Lab 01 Data Virtualization and Data Protection

1.1 Lab overview

Organizations are faced with 2 realities today:

a. They have data spread all over the place – multiple clouds, a plethora of onsite applications, data warehouses, spreadsheets, email systems, etc

b. Regulatory and confidentiality pressures are making it necessary for organizations to think of better, easier and more effective ways to protect that data, while at the same time getting the most value out of that data.

In this lab, we will be showing you how the capabilities of **Watson Query** and **Watson Knowledge Catalog**, when used together in Cloud Pak for Data, can solve these 2 vexing problems.

1.2 Personas represented in this lab

Persona (Role)	Capabilities
Data Steward	Data Stewards integrate and transform data as well as provide governance, lineage and classification of the data.

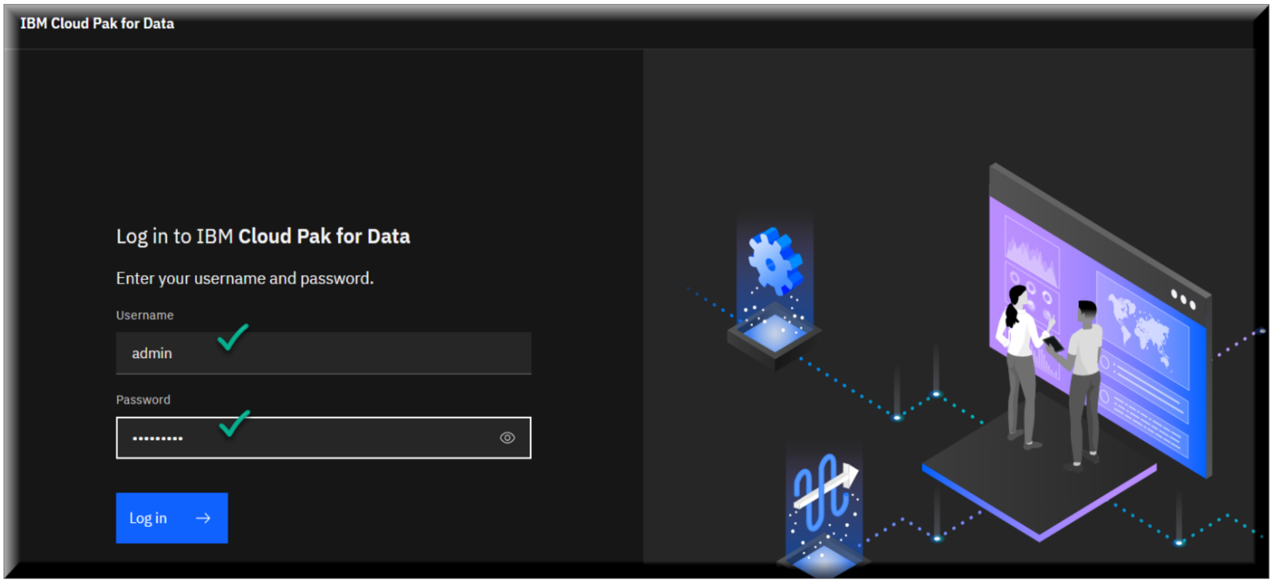
Persona (Role)	Capabilities
Business Analyst	Business Analysts deliver value by taking data, using it to answer questions, and communicating the results to help make better business decisions.

Persona (Role)	Capabilities
Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

1.3 Log into Cloud Pak for Data, Learn about Projects, Catalogs, and Governance Artifacts

__1) Open a browser and go to the URL given by your instructor

__2) Log in as admin user – password will be provided by your instructor



IBM Cloud Pak for Data

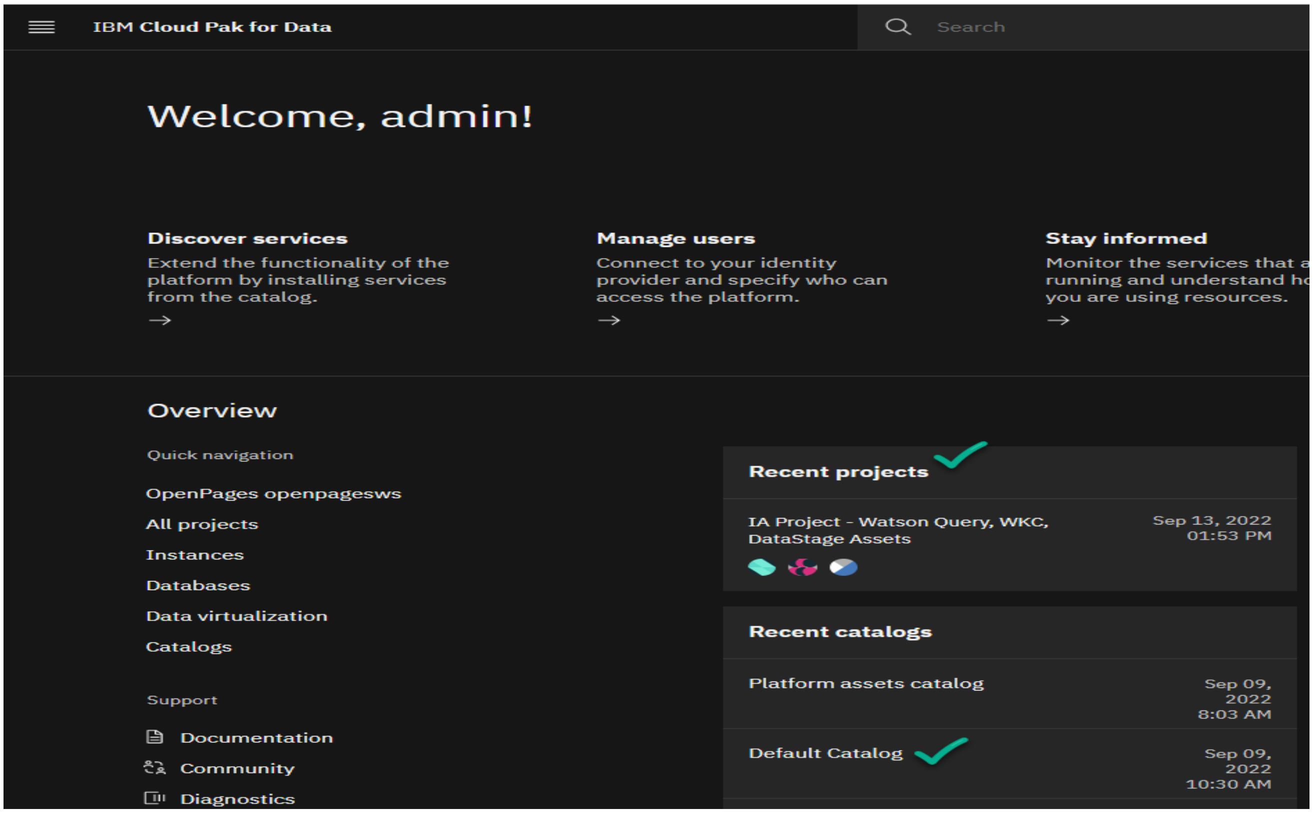
Log in to IBM **Cloud Pak for Data**

Enter your username and password.

Username: admin

Password: •••••••••

This will place you in the home screen for Cloud Pak for Data. This screen is completely customizable, allowing you to brand according to your organization, add and delete any of the tiles, etc. In addition, this screens' content will change based on the role of the user logging in.



IBM Cloud Pak for Data

Welcome, admin!

Discover services
Extend the functionality of the platform by installing services from the catalog.
→

Manage users
Connect to your identity provider and specify who can access the platform.
→

Stay informed
Monitor the services that a running and understand h you are using resources.
→

Overview

Quick navigation

OpenPages openpagesws
All projects
Instances
Databases
Data virtualization
Catalogs

Support

Documentation
Community
Diagnostics

Recent projects

IA Project - Watson Query, WKC, DataStage Assets	Sep 13, 2022 01:53 PM

Recent catalogs

Platform assets catalog	Sep 09, 2022 8:03 AM
Default Catalog	Sep 09, 2022 10:30 AM

Note the 'Recent Projects' and 'Recent Catalogs' areas – these are 2 important components of Cloud Pak for Data.

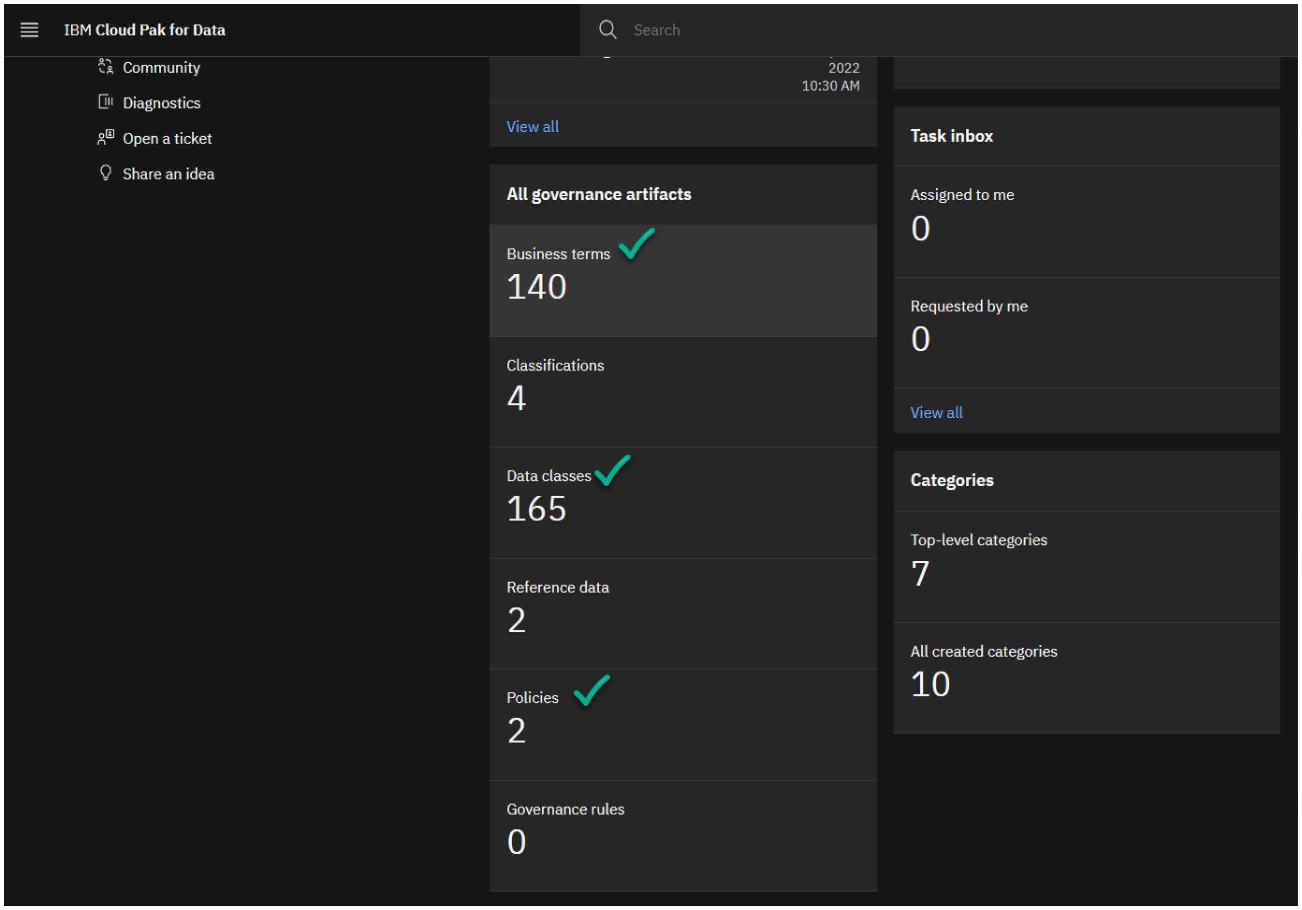
Projects are where the user develops assets – DataStage Jobs, Data Enrichment processes, Jupyter Notebooks, SPSS Modeler flows, and on and on are all housed within a project during development.

Catalogs are where users go to find assets that are 'in production'. These are assets that have been published through a workflow process and are governed to allow the right users to see the right data. For instance, a user can search for any data asset, business term, data connection, etc. from the search bar along the top of any Cloud Pak for Data screen.

Moving further down the home screen, we come across 'Governance Artifacts' – these are the components of Cloud Pak for Data that provide all the Data Governance capabilities of the platform.

You'll see we have 140 **Business terms** – these are English descriptions of data objects such as tables, columns, object storage files, and any related data assets or other governance artifacts. For example, a Business Term called 'Taxpayer ID' might be related to columns and tables that contain data on Taxpayer IDs, such as Social Security Number.

We also have 165 **Data Classes –** this is an incredibly powerful set of objects that are used to classify data, so continuing with our Taxpayer example, a column containing Social Security numbers can be automatically 'stamped' with the data class 'US Social Security Number', Any subsequent access to that column (or any column stamped with that data class) can then be governed by the platform, automatically.



