**Project Disha**

**Data Platform Strategy Project**

**Data Quality Framework**



17th Oct 2024

Version 1.0

**Document Control Information-**

**Document Information**

|  |  |
| --- | --- |
| **Summary** | **Detail** |
| Document name | Data Quality Framework |
| Project name | Disha – Platform Strategy Project |
| Client | Aavas Financiers ltd |
| Document Author |  |
| Document Version | Draft |
| Date released |  |

**Document Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Update by** | **Remarks** |
| Draft |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Document Approval**

|  |  |  |  |
| --- | --- | --- | --- |
| **##** | **Team** | **Name** | **Date** |
| 1 | Analytics Team | Pavithira Lakshmi |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

1. Purpose of the Document……………………………………………………………………..4
2. Business Requirement………………………………………………………………..............4
3. Department Coverage under Data Quality………………………………….......................5
4. Data Quality Design………………………………………………………………….
   1. Data Quality Source Identification……………………………
   2. Data Quality Rule Identification…………………………….
   3. Data Quality Dimensions……………………………………………
   4. Data Quality Profiling Summary………………………………….
5. Monitoring………………………………………
6. **Purpose of the Document**

The purpose of this document is to outline and define the data quality dimensions relevant to the industry and to provide detailed descriptions for each dimension. This report aims to establish a clear understanding of key data quality concepts to ensure accurate, reliable, and effective data management across various departments. By defining and exemplifying these data quality dimensions, this document serves as a foundational reference for assessing and improving data quality practices within the organization. By understanding and applying these data quality dimensions, the organization can enhance data stewardship, improve decision-making, and ensure compliance with regulatory requirements. Aid in the establishment of data quality assurance processes and controls. This document provides a basis for evaluating and improving data quality through systematic checks and validations, ultimately leading to more reliable and accurate data.

1. **Business Requirements**

It is utmost necessary for Aavas to maintain high quality data to ensure regulatory compliance, operational efficiency, and customer satisfaction. Accurate and reliable data will ensure that Aavas as an organization will meet regulatory standards thus avoiding any consequences. Consistent and accurate data across systems enhances operation efficiency, enables strategic decision-making, and improves risk management, minimizing financial losses.

Robust data quality is crucial for efficient and effective credit assessment, optimizing loan approval process and reduce default rates. Additionally, data privacy and security are paramount to protect sensitive information of customers as well as the organization, complying with data protection laws. Accurate data facilitates transparent stakeholder reporting and audit processes reinforcing trust and confidence among management, investors, regulatory bodies as well as customers. A comprehensive data quality framework is indispensable for the sustainable growth and success of Aavas.

Key requirements for data quality include:

* Enable data cleansing and transformation processes to ensure that data is standardized, consistent, and free from errors or duplications
* Define the data validation rules that data should adhere to when being ingested into the data platform. These include checks for data types, ranges, and formats, as well as any domain-specific rules
* Capability to generate data quality reports and dashboards that provide insights into the overall data quality status within the data platform
* Enable a flexible and configurable data validation framework to define and enforce data quality rules, including format, range, and referential integrity checks
* Implement custom validation rules to accommodate domain-specific requirements, developing, and scheduling automated data quality checks and validations
* Capability to track and manage different versions of datasets to enable data lineage and historical analysis
* Integrate data quality processes, roles, and responsibilities into the broader data governance framework
* Facilitate data quality monitoring and profiling capabilities for the data platform. It should automatically assess data quality, identify data anomalies, and highlight potential data issues for remediation

1. **Department Coverage under Data Quality**

This document coverage has been classified for below departments across Aavas

|  |  |
| --- | --- |
| # | Departments |
| 1. | Customer |
| 2. | Collateral |
| 3. | Employee |
| 4. | Branch |
| 5. | Product |
| 6. | Loan |
| 7. | Regulatory |
| 8. | KYC |

* 1. Customer

The Customer department is responsible for managing and maintaining all information related to individuals or entities who use Aavas’ services. This includes personal details, account information, transaction history, and any interactions with Aavas.

Ensuring high quality data in customer master will minimize the errors in transactions, reduces duplicated records, and enhances decision-making by providing a single, consistent view of each customer. This will also ensure effective communication and customer satisfaction.

