

Lab Center – Hands-On Lab

Session 5749

Classify Your Sensitive Data Assets Automatically
Using IBM Watson Knowledge Catalog



DISCLAIMER

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenShift is a trademark of Red Hat, Inc.

UNIX is a registered trademark of The Open Group in the United States and other countries.

© 2020 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

We Value Your Feedback

Don't forget to submit your Think 2020 session and speaker feedback! Your feedback is very important to us – we use it to continually improve the conference.

Access the Think 2020 agenda tool to quickly submit surveys from your smartphone or laptop.

Table of Contents

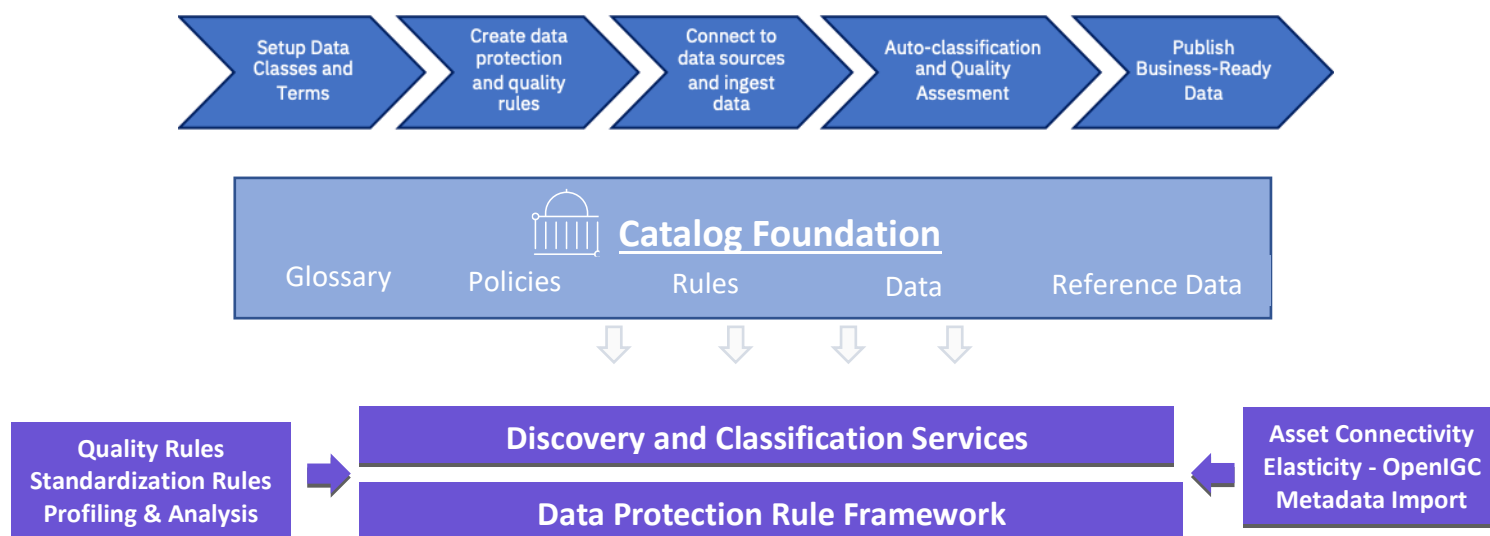
Introduction	5
Getting Started	6
Catalog Foundation	7
1. Define a New Data Class	7
2. Import and Annotate Glossary Terms	9
3. Create a Data Protection Rule	12
4. Create an Automation Rule	14
Data Discovery	15
1. Setting up a Discovery Job	15
2. Review and Publish Discovery Job Results	18
3. Search and Explore Discovered Assets	21

Introduction

IBM Watson Knowledge Catalog allows Data Citizens to search and explore meaningful, trusted and quality data; giving them insight and offering the ability to drive new analytics or support integration and data science.

First, we develop the foundational elements of the Catalog, Terms which give meaning to information; Classifiers for identifying information, including sensitive data; and Rules for enforcement of regulation and policies. Then, we ingest data into the Catalog, further enriching and preparing data through the available Discovery and Classification services and Metadata Curation experience. **Discovery and Classification leverage the Machine Learning (ML) capabilities of the platform to automate the process to assign meaning and identity to data, identify sensitive data and application of data protection rules, and calculate the Quality score and dimension.**

Ultimately, delivering business-ready-data to the Enterprise to facilitate the ability for the Data Citizen to search and explore meaningful, trusted and quality data with deeper insights and ability to advance Analytics and Data Science.

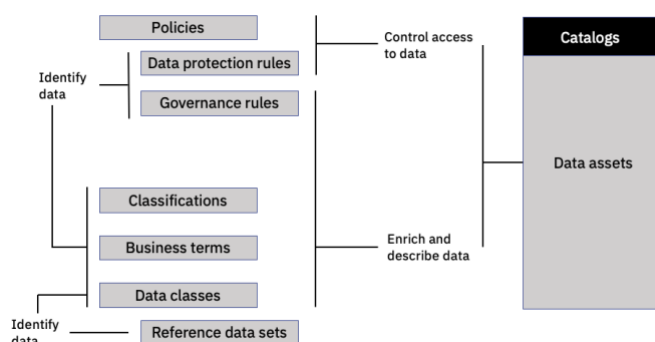


Getting Started

Governance is the process of curating, enriching, and controlling your data. You govern your data with governance artifacts.

You can create these types of governance artifacts:

- **Classifications:** You use classifications to describe the sensitivity of the data in data assets. Each data asset has one classification. Catalog collaborators assign the classification when they add data assets to a governed catalog. You can also assign classifications to governance artifacts, such as business terms, data classes, reference data, policies, and governance rules.
- **Data classes:** You use data classes to categorizes columns in relational data sets according to the type of the data and how the data is used. One data class is assigned to each column during profiling within a catalog. Catalog collaborators can change the data class assigned to a column. Users with the manage discovery permission can assign data classes to data set columns before adding the data to a catalog.
- **Reference data sets:** You create reference data sets to define values for specific types of columns. You can include a reference data set in the definition of a data class as part of the data matching criteria. During data profiling, if the values in a column match the reference data set and other criteria, that data class is assigned to the column. You can also use reference data sets in data quality analysis.
- **Business terms:** You create business terms to define business concepts in a standard way for your enterprise. Catalog collaborators can assign one or more terms to data assets and columns within relational data sets to describe the data. Users with the manage discovery permission can add business terms to data sets before adding them to a catalog.
- **Policies:** You create policies to describe how to govern data in catalogs. You can include data protection rules in policies to control access to data. You can also include governance rules in policies to describe data.
- **Data protection rules:** You create data protection rules to identify the data to control and to specify the method of control. Within data protection rules, you can include classifications, data classes, business terms, or tags to identify the data to control. You can choose to deny access to data or to mask sensitive data values.
- **Governance rules:** You create governance rules to provide a natural-language description of the criteria that are used to determine whether data assets are compliant with business objectives.




