Ch12

Boosting Your Confidence as a Data Engineer

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.

Quiz from the book: Reviewing all the concepts you've learned about

B1 **Designing data processing systems ? Question 1:** Your company already has some existing Spark jobs on an on-premise cluster. The company has decided to migrate to Google Cloud Platform (GCP). The company has asked you to migrate all the Spark jobs to GCP as soon as possible. What is the best way to handle this?

A. Re-create the Spark jobs on BigQuery.

B. Migrate the Spark jobs to Dataproc and use Google Cloud Storage (GCS) to store the data.

C. Re-create the Spark jobs on Dataflow.

D. Migrate the Spark jobs to Google Compute Engine (GCE) and install the Spark cluster.

B2 **Designing data processing systems ? Question 2:** You are a data engineer in an Internet of Things (IoT) company. The IoT devices produce logs in a streaming fashion to the server. You are asked to design a data pipeline to process the data and store it in a data warehouse for analytical purposes. Which GCP services will you choose?

A. Pub/Sub, Dataflow, and BigQuery

B. Cloud SQL, Cloud Composer, and Bigtable

C. Dataproc, Firestore, and Data Studio

D. GCS, Spanner, and Cloud Build

B3 **Designing data processing systems ? Question 3:** Your company's mobile application team needs a database to store data about a product catalog. The data is in JSON format and is considered semi-structured data. Currently, the data has a size of 400 GB. What is the best database service in GCP that you can recommend to the mobile application team?

A. BigQuery

B. Cloud SQL

C. Cloud Firestore in Datastore mode

D. GCS

B4 **Building and operationalizing data processing systems ? Question 4:** You are a data engineer at an e-commerce company. Every time a customer buys a product from the website, the order data will be published to Pub/Sub. There are three systems that need the order data: the billing system, the shipping system, and the analytical dashboard. What is the best way to configure the Pub/Sub topic and subscription?

A. Create one topic and one subscription. All three systems will use the same subscription.

B. Create one topic for each order and create one subscription for each system.

C. Create one topic and three subscriptions. Each system will use its own subscription.

D. Create three topics and one subscription.

B5 **Building and operationalizing data processing systems ? Question 5:** You are asked to build a batch data pipeline that needs to run every day at 5 A.M. The data pipeline needs to extract data from multiple Cloud SQL databases, store the files in GCS, load the data into BigQuery, and run some SQL transformations in BigQuery. Which GCP service is best suited to orchestrate this data pipeline?

A. Dataflow

B. Cloud Composer

C. Cloud Functions

D. Data Studio

B6 **Building and operationalizing data processing systems ? Question 6:** You work in an organization that has three user groups: Data Admins, Data Providers, and Data Readers. They all use GCS and BigQuery. You want to grant access to each user group according to their responsibilities. What is the best practice to grant IAM roles to these user groups?

A. Grant the Owner basic role to all user groups at the organization level.

B. Grant predefined roles for GCS and BigQuery to each user group based on the principle of least privilege.

C. Grant the Editor basic role to all user groups at the project level.

D. Create custom roles for all user groups.

B7 **Operationalizing machine learning models ? Question 7:** Your company is building a new web application that allows users to upload photos of receipts. The application needs to extract text from the photos and analyze the sentiment of the text. The company does not have a data science team and wants to launch the feature in 2 weeks. What is the best approach to build this capability on GCP?

A. Build custom ML models using TensorFlow on Vertex AI.

B. Use GCP's pre-built ML APIs: the Cloud Vision API and the Natural Language API.

C. Use AutoML Vision and AutoML Natural Language to train new models.

D. Hire a data science team first before building the feature.

B8 **Operationalizing machine learning models ? Question 8:** You are asked to create an ML model to predict the total sales of a store for next month based on historical data. The output of the prediction is a continuous numerical value. What type of ML model should you use?

A. K-means clustering

B. Regression

C. Classification

D. Logistic regression

B9 **Operationalizing machine learning models ? Question 9:** A data scientist on your team trained an ML model. During training, the model achieved 99% accuracy on the training dataset. However, when tested on a new test dataset, the model only achieved 60% accuracy. What is the most likely problem with the model?

A. The model is underfitting.

B. The model is overfitting.

C. The training dataset is too large.

D. The test dataset is corrupted.

B10 **Ensuring solution quality ? Question 10:** You have a BigQuery table that stores daily event logs for the past 2 years. The table is 50 TB in size. Most of the queries run by the analysts only look at the last 10 days of data. Currently, the BigQuery on-demand cost is very high. What is the best way to optimize the cost without losing historical data?

A. Split the table into multiple tables for every 10 days.

B. Recreate the table as a date-partitioned table based on the event date column and ask analysts to filter by the partition column.

C. Export data older than 10 days to GCS Archive storage and delete it from BigQuery.

D. Switch to flat-rate pricing immediately without changing the table.

B11 **Ensuring solution quality ? Question 11:** You are using a Bigtable cluster with three nodes to serve a real-time application. Suddenly, one of the nodes in the Bigtable cluster fails. What will happen to the data stored in Bigtable?

