Sep 14, 2021

# EDC Data Domain Best Practices

Hari Krishna Akula

Solutions Architect, Data Governance

Informatica™

# 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.

# Agenda

1 Data Domain Overview

2 Data Domain Types

3 Best Practices

# EDC Data Domain Overview

Informatica®

# Data Domain

## EDC Data Domain Overview

**A data domain** is a predefined or user-defined Model repository object that enables you to discover the functional meaning of column data or column names in an asset. Examples of data domains include Social Security number, account status, IP address, Part Numbers and UPC code. You can add one or more data domains into a data domain group.

You can use data domains to identify and understand the meaning of critical source data or undiscovered source data so that you can take measures, such as data masking, to work effectively on it. For example, you might have legacy data systems that contain Social Security numbers in a Comments field. You need to find this information so that you can take appropriate measures before you move it to new data systems.

Here data domain is an INFORMATICA Enterprise Data Catalog (EDC) terminology**, **it is **not to be confused with** common use of the term in the industry to denote Subject Areas such as Customer, Product, Vendor etc.

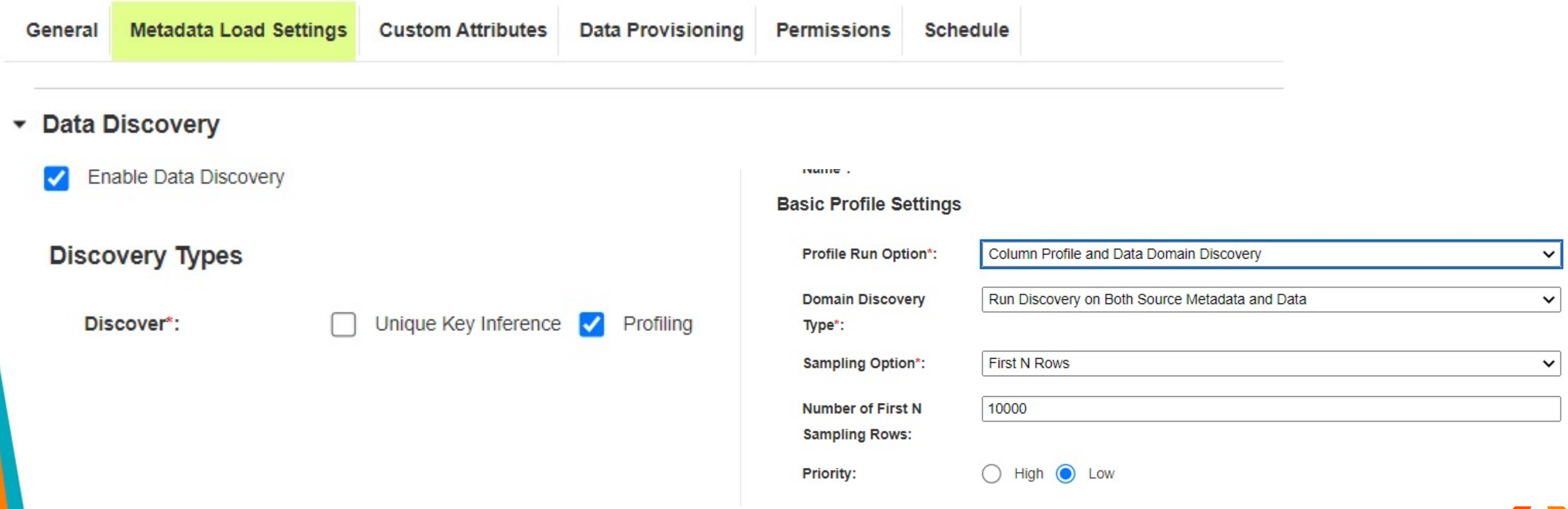
**Also not to be confused with** common use of the terminology ‘domain’ in Axon , which is a used to represent a type of a top-level glossary item to represent a subject area.

# How Data Domain works

## EDC Data Domain configuration

Data domain discovery is the process of discovering the functional meaning of data in the data sources based on the semantics of data. After you enable data domain discovery for a resource and run the resource, the profiling scanner uses the data domains to infer matching column data or column name patterns from the metadata extracted by the resources.

**EDC Administrator resource configuration UI**



Data Discovery

☑ Enable Data Discovery

Discovery Types

Discover*: ☐ Unique Key Inference ☑ Profiling

Basic Profile Settings

Profile Run Option*: Column Profile and Data Domain Discovery

Domain Discovery Type*: Run Discovery on Both Source Metadata and Data

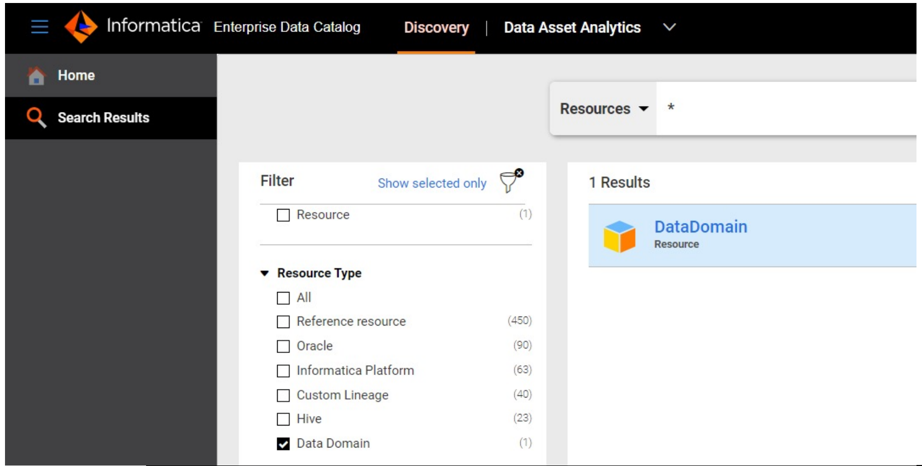
Sampling Option*: First N Rows

Number of First N Sampling Rows: 10000

Priority: ○ High ● Low

# Viewing Data Domains

In Enterprise Data Catalog 10.5 onwards, you can view DataDomain as a resource. Before 10.5 individual data domains are available as assets that can be searched for. Inferred data domains can be seen at table, column and field level in EDC.