All governance artifacts

Business terms ✓
140

Classifications
4

Data classes ✓
165

Reference data
2

Policies ✓
2

Governance rules
0

Task inbox

Assigned to me
0

Requested by me
0

Categories

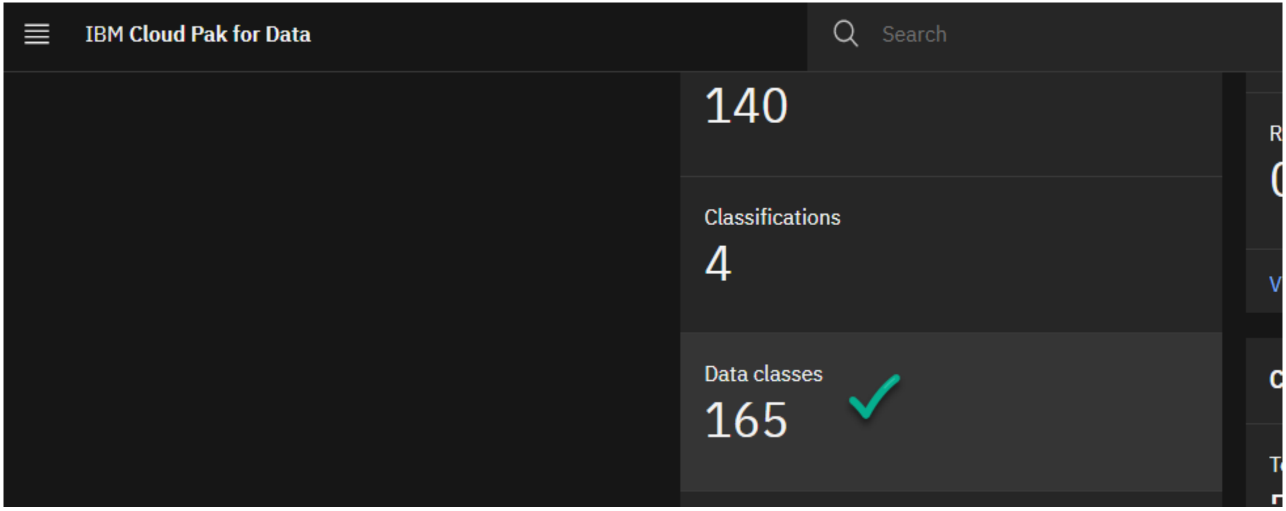
Top-level categories
7

All created categories
10

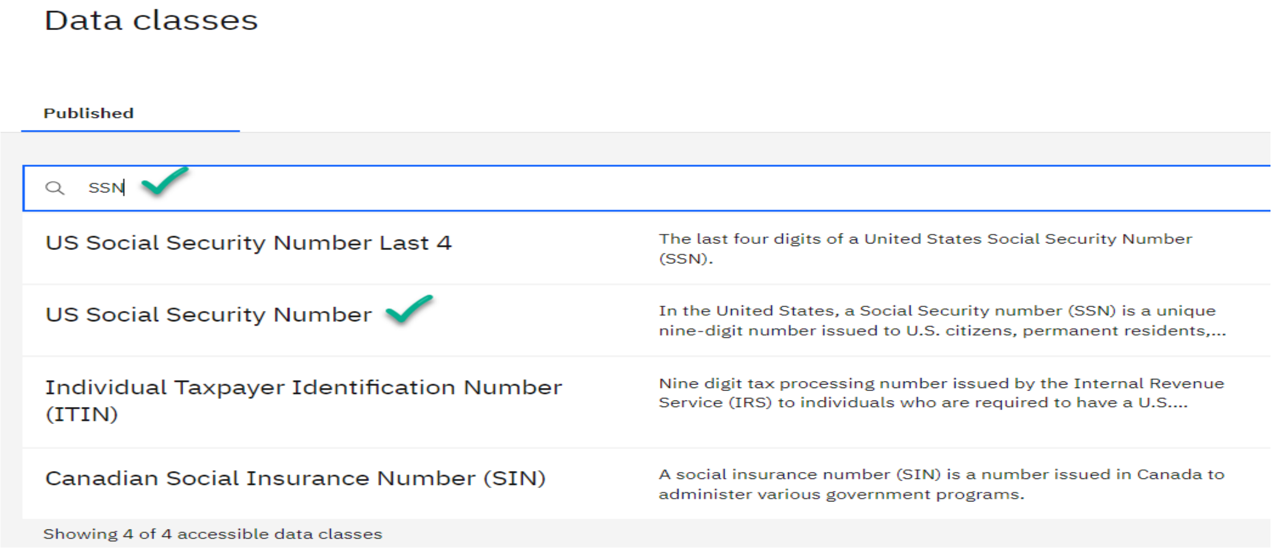
Lastly, **Policies** – these are actual organization policies around the use and handling of data, in software form. These policies allow organizations to actually codify their policies and use them as part of an active data governance practice, as opposed to just having them in paper form with no actual capabilities or enforcement mechanisms, beyond manual intervention.

Now, let's take a closer look at **Data Classes**

__3) Click on the area marked 'Data Classes'



__4) This will take you to a list of all published data classes in our environment. Click into the search bar on this page and type in 'SSN'.



SSN

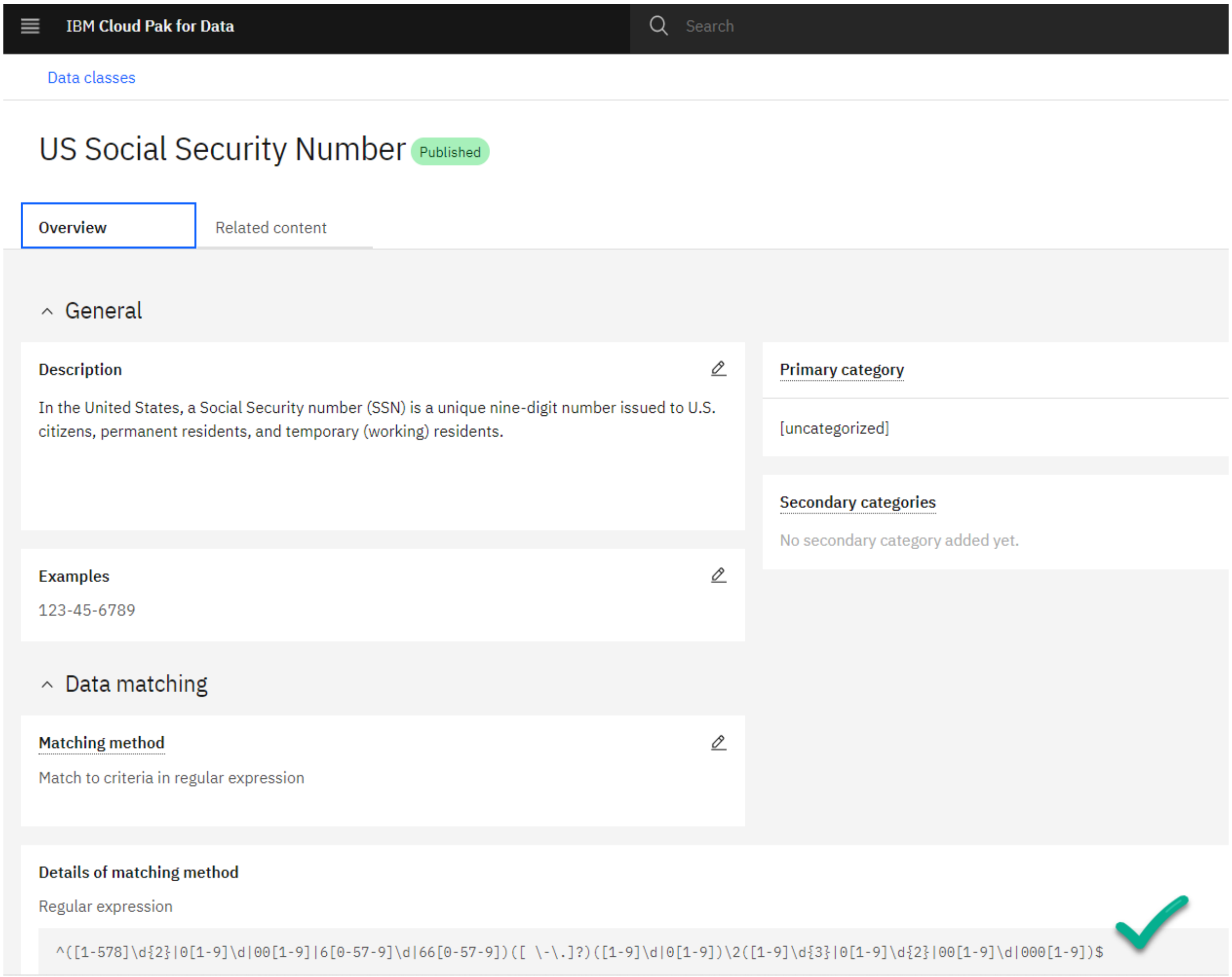
US Social Security Number Last 4	The last four digits of a United States Social Security Number (SSN).
US Social Security Number	In the United States, a Social Security number (SSN) is a unique nine-digit number issued to U.S. citizens, permanent residents,...
Individual Taxpayer Identification Number (ITIN)	Nine digit tax processing number issued by the Internal Revenue Service (IRS) to individuals who are required to have a U.S....
Canadian Social Insurance Number (SIN)	A social insurance number (SIN) is a number issued in Canada to administer various government programs.

Showing 4 of 4 accessible data classes

Now you see all data classes related to Social Security Number

__5) Click on '**US Social Security Number'**

This will bring you to a details screen showing everything about this data class. Notice the area called ‘Details of Matching Method’ – this shows a regular expression (a type of sophisticated pattern matching) – that will validate not just whether the pattern of a string is in a valid SSN format, but also whether the SSN is a valid SSN, as there are many constraints on the data in an SSN field. The reason this matching is so important, is because this is how the platform automatically determines if a particular field holds data for a particular data class.



US Social Security Number Published

General

Description

In the United States, a Social Security number (SSN) is a unique nine-digit number issued to U.S. citizens, permanent residents, and temporary (working) residents.

Primary category

[uncategorized]

Secondary categories

No secondary category added yet.

Examples

123-45-6789

Data matching

Matching method

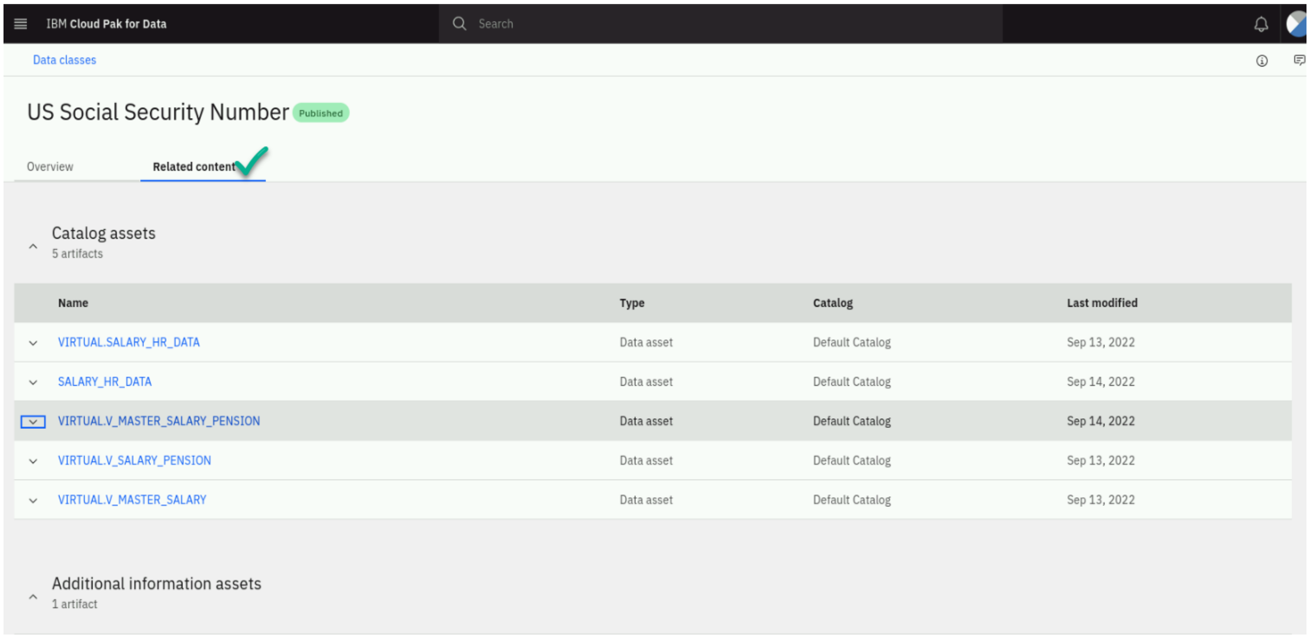
Match to criteria in regular expression

Details of matching method

Regular expression

```
^([1-578]\d{2}|0[1-9]\d|00[1-9]|6[0-57-9]\d|66[0-57-9])([ \-\.]?)([1-9]\d|0[1-9])\2([1-9]\d{3}|0[1-9]\d{2}|00[1-9]\d|000[1-9])$
```

__**6)** Click on the ‘Related content’ tab. This will show you a list of data assets (tables, etc) in the **catalog** that have this data class on at least one column. You will see that 5 items per page are being displayed.



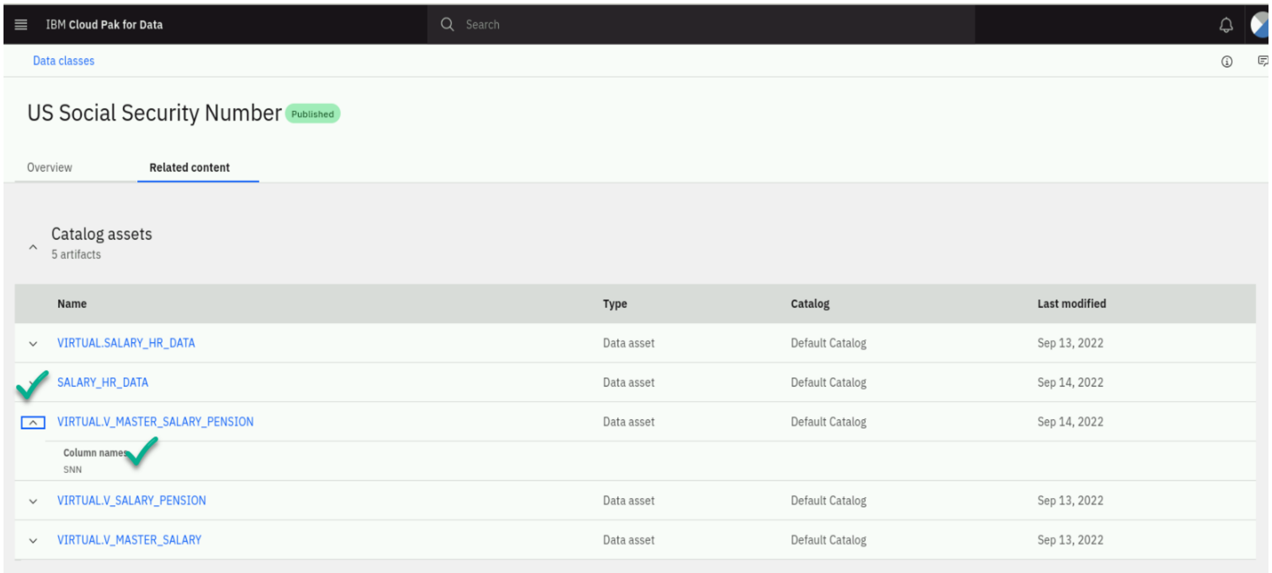
US Social Security Number Published

Catalog assets
5 artifacts

Name	Type	Catalog	Last modified
VIRTUAL.SALARY_HR_DATA	Data asset	Default Catalog	Sep 13, 2022
SALARY_HR_DATA	Data asset	Default Catalog	Sep 14, 2022
VIRTUAL.V_MASTER_SALARY_PENSION	Data asset	Default Catalog	Sep 14, 2022
VIRTUAL.V_SALARY_PENSION	Data asset	Default Catalog	Sep 13, 2022
VIRTUAL.V_MASTER_SALARY	Data asset	Default Catalog	Sep 13, 2022

Additional information assets
1 artifact

__7) **Now** click the dropdown next to **VIRTUAL.V_MASTER_SALARY_PENSION** you'll see a column name 'SNN'. This column has the data class of 'US Social Security number' stamped on it, even though the column name is incorrect.



