Free Questions for Professional-Data-Engineer by certscare

Shared by Brewer on 06-06-2022

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: **MultipleChoice**

You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons What should you do?

Options:

A- Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.

B- Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors

C- Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in
Cloud Storage

D- Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage

Answer:

C

Question 2

Question Type: MultipleChoice

You are migrating an application that tracks library books and information about each book, such as author or year published, from an on-premises data warehouse to BigQuery In your current relational database, the author information is kept in a separate table and joined to the book information on a common key Based on Google's recommended practice for schema design, how would you structure the data to ensure optimal speed of queries about the author of each book that has been borrowed?

Options:

A- Keep the schema the same, maintain the different tables for the book and each of the attributes, and query as you are doing today

B- Create a table that is wide and includes a column for each attribute, including the author's first name, last name, date of birth, etc

C- Create a table that includes information about the books and authors, but nest the author fields inside the author column

D- Keep the schema the same, create a view that joins all of the tables, and always query the view

Answer:

C

Question 3

Question Type: **MultipleChoice**

You need (o give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system How should you design your pipeline to minimize that backpressure?

Options:

A- Call out to the service via HTTP

B- Create the pipeline statically in the class definition

C- Create a new object in the startBundle method of DoFn

D- Batch the job into ten-second increments

Answer:

A

Question 4

Question Type: **MultipleChoice**

You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time What tools should you use to migrate the data and ensure that it continues to write to your warehouse?

Options:

A- Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates

B- BigQuery Data Transfer Service lor the migration, Pub/Sub and Dataproc for the real-time updates

C- gsutil for the migration; Pub/Sub and Dataflow for the real-time updates

D- gsutil for both the migration and the real-time updates

Answer:

C

Question 5

Question Type: **MultipleChoice**

You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API Following Google's best practices, you have both a staging and a production table for the data How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?

Options:

A- Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging

B- Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging

C- Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours

D- Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes

Answer:

D

Question 6

Question Type: **MultipleChoice**

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SideInputs to join data You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

Options:

A- Switch to compressed Avro files

B- Reduce the batch size

C- Retry records that throw an error

D- Use CoGroupByKey instead of the SideInput

Answer:

B

Question 7

Question Type: **MultipleChoice**

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in How should you design your row key and tables to ensure that you can access the data with the most simple query?

Options:

A- Create one unique table for all of the indices, and then use the index and timestamp as the row key design

B- Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.

C- For each index, have a separate table and use a timestamp as the row key design

D- For each index, have a separate table and use a reverse timestamp as the row key design

Answer:

A

Question 8

Question Type: MultipleChoice

Your company is implementing a data warehouse using BigQuery and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

Options:

A- Denormalize the data

B- Shard the data by customer ID

C- Materialize the dimensional data in views

D- Partition the data by transaction date

Answer:

D

Question 9

Question Type: **MultipleChoice**

You are building a teal-lime prediction engine that streams files, which may contain Pll (personal identifiable information) data, into Cloud Storage and eventually into BigQuery You want to ensure that the sensitive data is masked but still maintains referential Integrity, because names and emails are often used as join keys How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the Pll data is not accessible by unauthorized individuals?

Options:

A- Create a pseudonym by replacing the PII data with cryptogenic tokens, and store the non-tokenized data in a locked-down button.

B- Redact all PII data, and store a version of the unredacted data in a locked-down bucket

C- Scan every table in BigQuery, and mask the data it finds that has PII

D- Create a pseudonym by replacing PII data with a cryptographic format-preserving token

Answer:

A

Question 10

Question Type: **MultipleChoice**

You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

Options:

A- Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery

B- Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.

C- Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table

D- Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery

Answer:

A

Question 11

Question Type: MultipleChoice

You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app You have reviewed old chat logs and lagged each conversation for intent based on each customer's stated intention for contacting customer service About 70% of customer requests are simple requests that are solved within 10 intents The remaining 30% of inquiries require much longer, more complicated requests Which intents should you automate first?

Options:

A- Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests

B- Automate the more complicated requests first because those require more of the agents' time

C- Automate a blend of the shortest and longest intents to be representative of all intents

D- Automate intents in places where common words such as 'payment' appear only once so the software isn't confused

Answer:

A

20%
DISCOUNT