DSBDA UNIT-3 PYQ'S

➢ MAY / JUN 2022

Q1)

a) What is driving data deluge? Explain with one example. [9]

A **data deluge** refers to the rapid and overwhelming growth of data generated from various digital sources.

It is primarily driven by advancements in technology and increased digital activity in our daily lives.

Key Drivers of Data Deluge:

1. **Proliferation of IoT Devices:**
    - Billions of sensors and connected devices continuously collect and transmit data (e.g., smart homes, wearables, cars).
2. **Social Media Explosion:**
    - Platforms like Facebook, Twitter, Instagram generate petabytes of data daily through posts, likes, shares, images, and videos.
3. **Mobile Device Usage:**
    - Smartphones generate high volumes of location data, browsing history, app usage data, etc.
4. **Cloud Computing & Online Services:**
    - Services like Google Drive, YouTube, Netflix contribute heavily to the volume of stored and streamed data.
5. **E-commerce & Online Transactions:**
    - Websites like Amazon, Flipkart generate user data via clicks, purchases, reviews, and behavioral analytics.
6. **Scientific Research & Healthcare:**
    - Genomics, medical imaging, and patient records contribute significantly to big data in healthcare.

Example:

- **Example – Smart City Project:**
    - In smart cities, sensors placed in traffic signals, pollution monitors, surveillance cameras, and public transport continuously collect real-time data.
    - This data helps manage traffic flow, monitor air quality, and ensure public safety.
    - Such initiatives can generate **terabytes of data daily**, showing how urban development contributes to the data deluge.

The data deluge is a result of the digital transformation of industries and personal life.

Efficient handling and analysis of this data is crucial for informed decision-making, innovation, and optimization in various sectors.

b) What is data science? Differentiate between Business Intelligence and Data Science. [9]

Data Science:

- **Data Science** is an interdisciplinary field that uses techniques from statistics, computer science, and domain knowledge to extract meaningful insights and knowledge from structured and unstructured data.
- It involves **data collection, cleaning, exploration, modeling**, and **decision-making** using advanced algorithms.

Difference between Business Intelligence (BI) and Data Science:

Aspect	Data Science	Business Intelligence (BI)
Purpose	Predicts future trends and outcomes using data modeling.	Analyzes **historical data** to describe business performance.
Approach	Predictive and prescriptive analytics.	Descriptive and diagnostic analytics.
Tools Used	Python, R, Hadoop, TensorFlow, Scikit-learn.	SQL, Excel, Tableau, Power BI.
Techniques	Machine Learning, Data Mining, AI, Statistics.	Reporting, Dashboards, OLAP.
User	Data scientists, analysts, engineers.	Business analysts, managers.
Data Type	Structured, semi-structured, and unstructured data.	Mainly structured data.
Output	Predictive models, recommendations, actionable insights.	Reports, summaries, visual dashboards.

MORE POINTS :

S. No.	Factor	Data Science	Business Intelligence
1.	Concept	It is a field that uses mathematics, statistics and various other tools to discover the hidden patterns in the data.	It is basically a set of technologies, applications and processes that are used by the enterprises for business data analysis.
2.	Focus	It focuses on the future.	It focuses on the past and

			present.
3.	Data	It deals with both structured as well as unstructured data.	It mainly deals only with structured data.
4.	Flexibility	Data science is much more flexible as data sources can be added as per requirement.	It is less flexible as in case of business intelligence data sources need to be pre-planned.
5.	Method	It makes use of the scientific method.	It makes use of the analytic method.
6.	Complexity	It has a higher complexity in comparison to business intelligence.	It is much simpler when compared to data science.
7.	Expertise	It's expertise is data scientist.	It's expertise is the business user.

Q2)

a) What are the sources of Big Data. Explain model building phase with example. [9]

Sources of Big Data :

- **Social Media:**
  Generates massive user data (e.g., Facebook – 500+ TB/day from posts, videos, messages).
- **Stock Exchange:**
  Daily trade data of companies and users in terabytes; used in financial analysis.
- **Aviation Industry:**
  Jet engines generate ~10 TB per 30-min flight; used for predictive maintenance.
- **Survey Data:**
  Large volumes from online/offline surveys; useful in market research.
- **Compliance Data:**
  Data from hospitals, banks for legal/audit purposes (e.g., health records, financial reports).