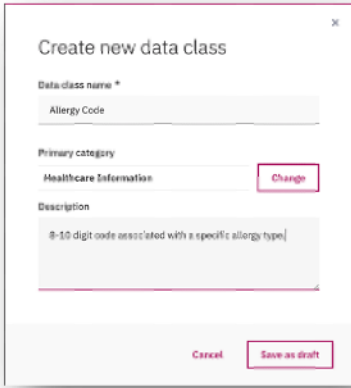
Above diagram shows an overview of how governance artifacts and governance tools work together to describe, enrich, improve, and protect data. Governance artifacts can have more relationships than are shown in this diagram with other artifacts and data assets.

Catalog Foundation

1. Define a New Data Class

This step will allow the user to explore the process for creating a new Data Class to aide in the discovery and classification based upon the newly created Reference Data Set for Country Codes and Names

- 1 Open the navigation menu by selecting the  action and expand the section Organize and further expand the section Data and AI Governance selecting the item Data Classes. This will open the Data Class view.
- 2 Here we can see a set of standard Data Classes that come pre-built in Cloud Pak for Data. We will be creating a new Data Class by selecting the action item *Create Data Class* to initiate the process to create a new Data Class. The create dialog will display.
- 3 Enter a name for the new Data Class *Allergy Code* <Your Name> and add a Description *8-10-digit code associated with specific allergy type*. Select the Category *Healthcare Information*.



- 4 Click *Save as Draft* to complete the creation process. The draft copy of the Data Class appears.
- 5 Select and Edit the *Matching Method* to define the actions of the Data Class:

The first screenshot shows the 'Matching method' selection screen. It has a progress bar at the top with three steps: 'Select matching method' (active), 'Define data matching', and 'Other matching criteria'. Below the progress bar, the title is 'Matching method'. The instruction says: 'Choose how to specify matching criteria. Most methods include data and column matching criteria.' There are five options: 'No automatic matching', 'Match to list of valid values', 'Match to reference data', 'Match to criteria in regular expression' (selected with a blue border and a checkmark), and 'Match to criteria in deployed Java class'. At the bottom right are 'Cancel' and 'Next' buttons.


The second screenshot shows the 'Match to criteria in regular expression' configuration screen. It has the same progress bar. The title is 'Match to criteria in regular expression'. The 'Match criteria for column value' field contains '^[0-9]{8,10}\$'. The 'Test value to match criteria of column value' field contains '232347008'. The 'Percentage match threshold' is set to '80'. Below these fields, it says 'Matches criteria of regular expression'. At the bottom right are 'Back' and 'Next' buttons.

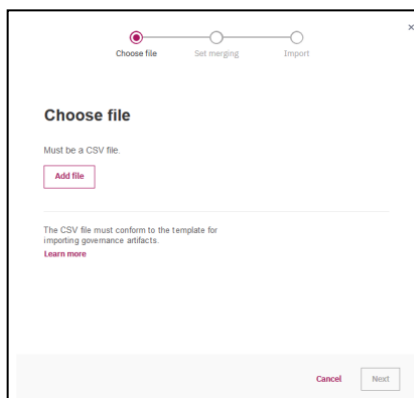
- a. Select the Matching Method *Match to criteria in Regular Expression* and click *Next* to continue.
 - b. Enter `^[0-9]{8,10}$` as the match criteria and use `232347008` as the Test Value. Click *Next*.
 - c. There are no *Other Matching Criteria* required. Click *Save* to complete the action.
- 6 Click *Publish* to publish the Data Class and complete the creation process.

This completes Step 2 and the creation of a Data Class

2. Import and Annotate Glossary Terms

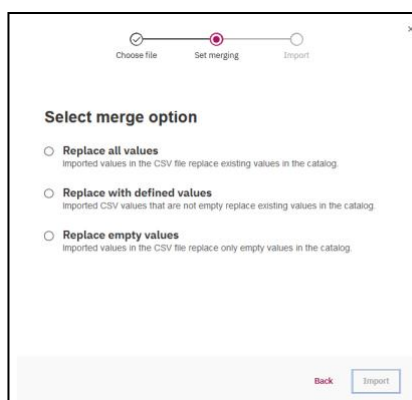
This step will allow the user to explore the process for creating a set of Glossary Terms with existing content, further annotating the created Terms

- 1 Open the navigation menu by selecting the  action (from the upper-left corner of the navigation bar). Expand the section *Organize* and further expand the section *Data and AI Governance* selecting the item *Business Terms*.
- 2 Select the action item *Import* to initiate the import of a set of Glossary Term. The import wizard will display.
- 3 Click *Add File* to browse to the directory *C:/THINK* and select the provided file *healthcare_terms.csv*. The import file has previously been prepared for you and is a

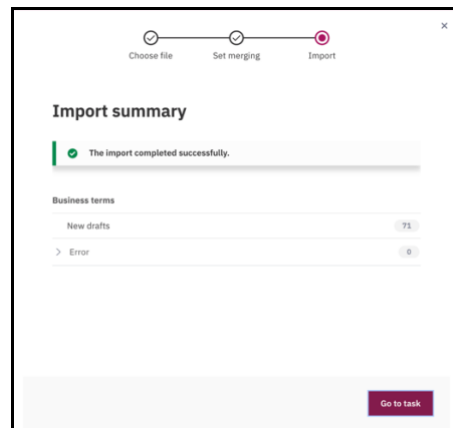


standard text delimited file that allows an author to easily author and import Terms into the Catalog. The format of the import file is documented [here](#). Click *Next* to continue.

- 4 Select the Merge Option *Replace all Values*. This will replace any existing Glossary Terms with the imported Terms.

