This Project aims to create an Ingestion framework to load Data (CSV files) from Local files system to a Distributed Big Data Infrastructure (Hadoop Cluster on GCP).

* MySQL
* Apache Sqoop
* Apache Hive

**Required**

Google Cloud Platform (GCP) -a suite of cloud computing services

.csv Files -Used as an Input Data

**Create Table in MySQL db**

Get schema of table using csv file with proper datatype

import pandas

import pymysql

import sqlalchemy

When writing data from a Pandas DataFrame to a SQL database, I used

* DataFrame.to\_sql method.
* A.to\_sql(con=engine, name='application\_test\_table', if\_exists='replace')

We could also use LOAD DATA INFILE

**Auditing**

Developed Audit component using Python + MySql

Checking Attributes count and Records count.

**Ingesting Data to Hadoop Cluster**

To Ingest Data From MySQL RDBMS to Hadoop Cluster, I used cli command

* hdfs dfs -put <Source File from LFS> <Destination on HDFS>

We could also use copyFromLocal command

**Hive and sqoop**

To ingest data in Hive table from MySQL

* query='''sqoop import --connect {c} --username {un} --password {pd} --table {tb} --m 1 --target-dir {td} --hive-import --create-hive-table'''.format(c=val['conn'],

un=val['username'],

pd=val['password'],

tb=val['tabName'],

td=val['tarDir'],)

We could also create External Table on top of HDFS file using

create external table if not exist <table name>() command

**Key Features**

* Designed Effective and Robust Data Ingestion Framework
* Importing full and Delta data load from source.
* Ingestion jobs using Sqoop Incremental Inputs
* Created Automated Ingestion Framework using configuration driven approach.
* Created Framework/ Automated script using Python & json file
* Project is in accordance with Error Handling, File auditing ( checking for data quality and consistency) and logging of steps.