Tutorial 3c – GroupBy, GroupByKey, CoGroupByKey and GroupIntoBatches Transform in Apache Beam

GroupBy:

- Takes a collection of elements and produces a collection grouped, by properties of those elements.
- Unlike GroupByKey, the key is dynamically created from the elements themselves.

GroupByKey:

- GroupByKey is a Beam transform for processing collections of key/value pairs.
- Takes a keyed collection of elements and produces a collection where each element consists of a key and all values associated with that key.
- It's a parallel reduction operation, analogous to the Shuffle phase of a Map/Shuffle/Reduce-style algorithm.
- For example, if you have a collection that stores records of customer orders, you might want to group together all the orders from the same postal code (wherein the "key" of the key/value pair is the postal code field, and the "value" is the remainder of the record).

CoGroupByKey:

- Aggregates all input elements by their key and allows downstream processing to consume all values associated with the key.
- While GroupByKey performs this operation over a single input collection and thus a single type of input values, CoGroupByKey operates over multiple input collections.
- **CoGroupByKey expects a dictionary of named keyed PCollections**, and produces elements joined by their keys. The values of each output element are dictionaries where the names correspond to the input dictionary, with lists of all the values found for that key.
- As a result, the result for each key is a tuple of the values associated with that key in each input collection.

GroupIntoBatches:

- Batches the input into desired batch size.

Resources:

- https://beam.apache.org/documentation/transforms/python/aggregation/groupby/
- https://beam.apache.org/documentation/programming-guide/#groupbykey
  - https://beam.apache.org/documentation/transforms/python/aggregation/groupbykey/
- https://beam.apache.org/documentation/programming-guide/#cogroupbykey
  - https://beam.apache.org/documentation/transforms/python/aggregation/cogroupbykey/

- https://beam.apache.org/documentation/transforms/python/aggregation/groupintobatches/