



Informatica Cloud Data Quality

Rob Marx - Field Engineering

Where data
& AI come to **LIFE**

Data Quality – Current State



Only 27% of data practitioners completely trust their data¹



Trust in data degrades as it moves further away from its origin²



Up to 30% of time spent on non-value-added tasks because of poor data quality and availability³



Existing hand coded approaches become error prone and do not scale

(1) Source: Data Culture Survey, IDC December 2020, N=455, Data Trust Survey, IDC December 2021 N=500

(2) Source: "In Data We Trust. Or Do We?", Stewart Bond, IDC Directions Conference, March 2022

(3) Source: Designing data governance that delivers value, McKinsey 2020

Put Simply

Garbage in, Garbage out



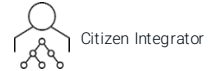
DATA CONSUMERS



ETL Developer



Data Engineer



Citizen Integrator



Data Scientist



Data Analyst



Business Users

Intelligent Data Management Cloud™

DISCOVER &
UNDERSTAND



DATA
CATALOG

ACCESS &
INTEGRATE



DATA INTEGRATION &
ENGINEERING

CONNECT &
AUTOMATE



API & APP
INTEGRATION

CLEANSE &
TRUST



DATA QUALITY &
OBSERVABILITY

MASTER &
RELATE



MDM & 360
APPLICATIONS

GOVERN &
PROTECT



GOVERNANCE &
PRIVACY

SHARE &
DEMOCRATIZE



DATA
MARKETPLACE

CLAIRE®

AI-Powered Metadata Intelligence & Automation

Connectivity

Metadata System of Record

DATA SOURCES



SaaS Apps
Sources

+



Mainframe



Applications



Databases

On-premises
Sources

+



IoT



Machine Data



Logs

Real-time /
Streaming Sources

Consistent Data Quality Process and Methodology



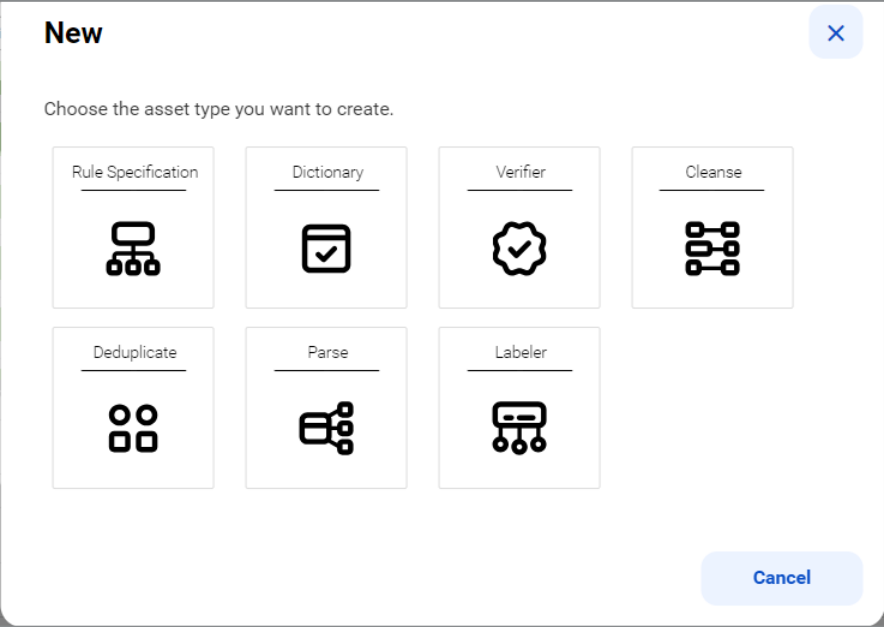
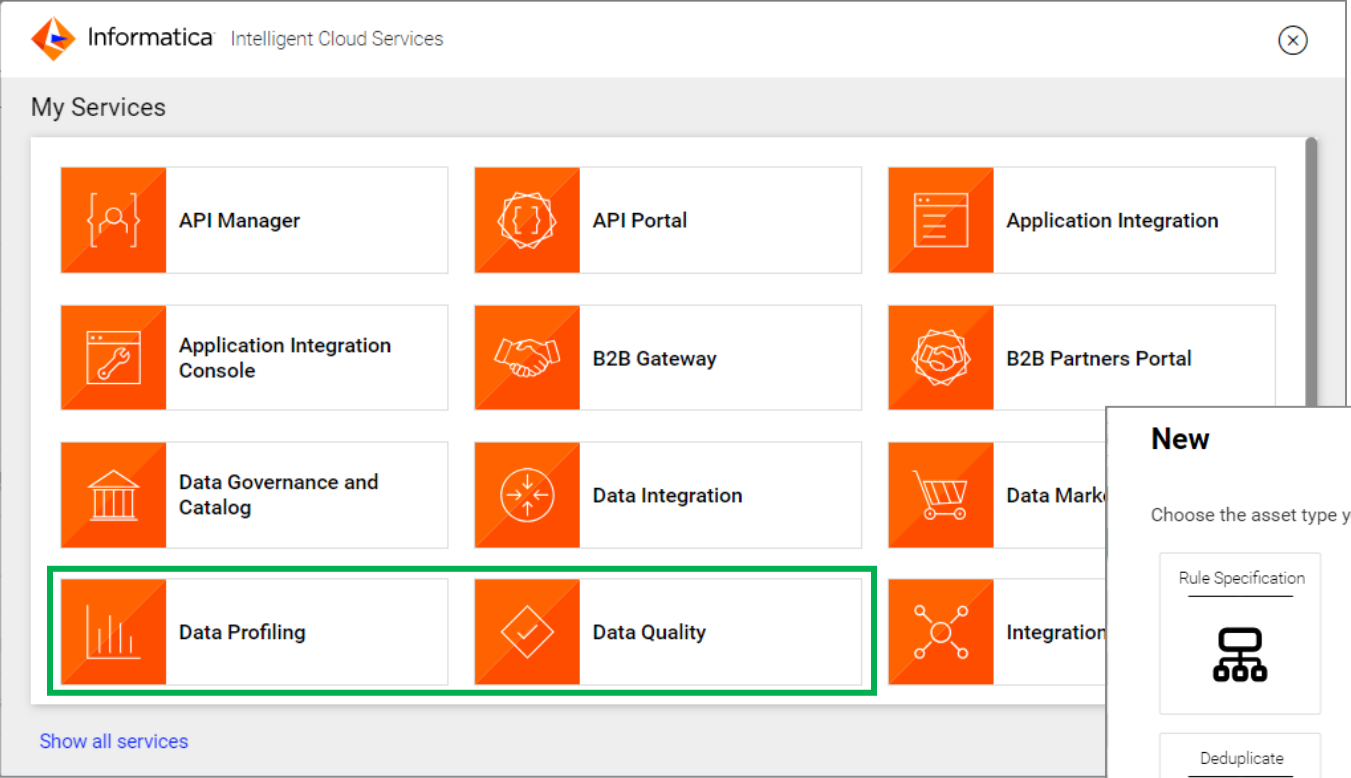
Consistent Process and Methodology



How do we get there?

