

[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 33 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 33

Topic #: 1

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

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud

Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

- A. Check the dashboard application to see if it is not displaying correctly.
- B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
- C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
- D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

[Show Suggested Answer](#)

by [deleted] at *March 20, 2020, 3:42 p.m.*

### Comments

[Submit](#)

🗑️ 👤 [Removed] Highly Voted 4 years, 1 month ago

Answer: C

Description: Stackdriver can be used to check the error like number of unack messages, publisher pushing messages faster

👍 ↩️ 🚩 upvoted 35 times

🗑️ 👤 snamburi3 3 years, 5 months ago

All messages are being published to Cloud Pub/Sub successfully. so Stackdriver might not help.

👍 ↩️ 🚩 upvoted 10 times

🗑️ 👤 kubosuke 2 years, 7 months ago

messages sent successfully to Topic, but not Subscription.

in this case, if Dataflow cannot handle messages correctly it might not return acknowledgments to the Pub/Sub, and these errors can be seen from Monitoring.

[https://cloud.google.com/pubsub/docs/monitoring#monitoring\\_exp](https://cloud.google.com/pubsub/docs/monitoring#monitoring_exp)

👍 ↩️ 🚩 upvoted 12 times

🗑️ 👤 Tanzu 2 years, 3 months ago

to be more precise, first to publisher,

- then forwards to topic, and persistence for a while
- then forwards to subscriber,
- then to subscription..
- then acknowledgement happens

so in every steps, there is possibly for errors.

👍 ↩️ 🚩 upvoted 1 times

🗑️ 👤 jkhong 1 year, 4 months ago

PubSub doesn't forward from subscriber to subscription. A topic sends it over to subscription first, then to subscriber

👍 ↩️ 🚩 upvoted 2 times

🗑️ 👤 [Removed] 4 years, 1 month ago

this will help us understand the reason, when we know that the data is not reaching subscriber then there is no point in checking it with dummy data

👍 ↩️ 🚩 upvoted 12 times

🗑️ 👤 tprashanth 3 years, 9 months ago

B.

Stack driver monitoring is for performance, not logging of missing data.

👍 ↩️ 🚩 upvoted 25 times

🗑️ 👤 jkhong 1 year, 4 months ago

Please refer to this PubSub specific Monitoring metrics

[https://cloud.google.com/pubsub/docs/monitoring#monitoring\\_the\\_backlog](https://cloud.google.com/pubsub/docs/monitoring#monitoring_the_backlog)

👍 ↩️ 🚩 upvoted 1 times

🗑️ 👤 mikey007 3 years, 9 months ago

<https://cloud.google.com/pubsub/docs/monitoring>

👍 ↩️ 🚩 upvoted 2 times

🗑️ 👤 ritinhabb 1 year, 10 months ago

Exactly!

👍 ↩️ 🚩 upvoted 1 times

🗑️ 👤 [Removed] Highly Voted 4 years, 1 month ago

Should be B

👍 ↩️ 🚩 upvoted 25 times

🗑️ 👤 [Removed] 4 years, 1 month ago

confused with D as well.

👍 ↩️ 🚩 upvoted 1 times

🗑️ 👤 Tanzu 2 years, 3 months ago

push or pull is about how target will handle the messages. pull mode gives flexibility when to get messages , so considering if target (or client) is slow, then it can make predictable choices.

dataflow is serverless. so if you need to awake it when necessary, you should use push mechanism. or leverage cloud composer/airflow and listen to pub/sub to trigger the dataflow.

👍 ↩️ 🚩 upvoted 3 times

  **jvg637** 4 years, 1 month ago

pushing is only for a https endpoint. So Dataflow just can pull messages

   upvoted 4 times

 **Rajokkiyam** 4 years, 1 month ago

Push or Pull guarantees the message to be delivered at-least once. So it doesn't make any difference.

   upvoted 6 times

  **gopinath\_k** 3 years, 1 month ago

## Push needs Https endpoint

   upvoted 3 times

  **rtcpo** **Most Recent**  6 months, 2 weeks ago

**Selected Answer: B**

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

\* By running a fixed dataset through the Cloud Dataflow pipeline, you can determine if the problem lies within the data processing stage. This allows you to identify any issues with data transformation, filtering, or processing in your pipeline.

