EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 218 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 218
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have built a custom model that performs several memory-intensive preprocessing tasks before it makes a prediction. You deployed the model to a Vertex AI endpoint, and validated that results were received in a reasonable amount of time. After routing user traffic to the endpoint, you discover that the endpoint does not autoscale as expected when receiving multiple requests. What should you do?

A. Use a machine type with more memory

B. Decrease the number of workers per machine

C. Increase the CPU utilization target in the autoscaling configurations.

D. Decrease the CPU utilization target in the autoscaling configurations

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 6:08 a.m.*

Comments

Type your comment...

Submit

b1a8fae Highly Voted 9 months, 1 week ago

Selected Answer: D

D.
The idea behind this question is getting autoscaling to handle well the fluctuating input of requests. Changing the machine (A) is not related to autoscaling, and you might not be using the full potential of the machine during the whole time, bur rather only during instances of peak traffic. You need to lower the autoscaling threshold (the target utilization metric mentioned in the options is CPU, so we will go with this) so you make use of more resources whenever too many memory-intensive requests are happening.

https://cloud.google.com/compute/docs/autoscaler/scaling-cpu#scaling_based_on_cpu_utilization
https://cloud.google.com/compute/docs/autoscaler#autoscaling_policy

upvoted 9 times

b1a8fae 9 months, 1 week ago

Addition: although memory-intensive is not directly related to CPU, for me the key is "the model does not autoscale as expected". To me this is addressing directly the settings of autoscaling, which won't change by changing the machine.

upvoted 2 times

pikachu007 Highly Voted 9 months, 1 week ago

Selected Answer: A

B. Decreasing Workers: This might reduce memory usage per machine but could also decrease overall throughput, potentially impacting performance.
C. Increasing CPU Utilization Target: This wouldn't directly address the memory bottleneck and could trigger unnecessary scaling based on CPU usage, not memory requirements.
D. Decreasing CPU Utilization Target: This could lead to premature scaling, potentially increasing costs without addressing the root cause.

upvoted 5 times

VinaoSilva Most Recent 3 months, 3 weeks ago

Selected Answer: D

"use autoscale" = deacrease cpu utilization target

upvoted 1 times

fitri001 6 months, 1 week ago

Selected Answer: D

D. Decrease the CPU utilization target: This is the most suitable approach. By lowering the CPU utilization target, the endpoint will scale up at a lower CPU usage level. This increases the likelihood of scaling up when the memory-intensive preprocessing tasks cause a rise in CPU utilization, even though memory is the root cause.

upvoted 2 times

fitri001 6 months, 1 week ago

A. Use a machine type with more memory: While this might seem logical, autoscaling in Vertex AI endpoints relies on CPU utilization as the metric, not directly on memory usage. Even with more memory, the endpoint might not scale up if CPU utilization remains below the threshold.
B. Decrease the number of workers per machine (Not applicable to Vertex AI Endpoints): This option might be relevant for some serving frameworks, but Vertex AI Endpoints don't typically use a worker concept. Scaling down workers wouldn't directly address the memory bottleneck.
C. Increase the CPU utilization target: This would instruct the endpoint to scale up only when CPU usage reaches a higher threshold. Since the issue is memory usage, increasing the CPU target wouldn't trigger scaling when memory is the limiting factor.

upvoted 1 times

guilhermebutzke 8 months, 1 week ago

Selected Answer: D

Option D, "Decrease the CPU utilization target in the autoscaling configurations," could be a valid approach to address the issue of autoscaling and anticipate spikes in traffic. By lowering the threshold, the autoscaling system would initiate scaling actions at a lower CPU utilization level, allowing for a more proactive response to increasing demands.

upvoted 3 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.