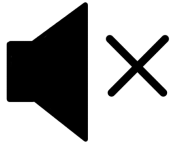


May 31st, 2022

Cloud Data Profiling - Using Rules and Dictionaries to Parse, Standardize, and Validate Data

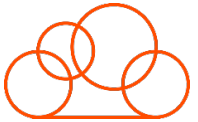
- Ranganadhamu Kusuma, Senior Product Manager

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.

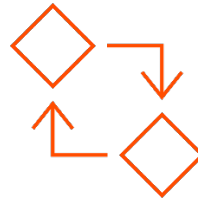
Feature Rich Success Portal



Bootstrap trial and
POC Customers



Enriched Customer
Onboarding
experience



Product Learning
Paths and Weekly
Expert Sessions



Informatica
Concierge



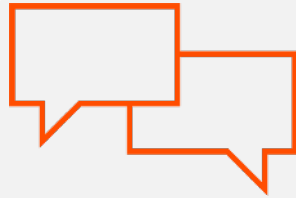
Tailored training and
content
recommendations

More Information



Success Portal

<https://success.informatica.com>



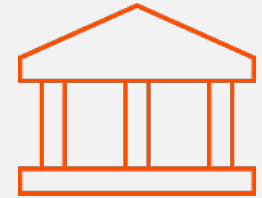
Communities & Support

<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.



Data Quality for Data Integration

Cloud Data Profiling - Using Rules and Dictionaries to Parse, Standardize, and Validate Data

Meet The Team

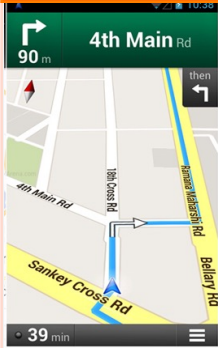


Ranganadhamu Kusuma

Product Management, Cloud Data Quality

Why Data Quality for Data Integration?

Profile



Turn by turn directions for data pipeline development instead of "heading west"

Standardize

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

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

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

Match/Consolidate

Source Data

System of Record

Land O Lakes	Land O'Lakes
Biluxi	Biloxi
Stephen Ahern	Steve Ahern

Validate

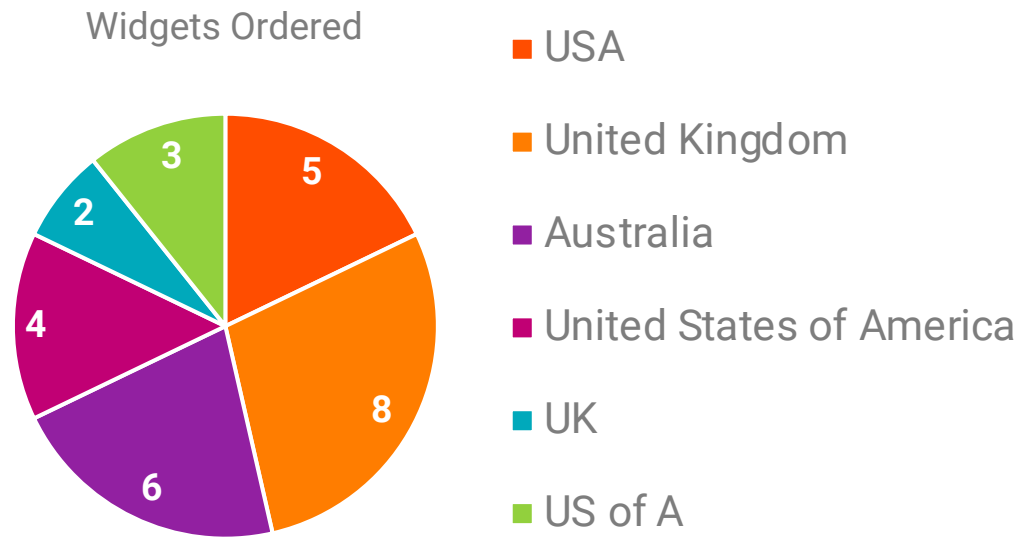
- ✓ Currency code must be consistent with country code
- ✓ Employee ID must be unique
- ✓ If customer tier is bronze then max credit is 1000
- ✓ All ICD10 codes must have a verified description

