Informatica

18 July, 2024

Data Quality and Observability Framework for Trusted Data Governance

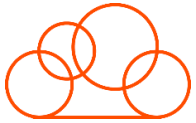
- Ronit Sen, Principal Consultant, IPS
- Faraz Nazeer, Consultant, IPS

Where data & AI come to LIFE

Housekeeping Tips

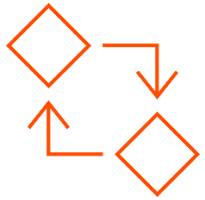
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

More Information

Success Portal

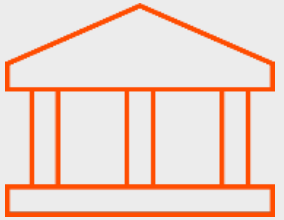
https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com



University

https://www.informatica.com/in/services-and-training/informatica-university.html

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

- Introduction
- Challenges & Solution Approach
- Feature Comparison - Manual development v/s Data Quality Reporting Framework
- Data Quality Implementation Process Flow
- Data Observability Framework – Design Principles & Key Functional Capabilities
- Data Observability Framework – High-Level Process Flow & Solution Architecture
- Data Observability Framework – Demo
- Data Observability Framework - Data Quality Reporting Dimensions


Informatica
Challenges & Solution Approach

Challenges with typical DQ Implementation

- Time consuming & cumbersome process to onboard many objects for DQ measurement
- Manual mapping development for DQ measurement for every single source object
- Higher IPU consumption costs on maintenance of large inventory of objects
- Absence of Best Practices – Reusability, Automation, Parametrization, Audit and Alerting
- Custom coding leads to lengthy testing cycles
- Deployment and maintenance overhead
- Operational and BI reporting integration challenges with existing exception handling process

High Cost	Time Consuming	Maintenance Overhead	Operational & Testing Overhead	Monitoring Challenges

Solution Approach for Enterprise Data Quality Implementation

Discover

1. Data Profiling
2. Identify Data Quality Issues
3. Derive Data Quality Goals

- Initial starting point to gather understanding of the data in terms of Unique values, Null values, patterns, primary keys, data domains
- Determine Data Validation, Cleansing, Standardization, Data Dictionaries, Parsing, Matching, Enrichment requirements and plan reusable DQ Rules
- Generate Data Quality Master Matrix – Systems, Datasets, Attributes, DQ Rule, DQ Dimension, Processing requirements

Define & Apply

4. Dictionaries and Rules
5. **Dynamic Data Quality Reporting Framework**
6. Data Governance Integration

- Create standard data dictionaries and reusable DQ rules/mapplets as derived in discovery phase.
- Deploy Data Quality Reporting Framework is a metadata-driven accelerator which includes set of reusable components deployed to provided fully automated Data Quality rules application, measurement, exception management, audit & alerting, Business Context/Governance Integration, designed with principles of Reusability, Parameterization, Automation and Codeless architecture.
- Automated DQ scores integration with Business Context in CDGC