*Example: Maintaining accurate records of a customer’s name, address, contact details and officially verified documents (PAN, Aadhar, etc.) for active, closed and other type to check valid and invalid data.*

* 1. Collateral

The Collateral department oversees the management and valuation of assets pledged by borrowers to secure loans. This includes ensuring that collateral is properly valued, documented, and managed throughout the life of the loan.

Reliable and consistent collateral data ensures the proper valuation and tracking of assets, reduces the risk of errors in loan processing, and supports compliance with regulatory requirements. Accurate collateral data is critical for making informed lending decisions, managing credit risk, and maintaining the integrity of Aavas’ balance sheet.

*Example: Maintaining accurate record for each collateral such as collateral associated with the loans, legal property owners and address information for active, closed and other type to check valid and invalid data.*

* 1. Employee

The Employee department manages all aspects related to the Aavas’ workforce. This includes recruitment, performance management, benefits administration, and employee records.

Ensuring high data quality in employee master data is vital for efficient human resource management and operational effectiveness. Reliable employee data also enhances decision-making related to workforce planning, talent management, and performance evaluations.

*Example: Maintaining accurate records of employee details such as employee on-boarding, job position, salary, and contact information for valid and invalid check.*

* 1. Branch

The Branch department is responsible for the operations and management of individual branches. This includes overseeing branch staff, managing branch-specific transactions, and ensuring compliance with regulations and policies at the branch level.

*Example: Managing accurate records for each branch such as complete address information, branch type, branch code information, valid geographic location and branch region information for valid and invalid check.*

* 1. Product

The Product department focuses on the development, management, and marketing of the Aavas’ financial products and services. This includes designing new products, managing existing product portfolios, and analysing product performance.

*Example: Developing a new product, managing its features and benefits, and analysing its market performance for valid and invalid check.*

* 1. Loan

The Loan department handles all aspects related to the issuance and management of loans. This includes processing loan applications, assessing credit worthiness, disbursing funds, and managing loan repayments and collections.

High data quality in loan master data is critical for managing loan portfolios effectively and ensuring accurate financial reporting. Reliable and consistent loan data enables precise tracking of loan terms, payment schedules, interest rates, and borrower information, reducing the risk of errors in billing and collections.

*Example: Managing accurate records for each loan such as loan sanction amount should be mandatory for active and closed loans, sum of schedule principal amount should be less than or equal to loan sanction amount, dates for each loan, loan number format check, loan account status, loan tenure and repayments check for active, closed and other type to check valid and invalid data.*

* 1. Regulatory

The Regulatory department is responsible for ensuring that the bank operates in compliance with legal and regulatory requirements set by governing authorities. This includes overseeing adherence to laws, industry standards, and internal policies to mitigate risks and avoid legal penalties.

The department ensures that all necessary reports, filings, and audits are completed accurately and submitted on time to regulators, safeguarding the bank’s integrity and reputation.

*Example: Managing records for a personal loan, including the loan amount, interest rate, loan approval and disbursement, repayment schedule, and borrower’s credit history for valid and invalid check.*

3.8. KYC

The KYC department is responsible for verifying the identity of customers. This involves collecting, validating, and maintaining customer identification documents such as PAN cards, Aadhaar numbers, passports, and driver's licenses. KYC processes are critical for ensuring regulatory compliance and maintaining the integrity of customer relationships in financial institutions.

*Example: Capturing all required KYC information, including customer name, address, and valid identification documents (PAN, Aadhaar, etc.).*

1. **Data Quality Design**

This section provides an overview of how data quality will be managed and monitored. It should explain the approach and methodology used to ensure high-quality data across various processes and systems.

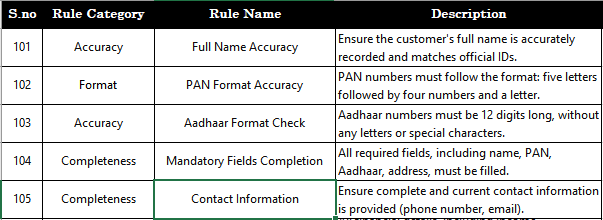
4.1. Data Quality Rule Identification

This section defines some specific rules and criteria that data must meet to be

considered high quality.

