EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 235 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 235
Topic #: 1
[All Professional Data Engineer Questions]

You want to schedule a number of sequential load and transformation jobs. Data files will be added to a Cloud Storage bucket by an upstream process. There is no fixed schedule for when the new data arrives. Next, a Dataproc job is triggered to perform some transformations and write the data to BigQuery. You then need to run additional transformation jobs in BigQuery. The transformation jobs are different for every table. These jobs might take hours to complete. You need to determine the most efficient and maintainable workflow to process hundreds of tables and provide the freshest data to your end users. What should you do?

A. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage, Dataproc, and BigQuery operators.
2. Use a single shared DAG for all tables that need to go through the pipeline.
3. Schedule the DAG to run hourly.

B. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage, Dataproc, and BigQuery operators.
2. Create a separate DAG for each table that needs to go through the pipeline.
3. Schedule the DAGs to run hourly.

C. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.
2. Use a single shared DAG for all tables that need to go through the pipeline.
3. Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG.

D. 1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.
2. Create a separate DAG for each table that needs to go through the pipeline.

2. Create a separate DAG for each table that needs to go through the pipeline.
3. Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG.

Show Suggested Answer

by scaenruy at *Jan. 3, 2024, 1:24 p.m.*

Comments

Type your comment...

Submit

cuadradobertolinisebastiancami Highly Voted 1 year, 2 months ago

D

* Transformations are in Dataproc and BigQuery. So you don't need operators for GCS (A and B can be discard)
* "There is no fixed schedule for when the new data arrives." so you trigger the DAG when a file arrives
* "The transformation jobs are different for every table. " so you need a DAG for each table.

Then, D is the most suitable answer

upvoted 8 times

choprat1 Most Recent 3 months ago

Selected Answer: D

managing indidivuals DAGs is the best way when they're too different

upvoted 1 times

f74ca0c 4 months, 2 weeks ago

Selected Answer: C

A single shared DAG is efficient to manage, and table-specific transformations can be handled using parameters (e.g., passing table names and configurations dynamically).
Triggering the DAG using a Cloud Storage object notification and a Cloud Function ensures the workflow starts immediately upon data arrival.
Event-driven architecture minimizes delays and provides the freshest data to users.
Efficient, maintainable, and event-driven.

upvoted 3 times

8ad5266 10 months, 1 week ago

Selected Answer: C

This explains why it's not D:
maintainable workflow to process hundreds of tables and provide the freshest data to your end users

How is creating a DAG for each of the hundreds of tables maintainable?

upvoted 3 times

> **plum21** 2 months, 3 weeks ago
>
> It's possible to generate multiple DAGs programatically. That's the reason for C.
> https://cloud.google.com/blog/products/data-analytics/optimize-cloud-composer-via-better-airflow-dags -> look at #5
>
> upvoted 1 times

JyoGCP 1 year, 2 months ago

Selected Answer: D

Option D

upvoted 1 times

Matt_108 1 year, 3 months ago

Selected Answer: D

Option D, which gets triggered when the data comes in and accounts for the fact that each table has its own set of transformations

upvoted 3 times

Jordan18 1 year, 4 months ago

why not C?

upvoted 3 times

cuadradobertolinisebastiancami 1 year, 2 months ago

It says that the transformations for each table are very different

upvoted 2 times

AllenChen123 1 year, 3 months ago

Same question, why not use single DAG to manage as there are hundreds of tables.

upvoted 5 times

raaad 1 year, 4 months ago

Selected Answer: D

- Option D: Tailored handling and scheduling for each table; triggered by data arrival for more timely and efficient processing.

upvoted 2 times

scaenruy 1 year, 4 months ago

Selected Answer: D

D.
1. Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.
2. Create a separate DAG for each table that needs to go through the pipeline.
3. Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG.

upvoted 1 times