EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 219 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 219
Topic #: 1
[All Professional Machine Learning Engineer Questions]

Your company manages an ecommerce website. You developed an ML model that recommends additional products to users in near real time based on items currently in the user's cart. The workflow will include the following processes:

1. The website will send a Pub/Sub message with the relevant data and then receive a message with the prediction from Pub/Sub
2. Predictions will be stored in BigQuery
3. The model will be stored in a Cloud Storage bucket and will be updated frequently

You want to minimize prediction latency and the effort required to update the model. How should you reconfigure the architecture?

A. Write a Cloud Function that loads the model into memory for prediction. Configure the function to be triggered when messages are sent to Pub/Sub.

B. Create a pipeline in Vertex AI Pipelines that performs preprocessing, prediction, and postprocessing. Configure the pipeline to be triggered by a Cloud Function when messages are sent to Pub/Sub.

C. Expose the model as a Vertex AI endpoint. Write a custom DoFn in a Dataflow job that calls the endpoint for prediction.

D. Use the RunInference API with WatchFilePattern in a Dataflow job that wraps around the model and serves predictions.

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 6:16 a.m.*

Comments

Type your comment...

Submit

guilhermebutzke Highly Voted 8 months, 1 week ago

Selected Answer: D

My answer: D

This Google Documentation explains “Instead of deploying the model to an endpoint, you can use the RunInference API to serve machine learning models in your Apache Beam pipeline. This approach has several advantages, including flexibility and portability.”
https://cloud.google.com/blog/products/ai-machine-learning/streaming-prediction-with-dataflow-and-vertex

This documentation uses RunInference and WatchFilePattern to “to automatically update the ML model without stopping the Apache Beam”.
https://cloud.google.com/dataflow/docs/notebooks/automatic_model_refresh

So, thinking in “minimize prediction latency”, its suggested use RunInfenrece, while “effort required to update the model” the **WatchFilePattern is the best approach.** I think D is the best option

upvoted 5 times

PhilipKoku Most Recent 4 months, 2 weeks ago

Selected Answer: C

C) Expose the model as Vertex AI End Point

upvoted 1 times

pinimichele01 6 months, 1 week ago

Selected Answer: D

agree with guilhermebutzke

upvoted 1 times

Yan_X 7 months, 2 weeks ago

Selected Answer: A

A for me.

upvoted 1 times

ddogg 8 months, 2 weeks ago

Selected Answer: D

Automatic Model Updates: WatchFilePattern automatically detects model changes in Cloud Storage, leading to seamless updates without managing endpoint deployments.

upvoted 3 times

pikachu007 9 months, 1 week ago

Selected Answer: A

Low Latency:

Serverless Execution: Cloud Functions start up almost instantly, reducing prediction latency compared to alternatives that require longer setup or deployment times.
In-Memory Model: Loading the model into memory eliminates disk I/O overhead, further contributing to rapid predictions.

upvoted 2 times

CHARLIE2108 8 months, 2 weeks ago

Cloud Functions offer low latency but it might not scale well.

upvoted 2 times

Start Learning for free

Social Media

Facebook , Twitter
YouTube , Reddit
Pinterest

We are the biggest and most updated IT certification exam material website.

Using our own resources, we strive to strengthen the IT professionals community for free.