Welcome to ExamTopics

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 170 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 170

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You need to deploy a scikit-leam classification model to production. The model must be able to serve requests 24/7, and you expect millions of requests per second to the production application from 8 am to 7 pm. You need to minimize the cost of deployment. What should you do?

A. Deploy an online Vertex AI prediction endpoint. Set the max replica count to 1

B. Deploy an online Vertex AI prediction endpoint. Set the max replica count to 100

C. Deploy an online Vertex AI prediction endpoint with one GPU per replica. Set the max replica count to 1

D. Deploy an online Vertex AI prediction endpoint with one GPU per replica. Set the max replica count to 100

Show Suggested Answer

by b1a8fae at *Jan. 8, 2024, 4:33 p.m.*

Comments

Type your comment...

Submit

⊟ AzureDP900 4 months ago

Option A (Deploying an online Vertex AI prediction endpoint. Set the max replica count to 1) is still a good choice for minimizing costs. By setting the max replica count to 1, you are allowing Vertex AI to scale up or down based on load, which means that during off-peak hours, you won't be paying for unnecessary instances.

upvoted 1 times

pinimichele01 6 months, 1 week ago

Selected Answer: B

see pikachu007

upvoted 1 times

36bdc1e 9 months, 1 week ago

B
we don't need GPU for scikit-learn

upvoted 2 times

BlehMaks 9 months, 2 weeks ago

Selected Answer: B

scikit-learn doesn't support GPU
https://scikit-learn.org/stable/faq.html#will-you-add-gpu-support

upvoted 2 times

pikachu007 9 months, 2 weeks ago

Selected Answer: B

B. Deploy an online Vertex AI prediction endpoint. Set the max replica count to 100:
This option provides a higher number of replicas (100) to handle the expected high volume of requests during peak hours. While it might result in increased costs, it provides the necessary scalability to manage the incoming traffic efficiently. During non-peak hours, you can consider scaling down the replicas to reduce costs, as Vertex AI allows dynamic scaling based on demand.

upvoted 4 times

b1a8fae 9 months, 2 weeks ago

B.
scikit-learn -> no need for GPU
max number of replicas -> 1 is too little if we are serving online predictions at such a massive scale (millions per second)

upvoted 1 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.

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

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.