and Keras offer extensive functionalities for building machine learning and deep learning models.**
2. **R Programming: R provides a wide range of packages such as caret, glmnet, and randomForest for statistical modeling and machine learning.**
3. **KNIME: A visual data analytics platform that allows users to build and execute machine learning workflows without writing code.**
4. **Apache Spark: Distributed computing framework with MLlib library for scalable machine learning tasks.**
5. **Microsoft Azure ML: Cloud-based platform offering tools and services for building, training, and deploying machine learning models.**