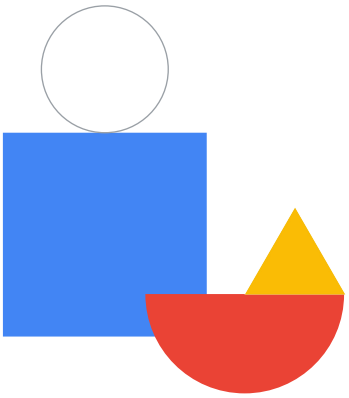
Google Cloud

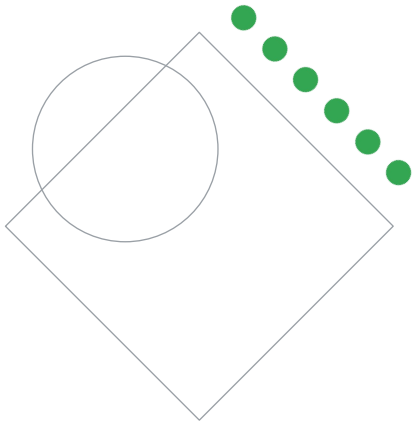


Preparing for Your Professional Data Engineer journey

Module 4: Preparing and Using Data for Analysis

Welcome to Module 4: Preparing and Using Data for Analysis.

Review and study planning



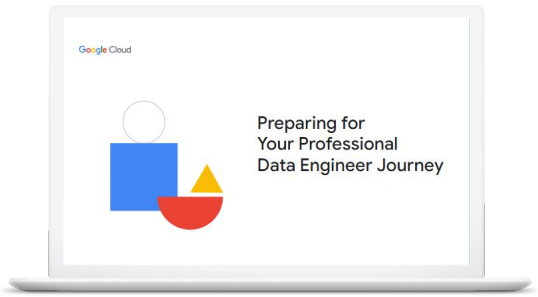
Google Cloud

Now let’s review how to use these diagnostic questions to help you identify what to include in your study plan.

As a reminder, this course isn’t designed to teach you everything you need to know for the exam—and the diagnostic questions don’t cover everything that could be on the exam. Instead, this activity is meant to give you a better sense of the scope of this section and the different skills you’ll want to develop as you prepare for the certification.

Your study plan:

Preparing and using data for analysis



4.1 Preparing data for visualization

4.2 Sharing data

4.3 Exploring and analyzing data

Google Cloud

You'll approach this review by looking at the objectives of this exam section and the questions you just answered about each one. Let's introduce an objective, briefly review the answers to the related questions, then explain where you can find out more in the learning resources and/or in Google documentation. As you go through each section objective, use the page in your workbook to mark the specific documentation, courses, and Skill Badges you'll want to emphasize in your study plan.

4.1 Preparing data for visualization

Considerations include:

- Connecting to tools
- Precalculating fields
- BigQuery materialized views (view logic)
- Determining granularity of time data
- Troubleshooting poor performing queries
- Identity and Access Management (IAM) and Cloud Data Loss Prevention (Cloud DLP)

Google Cloud

Effective data visualization is crucial to derive the most meaningful insights from the data. As a Professional Data Engineer, you will work on various tasks that prepare your data for visualization. Many times, the underlying queries are not efficient and as a result, your visualization will be slow and non-effective. You will have to deploy techniques such as pre-calculating the required fields before the query runs, creating materialized views from BigQuery tables, reducing the granularity of the time data, among others.

Question 1 tested your familiarity with data visualization and analysis tools that can connect to Google Cloud as a data source. Question 2 tested your ability to outline workflows to pre-calculate fields for data analysis. Question 3 asked about how to design and create views to support data analysis. Question 4 tested your ability to determine the granularity and organization of time-based data for analysis. Question 5 asked you about best practices for troubleshooting poor performance of queries.

4.1 Diagnostic Question 01 Discussion

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.

