EXAMTOPICS

- Expert Verified, Online, **Free.**

☰ MENU   🔍

← **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 202 DISCUSSI...**

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 202

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a startup that has multiple data science workloads. Your compute infrastructure is currently on-premises, and the data science workloads are native to PySpark. Your team plans to migrate their data science workloads to Google Cloud. You need to build a proof of concept to migrate one data science job to Google Cloud. You want to propose a migration process that requires minimal cost and effort. What should you do first?

- A. Create a n2-standard-4 VM instance and install Java, Scala, and Apache Spark dependencies on it.

- B. Create a Google Kubernetes Engine cluster with a basic node pool configuration, install Java, Scala, and Apache Spark dependencies on it.

- C. Create a Standard (1 master, 3 workers) Dataproc cluster, and run a Vertex AI Workbench notebook instance on it.

- D. Create a Vertex AI Workbench notebook with instance type n2-standard-4.

Show Suggested Answer

---

by 👤 **pikachu007** at *Jan. 13, 2024, 4:23 a.m.*

Comments

```
Type your comment...
```

Submit

👤 **baimus** 1 month, 1 week ago

C and D are both valid, as people point out you can technically have Spark preinstalled on D. But this is for a proof of concept for the real design. The concept is not proved by using a notebook, as it's not best practice. Therefore C makes more sense, and is still low effort as it's managed.

👍 ↩ 🚩 upvoted 1 times

👤 **AK2020** 2 months, 2 weeks ago

C is the answer

👍 ↩ 🚩 upvoted 1 times

👤 **TanTran04** 3 months, 2 weeks ago

I'm following option C. Please take a look the concept of 'Dataproc documentation' (ref: https://cloud.google.com/dataproc/docs)

With option D: doesn't provide a solution for managing and scaling the Spark environment, which is necessary for running PySpark workloads.

👍 ↩ 🚩 upvoted 2 times

👤 **fitri001** 6 months ago

Vertex AI Workbench notebook: This option provides a pre-configured environment with popular data science libraries like PySpark already installed. It allows you to focus on migrating your PySpark code with minimal changes.
n2-standard-4 instance type: This is a general-purpose machine type suitable for various data science tasks. It offers a good balance between cost and performance for initial exploration.

👍 ↩ 🚩 upvoted 1 times

   👤 **Jason_Cloud_at** 1 month, 3 weeks ago

   Option D doesnt provide Pyspark out of the box, you have to manually install it wherelse in C dataproc is managed spark and hadoop services which supports running pyspark services right away.

   👍 ↩ 🚩 upvoted 1 times

   👤 **pinimichele01** 6 months ago

   https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview
   why not c?

   👍 ↩ 🚩 upvoted 1 times

   👤 **fitri001** 6 months ago

   A. Create a n2-standard-4 VM instance: This option requires manually installing Java, Scala, and Spark dependencies, which is time-consuming and prone to errors. It also involves managing the VM instance lifecycle, increasing complexity.
   B. Create a Google Kubernetes Engine cluster: Setting up and managing a Kubernetes cluster for a single job is overkill for a proof of concept. It adds unnecessary complexity and cost.
   C. Create a Standard Dataproc cluster: While Dataproc is a managed Spark environment on GCP, setting up a full cluster (master and workers) might be more resource-intensive than needed for a single job, especially for a proof of concept.

   👍 ↩ 🚩 upvoted 1 times

👤 **gscharly** 6 months, 1 week ago

went with D: https://cloud.google.com/vertex-ai/docs/workbench/instances/create-dataproc-enabled

👍 ↩ 🚩 upvoted 2 times

   👤 **pinimichele01** 6 months ago

   https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview

   👍 ↩ 🚩 upvoted 1 times

👤 **pinimichele01** 6 months, 2 weeks ago

When you want to move your Apache Spark workloads from an on-premises environment to Google Cloud, we recommend using Dataproc to run Apache Spark/Apache Hadoop clusters.

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview

👍 ↩ 🚩 upvoted 1 times

👤 **Yan_X** 7 months, 2 weeks ago

D
Can use Notebook pre-installed libraries and tools, including PySpark.

👍 ↩ 🚩 upvoted 2 times

⊟ 👤 **Carlose2108** 7 months, 3 weeks ago

Selected Answer: D

My bad, I mean is Option D.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **Carlose2108** 7 months, 3 weeks ago

Selected Answer: C

I went with C.
For Proof Of Concept and requires minimal cost and effort. Furthermore, Vertex AI Workbench notebooks come pre-configured with PySpark.

👍 ↩ 🚩 upvoted 2 times

⊟ 👤 **guilhermebutzke** 8 months, 1 week ago

Selected Answer: C

My answer: C
C: This option leverages Google Cloud's Dataproc service, which is designed for running Apache Spark and other big data processing frameworks. By creating a Standard Dataproc cluster, you can easily scale resources as needed for your workload.

A. n2-standard-4 VM: This requires manual setup and ongoing maintenance, increasing cost and effort.

B. GKE cluster: While offering containerization benefits, it necessitates managing containers and Spark configurations, adding complexity.

D. With Vertex AI Workbench, your team can develop, train, and deploy machine learning models using popular frameworks like TensorFlow, PyTorch, and scikit-learn. However, while Vertex AI Workbench supports PySpark, it may not be the optimal choice for migrating existing PySpark workloads, as it's primarily focused on machine learning tasks.

👍 ↩ 🚩 upvoted 3 times

⊟ 👤 **Carlose2108** 7 months, 4 weeks ago

You're right but I have a doubt about in a part of Option D
"You need to build a proof of concept to migrate one data science job to Google Cloud"

👍 ↩ 🚩 upvoted 2 times

⊟ 👤 **ddogg** 8 months, 3 weeks ago

Selected Answer: C

Agree with BlehMaks https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview
Dataproc cluster seems more suitable

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **shadz10** 9 months, 1 week ago

Selected Answer: D

https://cloud.google.com/vertex-ai-notebooks?hl=en
Data Data Lake and Spark in one place

Whether you use TensorFlow, PyTorch, or Spark, you can run any engine from Vertex AI Workbench.

D is correct

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **BlehMaks** 9 months, 1 week ago

Selected Answer: C

https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **pikachu007** 9 months, 1 week ago

Selected Answer: D

Minimal setup: Vertex AI Workbench notebooks come pre-configured with PySpark and other data science tools, eliminating the need for manual installation and setup.
Cost-effectiveness: Vertex AI Workbench offers managed notebooks with pay-as-you-go pricing, making it a cost-efficient option for proof-of-concept testing.
Ease of use: Data scientists can directly run PySpark code in the notebook without managing infrastructure, streamlining the migration process.
Scalability: Vertex AI Workbench can easily scale to handle larger workloads or multiple users if the proof-of-concept is successful.

👍 ↩ ⚑ upvoted 1 times

👥

Social Media

Facebook , Twitter

YouTube , Reddit

Pinterest

EXAMTOPICS

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.

f 🐦 ▶ 🤖 📌