

# Module 4

## Data Domains and Curation

### 1- Data Domains :

What is Data Domain ?

A data domain describes the semantics of column data based on either the patterns in that data or the column name.

هو عبارة عن وصف وشرح نوع ومعني البيانات اللي في الاعمده اللي موجوده في الجدول.

Examples of data domains:

- Email address
- IP address
- Phone number
- Social security number

Why do we need Data Domains?

Use data domains to find specific types of data across many tables and columns in your enterprise

يستخدم للعثور علي انواع معينه من البيانات عن طريق معرفة نوع البيانات وهنعرف برضوا هنعرف فين البيانات الاكثر حساسية و نتجنبها.

مثال :

لو عندي في المؤسسة نظام قديم و فية عمود بيحتوي علي بيانات حساسة رقم الضمان الاجتماعي

(Social Security numbers)

وعايز انقل النظام القديم الي نظام جديد

الحل ان عن طريق الداتا دومين اقدر اعرف المعلومات دي واقدر احميها عن طريق التشفير قبل ما انقلها للنظام الجديد.

Where can you Find Data Domains?

يظهر الداتا دومين من خلال data asset اللي انا عايزها عن طريق لونين

لون اخضر و لون اصفر

اللون الاخضر : تم تعينه وتاكيدة علي المعلومات دي  
يمكن يكون ديفلوير عينه و أكد عليه  
او أن EDC هو عينه وحد من الديفلوير اكد عليه

● مثال: عمود تم تأكيد إنه يحتوي على "Social Security Numbers"، ف يظهر كمربع أخضر.  
اللون الاصفر

يعني إن EDC استنتج نوع البيانات بناءً على شكل البيانات أو اسم العمود، لكن ما تمش تأكيده

• هو مجرد تخمين ، لكنه محتاج مراجعة من المستخدم أو الـ Data Steward.

● مثال: عمود فيه بيانات شكلها شبه أرقام هاتف، فيظهر "Phone Number" كمربع أصفر لأنه مجرد تخمين  
لحد دلوقتي

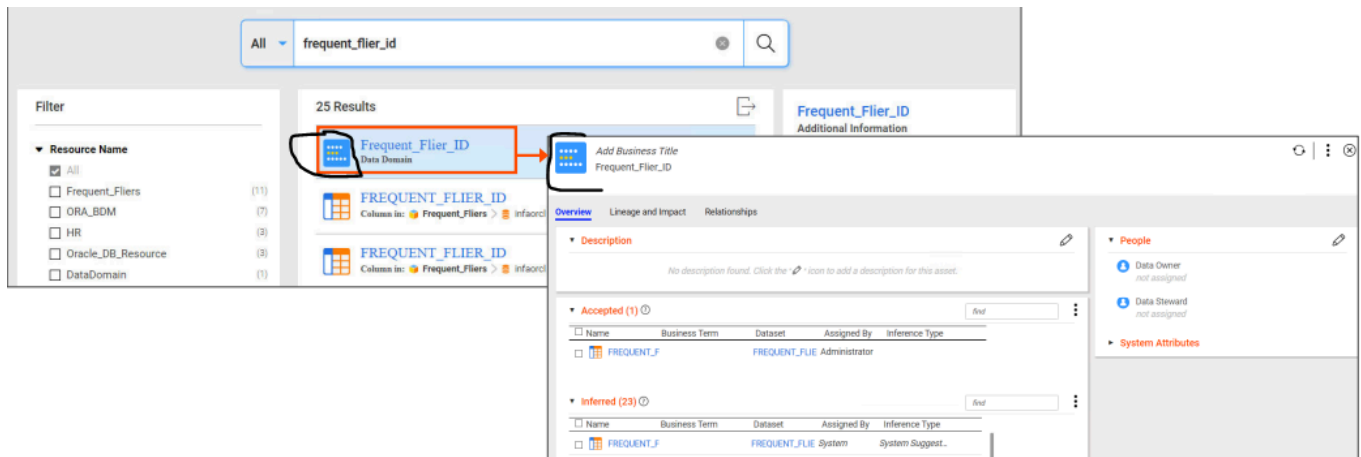
The left screenshot shows a table with columns: Name, Business Title, and Data Domains. The 'Data Domains' column has a red box around it, and a red bracket is next to it. The 'Data Domains' column has a red box around it, and a red bracket is next to it. The 'Data Domains' column has a red box around it, and a red bracket is next to it.

The right screenshot shows two asset cards, 'Mileage' and 'Frequent Flyer'. The 'Mileage' card has a yellow arrow pointing to its 'Data Domains' section. The 'Frequent Flyer' card has a green arrow pointing to its 'Data Domains' section.

