Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 138 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 138
Topic #: 1
[All Professional Data Engineer Questions]

You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

A. Create a Cloud Dataproc Workflow Template

B. Create an initialization action to execute the jobs

C. Create a Directed Acyclic Graph in Cloud Composer

D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Show Suggested Answer

by ducc at *Sept. 3, 2022, 6:39 a.m.*

Comments

Type your comment...

Submit

LP_PDE Highly Voted 2 years, 1 month ago
Correct answer is A. https://cloud.google.com/dataproc/docs/concepts/workflows/using-workflows
upvoted 5 times

sknaire Most Recent 3 months ago

A. Create a Cloud Dataproc Workflow Template
Dataproc Workflow Template can be used to run jobs concurrently and sequentially. DAG is an overkill.
https://cloud.google.com/dataproc/docs/concepts/workflows/use-workflows

upvoted 1 times

MaxNRG 10 months, 3 weeks ago

Selected Answer: C

The best option for automating your scheduled Spark jobs on Cloud Dataproc, considering sequential and concurrent execution, is:
C. Create a Directed Acyclic Graph (DAG) in Cloud Composer.

upvoted 3 times

MaxNRG 10 months, 3 weeks ago

Here's why:
DAG workflows: Cloud Composer excels at orchestrating complex workflows with dependencies, making it ideal for managing sequential and concurrent execution of your Spark jobs. You can define dependencies between tasks to ensure certain jobs only run after others finish.
Automation: Cloud Composer lets you schedule workflows to run automatically based on triggers like time intervals or data availability, eliminating the need for manual intervention.
Integration: Cloud Composer integrates seamlessly with Cloud Dataproc, allowing you to easily launch and manage your Spark clusters within the workflow.
Scalability: Cloud Composer scales well to handle a large number of jobs and workflows, making it suitable for managing complex data pipelines.

upvoted 2 times

MaxNRG 10 months, 3 weeks ago

While the other options have some merit, they fall short in certain aspects:
A. Cloud Dataproc Workflow Templates: While workflow templates can automate job submission on a cluster, they lack the ability to define dependencies and coordinate concurrent execution effectively.
B. Initialization action: An initialization action can only run a single script before a Dataproc cluster starts, not suitable for orchestrating multiple scheduled jobs with dependencies.
D. Bash script: A Bash script might work for simple cases, but it can be cumbersome to manage and lacks the advanced scheduling and error handling capabilities of Cloud Composer.
Therefore, utilizing a Cloud Composer DAG offers the most comprehensive and flexible solution for automating your scheduled Spark jobs with sequential and concurrent execution on Cloud Dataproc.

upvoted 3 times

emmylou 11 months, 2 weeks ago

Selected Answer: C

I thought it might be A but the templates can only run sequentially, not concurrently.

upvoted 1 times

barnac1es 1 year, 1 month ago

Selected Answer: C

Directed Acyclic Graph (DAG): Cloud Composer (formerly known as Cloud Composer) is a managed Apache Airflow service that allows you to create and manage workflows as DAGs. You can define a DAG that includes tasks for running Spark jobs in sequence or concurrently.

Scheduling: Cloud Composer provides built-in scheduling capabilities, allowing you to specify when and how often your DAGs should run. You can schedule the execution of your Spark jobs at specific times or intervals.

Dependency Management: In a DAG, you can define dependencies between tasks. This means you can set up tasks to run sequentially or concurrently based on your requirements. For example, you can specify that Job B runs after Job A has completed, or you can schedule jobs to run concurrently when there are no dependencies.

upvoted 1 times

midgoo 1 year, 7 months ago

Selected Answer: C

I would choose A if there was one more step to schedule the Template. It is like creating DAG without running it in Airflow. So only option C is correct here.

upvoted 2 times

AzureDP900 1 year, 10 months ago

C. Create a Directed Acyclic Graph in Cloud Composer

upvoted 3 times

saurabhsingh4k 1 year, 10 months ago

Selected Answer: A

Why go for an expensive Composer when you only have to schedule and create a DAG for Dataproc. A is sufficient

upvoted 2 times

captainbu 1 year, 10 months ago

I've would've gone for Workflow Templates as well. But those are lacking the scheduling capability. Hence you would need to use Cloud Composer (or Cloud Functions or Cloud Scheduler) anyway. Hence C seems to be the better solution.

Pls see here:
https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions

upvoted 5 times

zellck 1 year, 11 months ago

Selected Answer: C

C is the answer.

https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions#cloud_composer
Cloud Composer is a managed Apache Airflow service you can use to create, schedule, monitor, and manage workflows.
Advantages:
- Supports time- and event-based scheduling
- Simplified calls to Dataproc using Operators
- Dynamically generate workflows and workflow parameters
- Build data flows that span multiple Google Cloud products

upvoted 2 times

devaid 2 years ago

Selected Answer: C

C.
Composer fits better to schedule Dataproc Workflows, check the documentation:
https://cloud.google.com/dataproc/docs/concepts/workflows/workflow-schedule-solutions

Also A is not enough. Dataproc Workflow Template itself don't has a native schedule option.

upvoted 4 times

louisgcpde 2 years ago

Selected Answer: C

So that I thing the answer should be C (Composer).

upvoted 1 times

louisgcpde 2 years ago

To me, the point is "automate" the process, so that Composer DAG is needed and can be used with Dataproc Workflow Template.

upvoted 2 times

dmzr 2 years ago

Selected Answer: A

Ans A makes more sense, since a question is regarding Dataproc jobs only

upvoted 2 times

HarshKothari21 2 years, 1 month ago

Selected Answer: C

Option c

upvoted 1 times

ducc 2 years, 2 months ago

Selected Answer: C

You have streaming and batch job, so Composer is the choice for me

upvoted 1 times