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 297 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 297
Topic #: 1
[All Professional Data Engineer Questions]

You migrated your on-premises Apache Hadoop Distributed File System (HDFS) data lake to Cloud Storage. The data scientist team needs to process the data by using Apache Spark and SQL. Security policies need to be enforced at the column level. You need a cost-effective solution that can scale into a data mesh. What should you do?

A. 1. Deploy a long-living Dataproc cluster with Apache Hive and Ranger enabled.
2. Configure Ranger for column level security.
3. Process with Dataproc Spark or Hive SQL.

B. 1. Define a BigLake table.
2. Create a taxonomy of policy tags in Data Catalog.
3. Add policy tags to columns.
4. Process with the Spark-BigQuery connector or BigQuery SQL.

C. 1. Load the data to BigQuery tables.
2. Create a taxonomy of policy tags in Data Catalog.
3. Add policy tags to columns.
4. Process with the Spark-BigQuery connector or BigQuery SQL.

D. 1. Apply an Identity and Access Management (IAM) policy at the file level in Cloud Storage.
2. Define a BigQuery external table for SQL processing.
3. Use Dataproc Spark to process the Cloud Storage files.

Show Suggested Answer

by [redacted] at Jan. 4, 2024, 12:15 p.m.

by scaenruy at Jan. 4, 2024, 12:15 p.m.

Comments

Type your comment...

Submit

raaad Highly Voted 9 months, 4 weeks ago

Selected Answer: B

- BigLake Integration: BigLake allows you to define tables on top of data in Cloud Storage, providing a bridge between data lake storage and BigQuery's powerful analytics capabilities. This approach is cost-effective and scalable.
- Data Catalog for Governance: Creating a taxonomy of policy tags in Google Cloud's Data Catalog and applying these tags to specific columns in your BigLake tables enables fine-grained, column-level access control.
- Processing with Spark and SQL: The Spark-BigQuery connector allows data scientists to process data using Apache Spark directly against BigQuery (and BigLake tables). This supports both Spark and SQL processing needs.
- Scalability into a Data Mesh: BigLake and Data Catalog are designed to scale and support the data mesh architecture, which involves decentralized data ownership and governance.

upvoted 17 times

JyoGCP Most Recent 8 months, 2 weeks ago

Selected Answer: B

Going with 'B' based on the comments

upvoted 1 times

Matt_108 9 months, 3 weeks ago

Selected Answer: B

Option B, agree with comments explanation

upvoted 1 times

Jordan18 10 months ago

Selected Answer: B

BigLake leverages existing Cloud Storage infrastructure, eliminating the need for a dedicated Dataproc cluster, reducing costs significantly.

upvoted 4 times

scaenruy 10 months ago

Selected Answer: C

C.
1. Load the data to BigQuery tables.
2. Create a taxonomy of policy tags in Data Catalog.
3. Add policy tags to columns.
4. Process with the Spark-BigQuery connector or BigQuery SQL.

upvoted 1 times

raaad 9 months, 4 weeks ago

- Option B offers a serverless approach that integrates Cloud Storage (as a data lake), BigLake (for table definition), Data Catalog (for data mesh), and BigQuery (for analytics), all of which are essential components of a flexible, scalable, and secure data platform.

upvoted 7 times

EXAMTOPICS

Platform

Home

All Exams

Examtopics PRO
Training Courses

© 2024 ExamTopics