

Enterprise Data Catalog

Hands on Lab

EDC Environment Information

Using Google Chrome Browser on your assigned jumpbox:

EDC UI:

User Name: edc-user

Password: pocInfaAdmin@2018

Lesson 1: Data Discovery

Duration: 10 Minutes

Overview and Objective

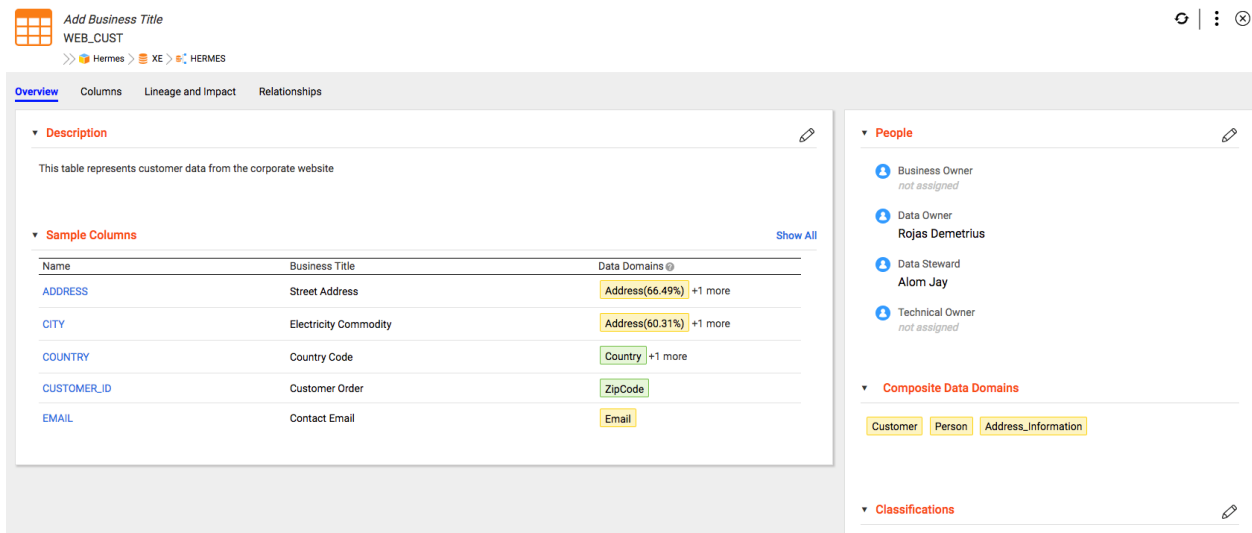
In this lesson, you will learn how to search for relevant data assets using Search and Dynamic Faceting capabilities in the Catalog. You will also learn to explore associated Data Profiling statistics to determine the quality of the assets. Finally, you will understand how to explore data asset relationships.

Lab 1: Search and Faceting

You can search for the required assets in the catalog. You can search based on the name of the asset or perform a generic search using the wildcard character asterisk (*) or the wildcard character question mark (?). Enterprise Data Catalog also displays probable matches when you type the name of a required asset.

From the search results displayed, you can sort the results based on the asset name, the relevance, or the system attributes or the custom attributes. You can use the search filters displayed to filter the search results and view additional details for the displayed assets. After searching and finding the required asset, you can annotate and enrich the required assets with custom attributes.

1. Log in to Enterprise Data Catalog using the following credentials: Username: edc-user / Password: pocInfAdmin@2018
2. Search for “Web Customers”
3. Filter search results by “Asset Type: Table”
4. Click on “WEB_CUST”. This will take you to the overview page of the asset. You can see the description and, if you have the permission, you can edit the description and the custom attributes such as people, domains, classification. You will also notice sample columns of the table.



The screenshot shows the 'Overview' page for the asset 'WEB_CUST'. The page includes a description, sample columns, people, composite data domains, and classifications.

Add Business Title
WEB_CUST

>> Hermes > XE > HERMES

Overview Columns Lineage and Impact Relationships

Description

This table represents customer data from the corporate website

Sample Columns [Show All](#)

Name	Business Title	Data Domains @
ADDRESS	Street Address	Address(66.49%) +1 more
CITY	Electricity Commodity	Address(60.31%) +1 more
COUNTRY	Country Code	Country +1 more
CUSTOMER_ID	Customer Order	ZipCode
EMAIL	Contact Email	Email

People

- Business Owner *not assigned*
- Data Owner **Rojas Demetrius**
- Data Steward **Alom Jay**
- Technical Owner *not assigned*

