

Lab Guide

Metadata Import and Discovery with Watson Knowledge

Marc Haber, Senior Offering Manager - Data Governance and Catalog

Kunjavihari (Kunju) Kashalikar, Program Director - Data Governance, Private Cloud



Guide

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.

Learn how to ingest data into the Catalog, further enriching and preparing data thru 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.

In this Lab, you will explore the following:

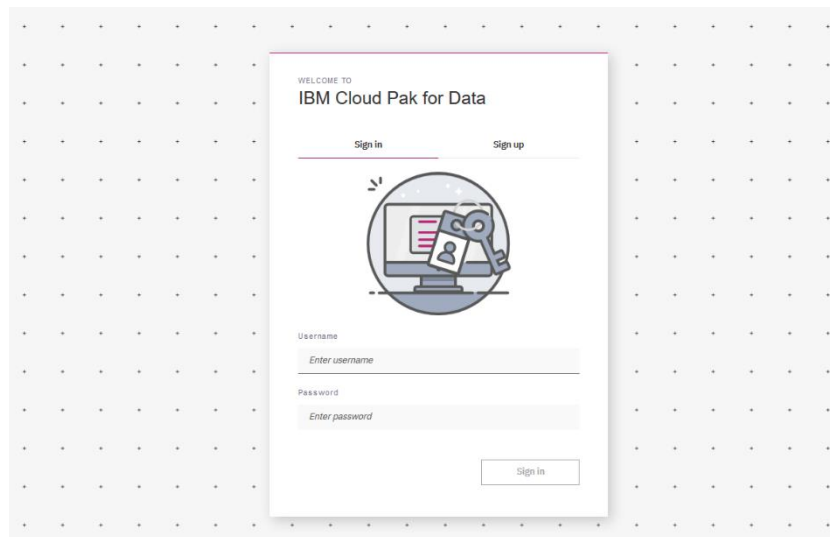
- Step 1: Configure Metadata Import and Discovery
 - Define Quality Workspace (Project) and Quality Dimension measurements
 - Define a Data Connection for the data Discovery and Classification
- Step 2: Data Discovery
 - Initiate Import and Data Discovery process
 - Review Data Discovery results
 - Publish Data Asset to the Catalog
- Step 3: Metadata Import
 - Initiate Import of a Data Model Asset
 - Review and Publish imported Data Model results to the Catalog
- Step 4: Data Quality
 - Review Data Quality dashboard and results
 - Initiate and update Data Quality assessment for Data Asset
 - Optional: Initiate and update Key Analysis for Data Assets
- Step 5: Search and explore business-ready-data within the Catalog


Step 1: Configure Metadata Import and Discovery

An initial step in the Metadata Import and Discovery process is the configuration of a Quality Workspace and Dimension, and definition of new Data Asset Connector.

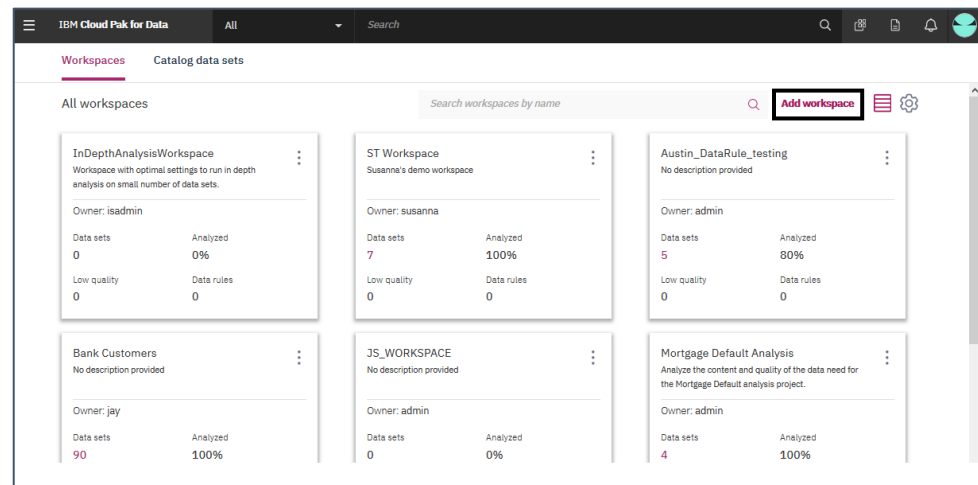
The following is for informational purposes only. The Lab has already configured and defined the required Data Quality Workspace and Data Asset Connector

1. Launch Watson Knowledge Catalog from the desktop-shortcut, or enter the provided URL into a supported Internet Browser.
2. Enter the credentials for Watson Knowledge Catalog, as provided.

