Ch02

Big Data Capabilities on 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 the Guru Jazzy chapter page, for classroom use.

Sorting activities from the chapter page

S1 Place each scenario

Options: **VM-based** · **Managed** · **Serverless**

No.	Scenario	Answer
1	You need Elasticsearch, and Google offers no managed version of it.	
2	You want Spark jobs to run without installing Hadoop, but you still want to choose the cluster size and network.	
3	A 1 TB CSV must be queryable as a table with SQL, and you want zero infrastructure to set up.	
4	Your workload needs one exact older version of a tool and custom OS-level packages.	

S2 Which identity should do the work?

Options: **User account** · **Service account**

No.	Scenario	Answer
1	A Cloud Composer job loads files from Cloud Storage into BigQuery every night.	
2	An analyst runs an ad-hoc query in the BigQuery console.	
3	A build pipeline creates a Docker image and pushes it whenever code changes.	

S3 Quotas versus cost

Options: **Quota** · **Cost**

No.	Scenario	Answer
1	A table design needs more than 10,000 columns in BigQuery.	
2	A Cloud Composer environment was left running all month and the bill is high.	
3	You try to store a single 6 TB object in Cloud Storage.	
4	Your monthly BigQuery queries pass the free tier and charges begin.	

Quiz from the chapter page: Check yourself

1. A fully managed (serverless) service exists and fits your need. What is the sensible default?
   A. Use it, and fall back to VMs only for compatibility, features or budget
   B. Always install the software yourself on VMs
   C. Wait for a managed version of every tool
   D. Use whichever service is cheapest today

2. What is the difference between a project name and a project ID?
   A. The name must be globally unique; the ID need not be
   B. The ID must be globally unique; the name is only a friendly label
   C. They are always identical and cannot differ
   D. Only the name appears in billing

3. Why is it common on GCP to create a Spark cluster per job and delete it afterwards?
   A. Clusters cannot run more than one job
   B. It is required by Dataproc licensing
   C. Resources are dedicated to the job, and you pay only while it runs
   D. Deleting the cluster speeds up the next job

4. Which service is the managed version of Apache Airflow?
   A. Dataflow
   B. Cloud Composer
   C. Data Fusion
   D. Dataprep

5. Why does Google set quotas on its services?
   A. Only to make customers pay more
   B. To keep shared services available to everyone and to flag questionable designs
   C. Because the hardware cannot store more
   D. To slow down free-tier users

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-02-big-data-capabilities-on-gcp/