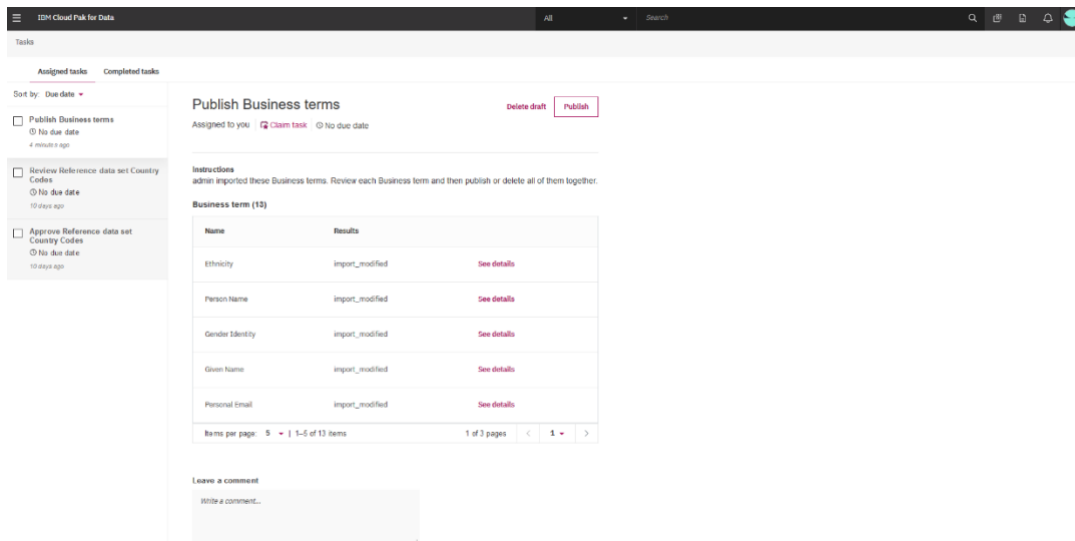


- 5 Click *Import* to continue and finalize the import action. Upon completion, a *Workflow* task is created to manage the reviewal and approval of the imported Terms into the Catalog.

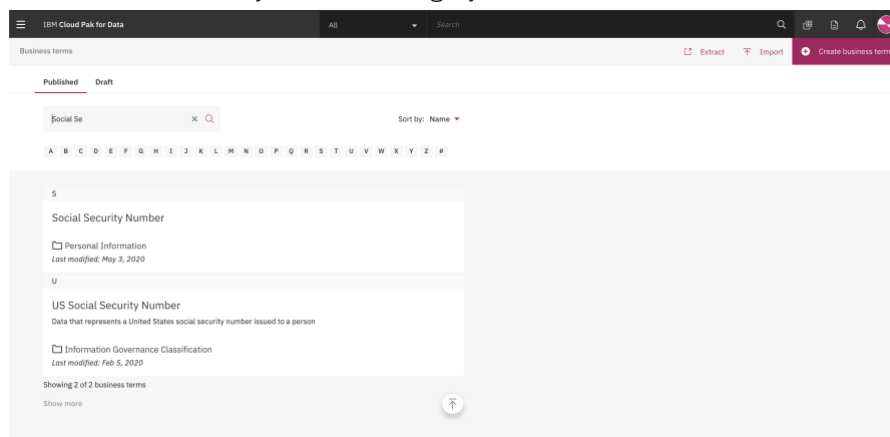


- 6 (Optional) Click *Go to Task* to open the created Workflow Task to publish changes. Note a single Workflow Task was created for the multiple Glossary Terms either created or modified in the Catalog.

--> STOP. For purposes of this Lab you will not publish the imported Terms. <--



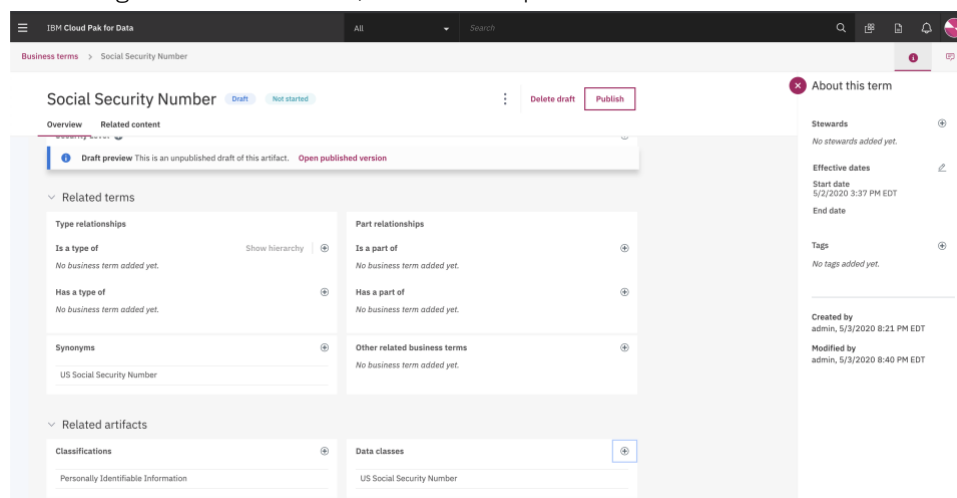
- 7 Navigate back to the Business Term view from the menu on the upper-left *Organize > Data and AI Governance > Business terms*. In the search bar enter *Social Security Number*. Click on the term *Social Security Number* in the *Personal Information* category.



- 8 Scroll down to the *Related Terms* section and add a Synonym by clicking on the \oplus action. In the window that comes up search for *Social security* and select the term *US Social Security Number* and click *Add*.




- 9 In the *Related Artifacts* section, we are going to classify this term as PII and relate it to a data class.
- Add a Classification by clicking on the \oplus action. Select the *Personally Identifiable Information* classification and click *Add*.
 - Add a related Data Class by clicking on the \oplus action. Search for *Social Security* and select the class *US Social Security Number*. Click *Add*. This step will help with auto-classification of data as when an asset meets the data class definition it will also be assigned the business term.
- 10 Once all the changes have been made, *Publish* the updated term.