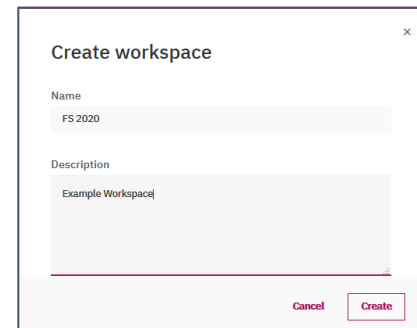


3. Open the navigation menu by selecting the  action (from the upper-left corner of the navigation bar)

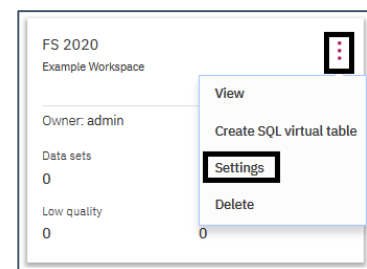
4. Expand the section *Organize* and Select the item *Data Quality*. This will open the *Data Quality* view. The view displays a list of existing Quality Workspaces.



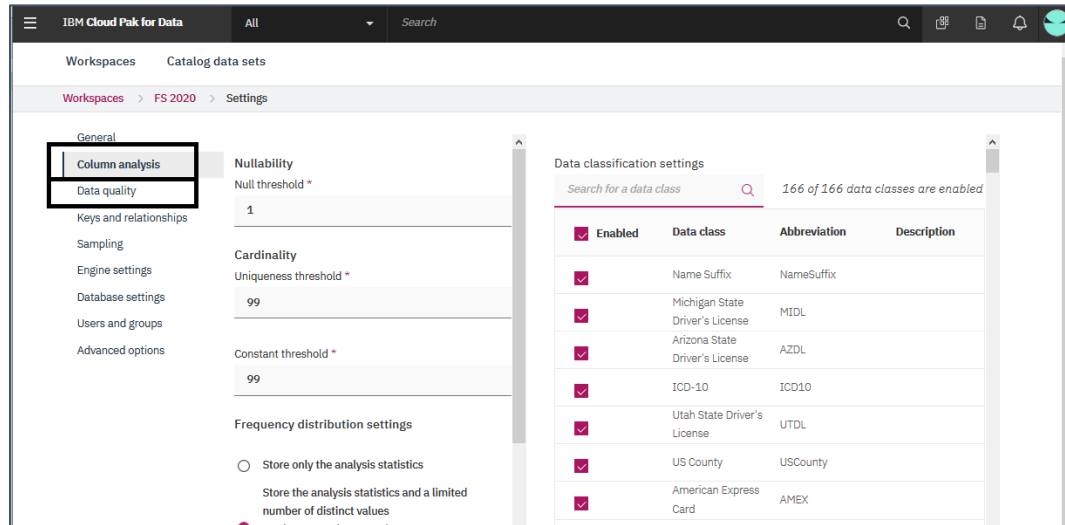
5. Click the action *Add Workspace* to create a new Quality Workspace. The create wizard appears.
 - a. Enter the name of the Workspace *FS 2020 <Your Name>*
 - b. Enter a description for the Workspace *Example Workspace*
 - c. Click Create to continue




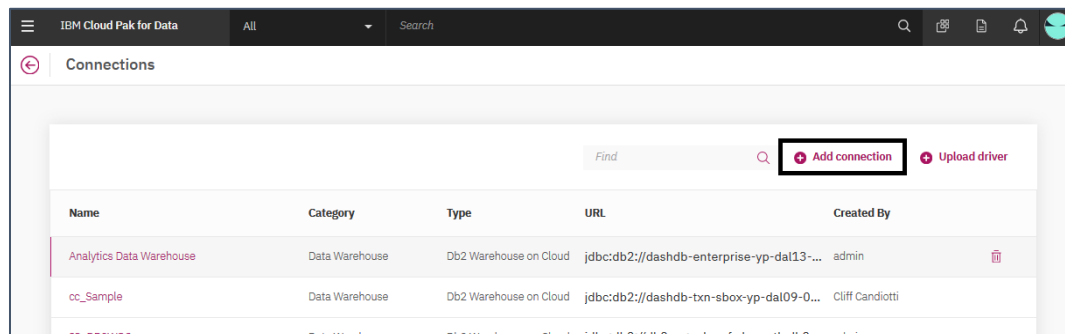
6. From the displayed list of existing Quality Workspaces, find the newly created Workspace *FS 2020 <Your Name>*. Expand the action menu and select *Settings* to review the Quality Dimension settings.