يمكن اقدر اوصل ل Assets عن طريق الداتا دومين وبحدد ان البحث و الفلتر يكون من خلال الداتا دومين

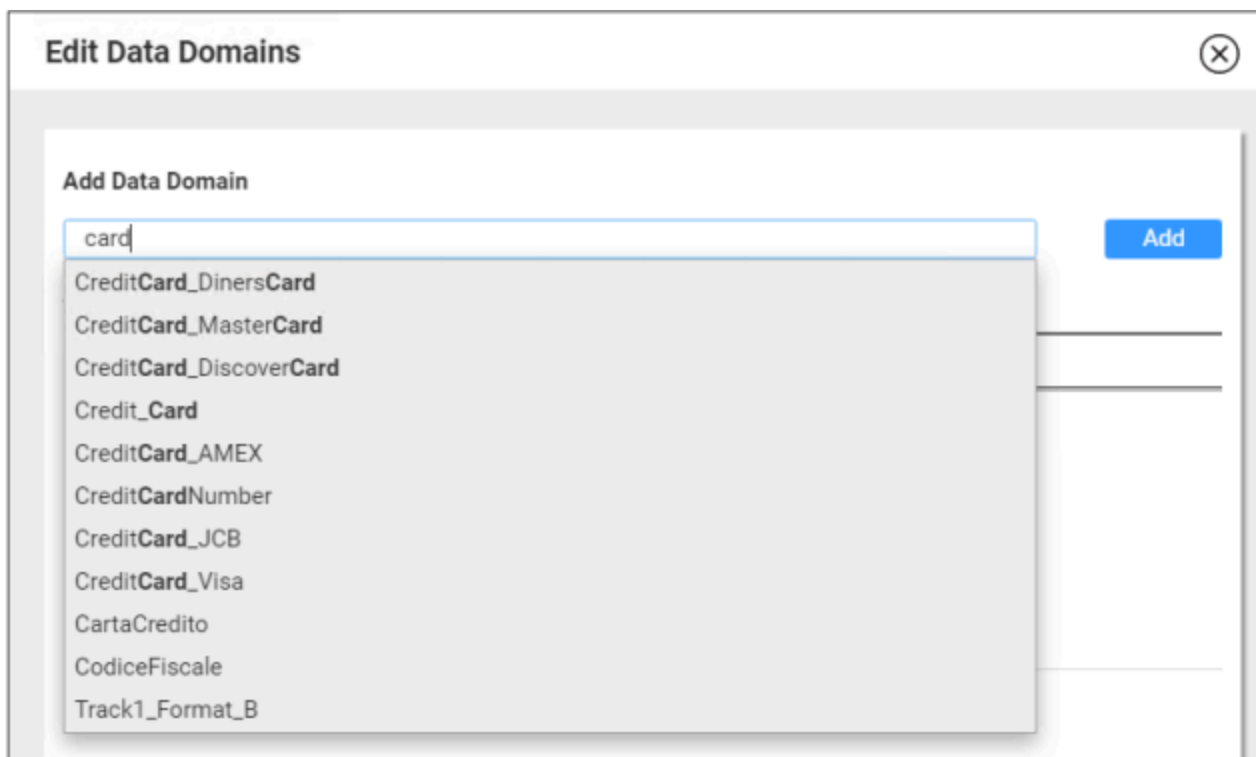
The search bar shows 'frequent\_flier\_id'. The results list shows 'Frequent\_Flier\_ID' as a Data Domain. A dropdown menu for 'Data Domain' is open, showing 'All' (checked), 'Frequent\_Flier\_ID' (24), and 'Flier\_ID' (21).