A. One-third of the data will be lost and needs to be restored from backup.

B. The data will be recovered using RAID configuration on the node's local disk.

C. No data is lost because Bigtable separates compute nodes from storage (Colossus).

D. The data will automatically fail over to BigQuery.

B12 **Ensuring solution quality ? Question 12:** A business user reports that their Data Studio dashboard, which is connected to a BigQuery table, is not showing the latest data that was loaded into BigQuery 5 minutes ago. What is the first thing you should suggest to the business user?

A. Recreate the BigQuery table.

B. Click the Refresh data button in Data Studio to clear the cache and fetch the latest data from BigQuery.

C. Contact Google Cloud Support.

D. Clear their browser history and cookies.

Sorting activities from the chapter page

S1 **Which application database?**

The exam expects you to know the positioning of a few more services. You need to know what each is for and how to choose, not every feature.

Options: **Cloud SQL** ? **Firestore / Datastore** ? **Spanner** ? **Bigtable**

No.	Scenario	Answer
1	Structured order data for an application, under 1 TB a day, needing SQL.	
2	Flexible product-catalog documents for a web app, modest volume, semi-structured.	
3	Very high volume, structured data that needs global consistency and SQL.	
4	Over 10 TB of sensor readings, NoSQL, with a real-time app that needs high throughput and low latency.	

S2 **Bigtable in a nutshell**

The exam expects you to know the positioning of a few more services. You need to know what each is for and how to choose, not every feature.

Options: **Good key** ? **Hotspot risk**

No.	Scenario	Answer
1	Row key is the website domain name, so writes pile onto neighbouring rows.	
2	Row key is the plain event timestamp.	
3	Row key is the reversed domain name.	
4	Row key is a device id followed by the timestamp.	

Quiz from the chapter page: Review quiz: twelve scenarios

1 A company has dozens of Spark jobs on its own servers and wants to move to GCP quickly with little rewriting. What is the natural fit?

A. Rewrite everything in BigQuery SQL first

B. Run them on Dataproc, with data in Cloud Storage

C. Rebuild them as Dataflow pipelines

D. Install Spark by hand on Compute Engine

2 Logs stream from thousands of devices and must be processed and stored for analytics. Which combination?

A. Pub/Sub, Dataflow, BigQuery

B. Cloud SQL, Composer, Bigtable

C. Dataproc, Firestore, Looker

D. Cloud Storage, Spanner, Cloud Build

---

3 A mobile app needs a managed database for a 400 GB catalogue of semi-structured product documents. Which is the best fit?

A. BigQuery

B. Cloud SQL

C. Firestore (Datastore mode)

D. Cloud Storage Archive

---

4 Billing, shipping and analytics systems must each receive every order event independently. How should Pub/Sub be set up?

A. One topic, one shared subscription

B. One topic per order, one subscription each

C. One topic, one subscription per system

D. Three topics, one subscription

---

5 Every night at 05:00 you must extract from several Cloud SQL databases, land files in Cloud Storage, load BigQuery and run transformations. What should orchestrate it?

A. A Dataflow streaming job

B. A Cloud Composer DAG using native operators

C. A cron script on one VM

D. A Looker Studio schedule

---

6 Data admins, data providers and data readers all use Cloud Storage and BigQuery. How should access be granted?

A. Owner for everyone at organisation level

B. Predefined roles per service for each group, at the narrowest sensible level

C. Editor for all three groups

D. Three custom roles, always

---

7 A web app must extract text from uploaded photos and score its sentiment, with no ML staff, within weeks. What approach?

A. Train a TensorFlow model for each task

B. Use pre-built vision and natural language APIs

C. Build an AutoML model for text detection

D. Hire a research team first

---

8 You want to predict tomorrow's electricity demand, a numeric value, from past data. Which type of model fits?

A. K-means clustering

B. Regression

C. Principal component analysis

D. Classification into "high" and "low" only

---

9 A model scores 99% on its training data and 60% on new data. What is the likely problem?

A. The cluster is too small

B. Overfitting: it has memorised the training data, so evaluate on held-out data and add data or simplify

C. The features are too few in the API

D. Online prediction is broken

---

10 A BigQuery table holds two years of events, but most queries need only the last 10 days, and cost keeps rising. What helps most?

A. A new table every 10 days joined together

B. Partition by date and make queries filter on the partition column

C. Move all old data to Archive storage

D. Buy more slots

---

11 One node in a Bigtable cluster fails. What happens to the data?

A. It is lost and must be restored from a backup

B. It is rebuilt from RAID disks

C. Nothing is lost, because storage is separate from the compute nodes

D. It moves to BigQuery

---

12 A BI report on BigQuery shows numbers that look stale. What is the first thing to try?

A. Recreate the BigQuery table

B. Refresh the report's data (or check its data-freshness setting) so it re-queries instead of serving cached results

C. Open a ticket with the infrastructure team

D. Clear browser cookies

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-12-certification-and-review/