

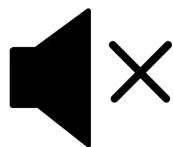
Apr 26<sup>th</sup>, 2022

# Cloud Data Integration with Cloud Data Quality

Shiv Patel, Principal Customer Success Technologist,  
Customer Success Management



# 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 **INFASupport YouTube channel** and [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.

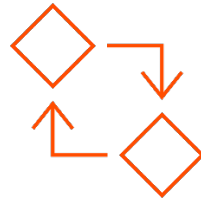
# Feature Rich Success Portal



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>



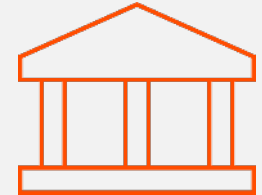
## Communities & Support

<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.

# Cloud Data Integration with Cloud Data Quality

# Agenda

1 Data Quality

---

2 Profiling

---

3 Data Quality  
Assets

---

4 Demo

---

# Data Quality

## Measure

- Review Progress
- Threshold Alerts
- Scorecards



## Apply

- Mapping Generation
- Standardization / Validation
- Matching / Consolidation

## Discover

- Profiling
- Identify Data Issues
- Set Data Quality Goals

## Define

- Dictionaries
- Implement Rules
- Build Cleanse, Parse, Verification, etc. processes

# Data Quality on IICS

Informatica Intelligent Cloud Services Log Out

My Services

B2B Partners Portal	Data Governance and Catalog	Data Integration
Data Marketplace	Data Profiling	Data Quality
Integration Hub	Mass Ingestion	PC to Cloud Conversion
Administrator	Monitor	Operational Insights

New ✕

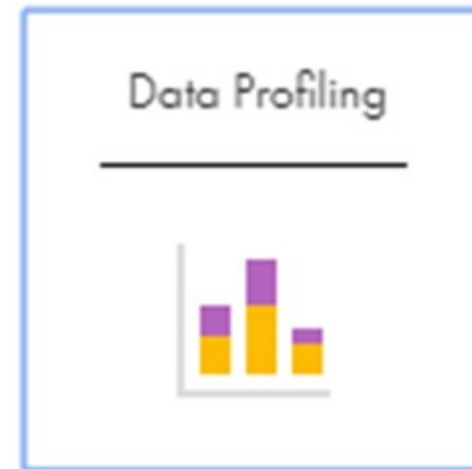
Select the type of asset you want.

Rule Specification	Dictionary	Verifier	Cleanse
Deduplicate	Parse	Labeler	

? Cancel

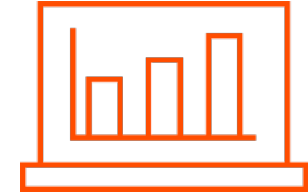
# Cloud Data Profiling

- Analyze data schemas
- Determine the quality of data across sources
- Understand the completeness, conformity, and consistency of data in the data sources.



# Cloud Data Profiling Statistics

Metrics Analyzed on the Profiled Data



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 (%/#) <b>NEW</b> (for numeric data types)		Pattern and Value Frequency Outlier

# Cloud Data Profiling

- Perform What-If scenarios by profiling outputs of Data Quality Assets and Mapplets
- Mapplets
- Parser
- Verifier
- Cleansing
- Rule Specification

Profile\_SingleInput... | Profile run 1 of 1 | 1 of 1 Columns | 1 of 1 Rules | 14 Rows (All rows) | Dec 6, 2021, 7:51:44 PM

Results | Definition | Rules | Schedule

View: Columns and Rules with: All Statistics

Columns	Value Distribution	% Null	# Null	% Distinct	# Distinct	% Non-dist...	# Non...
Columns							
▼ Columns							
COLUMN1		0%	0	35.71%	5	64.29%	9
▼ Mapplet_textDT	Input(s): COLUMN1						
outputText		0%	0	35.71%	5	64.29%	9

Profile\_Order | Profile run 1 of 1 | 33 of 33 Columns | 4 of 4 Rules | 10013 Rows (All rows) | Sep 3, 2020, 10:25:38 PM

Results | Definition | Rules | Schedule

View: Rules with: All Statistics

Columns	Value Distribution	% Null	# Null	% Distinct
Columns				
▶ v_discrete_all_op_ports	Input(s): COMPANY, COMPANYZIP, CUSTOMERADDRESS			
▼ Parse with dictionary	Input(s): COMPANYCITY			
City		0%	0	0.01%
Overflow		0%	0	0.01%
Unparsed		0%	0	10.89%
▼ cleanse_titlecase	Input(s): COMPANY			
cleanse_titlecase		0%	0	60.7%
▼ Validate_USCounty	Input(s): CUSTOMERCITY			
County_US_Validation		0%	0	0.02%