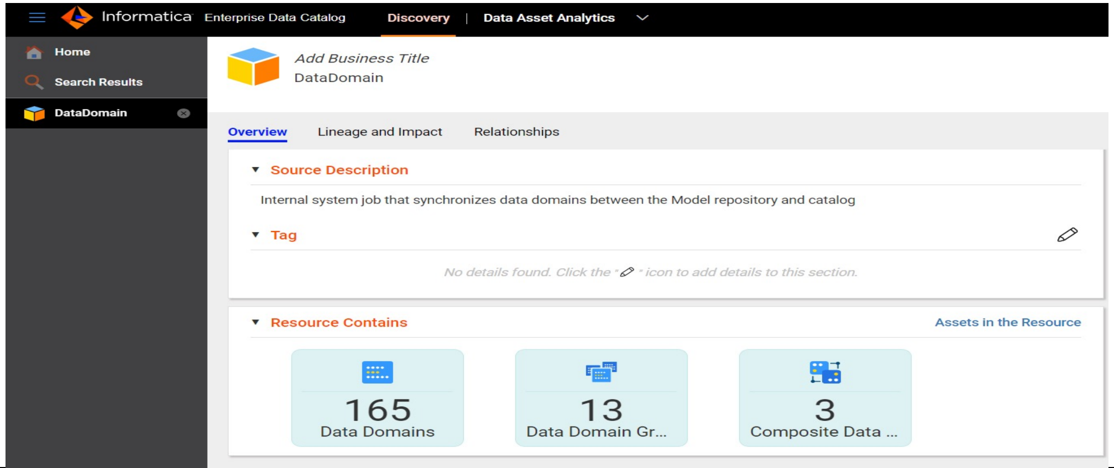
Informatica Enterprise Data Catalog
Discovery | Data Asset Analytics
Home
Search Results
Resources *
Filter Show selected only

□ Resource (1)

Resource Type
- □ All
- □ Reference resource (450)
- □ Oracle (90)
- □ Informatica Platform (63)
- □ Custom Lineage (40)
- □ Hive (23)
- ☑ Data Domain (1)

1 Results

DataDomain
Resource



Source Description

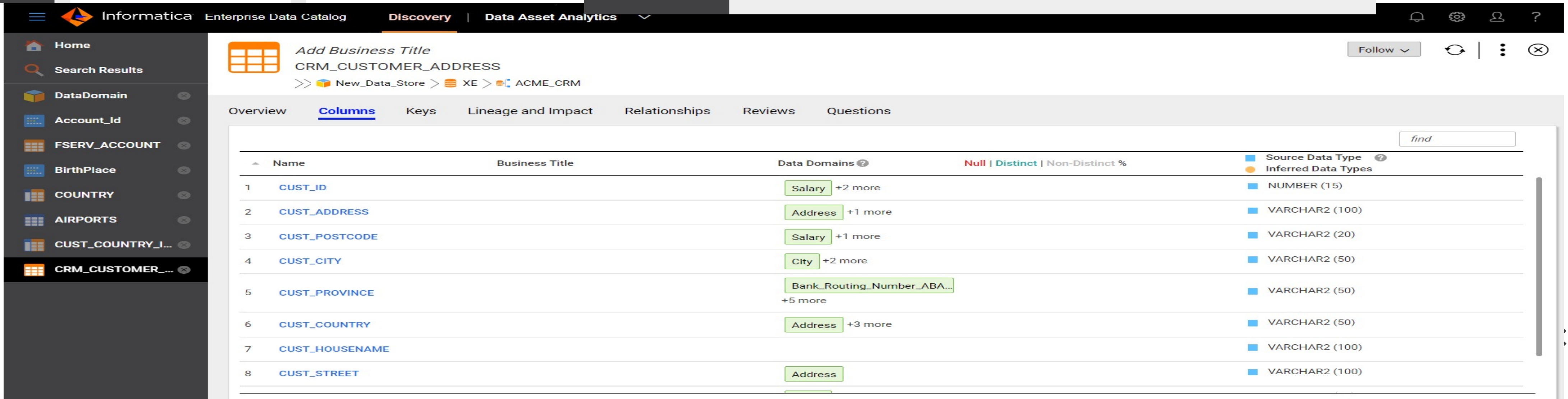
Internal system job that synchronizes data domains between the Model repository and catalog

Tag

No details found. Click the " " icon to add details to this section.

Resource Contains — Assets in the Resource

165 Data Domains | 13 Data Domain Gr... | 3 Composite Data ...

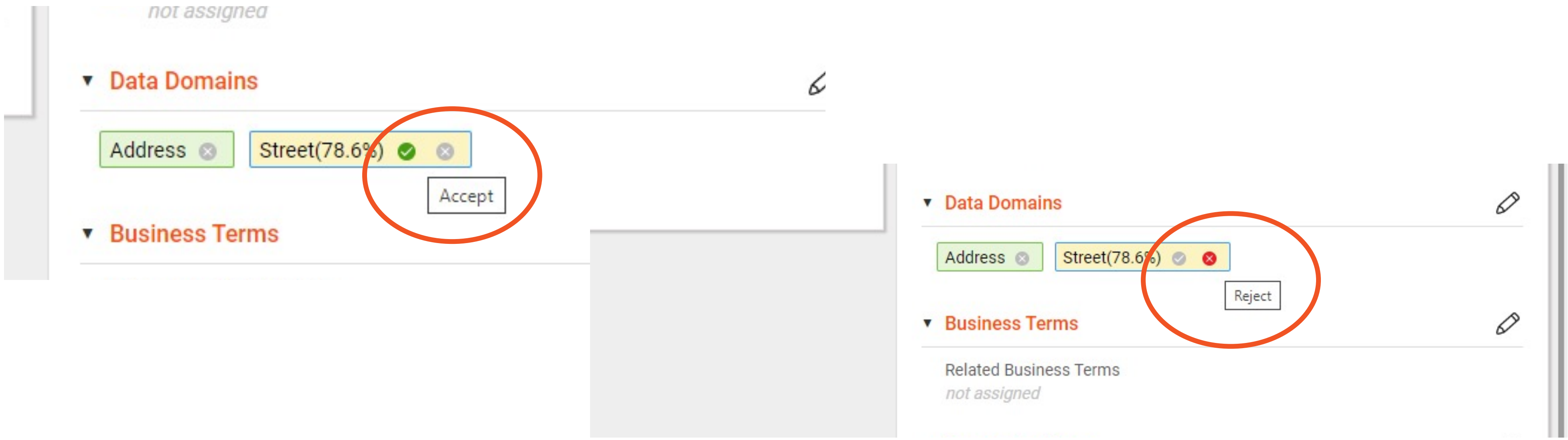


Add Business Title
CRM_CUSTOMER_ADDRESS
>> New_Data_Store > XE > ACME_CRM

| | Name | Business Title | Data Domains | Null \| Distinct \| Non-Distinct % | Source Data Type<br>Inferred Data Types |
|---|---|---|---|---|---|
| 1 | CUST_ID | | Salary +2 more | | NUMBER (15) |
| 2 | CUST_ADDRESS | | Address +1 more | | VARCHAR2 (100) |
| 3 | CUST_POSTCODE | | Salary +1 more | | VARCHAR2 (20) |
| 4 | CUST_CITY | | City +2 more | | VARCHAR2 (50) |
| 5 | CUST_PROVINCE | | Bank_Routing_Number_ABA...<br>+5 more | | VARCHAR2 (50) |
| 6 | CUST_COUNTRY | | Address +3 more | | VARCHAR2 (50) |
| 7 | CUST_HOUSENAME | | | | VARCHAR2 (100) |
| 8 | CUST_STREET | | Address | | VARCHAR2 (100) |

# Curating Data Domains

You can accept or reject the inferred assets for the data domain. You can also view the data domains for tabular, column, and field assets. These data domains are inferred for the asset from the profile results or from similar columns. You can accept or reject a inferred data domain for a tabular, column, or field asset.



