

[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 135 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 135

Topic #: 1

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

You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

- ⇒ Decoupling producer from consumer
- ⇒ Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
- ⇒ Near real-time SQL query
- ⇒ Maintain at least 2 years of historical data, which will be queried with SQL

Which pipeline should you use to meet these requirements?

- A. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
- B. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
- C. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
- D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.





















































[Show Suggested Answer](#)

by [deleted] at March 22, 2020, 10:49 a.m.

Comments

Type your comment...

Submit

-   **[Removed]** Highly Voted 5 years, 1 month ago
Correct - D
   upvoted 44 times
-   **[Removed]** Highly Voted 5 years, 1 month ago
Answer: D
Description: All the requirements meet with D
   upvoted 16 times
-   **edre** Most Recent 9 months, 2 weeks ago
Selected Answer: D
Google recommended approach
   upvoted 1 times
-   **julioirevk** 1 year, 7 months ago
Selected Answer: D
D because pub/sub decouples while dataflow processes; Cloud Storage can be used to store the raw ingested data indefinitely and BQ can be used to query.
   upvoted 2 times
-   **barnac1es** 1 year, 7 months ago
Selected Answer: D
Here's how this option aligns with your requirements:
Decoupling Producer from Consumer: Cloud Pub/Sub provides a decoupled messaging system where the producer publishes events, and consumers (like Dataflow) can subscribe to these events. This decoupling ensures flexibility and scalability.
Space and Cost-Efficient Storage: Storing data in Avro format is more space-efficient than JSON, and Cloud Storage is a cost-effective storage solution. Additionally, Cloud Pub/Sub and Dataflow allow you to process and transform data efficiently, reducing storage costs.
Near Real-time SQL Query: By using Dataflow to transform and load data into BigQuery, you can achieve near real-time data availability for SQL queries. BigQuery is well-suited for ad-hoc SQL queries and provides excellent query performance.
   upvoted 3 times
-   **FP77** 1 year, 8 months ago
Selected Answer: D
Should be D
   upvoted 1 times
-   **vaga1** 1 year, 11 months ago
Selected Answer: D
For sure D
   upvoted 1 times
-   **forepick** 1 year, 11 months ago
Selected Answer: D
D is the most suitable, however the stored format should be JSON, and AVRO isn't JSON...
   upvoted 1 times
-   **OberstK** 2 years, 3 months ago
Selected Answer: D
Correct - D
   upvoted 1 times
-   **desertlotus1211** 2 years, 3 months ago
I believe this was also on the GCP PCA exam as well! ;)
   upvoted 1 times
-   **AzureDP900** 2 years, 4 months ago
D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the

JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

   upvoted 1 times

  **zellick** 2 years, 5 months ago

Selected Answer: D

D is the answer.

   upvoted 1 times

  **mbacelar** 2 years, 5 months ago

Selected Answer: D

For sure D

   upvoted 1 times

  **clouditis** 2 years, 7 months ago

D it is!

   upvoted 1 times

  **Prasanna_kumar** 3 years, 2 months ago

Answer is D

   upvoted 2 times

  **MaxNRG** 3 years, 3 months ago

Selected Answer: D

D:

Cloud Pub/Sub, Cloud Dataflow, Cloud Storage, BigQuery <https://cloud.google.com/solutions/stream-analytics/>

   upvoted 4 times

  **medeis_jar** 3 years, 3 months ago

Selected Answer: D

OMG only D

   upvoted 1 times

[Load full discussion...](#)



Platform

> [Home](#)

> [All Exams](#)

> [Examtopics PRO](#)

> [Training Courses](#)

