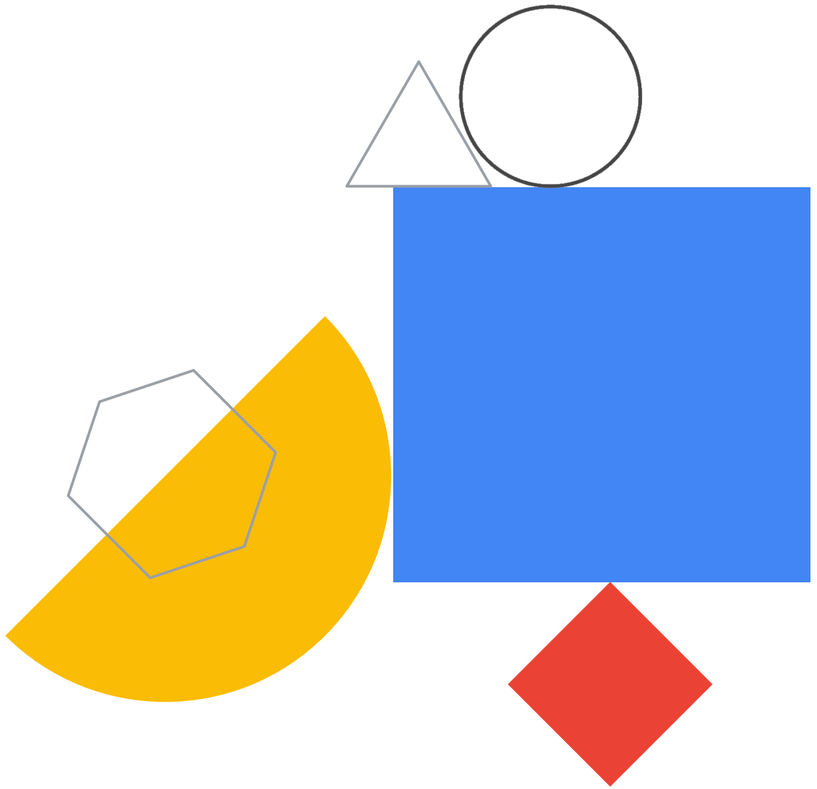
Google Cloud

Preparing for Your Professional Data Engineer Journey



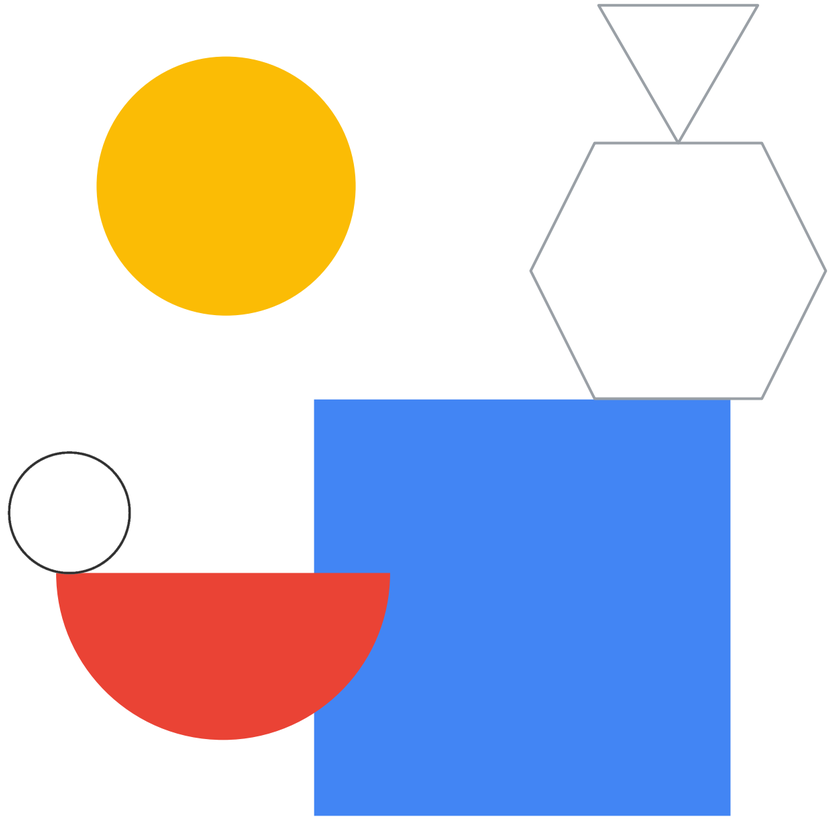
Course Workbook

Certification Exam Guide Sections

1	Designing Data Processing Systems
2	Ingesting and Processing the Data
3	Storing the Data
4	Preparing and Using Data for Analysis
5	Maintaining and Automating Data Workloads

Google Cloud

Section 1: Designing Data Processing Systems



1.1 Diagnostic Question 01

Business analysts in your team need to run analysis on data that was loaded into BigQuery. You need to follow recommended practices and grant permissions.

What role should you grant the business analysts?

A. bigquery.resourceViewer and bigquery.dataViewer

B. bigquery.user and bigquery.dataViewer

C. bigquery.dataOwner

D. storage.objectViewer and bigquery.user

1.1 Diagnostic Question 02

Cymbal Retail has acquired another company in Europe. Data access permissions and policies in this new region differ from those in Cymbal Retail's headquarters, which is in North America. You need to define a consistent set of policies for projects in each region that follow recommended practices.

What should you do?

A. Create a new organization for all projects in Europe and assign policies in each organization that comply with regional laws.

B. Implement a flat hierarchy, and assign policies to each project according to its region.

C. Create top level folders for each region, and assign policies at the folder level.

D. Implement policies at the resource level that comply with regional laws.

1.1 Diagnostic Question 03

You are migrating on-premises data to a data warehouse on Google Cloud. This data will be made available to business analysts. Local regulations require that customer information including credit card numbers, phone numbers, and email IDs be captured, but not used in analysis. You need to use a reliable, recommended solution to redact the sensitive data.

