Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 151 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 151
Topic #: 1
[All Professional Data Engineer Questions]

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 closing soon, so a rapid lift-and-shift migration is necessary. However, the data you've been using will be migrated to 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 Vertex AI for training existing Spark ML models

B. Rewrite your models on TensorFlow, and start using Vertex AI

C. Use 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

Show Suggested Answer

by ducc at *Sept. 3, 2022, 6:49 a.m.*

Comments

Type your comment...

Submit

vamgcp Highly Voted 1 year, 9 months ago

Selected Answer: C

Option C : It is the most rapid way to migrate your existing training pipelines to Google Cloud.
It allows you to continue using your existing Spark ML models.
It allows you to take advantage of the scalability and performance of Dataproc.
It allows you to read data directly from BigQuery, which is a more efficient way to process large datasets

upvoted 6 times

vaga1 Highly Voted 1 year, 12 months ago

Selected Answer: A

the question is: is it faster to move a SparkML job to a Vertex AI or to Dataproc? I am personally not sure, I would go for Dataproc as notebooks are not mentioned, but reading the Google article:

https://cloud.google.com/blog/topics/developers-practitioners/announcing-serverless-spark-components-vertex-ai-pipelines/

"Dataproc Serverless components for Vertex AI Pipelines that further simplify MLOps for Spark, Spark SQL, PySpark and Spark jobs."

upvoted 6 times

emmylou 1 year, 5 months ago

But you would need to re-write your models which can be a block

upvoted 3 times

Anudeep58 Most Recent 10 months ago

Selected Answer: C

Vertex AI is better suited for TensorFlow or scikit-learn models. Direct Spark ML support isn't native to Vertex AI, making this a less straightforward migration path.

upvoted 2 times

mothkuri 1 year, 2 months ago

C
Question is about rapid lift and shift. So code changes should be minimul

upvoted 1 times

GCP001 1 year, 3 months ago

Selected Answer: C

C looks more suitable as data is alerady on BigQuery.
Ref - https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml

upvoted 1 times

Matt_108 1 year, 3 months ago

Selected Answer: C

Option C, agreed with other comments

upvoted 1 times

MaxNRG 1 year, 4 months ago

Selected Answer: C

Use Cloud Dataproc, BigQuery, and Apache Spark ML for Machine Learning
https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml
Using Apache Spark with TensorFlow on Google Cloud Platform
https://cloud.google.com/blog/products/gcp/using-apache-spark-with-tensorflow-on-google-cloud-platform

upvoted 4 times

Nandababy 1 year, 4 months ago

Why not option D? To spin up the spark cluster on compute engine, considering rapid migration it potentially could be best approach as team wont have to re-work on model (may be only few configurational changes) and again to get data from Bigquery which is required periodically not all the time, could be easy.
With Dataproc it would have more code changes eventually can take more time.
With Vertex AI it doesn't support spark ML natively and also training would be black box.

For me Answer should be D.

upvoted 1 times

barnac1es 1 year, 7 months ago

Selected Answer: C

Dataproc for Spark: Google Cloud Dataproc is a managed Spark and Hadoop service that allows you to run Spark jobs seamlessly on Google Cloud. It provides the flexibility to run Spark jobs using Spark MLlib and other Spark libraries.

BigQuery Integration: You mentioned that your data is being migrated to BigQuery. Dataproc has native integration with

BigQuery, allowing you to read data directly from BigQuery tables. This eliminates the need to export data from BigQuery to another storage system before processing it with Spark.

Rapid Migration: This approach allows you to quickly migrate your existing Spark ML models and training pipelines without the need for a complete rewrite or extensive changes to your existing workflows. You can continue using your Spark ML models while adapting them to read data from BigQuery.

upvoted 2 times

DeepakVenkatachalam 1 year, 7 months ago

they are talking about rapid lift and shift, in which case Dataproc cluster will be right one for Spark ML models for lift and shift. so I think the answer is C.

upvoted 1 times

ckanaar 1 year, 7 months ago

Selected Answer: A

The updated answer seems A based on the following article:

https://cloud.google.com/blog/topics/developers-practitioners/announcing-serverless-spark-components-vertex-ai-pipelines/

upvoted 4 times

FP77 1 year, 8 months ago

Selected Answer: C

The answer is C. Spin up a Cloud Dataproc Cluster, migrate spark jobs to there, and link the Cluster to Bgquery with the connector. It's a straightforward solution.

upvoted 1 times

knith66 1 year, 9 months ago

Selected Answer: C

If you wanted to use Vertex AI for training Spark ML models, you would typically need to convert your Spark ML code to another supported machine learning framework like TensorFlow or scikit-learn. Then you could use Vertex AI's pre-built training and prediction services for those frameworks.

upvoted 3 times

wan2three 1 year, 9 months ago

Selected Answer: A

Through Vertex AI Workbench, Vertex AI is natively integrated with BigQuery, Dataproc, and Spark. You can use BigQuery ML to create and execute machine learning models in BigQuery using standard SQL queries on existing business intelligence tools and spreadsheets, or you can export datasets from BigQuery directly into Vertex AI Workbench and run your models from there.
https://cloud.google.com/vertex-ai#all-features:~:text=Data%20and%20AI%20integration

upvoted 1 times

blathul 1 year, 10 months ago

Selected Answer: C

Dataproc is a managed Spark and Hadoop service on Google Cloud, which makes it an ideal choice for migrating your existing Spark ML training pipelines. By using Dataproc, you can continue to leverage Spark and its ML capabilities without the need for significant code changes or rewriting your models.
By combining Dataproc and BigQuery, you can create Spark jobs or workflows in Dataproc that read data from BigQuery and train your existing Spark ML models. This approach allows you to quickly migrate your training pipelines to Google Cloud and take advantage of the scalability and performance benefits of both Dataproc and BigQuery.

upvoted 1 times

KC_go_reply 1 year, 10 months ago

Selected Answer: C

It is obviously C) Dataproc, since we don't want to rewrite the training from scratch, highly prefer Dataproc for anything Hadoop/Spark ecosystem, and Vertex AI doesn't support *training* with SparkML (but deploying existing models).

upvoted 4 times

Takshashila 1 year, 10 months ago

Selected Answer: C

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

upvoted 1 times

Load full discussion...