Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 39 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 39
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

[illegible] 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.
You need to compose visualizations for operations teams with the following requirements:
✑ The report must include telemetry data from all 50,000 installations for the most resent 6 weeks (sampling once every minute).
✑ The report must not be more than 3 hours delayed from live data.
✑ The actionable report should only show suboptimal links.
✑ Most suboptimal links should be sorted to the top.
✑ Suboptimal links can be grouped and filtered by regional geography.
✑ User response time to load the report must be <5 seconds.
Which approach meets the requirements?

A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.

B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.

C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.

D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a

metric, and then uses a filter expression to show only suboptimal rows in a table.

Show Suggested Answer

by [deleted] at *March 21, 2020, 4:22 a.m.*

Comments

Type your comment...

Submit

[Removed] Highly Voted 4 years, 7 months ago

Answer: D

upvoted 28 times

itche_scratche Highly Voted 4 years, 6 months ago

D; dataflow doesn't connect to datastore, and not really for reporting. BQ, and data studio is a better choice.

upvoted 13 times

ckanaar 1 year, 1 month ago

Dataflow does connect to Datastore, D is still the right answer though.

upvoted 1 times

rtcpost Most Recent 1 year ago

Selected Answer: D

D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Here's why this option is the most suitable:

Google BigQuery is a powerful data warehouse for processing and analyzing large datasets. It can efficiently handle the telemetry data from all 50,000 installations.
Google Data Studio 360 is designed for creating interactive and visually appealing reports and dashboards.
Using Google Data Studio allows you to connect to BigQuery, calculate the required metrics, and apply filters to show only suboptimal links.
It can provide real-time or near-real-time data updates, ensuring that the report is not more than 3 hours delayed from live data.
Google Data Studio can also be used to sort and group suboptimal links and display them based on regional geography.
With the right design, you can ensure that user response time to load the report is less than 5 seconds.
This approach leverages Google's cloud services effectively to meet the specified requirements.

upvoted 2 times

mark1223jkh 5 months, 2 weeks ago

Is Google Data studio 360 a product now?

upvoted 1 times

theseawillclaim 1 year, 3 months ago

Selected Answer: D

Why bother with a custom GAE app when you have Data Studio?

upvoted 3 times

DGames 1 year, 10 months ago

Selected Answer: C

Its think answer would be C because of telemetry data and response time is <5 second that force me to think about datastore,

upvoted 2 times

willymac2 2 years, 4 months ago

I believe the answer is C.
First requirement is that it must be a visualisation with, so A and B do not work (create a table and a spreadsheet).
Now the second constraint which I believe is important is that the report MUST load in less than 5 seconds. But we do not know how complex the metric computation is, thus I cannot assume that we can compute it when we want to load the report, making me think that it be must be pre-computed. Thus option D cannot work as it create the metric AFTER querying the

data (we are also not sure if we can really compute it in a query).

upvoted 4 times

gudguy1a 1 year, 2 months ago

Ummm, sorry @willymac2, but you have to account for size and growth which datastore cannot scale to.
Then, you have to worry about sub-second response time and datastore cannot do that as well as BigQuery...

upvoted 2 times

Raj0123 2 years, 5 months ago

Answer D

upvoted 1 times

CedricLP 2 years, 6 months ago

Selected Answer: D

DataStudio and BQ are the simpliest way to do it

upvoted 1 times

devric 2 years, 7 months ago

Selected Answer: D

They also can activate BI Engine feature to improve the response time.

upvoted 1 times

sraakesh95 2 years, 9 months ago

Selected Answer: D

D: Usually when a reporting tool is involved for GCP, DataStudio mostly goes by default due to it's no cost analytics and BigQuery joins it due to it's OLAP nature and the wonderful integration provided by GCP for these 2

upvoted 2 times

medeis_jar 2 years, 10 months ago

Selected Answer: D

as explained by JayZeeLee

upvoted 1 times

JayZeeLee 3 years ago

D.
A and B are incorrect, because Google Sheets are not the best fit to handle large amount of data.
C may work, but it requires building an application which equates to more work.
D is more efficient, therefore a better option.

upvoted 4 times

wubston 1 year, 11 months ago

I can't think of a single compelling reason to go with anything but D, given the scope definition in the question brief.

upvoted 1 times

Chelseajcole 3 years ago

Visualization = Data Studio 360

upvoted 1 times

Chelseajcole 3 years ago

Next question give you the answer: Question #40. They using Data Studio 360 and Bigquery as source

upvoted 1 times

anji007 3 years ago

Ans: D

upvoted 1 times

sumanshu 3 years, 3 months ago

Vote for D

upvoted 3 times

zosoabi 3 years, 5 months ago

just check the next question (#40) to get an idea about correct answer

upvoted 3 times

BhupiSG 3 years, 7 months ago

D
Correct

upvoted 2 times

Load full discussion...