← **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 143 DISCUSSION**

Actual exam question from Google's Professional Data Engineer

Question #: 143

Topic #: 1

[All Professional Data Engineer Questions]

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

    A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name

    B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name

    C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code

    D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code

Show Suggested Answer

by 👤 ducc at *Sept. 3, 2022, 6:44 a.m.*

Comments

Type your comment...

Submit

⊟ 👤 **odacir** `Highly Voted 👍` 1 year, 11 months ago

`Selected Answer: D`

It's D. → Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. New version is mayor changes. Stop and drain and then launch the new code is a lot is the safer way.
We recommend that you attempt only smaller changes to your pipeline's windowing, such as changing the duration of fixed- or sliding-time windows. Making major changes to windowing or triggers, like changing the windowing algorithm, might have unpredictable results on your pipeline output.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

👍 ↩ 🚩 upvoted 15 times

⊟ 👤 **maggieee** 1 year, 10 months ago

Since updating the job as in A does a compatibility check, wouldn't you want to try that first? Then if the compatibility check fails then you proceed to drain current pipeline and then launch new pipeline (Answer D)?

As in A would be correct answer, then if compatibility check fails, you proceed to D.

https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#CCheck

👍 ↩ 🚩 upvoted 2 times

⊟ 👤 **ckanaar** 1 year, 1 month ago

You're right in your reasoning, but since the documentation specifically uses this example for stopping and draining, it's safe to assume that the compatibility check will always fail with these adjustments. Therefore, we can go straight to D.

Furthermore, answer A doesn't state: "Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing name, if the compatibility check fails, THEN proceed to stopping the pipeline with the drain option", so in itself it is not the right answer if the check fails.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **patitonav** [Most Recent ⊙] 10 months, 1 week ago

Selected Answer: D

D seems the right way to go

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **TVH_Data_Engineer** 10 months, 2 weeks ago

Selected Answer: D

Option A is the first approach to try, as it allows for an in-flight update with minimal disruption. However, if the changes in the new version of the pipeline are not compatible with an in-flight update (due to significant changes in windowing or triggering), then option D should be used. The Drain option ensures a graceful shutdown of the existing pipeline, reducing the risk of data loss, and then a new job can be started with the updated code.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **MaxNRG** 10 months, 3 weeks ago

Selected Answer: D

A is not an option as "You want to ensure that no data is lost during the update. ":
Making major changes to windowing or triggers, like changing the windowing algorithm, might have unpredictable results on your pipeline output.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#change_windowing

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **barnac1es** 1 year, 1 month ago

Selected Answer: D

Drain Option: The "Drain" option allows the existing Dataflow job to complete processing of any in-flight data before stopping the job. This ensures that no data is lost during the transition to the new version.
Create a New Job: After draining the existing job, you create a new Cloud Dataflow job with the updated code. This new job starts fresh and continues processing data from where the old job left off.

Option A (updating the inflight pipeline with the --update option) may not guarantee no data loss, as the update could disrupt the existing job's operation and potentially cause data loss.

Option B (updating the inflight pipeline with the --update option and a new job name) is similar to option A and may not provide data loss guarantees.

Option C (stopping the pipeline with the Cancel option and creating a new job) will abruptly stop the existing job without draining, potentially leading to data loss.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **knith66** 1 year, 3 months ago

Look D after seeing some docs. please check the below link https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **vamgcp** 1 year, 3 months ago

I will go with option D - If you want to minimize the impact of the update, then option A is the best option. However, if you are not concerned about a temporary interruption in processing, then option D is also a valid option. Option Pros Cons
A Does not stop the pipeline, so no data is lost. Requires you to create a new version of the pipeline.
B Creates a new job with the updated code, so you do not have to update the running pipeline. Can lead to data loss if the new job does not process all of the data that was in the running pipeline.
C Stops the pipeline and drains any data that is currently in flight, so no data is lost. Causes a temporary interruption in processing.