## View Details of Data Domains :



## Associating Data Domains :

ربط الداتا دومين بال Assest اللي انا عايزها

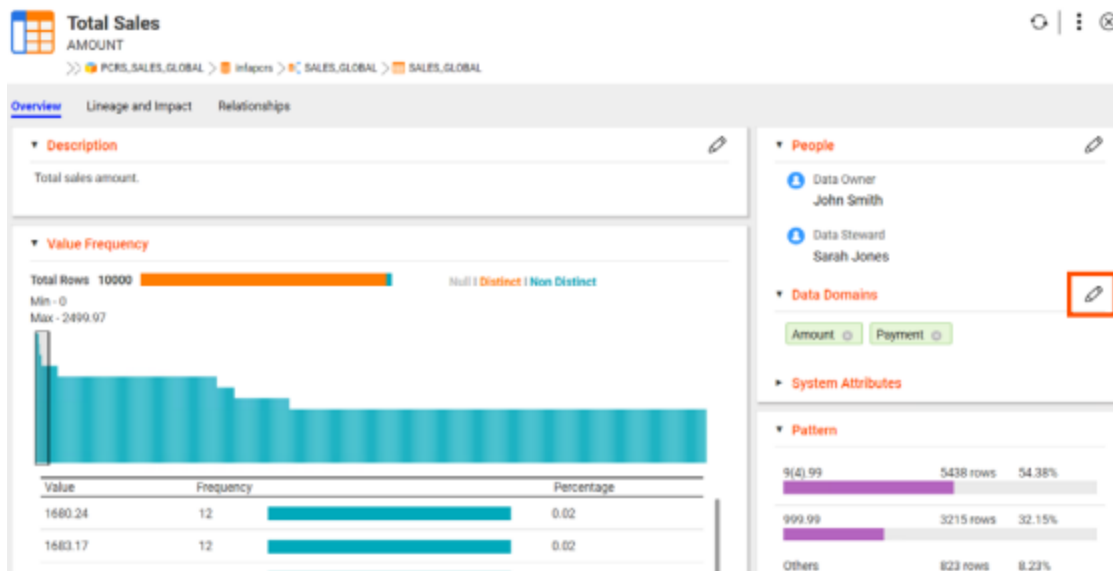


## How do you Associate a Data Domain?

1. Open a column or field asset
2. In the Data Domains section, click the Edit icon

3. Type the name of the existing data domain, and click Add

1-



2-

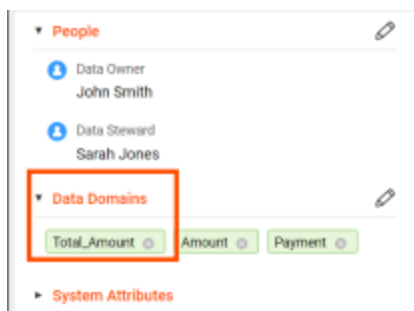
**Edit Data Domains**

**Add Data Domain**

Search: Total  
Suggestions: Total\_Amount

Add

3-



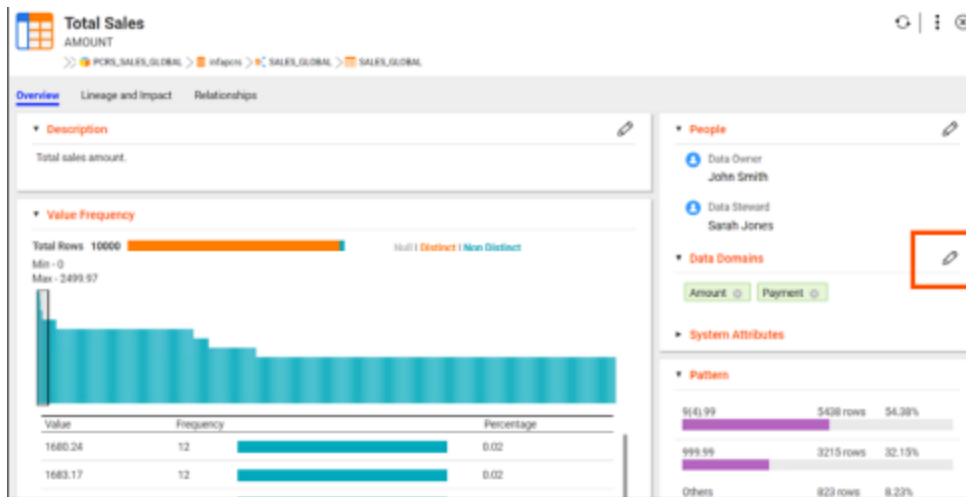
## Creating a Data Domain :

لو مافيش دومين مناسب ليا عشان اربط فيه الدومائيز الجديده اللي انا عايزها هعمل داتا دومين من اول وجديد.

1. Open the column or field asset for which the data domain is to be assigned.
2. In the Data Domains section, click the Edit icon.

3. Type the name of the data domain, and click Add. As the data domain does not exist, it opens a New Data Domain dialog box
4. . 4. Add a description to identify the data domain.
5. Select a data domain group, if required.
6. Click OK.

1-



2-

The screenshot shows the 'Edit Data Domains' dialog box. The 'Add Data Domain' button is highlighted with a red box. The 'New Data Domain' dialog box is also shown, with fields for Name, Description, and Data domain Group.

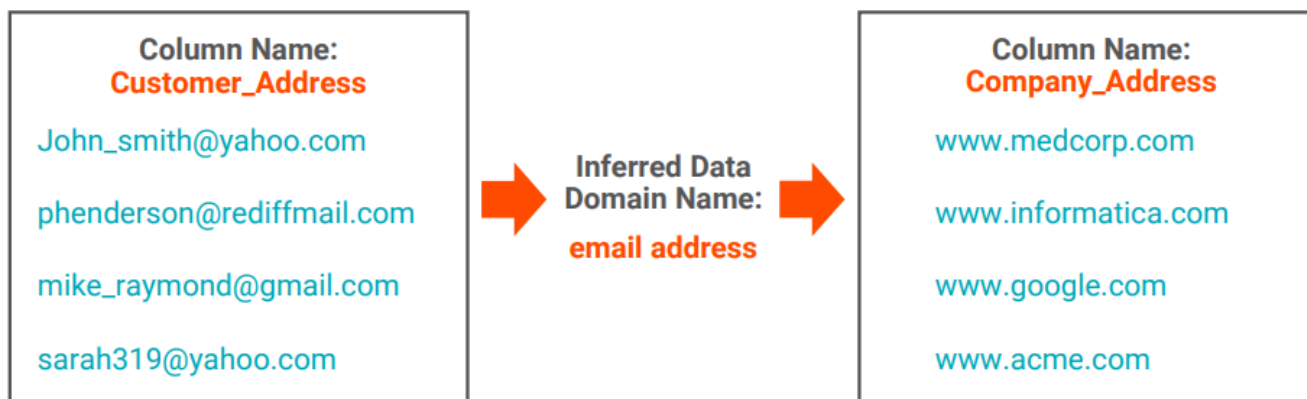
Name	Description	Assigned By	Inference Type	Status
Amount		Administrat...		Accepted
Payment		Administrat...		Accepted

## Curating Data Domains :

Curation is the process of validating and managing discovered metadata of a data source so that it is fit for use and for reporting.

