# Data Quality Scorecard Development

**Overview:**

The project is a financial institutional project with the goal of better understanding of data quality for different business lines. Different profiling rules (null checks, duplicate checks, data type validation, referential integrity) are implemented.

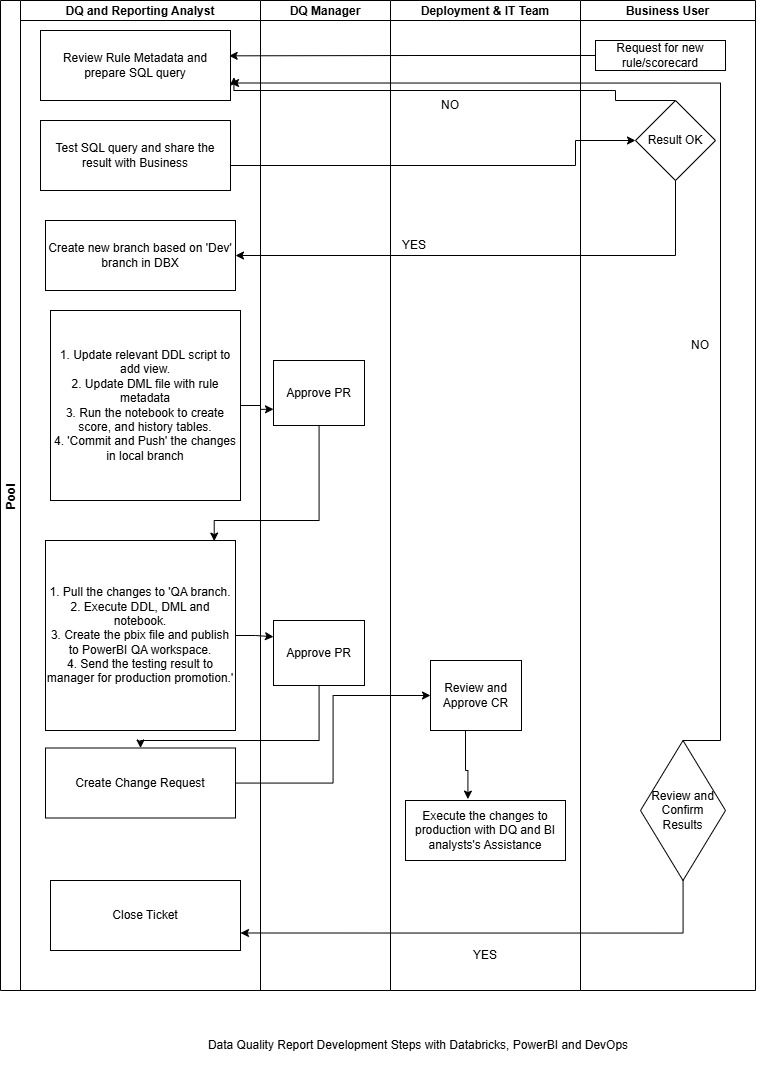
**Tech Stack:**

* SQL
* Databricks (dev,qa and prod environment)
* PowerBI
* Azure Devops
* PySpark

**Objectives**:

* Perform data profiling to detect anomalies in financial institution.
* Define data quality rules (missing values, duplicates, invalid/violation).
* Implement data cleansing techniques.
* Visualize data quality scorecard and trends in Power BI.

