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.

It separates data processing requirements from data source being batch or stream data.

Based off the Apache Beam open-source project.

Features

- No-ops managed service for executing scalable and reliable data piplines.
- Write code once and get batch and streaming.
- Clusters are sized for you (liquid sharing, autoscaling mid-job).
- 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.

Simple Pipeline:

```
Input => PCollection => Transformation => PCollection => Output
```

Multiple Transform Pipeline:

```
BigQuery => PCollection => Transform => PCollection 'A' Names
                                   \
                                    \  => PCollection 'B' Names
```

Merge Pipeline:

```
BigQuery => PCollection => Transform => PCollection 'A' Names => Transform => PCollection 'A/B' Names
                                   \                                    /
                                    \  => PCollection 'B' Names =>  /
```

Multiple Input Pipeline:

```
     BigQuery => PCollection Names & Addresses => Transform => PCollection Names/Addresses/Orders
                                                            /
Cloud Storage => PCollection Names & Orders    =>  /
```