**Assignment-1**

1. **Explain key steps of data engineering?**

* **Key steps of data Engineering:**
* **Data Pipeline:**A data pipeline is a series of processes and tools that collect, transform, and move data from various sources to a destination for storage or analysis. Data engineering often involves creating and managing data pipelines. The key steps in a data pipeline include:
* **Data Ingestion:** Collect data from various sources such as databases, APIs, logs, or files.
* **Data Transformation:** Clean, enrich, and reformat the data to prepare it for analysis or storage.
* **Data Storage:** Store the processed data in a data warehouse, data lake, or other storage systems.
* **Data Processing:** Perform batch or real-time processing on the data as needed.
* **Data Quality Assurance:** Implement data quality checks and validation to ensure data accuracy and consistency.
* **Data Monitoring:** Continuously monitor the pipeline for issues or failures.
* **Data Distribution:** Make the processed data available to data consumers, such as analysts or data scientists.
* **Data Warehouse Definition:** A data warehouse is a central repository for storing, organizing, and managing large volumes of structured data from various sources within an organization. The primary purpose of a data warehouse is to provide a unified and historical view of the data, which enables efficient querying and reporting. Key characteristics of data warehouses include:
* **Data Integration:** Data from different sources is integrated and transformed into a consistent format within the data warehouse.
* **Historical Data:** Data warehouses typically store historical data, allowing for trend analysis and historical comparisons.
* **Query Performance:** Data warehouses are optimized for complex queries, enabling business analysts to perform data analysis and generate reports quickly.
* **Data Quality:** Data quality is a focus in data warehouses, with processes in place to ensure the accuracy and consistency of data.
* **Data Marts:** Data marts are subsets of a data warehouse that are designed to serve the specific reporting and analytical needs of a particular department, team, or business unit within an organization. Data marts are created by selecting and summarizing relevant data from the central data warehouse. They provide a more focused and tailored view of data for specific user groups, making it easier for them to access and analyze data that is relevant to their roles.
* **OLAP and OLAP Cubes:** OLAP (Online Analytical Processing): OLAP is a technology that enables users to interactively analyze multidimensional data. It allows users to perform complex queries and aggregations on data quickly. OLAP databases are designed for analytical tasks and are optimized for read-heavy workloads.
* **OLAP Cubes:** OLAP cubes are a multidimensional representation of data that allows for efficient querying and analysis. They organize data into dimensions (e.g., time, geography, product) and measures (e.g., sales revenue, profit). Users can "slice and dice" data within cubes to gain insights and explore data from various angles.
* **Big Data Concepts:** Big data refers to the management and analysis of extremely large and complex datasets that traditional data processing tools and methods struggle to handle. Key concepts in big data include:
* **Volume:** Big data involves processing and storing massive amounts of data.
* **Velocity:** Data is generated and updated rapidly, requiring real-time or near-real-time processing.
* **Variety:** Big data can be structured, semi-structured, or unstructured, coming from various sources, including social media, sensors, and more.
* **Variability:** Data may exhibit irregular patterns or changes in behavior over time.
* **Veracity:** Dealing with data uncertainty and quality issues.
* **Value:** Extracting valuable insights and knowledge from big data for decision-making.
* **Data Architecture as a Service:** Data architecture as a service refers to a cloud-based approach to managing an organization's data architecture components and services. This includes data storage, data processing, data integration, and more. By leveraging cloud services, organizations can scale their data infrastructure, reduce maintenance overhead, and take advantage of managed data services provided by cloud providers like AWS, Azure, or Google Cloud.

