Google GCP-PDE

Google Professional Data Engineer Certification Questions & Answers

Get Instant Access to Vital Exam Acing Materials | Study Guide | Sample Questions | Practice Test

GCP-PDE

Google Cloud Platform - Professional Data Engineer (GCP-PDE)

50 Questions Exam – 70% Cut Score – Duration of 120 minutes

Table of Contents:

Discover More about the GCP-PDE Certification

Are you interested in passing the Google GCP-PDE exam? First discover, who benefits from the GCP-PDE certification. The GCP-PDE is suitable for a candidate if he wants to learn about Cloud. Passing the GCP-PDE exam earns you the Google Cloud Platform - Professional Data Engineer (GCP-PDE) title.

While preparing for the GCP-PDE exam, many candidates struggle to get the necessary materials. But do not worry; your struggling days are over. The GCP-PDE PDF contains some of the most valuable preparation tips and the details and instant access to useful **GCP-PDE study materials just at one click.**

Google GCP-PDE Professional Data Engineer Certification Details:

Exam Name	Google Professional Data Engineer
Exam Code	GCP-PDE
Exam Price	$200 USD
Duration	120 minutes
Number of Questions	50
Passing Score	Pass / Fail (Approx 70%)
Recommended Training / Books	**Google Cloud training** **Google Cloud documentation** **Google Cloud solutions**
Schedule Exam	**PEARSON VUE**
Sample Questions	**Google GCP-PDE Sample Questions**
Recommended Practice	**Google Cloud Platform - Professional Data Engineer (GCP-PDE) Practice Test**

GCP-PDE Syllabus:

Section	Objectives
Designing data processing systems	
Selecting the appropriate storage technologies. Considerations include:	- Mapping storage systems to business requirements - Data modeling - Trade-offs involving latency, throughput, transactions - Distributed systems - Schema design

Section	Objectives
Designing data pipelines. Considerations include:	- Data publishing and visualization (e.g., BigQuery) - Batch and streaming data (e.g., Dataflow, Dataproc, Apache Beam, Apache Spark and Hadoop ecosystem, Pub/Sub, Apache Kafka) - Online (interactive) vs. batch predictions - Job automation and orchestration (e.g., Cloud Composer)
Designing a data processing solution. Considerations include:	- Choice of infrastructure - System availability and fault tolerance - Use of distributed systems - Capacity planning - Hybrid cloud and edge computing - Architecture options (e.g., message brokers, message queues, middleware, service-oriented architecture, serverless functions) - At least once, in-order, and exactly once, etc., event processing
Migrating data warehousing and data processing. Considerations include:	- Awareness of current state and how to migrate a design to a future state - Migrating from on-premises to cloud (Data Transfer Service, Transfer Appliance, Cloud Networking) - Validating a migration
Building and operationalizing data processing systems	
Building and operationalizing storage systems. Considerations include:	- Effective use of managed services (Cloud Bigtable, Cloud Spanner, Cloud SQL, BigQuery, Cloud Storage, Datastore, Memorystore) - Storage costs and performance - Life cycle management of data
Building and operationalizing pipelines. Considerations include:	- Data cleansing - Batch and streaming - Transformation - Data acquisition and import - Integrating with new data sources
Building and operationalizing processing infrastructure. Considerations include:	- Provisioning resources - Monitoring pipelines - Adjusting pipelines - Testing and quality control

