EXAMTOPICS

MENU

Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 292 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 292

Topic #: 1

[All Professional Data Engineer Questions]

You have terabytes of customer behavioral data streaming from Google Analytics into BigQuery daily. Your customers' information, such as their preferences, is hosted on a Cloud SQL for MySQL database. Your CRM database is hosted on a Cloud SQL for PostgreSQL instance. The marketing team wants to use your customers' information from the two databases and the customer behavioral data to create marketing campaigns for yearly active customers. You need to ensure that the marketing team can run the campaigns over 100 times a day on typical days and up to 300 during sales. At the same time, you want to keep the load on the Cloud SQL databases to a minimum. What should you do?

A. Create BigQuery connections to both Cloud SQL databases. Use BigQuery federated queries on the two databases and the Google Analytics data on BigQuery to run these queries.

B. Create a job on Apache Spark with Dataproc Serverless to query both Cloud SQL databases and the Google Analytics data on BigQuery for these queries.

C. Create streams in Datastream to replicate the required tables from both Cloud SQL databases to BigQuery for these queries.

D. Create a Dataproc cluster with Trino to establish connections to both Cloud SQL databases and BigQuery, to execute the queries.

Show Suggested Answer

by scaenruy at *Jan. 4, 2024, 11:33 a.m.*

Comments

Type your comment...

Submit

raaad Highly Voted 1 year, 3 months ago

Selected Answer: C

- Datastream: It's a fully managed, serverless service for real-time data replication. It allows to stream data from various sources, including Cloud SQL, into BigQuery.
- Reduced Load on Cloud SQL: By replicating the required tables from both Cloud SQL databases into BigQuery, you minimize the load on the Cloud SQL instances. The marketing team's queries will be run against BigQuery, which is designed to handle high-volume analytics workloads.
- Frequency of Queries: BigQuery can easily handle the high frequency of queries (100 times daily, up to 300 during sales events) due to its powerful data processing capabilities.
- Combining Data Sources: Once the data is in BigQuery, you can efficiently combine it with the Google Analytics data for comprehensive analysis and campaign planning.

upvoted 11 times

SanjeevRoy91 1 year, 1 month ago

Why not A ? Federrated queries will downgrade Cloud SQL perf?

upvoted 1 times

Blackstile Most Recent 1 month, 3 weeks ago

Selected Answer: C

To Replication data, use datastream

upvoted 1 times

987af6b 9 months, 2 weeks ago

Selected Answer: C

Initially I said A, but this question was how I learned about Datastream, which I think would be the better solution in this scenario. So my answer is C

upvoted 3 times

AlizCert 11 months ago

Selected Answer: C

C, noting that federated queries on read replicas would be the ideal solution

upvoted 1 times

joao_01 1 year ago

Its option C.

"Performance. A federated query is likely to not be as fast as querying only BigQuery storage. BigQuery needs to wait for the source database to execute the external query and temporarily move data from the external data source to BigQuery. Also, the source database might not be optimized for complex analytical queries."

So, it will load the Cloud SQL external sources with the queries, impacting performance on those.

Link: https://cloud.google.com/bigquery/docs/federated-queries-intro

upvoted 1 times

datasmg 1 year, 1 month ago

Selected Answer: C

C is make sense

upvoted 1 times

JyoGCP 1 year, 2 months ago

Selected Answer: C

Option C

upvoted 1 times

scaenruy 1 year, 4 months ago

Selected Answer: C

C. Create streams in Datastream to replicate the required tables from both Cloud SQL databases to BigQuery for these queries.

upvoted 3 times

Smakyel79 1 year, 3 months ago

Smakyel79 1 year, 3 months ago

Datastream is a serverless, easy-to-use change data capture (CDC) and replication service. By replicating the necessary tables from the Cloud SQL databases to BigQuery, you can offload the query load from the Cloud SQL databases. The marketing team can then run their queries directly on BigQuery, which is designed for large-scale data analytics. This approach seems to balance both efficiency and performance, minimizing load on the Cloud SQL instances.

upvoted 3 times