Related Business Terms
not assigned

# EDC Data Domain types

Informatica

There are three categories of data domains.

- **Rule based**
- **Smart data domains**
- **Composite Data Domain**

**Rule based data domain**

Rule based data domains are the ones where the semantic meaning of a column can be discovered based on a rule.

- ***The rules can be "metadata based" or a "data rule".***
- The rules can be out of box shipped with Informatica or custom created as per the requirement.
- There are three general categories of rules.
  - **Regex-based rule:** Regex to determine if the metadata or data follows a pattern.
  - **Reference table-based rule:** If there is a finite set of data and typically non-overlapping, then it can be validated against a reference table.
  - **Mapplet Rules:** These are mapplets created using Informatica's Developer tool and can involve a combination of "regex", "reference table", "lookups" and several other expressions for example to check whether a value falls in a range or not.

# Smart Data Domains

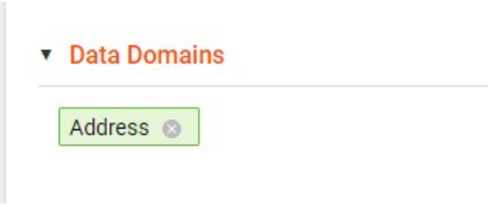
This data domain does not have any rule. It is also called “example based” data domain. This is more like a user “tagging” a column by observing various factors like the name of the column and the profiling statistics like “pattern” or min or max values.

- The users can in such cases can tag the column with a data domain.
- This data domain can be created on the fly. It does not necessarily have to exist in the catalog at the time of creation.

After creating smart data domain, propagate it to all the assets by running the **SimilarityDiscovery** scanner and **DataDomainPropagation** scanner in Catalog Administrator to discover assets that contain a similar pattern.

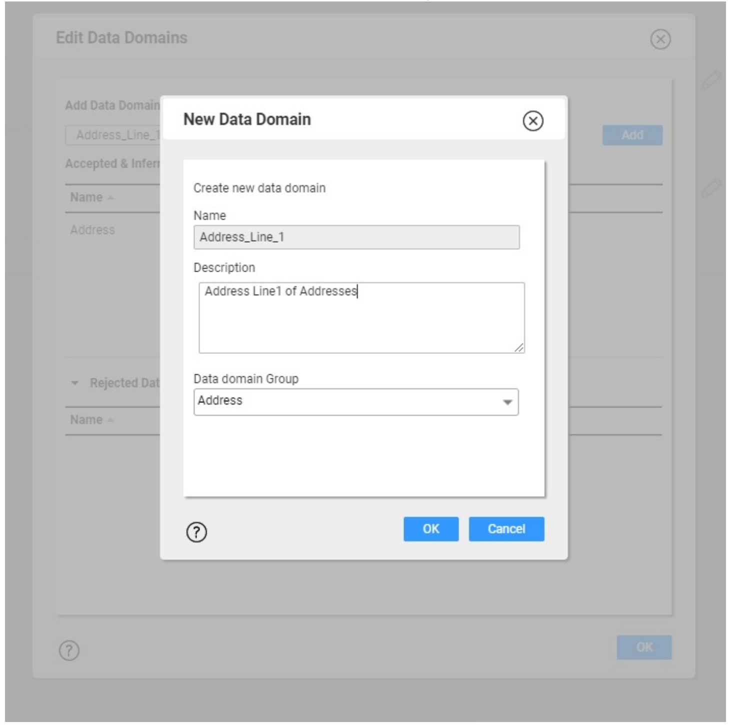
***smart data domain created by user in EDC UI on the fly***