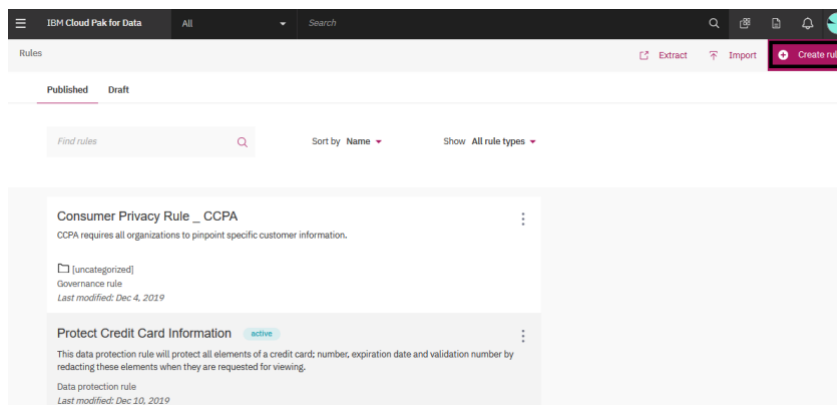


This completes Step 2 and the import and creation of Glossary Terms

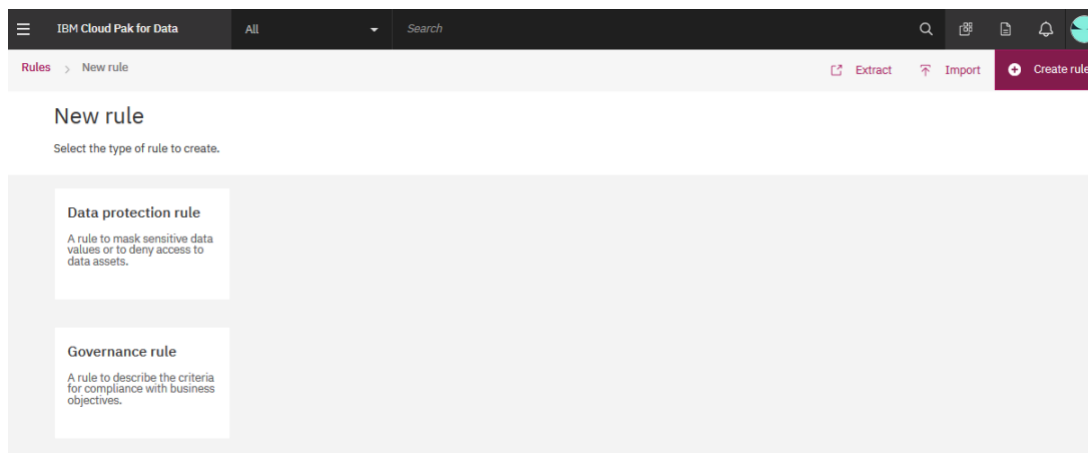
3. Create a Data Protection Rule

This step will allow the user to explore the process for creating a new Data Protection Rule for ensuring that sensitive or personal data is properly handled, and compliancy is met

- 1 Open the navigation menu by selecting the  action and expand the section *Organize* and further expand the section *Data and AI Governance* selecting the item *Rules*.



- 2 Select the action item *Create Rule* to initiate the process to create a new Governance Rule. The New




Rule selection type dialog will display.

- 3 Select *Data Protection Rule* to create a new rule of this type. The Rule create dialog will display.
- 4 Define the following information for the new Data Protection Rule:

The screenshot shows the 'New data protection rule' configuration page in the IBM Cloud Pak for Data interface. The page is divided into three main sections: Details, Criteria, and Action.




- Details:**
 - Name:** Mask SSN and DOB
 - Type:** Access
 - Business definition:** Mask all columns that have been identified to contain a Social Security Number or Date of Birth.
- Criteria:**
 - CONDITION 1:** If Data class contains any US Social Security Number X Date of Birth X.
 - Add condition:** A dashed box with a plus icon and the text 'Add condition'.
- Action:**
 - then:** mask data
 - in columns containing:** US Social Security Number X Date of Birth X
 - Select how to mask data:**
 - Redact:** Replace data with Xs. (Selected)
 - Substitute:** Replace data with values that don't match the original.
 - Obfuscate:** Replace data with similarly formatted values.

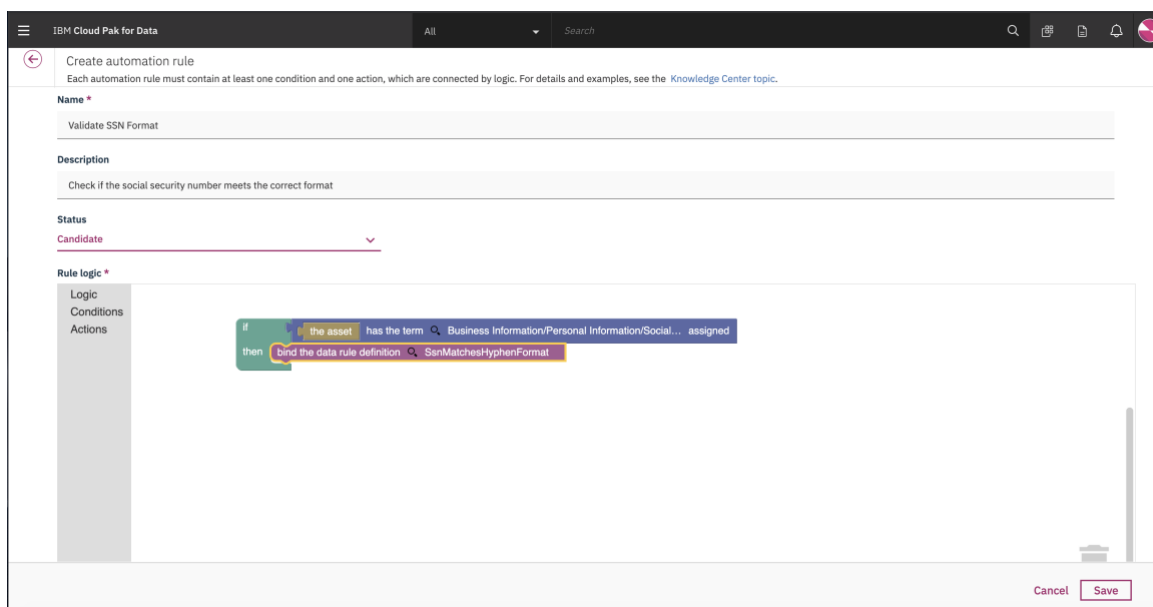
At the bottom right, there are 'Cancel' and 'Create' buttons.

- Set the Name of the Rule to *Mask SSN and DOB <Your Name>*
 - Set the Business Definition of the Rule to *Mask all columns that have been identified to contain a Social Security Number or Date of Birth.*
 - Using the Rule Builder, define a condition *If Data Class, US Social Security Number and Date of Birth, contains any*
 - Using the Action Builder, define an action *Then Mask Data in Columns Containing US Social Security Number and Date of Birth.* Select the masking option of *Redact*.
- Click *Create* to create the new Data Protection Rule. The Data Protection Rule is displayed for further modifications.
 - Click the return arrow  to return to the list of Rules.