Monitor

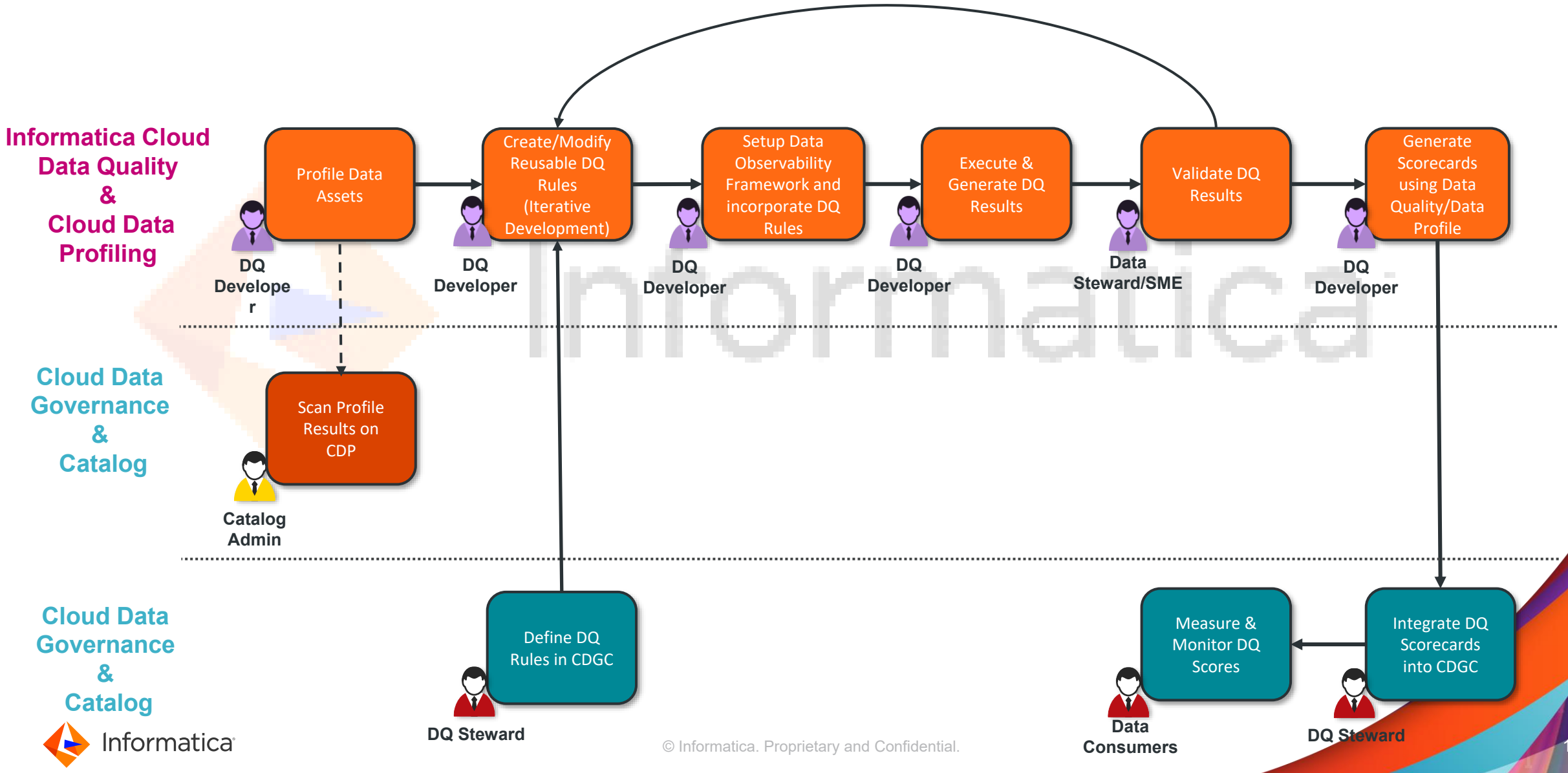
7. DQ Scorecards
8. External Reporting Integration
9. DQ Analytics & Exception Management

- DOF enables capture of data quality scores in a modelled exception database with business and technical context.
- DQ scorecards and dashboards generation in IDMC
- External reporting integration for DQ dashboards, DQ Audit and Operational dashboards.

Feature Comparison – Manual development v/s Data Quality Reporting Framework

Feature	Manual Development	Data Observability Framework
Automation of DQ Rules Application on Objects		
Ease of Scale – Parametrized framework		
DQ measurement on Incremental Data - Profiling		
DQ measurement on Incremental Data – DQ Mapping		
Exception Management Model		
Failure Alerts and Notifications		
Audit Batch Run Statistics		
Job Sequencing/Orchestration		
Restart and Recovery		
Code-less Architecture		

Data Quality Implementation Process Flow



Informatica

Data Observability Framework (DOF)

Key Design Principles

The proposed Informatica Data Quality Reporting Framework will be built based on the following design principles.

Metadata Driven	Reusability & Ease of Maintenance	Scalable (Extensible to Multiple Database Types)	Parallelism
Performance	Archival & Purging	Auditing	Failure Alerting

13

Key Functional Capabilities

1

Seeding

Seeding Module is an Excel based utility created to simplify data entry and maintain data integrity to framework control tables. The utility converts excel contents into CSV and the generated CSV files will act as an input to the Dynamic Data Quality Framework control tables.

2

Metadata Driven

Onboarding of any new data sources to the proposed framework will only require registry of attributes related to the new data sources to the framework metadata tables. No new code needs to be introduced to onboard new data sources.