Composite Data Domains

Customer Person Address_Information

Classifications

Lab 2: Data Profiling Statistics

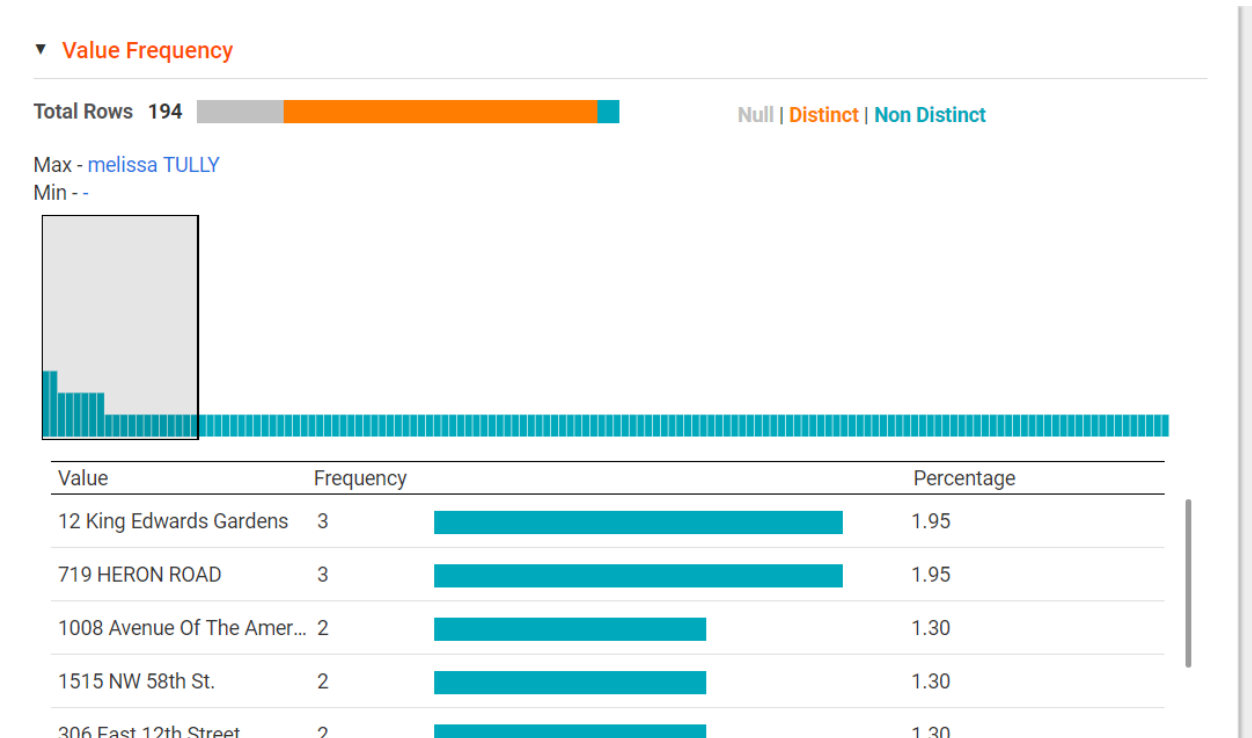
Continuing from above, alongside the overview, you can view column metadata and explore assets to verify the quality of data, such as profiling information.

The Columns section displays the following details for each column in the asset:

- Column name
 - Business Title, which allows you to assign business glossary resources to the asset
 - Data domains that a user has assigned to the column, that are inferred from profile results, or that are inferred from similar assets.
 - The percentage of null, distinct, and non-distinct values that are calculated from profile results.
 - The source data type if a data type is defined in the data source.
 - Data types that are inferred from profile results.
1. In this view, you can understand percentages of NULL, DISTINCT and NON_DISTINCT values in each column.

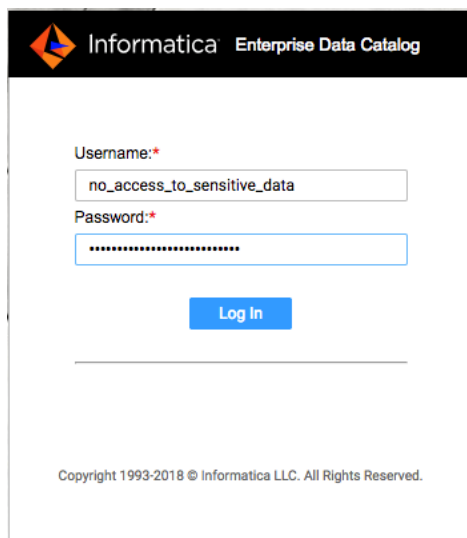
Overview		Columns	Lineage and Impact	Relationships	
Name	Business Title	Data Domains	Null Distinct Non-Distinct %	Source Data Type	Inferred Data Types
1 ADDRESS	Street Address	Address(66.49%) +1 more	20.61 74.22 5.17	VARCHAR2 (60)	String(51) 100.00%
2 CITY	Electricity Commodity	Address(60.31%) +1 more	1.54 56.70 41.76	VARCHAR2 (30)	String(22) 100.00%
3 COUNTRY	Country Code	Country +1 more	8.24 3.60 88.16	VARCHAR2 (20)	Fixed Length String(2) 91.01% String(13) 10...
4 CUSTOMER_ID	Customer Order	ZipCode Data Match : 65.98%	00 0	NUMBER (22)	Date(yymmdd) 1.54% +4 more
5 EMAIL	Contact Email	Email	47.42 51.54 1.04	VARCHAR2 (100)	String(30) 100.00%
6 NAME	Last Name	FullName(59.28%)	2.06 94.32 3.62	VARCHAR2 (60)	String(35) 100.00%
7 PHONE		PhoneNumber	17.01 78.35 4.64	VARCHAR2 (30)	Decimal(10) 55.27% +2 more
8 POSTALCODE	Postal Code	Postcode +1 more	27.83 47.93 24.24	VARCHAR2 (20)	Date(yydd) 38.57% +4 more
9 STATE	Stake	Address(47.42%) +2 more	1.03 28.86 70.11	VARCHAR2 (20)	Decimal(3) 11.45% +2 more

2. Within the column view you can get detailed information about Column Value Distribution, Data Patterns, Inferred Data Types and Auto inferred Data Domains.
3. Drill down into any column, for ex: "ADDRESS" to get more details about the column.

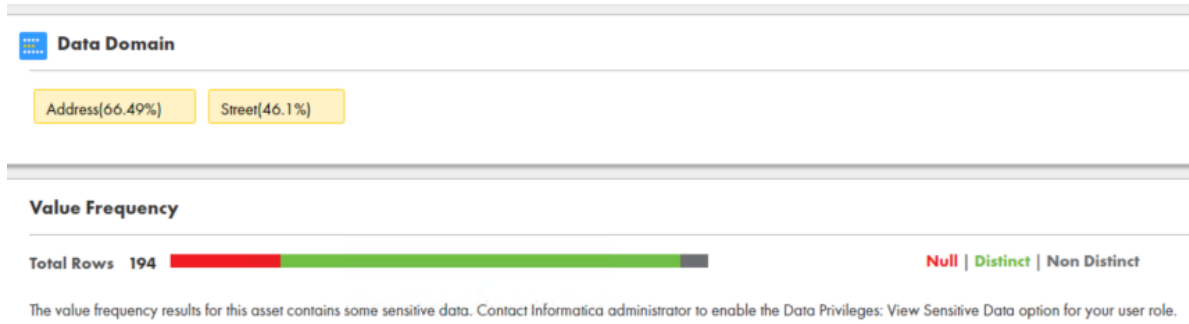


As you can see, we can view data content which may contain sensitive content. Access to data content can be restricted via user privileges. The user you are logged with has privileges to data content. We have created another user without privileges to data content.