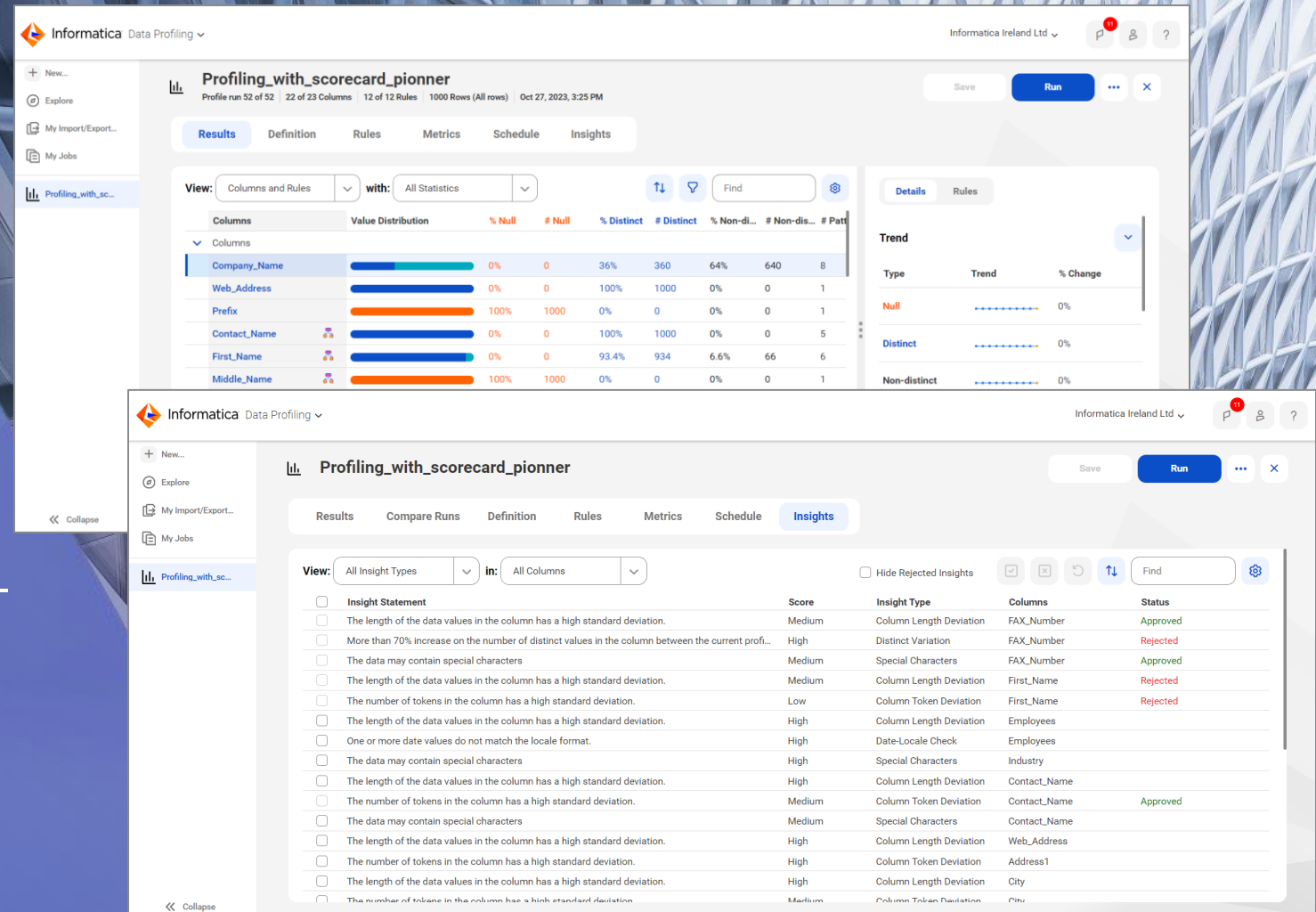


Cloud Data Quality



Automate and Scale Data Discovery and Identification

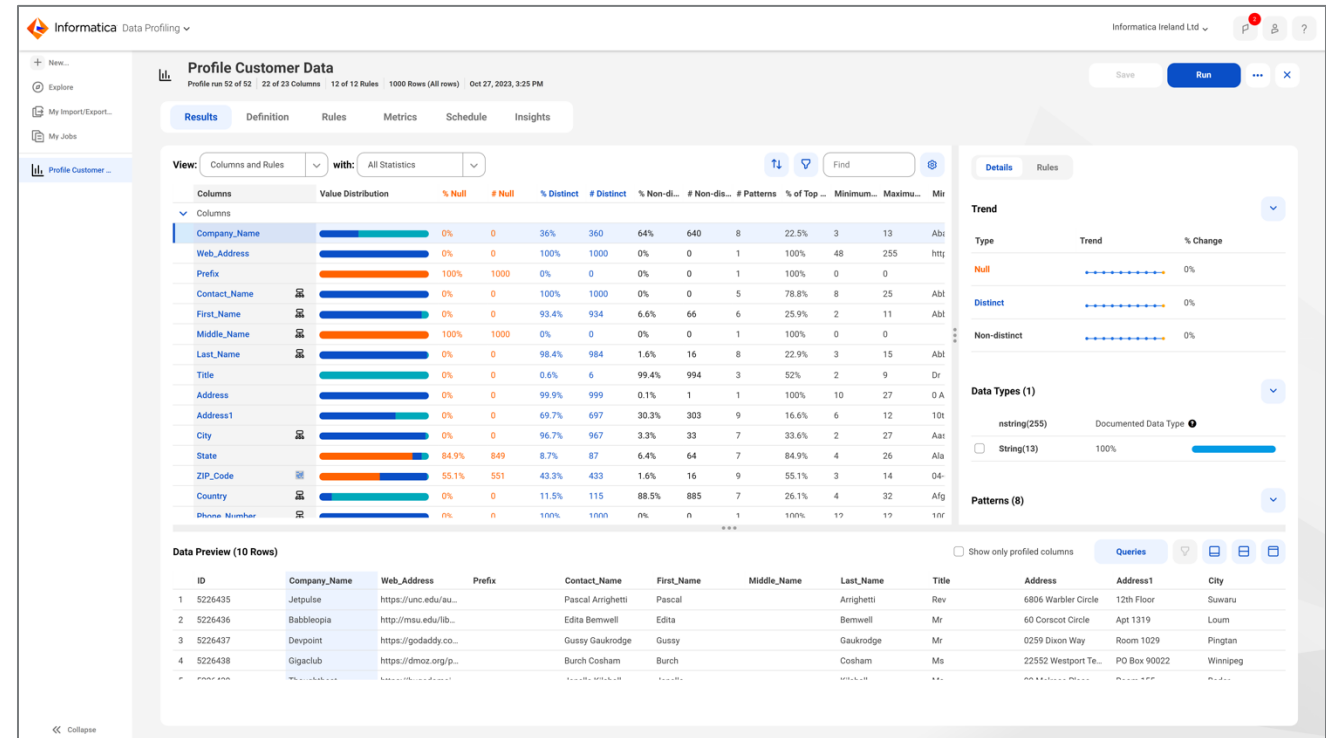
- Profile data to examine its structure and context using out-of-the-box templates
- Drill down to see details and filter on results
- Compare profile runs to identify trends over time
- CLAIRE Insights on profiling results



Cloud Data Profiling

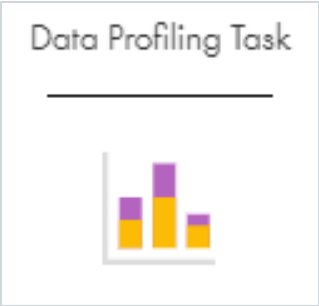
Get to Know Your Data

- Identify potential data issues
- Track aggregated changes on data over time
 - Provide statistics on data
 - Identify uniqueness or repeating values (value frequencies)
 - Identify patterns and formats
- Can profile Data Quality Assets
 - What-if analyses to determine rule fit



Cloud Data Profiling Statistics

Column profiling metrics



Null (%/#)	Min Value	Patterns
Distinct (%/#)	Max Value	Documented Data Type
Non-Distinct (%/#)	Min Length	Inferred Data Type
Blanks (%/#)	Max Length	Value Frequency
Average, Sum, Std Deviation, Zeros (%/#) (for numeric data types)	% Rows on Top Pattern	Pattern and Value Frequency Outlier

Augmented Data Quality – CLAIRE Insights for Profiling Results



Use case

- Automate the discovery of potential data problems by analyzing profiling results
- Results presented as **Insights**
- User acceptance of Insights automatically creates Data Quality Rules and assigns to profiling for monitoring
- **Scalable framework** allows expansion to advanced AI/ML algorithms of outlier detection

