CERT-PASS

Google GCP Professional Data Engineer

Free Practice Questions Preview

Here are 35 sample questions to help you get started. Unlock the full exam to access all 1100+ questions with detailed explanations.

Question 1 : Designing data processing systems

Case-based: A travel marketplace is designing a Google Cloud data platform for payment transactions. The team must separate development and production access while still allowing curated analytics. What should the data engineer recommend?

A. Create separate projects or datasets for environments, apply IAM at the narrowest practical scope, and expose curated BigQuery views with authorized views or policy tags.

B. Put all tables in one dataset and ask teams to use naming conventions for access control.

C. Grant BigQuery Admin to analysts and rely on audit logs to detect misuse.

D. Export curated data to CSV files in Cloud Storage and share bucket object ACLs with analysts.

Answer: A
Correct because separate environments, IAM, authorized views, and policy tags align governance with least privilege. The best wrong answer fails because naming conventions do not enforce access control.

Question 2 : Designing data processing systems

You need to choose the best architecture. A global energy utility must keep EU customer data in the EU and US customer data in the US. Analysts need aggregate reporting across both regions. Which design best satisfies the requirement?

A. Store all raw data in a multi-region US dataset because BigQuery is globally available.

B. Store regional datasets in matching BigQuery locations and create aggregate, non-sensitive derived datasets only where policy allows cross-region reporting.

C. Replicate all EU raw data to US to simplify analyst access.

D. Use one Cloud Storage bucket with dual-region placement and ignore dataset location.

Answer: B
Correct because data residency is enforced by dataset location and only approved aggregates should cross boundaries. The best wrong answer fails because global service availability does not override residency requirements.

Question 3 : Designing data processing systems

A data engineer is reviewing a proposed solution. A team is migrating an on-premises warehouse containing 20 TB per month to BigQuery. They need to validate row counts and critical aggregates during cutover. What is the best migration approach?

A. Use a one-time gsutil copy and immediately decommission the source system.

B. Rewrite all dashboards first and validate only after production traffic is moved.

C. Plan staged loads with BigQuery Data Transfer Service or Storage Transfer Service, run validation queries, reconcile counts and aggregates, then switch consumers.

D. Use Pub/Sub because it is optimized for historical bulk warehouse migration.

Answer: C
Correct because staged migration with validation reduces cutover risk. The best wrong answer fails because a one-time copy without reconciliation can hide missing or corrupted data.

Question 4 : Designing data processing systems

During production planning, a healthcare provider has pipelines that sometimes produce duplicate records after retries. The business requires accurate financial reporting. Which design principle should be emphasized?

A. Disable all retries so duplicates cannot happen.

B. Use only append-only tables and let dashboard users filter duplicates manually.

C. Increase Dataflow worker count so retries complete faster.

D. Design idempotent ingestion with deterministic keys, deduplication logic, and validation checks before publishing curated tables.

Answer: D
Correct because idempotency and deduplication preserve fidelity despite retries. The best wrong answer fails because disabling retries lowers reliability and does not address partial failures.

Question 5 : Designing data processing systems

A data platform must support future migration to multiple clouds and minimize lock-in for raw data. Query performance in BigQuery is still important. What should the architect choose?

A. Store raw data in open formats such as Parquet in Cloud Storage, govern it with Dataplex, and use BigLake or BigQuery external/native tables where appropriate.

B. Store all data only in proprietary application exports because they are easiest to ingest.

C. Use only Cloud SQL for all analytics so the schema is portable.

D. Avoid metadata catalogs because they make the design Google-specific.

Answer: A
Correct because open formats plus governance support portability while enabling Google Cloud analytics. The best wrong answer fails because Cloud SQL is not appropriate for large-scale analytical workloads.

Question 6 : Designing data processing systems

A healthcare provider wants automated data quality checks before data reaches trusted BigQuery tables. Which approach best fits Google Cloud best practices?

A. Let Looker dashboards show NULL values so users can decide if data is valid.

B. Add validation rules in Dataform or Dataflow, quarantine failed records, and publish only validated outputs to curated datasets.

C. Use Cloud Monitoring only; metrics replace row-level validation.

D. Grant analysts permission to edit production tables when they find bad records.

Answer: B
Correct because automated validation and quarantine prevent bad data from reaching trusted tables. The best wrong answer fails because monitoring does not validate individual records or enforce quality gates.

Question 7 : Designing data processing systems

A regulated gaming company wants encryption keys controlled by its security team for BigQuery and Cloud Storage datasets. What should be used?