This completes Step 3 and the creation of a Data Protection Rule

4. Create an Automation Rule

- 1 Open the navigation menu by selecting the  action (from the upper-left corner of the navigation bar). Expand the section *Organize* and further expand the section *Curation* selecting the item *Automation Rules*.
- 2 Click the *Create automation rule* button on the upper-right to open the rule builder.
- 3 Enter the rule name *Validate SSN Format <Your Name>* and the description *Check if the social security number meets the correct format*. Change the status to *Accepted*.
- 4 We will start building the rule using the rule builder dialog. This follows an If-Then logic – **if** a certain condition is met **then** the action is taken.
 - a. IF: From the sidebar click on *Conditions* and drag over the *asset has term assigned* option and drop it by the if space. Next, click on the magnifying glass icon () , search for and select the *Social Security Number* term. Click *Save*.
 - b. THEN: From the sidebar click on *Actions* and drag over the *bind the data rule* and drop it by the then space. Next, click on the magnifying glass icon () , search for and select the *SsnMatchesHyphenFormat* rule definition Click *Save*.



IBM Cloud Pak for Data

Create automation rule

Each automation rule must contain at least one condition and one action, which are connected by logic. For details and examples, see the [Knowledge Center topic](#).

Name *

Validate SSN Format

Description

Check if the social security number meets the correct format

Status

Candidate

Rule logic *

Logic

Conditions

Actions

if the asset has the term Business Information/Personal Information/Social... assigned

then bind the data rule definition SsnMatchesHyphenFormat

Cancel Save

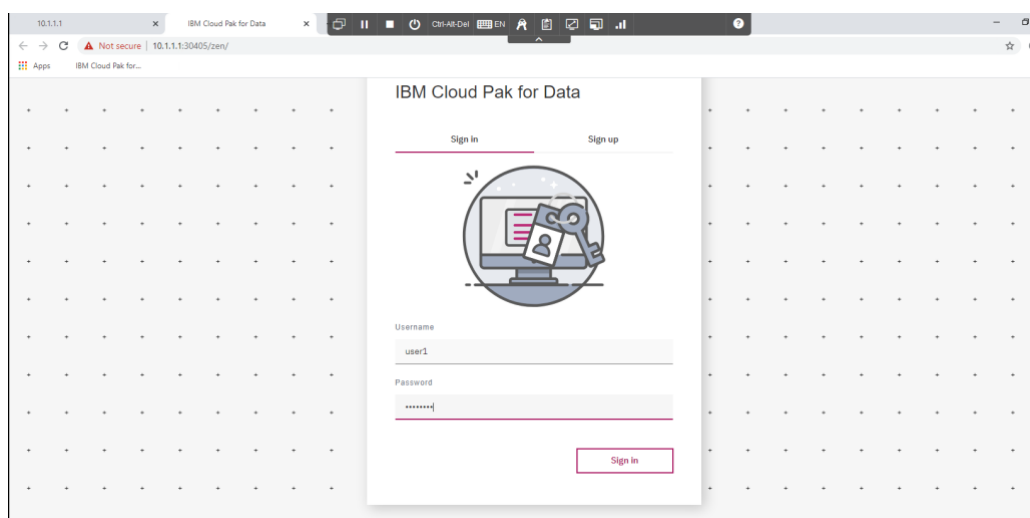
- 5 Click *Save* to create the new automation rule.


Data Discovery

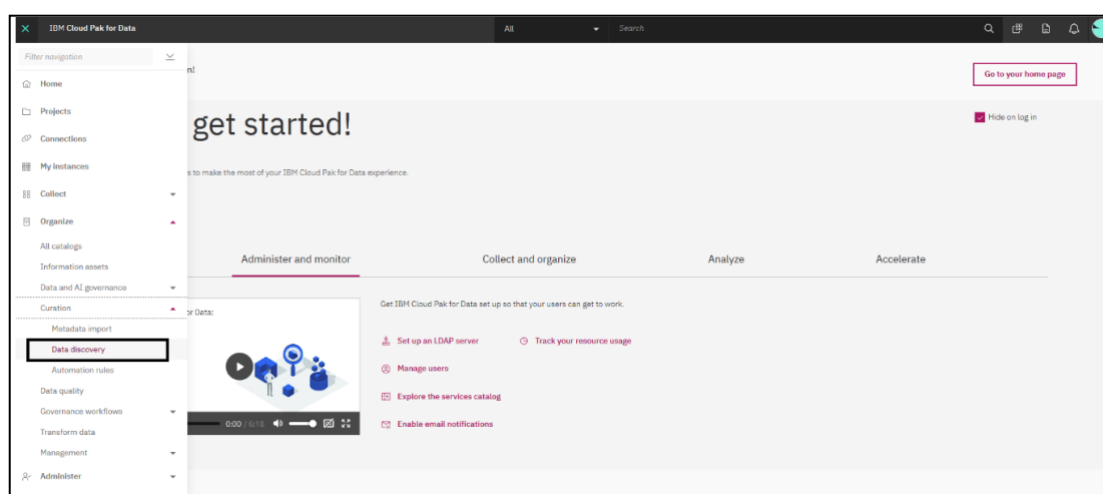
This step will allow the user to explore the process for initiating, reviewing and publishing the Data Discovery results

1. Setting up a Discovery Job

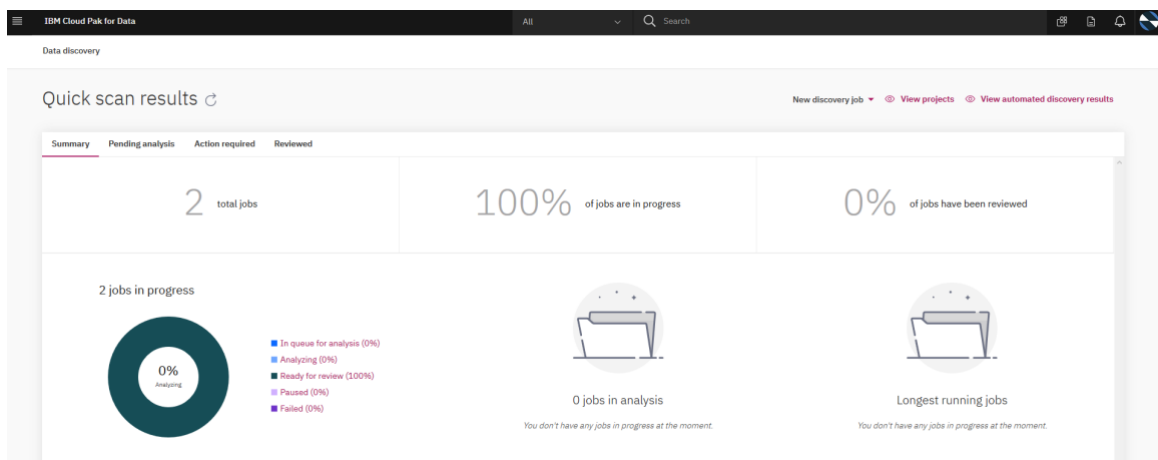
- 1 If not already logged-in, login Cloud Pak for Data using browser and your student ID and password



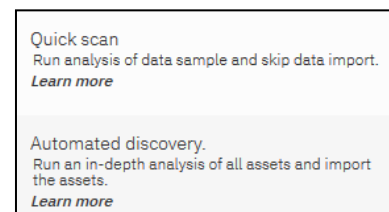
- 2 Open the navigation menu by selecting the  action (from the upper-left corner of the navigation bar)
- 3 Expand the section *Organize* and further expand the section *Curation*



- 4 Select the item *Discovery*. This will open the **Data Discovery** view. The view includes a dashboard of current and ongoing Discovery jobs.



