

🔗 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 62 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 62

Topic #: 1

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

Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period.

However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

- A. Set a single global window to capture all the data.
- B. Set sliding windows to capture all the lagged data.
- C. Use watermarks and timestamps to capture the lagged data.
- D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.




[Show Suggested Answer](#)

by [deleted] at *March 21, 2020, 3:16 p.m.*

Comments

Type your comment...

[Submit](#)

  **[Removed]** Highly Voted  4 years, 1 month ago

Answer: C

Description: A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

   upvoted 44 times

  **[Removed]** Highly Voted  4 years, 1 month ago

Answer: C

   upvoted 18 times

  **MikkelRev** Most Recent  7 months, 1 week ago

Selected Answer: C


option C: Use watermarks and timestamps to capture the lagged data.

   upvoted 1 times

  **MikkelRev** 7 months, 1 week ago

option C: Use watermarks and timestamps to capture the lagged data.

   upvoted 1 times

  **samdhimal** 1 year, 3 months ago

C: Use watermarks and timestamps to capture the lagged data.

Watermarks are a way to indicate that some data may still be in transit and not yet processed. By setting a watermark, you can define a time period during which Dataflow will continue to accept late or out-of-order data and incorporate it into your processing. This allows you to maintain a predictable time period for processing while still allowing for some flexibility in the arrival of data.

Timestamps, on the other hand, are used to order events correctly, even if they arrive out of order. By assigning timestamps to each event, you can ensure that they are processed in the correct order, even if they don't arrive in that order.

   upvoted 9 times



  **samdhimal** 1 year, 3 months ago

Option A: Set a single global window to capture all the data is not a good idea because it may not allow for late or out-of-order data to be processed.

Option B: Set sliding windows to capture all the lagged data is not suitable for the case where you want to process the data over a predictable time period. Sliding windows are used when you want to process data over a period of time that is continuously moving forward, not a fixed period.

Option D: Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data is a good practice but not a complete solution, because it only ensures that data is ordered correctly, but it does not account for data that may be late.



   upvoted 4 times

  **desertlotus1211** 1 year, 3 months ago

Answer is C:

There is no such thing as a sliding windows using by dataflow.

   upvoted 1 times

  **Mathew106** 9 months, 2 weeks ago

The naming in Apache Beam is: Fixed, Sliding, Session

In Dataflow it's: Tumbling, Hopping, Session.

I was very confused at first too when I saw "hopping" in a question.

   upvoted 1 times

  **DeeData** 1 year, 3 months ago

I highly doubt, DataFlow windowing is divided into three(3) types:

1. Fixed
2. Sliding
3. Session

   upvoted 1 times

  **AzureDP900** 1 year, 4 months ago

Answer is Use watermarks and timestamps to capture the lagged data.

A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

   upvoted 1 times

  **DeeData** 1 year, 4 months ago

📄 **UGames** 1 year, 4 months ago

Selected Answer: C

Watermark is use for late date,

👍 ↩ 🚩 upvoted 2 times

📄 **[Removed]** 1 year, 5 months ago

Watermark doesn't solve the out-of-order data problem. It only solves the problem of late data. However, with D, you can use the timestamps to solve both problems (for instance, if you're storing incoming data in a table, you can easily insert any late data to its correct place a time-partionned table using the timestamp of the element)

👍 ↩ 🚩 upvoted 3 times

📄 **ovokpus** 1 year, 5 months ago

with watermarks, when the late data arrives, it goes into its rightful window and gets in order

👍 ↩ 🚩 upvoted 1 times

📄 **ovokpus** 1 year, 5 months ago

C even says watermarks AND timestamps.

👍 ↩ 🚩 upvoted 1 times

📄 **FrankT2L** 1 year, 11 months ago

Selected Answer: B

Preemptible workers are the default secondary worker type. They are reclaimed and removed from the cluster if they are required by Google Cloud for other tasks. Although the potential removal of preemptible workers can affect job stability, you may decide to use preemptible instances to lower per-hour compute costs for non-critical data processing or to create very large clusters at a lower total cost

<https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vm>

👍 ↩ 🚩 upvoted 1 times

📄 **FrankT2L** 1 year, 11 months ago

delete this answer. The answer belongs to another question

👍 ↩ 🚩 upvoted 8 times

📄 **Tanzu** 2 years, 2 months ago

Selected Answer: A

That's why we have watermarks in apache beam.

👍 ↩ 🚩 upvoted 1 times

📄 **VishalBule** 2 years, 2 months ago

Answer is C Use watermarks and timestamps to capture the lagged data.

A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

👍 ↩ 🚩 upvoted 1 times

📄 **medeis_jar** 2 years, 4 months ago

Selected Answer: C

"Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event timestamp that is earlier than the watermark, the record is treated as late data."

👍 ↩ 🚩 upvoted 3 times

📄 **MaxNRG** 2 years, 4 months ago

Selected Answer: C

A is a direct No, if data don't have timestamp, we'll only have the procesing time and not the "event time".

B is not either, sliding windows are not for this. Hopping|sliding windowing is useful for taking running averages of data, but not to process late data.

D looks correct but has one concept missing, the watermark to know if the process time is ok with the event time or not. I'm not 100% sure is incorrect. If, since we have a "predictable time period", might be this will do. I mean, if our dashboard is shown after the last input data has arrived (single global window), this should be ok. We'd have a "perfect watermark". Anyway we'd need triggering .

👍 ↩ 🚩 upvoted 4 times

📄 **MaxNRG** 2 years, 4 months ago

C is, I think, the correct answer: Watermark is different from late data. Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event timestamp that is earlier than the watermark, the record is treated as late data.

I'll try to explain: Late data is inherent to Beam's model for out-of-order processing. What does it mean for data to be late? The definition and its properties are intertwined with watermarks that track the progress of each computation across the event time domain. The simple intuition behind handling lateness is this: only late input should result in late data anywhere

in the pipeline.

So, is not easy to decide between C and D. If you ask me I'd say C since for D we ought to make some suppositions.



   upvoted 3 times

  **Jlozano** 2 years, 4 months ago

Selected Answer: C

"Expert Verified" but >50% questions have random answer. "Sliding window" really? Please, this can be fixed easily with our most voted answer. Of course, the correct answer is C.



   upvoted 4 times

  **JG123** 2 years, 5 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: C

   upvoted 4 times

  **anji007** 2 years, 6 months ago

Ans: C

   upvoted 2 times

[Load full discussion...](#)



Platform

> [Home](#)

> [All Exams](#)

> [Examtopics PRO](#)

> [Training Courses](#)



© 2024 ExamTopics