**Process Flow of Scorecard Development:**

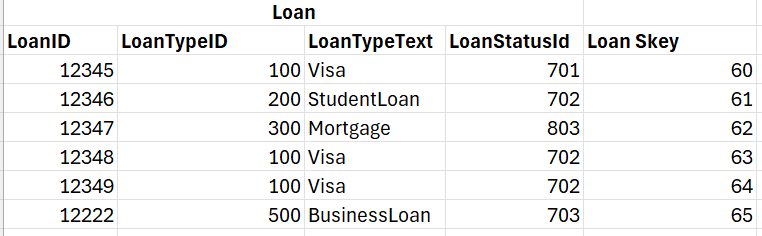
****

**Data Quality Scorecard Developing Steps**

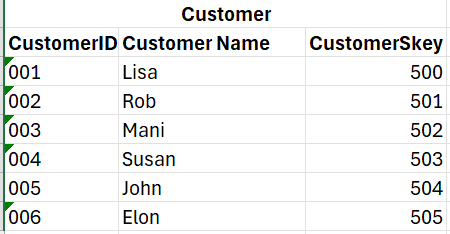
* Business line requests for developing a new report or adding a new profiling rule to the existing report. Then following steps can be done to create a new rule/new scorecard for data quality profiling.
* Upon receiving the request, sql query is prepared and tested against prod data in Databricks using spark SQL.
* If the total number of records and violation looks good, then it’s recommended to show the result to business user.
* If they agree with the result, then create a local branch in databricks based on ‘dev’ branch for development purpose. The query then incorporated into the appropriate DDL scripts to create the views in Databricks.
* Run the notebook in databricks which creates appropriate score and history table.
* Push the code to Dev. Create a pull request in Azure DevOps to pull the codes in QA branch.
* If the testing is successful in QA, create a Change Request (CR), get approval and push the code to ‘Production’ branch.
* Run the code in production branch, create the pbix file. Connect appropriate score and history tables according to the view and publish to PowerBI.

**Sample Data Fields**

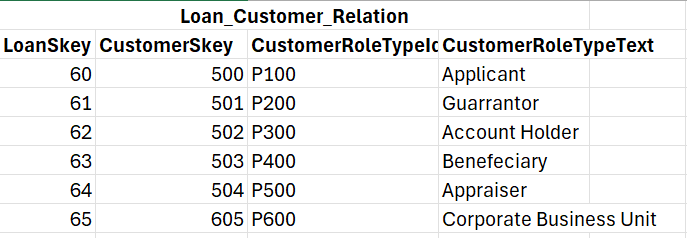
**Table 1**

****

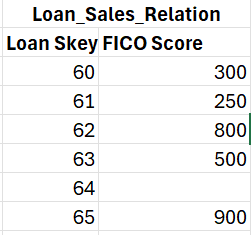
**Table 2**

****

**Table 3**

****

**Table 4**

****

**Sample Data profiling using SQL**

The requirement of business line is to find out the customer/loan whose FICO Score is not between 200 and 800 and FICO score is blank.

**SQL Code**

Select

Laon.LoanID AS Loan\_Number,

Loan.LoanStatusID AS Loan\_Status\_ID,

Loan.LoanTypeId AS Loan\_Type\_ID,

Customer.CustomerName AS Customer\_Name,

Customer.CustomerRoleTypeId AS Customer\_Role\_Type\_ID,

Customer.CustomerRoleTypeText AS Customer\_Role\_Type\_Text,

Loan\_Sales\_Relation.FICOScore AS FICO\_Score

FROM

Loan Loan

Left join Loan\_Customer\_Relation Loan\_Customer\_Relation

ON Loan.LoanSkey = Loan\_Customer\_Relation.LoanSkey

Left join Customer Customer

ON Loan\_Customer\_Relation.CustomerSkey = Customer.CustomerSkey

Left Join Loan\_Sales\_Relation Loan\_Sales\_Relation

ON Loan.LoanSkey = Loan\_Sales\_Relation.LoanSkey

Where Loan\_Customer\_Relation.CustomerRoleTypeId IN ('P100', 'P400')

AND Loan.LoanStatusId IN ('702','803')

AND LoanTypeId IN ('100','200','300')

AND

(

(Loan\_Sales\_Relation.FICOScore <> 0

AND Loan\_Sales\_Relation.FICOScore < 200

AND Loan\_Sales\_Relation.FICOScore > 800)

OR

Loan\_Sales\_Relation.FICOScore IS NULL

)

GROUP BY

Laon.LoanID,

Loan.LoanStatusID,

Loan.LoanTypeId,

Customer.CustomerName,

Customer.CustomerRoleTypeId,

Customer.CustomerRoleTypeText,

Loan\_Sales\_Relation.FICOScore

**PowerBI Dashboard:**

The above sql query will find out the loan/customer whose Credit Score is violating the rule. The dashboard can be built which shows the total number of records, total number of violations, trend and drilldown data to see the violation details.

Sample PowerBI dashboard will look like as follows:



**Data Cleansing in Source System:**

Once the report is published in PBI, then business can fix the data in the source system like SAP or other source systems.