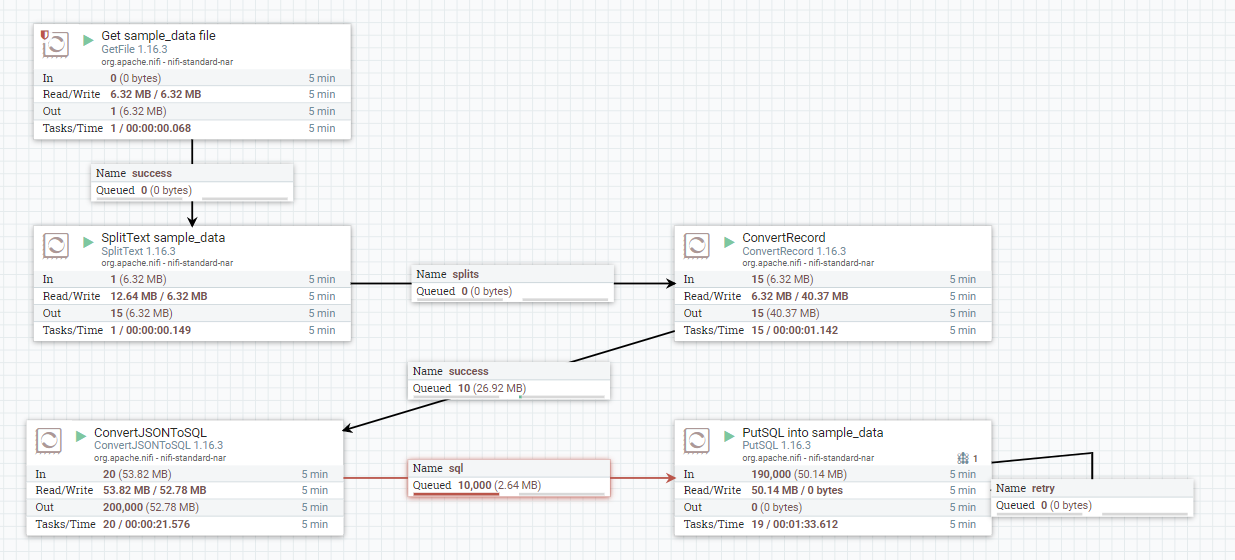
**ETL FLOW FOR DATA INGESTION – CREDIT RISK DATASET**

**1.Method:**

- I use Apache Nifi to build a PoC for a data ingestion system to make incoming CSV data easy to use / query by our Data Scientists as below:



**2.How to build this data ingestion:**

**-Step 1:**

+ I put sample\_data.csv in this folder on my local machine:

Folder: C:\Users\User\Desktop\data

FilePath: C:\Users\User\Desktop\data\sample\_data.csv

**-Step 2:**

+ I create a Process Group namely “sample\_data”

**-Step 3:**

+ I use Processor “GetFile” to load this file from the folder above, with configuration:

* Scheduling Strategy: CRON driven
* Run Schedule: 0 0 0/1 1/1 \* ? \*
* Execution: All nodes
* Input Directory: C:\Users\User\Desktop\data
* Keep Source File: false

**-Step 4:**

+ I use Processor “SplitText” to split the files into many files to run run parallel tasks, with configuration:

* Line Split Count: 10000
* Header Line Count: 1
* Remove Trailing Newlines: true
* Relationships: Auto Terminate on failure/original

**-Step 5:**

+ I use Processor “ConvertRecord” to convert data from csv to json, with configuration:

* Record Reader: CSVReader (this is created from Controller Service Details in sample\_data Process Group, and enabled after that).

**Note: In CSVReader, configure NullString = NA (this NA value in sample\_data.csv will be considered as null value)**

* Record Writer: JsonRecordSetWriter (this is created from Controller Service Details in sample\_data Process Group, and enabled after that)

**Note: In JsonRecordSetWriter, configure Suppress Null Values = Always Suppress (How the writer should handle a null value)**

* Include Zero Record FlowFiles: true
* Relationships: Auto Terminate on failure

**-Step 6:**

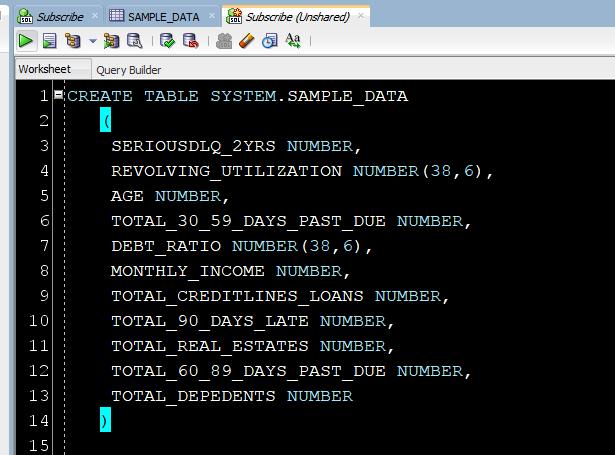
+ I use Processor “ConvertJSONToSQL” to convert json to sql, with configuration:

