PR OFESSIONAL-DATA-ENGINEER^{Q&As}

Professional Data Engineer on Google Cloud Platform

Pass Google PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.leads4pass.com/professional-data-engineer.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Google Official Exam Center

QUESTION 1

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system
that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the
data

periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a
different environment. You want to archive these snapshots for a long time.

Which two methods can accomplish this? Choose 2 answers.

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.

B. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace
reserved for that export.

C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete
temporary export files.

D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery
table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column
for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.

E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a
JSON file. Apply compression before storing the data in Cloud Source Repositories.

Correct Answer: AB

QUESTION 2

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the
application\\'s interface. You do not want individual users to authenticate to BigQuery and you do not want to give them
access to the dataset. You need to securely access BigQuery from your IT application.

What should you do?

A. Create groups for your users and give those groups access to the dataset

B. Integrate with a single sign-on (SSO) platform, and pass each user\\'s credentials along with the query request

C. Create a service account and grant dataset access to that account. Use the service account\\'s private key to access
the dataset

D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file
on the files system, and use those credentials to access the BigQuery dataset

Correct Answer: C

QUESTION 3

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?

A. Change the VM type to n2-highmem-32

B. Change the VM type to e2 standard-32

C. Train the model using a VM with a GPU hardware accelerator

D. Train the model using a VM with a TPU hardware accelerator

Correct Answer: C

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

There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead 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/a-wide-and-deep-model.html

---

QUESTION 5

Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

A. Enable data access logs in each Data Analyst\\'s project. Restrict access to Stackdriver Logging via Cloud IAM roles.

B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts\\' projects. Restrict access to the Cloud Storage bucket.

C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.

D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.

Correct Answer: D

QUESTION 6

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?

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

Correct Answer: D

https://cloud.google.com/architecture/building-production-ready-data-pipelines-using-dataflow-developing-and-testing#choose_correctly_between_side_inputs_or_cogroupbykey_for_joins

The CoGroupByKey transform is a core Beam transform that merges (flattens) multiple PCollection objects and groups elements that have a common key. Unlike a side input, which makes the entire side input data available to each worker,

CoGroupByKey performs a shuffle (grouping) operation to distribute data across workers. CoGroupByKey is therefore ideal when the PCollection objects you want to join are very large and don\\'t fit into worker memory.

Use CoGroupByKey if you need to fetch a large proportion of a PCollection object that significantly exceeds worker memory.

QUESTION 7

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: A

QUESTION 8

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

A. There are very few occurrences of mutations relative to normal samples.

B. There are roughly equal occurrences of both normal and mutated samples in the database.

C. You expect future mutations to have different features from the mutated samples in the database.

D. You expect future mutations to have similar features to the mutated samples in the database.

E. You already have labels for which samples are mutated and which are normal in the database.

Correct Answer: AD

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

https://en.wikipedia.org/wiki/Anomaly_detection

QUESTION 9

An organization maintains a Google BigQuery dataset that contains tables with user-level datA. 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 and ensure 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: A

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

QUESTION 10

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: B

QUESTION 11

Which of these is NOT a way to customize the software on Dataproc cluster instances?

A. Set initialization actions

B. Modify configuration files using cluster properties

C. Configure the cluster using Cloud Deployment Manager

D. Log into the master node and make changes from there

Correct Answer: C

You can access the master node of the cluster by clicking the SSH button next to it in the Cloud Console.

You can easily use the --properties option of the dataproc command in the Google Cloud SDK to modify many common configuration files when creating a cluster.

When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up. [https://

cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions]

Reference: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties

QUESTION 12

You are deploying a new storage system for your mobile application, which is a media streaming service. You decide

the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity `Movie\\' the property `actors\\' and the property `tags\\' have multiple values but the property `date released\\' does not. A typical query would ask for all movies with actor= ordered by date_released or all movies with tag=Comedy ordered by date_released.

How should you avoid a combinatorial explosion in the number of indexes?

A. Manually configure the index in your index config as follows:

```
Indexes:
-kind: Movie
      Properties:
      -name: actors
      name: date_released
-kind: Movie
      Properties:
      -name: tags
      name: date_released
```

B. Manually configure the index in your index config as follows:

```
Indexes:
      -kind: Movie
            Properties:
            -name: actors
            -name: tags
-name: date_published
```

C. Set the following in your entity options: exclude_from_indexes = `actors, tags`

D. Set the following in your entity options: exclude_from_indexes = `date_published`

A. Option A

B. Option B.

C. Option C

D. Option D

Correct Answer: A

QUESTION 13

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the

campaign, the data is growing exponentially every hour.

The data scientists have written the following code to read the data for a new key features in the logs.

BigQueryIO.Read

.named("ReadLogData")

.from("clouddataflow-readonly:samples.log_data")

You want to improve the performance of this data read. What should you do?

A. Specify the TableReference object in the code.

B. Use .fromQuery operation to read specific fields from the table.

C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.

D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.

Correct Answer: B

QUESTION 14

The Dataflow SDKs have been recently transitioned into which Apache service?

A. Apache Spark

B. Apache Hadoop

C. Apache Kafka

D. Apache Beam

Correct Answer: D

Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive Reference: https://cloud.google.com/dataflow/docs/

QUESTION 15

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

A. The wide model is used for memorization, while the deep model is used for generalization.

B. A good use for the wide and deep model is a recommender system.

C. The wide model is used for generalization, while the deep model is used for memorization.

D. A good use for the wide and deep model is a small-scale linear regression problem.

Correct Answer: AB

Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It\\'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 and Deep Learning. It\\'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

PROFESSIONAL-DATA-
ENGINEER VCE Dumps

PROFESSIONAL-DATA-
ENGINEER Practice Test

PROFESSIONAL-DATA-
ENGINEER Study Guide