3

Dynamic DQ Rule Assignment

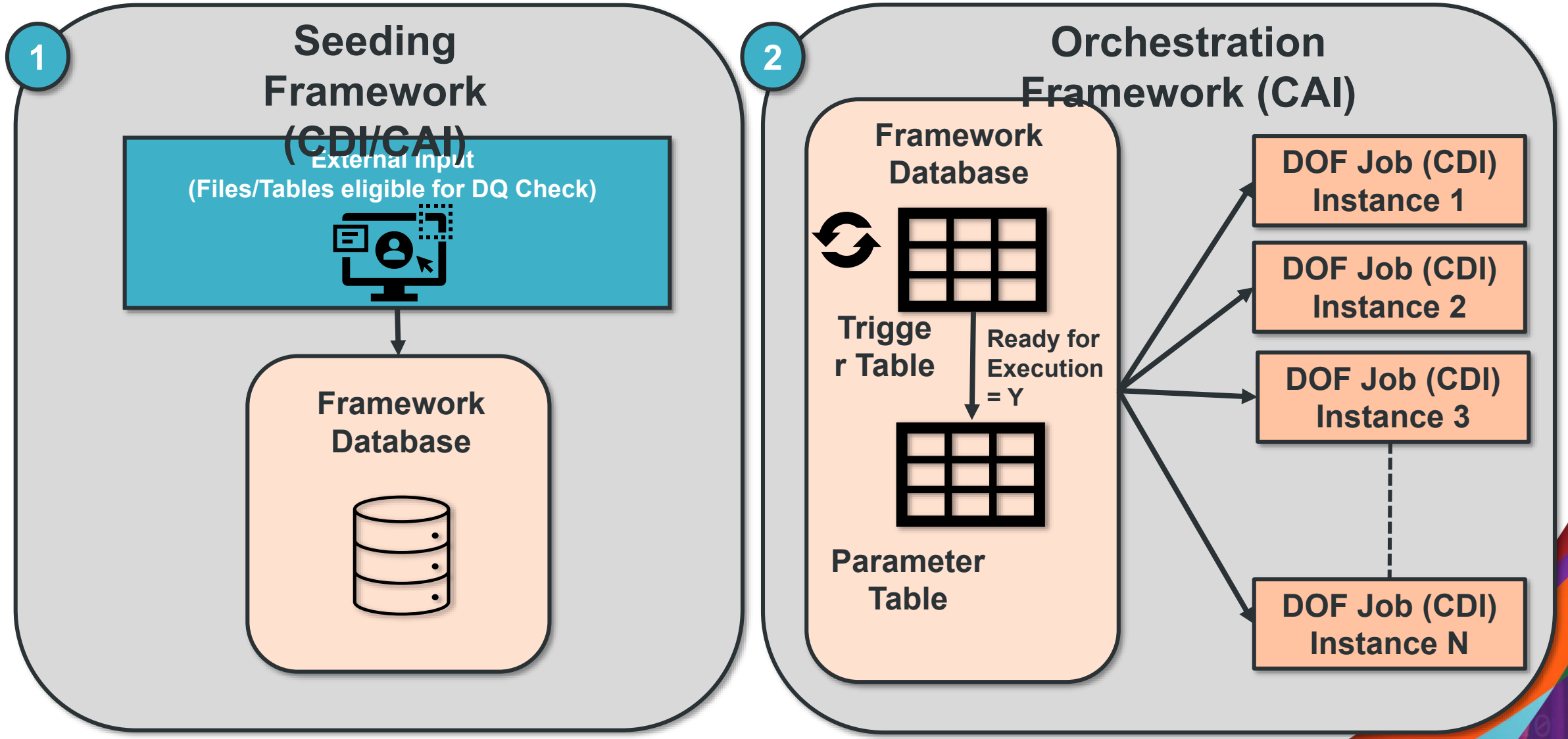
All the data quality rules will be built as Mapplets (Reusable DQ Rules) and integrated into a Mapping and deployed as an application. The Source Attributes will be assigned to DQ Rules dynamically using Parameters and each deployed application will have its own parameter file which will drive execution.

4

Business Metadata Enrichment

Data Quality mapping will perform REST API call to pull metadata information from CDGC which will be further used to add business context to data quality results.

