G 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 111 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 111

Topic #: 1

[All Professional Data Engineer Questions]

You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the

Data Science team runs a query filtered on a date column and limited to 30"90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries.

What should you do?

- A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type.
- B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.
- C. Modify your pipeline to maintain the last 3090"€\(\text{\lambda}\) days of data in one table and the longer history in a different table to minimize full table scans over the entire history.
- D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Show Suggested Answer

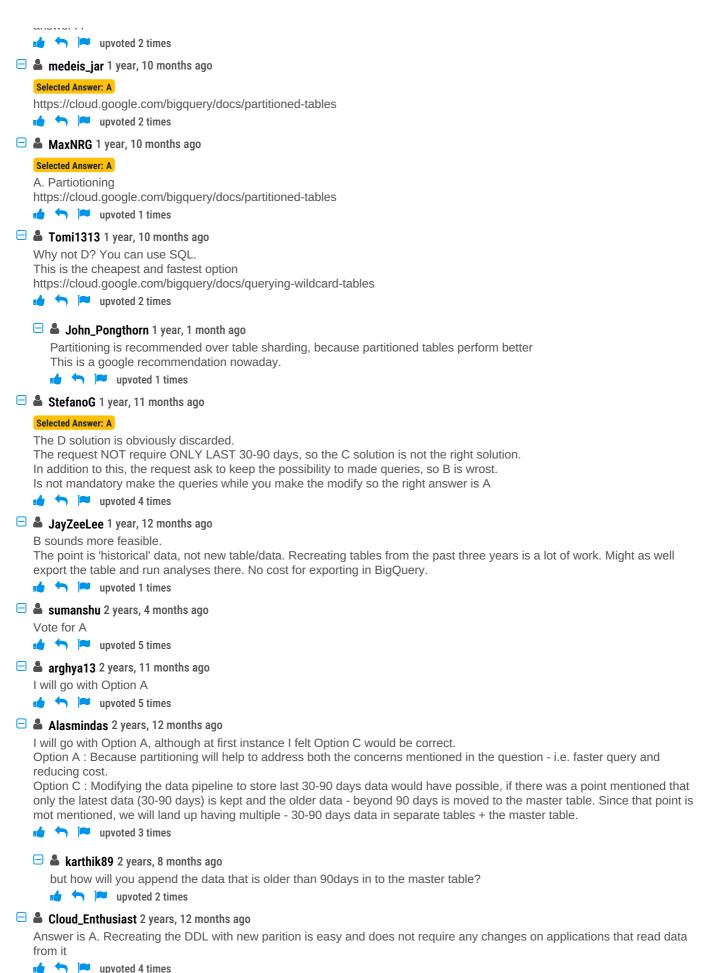
by [deleted] at March 22, 2020, 1:15 p.m.

Comments

	ype your comment
Submit	
	■ [Removed] Highly Voted 10 3 years, 7 months ago should be A □ upvoted 35 times
	♣ [Removed] Highly Voted 3 years, 7 months ago Answer: A Description: Partition is the solution for reducing cost and time ↓ □ upvoted 18 times
	 willbot 3 years, 5 months ago but how would recreating tables with 3 years of data, maintain the ability to conduct sql queries during that time? pupvoted 1 times
0	squishy_fishy 2 years ago Recreating the new table, the old table will still have new data coming, then append the difference to the new table. upvoted 2 times
	■ odacir Most Recent ② 11 months ago Selected Answer: A Answer: A, has no cost to reload the data, Also Partition is the solution for reducing cost and time □ □ □ upvoted 1 times
	Zelick 11 months ago Selected Answer: A A is the answer.
	https://cloud.google.com/bigquery/docs/partitioned-tables A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.
	You can partition BigQuery tables by: - Time-unit column: Tables are partitioned based on a TIMESTAMP, DATE, or DATETIME column in the table.
	A is right upvoted 1 times
	Li is not B in the sense of cost-effective certainly. read below in limitation https://cloud.google.com/bigquery/docs/querying-wildcard-tables#limitations Currently, cached results are not supported for queries against multiple tables using a wildcard even if the Use Cached Results option is checked. If you run the same wildcard query multiple times, you are billed for each query. □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
	▲ John_Pongthorn 1 year, 1 month ago
_	Selected Answer: A
	https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard Partitioning is recommended over table sharding, because partitioned tables perform better upvoted 1 times
	♣ John_Pongthorn 1 year, 1 month ago
	Selected Answer: A A AND D , they are the most likely choiced but the questionn want issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. 1 table may be cheaper so partition is better than wildcarf upvoted 1 times

Selected Answer: A
answer A

🖯 🏜 Didine_22 1 year, 6 months ago



Load full discussion...

