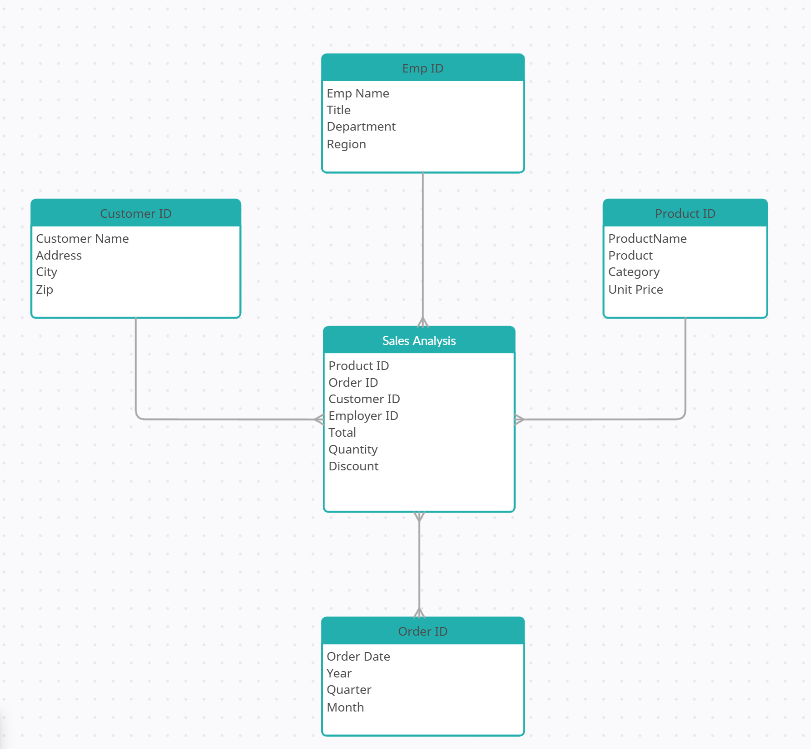


Assignment -1

Q1 SALES are a fact table having attributes i.e. (Product ID, Order ID, Customer ID, Employer ID, Total, Quantity, Discount) which references to the dimension tables. Employee dimension table contains the attributes: Emp ID, Emp Name, Title, Department and Region. Product dimension table contains the attributes: Product ID, Product Name, Product Category, Unit Price. Customer dimension table contains the attributes: Customer ID, Customer Name, Address, City, Zip. Time dimension table contains the attributes: Order ID, Order Date, Year, Quarter, and Month. Design the Star Schema for the above example

Ans.



Q2. Which schema (star or snowflake) is used in large organization? Justify your answer

Ans. Large organizations prefer the Snowflake Schema over the Star Schema for the following reasons:

1. Snowflake Schema:

* A normalized version of the Star Schema.
* Dimension tables are split into multiple related tables to reduce data redundancy.
* Example: Instead of storing City, State, and Country in the Customer Dimension Table, they are separated into individual tables.
* Used in enterprise-scale data warehouses.

Simplicity: The Star Schema is simpler to design and understand. It has a single layer of dimension tables directly connected to the fact table, making it easier for business users and analysts to query.

Query Performance: Star Schema requires fewer joins compared to Snowflake Schema, which results in faster query performance. This is crucial for large organizations dealing with massive datasets.

Scalability: Star Schema is highly scalable and can handle large volumes of data efficiently. It is optimized for read-heavy operations, which are common in data warehousing.

Maintenance: Maintaining a Star Schema is easier because it has fewer tables and relationships compared to Snowflake Schema, which normalizes dimension tables into multiple levels.

Q3: What do you understand by data exploration? Illustrate the answer with an example

Ans. Data Exploration is the process of analyzing and summarizing datasets to understand their structure, patterns, and relationships. It involves using statistical and visualization techniques to uncover insights and identify potential issues in the data.

Steps in Data Exploration:

1. Data Collection: Gather data from various sources.
2. Data Cleaning: Handle missing values, remove duplicates, and correct errors.
3. Descriptive Statistics: Calculate measures like mean, median, mode, standard deviation, etc.
4. Data Visualization: Use charts (e.g., histograms, scatter plots, box plots) to visualize data distributions and relationships.
5. Pattern Identification: Identify trends, correlations, and anomalies in the data.

Example: Checking Missing Values

df.isnull().sum()

Summary Statistics

df.describe()

Q4. Write the steps of data preprocessing?

Ans. Data preprocessing is the process of cleaning and transforming raw data into a usable format for analysis. It consists of several steps:

1. Data Collection
   * Gathering raw data from various sources (databases, CSV files, APIs).
2. Data Cleaning
   * Handling missing values (filling with mean, median, or removing rows).
   * Removing duplicates.
3. Data Transformation
   * Converting data into a uniform format (e.g., standardizing dates).
4. Data Reduction
   * Removing unnecessary columns or redundant data to optimize performance.
5. Data Normalization & Scaling
   * Rescaling numeric values to a common scale for machine learning models.
6. Feature Selection
   * Selecting only relevant variables for analysis.
7. Splitting Data
   * Dividing the dataset into training and testing sets (used in machine learning).

Q5. Explain Types of Classification Algorithms ?

Ans. Classification algorithms are used to predict the category or class of a data point. Here are the main types:

1. Decision Trees:
   * A tree-like model where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome.
   * Example: Classifying whether an email is spam or not based on features like subject line and sender.
2. Random Forest:
   * An ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting.
   * Example: Predicting customer churn by aggregating predictions from multiple decision trees.
3. Support Vector Machines (SVM):
   * Finds the hyperplane that best separates the classes in the feature space.
   * Example: Classifying images of cats and dogs based on pixel values.
4. K-Nearest Neighbors (KNN):
   * Classifies a data point based on the majority class among its k-nearest neighbors.
   * Example: Predicting whether a customer will buy a product based on similar customers' behavior.
5. Naive Bayes:
   * Based on Bayes' theorem, it assumes independence between features.
   * Example: Spam detection by calculating the probability of an email being spam based on word frequencies.
6. Logistic Regression:
   * Used for binary classification, it estimates the probability of a binary outcome using a logistic function.
   * Example: Predicting whether a student will pass or fail an exam based on study hours.

Q6 Explain ?

a) Divisive Clustering

b) BIRCH Clustering

c) Density-Based Clustering

Ans:

Divisive Clustering:

* A top-down hierarchical clustering method where all data points start in one cluster.
* The cluster is recursively split into smaller clusters based on dissimilarity until each data point is in its own cluster.
* Example: Dividing a dataset of customers into segments based on purchasing behavior.

BIRCH Clustering:

* Designed for large datasets, it builds a CF Tree (Clustering Feature Tree) to summarize the data.
* It performs clustering in two phases: building the CF Tree and applying clustering algorithms to the leaf nodes.
* Example: Clustering millions of online transactions for fraud detection.

Density-Based Clustering:

* Focuses on identifying regions of high density separated by regions of low density.
* DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular algorithm in this category.
* Example: Identifying clusters of GPS locations to find popular tourist spots.