US Social Security Number Published

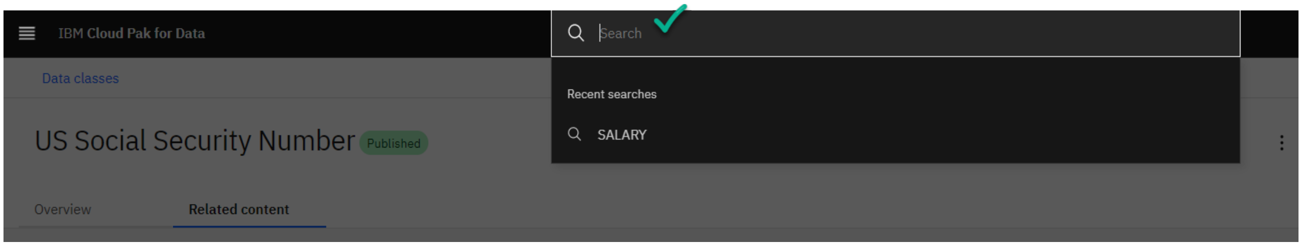
Catalog assets
5 artifacts

Name	Type	Catalog	Last modified
VIRTUAL.SALARY_HR_DATA	Data asset	Default Catalog	Sep 13, 2022
SALARY_HR_DATA	Data asset	Default Catalog	Sep 14, 2022
VIRTUAL.V_MASTER_SALARY_PENSION	Data asset	Default Catalog	Sep 14, 2022
Column names SNN			
VIRTUAL.V_SALARY_PENSION	Data asset	Default Catalog	Sep 13, 2022
VIRTUAL.V_MASTER_SALARY	Data asset	Default Catalog	Sep 13, 2022

1.4 A look at Searching and the Data Catalog

Now that we've learned a bit about some of the governance artifacts and how they work, we will next delve into the Data Catalog. The Data Catalog is at the heart of everything that happens in Cloud Pak for Data. It's the centralized repository for all the data assets and data connections within your organization.

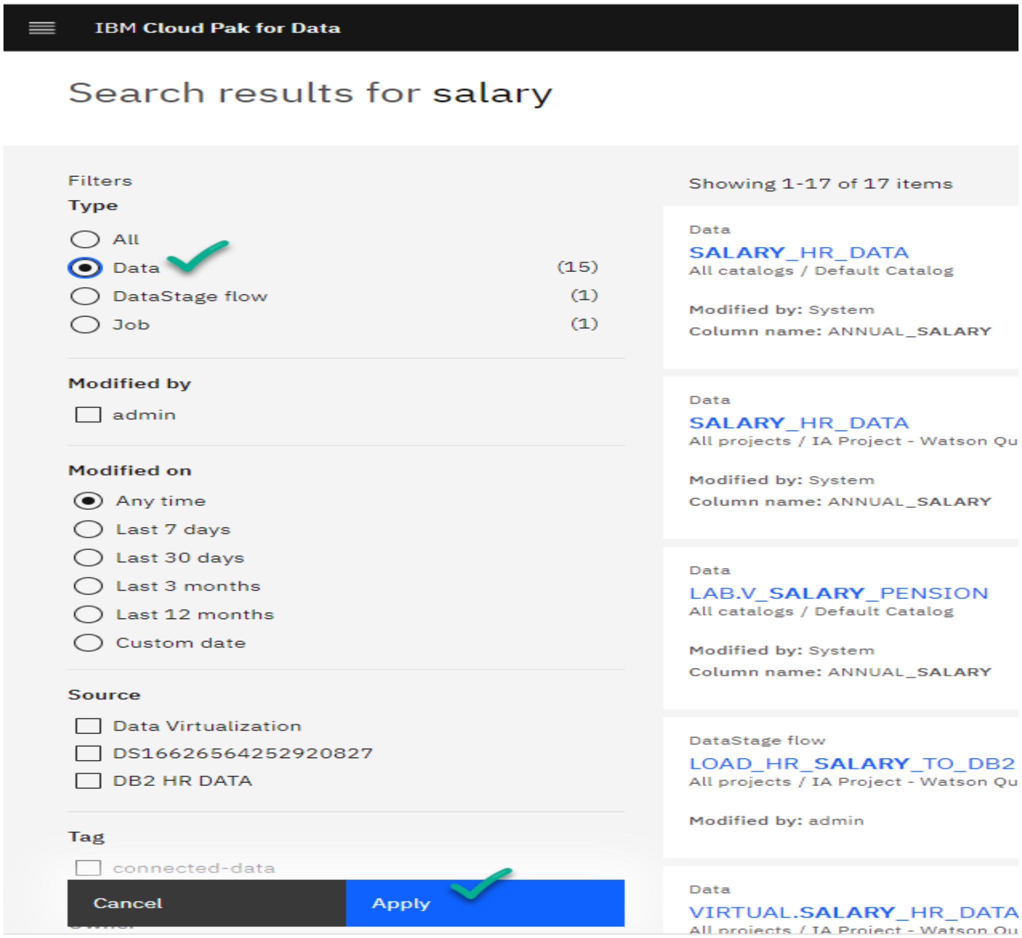
__8) Position your cursor in the Search Bar at the very top of the screen you're on.



__9) Type in 'salary', and then press 'enter'

This will bring you to a screen that shows all assets across the platform that have 'salary' somewhere in the name or description or tag.

__10) You'll see a amazon.com style search on the left side of the screen. Click the checkbox next to '**Data**' and then click **apply.**



Search results for salary

Filters

Type

○ All
● Data (15)
○ DataStage flow (1)
○ Job (1)

Modified by

□ admin

Modified on

● Any time
○ Last 7 days
○ Last 30 days
○ Last 3 months
○ Last 12 months
○ Custom date

Source

□ Data Virtualization
□ DS16626564252920827
□ DB2 HR DATA

Tag

□ connected-data

Showing 1-17 of 17 items

Data
SALARY_HR_DATA
All catalogs / Default Catalog
Modified by: System
Column name: ANNUAL_SALARY

Data
SALARY_HR_DATA
All projects / IA Project - Watson Qu
Modified by: System
Column name: ANNUAL_SALARY

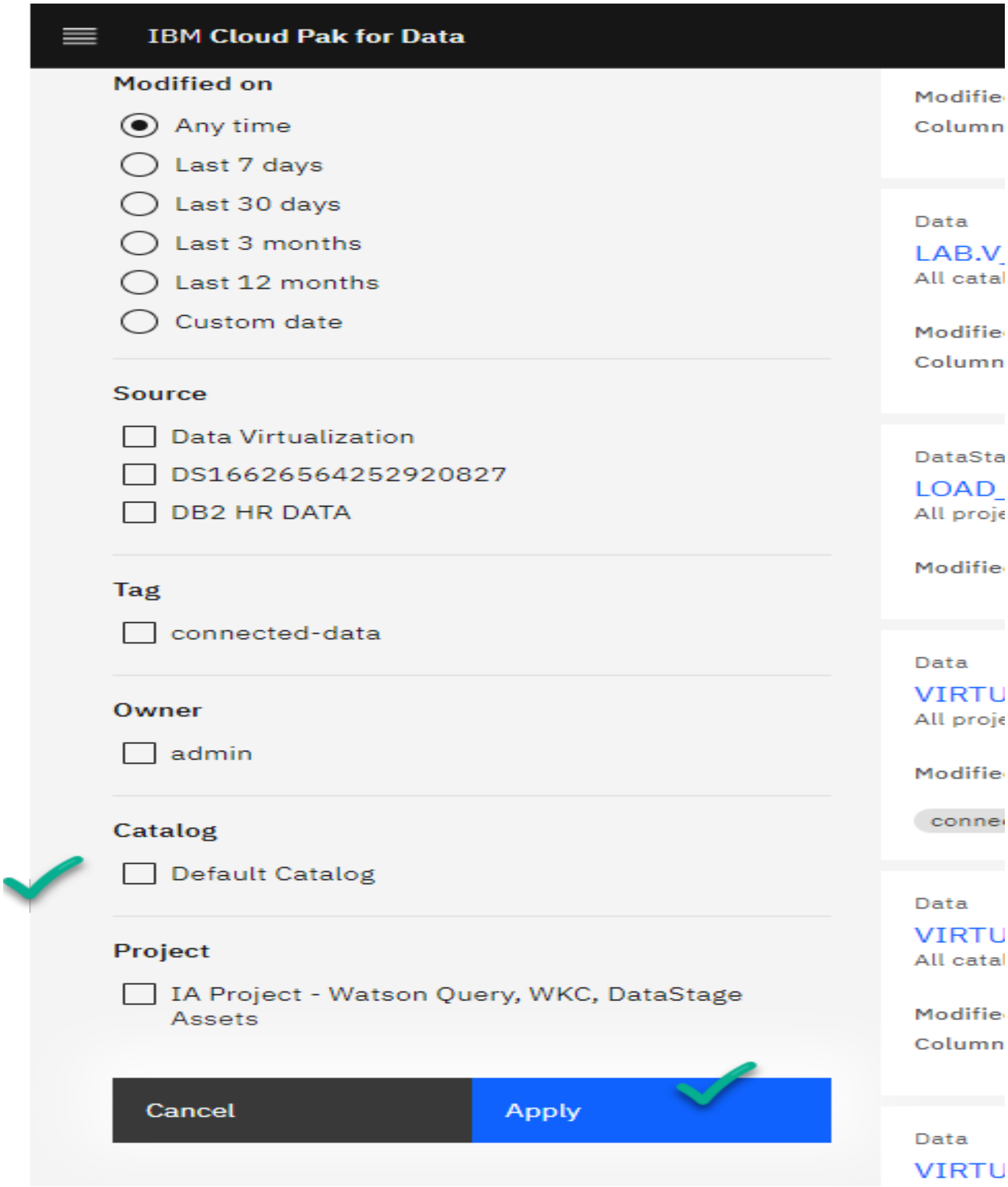
Data
LAB.V_SALARY_PENSION
All catalogs / Default Catalog
Modified by: System
Column name: ANNUAL_SALARY

DataStage flow
LOAD_HR_SALARY_TO_DB2
All projects / IA Project - Watson Qu
Modified by: admin

Data
VIRTUAL.SALARY_HR_DATA
All projects / IA Project - Watson Qu

Next, we'll want to filter to only show options in the Catalog (as Admin we can see all assets, in reality you would be logging in as a user with less authority and therefore, would not see all assets.)

__11) Scroll the window down until you see a checkbox for ‘**Default Catalog’**. Click that, and then click **‘Apply’**.



IBM Cloud Pak for Data

Modified on

● Any time
○ Last 7 days
○ Last 30 days
○ Last 3 months
○ Last 12 months
○ Custom date