Logout from Catalog and connect with user "no_access_to_sensitive_data" with password "no_access_to_sensitive_data".



Search for WEB_CUST, open it, and open attribute ADDRESS.



You get a clear message informing you have no access to value frequencies for this attribute as it may contain sensitive data.

Lesson 2: Data Domain Curation

Duration: 10 Minutes


A data domain is a predefined or user-defined asset based on data values or a column or field name. Some examples of data domains are Social Security number, account status, IP address, and UPC code. Assigning a data domain to a column or field makes the asset easier to identify and understand.

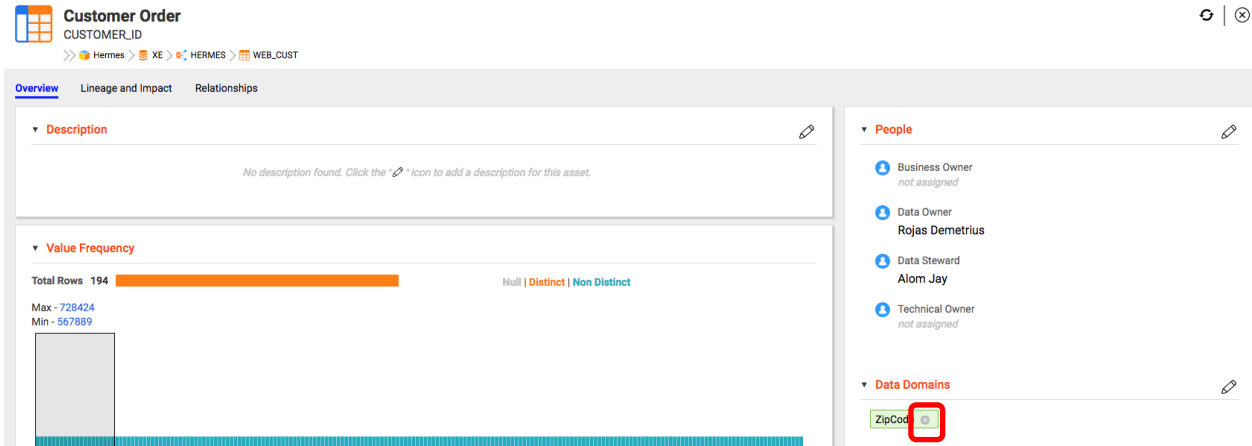
You can organize data domains that apply to similar types of data in data domain groups. For example, the Bank Account data domain group might contain data domains such as Account Status, Account Number, and Credit Card Number.

Enterprise Data Catalog infers data domains for columns and fields based on profile results. It also infers data domains for assets based on the data domains that are assigned to similar assets. You can accept and reject assets for a data domain.

When you curate a data domain, you make the data domain more accurately reflect the type of data that belongs in the data domain. Curating a data domain also makes Enterprise Data Catalog more accurate when it infers data domains for similar assets.

Lab 1: Data Domain Curation


1. Log back in as user **edc-user** with password **poclnfaAdmin@2018**
2. Search WEB_CUST and open it. Note that CUSTOMER_ID has been assigned to ZipCode, which is wrong. We would like to curate it and allocate it to the right domain.
3. Click CUSTOMER_ID column to open it
4. Under Data Domains, Reject ZipCode data domain (un-associate) by a click on 




Customer Order
CUSTOMER_ID

Overview | Lineage and Impact | Relationships

Description

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

Value Frequency


Total Rows 194  Null | Distinct | Non Distinct


Max - 728424
Min - 567889

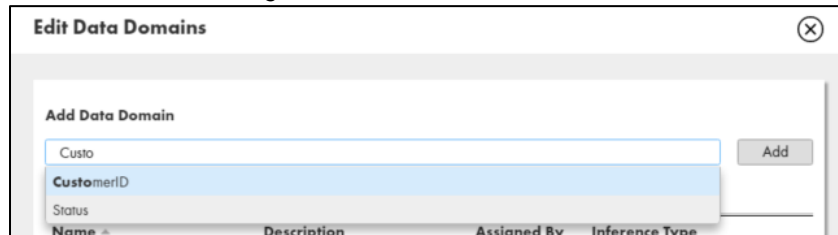
People

- Business Owner *not assigned*
- Data Owner **Rojas Demetrius**
- Data Steward **Alom Jay**
- Technical Owner *not assigned*

Data Domains

- ZipCode 

- Let's assign now the right domain. To assign or change the Data Domain assignments, click the Data Domain section Edit Properties icon ().
- When the Edit Data Domain dialog box appears, type "Custo", select "CustomerID" & click "Add". Now the column has been assigned the address data domain.



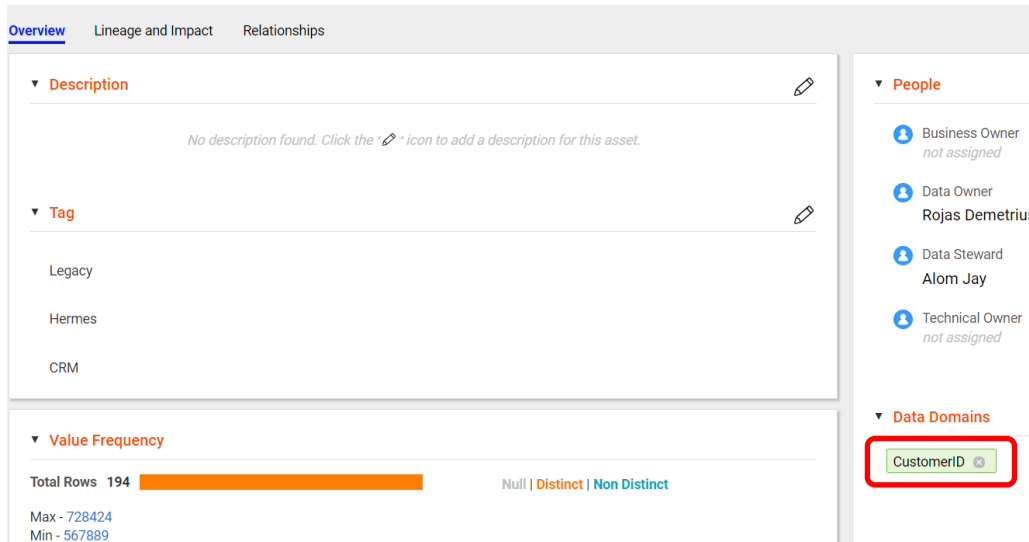
Edit Data Domains

Add Data Domain

Custo

Name	Description	Assigned By	Inference Type
CustomerID			
Status			


- Click "ok"
- Now you will notice "CustomerID" is assigned as a data domain to CUSTOMER_ID