1. **Describe Data Pipeline Challenges?**

* Data pipelines are beneficial to businesses in many ways, especially when it comes to real-time and predictive analytics. However, building a custom or standard data pipeline can be pretty daunting, especially for first-time organizations. Here are the five prevalent data pipeline challenges:
* **Data Placement:** Businesses need to store their data in the right format and in the right location to enhance seamless access, as well as usability. Making the right decision can be challenging, given that businesses must use multiple tools and connect them to numerous data stores and formats, especially if they are going to harness the full power of big data.
* **Data Scaling:** Data scaling can be challenging, given that modern businesses handle up to [2.5 quintillion bytes](https://seedscientific.com/how-much-data-is-created-every-day/)of data generated by consumers every day. At the same time, the number of data sources, whether sensors or IoT devices, may increase unexpectedly. With this in mind, organizations should have data storage options that are automatically scalable. However, data scaling issues are more challenging among organizations that use on-premise storage solutions. For instance, overwhelming data velocity and volume sharding and replication create more space for incoming data. In the long haul, these processes can prove costly in terms of operations because a single technical hitch can mean hours of troubleshooting the whole system.
* **Data Flexibility:** A whole data pipeline system relies on the entire ETL (Extract-Transform-Load) process. Although this process is often meticulous, a single hitch in one step can cause hours of downtime, something that can affect data quality. The situation even gets trickier if a business deals with dynamic data sources and events, which might mean setting up schemas for real-time data analytics. At the same time, an ETL data pipeline that is used for data analytics must be optimally elastic for compatibility with various data types and schemas.
* **Data Hosting:** Data hosting can either be done in the cloud or on-premise. Choosing the right hosting service can be challenging, especially if you have to modify the data into a specific format. That’s why some organizations choose to self-host their servers, but this option also comes with operating system, latency requirements, as well as memory and disk challenges.
* **Data Migration:** Data migration techniques depend on how an organization uses its data. However, most businesses choose to migrate their data during off-peak periods, such as at night, so as to minimize unnecessary downtime. Although this might sound convenient, it gets challenging when it comes to real-time analytics, as the migrated data will be from the previous day.

1. **Explain data types?**

* **Data Types:**
* **Categorical data** represent characteristics such as a person’s gender , marital status, hometown the types of movies they like. Categorical data can take on numerical values (such as “1” indicating male and “2” indicating female). But those numbers don’t have mathematical meaning. You couldn’t add them together.
* **Numerical data:** These data have meaning as a measurement, such as a person’s height, weight, IQ, or blood pressure; or they’re a count, such as the number of stock shares a person owns. How many teeth a dog has, or how many pages you can read of your favorite book before you fall asleep. Statisticians also call numerical data quantitative data. Numerical data can be further broken into two types:
* Discrete
* Continuous.
* **Ordinal data:** Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can’t do any arithmetical tasks on them.
* **Examples:** When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10. Letter grades in the exam (A, B, C, D, etc.). Ranking of people in a competition (First, Second, Third, etc.). Economic Status (High, Medium, and Low).
* **Time series analysis:** Time series analysis in statistics refers to a technique that involves conducting an analysis of various data points collected during a certain timeframe. It provides insights and consequences of a given data set’s features changing over time. Individuals analyzing [time series](https://www.wallstreetmojo.com/time-series/) data record the data points at fixed intervals over a particular period. This type of analysis has applications in multiple areas, including [economics](https://www.wallstreetmojo.com/economics/), sales, and [statistics](https://www.wallstreetmojo.com/statistics/).

1. **Explain EDA & it's types?**

* **Exploratory Data Analysis (EDA)** refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