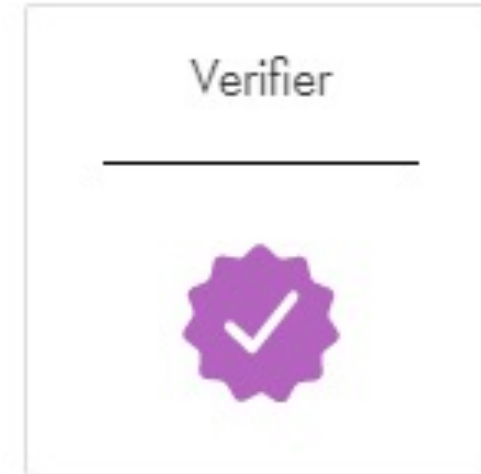
# Data Quality Assets

# Data Quality Assets

- Verifier
- Labeler
- Deduplicate
- Cleanse
- Dictionary
- Rule Specification
- Parse

# Verifier Assets

- Evaluates the accuracy and deliverability of address records
- Determine the accuracy of the input addresses
- Fix errors in the addresses
- Enhance the addresses where possible with additional information.
- Measure and report on the quality of each address



# Verifier Assets

- The verifier can perform the following operations
  - Verify
  - Format
  - Suggest
  - Measure
  - Enrich
  - Certify

Outputs ?

---

- ▶  Single address elements
- ▼  Preformatted data
  - ▶  Address Lines
  - ▶  Last Line
  - ▶  Formatted Address Lines
- ▼  Status Codes
  - Verification Status Code
  - Match Percentage
- ▼  Enrichments
  - ▶  Geo Coding
  - ▶  Certified
  - ▶  Global
  - ▶  Country Specific

# Labeler Assets


- Derives information about the content and structure of data.
- Perform token labeling or character labeling.
- Token labeling analyzes one or more tokens, or delimited values, in an input field
- Character labeling analyzes the individual characters in the input field.



# Labeler Assets

## When to use a labeler asset

- Verify business information with dictionaries
- Identify data by character format
- Review the structure of your input data

Regular Expression(365) ↓↑  

Name	Regular Expression	Outputs
Canadian Transit Number	<code>\d{5}\-\d{3} \d{9}</code>	1
Email	<code>.\@.\+[A-Za-z]+</code>	1
English Words		1
French Words		1
Email Alternate	<code>^[A-Za-z0-9!#\$%&amp;*+/?^_...</code>	1
Digits		1
Brazil Post Code	<code>^[0-9]{5}(-?[0-9]{3})?&amp;</code>	1

# Deduplicate Assets

- Calculates degree of similarity between records and generates output data for these calculations
- Based on the Match Transformation
  - Identity matching
  - Single source / single data set
- An identity is a set of data values that collectively identify a person or organization



# Deduplicate Assets

## Duplicate analysis methodology

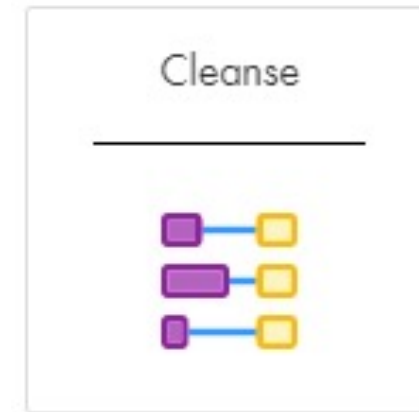
- It's a set of instructions for a Deduplicate transformation.
- Select the type of identity to search for at run time
- Define the search criteria to the input data.

The screenshot shows the configuration window for the 'Deduplicate' transformation in Informatica. The window has three tabs: 'Definition', 'Deduplication' (which is selected), and 'Consolidation'. Under the 'Deduplication' tab, there is a 'Configuration' section with the following settings:

Property	Value
Objective: ?	Wide Contact
Index Key: ?	Person Name
Data Locale: ?	United States
Optional Fields: ?	<input type="checkbox"/> Enable
Filter Exact Duplicates: ?	<input type="checkbox"/> Enable
Performance: ?	Fast and less specific

# Cleanse Assets

- It's a set of one or more data transformation steps that can standardize the form and content of your data
- Perform cleanse and merge operations on the input fields that a cleanse instance identifies in the asset



# Cleanse Assets

## Standardize the data

- Improve data consistency in a data set
- Fix errors in data
- Comply with regulatory standards
- Prepare for downstream data quality initiatives
- Merge data fields to effectively manage your data set

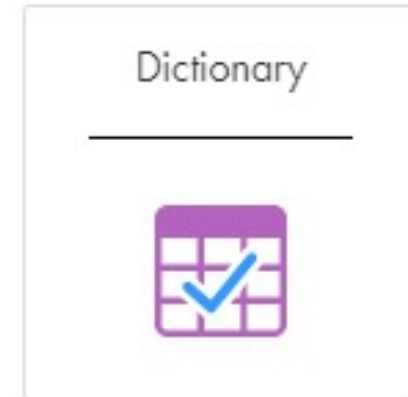
### Add Step

---

- Convert Case**  
Updates the character case of the data to the style that you select.
- Remove Values**  
Searches the data for one or more values and removes the values from the data.
- Remove Spaces**  
Removes redundant character spaces from the data.
- Replace Values**  
Searches the data for one or more values and replaces each value with another value.