هو عملية تحقق من صحة الميتا داتا ودقتها و تحقق من مصدر الداتا بحيث تكون صالحة للاستخدام و التقارير

عن طريق انه بيستنتج الدومين ده و انا اوفق عليه او ارفض



يمكن الاستنتاج بتاع EDC يكون محتاج لتعديل وتنسيق و ممكن يكون صح او افق عليه وممكن يكون غلط ارفضه

### Accepting or Rejecting Inferred Data Domains :

Data domains inferred for the asset are shown in a yellow box with an inference percentage displayed next to it.

الداتا دومين بتظهر في مربع اصفر و جيمها نسبة مؤية النسبة دي بتشير إلى نسبة تطابق البيانات داخل العمود مع الداتا دومين المقترح من EDC

لو عندك عمود فيه 1000 صف بيانات، و EDC استنتج إنه ممكن يكون فيه SSN(Social Security Number):

900	عدد القيم اللي شكلها SSN
1000	عدد القيم الكلي
90%	نسبة التطابق

فتلاقي المربع الأصفر مكتوب عليه:

**Social Security Number – 90%**

مثال مع الصور :

Location(97%) ✓ ✕

Accept



When you accept, the data domain box turns green to indicate that the domain is assigned to the asset

Reject






When you reject, the data domain no longer appears in the Data Domain section for the asset

## Methods to Curate Data Domains :

في 3 طرق مختلفة بشأن عمل Curation :

### 1. Curate Data Domain via the Asset that is Opened :

- فتح **Asset معين** (زي جدول أو عمود)
- تشوف الـ **Data Domain** اللي EDC استنتجه (inferred)
- تراجع النسبة المئوية للتطابق (مثلاً 88%)
- وتقرر:
-  تقبله (Accept)
-  ترفضه (Reject)
-  أو تغيّره لنوع ثاني

### 2-Open the Data Domain and Curate All Assets Associated With It :

يعني:

- تفتح **Data Domain معين** (زي Social Security Number)
- هنا بفتح الدومين نفسه مش نفس الطريقة الاول (الطريقة الاولى بفتح العمود او الجدول واشوف الاقتراح)
- تشوف **كل الأعمدة** أو الـ Assets اللي EDC ربطها بيه
- تراجعهم واحد واحد أو تعمل **bulk review** وتشوف:
- مين يستحق الاحتفاظ بـ Data Domain ده
- ومين فيه غلط

## Use Export/Import to Bulk Curate Data Domains

يعني:

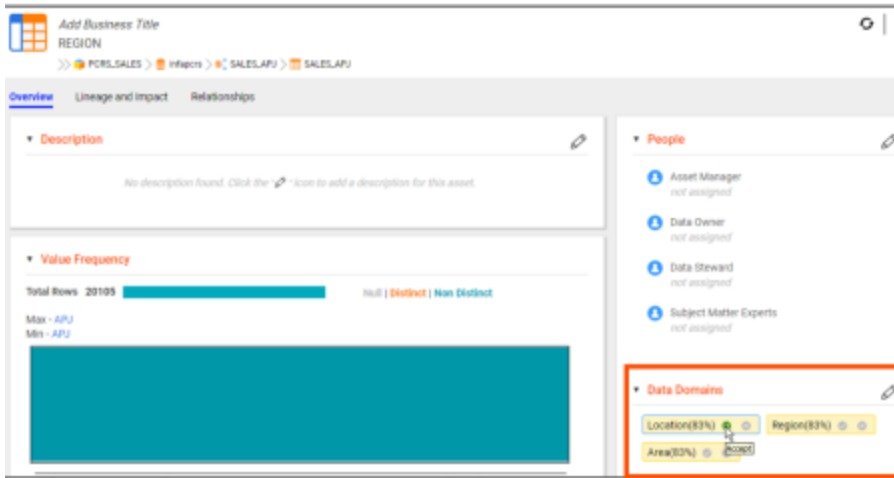
- تعمل تصدير (Export) لقائمة الأعمدة والـ Data Domains المرتبطة بيها على هيئة ملف (Excel أو CSV)
- تراجع البيانات براحة على جهازك
- تحدد في الملف إيه يتم قبوله ورفضه
- بعد كده ترجع الملف تاني إلى EDC (Import) علشان يتم تحديث الحالة Bulk

دي أحسن طريقة لو عندي آلاف الأعمدة وعازي تشتغل عليها دفعة واحدة.

