

🔗 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 296 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 296

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

Your infrastructure team has set up an interconnect link between Google Cloud and the on-premises network. You are designing a high-throughput streaming pipeline to ingest data in streaming from an Apache Kafka cluster hosted on- premises. You want to store the data in BigQuery, with as minimal latency as possible. What should you do?

- A. Setup a Kafka Connect bridge between Kafka and Pub/Sub. Use a Google-provided Dataflow template to read the data from Pub/Sub, and write the data to BigQuery.
- B. Use a proxy host in the VPC in Google Cloud connecting to Kafka. Write a Dataflow pipeline, read data from the proxy host, and write the data to BigQuery.
- C. Use Dataflow, write a pipeline that reads the data from Kafka, and writes the data to BigQuery.
- D. Setup a Kafka Connect bridge between Kafka and Pub/Sub. Write a Dataflow pipeline, read the data from Pub/Sub, and write the data to BigQuery.

[Show Suggested Answer](#)

by [rahulvin](#) at Dec. 30, 2023, 9:07 p.m.

Comments

Type your comment...



[Submit](#)

  **rajshiv** 3 weeks, 1 day ago

Selected Answer: A

I think it is A and not C. While Dataflow can read from Kafka directly, it works best for Kafka clusters hosted in Google Cloud. Reading from an on-prem Kafka over Interconnect directly from Dataflow is not recommended due to latency, firewall/NAT issues, and network complexity. Most important this option is Not optimal for performance and reliability across hybrid environments.

   upvoted 1 times


  **Pime13** 3 months, 4 weeks ago

Selected Answer: C

C. Use Dataflow, write a pipeline that reads the data from Kafka, and writes the data to BigQuery.

This approach allows you to directly connect Dataflow to your Kafka cluster, ensuring minimal latency by avoiding additional intermediaries like Pub/Sub. Dataflow is designed to handle high-throughput data processing and can efficiently ingest and process streaming data from Kafka, then write it to BigQuery. This setup leverages the interconnect link for a direct and efficient data flow


   upvoted 1 times

  **Pime13** 3 months, 3 weeks ago

https://cloud.google.com/dataflow/docs/kafka-dataflow#deploy_kafka

Alternatively, you might have an existing Kafka cluster that resides outside of Google Cloud. For example, you might have an existing workload that is deployed on-premises or in another public cloud.

   upvoted 1 times

  **meh_33** 8 months, 4 weeks ago

Going with C

   upvoted 1 times

  **Anudeep58** 10 months ago

Selected Answer: C

Latency: Option C, with direct integration between Kafka and Dataflow, offers lower latency by eliminating intermediate steps.

Flexibility: Custom Dataflow pipelines (Option C) provide more control over data processing and optimization compared to using a pre-built template.

   upvoted 2 times

  **anushree09** 1 year ago

per the text below at <https://cloud.google.com/dataflow/docs/kafka-dataflow> -

"Alternatively, you might have an existing Kafka cluster that resides outside of Google Cloud. For example, you might have an existing workload that is deployed on-premises or in another public cloud."

   upvoted 1 times

  **Moss2011** 1 year, 2 months ago

Selected Answer: C

From my point of view, the best option is C taking into account this doc: <https://cloud.google.com/dataflow/docs/kafka-dataflow>

   upvoted 2 times

  **MaxNRG** 1 year, 2 months ago

Selected Answer: D

Based on the key requirements highlighted:

- Interconnect link between GCP and on-prem Kafka
- High throughput streaming pipeline
- Minimal latency
- Data to be stored in BigQuery


D - The key reasons this meets the requirements:

- Kafka connect provides a reliable bridge to Pub/Sub over the interconnect
- Reading from Pub/Sub minimizes latency vs reading directly from Kafka
- Dataflow provides a high throughput streaming engine
- Writing to BigQuery gives scalable data storage

By leveraging these fully managed GCP services over the dedicated interconnect, a low latency streaming pipeline from on-prem Kafka into BigQuery can be implemented rapidly.

Options A/B/C have higher latencies or custom code requirements, so do not meet the minimal latency criteria as well as option D.

   upvoted 2 times

  **MaxNRG** 1 year, 2 months ago

Why choose option D over A?

The key advantage with option D is that by writing a custom Dataflow pipeline rather than using a Google provided template, there is more flexibility to customize performance tuning and optimization for lowest latency.