# Dictionary Assets

- Reference data set that you can use to evaluate data in a mapping
- Use dictionaries to verify that the data values on a data source or another object in a mapping are accurate and correctly formatted



# Dictionary Assets

## Dictionary Operations

- Compares the input field data to the data in the dictionary
- If match, the transformation performs an action based on definitions

Column 1	Column 2	Column 3	Column 4	Column 5
Mississippi	228	601	662	769
Missouri	314	417	573	636
Montana	406			
Nebraska	308	402	531	
Nevada	702	725	775	
New Hampshire	603			

# Parse Asset

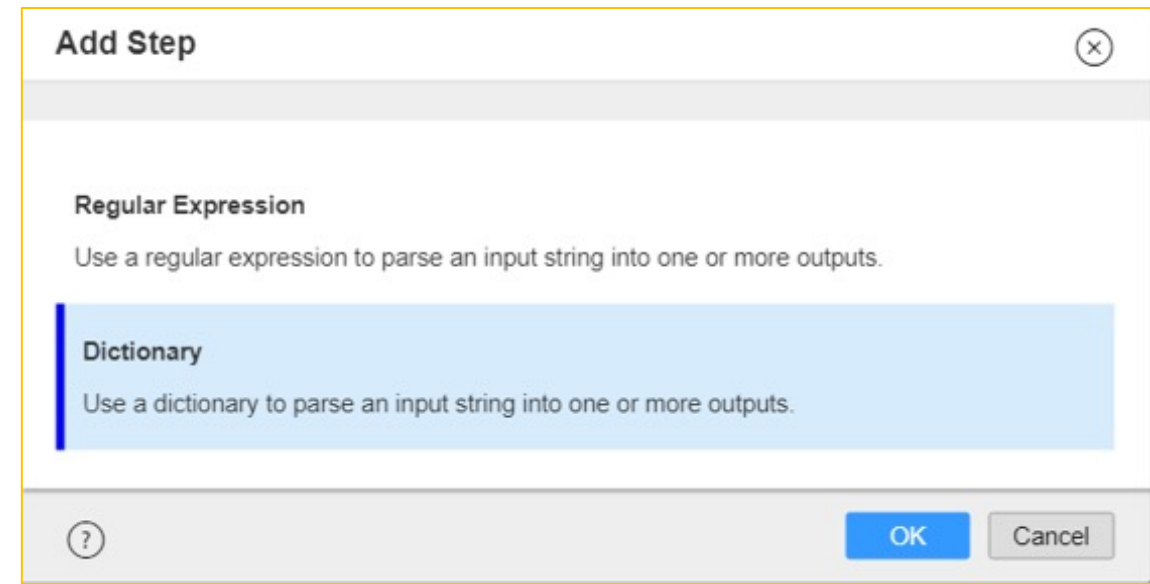
- Use a parse asset to improve the structure of your data
- A parse asset defines a set of operations that can identify discrete values in an input field and write the values to appropriate output fields



# Parse Assets

Parse operation to identify values in the following ways:

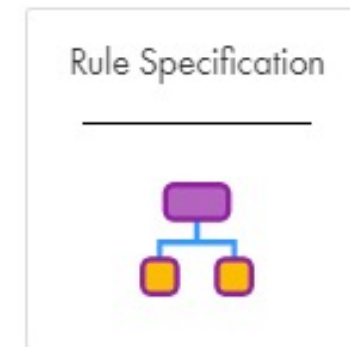
- Use a dictionary to identify values
- Use a regular expression to identify values
- Use a pre-built pattern to identify values



The screenshot shows a dialog box titled "Add Step" with a close button (X) in the top right corner. The dialog contains two main sections: "Regular Expression" and "Dictionary". The "Regular Expression" section has a description: "Use a regular expression to parse an input string into one or more outputs." The "Dictionary" section is highlighted with a blue background and has a description: "Use a dictionary to parse an input string into one or more outputs." At the bottom of the dialog, there is a help icon (question mark in a circle) on the left, and "OK" and "Cancel" buttons on the right.

