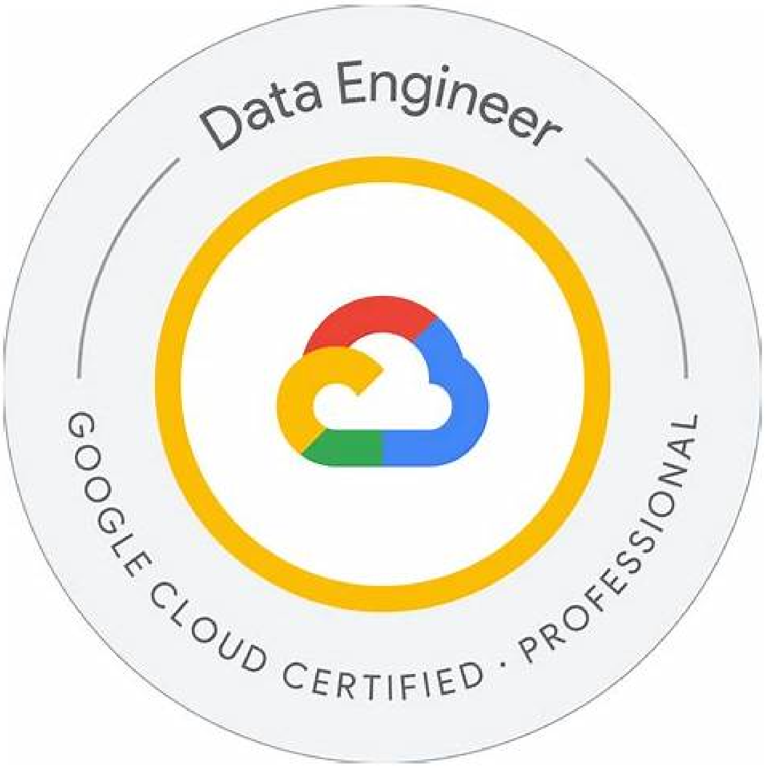
Pass Guaranteed Quiz 2026 High Pass-Rate Professional-Data-Engineer: Google Certified Professional Data Engineer Exam High Quality



P.S. Free 2026 Google Professional-Data-Engineer dumps are available on Google Drive shared by Test4Engine: https://drive.google.com/open?id=13mLZucVnflmfUfEavZWHpCNcdUhRj9kE

After you practice our study materials, you can master the examination point from the Professional-Data-Engineer exam torrent. Then, you will have enough confidence to pass your exam. We can succeed so long as we make efforts for one thing. As for the safe environment and effective product, why don't you have a try for our Professional-Data-Engineer Test Question, never let you down! Before your purchase, there is a free demo for you. You can know the quality of our Professional-Data-Engineer guide question earlier.

The Google Professional Data Engineer certification is designed to equip the individuals with the required knowledge and skills to enable data-driven decision-making through collecting, transforming, and publishing data. To earn this certificate, the candidates will be required to pass a single test measuring their skills in leveraging, deploying, and continuously training pre-existing machine learning models. The qualifying exam also evaluates the ability of the applicants to design, build, operationalize, monitor, and secure data processing systems.

>> Professional-Data-Engineer High Quality <<

Valid and Reliable Professional-Data-Engineer Exam Questions [2026]

The Google Professional-Data-Engineer desktop-based practice exam is compatible with Windows-based computers and only requires an internet connection for the first-time license validation. The web-based Google Certified Professional Data Engineer Exam (Professional-Data-Engineer) practice test is accessible on any browser without needing to install any separate software. Finally, the Google Certified Professional Data Engineer Exam (Professional-Data-Engineer) dumps pdf is easily portable and can be used on smart devices or printed out.

Google Certified Professional Data Engineer Exam Sample Questions (Q129-Q134):

NEW QUESTION # 129
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:
The user profile: What the user likes and doesn't like to eat
The user account information: Name, address, preferred meal times
The order information: When orders are made, from where, to whom
The database will be used to store all the transactional data of the product. You want to optimize the data schem
a. Which Google Cloud Platform product should you use?

- A. Cloud Datastore
- B. Cloud SQL
- C. Cloud Bigtable
- D. BigQuery

Answer: D

NEW QUESTION # 130
You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country You check the query plan for the query and see the following output in the Read section of Stage:1:
What is the most likely cause of the delay for this query?

