Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 177 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 177
Topic #: 1
[All Professional Data Engineer Questions]

You want to rebuild your batch pipeline for structured data on Google Cloud. You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run. To expedite development and pipeline run time, you want to use a serverless tool and SOL syntax. You have already moved your raw data into Cloud Storage. How should you build the pipeline on Google Cloud while meeting speed and processing requirements?

A. Convert your PySpark commands into SparkSQL queries to transform the data, and then run your pipeline on Dataproc to write the data into BigQuery.

B. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated quenes from BigQuery for machine learning.

C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table.

D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery.

Show Suggested Answer

by AWSandeep at *Sept. 2, 2022, 8:30 p.m.*

Comments

Type your comment...

Submit

devaid Highly Voted 2 years, 1 month ago

Selected Answer: C

The question is C but not because the SQL Syntax, as you can perfectly use SparkSQL on Dataproc reading files from GCS. It's because the "serverless" requirement.

upvoted 14 times

GCP001 Most Recent 9 months, 3 weeks ago

Selected Answer: A

A) Looks more suitable , serverless approach for handling and performance.

upvoted 2 times

MaxNRG 10 months, 2 weeks ago

Selected Answer: C

Option C is the best approach to meet the stated requirements. Here's why:

BigQuery SQL provides a fast, scalable, and serverless method for transforming structured data, easier to develop than PySpark.
Directly ingesting the raw Cloud Storage data into BigQuery avoids needing an intermediate processing cluster like Dataproc.
Transforming the data via BigQuery SQL queries will be faster than PySpark, especially since the data is already loaded into BigQuery.
Writing the transformed results to a new BigQuery table keeps the original raw data intact and provides a clean output.
So migrating to BigQuery SQL for transformations provides a fully managed serverless architecture that can significantly expedite development and reduce pipeline runtime versus PySpark. The ability to avoid clusters and conduct transformations completely within BigQuery is the most efficient approach here.

upvoted 3 times

MoeHaydar 1 year, 3 months ago

Selected Answer: C

Note: Dataproc by itself is not serverless
https://cloud.google.com/dataproc-serverless/docs/overview

upvoted 3 times

Prudvi3266 1 year, 6 months ago

Selected Answer: C

because of serverless nature

upvoted 3 times

musumusu 1 year, 8 months ago

Answer C: need to setup SQL based job means transformation in not very complex. And Biqquery sql are faster than spark sql context. (google claims)
However, i will make a test by myself to check it.

upvoted 1 times

maci_f 1 year, 9 months ago

Selected Answer: A

In the GCP Machine Learning Engineer practice question (Q4) there's the same question with similar answers and the correct answer is A since B "is incorrect, here transformation is done on Cloud SQL, which wouldn't scale the process" and C "is incorrect as this process wouldn't scale the data transformation routine. And, it is always better to transform data during ingestion": https://medium.com/@gcpguru/google-google-cloud-professional-machine-learning-engineer-practice-questions-part-1-3ee4a2b3f0a4

upvoted 2 times

evanfebrianto 1 year, 5 months ago

Dataproc is not a serverless tool unless it mentions "Dataproc Serverless" explicitly.

upvoted 2 times

Atnafu 1 year, 11 months ago

C
D-is incorrect because you are rebuild your batch pipeline for structured data on Google Cloud.

upvoted 1 times

Atnafu 1 year, 11 months ago

A could be answer if it was Dataproc serverless and no conversion of code. Dp serverless support: scala,pyspark,sparksql and SparkR

upvoted 2 times

TNT87 2 years, 1 month ago

Selected Answer: C

This same question is there on Google's Professional Machine Learning Engineer, Question 4
Answer is C.

upvoted 4 times

Wasss123 2 years, 1 month ago

Selected Answer: C

I choose C
BigQuery SQL is more performant but more expensive. Here, it's a performance issue (time reduction)
Source : https://medium.com/paypal-tech/comparing-bigquery-processing-and-spark-dataproc-4c90c10e31ac

upvoted 2 times

John_Pongthorn 2 years, 1 month ago

C is the most likely , bigquery is severless and sql
D is dataflow severless but it is wrong at using python sdk but using sql beam rthen it will be correct

upvoted 1 times

TNT87 2 years, 1 month ago

Answer C

upvoted 2 times

ducc 2 years, 2 months ago

Selected Answer: A

A
- You have to maintain PySpark Code -> Proc

upvoted 1 times

ducc 2 years, 2 months ago

After thinking a while, I think the question is not clear enough. To be honest

upvoted 1 times

ducc 2 years, 2 months ago

A or C. I go for C because they said they want to use SQL syntax...

upvoted 1 times

AWSandeep 2 years, 2 months ago

Selected Answer: C

C. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table.

Keys: "Serverless" and "SQL"

upvoted 3 times

ducc 2 years, 2 months ago

The question said "use SQL syntax"
C might still correct

upvoted 1 times

AWSandeep 2 years, 2 months ago

Changing answer to A as this is a new question referring to Dataproc Serverless. Dataproc Serverless for Spark batch workloads supports Spark SQL. Why modify ETL to ELT and convert PySpark to BigQuery SQL when it can be similar to a lift-and-shift?

upvoted 3 times

Atnafu 1 year, 11 months ago

Dataproc is diffrent than Dataproc Serveless. This question is talking about dataproc.
By the way dp serverless support both pyspark and sparkSql no need of conversion.
C is best answer

upvoted 3 times