Google GCP-PDE

GOOGLE PROFESSIONAL DATA ENGINEER CERTIFICATION QUESTIONS & ANSWERS

Exam Summary – Syllabus – Questions

GCP-PDE

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

50-60 Questions Exam – 70% Cut Score – Duration of 120 minutes

www.VMExam.com

Table of Contents

Know Your GCP-PDE Certification Well:

The GCP-PDE is best suitable for candidates who want to gain knowledge in the Google Professional. Before you start your GCP-PDE preparation you may struggle to get all the crucial Professional Data Engineer materials like GCP-PDE syllabus, sample questions, study guide.

But don't worry the GCP-PDE PDF is here to help you prepare in a stress free manner.

The PDF is a combination of all your queries like-

- What is in the GCP-PDE syllabus?
- How many questions are there in the GCP-PDE exam?
- Which Practice test would help me to pass the GCP-PDE exam at the first attempt?

Passing the GCP-PDE exam makes you Google Cloud Platform - Professional Data Engineer (GCP-PDE). Having the Professional Data Engineer certification opens multiple opportunities for you. You can grab a new job, get a higher salary or simply get recognition within your current organization.

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	50-60
Passing Score	Pass / Fail (Approx 70%)
Recommended Training / Books	**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 (22% of the exam)	
Designing for security and compliance. Considerations include:	- Identity and Access Management (e.g., Cloud IAM and organization policies) - Data security (encryption and key management) - Privacy (e.g., personally identifiable information, and Cloud Data Loss Prevention API) - Regional considerations (data sovereignty) for data access and storage - Legal and regulatory compliance
Designing for reliability and fidelity. Considerations include:	- Preparing and cleaning data (e.g., Dataprep, Dataflow, and Cloud Data Fusion) - Monitoring and orchestration of data pipelines - Disaster recovery and fault tolerance - Making decisions related to ACID (atomicity, consistency, isolation, and durability) compliance and availability - Data validation
Designing for flexibility and portability. Considerations include:	- Mapping current and future business requirements to the architecture - Designing for data and application portability (e.g., multi-cloud and data residency requirements) - Data staging, cataloging, and discovery (data governance)
Designing data migrations. Considerations include:	- Analyzing current stakeholder needs, users, processes, and technologies and creating a plan to get to desired state - Planning migration to Google Cloud (e.g., BigQuery Data Transfer Service, Database Migration Service, Transfer Appliance, Google Cloud networking, Datastream) - Designing the migration validation strategy - Designing the project, dataset, and table architecture to ensure proper data governance
Ingesting and processing the data (25% of the exam)	
Planning the data pipelines. Considerations include:	- Defining data sources and sinks - Defining data transformation logic - Networking fundamentals - Data encryption
Building the pipelines. Considerations include:	- Data cleansing - Identifying the services (e.g., Dataflow, Apache Beam, Dataproc, Cloud Data Fusion, BigQuery, Pub/Sub, Apache Spark, Hadoop ecosystem, and Apache Kafka) - Transformations

Section	Objectives
	• Batch • Streaming (e.g., windowing, late arriving data) • Language • Ad hoc data ingestion (one-time or automated pipeline) - Data acquisition and import - Integrating with new data sources
Deploying and operationalizing the pipelines. Considerations include:	- Job automation and orchestration (e.g., Cloud Composer and Workflows) - CI/CD (Continuous Integration and Continuous Deployment)
Storing the data (20% of the exam)	
Selecting storage systems. Considerations include:	- Analyzing data access patterns - Choosing managed services (e.g., Bigtable, Cloud Spanner, Cloud SQL, Cloud Storage, Firestore, Memorystore) - Planning for storage costs and performance - Lifecycle management of data
Planning for using a data warehouse. Considerations include:	- Designing the data model - Deciding the degree of data normalization - Mapping business requirements - Defining architecture to support data access patterns
Using a data lake. Considerations include:	- Managing the lake (configuring data discovery, access, and cost controls) - Processing data - Monitoring the data lake
Designing for a data mesh. Considerations include:	- Building a data mesh based on requirements by using Google Cloud tools (e.g., Dataplex, Data Catalog, BigQuery, Cloud Storage) - Segmenting data for distributed team usage - Building a federated governance model for distributed data systems
Preparing and using data for analysis (15% of the exam)	
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)

Section	Objectives
Sharing data. Considerations include:	- Defining rules to share data - Publishing datasets - Publishing reports and visualizations - Analytics Hub
Exploring and analyzing data. Considerations include:	- Preparing data for feature engineering (training and serving machine learning models) - Conducting data discovery
Maintaining and automating data workloads (18% of the exam)	
Optimizing resources. Considerations include:	- Minimizing costs per required business need for data - Ensuring that enough resources are available for business-critical data processes - Deciding between persistent or job-based data clusters (e.g., Dataproc)
Designing automation and repeatability. Considerations include:	- Creating directed acyclic graphs (DAGs) for Cloud Composer - Scheduling jobs in a repeatable way
Organizing workloads based on business requirements. Considerations include:	- Flex, on-demand, and flat rate slot pricing (index on flexibility or fixed capacity) - Interactive or batch query jobs
Monitoring and troubleshooting processes. Considerations include:	- Observability of data processes (e.g., Cloud Monitoring, Cloud Logging, BigQuery admin panel) - Monitoring planned usage - Troubleshooting error messages, billing issues, and quotas - Manage workloads, such as jobs, queries, and compute capacity (reservations)
Maintaining awareness of failures and mitigating impact. Considerations include:	- Designing system for fault tolerance and managing restarts - Running jobs in multiple regions or zones - Preparing for data corruption and missing data - Data replication and failover (e.g., Cloud SQL, Redis clusters)

Google GCP-PDE Sample Questions:

Question: 1

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.

The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization.

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: 2

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: 3

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: 4

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: 5

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: 6

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.

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: 7

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: 8

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

Question: 9

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: 10

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

Study Guide to Crack Google Professional Data Engineer GCP-PDE Exam:

- Getting details of the GCP-PDE syllabus, is the first step of a study plan. This pdf is going to be of ultimate help. Completion of the syllabus is must to pass the GCP-PDE exam.
- Making a schedule is vital. A structured method of preparation leads to success. A candidate must plan his schedule and follow it rigorously to attain success.

- Joining the Google provided training for GCP-PDE exam could be of much help. If there is specific training for the exam, you can discover it from the link above.
- Read from the GCP-PDE sample questions to gain your idea about the actual exam questions. In this PDF useful sample questions are provided to make your exam preparation easy.
- Practicing on GCP-PDE practice tests is must. Continuous practice will make you an expert in all syllabus areas.