Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 40 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 40
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.
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)

A. Ensure all the tables are included in global dataset.

B. Ensure each table is included in a dataset for a region.

C. Adjust the settings for each table to allow a related region-based security group view access.

D. Adjust the settings for each view to allow a related region-based security group view access.

E. Adjust the settings for each dataset to allow a related region-based security group view access.

Show Suggested Answer

by [deleted] at *March 21, 2020, 4:25 a.m.*

Comments

Type your comment...

Submit

[Removed] Highly Voted 5 years, 1 month ago

Answer: B E

upvoted 48 times

shanjin14 Highly Voted 3 years, 11 months ago

C is correct starting 2020, as BigQuery come with table level access control
https://cloud.google.com/blog/products/data-analytics/introducing-table-level-access-controls-in-bigquery

upvoted 36 times

samstar4180 3 years, 8 months ago

Yes, the correct answer should be BC - since we can have table-level access and each region represents a table.

upvoted 11 times

willyunger Most Recent 1 month, 2 weeks ago

Selected Answer: BE

Less effort to assign access by dataset, not by table or view.

upvoted 1 times

iooj 9 months, 1 week ago

Selected Answer: BE

C doesn't make sense, if you have already selected B

upvoted 1 times

suwalsageen12 11 months, 2 weeks ago

Selected Answer: BE

If I choose Option E, then Option C and D are eliminated because, Once you provide the access level in dataset for the use group, it applies for both Table and view.

Now for remaining options A and B.
The Question itself has stated to include the table region wise in dataset, So Option A is eliminated.
So, B and E are Correct answers.

upvoted 3 times

niru12376 1 year, 2 months ago

BE... If we've already split the tables into regions via datasets then why give regional access through tables again.. If not, then AC but definitely not BC.. Please help

upvoted 6 times

rocky48 1 year, 5 months ago

Answer : BE

B. Ensure each table is included in a dataset for a region.
This means that you should organize your data in BigQuery into separate datasets, one for each region. Each dataset contains the tables specific to that region. This ensures that data is segregated by region.
E. Adjust the settings for each dataset to allow a related region-based security group view access.

upvoted 3 times

rtcpost 1 year, 6 months ago

Selected Answer: BE

B. Ensure each table is included in a dataset for a region.

This means that you should organize your data in BigQuery into separate datasets, one for each region. Each dataset contains the tables specific to that region. This ensures that data is segregated by region.
E. Adjust the settings for each dataset to allow a related region-based security group view access.

You should set the access controls at the dataset level in BigQuery. This means configuring access permissions for each dataset based on regional security groups. This way, you can enforce the regional access policy to the data, ensuring that users from different regions can only access the data associated with their region.
Option A is not necessary because you don't need to include all the tables in a global dataset. Segregating data into region-specific datasets is a better approach for enforcing access controls.

Options C and D are not typical actions in BigQuery. Access control and permissions are usually managed at the dataset

Options C and D are not typical actions in BigQuery. Access control and permissions are usually managed at the dataset level, and you can grant access to specific groups at that level.

upvoted 1 times

Mathew106 1 year, 9 months ago

It's B and E or B and C. However, B and E makes some more sense because if you have one dataset for each region and they need to access the data for each region then why not allow them access to the whole dataset? What if you want to add other supplementary tables later? If you did that on a table level you would have to add access to every table separately.

Still, I think both are valid because we don't have any extra requirement, but B E makes more sense.

upvoted 4 times

Jarek7 2 years ago

Selected Answer: C

The intended answer was for sure BE. If C or D would be the right answers there is absolutely no reason to do B, right? Why should you put each table into separate dataset if you then set the accesss on table/view level? What is more the question is about tables not views, so I have no idea why would anybody take D.
The issue is that this question is out of date and now the right answer would be sole C.

upvoted 4 times

Oleksandr0501 2 years ago

The two actions that should be taken are B and E.
B. Ensure each table is included in a dataset for a region: By creating separate datasets for each region and including only the tables associated with that region, you can enforce the regional access policy.

E. Adjust the settings for each dataset to allow a related region-based security group view access: By adjusting the settings for each dataset to allow only the related region-based security group view access, you can ensure that employees can only view data associated with their region.

A is incorrect because including all tables in a global dataset would not enforce the regional access policy.

C is incorrect because adjusting the settings for each table is not a scalable solution, especially as the number of tables grows.

D is incorrect because adjusting the settings for each view does not ensure that employees can only view data associated with their region.

upvoted 2 times

sjtesla 2 years ago

Selected Answer: BC

B: Location is on dataset level: https://cloud.google.com/bigquery/docs/datasets#dataset_limitations
C: IAM can be set on table level

upvoted 2 times

Lestrang 2 years, 3 months ago

Guys,
there are 2 possible combinations
If you think that each table represents a region, then they should all be in a global dataset and you should apply table access control to them.
So A+C

Otherwise you would put each table in a regional dataset, and apply access control to the dataset. Why would you create a dataset for the purpose of controlling regional access, and then only apply the controls to a table inside it? that is not extensible in the future.
Anyway create dataset+access control for dataset (B+E) is also valid.

Which to choose? I dont know.

upvoted 3 times

PolyMoe 2 years, 3 months ago

Selected Answer: BE

First put tables in region-dedicated dataset (B)
Then, ensure access control at dataset level (by creating region-based security groups) (E)

upvoted 2 times

korntewin 2 years, 3 months ago

Selected Answer: AC

I would vote for AC. As we already split the table for each region, why do we need to split the dataset per region?
Furthermore, the access control will be provided to the users based on table level anyway.

upvoted 3 times

korntewin 2 years, 3 months ago

Oh, the location should be specified in the dataset level! Then, the dataset should be splitted by region, my bad!

upvoted 1 times

MisuLava 2 years, 6 months ago

Selected Answer: BE

if you create table-level access control and grant it to different groups for different tables, what is the point of putting tables in different datasets and different regions?
So i choose BE

upvoted 7 times

svkds 2 years, 7 months ago

Selected Answer: BC

BigQuery come with table level access control. Since we can have table-level access and each region represents a table, B & C is correct answer.

upvoted 2 times

Load full discussion...