Informatica Data Profiling

Informatica Ireland Ltd

Profiling_with_scorecard_pionner

Results Compare Runs Definition Rules Metrics Schedule Insights

View: All Insight Types in: All Columns

Hide Rejected Insights

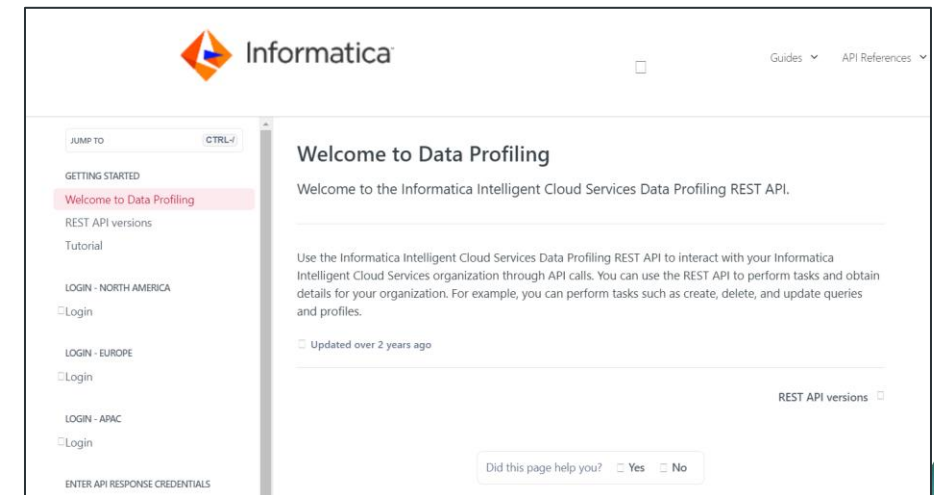
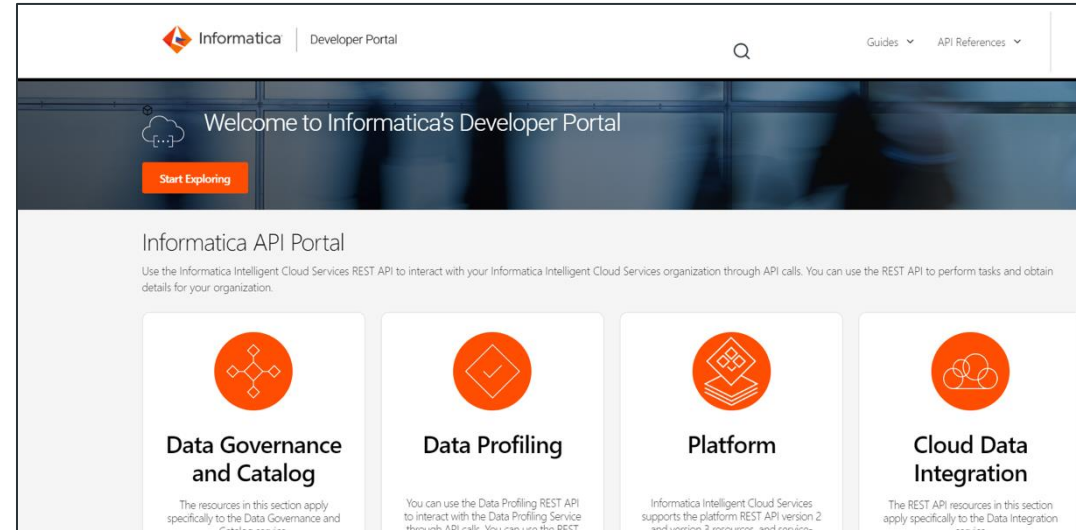
Insight Statement	Score	Insight Type	Columns	Status
<input type="checkbox"/> The length of the data values in the column has a high standard deviation.	Medium	Column Length Deviation	FAX_Number	Approved
<input type="checkbox"/> More than 70% increase on the number of distinct values in the column between the current profi...	High	Distinct Variation	FAX_Number	Rejected
<input type="checkbox"/> The data may contain special characters	Medium	Special Characters	FAX_Number	Approved
<input type="checkbox"/> The length of the data values in the column has a high standard deviation.	Medium	Column Length Deviation	First_Name	Rejected
<input type="checkbox"/> The number of tokens in the column has a high standard deviation.	Low	Column Token Deviation	First_Name	Rejected
<input type="checkbox"/> The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	Employees	
<input type="checkbox"/> One or more date values do not match the locale format.	High	Date-Locale Check	Employees	
<input type="checkbox"/> The data may contain special characters	High	Special Characters	Industry	
<input type="checkbox"/> The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	Contact_Name	
<input type="checkbox"/> The number of tokens in the column has a high standard deviation.	Medium	Column Token Deviation	Contact_Name	Approved
<input type="checkbox"/> The data may contain special characters	Medium	Special Characters	Contact_Name	
<input type="checkbox"/> The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	Web_Address	
<input type="checkbox"/> The number of tokens in the column has a high standard deviation.	High	Column Token Deviation	Address1	
<input type="checkbox"/> The length of the data values in the column has a high standard deviation.	High	Column Length Deviation	City	
<input type="checkbox"/> The number of tokens in the column has a high standard deviation.	Medium	Column Token Deviation	City	

<< Collapse

Cloud Data Profiling

Public API for the complete lifecycle

- Automate and scale your ability to profile your data
- Set of REST APIs allow:
 - CRUD operation for profiling tasks
 - Scheduling and Execution of profiling tasks
 - Monitoring
 - Retrieval of profiling results
 - Creation and execution of profiling queries
- Documentation available at:
 - API Portal [here](#)
 - H2L [here](#)



<https://developer.informatica.com/>

Empower Self-service and Business Ownership

- Define metrics to measure the quality of data within your application's key data elements
- Apply DQ rules to business processes and applications
- Promote business and IT collaboration

The screenshot displays the Informatica Data Quality console interface. The top navigation bar includes the Informatica logo, 'Data Quality', and a user profile. The left sidebar contains navigation links: 'New...', 'Home', 'Explore', 'My Jobs', 'My Import/Export...', and 'SetCustomerStat...'. The main workspace is titled 'SetCustomerStatus' and features tabs for 'Definition' and 'Configuration'. The 'Design' view shows a hierarchical diagram with a parent node 'SetCustomerStatus' containing two child nodes: 'CheckLastOrderDate' (with input 'LastOrderDate') and 'CheckCurrentStatus' (with input 'CustomerStatus'). Below the diagram, the 'Properties: SetCustomerStatus' panel is visible, showing the 'Rule logic definition' table.

Rule logic definition						
		Input	Operator	Condition	Action	
Inputs	if	CheckLastOrderDate	is	'nodate'	then 'Inactive'	
	or	if	CheckLastOrderDate	is	'ok'	then 'Active'
Test	or	if	No rule statement is valid		then	CheckCurrentStatus

Informatica Cloud Data Quality

Profile



Turn by turn
directions for data
management instead
of “heading west”

Validate

- ✓ Currency code must be consistent with country code
- ✓ Employee ID must be unique
- ✓ If customer tier is bronze, then max credit is 1000
- ✓ All ICD10 codes must have a verified description

Standardize

Example 1: KitKat
Kit-Cat
Kit Kat } **Kit-Kat**

Example 2: North Central
Midwest
Great Lakes } **Central**

Fuzzy Lookup/Match

Source Data

Land O Lakes

Biluxi

KitKat

System of Record

Land O'Lakes

Biloxi

Kit-Kat

Parse

Text Data: Call Center Comments, Social Media, Product Descriptions

I love my Pink 64GB iPad Pro!!!

Color	Size	Product	Model
Pink	64 GB	iPad	Pro

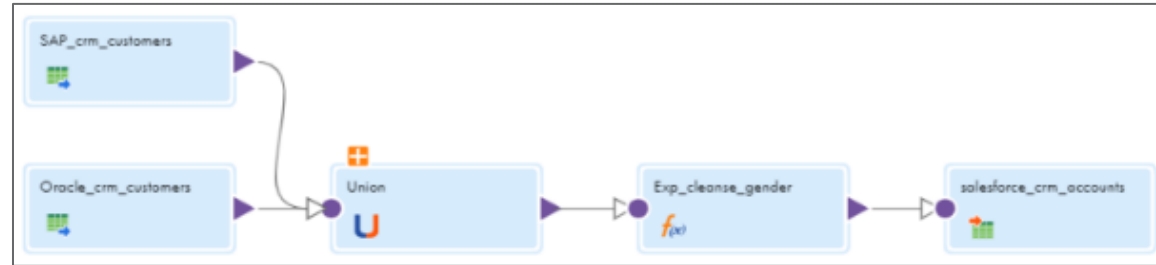
Enrich

AddressL1: 1008 Avenue of the Americas
AddressL2: Suite 7
City: New York
State: NY
Zip Code: 10018-5402
Longitude: 40.7325525
Latitude: -74.004970

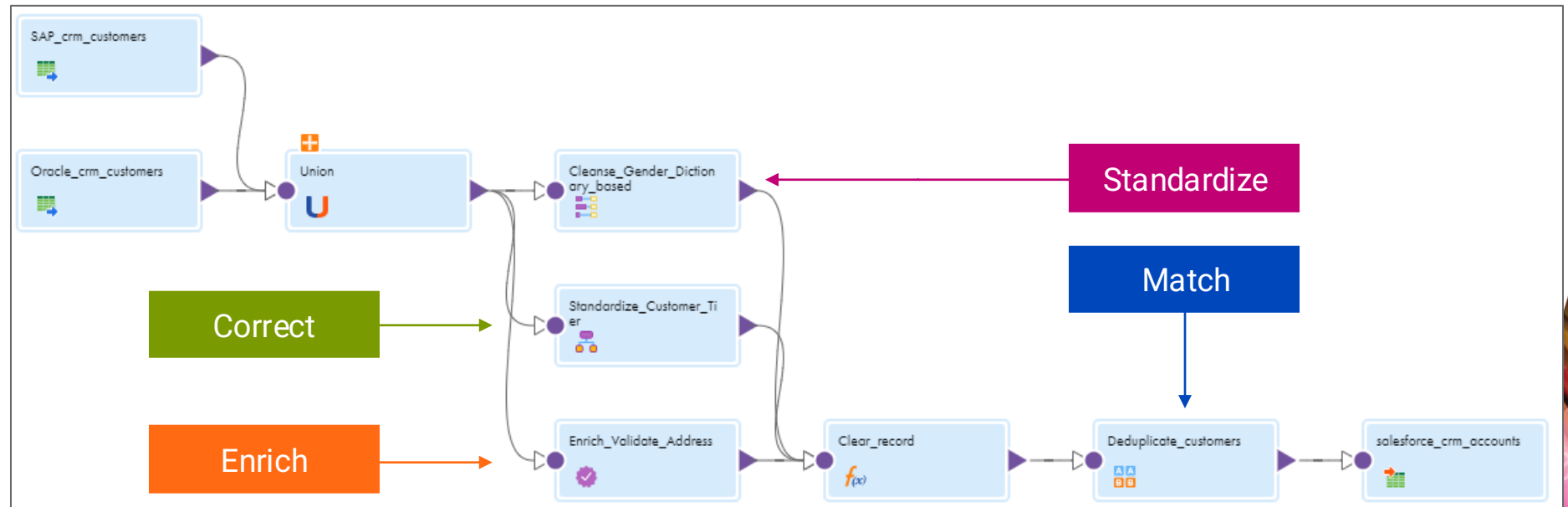
Drag and Drop No Code Data Quality into Your Pipeline