Model Building Phase (Phase 4 of Data Analytic Lifecycle):

Definition:

- In this phase, **actual model development takes place** using the techniques chosen in the model planning phase.
- Data scientists **train, test, and validate models** using machine learning or statistical methods.

Steps Involved:

1. **Data Collection & Preprocessing:**
    - Gather data from relevant sources (e.g., customer purchase history).
    - Clean data (handle missing values, remove duplicates).
    - Normalize/transform data (scaling, encoding categorical variables).
2. **Feature Selection & Engineering**:
    - Identify key variables (e.g., age, location, purchase frequency).
    - Create new features (e.g., "average spending per month").
3. **Model Selection & Training:**
    - Choose an algorithm (e.g., Random Forest for classification).
    - Split data into training & testing sets.
    - Train the model on historical data.
4. **Model Evaluation & Optimization:**
    - Test performance using metrics (accuracy, precision, recall).
    - Fine-tune hyperparameters (e.g., tree depth in Random Forest).
5. **Deployment & Monitoring:**
    - Deploy the model in production (e.g., recommendation engine).
    - Continuously monitor and retrain as needed.

Example:

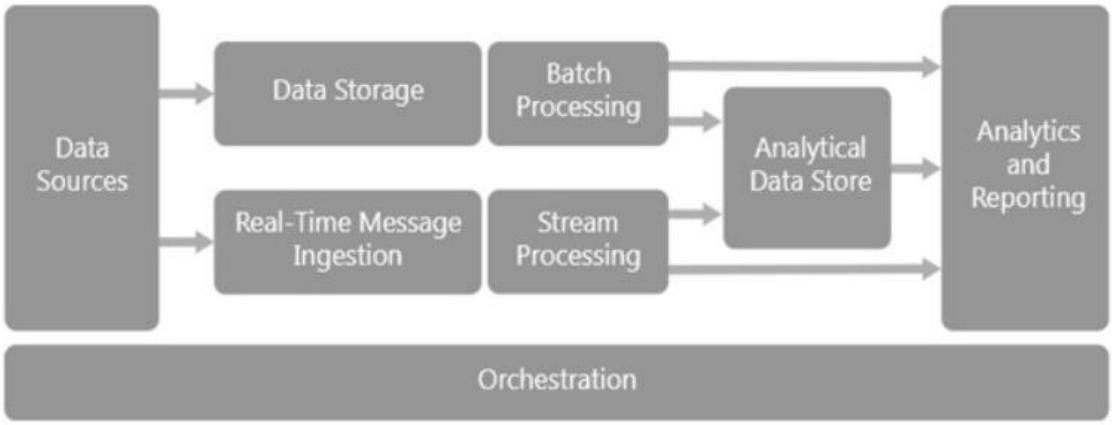
Predictive Maintenance in Manufacturing

Problem: A factory wants to predict machine failures using sensor data.

- **Data Sources:** IoT sensors (vibration, temperature logs), maintenance records.
- **Model Building Steps:**
    1. **Preprocess:** Remove noise, align time-series data.
    2. **Feature Engineering:** Extract rolling averages, peak values.
    3. **Model Training:** Use a Random Forest to classify "failure" vs "normal."

4. **Evaluation:** Achieves 95% accuracy in predicting breakdowns.
5. **Deployment:** Alerts technicians before failures occur.

b) Explain big data analytics architecture with diagram. What is data discovery phase. Explain with example. [9]



Big Data Analytics Architecture

Big Data Analytics architecture organizes the flow of data from its source to actionable insights through several key components:

Components:

1. **Data Sources**
   - Raw data originates from multiple sources like databases, IoT devices, APIs, social media, etc.
2. **Data Storage**
   - This includes **Data Lakes** (storing raw, unstructured data) and **Data Warehouses** (storing cleaned and structured data).
3. **Batch & Stream Processing**
   - Frameworks and tools to process data in large batches or real-time streams (e.g., **Hadoop** for batch, **Spark** for in-memory processing, **Kafka** for streaming).
4. **Analytical Data Store**
   - Specialized storage like **OLAP databases** or **data marts** optimized for fast querying and multidimensional analysis.
5. **Analytics & Reporting**
   - BI tools and dashboards that visualize processed data, generate reports, and enable decision-making.
6. **Orchestration**
   - Tools like **Apache Airflow** or **Kubernetes** that manage and schedule workflows, ensuring smooth, automated data pipeline operations.

