

🔗 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 110 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 110

Topic #: 1

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

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

- A. Add a SideInput that returns a Boolean if the element is corrupt.
- B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
- C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
- D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Show Suggested Answer

by [deleted] at March 22, 2020, 2:51 p.m.

### Comments

Type your comment...

Submit

🗨️ 👤 [Removed] Highly Voted 👍 3 years, 7 months ago

Correct - B

👍 🗳️ 🗳️ upvoted 16 times

upvoted 10 times

**[Removed]** **Highly Voted** 3 years, 7 months ago

Answer: B

Description: ParDo is used to do transformation and create side output

upvoted 12 times

**midgoo** **Most Recent** 7 months, 3 weeks ago

**Selected Answer: B**

A - SideInput is often used to validate data, however, we need to create the SideInput first. When using SideInput to filter data, it is actually another ParDo call.

C, D - This is common way to filter too, but we will need the key in order to partition or GroupByKey

B - ParDo is the most basic method, it can do anything to the PCollection

upvoted 3 times

**AzureDP900** 10 months, 1 week ago

B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.

upvoted 1 times

**zellick** 11 months ago

**Selected Answer: B**

B is the answer.

<https://cloud.google.com/dataflow/docs/concepts/beam-programming-model#concepts>

ParDo is the core parallel processing operation in the Apache Beam SDKs, invoking a user-specified function on each of the elements of the input PCollection. ParDo collects the zero or more output elements into an output PCollection. The ParDo transform processes elements independently and possibly in parallel.

upvoted 3 times

**Pime13** 1 year, 4 months ago

**Selected Answer: B**

vote B :<https://beam.apache.org/documentation/programming-guide/#pardo>

Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection or discard it.

Formatting or type-converting each element in a data set. If your input PCollection contains elements that are of a different type or format than you want, you can use ParDo to perform a conversion on each element and output the result to a new PCollection.

Extracting parts of each element in a data set. If you have a PCollection of records with multiple fields, for example, you can use a ParDo to parse out just the fields you want to consider into a new PCollection.

Performing computations on each element in a data set. You can use ParDo to perform simple or complex computations on every element, or certain elements, of a PCollection and output the results as a new PCollection.

upvoted 4 times

**medeis\_jar** 1 year, 10 months ago

**Selected Answer: B**

Filtering with ParDo. ParDo is a Beam transform for generic parallel processing. ParDo is useful for common data processing operations/

upvoted 2 times

**AzureDP900** 10 months, 1 week ago

I agree with B

upvoted 1 times

**MaxNRG** 1 year, 10 months ago

**Selected Answer: B**

B: ParDo is a Beam transform for generic parallel processing. ParDo is useful for common data processing operations, including:

a. Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection, or discard it.

b. Formatting or type-converting each element in a data set.

c. Extracting parts of each element in a data set.

d. Performing computations on each element in a data set.

A does not help

C Partition is a Beam transform for PCollection objects that store the same data type. Partition splits a single PCollection into a fixed number of smaller collections. Again, does not help

D GroupByKey is a Beam transform for processing collections of key/value pairs. GroupByKey is a good way to aggregate data that has something in common

upvoted 6 times

**sumanshu** 2 years, 4 months ago

vote for 'B', ParDo can discard the elements.

<https://beam.apache.org/documentation/programming-guide/>

   upvoted 4 times

  **DeepakKhattar** 2 years, 9 months ago

B - seems to be better option since we need to filter out, question does not specify that we do need to store it into different PCollection.

<https://beam.apache.org/documentation/transforms/python/overview/>

ParDo is general purpose whereas partition splits the elements into do different pcollections.



<https://beam.apache.org/documentation/transforms/python/elementwise/partition/>

   upvoted 3 times

  **arghya13** 2 years, 11 months ago

B is correct

   upvoted 3 times

  **SteelWarrior** 3 years, 1 month ago


Should be B. The Partition transform would require the element identifying the valid/invalid records for partitioning the pcollection that means there is some logic to be executed before the Partition transformation is invoked. That logic can be implemented in a ParDO transform and which can both identify valid/invalid records and also generate two PCollections one with valid records and other with invalid records.

   upvoted 7 times

  **haroldbenites** 3 years, 2 months ago

B is correct

   upvoted 3 times

  **Archy** 3 years, 3 months ago

B, ParDo is useful for a variety of common data processing operations, including:

Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection or discard it.

   upvoted 4 times

  **tprashanth** 3 years, 3 months ago

Looks like C it is

<https://beam.apache.org/documentation/programming-guide/>

   upvoted 2 times


  **atnafu2020** 3 years, 2 months ago

according this link its

Pardo

\* Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection or discard it.

\* But Partition just splitting which is is a Beam transform for PCollection objects that store the same data type. Partition splits a single PCollection into a fixed number of smaller collections.

   upvoted 5 times

  **xrun** 2 years, 10 months ago

Seems like two answers may be correct. With ParDo you can discard corrupt data. With Partition you can split the data into two PCollections: corrupt and ok. You stream ok data further to BigQuery and corrupt data to some other storage for analysis. If one is not interested in analysis, then ParDo is enough.

   upvoted 1 times

  **dg63** 3 years, 3 months ago

Correct answer should be "C". A Pardo transform will allow the processing to happening in parallel using multiple workers. Partition transform will allow data to be partitions in two different Pcollections according to some logic. Using partition transform once can split the corrupted data and finally discard it.

   upvoted 5 times

  **Rajuuu** 3 years, 3 months ago

Correct B.

   upvoted 4 times

[Load full discussion...](#)



## Platform

> [Home](#)

> [Examtopics PRO](#)

> [All Exams](#)

> [Training Courses](#)