Source

□ Data Virtualization
□ DS1662564252920827
□ DB2 HR DATA

Tag

□ connected-data

Owner

□ admin

Catalog

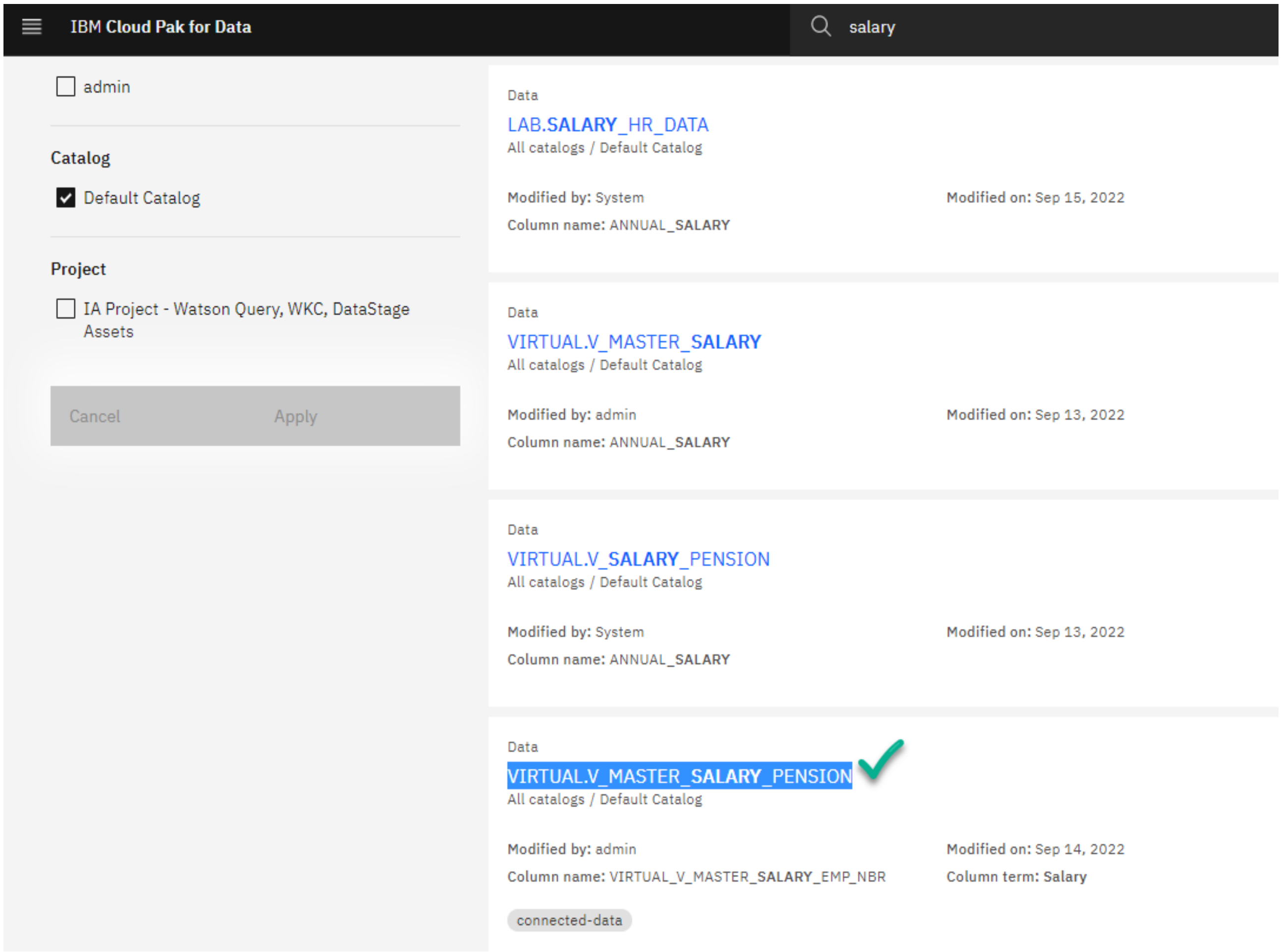
□ Default Catalog

Project

□ IA Project - Watson Query, WKC, DataStage Assets

Cancel | Apply

__12) Now, find **VIRTUAL.V_MASTER_SALARY_PENSION** in the list, and click on it.



□ admin

Catalog

☑ Default Catalog

Project

□ IA Project - Watson Query, WKC, DataStage Assets

Data
LAB.SALARY_HR_DATA
All catalogs / Default Catalog
Modified by: System **Modified on:** Sep 15, 2022
Column name: ANNUAL_SALARY

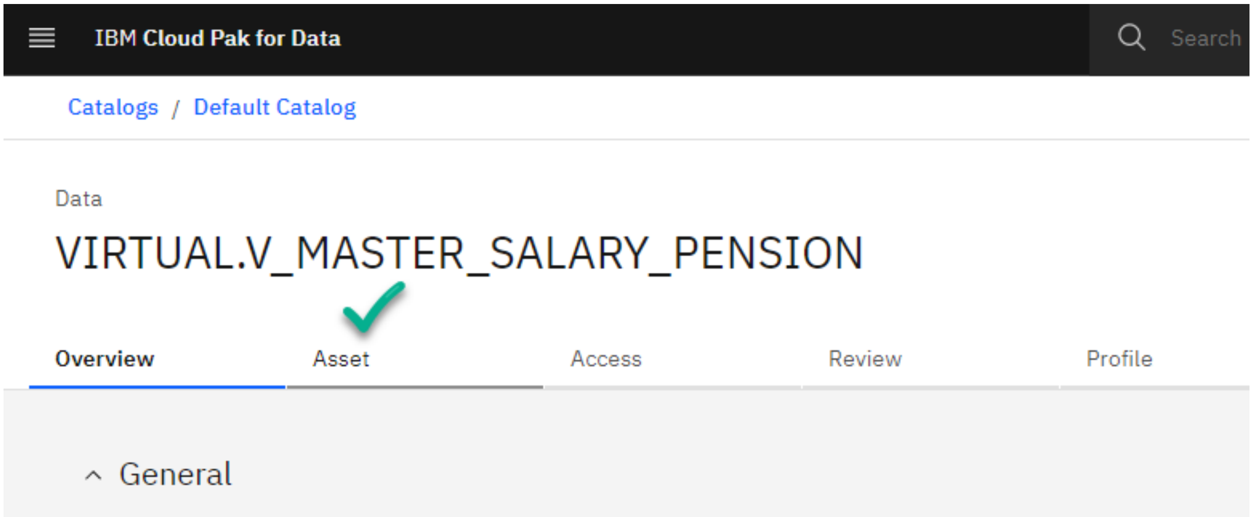
Data
VIRTUAL.V_MASTER_SALARY
All catalogs / Default Catalog
Modified by: admin **Modified on:** Sep 13, 2022
Column name: ANNUAL_SALARY

Data
VIRTUAL.V_SALARY_PENSION
All catalogs / Default Catalog
Modified by: System **Modified on:** Sep 13, 2022
Column name: ANNUAL_SALARY

Data
VIRTUAL.V_MASTER_SALARY_PENSION ✓
All catalogs / Default Catalog
Modified by: admin **Modified on:** Sep 14, 2022
Column name: VIRTUAL_V_MASTER_SALARY_EMP_NBR **Column term:** Salary
connected-data

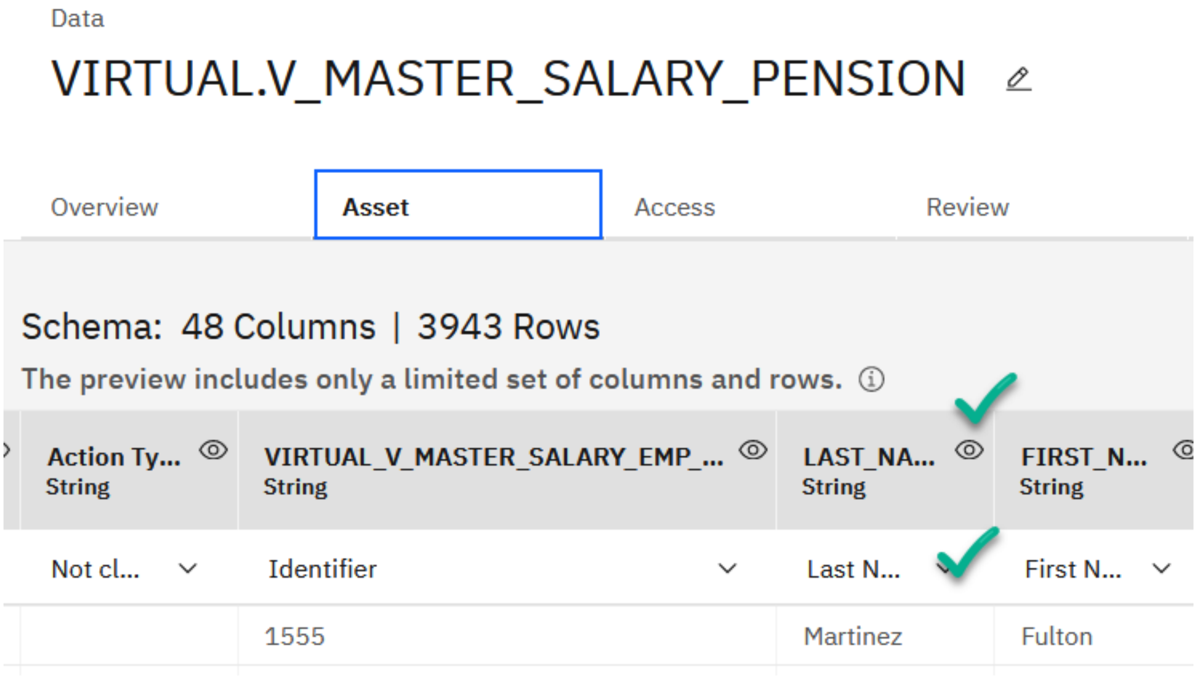
This will bring you into the asset within the catalog.

__13) Click on the '**Asset**' tab



This will bring you into the data preview screen. Here you'll see the first use of the **data classes**. If you look at the row of dropdowns below the column names, you'll see various data classes. These were automatically stamped onto this asset through the '**Automated Data Enrichment'** process. Recall earlier, we looked at data classes and their related assets and **VIRTUAL.V_MASTER_SALARY_PENSION** was in that list of related assets for the **US Social Security Number** class.

Scroll all the way over until you see the 'LAST_NA...' column. Click the 👁 symbol. This will pop up a window showing that there is a Business Term associated with this column – not surprisingly, it is 'LAST NAME'. This was also automatically stamped to this asset using 'Automated Data Enrichment'. No need to click on it, but if you want it will open up a new tab with the details of the 'LAST NAME' glossary term.



Schema: 48 Columns | 3943 Rows

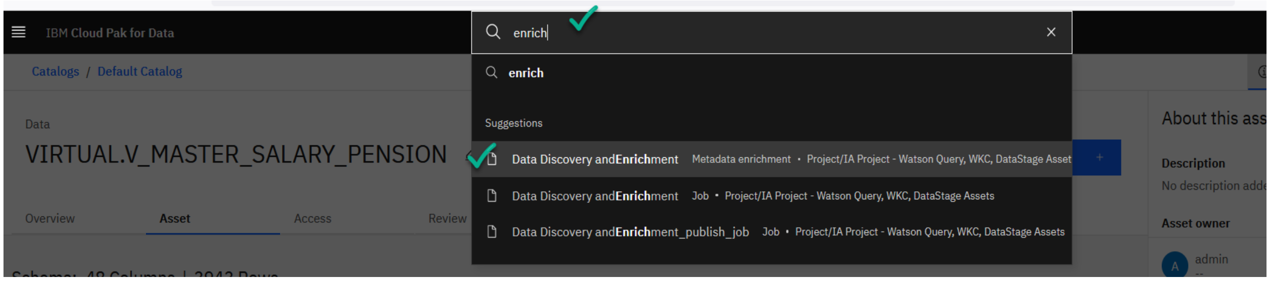
The preview includes only a limited set of columns and rows.

Action Ty... String	VIRTUAL_V_MASTER_SALARY_EMP_... String	LAST_NA... String	FIRST_N... String
Not cl...	Identifier	Last N...	First N...
	1555	Martinez	Fulton

1.5 Automated Data Discovery and Enrichment

Now let’s explore the process that automatically ‘stamps’ those Data Classes and Business Terms onto our Data Assets

__14) Position the cursor back on the search bar at the top of the page, and type ‘**enrich’,** then click on the 1st entry in the dropdown.

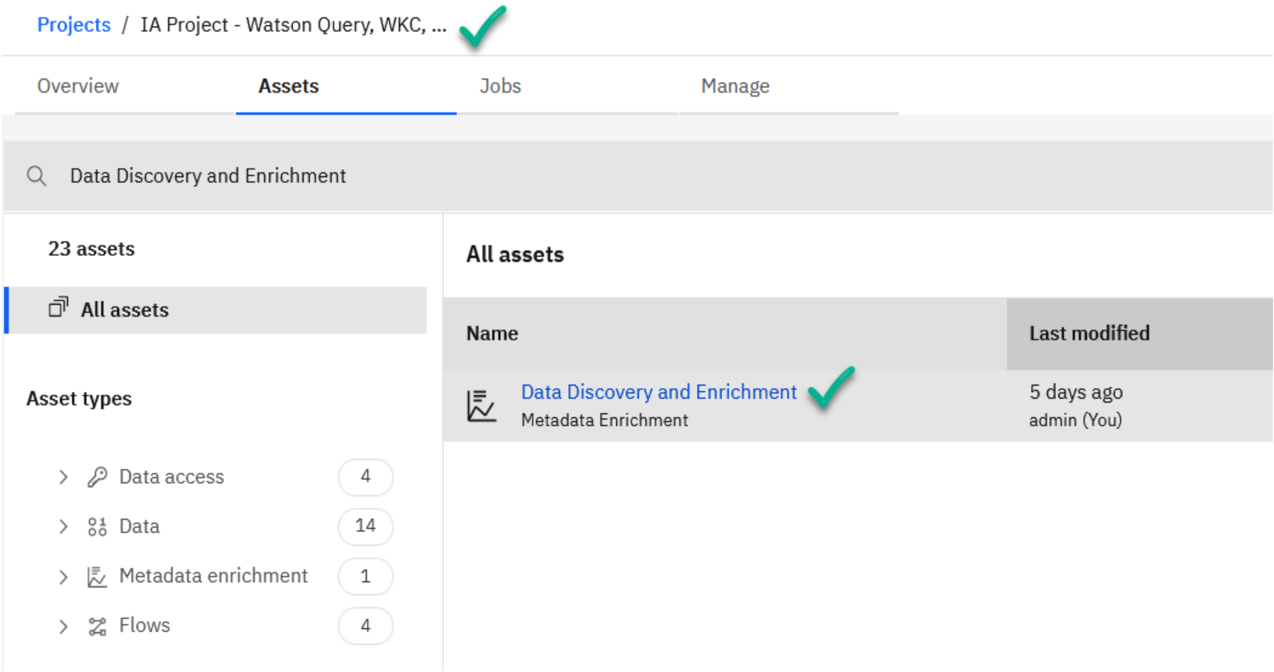


Data Discovery and**Enrich**ment Metadata enrichment • Project/IA Project - Watson Query, WKC, DataStage Asset

Data Discovery and**Enrich**ment Job • Project/IA Project - Watson Query, WKC, DataStage Assets

Data Discovery and**Enrich**ment_publish_job Job • Project/IA Project - Watson Query, WKC, DataStage Assets

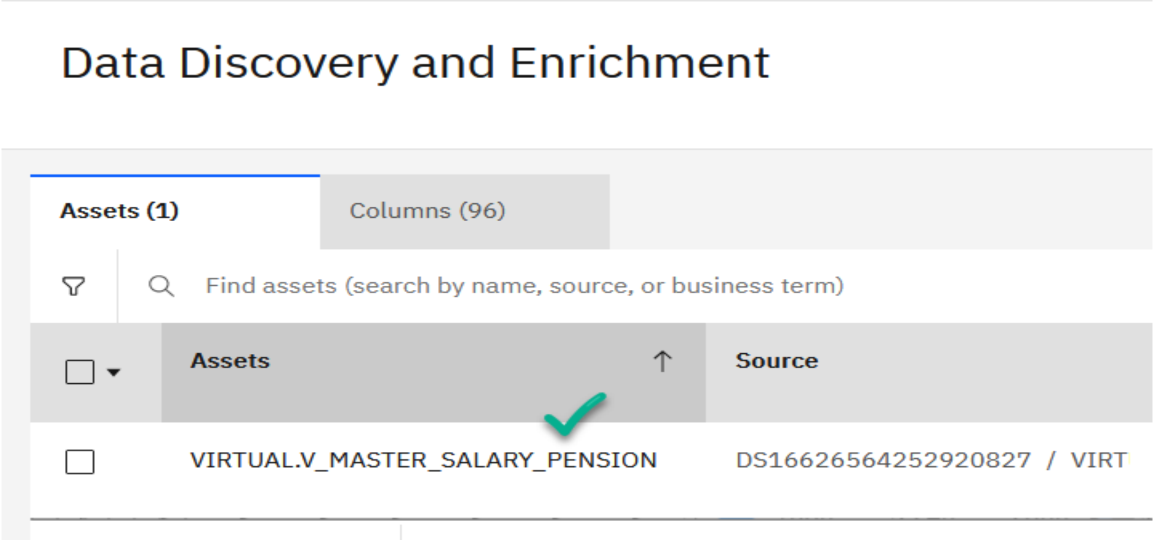
__15) This will take you to the metadata enrichment job. Note that we are now inside a project called **‘IA Watson Query, WKC…**’ Click on the ‘**Data Discovery and Enrichment’** link in the middle of the screen.


