This Project aims to build and integrate a Validation framework using Python + Sqoop/SQL + HDFS which validates data loaded between source and target systems.

* Python
* MySQL
* Apache Sqoop
* Apache Hive

**Required**

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

My SQL Tables -Used as an Input Data

**Ingest Data into HIVE Table**

Used sqoop import --hive-import to import data seamlessly into Hive tables from RDBMS systems. import sys

import json

import subprocess

Used subprocess module and parameterized the script using Json file

cmd='''sqoop import \

--connect {conn} \

--username {usr} \

--password {pwd} \

--table {tnm} \

--target-dir {tdr} \

--create-hive-table \

--hive-import \

--hive-table {htn}'''.format(conn=val['connectionURL'],

usr=val['usrName'],

pwd=val['pWord'],

tnm=val['mysqlName'],

tdr=val['targetDir'],

htn=val['hiveName'])

**Validating data between source & target systems**

Generating Row counts (Number of records) for MySQL table, before creating hive table.

Generating Row counts (Number of records) for HIVE table

Designing validation framework by building validation pipelines using hive, sqoop and mysql

using configuration driven approach

def getCount(std):

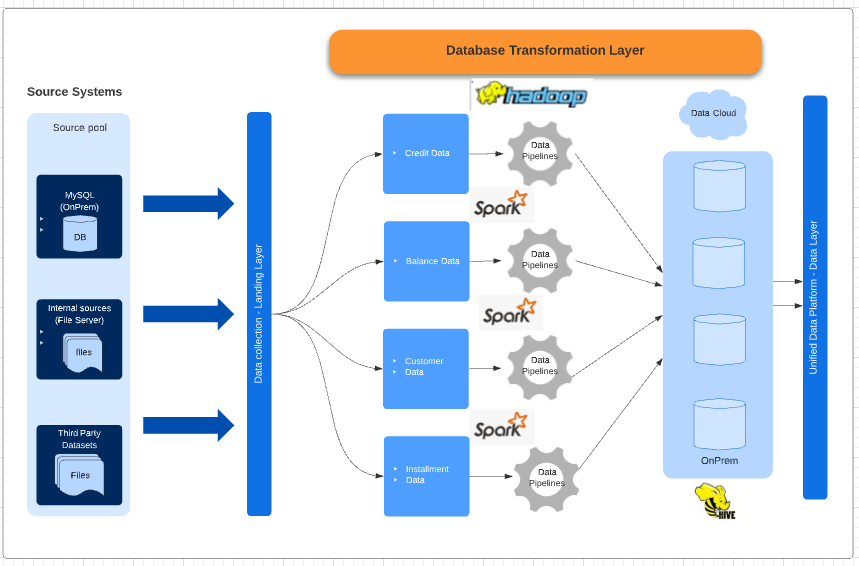
count = re.findall(r'\d+', std)[0]

count=int(count)

return count

This validates data loaded between source and target systems

**Project Architecture**



**Key Features**

* Designed Effective and Robust Validation Framework
* Importing full load from source.
* Ingestion jobs using Sqoop.
* Created Automated Validation 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.