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 166 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 166

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 implement a change that would improve query performance in BigQuery. What should you do?

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

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

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

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

Show Suggested Answer

by AWSandeep at *Sept. 4, 2022, 10:50 p.m.*

## Comments

Type your comment...

Submit

⊟ zellck Highly Voted 1 year, 11 months ago

**Selected Answer: B**

B is the answer.

https://cloud.google.com/bigquery/docs/clustered-tables
Clustered tables in BigQuery are tables that have a user-defined column sort order using clustered columns. Clustered tables can improve query performance and reduce query costs.

In BigQuery, a clustered column is a user-defined table property that sorts storage blocks based on the values in the clustered columns. The storage blocks are adaptively sized based on the size of the table. A clustered table maintains the sort properties in the context of each operation that modifies it. Queries that filter or aggregate by the clustered columns only scan the relevant blocks based on the clustered columns instead of the entire table or table partition.

upvoted 10 times

**AzureDP900** 1 year, 10 months ago

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

upvoted 1 times

**AlizCert** Most Recent 5 months ago

**Selected Answer: B**

Though I almost fell for D, but the delivery date information is only available on the event(s) that happen after the delivery, but not on the ones before where it will be NULL I guess. The only other option that can make some sense is B, though high cardinality is not recommended for clustering.

upvoted 2 times

**MaxNRG** 10 months, 2 weeks ago

**Selected Answer: B**

B as Clustering the data on the package Id can greatly improve the performance.
Refer GCP documentation - BigQuery Clustered Table:https://cloud.google.com/bigquery/docs/clustered-tables

upvoted 1 times

**MaxNRG** 10 months, 2 weeks ago

Clustering can improve the performance of certain types of queries such as queries that use filter clauses and queries that aggregate data. When data is written to a clustered table by a query job or a load job, BigQuery sorts the data using the values in the clustering columns. These values are used to organize the data into multiple blocks in BigQuery storage. When you submit a query containing a clause that filters data based on the clustering columns, BigQuery uses the sorted blocks to eliminate scans of unnecessary data.
Currently, BigQuery allows clustering over a partitioned table. Use clustering over a partitioned table when:
- Your data is already partitioned on a date, timestamp, or integer column.
- You commonly use filters or aggregation against particular columns in your queries.
Table clustering is possible for tables partitioned by:
- ingestion time
- date/timestamp
- integer range

upvoted 1 times

**MaxNRG** 10 months, 2 weeks ago

In a table partitioned by a date or timestamp column, each partition contains a single day of data. When the data is stored, BigQuery ensures that all the data in a block belongs to a single partition. A partitioned table maintains these properties across all operations that modify it: query jobs, Data Manipulation Language (DML) statements, Data Definition Language (DDL) statements, load jobs, and copy jobs. This requires BigQuery to maintain more metadata than a non-partitioned table. As the number of partitions increases, the amount of metadata overhead increases.

upvoted 1 times

**MaxNRG** 10 months, 2 weeks ago

Although more metadata must be maintained, by ensuring that data is partitioned globally, BigQuery can more accurately estimate the bytes processed by a query before you run it. This cost calculation provides an upper bound on the final cost of the query.
In a clustered table, BigQuery automatically sorts the data based on the values in the clustering columns and organizes them in optimally sized storage blocks. You can achieve more finely grained sorting by creating a table that is clustered and partitioned. A clustered table maintains the sort properties in the context of each operation that modifies it. As a result, BigQuery may not be able to accurately estimate the bytes processed by the query or the query costs. When blocks of data are eliminated during query execution, BigQuery provides a best effort reduction of the query costs.

upvoted 1 times

**Aman47** 10 months, 3 weeks ago

**Selected Answer: B**

Package Tracking mostly contains, geospatial prefixes, Like HK0011, US0022, etc, this can help in clustering.

upvoted 2 times

**kcl10** 1 year, 1 month ago

Selected Answer: D

D is the correct answer

requirements: analyze geospatial trends in the lifecycle of a package

cuz the data of the lifecycle of the package would span across ingest-date-based partition table, it would degrade the performance.

hence, re-partitoning by package delivery date, which is the package initially delivered, would improve the performance when querying such table.

upvoted 4 times

**sdi_studiers** 1 year, 4 months ago

Selected Answer: D

I vote D
Queries to analyze the package lifecycle will cross partitions when using ingest date. Changing this to delivery date will allow a query to full a package's full lifecycle in a single partition.

upvoted 3 times

**cloudmon** 2 years ago

Selected Answer: B

B. https://cloud.google.com/bigquery/docs/clustered-tables

upvoted 1 times

**John_Pongthorn** 2 years, 1 month ago

D is not correct becsuse This Is problem Is The Real Time so ingested date is the same as delivery date.

upvoted 3 times

**kenanars** 2 years, 1 month ago

why not D ?

upvoted 1 times

> **John_Pongthorn** 2 years, 1 month ago
>
> There are several rows that represent movement of life cycle of 1 package-tracking ID
> package delivery date = ingestion date , i suppose
>
> upvoted 1 times
>
> **jkhong** 1 year, 10 months ago
>
> The table has already been partitioned
>
> upvoted 2 times

**pluiedust** 2 years, 1 month ago

Selected Answer: B

B;
As the table has already created with ingest-date partitioning.

upvoted 2 times

**AWSandeep** 2 years, 2 months ago

Selected Answer: B

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

upvoted 1 times