Projects / IA Project - Watson Query, WKC, ...
Overview
Assets
Jobs
Manage
Data Discovery and Enrichment
23 assets
All assets
Asset types
All assets

Asset types	
Data access	4
Data	14
Metadata enrichment	1
Flows	4

Name	Last modified
Data Discovery and Enrichment Metadata Enrichment	5 days ago admin (You)

__16) Here, you'll see any assets that are in this Data Discovery job. You can do discovery on schemas, databases, object storage, and tables – here we've done it on a single table **- VIRTUAL.V_MASTER_SALARY_PENSION.** Click on the link.

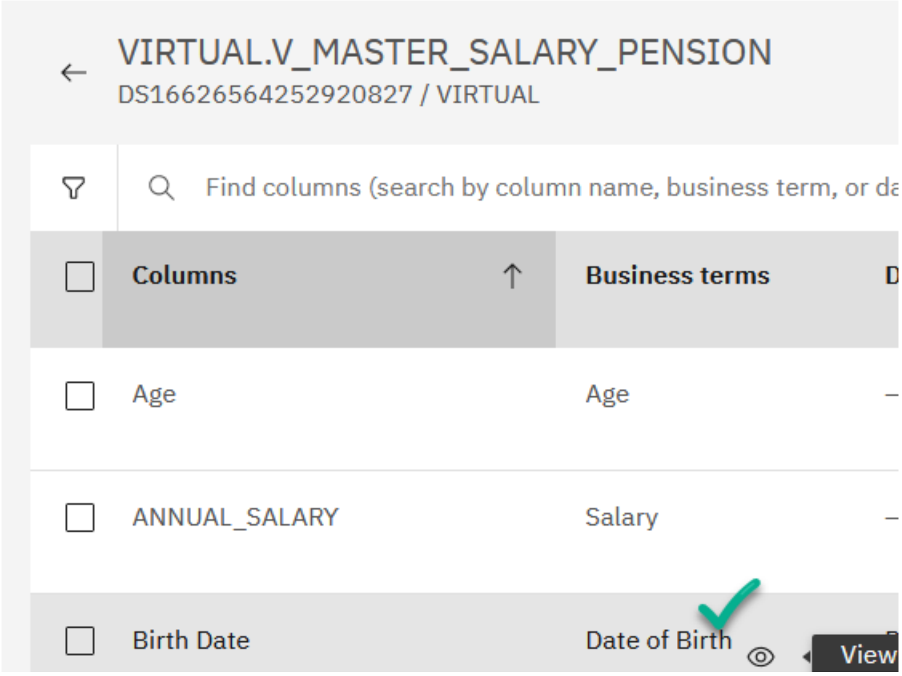


Data Discovery and Enrichment

☐	Assets ↑	Source
☐	VIRTUAL.V_MASTER_SALARY_PENSION	DS16626564252920827 / VIRT

Here we see a lot of information about our data. There is a column that indicates if any **Business Term** have been auto-assigned, as well as whether and **Data Classes** have been assigned.

__17) Click on the 👁 symbol next to the '**Date of Birth**' business term.



VIRTUAL.V_MASTER_SALARY_PENSION
DS16626564252920827 / VIRTUAL

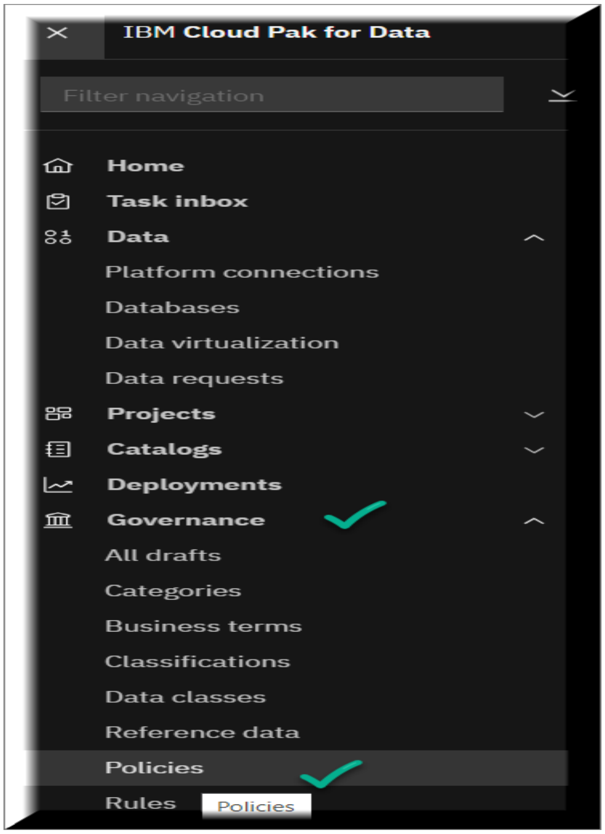
☐	Columns ↑	Business terms	D
☐	Age	Age	–
☐	ANNUAL_SALARY	Salary	–
☐	Birth Date	Date of Birth 👁	

This will open up a side panel that shows more details about the Term and the Class, and allows you to modify, add or delete terms and data classes. This is at the heart of the ML/AI capabilities built into Cloud Pak – if I change these to terms or classes that are more accurate for my organization, that will update the machine learning algorithms that decide what data should be matched with what terms and categories, which will make the algorithm more tuned to my organizations' naming conventions, business practices, etc.

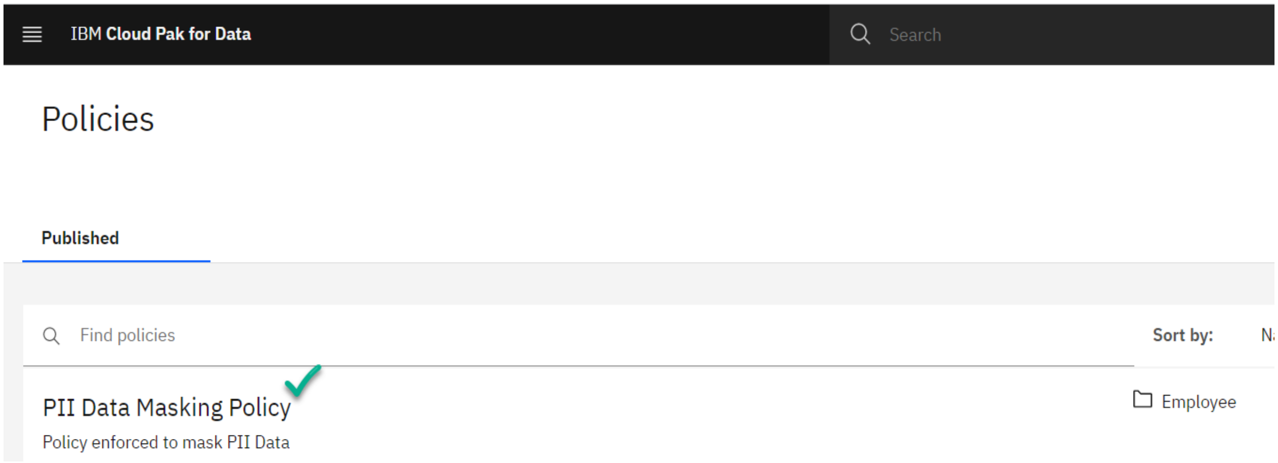
1.6 Data Policies and Data Protection Rules – The Enforcers

OK so now that we've covered governance artifacts like Business Terms, etc, the Data Catalog, and Data Discovery and Enrichment, we now get to the step where we will actually enforce data privacy, by user group, to ensure that only the correct users will have access to the PII data embedded in our data sources.

__18) Start by clicking the hamburger menu icon ≡ in the upper left corner of the screen. Then choose Governance->Policies



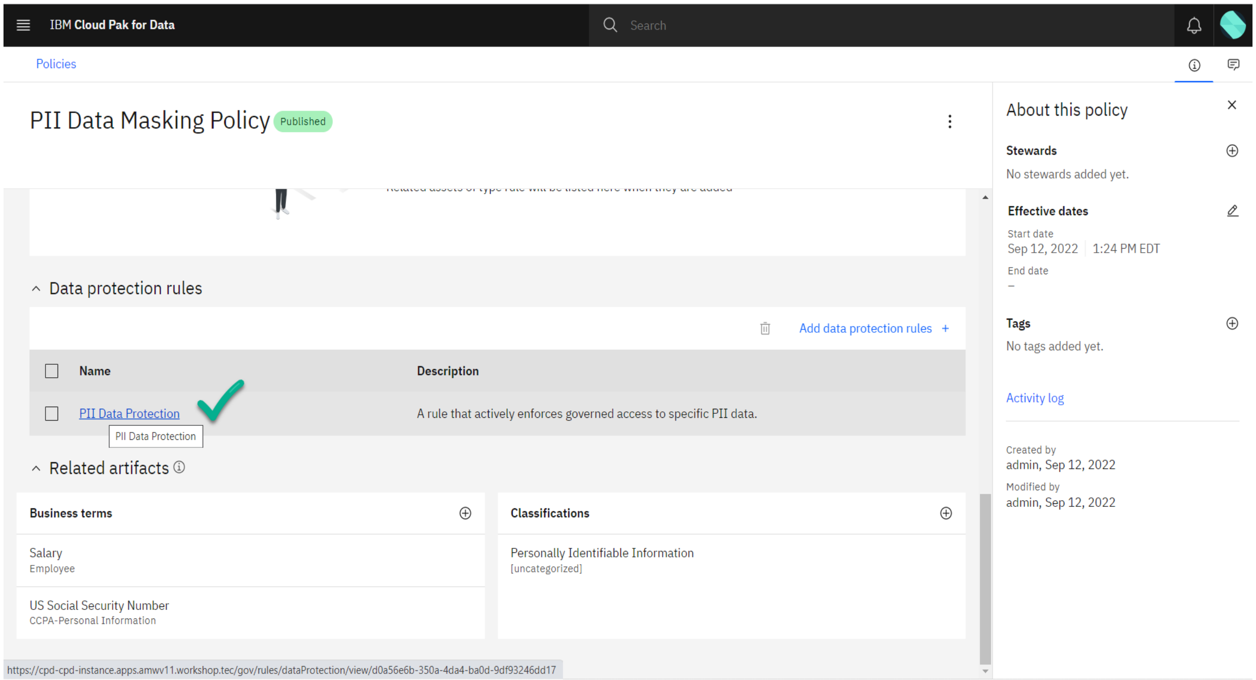
__19) This will bring you into a list of policies that have been developed for our organization. Choose ‘PII Data Masking Policy’



PII Data Masking Policy ✓
Policy enforced to mask PII Data
Employee

When you open the policy, you will see a brief description, and if you scroll down you will see an entry in the ‘Data Protection Rules’ section for **PII Data Protection**. This is the rule that actively enforces data protection. You’ll also notice that there are Business Terms associated with this policy; that is to aid in searching for the Policy, and understanding what the policy relates to. Now we’ll take a look at the rule.

__20) Click on the **PII Data Protection** link –
Note: Scroll down on the page to find the PII Data Protection Link; it is near the bottom.



Name	Description
□ PII Data Protection ✓	A rule that actively enforces governed access to specific PII data.

PII Data Protection

Business terms	Classifications
Salary Employee	Personally Identifiable Information [uncategorized]
US Social Security Number CCPA-Personal Information	

Stewards: No stewards added yet.

Effective dates: Start date Sep 12, 2022 1:24 PM EDT; End date –

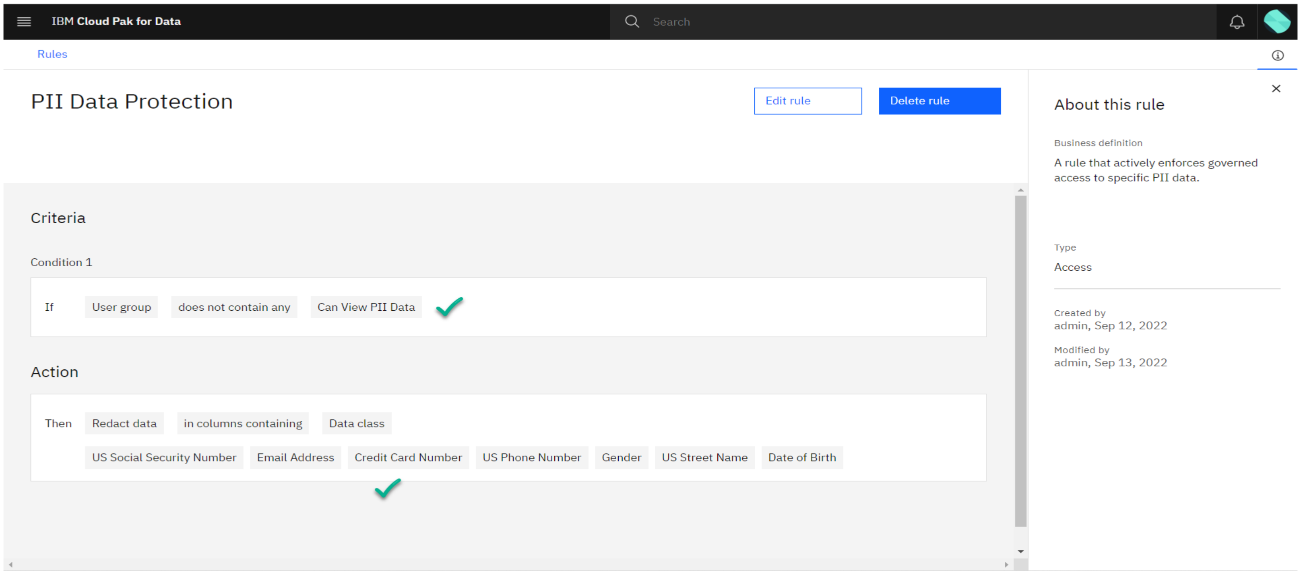
Tags: No tags added yet.