حالة Bulk : معناها دفعة واحدة او كمية كبيره يعني بدل منا يشتغل علي عمود عمود لا بتعامل مع كمية كبيرة من الاعمدة مع دفعة واحدة

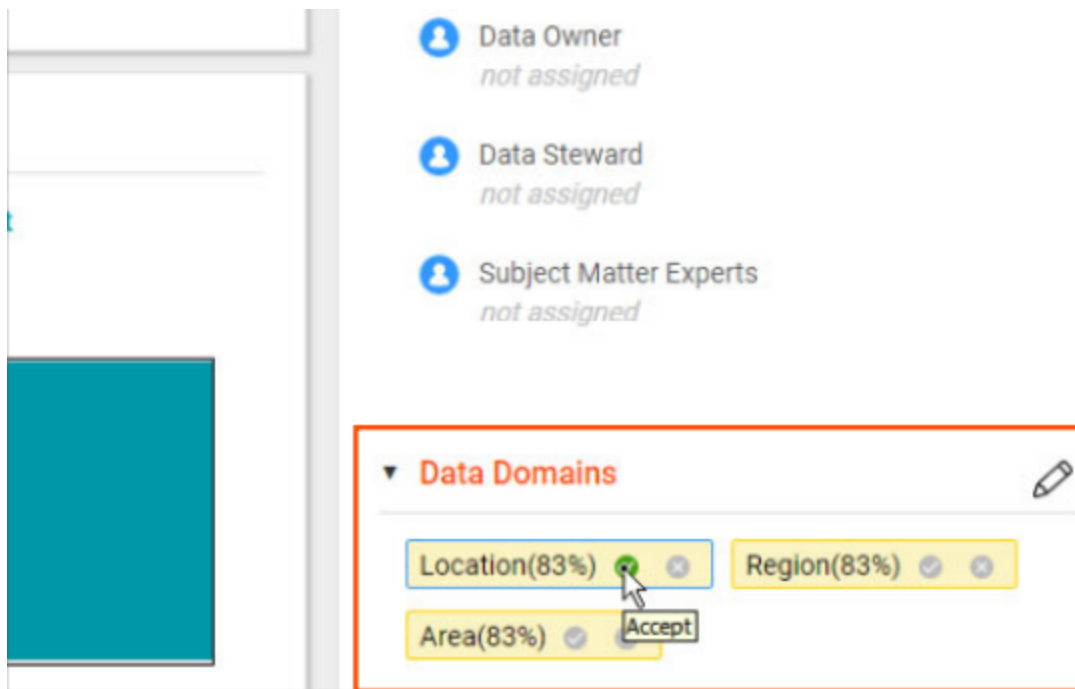
## Curating a Data Domain for an Asset :

1. Open the asset for which you are curating the data domain



2. Accept or reject the inferred data domain(s)





## Curating Assets through Data Domains :

1. Open the Data Domain
2. Select the assets that you want to reject for the data domain
3. Click the Options icon, and click Reject
4. Select the assets that you want to accept for the data domain
5. Click the Options icon, and click Accept

هنا انا بفتح الداتا دومين نفسها كلها وبشوف كل اقتراح ال EDC مدهوني وتبع انهبي Data Set ومين اللي مسؤل عنها وبختار بقا اية اللي ممكن اوافق عليه واية اللي ممكن ارفضه وممكن اختار كل الاقتراحات و اوافق عليها والعكس صحيح

مثال

Add Business Title

Location

Overview

Lineage and Impact

Relationships

Description

Data domain that verifies a location.

Assigned ?

find

Name	Business Term	Dataset	Assigned By	Inferred
<input checked="" type="checkbox"/> COUNTRY		SALES_GLOBAL	Administrator	
<input checked="" type="checkbox"/> REGION		SALES_GLOBAL	Megan_R	
<input checked="" type="checkbox"/> CITY		SALES_GLOBAL	Megan_R	
<input checked="" type="checkbox"/> COUNTRY		CARD_W	System	System Suggeste...
<input checked="" type="checkbox"/> COUNTRY_NAM		EMP_DETAILS_VIE	System	System Suggeste...

Accept

Reject

## Curating Data Domains in Bulk :

1. Export the assets from a resource into a CSV file
2. In the CSV file, move the inferred data domains to either the Accepted or Rejected columns
3. Import the updated file back into EDC

	A	B	C	D	E	F
1	id	core_name	core_classType	bulkimport-DataDomainsAccepted	bulkimport-DataDomainsInferred	bulkimport-DataDomainsRejected
2	id	name	classType	DataDomains Accepted	DataDomains Inferred(Read Only)	DataDomains Rejected
3	HR://infoacrl/HR/LOCATIONS/COUNTRY_ID	COUNTRY_ID	Column		/Location	
4	HR://infoacrl/HR/LOCATIONS/POSTAL_CODE	POSTAL_CODE	Column			
5	HR://infoacrl/HR/COUNTRIES/COUNTRY_NAME	COUNTRY_NAME	Column		/Location	
6	HR://infoacrl/HR/DEPARTMENTS/LOCATION_ID	LOCATION_ID	Column		"/Flier_ID,/Frequent_Flier_ID"	
7	HR://infoacrl/HR/EMPLOYEES/COMMISSION_PCT	COMMISSION_PCT	Column			
8	HR://infoacrl/HR/JOBS/MIN_SALARY	MIN_SALARY	Column			
9	HR://infoacrl/HR/JOB_HISTORY/START_DATE	START_DATE	Column			
10	HR://infoacrl/HR/LOCATIONS/LOCATION_ID	LOCATION_ID	Column		"/Frequent_Flier_ID,/Flier_ID"	
11	HR://infoacrl/HR/JOB_HISTORY/DEPARTMENT_ID	DEPARTMENT_ID	Column			
12	HR://infoacrl/HR/JOBS/MAX_SALARY	MAX_SALARY	Column			
13	HR://infoacrl/HR/EMPLOYEES/EMAIL	EMAIL	Column			
14	HR://infoacrl/HR/EMPLOYEES/EMPLOYEE_ID	EMPLOYEE_ID	Column			
15	HR://infoacrl/HR/COUNTRIES/COUNTRY_ID	COUNTRY_ID	Column		/Location	
16	HR://infoacrl/HR/COUNTRIES/REGION_ID	REGION_ID	Column			
17	HR://infoacrl/HR/DEPARTMENTS/MANAGER_ID	MANAGER_ID	Column			
18	HR://infoacrl/HR/LOCATIONS/STREET_ADDRESS	STREET_ADDRESS	Column			