A. Default Google-owned encryption keys only, because customer control is not possible.

B. Hard-code encryption keys in Dataflow pipeline options.

C. Customer-managed encryption keys in Cloud KMS with IAM separation and key rotation policies.

D. Store keys in a BigQuery table and join them during processing.

Answer: C
Correct because CMEK with Cloud KMS provides customer-controlled key management. The best wrong answer fails because default encryption protects data but does not give the customer direct key control.

Question 8 : Designing data processing systems

Case-based: A platform team must design project structure for many product teams. Each team owns data products but central governance must enforce policy. Which design is best?

A. Put every workload in one project to make billing simple.

B. Create one service account shared by every team to avoid permission complexity.

C. Let each team create unmanaged buckets and datasets without central standards.

D. Use separate projects or folders for teams, shared governance policies at folder or organization level, and Dataplex zones for discoverability and policy consistency.

Answer: D
Correct because project/folder boundaries plus central policy support federated governance. The best wrong answer fails because one shared project makes isolation, quota management, and least privilege difficult.

Question 9 : Designing data processing systems

You need to choose the best architecture. A company wants to use generative AI to help analysts translate natural language to SQL, but the data contains sensitive fields. What is the safest design?

A. Restrict model access to governed semantic layers or authorized views and mask sensitive fields using policy tags or data masking before query generation.

B. Give the LLM service account BigQuery Admin so it can discover all schemas.

C. Send raw tables including PII to a public prompt for better SQL quality.

D. Disable IAM because the generated SQL will be reviewed by analysts.

Answer: A
Correct because AI-assisted query generation must respect governed views and masking. The best wrong answer fails because broad admin access exposes sensitive data unnecessarily.

Question 10 : Designing data processing systems

A data engineer is reviewing a proposed solution. A migration plan must move an Oracle database to Google Cloud with minimal downtime and continuous replication during testing. Which service should be considered?

A. Transfer Appliance because it continuously replicates database changes.

B. Datastream for change data capture into Google Cloud, combined with validation and a controlled cutover plan.

C. Cloud Composer because it is a database replication engine.

D. BigQuery BI Engine because it migrates operational databases.

Answer: B
Correct because Datastream supports CDC-style replication for migration patterns. The best wrong answer fails because Transfer Appliance is for offline bulk transfer, not continuous database change replication.

Question 11 : Designing data processing systems

During production planning, a energy utility is designing a governed lakehouse. Which option lists the best combination of design actions?

A. Use one public bucket, skip cataloging, and give all users Storage Admin.

B. Store only dashboard extracts and delete raw data after every load.

C. Define domain ownership, use Dataplex for catalog and governance, separate raw and curated zones, and apply IAM and policy tags.

D. Use a single VM with local disks as the central platform.

Answer: C
Correct because governance, zones, and fine-grained access align with a governed lakehouse. The best wrong answer fails because public buckets and broad admin roles are insecure.

Question 12 : Designing data processing systems

A bank is designing a Google Cloud data platform for shipment status events. The team must prevent analysts from seeing raw PII while still allowing curated analytics. What should the data engineer recommend?

A. Put all tables in one dataset and ask teams to use naming conventions for access control.

B. Grant BigQuery Admin to analysts and rely on audit logs to detect misuse.

C. Export curated data to CSV files in Cloud Storage and share bucket object ACLs with analysts.

D. Create separate projects or datasets for environments, apply IAM at the narrowest practical scope, and expose curated BigQuery views with authorized views or policy tags.

Answer: D
Correct because separate environments, IAM, authorized views, and policy tags align governance with least privilege. The best wrong answer fails because naming conventions do not enforce access control.

Question 13 : Designing data processing systems

A global public agency must keep EU customer data in the EU and US customer data in the US. Analysts need aggregate reporting across both regions. Which design best satisfies the requirement?

A. Store regional datasets in matching BigQuery locations and create aggregate, non-sensitive derived datasets only where policy allows cross-region reporting.

B. Store all raw data in a multi-region US dataset because BigQuery is globally available.

C. Replicate all EU raw data to US to simplify analyst access.

D. Use one Cloud Storage bucket with dual-region placement and ignore dataset location.

Answer: A
Correct because data residency is enforced by dataset location and only approved aggregates should cross boundaries. The best wrong answer fails because global service availability does not override residency requirements.

Question 14 : Designing data processing systems

A team is migrating an on-premises warehouse containing 200 GB per hour to BigQuery. They need to validate row counts and critical aggregates during cutover. What is the best migration approach?

A. Use a one-time gsutil copy and immediately decommission the source system.

B. Plan staged loads with BigQuery Data Transfer Service or Storage Transfer Service, run validation queries, reconcile counts and aggregates, then switch consumers.

