

🔗 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 173 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 173

Topic #: 1

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

You are designing a pipeline that publishes application events to a Pub/Sub topic. Although message ordering is not important, you need to be able to aggregate events across disjoint hourly intervals before loading the results to BigQuery for analysis. What technology should you use to process and load this data to BigQuery while ensuring that it will scale with large volumes of events?

- A. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.
- B. Schedule a Cloud Function to run hourly, pulling all available messages from the Pub/Sub topic and performing the necessary aggregations.
- C. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub/Sub topic and performing the necessary aggregations.
- D. Create a streaming Dataflow job that reads continually from the Pub/Sub topic and performs the necessary aggregations using tumbling windows.




[Show Suggested Answer](#)

by [AWSandeep](#) at *Sept. 2, 2022, 7:51 p.m.*

### Comments

Type your comment...

Submit

  **Atnafu** Highly Voted  2 years, 4 months ago

D

TUMBLE=> fixed windows.  
HOP=> sliding windows.  
SESSION=> session windows.

   upvoted 11 times

  **musumusu** Highly Voted  2 years, 2 months ago



why not c ? as data is arriving hourly why we can use batch processing rather than streaming with 1 hour fixed window?

   upvoted 7 times

  **MrMone** 2 years ago



"you need to be able to aggregate events across disjoint hourly intervals" does not means data is arriving hourly. however, it's tricky! Answer D

   upvoted 2 times

  **ga8our** 1 year, 6 months ago

I second your question. Noone who suggests Dataflow streaming (D) has given an explanation why an hourly batch job is insufficient.

   upvoted 2 times

  **ga8our** 1 year, 6 months ago

I second your question. Noone who suggests C has given an explanation why an hourly batch job is insufficient.

   upvoted 2 times

  **baimus** Most Recent  7 months ago

**Selected Answer: D**

Just to provide clarity to people asking "why not C" - the source is a pub/sub. Pub/Sub has a limit of 10 MB or 1000 messages for a single batch publish request, which means that batch dataflow will not necessarily be able to retrieve all messages. If the question had said "there will always be less than 1000 messages and less than 10mb", only then would batch be acceptable.



   upvoted 4 times

  **mayankazyour** 8 months ago

**Selected Answer: D**

The question asks for future scalability for large volumes of events, its better to go with streaming dataflow job.

   upvoted 1 times

  **emmylou** 1 year, 5 months ago

I just do not understand why this needs to be streamed. I understand that there might be a slight delay using batch processing but there is no indication this is critical data. Can someone please provide your thinking?

   upvoted 2 times

  **vamgcp** 1 year, 9 months ago

We can use TUMBLE(1 HOUR) to create hourly windows, where each window contains events from a specific hour.


   upvoted 1 times

  **vamgcp** 1 year, 9 months ago

**Selected Answer: D**

Option D : A streaming Dataflow job is the best way to process and load data from Pub/Sub to BigQuery in real time. This is because streaming Dataflow jobs can scale to handle large volumes of data, and they can perform aggregations using tumbling windows.

   upvoted 1 times

  **zelck** 2 years, 5 months ago

**Selected Answer: D**

D is the answer.

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#tumbling-windows>

   upvoted 3 times

  **devaid** 2 years, 6 months ago



**Selected Answer: D**

Answer D

Tumbling Windows = Fixed Windows

Tumbling windows – Fixed windows

   upvoted 2 times

  **TNT87** 2 years, 7 months ago

**Selected Answer: D**

Answer D

   upvoted 2 times

  **AWSandeep** 2 years, 8 months ago

**Selected Answer: D**

D. Create a streaming Dataflow job that reads continually from the Pub/Sub topic and performs the necessary aggregations using tumbling windows.

A tumbling window represents a consistent, disjoint time interval in the data stream.

Reference:

<https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines#tumbling-windows>

   upvoted 2 times



## Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses



© 2024 ExamTopics