What should you do?

A. Use the Cloud Data Loss Prevention (DLP) API to identify and redact data that matches infoTypes like credit card numbers, phone numbers, and email IDs.

B. Delete all columns with a title similar to "credit card," "phone," and "email."

C. Create a regular expression to identify and delete patterns that resemble credit card numbers, phone numbers, and email IDs.

D. Use the Cloud Data Loss Prevention (DLP) API to perform date shifting of any entries with credit card numbers, phone numbers, and email IDs.

1.1 Diagnostic Question 04

Your data and applications reside in multiple geographies on Google Cloud. Some regional laws require you to hold your own keys outside of the cloud provider environment, whereas other laws are less restrictive and allow storing keys with the same provider who stores the data. The management of these keys has increased in complexity, and you need a solution that can centrally manage all your keys.

What should you do?

A. Enable confidential computing for all your virtual machines.

B. Store keys in Cloud Key Management Service (KMS), and reduce the number of days for automatic key rotation.

C. Store your keys in Cloud Hardware Security Module (HSM), and retrieve keys from it when required.

D. Store your keys on a supported external key management partner, and use Cloud

External Key Manager (EKM) to get keys when required.

1.1 Designing for security and compliance

Courses

Modernizing Data Lakes and Data Warehouses with Google Cloud
- Introduction to Data Engineering
- Building a Data Lake
- Building a Data Warehouse

Smart Analytics, Machine Learning, and AI on Google Cloud
- Prebuilt ML Model APIs for Unstructured Data

Serverless Data Processing with Dataflow: Foundations
- IAM, Quotas, and Permissions
- Security

Skill Badges

Create and Manage Cloud Resources

Perform Foundational Data, ML, and AI Tasks in Google Cloud

Documentation

Secure a BigQuery data warehouse that stores confidential data
IAM basic and predefined roles reference
Creating and managing Folders
Resource hierarchy
Cloud Data Loss Prevention
InfoType detector reference
Cloud External Key Manager
Hold your own key with Google Cloud External Key Manager
Evolving Cloud External Key Manager – What's new with Cloud EKM | Google Cloud Blog

1.2 Diagnostic Question 05

Cymbal Retail has a team of business analysts who need to fix and enhance a set of large input data files. For example, duplicates need to be removed, erroneous rows should be deleted, and missing data should be added. These steps need to be performed on all the present set of files and any files received in the future in a repeatable, automated process. The business analysts are not adept at programming.

What should they do?

A. Load the data into Dataprep, explore the data, and edit the transformations as needed.

B. Create a Dataproc job to perform the data fixes you need.

C. Create a Dataflow pipeline with the data fixes you need.

D. Load the data into Google Sheets, explore the data, and fix the data as needed.

1.2 Diagnostic Question 06

You have a Dataflow pipeline that runs data processing jobs. You need to identify the parts of the pipeline code that consume the most resources.

What should you do?

A. Use Cloud Monitoring

B. Use Cloud Logging

C. Use Cloud Profiler

D. Use Cloud Audit Logs

1.2 Designing for reliability and fidelity

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud]
- Building a Data Warehouse

[Building Batch Data Pipelines on Google Cloud]
- Introduction to Building Batch Data Pipelines
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

[Building Resilient Streaming Analytics Systems on Google Cloud]
- Serverless Messaging with Pub/Sub

[Serverless Data Processing with Dataflow: Develop Pipelines]
- Best Practices

[Serverless Data Processing with Dataflow: Operations]
- Monitoring
- Logging and Error Reporting
- Troubleshooting and Debug
- Testing and CI/CD
- Reliability

Skill Badges

[Perform Foundational Data, ML, and AI Tasks in Google Cloud]

[Engineer Data with Google Cloud]

Documentation

[Clean and Enhance Your Data]

[Introduction to Data Wrangling]

[Monitoring pipeline performance using Cloud Profiler | Cloud Dataflow]

1.3 Diagnostic Question 07

You are using Dataproc to process a large number of CSV files. The storage option you choose needs to be flexible to serve many worker nodes in multiple clusters. These worker nodes will read the data and also write to it for intermediate storage between processing jobs.

What is the recommended storage option on Google Cloud?

A. Cloud SQL

B. Zonal persistent disks

C. Local SSD

D. Cloud Storage

1.3 Diagnostic Question 08

You are managing the data for Cymbal Retail, which consists of multiple teams including retail, sales, marketing, and legal. These teams are consuming data from multiple producers including point of sales systems, industry data, orders, and more. Currently, teams that consume data have to repeatedly ask the teams that produce it to verify the most up-to-date data and to clarify other questions about the data, such as source and ownership. This process is unreliable and time-consuming and often leads to repeated escalations. You need to implement a centralized solution that gains a unified view of the organization's data and improves searchability.

What should you do?

A. Implement a data mesh with Dataplex and have producers tag data when created.

B. Implement a data lake with Cloud Storage, and create buckets for each team such as retail, sales, marketing.

C. Implement a data warehouse by using BigQuery, and create datasets for each team such as retail, sales, marketing.

D. Implement Looker dashboards that provide views of the data that meet each teams' requirements.

1.3 Designing for flexibility and portability

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud](#)
- Introduction to Data Engineering
- Building a Data Lake

[Building Batch Data Pipelines on Google Cloud](#)
- Introduction to Building Batch Data Pipelines

[Serverless Data Processing with Dataflow: Foundations](#)
- Beam Portability

Skill Badges

[Get Started with Dataplex](#)

Documentation

[Dataproc best practices | Google Cloud Blog](#)
[HDFS vs. Cloud Storage: Pros, cons and migration tips | Google Cloud Blog](#)
[Dataplex overview](#)

1.4 Diagnostic Question 09

Laws in the region where you operate require that files related to all orders made each day are stored immutably for 365 days. The solution that you recommend has to be cost-effective.

What should you do?