Google Cloud

Feedback:

A. Incorrect. Using Looker requires additional training for the team and incurs additional cost.
B. Correct. Connected Sheets is an easy way to integrate BigQuery data with Google Sheets. Because the leadership team is familiar with Workspace and Sheets, this satisfies their requirement.
C. Incorrect. Using Tableau requires additional training for the team and incurs additional cost.
D. Incorrect. Using Looker Studio requires additional training for the team.

Links:
https://cloud.google.com/bigquery/docs/data-analysis-tools-intro

More information:
Courses:
Google Cloud Big Data and Machine Learning Fundamentals

- Data Engineering for Streaming Data

Summary:
Google Cloud supports multiple visualization tools. You can also integrate third party tools. As a Professional Data Engineer, you have to evaluate the needs of the business and choose an appropriate tool depending on the level of expertise of the user and other factors like cost. Connected Sheets is an easy integration between

Google Sheets and BigQuery that provides convenient analysis and visualization with familiar tools.

4.1 Diagnostic Question 02 Discussion

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.

Google Cloud

Feedback:

A. Correct. Because the data is queried together, the schema design can group the data together with nested and repeated fields.
B. Incorrect. Normalized data is suitable for transactional databases, but is not efficient for analytics, especially for the kind of data in this requirement.
C. Incorrect. Federated queries are not efficient for analytics, because they have to join data across multiple data storage systems.
D. Incorrect. The inefficiency of multiple, separate, normalized tables still exists even if you partition the table.

Links:
https://cloud.google.com/bigquery/docs/best-practices-performance-nested

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud
- Building a Data Warehouse

Skill badges:
Prepare Data for ML APIs on Google Cloud
Engineer Data for Predictive Modeling with BigQuery ML

Summary:

A Professional Data Engineer needs to design the data schema efficiently. The normalized form is suitable for transactional databases, but unsuitable for analytical databases. Joins take time. Collecting related data together with nested and repeated fields can make the data more efficient to read.

4.1 Diagnostic Question 03 Discussion

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

Google Cloud

Feedback:

A. Incorrect. Views rerun the query each time on the source data; therefore, is not optimal.
B. Correct. Materialized views will improve query performance by precomputing and periodically caching query results.
C. Incorrect. Federated queries are not optimal, because they are used to query data that is stored outside BigQuery.
D. Incorrect. Partitions can improve the query performance by only scanning a portion of a table. However, scanning a portion of the table might not suit the business need in this scenario.

Links:
https://cloud.google.com/bigquery/docs/materialized-views-intro

More information:
Courses:
Modernizing Data Lakes and Data Warehouses on Google Cloud

- Building a Data Warehouse

Summary:
As data volumes increase, performance efficiency becomes an important consideration for the Professional Data Engineer. Knowledge of the various ways to abstract the data can help you determine the right approach for different business use

cases.

4.1 Diagnostic Question 04 Discussion

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.

Google Cloud

Feedback:

A. Incorrect. An aggregate function gives you one result for all the rows.
B. Incorrect. Using a UDF function entails extra work; therefore, it is not the most efficient solution.
C. Incorrect. BigQuery ML does not have any inbuilt features that provide this functionality.
D. Correct. A window function defines rows around each row. As a result, this approach can produce a separate output for each row.

Links:
https://cloud.google.com/bigquery/docs/reference/standard-sql/window-function-calls

More information:
Courses:
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

Summary:
BigQuery is a powerful tool for analytics that supports SQL. Knowledge of the

supported SQL clauses is critical for a Professional Data Engineer. Windowing functions save considerable effort in data analytics that would otherwise require complex queries.

4.1 Diagnostic Question 05 Discussion

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.

Google Cloud

Feedback:

A. Correct. Updating and inserting rows one at a time is not performant. Batch them instead.
B. Incorrect. A limit clause is applied at the end of the query, which does not make it performant.
C. Incorrect. Filtering data as early as possible reduces the amount of data processed across stages.
D. Incorrect. Self-joins are not performant. Use window analytical functions instead.