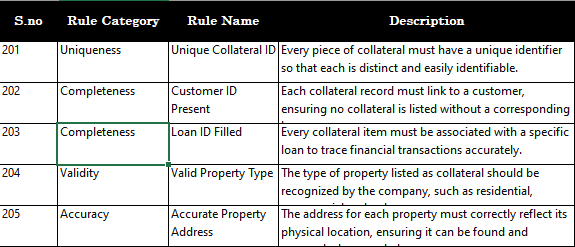
4.1.1. Customer:

A total of 24 data quality rules have been created to check the standard/quality of customer related data fields. Sample of the rules covered in Customer area are:



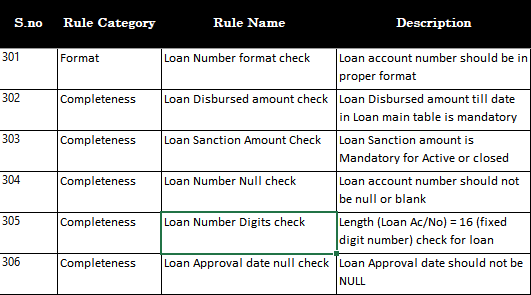
4.2.2. Collateral:

A total of 13 data quality rules have been created to check the standard/quality of collateral related data fields. Sample of the rules covered in Collateral area are:



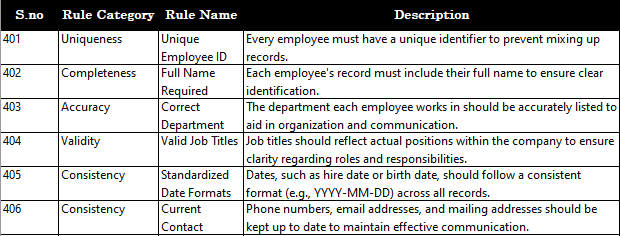
4.2.3. Loan

A total of 30 data quality rules have been created to check the standard/quality of loan related data fields. Sample of the rules covered in Loan area are:



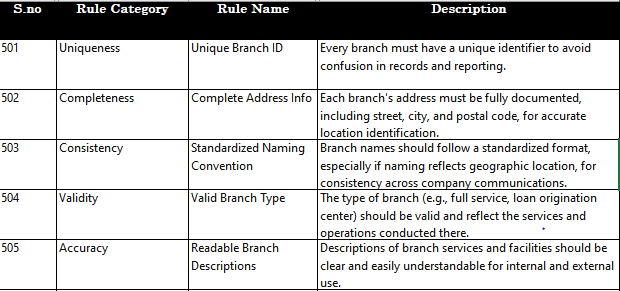
4.2.4. Employee

A total of 9 data quality rules have been created to check the standard/quality of employee related data fields. Sample of the rules covered in Employee area are:



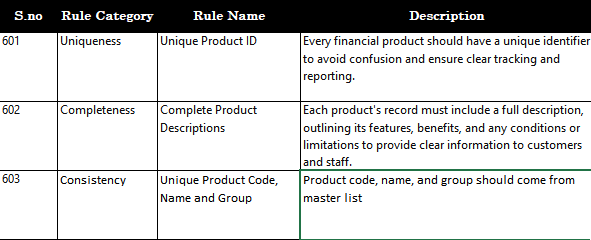
4.2.5. Branch

A total of 9 data quality rules have been created to check the standard/quality of branch related data fields. Sample of the rules covered in Branch area are:



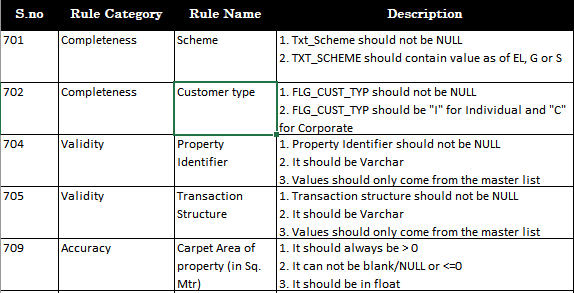
4.2.6. Product

A total of 3 data quality rules have been created to check the standard/quality of product related data fields. Sample of the rules covered in Product area are:



4.2.7. Regulatory

A total of 21 data quality rules have been created to check the standard/quality of regulatory related data fields. Sample of the rules covered in Regulatory area are:



4.2.8. KYC

A Waterfall (Funnel) structure will be created for KYC related DQ checks. The Waterfall structure will cover a set of given rules in an order for the specific KYC document (PAN/Aadhar/DL).