Section	Objectives
Operationalizing machine learning models	
Leveraging pre-built ML models as a service. Considerations include:	- ML APIs (e.g., Vision API, Speech API) - Customizing ML APIs (e.g., AutoML Vision, Auto ML text) - Conversational experiences (e.g., Dialogflow)
Deploying an ML pipeline. Considerations include:	- Ingesting appropriate data - Retraining of machine learning models (AI Platform Prediction and Training, BigQuery ML, Kubeflow, Spark ML) - Continuous evaluation
Choosing the appropriate training and serving infrastructure. Considerations include:	- Distributed vs. single machine - Use of edge compute - Hardware accelerators (e.g., GPU, TPU)
Measuring, monitoring, and troubleshooting machine learning models. Considerations include:	- Machine learning terminology (e.g., features, labels, models, regression, classification, recommendation, supervised and unsupervised learning, evaluation metrics) - Impact of dependencies of machine learning models - Common sources of error (e.g., assumptions about data)
Ensuring solution quality	
Designing for security and compliance. Considerations include:	- Identity and access management (e.g., Cloud IAM) - Data security (encryption, key management) - Ensuring privacy (e.g., Data Loss Prevention API) - Legal compliance (e.g., Health Insurance Portability and Accountability Act (HIPAA), Children's Online Privacy Protection Act (COPPA), FedRAMP, General Data Protection Regulation (GDPR))
Ensuring scalability and efficiency. Considerations include:	- Building and running test suites - Pipeline monitoring (e.g., Cloud Monitoring) - Assessing, troubleshooting, and improving data representations and data processing infrastructure - Resizing and autoscaling resources
Ensuring reliability and fidelity. Considerations include:	- Performing data preparation and quality control (e.g., Dataprep) - Verification and monitoring - Planning, executing, and stress testing data recovery (fault tolerance, rerunning failed jobs, performing retrospective re-

Section	Objectives
	analysis) - Choosing between ACID, idempotent, eventually consistent requirements
Ensuring flexibility and portability. Considerations include:	- Mapping to current and future business requirements - Designing for data and application portability (e.g., multicloud, data residency requirements) - Data staging, cataloging, and discovery

Broaden Your Knowledge with Google GCP-PDE Sample Questions:

Question: 1

You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. You intend to use ANSI SQL to run queries for your analysts.

How should you transform the input data?

a) Use BigQuery for storage. Use Dataflow to run the transformations.
b) Use BigQuery for storage. Use Dataproc to run the transformations.
c) Use Cloud Storage for storage. Use Dataflow to run the transformations.
d) Use Cloud Storage for storage. Use Dataproc to run the transformations.

Answer: b

Question: 2

You are building storage for files for a data pipeline on Google Cloud. You want to support JSON files. The schema of these files will occasionally change.

Your analyst teams will use running aggregate ANSI SQL queries on this data. What should you do?

a) Use BigQuery for storage. Provide format files for data load. Update the format files as needed.
b) Use BigQuery for storage. Select "Automatically detect" in the Schema section.
c) Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.
d) Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

Answer: b

Question: 3

You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world.

You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?

a) Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.
b) Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.
c) Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.
d) Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

Answer: a

Question: 4

You are working on a project with two compliance requirements. The first requirement states that your developers should be able to see the Google Cloud billing charges for only their own projects.

a) The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization.
b) The finance team should not be able to view the project contents. You want to set permissions. What should you do?
a) Add the finance team members to the Billing Administrator role for each of the billing accounts that they need to manage. Add the developers to the Viewer role for the Project.
b) Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their own spend only.
c) Add the developers and finance managers to the Viewer role for the Project.
d) Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.

Answer: a

Question: 5

Your company is loading comma-separated values (CSV) files into BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file.

a) What is the most likely cause of this problem?
a) The CSV data loaded in BigQuery is not flagged as CSV.
b) The CSV data had invalid rows that were skipped on import.
c) The CSV data has not gone through an ETL phase before loading into BigQuery.
d) The CSV data loaded in BigQuery is not using BigQuery's default encoding.

Answer: d

Question: 6

You want to publish system metrics to Google Cloud from a large number of on-prem hypervisors and VMs for analysis and creation of dashboards.

You have an existing custom monitoring agent deployed to all the hypervisors and your on-prem metrics system is unable to handle the load. You want to design a system that can collect and store metrics at scale. You don't want to manage your own time series database.