***Before smart data domain***



▾ Data Domains

Address



Edit Data Domains

Add Data Domain

Address_Line_1

Accepted & Inferred

Name

Address

Rejected Data Domains

Name

New Data Domain

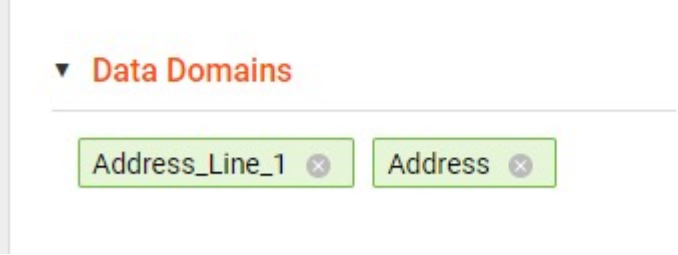
Create new data domain

Name: Address_Line_1

Description: Address Line1 of Addresses

Data domain Group: Address

***After smart data domain creation***



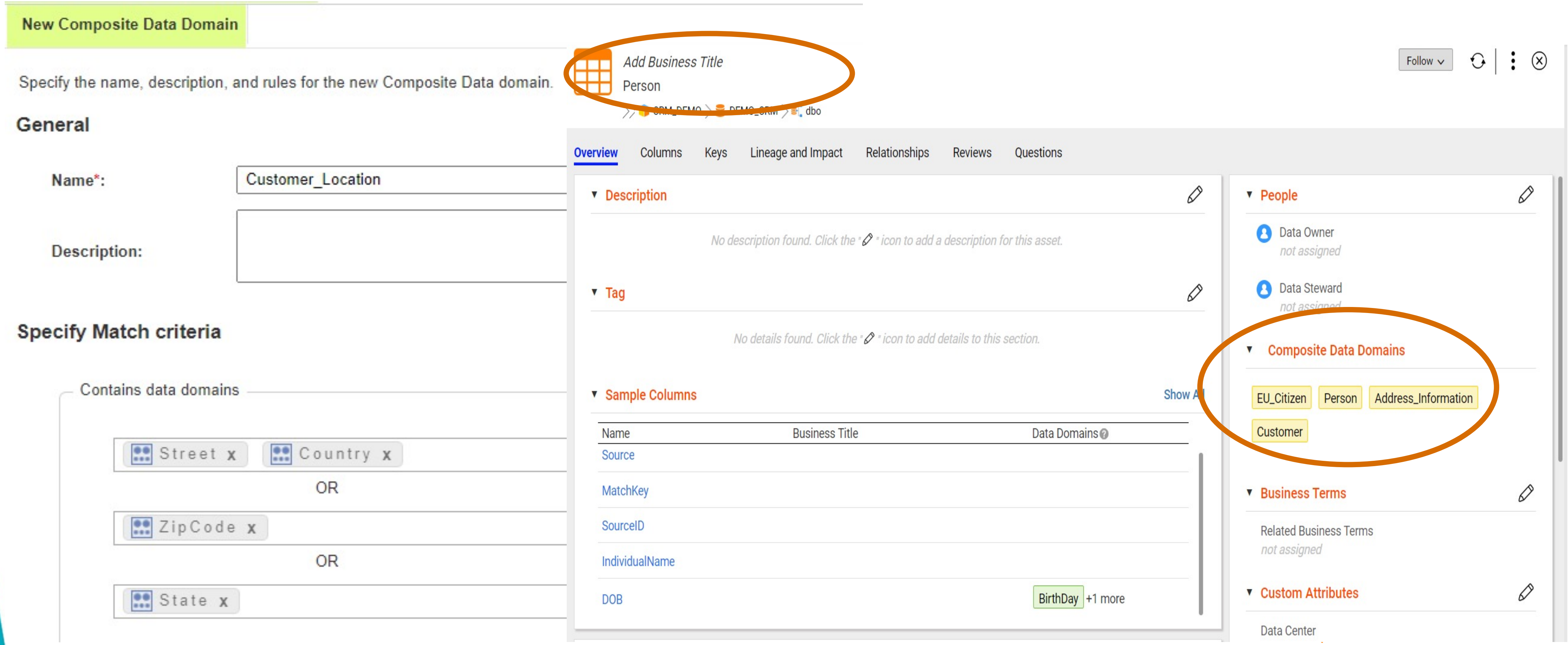
▾ Data Domains

Address_Line_1

Address

# Composite Data Domain

- A composite data domain is a collection of data domains or other composite data domains linked using rules.
- You can view composite data domains for tabular assets in the Asset Details view after you create and enable composite data domain discovery for resources in the Catalog Administrator. You can also search for composite data domains and view details of the composite data domains in the Asset Details view.



**New Composite Data Domain**

Specify the name, description, and rules for the new Composite Data domain.

**General**

Name*: Customer_Location

Description:

**Specify Match criteria**

Contains data domains

Street x Country x

OR

ZipCode x

OR

State x

*Add Business Title*

Person

CRM_DEMO > DEMO_CRM > dbo

Follow

Overview | Columns | Keys | Lineage and Impact | Relationships | Reviews | Questions

Description

*No description found. Click the "✎" icon to add a description for this asset.*

Tag

*No details found. Click the "✎" icon to add details to this section.*

Sample Columns — Show A

| Name | Business Title | Data Domains |
|---|---|---|
| Source | | |
| MatchKey | | |
| SourceID | | |
| IndividualName | | |
| DOB | | BirthDay +1 more |