Example: PAN Waterfall structure –  
 - Total PAN  
 - PAN Present number  
 - PAN Format check  
 - PAN Duplicity check  
 - PAN match between LOS and LMS  
 - PAN Consistency across Applicant/Co-Applicant/Borrower  
 - API Verified  
 - PAN Name Match  
 - PAN DOB Match

**4.3. Data Quality Dimensions**

* + 1. **Accuracy**

Accuracy is about how closely data values match the real-world entity or event they represent. Accurate data is correct and free from errors.

*Example:*

*1.* The address for each property must correctly reflect its physical location, ensuring it can be found and assessed when needed.

*2.* Reported income should be verified against official documents like tax returns.

* + 1. **Format**:

Format validation ensures that data adheres to a specified format or pattern, as defined by the system or business requirements.

*Example:*

1. *PAN numbers must follow the format: five letters followed by four numbers and a letter.*
   * 1. **Completeness**:

Completeness measures whether all required data is present. It ensures that no essential information is missing.

*Example:*

1. *All financial details, including income, employment, and bank account numbers, are provided.*
2. *The name on the record should be consistent across PAN and Aadhaar documents.*
   * 1. **Consistency**:

Consistency checks that data remains the same across different datasets or systems, ensuring no contradictions.

*Example:*

1. *All dates (DOB, loan application date) follow the same format (e.g., DD-MM-YYYY).*
2. *The name on the record should be consistent across PAN and Aadhaar documents.*
   * 1. **Uniqueness**:

Uniqueness ensures that each data record is distinct and not duplicated within a dataset.

*Example:*

1. *Each customer should have a unique identifier.*
2. *Each PAN number in the database should be unique to a customer.*
   * 1. **Validity**:

Validity checks that data conforms to defined rules or constraints. It ensures that data falls within acceptable values or formats.

*Example:*

1. *Email addresses should follow a valid format (e.g., name@domain.com).*
2. *Age should be within a reasonable and valid range for adults (e.g., 18-65).*
   * 1. **Referential Integrity**:

Integrity refers to the overall correctness, consistency, and reliability of data throughout its lifecycle.

*Example:*

1. *Every loan record should be correctly linked to a valid customer record.*
2. *Physical or digital verification of the customer's address should be completed.*
   * 1. **Range**:

Range ensures that data values fall within a specified set of acceptable limits or boundaries.

*Example:*

1. *For a field that records ages, a valid range might be between 18 and 65. An age value of 66 and above would fall outside this range and would be flagged as invalid.*
   * 1. **Data Standardization**

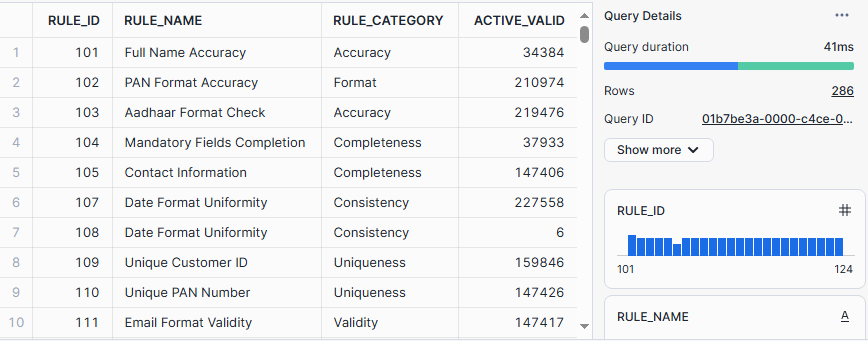
Defining standardized meanings and formats for each data element, ensuring consistent interpretation of data across systems.