Data Discovery Phase

The **Data Discovery Phase** is the initial step in the analytics process where data scientists or analysts explore and understand the raw data to prepare it for deeper analysis.

- **Goal:** Understand data structure, quality, and key attributes.
- **Activities:** Data profiling, identifying missing or inconsistent data, detecting relationships, and generating initial visualizations.

Example of Data Discovery Phase

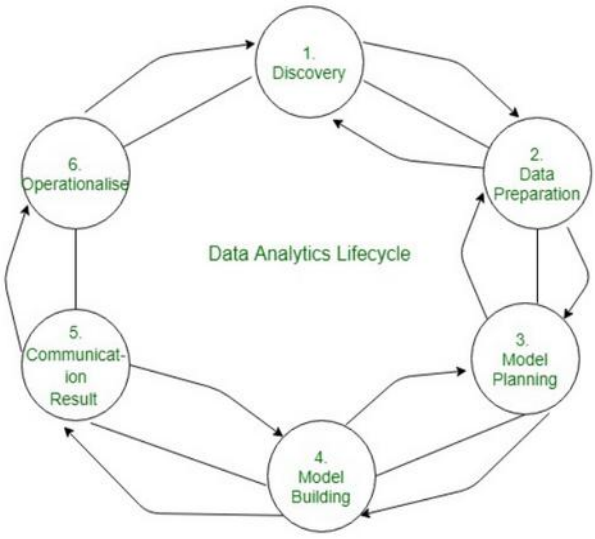
Suppose a transportation company wants to analyze vehicle sensor data collected via IoT devices.

- In the **data discovery phase**, they:
  - Profile sensor data streams to check for missing readings or irregular patterns.
  - Visualize speed, fuel consumption, and engine temperature distributions.
  - Detect anomalies such as sudden spikes in temperature.
  - Identify relevant sensor features for predicting maintenance needs.

➢ NOV / DEC 2022

Q1)

a) Draw the diagram of data analytics life cycle in big data and briefly explain its phases. [8]



Phases of Data Analytics Life Cycle:

1. Discovery:

- The data science team understands the problem context and investigates the project requirements.
- They identify relevant data sources and formulate initial hypotheses to be tested later with data.

2. Data Preparation:

- This phase involves exploring, cleaning, and transforming raw data to prepare it for analysis.
- Data is loaded into an analytic sandbox where multiple iterations of cleaning and transformation occur.
- Common tools: Hadoop, Alpine Miner, OpenRefine.

3. Model Planning:

- The team analyzes relationships among variables and selects important features.
- They decide on suitable modeling techniques and prepare datasets for training, testing, and production.
- Common tools: MATLAB, STASTICA.

4. Model Building:

- The actual models are developed and tested using the prepared datasets.
- Teams evaluate whether existing tools suffice or if more robust environments are needed.
- Tools include open-source ones like R, Octave, WEKA, and commercial software such as MATLAB.

5. Communicating Results:

- Model outcomes are compared against success criteria.
- The team summarizes key findings, quantifies business impact, and prepares presentations or reports for stakeholders.
- Clear communication ensures understanding of assumptions, limitations, and insights.

6. Operationalize:

- The validated model is deployed in a controlled pilot environment to assess real-world performance.
- Based on feedback, adjustments are made before full-scale deployment.
- Deliverables include final reports, codes, and documentation.
- Tools like Octave, WEKA, SQL, MADlib are used.

b) Explain in detail how the model-building phase is executed by a team in the data analytics lifecycle.

The **model-building phase** is a critical stage in the **data analytics lifecycle**, where a cross-functional team collaborates to develop, train, and validate predictive or analytical models. Below is a **detailed breakdown** of how teams execute this phase

Team Roles in Model Building :

- **Data Engineers** (Prepare data pipelines)
- **Data Scientists** (Develop & train models)
- **ML Engineers** (Optimize & deploy models)
- **Business Analysts** (Define problem & success metrics)
- **Domain Experts** (Provide industry insights)

Step-by-Step Model-Building Process

1: Problem Definition & Data Collection

- **Business Analysts & Domain Experts** define the problem (e.g., "Predict customer churn").
- **Data Engineers** gather structured (SQL databases) and unstructured (logs, social media) data.

