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 242 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 242
Topic #: 1
[All Professional Data Engineer Questions]

You have designed an Apache Beam processing pipeline that reads from a Pub/Sub topic. The topic has a message retention duration of one day, and writes to a Cloud Storage bucket. You need to select a bucket location and processing strategy to prevent data loss in case of a regional outage with an RPO of 15 minutes. What should you do?

A. 1. Use a dual-region Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 15 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

B. 1. Use a multi-regional Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

C. 1. Use a regional Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by one day to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region and write in a bucket in the same region.

D. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

Show Suggested Answer

Show Suggested Answer

by scaenruy at *Jan. 3, 2024, 2:15 p.m.*

Comments

Type your comment...

Submit

datapassionate Highly Voted 9 months, 3 weeks ago

Selected Answer: D

D. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

RPO of 15 minutes is guaranteed when turbo replication is used
https://cloud.google.com/storage/docs/availability-durability

upvoted 8 times

ashdam 8 months, 2 weeks ago

Why multi-region is not correct. There is no downtime in case a region goes down.

upvoted 1 times

JyoGCP Highly Voted 8 months, 2 weeks ago

Selected Answer: D

Option D is correct.

Not A, because dual-region bucket WITHOUT turbo replication takes atleast 1 hour to sync data between regions. SLA for 100% data sync is 12 hours as per google.

upvoted 5 times

0dd4e0c Most Recent 4 days, 8 hours ago

Selected Answer: D

it's D, keyword "Turbo replication" for RPO recovery

upvoted 1 times

LP_PDE 3 months, 1 week ago

Selected Answer: D

Could be A or D. The choice between a 15-minute seek and a 60-minute seek depends on your specific requirements and priorities. If a very low RPO is critical, a 60-minute seek might be necessary to ensure data completeness.If minimizing cost and processing time is more important, a 15-minute seek might be sufficient, especially if you're confident in the reliability of Turbo Replication.

upvoted 1 times

m_a_p_s 4 months, 1 week ago

Selected Answer: A

An RPO of 15 minutes seemingly suggests using Turbo Replication. But here's the thing - why would you want to seek the subscription back in time by 60 minutes and run the Dataflow job? Thus, if turbo replication is enabled, steps 3 & 4 are completely redundant and unnecessary. Which is why option A is correct. This was a tricky one!

upvoted 1 times

shangning007 4 months, 2 weeks ago

Selected Answer: A

I don't like answer D. If we have turbo replication can ensure that change within 15min can be replicated, why do we still need to seek the subscription back in time by 60min?

upvoted 1 times

SVGoogle89 5 months, 3 weeks ago

D
https://cloud.google.com/storage/docs/availability-durability#cross-region-redundancy

upvoted 1 times

lipa31 9 months, 2 weeks ago

Selected Answer: D

https://cloud.google.com/storage/docs/availability-durability#turbo-replication says : "When enabled, turbo replication is designed to replicate 100% of newly written objects to both regions that constitute the dual-region within the recovery point objective of 15 minutes, regardless of object size."
so seems D to me

upvoted 4 times

raaad 10 months ago

Selected Answer: A

- Low RPO: Dual-region buckets offer synchronous replication, ensuring data is immediately available in both regions, aligning with the 15-minute RPO.
- Turbo Replication: enabling turbo replication can further reduce replication latency to near-real-time for even stricter RPO requirements.
- Resilient Data Storage: Dual-region buckets ensure data availability even during regional outages, protecting processed data.
- Fast Recovery: Reprocessing from the last 15 minutes of acknowledged messages minimizes data loss and downtime.

upvoted 1 times

qq589539483084gfrgrgfr 9 months, 3 weeks ago

why not D then, if turbo replication improves RPO??

upvoted 2 times

scaenruy 10 months ago

Selected Answer: A

A.
1. Use a dual-region Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 15 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

upvoted 1 times