Customer Order
CUSTOMER_ID

Overview | Lineage and Impact | Relationships

Description

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


Tag

Legacy

Hermes

CRM

Value Frequency


Total Rows 194  Null | Distinct | Non Distinct

Max - 728424
Min - 567889

People

- Business Owner *not assigned*
- Data Owner **Rojas Demetrius**
- Data Steward **Alom Jay**
- Technical Owner *not assigned*

Data Domains

- CustomerID 

Lesson 3: Lineage and Impact Analysis

Overview and Objective

In this lesson, you will learn how to use the new drill down lineage views in the Catalog to visualize data provenance. You will also learn how to use the detailed impact analysis reports in the catalog to understand impact due to change in data assets or ETL flows.

Lineage and impact describes the end-to-end data flow of data for an asset. The data flow for an asset has two components, the lineage and the impact.

Lineage describes the flow of data from the origins to an asset. Lineage shows you where the data for an asset comes from and which assets affect the asset that you are studying. When you view an asset in a lineage and impact diagram, the lineage includes the asset that you are viewing and all of the upstream assets in the data flow.

Impact describes the flow of data from an asset to the destinations. Impact shows you where the data is used and which assets might be affected if you change the asset that you are studying. When you view an asset in a lineage and impact diagram, the impact includes the asset that you are viewing and all of the downstream assets in the data flow.

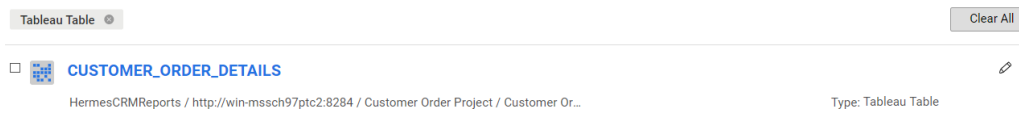
Objectives

- Understand Drill Down Lineage Views in the Catalog
- Perform Impact Analysis on Data Assets
- Export Lineage Views and Impact Analysis Reports

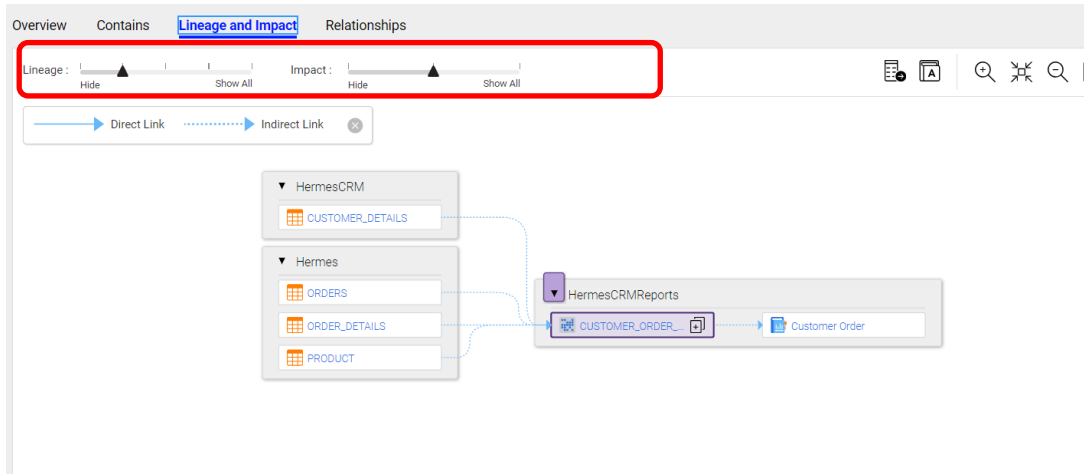
Duration: 10 Minutes

Lab 1: Drill Down Lineage

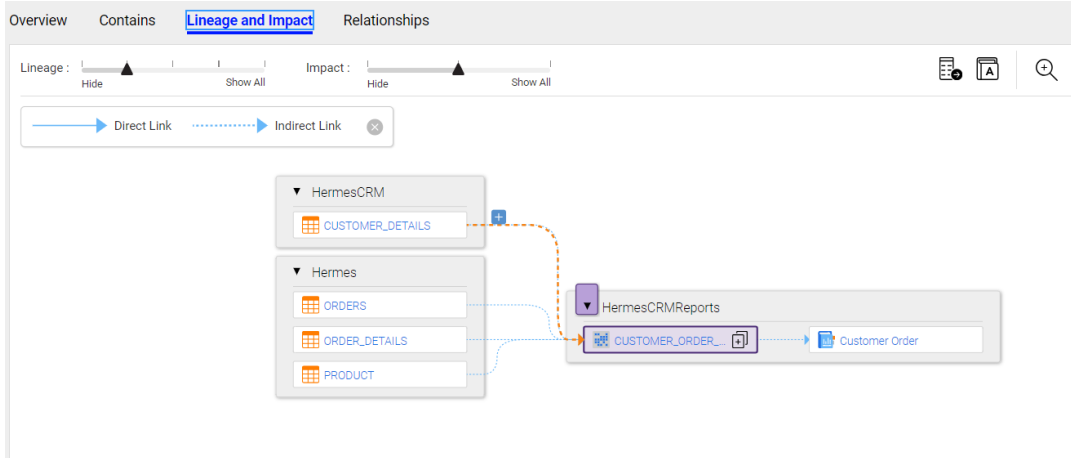
1. Search for “customer order”
2. Filter by Asset Type: Tableau Table
 - a. If it doesn't show in the list, click on “Show All”, search for the asset type “Tableau Table”
3. Open Customer_Order_Details report.




