← 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 209 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 209

Topic #: 1

[All Professional Data Engineer Questions]

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to copy all the data to a new clustered table. What should you do?

A. Re-create the table using data partitioning on the package delivery date.

B. Implement clustering in BigQuery on the package-tracking ID column.

C. Implement clustering in BigQuery on the ingest date column.

D. Tier older data onto Cloud Storage files and create a BigQuery table using Cloud Storage as an external data source.

**Show Suggested Answer**

by 👤 e70ea9e at *Dec. 30, 2023, 9:31 a.m.*

## Comments

Type your comment...

Submit

⊟ 👤 **desertlotus1211** 3 weeks, 3 days ago

Almost the same at #166:

👍 ↩ 🏳 **upvoted 1 times**

☐ 👤 **apoio.certificacoes.closer** 4 months, 3 weeks ago

You don't need to recreate a table to cluster it, contrary to partitioning, where you have to create a new table with the old data (migration)

> If you alter an existing non-clustered table to be clustered, the existing data is not automatically clustered. Only new data that's stored using the clustered columns is subject to automatic reclustering.
https://cloud.google.com/bigquery/docs/clustered-tables#limitations

👍 ↩ 🏳 **upvoted 1 times**

☐ 👤 **JyoGCP** 8 months, 3 weeks ago

B. Implement clustering in BigQuery on the package-tracking ID column.

👍 ↩ 🏳 **upvoted 1 times**

☐ 👤 **datapassionate** 9 months, 3 weeks ago

B. Implement clustering in BigQuery on the package-tracking ID column.

👍 ↩ 🏳 **upvoted 1 times**

☐ 👤 **Matt_108** 9 months, 3 weeks ago

Definitely B

👍 ↩ 🏳 **upvoted 1 times**

☐ 👤 **MaxNRG** 10 months ago

This looks like Question #166

Option B, implementing clustering in BigQuery on the package-tracking ID column, seems the most appropriate. It directly addresses the query slowdown issue by reorganizing the data in a way that aligns with the analysts' query patterns, leading to more efficient and faster query execution.

👍 ↩ 🏳 **upvoted 3 times**

☐ 👤 **raaad** 10 months, 1 week ago

Answer is B

👍 ↩ 🏳 **upvoted 3 times**

☐ 👤 **e70ea9e** 10 months, 1 week ago

Query Focus: Analysts are interested in geospatial trends within individual package lifecycles. Clustering by package-tracking ID physically co-locates related data, significantly improving query performance for these analyses.

Addressing Slow Queries: Clustering addresses the query slowdown issue by optimizing data organization for the specific query patterns.

Partitioning vs. Clustering:

Partitioning: Divides data into segments based on a column's values, primarily for managing large datasets and optimizing query costs.
Clustering: Organizes data within partitions for faster querying based on specific columns.

👍 ↩ 🏳 **upvoted 4 times**