# 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 & Order =>  /
```

### Pipeline

In the Dataflow SDKs, a pipeline represents a data processing job. You build a pipeline by writing a program using a Dataflow SDK. A pipeline consists of a set of operations that can read a source of input data, transform that data, and write out the resulting output.

### PCollection

A PCollection represents a potentially large, immutable "bag" of elements. There is no upper limit on how many elements a PCollection can contain; any given PCollection might fit in memory, or it might represent a very large data set backed by a persistent data store.

A PCollection has several key aspects in which it differs from a regular collection class:

- A PCollection is immutable. Once created, you cannot add, remove, or change individual elements.
- A PCollection does not support random access to individual elements.
- A PCollection belongs to the pipeline in which it is created. You cannot share a PCollection between Pipeline objects.

### Transform

In a Dataflow pipeline, a transform represents a step, or a processing operation that transforms data. A transform can perform nearly any kind of processing operation, including performing mathematical computations on data, converting data from one format to another, grouping data together, reading and writing data, filtering data to output only the elements you want, or combining data elements into single values.