Google Discussions

Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

Go to Exam

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 193 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 193

Topic #: 1

[All Professional Data Engineer Questions]

An aerospace company uses a proprietary data format to store its flight data. You need to connect this new data source to BigQuery and stream the data into
BigQuery. You want to efficiently import the data into BigQuery while consuming as few resources as possible. What should you do?

A. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source.

B. Use a standard Dataflow pipeline to store the raw data in BigQuery, and then transform the format later when the data is used.

C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format.

D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format.

Show Suggested Answer

by ducc at *Sept. 3, 2022, 3:52 a.m.*

Comments

Type your comment...

Submit

beanz00 Highly Voted 2 years ago

This has to be D. How could it even be B? The source is a proprietary format. Dataflow wouldn't have a built-in template to ead the file. You will have to create something custom.

upvoted 17 times

devaid Highly Voted 2 years ago

Selected Answer: D

For me it's clearly D
It's between B and D, but read B, store raw data in Big Query? Use a Dataflow pipeline just to store raw data into Big Query, and transform later? You'd need to do another pipeline for that, and is not efficient.

upvoted 12 times

MaxNRG Most Recent 10 months, 2 weeks ago

Selected Answer: D

Option D is the best approach given the constraints - use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format.
The key reasons:
• Dataflow provides managed resource scaling for efficient stream processing
• Avro format has schema evolution capabilities and efficient serialization for flight telemetry data
• Apache Beam connectors avoid having to write much code to integrate proprietary data sources
• Streaming inserts data efficiently compared to periodic batch jobs
In contrast, option A uses Cloud Functions which lack native streaming capabilities. Option B stores data in less efficient JSON format. Option C uses Dataproc which requires manual cluster management.
So leveraging Dataflow + Avro + Beam provides the most efficient way to stream proprietary flight data into BigQuery while using minimal resources.

upvoted 3 times

Aman47 10 months, 3 weeks ago

Its talking about streaming? none of the options talk about triggering a load to begin. We need a trigger or schedule to run first.

upvoted 1 times

AjoseO 1 year ago

Selected Answer: D

Option D allows you to use a custom connector to read the proprietary data format and write the data to BigQuery in Avro format.

upvoted 2 times

sergiomujica 1 year, 1 month ago

Selected Answer: D

the keyword is streaming

upvoted 1 times

knith66 1 year, 3 months ago

Between B and D. Firstly transformation is not mentioned in the question, So B is less probable. Then Efficient import is mentioned in the question, Converting to Avro will consume less space. I am going with D

upvoted 3 times

musumusu 1 year, 8 months ago

Answer is D ,
Why not B, changing data format before uploading to bigquery is good approach.

upvoted 1 times

cetanx 1 year, 9 months ago

Selected Answer: B

I believe keyword here is "An aerospace company uses a proprietary data format"
So if we list the connectors available in Apache Beam, we are listed with these options;
https://beam.apache.org/documentation/io/connectors/

So I believe, we have to create our own custom connector to read from the proprietary data format hence the answer should be B

upvoted 1 times

> **cetanx** 1 year, 9 months ago
>
> sorry the answer should be D
>
> upvoted 1 times

AzureDP900 1 year, 10 months ago

D is right

upvoted 1 times

upvoted 1 times

zellck 1 year, 11 months ago

Selected Answer: D

D is the answer.

upvoted 3 times

TNT87 1 year, 9 months ago

There is dataflow connector and D isnt cost effective

upvoted 1 times

hauhau 1 year, 11 months ago

Selected Answer: B

B is the most efficient

upvoted 2 times

TNT87 2 years ago

https://cloud.google.com/spanner/docs/change-streams/use-dataflow#core-concepts

upvoted 1 times

TNT87 2 years, 1 month ago

Ans B
https://cloud.google.com/architecture/streaming-avro-records-into-bigquery-using-dataflow
Is there a reason to use apache beam connector yet there is dataflow which is a standard solution for that scenario?

upvoted 2 times

TNT87 2 years, 1 month ago

https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-data-ingestion

upvoted 1 times

learner2610 2 years, 1 month ago

Can standard dataflow be used to ingest any proprietary format of file ?
shouldn't we use custom apache beam connector ?
So I think it is D ,though it isn't simple ,But in this scenario they have asked to use less resources to import data

upvoted 1 times

TNT87 2 years, 1 month ago

Do you mind reading the links i provided and revisiting the question, then you will understand why D isnt the best.
Why use Apache beam yet there is Dataflow

upvoted 1 times

John_Pongthorn 2 years, 1 month ago

D: just have your team develop custom connector.
https://cloud.google.com/architecture/bigquery-data-warehouse#storage_management
Internally, BigQuery stores data in a proprietary columnar format called Capacitor, which has a number of benefits for data warehouse workloads. BigQuery uses a proprietary format

I suppose this matter , it mean BQ use proprietary format by itself to work internally
but the question means data as proprietary format as input for ingesting into BQ.

upvoted 2 times

Load full discussion...

TNT87 2 years, 1 month ago

Option D streams, thats not cost effective. We need something that is cost effectictive, hence B is the option

upvoted 1 times

TNT87 2 years, 1 month ago

I mean that consumes fewer resources

upvoted 1 times

AWSandeep 2 years, 2 months ago

Selected Answer: D

D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format.
Reveal Solution

upvoted 3 times

ducc 2 years, 2 months ago

Selected Answer: B

B is the most efficient for me

D is the most efficient for me.

upvoted 2 times

ducc 2 years, 2 months ago

Sorry, D is correct

upvoted 2 times

EXAMTOPICS

Platform

- Home
- Examtopics PRO
- All Exams
- Training Courses