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

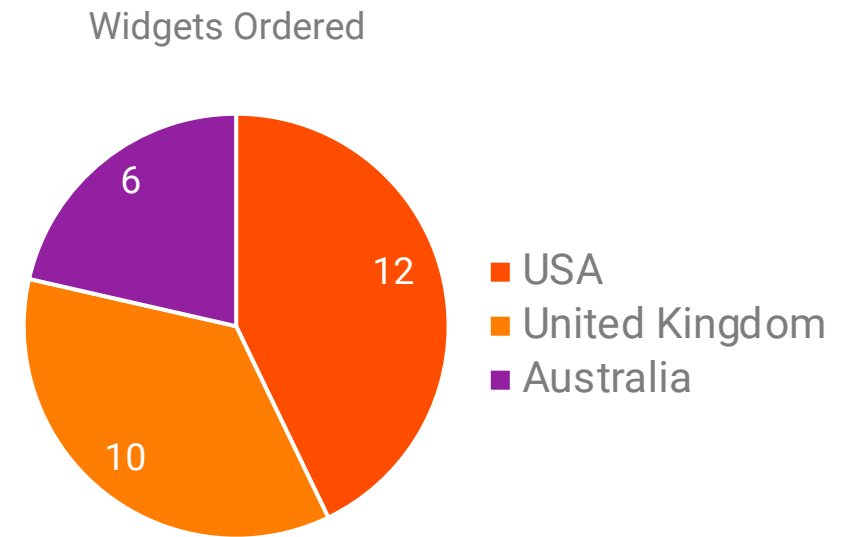
Why Data Quality for Business Intelligence?

Without Standardization



6 customers, United Kingdom is largest, Australia is second

With Standardization



3 customers, USA is largest, United Kingdom is second

Why Do You Need Data Quality for Artificial Intelligence?

The enormous power of AI can be crippled by poor quality data!

POOR QUALITY DATA



BEST-EVER, AI-DRIVEN



DATA SCIENCE PROJECT

POOR QUALITY RESULTS



Garbage In

Garbage Out

Hand Coding Data Quality is Time Consuming, Error Prone, Hard to Maintain, and Does Not Scale

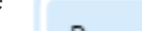
Parse Product Description

```
my_string = 'Names: Widget, Cog'
```

```
# split the string at ':'
```

ste # g ste # c

Parse_Product_Name



```
step_2 = step_1.split(',')
```

strip leading and trailing edge spaces of each item of the list

```
step_3 = [name.strip() for name in step_2]
```

```
# do all the above operations in one go
```

```
one_go = [name.strip() for name in my_string.split(':')[1].split(',')]
```

```
for idx, item in enumerate([step_0, step_1, step_2, step_3]):
```

```
print("Step {}: {}".format(idx, item))
```

```
print("Final result in one go: {}".format(one_go))
```

Verify Email

NOT(REGEX(Custom_Email_Field__c, '([a-zA-Z0-9_+]{1,254})'))

Standardize Country

IIF(OR(UPPER(Country)="UNITED STATES OF AMERICA",

Standardize_Country

try)="

=\"OZ\",

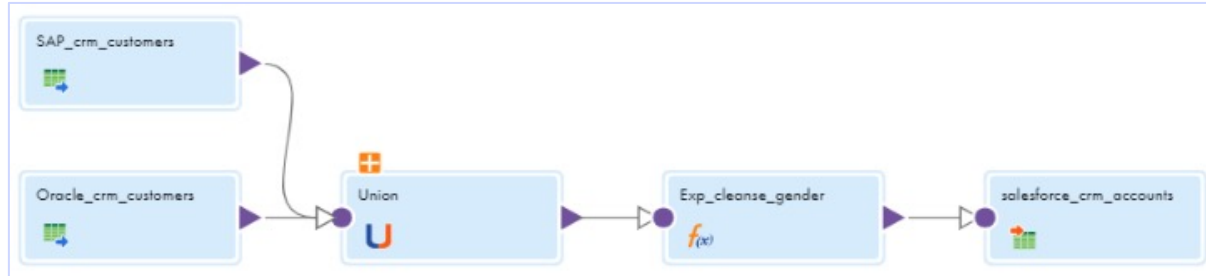
"Australia",Country)))