Quickly Identify & Resolve Data Inconsistencies With Integrated Cloud Data Quality

Data Pipeline:
Integration only



Data Pipeline:
Integration and
Data Quality




Hundreds of Out-of-the-Box Accelerators

Rules, Dictionaries, Verifications and Mapplets













Dictionaries

Name ▲	Type
 character_sets	Dictionary
 diacritics_translate_info	Dictionary
 domain_names_info	Dictionary
 dq_av_mailabilityscores_info	Dictionary
 dq_av_match_code_descriptions_i...	Dictionary
 iso_full_country_name_info	Dictionary
 numeric_default_values_info	Dictionary
 string_default_values_info	Dictionary
 suspect_names_info	Dictionary
 user_defined_dates_info	Dictionary



Cleanse

Name	Type
 c_SSN_Max_Index_For_Group	Cleanse
 c_Standardize_rplce_Back_Slas...	Cleanse
 c_standardize_rplce_bck_slash...	Cleanse
 c_Std_Name	Cleanse
 c_std_terms_mpllt_association	Cleanse
 c_Titlecase	Cleanse
 c_Uppercase	Cleanse
 c_US_Company_Name_Std	Cleanse
 c_US_FullName	Cleanse
 c_US_Get_Company_Acronym	Cleanse













Deduplicate

Name ▼	Type
 dedupe_USA_Address_Match	Deduplicate
 dedupe_US_IMO_Personal_Nam...	Deduplicate
 dedupe_US_IMO_Match_Compa...	Deduplicate
 dedupe_US_IMO_Indiv_Name_A...	Deduplicate
 dedupe_US_IMO_Fam_Name_A...	Deduplicate
 dedupe_CompanyName_Match	Deduplicate
 dedupe_BRA_Match_Name_Perso...	Deduplicate
 dedupe_BRA_Match_Indiv_Name...	Deduplicate
 dedupe_BRA_Match_Company_a...	Deduplicate
 dedupe_BRA_Individual_Name_a...	Deduplicate













Parse

Name	Type
 p_Parse_SSN	Parse
 p_Parse_US_FullName	Parse
 P_ParseFullName	Parse
 p_SN_metrics_to_ind_fields	Parse
 p_SSN	Parse
 p_Suspect_Names	Parse
 p_Text_bt看_Parentheses	Parse
 p_Text_bt看_Single_Quotes	Parse
 p_top_level_domain	Parse
 p_usa_gender_assignemn...	Parse



Validate

Name ▼	Type
 v_USA_Validation_Discrete_Addre...	Verifier
 v_USA_Validation_Discrete_Address	Verifier
 v_USA_AddressValidation_Multiline	Verifier
 v_USA_AddressValidation_Geo...	Verifier
 v_US_AddressValidation_Hybrid	Verifier
 v_US_AddressValidation_Hyb_w...	Verifier
 v_CAN_AddressVerification_Hybrid	Verifier
 v_BRA_Validation_Discrete_Addre...	Verifier
 v_BRA_Validation_Discrete_Address	Verifier
 v_BRA_Parse_Multiline_Address...	Verifier

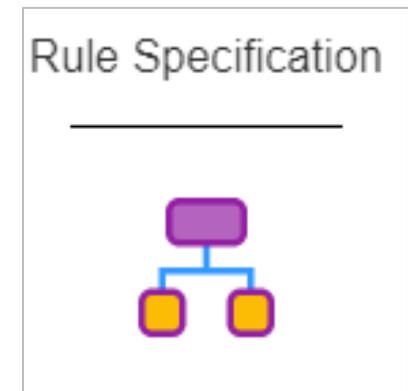
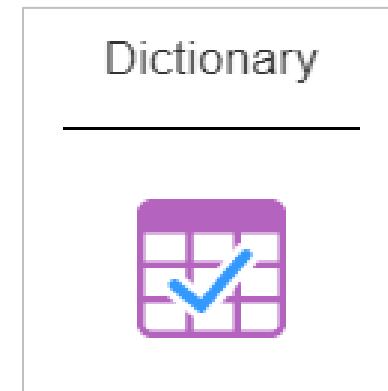
Rule Specification + Dictionary

- Enable business users to create Rule Specifications to perform a variety of Data Quality Actions including Completeness, Validation and Standardization
- Rule Specifications Create business logic to validate assumptions and statements – based on decision trees and IF/THEN/ELSE constructs
- Rules can consume dictionaries and allow for simple standardization
- Dictionaries

Create and manage value lists for data quality evaluation and operations using Dictionaries.

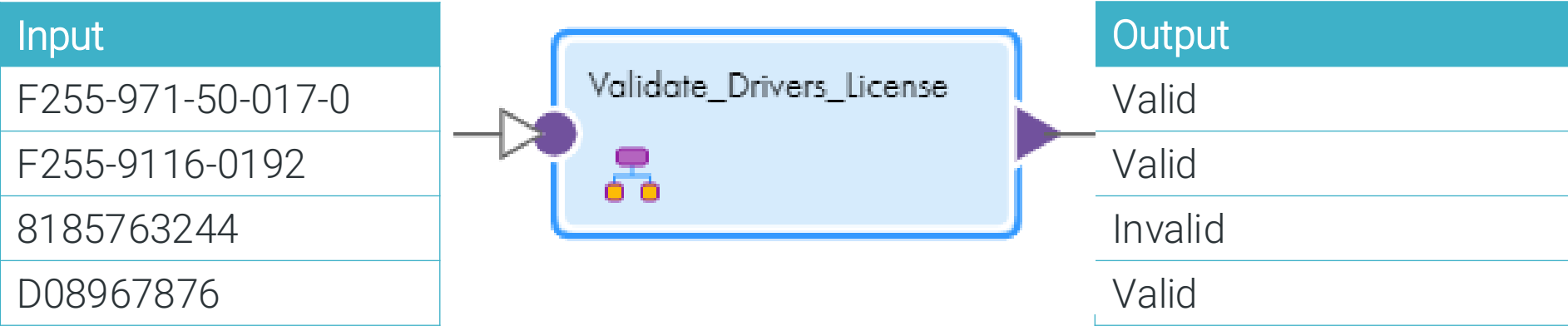
Dictionaries can be used to identify, validate and standardize data as part of Rule Specifications

Consumed by other assets (Rules, Cleanse, Parse)



