Ch03

Building a Data Warehouse in BigQuery

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.

Scenarios and exercise from the book

Sc1 **Data warehouse in BigQuery – Requirements for scenario 1**

In this scenario 1, we met with business users, and here are their requirements:

1. As a regional manager user, I want to know the top two region IDs, ordered by the total capacity of the stations in that region.

2. As a regional manager, I want to download the answers to my questions as CSV files to my local computer.

3. The data source table is the station table, which is located in the CloudSQL-MySQL database.

4. There will be more data sources in the future besides the station table.

Given the requirements, I suggest you take a moment to imagine yourself in this situation:

(a) What will you do?

(b) What GCP services will you use?

(c) How will you do it?

Sc2 **Data warehouse in BigQuery – Requirements for scenario 2**

In the second scenario, we are going to load two more tables – bike trips and regions. In this scenario, we want to simulate how to handle new data loading and think about data modeling.

Here are some requirements:

1. As an operational business user, I need to access the following information:
   - How many bike trips take place daily?
   - What is the daily average trip duration?
   - The top five station names as the starting station that has the longest trip duration.
   - The top five region names that have the shortest total trip durations.
2. The bike trips data is in the GCS bucket, and each bucket folder contains daily data.
3. The regions data is from the BigQuery public dataset.
4. New data will be updated daily in the GCS bucket for stations and trips tables.

Similar to scenario 1, I suggest you take a moment to imagine yourself in the situation:

(a) What will you do?

(b) What services will you use?

(c) How will you do that?

Sc3 **Exercise – Scenario 3**

As a final activity in this chapter, you can do a self-assessment exercise to solve an additional business question from business users. Our operational user from scenario 2 wants to ask this additional question:

Show me the top three regions that have the most female riders as of the most recent date (2018-01-02).

Because the gender of members is not yet included in our fact and dimension table, you need to create a different fact and dimension table for this.

Remember that the data model is subjective, especially in the Kimball method. There is no right or wrong answer to the question. As we've discussed in this chapter, everyone can have different data models to represent the real world.

Try to solve it yourself and compare it to the solution in the Git code example:

- `Chapter-3\code\bigquery_self_excercise_create_dim_table_regions.py`
- `Chapter-3\code\bigquery_self_excercise_create_fact_table_daily_by_gender_region.py`
- `chapter-3\code\bigquery_self_excercise_query_answer.sql`

Sorting activities from the chapter page

S1 Event or snapshot?

Options: **Event** · **Snapshot**

No.	Scenario	Answer
1	A customer taps a bike out of a dock at 08:03.	
2	A docking station has a name, a region and a capacity.	
3	A sensor reports 21.4 °C at 12:00:05.	

S2 Pick the write mode

Options: **WRITE_APPEND** · **WRITE_TRUNCATE** · **WRITE_EMPTY**

No.	Scenario	Answer
1	Each day new bike trips arrive. A trip never changes once it has happened.	
2	A stations file arrives daily as a full copy. Stations can be added, changed or removed, and you only need the current picture.	
3	A one-off load must never overwrite a table that already has data.	

S3 Views or tables?

Options: **View** · **Table**

No.	Scenario	Answer
1	A small stations table, and a manager wants an always-fresh top-two-regions answer without storing anything extra.	
2	1,000 dashboards each use a view that joins and aggregates five multi-petabyte tables.	
3	You want zero extra storage and results that update the moment the source table does.	

S4 Inmon or Kimball?

Options: **Inmon** · **Kimball**

No.	Scenario	Answer
1	Finance needs statements that cannot contain mistakes, from sources that rarely change, and accuracy matters more than build speed.	
2	New sources and columns appear every week and the business wants answers quickly, even if some inefficiency remains.	
3	You are building one enterprise-wide source of truth first, with every data mart obeying its rules.	

S5 **Facts and dimensions**

Options: **Fact** · **Dimension**

No.	Scenario	Answer
1	Total trips and total trip seconds per station per day.	
2	Each station's name, region and capacity.	
3	A customer's name, segment and signup date.	

Quiz from the chapter page: Check yourself

1 Why stage database extracts as files in Cloud Storage before loading into BigQuery?

A. BigQuery cannot connect to databases in any way

B. Cloud Storage is a standard, highly scalable landing area that works for almost any source

C. Files are always cheaper than tables

D. It removes the need for a schema

2 A load from a bucket in one region fails against a dataset in another. What is the most likely rule you broke?

A. The file was too small

B. CSV files must be zipped

C. Bucket and dataset must be in the same location

D. The dataset name contained a number

3 Why can BigQuery keep timestamps only in UTC, and what does that force on you?

A. It saves storage, so convert after loading

B. For consistency across regions, so convert local times to UTC before loading

C. Only for US customers

D. It has no effect on data engineers

4 Which best describes a BigQuery dataset?

A. A single table with many partitions

B. A group of related tables and views, ideally named for a business meaning

C. A copy of a bucket

D. A billing account

5 In BigQuery, which trade-off does the book argue you can lean toward?

A. Save storage and compute over consistency and user-friendliness

B. Favour consistency and a real-world model over storage and compute efficiency

C. Always denormalise everything

D. Skip modelling because storage is cheap

---

6 What does BigQuery's nested and repeated data offer?

A. The convenience of denormalised data with less repetition, at the price of harder SQL for newcomers

B. Automatic partitioning

C. Faster streaming inserts

D. Removal of the need for a warehouse

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-03-building-a-data-warehouse-in-bigquery/