Google Discussions

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 142 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 142

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have built a model that is trained on data stored in Parquet files. You access the data through a Hive table hosted on Google Cloud. You preprocessed these data with PySpark and exported it as a CSV file into Cloud Storage. After preprocessing, you execute additional steps to train and evaluate your model. You want to parametrize this model training in Kubeflow Pipelines.

What should you do?

A. Remove the data transformation step from your pipeline.

B. Containerize the PySpark transformation step, and add it to your pipeline.

C. Add a ContainerOp to your pipeline that spins a Dataproc cluster, runs a transformation, and then saves the transformed data in Cloud Storage.

D. Deploy Apache Spark at a separate node pool in a Google Kubernetes Engine cluster. Add a ContainerOp to your pipeline that invokes a corresponding transformation job for this Spark instance.

Show Suggested Answer

by mil_spyro at *Dec. 13, 2022, 7:02 p.m.*

Comments

Type your comment...

Submit

mil_spyro Highly Voted 1 year, 10 months ago

Selected Answer: C

This will allow to reuse the same pipeline for different datasets without the need to manually preprocess and transform the data each time.

upvoted 7 times

tavva_prudhvi Highly Voted 1 year, 4 months ago

Selected Answer: C

Since the data is stored in Parquet format, it's more efficient to use Spark to transform it. Containerizing the PySpark transformation step and adding it to the pipeline may not be the optimal solution since it may require additional resources to run this container. Deploying Apache Spark at a separate node pool in a Google Kubernetes Engine cluster and adding a ContainerOp to invoke a corresponding transformation job for this Spark instance is also a possible solution, but it may require more setup and configuration.

Using Dataproc can simplify this process since it's a fully managed service that simplifies running Apache Spark and Hadoop clusters. A ContainerOp can be added to the pipeline to spin up a Dataproc cluster, run the transformation using PySpark, and save the transformed data in Cloud Storage. This solution is more efficient since Dataproc can scale the cluster based on the size of the data and the complexity of the transformation.

upvoted 6 times

momosoundz Most Recent 1 year, 4 months ago

Selected Answer: B

you can conteinerize the transformation and then save to google storage

upvoted 1 times

> **tavva_prudhvi** 1 year, 3 months ago
>
> it is not the most efficient and scalable solution when working with big data in the context of Google Cloud.
>
> upvoted 1 times

M25 1 year, 5 months ago

Selected Answer: C

https://kubeflow-pipelines.readthedocs.io/en/stable/source/kfp.dsl.html#kfp.dsl.ContainerOp
https://medium.com/@vignesh093/running-preprocessing-and-ml-workflow-in-kubeflow-with-google-dataproc-84103a9ef67e

upvoted 1 times

TNT87 1 year, 8 months ago

Selected Answer: C

C. Add a ContainerOp to your pipeline that spins a Dataproc cluster, runs a transformation, and then saves the transformed data in Cloud Storage.

The recommended approach to parametrize the model training in Kubeflow Pipelines would be to add a ContainerOp to the pipeline that spins up a Dataproc cluster, runs the PySpark transformation step, and saves the transformed data in Cloud Storage. This approach allows for easy integration of PySpark transformations with Kubeflow Pipelines while taking advantage of the scalability and efficiency of Dataproc.

upvoted 2 times

chidstar 1 year, 8 months ago

Selected Answer: B

All the wrong answers on this site really baffle me...correct answer is B... you must containerize your component for Kubeflow to run it.

https://www.kubeflow.org/docs/components/pipelines/v1/sdk/component-development/#containerize-your-components-code

upvoted 6 times

> **TNT87** 1 year, 8 months ago
>
> C. Add a ContainerOp to your pipeline that spins a Dataproc cluster, runs a transformation, and then saves the transformed data in Cloud Storage.
>
> The recommended approach to parametrize the model training in Kubeflow Pipelines would be to add a ContainerOp to the pipeline that spins up a Dataproc cluster, runs the PySpark transformation step, and saves the transformed data in Cloud Storage. This approach allows for easy integration of PySpark transformations with Kubeflow Pipelines while taking advantage of the scalability and efficiency of Dataproc.
>
> upvoted 3 times

TNT87 1 year, 10 months ago

Selected Answer: C

Selected Answer: C

Answer C

upvoted 2 times

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.