A. Store the data in a Cloud Storage bucket, and enable object versioning and delete any version older than 365 days.

B. Store the data in a Cloud Storage bucket, and specify a retention period.

C. Store the data in a Cloud Storage bucket, and set a lifecycle policy to delete the file after 365 days.

D. Store the data in a Cloud Storage bucket, enable object versioning, and delete any version greater than 365.

1.4 Diagnostic Question 10

Cymbal Retail is migrating its private data centers to Google Cloud. Over many years, hundreds of terabytes of data were accumulated. You currently have a 100 Mbps line and you need to transfer this data reliably before commencing operations on Google Cloud in 45 days.

What should you do?

A. Store the data in an HTTPS endpoint, and configure Storage Transfer Service to copy the data to Cloud Storage.

B. Upload the data to Cloud Storage by using gsutil.

C. Zip and upload the data to Cloud Storage buckets by using the Google Cloud console.

D. Order a transfer appliance, export the data to it, and ship it to Google.

1.4 Designing data migrations

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Lake
- Building a Data Warehouse

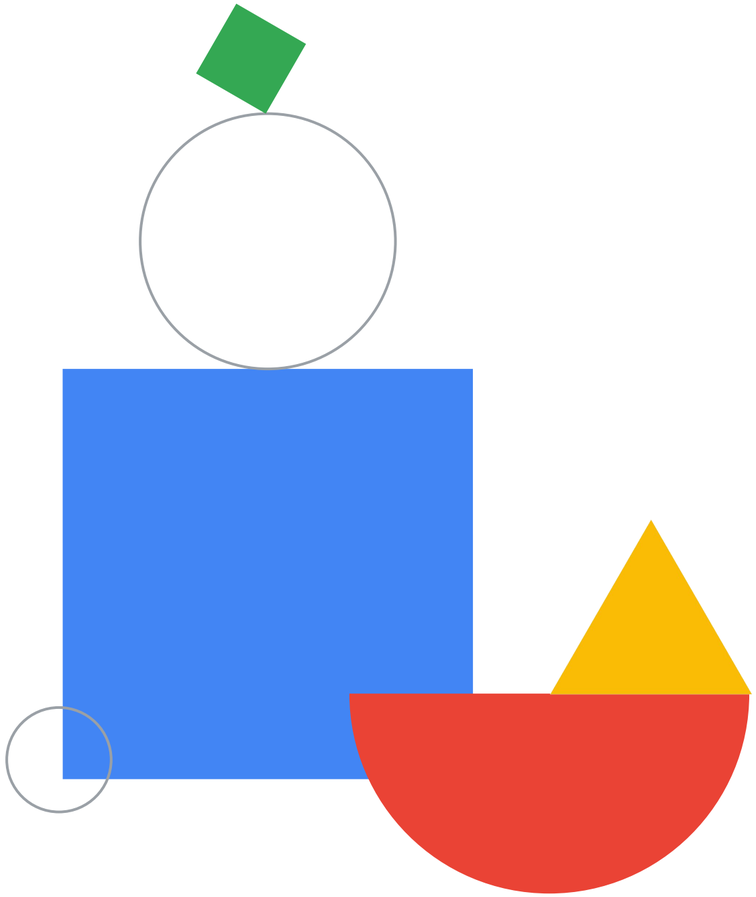
BigQuery Fundamentals for Redshift Professionals
- BigQuery and Google Cloud IAM

Documentation

Retention policies and retention policy locks | Cloud Storage

Migration to Google Cloud: Transferring your large datasets

Google Cloud



Section 2: Ingesting and Processing the Data

2.1 Diagnostic Question 01

Your data engineering team receives data in JSON format from external sources at the end of each day. You need to design the data pipeline.

What should you do?

A. Store the data in Cloud Storage and create an extract, transform, and load (ETL) pipeline.

B. Make your BigQuery data warehouse public and ask the external sources to insert the data.

C. Create a public API to allow external applications to add the data to your warehouse.

D. Store the data in persistent disks and create an ETL pipeline.

2.1 Diagnostic Question 02

The first stage of your data pipeline processes tens of terabytes of financial data and creates a sparse, time-series dataset as a key-value pair.

Which of these is a suitable sink for the pipeline's first stage?

A. Cloud Storage
B. Cloud SQL
C. AlloyDB
D. Bigtable

2.1 Diagnostic Question 03

You are processing large amounts of input data in BigQuery. You need to combine this data with a small amount of frequently changing data that is available in Cloud SQL.

What should you do?

A. Copy the data from Cloud SQL to a new BigQuery table hourly.

B. Copy the data from Cloud SQL and create a combined, normalized table hourly.

C. Use a federated query to get data from Cloud SQL.

D. Create a Dataflow pipeline to combine the BigQuery and Cloud SQL data when the Cloud SQL data changes.

2.1 Planning the data pipelines

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Introduction to Data Engineering
- Building a Data Lake
- Building a Data Warehouse

Building Batch Data Pipelines on Google Cloud
- Executing Spark on Dataproc
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Building Resilient Streaming Analytics Systems on Google Cloud
- High-Throughput BigQuery and Bigtable Streaming Features

Serverless Data Processing with Dataflow: Develop Pipelines
- Beam Concepts Review
- Sources and Sinks
- Schemas

Skill Badges

Perform Foundational Data, ML, and AI Tasks in Google Cloud
Engineer Data with Google Cloud

Documentation

What Data Pipeline Architecture should I use? | Google Cloud Blog

Bigtable overview

Cloud SQL federated queries | BigQuery

Exploring new features in BigQuery federated queries | Google Cloud Blog

2.2 Diagnostic Question 04

Your company has multiple data analysts but a limited data engineering team. You need to choose a tool where the analysts can build data pipelines themselves with a graphical user interface.

Which of these products is the most appropriate?

A. Dataflow
B. Cloud Data Fusion
C. Dataproc
D. Cloud Composer

2.2 Diagnostic Question 05

You manage a PySpark batch data pipeline by using Dataproc. You want to take a hands-off approach to running the workload, and you do not want to provision and manage your own cluster.

What should you do?