Links:

https://cloud.google.com/bigquery/docs/best-practices-performance-compute#avoid-anti-sql-patterns
https://cloud.google.com/bigquery/docs/best-practices-performance-compute#avoid_self_joins

More information:

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

Summary:
BigQuery performance tuning is a key function that the data engineer needs to perform. You should identify bottlenecks and apply various performance tuning techniques such as partitioning and clustering, batch updates, rewriting queries to filter data as early as possible, avoiding SQL anti-patterns, and other options.

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

Prepare Data for ML APIs on Google Cloud

Engineer Data for Predictive Modeling with BigQuery ML

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

You just reviewed several diagnostic questions that addressed different aspects of preparing data for visualization. These are some courses, Skill Badges, and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

https://cloud.google.com/bigquery/docs/data-analysis-tools-intro
https://cloud.google.com/bigquery/docs/best-practices-performance-nested
https://cloud.google.com/bigquery/docs/materialized-views-intro
https://cloud.google.com/bigquery/docs/reference/standard-sql/window-function-calls
https://cloud.google.com/bigquery/docs/best-practices-performance-compute#avoid-anti-sql-patterns
https://cloud.google.com/bigquery/docs/best-practices-performance-compute#avoid_self_joins

4.2 Sharing data

Considerations include:

- Defining rules to share data
- Publishing datasets
- Publishing reports and visualizations
- Analytics Hub

Google Cloud

As a Professional Data Engineer, you need to know how to share your own data with third parties and also how to access the data that is provided by third parties and use it inside your organization. Google Cloud Analytics Hub is an exchange platform that allows you to efficiently and securely exchange data, ML models and other analytics assets across organizations. Analytics Hub builds on the scalability and flexibility of BigQuery to streamline how you publish, discover, and subscribe to data exchanges and incorporate into your analysis. Data shared within Analytics Hub automatically includes in-depth governance, encryption, and security.

Question 6 tested your ability to outline rules to securely share data with stakeholders. Question 7 tested your familiarity with options for publishing datasets across organizational boundaries. Question 8 tested your familiarity with options for options for publishing reports and visualizations across organizational boundaries.

4.2 Diagnostic Question 06 Discussion

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.

Google Cloud

Feedback:

A. Incorrect. Creating a separate dataset with the required column is not a viable approach, because the original data might change.
B. Incorrect. Creating a separate table with the required column is not a reasonable approach, because the original data might change.
C. Correct. In BigQuery, policy tags give you fine-grained column level access control.
D. Incorrect. IAM permissions do not provide the column-level access control granularity that this scenario requires.

Links:
https://cloud.google.com/bigquery/docs/column-level-security-intro

More information:
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 badge:
Data Catalog Fundamentals

Summary:
Controlling access to data on a need-to-see basis is important for businesses. This requirement trickles down to the Professional Data Engineer who must know the various access control mechanisms. BigQuery, like other Google Cloud products, has some of its access controls set with IAM. For more granular control, such as column-level access, you can use policy tags.

4.2 Diagnostic Question 07 Discussion

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.

Google Cloud

Feedback:

A. Incorrect. Cloud Storage does not provide the greatest control over external data access for data sharing, and monetization will have to be done separately.
B. Correct. Analytics Hub has the built-in options to connect publishers and subscribers with access control and to monetize data access.
C. Incorrect. Sharing data through FTP endpoints has very limited access control and is not recommended.
D. Incorrect. Cloud SQL does not provide convenient access control or built-in monetization options.

Links:

https://cloud.google.com/analytics-hub
https://cloud.google.com/bigquery/docs/analytics-hub-introduction
https://www.youtube.com/watch?v=KOzo1tkScTs

More information:

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

Summary:
Analytics Hub is a convenient tool to share data with partners. Data engineers will have control over who can do what with the data. Analytics Hub can also turn a cost center into a profit center with the monetization of data.

