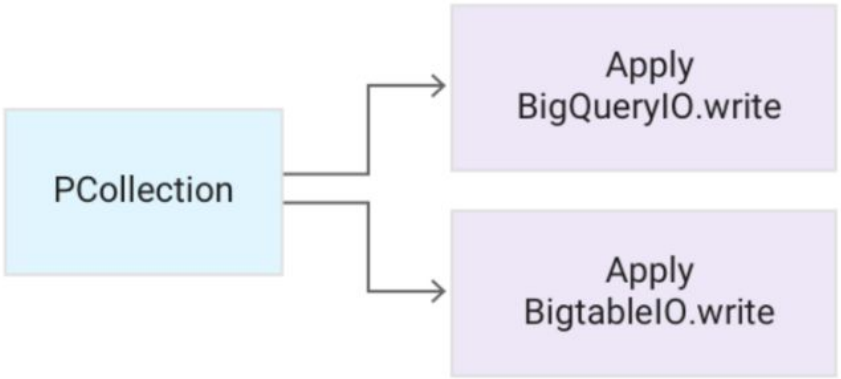
CS 327E Class 10

November 18, 2019

1) What is meant by the following usage pattern?



A. The elements in the PCollection are split up such that 1/2 of the elements are written to BigQuery and 1/2 are written to Bigtable.
B. The same PCollection can be written to multiple data sinks including BigQuery and Bigtable.
C. The PCollection can only be written to BigQuery or Bigtable.

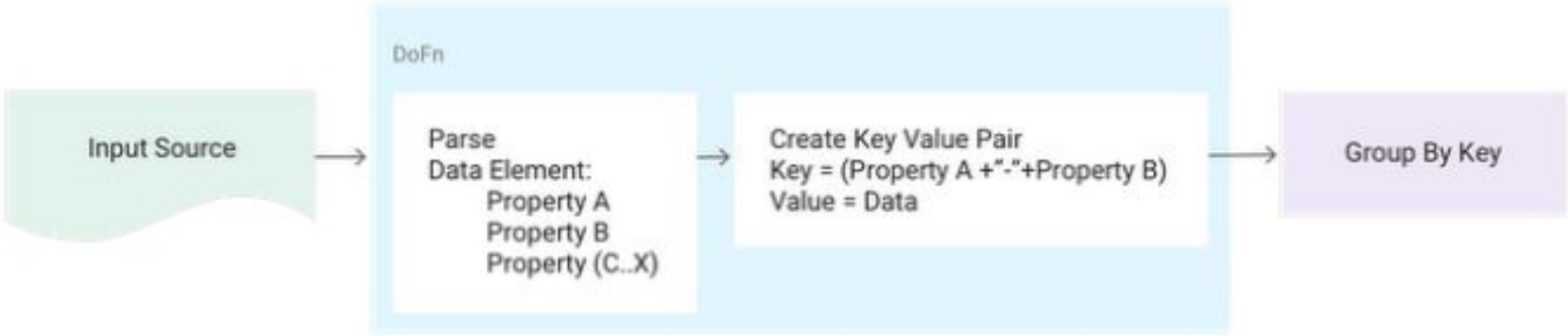
2) How do the authors suggest handling bad data?

A. Send the bad data out of the DoFn as a SideOutput.
B. Send the bad data into the DoFn as a SideInput.
C. Write the bad data to an error log, but don’t write it to a back-end database.

3) What method do the authors suggest for triggering a Dataflow pipeline that needs to start after a file has been uploaded to Google Cloud Storage?

A. Use a simple REST endpoint to trigger the pipeline.
B. Open CloudShell and run the pipeline from the command-line.
C. Trigger the pipeline from Google Cloud Storage.

4) What is meant by the following usage pattern?



A. GroupByKey requires a preceding DoFn step in the pipeline.
B. GroupByKey requires a composite key as input.
C. Create a composite key to group by multiple properties with GroupByKey.

5) What method do the authors suggest for joining two PCollections in which one of the PCollections is small?

A. Use a CoGroupByKey transform
B. Use a SideInput to a ParDo
C. Use a SQL Join

Common Beam Errors

1. `Table name XYZ cannot be resolved: dataset name is missing.`
2. `RuntimeError: Transform XYZ does not have a stable unique label.`
3. `IndexError: list index out of range while running ParDo(DoFn)`
4. `ValueError: need more than 1 value to unpack while running ParDo(DoFn)`
5. `TypeError: object of type '_UnwindowedValues' has no len()`
6. `AttributeError: 'set' object has no attribute 'iteritems'`
7. `RuntimeError: Could not successfully insert rows to BigQuery table… This field is not a record and Array specified for non-repeated field`

Hands-on Lab

1) Set up Jupyter to run Beam & Dataflow: how-to guide

2) Debug several Beam pipelines :)

Practice Problem 1

Run and fix `oscars_6.py`

Practice Problem 1

Run and fix `oscars_6.py`

What was the cause of the error?

A. Syntax error
B. Logic error
C. All of the above

Practice Problem 2

Run and fix `oscars_8.py`

Practice Problem 2

Run and fix `oscars_8.py`

What was the cause of the error?

A. Syntax error
B. Logic error
C. All of the above

Practice Problem 3

Run and fix `oscars_9.py`

Practice Problem 3

Run and fix `oscars_9.py`

What was the cause of the error?

A. Syntax error
B. Logic error
C. All of the above

Milestone 10

http://www.cs.utexas.edu/~scohen/milestones/Milestone10.pdf