2: Data Preprocessing & Feature Engineering

- **Data Engineers & Scientists** clean data (handle missing values, remove duplicates).
- **Feature Engineering:** Transform raw data into meaningful features (e.g., "Average session duration" for churn prediction).

3: Model Selection & Training

- **Data Scientists** experiment with algorithms:
    - **Supervised Learning** (Regression, Classification)
    - **Unsupervised Learning** (Clustering, Dimensionality Reduction)
    - **Deep Learning** (Neural Networks for complex patterns)
- **Train-Test Split:** Data is divided into training (70-80%) and testing (20-30%) sets.

4: Model Evaluation & Optimization

- **Performance Metrics:**
    - Classification: Accuracy, Precision, Recall, F1-Score
    - Regression: RMSE, MAE, R^2

- **Hyperparameter Tuning:** Adjust parameters (e.g., learning rate, tree depth) for better accuracy.

5: Model Deployment & Monitoring

- **ML Engineers** deploy models using:
    - APIs (Flask, FastAPI)
    - Cloud services (AWS SageMaker, Google Vertex AI)
- **Continuous Monitoring:** Track model drift & retrain as needed.

Example: Fraud Detection in Banking

Team Execution:

1. **Business Analysts** define fraud detection KPIs (e.g., reduce false positives).
2. **Data Engineers** collect transaction logs, user behavior data.
3. **Data Scientists** preprocess data, engineer features ("transaction frequency").
4. **Model Training:** Use Random Forest for anomaly detection.
5. **Evaluation:** Achieves 98% recall (minimizing missed fraud cases).
6. **Deployment:** Model integrated into real-time payment processing.

Q2)

a) List and explain the steps in data preparation phase of data analytics life cycle. [8]

Data Preparation Phase in Data Analytics Life Cycle
The **Data Preparation Phase** is crucial because the quality and format of data directly impact the accuracy and effectiveness of analytics and modeling. This phase involves transforming raw data into a clean, consistent, and usable form.

Key Steps in Data Preparation Phase:

1. **Data Collection and Ingestion:**
    - Gather data from multiple sources such as databases, APIs, IoT devices, files, and logs.
    - Import data into a centralized location called the analytic sandbox for processing.
2. **Data Exploration and Profiling:**
    - Understand the data by analyzing its structure, types, distributions, and relationships.
    - Identify missing values, outliers, inconsistencies, and duplicate records.
3. **Data Cleaning:**

- Handle missing data by imputation, removal, or estimation.
- Correct errors and inconsistencies in the data (e.g., fixing typos, formatting issues).
- Remove duplicate or irrelevant records.

4. **Data Transformation:**
   - Convert data into the required format or structure (e.g., normalization, scaling, encoding categorical variables).
   - Aggregate or filter data to focus on relevant subsets.
5. **Data Integration:**
   - Combine data from different sources and formats into a cohesive dataset.
   - Resolve schema conflicts and ensure consistency.
6. **Feature Engineering:**
   - Create new variables or features that improve model performance by deriving meaningful attributes from raw data.
   - Example: Extracting "day of week" from a timestamp.
7. **Data Reduction:**
   - Reduce dataset size for efficiency by techniques like sampling, dimensionality reduction, or selecting important features.
8. **Data Loading into Sandbox:**
   - Load the cleaned and transformed data into the analytic sandbox environment where further analysis and modeling occur.

b) Write short note on the following: [9]

i) ETL (Extract, Transform, Load)

ETL is a fundamental process in data analytics and data warehousing that involves:

- **Extract:** Collecting data from various heterogeneous sources such as databases, APIs, files, and IoT devices.
- **Transform:** Cleaning, filtering, and converting raw data into a consistent and usable format by applying rules, calculations, or aggregations.
- **Load:** Loading the transformed data into a target system like a data warehouse, data lake, or analytical database for further processing and analysis.

ETL ensures that data is accurate, integrated, and ready for analytics tasks.

ii) Common Tools for Model Building

Common tools used by data science teams for building models include:

- **Open-source Tools:**
  - **R**: Popular for statistical analysis and visualization.
  - **Python Libraries:** Scikit-learn, TensorFlow, Keras, PyTorch for machine learning and deep learning.
  - **WEKA:** A collection of machine learning algorithms for data mining tasks.
  - **Octave:** Open-source alternative to MATLAB for numerical computations.