1. **Univariate Analysis:**This sort of evaluation makes a speciality of analyzing character variables inside the records set. It involves summarizing and visualizing a unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Techniques like histograms, field plots, bar charts, and precis information are generally used in univariate analysis.
2. **Bivariate Analysis:**Bivariate evaluation involves exploring the connection between  variables. It enables find associations, correlations, and dependencies between pairs of variables. Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.
3. **Multivariate Analysis:**Multivariate analysis extends bivariate evaluation to encompass greater than  variables. It ambitions to apprehend the complex interactions and dependencies among more than one variables in a records set. Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.
4. **Time Series Analysis:**This type of analysis is mainly applied to statistics sets that have a temporal component. Time collection evaluation entails inspecting and modeling styles, traits, and seasonality inside the statistics through the years. Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA ( AutoRegressive Integrated Moving Average) fashions are generally utilized in time series analysis.
5. **Missing Data Analysis:**Missing information is a not unusual issue in datasets, and it may impact the reliability and validity of the evaluation. Missing statistics analysis includes figuring out missing values, know-how the patterns of missingness , and using suitable techniques to deal with missing data. Techniques along with lacking facts styles, imputation strategies, and sensitivity evaluation are employed in lacking facts evaluation.
6. **Outlier Analysis:**Outliers are statistics factors that drastically deviate from the general sample of the facts. Outlier analysis includes identifying and knowledge the presence of outliers, their capability reasons, and their impact at the analysis.Techniques along with box plots, scatter plots, z-rankings, and clustering algorithms are used for outlier evaluation.
7. **Data Visualization:**Data visualization is a critical factor of EDA that entails creating visible representations of the statistics to facilitate understanding and exploration. Various visualization techniques, inclusive of bar charts, histograms, scatter plots, line plots, heatmaps, and interactive dashboards, are used to represent exclusive kinds of statistics.
8. **Describe the Steps In Data Preprocessing?**

* **Steps In Data Preprocessing:**
* **Gathering the data:**
* **Here i am sharing some website with you to get the dataset :**
* **Kaggle:** Kaggle is my personal favorite one to get the dataset. <https://www.kaggle.com/datasets>
* **UCI Machine Learning Repository:** One of the oldest sources on the web to get the dataset.  
  <http://mlr.cs.umass.edu/ml/>
* This awesome GitHub repository has high-quality datasets. <https://github.com/awesomedata/awesome-public-datasets>
* If you are looking for Government’s Open Data then here is few of them:
* Indian Government: [http://data.gov.in](http://data.gov.in/)
* US Government: <https://www.data.gov/>
* British Government: <https://data.gov.uk/>
* France Government: <https://www.data.gouv.fr/en/>
* **Import the dataset & Libraries.**
* **Dealing with Missing Values:** Check for null values, Drop Null values, Replacing Null values with Strategy
* **Divide the dataset into Dependent & Independent variable:** After importing the dataset, the next step would be to identify the independent variable (X) and the dependent variable (Y). Basically dataset might be labeled or unlabeled.
* **dealing with Categorical values:** Use libraries.
* **Split the dataset into training and test set:** In machine learning we usually splits the data into Training and Testing data for applying models.
* **Feature Scaling:** Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.

1. **Explain the terms:**
2. **Data warehouse and Data Warehouse Storage:**

|  |
| --- |
| **Data warehouse architrecture** |

* **Data warehouse:** A typical data warehouse architecture includes three basic components. A data warehouse is a central repository, usually, a relational database, modified and optimized to support data reading, aggregation, and querying. Although traditional data warehouses only supported structured data formatted in tables, modern applications can support both structured and unstructured data formats. Unstructured data in this case include information formatted and presented as images, PDF files, or even audio files. Data warehouse concepts as a single point of truth and information in an organization. As opposed to retrieving data from multiple storage, data warehousing allows business analysts to report similar results and create near-accurate metrics for predictive analytics.
* **Data Warehouse Storage:** A central repository or a database is the bloodline of a custom or standard data warehouse architect, as all business data is stored where Business leaders and other employees can then access the data warehouse storage to draw valuable insights from its contents. Businesses have the option of either an on-premise or cloud-based data warehouse storage. The former option is ideal for organizations that want to process their data at high querying speeds and uncompromised security. On the other hand, cloud-based data warehouses support automatic scalability and any data structure. They are also relatively affordable than their on-premise counterparts. Some data architects might also help you build collective storage options that run parallel as a centralized warehouse. This approach is usually ideal when enhancing scalability.
* **Metadata :** Metadata contains the information and guidelines for changing and processing data when loading it into a warehouse environment.
* **Access Tools:** These are tools that are integrated into the warehouse architecture to facilitate access, as well as interactions of the stored data with end-users. These tools might include querying, reporting, or data mining tools, based on the model of the data warehouse model.
* **Management Tools:** Data warehouse tools help businesses perform automated administrative duties.

