C

G 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 137 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 137

Topic #: 1

[All Professional Data Engineer Questions]

You have a data pipeline with a Dataflow job that aggregates and writes time series metrics to Bigtable. You notice that data is slow to update in Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

- A. Configure your Dataflow pipeline to use local execution
- B. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions
- C. Increase the number of nodes in the Bigtable cluster
- D. Modify your Dataflow pipeline to use the Flatten transform before writing to Bigtable
- E. Modify your Dataflow pipeline to use the CoGroupByKey transform before writing to Bigtable

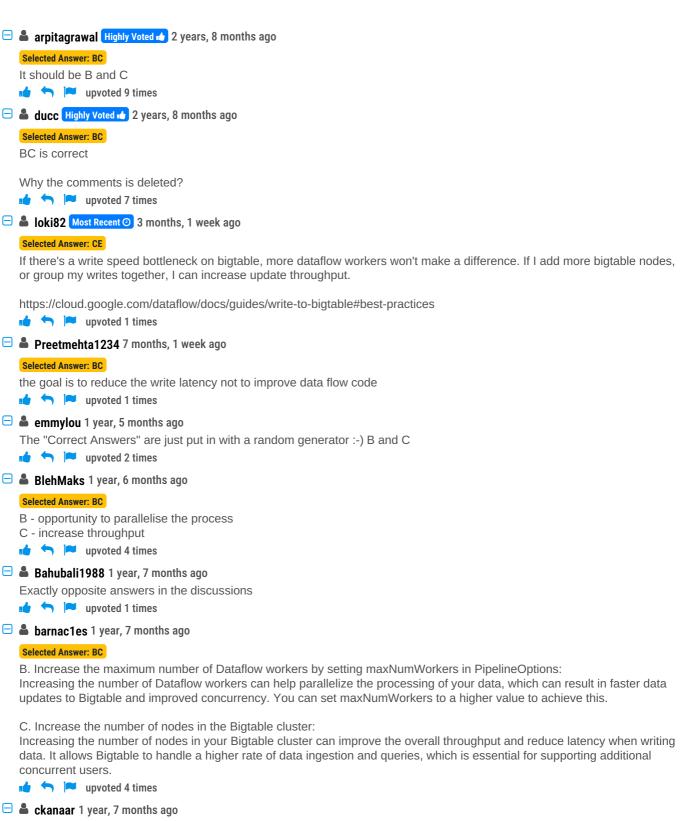
Show Suggested Answer

by 8 ducc at Sept. 3, 2022, 6:36 a.m.

Comments

Type your comment...

Submit



Selected Answer: CD

C definetely is correct, as it improves the read and write performance of Bigtable.

However, I do think that the second option is actually D instead of B, because the question specifically states that the pipeline aggregates data. Flatten merges multiple PCollection objects into a single logical PCollection, allowing for faster aggregation of time series data.

upvoted 2 times

■ NewDE2023 1 year, 9 months ago

Selected Answer: BE

B - I believe it is consensus.

D - The question mentions "a Dataflow job that "aggregates" and writes time series metrics to Bigtable". So CoGroupByKey performs a shuffle (grouping) operation to distribute data across workers.

https://cloud.google.com/dataflow/docs/guides/develop-and-test-pipelines





Selected Answer: DE

I read this question as: BigTable Write operations are all over the place (key-wise), and BigTable doesn't like that. When creating groups (batch writes), of similar keys (close to each other), BigTable is happy again, which I loosely translate into DE.

- upvoted 1 times
- 😑 🏜 vaga1 2 years ago

B is correct. But I don't see how you increase the write throughput of Bigtable increasing its cluster size. It should be dataflow instance resources that have to be increased

- upvoted 1 times
- 🖃 🏜 juliobs 2 years, 1 month ago

Selected Answer: BC

BC make sense

- upvoted 1 times
- NamitSehgal 2 years, 3 months ago

BC only makes sense here , no mention of data, no mention of keeping cost low

- upvoted 1 times
- 🖃 🚨 AzureDP900 2 years, 4 months ago
 - B. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions Most Voted
 - C. Increase the number of nodes in the Bigtable cluster
 - upvoted 1 times
- 🖃 🏜 ovokpus 2 years, 5 months ago

Selected Answer: BC

Increase max num of workers increases pipeline performance in Dataflow Increase number of nodes in Bigtable increases write throughput

upvoted 2 times

