Aug 2023

Informatica Data Quality & Observability

Helen Kinsella

Principal Product Specialist

hkinsella@informatica.com

Informatica®

Consistent Data Quality Process and Methodology

Discover

Data Quality Process

Define

Measure & Monitor

Apply

Informatica

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	System of Record
Land O Lakes	Land O'Lakes
Biluxi	Biloxi
KitKat	Kit-Kat

Parse

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

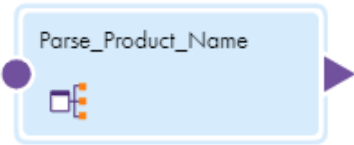I love my Pink 64GB iPad Pro!!!

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

Informatica®

Hand Coding Data Quality Is Time Consuming, Error Prone, Hard to Maintain, and Does Not Scale
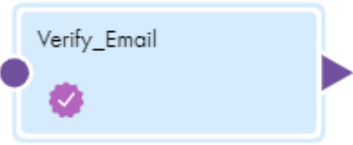
Parse Product Description

```
my_string = 'Names: Widget, Cog
# split the string at ':'
step_0 = my_string.split(':')
# get the first slice of
the list
step_1 = step_0[1]
# split the string at ','
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))
```
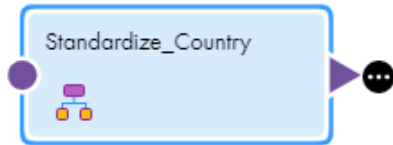
Parse_Product_Name

Verify Email

```
NOT(REGEX( Custom_Email_Field__c ,'([a-zA-
Z0-9_\\-\\.]+)@((\\[a-z]{1,3}\\.[a-z]{1,3}\\.[a-
z]{1,3}\\.)|(([a-zA-Z0-9\\-]+\\.)+))([a-zA-
Z]{2,4}|[0-9]{1,3})'))
```

Verify_Email

Standardize Country

```
IIF(OR(UPPER(Country)="UNITED STATES OF
AMERICA", UPPER(Country)="US OF
A"),"USA",IIF(OR(UPPER(Country)="UK",UPPER
(Country)="GREAT BRITAIN"),"United
Kingdom",IIF(UPPER(Country)="OZ","Australia",
Country)))
```
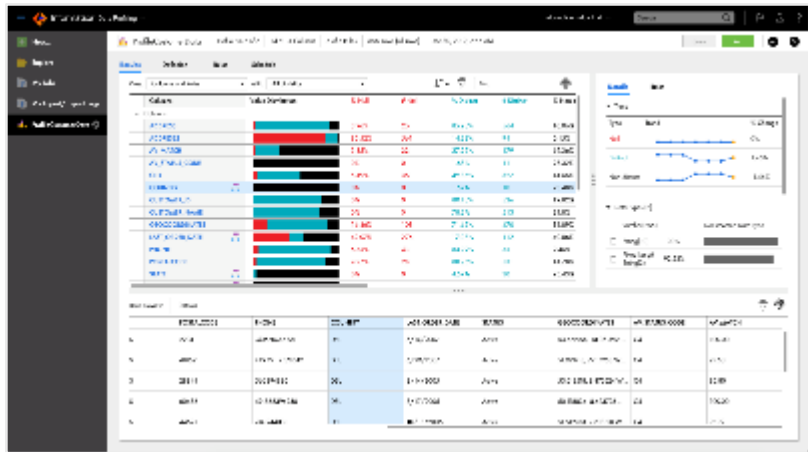
Standardize_Country

Standardized	Variantion_1	Variation_2	Variation_3
USA	United States	US	America
United Kingdom	Great Britan	UK	