Created by admin, Sep 12, 2022

Modified by admin, Sep 12, 2022

https://cpd-cpd-instance.apps.amwv11.workshop.tec/gov/rules/dataProtection/view/d0a56e6b-350a-4da4-ba0d-9df93246dd17

Here you’ll see an if-then rule that defines in what cases this rule needs to be enforced, and what are the actions to enforce. This rule will enforce redaction of any data that has the listed data classes, IF the user does not belong to the listed user group.



IBM Cloud Pak for Data

Rules

PII Data Protection

Criteria

Condition 1

If User group does not contain any Can View PII Data

Action

Then Redact data in columns containing Data class

US Social Security Number Email Address Credit Card Number US Phone Number Gender US Street Name Date of Birth

About this rule

Business definition

A rule that actively enforces governed access to specific PII data.

Type

Access

Created by

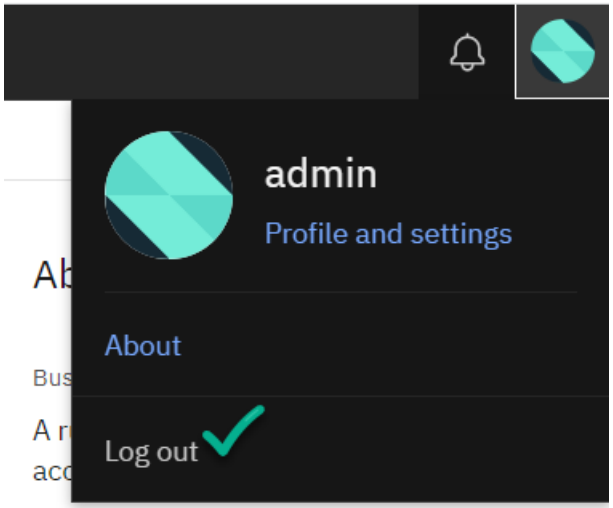
admin, Sep 12, 2022

Modified by

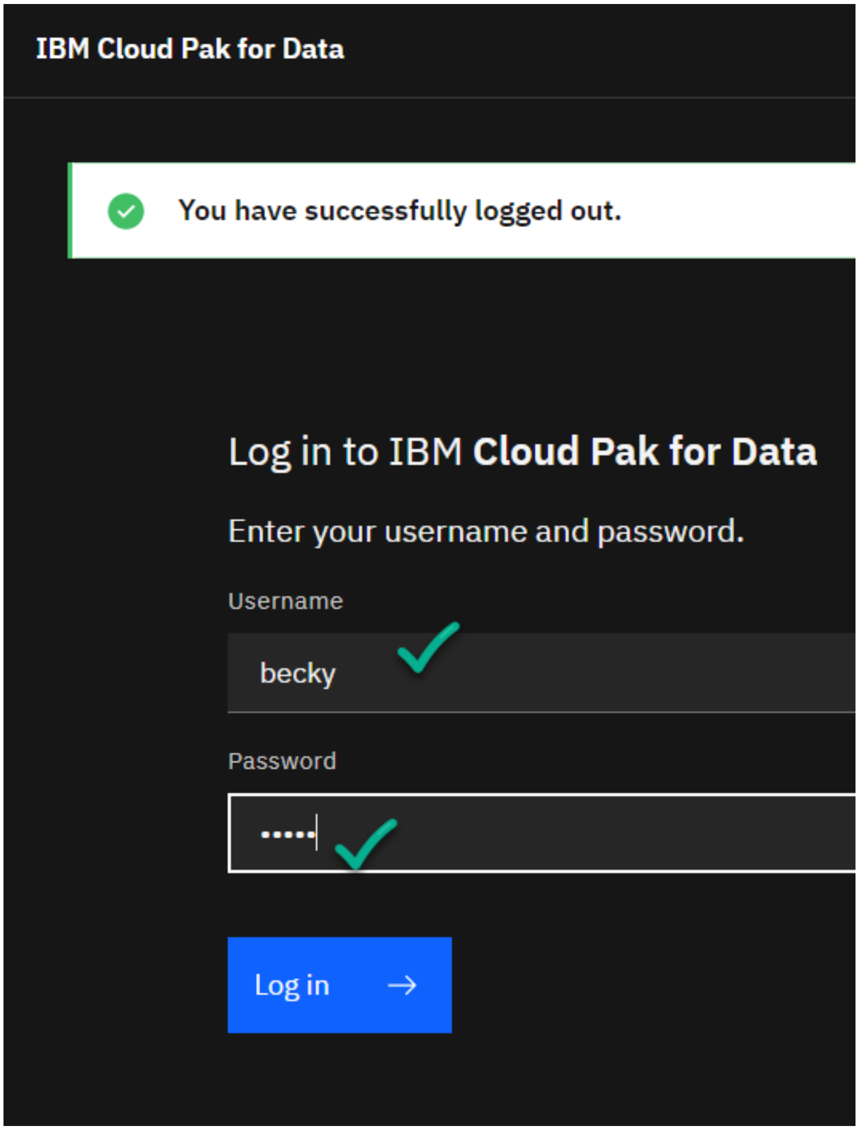
admin, Sep 13, 2022

Remember our exercise around looking at data classes on the data in our table named **VIRTUAL.V_MASTER_SALARY_PENSION?** There were classes for **US Social Security Number**, **Phone Number**, etc. Now those classes stamped on that data will be automatically checked, and if necessary for a given user that data will be redacted. Let's take a look.

__21) Log out of the Admin user by clicking on the icon in the upper right hand corner of the screen, and choosing 'Log Out'.



__22) We will be logging back in as a different user, one who does not belong to the **Can View PII Data** user group. You will find yourself back at the login screen. Login as user **becky,** password **becky.**



IBM Cloud Pak for Data

You have successfully logged out.

Log in to IBM **Cloud Pak for Data**

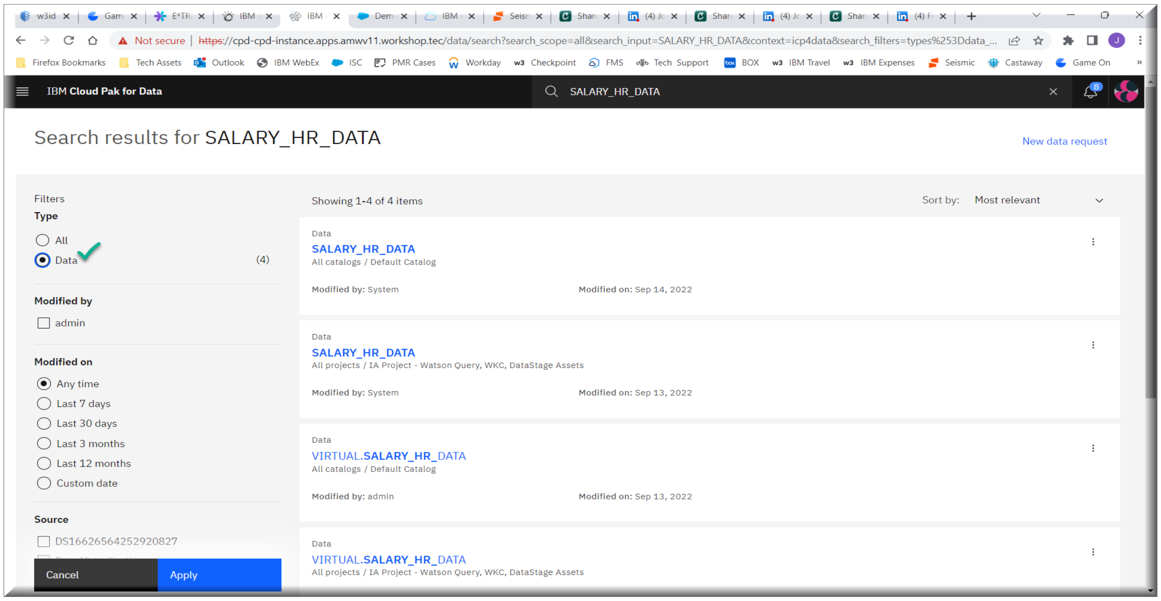
Enter your username and password.

Username: becky

Password: •••••

__23) Now, let's go find **SALARY_HR_DATA.** Go to your search bar, and type in or copy paste '**SALARY_HR_DATA'**. Then press the enter key.

__24) You'll see a list of objects that satisfy our search. Let's filter down to just data in our catalog. So click the radio button next to 'Data' on the left hand filters area, then scroll down and click 'Default Catalog'. Then click 'Apply'.



Search results for SALARY_HR_DATA

Filters

Type

○ All
● Data (4)

Modified by

□ admin

Modified on

● Any time
○ Last 7 days
○ Last 30 days
○ Last 3 months
○ Last 12 months
○ Custom date

Source

□ DS16626564252920827

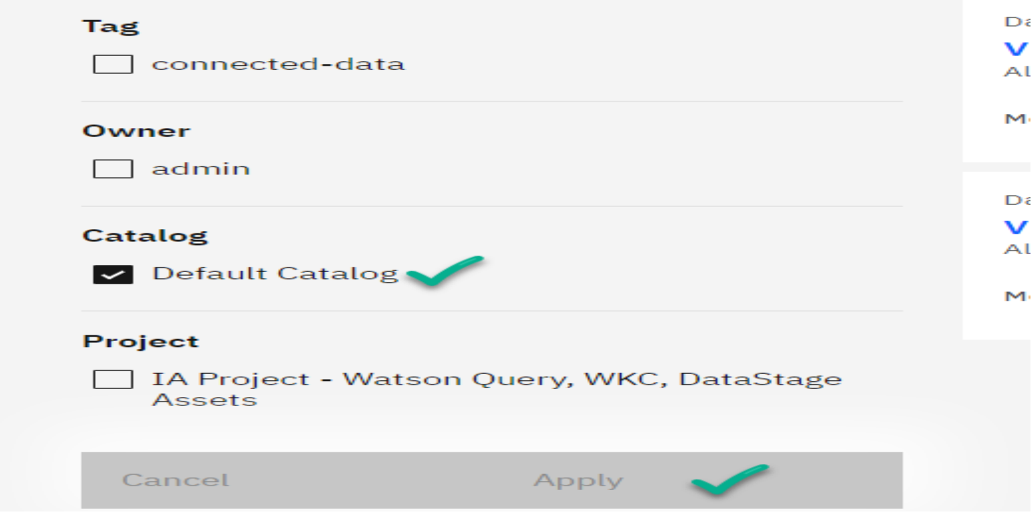
Showing 1-4 of 4 items — Sort by: Most relevant

Data
SALARY_HR_DATA
All catalogs / Default Catalog
Modified by: System — Modified on: Sep 14, 2022

Data
SALARY_HR_DATA
All projects / IA Project - Watson Query, WKC, DataStage Assets
Modified by: System — Modified on: Sep 13, 2022

Data
VIRTUAL.SALARY_HR_DATA
All catalogs / Default Catalog
Modified by: admin — Modified on: Sep 13, 2022

Data
VIRTUAL.SALARY_HR_DATA
All projects / IA Project - Watson Query, WKC, DataStage Assets



Tag

□ connected-data

Owner

□ admin

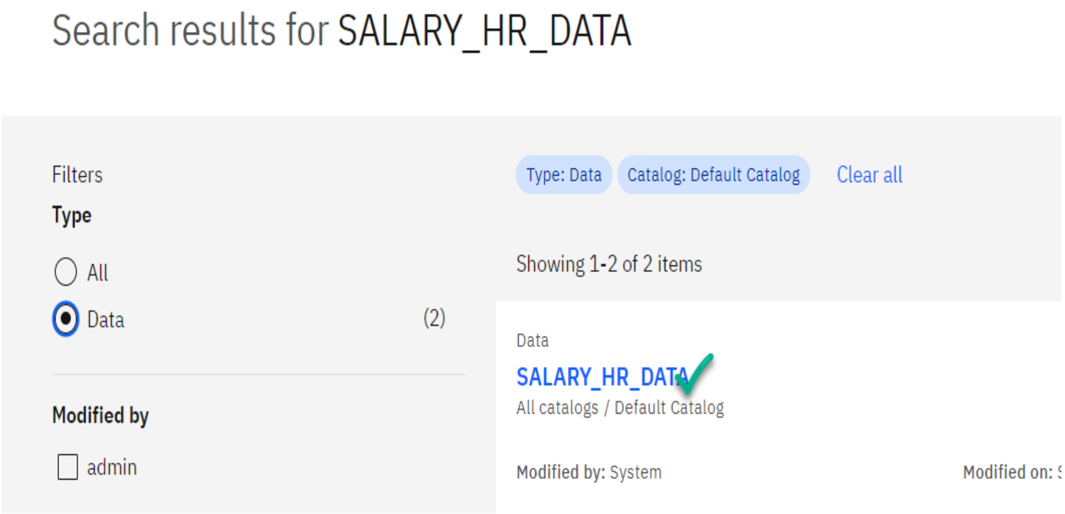
Catalog

☑ Default Catalog

Project

□ IA Project - Watson Query, WKC, DataStage Assets

__25) Now click the entry named **SALARY_HR_DATA**



Search results for SALARY_HR_DATA

Filters

Type

○ All

● Data (2)

