11/5/18 Notes

Terminology

- **Dataflow:** Distributed system to handle computations on large amounts of data.
- **Apache Beam:** Programming model behind Dataflow, which was developed by Google and then open sourced by donating to Apache.
- **Pipeline:** A data processing task from start to finish. Sequence if transforms that takes in and output a PCollection.
- **PCollection:** Collection of data elements. Normally, PCollections are split up and distributed on different processing locations.
- **Transform:** A data transform operation. The Transforms are run on the different subsets of the PCollection distributed among different machines.

Pipeline

A directed acyclic graph that has Transforms as nodes and the Transforms are linked through the PCollections.
The Pipeline takes the input PCollection from data source(s).

Process:

- Read in data as a PCollection from data source(s)
- Apply Transforms on the PCollection. Transforms read in PCollections and output PCollections
- Outputs one or more PCollections as one or more data sinks.

PCollection

- A collection of data elements.
- The elements can be of any type.
- Subsets are formed from the PCollection and distributed across machines and this segmentation happens in the background over which the user has no control over.
- A transform specified in the Pipeline must execute independently on every element of the PCollection.

Transforms

Types:

- **Element-wise:** Takes in one input and can output 0, 1 or many outputs. Eg: ParDo, Map
- **Aggregation:** Takes in many inputs and outputs one or fewer inputs. Eg: GroupByKey
- **Composite:** Combines Element-wise & Aggregation. Eg: GroupByKey & ParDo

Properties:

- **Serializable:** Data can be formatted to transfer it between machines.

- **Parallelizable:** Can run independently on different machines on different subsets of the data.
- **Idempotent:** Reliably output the same result every time the Transform is run on the same data.

ParDo (Parallel Do)

- Takes in one input and outputs 1, 0 or many outputs.
- User defines the operation to be performed during the ParDo.
- Outputs is a PCollection