

[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 68 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 68

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

- A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
- B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
- C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
- D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.


[Show Suggested Answer](#)

by [madhu1171](#) at March 13, 2020, 2:01 p.m.

Comments

Type your comment...

[Submit](#)

 **madhu1171** Highly Voted 4 years, 7 months ago

Answer should be C

   upvoted 36 times

 **[Removed]** Highly Voted 4 years, 7 months ago

Answer: C - best suitable for the purpose with autoscaling and google recommended transform engine between pubsub n bq

   upvoted 26 times

 **Abizi** Most Recent 2 months ago

Selected Answer: C

C for me

   upvoted 1 times

 **yassoraa88** 5 months, 3 weeks ago

Selected Answer: C

using Cloud Dataflow for transformations with monitoring via Stackdriver and leveraging its default autoscaling settings, is the best choice. Cloud Dataflow is purpose-built for this type of workload, providing seamless scalability and efficient processing capabilities for streaming data. Its autoscaling feature minimizes manual intervention and helps manage costs by dynamically adjusting resources based on the actual processing needs, which is crucial for handling fluctuating data volumes efficiently and cost-effectively.

   upvoted 2 times

 **barnac1es** 1 year, 1 month ago

Selected Answer: D

Option C suggests using Cloud Dataflow to run the transformations and monitoring the job system lag with Stackdriver while using the default autoscaling setting for worker instances.

While using Cloud Dataflow is a suitable choice for processing data from Cloud Pub/Sub to BigQuery, and monitoring with Stackdriver provides valuable insights, the specific emphasis on configuring non-default Compute Engine machine types (as mentioned in option D) gives you more control over cost optimization and performance tuning.

By configuring non-default machine types, you can precisely tailor the computational resources to match the specific requirements of your workload. This fine-grained control allows you to optimize costs further by avoiding over-provisioning of resources, especially if your workload is memory-intensive, CPU-bound, or requires specific configurations that are not met by the default settings.

   upvoted 1 times

 **barnac1es** 1 year, 1 month ago

Additionally, having the flexibility to adjust machine types based on workload characteristics ensures that you can achieve the desired performance levels without overspending on unnecessary resources. This level of customization is not provided by simply relying on the default autoscaling settings, making option D a more comprehensive and cost-effective solution for managing varying data volumes.

   upvoted 2 times

 **Mathew106** 1 year, 3 months ago

Selected Answer: B

At first I answered C. However, Dataproc is indeed cheaper than Dataflow. And both of them can scale automatically horizontally.

Dataflow horizontal scaling applies to both primary and secondary nodes. Scaling secondary nodes scales up CPU/compute and scaling primary nodes scales up both memory and CPU/compute.

I don't quite understand the second part of answer B where it says I should allocate resources accordingly. I guess I could do that, but auto-scaling should be enough.

   upvoted 1 times

 **AbdullahAnwar** 1 year, 8 months ago

Answer should be C

   upvoted 2 times


 **samdhimal** 1 year, 9 months ago

C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.

Cloud Dataflow is a managed service that allows you to write and execute data transformations in a highly scalable and fault-tolerant way. By default, it will automatically scale the number of worker instances based on the input data volume and job performance, which can help minimize costs. Monitoring the job system lag with Stackdriver can help you identify any issues that may be impacting performance and take action as needed. Additionally, using the default autoscaling setting for worker

that may be impacting performance and take action as needed. Additionally, using the default autoscaling setting for worker instances can help you minimize manual intervention and ensure that resources are used efficiently.



   upvoted 3 times

  **zelck** 1 year, 11 months ago

Selected Answer: C

C is the answer.

   upvoted 1 times



  **odacir** 1 year, 11 months ago

Selected Answer: C

@admin why all the answers are wrong. I paid 30 euros for this web and its garbage.

Dataproc has no sense in this scenario, because you want to have minimal intervention/operation. D is not a good practice, the answer is C.

   upvoted 11 times

  **zelck** 1 year, 11 months ago

you need to look at community vote distribution and comments, and not the suggested answer.

   upvoted 9 times

  **medeis_jar** 2 years, 10 months ago

Selected Answer: C

C only as referred by MaxNRG

   upvoted 4 times

  **MaxNRG** 2 years, 10 months ago

Selected Answer: C

C.

Dataproc does not seem to be a good solution in this case as it always require a manual intervention to adjust resources. Autoscaling with dataflow will automatically handle changing data volumes with no manual intervention, and monitoring through Stackdriver can be used to spot bottleneck. Total execution time is not good there as it does not provide a precise view on potential bottleneck.

   upvoted 9 times

  **StefanoG** 2 years, 11 months ago

Selected Answer: C

Dataflow, Stackdriver and autoscaling

   upvoted 3 times

  **victorlie** 2 years, 11 months ago



Admin, please take a look on the comments. Almost all answers are wrong

   upvoted 4 times

  **nguyenmoon** 3 years, 1 month ago

Answer should be C as dataflow is unpredictable size (input that will vary in size), dataproc is with known size

   upvoted 4 times

  **Tanzu** 2 years, 8 months ago

dataflow over dataproc is always the preferred way in gcp. use dataproc only there is specific client requirements such as existing hadoop workloads, etc..

   upvoted 1 times

  **sandipk91** 3 years, 2 months ago

Option C is the answer

   upvoted 3 times

  **sumanshu** 3 years, 4 months ago

Vote for C

   upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses



© 2024 ExamTopics