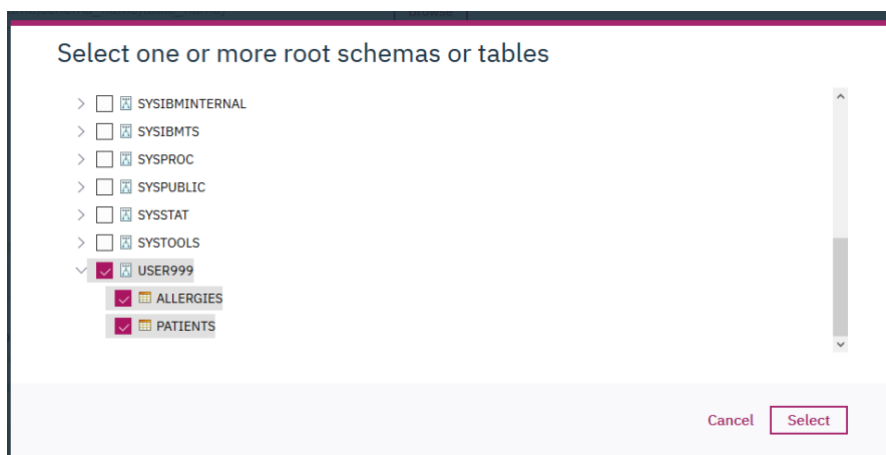
- 5 Expand the action item **New Discovery Job** to open the Discovery menu. Select *Automated Discovery* to initiate an in-depth analysis of a Data Set. The Automated Discovery Job dialog displays.



- 6 Click the action *Select a Connection* and then
 - a. From the list of available connections select **DB2THINK2020**



- 7 Click on **Browse** and expand BLUDB to show all schemas available to discover
 - a. Select **USER999** schema and **2 tables** as shown below and Click **Select**



- 8 This will bring back Discovery Job screen. Select the following Discovery Options:
- Analyze Columns. Examine the characteristics and identify the matching Data Classes.
 - Analyze Data Quality. Calculate the Quality Score based upon the Quality Dimensions.
 - Assign Terms. Suggest and assign Business Terms.
 - Use Data Sampling. Set the *Maximum Number of Records* to 500
 - Do not select *Publish Results to Catalog*. The user will review and further annotate the Data Asset prior to publishing results to the Catalog.
- 9 Select the Workspace *DataLakeWarehouse*

STOP. Due to constraints, you will not be completing the analysis process and viewing the calculated results. Rather, you will be directed to review previous analysis results and take the same steps to publish the results.

IBM Cloud Pak for Data

All

Search

Automated discovery job

Connection *

DB2THINK2020

Discovery root ⓘ

Examples: schema[db_name/schema_name]; table[db_name/schema_name/table_name]

Browse

Discovery options

☒ Analyze columns

☒ Analyze data quality

☒ Assign terms

☐ Publish results to catalog

☐ Use data sampling

Set the maximum number of records that you want to include in your data set sample:

Example: 2000

Select the method that you want to use to create your sample:

☒ Use the first x number of rows (where x = maximum number of records allowed)

☐ Use every Nth value (up to maximum number of records)

Nth interval

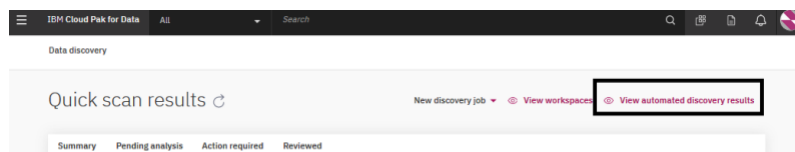
Example: 1000

☐ Use a random sampling

- 10 Click the action *Cancel* to cancel the process and return to the previous screen.

2. Review and Publish Discovery Job Results

- 1 From the menu, select the tab *View Automated Discovery Results*



- 2 From the list of previous results, identify the result for the Database Schema USER999 and click the *Discovery ID* to view its details

Discovery job **1588624840875** Finished

General information

Start	May 4, 2020, 4:40 PM	End	May 4, 2020, 4:44 PM
Started by	user2	Assets included in the discovery	1

Discover options

Workspace	DataLakeWorkspace	Connection	DB2THINK2020
Discovery options used	Column analysis, Term assignment, Data quality analysis	Discovery root	schema[BLUDB[USER999]]
Source asset import	All assets	Analysis	All assets


Sampling options

Sample size	500	Sample type	SEQUENTIAL
-------------	-----	-------------	------------

Discovered assets

Number of schemas: 1 Number of tables: 2

Asset name	Asset type	Tables	Status	Actions
USER999	Schema	2	<div> <div>Phase: Import Finished</div> <div>Phase: Analyze Finished</div> <div>Start: May 4, 2020, 4:40 PM</div> <div>End: May 4, 2020, 4:41 PM</div> <div>Done: 100%</div> <div>Successful: 100%</div> <div>Cancelled: 0%</div> <div>Failed: 0%</div> </div>	<div> <div>Review</div> <div>Details</div> </div>

- 3 Click the action *Review Discovery Results*  to view the results for the Schema **USER999** and its included Tables.

Analysis: All assets

Number of tables: 2

Asset name	Asset type	Tables	Status	Actions
USER999	Schema	2	<div> <div>Phase: Import Finished</div> <div>Phase: Analyze Finished</div> <div>Start: May 4, 2020, 2:12 PM</div> <div>End: May 4, 2020, 2:15 PM</div> <div>Done: 100%</div> <div>Successful: 100%</div> <div>Cancelled: 0%</div> <div>Failed: 0%</div> </div>	<div> <div>Review</div> <div>Details</div> </div>

[Review discovery results](#)

- 4 Expand the Table **Patient** and **Allergies** view the Quality Score, Data Class and Assigned Term of its columns.