Informatica

Data Quality Collaboration

1 **Automatically** identify all data variations and **build dictionary**

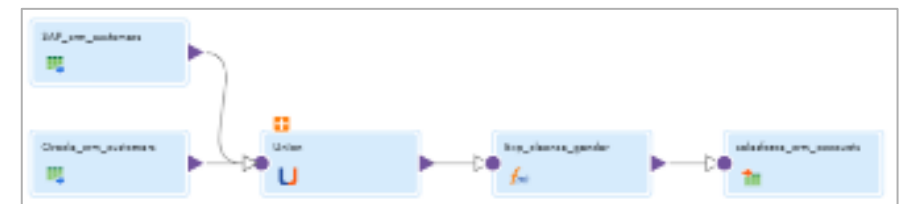**2** Business Analysts own/ update the dictionary and define business rules
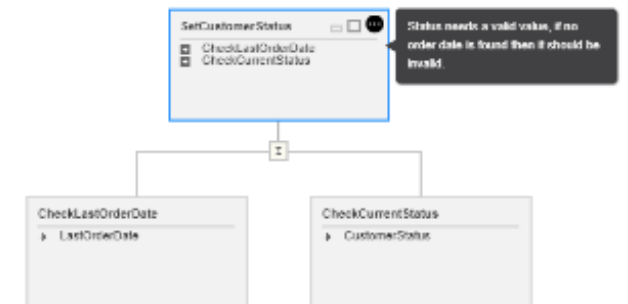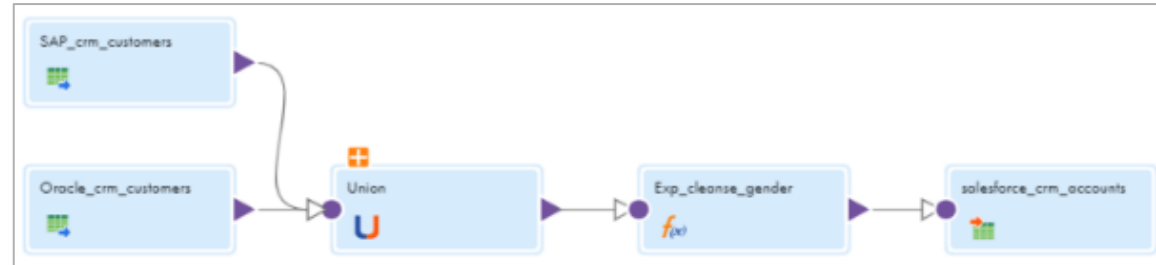
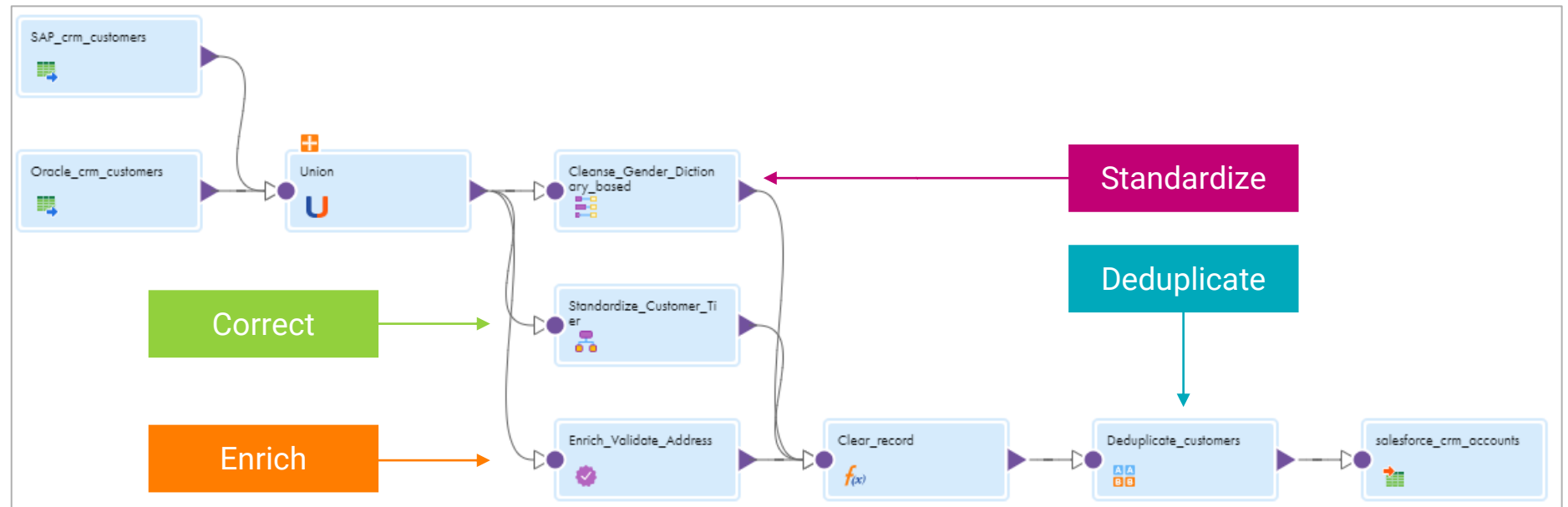**3** Data Engineers and Data Pipelines reference Dictionaries and rules