A. Configure the job to run on Dataproc Serverless.
B. Configure the job to run with Spot VMs.
C. Rewrite the job in Spark SQL.
D. Rewrite the job in Dataflow with SQL.

2.2 Diagnostic Question 06

You need to run batch jobs, which could take many days to complete. You do not want to manage the infrastructure provisioning.

What should you do?

A. Use Cloud Scheduler to run the jobs.
B. Use Workflows to run the jobs.
C. Run the jobs on Batch.
D. Use Cloud Run to run the jobs.

2.2 Diagnostic Question 07

You are creating a data pipeline for streaming data on Dataflow for Cymbal Retail's point of sales data. You want to calculate the total sales per hour on a continuous basis.

Which of these windowing options should you use?

A. Hopping windows (sliding windows in Apache Beam)
B. Session windows
C. Global window
D. Tumbling windows (fixed windows in Apache Beam)

2.2 Diagnostic Question 08

You want to build a streaming data analytics pipeline in Google Cloud. You need to choose the right products that support streaming data.

Which of these would you choose?

A. Pub/Sub, Dataflow, BigQuery
B. Pub/Sub, Dataprep, BigQuery
C. Cloud Storage, Dataflow, Cloud SQL
D. Cloud Storage, Dataprep, AlloyDB

2.2 Building the pipelines

Courses

Building Batch Data Pipelines on Google Cloud
- Introduction to Building Batch Data Pipelines
- Executing Spark on Dataproc
- Serverless Data Processing with Dataflow
- Manage Data Pipelines with Cloud Data Fusion and Cloud Compose

Building Resilient Streaming Analytics Systems on Google Cloud
- Serverless Messaging with Pub/Sub
- Dataflow Streaming Features

Serverless Data Processing with Dataflow: Foundations
- Separating Compute and Storage with Dataflow

Serverless Data Processing with Dataflow: Develop Pipelines
- Windows, Watermarks, and Triggers
- States and Timers
- Dataflow SQL and DataFrames

Serverless Data Processing with Dataflow: Operations
- Performance
- Testing and CI/CD
- Flex Templates

Skill Badges

Perform Foundational Data, ML, and AI Tasks in Google Cloud

Documentation

Cloud Data Fusion overview

What is Dataproc Serverless?

Introduction to Google Batch

Get started with Batch | Google Cloud

Streaming pipelines | Cloud Dataflow

Basics of the Beam model

Streaming analytics solutions | Google Cloud

2.3 Diagnostic Question 09

You have a data pipeline that requires you to monitor a Cloud Storage bucket for a file, start a Dataflow job to process data in the file, run a shell script to validate the processed data in BigQuery, and then delete the original file. You need to orchestrate this pipeline by using recommended tools.

Which product should you choose?

A. Cloud Tasks
B. Cloud Composer
C. Cloud Scheduler
D. Cloud Run

2.3 Diagnostic Question 10

You are running Dataflow jobs for data processing. When developers update the code in Cloud Source Repositories, you need to test and deploy the updated code with minimal effort.

A. Terraform
B. Compute Engine
C. Cloud Code
D. Cloud Build

Which of these would you use to build your continuous integration and delivery (CI/CD) pipeline for data processing?

2.3 Deploying and operationalizing the pipelines

Courses

[Building Batch Data Pipelines on Google Cloud]
- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

[Serverless Data Processing with Dataflow: Operations]
- Testing and CI/CD

Skill Badges

[Engineer Data with Google Cloud]

Documentation

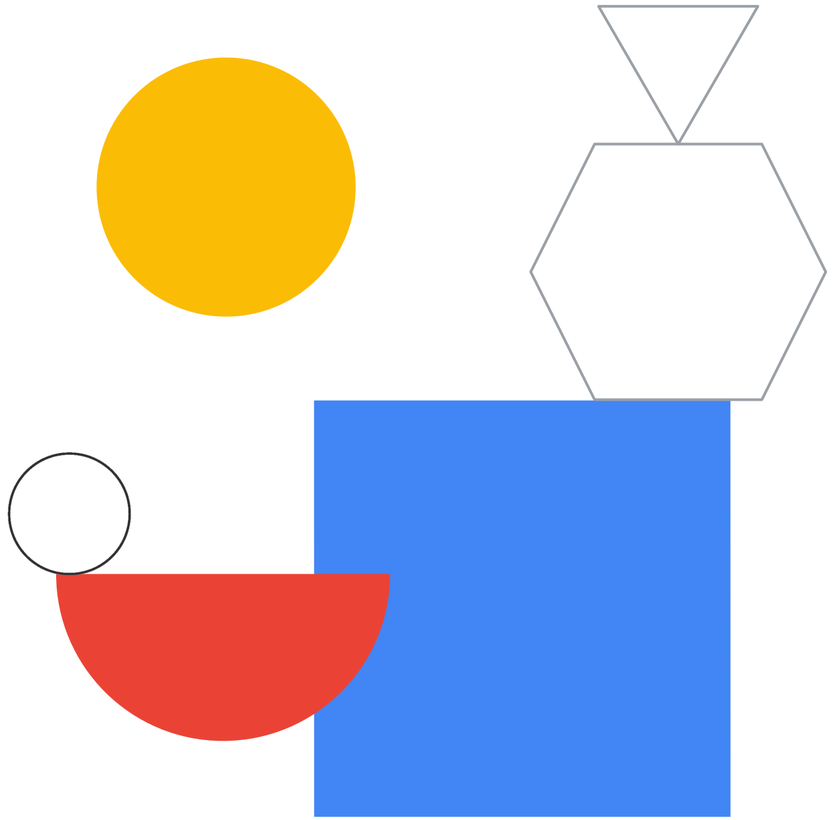
[How to use Cloud Composer for data orchestration]

[Cloud Composer overview]

[Use a CI/CD pipeline for data-processing workflows | Google Cloud]

Google Cloud

Section 3:
Storing the Data



3.1 Diagnostic Question 01

You need to choose a data storage solution to support a transactional system. Your customers are primarily based in one region. You want to reduce your administration tasks and focus engineering effort on building your business application.

