EXAMTOPICS

- Expert Verified, Online, **Free.**

☰ MENU   🔍

⬅ Google Discussions

Exam Professional Machine Learning Engineer All Questions
View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

📄 **EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 143 DISCUSSI...**

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 143
Topic #: 1

[All Professional Machine Learning Engineer Questions]

You have developed an ML model to detect the sentiment of users' posts on your company's social media page to identify outages or bugs. You are using Dataflow to provide real-time predictions on data ingested from Pub/Sub. You plan to have multiple training iterations for your model and keep the latest two versions live after every run. You want to split the traffic between the versions in an 80:20 ratio, with the newest model getting the majority of the traffic. You want to keep the pipeline as simple as possible, with minimal management required. What should you do?

   A. Deploy the models to a Vertex AI endpoint using the traffic-split=0=80, PREVIOUS_MODEL_ID=20 configuration.

   B. Wrap the models inside an App Engine application using the --splits PREVIOUS_VERSION=0.2, NEW_VERSION=0.8 configuration

   C. Wrap the models inside a Cloud Run container using the REVISION1=20, REVISION2=80 revision configuration.

   D. Implement random splitting in Dataflow using beam.Partition() with a partition function calling a Vertex AI endpoint.

Show Suggested Answer

by 👤 **FherRO** at *Feb. 22, 2023, 11:41 p.m.*

Comments

Type your comment...

TNT87 `Highly Voted 👍` 1 year, 8 months ago

`Selected Answer: A`

A. Deploy the models to a Vertex AI endpoint using the traffic-split=0=80, PREVIOUS_MODEL_ID=20 configuration.

The recommended approach to achieve the desired outcome would be to deploy the ML models to a Vertex AI endpoint and configure the traffic splitting using the traffic-split parameter. The traffic-split parameter enables you to split traffic between multiple versions of a model based on a percentage split. In this case, the newest model should receive the majority of the traffic, which can be achieved by setting the traffic-split parameter to 0=80. The previous version of the model should receive the remaining 20% of the traffic, which can be achieved by setting the PREVIOUS_MODEL_ID parameter to 20.

👍 ↩ ⚑ upvoted 8 times

fitri001 `Most Recent ⊙` 6 months, 2 weeks ago

`Selected Answer: A`

Vertex AI Traffic Splitting: Vertex AI natively supports traffic splitting between deployed models through the traffic-split parameter. This allows you to specify the desired traffic distribution (80% to the newest model, 20% to the previous one) during deployment.

👍 ↩ ⚑ upvoted 1 times

M25 1 year, 5 months ago

`Selected Answer: A`

Went with A

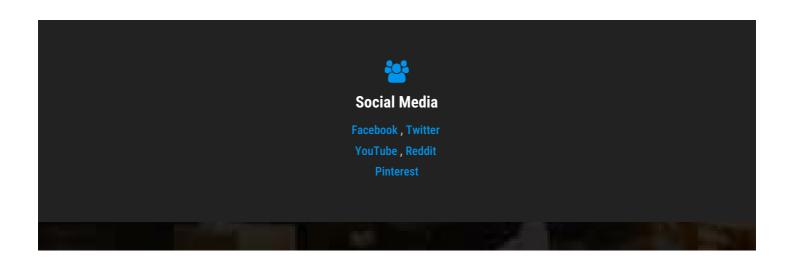👍 ↩ ⚑ upvoted 2 times

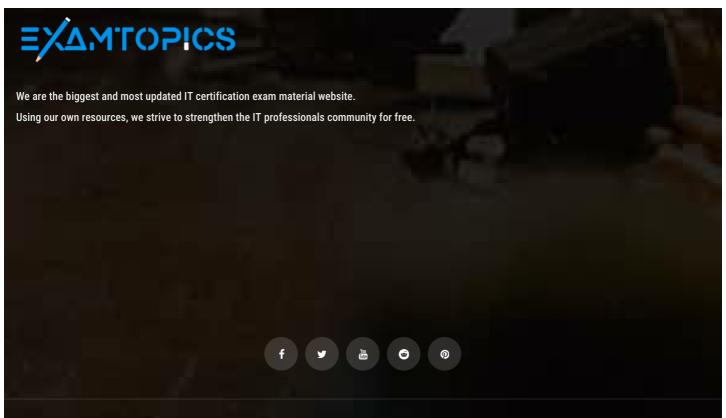**FherRO** 1 year, 8 months ago

`Selected Answer: A`

I think is A because traffic can be split across different versions when using Endpoints https://cloud.google.com/vertex-ai/docs/general/deployment#models-endpoint.
The --trafic-split flag does exist, but in the question the syntax is incorrect, it should be "--traffic-split = [MODEL_ID_1=value, MODEL_ID_2=value]" as explained in https://cloud.google.com/sdk/gcloud/reference/ai/endpoints/deploy-model

👍 ↩ ⚑ upvoted 4 times

EXAMTOPICS

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.