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.pass4itsure.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\\'ve migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you\\'d like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you\\'d like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.

What should you do?

A. Increase the size of your parquet files to ensure them to be 1 GB minimum.

B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.

C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.

D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Correct Answer: D

QUESTION 2

You\\'ve migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffling operations and initial data are parquet files (on average 200-400 MB size each).

You see some degradation in performance after the migration to Dataproc, so you\\'d like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you\\'d like to continue using Dataproc on preemptibles (with 2

non-preemptible workers only) for this workload.

What should you do?

A. Increase the size of your parquet files to ensure them to be 1 GB minimum.

B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.

C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.

D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Correct Answer: D

It\\'s the recommended best practice for this scenario. https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#switch_to_ssd_disks If you perform many shuffling operations or partitioned writes, switch to SSDs to boost performance.

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#use_preemptible_vms As

a default, preemptible VMs are created with a smaller boot disk size, and you might want to override this configuration if you are running shuffle-heavy workloads. For details, see the page on preemptible VMs in the Dataproc documentation.

Elimination Strategy:

A. Increase the size of your parquet files to ensure them to be 1 GB minimum (doesn\\'t make sense as the file size are fit for migration to proceed with given scenario, recommended size is between 128 MB to 1 GB.) B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files(doesn\\'t make sense to make changes to file format) C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS(doesn\\'t make sense to copy the file from GCS to HDFS as the workload that consists of many shuffling operations) D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size(perfect fit as the workload that consists of many shuffling operations which requires attention to increase the performance reference doc:- https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance)

QUESTION 3

If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?

A. Unsupervised learning

B. Regressor

C. Classifier

D. Clustering estimator

Correct Answer: B

Regression is the supervised learning task for modeling and predicting continuous, numeric variables. Examples include predicting real-estate prices, stock price movements, or student test scores.

Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.

Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset. Examples include customer segmentation, grouping similar items in e-commerce, and

social network analysis.

Reference: https://elitedatascience.com/machine-learning-algorithms

---

QUESTION 4

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?

A. 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.

B. 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.

C. 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\\'s viewing history to generate preferences.

D. 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\\'s viewing history to generate preferences.

Correct Answer: C

QUESTION 5

Which of these statements about BigQuery caching is true?

A. By default, a query\\'s results are not cached.

B. BigQuery caches query results for 48 hours.

C. Query results are cached even if you specify a destination table.

D. There is no charge for a query that retrieves its results from cache.

Correct Answer: D

When query results are retrieved from a cached results table, you are not charged for the query.

BigQuery caches query results for 24 hours, not 48 hours.

Query results are not cached if you specify a destination table.

A query\\'s results are always cached except under certain conditions, such as if you specify a destination table.

Reference: https://cloud.google.com/bigquery/querying-data#query-caching