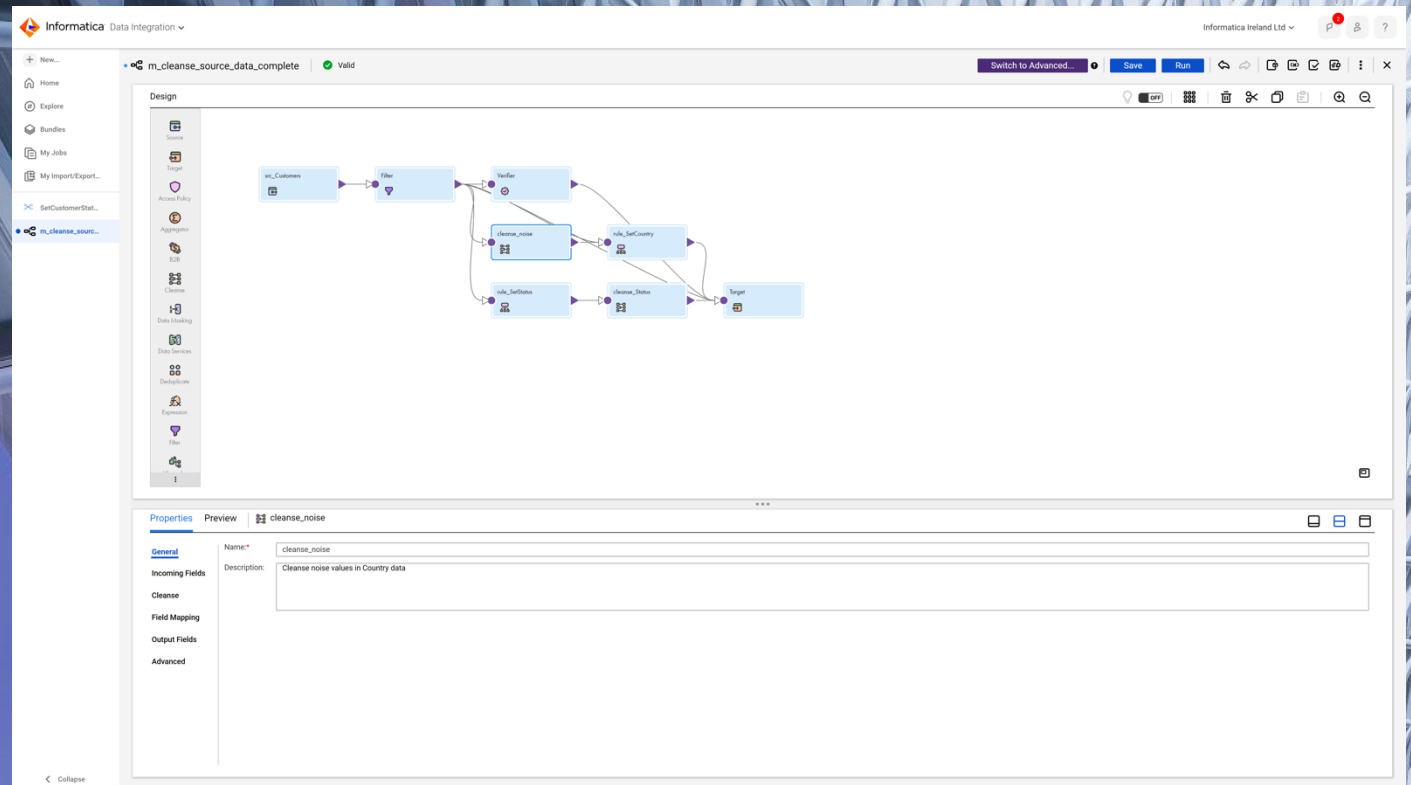
Rule Specification + Dictionary

Validation Rule to Validate Drivers License



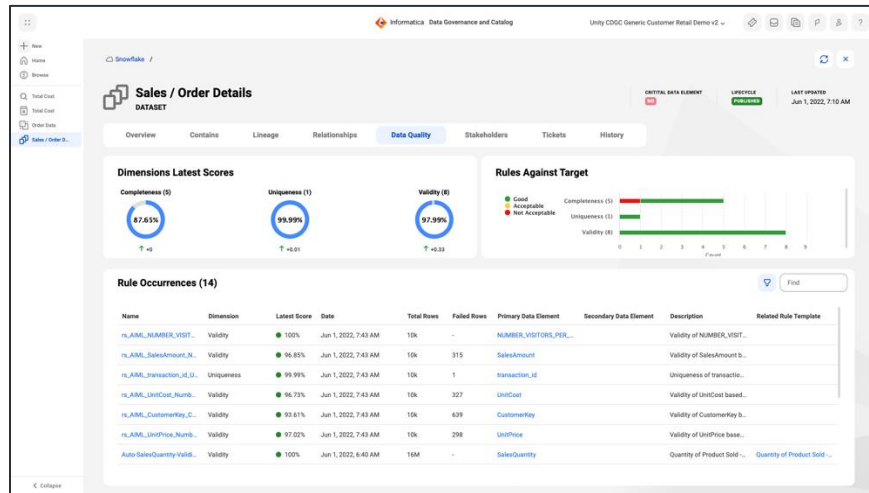
Build Once and Re-use Everywhere Across Cloud and On-premises

- Ensure consistency of data quality with centralized rules management and execution
- Automate application of data quality rules across sources in on-premises and multi-cloud hybrid environments
- Embed enforcement of quality rules into data pipelines and business processes



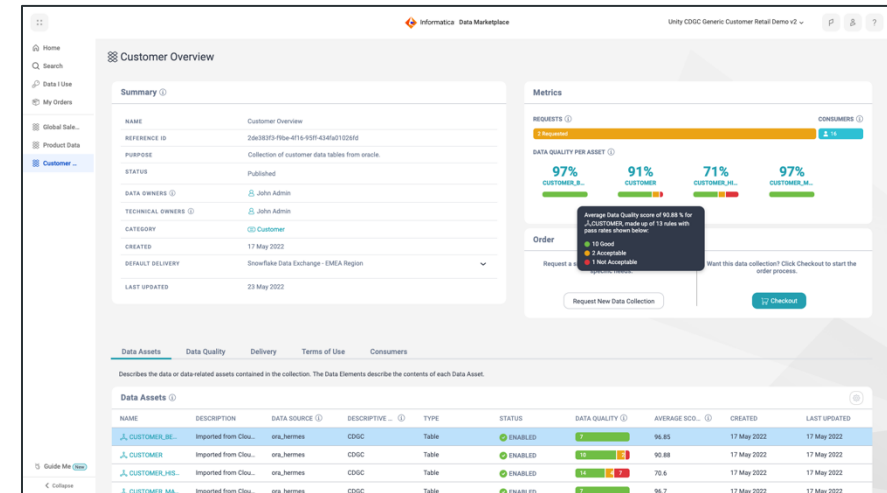
Innovation—DQ for Business Personas

Data Steward and Data Analyst



- Top Down and Bottom Up Data Quality Metrics Integrations aligned to Business Contexts
- Full data quality automation from Cloud Data Governance and Catalog
- Scorecard Metrics created by Data Quality users surfaced to Governance consoles automatically and aligned to business contexts

Data consumer



- Easily find and request data required for business objectives
- Get Data Quality insights and measurements direct on data assets to support trust considerations
- Interact with Business and Technical Data Owners

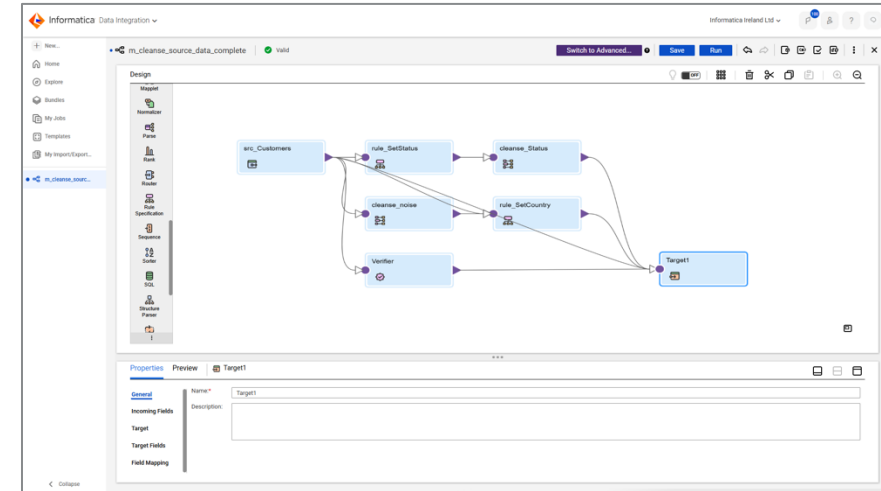
Innovation—DQ for Business Personas

Master Data Steward

The screenshot shows the Informatica Customer 360 interface. At the top, a yellow banner states: "The record has validation errors. Fix them, and then click Submit." Below this, the record details for "Johnathan Sullivan Cordeiro" are displayed. The "Address" section has a red error icon and message: "Some of the values in an entry are not valid. Enter valid values." The "Billing" section also has a red error icon and message: "The value in 1 field City is not valid. Enter a valid value." The form fields include: Address Line 1 (73 State Road 434 E), Address Line 2, Country (United States), State (Arizona), City (Phoenix), County (Maricopa), Postal Code (85013), Postal Code Extension, Location Coordinate Desc, Start Date (02/15/2004 02:33:00 PM), End Date (06/28/2020 06:57:00 AM), Latitude (33.511087), and Longitude (-112.084749). A tooltip for the City field shows "City Not Found".