- **Commercial Tools:**
  - **MATLAB:** Widely used for advanced modeling, simulations, and algorithm development.
  - **SAS:** Provides comprehensive analytics and modeling capabilities.
  - **STASTICA:** Specialized tool for statistical analysis.

  These tools assist in data manipulation, algorithm implementation, model training, and evaluation.

iii) Model Selection for Data Analytics

Model Selection is the process of choosing the most appropriate algorithm or statistical model based on:

- The **problem type** (classification, regression, clustering, etc.).
- The **nature of the data** (size, features, noise).
- Performance on training and validation datasets measured by metrics like accuracy, precision, recall, F1-score, RMSE, etc.
- Computational efficiency and interpretability.
- Ability to generalize well to unseen data (avoiding overfitting and underfitting).
  Proper model selection ensures reliable and accurate analytics outcomes, ultimately supporting better decision-making.

➢ **MAY / JUN 2023**

Q1)

a) What is Model Building elaborate this phase of data analytics with the help of suitable example.[9]

➔ **REPEATED !**

b) Explain any three sources of Big Data. Differentiate BI versus Data science [8]

Three Sources of Big Data

1. Social Media:
Platforms like Facebook and Twitter generate huge amounts of data from user posts, comments, and interactions.
This data is mostly unstructured and used for analyzing trends and sentiments.
It helps businesses understand customer behavior and preferences.

2. Internet of Things (IoT):
IoT devices like sensors and wearables produce continuous streams of real-time data.
This data includes measurements like temperature, location, and health stats.
It is crucial for automation, monitoring, and predictive maintenance.

3. Transactional Data:
Generated from everyday business operations such as sales, payments, and orders.
It is structured data stored in databases for tracking and analysis.
Used for financial reporting, customer insights, and operational decisions.

Differentiation is already done !!

Q2)

a) What are the three characteristic of Big Data and what are the main consideration in processing Big Data. [8]

Characteristics of Big Data (The 3 Vs)

1. **Volume:**
   Refers to the massive amount of data generated every second from sources like social media, IoT devices, and transactions.
   Big Data involves data sizes ranging from terabytes to petabytes and beyond.
2. **Velocity:**
   The speed at which data is generated, collected, and processed in real-time or near real-time.
   Examples include streaming data from sensors, financial markets, and social media feeds.
3. **Variety:**
   Refers to the different types and formats of data such as structured (databases), semi-structured (XML, JSON), and unstructured data (videos, images, text).

Main Considerations in Processing Big Data

1. **Data Storage and Management:**
   Efficient storage solutions like data lakes and distributed file systems (e.g., HDFS) are needed to handle large volumes and different data types.
2. **Scalability:**
   Processing systems must scale horizontally (adding more machines) to handle increasing data volume and velocity without performance loss.

3. **Data Quality and Cleaning:**
   Ensuring accuracy, consistency, and completeness is crucial since Big Data often contains noise, errors, and missing values.
4. **Processing Speed:**
   Choosing appropriate tools and architectures (batch vs. stream processing) to meet real-time or near real-time analysis requirements.
5. **Security and Privacy:**
   Protecting sensitive data from unauthorized access and ensuring compliance with data protection laws.

b) Explain Descriptive, Diagnostic, Predictive analytics. [9]

Types of Analytics

1. Descriptive Analytics

- Descriptive analytics answers the question "What happened?" by summarizing past data to provide insights into historical performance.
- It uses data aggregation, reporting, and visualization techniques like dashboards, charts, and summary statistics to reveal trends and patterns.
- **For example,** a retail company uses descriptive analytics to generate monthly sales reports showing total sales, revenue, and customer demographics. This helps understand overall business performance but does not explain why changes occurred.

2. Diagnostic Analytics

- Diagnostic analytics goes deeper by answering "Why did it happen?" It investigates the causes behind past outcomes through detailed analysis.
- Techniques such as drill-down, data discovery, correlations, and root cause analysis are used to identify factors influencing the results.
- **For example**, if sales dropped last quarter, diagnostic analytics might examine customer feedback, inventory levels, or marketing campaigns to find reasons for the decline. It helps organizations identify problems and opportunities for improvement**.**

3. Predictive Analytics

- Predictive analytics focuses on forecasting "What is likely to happen?" in the future by using historical data, statistical models, and machine learning algorithms.
- It helps in anticipating future events, trends, and behaviors, enabling proactive decision-making. Common techniques include regression analysis, classification, and time series forecasting.
- **For example**, a telecom company uses predictive analytics to identify customers likely to churn by analyzing usage patterns and customer service interactions, allowing targeted retention strategies. It transforms data into actionable foresight.

