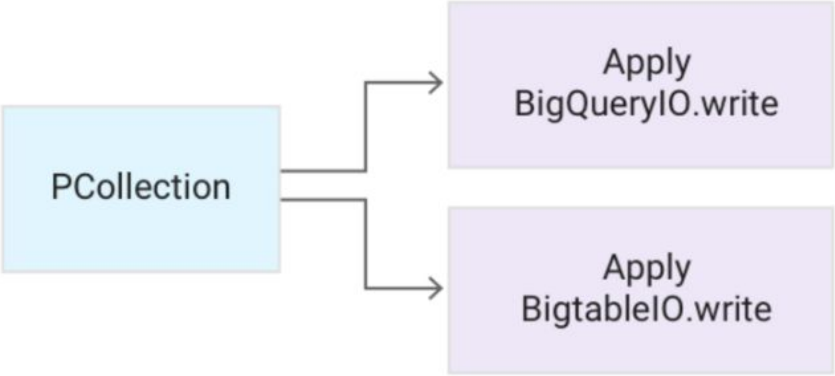
CS 327E Class 10

April 15, 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.

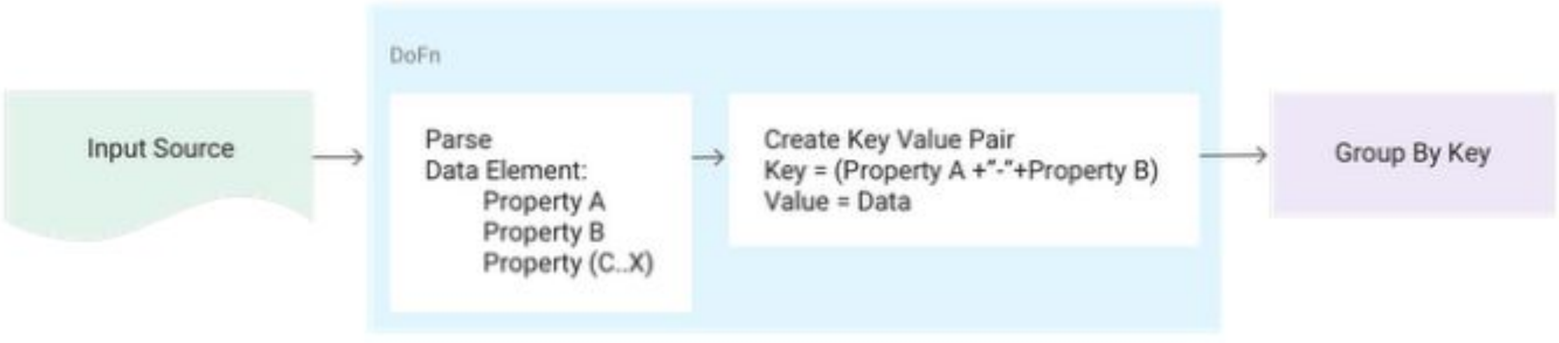
3) What method do the authors suggest for triggering a Dataflow pipeline that needs to start after a file has been uploaded to Google Cloud Storage?

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. HttpUnauthorizedError()}
2. RuntimeError: Transform "Write File" 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. NameError: global name 'pvalue' is not defined
8. RuntimeError: Could not successfully insert rows to BigQuery table

