Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 207 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 207
Topic #: 1
[All Professional Data Engineer Questions]

You are collecting IoT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data, filtered by location_id and device_version with the following query:

```
SELECT
    MAX(temperature)
FROM
    acme_iot_data.sensors
WHERE
    create_date > DATE_SUB(CURRENT_DATE(), INTERVAL 7 day)
    AND location_id = "SW1W9TQ"
    AND device_version = "202007r3"
```

You want to optimize your queries for cost and performance. How should you structure your data?

A. Partition table data by create_date, location_id, and device_version.

B. Partition table data by create_date, cluster table data by location_id, and device_version.

C. Cluster table data by create_date, location_id, and device_version.

D. Cluster table data by create_date, partition by location_id, and device_version.

Show Suggested Answer

by e70ea9e at *Dec. 30, 2023, 9:29 a.m.*

Comments

Type your comment...

Submit

JyoGCP 8 months, 3 weeks ago

Selected Answer: B

B. Partition table data by create_date, cluster table data by location_id, and device_version.

upvoted 1 times

datapassionate 9 months, 3 weeks ago

Selected Answer: B

B. Partition table data by create_date, cluster table data by location_id, and device_version.

upvoted 1 times

Matt_108 9 months, 3 weeks ago

Selected Answer: B

B: Partitioning makes date-related querying efficient, clustering will keep relevant data close together and optimize the performance of filters for the cluster columns

upvoted 2 times

MaxNRG 10 months ago

Selected Answer: B

1. Partitioning the data by create_date will allow BigQuery to prune partitions that are not relevant to the query by date.
2. Clustering the data by location_id and device_version within each partition will keep related data close together and optimize the performance of filters on those columns.
This provides both the pruning benefits of partitioning and locality benefits of clustering for filters on multiple columns.
The query provided indicates that the access pattern is primarily based on the most recent data (within the last 7 days), filtered by location_id and device_version. Given this pattern, you would want to optimize your table structure in such a way that queries scanning through the data will process the least amount of data possible to reduce costs and improve performance.

upvoted 3 times

Smakyel79 10 months ago

Selected Answer: B

Only correct answer is B, you can only partition by one field, and you can only cluster on partitioned tables

upvoted 1 times

raaad 10 months, 1 week ago

Selected Answer: B

Answer is B:
- Partitioning the table by create_date allows us to efficiently query data based on time, which is common in access patterns that prioritize recent data.
- Clustering the table by location_id and device_version further organizes the data within each partition, making queries filtered by these columns more efficient and cost-effective.

upvoted 2 times

e70ea9e 10 months, 1 week ago

Selected Answer: B

The best answer is B. Partition table data by create_date, cluster table data by location_id, and device_version.

Here's a breakdown of why this structure is optimal:

Partitioning by create_date:

Aligns with query pattern: Filters for recent data based on create_date, so partitioning by this column allows BigQuery to quickly narrow down the data to scan, reducing query costs and improving performance.
Manages data growth: Partitioning effectively segments data by date, making it easier to manage large datasets and optimize

storage costs.
Clustering by location_id and device_version:

Enhances filtering: Frequently filtering by location_id and device_version, clustering physically co-locates related data within partitions, further reducing scan time and improving performance.

upvoted 2 times