EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 256 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 256
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for an online grocery store. You recently developed a custom ML model that recommends a recipe when a user arrives at the website. You chose the machine type on the Vertex AI endpoint to optimize costs by using the queries per second (QPS) that the model can serve, and you deployed it on a single machine with 8 vCPUs and no accelerators.

A holiday season is approaching and you anticipate four times more traffic during this time than the typical daily traffic. You need to ensure that the model can scale efficiently to the increased demand. What should you do?

A. 1. Maintain the same machine type on the endpoint.
2. Set up a monitoring job and an alert for CPU usage.
3. If you receive an alert, add a compute node to the endpoint.

B. 1. Change the machine type on the endpoint to have 32 vCPUs.
2. Set up a monitoring job and an alert for CPU usage.
3. If you receive an alert, scale the vCPUs further as needed.

C. 1. Maintain the same machine type on the endpoint Configure the endpoint to enable autoscaling based on vCPU usage.
2. Set up a monitoring job and an alert for CPU usage.
3. If you receive an alert, investigate the cause.

D. 1. Change the machine type on the endpoint to have a GPU. Configure the endpoint to enable autoscaling based on the GPU usage.
2. Set up a monitoring job and an alert for GPU usage.
3. If you receive an alert, investigate the cause.

Show Suggested Answer

by kalle_balle at *Jan. 9, 2024, 11:07 p.m.*

Comments

Type your comment...

Submit

fitri001 Highly Voted 6 months, 1 week ago

Selected Answer: C

Option A: Manually adding compute nodes after an alert might lead to delays and potential outages during peak traffic.
Option B: Upgrading to 32 vCPUs upfront might be an overkill if the current machine type with 8 vCPUs can handle the typical daily traffic. Vertical scaling (more vCPUs) might be suitable only if the model can benefit from additional CPU power.
Option D: Using a GPU is unlikely to benefit a recipe recommendation model, which likely doesn't involve intensive graphical processing. Additionally, monitoring GPU usage wouldn't be relevant.

upvoted 5 times

AzureDP900 Most Recent 3 months, 2 weeks ago

C is right because
1)Since you've already optimized your model's deployment on a single machine with 8 vCPUs, it makes sense to maintain the same machine type to avoid any potential performance issues.
2)Enabling autoscaling based on vCPU usage will allow your endpoint to automatically add more machines as needed to handle the increased traffic during the holiday season. This approach is more efficient and cost-effective than scaling up individual machines or adding new machines manually.
3)Monitoring CPU usage with a job and alerting when thresholds are exceeded allows you to detect potential issues before they impact performance.

upvoted 1 times

omermahgoub 6 months, 1 week ago

Selected Answer: C

C: Use Autoscaling Based on vCPU Usage

upvoted 1 times

emsherff 6 months, 2 weeks ago

Selected Answer: C

Autoscaling based on vCPU usage aligns well with the workload.

upvoted 1 times

emsherff 6 months, 2 weeks ago

Option A is manual intervention
Option B is overprovisioning preemptively, which is an overkill (autoscaling should be preferred)
Option D - Unless the recipe recommendation model uses GPU-accelerated computations (e.g., some deep learning models), adding a GPU won't be beneficial and will increase costs.
I would go with C - Autoscaling based on vCPU usage which aligns well with the workload.

upvoted 2 times

daidai75 9 months ago

Selected Answer: C

Option B can only support exact 4x times traffic, but the requirement is four times "more", so B is not the best at least for me.

upvoted 1 times

b1a8fae 9 months ago

Selected Answer: C

I would go for C as it enables autoscaling when exceeding a determined CPU usage threshold.

upvoted 1 times

pikachu007 9 months, 1 week ago

Selected Answer: C

Cost Optimization: It starts with the current machine type, avoiding unnecessary upfront costs, and scales only when needed.
Autoscaling: It automatically adjusts compute resources based on vCPU usage, ensuring the endpoint can handle traffic spikes without manual intervention.

Monitoring and Alerting: It provides visibility into resource usage and triggers alerts for potential issues, enabling proactive actions.
Investigation: It encourages investigation of alerts to identify any underlying problems beyond expected traffic growth, ensuring overall system health.

upvoted 1 times

kalle_balle 9 months, 2 weeks ago

Selected Answer: B

Voting for B as it's the only option to autoscale even though the cost will go up. All other options include manual intervention.

upvoted 1 times

b1a8fae 9 months ago

Wouldn't scaling up the vCPUs after receiving the alert also be manual? It comes across as such to me at least.

upvoted 1 times