- a. For the column **SSN**, **ADDRESS**, **ZIP** review the Term Suggestion.

Column	Quality Score	Data Class	Assigned Term	Term Suggestion	Time
BIRTHDATE	100%	Date of Birth	100%	Date of Birth 82% X	May 2, 2020, 2:15 PM
BIRTHPLACE	99%	Text	100%	City X	May 2, 2020, 2:15 PM
CITY	98%	Text	100%	City 100% X	May 2, 2020, 2:15 PM
DEATHDATE	100%	Date	100%	Date of Death 82% X	May 2, 2020, 2:15 PM
DRIVERS	100%	Missouri State Driver's...	100%	Drivers License Num... 58% X	May 2, 2020, 2:15 PM
ETHNICITY	100%	NoClassDetected	56%	-	May 2, 2020, 2:15 PM
FIRST	100%	NoClassDetected	100%	First Name 73% ✓	May 2, 2020, 2:15 PM
GENDER	100%	Gender	100%	Gender 100% X	May 2, 2020, 2:15 PM
ID	97%	Text	100%	Age 82% X	May 2, 2020, 2:15 PM
LAST	98%	NoClassDetected	100%	Last Name 73% ✓	May 2, 2020, 2:15 PM
MADEN	99%	Code	100%	Maiden Name 73% ✓	May 2, 2020, 2:15 PM
MARITAL	86%	NoClassDetected	46%	-	May 2, 2020, 2:15 PM
PASSPORT	99%	Spanish Fiscal Identific...	99%	Passport Number 73% ✓	May 2, 2020, 2:15 PM
PREFIX	100%	Honorific	100%	-	May 2, 2020, 2:15 PM
RACE	100%	Ethnicity	100%	-	May 2, 2020, 2:15 PM
SSN	100%	Identifier	100%	SSN 100% X	May 2, 2020, 2:15 PM
STATE	100%	US State Name	100%	State 100% X	May 2, 2020, 2:15 PM
SUFFIX	100%	Name Suffix	100%	-	May 2, 2020, 2:15 PM
ZIP	94%	Code	100%	Postal Code X	May 2, 2020, 2:15 PM

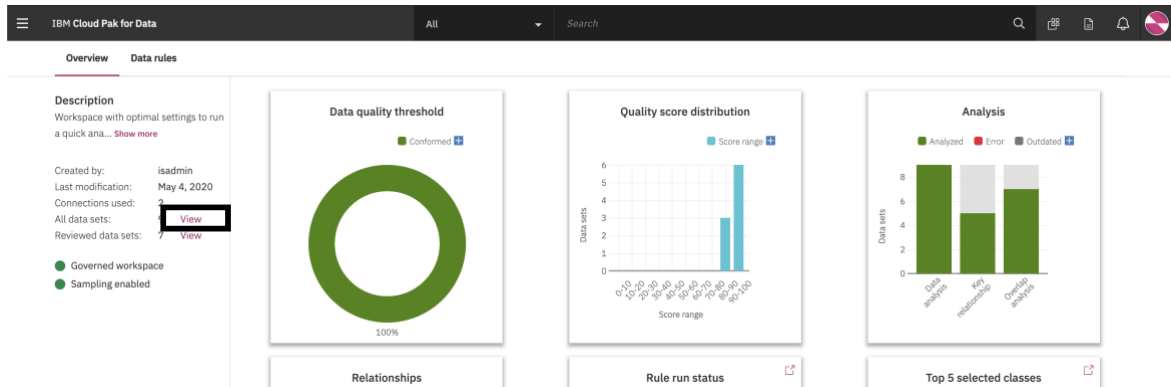
- 5 If not published, select all Tables and Columns by clicking the checkbox within the header section and click the action **Publish** to publish the Table and their Quality Score and assigned Data Class and Term to the Catalog. The Publish summary dialog appears.

Column	Quality Score	Data Class	Assigned Term	Term Suggestion	Time
BIRTHDATE	100%	Date of Birth	100%	Date of Birth 82% X	May 2, 2020, 2:15 PM
BIRTHPLACE	99%	Text	100%	City X	May 2, 2020, 2:15 PM
CITY	98%	Text	100%	City 100% X	May 2, 2020, 2:15 PM
DEATHDATE	100%	Date	100%	Date of Death 82% X	May 2, 2020, 2:15 PM
DRIVERS	100%	Missouri State Driver's...	100%	Drivers License Num... 58% X	May 2, 2020, 2:15 PM
ETHNICITY	100%	NoClassDetected	56%	-	May 2, 2020, 2:15 PM
FIRST	100%	NoClassDetected	100%	First Name 73% ✓	May 2, 2020, 2:15 PM
GENDER	100%	Gender	100%	Gender 100% X	May 2, 2020, 2:15 PM
ID	97%	Text	100%	Age 82% X	May 2, 2020, 2:15 PM
LAST	98%	NoClassDetected	100%	Last Name 73% ✓	May 2, 2020, 2:15 PM
MADEN	99%	Code	100%	Maiden Name 73% ✓	May 2, 2020, 2:15 PM
MARITAL	86%	NoClassDetected	46%	-	May 2, 2020, 2:15 PM
PASSPORT	99%	Spanish Fiscal Identific...	99%	Passport Number 73% ✓	May 2, 2020, 2:15 PM
PREFIX	100%	Honorific	100%	-	May 2, 2020, 2:15 PM
RACE	100%	Ethnicity	100%	-	May 2, 2020, 2:15 PM
SSN	100%	Identifier	100%	SSN 100% X	May 2, 2020, 2:15 PM
STATE	100%	US State Name	100%	State 100% X	May 2, 2020, 2:15 PM
SUFFIX	100%	Name Suffix	100%	-	May 2, 2020, 2:15 PM
ZIP	94%	Code	100%	Postal Code X	May 2, 2020, 2:15 PM

- 6 Click **Submit** to complete the publication process. Click on the icon on the top left to return to the job view. In the upper panel, click on the workspace name **DataLakeWorkspace** to go into the workspace.

General information
Start May 4, 2020, 4:40 PM
Started by user2
Discover options
Workspace DataLakeWorkspace
Discovery options used Column analysis, Term assignment, Data quality analysis
Source asset import All assets
Sampling options
Sample size 500

- Once in the workspace, you a dashboard that gives you an overview of all the data contained in it. From the left panel click *View* next to *All data sets*. From the list of data sets click in the *PATIENTS* table to examine it further.