People

Data Owner
*not assigned*

Data Steward
*not assigned*

Composite Data Domains

EU_Citizen | Person | Address_Information | Customer

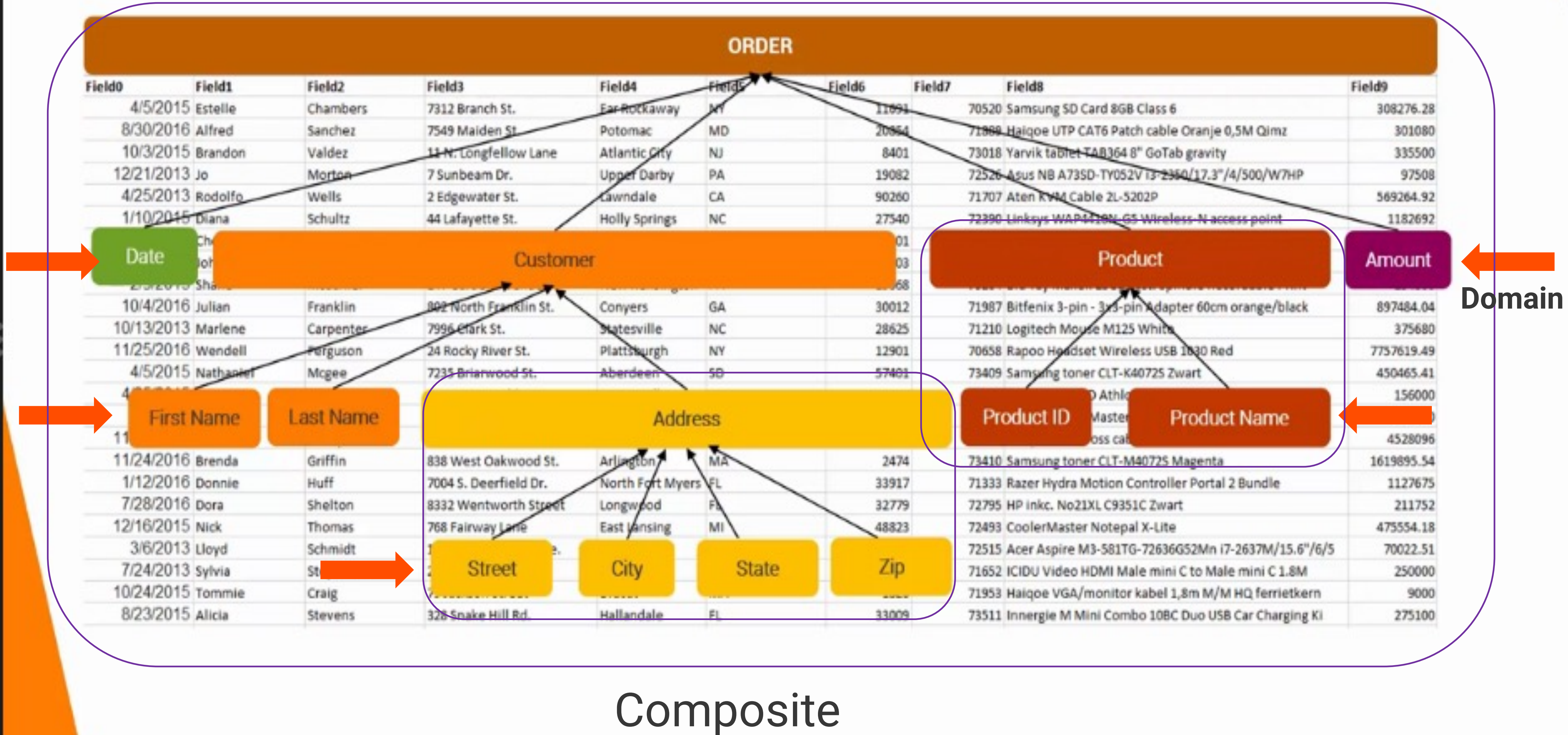
Business Terms

Related Business Terms
*not assigned*

Custom Attributes

Data Center

# Entity Recognition using Composite Domains

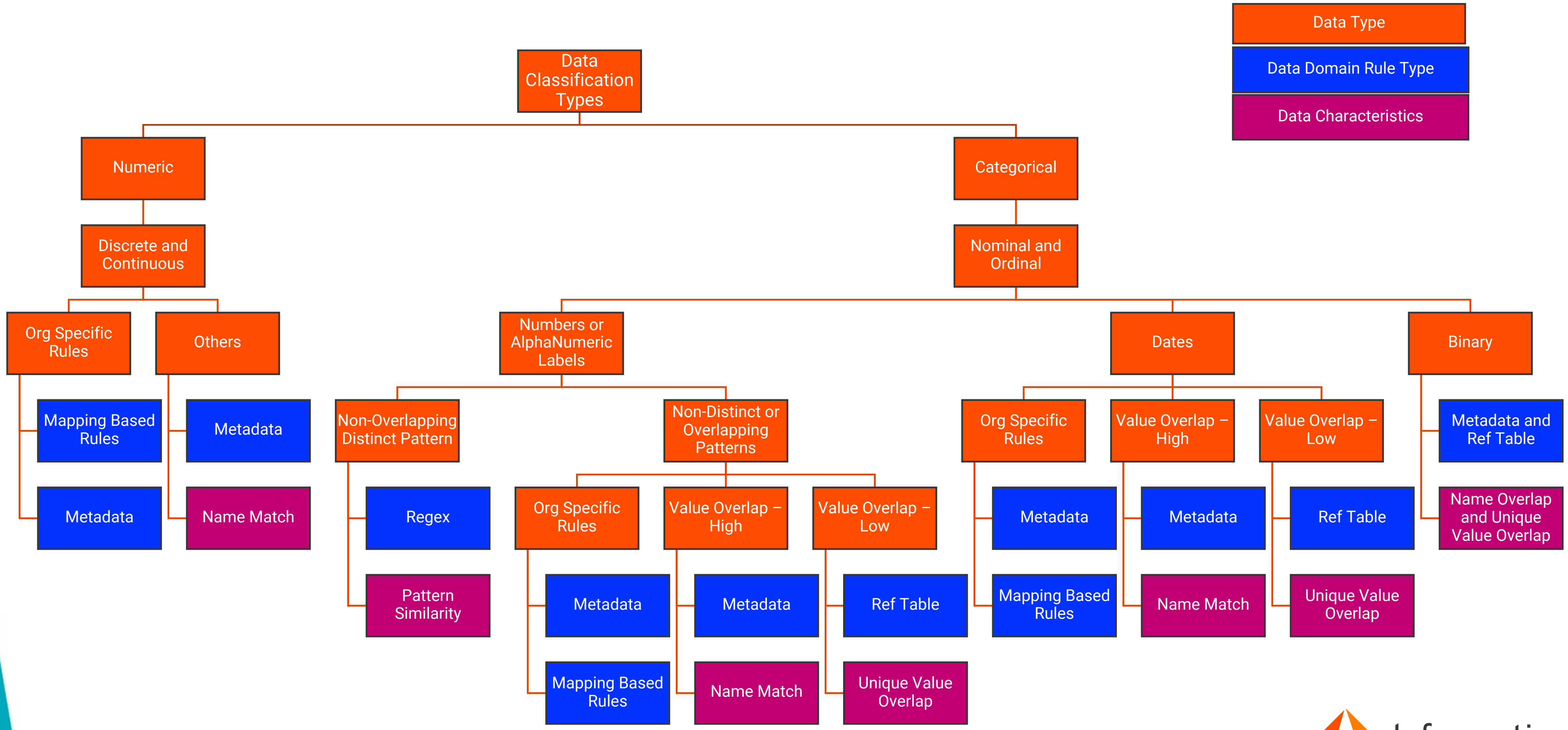


| Field0 | Field1 | Field2 | Field3 | Field4 | Field5 | Field6 | Field7 | Field8 | Field9 |
|---|---|---|---|---|---|---|---|---|---|
| 4/5/2015 | Estelle | Chambers | 7312 Branch St. | Far Rockaway | NY | 11691 | 70520 | Samsung SD Card 8GB Class 6 | 308276.28 |
| 8/30/2016 | Alfred | Sanchez | 7549 Maiden St. | Potomac | MD | 20854 | 71888 | Haiqoe UTP CAT6 Patch cable Oranje 0,5M Qimz | 301080 |
| 10/3/2015 | Brandon | Valdez | 11 N. Longfellow Lane | Atlantic City | NJ | 8401 | 73018 | Yarvik tablet TAB364 8" GoTab gravity | 335500 |
| 12/21/2013 | Jo | Morton | 7 Sunbeam Dr. | Upper Darby | PA | 19082 | 72526 | Asus NB A73SD-TY052V i3-2350/17.3"/4/500/W7HP | 97508 |
| 4/25/2013 | Rodolfo | Wells | 2 Edgewater St. | Lawndale | CA | 90260 | 71707 | Aten KVM Cable 2L-5202P | 569264.92 |
| 1/10/2015 | Diana | Schultz | 44 Lafayette St. | Holly Springs | NC | 27540 | 72390 | Linksys WAP4410N-G5 Wireless-N access point | 1182692 |
| 10/4/2016 | Julian | Franklin | 802 North Franklin St. | Conyers | GA | 30012 | 71987 | Bitfenix 3-pin - 3x3-pin Adapter 60cm orange/black | 897484.04 |
| 10/13/2013 | Marlene | Carpenter | 7996 Clark St. | Statesville | NC | 28625 | 71210 | Logitech Mouse M125 White | 375680 |
| 11/25/2016 | Wendell | Ferguson | 24 Rocky River St. | Plattsburgh | NY | 12901 | 70658 | Rapoo Headset Wireless USB 1030 Red | 7757619.49 |
| 4/5/2015 | Nathaniel | Mcgee | 7235 Briarwood St. | Aberdeen | SD | 57401 | 73409 | Samsung toner CLT-K40725 Zwart | 450465.41 |
| 11/24/2016 | Brenda | Griffin | 838 West Oakwood St. | Arlington | MA | 2474 | 73410 | Samsung toner CLT-M40725 Magenta | 1619895.54 |
| 1/12/2016 | Donnie | Huff | 7004 S. Deerfield Dr. | North Fort Myers | FL | 33917 | 71333 | Razer Hydra Motion Controller Portal 2 Bundle | 1127675 |
| 7/28/2016 | Dora | Shelton | 8332 Wentworth Street | Longwood | FL | 32779 | 72795 | HP inkc. No21XL C9351C Zwart | 211752 |
| 12/16/2015 | Nick | Thomas | 768 Fairway Lane | East Lansing | MI | 48823 | 72493 | CoolerMaster Notepal X-Lite | 475554.18 |
| 3/6/2013 | Lloyd | Schmidt | [illegible] | [illegible] | [illegible] | [illegible] | 72515 | Acer Aspire M3-581TG-72636G52Mn i7-2637M/15.6"/6/5 | 70022.51 |
| 7/24/2013 | Sylvia | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 71652 | ICIDU Video HDMI Male mini C to Male mini C 1.8M | 250000 |
| 10/24/2015 | Tommie | Craig | [illegible] | [illegible] | [illegible] | [illegible] | 71953 | Haiqoe VGA/monitor kabel 1,8m M/M HQ ferrietkern | 9000 |
| 8/23/2015 | Alicia | Stevens | 328 Snake Hill Rd. | Hallandale | FL | 33009 | 73511 | Innergie M Mini Combo 10BC Duo USB Car Charging Ki | 275100 |

# Data Domain Best Practices

Informatica®

# Understand your data / Apply right methodology



- Data Type
- Data Domain Rule Type
- Data Characteristics

Data Classification Types

- Numeric
  - Discrete and Continuous
    - Org Specific Rules
      - Mapping Based Rules
      - Metadata
    - Others
      - Metadata
      - Name Match
- Categorical
  - Nominal and Ordinal
    - Numbers or AlphaNumeric Labels
      - Non-Overlapping Distinct Pattern
        - Regex
        - Pattern Similarity
      - Non-Distinct or Overlapping Patterns
        - Org Specific Rules
          - Metadata
          - Mapping Based Rules
        - Value Overlap – High
          - Metadata
          - Name Match
        - Value Overlap – Low
          - Ref Table
          - Unique Value Overlap
    - Dates
      - Org Specific Rules
        - Metadata
        - Mapping Based Rules
      - Value Overlap – High
        - Metadata
        - Name Match
      - Value Overlap – Low
        - Ref Table
        - Unique Value Overlap
    - Binary
      - Metadata and Ref Table
      - Name Overlap and Unique Value Overlap

- Understand your data types – use the best suited method (rule) of discovery
  - For example: Credit card numbers are best identified by a regular expression
  - Another example is Part Numbers / Order Numbers created in an organization based on organization specific rules works best with Mapplet rule and Metadata Rule combination.