C. Rewrite all dashboards first and validate only after production traffic is moved.

D. Use Pub/Sub because it is optimized for historical bulk warehouse migration.

Answer: B
Correct because staged migration with validation reduces cutover risk. The best wrong answer fails because a one-time copy without reconciliation can hide missing or corrupted data.

Question 15 : Designing data processing systems

Case-based: A insurance company has pipelines that sometimes produce duplicate records after retries. The business requires accurate financial reporting. Which design principle should be emphasized?

A. Disable all retries so duplicates cannot happen.

B. Use only append-only tables and let dashboard users filter duplicates manually.

C. Design idempotent ingestion with deterministic keys, deduplication logic, and validation checks before publishing curated tables.

D. Increase Dataflow worker count so retries complete faster.

> **Answer: C**
> Correct because idempotency and deduplication preserve fidelity despite retries. The best wrong answer fails because disabling retries lowers reliability and does not address partial failures.

Question 16 : Designing data processing systems

You need to choose the best architecture. A data platform must support future migration to multiple clouds and minimize lock-in for raw data. Query performance in BigQuery is still important. What should the architect choose?

A. Store all data only in proprietary application exports because they are easiest to ingest.

B. Use only Cloud SQL for all analytics so the schema is portable.

C. Avoid metadata catalogs because they make the design Google-specific.

D. Store raw data in open formats such as Parquet in Cloud Storage, govern it with Dataplex, and use BigLake or BigQuery external/native tables where appropriate.

> **Answer: D**
> Correct because open formats plus governance support portability while enabling Google Cloud analytics. The best wrong answer fails because Cloud SQL is not appropriate for large-scale analytical workloads.

Question 17 : Designing data processing systems

A data engineer is reviewing a proposed solution. A manufacturer wants automated data quality checks before data reaches trusted BigQuery tables. Which approach best fits Google Cloud best practices?

A. Add validation rules in Dataform or Dataflow, quarantine failed records, and publish only validated outputs to curated datasets.

B. Let Looker dashboards show NULL values so users can decide if data is valid.

C. Use Cloud Monitoring only; metrics replace row-level validation.

D. Grant analysts permission to edit production tables when they find bad records.

> **Answer: A**
> Correct because automated validation and quarantine prevent bad data from reaching trusted tables. The best wrong answer fails because monitoring does not validate individual records or enforce quality gates.

Question 18 : Designing data processing systems

During production planning, a regulated retailer wants encryption keys controlled by its security team for BigQuery and Cloud Storage datasets. What should be used?

A. Default Google-owned encryption keys only, because customer control is not possible.

B. Customer-managed encryption keys in Cloud KMS with IAM separation and key rotation policies.

C. Hard-code encryption keys in Dataflow pipeline options.

D. Store keys in a BigQuery table and join them during processing.

Answer: B
Correct because CMEK with Cloud KMS provides customer-controlled key management. The best wrong answer fails because default encryption protects data but does not give the customer direct key control.

Question 19 : Designing data processing systems

A platform team must design project structure for many product teams. Each team owns data products but central governance must enforce policy. Which design is best?

A. Put every workload in one project to make billing simple.

B. Create one service account shared by every team to avoid permission complexity.

C. Use separate projects or folders for teams, shared governance policies at folder or organization level, and Dataplex zones for discoverability and policy consistency.

D. Let each team create unmanaged buckets and datasets without central standards.

Answer: C
Correct because project/folder boundaries plus central policy support federated governance. The best wrong answer fails because one shared project makes isolation, quota management, and least privilege difficult.

Question 20 : Designing data processing systems

A company wants to use generative AI to help analysts translate natural language to SQL, but the data contains sensitive fields. What is the safest design?

A. Give the LLM service account BigQuery Admin so it can discover all schemas.

B. Send raw tables including PII to a public prompt for better SQL quality.

C. Disable IAM because the generated SQL will be reviewed by analysts.

D. Restrict model access to governed semantic layers or authorized views and mask sensitive fields using policy tags or data masking before query generation.

Answer: D
Correct because AI-assisted query generation must respect governed views and masking. The best wrong answer fails because broad admin access exposes sensitive data unnecessarily.

Question 21 : Designing data processing systems

A migration plan must move an Oracle database to Google Cloud with minimal downtime and continuous replication during testing. Which service should be considered?

A. Datastream for change data capture into Google Cloud, combined with validation and a controlled cutover plan.

B. Transfer Appliance because it continuously replicates database changes.

C. Cloud Composer because it is a database replication engine.

D. BigQuery BI Engine because it migrates operational databases.

Answer: A
Correct because Datastream supports CDC-style replication for migration patterns. The best wrong answer fails because Transfer Appliance is for offline bulk transfer, not continuous database change replication.

