← 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 254 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 254

Topic #: 1

[All Professional Data Engineer Questions]

You are running a Dataflow streaming pipeline, with Streaming Engine and Horizontal Autoscaling enabled. You have set the maximum number of workers to 1000. The input of your pipeline is Pub/Sub messages with notifications from Cloud Storage. One of the pipeline transforms reads CSV files and emits an element for every CSV line. The job performance is low, the pipeline is using only 10 workers, and you notice that the autoscaler is not spinning up additional workers. What should you do to improve performance?

A. Enable Vertical Autoscaling to let the pipeline use larger workers.

B. Change the pipeline code, and introduce a Reshuffle step to prevent fusion.

C. Update the job to increase the maximum number of workers.

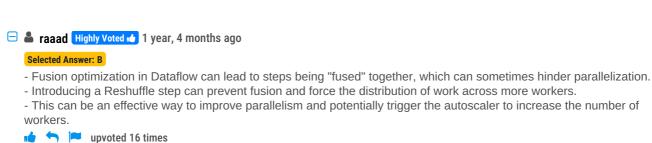D. Use Dataflow Prime, and enable Right Fitting to increase the worker resources.

Show Suggested Answer

by 👤 **scaenruy** at *Jan. 3, 2024, 4:34 p.m.*

Comments

Type your comment...

Submit

- **raaad** `Highly Voted` 👍 1 year, 4 months ago

Selected Answer: B

- Fusion optimization in Dataflow can lead to steps being "fused" together, which can sometimes hinder parallelization.
- Introducing a Reshuffle step can prevent fusion and force the distribution of work across more workers.
- This can be an effective way to improve parallelism and potentially trigger the autoscaler to increase the number of workers.

👍 ↩ 🚩 upvoted 16 times

- **meh_33** `Most Recent ⊘` 8 months, 4 weeks ago

Selected Answer: B

https://cloud.google.com/dataflow/docs/pipeline-lifecycle#prevent_fusion

👍 ↩ 🚩 upvoted 1 times

- **Lestrang** 11 months ago

Selected Answer: C

Right fitting is for declaration, declaring the correct resources will not help. Reshuffling step is what can prevent fusion which can lead to unused workers.

👍 ↩ 🚩 upvoted 1 times

- **ML6** 1 year, 2 months ago

Selected Answer: B

Fusion occurs when multiple transformations are fused into a single stage, which can limit parallelism and hinder performance, especially in streaming pipelines. By introducing a Reshuffle step, you break fusion and allow for better parallelism.

👍 ↩ 🚩 upvoted 3 times

- **srivastavas08** 1 year, 2 months ago

https://cloud.google.com/dataflow/docs/guides/right-fitting

👍 ↩ 🚩 upvoted 2 times

- **GCP001** 1 year, 3 months ago

Selected Answer: B

Problem is performnace and not using all workers properly, https://cloud.google.com/dataflow/docs/pipeline-lifecycle#fusion_optimization

👍 ↩ 🚩 upvoted 3 times

- **scaenruy** 1 year, 4 months ago

Selected Answer: D

D. Use Dataflow Prime, and enable Right Fitting to increase the worker resources.

👍 ↩ 🚩 upvoted 1 times