

🔗 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 5 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 5

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

- A. Use federated data sources, and check data in the SQL query.
- B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
- C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
- D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Show Suggested Answer

by [deleted] at March 15, 2020, 8:14 a.m.

Comments

Type your comment...

Submit


 **Radhika7983** Highly Voted 4 years, 6 months ago

The answer is D. An ETL pipeline will be implemented for this scenario. Check out handling invalid inputs in cloud data flow

<https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow>

ParDos . . . and don'ts: handling invalid inputs in Dataflow using Side Outputs as a "Dead Letter" file


   upvoted 15 times

 **jkhong** 2 years, 5 months ago

The sources you've provided cannot be accessed. Here is an updated best practice.

https://cloud.google.com/architecture/building-production-ready-data-pipelines-using-dataflow-developing-and-testing#use_dead_letter_queues

   upvoted 5 times

 **nadavw** 8 months, 3 weeks ago

<https://cloud.google.com/dataflow/docs/guides/write-to-bigquery#:~:text=It%27s%20a%20good%20practice%20to%20send%20the%20errors%20to%20a%20dead%20letter%20queue%20or%20table%2C%20for%20later%20processing.%20For%20more%20information%20about%20this%20pattern%2C%20see%20BigQueryIO%20dead%20letter%20pattern.>

It's a good practice to send the errors to a dead-letter queue or table, for later processing. For more information about this pattern, see BigQueryIO dead letter pattern.

   upvoted 1 times

 **fire558787** Highly Voted 3 years, 8 months ago

Disagree a bit here. Could well be A. In one Coursera video course (<https://www.coursera.org/learn/batch-data-pipelines-gcp/lecture/SkDus/how-to-carry-out-operations-in-bigquery>), they do have a video about when to just use an SQL query to find wrong data without creating a Dataflow pipeline. The question says "SQL" as a language, not Cloud SQL as a service. Federated Sources is great because you can federate a CSV file in GCS with BigQuery. From the video: "In this section, we'll take a look at exactly how BigQuery can help with some of those data quality issues we just described. Let's start with validity, what do we mean by invalid? It can mean things like corrupted data maybe data that is missing a timestamp"

   upvoted 5 times


 **kelvinksau** 3 years, 8 months ago

<https://cloud.google.com/bigquery/external-data-sources>

Use cases for external data sources include:

For ETL workloads, loading and cleaning your data in one pass and writing the cleaned result into BigQuery storage. Joining BigQuery tables with frequently changing data from an external data source. By querying the external data source directly, you don't need to reload the data into BigQuery storage every time it changes.

   upvoted 2 times

 **willyunger** Most Recent 1 month, 2 weeks ago

Selected Answer: D

"you want to keep the data"

   upvoted 1 times

 **Ahamada** 2 months, 1 week ago

Selected Answer: D

You should transform the raw data by eliminate the error before analyse.

   upvoted 1 times

 **rocky48** 7 months, 1 week ago

Selected Answer: D

Option A is incorrect because federated data sources do not provide any data validation or cleaning capabilities and you'll have to do it on the SQL query, which could slow down the performance.

Option B is incorrect because Stackdriver monitoring can only monitor the performance of the pipeline, but it can't handle corrupted or incorrectly formatted data.

Option C is incorrect because using gcloud CLI and setting max_bad_records to 0 will ignore the corrupted or incorrectly formatted data and continue the load process, this will lead to incorrect analysis.

Answer D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

   upvoted 4 times

 **rtcpst** 7 months, 1 week ago

Selected Answer: D

Google Cloud Dataflow allows you to create a data pipeline that can preprocess and transform data before loading it into

Google Cloud Dataflow allows you to create a data pipeline that can preprocess and transform data before loading it into BigQuery. This approach will enable you to handle problematic rows, push them to a dead-letter table for later analysis, and load the valid data into BigQuery.

Option A (using federated data sources and checking data in the SQL query) can be used but doesn't directly address the issue of handling corrupted or incorrectly formatted rows.

Options B and C are not the best choices for handling data quality and error issues. Enabling monitoring and setting `max_bad_records` to 0 in BigQuery may help identify errors but won't store the problematic rows for further analysis, and it might prevent loading any data with issues, which may not be ideal.

   upvoted 2 times

  **samdhimal** 7 months, 1 week ago

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

By running a Cloud Dataflow pipeline to import the data, you can perform data validation, cleaning and transformation before it gets loaded into BigQuery. Dataflow allows you to handle corrupted or incorrectly formatted rows by pushing them to another dead-letter table for analysis. This way, you can ensure that only clean and correctly formatted data is loaded into BigQuery for analysis.

   upvoted 2 times

  **samdhimal** 2 years, 3 months ago

Option A is incorrect because federated data sources do not provide any data validation or cleaning capabilities and you'll have to do it on the SQL query, which could slow down the performance.

Option B is incorrect because Stackdriver monitoring can only monitor the performance of the pipeline, but it can't handle corrupted or incorrectly formatted data.

Option C is incorrect because using `gcloud` CLI and setting `max_bad_records` to 0 will ignore the corrupted or incorrectly formatted data and continue the load process, this will lead to incorrect analysis.

   upvoted 5 times

  **hamza101** 1 year, 9 months ago

for Option C i think when setting `max_bad_records` to 0 this will prevent the loading to be achieved since the condition will cut off the loading if we have at least 1 corrupted row


   upvoted 2 times

  **RT_G** 7 months, 1 week ago

Selected Answer: D

All other options only alert or error out bad data. As the question requires, option D sends bad data to the dead letter table for further analysis while valid data is loaded to the table

   upvoted 1 times

  **vaga1** 1 year, 11 months ago

Selected Answer: D

Agreed: D

   upvoted 1 times

  **odiez3** 2 years, 1 month ago

D because you need Transform the data

   upvoted 1 times

  **Morock** 2 years, 2 months ago

Selected Answer: D

D. The question is asking pipeline, then let's build a pipeline.

   upvoted 3 times

  **Besss** 2 years, 6 months ago

Selected Answer: D

Agreed: D

   upvoted 1 times

  **Dip1994** 2 years, 9 months ago

The correct answer is D

   upvoted 1 times

  **Arkon88** 3 years, 2 months ago

Selected Answer: D

Correct - D (as we need to create Pipeline) which possible via 'D'

   upvoted 1 times

  **MaxNRG** 3 years, 5 months ago

Looks like D, with C you will not import anything, stackdriver alerts will not help you with this and with federated resources you won't know what happened with those bad records. D is the most complete one.
<https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow>

   upvoted 3 times

  **anji007** 3 years, 6 months ago

Ans: D

   upvoted 1 times

  **nickozz** 3 years, 7 months ago

D seems to be correct. explained here how combined with Pub/Sub, this can be achieved.
<https://cloud.google.com/pubsub/docs/handling-failures>

   upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses



© 2024 ExamTopics