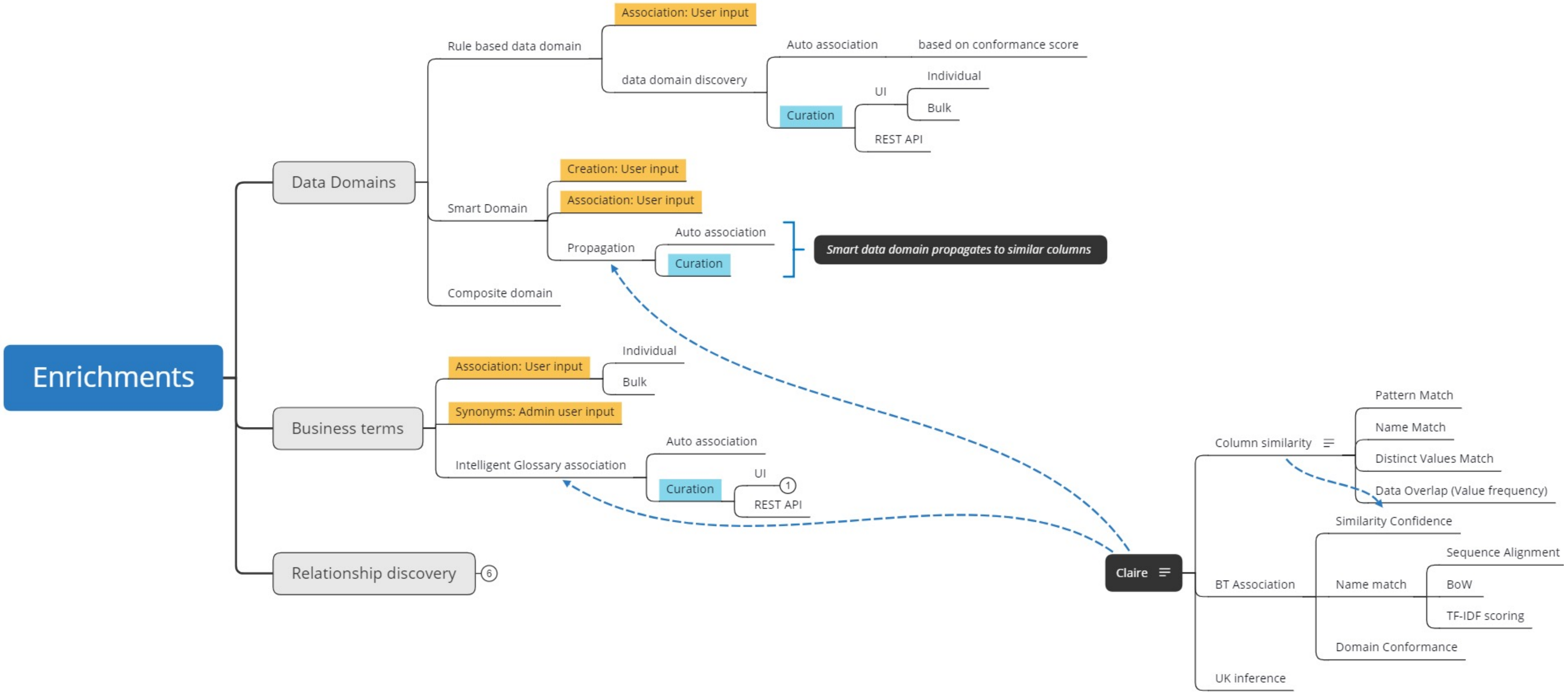
| Type | Sub Type | data type | Characteristic | Discovery methodology |
|---|---|---|---|---|
| **Pure number values** | continuous or discrete | **Numeric**<br>- Discrete<br>Ex. Order Number<br>- Continuous<br>Ex. Age / Flight status<br>No of Arrivals that gives trend | Organization specific rule | Mapplet rule & Metadata rule |
| | | | General/Others | Metadata rule |
| **Categorical** | Nominal and Ordinal | **Alphanumeric**<br>- Nominal data<br>Ex. Nationality / Gender<br>- Ordinal data<br>Ex. Rating (agree / disagree)<br>FICO Score | Unique Value overlap with other data elements = None | Regex rule |
| | | | Unique Value overlap with other data elements = High | Metadata rule |
| | | | Unique Value overlap with other data elements = Low | Reference table rule |
| | | | Organization specific rule | Mapplet rules & Metadata rule |
| | | **Dates**<br>Ex. Date of Birth<br>Order Date<br>Purchase Date | Unique Value overlap with other data elements = High | Metadata rule |
| | | | Unique Value overlap with other data elements = Low | Reference table rule |
| | | | Organization specific rule | Mapplet rule & Metadata rule |
| | | **Binary**<br>Ex. Flags ( Y / N )<br>True / False<br>0 / 1 | Not Applicable | Metadata rule & Reference table rule |

# Understand Column Similarity and its role in smart data domain propagation

- Manual curation of labels or tags will take forever and tedious. Identification of, therefore, similar columns in an automated fashion becomes important. One column gets a tag(s) and that gets propagated to the similar columns. This will save a lot of curation effort.