4.2 Diagnostic Question 08 Discussion

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

Google Cloud

Feedback:

A. Incorrect. Looker Studio (previously Data Studio) is a visualization tool that does not have the data abstraction capabilities of Looker modeling.
B. Incorrect. Connected Sheets gives you access to data in BigQuery, but not other data sources.
C. Incorrect. To use the D3.js library, the analysts need to be adept at Javascript and complex details of creating visualizations with the library. This is not usually feasible for analysts.
D. Correct. Looker lets you model the underlying data sources and create an abstraction that analysts can easily build on.

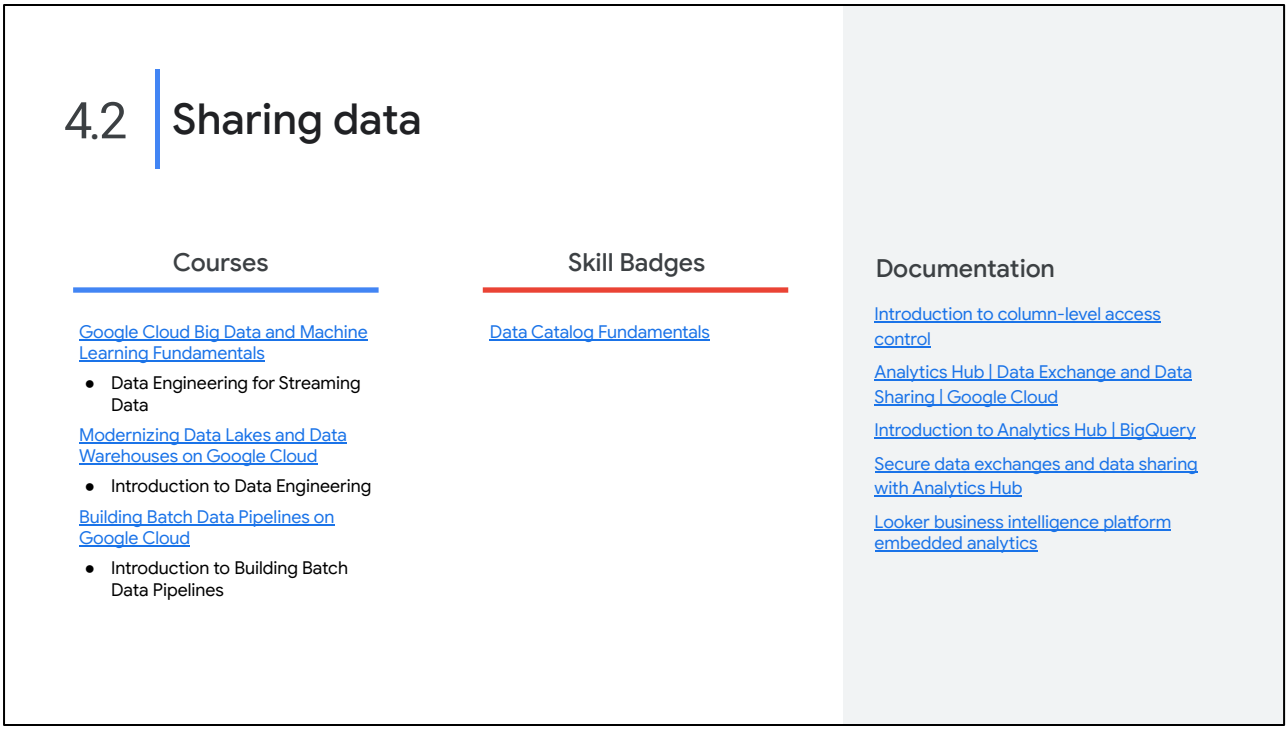
Links:
https://cloud.google.com/looker

More information:
Courses:
Google Cloud Big Data and Machine Learning Fundamentals

- Data Engineering for Streaming Data

Summary:
Google Cloud has a variety of different visualization options for the data engineer to choose from depending on the level of complexity users want to deal with. The Looker family of visualization products itself supports ease of use based on simple tables and

also complex, multi-source data that can be abstracted for end users.