Modified by

□ admin

Type: Data | Catalog: Default Catalog | Clear all

Showing 1-2 of 2 items

Data

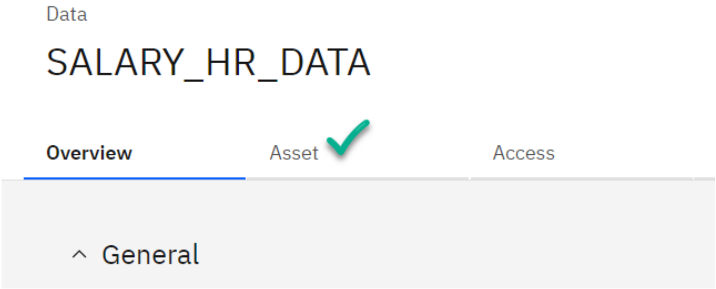
SALARY_HR_DATA ✓

All catalogs / Default Catalog

Modified by: System

Modified on: S

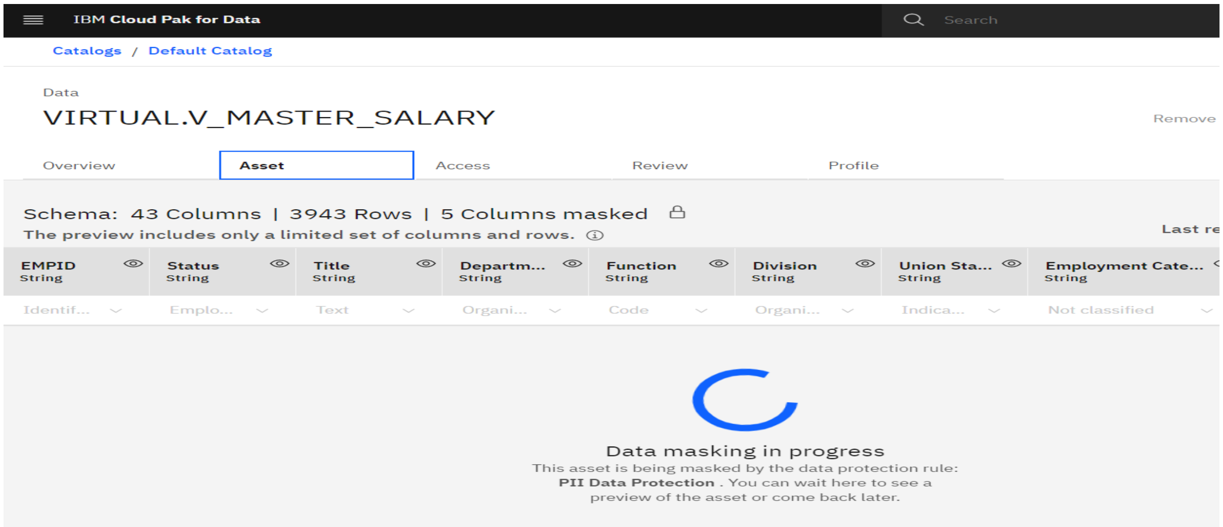
__26) This will bring up the table in our catalog. Click on the 'Asset' tab to get to the data.



Data

SALARY_HR_DATA

__27) At this time, you may get a screen indicating that the data is being masked, wait a few moments and that will finish. That only happens the first time a new user accesses a new catalog data set.



Data

VIRTUAL.V_MASTER_SALARY

Schema: 43 Columns | 3943 Rows | 5 Columns masked

The preview includes only a limited set of columns and rows.

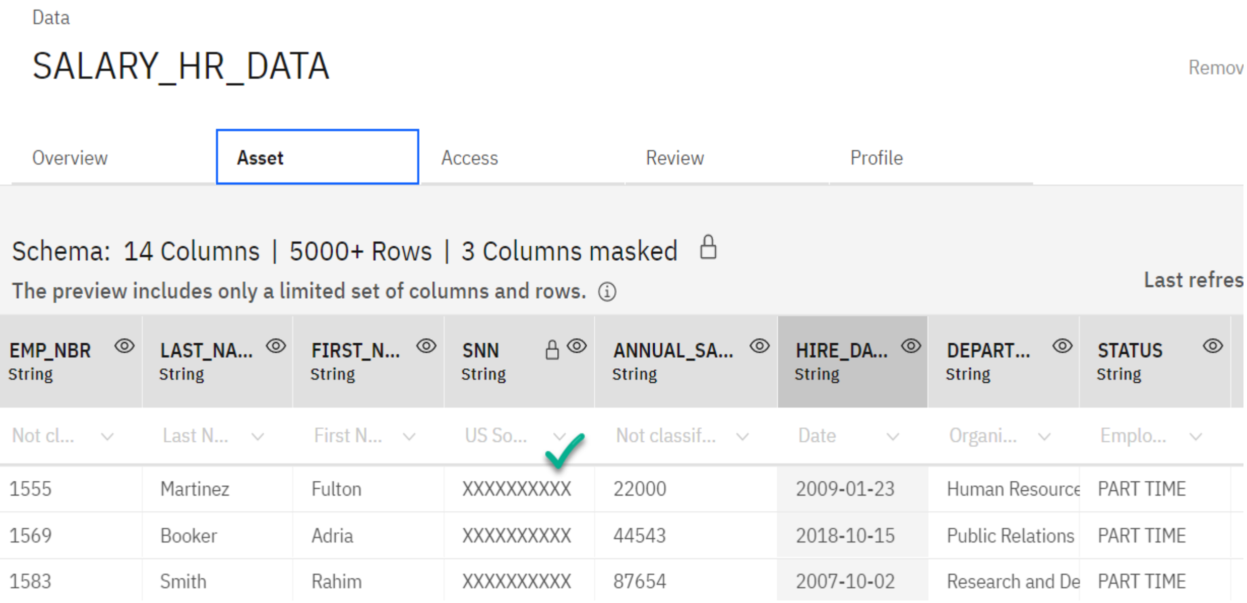
EMPID String	Status String	Title String	Departm... String	Function String	Division String	Union Sta... String	Employment Cate... String
Identif...	Emplo...	Text	Organi...	Code	Organi...	Indica...	Not classified

Data masking in progress

This asset is being masked by the data protection rule: **PII Data Protection** . You can wait here to see a preview of the asset or come back later.

If the data does not come back after the ‘Data Masking’ message disappears, just refresh the page.

When the data comes back, you’ll notice right away that the ‘SNN’ column is all X’s. This is the redaction, because user becky doesn’t have access to PII data, because of our Data Protection rule. Feel free to scroll right, you’ll notice the Phone Number and Street Address fields are similarly redacted.



SALARY_HR_DATA

Schema: 14 Columns | 5000+ Rows | 3 Columns masked

The preview includes only a limited set of columns and rows.

Last refres

EMP_NBR String	LAST_NA... String	FIRST_N... String	SNN String	ANNUAL_SA... String	HIRE_DA... String	DEPART... String	STATUS String
Not cl...	Last N...	First N...	US So...	Not classif...	Date	Organi...	Emplo...
1555	Martinez	Fulton	XXXXXXXXXX	22000	2009-01-23	Human Resource	PART TIME
1569	Booker	Adria	XXXXXXXXXX	44543	2018-10-15	Public Relations	PART TIME
1583	Smith	Rahim	XXXXXXXXXX	87654	2007-10-02	Research and De	PART TIME

Now, let’s move on to see how we can distribute this clean, protected, governed data to our end users, quickly and easily…

NOTE: The grid view showing the data above may not populate immediately; it won’t affect any of the next steps. If you want to return to this step later, you will see the results shown.

1.7 Data Virtualization with Watson Query

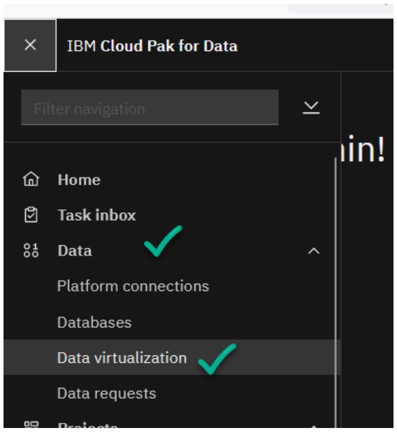
Now we have a little confession to make. Throughout this lab we’ve been working with a ‘table’ called **VIRTUAL.V_MASTER_SALARY_PENSION.** Truth is, this object isn’t a table at all. It’s a view. But the really cool part is this – it’s a view that’s joining data from DB2 and Postgres ***without moving or copying data from either of those sources***. This process is known as ‘Data Virtualization’. The product name is Watson Query.

Here's what’s really unique about Watson Query: It works in conjunction with Watson Knowledge Catalog. It allows you to bring the power of WKC’s data governance capabilities - Data Discovery and Enrichment, Business Terms, and Data Classes – that we’ve been working with in this lab - to any virtualized data source you’ve created in Watson Query *from any tool.* As long as the product (reporting tools, ETL tools, data visualization, etc) can make a jdbc connection, you can have enforced data governance within that tool. This is a game changer: most governance solutions require you to use tools provided by the vendor, and specific, often proprietary, access mechanisms to the data, in order to get the enforcement and governance capabilities of their solution. And even then it usually is not granular to the row and column level. With Watson Query and Watson Knowledge Catalog, you can use the tools you're already comfortable with in your organization.

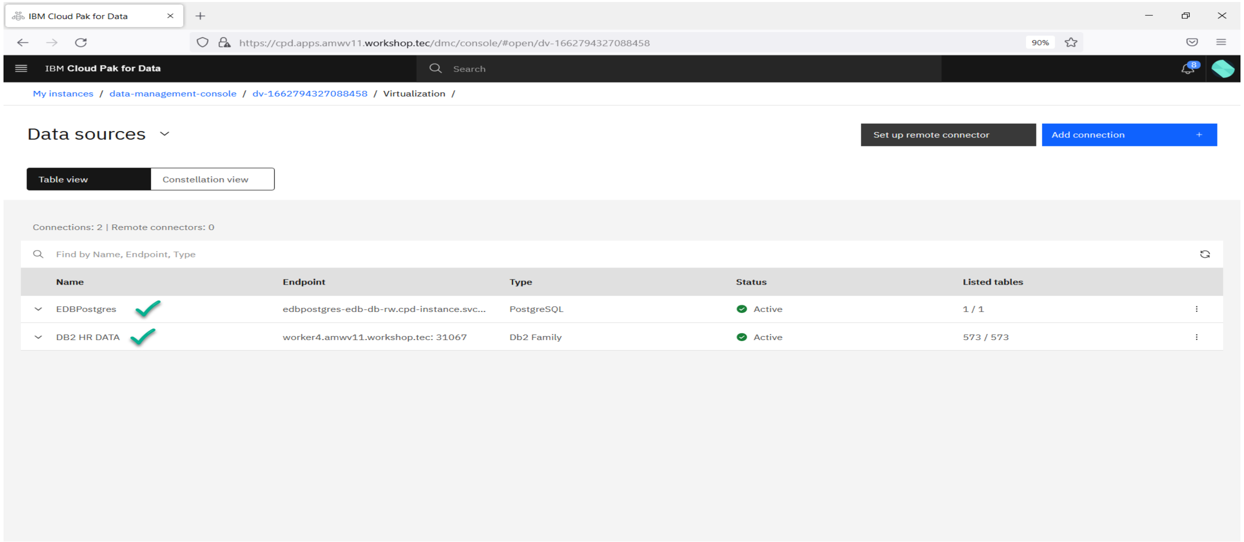
Now let's see how easy it is to virtualize some data.

__28) If you are logged in as Becky, logoff and log in as Admin.

Click on the hamburger icon in the upper left corner of your screen, then choose **Data->Data Virtualization**



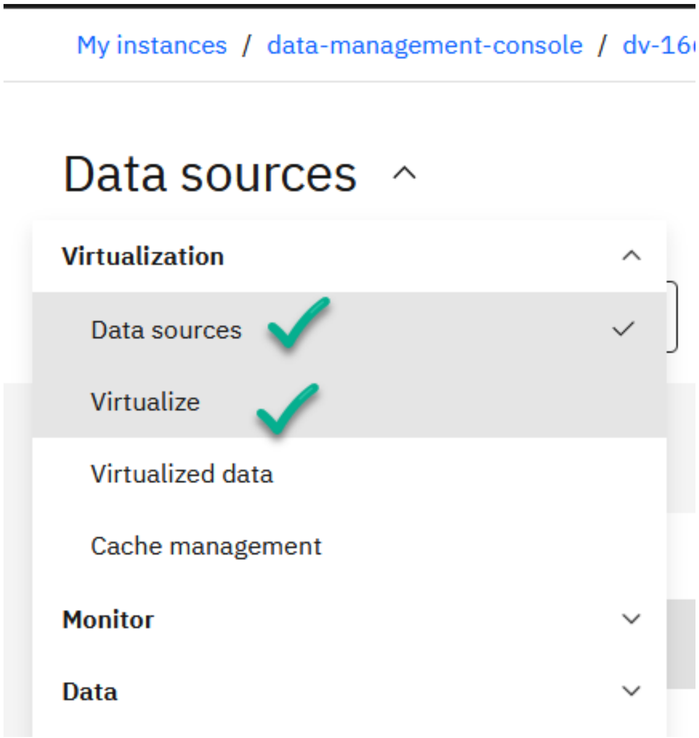
This will open a screen showing you the Data Sources that are currently defined in Watson Query (AKA Data Virtualization). Notice we have 2 – one for DB2 and one for Postgres. Our data will be a join from these 2 separate sources


IBM Cloud Pak for Data
https://cpd.apps.amwv11.workshop.tec/dmc/console/#open/dv-1662794327088458
90%
IBM Cloud Pak for Data
Search
My instances / data-management-console / dv-1662794327088458 / Virtualization /
Data sources
Set up remote connector
Add connection +
Table view
Constellation view
Find by Name, Endpoint, Type

Connections: 2 | Remote connectors: 0

Name	Endpoint	Type	Status	Listed tables
EDBPostgres	edbpostgres-edb-db-rw.cpd-instance.svc...	PostgreSQL	Active	1 / 1
DB2 HR DATA	worker4.amwv11.workshop.tec: 31067	Db2 Family	Active	573 / 573

__29) Next we'll navigate to where we define our virtualized data, using these sources. Drop down he 'Data Sources' and choose 'Virtualize'



My instances / data-management-console / dv-16

Data sources ^

Virtualization ^

Data sources ✓

Virtualize

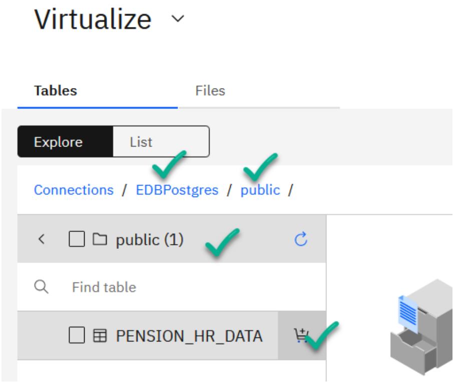
Virtualized data

Cache management

Monitor ˅

Data ˅

__30) Next, click on the **EDB Postgres,** then the **public schema,** and then click the shopping cart icon next to **PENSION_HR_DATA.** This will add our first table to the shopping cart.



