Practice with Professional-Data-Engineer Dumps for Google Cloud Certified Certified Exam Questions & Answer [Q41-Q64

Practice with Professional-Data-Engineer Dumps for Google Cloud Certified Certified Exam Questions & Answer [Q41-Q64]

Practice with Professional-Data-Engineer Dumps for Google Cloud Certified Certified Exam Questions & Answer
REAL Professional-Data-Engineer Exam Questions With 100% Refund Guarantee

Google Professional-Data-Engineer exam is a certification offered by Google to professionals who specialize in data engineering. Professional-Data-Engineer exam is designed to test the candidate's understanding of data processing systems, data modeling, data governance, and data transformation. Google Certified Professional Data Engineer Exam certification aims to validate the candidate's expertise in Google Cloud Platform's data engineering technologies and their ability to design and develop effective data solutions.

Q41. You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query &#8211; -dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

* Create a separate table for each ID.
* Use the LIMIT keyword to reduce the number of rows returned.

* Recreate the table with a partitioning column and clustering column.
* Use the bq query &#8211; -maximum_bytes_billed flag to restrict the number of bytes billed.

Q42. The marketing team at your organization provides regular updates of a segment of your customer dataset.

The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?
* Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
* Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
* Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
* Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.
https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery

Q43. Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)
* The wide model is used for memorization, while the deep model is used for generalization.
* A good use for the wide and deep model is a recommender system.
* The wide model is used for generalization, while the deep model is used for memorization.
* A good use for the wide and deep model is a small-scale linear regression problem.
Explanation

Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It&#8217;s not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It&#8217;s useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.

Reference: https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

Q44. After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You&#8217;ve loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.

What should you do?
* Select random samples from the tables using the RAND() function and compare the samples.
* Select random samples from the tables using the HASH() function and compare the samples.
* Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.
* Create stratified random samples using the OVER() function and compare equivalent samples from each table.
Full comparison with this option, rest are comparison on sample which doesn&#8217;t ensure all the data will be ok.

Q45. Which is not a valid reason for poor Cloud Bigtable performance?
* The workload isn&#8217;t appropriate for Cloud Bigtable.
* The table&#8217;s schema is not designed correctly.
* The Cloud Bigtable cluster has too many nodes.
* There are issues with the network connection.
Explanation

The Cloud Bigtable cluster doesn&#8217;t have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance. Use the monitoring tools to check whether the cluster is overloaded.

Reference: https://cloud.google.com/bigtable/docs/performance

Q46. You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the &#8220;Trust No One&#8221; (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?
* Use gcloud kms keys createto create a symmetric key. Then use gcloud kms encryptto encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
* Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encryptto encrypt each archival file with the key. Use gsutil cpto upload each encrypted file to the Cloud Storage bucket.

Manually destroy the key previously used for encryption, and rotate the key once.
* Specify customer-supplied encryption key (CSEK) in the .botoconfiguration file. Use gsutil cpto upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
* Specify customer-supplied encryption key (CSEK) in the .botoconfiguration file. Use gsutil cpto upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

Q47. You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients&#8217; personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems.

What should you do?
* Create an authorized view in BigQuery to restrict access to tables with sensitive data.
* Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
* Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
* Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API.

Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Q48. Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?
* Issue a command to restart the database servers.
* Retry the query with exponential backoff, up to a cap of 15 minutes.
* Retry the query every second until it comes back online to minimize staleness of data.
* Reduce the query frequency to once every hour until the database comes back online.
Explanation/Reference:

Q49. Your company&#8217;s customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations.

What should you do?
* Add a node to the MySQL cluster and build an OLAP cube there.
* Use an ETL tool to load the data from MySQL into Google BigQuery.
* Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.

* Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Q50. You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.

What should you do?
* Use Cloud Dataflow with Beam to detect errors and perform transformations.
* Use Cloud Dataprep with recipes to detect errors and perform transformations.
* Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
* Use federated tables in BigQuery with queries to detect errors and perform transformations.

Q51. Which of the following is not true about Dataflow pipelines?
* Pipelines are a set of operations
* Pipelines represent a data processing job
* Pipelines represent a directed graph of steps
* Pipelines can share data between instances
The data and transforms in a pipeline are unique to, and owned by, that pipeline. While your program can create multiple pipelines, pipelines cannot share data or transforms

Q52. 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&#8217;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 schema. Which Google Cloud Platform product should you use?
* BigQuery
* Cloud SQL
* Cloud Bigtable
* Cloud Datastore

Q53. Cloud Bigtable is a recommended option for storing very large amounts of

____________________________?
* multi-keyed data with very high latency
* multi-keyed data with very low latency
* single-keyed data with very low latency
* single-keyed data with very high latency
Explanation

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.

