🔍

← 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 94 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 94

Topic #: 1

[All Professional Data Engineer Questions]

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

A. Batch job, PubSubIO, side-inputs

B. Streaming job, PubSubIO, JdbcIO, side-outputs

C. Streaming job, PubSubIO, BigQueryIO, side-inputs

D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Show Suggested Answer
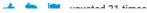
by 👤 rickywck at *March 17, 2020, 9:34 a.m.*

Comments

Type your comment...

Submit

🗆 👤 rickywck [Highly Voted 👍] 4 years, 7 months ago

Why not C? Without BigQueryIO how can data be written back to BigQuery?

 xq 4 years, 7 months ago

C should be right

 [Removed] `Highly Voted` 4 years, 7 months ago

Answer: C

Description: Sideinput for Bigquery data

 JOKKUNO `Most Recent ⊙` 10 months, 1 week ago

Side inputs

In addition to the main input PCollection, you can provide additional inputs to a ParDo transform in the form of side inputs. A side input is an additional input that your DoFn can access each time it processes an element in the input PCollection. When you specify a side input, you create a view of some other data that can be read from within the ParDo transform's DoFn while processing each element.

Side inputs are useful if your ParDo needs to inject additional data when processing each element in the input PCollection, but the additional data needs to be determined at runtime (and not hard-coded). Such values might be determined by the input data, or depend on a different branch of your pipeline.

 JOKKUNO 10 months, 1 week ago

https://beam.apache.org/documentation/programming-guide/#side-inputs

 piyush7777 1 year, 2 months ago

Why not side-output?

 TQM__9MD 1 year, 3 months ago

Selected Answer: B

B. Use multi-cluster routing to add a second cluster to the existing instance, utilizing a live traffic app profile for the regular workload and a batch analytics profile for the analytical workload.

 Mathew106 1 year, 3 months ago

Selected Answer: C

The answer is C. It's a trap so that you answer A because of batch vs streaming but you need BigQueryIO. On the other hand, streaming is absolutely redundant here and will incur extra costs. C is right but would be better with batch.

 Siadd 1 year, 10 months ago

A is the Answer.

A. Batch job, PubSubIO, side-inputs

 zellck 1 year, 11 months ago

Selected Answer: C

C is the answer.

https://cloud.google.com/dataflow/docs/tutorials/ecommerce-java#side-input-pattern

In streaming analytics applications, data is often enriched with additional information that might be useful for further analysis. For example, if you have the store ID for a transaction, you might want to add information about the store location. This additional information is often added by taking an element and bringing in information from a lookup table.

 sedado77 2 years, 1 month ago

Selected Answer: C

I got this question on sept 2022. Answer is C

 chrismayola 2 years ago

dear can you please help, i have some questions about how to prepare the cerification exam using this questionnaire. this is my email cmayola@yahoo.fr, ping me to have some conversation

 alex12441 2 years, 9 months ago

Selected Answer: C

Answer: C

👍 ↩ 🚩 **upvoted 1 times**

⊟ 👤 **medeis_jar** 2 years, 10 months ago

Selected Answer: C

I vote for C, because data will come from Pub/Sub, so it should be streaming, we'll need PubSubIO to be able to read from PubSub and BigQueryIO to be able to write to BigQuery, finally the side-inputs pattern let us enrich data

👍 ↩ 🚩 **upvoted 5 times**

⊟ 👤 **MaxNRG** 2 years, 10 months ago

Selected Answer: C

Static reference data from BigQuery will go as side-inputs and data from pub-sub will go as streaming data using PubSubIO and finally BigQueryIO is required to push the final data to BigQuery

👍 ↩ 🚩 **upvoted 4 times**

⊟ 👤 **JG123** 2 years, 11 months ago

Ans: C

👍 ↩ 🚩 **upvoted 1 times**

⊟ 👤 **pals_muthu** 3 years, 2 months ago

Answer is C,
You need pubsubIO and BigQueryIO for streaming data and writing enriched data back to BigQuery. side-inputs are a way to enrich the data
https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs

👍 ↩ 🚩 **upvoted 6 times**

⊟ 👤 **Meuter** 3 years, 2 months ago

I choose C, because data will come from Pub/Sub, so it should be streaming, we'll need PubSubIO to be able to read from PubSub y BigQueryIO to be able to write to BigQuery, finally the side-inputs pattern let us enrich data
https://beam.apache.org/releases/javadoc/2.4.0/org/apache/beam/sdk/io/gcp/pubsub/PubsubIO.html
https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs
https://beam.apache.org/releases/javadoc/2.3.0/org/apache/beam/sdk/io/gcp/bigquery/BigQueryIO.html

👍 ↩ 🚩 **upvoted 3 times**

⊟ 👤 **daghayeghi** 3 years, 7 months ago

C:
we have to use Streaming job because of Pub/Sub, and side-input thanks to static reference data. and we have to leverage BigQueryIO since finally we want to write data to BigQuery. then C is the correct answer.

👍 ↩ 🚩 **upvoted 2 times**

⊟ 👤 **someshsehgal** 3 years, 8 months ago

Correct A. batch is cost-effective and no need to go for streaming

👍 ↩ 🚩 **upvoted 1 times**

⊟ 👤 **funtoosh** 3 years, 8 months ago

How you are going to write back to BQ?

👍 ↩ 🚩 **upvoted 1 times**

Load full discussion...