👍 ↩ 🏴 upvoted 1 times

👤 **midgoo** 1 year, 7 months ago

A is not recommeded for major changes in pipeline.

👍 ↩ 🏴 upvoted 3 times

👤 **musumusu** 1 year, 8 months ago

Answer A:
```gcloud dataflow jobs update <JOB_ID> --update <GCS_PATH_TO_UPDATED_PIPELINE> --region <REGION>```
--update flag does not miss any data and you can execute this command even yourpipeline is running. Its safe any fast, you can continuously make some change and update this command. no problem.
Stop and Drain, is required when you want to test the pipeline and stop it without losing the data.

👍 ↩ 🏴 upvoted 1 times

> 👤 **musumusu** 1 year, 8 months ago
>
> Answer D: as per latest documents 02/2023 google has removed update flag.
>
> 👍 ↩ 🏴 upvoted 3 times

👤 **jkhong** 1 year, 10 months ago

agree with odacir

👍 ↩ 🏴 upvoted 4 times

👤 **hauhau** 1 year, 11 months ago

vote A
D: drain doesn't mention about update dataflow job just stop and preserve data
A: replace existing job and preserve data
(When you update your job, the Dataflow service performs a compatibility check between your currently-running job and your potential replacement job. The compatibility check ensures that things like intermediate state information and buffered data can be transferred from your prior job to your replacement job.)

https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

👍 ↩ 🏴 upvoted 2 times

👤 **zellck** 1 year, 11 months ago

A is the answer.

https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#Launching
To update your job, launch a new job to replace the ongoing job. When you launch your replacement job, set the following pipeline options to perform the update process in addition to the job's regular options:
- Pass the --update option.
- Set the --jobName option in PipelineOptions to the same name as the job you want to update.

👍 ↩ 🏴 upvoted 1 times

> 👤 **odacir** 1 year, 11 months ago
>
> Are mayor changes. It's not safe to update. I vote D.
>
> 👍 ↩ 🏴 upvoted 1 times

👤 **Atnafu** 1 year, 11 months ago

D
A-is not because The Dataflow service retains the job name, but runs the replacement job with an updated Job ID.
Description:
When you update a job on the Dataflow service, you replace the existing job with a new job that runs your updated pipeline code. The Dataflow service retains the job name, but runs the replacement job with an updated Job ID. This process can cause downtime while the existing job stops, the compatibility check runs, and the new job starts.'
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#python:~:text=When%20you%20update%20a,has%20the%20following%20transforms%3A
D is correct
Drain ->clone -> update -> run

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **Atnafu** 1 year, 11 months ago

Changed my mind to A
https://cloud.google.com/dataflow/docs/guides/updating-a-
pipeline#python_2:~:text=Set%20the%20%2D%2Djob_name,%2D%2Dtransform_name_mapping%20option.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **drunk_goat82** 1 year, 11 months ago

Selected Answer: D

Changing windowing algorithm may break the pipeline.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

👍 ↩ 🚩 upvoted 3 times

⊟ 👤 **ovokpus** 1 year, 11 months ago

Selected Answer: A

No, do not drain the current job.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **dish11dish** 1 year, 11 months ago

Selected Answer: D

in this scenario pipline is streaming pipline with windowing algorithm and triggering strategy changes to new one without loss
of data,so better to go with Drain option as it fullfile all precondition described in scenario which is :-
1.streaming
2.code changes with windowing algorithm and triggering strategy to new way
3.no loss of data during update

Referances:-
https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline#drain
Drain a job. This method applies only to streaming pipelines. Draining a job enables the Dataflow service to finish processing
the buffered data while simultaneously ceasing the ingestion of new data. For more information, see Draining a job.

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **dish11dish** 1 year, 11 months ago

If the pipeline was batch then ans would been A

👍 ↩ 🚩 upvoted 1 times

⊟ 👤 **Mcloudgirl** 1 year, 12 months ago

D: They want to preserve data and updates might not be predictable.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

👍 ↩ 🚩 upvoted 3 times

Load full discussion…