*Example:*

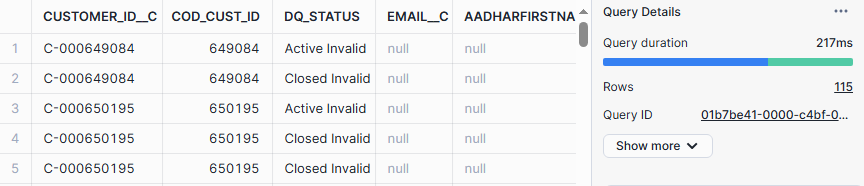
1. *Ensuring that every department refers to "Revenue" in the same way (e.g., as “Total Sales Amount”), so there is no confusion during analysis.*
   1. **Data Quality Results (Schema Name – Data\_Quality)**

**4.4.1. Customer**

4.4.1.2. Profiling Result (Snowflake Table – Customer)

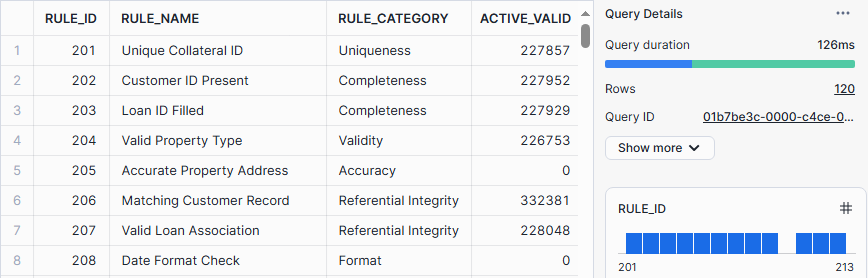


4.4.1.2. Detailed Result (Snowflake Table – Detailed\_Customer)

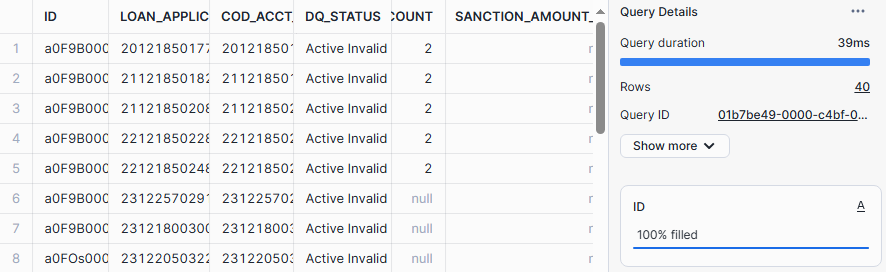


