Class of 03/04/2019

Review of ParDo

- Maps 1 element to any number of output elements
- Invokes a user-defined function on each element in a PCollection independently
- The function belongs to a user-defined subclass of DoFn ("Do Function")
    - The function looks like `process(self, element)`
- side input - an extra argument passed into a ParDo function to allow access to that resource during computation in the ParDo

GroupByKey

- Takes a PCollection where the elements of the PCollection are tuples of two values
    - the first value is the key, and the second value is the value
- Returns a PCollection where the elements of the new PCollection are tuples of two values
    - the first value is the key the transform grouped by, and the second value is a list containing all the values with that key
- Related to `GROUP BY` from SQL

Flatten

- Merges PCollections of matching schemas.

CoGroupByKey

- Maps a collection of different key-value pairs to a schema similar to **GroupByKey**
- Multiple key-value pair PCollections are combined and result in a key-value pair

Running with DataflowRunner

- Runs on distributed computers, unlike DirectRunner which runs on our VM
    - No access to the local file system of our VM, output files now write to our buckets
    - print statements no longer will print
- Specify a bucket with folders staging/ temp/ output/