- **PCollection:** It is an abstraction represents a potentially distributed, multi-element data set. It represents a distributed data set that our beam pipeline operates on.
  - **Immutability:** Pcollections are immutable in nature. Applying a transformations on a pcollection results in creation of new pcollection.
  - **Element type:** The elements in pcollection may be of any type, but all must be of same type.
  - **Operation type:** Pcollection does not support grained operations. We cannot apply transformations on specific elements in pcollection.
  - **Timestamps:** Each element in pcollection has an associated timestamp with it.
  - **Unbounded pcollections:** An unbounded PCollection represents a data set of unlimited size. Source assigns the timestamps.
  - **Bounded pcollections:** A bounded PCollection represents a data set of a known fixed size. Every element is set to same timestamp.
  - **No Random access:** Can't access data using index or some specific element. No size restriction.
  - **Ptransform:** Ptransform represent a data processing operation, or a step in our pipeline. Ex., Map, Groupby, FlatMap, ParDo, filter, flatten, combine etc.

- **PCollection characteristics:**
  - A PCollection is owned by the specific Pipeline object for which it is created; multiple pipelines cannot share a PCollection.
- Resources:
  - https://beam.apache.org/documentation/programming-guide/#pcollections
  - https://beam.apache.org/releases/pydoc/2.36.0/apache_beam.io.textio.html?highlight=readfromtext#apache_beam.io.textio.ReadFromText

Cloud & AI Analytics