Question 22 : Designing data processing systems

Case-based: A retailer is designing a governed lakehouse. Which option lists the best combination of design actions?

A. Use one public bucket, skip cataloging, and give all users Storage Admin.

B. Define domain ownership, use Dataplex for catalog and governance, separate raw and curated zones, and apply IAM and policy tags.

C. Store only dashboard extracts and delete raw data after every load.

D. Use a single VM with local disks as the central platform.

Answer: B
Correct because governance, zones, and fine-grained access align with a governed lakehouse. The best wrong answer fails because public buckets and broad admin roles are insecure.

Question 23 : Designing data processing systems

You need to choose the best architecture. A bank is designing a Google Cloud data platform for customer profiles. The team must separate development and production access while still allowing curated analytics. What should the data engineer recommend?

A. Put all tables in one dataset and ask teams to use naming conventions for access control.

B. Grant BigQuery Admin to analysts and rely on audit logs to detect misuse.

C. Create separate projects or datasets for environments, apply IAM at the narrowest practical scope, and expose curated BigQuery views with authorized views or policy tags.

D. Export curated data to CSV files in Cloud Storage and share bucket object ACLs with analysts.

Answer: C
Correct because separate environments, IAM, authorized views, and policy tags align governance with least privilege. The best wrong answer fails because naming conventions do not enforce access control.

Question 24 : Designing data processing systems

A data engineer is reviewing a proposed solution. A global telecom provider must keep EU customer data in the EU and US customer data in the US. Analysts need aggregate reporting across both regions. Which design best satisfies the requirement?

A. Store all raw data in a multi-region US dataset because BigQuery is globally available.

B. Replicate all EU raw data to US to simplify analyst access.

C. Use one Cloud Storage bucket with dual-region placement and ignore dataset location.

D. Store regional datasets in matching BigQuery locations and create aggregate, non-sensitive derived datasets only where policy allows cross-region reporting.

Answer: D
Correct because data residency is enforced by dataset location and only approved aggregates should cross boundaries. The best wrong answer fails because global service availability does not override residency requirements.

Question 25 : Designing data processing systems

During production planning, a team is migrating an on-premises warehouse containing 5 TB per day to BigQuery. They need to validate row counts and critical aggregates during cutover. What is the best migration approach?

A. Plan staged loads with BigQuery Data Transfer Service or Storage Transfer Service, run validation queries, reconcile counts and aggregates, then switch consumers.

B. Use a one-time gsutil copy and immediately decommission the source system.

C. Rewrite all dashboards first and validate only after production traffic is moved.

D. Use Pub/Sub because it is optimized for historical bulk warehouse migration.

Answer: A
Correct because staged migration with validation reduces cutover risk. The best wrong answer fails because a one-time copy without reconciliation can hide missing or corrupted data.

Question 26 : Designing data processing systems

A telecom provider has pipelines that sometimes produce duplicate records after retries. The business requires accurate financial reporting. Which design principle should be emphasized?

A. Disable all retries so duplicates cannot happen.

B. Design idempotent ingestion with deterministic keys, deduplication logic, and validation checks before publishing curated tables.

C. Use only append-only tables and let dashboard users filter duplicates manually.

D. Increase Dataflow worker count so retries complete faster.

Answer: B
Correct because idempotency and deduplication preserve fidelity despite retries. The best wrong answer fails because disabling retries lowers reliability and does not address partial failures.

Question 27 : Designing data processing systems

A data platform must support future migration to multiple clouds and minimize lock-in for raw data. Query performance in BigQuery is still important. What should the architect choose?

A. Store all data only in proprietary application exports because they are easiest to ingest.

B. Use only Cloud SQL for all analytics so the schema is portable.

C. Store raw data in open formats such as Parquet in Cloud Storage, govern it with Dataplex, and use BigLake or BigQuery external/native tables where appropriate.

D. Avoid metadata catalogs because they make the design Google-specific.

Answer: C
Correct because open formats plus governance support portability while enabling Google Cloud analytics. The best wrong answer fails because Cloud SQL is not appropriate for large-scale analytical workloads.

Question 28 : Designing data processing systems

A healthcare provider wants automated data quality checks before data reaches trusted BigQuery tables. Which approach best fits Google Cloud best practices?

A. Let Looker dashboards show NULL values so users can decide if data is valid.

B. Use Cloud Monitoring only; metrics replace row-level validation.

C. Grant analysts permission to edit production tables when they find bad records.

D. Add validation rules in Dataform or Dataflow, quarantine failed records, and publish only validated outputs to curated datasets.