High-Level Process Flow

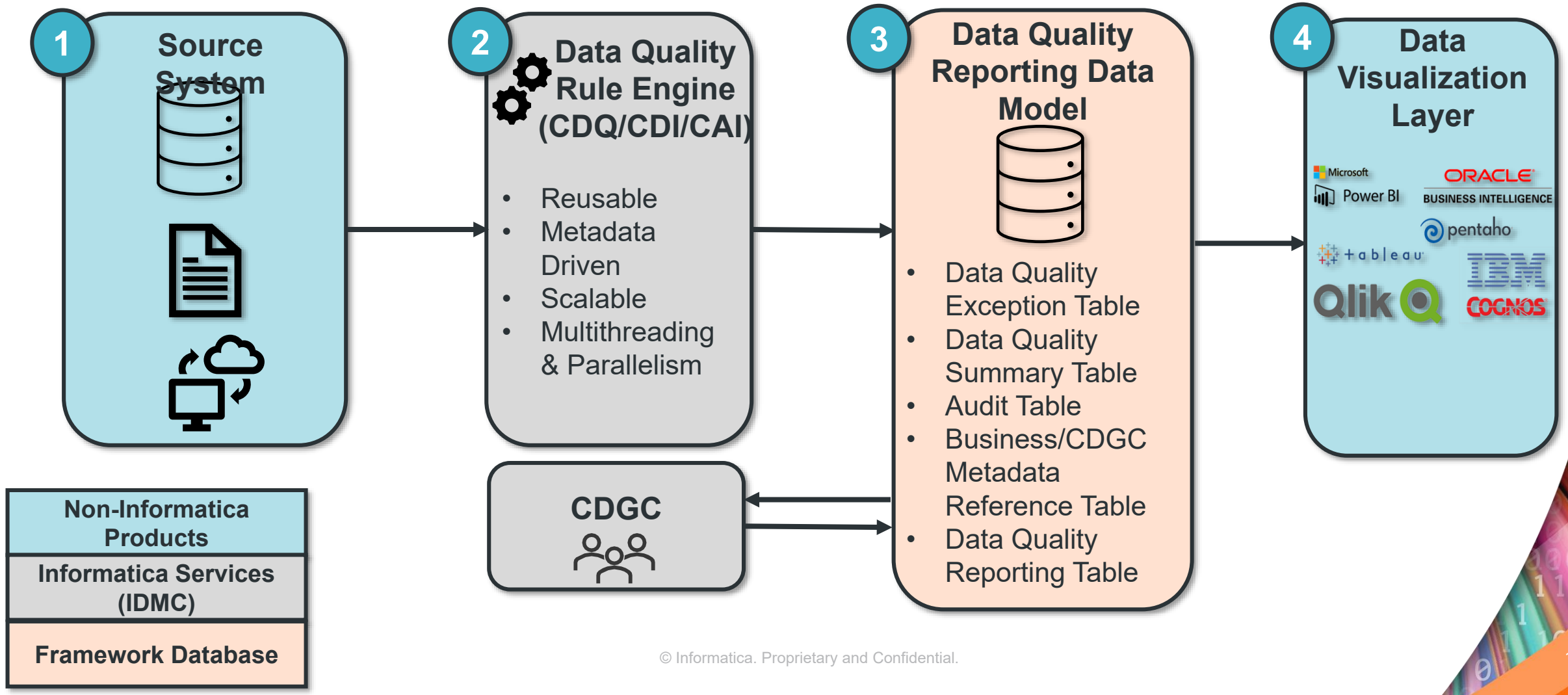


Orchestrated Flow as per Slide 10

1. The UI will be designed and developed with adherence to design principles - **Metadata Driven and Code Less, Reusability and Ease of Maintenance, Scalable.** This utility will be created to simplify data entry and maintain data integrity to framework control tables.