**4.4.2. Collateral**

4.4.2.1. Profiling Result (Snowflake Table – Collateral)

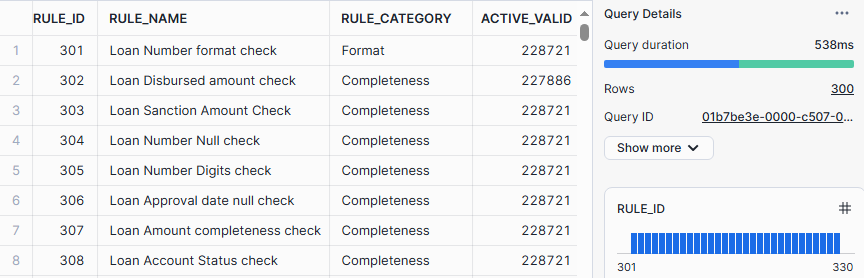


4.4.2.2. Detailed Result (Snowflake Table – Detailed\_Collateral)

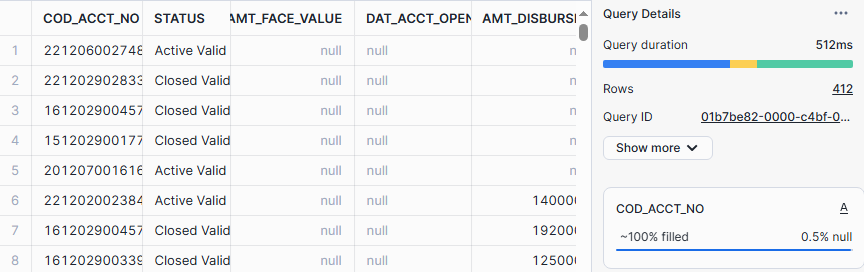


**4.4.3. Loan**

4.4.3.1. Profiling Result (Snowflake Table – Loan)

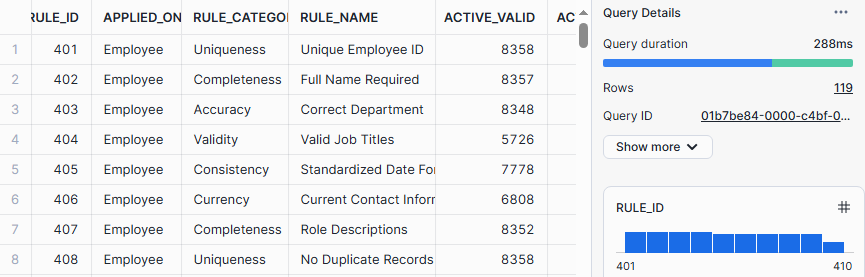


4.4.3.2. Detailed Result (Snowflake Table – Detailed\_Loan)

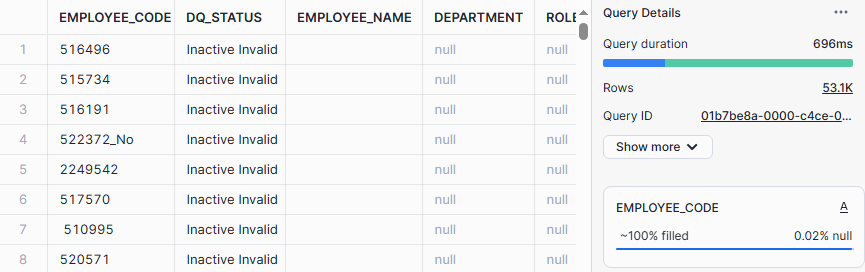


4.4.4. Employee

4.4.4.1. Profiling Result (Snowflake Table – Employee)

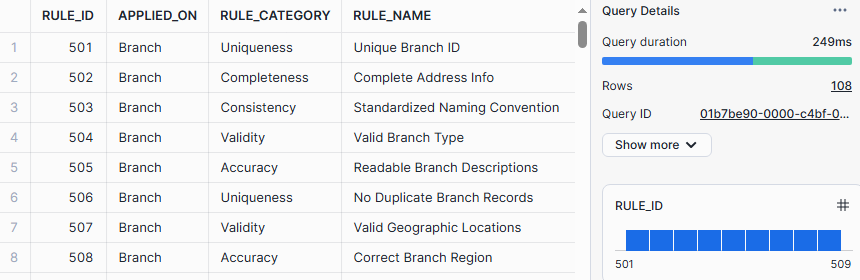


