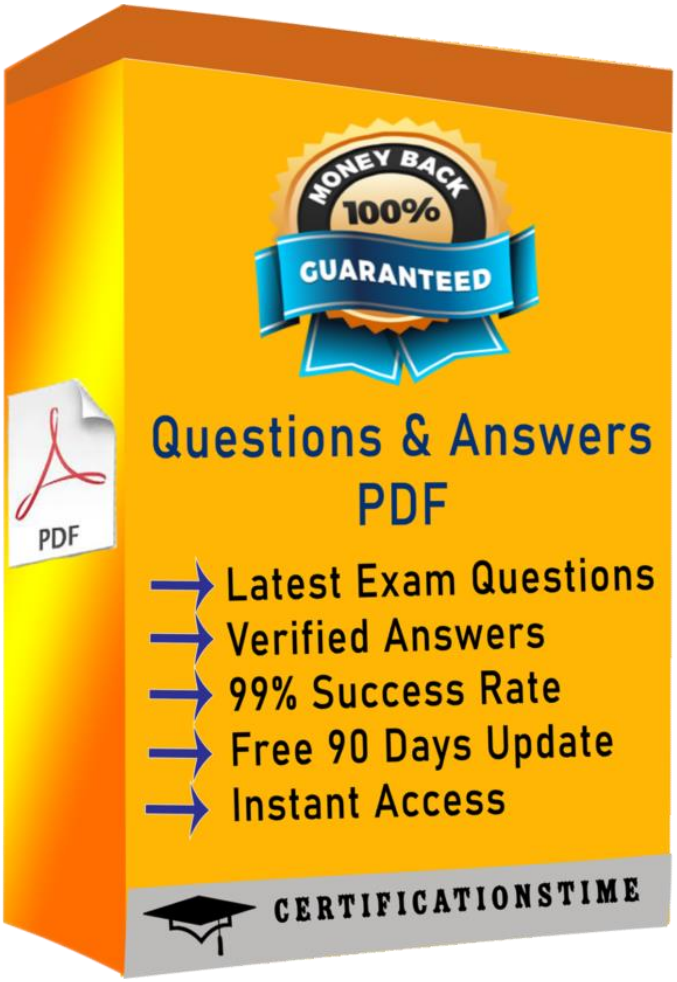
Exam Questions Professional-Data-Engineer

Cloud Certified Professional Data Engineer

https://certificationstime.com/



QUESTION 1

Your company built a TensorFlow neutral-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction

Correct Answer: D

Explanation/Reference:

Reference https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-predictionusing-tensorflow-30505541d877

QUESTION 2

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data whenmaking this update. What should you do?

A. . Update the current pipeline and use the drain flag.
B. Update the current pipeline and provide the transform mapping JSON object.
C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Correct Answer: D

QUESTION 3

What are two of the benefits of using denormalized data structures in BigQuery?

A. . Reduces the amount of data processed, reduces the amount of storage required
B. Increases query speed, makes queries simpler
C. Reduces the amount of storage required, increases query speed
D. Reduces the amount of data processed, increases query speed

Correct Answer: B

Explanation/Reference:

Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINs on large tables, but with a denormalized data structure, you don't have to use JOINs, since all of the data has been combined into one table. Denormalization also makes queries simpler because you do not have to use JOIN clauses. Denormalization increases the amount of data processed and the amount of storage required because it creates redundant data.
Reference:
https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

QUESTION 4

How can you get a neural network to learn about relationships between categories in a categorical feature?