4. Click on the lineage and impact tab.
5. Use the Lineage Sliders to add all levels in the lineage diagram.

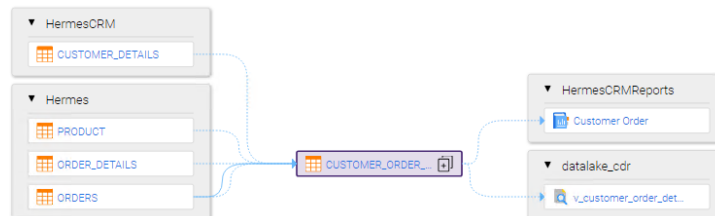


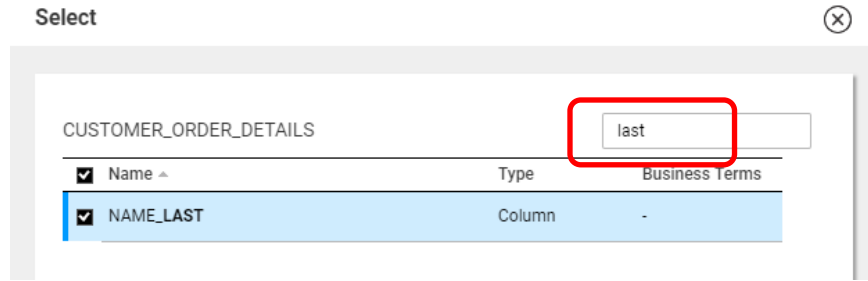
- Alternatively, you can also expand individual dotted links by clicking on the (+) icon on the link to only expand a path.



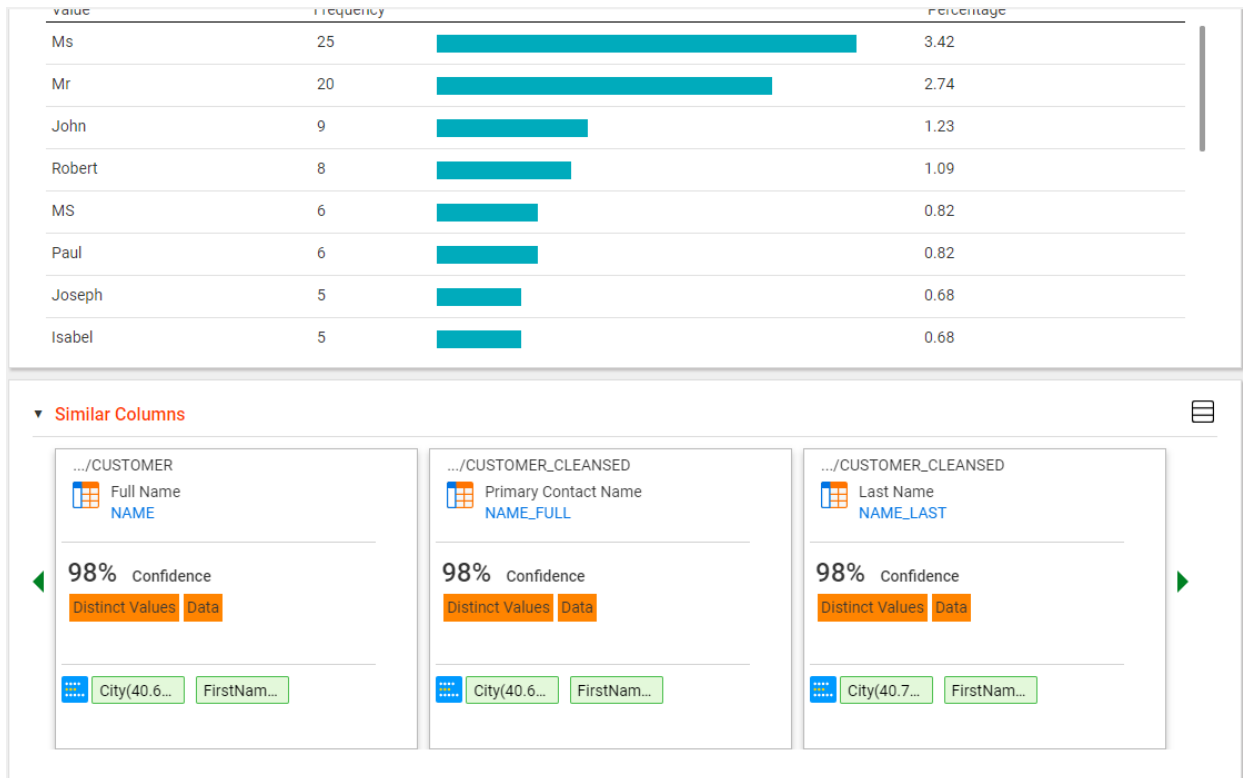
- Click on icon  to display business terms in the lineage diagram.

- In the middle of the lineage diagram, you will see customer order detail (you can hover over it). Click the "Customer_Order_Details" table to see lineage for the asset. Now click on the plus icon on "Customer Order Detail" and search for "Last", then select the "NAME_LAST" column. Click on "OK". This step will expand all the columns in the tables in the lineage diagram that directly affect this metric





9. Click (X) to view the lineage diagram. Notice that the lineage diagram shows the column name NAME_LAST from upstream to downstream applications.
10. Click on LASTNAME column in the resource HermesCRM This will take you to the Lineage and Impact diagram of the LASTNAME column
11. Click the "Overview" tab to view value frequencies as well as a new feature, similar columns.