What should you do?

A. Use Cloud Spanner.
B. Use Cloud SQL.
C. Install a database of your choice on a Compute Engine VM.
D. Create a Cloud Storage bucket with a regional bucket.

3.1 Diagnostic Question 02

You need to store data long term and use it to create quarterly reports.

What storage class should you choose?

A. Standard storage class is the recommended option when the data is accessed frequently, such as daily or weekly.

B. Nearline storage class is the recommended option when the data is accessed less frequently, such as once a month.

C. Coldline storage class is the recommended option when the data is accessed infrequently, such as once a quarter.

D. Archive storage class is the recommended option when the data is accessed rarely, like once a year or less.

3.1 Selecting storage systems

Courses

Google Cloud Big Data and Machine Learning Fundamentals

- Big Data and Machine Learning on Google Cloud

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to data engineering
- Building a data lake
- Building a data warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- High-Throughput BigQuery and Bigtable Streaming Features

Documentation

Cloud SQL for MySQL, PostgreSQL, and SQL Server

What is Cloud SQL?

Storage classes | Google Cloud

3.2 Diagnostic Question 03

You have several large tables in your transaction databases. You need to move all the data to BigQuery for the business analysts to explore and analyze the data.

How should you design the schema in BigQuery?

A. Retain the data on BigQuery with the same schema as the source.
B. Combine all the transactional database tables into a single table using outer joins.
C. Redesign the schema to normalize the data by removing all redundancies.
D. Redesign the schema to denormalize the data with nested and repeated data.

3.2 Diagnostic Question 04

You are ingesting data that is spread out over a wide range of dates into BigQuery at a fast rate. You need to partition the table to make queries performant.

What should you do?

A. Create an ingestion-time partitioned table with daily partitioning type.

B. Create an ingestion-time partitioned table with yearly partitioning type.

C. Create an integer-range partitioned table.

D. Create a time-unit column-partitioned table with yearly partitioning type.

3.2 Diagnostic Question 05

Your analysts repeatedly run the same complex queries that combine and filter through a lot of data on BigQuery. The data changes frequently. You need to reduce the effort for the analysts.

What should you do?

A. Create a dataset with the data that is frequently queried.

B. Create a view of the frequently queried data.

C. Export the frequently queried data into a new table.

D. Export the frequently queried data into Cloud SQL.

3.2 Planning for using a data warehouse

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud]

- Building a data warehouse

[Building Resilient Streaming Analytics Systems on Google Cloud]

- Advanced BigQuery functionality and performance

Skill Badges

[Build and Optimize Data Warehouses with BigQuery]

Documentation

[Introduction to optimizing query performance | BigQuery | Google Cloud]

[Introduction to partitioned tables | BigQuery | Google Cloud]

[Creating partitioned tables | BigQuery | Google Cloud]

[Introduction to views | BigQuery | Google Cloud]

3.3 Diagnostic Question 06

You have data that is ingested daily and frequently analyzed in the first month. Thereafter, the data is retained only for audits, which happen occasionally every few years. You need to configure cost-effective storage.

What should you do?

A. Create a bucket on Cloud Storage with object versioning configured.
B. Create a bucket on Cloud Storage with Autoclass configured.
C. Configure a data retention policy on Cloud Storage.
D. Configure a lifecycle policy on Cloud Storage.

3.3 Diagnostic Question 07

You have data stored in a Cloud Storage bucket. You are using both Identity and Access Management (IAM) and Access Control Lists (ACLs) to configure access control. Which statement describes a user's access to objects in the bucket?

Which statement describes a user's access to objects in the bucket?

A. The user has no access if IAM denies the permission.
B. The user only has access if both IAM and ACLs grant a permission.
C. The user has access if either IAM or ACLs grant a permission.
D. The user has no access if either IAM or ACLs deny a permission.

3.3 Diagnostic Question 08

A manager at Cymbal Retail expresses concern about unauthorized access to objects in your Cloud Storage bucket. You need to evaluate all access on all objects in the bucket.

What should you do?

A. Review the Admin Activity audit logs.

B. Enable and then review the Data Access audit logs.

C. Route the Admin Activity logs to a BigQuery sink and analyze the logs with SQL queries.

D. Change the permissions on the bucket to only trusted employees.

3.3 Using a data lake

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a data lake

Documentation

Cloud Storage

Object Lifecycle Management | Cloud Storage

Overview of access control | Cloud Storage

Cloud Audit Logs with Cloud Storage | Google Cloud

3.4 Diagnostic Question 09

Cymbal Retail has accumulated a large amount of data. Analysts and leadership are finding it difficult to understand the meaning of the data, such as BigQuery columns. Users of the data don't know who owns what. You need to improve the searchability of the data.

What should you do?

A. Create tags for data entries in Cloud Catalog.
B. Rename BigQuery columns with more descriptive names.
C. Export the data to Cloud Storage with descriptive file names.
D. Add a description column corresponding to each data column.

3.4 Diagnostic Question 10

You have large amounts of data stored on Cloud Storage and BigQuery. Some of it is processed, but some is yet unprocessed. You have a data mesh created in Dataplex. You need to make it convenient for internal users of the data to discover and use the data.

What should you do?

