Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 136 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 136
Topic #: 1
[All Professional Data Engineer Questions]

You are running a pipeline in Dataflow that receives messages from a Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

A. Increase the number of max workers

B. Use a larger instance type for your Dataflow workers

C. Change the zone of your Dataflow pipeline to run in us-central1

D. Create a temporary table in Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Bigtable to BigQuery

E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Show Suggested Answer

by jvg637 at *March 18, 2020, 4:39 p.m.*

Comments

Type your comment...

Submit

jvg637 Highly Voted 4 years, 7 months ago

A & B
instance n1-standard-1 is low configuration and hence need to be larger configuration, definitely B should be one of the option.
Increase max workers will increase parallelism and hence will be able to process faster given larger CPU size and multi core processor instance type is chosen. Option A can be a better step.

upvoted 50 times

AzureDP900 1 year, 10 months ago

Agreed

upvoted 2 times

sumanshu Highly Voted 3 years, 4 months ago

A & B.

With autoscaling enabled, the Dataflow service does not allow user control of the exact number of worker instances allocated to your job. You might still cap the number of workers by specifying the --max_num_workers option when you run your pipeline. Here as per question CAP is 3, So we can change that CAP.

For batch jobs, the default machine type is n1-standard-1. For streaming jobs, the default machine type for Streaming Engine-enabled jobs is n1-standard-2 and the default machine type for non-Streaming Engine jobs is n1-standard-4. When using the default machine types, the Dataflow service can therefore allocate up to 4000 cores per job. If you need more cores for your job, you can select a larger machine type.

upvoted 14 times

et2137 Most Recent 8 months, 2 weeks ago

Selected Answer: AB

A & B is correct

upvoted 1 times

kcl10 1 year, 1 month ago

Selected Answer: AB

A & B is correct

upvoted 1 times

juliorevk 1 year, 1 month ago

Selected Answer: AB

A because more workers improves performance through parallel work
B because the current instance size is too small

upvoted 1 times

barnac1es 1 year, 1 month ago

Selected Answer: AB

A. Increase the number of max workers:
By increasing the number of maximum workers, you allow Dataflow to allocate more computing resources to handle the peak load of incoming data. This can help improve processing speed and reduce CPU utilization per worker.

B. Use a larger instance type for your Dataflow workers:
Using a larger instance type with more CPU and memory resources can help your Dataflow workers handle a higher volume of data and processing tasks more efficiently. It can address CPU bottlenecks during peak periods.

upvoted 3 times

zellck 1 year, 11 months ago

Selected Answer: AB

AB is the answer.

upvoted 1 times

mbacelar 1 year, 11 months ago

Selected Answer: AB

Scale in and Scale Out

upvoted 1 times

FrankT2L 2 years, 4 months ago

Selected Answer: AB

maximum of 3 workers: Increase the number of max workers (A)
instance type n1-standard-1: Use a larger instance type for your Cloud Dataflow workers (B)

instance type n1-standard-1. Use a larger instance type for your Cloud Dataflow workers (B)

upvoted 2 times

MaxNRG 2 years, 10 months ago

Selected Answer: AB

A & B, other options don't make sense

upvoted 4 times

medeis_jar 2 years, 10 months ago

Selected Answer: AB

Only A & B make sense for improving pipeline performance.

upvoted 2 times

Mjvsj 2 years, 10 months ago

Selected Answer: AB

Should be A & B

upvoted 2 times

daghayeghi 3 years, 8 months ago

B, E:
B: Dataflow manage number of worker automatically, then we only can define machine type worker.
https://cloud.google.com/dataflow/docs/guides/deploying-a-pipeline
E: and adding a horizontally scale-able database like cloud spanner will reduce pressure on dataflow as it don't have to move data to specific zone and can be remain in same zone of EU, then E is correct.

upvoted 2 times

Vasu_1 3 years, 5 months ago

A & B is the right answer: You can set disable auto-scaling by setting the option --numWorkers (default is 3) and select the machine type by setting --workerMachineType at the time of creation of the pipeline (this applies to both auto and manual scaling)

upvoted 3 times

kavs 3 years, 11 months ago

Dataset is in EU so data can't be moved outside EU due to privacy law so zone option is ruled out. AB is Ok but intermediate table will boost perf apanee ruled out not sure of bigtable

upvoted 3 times

Alasmindas 3 years, 11 months ago

Option A and B for sure,
Option C : Changing Zone has nothing to do in improving performance
Option D and E : Adding BQ and BT is waste of many and does not solve the purpose of the question.

upvoted 3 times

SureshKotla 4 years, 1 month ago

B & D
DF will automatically take care of increasing workers. Developers won't need to access the settings .
https://cloud.google.com/dataflow/docs/guides/deploying-a-pipeline#autoscaling

upvoted 2 times

sumanshu 3 years, 4 months ago

automatically taking care of workers up to 3 (as the maximum worker is 3 set as per questions)

upvoted 1 times

SureshKotla 4 years, 1 month ago

On second thought, A B is looking right

upvoted 2 times

atnafu2020 4 years, 2 months ago

AB
is correct

upvoted 2 times

Load full discussion...