## Propagating Curated Data Domains :

ده معناه نشر البيانات والمسميات اللي تم التأكيد عليها علي أعمدة اخري مشابهها من حيث مضمون العمود

لما تراجع (**curate**) نطاق بيانات (Data Domain) في EDC — سواء قبلته (**Accept**) أو رفضته (**Reject**) — الـ EDC\*\* بيقا في خطوات بتم بعد كده زي :

بيستخدم القرار اللي الديفولبر قررها علشان يتعلم وينشر (**propagate**) النوع ده من البيانات على أعمدة تانية مشابهة في قواعد البيانات التانية.

## مثال عملي:

تخيل إنك قبلت إن العمود `user_email` يحتوي على Data Domain اسمه `Email Address`

بعد شوية، EDC يلاقي عمود تاني في جدول مختلف اسمه `contact_email`

- شكله شبه العمود الأول
  - نمط البيانات شبه بعض
- يقوم EDC يقترح أو يعين تلقائيًا نفس الـ Data Domain (Email Address) على العمود التاني.
- أو انا اعمل Curate سواء بالتأكد أو الرفض و هو يتعلم تاني

### Example of Data Domain Propagation :

The screenshot shows a table named 'FREQUENT\_FLIERS' with columns: FLIER\_ID, FREQUENT\_FLIER\_ID, FREQUENT\_FLIER\_LEVEL\_ID, FREQUENT\_FLIER\_MILES, FREQUENT\_FLIER\_MILES\_AVAILABLE, FREQUENT\_FLIER\_NUMBER, and JOIN\_DATE. The 'Data Domains' column shows how domains are propagated from 'FLIER\_ID' to other columns. For example, 'FREQUENT\_FLIER\_ID' has a domain of 'Frequent\_Flier\_ID' (+1 more), and 'FREQUENT\_FLIER\_LEVEL\_ID' has a domain of 'Frequent\_Flier\_ID(80%)' (+1 more). The 'FREQUENT\_FLIER\_MILES' column has a domain of 'Frequent\_Flier\_ID(67%)'. The 'FREQUENT\_FLIER\_MILES\_AVAILABLE' column has a domain of 'Frequent\_Flier\_ID(65%)'. The 'FREQUENT\_FLIER\_NUMBER' column has a domain of 'Frequent\_Flier\_ID(65%)'. The 'JOIN\_DATE' column has a domain of 'Date' (7).

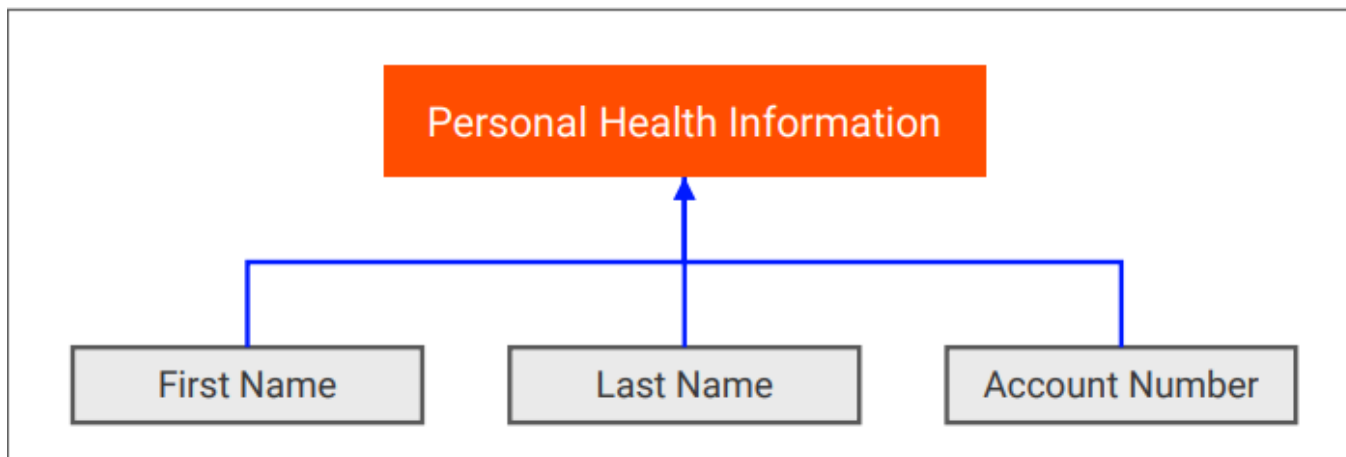
Name	Business Title	Data Domains	Null   Distinct   Non-Distinct %	Source Data Type   Inferred Data Types
1 FLIER_ID		Flier_ID	0   10.12   89.88	NUMBER (10) Fixed Length String(4)   100.00% +3 more
2 FREQUENT_FLIER_ID		Frequent_Flier_ID +1 more	0   100   0	NUMBER (10) Fixed Length String(4)   100.00% +3 more
3 FREQUENT_FLIER_LEVEL_ID		Frequent_Flier_ID(80%) +1 more	0   0.50   99.50	NUMBER (10) Fixed Length String(4)   100.00% +3 more
4 FREQUENT_FLIER_MILES	Mileage	Frequent_Flier_ID(67%)	0   26.25   73.75	NUMBER (22) Date(yywwd)   6.25% +5 more
5 FREQUENT_FLIER_MILES_AVAILABLE			0   1.25   98.75	NUMBER (22) String(5)   100.00% +2 more
6 FREQUENT_FLIER_NUMBER		Frequent_Flier_ID(65%)	0   100   0	NUMBER (22) Decimal(7)   100.00% +3 more
7 JOIN_DATE			0   0.87   99.13	DATE (7) Date Time(dd/mm/year HH24:mi:ss)   33.25% +3 r