\* Analyzing the output from this fixed dataset will help you isolate the problem and confirm whether it's related to data processing or the dashboard application.

   upvoted 1 times

  **ruben82** 6 months ago

You must know what kind of data causes errors. I think, the first step is to get erroneous data and then test with sample of it.

   upvoted 1 times

  **imran79** 7 months ago

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output. If this results in the expected output, then the problem might be with the dashboard application (Option A), and that should be checked next.

   upvoted 1 times

  **WillemHendr** 11 months ago

**Selected Answer: B**

"...to find the missing messages"

Up to that remark, Monitoring was a valid option as well. But missing messages cannot be found with monitoring.

It is simply not possible to find the exact missing message. I read this remark as a test if you know what is, and what isn't possible with monitoring.

   upvoted 3 times

  **izekc** 1 year ago

**Selected Answer: B**

here is to determine next step. Not better way to optimize the workload. So B is the correct next step

   upvoted 2 times

  **Jarek7** 1 year ago



B is not the next step. The next step is between pub/sub and dataflow(C). B will not help with it at all. However it could show the issue if it is the pipeline or the view. But also it could not show it - you have no idea why some messages are not shown, so most probably it wouldnt get you any info. Definitely next step is to chek if the issue is between pubsub and dataflow. Then you could go with B.

   upvoted 1 times

  **Adswerve** 1 year ago

**Selected Answer: D**

Pull subscription is the correct one. Push subscription means Dataflow cannot keep up with the topic.

   upvoted 1 times

  **Jarek7** 1 year ago

It could be the issue. But C would reveal it if this is the real issue - if you will not check stackdriver, you cannot be sure if you really resolved the issue, as even if it seems to be working properly after switch to pull you cannot be sure if it is because of some other temporal factor.

   upvoted 1 times

  **midgoo** 1 year, 2 months ago

**Selected Answer: B**

If the Dataflow does not have the expected output, it is either wrong at the input or at the pipelines. The chance that the issue is at the input (PubSub) is very low. For this case, it is likely the pipelines got some mistakes (e.g. JSON parsing failed). So we should follow B to debug the pipelines (using snapshot as test dataset for example)







   upvoted 4 times

  **ploer** 1 year, 3 months ago

**Selected Answer: B**

The most efficient solution would be to run a fixed dataset through the Cloud Dataflow pipeline and analyze the output (Option B). This will allow you to determine if the issue is with the pipeline or with the dashboard application. By analyzing the output, you can see if the messages are being processed correctly and determine if there are any discrepancies or missing messages. If the issue is with the pipeline, you can then debug and make any necessary updates to ensure that all messages are processed correctly. If the issue is with the dashboard application, you can then focus on resolving that issue. This approach allows you to isolate and identify the root cause of the missing messages in a controlled and efficient manner.

   upvoted 7 times

  **Lestrang** 1 year, 3 months ago

**Selected Answer: B**


I've just skimmed over the Stackdriver docs, yes guys, it helps you check the number and age of messages that were not received/acknowledged, excellent, hurray.

So first off, c will not give us the missing messages, it will give us the count and age.  
that means that c is inherently incorrect.

Additionally, will knowledge of the number of messages make resolving the problem any easier? No, it is just confirming what we already know.

Meanwhile, approach B, will allow us to see HOW and WHY it is missing some messages, which is the step that proceeds the fix.

   upvoted 7 times

  **PolyMoe** 1 year, 3 months ago

**Selected Answer: C**

Here is ChatGPT answer :