Metrics from all agents should be written to the same table but agents must not have permission to modify or read data written by other agents. What should you do?

a) Modify the monitoring agent to write protobuf messages directly to BigTable.
b) Modify the monitoring agent to publish protobuf messages to Pub/Sub. Use a Dataproc cluster or Dataflow job to consume messages from Pub/Sub and write to BigTable.
c) Modify the monitoring agent to write protobuf messages to HBase deployed on Compute Engine VM Instances
d) Modify the monitoring agent to write protobuf messages to Pub/Sub. Use a Dataproc cluster or Dataflow job to consume messages from Pub/Sub and write to Cassandra deployed on Compute Engine VM Instances.

Answer: b

Question: 7

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance.

How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

a) Use a row key of the form <timestamp>.
b) Use a row key of the form <sensorid>.
c) Use a row key of the form <timestamp>#<sensorid>.
d) Use a row key of the form <sensorid>#<timestamp>.

Answer: d

Question: 8

You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?

a) Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and displays device outlier data based on your business requirements.
b) Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.
c) Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements.
d) Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements.

Answer: c

Question: 9

You are using Pub/Sub to stream inventory updates from many point-of-sale (POS) terminals into BigQuery.

Each update event has the following information: product identifier "prodSku", change increment "quantityDelta", POS identification "termId", and "messageId" which is created for each push attempt from the terminal.

During a network outage, you discovered that duplicated messages were sent, causing the inventory system to over-count the changes. You determine that the terminal application has design problems and may send the same event more than once during push retries.

You want to ensure that the inventory update is accurate. What should you do?

a) Add another attribute orderId to the message payload to mark the unique check-out order across all terminals. Make sure that messages whose "orderId" and "prodSku" values match corresponding rows in the BigQuery table are discarded.
b) Inspect the "messageId" of each message. Make sure that any messages whose "messageId" values match corresponding rows in the BigQuery table are discarded.
c) Instead of specifying a change increment for "quantityDelta", always use the derived inventory value after the increment has been applied. Name the new attribute "adjustedQuantity".
d) Inspect the "publishTime" of each message. Make sure that messages whose "publishTime" values match rows in the BigQuery table are discarded.

Answer: a

Question: 10

You need to stream time-series data in Avro format, and then write this to both BigQuery and Cloud Bigtable simultaneously using Dataflow. You want to achieve minimal end-to-end latency.

Your business requirements state this needs to be completed as quickly as possible. What should you do?

a) Create a pipeline and use ParDo transform.
b) Create a pipeline that groups the data into a PCollection and uses the Combine transform.
c) Create a pipeline that groups data using a PCollection, and then use Avro I/O transform to write to Cloud Storage. After the data is written, load the data from Cloud Storage into BigQuery and Bigtable.
d) Create a pipeline that groups data using a PCollection and then uses Bigtable and BigQueryIO transforms.

Answer: d

Avail the Study Guide to Pass Google GCP-PDE Professional Data Engineer Exam:

- Find out about the GCP-PDE syllabus topics. Visiting the official site offers an idea about the exam structure and other important study resources. Going through the syllabus topics help to plan the exam in an organized manner.
- Once you are done exploring the **GCP-PDE syllabus**, it is time to plan for studying and covering the syllabus topics from the core. Chalk out the best plan for yourself to cover each part of the syllabus in a hassle-free manner.
- A study schedule helps you to stay calm throughout your exam preparation. It should contain your materials and thoughts like study hours, number of topics for daily studying mentioned on it. The best bet to clear the exam is to follow your schedule rigorously.
- The candidate should not miss out on the scope to learn from the GCP-PDE training. Joining the Google provided training for GCP-PDE exam helps a candidate to strengthen his practical knowledge base from the certification.
- Learning about the probable questions and gaining knowledge regarding the exam structure helps a lot. Go through the **GCP-PDE sample questions** and boost your knowledge

- Make yourself a pro through online practicing the syllabus topics. GCP-PDE practice tests would guide you on your strengths and weaknesses regarding the syllabus topics. Through rigorous practicing, you can improve the weaker sections too. Learn well about time management during exam and become confident gradually with practice tests.

Career Benefits:

- Passing the GCP-PDE exam, helps a candidate to prosper highly in his career. Having the certification on the resume adds to the candidate's benefit and helps to get the best opportunities.