- Re-use data quality rules for all MDM scenarios automatically
- Configure data quality at business entity level, enable for:
 - Batch Data Ingestion
 - Data Steward editing and maintenance
 - Incremental Data Ingestion and Updates

Citizen Integrator



- Easy use and implementation of re-usable Data Quality Rules for data assessment and correction
- Fully integrated capabilities with access to Profiling and Insights to support mapping development
- CLAIRE Next Transformation suggestions and intelligence

Lightweight Execution for Data Quality



Feature Highlights

- Portable data quality rules for embedding in any application or ecosystem
- Data quality services for application orchestration and point-of-entry scenarios
- Governed by IDMC API Gateway



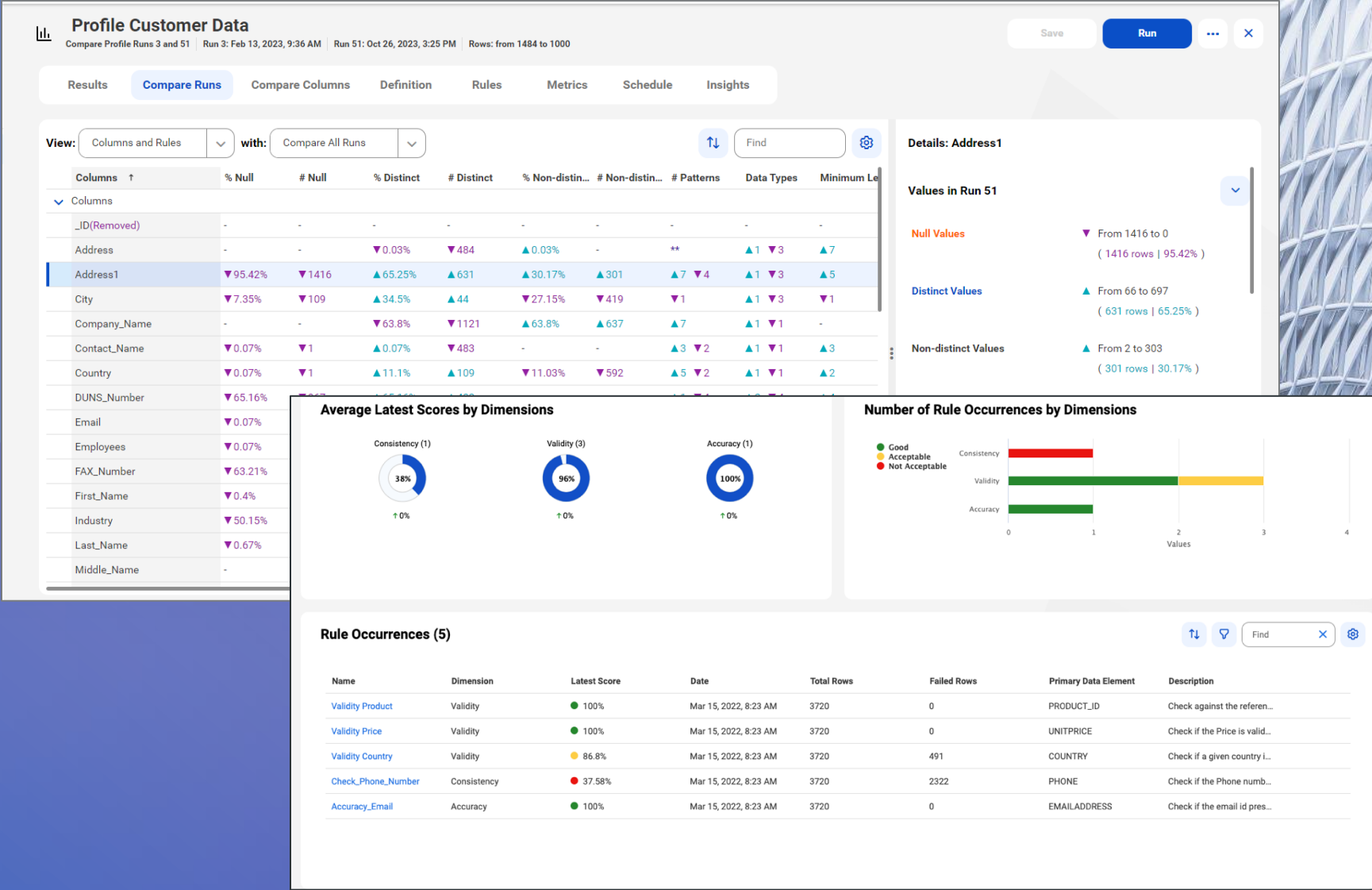
Benefits

- Facilitate scaling outside of Informatica run-times
- Low latency / low volume alternative to a DQ mapping



Provide Continuous Insight

- Track and remediate data quality changes over time
- Align data quality activities with data governance and privacy activities
- Compare profile runs to identify trends over time



Cloud Data Quality Scorecard

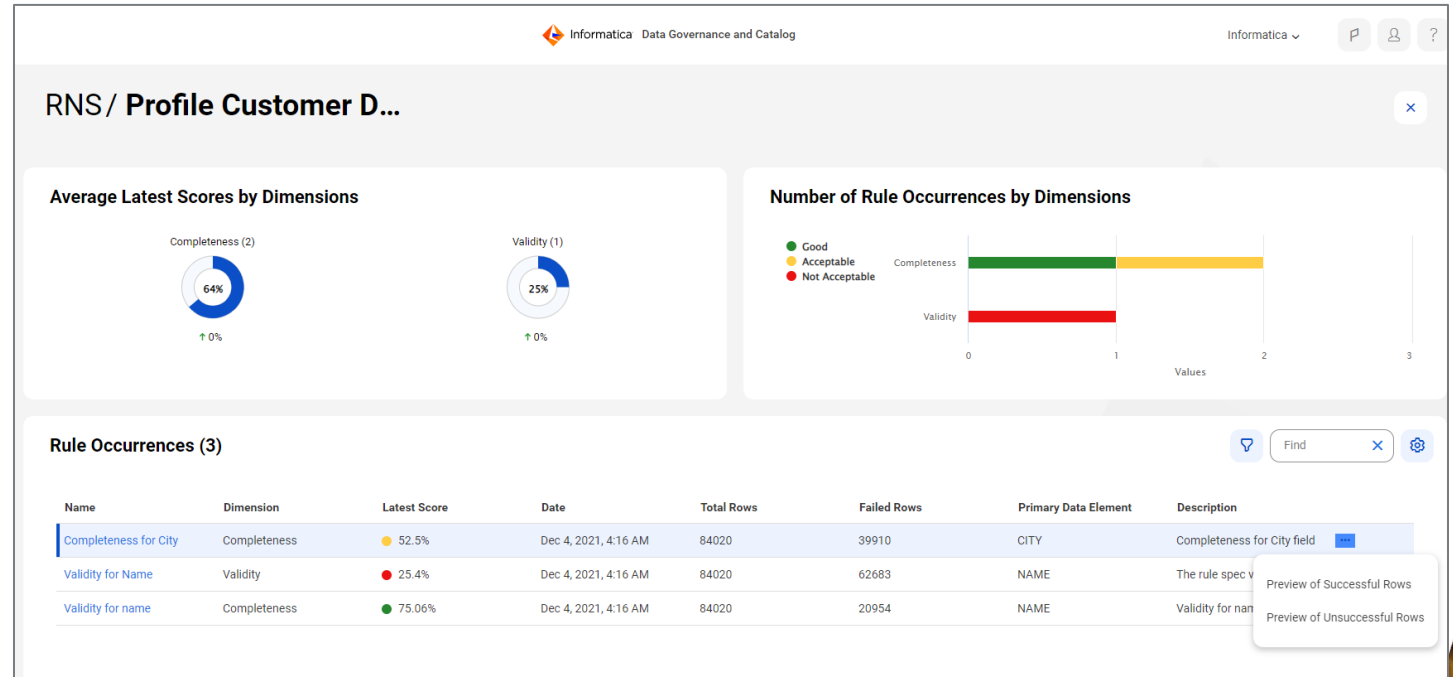
- **Definition:**

Scorecards are a graphical representation data quality scores calculated when a user profiles a source dataset.

- **Use case:**

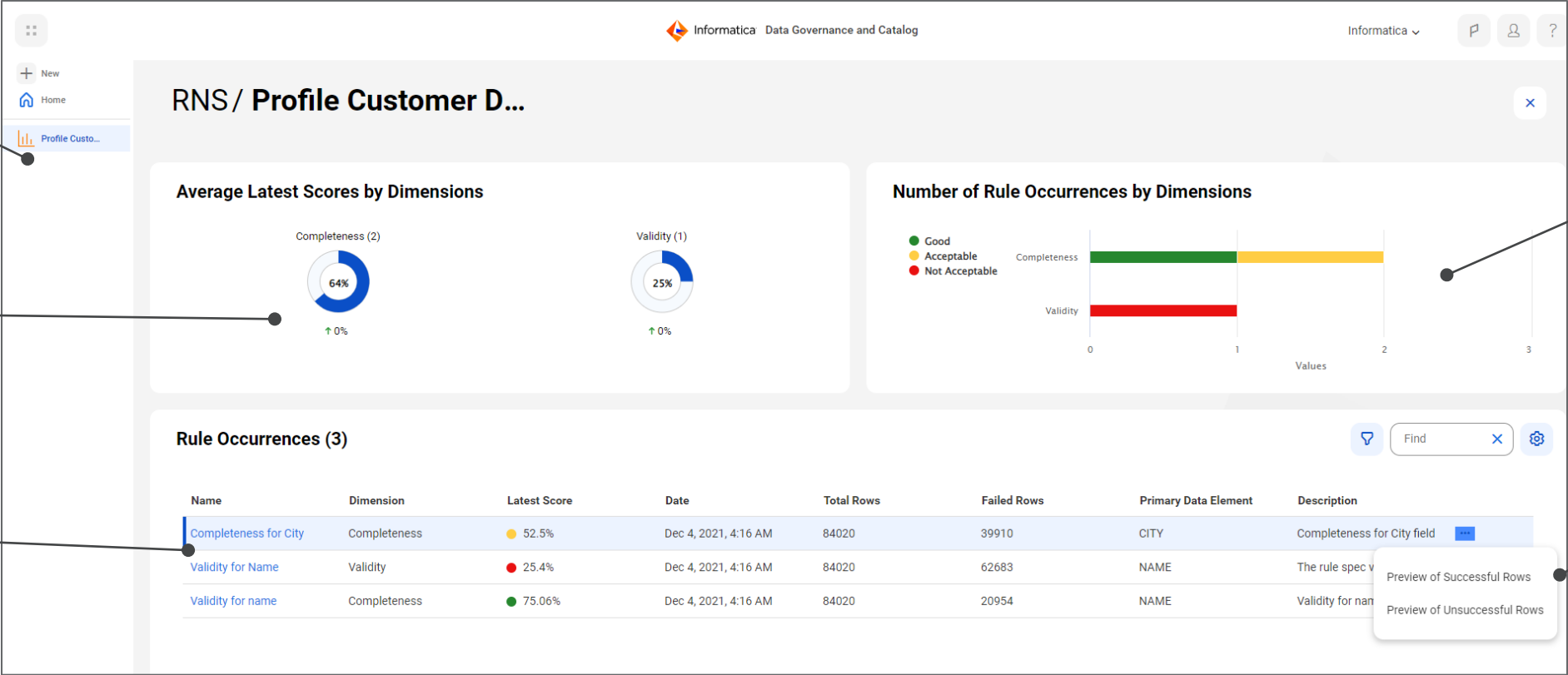
Use scorecards to measure and monitor data quality scores progress over time.