A. Create a lake for Cloud Storage data and a zone for BigQuery data.

B. Create a lake for BigQuery data and a zone for Cloud Storage data.

C. Create a lake for unprocessed data and assets for processed data.

D. Create a raw zone for the unprocessed data and a curated zone for the processed data.

3.4 Designing for a data mesh

Courses

[Modernizing Data Lakes and Data Warehouses on Google Cloud]

- Introduction to data engineering

[Building Batch Data Pipelines on Google Cloud]

- Introduction to building batch data pipelines

Skill Badges

[Data Catalog Fundamentals]

Documentation

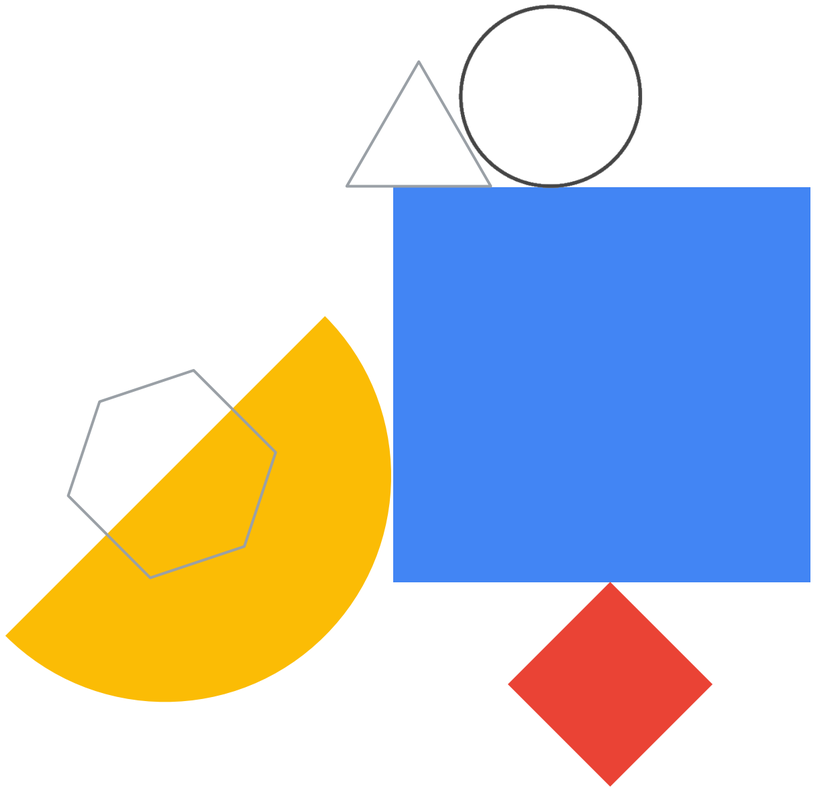
[Tags and tag templates | Data Catalog Documentation | Google Cloud]

[Quickstart: Tag a BigQuery table by using Data Catalog]

[Dataplex overview | Google Cloud]

Google Cloud

Section 4: Preparing and Using Data for Analysis



4.1 Diagnostic Question 01

Your company uses Google Workspace and your leadership team is familiar with its business apps and collaboration tools. They want a cost-effective solution that uses their existing knowledge to evaluate, analyze, filter, and visualize data that is stored in BigQuery.

What should you do to create a solution for the leadership team?

A. Create models in Looker.
B. Configure Connected Sheets.
C. Configure Tableau.
D. Configure Looker Studio.

4.1 Diagnostic Question 02

You have data in PostgreSQL that was designed to reduce redundancy. You are transferring this data to BigQuery for analytics. The source data is hierarchical and frequently queried together. You need to design a BigQuery schema that is performant.

What should you do?

A. Use nested and repeated fields.

B. Retain the data in normalized form always.

C. Copy the primary tables and use federated queries for secondary tables.

D. Copy the normalized data into partitions.

4.1 Diagnostic Question 03

You repeatedly run the same queries by joining multiple tables. The original tables change about ten times per day. You want an optimized querying approach.

Which feature should you use?

A. Views
B. Materialized views
C. Federated queries
D. Partitions

4.1 Diagnostic Question 04

You have analytics data stored in BigQuery. You need an efficient way to compute values across a group of rows and return a single result for each row.

What should you do?

A. Use an aggregate function.
B. Use a UDF (user-defined function).
C. Use BigQuery ML.
D. Use a window function with an OVER clause.

4.1 Diagnostic Question 05

You need to optimize the performance of queries in BigQuery. Your tables are not partitioned or clustered.

What optimization technique can you use?

A. Batch your updates and inserts.
B. Use the LIMIT clause to reduce the data read.
C. Filter data as late as possible.
D. Perform self-joins on data.

4.1 Preparing data for visualization

Courses

Google Cloud Big Data and Machine Learning Fundamentals

- Data Engineering for streaming data

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a data warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- Dataflow streaming features
- Advanced BigQuery functionality and performance

Serverless Data Processing with Dataflow: Develop Pipelines

- Windows, watermarks, and triggers

Skill Badges

Perform Foundational Data, ML, and AI Tasks in Google Cloud

Engineer Data with Google Cloud

Documentation

Introduction to analysis and business intelligence tools

Use nested and repeated fields

Introduction to materialized views

Window function calls

Optimize query computation

Optimize query computation

4.2 Diagnostic Question 06

Your data in BigQuery has some columns that are extremely sensitive. You need to enable only some users to see certain columns.

What should you do?

A. Create a new dataset with the column's data.
B. Create a new table with the column's data.
C. Use policy tags.
D. Use Identity and Access Management (IAM) permissions.

4.2 Diagnostic Question 07

Your business has collected industry-relevant data over many years. The processed data is useful for your partners and they are willing to pay for its usage. You need to ensure proper access control over the data.

What should you do?

A. Export the data to zip files and share it through Cloud Storage.
B. Host the data on Analytics Hub.
C. Export the data to persistent disks and share it through an FTP endpoint.
D. Host the data on Cloud SQL.

4.2 Diagnostic Question 08

You have a complex set of data that comes from multiple sources. The analysts in your team need to analyze the data, visualize it, and publish reports to internal and external stakeholders. You need to make it easier for the analysts to work with the data by abstracting the multiple data sources.

What tool do you recommend?

A. Looker Studio
B. Connected Sheets
C. D3.js library
D. Looker

4.2 Sharing data

Courses

Google Cloud Big Data and Machine Learning Fundamentals

- Data Engineering for Streaming Data

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to Data Engineering

Building Batch Data Pipelines on Google Cloud

- Introduction to Building Batch Data Pipelines

Skill Badges

Data Catalog Fundamentals

Documentation

Introduction to column-level access control

Analytics Hub | Data Exchange and Data Sharing | Google Cloud

Introduction to Analytics Hub | BigQuery

Secure data exchanges and data sharing with Analytics Hub

Looker business intelligence platform embedded analytics

4.3 Diagnostic Question 09

You built machine learning (ML) models based on your own data. In production, the ML models are not giving satisfactory results. When you examine the data, it appears that the existing data is not sufficiently representing the business goals. You need to create a more accurate machine learning model.

