Ch10

Cost Strategy in GCP

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 and examples from the book

Ex1 **Example ? estimating data engineering use case**

The Chief Data Officer (CDO) asks you to estimate the monthly cost of a planned GCP data architecture in `us-central1` using the Google Cloud Pricing Calculator (`https://cloud.google.com/products/calculator`):

? **Google Cloud Storage (GCS):** Store 1 TB of data in a regional (`us-central1`) Standard bucket; data transfer can be ignored.

? **Dataproc:** 1 cluster with 1 master node (`n1-standard-2`) and 10 worker nodes (`n1-standard-4`), 500 GB standard persistent disk per node, running 24 hours a day.

? **BigQuery:** 10 TB active storage (`us-central1`), 100 queries per day scanning an average of 500 GB per query on-demand, plus 2 GB of BI Engine reserved capacity.

? **Cloud Composer:** 1 small environment in `us-central1` with 4 vCPUs, 15 GB memory, and 100 GB storage running 24 hours a day (minimum 3 nodes).

? **Pub/Sub:** 10 GB of messages per day.

? **Dataflow:** 1 streaming job in `us-central1` with 2 worker nodes (`n1-standard-2`) and 100 GB standard persistent disk per worker running 24 hours a day.

(1) Record the estimated monthly cost for each of the six services and the total monthly estimate.

(2) Dataproc and BigQuery represent the largest portion of the estimate. What architectural or pricing changes can you propose to reduce the Dataproc cost (e.g., ephemeral clusters) and the BigQuery cost (e.g., flat-rate slots or table partitioning/clustering)?

Ex2 **Exercise ? optimizing BigQuery on-demand cost**

Compare the billed bytes across three table types using `bigquery-public-data.stackoverflow.stackoverflow_posts` (`WHERE creation_date BETWEEN '2016-01-01' AND '2016-04-01'`):

(1) Step 1 ? Create a standard BigQuery table `chapter_10.stackoverflow_posts_std`.

(2) Step 2 ? Create a date-partitioned table `chapter_10.stackoverflow_posts_partitioned` (`PARTITION BY DATE(creation_date)`).

(3) Step 3 ? Create a partitioned and clustered table `chapter_10.stackoverflow_posts_partitioned_clustered` (`PARTITION BY DATE(creation_date) CLUSTER BY tags, title, owner_user_id`).

(4) Step 4 ? With cached results disabled, run `SELECT DATE(creation_date) creation_date, count(1) total FROM ... WHERE DATE(creation_date) = '2016-01-05' AND tags = 'android' AND title LIKE 'a%' GROUP BY creation_date;` on each of the three tables. Record the billed bytes (MB processed) for each table and explain why they differ.

Sorting activities from the chapter page

S1 Three ways GCP charges you

Every service prices differently, but they fall into three families. Knowing which one a service belongs to tells you what to put in the calculator.

Options: **Machine-based** ? **Usage-based** ? **Flat rate**

No.	Scenario	Answer
1	A Cloud Composer environment sat idle over the weekend, and the bill still counted every hour.	
2	Cloud Storage is billed by the total gigabytes you keep.	
3	You reserve a fixed amount of BI Engine capacity for a month.	
4	Pub/Sub cost depends on how many bytes are published.	

S2 Partition or cluster?

On-demand cost follows bytes processed. Selecting fewer columns helps; partitioning and clustering reduce the rows read. Below is a table of 30 daily partitions, each split into 26 storage blocks. Dark cells are what BigQuery must read.

Options: **Partition** ? **Cluster**

No.	Scenario	Answer
1	Almost every query filters to one day of a table holding months of events.	
2	Within each day, queries also filter by title prefix or customer id, and you want to skip most of the day’s blocks.	
3	Sort records by country, then region, then city, then postal code (least to most granular).	
4	You want to divide a 1 TB, five-day table so that one day costs roughly a fifth to query.	

Quiz from the chapter page: Check yourself

1. Which pricing model bills for an idle Cloud Composer environment?

   A. Usage-based

   B. Machine-based: machines × size × hours

   C. Flat rate

   D. None

---

2. When should a new BigQuery user usually start?

   A. With flat-rate slots, always

   B. With on-demand pricing, and reconsider once spend grows large

   C. With neither: BigQuery is free

   D. With whichever is more expensive

---

3 How do partitioned and clustered tables help under flat-rate pricing?

A. They do not help at all

B. They reduce the slots your queries need, so you may buy fewer

C. They make storage free

D. They remove the need for slots

---

4 You cluster on country, region, city and postal code. Why that order?

A. Alphabetical order

B. Least to most granular: country is the broadest, postal code the finest

C. Random order is fine

D. Most granular first

---

5 What must you do when the estimate shows Dataproc as by far the largest line?

A. Ignore it: estimates are inaccurate

B. Discuss whether to substitute the component or change the design, for example with ephemeral clusters

C. Delete the estimate

D. Buy more slots

---

6 Why check the partition quota before choosing hourly partitions?

A. Hourly partitions are not allowed

B. A table can hold only so many partitions, so hourly ones cover a much shorter history than daily

C. Partitions cost extra per hour

D. They disable clustering

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-10-cost-strategy-in-gcp/