Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 82 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 82
Topic #: 1
[All Professional Data Engineer Questions]

MJTelco Case Study -

Company Overview -
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
✑ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
✑ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments `" development/test, staging, and production `" to meet the needs of

running experiments, deploying new features, and serving production customers.

Business Requirements -
✑ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
✑ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
✑ Provide reliable and timely access to data for analysis from distributed research workers
✑ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -
Ensure secure and efficient transport and storage of telemetry data
Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

A. Create a table called tracking_table and include a DATE column.

B. Create a partitioned table called tracking_table and include a TIMESTAMP column.

C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.

D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Show Suggested Answer

by [deleted] at *March 22, 2020, 6:17 p.m.*

Comments

Type your comment

Type your comment...

Submit

[Removed] Highly Voted 5 years, 1 month ago

Correct - B

upvoted 19 times

SamuelTsch Most Recent 6 months, 2 weeks ago

Selected Answer: B

partitioned table is more performancer than sharded tables

upvoted 1 times

sspsp 1 year, 9 months ago

Selected Answer: B

B, Partition tables in BQ have different cost. If a partition is not modified (DML) for 90 days then cost will be less by 50%, while querying will be efficient since its single large table.

upvoted 1 times

piotrpiskorski 2 years, 5 months ago

Selected Answer: B

always partion large tables

upvoted 1 times

Thierry_1 3 years, 5 months ago

B for sure

upvoted 3 times

nguyenmoon 3 years, 7 months ago

Correct is B

upvoted 3 times

sandipk91 3 years, 9 months ago

Option B for sure

upvoted 2 times

awssp12345 3 years, 10 months ago

https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard - Supports B

upvoted 2 times

sumanshu 3 years, 10 months ago

Vote for 'B' Partitioned Table for Faster Query and Low cost (because it will process less data)

upvoted 1 times

alonsoRios 4 years, 1 month ago

B is correct

upvoted 2 times

fabenavideso 4 years, 4 months ago

Correct - B

upvoted 2 times

ceak 4 years, 4 months ago

should be C

upvoted 1 times

lammingtons 4 years, 3 months ago

They're using BigQuery so partitioning is the better choice here. B

upvoted 3 times

haroldbenites 4 years, 8 months ago

B is correct

upvoted 3 times