## Data Domain Groups :

هو عبارة عن مجموعة (Group) بتضم أنواع بيانات (Data Domains) متشابهة أو مرتبطة ببعضها.

يعني بدل ما تتعامل مع Data Domains كأنهم أنواع منفصلة وكثير ،  
تقدر تجمعهم في "مجموعة" واحدة علشان التنظيم و بيقا اسهل في Search .

مثال بصور :



مثال ثاني :

Group Name	يحتوي على Data Domains مثل:
Personal Data	Email Address, Phone Number, SSN
Financial Data	Credit Card Number, Bank Account
Location Data	IP Address, Zip Code

ويمكن احط نفس الدومين في اكثر من Group :

يعني ممكن بيقا عندي داتا دومين اسمها Phone Number موجوده في اكثر من جروب زي

- 1- Personal Data
- 2- Contact Info

الديفلوبر بقا ممكن يشيل الدومين من الجروبات اللي هو مش عايز بيقا الدومين ده موجود فيها

## Why do we need to group Data Domains?

علشان نسهل عملية البحث، والتنظيم في السيطرة علي البيانات (الادارة )، والتحليل للبيانات اللي ليها علاقة ببعض، بدل ما نبحت عن كل نوع بيانات على لوحده.

مثال :

عايز اوصل لكل البيانات اللي ليها علاقة بـ الحسابات البنكية داخل المؤسسة.  
البيانات دي ممكن تكون متفرقة في قواعد بيانات مختلفة، وبأشكال متعددة:

- Bank Account Number

- IBAN
- Routing Number
- SWIFT Code
- Account Holder Name

كل واحدة من دول ممكن تكون **Data Domain** مستقلة.

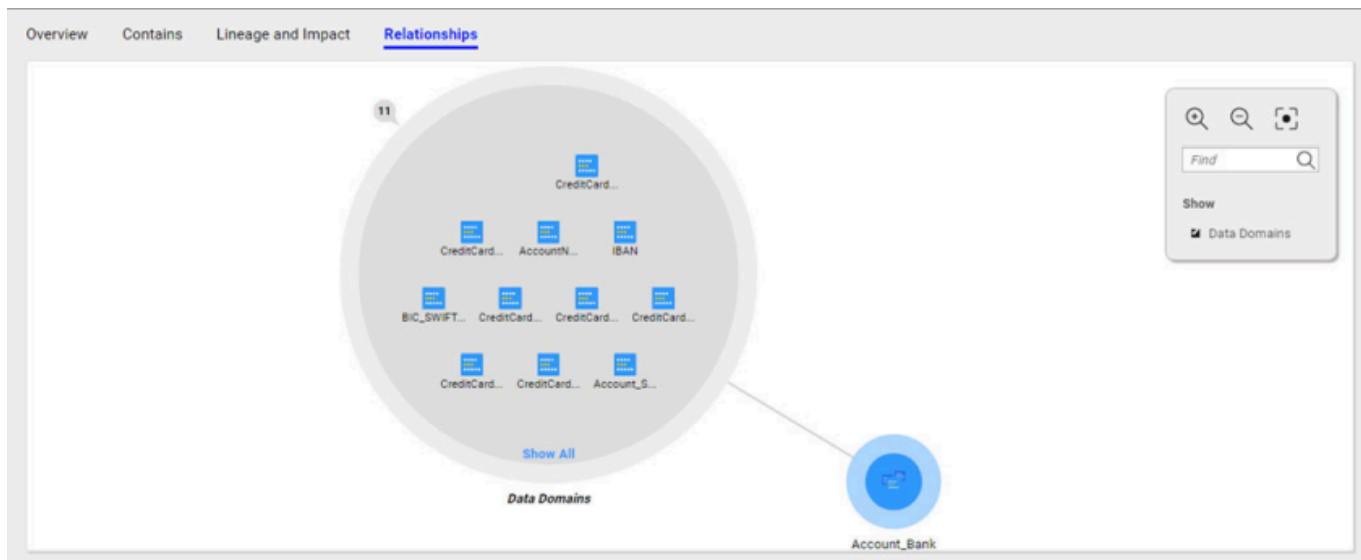
بدل ما تبحث عن كل واحدة منهم لوحدها،  
الديفلوهر \*\* يقدر يعمل مجموعة اسمها مثلاً:

◆ Account\_Bank

ويضيف جواها كل الـ Data Domains دي.