- A. The [myproject:mydataset.mytable] table has too many partitions
- B. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
- C. Users are running too many concurrent queries in the system
- D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Answer: C

NEW QUESTION # 131
Flowlogistic Case Study
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.
Solution Concept
Flowlogistic wants to implement two concepts using the cloud:
* Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
* Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.
Existing Technical Environment
Flowlogistic architecture resides in a single data center:
* Databases
- 8 physical servers in 2 clusters
- SQL Server - user data, inventory, static data
- 3 physical servers
- Cassandra - metadata, tracking messages
10 Kafka servers - tracking message aggregation and batch insert
* Application servers - customer front end, middleware for order/customs
- 60 virtual machines across 20 physical servers

- Tomcat - Java services
- Nginx - static content
- Batch servers
* Storage appliances
- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
Network-attached storage (NAS) image storage, logs, backups
* 10 Apache Hadoop /Spark servers
- Core Data Lake
- Data analysis workloads
* 20 miscellaneous servers
- Jenkins, monitoring, bastion hosts,
Business Requirements
* Build a reliable and reproducible environment with scaled panty of production.
* Aggregate data in a centralized Data Lake for analysis
* Use historical data to perform predictive analytics on future shipments
* Accurately track every shipment worldwide using proprietary technology
* Improve business agility and speed of innovation through rapid provisioning of new resources
* Analyze and optimize architecture for performance in the cloud
* Migrate fully to the cloud if all other requirements are met
Technical Requirements
* Handle both streaming and batch data
* Migrate existing Hadoop workloads
* Ensure architecture is scalable and elastic to meet the changing demands of the company.
* Use managed services whenever possible
* Encrypt data flight and at rest
Connect a VPN between the production data center and cloud environment
SEO Statement
We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

- A. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
- B. Cloud Dataflow, Cloud SQL, and Cloud Storage
- C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
- D. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
- E. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Answer: C

NEW QUESTION # 132
You are preparing an organization-wide dataset. You need to preprocess customer data stored in a restricted bucket in Cloud Storage. The data will be used to create consumer analyses. You need to follow data privacy requirements, including protecting certain sensitive data elements, while also retaining all of the data for potential future use cases. What should you do?

- A. Use Dataflow and the Cloud Data Loss Prevention API to mask sensitive data. Write the processed data in BigQuery.
- B. Use Dataflow and Cloud KMS to encrypt sensitive fields and write the encrypted data in BigQuery.
  Share the encryption key by following the principle of least privilege.

- C. Use the Cloud Data Loss Prevention API and Dataflow to detect and remove sensitive fields from the data in Cloud Storage. Write the filtered data in BigQuery.
- D. Use customer-managed encryption keys (CMEK) to directly encrypt the data in Cloud Storage. Use federated queries from BigQuery. Share the encryption key by following the principle of least privilege.

Answer: A

Explanation:
The core requirements are to protect sensitive data elements (data privacy) while retainingalldata for potential future use, and then using this preprocessed data for consumer analyses.
Retaining All Data:This immediately makes option B (remove sensitive fields) unsuitable because it involves data loss.
Protecting Sensitive Data for Analysis & Future Use:Masking is a de-identification technique that redacts or replaces sensitive data with a substitute, allowing the data structure and usability for analysis to be maintained without exposing the original sensitive values. This aligns with protecting data while still making it usable.
Cloud Data Loss Prevention (DLP) API:This service is specifically designed to discover, classify, and protect sensitive data. It offers various de-identification techniques, including masking.
Dataflow:This is a serverless, fast, and cost-effective service for unified stream and batch data processing. It's well-suited for transforming large datasets, such as those read from Cloud Storage, and can integrate with the DLP API for de-identification.
Writing to BigQuery:BigQuery is an ideal destination for an organization-wide dataset for consumer analyses.
Therefore, using Dataflow to read the data from Cloud Storage, leveraging the Cloud DLP API tomask(a form of de-identification) the sensitive elements, and then writing the processed (masked) data to BigQuery is the most appropriate solution. This approach protects privacy for the consumer analyses dataset while the original, unaltered data can still be retained in the restricted Cloud Storage bucket for future use cases that might require access to the original sensitive information (under strict governance).
Let's analyze why other options are less suitable:
Option B:"Remove sensitive fields" means data loss, which contradicts the requirement to retain all data for potential future use cases.
Option C:Encrypting sensitive fields with Cloud KMS and writing them to BigQuery is a valid way to protect data. However, for "consumer analyses," masked data is generally more directly usable than encrypted data.
Analysts would typically work with de-identified (e.g., masked) data rather than directly querying encrypted fields and managing decryption keys for analytical purposes. While decryption is possible, masking often provides a better balance of privacy and utility for broad analysis. The question also implies creating a datasetforanalysis, where masking makes the data ready-to-use for that purpose. The original data remains in Cloud Storage.
Option D:Using CMEK encrypts the entire object in Cloud Storage at rest. While this protects the data in Cloud Storage, federated queries from BigQuery would access the raw, unmasked data (assuming decryption occurs seamlessly). This doesn't address the preprocessing requirement of protectingcertain sensitive data elementswithin the data itself for theconsumer analysesdataset. The goal is to create a de-identified dataset for analysis, not just secure the raw data at rest.
Reference:
Google Cloud Documentation: Cloud Data Loss Prevention > De-identification overview. "De-identification is the process of removing identifying information from data. Cloud DLP uses de-identification techniques such as masking, tokenization, pseudonymization, date shifting, and more to help you protect sensitive data." Google Cloud Documentation: Cloud Data Loss Prevention > Basic de-identification > Masking. "Masking hides parts of data by replacing characters with a symbol, such as an asterisk (*) or hash (#)." Google Cloud Documentation: Dataflow > Overview. "Dataflow is a fully managed streaming analytics service that minimizes latency, processing time, and cost through autoscaling and batch processing." Google Cloud Solution: Automating the de-identification of PII in large-scale datasets using Cloud DLP and Dataflow. This solution guide explicitly outlines using Dataflow and DLP API for de-identifying (including masking) data from Cloud Storage and loading it into BigQuery. "You can use Cloud DLP to scan data for sensitive elements andthen apply de-identification techniques such as redaction, masking, or tokenization." and "This tutorial uses Dataflow to orchestrate the de-identification process."