Reference: https://cloud.google.com/bigtable/docs/overview

Q54. Which of the following statements about Legacy SQL and Standard SQL is not true?
* Standard SQL is the preferred query language for BigQuery.
* If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
* One difference between the two query languages is how you specify fully-qualified table names (i.e. table names that include their associated project name).
* You need to set a query language for each dataset and the default is Standard SQL.
You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.

Standard SQL has been the preferred query language since BigQuery 2.0 was released.

In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.

Due to the differences in syntax between the two query languages (such as with project- qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

Reference:

https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

Q55. You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
* Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
* Use Cloud Dataproc to run your transformations. Use the diagnosecommand to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
* Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
* Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs.

Configure the job to use non-default Compute Engine machine types when needed.
Explanation

Q56. You are building a model to make clothing recommendations. You know a user&#8217;s fashion pis likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?
* Continuously retrain the model on just the new data.
* Continuously retrain the model on a combination of existing data and the new data.
* Train on the existing data while using the new data as your test set.
* Train on the new data while using the existing data as your test set.

Q57. Case Study 1 &#8211; Flowlogistic

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

&#8211; SQL Server &#8211; user data, inventory, static data

3 physical servers

&#8211; Cassandra &#8211; metadata, tracking messages

10 Kafka servers &#8211; tracking message aggregation and batch insert

* Application servers &#8211; customer front end, middleware for order/customs

60 virtual machines across 20 physical servers

&#8211; Tomcat &#8211; Java services

&#8211; Nginx &#8211; static content

&#8211; Batch servers

* Storage appliances

&#8211; iSCSI for virtual machine (VM) hosts

&#8211; Fibre Channel storage area network (FC SAN) &#8211; SQL server storage

&#8211; Network-attached storage (NAS) image storage, logs, backups

* 10 Apache Hadoop /Spark servers

&#8211; Core Data Lake

&#8211; Data analysis workloads

* 20 miscellaneous servers

&#8211; 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&#8217; 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&#8217;t want to commit capital to building out a server environment.

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?
* Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
* Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.
* Use the NOW () function in BigQuery to record the event&#8217;s time.
* Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Q58. MJTelco Case Study

Company Overview

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world.

The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.

* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments &#8211; development/test, staging, and production &#8211; to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements

* Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

* Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

* Provide reliable and timely access to data for analysis from distributed research workers

* Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements

* Ensure secure and efficient transport and storage of telemetry data

* Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

* Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately

100m records/day

* Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud&#8217;s machine learning will allow our quantitative researchers to work on our high- value problems instead of problems with our data pipelines.

You need to compose visualizations for operations teams with the following requirements:

* The report must include telemetry data from all 50,000 installations for the most resent 6 weeks (sampling once every minute).

* The report must not be more than 3 hours delayed from live data.

* The actionable report should only show suboptimal links.

* Most suboptimal links should be sorted to the top.

* Suboptimal links can be grouped and filtered by regional geography.

* User response time to load the report must be <5 seconds.

Which approach meets the requirements?
* Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
* Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
* Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
* Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Q59. A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete framing. The model has custom TensorFlow operations that must run partially on a CPU You want to reduce the training time in a cost-effective manner. What should you do?
* Change the VM type to n2-highmem-32
* Change the VM type to e2 standard-32
* Train the model using a VM with a GPU hardware accelerator
* Train the model using a VM with a TPU hardware accelerator

Q60. You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users&#8217; privacy?
* Grant the consultant the Viewer role on the project.
* Grant the consultant the Cloud Dataflow Developer role on the project.
* Create a service account and allow the consultant to log on with it.
* Create an anonymized sample of the data for the consultant to work with in a different project.
A service account is a special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs.

https://cloud.google.com/iam/docs/understanding-service-accounts

Q61. Which Google Cloud Platform service is an alternative to Hadoop with Hive?
* Cloud Dataflow
* Cloud Bigtable
* BigQuery
* Cloud Datastore

Explanation

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.

Google BigQuery is an enterprise data warehouse.

Reference: https://en.wikipedia.org/wiki/Apache_Hive

Q62. You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

* Use the TABLE_DATE_RANGE function
* Use the WHERE_PARTITIONTIME pseudo column
* Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
* Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD

Legacy sql uses table date range whereas standard sql uses table_sufix for wildcard.

Q63. Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

* The wide model is used for memorization, while the deep model is used for generalization.
* A good use for the wide and deep model is a recommender system.
* The wide model is used for generalization, while the deep model is used for memorization.
* A good use for the wide and deep model is a small-scale linear regression problem.

Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It&#8217;s not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It&#8217;s useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.

Q64. The __________ for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

* Cloud Dataflow connector
* DataFlow SDK
* BiqQuery API
* BigQuery Data Transfer Service

Explanation

The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.

Reference: https://cloud.google.com/bigtable/docs/dataflow-hbase

PDF Download Google Test To Gain Brilliante Result!:

https://www.dumpsmaterials.com/Professional-Data-Engineer-real-torrent.html]