Tutorial 3 b – ParDo, Keys, Kvswap, Values, ToString Transform in Apache Beam:

- **Core beam transform:** parDo
- **Element wise:** Keys, Kvswap, Values, ToString

ParDo:

- ParDo is a Beam transform for generic parallel processing.
- The ParDo processing paradigm is similar to the "Map" phase of a Map/Shuffle/Reduce-style algorithm: a ParDo transform considers each element in the input PCollection, performs some processing function (your user code) on that element, and emits zero, one, or multiple elements to an output PCollection.
- ParDo is useful for a variety of common data processing operations, including:
- Filtering a data set.
- Formatting or type-converting each element in a data set.
- Extracting parts of each element in a data set.
- Performing computations on each element in a data set.
- When apply a ParDo transform, need to provide user code in the form of a DoFn object.

Dofn:

- DoFn is a Beam SDK class that defines a distributed processing function.
- The DoFn object that you pass to ParDo contains the processing logic that gets applied to the elements in the input collection.
- You don't need to manually extract the elements from the input collection; the Beam SDKs handle that for you.
- Your process method should accept an argument element, which is the input element, and return an iterable with its output values.
- A given DoFn instance generally gets invoked one or more times to process some arbitrary bundle of elements.
- Your method should meet the following requirements:
- You should not in any way modify the element argument provided to the process method, or any side inputs.
- Once you output a value using yield or return, you should not modify that value in any way.

Keys:

- Takes a collection of key-value pairs and returns the key to each element.

Values:

- Takes a collection of key-value pairs and returns the value of each element.

ToString:

- Transforms every element in an input collection to a string.
- Any non-string element can be converted to a string using standard Python functions and methods.
- Many I/O transforms, such as textio.WriteToText, expect their input elements to be strings.
    - Key-value pairs to string
    - Elements to string
    - Iterables to string

Kvswap:

- Takes a collection of key-value pairs and returns a collection of key-value pairs which has each key and value swapped.

Resources:

- https://beam.apache.org/documentation/programming-guide/#pardo
- https://beam.apache.org/documentation/transforms/python/elementwise/keys/
- https://beam.apache.org/documentation/transforms/python/elementwise/values/
- https://beam.apache.org/documentation/transforms/python/elementwise/tostring/