Google Cloud Dataflow

Cloud Dataflow is a unified programming model and a managed service for developing and executing a wide variety of data processing patterns.

Features

- Managed service for executing scalable and reliable data piplines.
- Write code once and get batch and streaming.
- Clusters are sized for you.
- Processes data using Comute Engine instances.
- Integrates with GCP services like Cloud Storage, Cloud Pub/Sub, BigQuery, Bigtable.
- Open source Java and Python SDKs.

Concepts

- Pipeline: Data processing and transformation flow.
- PCollection: A set of data in your Pipeline.
- Transformation: Processing done on the PCollections in the Pipeline.

PCollection => Pipeline => Transformation => Pipeline => PCollection