What should you do?

A. Train the model with more of similar data.

B. Perform L2 regularization.

C. Perform feature engineering, and use domain knowledge to enhance the column data.

D. Train the model with the same data, but use more epochs.

4.3 Diagnostic Question 10

You used Dataplex to create lakes and zones for your business data. However, some files are not being discovered.

What could be the issue?

A. You have an exclude pattern that matches the files.
B. You have scheduled discovery to run every hour.
C. The files are in ORC format.
D. The files are in Parquet format.

4.3 Exploring and analyzing data

Courses

[Google Cloud Big Data and Machine Learning Fundamentals]

- Big Data with BigQuery
- The machine learning workflow with Vertex A

[Modernizing Data Lakes and Data Warehouses on Google Cloud]

- Introduction to Data Engineering

[Building Batch Data Pipelines on Google Cloud]

- Introduction to building batch data pipelinesi

[Smart Analytics, Machine Learning, and AI on Google Cloud]

- Custom model building with SQL in BigQuery ML

Skill Badges

[Engineer Data with Google Cloud]

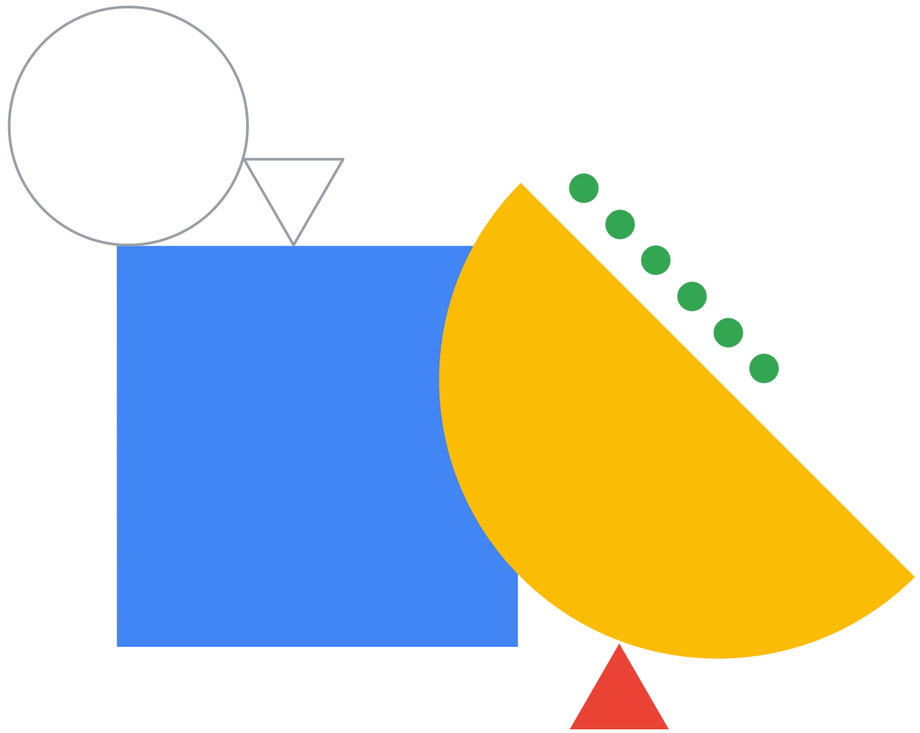
Documentation

[Use the BigQuery ML TRANSFORM clause for feature engineering | Google Cloud]

[Feature preprocessing overview | BigQuery | Google Cloud]

[Discover data | Dataplex | Google Cloud]

Google Cloud



Section 5: Maintaining and Automating Data Workloads

5.1 Diagnostic Question 01

You need to design a Dataproc cluster to run multiple small jobs. Many jobs (but not all) are of high priority.

What should you do?

A. Reuse the same cluster and run each job in sequence.
B. Reuse the same cluster to run all jobs in parallel.
C. Use ephemeral clusters.
D. Use cluster autoscaling.

5.1 Optimizing resources

Courses

[Building Batch Data Pipelines on Google Cloud]
- Executing Spark on Dataproc

Documentation

[Dataproc Job Optimization How-to Guide | Google Cloud Blog]

5.2 Diagnostic Question 02

You need to create repeatable data processing tasks by using Cloud Composer. You need to follow best practices and recommended approaches.

What should you do?

A. Write each task to be responsible for one operation.

B. Use current time with the `now( )` function for computation.

C. Update data with `INSERT` statements during the task run.

D. Combine multiple functionalities in a single task execution.

5.2 Designing automation and repeatability

Courses

Building Batch Data Pipelines on Google Cloud

- Manage Data Pipelines with Cloud Data Fusion and Cloud Composer

Serverless Data Processing with Dataflow: Develop Pipelines

- Best Practices

Skill Badges

Engineer Data with Google Cloud

Documentation

Write Airflow DAGs | Cloud Composer

DAGs — Airflow Documentation

DAG writing best practices in Apache Airflow | Astronomer Documentation

5.3 Diagnostic Question 03

Multiple analysts need to prepare reports on Monday mornings due to which there is heavy utilization of BigQuery. You want to take a cost-effective approach to managing this demand.

What should you do?

A. Use on-demand pricing.
B. Use Flex Slots.
C. Use BigQuery Enterprise edition with a one-year commitment.
D. Use BigQuery Enterprise Plus edition with a three-year commitment.

5.3 Diagnostic Question 04

You have a team of data analysts that run queries interactively on BigQuery during work hours. You also have thousands of report generation queries that run simultaneously. You often see an error: *Exceeded rate limits: too many concurrent queries for this project_and_region*.