$$\text{Score (\%)} = \frac{\# \text{ Valid records} * 100}{\# \text{ Total records}}$$



Cloud Data Quality Scorecard

Details



Tab with Scorecard

Aggregation of rule occurrences per dimension

Individual scores per rule occurrence

Aggregation of rule occurrences per dimension

Drill down to Valid and Invalid

Cloud Data Quality Scorecard

Alerting/Notifications for scorecards



Use case:

- Notify user when a Data Quality score worsens
- Allows monitoring of scores without checking scorecards in UI



Features:

- Triggered when scores changes from:
 - Good -> Acceptable
 - Acceptable -> Not Acceptable
 - Good -> Not Acceptable


Data Quality Failure Notification for NotificationRO

 **noreply@informaticacloud.com** <noreply@informaticacloud.com>
To:  Kumar, Noreply




Retention: Exchange online 3 Years delete Expires: 29/01/2026.

 **Informatica** Organization:  Informatica
Date: 2023-01-30 19:05:33 IST

Hi Manoj Mishra,

 **Data Quality Status change from Acceptable to Not Acceptable**

Rule Occurrence Name:	NotificationRO
Description:	NotificationRO
Rule Dimension:	Accuracy

Rule Occurrence Thresholds:  Not Acceptable(Min 0%,Max 60%)  Acceptable(Min 60%,Max 80%)  Good(Min 80%, Max 100%)

<div style="border: 1px solid #ccc; padding: 5px;"><p>Current Score:</p><p style="font-size: 24px; color: red; margin: 0;">20.0</p><p style="font-size: 12px; margin: 0;">▼-66.67</p><div style="font-size: 10px; margin-top: 5px;">Score Status: Not Acceptable Total Rows: 100 Failed Rows: 10</div></div>	<div style="border: 1px solid #ccc; padding: 5px;"><p>Previous Score:</p><p style="font-size: 24px; color: orange; margin: 0;">60.0</p><p style="font-size: 12px; margin: 0;">▼-25.0</p><div style="font-size: 10px; margin-top: 5px;">Score Status: Acceptable Total Rows: 94 Failed Rows: 9</div></div>
--	---

Primary Data Element: [NAME](#)

[Open Data Quality Rule](#)

Cloud Data Quality Scorecard

Alerting/Notifications for scorecards

Features:

- Notification recipients: all stakeholders of metrics/Rule Occurrences
- In app and email notification
- Ability to:
 - Turn notifications on/off at stakeholder/IDMC user level
 - Receive Realtime notifications or digest

The screenshot displays the Informatica Data Governance and Catalog interface. The main section is titled "Settings" and contains "Notification Settings". A table lists notification types with checkboxes for "Email Summary" and "Email Event". The "Data Quality" row is highlighted, showing both checkboxes checked. To the right, "Notification Summary Email Settings" shows an "Email Summary Frequency" of "24 Hours". An inset window titled "Notifications" shows a list of alerts, including status changes from "Good" to "Not Acceptable" for Data Quality rules, with timestamps and "Open Rule" links.

Notification Type	Email Summary	Email Event
Collaboration	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Data Quality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Ticket Manager	<input type="checkbox"/>	<input checked="" type="checkbox"/>
User Jobs	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Workflow Tasks	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Notification Summary Email Settings

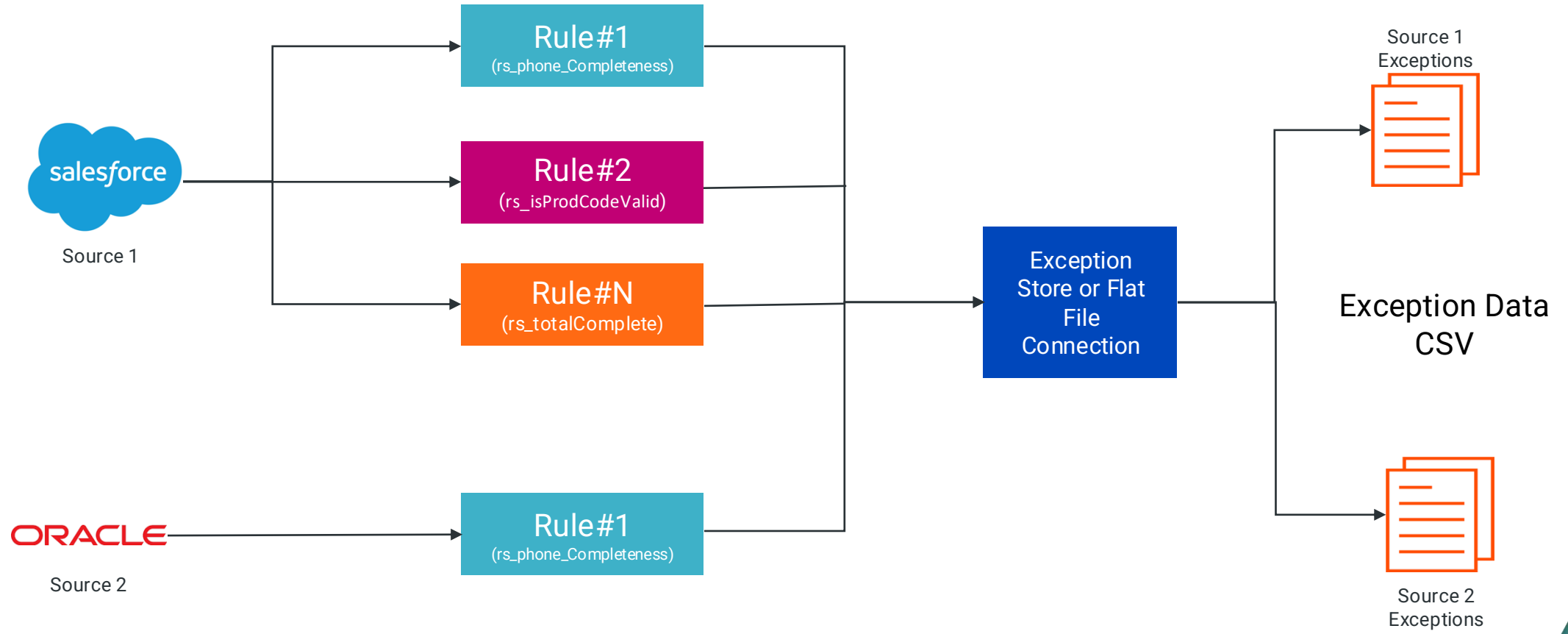
Email Summary Frequency: 24 Hours

Notifications

- Status change from "Good" to "Not Acceptable" for Data Quality : NotificationRO
Open Rule Jan 30, 2023, 7:11 PM
- Status change from "Acceptable" to "Not Acceptable" for Data Quality : NotificationRO
Open Rule Jan 30, 2023, 7:05 PM
- Status change from "Good" to "Not Acceptable" for Data Quality : DQ Rule Occurrence
Open Rule Jan 30, 2023, 1:35 PM
- Status change from "Good" to "Not Acceptable" for Data Quality : DQ Rule Occurrence
Open Rule Jan 24, 2023, 3:05 PM
- Status change from "Good" to "Not Acceptable" for Data Quality : DQ Rule Occurrence
Open Rule Jan 24, 2023, 2:36 PM

