Hive Assignment 1

Car Insurance Cold Calls Data Analysis

Problem 1: Data Loading and Inspection

1.1. Load the car insurance data into your HDFS (Hadoop Distributed File System).

1.2. Create a Hive table Insurance_Cold_Calls that matches the schema of the loaded data. Make sure to choose the correct data types for each column.

1.3. Load data into the Insurance_Cold_Calls table.

1.4. Write a HiveQL query to display the first 10 records.

Problem 2: Data Exploration

2.1. Write a HiveQL query to count the total number of records in the dataset.

2.2. Write a HiveQL query to find out the count of each outcome (Success, Failure, Other, No previous contact).

2.3. Write a HiveQL query to find the count of customers grouped by Education Level.

Problem 3: Complex Queries

3.1. Write a HiveQL query to find out the average call duration for each outcome.

3.2. Find out the maximum number of contacts performed for a single customer before this campaign and for a single customer during this campaign.

3.3. Write a HiveQL query to find the number of days that passed by after the client was last contacted from a previous campaign, grouped by the outcome of the campaign.

Problem 4: Advanced Analysis

4.1. Write a HiveQL query to find the age group that has the highest number of insurance purchases. Assume age groups as follows: 18-30, 31-40, 41-50, 51-60, 61-70, 70+.

4.2. Write a HiveQL query to find the correlation between the call duration and the success of the campaign.

4.3. Write a HiveQL query to find out if the month of contact affects the success of the campaign. Provide the success rate for each month.

Problem 5: Optimization

5.1. Partition the table by the outcome column and measure the query performance improvement.

5.2. Write a HiveQL query to create a bucketed table by 'Education' with 4 buckets and load data into it.

Problem 6: Derived Insights

6.1. Write a HiveQL query to identify the top 5 professions that are most likely to buy insurance.

6.2. Write a HiveQL query to find out the total calls made, grouped by the day of the week.

6.3. Write a HiveQL query to find out the hour of the day with the highest call success rate.

Problem 7: Advanced Data Manipulation

7.1. Create a new column age_group in the Insurance_Cold_Calls table based on the age of the customer as follows: 18-30, 31-40, 41-50, 51-60, 61-70, 70+. Write a HiveQL query to achieve this.

7.2. Write a HiveQL query to identify and replace any null or 'unknown' values in the job and education columns with the most frequent value in each column.

7.3. Write a HiveQL query to delete records where the duration of the call is zero.

Problem 8: Advanced Data Exploration

8.1. Write a HiveQL query to find the count of insurance purchased by each age group and sort them in descending order.

8.2. Write a HiveQL query to find out the number of calls made, grouped by the marital status and job of the customers.

8.3. Write a HiveQL query to find the customers (identified by customer numbers) who have been contacted more than twice and still did not purchase the insurance.

Problem 9: Optimization & Performance

9.1. Compare the performance of a join operation before and after enabling map joins (hint: set hive.auto.convert.join to true).

9.2. Write a HiveQL query to create a view that includes customers' job, marital status, education, and outcome of the campaign. Use this view in subsequent queries and measure any performance improvements.

9.3. Test the performance of your Hive queries with different file formats (e.g., text file, SequenceFile, Avro, Parquet). Document your observations on read and write times.

Please make sure to submit the HiveQL queries, along with their results and your observations. This assignment not only tests your understanding of Apache Hive but also requires you to derive meaningful insights from the data.