- In EDC, column similarity is CLAIRE functionality, and it is based on unsupervised clustering and computed based on 4 factors:
  - Name similarity
    - ✓ Match fields based on names and descriptions and identify fields even if abbreviated
    - ✓ CNTRY-CD and COUNTRY_CODE
  - Pattern similarity
    - ✓ Data patterns identified from the profiling results will be used to find similar columns and propagate smart data domains
  - Value frequency similarity (aka data similarity)
  - Unique value similarity

# Understanding CLAIRE in Data Domain Propagation

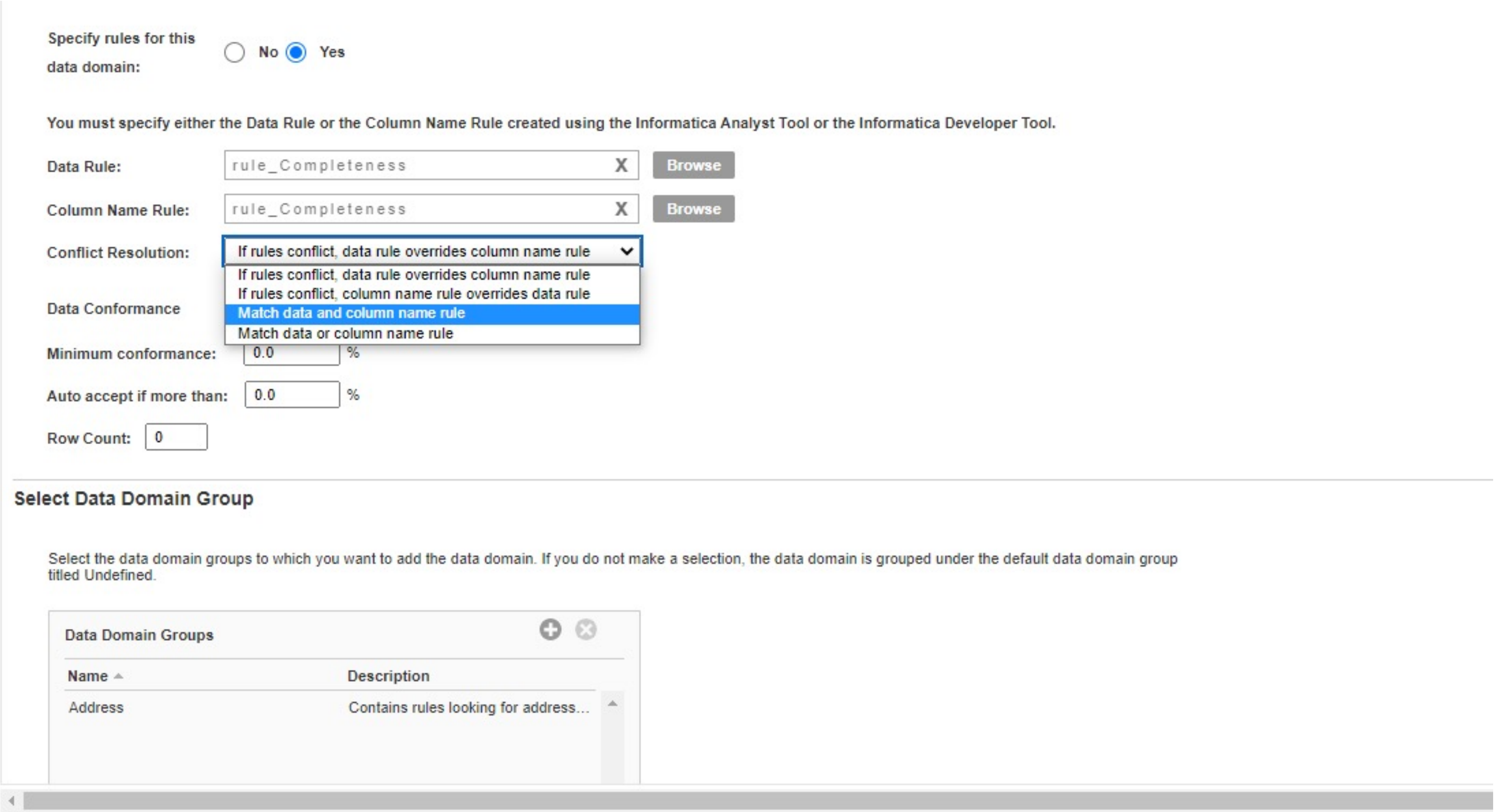


*Smart data domain propagates to similar columns*

# Best Practices continued…

- Set Profiling sampling size to the max recommended value i.e., 17K for better results. Profiling and Data domain discovery is resource intensive, we recommend to do sizing and tuning of your EDC environment.
- Do data domain discovery against test data / non prod data only if your use case demands, expect false positives for data domain discovery for test/non prod data.
- Use what you need
  - Instead of "All Data domains", choose what is needed. Specific groups or data domains
  - It simply adds to time and curation effort.
- Use "bulk curation" creatively and wherever applicable
  - Data lands in Zone A – EDC scans Zone A and data domains are curated.
  - Zone A data moves to Zone B encrypted. Zone B's objects are same as "Zone A"
  - One can simply curate Zone B using the bulk import
- Override rules for data domains
  - Use conflict resolution option to minimize false positives
  - Use Proximity Data Domains for tie brakers

# Conflict resolution / Override rules



Specify rules for this data domain: ○ No ● Yes

You must specify either the Data Rule or the Column Name Rule created using the Informatica Analyst Tool or the Informatica Developer Tool.

Data Rule: rule_Completeness X Browse

Column Name Rule: rule_Completeness X Browse

Conflict Resolution: If rules conflict, data rule overrides column name rule

- If rules conflict, data rule overrides column name rule
- If rules conflict, column name rule overrides data rule
- Match data and column name rule
- Match data or column name rule

Data Conformance

Minimum conformance: 0.0 %

Auto accept if more than: 0.0 %

Row Count: 0

**Select Data Domain Group**

Select the data domain groups to which you want to add the data domain. If you do not make a selection, the data domain is grouped under the default data domain group titled Undefined.

Data Domain Groups

| Name ▲ | Description |
|---|---|
| Address | Contains rules looking for address... |