➢ **NOV / DEC 2023**

Q1)

a) **Explain Data Analytics Cycle with suitable diagram and its phases. [8]**

➔ **REPEATED**

b) **List and Explain the various activities involved in identifying potential data resources as a part of discovery phase in Data Analytics Life Cycle? [9]**

Key Activities:

1. Problem Understanding:

- Understand the business problem or question to be answered.
- Engage with stakeholders to gather requirements, objectives, and constraints.
- Define clear goals that the analytics project aims to achieve.

2. Data Source Identification:

- Identify all possible sources of data relevant to the problem.
- These can be internal (databases, ERP systems, CRM, logs) or external (social media, public datasets, IoT devices).
- Consider structured, semi-structured, and unstructured data sources.

3. Data Availability Assessment:

- Check the accessibility and availability of the identified data sources.
- Assess if data can be extracted easily or if there are any access restrictions.
- Identify any missing data or gaps that need to be addressed.

4. Data Quality and Relevance Evaluation:

- Perform an initial assessment of data quality aspects such as completeness, accuracy, and consistency.
- Evaluate the relevance of data for addressing the analytics objectives.
- Determine if the data is sufficient or if additional data collection is necessary.

5. Data Collection Planning:

- Plan how data will be collected, extracted, or ingested for analysis.
- Consider formats, volume, frequency, and tools required for data acquisition.
- Decide on data storage options like data lakes or warehouses.

6. Hypothesis Formulation:

- Formulate initial hypotheses or assumptions about the problem based on domain knowledge and data understanding.
- These hypotheses guide further data exploration and model building.

7. Tool and Technology Assessment:

- Identify tools and technologies needed for data extraction, exploration, and initial analysis (e.g., SQL, Hadoop, APIs).
- Evaluate infrastructure needs for handling data volume and complexity.

Q2)

a) List and explain the key roles for successful analytics project. [8]

Key Roles in a Successful Analytics Project

For a data analytics project to succeed, various specialized roles must work collaboratively. Each role contributes unique skills to ensure that business objectives are met using data-driven approaches.

1. Business Analyst

- Acts as the bridge between stakeholders and the technical team.
- Understands the business problem, defines project goals, and gathers requirements.
- Helps ensure that the analytics outcomes align with business needs.

2. Data Engineer

- Responsible for collecting, cleaning, transforming, and preparing data for analysis.
- Builds and manages data pipelines, warehouses, and ensures data quality and availability.
- Works with big data tools like Hadoop, Spark, or ETL platforms.

3. Data Scientist

- Applies statistical, machine learning, and analytical techniques to extract insights.
- Builds predictive models, identifies patterns, and validates hypotheses.
- Communicates findings through visualizations and reports.

4. Data Architect

- Designs and manages the overall data architecture including storage, flow, and integration of data.
- Ensures scalability, performance, and security of the data systems.
- Selects suitable platforms and tools for the analytics ecosystem.

5. Project Manager

- Plans, coordinates, and oversees project execution from start to finish.
- Manages team communication, timelines, resources, and ensures timely delivery.
- Tracks progress and mitigates risks to keep the project on track.

6. Subject Matter Expert (SME)

- Provides domain-specific knowledge and context needed for interpreting data.
- Assists in validating findings and shaping relevant hypotheses.
- Helps the team align insights with real-world applications.

7. Visualization Expert / BI Developer

- Creates dashboards and visual reports to present insights in a user-friendly manner.
- Helps stakeholders understand analytics results through interactive tools (e.g., Power BI, Tableau).
- Focuses on storytelling with data.

b) Write short note on : [9]

i) Common Tools for Model Building :

Model building tools help data scientists train, test, and deploy machine learning and statistical models.

- **Open-source tools**:
    - **R** – Used for statistical analysis and visualization.
    - **Python (scikit-learn, TensorFlow, Keras)** – Popular for machine learning and deep learning.
    - **WEKA** – GUI-based tool for beginners in machine learning.
- **Commercial tools**:
    - **SAS, IBM SPSS Modeler, MATLAB** – Provide advanced modeling and analytics features with enterprise support.

These tools support regression, classification, clustering, and time-series models.

ii) Model Selection for Data Analytics

Model selection is the process of choosing the best algorithm or approach for a specific problem based on data characteristics and goals.

