(2022) Professional-Data-Engineer Dumps and Practice Test (270 Questions) [Q97-Q113

(2022) Professional-Data-Engineer Dumps and Practice Test (270 Questions)
Guide (New 2022) Actual Google Professional-Data-Engineer Exam Questions

Operationalizing Machine Learning Models

Here the candidates need to demonstrate their expertise in using pre-built Machine Learning models as a service, including Machine Learning APIs (for instance, Speech API, Vision API, etc.), customizing Machine Learning APIs (for instance, Auto ML text, AutoML Vision, etc.), conversational experiences (for instance, Dialogflow). The applicants should also have the skills in deploying the Machine Learning pipeline. This involves the ability to ingest relevant data, perform retraining of machine learning models (BigQuery ML, Cloud Machine Learning Engine, Spark ML, Kubeflow), as well as execute continuous evaluation. Additionally, the students should be able to choose the relevant training & serving infrastructure as well as know how to fulfill measuring, monitoring, and troubleshooting of Machine Learning models.

QUESTION 97

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?
* Use a row key of the form <timestamp>.
* Use a row key of the form <sensorid>.
* Use a row key of the form <timestamp>#<sensorid>.
* Use a row key of the form >#<sensorid>#<timestamp>.

QUESTION 98

When you design a Google Cloud Bigtable schema it is recommended that you _________.
* Avoid schema designs that are based on NoSQL concepts
* Create schema designs that are based on a relational database design
* Avoid schema designs that require atomicity across rows
* Create schema designs that require atomicity across rows
Explanation

All operations are atomic at the row level. For example, if you update two rows in a table, it&#8217;s possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows.

Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

QUESTION 99

Which action can a Cloud Dataproc Viewer perform?
* Submit a job.
* Create a cluster.
* Delete a cluster.
* List the jobs.
A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.

Reference:

https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations

_summary

QUESTION 100

You are responsible for writing your company&#8217;s ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?
* PigLatin using Pig
* HiveQL using Hive
* Java using MapReduce
* Python using MapReduce
Pig is scripting language which can be used for checkpointing and splitting pipelines.

QUESTION 101

Which of the following is NOT one of the three main types of triggers that Dataflow supports?
* Trigger based on element size in bytes
* Trigger that is a combination of other triggers
* Trigger based on element count
* Trigger based on time
There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers. You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way

QUESTION 102

Which of these is NOT a way to customize the software on Dataproc cluster instances?
* Modify configuration files using cluster properties
* Set initialization actions
* Configure the cluster using Cloud Deployment Manager
* Log into the master node and make changes from there

QUESTION 103

Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the dat
* Which three machine learning applications can you use? (Choose three.)
* Supervised learning to determine which transactions are most likely to be fraudulent.
* Unsupervised learning to determine which transactions are most likely to be fraudulent.
* Clustering to divide the transactions into N categories based on feature similarity.
* Supervised learning to predict the location of a transaction.
* Reinforcement learning to predict the location of a transaction.
* Unsupervised learning to predict the location of a transaction.

QUESTION 104

Which SQL keyword can be used to reduce the number of columns processed by BigQuery?
* BETWEEN
* WHERE
* SELECT
* LIMIT
SELECT allows you to query specific columns rather than the whole table. LIMIT, BETWEEN, and WHERE clauses will not reduce the number of columns processed by BigQuery.

Reference: https://cloud.google.com/bigquery/launch-

checklist#architecture_design_and_development_checklist

QUESTION 105

Government regulations in the banking industry mandate the protection of client&#8217;s personally identifiable information (PII). Your company requires PII to be access controlled encrypted and compliant with major data protection standards In addition to using Cloud Data Loss Prevention (Cloud DIP) you want to follow Google-recommended practices and use service accounts to control access to PII. What should you do?
* Assign the required identity and Access Management (IAM) roles to every employee, and create a single service account to access

protect resources
* Use one service account to access a Cloud SQL database and use separate service accounts for each human user
* Use Cloud Storage to comply with major data protection standards. Use one service account shared by all users
* Use Cloud Storage to comply with major data protection standards. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group

QUESTION 106

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?
* Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
* Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
* Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
* Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

QUESTION 107

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the dat

a. Which two actions should you take? (Choose two.)
* Configure your Cloud Dataflow pipeline to use local execution
* Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions
* Increase the number of nodes in the Cloud Bigtable cluster
* Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable
* Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable
References:

QUESTION 108

Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?
* An hourly watermark
* An event time trigger
* The with Allowed Lateness method
* A processing time trigger
When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.

Processing time triggers. These triggers operate on the processing time &#8211; the time when the data element is processed at any given stage in the pipeline.

Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam&#8217;s default trigger is event time-based.

Reference: https://beam.apache.org/documentation/programming-guide/#triggers

QUESTION 109

Case Study: 3,

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 ?development/test, staging, and production ?

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.

Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?
* The zone
* The number of workers
* The disk size per worker
* The maximum number of workers

QUESTION 110

You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?
* Build and train a complex classification model with Spark MLlib to generate labels and filter the results.

Deploy the models using Cloud Dataproc. Call the model from your application.
* Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.
* Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user&#8217;s viewing history to generate preferences.
* Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user&#8217;s viewing history to generate preferences.

QUESTION 111

You are building a model to make clothing recommendations. You know a user&#8217;s fashion preference is 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.
Explanation

https://cloud.google.com/automl-tables/docs/prepare

QUESTION 112

You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?
* Create an API using App Engine to receive and send messages to the applications
* Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them
* Create a table on Cloud SQL, and insert and delete rows with the job information
* Create a table on Cloud Spanner, and insert and delete rows with the job information
Pubsub is used to transmit data in real time and scale automatically.

QUESTION 113

How can you get a neural network to learn about relationships between categories in a categorical feature?
* Create a multi-hot column
* Create a one-hot column
* Create a hash bucket
* Create an embedding column
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&#8217;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&#8217;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

Understanding functional and technical aspects of Google Professional Data Engineer Exam Building and operationalizing data processing systems

The following will be discussed here:

- Batch and streaming- Transformation- Lifecycle management of data- Adjusting pipelines- Building and operationalizing processing infrastructure- Testing and quality control- Effective use of managed services (Cloud Bigtable, Cloud Spanner, Cloud SQL, BigQuery, Cloud Storage, Cloud Datastore, Cloud Memorystore)- Migrating from on-premises to cloud (Data Transfer Service, Transfer Appliance, Cloud Networking)- Building and operationalizing data processing systems- Awareness of current state and how to migrate a design to a future state **Professional-Data-Engineer Exam Dumps Pass with Updated 2022 Certified Exam Questions:** https://www.validexam.com/Professional-Data-Engineer-latest-dumps.html]