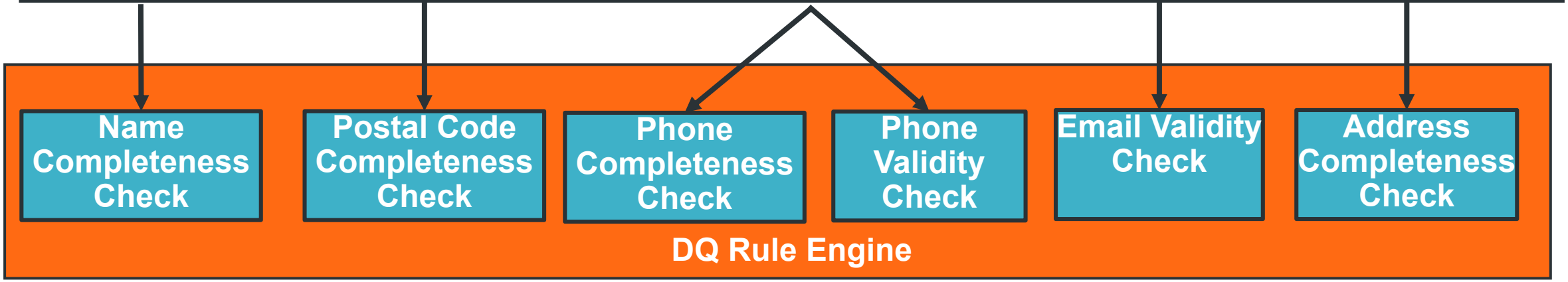
2. Poll based design will Trigger the Framework Jobs as and when corresponding Source Table/Files are ready for execution. Each Table/File will have its own parameter file which will get generated and will drive dynamic execution of the Data Quality Rules.

Data Observability Framework – Solution Architecture



Data Observability Framework – Process Flow

Name	Postal Code	Phone	Email	Address
Je,,rry	3362	(258 483-5383	libero.morbi@protonmail.edu	null



Exception Value
(258 483-5383
Null

CDE	DQ Dimension	Pass %	Fail %
Name	Completeness	70	30
Postal Code	Completeness	65	35
Phone	Completeness	70	30
Phone	Validity	66	34
Email	Validity	78	22
Address	Completeness	51	49

Orchestrated Flow as per Slide 12

1. Source systems data will be read from various parameterized data sources like RDBMS, Files, Cloud Data Lakes etc.,

2. Data Quality solution will be built using features and functionalities of Informatica Cloud Data Quality. All the data quality rules will be built as Mapplets (Reusable DQ Rules) and integrated into a Mapping and deployed as taskflow. Each instance of the published taskflow will have its own parameter file which will drive dynamic execution.

3. Below Target tables will be created in relational database like Azure SQL DB:
   I. Data Quality Exception Table
   II. Data Quality Summary Table
   III. Audit Table
   IV. DG Metadata Reference Table (Optional)
   V. Data Quality Reporting Table (Optional)