Standardized	Variation_1	Variation_2	Variation_3
USA	United States	US	America
United Kingdom	Great Britan	UK	

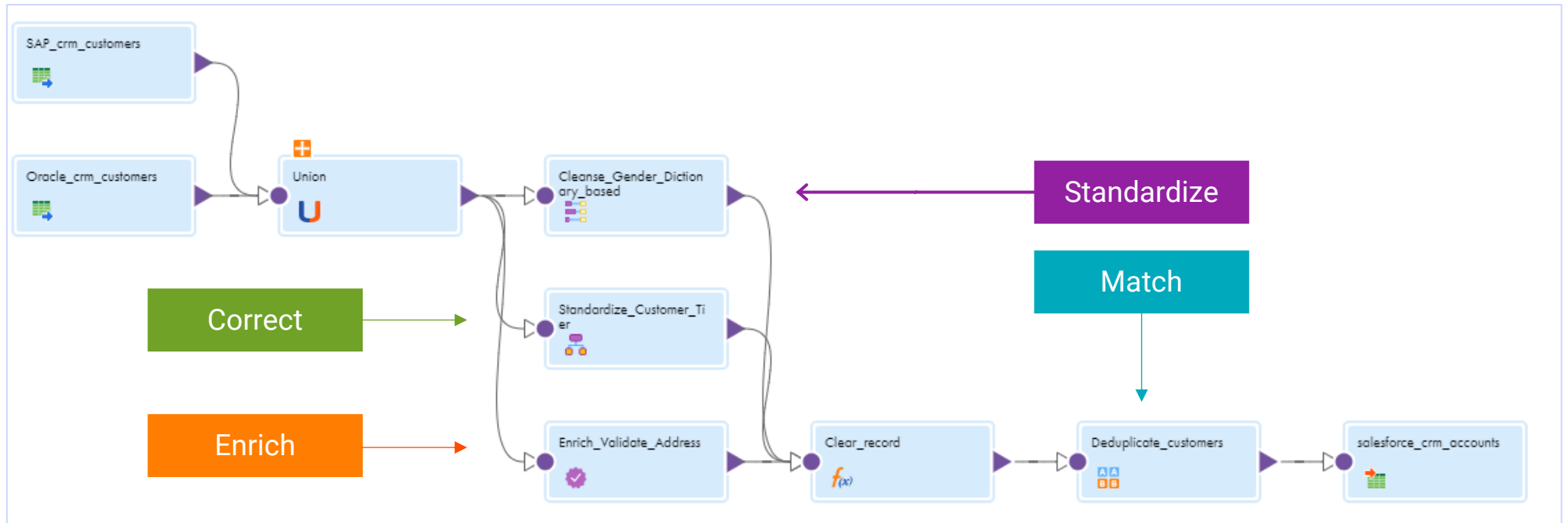
Drag and Drop No Code Data Quality into your Pipeline

