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 263 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 263
Topic #: 1
[All Professional Data Engineer Questions]

You maintain ETL pipelines. You notice that a streaming pipeline running on Dataflow is taking a long time to process incoming data, which causes output delays. You also noticed that the pipeline graph was automatically optimized by Dataflow and merged into one step. You want to identify where the potential bottleneck is occurring. What should you do?

A. Insert a Reshuffle operation after each processing step, and monitor the execution details in the Dataflow console.

B. Insert output sinks after each key processing step, and observe the writing throughput of each block.

C. Log debug information in each ParDo function, and analyze the logs at execution time.

D. Verify that the Dataflow service accounts have appropriate permissions to write the processed data to the output sinks.

Show Suggested Answer

by scaenruy at *Jan. 3, 2024, 6:09 p.m.*

Comments

Type your comment...

Submit

⊟ raaad Highly Voted 👍 1 year, 4 months ago

Selected Answer: A

- The Reshuffle operation is used in Dataflow pipelines to break fusion and redistribute elements, which can sometimes help improve parallelization and identify bottlenecks.
- By inserting Reshuffle after each processing step and observing the pipeline's performance in the Dataflow console, you can potentially identify stages that are disproportionately slow or stalled.
- This can help in pinpointing the step where the bottleneck might be occurring.

upvoted 9 times

srivastavas08 1 year, 2 months ago

ince we don't know for sure if fusion is the culprit, detailed debug logging is still the top choice to find the precise slow operation(s).

upvoted 1 times

Blackstile Most Recent 1 month, 3 weeks ago

Selected Answer: A

Reshuffle is the key.

upvoted 1 times

m_a_p_s 4 months, 3 weeks ago

Selected Answer: A

Looks like A. However, this option does not provide any option of identifying the underlying cause.
https://cloud.google.com/dataflow/docs/pipeline-lifecycle#prevent_fusion

upvoted 1 times

f6bc4a0 7 months ago

Selected Answer: B

B identifies where the problem lies.

upvoted 1 times

JyoGCP 1 year, 2 months ago

Selected Answer: A

Option A

upvoted 1 times

srivastavas08 1 year, 2 months ago

It should be C

upvoted 2 times

tibuenoc 1 year, 3 months ago

Selected Answer: B

The best option is B
Because create additional output to capturing and processing error data, will get error each step that allows you to observe the writing throughput of each block, which can help identify specific processing steps causing bottlenecks.

Option A also is valid but can not directly address all bottlenecks, especially if the graph was merged.

upvoted 1 times

Sofiia98 1 year, 3 months ago

Selected Answer: A

From the Dataflow documentation: "There are a few cases in your pipeline where you may want to prevent the Dataflow service from performing fusion optimizations. These are cases in which the Dataflow service might incorrectly guess the optimal way to fuse operations in the pipeline, which could limit the Dataflow service's ability to make use of all available workers.
You can insert a Reshuffle step. Reshuffle prevents fusion, checkpoints the data, and performs deduplication of records. Reshuffle is supported by Dataflow even though it is marked deprecated in the Apache Beam documentation."

upvoted 4 times

scaenruy 1 year, 4 months ago

Selected Answer: A

A. Insert a Reshuffle operation after each processing step, and monitor the execution details in the Dataflow console.

upvoted 2 times