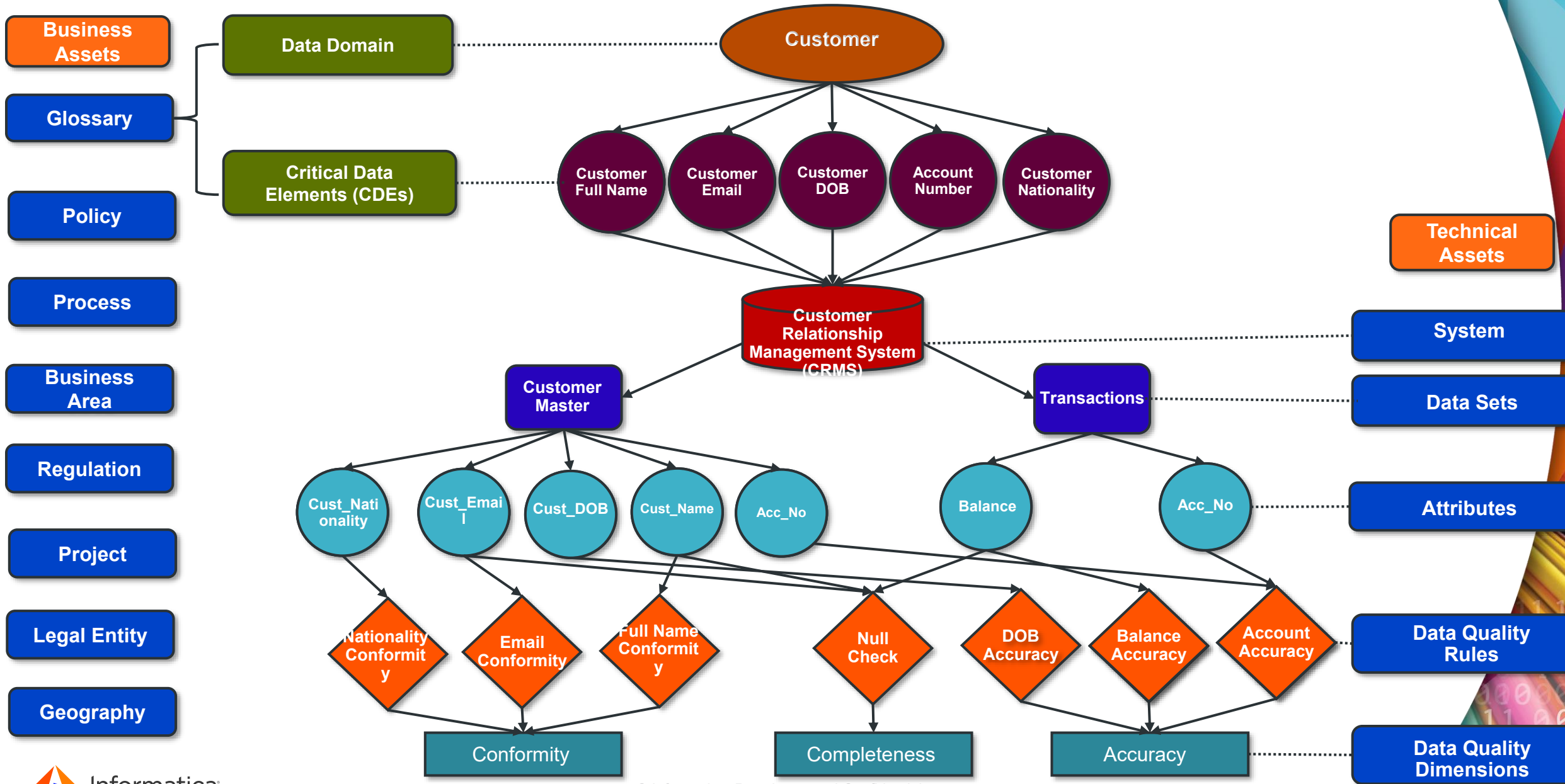
4. REST API calls will be made for the below functionalities:
   I. Pull metadata present in CDGC and insert into DG Metadata Reference table.
   II. Data Quality Scores from the Data Quality Reporting Table will be pushed to CDGC through REST API calls.

5. Data Quality Reporting table containing enriched (business information) data quality results at lowest granularity can be exposed through Data Visualization tools like Power BI to create custom reports with roll-up/roll-down, slice/dice features to report data quality scores.