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

© Informatica. Proprietary and Confidential.

Hundreds of Out-of-the-Box Accelerators
Rules, Dictionaries, Verifications and Mapplets

Dictionaries	Cleanse	Deduplicate	Parse	Validate

Name ▲	Type
character_sets	Dictionary
diacritics_translate_infa	Dictionary
domain_names_infa	Dictionary
dq_av_mailabilityscores_infa	Dictionary
dq_av_match_code_descriptions_i...	Dictionary
iso_full_country_name_infa	Dictionary
numeric_default_values_infa	Dictionary
string_default_values_infa	Dictionary
suspect_names_infa	Dictionary
user_defined_dates_infa	Dictionary

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

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

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_btw_Parentheses	Parse
p_Text_btw_Single_Quotes	Parse
p_top_Level_domain	Parse
p_usa_gender_assignemn...	Parse

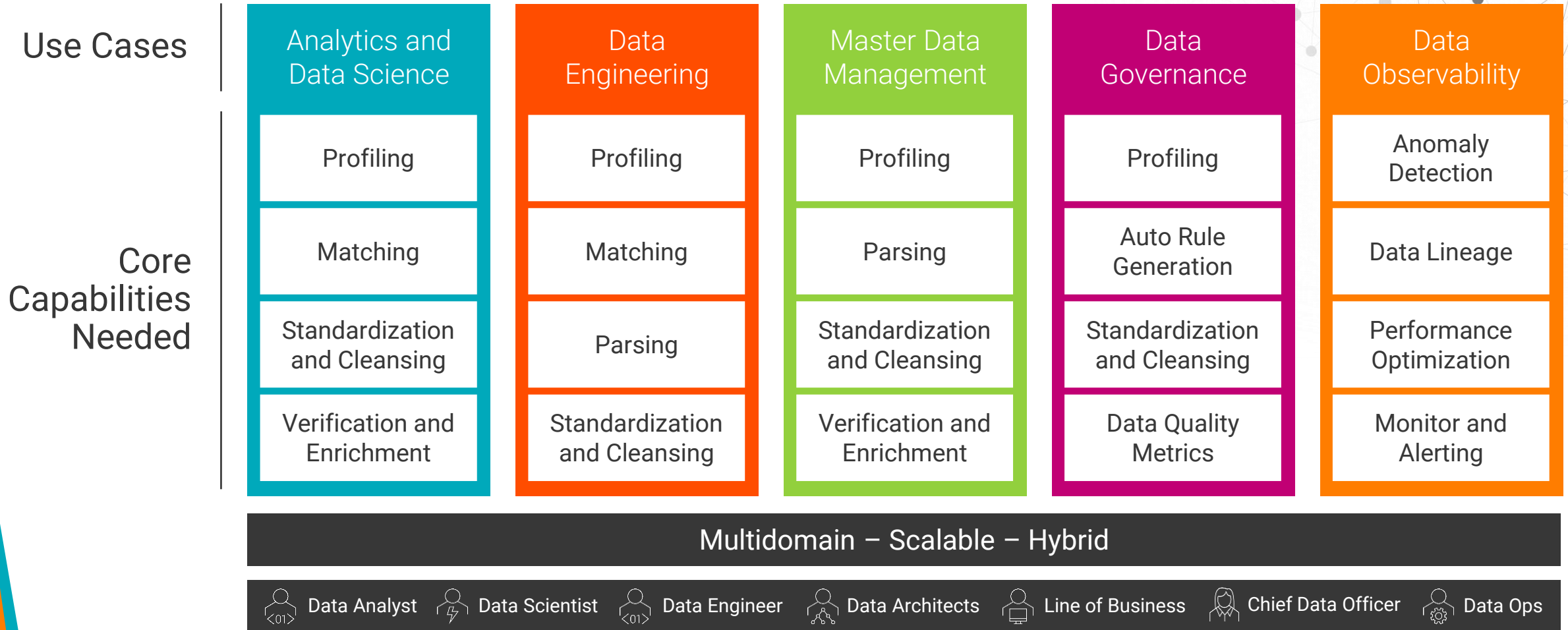| Name ▼	Type
v_USA_Validation_Discrete_Addre...	Verifier
v_USA_Validation_Discrete_Address	Verifier
v_USA_AddressValidation_Multiline	Verifier
v_USA_AddressValidation_Geo_...	Verifier
v_US_AddressValidation_Hybrid	Verifier
v_US_AddressValidation_Hyb_w_...	Verifier
v_CAN_AddressVerification_Hybrid	Verifier
v_BRA_Validation_Discrete_Addre...	Verifier
v_BRA_Validation_Discrete_Address	Verifier
v_BRA_Parse_Multiline_Address_...	Verifier