Quickly Find and Fix Data Inconsistencies With Integrated Cloud Data Quality

Data Pipeline:
Integration only



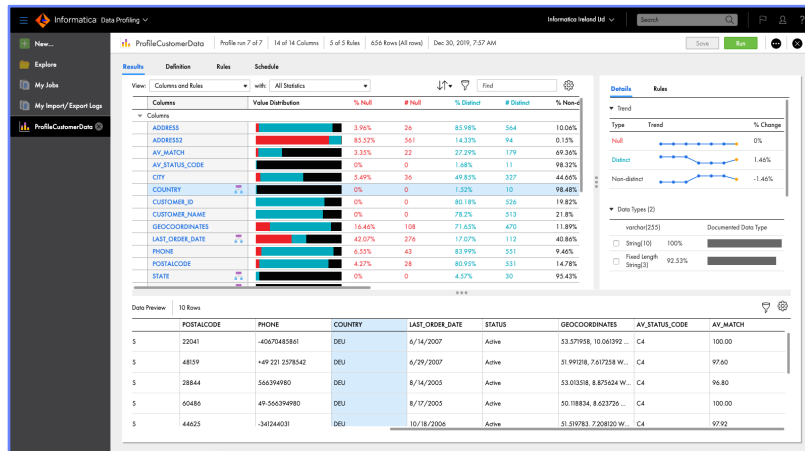
Data Pipeline:
Integration
and Data
Quality



Data Quality Collaboration

1

Automatically identify all data variations & build dictionary



2

Business Analysts own and update the dictionary, and define business rules

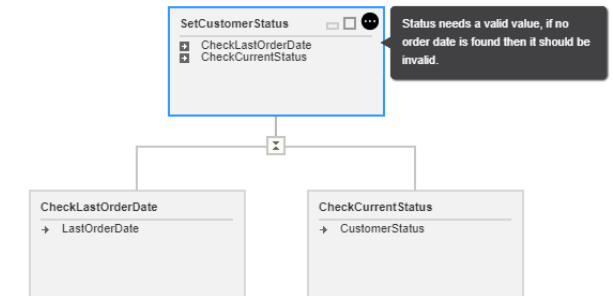


colors_info

Definition Configuration

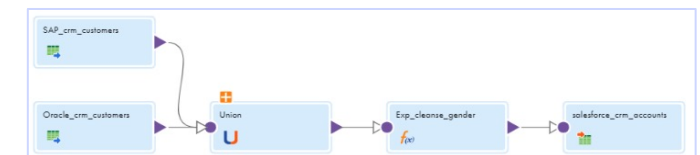
Items (1131)

	Column 1	Column 2	Column 3	Column 4	Column 5
338	Cobalt	Cobalt	CBL	CBLT	COB
339	Copper	Copper	CPR	COP	



3











Data Engineers & Data Pipelines reference Dictionaries and rules




Hundreds of Out-of-the-box Bundles/Accelerators

Rules, Dictionaries, Verifications And Mapplets

Dictionaries

Name	Type
 address_street_line_info	Dictionary
 blood_type_info	Dictionary
 bra_business_words_info	Dictionary
 bra_city_from_st_area_info	Dictionary
 bra_company_suffix_std...	Dictionary
 bra_gender_info	Dictionary
 bra_name_designators_i...	Dictionary
 bra_name_prefix_info	Dictionary
 bra_name_suffix_info	Dictionary
 bra_names_mispelled_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_mplt_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_BRA_CompanyName	Deduplicate
 dedupe_BRA_individual_name_...	Deduplicate
 dedupe_BRA_Individual_Name...	Deduplicate
 dedupe_BRA_Match_Company...	Deduplicate
 dedupe_BRA_Match_Indiv_Na...	Deduplicate
 dedupe_BRA_Match_Name_P...	Deduplicate
 dedupe_CompanyName_Match	Deduplicate
 dedupe_US_IMO_Fam_Name...	Deduplicate
 dedupe_US_IMO_Indiv_Name...	Deduplicate
 dedupe_US_IMO_Match_Com...	Deduplicate

Parse

Name	Type
 p_Parse_SSN	Parse
 p_Parse_US_FullName	Parse
 P_ParseFullName	Parse
 p_SSN_metrics_to_ind_fields	Parse
 p_SSN	Parse
 p_Suspect_Names	Parse
 p_Text_btwn_Parentheses	Parse
 p_Text_btwn_Single_Quotes	Parse
 p_top_level_domain	Parse
 p_usa_gender_assignemn...	Parse

