



- 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 164 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 164

Topic #: 1

[\[All Professional Machine Learning Engineer Questions\]](#)

You work for a small company that has deployed an ML model with autoscaling on Vertex AI to serve online predictions in a production environment. The current model receives about 20 prediction requests per hour with an average response time of one second. You have retrained the same model on a new batch of data, and now you are canary testing it, sending ~10% of production traffic to the new model. During this canary test, you notice that prediction requests for your new model are taking between 30 and 180 seconds to complete. What should you do?

- A. Submit a request to raise your project quota to ensure that multiple prediction services can run concurrently.
- B. Turn off auto-scaling for the online prediction service of your new model. Use manual scaling with one node always available.
- C. Remove your new model from the production environment. Compare the new model and existing model codes to identify the cause of the performance bottleneck.
- D. Remove your new model from the production environment. For a short trial period, send all incoming prediction requests to BigQuery. Request batch predictions from your new model, and then use the Data Labeling Service to validate your model's performance before promoting it to production.

Show Suggested Answer

by [kalle_balle](#) at Jan. 7, 2024, 3:56 a.m.

Comments

Type your comment...

Submit

  **YushiSato** 2 months, 1 week ago

I don't see B as the right answer.
The Vertex AI Endpoint cannot scale to 0 for newer version of the model.

> When you configure a DeployedModel, you must set `dedicatedResources.minReplicaCount` to at least 1. In other words, you cannot configure the DeployedModel to scale to 0 prediction nodes when it is unused.



<https://cloud.google.com/vertex-ai/docs/general/deployment#scaling>

   upvoted 1 times

  **YushiSato** 2 months, 1 week ago

I was convinced that the machines that are autoscaled by the Vertex AI Endpoint seem to be tied to the endpoint, not the model in which they are deployed.

   upvoted 1 times

  **AnnaR** 5 months, 4 weeks ago

Selected Answer: B


B can be effective in controlling resources available to the new model, ensuring that it is not delayed by the autoscaling trying to scale up from 0.

Not A: there is no indication in the description that quota limits cause the slowdown and does not address issue where new model is performing poorly on canary testing.

Not C : when you pull the new model from prod environment, you could affect end-user experience

Not D: Same as C plus you rely on batch predictions which does not align with the need for online, real-time predictions in the prod environment. Data Labeling Service is more about assessing accuracy and less about resolving latency issues.

   upvoted 2 times

  **pinimichele01** 6 months, 2 weeks ago

Selected Answer: B


You have retrained the same model on a new batch of data

   upvoted 1 times

  **pinimichele01** 6 months, 1 week ago

the new model has too few requests per hour and therefore scales down to 0. Which means it has to create the an instance every time it serves a request, and this takes time.

By manually setting the number of nodes, the nodes will always be running, whether or not they are serving predictions

   upvoted 3 times

  **VipinSingla** 7 months ago

Selected Answer: B

bottleneck seems to be start of node as there are very low number of requests so having one node always available will help in this case.

   upvoted 1 times

  **Aastha_Vashist** 7 months ago

Selected Answer: C

went with c

   upvoted 1 times

  **Carlose2108** 7 months, 3 weeks ago

Selected Answer: C

I went C.
Diagnosing the root cause.

   upvoted 1 times

  **guilhermebutzke** 8 months, 2 weeks ago

Selected Answer: C

Choose C.

The significant increase in response time from 1 second to between 30 and 180 seconds indicates a performance issue with the new model. Before making any further changes or decisions, it's crucial to identify the root cause of this performance bottleneck. By comparing the code of the new model with the existing model, you can pinpoint any differences that might be

bottomline. By comparing the code of the new model with the existing model, you can pinpoint any differences that might be causing the slowdown.

In A, This may not be the root cause and could incur unnecessary costs without addressing the performance issue. In B, it doesn't address the underlying issue causing the significant increase in response time observed during canary testing. In D, This would significantly increase latency and hinder real-time predictions, negatively impacting user experience.

   upvoted 1 times

  **vaibavi** 8 months, 1 week ago

But in the question it says "You have retrained the same model on a new batch of data" it's just the data that changed so no need to check for the code check.

   upvoted 2 times

  **sonicclasps** 8 months, 3 weeks ago

Selected Answer: B

sounds to me that the new model has too few requests per hour and therefore scales downs to 0. Which means it has to create the an instance every time it serves a request, and this takes time.

By manually setting the number of nodes, the nodes will always be running, whether or not they are serving predictions

   upvoted 4 times

  **b1a8fae** 9 months, 2 weeks ago

Unsure on this one, but I would go with A.


B. Turning off auto-scaling is a good measure when dealing with datasets with steep spikes of requests traffic (here we are dealing with avg. 20 request per hour) "The service may not be able to bring nodes online fast enough to keep up with large spikes of request traffic." <https://cloud.google.com/blog/products/ai-machine-learning/scaling-machine-learning-predictions>

C. You retrain the SAME model on a different batch of data. It is implied that the code is the same too?

D. Actual quality of the model is not in question here, but rather the long prediction time per request.


Even if the requests traffic is very low, I can only consider option A: the selected quota cannot deal with the amount of concurrent prediction requests.

   upvoted 1 times

  **kalle_balle** 9 months, 2 weeks ago

Selected Answer: C

Option B or D is completely wrong. Option A to raise the quota might be necessary in some situations but doesn't necessarily deal with the performance issue at the test. Option C seems like the most suitable option.

   upvoted 1 times

  **edoo** 7 months, 2 weeks ago

You only retrained the same model, your code hasn't changed, you won't find anything with C.
It's B.

   upvoted 1 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.



© 2024 ExamTopics

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.