Informatica

DEMO

Data Quality for Everyone, Everywhere

Use Cases

Analytics and Data Science	Data Engineering	Master Data Management	Data Governance	Data Observability
Profiling	Profiling	Profiling	Profiling	Anomaly Detection
Matching	Matching	Parsing	Auto Rule Generation	Data Lineage
Standardization and Cleansing	Parsing	Standardization and Cleansing	Standardization and Cleansing	Performance Optimization
Verification and Enrichment	Standardization and Cleansing	Verification and Enrichment	Data Quality Metrics	Monitor and Alerting

Core Capabilities Needed

Multidomain – Scalable – Hybrid

Data Analyst   Data Scientist   Data Engineer   Data Architects   Line of Business   Chief Data Officer   Data Ops
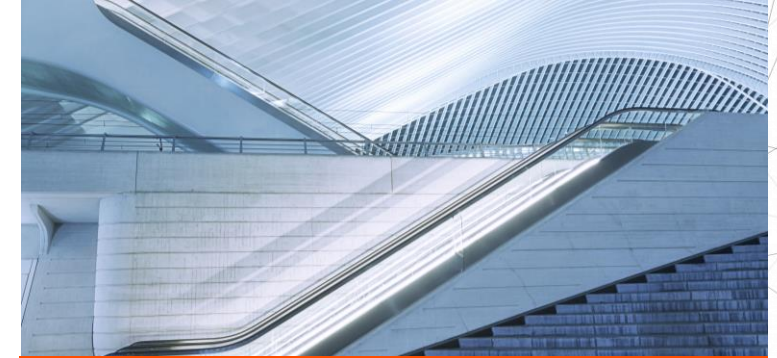
Informatica®

The Value

Reduce Risk

A data quality strategy identifies issues early and allows you to resolve them faster, ensuring projects are not delayed

Accelerate Development

Low-code / no-code data profiling, cleansing, standardization and enrichment within a familiar development environment

Scale & Reuse

Build, deploy, scale and re-use centrally manages data quality services to support all applications and data pipelines

Informatica

Visit the Informatica Experience Lounge to Learn More