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 277 DISCUSSI…

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 277

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a large bank that serves customers through an application hosted in Google Cloud that is running in the US and Singapore. You have developed a PyTorch model to classify transactions as potentially fraudulent or not. The model is a three-layer perceptron that uses both numerical and categorical features as input, and hashing happens within the model.

You deployed the model to the us-central1 region on nl-highcpu-16 machines, and predictions are served in real time. The model's current median response latency is 40 ms. You want to reduce latency, especially in Singapore, where some customers are experiencing the longest delays. What should you do?

A. Attach an NVIDIA T4 GPU to the machines being used for online inference.

B. Change the machines being used for online inference to nl-highcpu-32.

C. Deploy the model to Vertex AI private endpoints in the us-central1 and asia-southeast1 regions, and allow the application to choose the appropriate endpoint.

D. Create another Vertex AI endpoint in the asia-southeast1 region, and allow the application to choose the appropriate endpoint.

Show Suggested Answer

by guilhermebutzke at *Feb. 19, 2024, 3:05 p.m.*

Comments

Comments

Type your comment...

Submit

guilhermebutzke Highly Voted 8 months ago

Selected Answer: C

My Answer: C
The bottleneck is network latency. So,
A: Not Correct: might improve performance, but it's an expensive solution and may not be necessary if the bottleneck is network latency.
B: Not Correct: might offer slight improvement, but the primary issue is geographical distance between users and the model.
C: CORRECT: This approach leverages the geographical proximity of the endpoints to the users, reducing latency for customers in Singapore without neglecting customers in the US. Additionally, using Vertex AI private endpoints ensures secure and efficient communication between the application and the model.
D: Not Correct: it's not the most efficient approach because it does not utilize the existing infrastructure in the us-central1 region, and managing multiple endpoints might introduce additional complexity.

upvoted 6 times

> **tavva_prudhvi** 6 months, 3 weeks ago
>
> Deploying in additional regions (D) does not necessarily negate or underutilize existing deployments but rather complements them to provide a better global service.
>
> upvoted 3 times

wences Most Recent 1 month, 1 week ago

Selected Answer: D

I don't have any link to support this other than a simple analysis; if you want the data or process to be low latency, you need to deploy closes where it is required, in this case, to Singapore customers, which reduces latetency addressing the requirement.

upvoted 1 times

inc_dev_ml_001 4 months, 2 weeks ago

Selected Answer: D

I think it's D because C and D should work in the same way, but ensuring the connection through a private endpoint it's not necessary because in the question there's nothing about security or sensitive informations. So the scope for a generic endpoint is "Accessible from anywhere", the scope for a private endpoint is "Accessible only within VPC or private connections". Don't see why to do that, it's only a matter of latency, not a matter of safety.

upvoted 2 times

GuineaPigHunter 5 months, 1 week ago

Selected Answer: D

Not sure why I'd choose C over D, my choice is D.
Model is already deployed to us-central1 so now it's only a matter of deploying it to asia-southeast1 and letting the app choose the closer endpoint.
Why the need for private endpoints and what will happen with the current already deployed model in us-central1?

upvoted 2 times

omermahgoub 6 months, 1 week ago

Selected Answer: C

Deploying the model to a Vertex AI private endpoint in the Singapore region brings the model closer to users in that region. This significantly reduces network latency for those users compared to accessing the model hosted in us-central1.
Allowing the application to choose the appropriate endpoint based on user location (through private endpoints) ensures users access the geographically closest model replica, optimizing latency.
Why not D: creating a separate endpoint in Singapore would allow regional deployment, it wouldn't automatically route users to the closest endpoint. You still need additional logic within the application for regional routing, increasing complexity.

upvoted 1 times

tavva_prudhvi 6 months, 3 weeks ago

Selected Answer: D

By having an endpoint in the asia-southeast1 region (Singapore), the data doesn't have to travel as far, significantly reducing the round-trip time. Allowing the application to choose the appropriate endpoint based on the user's location ensures that requests are handled by the nearest available server, optimizing response times for users in different regions.

upvoted 2 times

shuvs 6 months, 3 weeks ago

Selected Answer: D

I think it is D. C is questionable as why do you need a private endpoint?