- لما اعمل بحث في EDC باستخدام مجموعة Account\_Bank هيطالعك كل الأعمدة أو الأصول المرتبطة بأي واحد من الداتا دومينز دي EDC
- من غير ما تكتب أسماءهم واحد واحد

مثال بالصورة :



## Composite Data Domains :

هو نوع خاص من **Data Domains** بيتكوّن من:

- مجموعة من **Data Domains** عادية
- أو حتى من **Composite Data Domains** ثانية
- مترابطين مع بعض عن طريق قواعد (Rules)

بمعني عندي بيانات بتمثل البنوك بس البيانات دي متوزعه علي اجزاء مختلفة مثلا في جدول ثاني او another Schema

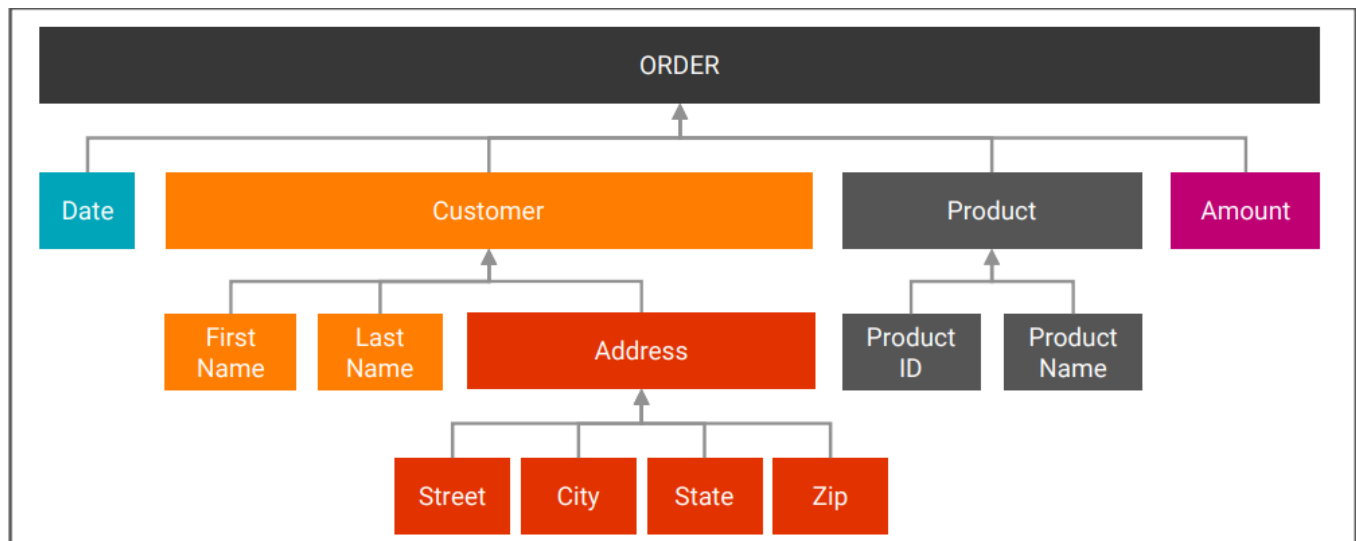
انا اقدر اجمع الداتا دومينز دي اللي في أماكن مختلفة علي أساس انها وحدة وحده مترابطه ببعض بقواعد معينه

## الفرق بين Composite Data Domains & Domain Group

	Data Domain Group 📁	Composite Data Domain 🌿
الهدف	تنظيم وتصنيف Data Domains تحت عنوان مشترك	تمثيل "كيان" مكون من بيانات متفرقة
الفكرة	مجرد قائمة ثابتة من Data Domains	مبني على قواعد تربط Data Domains ببعض
امكان الداتا	مش شرط مترابطة – مجرد أنواع بيانات متشابهة	غالبًا مترابطة بس متوزعة على أعمدة مختلفة
مثال	مجموعة Personal Info فيها: Email, Phone, SSN	Composite Customer_Profile فيه: Name + Email + Address

تبسيط الفكرة لل Composite Data Domains هنعبر انها قطع puzzle ومتفرقه وانا بجمعها مع بعضها بناء علي انها دي اللي هتكون الصورة الاساسية بتعتي اللي هيا هتعتبر الدتا ولما تتجمع مع بعضها هتديني الشكل العام

مثال بالصور :



## Module Summary :

- Describe data domains
- Identify the need for data domains
- List the types of data domains

- Describe the curation process
- List the methods to curate data domains