Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 237 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 237
Topic #: 1
[All Professional Data Engineer Questions]

You are planning to load some of your existing on-premises data into BigQuery on Google Cloud. You want to either stream or batch-load data, depending on your use case. Additionally, you want to mask some sensitive data before loading into BigQuery. You need to do this in a programmatic way while keeping costs to a minimum. What should you do?

A. Use Cloud Data Fusion to design your pipeline, use the Cloud DLP plug-in to de-identify data within your pipeline, and then move the data into BigQuery.

B. Use the BigQuery Data Transfer Service to schedule your migration. After the data is populated in BigQuery, use the connection to the Cloud Data Loss Prevention (Cloud DLP) API to de-identify the necessary data.

C. Create your pipeline with Dataflow through the Apache Beam SDK for Python, customizing separate options within your code for streaming, batch processing, and Cloud DLP. Select BigQuery as your data sink.

D. Set up Datastream to replicate your on-premise data on BigQuery.

Show Suggested Answer

by scaenruy at *Jan. 3, 2024, 1:42 p.m.*

Comments

Type your comment...

Submit

raaad Highly Voted 1 year, 4 months ago

Selected Answer: C

- Programmatic Flexibility: Apache Beam provides extensive control over pipeline design, allowing for customization of data transformations, including integration with Cloud DLP for sensitive data masking.
- Streaming and Batch Support: Beam seamlessly supports both streaming and batch data processing modes, enabling flexibility in data loading patterns.
- Cost-Effective Processing: Dataflow offers a serverless model, scaling resources as needed, and only charging for resources used, helping optimize costs.
- Integration with Cloud DLP: Beam integrates well with Cloud DLP for sensitive data masking, ensuring data privacy before loading into BigQuery.

upvoted 11 times

qq589539483084gfrgrgfr 1 year, 3 months ago

In correct Option is A because you want a programatic way whereas datafusion is codeless solution and also dataflow is cost effective

upvoted 2 times

AllenChen123 1 year, 3 months ago

You are saying Option C

upvoted 2 times

JyoGCP Most Recent 1 year, 2 months ago

Selected Answer: C

Option C

upvoted 1 times

tibuenoc 1 year, 3 months ago

Selected Answer: C

C is correct. Using Dataflow as Python as programming and BQ as sink.

A is incorrect - DataFusion is Code-free as the main propose

upvoted 2 times

scaenruy 1 year, 4 months ago

Selected Answer: A

A.
Use Cloud Data Fusion to design your pipeline, use the Cloud DLP plug-in to de-identify data within your pipeline, and then move the data into BigQuery.

upvoted 1 times

ggg24 6 months, 1 week ago

Data Fusion support only Batch and Streaming is required

upvoted 1 times

chrissamharris 11 months, 3 weeks ago

Incorrect, that's a low-code solution. Doesnt meet this specific requirement: "You need to do this in a programmatic way"

upvoted 1 times