I think it is D. C is questionable as why do you need a private endpoint?

upvoted 1 times

AzureDP900 3 months, 2 weeks ago

Yes, using private endpoints does introduce some overhead.
Additional latency: Establishing a connection to a private endpoint may add some latency compared to using the public endpoint.
Increased complexity: Managing private endpoints requires additional configuration and management, which can increase the overall complexity of your deployment.
However, in this scenario, the benefits of using private endpoints (security, control, and isolation) outweigh the potential overhead. The goal is to reduce latency for users in Singapore, and by deploying a private endpoint closer to them, you can achieve this while maintaining security and control over access to your model.

upvoted 1 times

AzureDP900 3 months, 2 weeks ago

I will go with C.
In this scenario, deploying the model to Vertex AI private endpoints in both us-central1 and asia-southeast1 regions is necessary because:

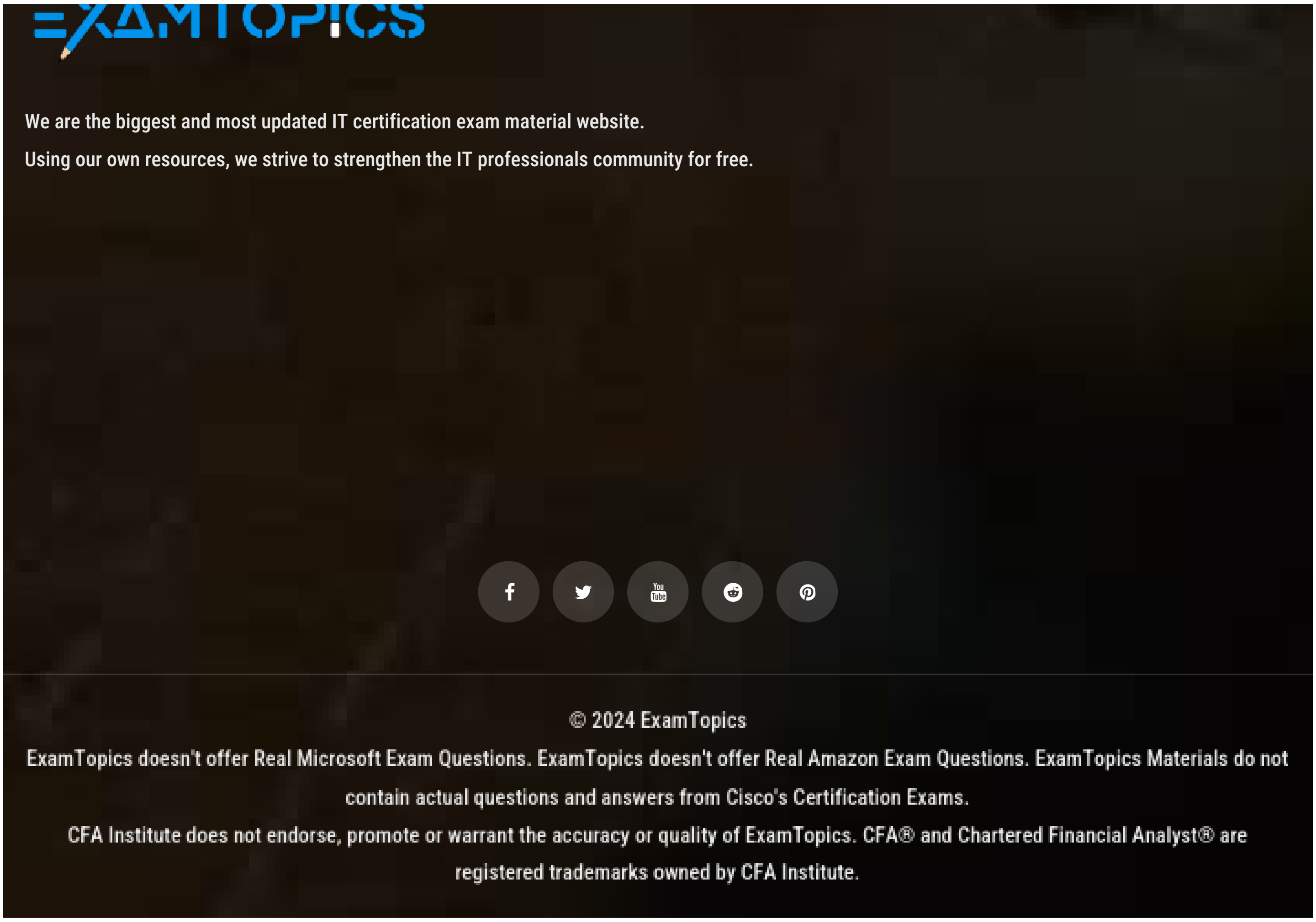
The application is hosted in Google Cloud and serves customers through APIs.
By using private endpoints, you can create a secure connection between your application and the Vertex AI endpoint without exposing the model or data to the public internet. This ensures that sensitive information remains within the cloud.
Private endpoints provide an IP address that is unique to your project, making it easier to manage access control and network policies.
Without private endpoints, you would need to expose your model or data to the public internet, which increases the risk of unauthorized access and security breaches. Private endpoints provide a secure and controlled environment for hosting your model, ensuring that only authorized users can access it.

upvoted 1 times

pinimichele01 6 months, 1 week ago

see guilhermebutzke

upvoted 1 times



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.