EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 282 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 282
Topic #: 1
[All Professional Data Engineer Questions]

You are using a Dataflow streaming job to read messages from a message bus that does not support exactly-once delivery. Your job then applies some transformations, and loads the result into BigQuery. You want to ensure that your data is being streamed into BigQuery with exactly-once delivery semantics. You expect your ingestion throughput into BigQuery to be about 1.5 GB per second. What should you do?

A. Use the BigQuery Storage Write API and ensure that your target BigQuery table is regional.

B. Use the BigQuery Storage Write API and ensure that your target BigQuery table is multiregional.

C. Use the BigQuery Streaming API and ensure that your target BigQuery table is regional.

D. Use the BigQuery Streaming API and ensure that your target BigQuery table is multiregional.

Show Suggested Answer

by Ed_Kim at *Jan. 3, 2024, 3:06 a.m.*

Comments

Type your comment...

Submit

AlizCert Highly Voted 11 months ago

Selected Answer: B

It should B, Storage Write API has "3 GB per second throughput in multi-regions; 300 MB per second in regions"

upvoted 16 times

rajshiv 3 weeks, 1 day ago

B is incorrect. Multiregional tables are not supported by the Storage Write API for exactly-once delivery. This option is invalid.

upvoted 1 times

raaad Highly Voted 1 year, 3 months ago

Selected Answer: A

- BigQuery Storage Write API: This API is designed for high-throughput, low-latency writing of data into BigQuery. It also provides tools to prevent data duplication, which is essential for exactly-once delivery semantics.
- Regional Table: Choosing a regional location for the BigQuery table could potentially provide better performance and lower latency, as it would be closer to the Dataflow job if they are in the same region.

upvoted 15 times

AllenChen123 1 year, 3 months ago

Agree.
https://cloud.google.com/bigquery/docs/write-api#advantages

upvoted 4 times

aditya_ali Most Recent 1 day, 10 hours ago

Selected Answer: A

You need a write latency of 1.5 GBs per second. Given the high throughput requirement, a regional BigQuery table (Option A) is generally preferred over a multi-regional table due to potentially lower write latency in multi-region. Simple.

upvoted 1 times

Aungshuman 5 days, 7 hours ago

Selected Answer: B

As per GCP document multi-region meets the troughput requirement.

upvoted 1 times

gabbferreira 1 week, 5 days ago

Selected Answer: A

It's A

upvoted 1 times

Siahara 2 months, 4 weeks ago

Selected Answer: A

A. Implement the BigQuery Storage Write API and guarantee that the target BigQuery table is regional.

Here's the breakdown:

Why Option A is Superior

Exactly-Once Delivery: The BigQuery Storage Write API intrinsically supports exactly-once delivery using stream offsets. This guarantees that each message is written to BigQuery exactly one time, even in the case of retries due to the lack of native exactly-once support in your message bus.

High Throughput: The Storage Write API is optimized for high-throughput scenarios. It can handle the expected ingestion throughput of 1.5 GB per second.

Regional Tables: Using a regional BigQuery table aligns with best practices when utilizing the Storage Write API, as it helps to minimize latency and reduce potential cross-region communication costs.

upvoted 4 times

gord_nat 1 month, 1 week ago

Has to be multi-regional (B)
Max throughput for regional currently only 300 MB/s
https://cloud.google.com/bigquery/quotas

upvoted 1 times

juliorevk 3 months ago

Selected Answer: B

- BigQuery Storage Write API: This API is designed for high-throughput, low-latency writing of data into BigQuery. It also provides tools to prevent data duplication, which is essential for exactly-once delivery semantics.
- The multiregional table ensures that your data is highly available and can be streamed into BigQuery across multiple regions. It is better suited for high-throughput and low-latency workloads, as it provides distributed write capabilities that can

regions. It is better suited for high-throughput and low-latency workloads, as it provides distributed write capabilities that can handle large data volumes, such as the 1.5 GB per second you expect to stream.

upvoted 1 times

Pime13 3 months, 4 weeks ago

Selected Answer: A

https://cloud.google.com/bigquery/docs/streaming-data-into-bigquery
For new projects, we recommend using the BigQuery Storage Write API instead of the tabledata.insertAll method. The Storage Write API has lower pricing and more robust features, including exactly-once delivery semantics
https://cloud.google.com/bigquery/docs/write-api#advantages

upvoted 2 times

hussain.sain 4 months, 1 week ago

Selected Answer: B

B is correct.
When aiming for exactly-once delivery in a Dataflow streaming job, the key is to use the BigQuery Storage Write API, as it provides the capability to handle large-scale data ingestion with the correct semantics, including exactly-once delivery.

upvoted 1 times

himadri1983 4 months, 3 weeks ago

Selected Answer: B

3 GB per second throughput in multi-regions; 300 MB per second in regions
https://cloud.google.com/bigquery/quotas#write-api-limits

upvoted 2 times

m_a_p_s 4 months, 3 weeks ago

Selected Answer: B

streamed into BigQuery with exactly-once delivery semantics >>> Storage Write API

ingestion throughput into BigQuery to be about 1.5 GB per second >>> multiregional (check throughput rate here >>> https://cloud.google.com/bigquery/quotas#write-api-limits)

upvoted 2 times

NatyNogas 5 months ago

Selected Answer: A

- Choosing a regional target BigQuery table ensures that data is stored redundantly in a single region, providing high availability and durability.

upvoted 2 times

CloudAdrMX 5 months, 1 week ago

Selected Answer: B

According to this documentation, its B
https://cloud.google.com/bigquery/quotas#write-api-limits

upvoted 2 times

imazy 5 months, 3 weeks ago

Selected Answer: A

Write API support 2.5 GB / sec speed and support exactly-once delivery semantics
https://cloud.google.com/bigquery/docs/write-api#connections

whereas in streaming duplicates can come and needed to remove them manually
https://cloud.google.com/bigquery/docs/streaming-data-into-bigquery#dataavailability

upvoted 1 times

SamuelTsch 6 months ago

Selected Answer: B

looking for this documentation https://cloud.google.com/bigquery/quotas#write-api-limits. 3 GB/s in multi-regions; 300MB/s in regions

upvoted 4 times

HermanTan 7 months, 1 week ago

To ensure that analysts do not see customer data older than 30 days while minimizing cost and overhead, the best option is:
B. Use a timestamp range filter in the query to fetch the customer's data for a specific range.

This approach directly addresses the issue by filtering out data older than 30 days at query time, ensuring that only the relevant data is retrieved. It avoids the overhead and potential delays associated with garbage collection and manual deletion processes

upvoted 2 times

hanoverquay 1 year, 1 month ago

Selected Answer: D

option D

upvoted 1 times

BennyXu 1 year ago

you are wrong!!!!!!!!!!!!

upvoted 1 times

Load full discussion...