- Some potential optimizations:
- Fine tuning number of workers, machine types to meet specific throughput targets
- Custom data parsing/processing logic if applicable
- Experimenting with autoscaling parameters or triggers

   upvoted 1 times

  **MaxNRG** 1 year, 2 months ago

The Google template may be easier to set up initially, but a custom pipeline provides more control over optimizations specifically for low latency requirements stated in the question.

That being said, option A would still work reasonably well - but option D allows squeezing out that extra bit of performance if low millisecond latency is absolutely critical in the pipeline through precise tuning.

So in summary, option A is easier to implement but option D provides more optimization flexibility for ultra low latency streaming requirements.

   upvoted 1 times

  **MaxNRG** 1 year, 2 months ago

Why not C:

At first option C (using a Dataflow pipeline to directly read from Kafka and write to BigQuery) seems reasonable.

However, the key requirement stated in the question is to have minimal latency for the streaming pipeline.

By reading directly from Kafka within Dataflow, there can be additional latency and processing overhead compared to reading from Pub/Sub, for a few reasons:

1. Pub/Sub acts as a buffer and handles scaling/reliability of streaming data automatically. This reduces processing burden on the pipeline.
2. Network latency can be lower by leveraging Pub/Sub instead of making constant pull requests for data from Kafka within the streaming pipeline.
3. Any failures have to be handled within the pipeline code itself when reading directly from Kafka. With Pub/Sub, reliability is built-in.

   upvoted 2 times

  **SanjeevRoy91** 1 year, 1 month ago

You are adding an intermediate hop in between on prem kafka and Dataflow (pubsub). Why won't this add additional latency.

   upvoted 3 times

  **MaxNRG** 1 year, 2 months ago

So in summary, while option C is technically possible, option D introduces Pub/Sub as a streaming buffer which reduces overall latency for the pipeline, allowing the key requirement of minimal latency to be better satisfied.

   upvoted 2 times

  **JyoGCP** 1 year, 2 months ago

A Vs C -- Not sure which one would have low latency.

Points related to option C:

"Yes, Dataflow can read events from Kafka. Dataflow is a fully-managed, serverless streaming analytics service that supports both batch and stream processing. It can read events from Kafka, process them, and write the results to a BigQuery table for further analysis. "

"Dataflow supports Kafka support, which was added to Apache Beam in 2016. Google provides a Dataflow template that configures a Kafka-to-BigQuery pipeline. The template uses the BigQueryIO connector provided in the Apache Beam SDK. "

   upvoted 2 times

  **JyoGCP** 1 year, 2 months ago

Going with C

   upvoted 2 times

  **DarkLord2104** 1 year, 2 months ago

Final???

   upvoted 2 times

  **T2Clubber** 1 year, 3 months ago

Selected Answer: C

Option C makes more sense to me because of the "minimal latency as possible".

I would have chosen option A if it were "less CODING as possible".



   upvoted 3 times

  **Matt_108** 1 year, 3 months ago

Selected Answer: A

Option A, leverage dataflow template for Kafka <https://cloud.google.com/dataflow/docs/kafka-dataflow>

   upvoted 4 times

  **AllenChen123** 1 year, 3 months ago

Agree. "Google provides a Dataflow template that configures a Kafka-to-BigQuery pipeline. The template uses the BigQueryIO connector provided in the Apache Beam SDK."

   upvoted 1 times

  **ML6** 1 year, 2 months ago

But it includes setting up a Kafka Connect bridge while an interconnect link has already been set up.
https://cloud.google.com/dataflow/docs/kafka-dataflow#connect_to_an_external_cluster

   upvoted 1 times

  **scaenrui** 1 year, 4 months ago

Selected Answer: C

C. Use Dataflow, write a pipeline that reads the data from Kafka, and writes the data to BigQuery.

   upvoted 4 times

  **rahulvin** 1 year, 4 months ago

Selected Answer: C

Dataflow has templates to read from Kafka. Other options are too complicated
<https://cloud.google.com/dataflow/docs/kafka-dataflow>

   upvoted 3 times

  **Sofiia98** 1 year, 3 months ago

so, this is the answer A, whe C?

   upvoted 2 times

  **Matt_108** 1 year, 3 months ago

Yeah, the answer is A. C requires you to develop the pipeline yourself and ensure minimal latency, which means that you perform better than a pre-built template from Google...not really the case most of the times :)

   upvoted 1 times

  **saschak94** 1 year, 3 months ago

but Option A introduces additional replication into Pub/Sub and the question states with minimal latency. In my opinion subscribing to Kafka via Dataflow has a lower latency than replicating the messages first to Pub/Sub and subscribing with Dataflow to it.

   upvoted 7 times



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses

