CS 327E Class 10

April 20, 2020

Announcements

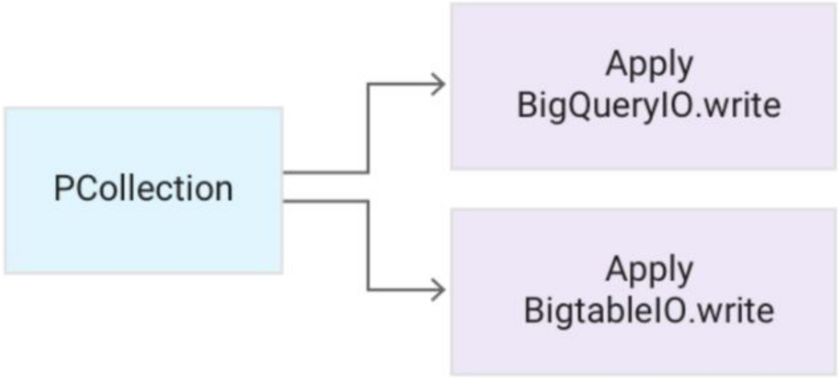
Milestone 12:

- What: Final Demos and Presentations.
- When: Week of May 4th. M-F 6:00pm - 8:00pm.
- Where: Zoom.
- Requested Action: Email me your preferred time(s) by EOD Friday.

Extra Credit Project:

- Request your repo by email no later than EOD Friday.

1) What does this Dataflow usage pattern mean?



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 PCollection can be written to multiple data sinks including BigQuery and Bigtable.

C. The PCollection can only be written to BigQuery or Bigtable.

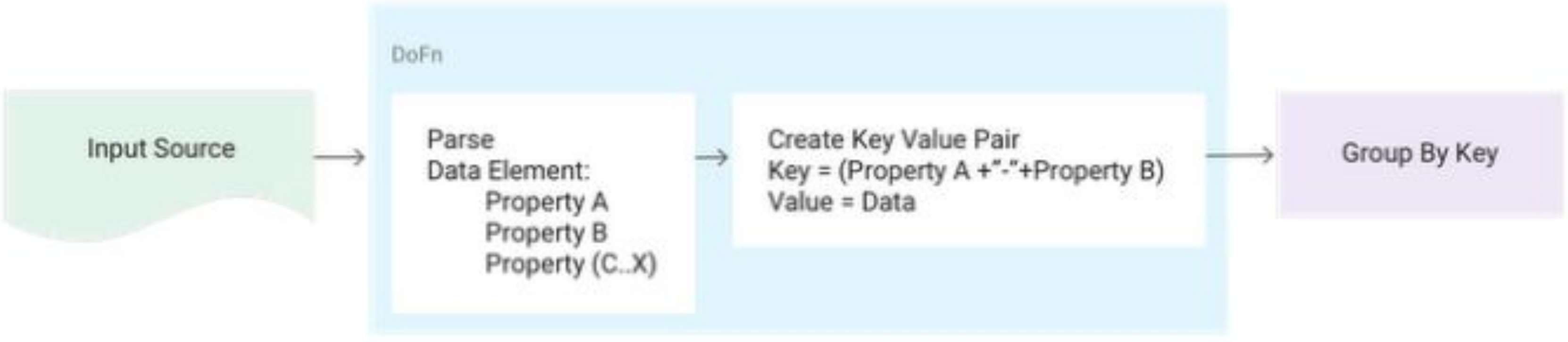
2) How do the authors suggest handling bad data?

A. Send it out of the DoFn as a SideOutput.
B. Send it into the DoFn as a SideInput.
C. Write it to an error log without writing 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. Create a Jupyter notebook instance and run the pipeline from a notebook.
C. Trigger the pipeline from Apache Airflow.

4) What does this Dataflow usage pattern mean?



A. The GroupByKey step requires a preceding DoFn step in the pipeline.

B. The GroupByKey step 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 a PCollection of any size with another PCollection that is small?

A. Use a CoGroupByKey transform
B. Use a SideInput to a ParDo
C. Use a GroupByKey transform
D. None of the above

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) Load Oscars data into BQ
2) Start up Jupyter notebook
3) Pull down latest code snippets
4) Practice debugging a few Beam pipelines

Practice Problem 1

Run and fix `oscars_1.py` and `oscars_2.py`.

Practice Problem 1

Run and fix `oscars_1.py` and `oscars_2.py`.

What were the cause of the errors?

A. Syntax errors
B. Logic errors
C. Syntax + logic errors

Practice Problem 2

Run and fix `oscars_3.py` and `oscars_4.py`.

Practice Problem 2

Run and fix `oscars_3.py` and `oscars_4.py`.

What was the cause of the bugs?

A. Syntax errors
B. Logic errors
C. Syntax + logic errors.

Code Review:

Data transforms implemented in both Beam and SQL

1) `covid_19_ingest.ipynb`
2) `covid_19_model.ipynb`
3) `Event_beam.py` and `Event_beam_dataflow.py`
4) `Location_beam.py` and `Location_beam_dataflow.py`

Observations:

- SQL transforms require several intermediate tables and more steps than Beam
- Beam pipelines require more effort to implement than SQL, but more maintainable

Milestone 10

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