Hands-on Exercise

`git clone https://github.com/cs327e-spring2019/snippets.git`

or

`git pull origin master` to pull down the latest

Let’s start with: `nomination_count_6.py`

Practice Problem

Debug and fix the code in `nomination_count_9.py`

Practice Problem

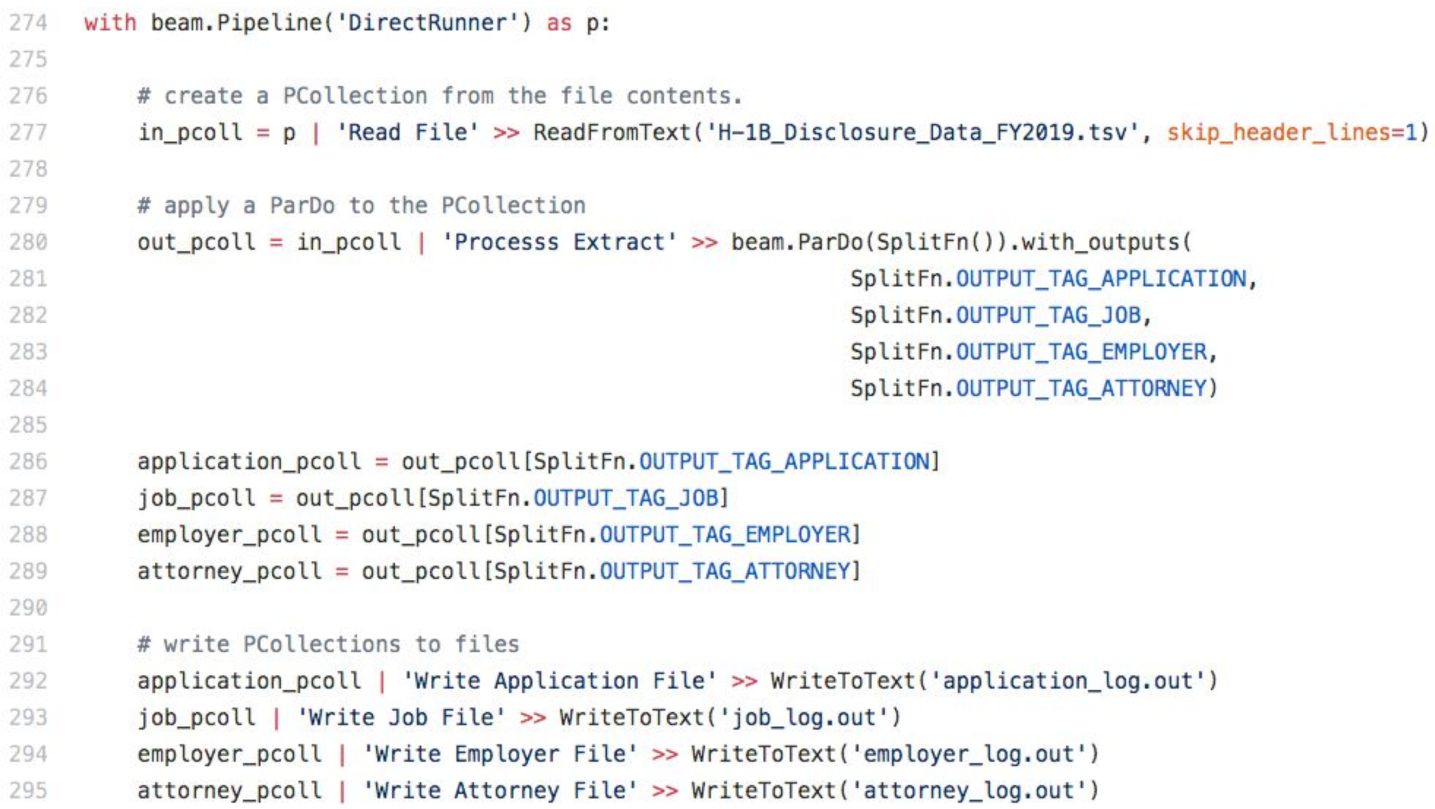
Debug and fix the code in `nomination_count_9.py`

What was the cause of the error?

A. Invalid record format for writing to BQ
B. Invalid table schema specification
C. BQ tables don’t exist
D. BQ tables already exist

ETL vs ELT

Transform-Load Example



```python
with beam.Pipeline('DirectRunner') as p:

    # create a PCollection from the file contents.
    in_pcoll = p | 'Read File' >> ReadFromText('H-1B_Disclosure_Data_FY2019.tsv', skip_header_lines=1)

    # apply a ParDo to the PCollection
    out_pcoll = in_pcoll | 'Processs Extract' >> beam.ParDo(SplitFn()).with_outputs(
                                                          SplitFn.OUTPUT_TAG_APPLICATION,
                                                          SplitFn.OUTPUT_TAG_JOB,
                                                          SplitFn.OUTPUT_TAG_EMPLOYER,
                                                          SplitFn.OUTPUT_TAG_ATTORNEY)

    application_pcoll = out_pcoll[SplitFn.OUTPUT_TAG_APPLICATION]
    job_pcoll = out_pcoll[SplitFn.OUTPUT_TAG_JOB]
    employer_pcoll = out_pcoll[SplitFn.OUTPUT_TAG_EMPLOYER]
    attorney_pcoll = out_pcoll[SplitFn.OUTPUT_TAG_ATTORNEY]

    # write PCollections to files
    application_pcoll | 'Write Application File' >> WriteToText('application_log.out')
    job_pcoll | 'Write Job File' >> WriteToText('job_log.out')
    employer_pcoll | 'Write Employer File' >> WriteToText('employer_log.out')
    attorney_pcoll | 'Write Attorney File' >> WriteToText('attorney_log.out')
```

Source File: https://github.com/shirleycohen/h1b_analytics/blob/master/transform_load_h1b_data_extract.py

Milestone 10

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