EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 190 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 190

Topic #: 1

[All Professional Data Engineer Questions]

You are loading CSV files from Cloud Storage to BigQuery. The files have known data quality issues, including mismatched data types, such as STRINGs and
INT64s in the same column, and inconsistent formatting of values such as phone numbers or addresses. You need to create the data pipeline to maintain data quality and perform the required cleansing and transformation. What should you do?

A. Use Data Fusion to transform the data before loading it into BigQuery.

B. Use Data Fusion to convert the CSV files to a self-describing data format, such as AVRO, before loading the data to BigQuery.

C. Load the CSV files into a staging table with the desired schema, perform the transformations with SQL, and then write the results to the final destination table.

D. Create a table with the desired schema, load the CSV files into the table, and perform the transformations in place using SQL.

Show Suggested Answer

by AWSandeep at *Sept. 2, 2022, 11:15 p.m.*

Comments

Type your comment...

Submit

saurabhsingh4k Highly Voted 1 year, 10 months ago

Selected Answer: A

I'm kinda inclined towards C as SQL seems a powerful option to treat this kind of use case.

Also, I didn't get how the transformations mentioned on this page will help to clean the data (https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported_transformations)

But I guess using Wrangler plugin, this kind of stuff can be done on DataFusion, also the question talks about an pipeline, so A is the final choice.

upvoted 6 times

MaxNRG Most Recent 10 months, 2 weeks ago

Selected Answer: A

Data Fusion's advantages:

Visual interface: Offers a user-friendly interface for designing data pipelines without extensive coding, making it accessible to a wider range of users.
Built-in transformations: Includes a wide range of pre-built transformations to handle common data quality issues, such as:
Data type conversions
Data cleansing (e.g., removing invalid characters, correcting formatting)
Data validation (e.g., checking for missing values, enforcing constraints)
Data enrichment (e.g., adding derived fields, joining with other datasets)
Custom transformations: Allows for custom transformations using SQL or Java code for more complex cleaning tasks.
Scalability: Can handle large datasets efficiently, making it suitable for processing CSV files with potential data quality issues.
Integration with BigQuery: Integrates seamlessly with BigQuery, allowing for direct loading of transformed data.

upvoted 3 times

> **MaxNRG** 10 months, 2 weeks ago
>
> Why other options are less suitable:
>
> B. Converting to AVRO: While AVRO is a self-describing format, it doesn't inherently address data quality issues. Transformations would still be needed, and Data Fusion provides a more comprehensive environment for this.
> C. Staging table: Requires manual SQL transformations, which can be time-consuming and error-prone for large datasets with complex data quality issues.
> D. Transformations in place: Modifying data directly in the destination table can lead to data loss or corruption if errors occur. It's generally safer to keep raw data intact and perform transformations separately.
> By using Data Fusion, you can create a robust and efficient data pipeline that addresses data quality issues upfront, ensuring that only clean and consistent data is loaded into BigQuery for accurate analysis and insights.
>
> upvoted 1 times

squishy_fishy 1 year ago

The answer is C. That is what we do at work. We have landing/staging table, sort table and deliver table,

upvoted 4 times

> **squishy_fishy** 1 year ago
>
> Okay, second thought, it is asking for a pipeline, so the answer should be A. At work, we use dataflow inside the composer to build a pipeline injecting data into landing/staging table, then transform/clean data in the sort table, then send the cleaned data to deliver table.
>
> upvoted 4 times

phidelics 1 year, 4 months ago

Selected Answer: A

Keyword: Data Pipeline

upvoted 4 times

mialll 1 year, 5 months ago

Selected Answer: A

same as @saurabhsingh4k

upvoted 2 times

Adswerve 1 year, 6 months ago

Selected Answer: C

C is the right answer. Do ELT in BigQuery. Data Fusion is not the right too for this job.

upvoted 4 times

musumusu 1 year, 8 months ago

musumusu 1 year, 8 months ago

Answer C,
Datafusion is costly and current transformation is just a cast transformation in a column.
I guess no one wanna pay for datafusion for this little transformation and Staging table processing handles such minor cleaning.

upvoted 4 times

maci_f 1 year, 9 months ago

Selected Answer: A

Data Fusion enables changing the data type directly as shown in this lab:
https://www.cloudskillsboost.google/focuses/25335?parent=catalog
Wrangler is the feature to enable that, as already mentioned: https://stackoverflow.com/questions/58699872/google-cloud-data-fusion-how-to-change-datatype-from-string-to-date

upvoted 4 times

Mike422 1 year, 10 months ago

Apparently chatGPT thinks C is the correct answer just sayin (for the same reason that @saurabhsingh4k wrote)

upvoted 2 times

Atnafu 1 year, 10 months ago

A
https://cloud.google.com/data-fusion/docs/concepts/overview#:~:text=The%20Cloud%20Data%20Fusion%20web%20UI%20lets%20you%20to%20build%20scalable%20data%20integration%20solutions%20to%20clean%2C%20prepare%2C%20blend%2C%20transfer%2C%20and%20transform%20data%2C%20without%20having%20to%20manage%20the%20infrastructure.

upvoted 1 times

zellck 1 year, 11 months ago

Selected Answer: A

A is the answer.

https://cloud.google.com/data-fusion/docs/concepts/overview
loud Data Fusion is a fully managed, cloud-native, enterprise data integration service for quickly building and managing data pipelines.

The Cloud Data Fusion web UI lets you to build scalable data integration solutions to clean, prepare, blend, transfer, and transform data, without having to manage the infrastructure.

upvoted 4 times

AzureDP900 1 year, 10 months ago

thx for sharing link

upvoted 1 times

samirzubair 1 year, 11 months ago

The Correct Ans is C

upvoted 2 times

jkhong 1 year, 10 months ago

although this is my preferred answer. this doesn’t satisfy how this becomes a pipeline.

upvoted 1 times

hiromi 1 year, 11 months ago

Selected Answer: A

Data Fusion

upvoted 1 times

TNT87 2 years, 1 month ago

Selected Answer: A

Ans A
https://cloud.google.com/data-fusion/docs/concepts/transformation-pushdown#supported_transformations

upvoted 1 times

ducc 2 years, 2 months ago

Selected Answer: A

A is correct for me

upvoted 1 times

AWSandeep 2 years, 2 months ago

Selected Answer: A

A. Use Data Fusion to transform the data before loading it into BigQuery.

A. Use Data Fusion to transform the data before loading it into BigQuery.

upvoted 1 times

Platform

- Home
- All Exams
- Examtopics PRO
- Training Courses