- Depends on **problem type**: classification, regression, clustering, etc.
- Influenced by factors like **data size, accuracy, interpretability**, and **computational efficiency**.
- Example:
    - Use **Linear Regression** for predicting continuous values.
    - Use **Decision Trees/Random Forest** for classification problems.

Proper model selection ensures better performance and insights.

➢ **MAY / JUN 2024**

Q1)

a) What is the data Preparation phase in Data Analytics Lifecycle. What is the Analytics Sandbox and ETLT process in this phase? [8]

DATA PREPARATION PHASE ALREADY DONE !

Analytics Sandbox:

An **Analytics Sandbox** is a secure, isolated environment where data scientists can freely explore, preprocess, and model data without impacting live production systems.

- It provides the flexibility to experiment with different datasets, transformations, and algorithms.

- Often built using cloud platforms, Hadoop clusters, or local systems, allowing large-scale processing.
- The data used in the sandbox is typically a curated subset of production data or a fully anonymized version.

ETLT Process (Extract, Transform, Load, Transform):

ETLT is a variant of the traditional ETL process used in big data environments.

1. **Extract**: Raw data is collected from multiple sources like databases, sensors, logs, APIs, etc.
2. **Transform (1st stage)**: Initial cleaning and formatting is done to structure the data for staging.
3. **Load**: The semi-processed data is loaded into the analytics sandbox or storage layer (e.g., data lake).
4. **Transform (2nd stage)**: Advanced transformation is performed inside the sandbox to prepare final datasets for modeling.

b) List out different stakeholders of an analytics project. What they usually expect at the conclusion (key outputs) of a project? [8]

Different Stakeholders in an Analytics Project:

1. **Business Executives / Sponsors**
   - Provide funding and strategic direction.
   - Expect business value, ROI, and actionable insights to support decision-making.
2. **Project Managers**
   - Oversee timelines, resource allocation, and risk management.
   - Expect clear deliverables, progress reports, and adherence to the project plan.
3. **Data Scientists / Analysts**
   - Responsible for data modeling, analysis, and interpreting results.
   - Expect quality data, technical clarity, and feedback on model impact.
4. **IT/Data Engineers**
   - Manage data infrastructure, pipelines, and security.
   - Expect documentation of data usage and clarity on integration or deployment requirements.
5. **End Users / Business Teams**
   - Use the outcomes to improve operations or customer engagement.
   - Expect easy-to-understand visualizations, dashboards, and tools they can use effectively.
6. **Compliance Officers / Legal Team**
   - Ensure data usage follows regulatory guidelines (e.g., GDPR).

- Expect transparency, audit trails, and adherence to privacy and data governance policies.

Key Outputs Expected at the Conclusion:

- Final report with findings and recommendations.
- Performance metrics of the model (accuracy, precision, etc.).
- Dashboards or BI tools for ongoing insights.
- Deployment or integration plan (if applicable).
- Business impact summary or ROI analysis.
- Code, data documentation, and knowledge transfer sessions

Q2)

a) List out the activities to be carried out in model planning and model building phase. What are different tools used for these phases? [8]

Model Planning Phase – Activities:

1. **Select Analytical Techniques** – Choose suitable statistical or machine learning methods based on the problem (e.g., classification, regression).
2. **Identify Key Variables** – Find important predictors and relationships among features.
3. **Define Data Splits** – Plan data partitioning into training, validation, and test sets.
4. **Assess Tools & Environment** – Decide the tools and compute resources for modeling.
5. **Document Assumptions** – Record assumptions, constraints, and success criteria.

Model Building Phase – Activities:

1. **Develop Training & Testing Datasets** – Prepare datasets based on planning phase.
2. **Build Models** – Implement algorithms using selected techniques.
3. **Train Models** – Fit models to training data and tune parameters.
4. **Validate Performance** – Evaluate model accuracy, precision, recall, etc., using test data.
5. **Iterate & Refine** – Improve model based on feedback and results.