NEW QUESTION # 133
You have terabytes of customer behavioral data streaming from Google Analytics into BigQuery daily Your customers' information, such as their preferences, is hosted on a Cloud SQL for MySQL database Your CRM database is hosted on a Cloud SQL for PostgreSQL instance. The marketing team wants to use your customers' information from the two databases and the customer behavioral data to create marketing campaigns for yearly active customers. You need to ensure that the marketing team can run the campaigns over 100 times a day on typical days and up to 300 during sales. At the same time you want to keep the load on the Cloud SQL databases to a minimum. What should you do?

- A. Create streams in Datastream to replicate the required tables from both Cloud SQL databases to BigQuery for these queries.
- B. Create a Dataproc cluster with Trino to establish connections to both Cloud SQL databases and BigQuery, to execute the queries.

- C. Create a job on Apache Spark with Dataproc Serverless to query both Cloud SQL databases and the Google Analytics data on BigQuery for these queries.
- D. Create BigQuery connections to both Cloud SQL databases Use BigQuery federated queries on the two databases and the Google Analytics data on BigQuery to run these queries.

Answer: D

Explanation:
Datastream is a serverless Change Data Capture (CDC) and replication service that allows you to stream data changes from Oracle and MySQL databases to Google Cloud services such as BigQuery, Cloud Storage, Cloud SQL, and Pub/Sub. Datastream captures and delivers database changes in real-time, with minimal impact on the source database performance. Datastream also preserves the schema and data types of the source database, and automatically creates and updates the corresponding tables in BigQuery.
By using Datastream, you can replicate the required tables from both Cloud SQL databases to BigQuery, and keep them in sync with the source databases. This way, you can reduce the load on the Cloud SQL databases, as the marketing team can run their queries on the BigQuery tables instead of the Cloud SQL tables. You can also leverage the scalability and performance of BigQuery to query the customer behavioral data from Google Analytics and the customer information from the replicated tables. You can run the queries as frequently as needed, without worrying about the impact on the Cloud SQL databases.
Option A is not a good solution, as BigQuery federated queries allow you to query external data sources such as Cloud SQL databases, but they do not reduce the load on the source databases. In fact, federated queries may increase the load on the source databases, as they need to execute the query statements on the external data sources and return the results to BigQuery. Federated queries also have some limitations, such as data type mappings, quotas, and performance issues.
Option C is not a good solution, as creating a Dataproc cluster with Trino would require more resources and management overhead than using Datastream. Trino is a distributed SQL query engine that can connect to multiple data sources, such as Cloud SQL and BigQuery, and execute queries across them. However, Trino requires a Dataproc cluster to run, which means you need to provision, configure, and monitor the cluster nodes. You also need to install and configure the Trino connector for Cloud SQL and BigQuery, and write the queries in Trino SQL dialect. Moreover, Trino does not replicate or sync the data from Cloud SQL to BigQuery, so the load on the Cloud SQL databases would still be high.
Option D is not a good solution, as creating a job on Apache Spark with Dataproc Serverless would require more coding and processing power than using Datastream. Apache Spark is a distributed data processing framework that can read and write data from various sources, such as Cloud SQL and BigQuery, and perform complex transformations and analytics on them. Dataproc Serverless is a serverless Spark service that allows you to run Spark jobs without managing clusters. However, Spark requires you to write code in Python, Scala, Java, or R, and use the Spark connector for Cloud SQL and BigQuery to access the data sources. Spark also does not replicate or sync the data from Cloud SQL to BigQuery, so the load on the Cloud SQL databases would still be high. References: Datastream overview | Datastream | Google Cloud, Datastream concepts | Datastream | Google Cloud, Datastreamquickstart | Datastream | Google Cloud, Introduction to federated queries | BigQuery | Google Cloud, Trino overview | Dataproc Documentation | Google Cloud, Dataproc Serverless overview | Dataproc Documentation | Google Cloud, Apache Spark overview | Dataproc Documentation | Google Cloud.

NEW QUESTION # 134
......