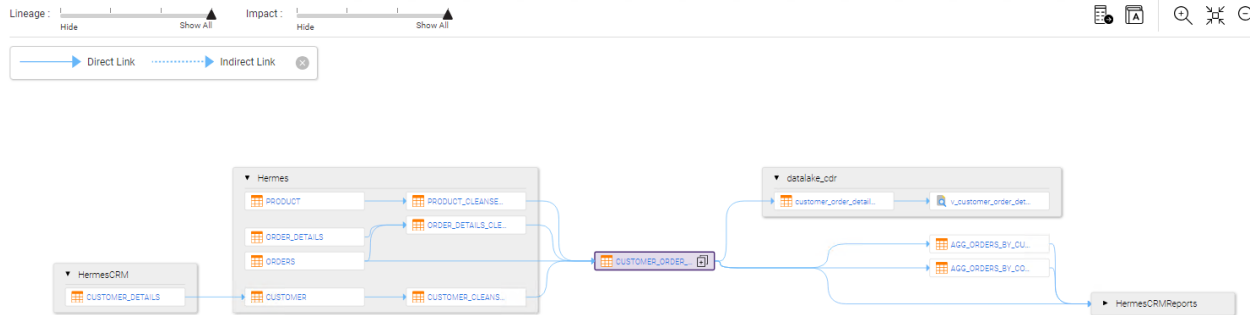
In an organization, a column name such as Last Name might exist across multiple data sources. To identify the data sources that contain such columns, you can use column similarity in Enterprise Data Catalog. It uses unsupervised clustering which is a machine learning technique to identify the similar columns. Enterprise Data Catalog performs unsupervised clustering across multiple data sources based on several factors, such as data overlap, distinct value match, pattern match, and name match. It then assigns an overall similarity score as well as the match likelihood for each factor.

You can click on any of the columns to view metadata associated with it.

12. Go back to "CUSTOMER_ORDER_DETAILS" asset.

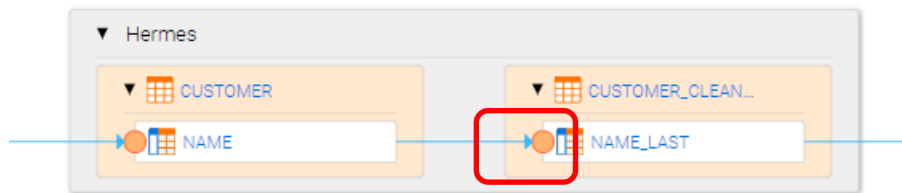
13. Expand the lineage to "show all"

14. Click on icon  to show transformation details, and extend lineage to get full lineage as below

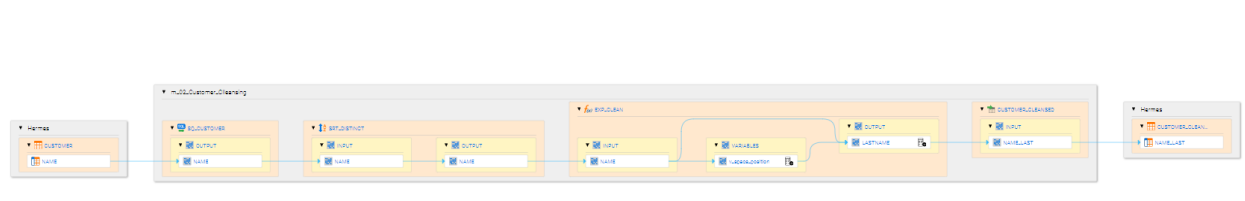



Each orange bubble represents detailed transformation available coming from PowerCenter, Informatica Cloud, Informatica Big Data Management or even Cloudera Navigator.

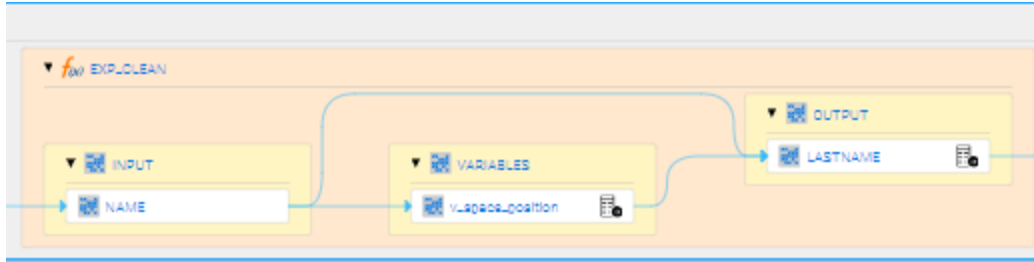
Click on orange bubble in front of table Hermes/CUSTOMER_CLEANSSED to get more details on transformations used to load this table.



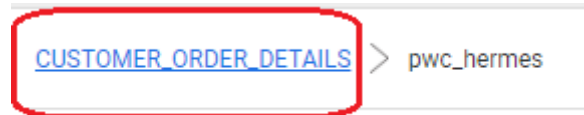
It shows mapping details



The icon  indicates that a transformation has been applied on the attributes. Move your mouse over or click on it to get more details on functions used



Click on "CUSTOMER_ORDER_DETAILS" to leave detailed transformation.



Lab 2: Understanding Relationships

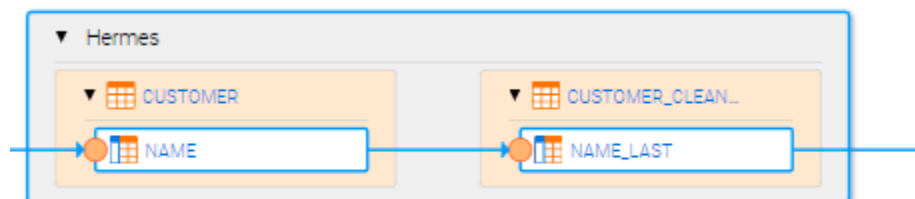
Relationships describe the relationships between the selected asset and other assets in Enterprise Data Catalog. The view displays the asset relationships in a graphical form.

When you open an asset in the Relationships view, the selected asset appears at the center of the Relationships view, and the related assets appear around the selected asset.

The Relationships view shows different circles that represent a specific asset or a group of assets. By default, the selected asset is highlighted in blue and the related assets of the same type are represented as small icons within the asset type circle.