The diagnostic questions you just reviewed explored some aspects of planning for sharing data. These are some courses and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

https://cloud.google.com/bigquery/docs/column-level-security-intro
https://cloud.google.com/analytics-hub
https://cloud.google.com/bigquery/docs/analytics-hub-introduction
https://www.youtube.com/watch?v=KOzo1tkScTs
https://cloud.google.com/looker

4.3 Exploring and analyzing data

Considerations include:

- Preparing data for feature engineering (training and serving machine learning models
- Conducting data discovery

Google Cloud

As a Professional Data Engineer, you will work closely with machine learning engineers and provide them with the necessary data to build ML models. This data needs to be processed so that ML engineers can perform feature engineering while training and serving machine learning models.

Question 9 asked you to describe the process of feature engineering and how it supports machine learning workflows. Question 10 tested your knowledge of how to support data access and discovery for machine learning workflows.

4.3 Diagnostic Question 09 Discussion

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.

Google Cloud

Feedback:
A: Incorrect. The type of data seems to be insufficient in representing the business requirement. Having more of the same data does not help.
B: Incorrect. It seems that overfitting is not the issue. L2 regularization does not improve the model's predictions.
C: Correct. Feature engineering can pick and choose relevant data columns and also enhance a model by combining columns. For this requirement, feature engineering improves the ML model.
D: Incorrect. Repeating training for longer with the same data might not improve the model.

Links:
https://cloud.google.com/bigquery/docs/bigqueryml-transform
https://cloud.google.com/bigquery/docs/preprocess-overview

More information:
Courses:
Google Cloud Big Data and Machine Learning Fundamentals
- Big Data with BigQuery
- The Machine Learning Workflow with Vertex AI

Smart Analytics, Machine Learning, and AI on Google Cloud
- Custom Model Building with SQL in BigQuery ML

Skill badge:
Engineer Data for Predictive Modeling with BigQuery ML

Summary:
Machine learning is vital for businesses. However, getting satisfactory results requires fine-tuning the model in different ways. A Professional Data Engineer can improve model performance with techniques such as feature engineering, where you choose the relevant columns and combine them to make the data relevant.

4.3 Diagnostic Question 10 Discussion

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.

Google Cloud

Feedback:
A: Correct. An exclude pattern will skip the file and therefore, this could be the reason the data is not discovered.
B: Incorrect. Even if the delivery runs only every hour, the files should be discovered then, which is not the case here. So, this cannot be the issue.
C: Incorrect. ORC format is supported and the file should be discovered.
D: Incorrect. Parquet format is supported and the file should be discovered.

Links:
https://cloud.google.com/dataplex/docs/discover-data

More information:
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

Summary:
Dataplex is able to automatically discover and catalog data in some file formats. However, you have control over which files can be discovered, and you can configure exclude patterns to intentionally ignore certain data.

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 pipelines

Smart Analytics, Machine Learning, and AI on Google Cloud

- Custom model building with SQL in BigQuery ML

Skill Badges

Engineer Data for Predictive Modeling with BigQuery ML

Documentation

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

Feature preprocessing overview | BigQuery | Google Cloud

Discover data | Dataplex | Google Cloud

You just reviewed diagnostic questions that addressed considerations related to using a data lake. These are some courses, Skill Badges, and links to learn more about the concepts in these questions. They provide a starting point to explore Google-recommended practices.

https://cloud.google.com/bigquery/docs/bigqueryml-transform
https://cloud.google.com/bigquery/docs/preprocess-overview
https://cloud.google.com/dataplex/docs/discover-data