EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 238 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 238
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You have deployed a scikit-team model to a Vertex AI endpoint using a custom model server. You enabled autoscaling: however, the deployed model fails to scale beyond one replica, which led to dropped requests. You notice that CPU utilization remains low even during periods of high load. What should you do?

A. Attach a GPU to the prediction nodes

B. Increase the number of workers in your model server

C. Schedule scaling of the nodes to match expected demand

D. Increase the minReplicaCount in your DeployedModel configuration

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 8:25 a.m.*

Comments

Type your comment...

Submit

sonicclasps Highly Voted 8 months, 3 weeks ago

Selected Answer: A

"We generally recommend starting with one worker or thread per core. If you notice that CPU utilization is low, especially under high load, or your model is not scaling up because CPU utilization is low, then increase the number of workers."
https://cloud.google.com/vertex-ai/docs/general/deployment

upvoted 6 times

sonicclasps 8 months, 3 weeks ago

sorry clicked wrong, answer is B

upvoted 2 times

fitri001 Most Recent 6 months, 1 week ago

Selected Answer: B

agree with sonicclasps -> B

upvoted 1 times

pinimichele01 6 months, 1 week ago

Selected Answer: B

agree with sonicclasps -> B

upvoted 1 times

pinimichele01 6 months ago

NOT D: This might help ensure at least one replica is always available, but it won't address the issue of not scaling up during high load.

upvoted 1 times

Carlose2108 7 months, 4 weeks ago

Selected Answer: B

I went B

upvoted 2 times

guilhermebutzke 8 months, 1 week ago

Selected Answer: C

My answer: C

The problem is in scale. The provided resources areok. So,

A: Not correct, because CPU is enough.

B: Not correct, because increasing the number of workers will accelerate the process in a single replica, and make the time of prediction faster for example, but not will happen in scale problem.

C:Correct: This option involves adjusting the scaling of resources to match the expected demand, ensuring that the system can handle increased loads effectively

D: This might help ensure at least one replica is always available, but it won't address the issue of not scaling up during high load.

upvoted 1 times

pikachu007 9 months, 1 week ago

Selected Answer: B

Low CPU Utilization: Despite high load, low CPU utilization indicates underutilization of available resources, suggesting a bottleneck within the model server itself, not overall compute capacity.
Worker Concurrency: Increasing the number of workers within the model server allows it to handle more concurrent requests, effectively utilizing available CPU resources and addressing the bottleneck.

upvoted 3 times

BlehMaks 9 months, 1 week ago

i don't get it. The autoscaling system should increase/decrease the number of workers itself. if we do it instead of the autoscaling system, why do we need it?

upvoted 1 times

guilhermebutzke 8 months, 1 week ago

Increase the number of workers within the model server will distribute the load within the single replica, but it wouldn't address the problem of not scaling beyond one replica. Increasin worker will be a good option for delay in prediction.

upvoted 1 times

asmgi 3 months, 1 week ago

Not scaling beyond one replica is symptom and not the source of the problem. The problem is low CPU utilization.

upvoted 1 times