4.4.4.2. Detailed Result (Snowflake Table – Detailed\_Employee)

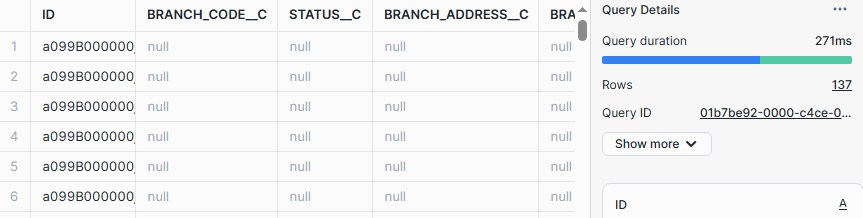


4.4.5. Branch

4.4.5.1. Profiling Result (Snowflake Table – Branch)

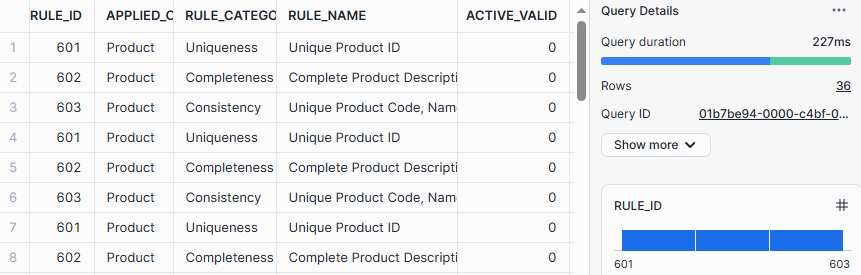


4.4.6.2. Detailed Result (Snowflake Table – Detailed\_Branch)

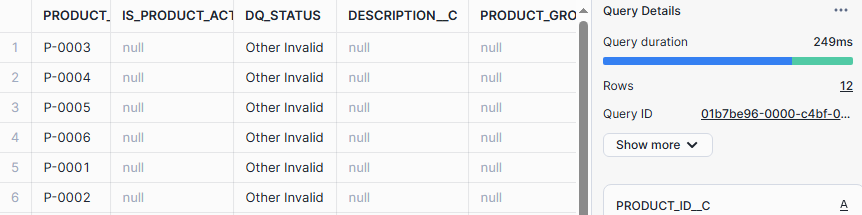


4.4.6. Product

4.4.6.1. Profiling Result (Snowflake Table – Product)

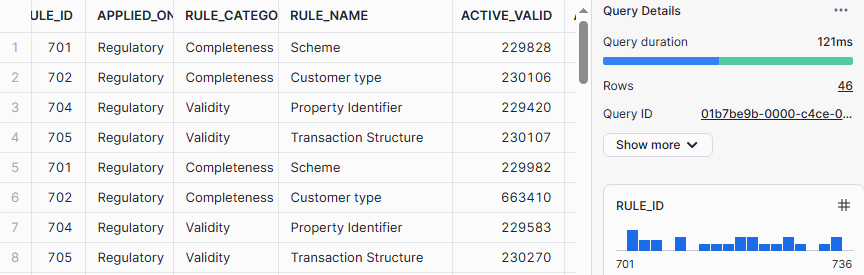


4.4.6.2. Detailed Result (Snowflake Table – Detailed\_Product)