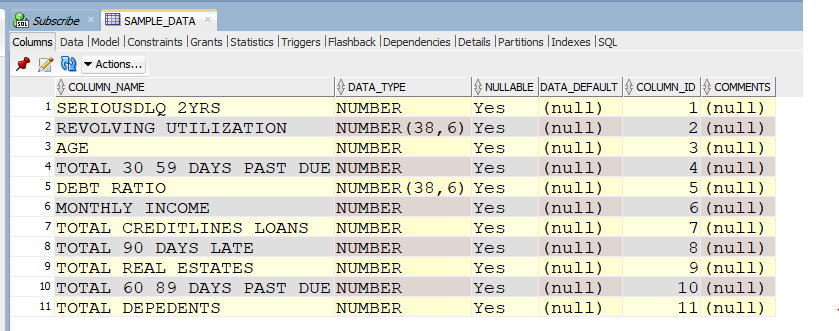
* JDBC Connection Pool: oracle database (this is created from Controller Service Details in sample\_data Process Group, and enabled after that)

Note: need to configure **Database Connection URL, Database Driver Class Name,** Database Driver Location(s) for this JDBC Connection Pool

* Statement Type: INSERT
* Table Name: SAMPLE\_DATA – this is table name on oracle database
* Catalog Name: SYSTEM – this is schema name on oracle database for this table
* Translate Field Names: true
* Relationships: Auto Terminate on failure/original

**-Step 7:** Create table on oracle database:





**-Step 8:**

+ I use Processor “PutSQL” to put sql into oracle database (in this case), with configuration:

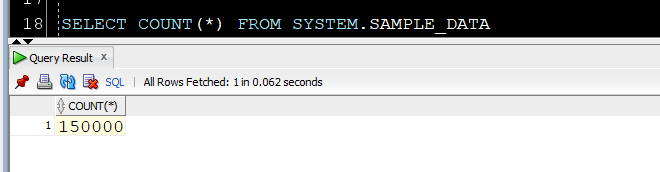
* JDBC Connection Pool: oracle database (this is created from Controller Service Details in sample\_data Process Group, and enabled after that)
* Relationships: Auto Terminate on failure/success

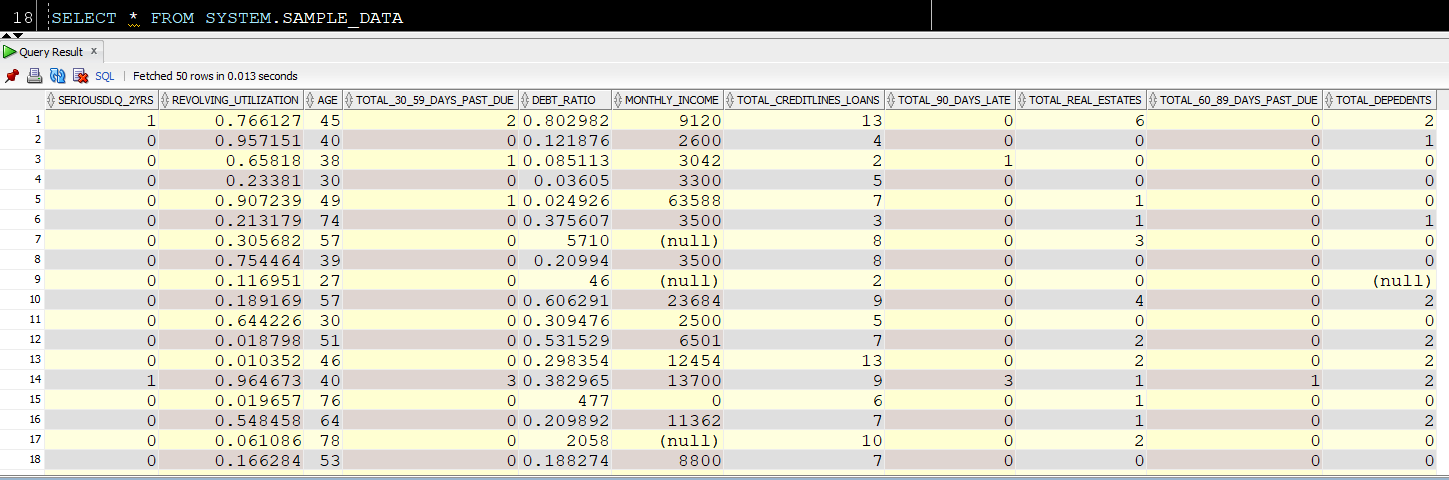
**NOTE: For the configurations that I didn’t meantion here, they will be set as default values**

**-Step 9:**

+ Compare data between source and target table -> this is correct

+ All the NA values are now considered as Null values





**3. Result: Pls find flow.xml.gz file in my github – This is ETL Flow designed for Data Ingestion System in this Project**

<https://github.com/nhungluong1989/paidy_assignment/blob/main/flow.xml.gz>

------------------------------------------------------ALL DONE-----------------------------------------------------------------