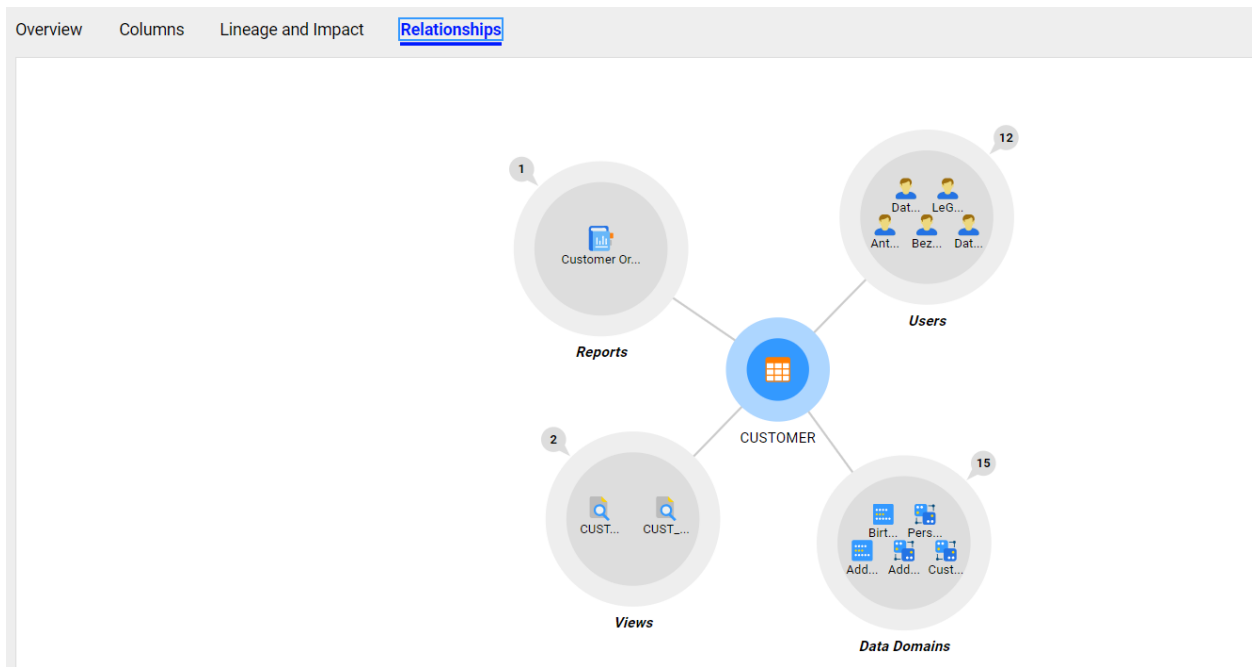
The assets that you see in the Relationships view vary based on the selected asset type. For example, if you select a table, the Relationships view displays related assets such as data domains, business term, reports, and synonyms.

1. Continuing from the previous lab, close the impact summary to return to the lineage diagram of the "CUSTOMER_ORDER_DETAILS" report.
2. Click on "Hermes / CUSTOMER" table, which takes you the lineage diagram of the table. Optionally, you can click on asset details, to view metadata & statistics, of the table.



3. Click on relationships

The Relationship tab displays the relationship in a diagram that shows how the selected data asset is related to other data assets. You can see how Hermes Customer table is associated to other views, data domains, reports and users.



1. Click on the Number callouts to expand or minimize the relationship bubbles.
2. Expand the data domains relationships to see more details about the relationships.
3. Hover on any data domain icon and click the domain name for it to pop-up.

Lesson 4: Data Classification

Duration: 10 Minutes

Overview

In this lesson, you will learn how the Catalog automatically classifies data based on known domains. You will also learn how you can annotate datasets to further classify data assets along multiple dimensions.

Objectives

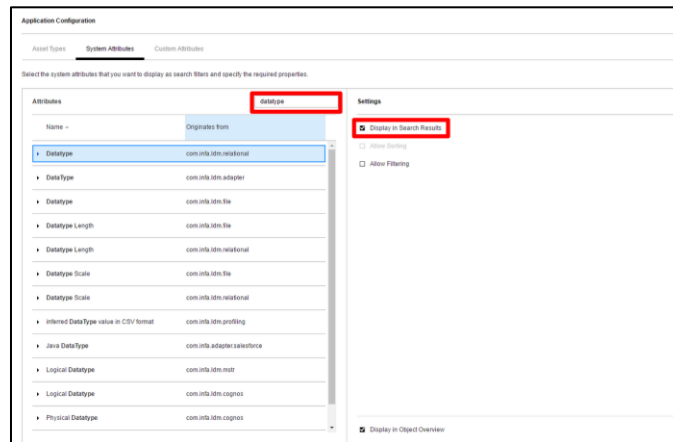
- Search Classified Columns
- Domain Overview
- Annotate Data Assets

Lab 1: Search Configuration

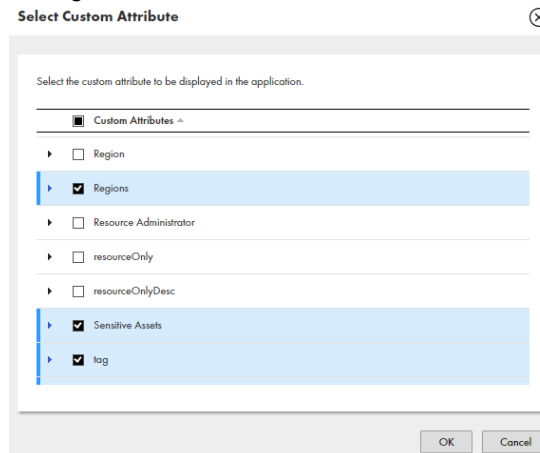
1. Click the Settings icon, then select Application Configuration on the top right bar.



