**Cloud Report Factory – Proof of Concept Requirements**

**Purpose:**

The Solution for the PoC is to create a data pipeline in python which supports both BATCH and STREAMING with Unified BEAM Model which uses runner as DataFlow and orchestrated on Apache Airflow.

**Technologies:**

* Source System – Oracle Database 12c
* Target System – Google Big Query
* Real time Replication (CDC) – STRIIM
* Apache BEAM MODEL – For BATCH & STREAMING MODE
* Runner: Cloud Data Flow
* Orchestration Tool: Apache Airflow

**Tables for the PoC :**

* GL\_BALANCES – 228Mb (STREAMING CDC)
* GL\_JE\_LINES – 136Mb
* GL\_JE\_HEADERS – 25Mb
* GL\_CODE\_COMBINATION

**Considerations:**

* Streaming CDC is only for GL\_BALANCES where STRIIM trail version will be used.
* The oracle tables will be hosted in a EC2 Instance
* STRIIM can be installed in AWS EC2 or GCP.
* AWS and GCP accounts for development will be issued.

**SUCCESS CRITERIA:**

* Build seamless pipeline with orchestration in Airflow for batch and real-time streaming using Striim.
* Documentation of Data Pipelines with configuration.
* Use of Open Source Software as much as possible.
* Support of Multi Zone Batch loads (parameterised)