7. From the settings dialog, select the tab *Column Analysis*. Optionally:
 - a. Enable or disable the Data Classes to be used when analyzing and evaluating data
 - b. Adjust the threshold values for Cardinality and Nullability
8. Select the tab *Data Quality*. Optionally:
 - a. Enable or disable the Dimension to be used when calculating the Quality Score
 - b. Enable the use of Automation Rules to automatically execute Quality Rules upon Discovery



9. Click the action *Save* to save changes to the Quality Workspace.
10. Open the navigation menu by selecting the  action and select *Connections*. The Connection management dialog displays.



11. Click the action item *Add Connection* to create a new data connection.
 - a. Enter the Connection Name *FS 2020 <Your Name>*
 - b. Enter the Connection description *DB2 Data Sample*
 - c. Select the Connection Type *DB2*
 - d. Enter the Host for Database *dashdb-enterprise-yp-dal13-16.services.dal.ibmcloud.net*
 - e. Enter the Database Port *50000*
 - f. Enter the Database Name *BLUDB*

- g. Enter the Username to access the Database *watsondemo*
- h. Enter the Password for the user *WatsOnDataandAI!*
- i. Click *Test Connection* to validate the connection to the Database. Wait for the test to return successfully.
- j. Click *Add* to define the new Database Connection


The screenshot shows the 'Add connection' dialog in the IBM Cloud Pak for Data interface. The dialog is titled 'Add connection' and contains several input fields and options. The 'Connection name' field is labeled 'Data Warehouse'. The 'Description' field is labeled 'DB2 Data Sample'. The 'Connection type' is set to 'Db2'. The 'Host' field is labeled 'dashdb-enterprise-yp-dal13-16.services.dal.bluemix.net'. The 'Port' field is labeled '50000'. The 'Database' field is labeled 'BLUDB'. The 'Username' field is labeled 'watsondemo'. There are also fields for 'JDBC URL' and 'Options'. The 'JDBC URL' field is labeled 'jdbc:db2://dashdb-enterprise-yp-dal13-16.services.dal.bluemix.net:50000/BLUDB'. The 'Options' field is labeled 'Type additional options here'. There are checkboxes for 'Use SSL' and 'Verify server SSL certificate'. There is a 'Select file' button for uploading a certificate. At the bottom of the dialog are buttons for 'Cancel', 'Test connection', and 'Add'.