It's always a good practice to start by checking the logs and monitoring tools to see if there is any indication of an issue with the messages being published to Cloud Pub/Sub. In this case, you should use Google Stackdriver Monitoring to investigate if the missing messages have been published or not. You can also run a fixed dataset through the Cloud Dataflow pipeline to see if the pipeline is processing the messages correctly. If there is no issue found on the Cloud Pub/Sub and Cloud Dataflow, then you can check the dashboard application to see if it is not displaying the messages correctly. As a last resort, you can switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

   upvoted 2 times

  **Jarek7** 1 year ago

I'm tired with these responses about what chatGPT says. Most probably you've used the free 3.5 version which is absolute disaster regarding being all knowing oracle. BTW in this case I wouldn't believe even GPT4. It is a difficult question that needs a specific knowledge and experience which might be not available in the GPT training data. You cannot use any GPT up to 4 as an argument in such cases.

   upvoted 2 times

  **axantroff** 5 months, 2 weeks ago

Exactly. Sometimes it is total garbage

   upvoted 1 times



  **Lestrang** 1 year, 3 months ago

I provided it with the question as input but added the metrics available in Stackdriver, here is the response:

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

If messages are being published successfully to Cloud Pub/Sub but are missing in the dashboard, the issue is likely to be with the Cloud Dataflow pipeline that processes the messages. To find the root cause of the problem, you should run a fixed dataset through the pipeline and analyze the output. This will allow you to see if the pipeline is correctly processing all messages, and identify any processing errors that might be causing messages to be lost. The output can be compared to the expected results to identify any discrepancies and resolve the issue.

   upvoted 4 times

  **desertlotus1211** 1 year, 3 months ago

The question is really not asking for a solution to the problem, per se - but more of what would the next step in the situation to triage the issue....

Answer would be B over D. Answer D would be the recommended solution IF the question asked to rectify/fixed the issue.

Thoughts?

   upvoted 4 times

🗨️ 👤 **izekc** 1 year ago

Agree with u

👍 ↩️ 🚩 upvoted 1 times

🗨️ 👤 **Prakzz** 1 year, 4 months ago

**Selected Answer: D**

D. Dataflow must PULL the data to process it in real-time. Missing messages in the dashboard, means that the Pub/Sub to Dataflow was misconfigured as PUSH.

👍 ↩️ 🚩 upvoted 3 times

🗨️ 👤 **hasoweh** 1 year, 3 months ago

Pull will lead to latency as new data will not be streamed upon arrival, but instead will only be passed on when Dataflow makes a pull request. So if data comes in at time 0:01 but pull requests are only happening every 10 seconds, we have 9 second delay. Push will automatically push the data to any subscribers as soon as the data comes, and thus is closer to real-time.

👍 ↩️ 🚩 upvoted 1 times

🗨️ 👤 **Krish6488** 1 year, 4 months ago

**Selected Answer: B**

To me, B sounds more logical for the below reason.

Option C would have been ideal because any debugging starts with checking the logs, however the option says, check stackdriver for missing messages. Had it been, check stackdriver to figure out the number of undelivered messages, C would have been more suitable. Given the slight bit of dodginess in option c, I would go with B

👍 ↩️ 🚩 upvoted 3 times

🗨️ 👤 **Nirca** 1 year, 4 months ago

**Selected Answer: B**

Why checking Pub/Sub again when this is already verified to be fine according to the question. Shouldn't you be checking the next stage in the flow which is Dataflow?

Option - B

👍 ↩️ 🚩 upvoted 1 times

🗨️ 👤 **DGames** 1 year, 4 months ago

**Selected Answer: B**

Answer - B. Because already we know message is missing so better to test with fixed dataset and check code .

👍 ↩️ 🚩 upvoted 1 times

🗨️ 👤 **Atnafu** 1 year, 5 months ago

C

[https://cloud.google.com/pubsub/docs/monitoring#:~:text=the%20specific%20metrics.-](https://cloud.google.com/pubsub/docs/monitoring#:~:text=the%20specific%20metrics.-,Monitor%20message%20backlog,information%20about%20this%20metric%2C%20see%20the%20relevant%20section%20of%20this%20document.,-Create%20alerting%20policies)

[,Monitor%20message%20backlog,information%20about%20this%20metric%2C%20see%20the%20relevant%20section%20of%20this%20document.,-Create%20alerting%20policies](#)

👍 ↩️ 🚩 upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses



