BigQuery Assignment 1.1

Introduction

This assignment is designed to help you practice key BigQuery concepts, including:

1. Using DDL to create and manage tables
2. Understanding and using Materialized Views
3. Implementing column and row-level data masking

All tasks should be completed within the BigQuery Sandbox, ensuring you stay within the free usage limits.

Dataset Information

You will use a simplified version of the Google Analytics Sample dataset:

1. Project: `bigquery-public-data`
2. Dataset: `google_analytics_sample`
3. Table: `ga_sessions_20170801`

Dataset Setup

Create a Dataset: In your BigQuery project, create a new dataset named **`bq_assignment`**. All your tables and views for this assignment will be stored here.

Tasks

Question 1: Creating a Table Using DDL

Objective: Create a new table from a public dataset using DDL statements.

Instructions:

1. Table Creation: Create a table named `ga_sessions_copy` in your `bq_assignment` dataset.
2. Data Selection: Select the following fields from `bigquery-public-data.google_analytics_sample.ga_sessions_20170801`:
   - `fullVisitorId`
   - `visitId`
   - `visitNumber`
   - `visitStartTime`
   - `date`
   - `totals.visits` as `visits`
   - `totals.hits` as `hits`
   - `totals.pageviews` as `pageviews`
   - `trafficSource.medium` as `medium`
   - `device.deviceCategory` as `device_category`

- `geoNetwork.country` as `country`
- Row Limitation: Limit the number of rows to 100 to stay within free usage limits.

Hints:

- Use `CREATE TABLE AS SELECT` syntax.
- Use aliases to simplify nested fields (e.g., `totals.visits AS visits`).
- Apply a `LIMIT` clause to control the data size.

Question 2: Creating and Querying a Materialized View

Objective: Learn how to create and use a Materialized View to optimize query performance.

Instructions:

1. Create a Materialized View:
   - Name: `mv_country_sessions`
   - Location: In your `bq_assignment` dataset.
   - Based on: The `ga_sessions_copy` table.
   - Content: Calculate the total number of sessions (`session_count`) per `country`.
   - Filter: Include only countries with more than 10 sessions.
2. Query the Materialized View:
   - Retrieve all columns from `mv_country_sessions`.
   - Order the results by `session_count` in descending order.

Hints:

- Use `CREATE MATERIALIZED VIEW` syntax.
- Utilize `GROUP BY`, `HAVING`, and `ORDER BY` clauses.

Question 3: Implementing Column and Row-Level Data Masking

Objective: Apply data masking techniques to protect sensitive information.

Instructions:

1. Column-Level Data Masking:
   - Goal: Mask the `fullVisitorId` to protect user privacy.
   - Method: Replace the middle digits of `fullVisitorId` with asterisks (`*`), keeping the first 3 and last 3 digits visible.
   - Select: `fullVisitorId`, `masked_fullVisitorId`, `visitId`, and `date`.
   - Limit: Retrieve 10 rows.
2. Row-Level Data Masking:
   - Goal: Mask sensitive data based on a condition.
   - Condition: For rows where `country` is `'United States'`, set `hits` and

`pageviews` to `NULL`.

- Select: `fullVisitorId`, `country`, `hits`, and `pageviews`.
- Limit: Retrieve 10 rows.

Hints:

- Use string functions like `SUBSTR`, `LENGTH`, and `CONCAT` for masking.
- Use `CASE WHEN` statements for conditional masking.

Submission Instructions

1. Share your dataset with the TA: 113356042@g.nccu.edu.tw

BigQuery Resources: For each task, save your queries and any created tables or views in your `bq_assignment` dataset.

Sharing:

- Go to your BigQuery console.
- Navigate to your `bq_assignment` dataset.
- Click on the "Share Dataset" button.
- Add the TA's email address with the permissions (`Viewer`).
- Ensure that the TA has access to all the resources needed to review your work.

2. Share your queries with the TA: 113356042@g.nccu.edu.tw

Saving Your Queries:

A. Compose and Run Your Query:
   - In the Query Editor, write and execute your query.

B. Save the Query:
   - Click the "Save" button located above the Query Editor.
   - In the dialog that appears, enter a query name (e.g., Question1).
   - Choose a save location (it's recommended to save it within your project).
   - Click "Save".

Sharing Your Queries:

C. Find the Saved Query:
   - In the left-hand panel under "Saved Queries", locate the query you want to share.

D. Share the Query:
   - Click the three-dot icon (⋮) next to the query name and select "Share".
   - In the sharing settings: Add the TA's email address.
   - Click "Done".

3. Please complete and share your assignment by October 2, 2024, at 11:59 PM