Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 313 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 313
Topic #: 1
[All Professional Data Engineer Questions]

You want to migrate an Apache Spark 3 batch job from on-premises to Google Cloud. You need to minimally change the job so that the job reads from Cloud Storage and writes the result to BigQuery. Your job is optimized for Spark, where each executor has 8 vCPU and 16 GB memory, and you want to be able to choose similar settings. You want to minimize installation and management effort to run your job. What should you do?

A. Execute the job as part of a deployment in a new Google Kubernetes Engine cluster.

B. Execute the job from a new Compute Engine VM.

C. Execute the job in a new Dataproc cluster.

D. Execute as a Dataproc Serverless job.

Show Suggested Answer

by mcdaley at *Dec. 7, 2024, 2:45 p.m.*

Comments

Type your comment...

Submit

chicity_de Highly Voted 4 months, 3 weeks ago

Selected Answer: D

Priority is "minimize installation and management effort" which is done via Dataproc Serverless. Furthermore, with Dataproc serverless you can still specify resource settings for your job, such as the number of vCPUs and memory per executor (https://cloud.google.com/dataproc-serverless/docs/concepts/properties)

upvoted 9 times

gabbferreira Most Recent 1 week, 5 days ago

Selected Answer: C

" where each executor has 8 vCPU and 16 GB memory, and you want to be able to choose similar settings."

minimize effort by using dataproc not GKE or vms..

upvoted 1 times

rajshiv 3 weeks, 1 day ago

Selected Answer: D

D is the best answer - It allows Dataproc serverless allows to specify executor configurations like vCPU and memory settings (e.g., executor cores and memory) to match the current setup as is specified.
C is a valid but sub-optimal choice — while we can specify vCPUs and memory similar to our on-prem setup but it requires provisioning and managing clusters, which we want to avoid. it requires slightly more effort compared to Dataproc Serverless.

upvoted 1 times

gabazzzo 2 months, 1 week ago

I agree that minimizing installation and management means using Dataproc Serverles.
Also, Serverles can be configured with up to 16 VPU and up to 29696m of memory in for the premium tier.
https://cloud.google.com/dataproc-serverless/docs/concepts/properties#:~:text=Total%20driver%20memory%20per%20driver%20core%2C%20including%20driver%20memory%20overhead%2C%20which%20must%20be%20between%201024m%20and%207424m%20for%20the%20Standard%20compute%20tier%20(24576m%20for%20the%20Premium%20compute%20tier).%20For%20example%2C%20if%20spark.driver.cores%20%3D%204%2C%20then%204096m%20%3C%3D%20spark.driver.memory%20%2B%20spark.driver.memoryOverhead%20%3C%3D%2029696m.

upvoted 1 times

a494e30 2 months, 3 weeks ago

Selected Answer: C

Needs to be able to configure "similar settings"

upvoted 1 times

plum21 3 months ago

Selected Answer: C

It's not possible to specify a machine type using Dataproc Serverless

upvoted 1 times

marlon.andrei 3 months, 2 weeks ago

Selected Answer: C

I choice "C", just: "where each executor has 8 vCPU and 16 GB memory, and you want to be able to choose similar settings"

upvoted 1 times

Pime13 3 months, 4 weeks ago

Selected Answer: D

Dataproc Serverless allows you to run Spark jobs without needing to manage the underlying infrastructure. It automatically handles resource provisioning and scaling, which simplifies the process and reduces management overhead

upvoted 1 times

mcdaley 4 months, 4 weeks ago

Selected Answer: C

Dataproc supports Spark 3, ensuring compatibility with your existing job.

It also allows you to customize the cluster configuration, including the number of executors, vCPUs, and memory per executor, to match your on-premises setup (8 vCPU and 16 GB memory)

upvoted 1 times

Platform

- Home
- Examtopics PRO
- All Exams
- Training Courses