Ch07

Visualizing Data with Data Studio (now Looker Studio)

CHAPTER ASSIGNMENT

Name _____ Class _____ No. _____ Date _____

> Answer every question in the space given. Show your working. You may use the chapter page to help you.

Questions from *Data Engineering with Google Cloud Platform*, 1st edition, Adi Wijaya (Packt, 2022), and from the Guru Jazzy chapter page, for classroom use.

Exercises from the book

Ex1 **Exercise ? Exploring the BigQuery `INFORMATION_SCHEMA` table using Data Studio**

In this exercise, we will explore the metadata of a public dataset from BigQuery (`bigquery-public-data.austin_bikeshare.INFORMATION_SCHEMA.COLUMNS`) using Data Studio. Complete the following steps:

(1) Querying the `INFORMATION_SCHEMA` table from BigQuery: run `SELECT table_name, column_name, is_nullable, data_type FROM bigquery-public-data.austin_bikeshare.INFORMATION_SCHEMA.COLUMNS;` in the BigQuery console and record what the query returns.

(2) Exploring the query results using Data Studio: click **EXPLORE DATA → Explore with Data Studio** and describe the default table visualization generated by Data Studio.

(3) Modifying the visualization in Data Studio to find new insights: change the chart to a bar chart with `table_name` as Dimension and `data_type` as Breakdown Dimension, add a pie chart of `data_type`, and state the two main limitations of this quick exploration approach.

Ex2 **Exercise ? Creating a Data Studio report using data from a bike-sharing data warehouse**

Create a dashboard report from scratch on the Data Studio console (`https://datastudio.google.com/`) using the star-schema tables from Chapter 3 (`fact_trips_daily_summary`, `dim_stations`, `dim_bikeshare_type`):

(1) Creating a new empty report in Data Studio: create a Blank Report named `Chapter 7 - Exercise 2` connected to `chapter_3_dwh_bikeshare.fact_trips_daily_summary`.

(2) Creating a score card and date range control: add a Date range control (`Jan 1, 2018` to `Dec 31, 2018`) and a Scorecard for `total_trips`. Record the total trips in 2018.

(3) Creating a time series chart: add a Time series chart with `trip_date` as Dimension and `total_trips` as Metric, and describe the trend across 2018.

(4) Blending data: blend `fact_trips_daily_summary` with `dim_stations` (on `start_station_id = station_id`) and `dim_bikeshare_type` (on `bikeshare_type_id`), rename the data source `Fact - Dim Blend`, and build a table of `name` (station name) and `total_trips`.

(5) Sharing the Data Studio report: compare the sharing options (Invite people, Schedule email delivery, Get report link, Embed report, Download report) and explain how Data Studio Data Credentials (**Owner's credentials** vs **Viewer's credentials**) control whether a viewer without BigQuery access can see the data.

Ex3 Creating a materialized view and understanding how BI Engine works

(1) Materialized views: create `chapter_7.fact_table_from_public` from `bigquery-public-data.austin_bikeshare.bikeshare_trips`, then create the materialized view `chapter_7.mv_bikeshare_station_Lockhart` aggregating `COUNT(1) as total_trips` and `SUM(duration_minutes) as total_duration_minutes` grouped by `start_station_name`. Compare the bytes processed when querying the materialized view (`WHERE start_station_name = '8th & Congress'`) versus running the same aggregation directly on `fact_table_from_public`.

(2) BigQuery BI Engine: explain how BI Engine caches BigQuery results in memory for Data Studio, why a reservation is required per project and location, and why you should not leave a BI Engine reservation running in a practice project.

Sorting activities from the chapter page

S1 Two reasons to visualise

Visualisation is not usually your main job, but you will do it for two different reasons, and they need different amounts of care.

Options: **Exploration** ? **Report**

No.	Scenario	Answer
1	You want a quick bar chart of yesterday's pipeline job durations to find the slow one.	
2	Regional managers need a weekly dashboard with a clear layout, shared by link.	
3	You want to see how much you were billed per day on BigQuery this month, once.	

S2 Big multiplier check

On-demand BigQuery cost is roughly bytes scanned times price. A shared dashboard multiplies it twice: by the number of users and by how often each one refreshes. Play with the numbers.

Options: **Costly mistake** ? **Fine**

No.	Scenario	Answer
1	A 1 TB table sits behind a dashboard that 10 people each refresh every half hour.	
2	A large table is queried by only a handful of analysts, a few times a month.	
3	Hundreds of people use dashboards built on several small, aggregated tables.	
4	A dashboard for hundreds of people is connected straight to a raw transactional table.	

S3 Table, view or materialized view?

Two BigQuery features reduce repeated processing. Both help most when the same dashboard queries run over and over.

Options: **Table** ? **View** ? **Materialized view**

No.	Scenario	Answer
1	Store a transformation's output so downstream users read stored rows, refreshed by your own job.	
2	Always-current results, but you accept the query being recomputed every time it is read.	
3	Always-current aggregate results, with the heavy computation precomputed so far fewer bytes are processed.	

Quiz from the chapter page: Check yourself

1 Why should a data engineer care about how dashboards are built?

A. They are your only deliverable

B. They are the last downstream of your work, and their users' satisfaction measures it, and they drive BigQuery cost

C. Dashboards replace warehouses

D. BigQuery cannot be used without them

2 BigQuery cost is driven by data size times what?

A. Number of columns

B. Amount of usage (how many queries)

C. Dataset name length

D. Region count

---

3 Where in the layered warehouse are tables least likely to be huge?

A. The raw layer

B. The access layer, which is filtered and aggregated for reports

C. The staging bucket

D. Nowhere: all tables are huge

---

4 A report is shared with a colleague who has no BigQuery access. What decides whether they see data?

A. Their own BigQuery permissions

B. The data source’s credentials, by default the owner’s

C. A Pub/Sub subscription

D. They cannot see it

---

5 What is a materialized view for?

A. Exporting to CSV

B. A cache-like object that keeps results fresh but avoids reprocessing, cutting bytes processed

C. Renaming datasets

D. Streaming inserts

---

6 What must you remember about BI Engine while practising?

A. It is free

B. It caches results in reserved memory, and a reservation is not free, so do not create one for practice

C. It replaces materialized views

D. It only works on Sundays

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-07-visualizing-data-with-looker-studio/