Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 61 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 61
Topic #: 1
[All Professional Data Engineer Questions]

Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud
Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google
BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

A. Migrate the workload to Google Cloud Dataflow

B. Use pre-emptible virtual machines (VMs) for the cluster

C. Use a higher-memory node so that the job runs faster

D. Use SSDs on the worker nodes so that the job can run faster

Show Suggested Answer

by jvg637 at *March 16, 2020, 3:32 p.m.*

Comments

Type your comment...

Submit

jvg637 Highly Voted 5 years, 1 month ago

B. (Hadoop/Spark jobs are run on Dataproc, and the pre-emptible machines cost 80% less)

upvoted 47 times

rickywck Highly Voted 5 years, 1 month ago

I think the answer should be B:

https://cloud.google.com/dataproc/docs/concepts/compute/preemptible-vms

upvoted 18 times

AmitK121981 Most Recent 4 months, 3 weeks ago

Selected Answer: B

all are saying its pre-emptibles. but spot VMs can only be used in secondary worker, not on master and primary worker so not sure why this will cause savings, and secondary workers are not mandatory too

upvoted 1 times

theseawillclaim 1 year, 9 months ago

I believe it might be "B", but what if the job is mission critical?
Pre-emptible VMs would be of no use.

upvoted 2 times

enivid007 9 months, 1 week ago

Mission critical workloads can't be needed "weekly"

upvoted 1 times

abi01a 2 years ago

I believe Exam Topics ought to provide brief explanation or supporting link to picked correct answers such as this one. Option A may be correct from the view point that Dataflow is a Serverless service that is fast, cost-effective and the fact that Preemptible VMs though can give large price discount may not always be available. It will be great to know the reason(s) behind Exam Topic selected option.

upvoted 8 times

samdhimal 2 years, 3 months ago

B. Use pre-emptible virtual machines (VMs) for the cluster

Using pre-emptible VMs allows you to take advantage of lower-cost virtual machine instances that may be terminated by Google Cloud after a short period of time, typically after 24 hours. These instances can be a cost-effective way to handle workloads that can be interrupted, such as batch processing jobs like the one described in the question.

Option A is not ideal, as it would require you to migrate the workload to Google Cloud Dataflow, which may cause additional complexity and would not address the issue of cost optimization.
Option C is not ideal, as it would require you to use a higher-memory node which would increase the cost.
Option D is not ideal, as it would require you to use SSDs on the worker nodes which would increase the cost.

Using pre-emptible VMs is a better option as it allows you to take advantage of lower-cost virtual machine instances and handle workloads that can be interrupted, which can help to optimize the cost of the cluster.

upvoted 4 times

Rodolfo_Marcos 2 years, 4 months ago

What is happening with this test "correct answer" a lot of times it doesn't make any sense. As this one... Clear it's B

upvoted 2 times

DipT 2 years, 4 months ago

Selected Answer: B

Using preemtible machines are cost effective , and because is suitable for a job mentioned here as it is fault tolerant .

upvoted 2 times

DGames 2 years, 4 months ago

Selected Answer: B

User Pre-emptible VM machine and save process cost, and question want simple solution.

upvoted 1 times

odacir 2 years, 4 months ago

Selected Answer: B

A- Data flow it's not cost-effective in comparison with dataproc
B- Preemptible VM instances are available at much lower price—a 60-91% discount—compared to the price of standar, so this is the answer
C and D are more expensive.

upvoted 1 times

upvoted 1 times

Remi2021 2 years, 8 months ago

Selected Answer: B

B is right way to go

upvoted 1 times

FrankT2L 2 years, 11 months ago

Selected Answer: B

Preemptible workers are the default secondary worker type. They are reclaimed and removed from the cluster if they are required by Google Cloud for other tasks. Although the potential removal of preemptible workers can affect job stability, you may decide to use preemptible instances to lower per-hour compute costs for non-critical data processing or to create very large clusters at a lower total cost

https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

upvoted 1 times

Remi2021 3 years, 1 month ago

B is teh right answer. examtopics update your answers or make your site free again.

upvoted 4 times

OmJanmeda 3 years, 1 month ago

Selected Answer: B

B is right answer.
my experience is not good with Examtopics, so many wrong answers.

upvoted 4 times

Yaa 3 years, 3 months ago

Selected Answer: B

B should be the right answer.
I am amazed that almost 60% of the marked answers on the site are wrong.

upvoted 2 times

byash1 3 years, 3 months ago

Ans : B,
here we are checking on reducing cost, so pre-emptiable machines are best choice

upvoted 1 times

medeis_jar 3 years, 4 months ago

Selected Answer: B

"this workload can run in approximately 30 minutes on a 15-node cluster,"
so you need performance for only 30 mins -> preemptible VMs

https://cloud.google.com/dataproc/docs/concepts/compute/preemptible-vms

upvoted 4 times

Load full discussion...