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.passapply.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 work for an advertising company, and you\\'ve developed a Spark ML model to predict click-through rates at advertisement blocks. You\\'ve been developing everything at your on- premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models

B. Rewrite your models on TensorFlow, and start using Cloud ML Engine

C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

Correct Answer: C

https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml

QUESTION 2

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster ____.

A. application node

B. conditional node

C. master node

D. worker node

Correct Answer: C

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix--for example, if your cluster is named "my- cluster", the master-host-name would be "my-cluster-m". Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web- interfaces#interfaces

QUESTION 3

You need to look at BigQuery data from a specific table multiple times a day. The underlying table you are querying is several petabytes in size, but you want to filter your data and provide simple aggregations to downstream users. You want to run queries faster and get up-to-date insights quicker. What should you do?

A. Run a scheduled query to pull the necessary data at specific intervals daily.

B. Create a materialized view based off of the query being run.

C. Use a cached query to accelerate time to results.

D. Limit the query columns being pulled in the final result.

Correct Answer: B

Materialized views are precomputed views that periodically cache the results of a query for increased performance and efficiency. BigQuery leverages precomputed results from materialized views and whenever possible reads only changes from the base tables to compute up-to-date results. Materialized views can significantly improve the performance of workloads that have the characteristic of common and repeated queries. Materialized views can also optimize queries with high computation cost and small dataset results, such as filtering and aggregating large tables. Materialized views are refreshed automatically when the base tables change, so they always return fresh data. Materialized views can also be used by the BigQuery optimizer to process queries to the base tables, if any part of the query can be resolved by querying the materialized view. References: Introduction to materialized views Create materialized views BigQuery Materialized View Simplified: Steps to Create and 3 Best Practices Materialized view in Bigquery

QUESTION 4

You are migrating a large number of files from a public HTTPS endpoint to Cloud Storage. The files are protected from unauthorized access using signed URLs. You created a TSV file that contains the list of object URLs and started a transfer job by using Storage Transfer Service. You notice that the job has run for a long time and eventually failed Checking the logs of the transfer job reveals that the job was running fine until one point, and then it failed due to HTTP 403 errors on the remaining files You verified that there were no changes to the source system You need to fix the problem to resume the migration process. What should you do?

A. Set up Cloud Storage FUSE, and mount the Cloud Storage bucket on a Compute Engine Instance Remove the completed files from the TSV file Use a shell script to iterate through the TSV file and download the remaining URLs to the FUSE mount point.

B. Update the file checksums in the TSV file from using MD5 to SHA256. Remove the completed files from the TSV file and rerun the Storage Transfer Service job.

C. Renew the TLS certificate of the HTTPS endpoint Remove the completed files from the TSV file and rerun the Storage Transfer Service job.

D. Create a new TSV file for the remaining files by generating signed URLs with a longer validity period. Split the TSV file into multiple smaller files and submit them as separate Storage Transfer Service jobs in parallel.

Correct Answer: D

A signed URL is a URL that provides limited permission and time to access a resource on a web server. It is often used to grant temporary access to protected files without requiring authentication. Storage Transfer Service is a service that allows you to transfer data from external sources, such as HTTPS endpoints, to Cloud Storage buckets. You can use a TSV file to specify the list of URLs to transfer. In this scenario, the most likely cause of the HTTP 403 errors is that the signed URLs have expired before the transfer job could complete. This could happen if the signed URLs have a short validity period or the transfer job takes a long time due to the large number of files or network latency. To fix the problem, you need to create a new TSV file for the remaining files by generating new signed URLs with a longer validity period. This will ensure that the URLs do not expire before the transfer job finishes. You can use the Cloud Storage tools or your own program to generate signed URLs. Additionally, you can split the TSV file into multiple smaller files and submit them as separate Storage Transfer Service jobs in parallel. This will speed up the transfer process and reduce the risk of errors. References: Signed URLs | Cloud Storage Documentation V4 signing process with Cloud Storage tools V4 signing process with your own program Using a URL list file What Is a 403 Forbidden Error (and How Can I Fix It)?

QUESTION 5

Your company currently runs a large on-premises cluster using Spark Hive and Hadoop Distributed File System (HDFS)

in a colocation facility. The duster is designed to support peak usage on the system, however, many jobs are batch n nature, and usage of the cluster fluctuates quite dramatically. Your company is eager to move to the cloud to reduce the overhead associated with on- premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more servers offerings m order to take advantage of the cloud Because of the tuning of their contract renewal with the colocation facility they have only 2 months for their initial migration How should you recommend they approach thee upcoming migration strategy so they can maximize their cost savings in the cloud will still executing the migration in time?

A. Migrate the workloads to Dataproc plus HOPS, modernize later

B. Migrate the workloads to Dataproc plus Cloud Storage modernize later

C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery

D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery

Correct Answer: D