**Google Cloud DataFlow real-time service for batch and stream processing**

Shan Zhou, Jay Upadhyay,Wendy Jiang

## Summary

Google Cloud Dataflow is the new cloud service that is designed to simplify the mechanics of large-scale data processing, it allows people to concentrate on the logical composition of data processing job, rather than the physical orchestration of parallel processing.

Why use Google Dataflow:

1. It automates the management of processing resources and frees people from operational tasks.
2. On demand, no need to buy reserved compute instances.
3. Automated and optimized work partitioning.
4. Auto scaling of worker resources.
5. Good monitoring using UI and command-line and Stackdriver.
6. Integrating with Cloud Storage, Cloud Pub/Sub, Cloud Datastore, Cloud Bigtable and BigQuery and can be extended to interact with other sources and sinks like Kafka and HDFS

**The goal** of our project is to provide an overview of the Google Cloud Dataflow and to demonstrate how to build and execute a simple pipeline.

What we have done:

1. Create Storage bucket and installed Cloud SDK in Mac and run an example pipeline remotely using Python.
2. Installed Cloud SDK in Windows and run an example pipeline on Cloud Dataflow Service using Java and Apache Maven
3. Run a mobile gaming pipeline to experience processing in batch and windowing and streaming with Real-Time Game Data. Input source are from Cloud data storage for batch and Pub/Sub for streaming. Results are stored locally and Cloud storage and BigQuery tables.
4. Created an own pipeline using Java and applied pipeline transformation and used google console for monitoring and logs for debugging.

**Comparison between spark and google cloud dataflow**:

We used a mobile gaming scenario as an example to compare dataflow vs spark in detail using three different kinds of pipelines:

- classic batch pipeline
- window batch pipeline
- streaming pipeline

For more details of this part, please look at comparison dataflow vs spark.pdf

Reference: https://cloud.google.com/dataflow/model/programming-model
https://cloud.google.com/dataflow/docs/

YouTube URL of the full presentation video: https://youtu.be/-2sF5Q0TplA
YouTube URL of the 2min preview presentation video: https://youtu.be/l2eHgQAWdio