Apache Beam and Dataflow

What is Apache Beam?

Apache Beam is a unified programming model for batch and streaming data processing jobs. It provides different api’s to interact with different data sources and process data using variety of backend’s be it Spark or Dataflow. In this way the data can reside somewhere else and compute can happen on it in serverless manner or on specified backend.

Favorable Distributed Processing Engines

Apache beam supports below mentioned distributed engines.

- Apache Apex
- Apache Flink
- Apache Gearpump
- Apache Samza
- Apache Spark
- Google Cloud Dataflow
- Hazelcast Jet

Concepts in Apache Beam

With Apache Beam, one can build workflow graphs (pipelines) and execute them. The key concepts to understand are below:

- *PCollection* – represents a data set which can be a fixed batch or a stream of data (Similar to RDD/DF in Spark)
- *PTransform* – a data processing operation that takes one or more *PCollection*s and outputs zero or more *Pcollection*s (Similar to transformation in Spark)
- *Pipeline* – represents a directed acyclic graph of *PCollection* and *PTransform*, and hence, encapsulates the entire data processing job. (Lineage DAG in Spark)
- *PipelineRunner* – executes a *Pipeline* on a specified distributed processing backend

In simple words, a *PipelineRunner* executes a *Pipeline,* and a *Pipeline* consists of *PCollection* and *PTransform*.

What is DataFlow?

Google Cloud Dataflow is a cloud-based data processing service for both batch and real-time data streaming applications. It enables developers to set up processing pipelines for integrating, preparing and analyzing large data sets, typically required for big data processing.

Dataflow Features

Some of the important features of Dataflow which makes it a great choice for big data computing are as follows.

- Fully managed by Google
- Completely serverless
- Autoscaling based on throughput
- Cost optimized
- Interactive GUI