2. The Application Configuration dialog box appears. Navigate to the “System Attributes” tab
3. Search for “Datatype”.
4. For Datatype that originates from relational objects(“com.infa.ldm.relational”), enable “Display in Search Results”



5. Navigate to the “Custom Attributes” tab and click “Add”
6. Select the following custom attributes from the list: “Application Source”, “Data Migration to Azure”, “Sensitive Assets”, “Tag”, then click on OK.



7. Enable “Allow Searching”, “Allow Filtering” and “Display in Search Results” for the custom attribute “Application Source” and “Data Migration to Azure”.

Select the custom attributes that you want to display as search filters and specify the required properties.

Attributes Add

Name ^
▶ Application Source
▶ Asset Links
▶ Business Unit
▶ Business Description
▶ Business Owner
▶ Data Center
▶ Data Owner

Settings

- Display in Search Results
- Allow Searching
- Search Rank Medium ▼
- Allow Sorting
- Allow Filtering
- Allow Editing

8. Click Save and Refresh the browser.

Note: it may take a few minutes for the attributes to appear for objects.

Lab 2: Classifying data for migration to Azure DWH.

You can annotate and enrich the required assets with custom attributes. Enriching the assets with the attributes makes the data asset easily discoverable. In this case, you are enriching the asset with the attributes to make it easily discoverable data for migration to Azure DWH.

1. Search for search for “staging customer data from oracle data warehouse”
2. Under the asset overview page, scroll to down to the Custom Attributes pane, click on the Edit Icon (Pencil Icon).

▼ **Custom Attributes** ✎

Application Source
not assigned

Asset Links
not assigned

Business Unit
not assigned

Business Description
staging customer data from oracle data warehouse

Data Center
not assigned

Data Migration to Azure
not assigned

3. Add a new value for “**Application Source**” - “EDW”
4. Add another value for “Data Center” – “On Prem”
5. Add Data Migration to Azure – “Yes”
(note: if you don’t see the Data Migration to Azure attribute, please wait a few minutes)
6. Click the check mark next to custom attribute tag to save the values.
7. You should have the following custom attribute set:

Application Source
EDW

Asset Links
not assigned

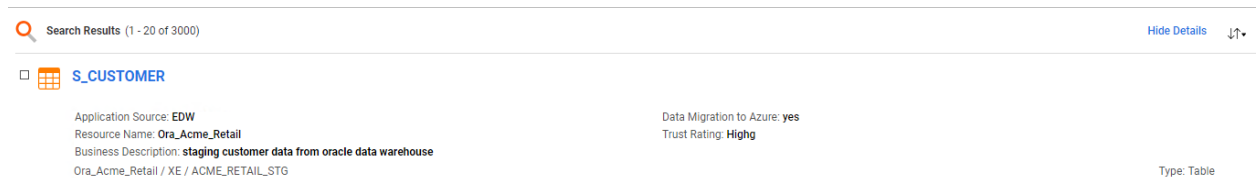
Bunissess Unit
not assigned

Business Description
staging customer data from oracle data warehouse

Data Center
ON-PREM

Data Migration to Azure
yes

8. In the search bar, type: “staging customer data from oracle data warehouse”
9. Click on “show details”. Notice, the results, based on the classification you identified:



Search Results (1 - 20 of 3000) Hide Details

S_CUSTOMER

Application Source: EDW	Data Migration to Azure: yes
Resource Name: Ora_Acme_Retail	Trust Rating: High
Business Description: staging customer data from oracle data warehouse	
Ora_Acme_Retail / XE / ACME_RETAIL_STG	

Type: Table

Note, in most cases, you will have thousands of tables to migrate. To efficiently tag and identify the assets for data migration, Enterprise Data Catalog allows the use of rest api to update bulk objects and retrieve objects with associated tags. (Please note, we will not have a hands-on lab on Rest API, however, if you have questions please reach out to the instructor).

Lesson 5: Understanding Lineage from Azure SQL DW (optional)

Overview and Objective

In this optional lesson, you will learn how to use the new drill down lineage views in the Catalog to visualize Azure ecosystem objects. You will use the same techniques in lesson #3, “Lineage and Impact Analysis”. As we are in a lab environment, we have already ingested metadata from Azure & Informatica Cloud Data Integration. (In the next lab, you will learn how to move data to cloud using Intelligent Cloud Services).

1. In EDC, search for “retail store dimension table”
2. Filter by Trust Rating: HIGH
3. Open “DIM_RETAIL_STORE_INSTRUCTOR” table.
4. Review the Overview page. Notice how enriched the asset is, but reviewing the description, the custom attributes, this allow the dimension table to be discoverable by the catalog and user.


5. Click on the lineage and impact tab.

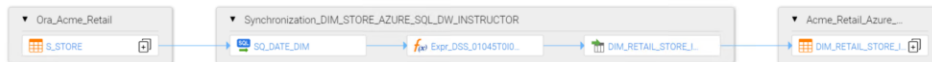
Notice that the Azure dimension table was loaded from a flat file.

6. Use the Lineage Sliders to add all levels in the lineage diagram.

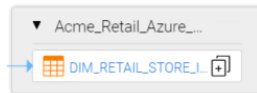
On expanding the lineage, you will notice that the flat file was loaded to an oracle staging table, then loaded to Azure Blob and the finally to an Azure SQL DW Table.

Once you migrated data into Azure SQL DW using Informatica cloud, you can capture the metadata and lineage in Enterprise Data Catalog. (In the next lab, you will learn how to move data to cloud using Intelligent Cloud Services).

7. Click on icon  to show transformation details, and extend lineage to get full lineage as below
8. You will see an orange bubble next to the DIM_RETAIL_STORE table. Click on the bubble to show detailed transformation used to move data from the Oracle table to Azure SQL DW.



9. Now click on the plus icon next to "DIM_RETAIL_STORE_INSTRUCTOR"



10. Select ALL the columns. Click on "OK". This step will expand all the columns in the tables in the lineage diagram that directly affect this metric.
11. You close this detailed lineage diagram to return to the dimension table lineage

You have now seen how metadata is captured from from Azure and Informatica Cloud Data Integration, and how the catalog will be build lineage can to show how an architect can track their data migration to Azure.