This Project aims to create an spark job -Pyspark pipeline to join 4 csv Files (from hadoop cluster) and 4 Hive Tables as per ER Diagram.

* PySpark
* Scala
* PySQL

**Required**

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

.csv Files -Used as an Input Data

Hive Tables -Used as an Input

**Read Files form HDFS**

Created Spark Session

Read csv files stored on Hadoop cluster and stored in respective Dataframes

* self.hd1\_bureauDF = self.spark.read.csv("/user/csv\_data/bureau.csv", header=True, inferSchema=True)
* self.hd2\_application\_testDF = self.spark.read.csv("/user/csv\_data/application\_test.csv", header=True, inferSchema=True)
* self.hd3\_application\_trainDF = self.spark.read.csv("/user/csv\_data/application\_train.csv", header=True, inferSchema=True)
* self.hd4\_bureau\_balanceDF = self.spark.read.csv("/user/csv\_data/bureau\_balance.csv", header=True, inferSchema=True)

We could also use

val customerDF = spark.read.csv("Filename with path")

**Auditing**

Developed Audit component

Checking Attributes count and Records count.

**Read HIVE Tables**

Read Hive Tables and stored inside respective DataFrames

* self.hi5\_credit\_card\_balanceDF = self.spark.sql("select \* from credit\_card\_balance")
* self.hi6\_installments\_paymentsDF = self.spark.sql("select \* from installments\_payments")
* self.hi7\_pos\_cash\_balanceDF = self.spark.sql("select \* from pos\_cash\_balance")
* self.hi8\_previous\_applicationDF = self.spark.sql("select \* from previous\_application")

Project is in accordance with Error Handling (try-except blocks), File auditing ( checking for data quality and consistency)

**Creating Pipeline To Join the Tables**

Joining is performed between

* csv to csv
* csv to hive tables
* Hive to csv

Inner joins operations are performed on various sets of files and tables as per ER Diagram.

We could also use Outer joins, Left Joins and Right Joins

**ER Diagram For Joining Keys**



**Key Features**

* Designed Effective and Robust Spark Job
* Read Files from Hadoop Cluster and Hive Tables.
* Created Pipeline for joining tables and files using modular programming approach.
* Created Spark Job using Pyspark & SQL Queries.
* Project is in accordance with Error Handling, File auditing ( checking for data quality and consistency)
* Inner Joins are performed as per ER Diagram