Validate

Name	Type
 v_BRA_Address_Val_Hy...	Verifier
 v_BRA_AddressValidatio...	Verifier
 v_BRA_AddressValidatio...	Verifier
 v_BRA_Parse_Multiline_...	Verifier
 v_BRA_Validation_Discr...	Verifier
 v_BRA_Validation_Discr...	Verifier
 v_CAN_AddressVerificat...	Verifier
 v_Global_AddressValid...	Verifier
 v_US_AddressValidation...	Verifier
 v_US_AddressValidation...	Verifier



Demo

Informatica Cloud Data Quality

Deliver fit-for-purpose data throughout the data lifecycle

Simplicity



Low-code / No-code data profiling, cleansing, standardization and enrichment for all data

Productivity



Accelerate data quality projects within a familiar development environment

Scale



Cloud First, Cloud Native Data Management at Enterprise Scale


Leader



14-Time Gartner Magic Quadrant Leader Data Quality Solutions

Start Your Free Data Quality Trial Today

FREE TRIAL

 Informatica

Cloud Data Quality

Home > Trials > Cloud Data Quality

Free 30-Day Trial: Cloud Data Quality

Experience a no-code, visual environment for creating and managing data quality capabilities—all built on a leading next-generation iPaaS.

In this trial, you can:

- Create business logic to validate assumptions and statements about your data.
- Verify the accuracy and formatting of data values on a data source or another object in a mapping.
- Publish and consume reusable data quality artifacts across all data integration initiatives, including cloud data warehousing and cloud data migrations.

Interested in more Informatica Cloud Services? [Explore our other free trials.](#)

Work Email

☒ Use my email address as my username.

First NameLast Name

Title

Select User Role

Phone Number

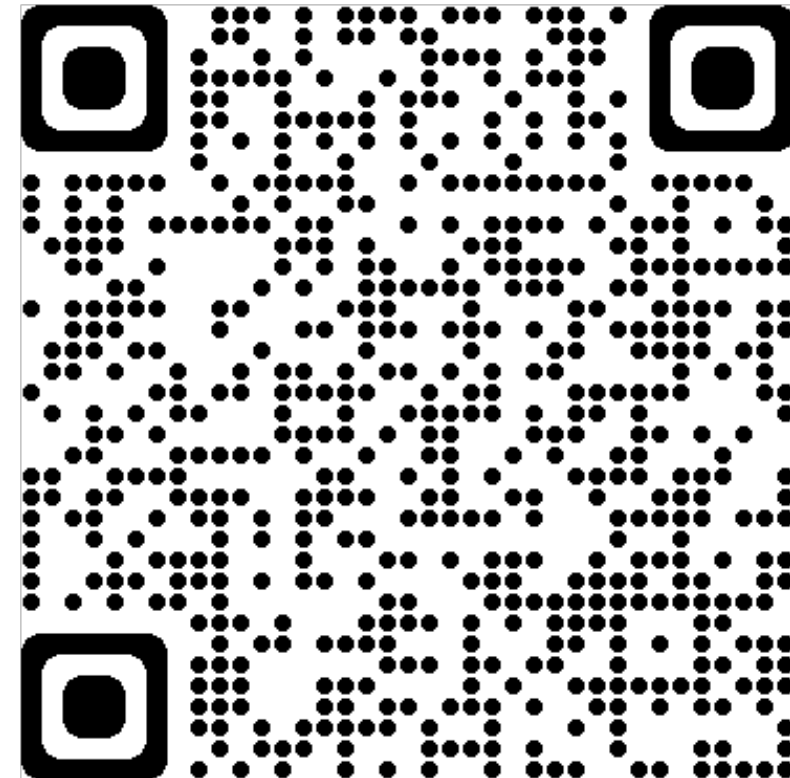
Organization
Informatica Ireland Ltd

Ireland

State / Province

Data Center Location
North America

☐ Yes, I would like to receive marketing communications from Informatica about products, solutions and events of Informatica and its partners. You can opt out from receiving these communications at any time.



Data Quality Trial: <https://www.informatica.com/trials/cloud-data-quality.html>

Thank you