Ch05

Building a Data Lake with Dataproc

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 – Building a data lake on a Dataproc cluster**

In this exercise, we will use Dataproc to store and process log data. Log data is a good representation of unstructured data. Organizations often need to analyze log data to understand their users' behavior.

In the exercise, we will learn how to use HDFS and PySpark using different methods. In the beginning, we will use Cloud Shell to get a basic understanding of the technologies. In the later sections, we will use Cloud Shell Code Editor and submit the jobs to Dataproc. But for the first step, let's create our Dataproc cluster.

Ex2 **Exercise: Creating and running jobs on a Dataproc cluster**

In this exercise, we will try two different methods to submit a Dataproc job. In the previous exercise, we used the Spark shell to run our Spark syntax, which is common when practicing but not common in real development. Usually, we would only use the Spark shell for initial checking or testing simple things. In this exercise, we will code Spark jobs in editors and submit them as jobs.

Here are the scenarios that we want to try:

(a) Preparing log data in GCS and HDFS

(b) Developing Spark ETL from HDFS to HDFS

(c) Developing Spark ETL from GCS to GCS

(d) Developing Spark ETL from GCS to BigQuery

Ex3 **Practicing using a workflow template on Dataproc**

If we decide to use the ephemeral cluster approach, then a question may come up – how do we sync the cluster creation and deletion with a job?

There are many ways; you can do it manually and automatically. One automatic option is using the Dataproc workflow template. Let's practice using it to understand the next section.

A workflow template is a feature in Dataproc that can help you manage simple job dependencies. Using a workflow template, you can also define an ephemeral cluster as a part of jobs.

Ex4 **Building an ephemeral cluster using Dataproc and Cloud Composer**

In this section, we will use Airflow to create a Dataproc cluster, submit a pyspark job, and delete the cluster when finished.

Sorting activities from the chapter page

S1 **Where should this data live?**

Options: **HDFS** · **Cloud Storage** · **BigQuery table**

No.	Scenario	Answer
1	You want output files that BigQuery, AI tools and other GCP services can read, with the least maintenance.	
2	Your output is fully structured with a clear schema and analysts will query it with SQL.	
3	Job latency is critical and you are willing to run a cluster with SSD-backed disks.	

S2 **RDD or DataFrame?**

Options: **RDD** · **DataFrame**

No.	Scenario	Answer
1	Raw web-server log lines that have no clean delimiter or schema yet, and need parsing first.	
2	After parsing, you want to run SQL over the ip, date, method and url columns.	
3	A well-formed CSV with a known schema.	

S3 **Ephemeral or permanent?**

Options: **Ephemeral** · **Permanent**

No.	Scenario	Answer
1	A nightly Spark job reads and writes Cloud Storage. You want dedicated resources and no idle cost.	
2	Analysts use Jupyter notebooks and Hue on the cluster all day.	
3	Hive tables stored in HDFS are the lake's main data store.	
4	One job needs a specific OS version and library set that would conflict with other jobs.	

S4 **Automating the create-run-delete cycle**

Options: **Workflow template** · **Cloud Composer**

No.	Scenario	Answer
1	One simple PySpark job that just needs a cluster created before it and deleted after.	
2	A pipeline with many branches, backfill logic and dependencies on BigQuery and Cloud Storage tasks.	

Quiz from the chapter page: Check yourself

1 On GCP, what is usually the best replacement for HDFS as a lake's storage?

A. Cloud SQL

B. Cloud Storage

C. Compute Engine local disks

D. Memorystore

---

2 What does "scale out" mean?

A. Upgrade to a bigger single server

B. Add more servers

C. Delete unused servers

D. Move to another region

---

3 You write rdd.filter(...).map(...).map(...) but nothing runs until you call sum(). Why?

A. The cluster is idle

B. Spark evaluates lazily: transformations build a plan, actions execute it

C. The code has a bug

D. RDDs cannot be filtered

---

4. Which is a genuine reason NOT to use an ephemeral cluster?

A. You want to save money

B. Your data lives in HDFS or Hive on the cluster

C. The job reads Cloud Storage

D. You want isolated resources

---

5. Why give an ephemeral cluster a name suffixed with the run date in Airflow?

A. It makes the cluster cheaper

B. It prevents name collisions when several dates run at once, such as during backfills

C. Dataproc requires dates

D. It speeds up Spark

---

6. Why set an idle-delete time-to-live on a cluster created by a DAG?

A. To keep the cluster forever

B. So the cluster deletes itself if the DAG fails and never reaches the delete task

C. To enable autoscaling

D. To make jobs faster

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-05-building-a-data-lake-with-dataproc/