1. **Data Marts:** Data mart is a smaller data warehouse (their size is usually less than 100Gb.). They become necessary when the company and the amount of its data grow and it becomes too long and ineffective to search for information in an enterprise DW. Instead, data marts are built to allow different departments (e.g., sales, marketing, C-suite) to access relevant information quickly and easily. Data marts exist in three prevalent types, including:

|  |
| --- |
| **Data infrastrcuture with data marts** |

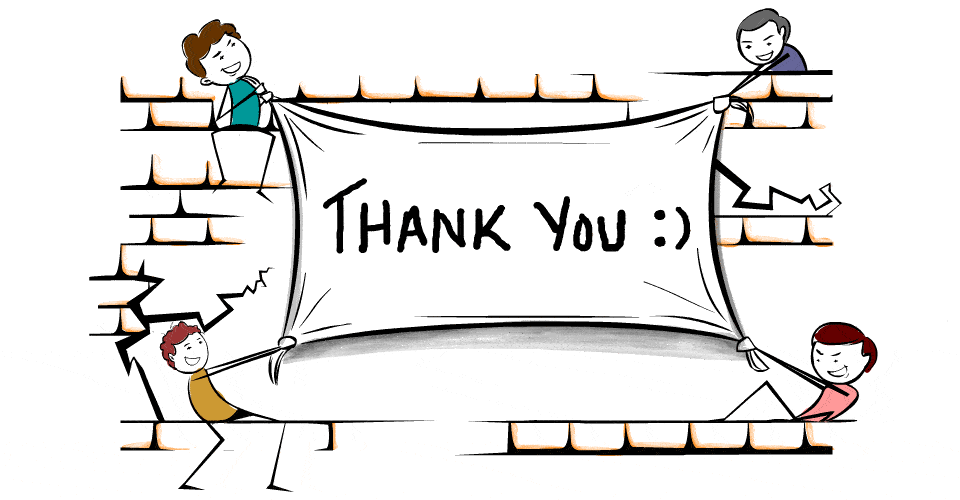
* **Dependent data marts**are created from an enterprise DW and used as a primary source of information (also known as a top-down approach).
* **Independent data marts**are standalone systems that function without DWs extracting information from various external and internal sources (it’s also known as a top-down approach).
* **Hybrid data** marts combine information from both DW and other operational systems.

|  |
| --- |
| **Illustration showing the OLAP Process** |

1. **OLAP and OLAP Cubes:** OLAP is a simple abbreviation for Online Analytical Processing. This computational program allows business analysts and data engineers to take a multidimensional approach to data analysis. In other words, this tech helps organizations vast data from different angles, as opposed to OLTP.

* OLAP cubes are the multidimensional structures that represent data. However, unlike traditional database representation (usually in rows and columns), which can be generated automatically, OLAP cubes must be custom-built for individual reporting and analytical querying. There are multiple steps of OLAP:
* First, data is first extracted from various data sources and formats, like text files and spreadsheets. This data is then stored in the Data Warehouse.
* Next, the data is cleaned, transformed, and stored in OLAP Cubes
* Once in the OLAP cubes, information is then pre-calculated and pre-aggregated in advance for further analysis
* Lastly, the user gets the data from the OLAP cubes by running queries against them

**(PPT Give you more understanding than PDF)** The material for the PDF has been compiled from various sources such as books, tutorials (offline and online), lecture notes, several resources available on Internet. The information contained in this PDF is for general information and education purpose only. While we endeavor to keep the information up to date and correct, **we make no representation of any kind about the completeness and accuracy of the material.** The information shared through this PDF material should be used for educational purpose only.

****