Informatica

# Proximity Data Domains

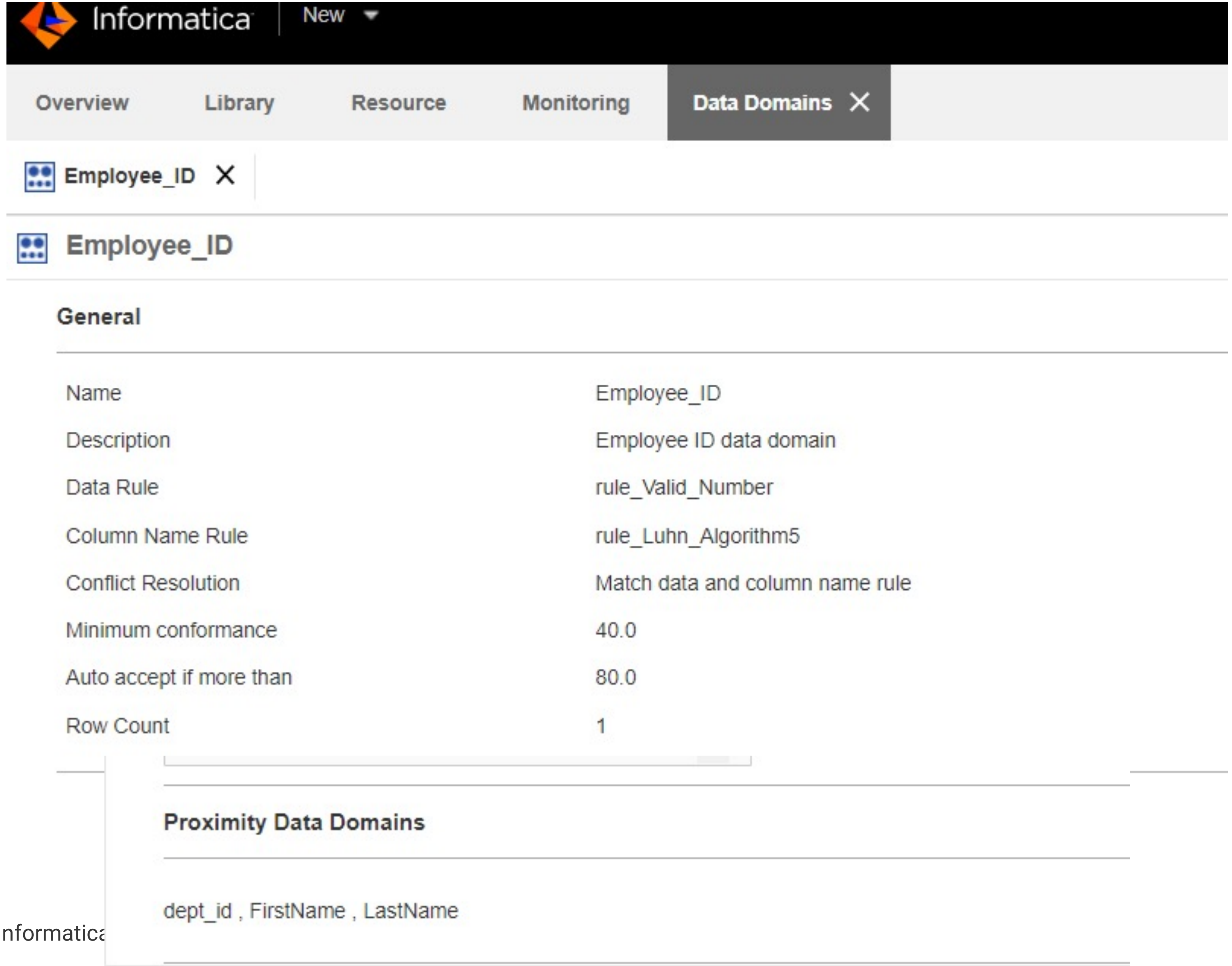
## Tie-breaker

- EDC uses proximity data domains to narrow down the inferred results to identify close-to-identical columns or fields for a data domain.
- EDC displays the results as a match score for the data domain. The match score is the ratio of number proximal data domains discovered in the data source to the number of configured proximal data domains for an inferred data domain.
- To use proximity data domains in the data domain discovery process, perform the following tasks
  - When you create or edit a data domain, add one or more data domains as proximity data domains.
  - When you create or edit a resource and enable data domain discovery, add the proximity data domains to the data domain.
  - When you enable data discovery and run the resource, the profiling scanner scans the data source for the data domain and the proximity data domains in the resource and displays a match score in Enterprise Data Catalog

# Proximity Domain - Example

- Proximity Domains are one of the factors that help determine the data type of a column. They are useful as a tie-breaker when all the other factors result in inference of two or more domains with equal conformance/probability.
- Imagine three data domains: EMPLOYEE ID, NODE ID and CUSTOMER ID. All three are 7 digit numbers that are each picked randomly from a distribution of all 7 digit numbers.
- Now imagine a file called “data.csv”. It has the following columns:
- Col1 – All 7 digit numbers
- Col2 – First Names
- Col3- Last Names
- Col4- Department Names

# Proximity Domain - configuration



## Employee_ID

### General

| | |
|---|---|
| Name | Employee_ID |
| Description | Employee ID data domain |
| Data Rule | rule_Valid_Number |
| Column Name Rule | rule_Luhn_Algorithm5 |
| Conflict Resolution | Match data and column name rule |
| Minimum conformance | 40.0 |
| Auto accept if more than | 80.0 |
| Row Count | 1 |

### Proximity Data Domains

dept_id , FirstName , LastName

# Proximity Domain – Example

- Intuitively we can recognize that data.csv may contain EMPLOYEE information and Col1 is EMPLOYEE ID from the description above. How did we come to that conclusion? It is because of the presence of:
- Col2 and Col3: Contains data about people – either employees, customers, shareholders etc. Certainly not nodes, as they do not have first and last names.
- Col4 – Department Name which narrows it down to employees. Customers and Shareholders generally may not have department names assigned to them.

- Proximity domains are a mechanism to bring the above intuition into the discovery process. While defining EMPLOYEE ID, users indicate that it generally occurs with FIRST NAMES, LAST NAMES and DEPARTMENT NAMES. The last three are also data domains. During discovery, the system associates Col1 with all the three domains EMPLOYEE ID, NODE ID and CUSTOMER ID because of the equal conformance score. However because of the identified proximity domains defined it will nudge the user to mark it as an EMPLOYEE ID because it found all the proximal domains for it.

# General information about custom data domain creation

## Data Domain Sync between EDC and IDQ

- When custom data domains are created in Analyst or Developer tool and not in Catalog administrator, make sure to run the DataDomain internal resource in Catalog Administrator to sync data domains from the Model Repository Service (MRS) with the Catalog. This is valid when same MRS is used for EDC and DQ.
- If MRS for DQ is separate from MRS for EDC, then export the data domains from DQ MRS and import the data domains in EDC MRS to make sure that the custom data domains are available for use in EDC.

## Options to create rules and using them in custom data domains

- Use Informatica ***Developer Tool*** to create Rules which can be associated with custom data domains in Analyst, Developer or Catalog Administrator tool.
- The business user can create rule specifications in the ***Analyst tool*** and then generate the rule to save as mapplet in the Developer tool. The mapplet can be validated as a Rule in the Developer tool which will then be available as a reusable rule for Data Domains.
- In the ***Catalog administrator***, custom data domains can be created either by selecting predefined rules created in Developer, or using RegEx, or using Reference table.
- If the user is a business user and has access only to the Analyst tool, then  the user can use pre-defined rules from the Developer tool which are available as ***reusable rules*** in the Analyst tool to create custom data domains.
- To modify any rule logic for the data domains, use the Developer tool to edit the mapplet for the rule.

# References

**EDC Data domain FAQ**

- https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1230-using-data-domains-in-enterprise-data-catalog/using-data-domains-in-enterprise-data-catalog.html
- https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1230-using-data-domains-in-enterprise-data-catalog/appendix/faq.html

**Composite Data Domain Article**

- https://knowledge.informatica.com/s/article/Usage-of-Composite-Data-Domains-in-EDC?language=en_US&type=external

**EDC Data domain article for quick reference**

- https://knowledge.informatica.com/s/article/Data-Domain-Discovery-in-EDC?language=en_US&type=external

**EDC Performance guide**

- https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1505-tuning-enterprise-data-catalog-performance-in-10-4-1/abstract.html

**EDC sizing guide**

- https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1505-tuning-enterprise-data-catalog-performance-in-10-4-1/tuning-enterprise-data-catalog-performance-in-10-4-1/enterprise-data-catalog-sizing-recommendations.html

# Q & A

Informatica

# Thank you

Hari Krishna Akula | hariakula@informatica.com

Customer Success Technology (CST) team

dl_cst@informatica.com

Informatica®