Data Quality Exception Management

Use Cases



Sample Exception Data

Exception Metadata Columns

Row Id	Email	Phone	SSN	Exception Message	Priority	Rule Name	Rule Input Columns	JobId	Creation Date
2		123		SSN can not be empty	Critical	rs_checkContact	SSN,Email,Phone	ZGE1OTdkNjQtNDFIYS00Zj	01/12/2022
23	xyz@gmail.com	456		SSN can not be empty	Critical	rs_checkContact	SSN,Email,Phone	ZGE1OTdkNjQtNDFIYS00Zj	01/12/2022
49			2345678	Missing phone number	Major	rs_checkContact	SSN,Email,Phone	ZGE1OTdkNjQtNDFIYS00Zj	01/12/2022
51	house@gmail.com	6768		SSN can not be empty	Critical	rs_checkContact	SSN,Email,Phone	ZGE1OTdkNjQtNDFIYS00Zj	01/12/2022
56	villa@gmail.com		45678	Missing phone number	Major	rs_checkContact	SSN,Email,Phone	ZGE1OTdkNjQtNDFIYS00Zj	01/12/2022
70	qwerty@gmail.com			SSN can not be empty	Critical	rs_checkContact	SSN,Email,Phone	ZGE1OTdkNjQtNDFIYS00Zj	01/12/2022

Source Data

Exception Management

New email notification with Job summary

❑ Revised / richer email update

- ❑ Provides Job summary
 - ❑ Source rows processed
 - ❑ Exception rows identified
 - ❑ Number of issues found
- ❑ Download Exception Data file link
- ❑ Top 10 Issues
- ❑ Top 10 rules identifying Issues

Job Summary

Source Rows
Exception rows
Number of
issues

Download
exception Data
file

Top 10 issues

Top 10 Rules
identifying
issues

Job Summary: Exception task ran successfully: Identified 86.3% exceptions

Exception Task Job: [Exception_Multiple_Rules_P0_43680-3](#)
Start Date: Aug 03, 2023, 1:02 PM
End Date: Aug 03, 2023, 1:02 PM
Started By: Hawkdgre1
Status: **Success**

Source Rows: Indicates the number of rows in the data source that the exception task job read.	183	Percentage of Exceptions Found: 86.3%
Exception Rows: Indicates the number of rows in which the exception task job found data quality issues.	158	
Number of Issues: Indicates the number of data quality issues that the exception task job found in the exception rows.	292	

Download Exception Data: [Exception_Multiple_Rules_P0_43680-3](#)

Counts per priority:

! Critical 3 ⬆ Major 123 ⬇ Minor 166

Top 10 Issues
Shows the top 10 issues, each with a count of one or more

Description:	Count
Countries - USP, GBR, UK is excluded	131
Rule_Escalations_Exception - exceeds 10	54
more then 10 escalations	54
overpaid	24
Name is null, so entry seems to be invalid	11
name is null	8
Country name cannot be empty	7
Invalid record or bad record, all the entries are null	3

Top 10 Rules identifying issues
Shows the top 10 rules that identify issues, each with a count of one or more

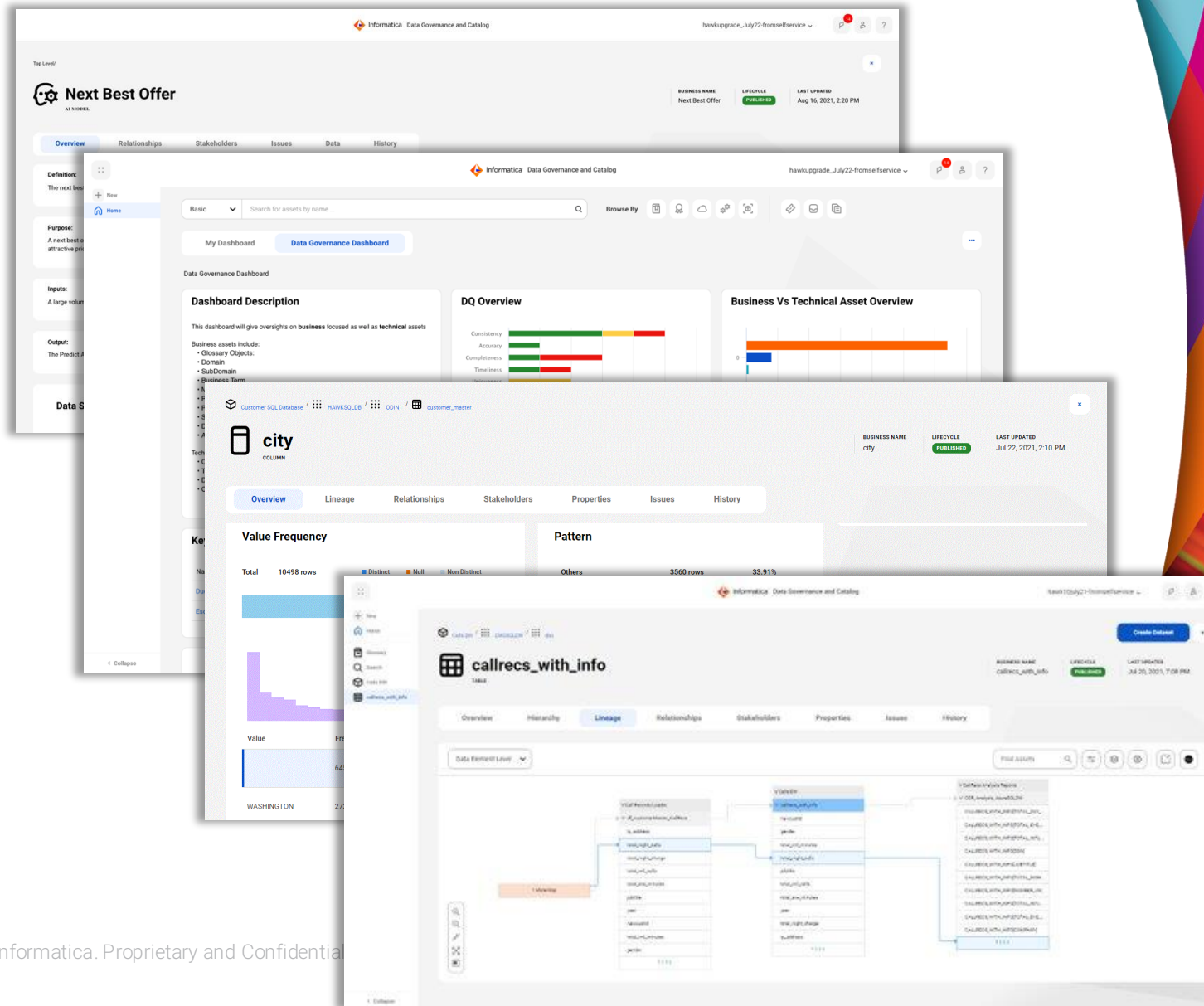
Description:	Count
Rule_List_of_Values_33767	138
Rule_String_Int_Float_Exception_40479	89
Rule_Escalation_Exception_2106	54

Data Quality for Data Governance

Where data & AI come to **LIFE**

Data Quality for Governance and Catalog

- Unified Governance & Catalog Capabilities
- Powerful Search Capabilities with Browsable Hierarchies
- Goal Oriented Workflows and Dashboards
- AI Model Governance
- Insights from Metadata with AI /ML
- Integrated Data Quality
- Hassle free Deployment & Serverless Execution



Natural Language Processing (NLP) for Data Quality Rule Generation

Use Case

Automatically generate data quality rules based on natural language and apply rules to all business entities

How It Works

Translate natural language sentences to executable Data Quality Rules - run pre-processor to generate primitives, use deep learning to tag entities, build data quality rule

Benefits:

Business can quickly define and execute data quality rules per governance policies automatically and consistently across the enterprise

Select Technical Rule Reference

☒ Create a Rule ☐ Pick an existing rule

Describe the rule [Click here to use the description from the rule template.](#)

CVV number must be between 001 and 999

[View Recommendations](#)

CLAIRE Recommendations

Select to view details and create a new rule:

☐ IF (CVV_CVV2 > '1') AND (CVV_CVV2 < '999') THEN TRUE

☒ IF (CVV_CVV2 >= '1') AND (CVV_CVV2 <= '999') THEN TRUE

CLAIRE will generate and attach the following rule:

Rule Name: BT_CVV__Val_Mar 14 2022 161302

Description: This is a CLAIRE generated rule based on the description "CVV number must be between 001 and 999"

Input: CVV_CVV2

☐ IF (CVV_CVV2 >= '1') AND (CVV_CVV2 < '999') THEN TRUE

☐ IF (CVV_CVV2 > '1') AND (CVV_CVV2 <= '999') THEN TRUE

[?](#) [OK](#) [Cancel](#)

Demonstration

Where data & AI come to



Informatica