Informatica

Demo

Data Observability Framework - Data Quality Reporting Dimensions



Thank You

Where data & AI come to

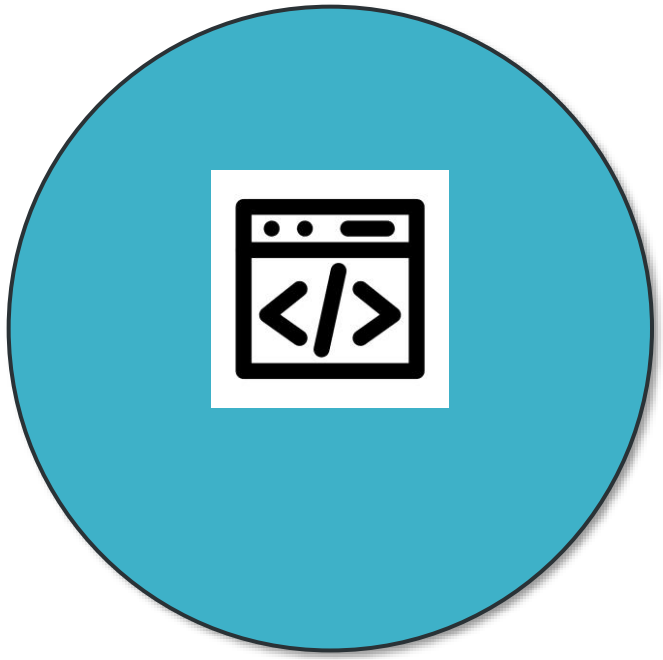
Informatica

Informatica

Reporting & Dashboards

Target Tables

Data Quality Exception Table



DG Metadata Reference Table

Data Quality Reporting Table

Data Quality Exception Table

S.No.	Exception Table Columns	Definition	Example
1	Source_Record_Date	It refers to Business Date of that record.	2021_10_04
2	System_Name	It refers to Source System Name where Data Quality rule has been measured.	CRMS
3	Data_Set_Name	It refers to Table Name where Data Quality rule has been measured.	Customer_Master
4	Attribute_Name	It refers to Attribute Name where Data Quality rule has been measured.	Birth Date
5	Attribute_Value	It refers to exception value present in the source table.	18000101
6	DQ_Rule_Name	It refers to Data Quality Rule Name.	Birth_Date_Completeness_Check
7	DQ_Dimension	It refers to Data Quality Dimension related to the Data Quality rule executed.	Completeness
8	Execution_Date	It refers to Date when Data Quality rule has been executed.	2021_10_05

Data Quality Reporting Table exposed to Data Visualization Tools

S.No.	Data Provider View Columns	Definition	Example
1	Source_Record_Date	It refers to Business Date of that record.	2021_10_04
2	Glossary_Name	It refers to Glossary value related to that CDE.	Birth Date
3	System_Name	It refers to Source System Name where Data Quality rule has been measured.	CRMS
4	Data_Set_Name	It refers to Table Name where Data Quality rule has been measured.	Customer_Master
5	Attribute_Name	It refers to CDE Name where Data Quality rule has been measured.	Birth Date
6	DQ_Rule_Name	It refers to Data Quality Rule Name.	Birth_Date_Completeness_Check
7	DQ_Dimension	It refers to Data Quality Dimension related to the Data Quality rule executed.	Completeness
8	Threshold_Amber_Target	It will contain the threshold value for Amber target.	85
9	Threshold_Green_Target	It will contain the threshold value for Green target.	95
10	Fail_Count	It refers to number of records which failed against the Data Quality rule check.	1000
11	Pass_Count	It refers to number of records which passed against the Data Quality rule check.	9000

DG Metadata Reference Table

S.No.	Axon Reference Table Columns	Definition	Example
1	Glossary_Name	It will contain the Glossary Name related to Data Quality rule.	Birth Date
2	System_Name	It will contain the System Name related to Data Quality rule.	CRMS
3	Data_Set_Name	It will contain the Data Set Name related to Data Quality rule.	Customer Master
4	Attribute_Name	It will contain the Attribute Name related to Data Quality rule.	DOB
5	DQ_Rule_Name	It will contain the Data Quality Rule Name.	Birth_Date_Completeness_Check
6	Threshold_Amber_Target	It will contain the threshold value for Amber target.	85
7	Threshold_Green_Target	It will contain the threshold value for Green target.	95
8	Policy_Name	It will contain the Policy Name related to Data Quality rule.	Privacy of Personal Data Policy
9	Process_Name	It will contain the Process Name related to Data Quality rule.	KYC Check Process
10	Business_Area_Name	It will contain the Business Area Name related to Data Quality rule.	Retail Banking
11	Regulation_Name	It will contain the Regulation Name related to Data Quality rule.	Personal Data Protection Act

Data Quality Reporting Table exposed to Data Visualization Tools

S.No.	Data Provider View Columns	Definition	Example
12	Total_Count	It refers to total count of records against which Data Quality rule was executed.	10000
13	DQ_Percentage	It refers to percentage of records that have passed the Data Quality rule check.	90
14	Execution_Date	It refers to Date when Data Quality rule has been executed.	2021_10_05
15	DQ_System	It refers to System where Data Quality rule engine is present.	Informatica_Data_Quality
16	Country	It refers to the Country associated to the System where Data Quality has been measured. Its value will come from Source System Data.	Singapore
17	Control_ID	It refers to the Control ID associated with the Data Quality Rule. It can be empty where not applicable.	A123456B
18	Policy_Name	It refers to the Policy associated to the System where Data Quality has been measured.	Privacy of Personal Data Policy
19	Process_Name	It refers to the Process associated to the System where Data Quality has been measured.	Know Your Customer
20	Regulation_Name	It refers to the Regulation associated to the System where Data Quality has been measured.	Personal Data Protection Act
21	Department_Name	It refers to the Department associated to the System where Data Quality has been measured.	Retail Banking

Data Quality Dashboard using Power BI & Tableau