- Here we can get a more detailed view of all the analysis done on the table. We can see all the Data Classes and Business Terms that have been automatically assigned to each column.

The interface displays a detailed view of the **PATIENTS** dataset analysis. It includes a sidebar with dataset statistics and a main table showing columns, their analysis status, last analyzed time, data class, term, format, nullability, uniqueness, minimum, maximum, and distinct values.

Column	Analysis status	Last analyzed	Data class	Term	Format	Nullability	Uniqueness	Minimum	Maximum	Distinct values
ID	Completed	10 minutes ago	Text	Age	NA			001d6674-ed60-463f-9fcb-42f359667053	Redd134-82e2-45a1-a22b-7761b391eeab	1167
BIRTHDATE	Completed	10 minutes ago	Date of Birth	Date of...	NATIVE/DEFAULT			1547-01-25	2019-01-24	989
DEATHDATE	Completed	10 minutes ago	Date	Date of...	NATIVE/DEFAULT			1547-04-28	2025-11-26	157
SSN	Completed	10 minutes ago	Identifier	Social...	999-99-9999			547-10-1126	547-99-9759	1167
DRIVERS	Completed	10 minutes ago	Missouri State Driver's License	Drivers...	A99999999			S99910183	S99999947	956
PASSPORT	Completed	10 minutes ago	Spanish Fiscal Identification Number	Passpo...	A99999999A			X10048914X	X9973766X	891
PREFIX	Completed	10 minutes ago	Honorific		As.			Mr.	Ms.	3
FIRST	Completed	10 minutes ago		First N...	NA			Abby752	Oscar156	957


- If we select the Rules tab and then the Quality Rules tab, we can see that the automation rule we created earlier to validate the format of a social security number has been automatically applied to the SSN column once it was assigned the business term Social Security Number. We can identify if any data points do not match the format.

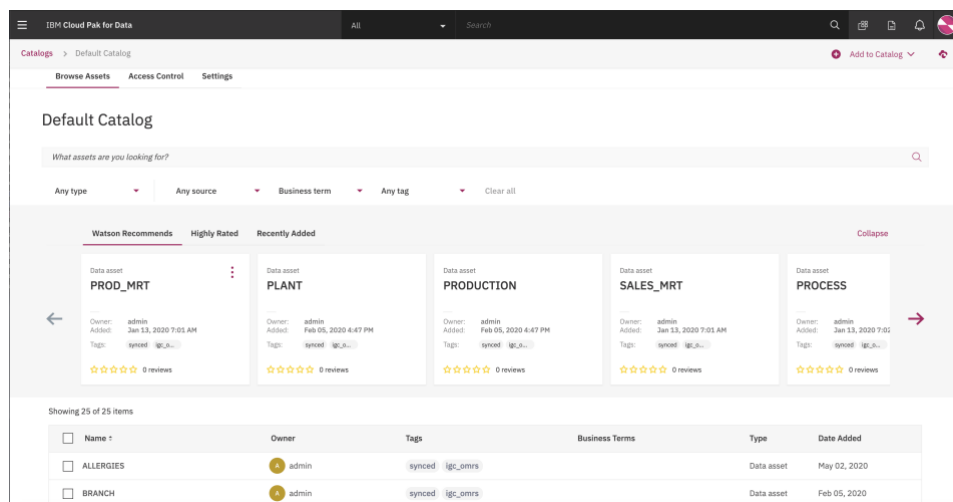
The interface displays the **Rules** tab, specifically the **Quality rules (1)** sub-tab. It shows a table with rule details, status, execution status, bindings, last run time, number met, percentage met, number not met, and percentage not met.

Rule details	Rule status	Execution status	Bindings	Last run time	Number met	Percentage met	Number not met	Percentage not met
Quality rule	--	Successful	SSN	5/4/2020, 10:14 PM	1167	100%	0	0%

Rule name: SsnMatchesHyphenFormat
Expression: SSN MATCHES_FORMAT '999-99-9999'

3. Search and Explore Discovered Assets

- 1 Optionally, open the navigation menu by selecting the  action (from the upper-left corner of the navigation bar) and expand the section **Organize** and select the item **All Catalogs**. The list of Catalogs will display. Catalogs include a sub-set of Data Sets which specific users have been granted access to preview and access.
- 2 Select the catalog **Default Catalog**. The Catalog view will display.



- 3 Browse the list of items and select the Table **Patient**, this is the same Table previously imported and discovered. The Data Set preview is displayed.

Catalogs > Default Catalog > PATIENTS

Overview

Access

Review

Profile

Lineage

DATA ASSET

PATIENTS

Remove

Download

Add to Project

Description

There is no description available for this asset.

Added: May 02, 2020 3:13 PM

Format: application/octet-stream

Business Terms

There are no terms available for this asset.

Tags

lgc_omrs synced

Reviews

☆☆☆☆ 0 reviews

Connection

Source: DB2THINK2020

Source type: db2

Classification

Personally Identifiable Information

Schema: 20 Columns

Preview: 1000 rows

Last refresh: 2 days ago

Refresh

ID	BIRTHDATE	DEATHDA...	SSN	DRIVERS	PASSPORT	PREFIX	FIRST	LAST	SUFFIX	MAIDE
Type: Char	Type: Date	Type: Date	Type: Char	Type: Char	Type: Char	Type: Char	Type: Char	Type: Char	Type: Char	Type: C
c8bbe7af-15fc-4	1991-06-07		547-73-8408	S99988154	X67639959X	Mrs.	Seema671	Stroman228		Deckor
0813f43e-7f65-	1922-11-06	1992-12-26	547-93-4724	S99912515	X56024572X	Ms.	Antonia30	Bañuelos542		
ea17bae9-8d26-	1961-01-06	1964-04-10	547-22-8004				Nick1254	VonRueden376		
15746215-3574	1972-09-13		547-53-7106	S99958010	X73159838X	Mrs.	Sharilyn202	Kuhic920		Waelcl
2d9b1e72-5179	1920-12-12	1982-06-30	547-21-9834	S99929022	X3330261X	Mr.	Emery884	Fritsch593		
761f9388-4f4b-4	2006-08-01		547-92-3627				Lilia791	Corona300		
7846c065-79cd-	1953-07-14		547-57-2251	S99976771	X84559834X	Mrs.	Rebeca548	Olivares593		Pichan
ed63a17e-0bea-	1956-06-15		547-17-6198	S99998503	X5343060X	Mrs.	Maricarmen445	Jiminez732		Romer
141ba9b4-9992	1935-10-06	1957-03-24	547-10-1904	S99981880	X61568211X	Mr.	Wilton999	Mosciski958		