Tools Used in Model Planning:

- **R / Python (Pandas, NumPy, Matplotlib)**
- **Excel**
- **Statistical Tools (SPSS, SAS, STATA)**
- **SQL / HiveQL**

Tools Used in Model Building:

- **Machine Learning Libraries** – scikit-learn, TensorFlow, Keras, PyTorch
- **Data Mining Tools** – WEKA, RapidMiner
- **Programming Languages** – R, Python

- **Commercial Platforms** – SAS Enterprise Miner, IBM SPSS Modeler

b) What is linear regression, and what are its primary objectives? What is the difference between simple linear regression and multiple linear regression? How do you evaluate the performance of linear regression?[8]

What is Linear Regression

Linear regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation. It predicts the value of the dependent variable based on the values of independent variable(s).

Primary Objectives of Linear Regression:

1. **Prediction** – Estimate or forecast future values of the dependent variable.
2. **Understanding Relationships** – Analyze how changes in predictors affect the response variable.
3. **Quantifying Impact** – Measure the strength and direction of the relationship (using coefficients).

Difference between Simple and Multiple Linear Regression:

Feature	**Simple Linear Regression**	**Multiple Linear Regression**
Number of Independent Variables	One	Two or more
Equation Format	$Y = \beta_o + \beta_1 X + \varepsilon$	$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta n Xn + \varepsilon$
Use Case	Basic trend prediction	Modeling complex relationships

Performance Evaluation of Linear Regression:

1. **R^2 (Coefficient of Determination)** – Measures how well the model explains variability in the data.
2. **Mean Absolute Error (MAE)** – Average of absolute errors between predicted and actual values.
3. **Mean Squared Error (MSE)** – Average of squared prediction errors.
4. **Root Mean Squared Error (RMSE)** – Square root of MSE, more sensitive to large errors.
5. **Residual Analysis** – Checks if errors are randomly distributed (assumption of linear regression).

➢ **NOV / DEC 2024**

Q1)

a) Draw data analytics life cycle diagram and briefly explain its phases. [6]
REPEATED

b) Explain the various key roles for a successful analytics project. [6]

A successful analytics project involves collaboration between various roles with distinct responsibilities:

1. **Project Sponsor:** Senior executive who defines business goals, allocates budget, and ensures stakeholder alignment.
2. **Business Analyst:** Bridges the gap between business and technical teams by gathering requirements and defining success metrics.
3. **Data Engineer:** Prepares and manages data pipelines, ensuring clean, reliable, and accessible data for analysis.
4. **Data Scientist / Analyst:** Performs data exploration, modeling, and derives insights to solve business problems using statistical and ML techniques.
5. **IT/System Architect:** Ensures the technical infrastructure (hardware, software, security) supports analytics operations effectively.
6. **Domain Expert:** Provides context and expertise about the business domain to guide data interpretation and validate outcomes.

c) What are various sources of Big data. [6]

Big Data is generated from a wide variety of sources. Key sources include:

1. **Social Media Data**
    - Platforms like Facebook, Twitter, and Instagram generate massive volumes of text, images, videos, and user interactions.
2. **Machine/IoT Data**
    - Data from sensors, smart devices, and industrial machines (e.g., temperature logs, GPS, RFID, manufacturing systems).
3. **Transactional Data**
    - Includes data from online purchases, banking systems, retail sales, and billing records.

4. **Web and Server Logs**
    - Logs generated by websites, servers, and applications that track user activity, performance, and security events.
5. **Public Data Sources**
    - Government portals, research publications, weather databases, and census data available for public use.
6. **Multimedia Content**
    - Audio, video, and image files from cameras, surveillance systems, or media platforms like YouTube and Netflix.

Q2)

a) Describe few applications of Big Data Analytics. [6]

1. **Healthcare**
    - Big Data helps in disease prediction, patient monitoring, personalized treatments, and managing hospital records efficiently.
2. **Retail & E-commerce**
    - Used for customer behavior analysis, recommendation systems, inventory optimization, and dynamic pricing.
3. **Banking & Finance**
    - Detects fraud, performs risk analysis, credit scoring, and enhances customer experience with personalized offers.
4. **Transportation & Logistics**
    - Optimizes route planning, traffic prediction, fleet management, and real-time vehicle tracking.
5. **Social Media & Marketing**
    - Analyzes user interactions, sentiment analysis, campaign performance, and trend predictions.
6. **Smart Cities & Utilities**
    - Enables efficient energy usage, waste management, public safety, and infrastructure planning using IoT data.

b) Explain Data preparation phase of data analytics lifecycle. [6]
REPEATED !

c) List common tools used for model building phase of data analytic. [6]
REPEATED !