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 187 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 187

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You recently deployed a scikit-learn model to a Vertex AI endpoint. You are now testing the model on live production traffic. While monitoring the endpoint, you discover twice as many requests per hour than expected throughout the day. You want the endpoint to efficiently scale when the demand increases in the future to prevent users from experiencing high latency. What should you do?

A. Deploy two models to the same endpoint, and distribute requests among them evenly

B. Configure an appropriate minReplicaCount value based on expected baseline traffic

C. Set the target utilization percentage in the autoscailngMetricSpecs configuration to a higher value

D. Change the model’s machine type to one that utilizes GPUs

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 3:10 a.m.*

Comments

Type your comment...

Submit

fitri001 6 months, 2 weeks ago

Selected Answer: B

Autoscaling based on baseline: Vertex AI endpoints have built-in autoscaling capabilities. Setting a minReplicaCount ensures there are always at least that many replicas running, handling the baseline traffic efficiently. When demand increases above the baseline, autoscaling will automatically provision additional replicas to maintain performance.
Efficient scaling: This approach allows the endpoint to scale up smoothly as traffic increases, preventing sudden spikes in latency for users.
Targeted resource allocation: Unlike option A (deploying multiple models), this method avoids redundant resources when traffic is low. Additionally, option D (switching to GPUs) might be unnecessary if the bottleneck isn't processing power.

upvoted 4 times

fitri001 6 months, 2 weeks ago

A. Deploying multiple models: This creates additional overhead and resource usage without directly addressing autoscaling. Traffic distribution may also not be perfectly even.
C. Increasing target utilization: Raising the target utilization could lead to under-provisioning during peak hours, causing latency issues. It's better to set a baseline with minReplicaCount and let autoscaling handle peak loads.
D. Switching to GPUs: While GPUs can be beneficial for computationally intensive models, it might be an unnecessary expense if the current model doesn't heavily utilize the CPU. Analyze the CPU usage before switching to a GPU-based machine type.

upvoted 1 times

guilhermebutzke 8 months, 3 weeks ago

Selected Answer: B

My Answer B

The letter C would be the correct answer if the target were set lower to anticipate traffic spikes, not set higher as the answer says. However, considering that the minReplicaCount is now twice the known value, letter B is the most appropriate answer as it suggests considering setting a new minReplicaCount, which could be the best choice.

upvoted 2 times

Yan_X 9 months, 1 week ago

Selected Answer: B

Not C, if set to a higher value, it is less easier to autoscale to another instance, as it will wait the utilisation to a even higher value.

upvoted 4 times

b1a8fae 9 months, 3 weeks ago

Selected Answer: C

I go with C. It calculates the number of replicas based on CPU utilization.
https://cloud.google.com/python/docs/reference/aiplatform/latest/google.cloud.aiplatform_v1.types.AutoscalingMetricSpec

upvoted 1 times

36bdc1e 9 months, 3 weeks ago

B
This
option allows you to leverage the power and simplicity of Vertex AI to automatically scale your endpoint resources according to the traffic patterns.

upvoted 2 times

pikachu007 9 months, 3 weeks ago

Selected Answer: C

c as it is dynamic

upvoted 1 times

sonicclasps 9 months, 1 week ago

yes it's dynamic, but the target should be set lower, not higher, if you want to anticipate traffic spikes.

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.

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.