A. Create a multi-hot column
B. Create a one-hot column
C. Create a hash bucket
D. Create an embedding column

Correct Answer: D

Explanation/Reference:

There are two problems with one-hot encoding. First, it has high dimensionality, meaning thatinstead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.
Both of these problems can be solved by representing a categorical feature with an embedding column. The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).
You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.
Reference:
https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/awide-and-deep-model.html

QUESTION 5

When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and _____

A. zone
B. node
C. label
D. type

Correct Answer: A

Explanation/Reference:

At a minimum, you must specify four values when creating a new cluster with the projects.regions.clusters.create operation:
The project in which the cluster will be created
The region to use
The name of the cluster
The zone in which the cluster will be created
You can specify many more details beyond these minimum requirements. For example, you can also specify the number of workers, whether preemptible compute should be used, and the network settings.
Reference:
https://cloud.google.com/dataproc/docs/tutorials/python-libraryexample#create_a_new_cloud_dataproc_cluste

QUESTION 6

An organization maintains a Google BigQuery dataset that contains tables with user-level dat A. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost andensure the analysis cost for other projects is assigned to those projects. What should they do?

A. Create and share an authorized view that provides the aggregate results.
B. . Create and share a new dataset and view that provides the aggregate results.
C. Create and share a new dataset and table that contains the aggregate results.
D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Correct Answer: D

Explanation/Reference:

Reference:
https://cloud.google.com/bigquery/docs/access-control

QUESTION 7

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

A. . Batch job, PubSubIO, side-inputs
B. Streaming job, PubSubIO, JdbcIO, side-outputs
C. Streaming job, PubSubIO, BigQueryIO, side-inputs
D. . Streaming job, PubSubIO, BigQueryIO, side-outputs

Correct Answer: A

QUESTION 8

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code
D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code

Correct Answer: A

QUESTION 9

You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

A. PCollection

B. Transform
C. Pipeline
D. Sink API

Correct Answer: B

Explanation/Reference:

In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.
Reference:
https://cloud.google.com/dataflow/model/programming-model

QUESTION 10

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 .
B. Use a row key of the form .
C. . Use a row key of the form #.
D. . Use a row key of the form >##.

Correct Answer: A

QUESTION 11

Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

A. . They have not assigned the timestamp, which causes the job to fail
B. They have not set the triggers to accommodate the data coming in late, which causes the job to fai
C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Correct Answer: C

QUESTION 12

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Whichtwo actions can you take to accomplish this? Choose 2 answers.

A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
C. . Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

Correct Answer: A,C

QUESTION 13

Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:
Syntax error : Expected end of statement but got "-" at [4:11]
SELECT age
FROM
bigquery-public-data.noaa_gsod.gsod
WHERE
age != 99
AND_TABLE_SUFFIX = '1929'
ORDER BY
age DESC
Which table name will make the SQL statement work correctly?

A. 'bigquery-public-data.noaa_gsod.gsod'
B. . bigquery-public-data.noaa_gsod.gsod*
C. 'bigquery-public-data.noaa_gsod.gsod'*
D. 'bigquery-public-data.noaa_gsod.gsod*`

Correct Answer: D

QUESTION 14

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Correct Answer: C

QUESTION 15

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, thescope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

A. Add capacity (memory and disk space) to the database server by the order of 200
B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Correct Answer: B

QUESTION 16

Your company needs to upload their historic data to Cloud Storage. The security rules don’t allow access from external IPs to their on-premises resources. After an initial upload, they will add new data

from existing on-premises applications every day. What should they do?

A. Execute gsutil rsync from the on-premises servers.
B. Use Cloud Dataflow and write the data to Cloud Storage.
C. Write a job template in Cloud Dataproc to perform the data transfer
D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Correct Answer: B

QUESTION 17

You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on nonkey columns. What should you do?

A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.
D. . Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Correct Answer: D

Explanation/Reference:

Reference:
https://cloud.google.com/solutions/data-lifecycle-cloud-platform

QUESTION 18

You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Correct Answer: A

QUESTION 19

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern. Which service do you select for storing and serving your data?

A. Cloud Spanner
B. Cloud Bigtable
C. Cloud Firestore
D. Cloud SQL

Correct Answer: D

QUESTION 20

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-todate YouTube channels log data. How should you set up the log data transfer into Google Cloud?

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
B. . Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
C. . Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage MultiRegional storage bucket as a final destination.
D. . Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Correct Answer: B

QUESTION 21

You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?

A. Load the data every 30 minutes into a new partitioned table in BigQuery.
B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery
C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore
D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Correct Answer: A