Virtualize ˅

Tables | Files

Explore | List

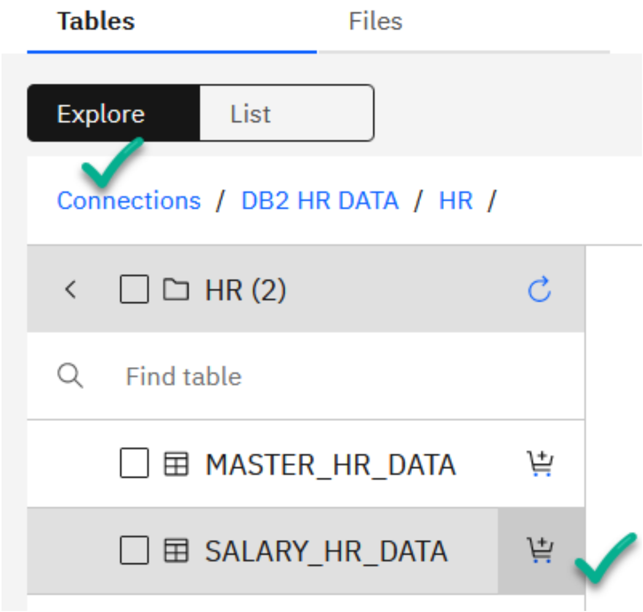
Connections / EDBPostgres / public /

< □ public (1)

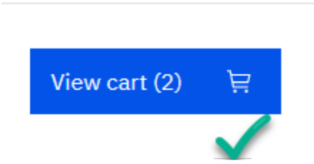
Find table

□ PENSION_HR_DATA

__31) Now, repeat the process – click the link for **'Connections',** then choose **DB2 HR Data**, then the **HR** schema, and then click the shopping cart next to **SALARY_HR_DATA.**

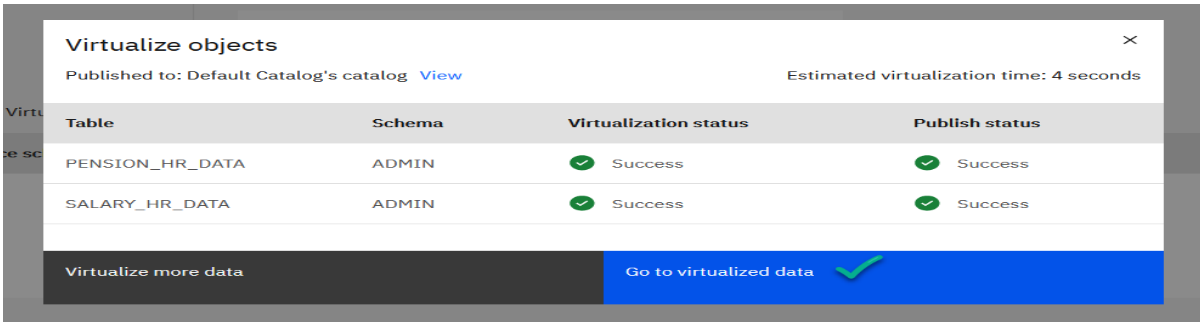


__32) We now have added 2 tables to our shopping cart. Click on the shopping cart in the upper left corner of your screen.



__33) Next, choose the radio button next to **'Virtualized Data',** and then click **'Virtualize'.**

__34) This will virtualize those 2 tables, meaning pointers are created within the Watson Query layer to those 2 tables, and then it will publish those assets to the Catalog. When that is finished, click 'Go to Virtualized Data'.



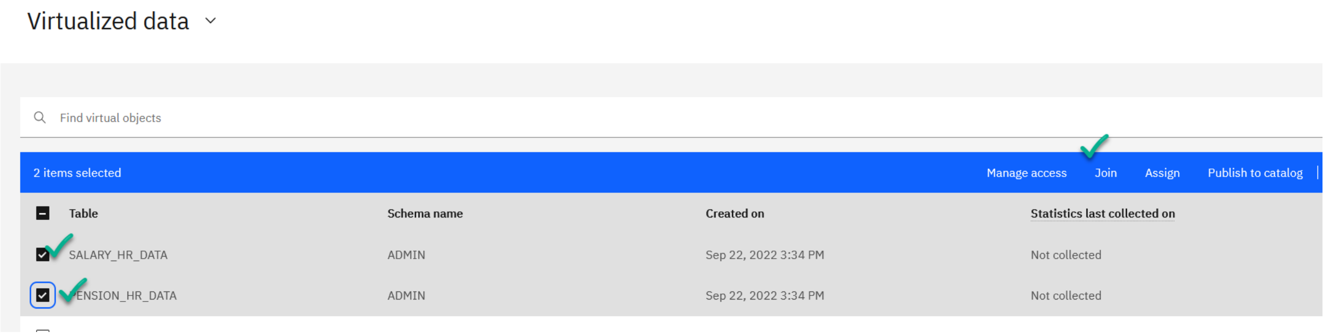
Virtualize objects

Published to: Default Catalog's catalog View

Estimated virtualization time: 4 seconds

Table	Schema	Virtualization status	Publish status
PENSION_HR_DATA	ADMIN	✓ Success	✓ Success
SALARY_HR_DATA	ADMIN	✓ Success	✓ Success

__35) This will bring you to a screen with all the tables and views that have been set up. The 2 you just virtualized are at the top. Click the check box next to each of those tables, and then click the **Join** link at the top of the grid.



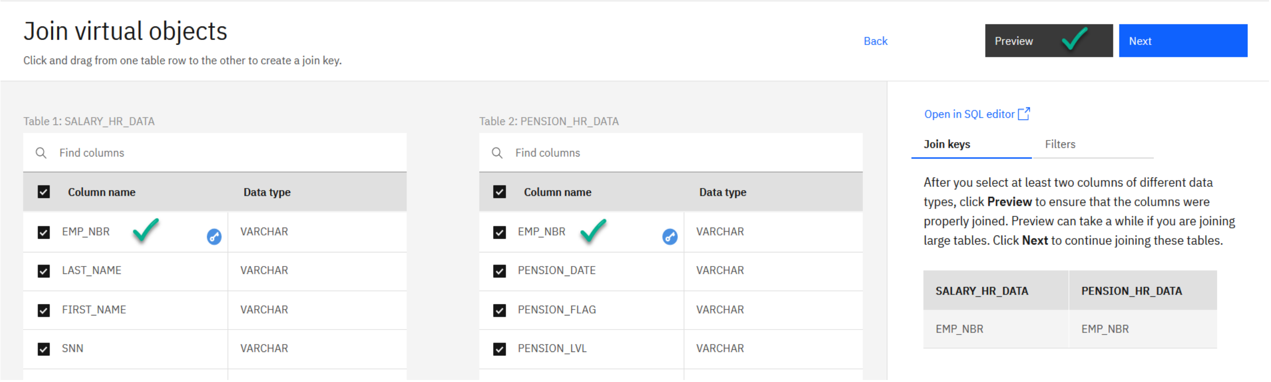
Virtualized data

Find virtual objects

2 items selected

	Table	Schema name	Created on	Statistics last collected on
☑	SALARY_HR_DATA	ADMIN	Sep 22, 2022 3:34 PM	Not collected
☑	PENSION_HR_DATA	ADMIN	Sep 22, 2022 3:34 PM	Not collected

__36) Now you'll define the join criteria. Click and hold the mouse button over **EMP_NBR** on the left table, and drag and drop the link onto **EMP_NBR** on the right table. Then click **PREVIEW.**



Join virtual objects

Click and drag from one table row to the other to create a join key.

Table 1: SALARY_HR_DATA

Find columns

	Column name	Data type
☑	EMP_NBR	VARCHAR
☑	LAST_NAME	VARCHAR
☑	FIRST_NAME	VARCHAR
☑	SNN	VARCHAR

Table 2: PENSION_HR_DATA

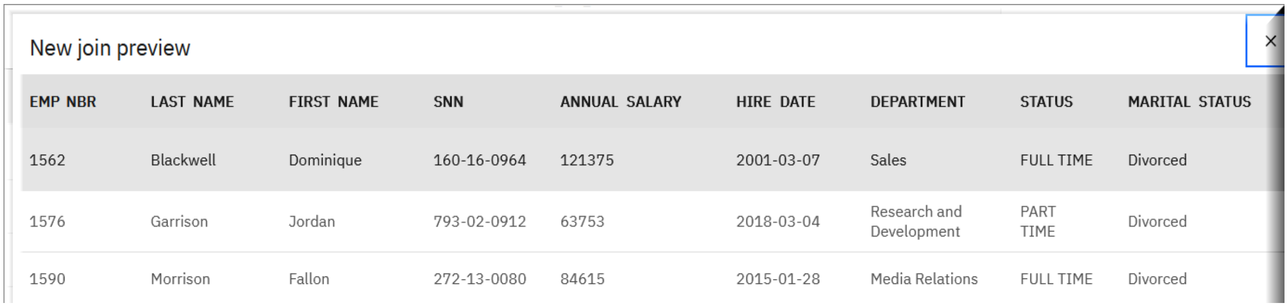
Find columns

	Column name	Data type
☑	EMP_NBR	VARCHAR
☑	PENSION_DATE	VARCHAR
☑	PENSION_FLAG	VARCHAR
☑	PENSION_LVL	VARCHAR

After you select at least two columns of different data types, click **Preview** to ensure that the columns were properly joined. Preview can take a while if you are joining large tables. Click **Next** to continue joining these tables.

SALARY_HR_DATA	PENSION_HR_DATA
EMP_NBR	EMP_NBR

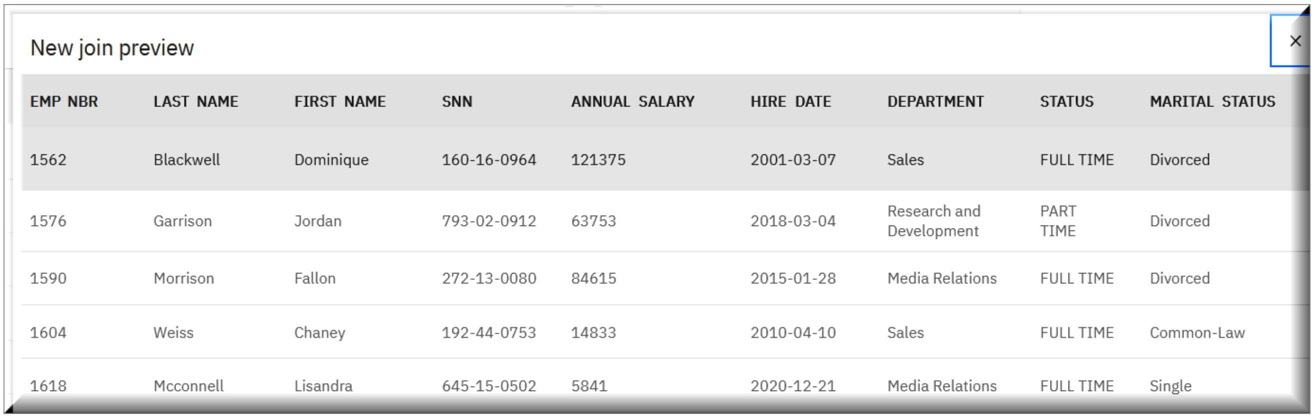
This will then show our resulting join, validate that all the columns have data.



New join preview

EMP NBR	LAST NAME	FIRST NAME	SNN	ANNUAL SALARY	HIRE DATE	DEPARTMENT	STATUS	MARITAL STATUS
1562	Blackwell	Dominique	160-16-0964	121375	2001-03-07	Sales	FULL TIME	Divorced
1576	Garrison	Jordan	793-02-0912	63753	2018-03-04	Research and Development	PART TIME	Divorced
1590	Morrison	Fallon	272-13-0080	84615	2015-01-28	Media Relations	FULL TIME	Divorced

If we published this, it would then be available to any users who are given access, and the data protection rules would be enforced automatically.



New join preview

EMP NBR	LAST NAME	FIRST NAME	SNN	ANNUAL SALARY	HIRE DATE	DEPARTMENT	STATUS	MARITAL STATUS
1562	Blackwell	Dominique	160-16-0964	121375	2001-03-07	Sales	FULL TIME	Divorced
1576	Garrison	Jordan	793-02-0912	63753	2018-03-04	Research and Development	PART TIME	Divorced
1590	Morrison	Fallon	272-13-0080	84615	2015-01-28	Media Relations	FULL TIME	Divorced
1604	Weiss	Chaney	192-44-0753	14833	2010-04-10	Sales	FULL TIME	Common-Law
1618	Mcconnell	Lisandra	645-15-0502	5841	2020-12-21	Media Relations	FULL TIME	Single

1.8 Lab Conclusion

You've learned a lot today. Good job. Cloud Pak for Data has a *lot* more capabilities, which you will see if you continue working through the labs in this workshop. This lab is just the beginning!

Speaking of which, if you are continuing on to the next lab, you will see that the business analyst using Cognos to gain some insights into the business will be using the same virtualized tables we used here, and you'll be able to see the automated masking and data privacy up close in a real world situation, using Cognos Reporting.