GROW DATA SKILLS

Hive Assignment 2

Data Analysis with Apache Hive on Telecom Data

1. Data Loading (Beginner)
    a. Download the dataset and load it into a Hive table.
    b. Write a query to display the top 10 rows of the table.

2. Data Exploration (Beginner)
    a. Write a HiveQL query to find the total number of customers in the dataset.
    b. Write a HiveQL query to find the number of customers who have churned.
    c. Analyze the distribution of customers based on gender and SeniorCitizen status.
    d. Determine the total charge to the company due to churned customers.

3. Data Analysis (Intermediate)
    a. Write a HiveQL query to find the number of customers who have churned, grouped by their Contract type.
    b. Write a HiveQL query to find the average MonthlyCharges for customers who have churned vs those who have not.
    c. Determine the maximum, minimum, and average tenure of the customers.
    d. Find out which PaymentMethod is most popular among customers.
    e. Analyze the relationship between PaperlessBilling and churn rate.

4. Partitioning (Intermediate)
    a. Create a partitioned table by Contract and load the data from the original table.
    b. Write a HiveQL query to find the number of customers who have churned in each Contract type using the partitioned table.
    c. Find the average MonthlyCharges for each type of Contract using the partitioned table.
    d. Determine the maximum tenure in each Contract type partition.

5. Bucketing (Advanced)
    a. Create a bucketed table by tenure into 6 buckets.
    b. Load the data from the original table into the bucketed table.

c. Write a HiveQL query to find the average MonthlyCharges for customers in each bucket.
d. Find the highest TotalCharges in each tenure bucket.

6. Performance Optimization with Joins (Advanced)

Assume another dataset, CustomerDemographics.csv, that contains the details of the demographic data of each customer.

a. Load the demographics dataset into another Hive table.
b. Write HiveQL queries to join the customer churn table and the demographics table on customerID using different types of joins - common join, map join, bucket map join, and sorted merge bucket join.
c. Observe and document the performance of each join type.

7. Advanced Analysis (Expert)

a. Find the distribution of PaymentMethod among churned customers.
b. Calculate the churn rate (percentage of customers who left) for each InternetService category.
c. Find the number of customers who have no dependents and have churned, grouped by Contract type.
d. Find the top 5 tenure lengths that have the highest churn rates.
e. Calculate the average MonthlyCharges for customers who have PhoneService and have churned, grouped by Contract type.
f. Identify which InternetService type is most associated with churned customers.
g. Determine if customers with a partner have a lower churn rate compared to those without.
h. Analyze the relationship between MultipleLines and churn rate.