# Rule Specification Asset

- A rule specification is an asset that represents the data requirements of a business rule in logical form



# Rule Specification Assets

## Rule Specification operation:

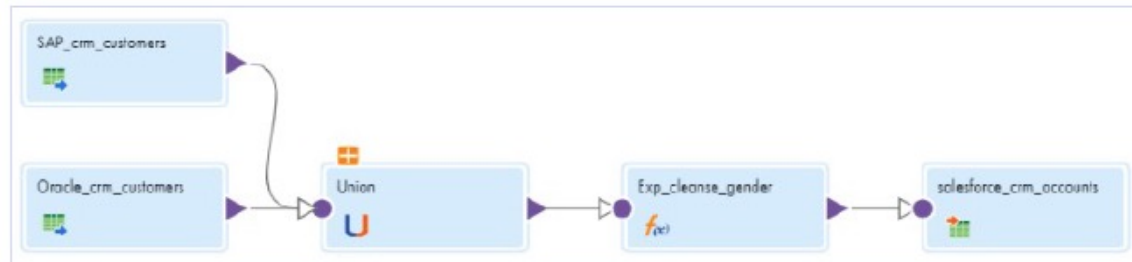
- Define the types of data that a business data set contains
- Define a set of conditions that the business data must satisfy
- Define the actions to take when the data satisfies the conditions of the business rule
- Define the actions to take when the data fails to satisfy the conditions of the business rule

The screenshot displays the 'Rule Specification' application window, currently in the 'Configuration' tab. The main design area shows a 'PrimaryRuleSet' component. Below the design area, the 'Properties: PrimaryRuleSet' section is visible, with the 'General' tab selected. The 'Rule logic definition' is shown in a table format.

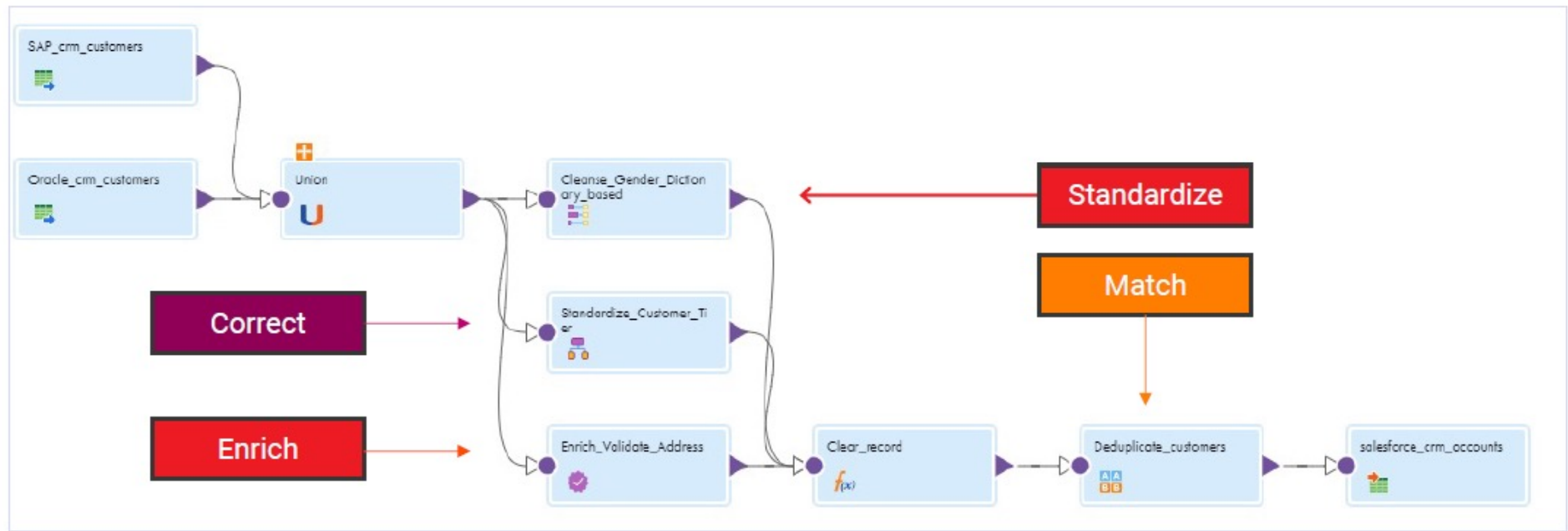
Inputs	Input	Operator	Condition	Action
Rule Logic	if		+ Add Input	then function
Test	or if		No rule statement is valid	then Do Nothing

# Data Quality in Data Integration Pipeline

Data Pipeline:  
Integration  
only



Data Pipeline:  
Integration  
and Data  
Quality



Demo

Q&A