How would you resolve this issue?

A. Run all queries in interactive mode.
B. Create a yearly reservation of BigQuery slots.
C. Run the report generation queries in batch mode.
D. Create a view to run the queries.

5.3 Organizing workloads based on business requirements

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Introduction to Data Engineering
- Building a Data Warehouse

Building Resilient Streaming Analytics Systems on Google Cloud

- Advanced BigQuery Functionality and Performance

Documentation

Scale cloud data warehouse up and down quickly

Introduction to reservations | BigQuery | Google Cloud

Introduction to BigQuery editions | Google Cloud

Run a query | BigQuery | Google Cloud

Troubleshoot quota and limit errors | BigQuery | Google Cloud

5.4 Diagnostic Question 05

You have a Dataflow pipeline in production. For certain data, the system seems to be stuck longer than usual. This is causing delays in the pipeline execution. You want to reliably and proactively track and resolve such issues.

What should you do?

A. Review the Dataflow logs regularly.

B. Set up alerts with Cloud Functions code that reviews the audit logs regularly.

C. Review the Cloud Monitoring dashboard regularly.

D. Set up alerts on Cloud Monitoring based on system lag.

5.4 Diagnostic Question 06

When running Dataflow jobs, you see this error in the logs: *"A hot key HOT_KEY_NAME was detected in..."*. You need to resolve this issue and make the workload performant.

What should you do?

A. Disable Dataflow shuffle.

B. Increase the data with the hot key.

C. Ensure that your data is evenly distributed.

D. Add more compute instances for processing.

5.4 Diagnostic Question 07

A colleague at Cymbal Retail asks you about the configuration of Dataproc autoscaling for a project.

What would be the Google-recommended situation when you should enable autoscaling?

A. When you want to scale on-cluster Hadoop Distributed File System (HDFS).
B. When you want to scale out single-job clusters.
C. When you want to down-scale idle clusters to minimum size.
D. When there are different size workloads on the cluster.

5.4 Monitoring and troubleshooting processes

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud
- Introduction to Data Engineering

Building Batch Data Pipelines on Google Cloud
- Executing Spark on Dataproc

Building Resilient Streaming Analytics Systems on Google Cloud
- Serverless Messaging with Pub/Sub
- Advanced BigQuery Functionality and Performance

Serverless Data Processing with Dataflow: Foundations
- IAM, Quotas, and Permissions

Serverless Data Processing with Dataflow: Develop Pipelines
- State and Timers
- Best Practices

Serverless Data Processing with Dataflow: Operations
- Monitoring
- Troubleshooting and Debug
- Reliability

Skill Badges

Perform Foundational Data, ML, and AI Tasks in Google Cloud

Documentation

Use Cloud Monitoring for Dataflow pipelines

Troubleshoot Dataflow errors | Google Cloud

Troubleshoot stragglers in batch jobs | Cloud Dataflow

Autoscaling clusters | Dataproc Documentation | Google Cloud

5.5 Diagnostic Question 08

Cymbal Retail processes streaming data on Dataflow with Pub/Sub as a source. You need to plan for disaster recovery and protect against zonal failures.

What should you do?

A. Take Dataflow snapshots periodically.
B. Create Dataflow jobs from templates.
C. Enable vertical autoscaling.
D. Enable Dataflow shuffle.

5.5 Diagnostic Question 09

You run a Cloud SQL instance for a business that requires that the database is accessible for transactions. You need to ensure minimal downtime for database transactions.

What should you do?

A. Configure replication.
B. Configure high availability.
C. Configure backups.
D. Configure backups and increase the number of backups.

5.5 Diagnostic Question 10

You are running a Dataflow pipeline in production. The input data for this pipeline is occasionally inconsistent. Separately from processing the valid data, you want to efficiently capture the erroneous input data for analysis.

What should you do?

A. Re-read the input data and create separate outputs for valid and erroneous data.

B. Read the data once, and split it into two pipelines, one to output valid data and another to output erroneous data.

C. Check for the erroneous data in the logs.

D. Create a side output for the erroneous data.

5.5 Maintaining awareness of failures and mitigating impact

Courses

Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Lake

Serverless Data Processing with Dataflow: Develop Pipelines

- State and Timers
- Best Practices

Serverless Data Processing with Dataflow: Operations

- Troubleshooting and Debug
- Reliability

Documentation

Use Dataflow snapshots | Google Cloud

Serverless Data Processing with Dataflow: Operations | Google Cloud Skills Boost

About high availability | Cloud SQL for MySQL

Design Your Pipeline

Plan time to prepare

When will you take the exam?

How many weeks do you have to prepare?

How many hours will you spend preparing for the exam each week?

How many total hours will you prepare?

Weekly study plan

Now, consider what you've learned about your knowledge and skills through the diagnostic questions in this course. You should have a better understanding of what areas you need to focus on and what resources are available.

Use the template that follows to plan your study goals for each week. Consider:

- What exam guide section(s) or topic area(s) will you focus on?
- What courses (or specific modules) will help you learn more?
- What Skill Badges or labs will you work on for hands-on practice?
- What documentation links will you review?
- What additional resources will you use - such as sample questions?
- What will you do to prepare for the case studies?

You may do some or all of these study activities each week.

Duplicate the weekly template for the number of weeks in your individual preparation journey.

Weekly study template (example)

Area(s) of focus:	Using BigQuery as a data warehouse
Courses/modules to complete:	Modernizing Data Lakes and Data Warehouses with Google Cloud • Building a data warehouse
Skill Badges/labs to complete:	Build and Optimize Data Warehouses with BigQuery
Documentation to review:	Overview of BigQuery storage | Google Cloud Overview of BigQuery analytics | Google Cloud Introduction to BigQuery administration | Google Cloud Organizing BigQuery resources | Google Cloud
Additional study:	Sample Questions 1- 5

Weekly study template

Area(s) of focus:	
Courses/modules to complete:	
Skill Badges/labs to complete:	
Documentation to review:	
Additional study:	