Answer: D
Correct because automated validation and quarantine prevent bad data from reaching trusted tables. The best wrong answer fails because monitoring does not validate individual records or enforce quality gates.

Question 29 : Designing data processing systems

Case-based: A regulated energy utility wants encryption keys controlled by its security team for BigQuery and Cloud Storage datasets. What should be used?

A. Customer-managed encryption keys in Cloud KMS with IAM separation and key rotation policies.

B. Default Google-owned encryption keys only, because customer control is not possible.

C. Hard-code encryption keys in Dataflow pipeline options.

D. Store keys in a BigQuery table and join them during processing.

Answer: A
Correct because CMEK with Cloud KMS provides customer-controlled key management. The best wrong answer fails because default encryption protects data but does not give the customer direct key control.

Question 30 : Designing data processing systems

You need to choose the best architecture. A platform team must design project structure for many product teams. Each team owns data products but central governance must enforce policy. Which design is best?

A. Put every workload in one project to make billing simple.

B. Use separate projects or folders for teams, shared governance policies at folder or organization level, and Dataplex zones for discoverability and policy consistency.

C. Create one service account shared by every team to avoid permission complexity.

D. Let each team create unmanaged buckets and datasets without central standards.

Answer: B
Correct because project/folder boundaries plus central policy support federated governance. The best wrong answer fails because one shared project makes isolation, quota management, and least privilege difficult.

Question 31 : Designing data processing systems

A data engineer is reviewing a proposed solution. A company wants to use generative AI to help analysts translate natural language to SQL, but the data contains sensitive fields. What is the safest design?

A. Give the LLM service account BigQuery Admin so it can discover all schemas.

B. Send raw tables including PII to a public prompt for better SQL quality.

C. Restrict model access to governed semantic layers or authorized views and mask sensitive fields using policy tags or data masking before query generation.

D. Disable IAM because the generated SQL will be reviewed by analysts.

Answer: C
Correct because AI-assisted query generation must respect governed views and masking. The best wrong answer fails because broad admin access exposes sensitive data unnecessarily.

Question 32 : Designing data processing systems

During production planning, a migration plan must move an Oracle database to Google Cloud with minimal downtime and continuous replication during testing. Which service should be considered?

A. Transfer Appliance because it continuously replicates database changes.

B. Cloud Composer because it is a database replication engine.

C. BigQuery BI Engine because it migrates operational databases.

D. Datastream for change data capture into Google Cloud, combined with validation and a controlled cutover plan.

Answer: D
Correct because Datastream supports CDC-style replication for migration patterns. The best wrong answer fails because Transfer Appliance is for offline bulk transfer, not continuous database change replication.

Question 33 : Designing data processing systems

A travel marketplace is designing a governed lakehouse. Which option lists the best combination of design actions?

A. Define domain ownership, use Dataplex for catalog and governance, separate raw and curated zones, and apply IAM and policy tags.

B. Use one public bucket, skip cataloging, and give all users Storage Admin.

C. Store only dashboard extracts and delete raw data after every load.

D. Use a single VM with local disks as the central platform.

Answer: A
Correct because governance, zones, and fine-grained access align with a governed lakehouse. The best wrong answer fails because public buckets and broad admin roles are insecure.

Question 34 : Designing data processing systems

A energy utility is designing a Google Cloud data platform for application logs. The team must enforce least privilege while still allowing curated analytics. What should the data engineer recommend?

A. Put all tables in one dataset and ask teams to use naming conventions for access control.

B. Create separate projects or datasets for environments, apply IAM at the narrowest practical scope, and expose curated BigQuery views with authorized views or policy tags.

C. Grant BigQuery Admin to analysts and rely on audit logs to detect misuse.

D. Export curated data to CSV files in Cloud Storage and share bucket object ACLs with analysts.

Answer: B
Correct because separate environments, IAM, authorized views, and policy tags align governance with least privilege. The best wrong answer fails because naming conventions do not enforce access control.

Question 35 : Designing data processing systems

A global healthcare provider must keep EU customer data in the EU and US customer data in the US. Analysts need aggregate reporting across both regions. Which design best satisfies the requirement?

A. Store all raw data in a multi-region US dataset because BigQuery is globally available.

B. Replicate all EU raw data to US to simplify analyst access.

C. Store regional datasets in matching BigQuery locations and create aggregate, non-sensitive derived datasets only where policy allows cross-region reporting.

D. Use one Cloud Storage bucket with dual-region placement and ignore dataset location.

Answer: C
Correct because data residency is enforced by dataset location and only approved aggregates should cross boundaries. The best wrong answer fails because global service availability does not override residency requirements.