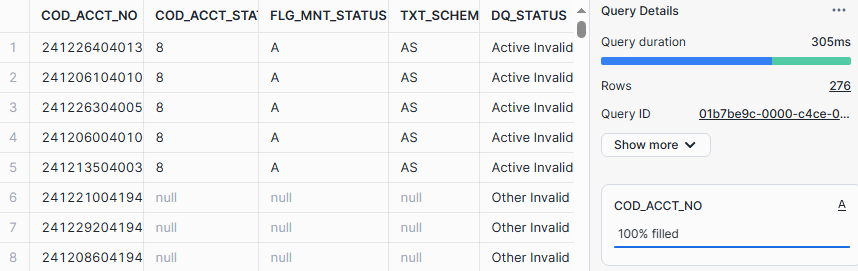


4.4.7. Regulatory

4.4.7.1. Profiling Result (Snowflake Table – Regulatory)

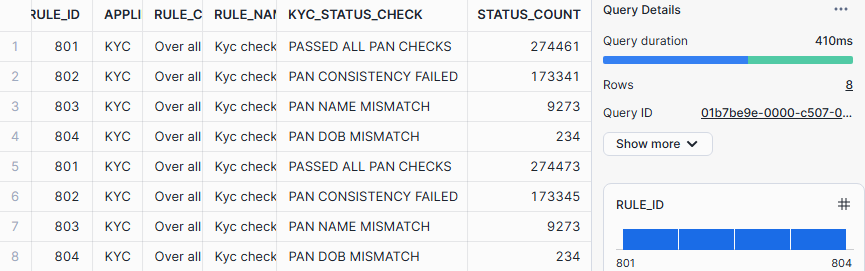


4.4.7.2. Detailed Result (Snowflake Table – Detailed\_Regulatory)



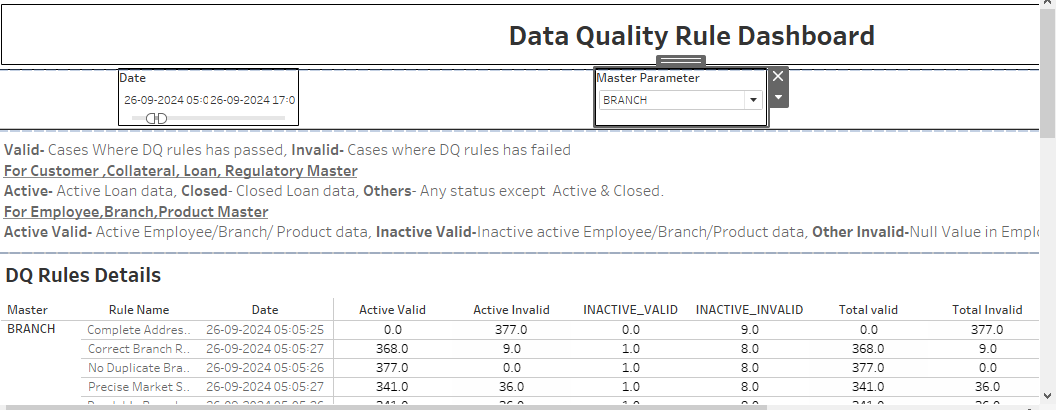
4.4.8. KYC

4.4.8.1. Profiling Result (Snowflake Table – KYC)

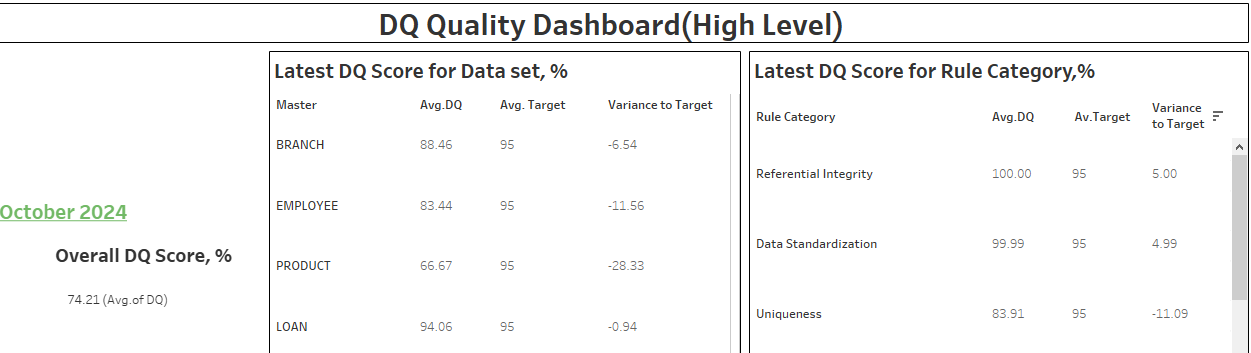


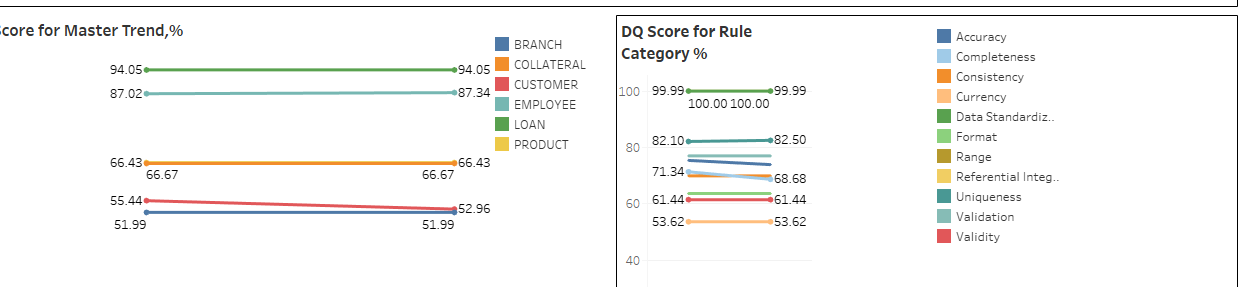
* 1. **Dashboard Summary**

4.5.1. Profiling Dashboard



* + 1. Department and Rule Category wise analysis





* + 1. KYC