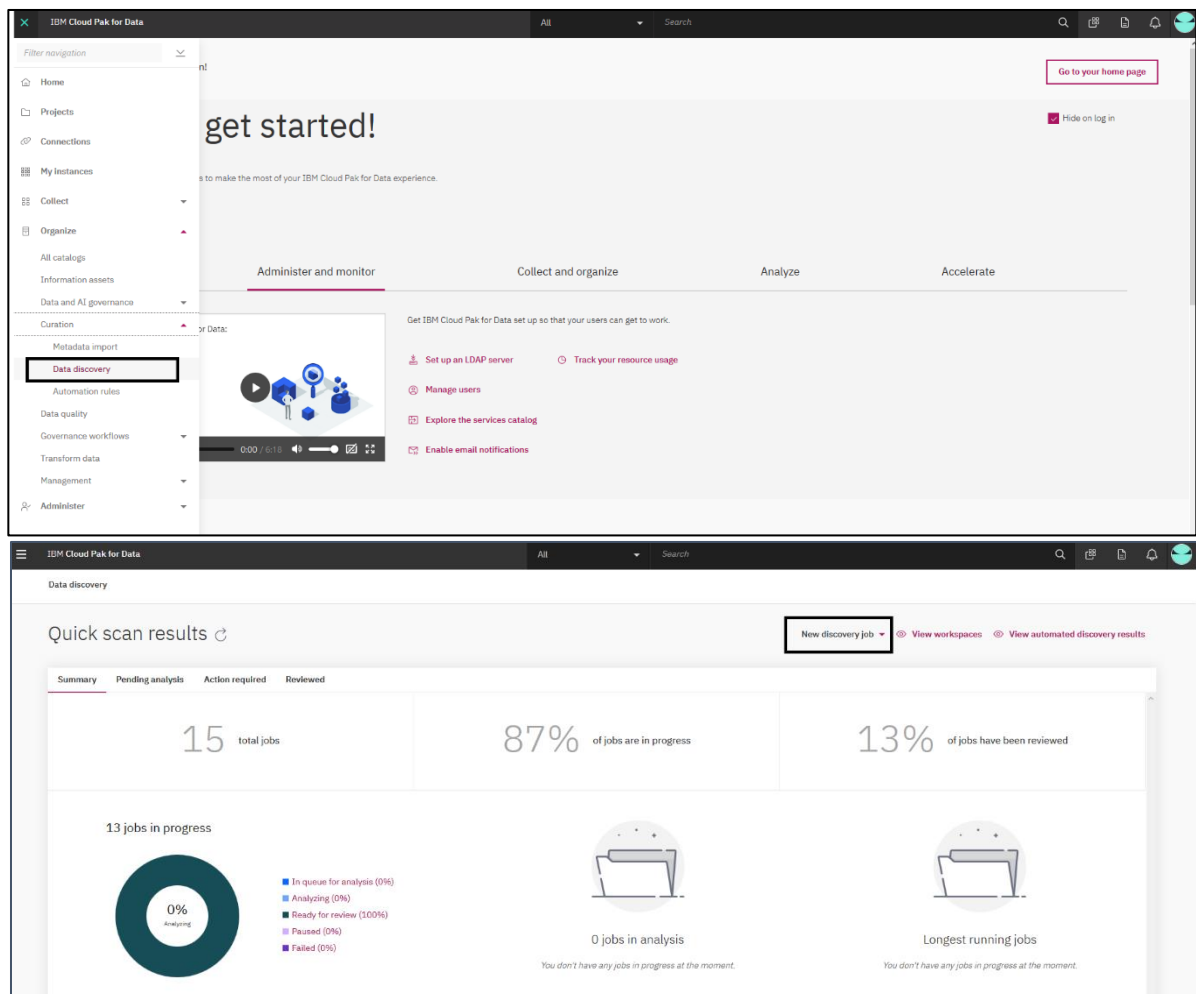
12. The list of Connections is displayed, and includes the newly created Database Connection

This completes Step 1 and the configuration for Metadata Import and Discovery

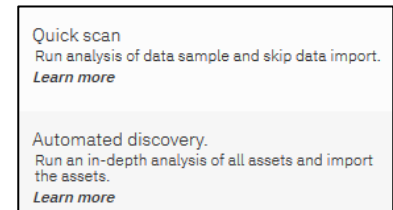
Step 2: Data Discovery

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

1. If not already open, launch Watson Knowledge Catalog from the desktop-shortcut, or enter the provided URL into a supported Internet Browser.
2. Enter the credentials for Watson Knowledge Catalog, as provided
3. Open the navigation menu by selecting the  action (from the upper-left corner of the navigation bar)
4. Expand the section *Organize* and further expand the section *Curation* selecting the item *Discovery*. This will open the *Data Discovery* view. The view includes a dashboard of current and ongoing Discovery jobs.



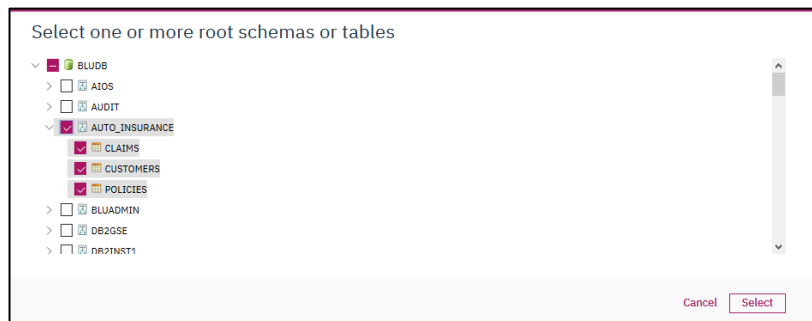
5. Expand the action item *New Discovery Job* to open the Discovery menu. Select *Automated Discovery* to initiate an in-depth analysis of a Data Set. The Automated Discovery Job dialog displays.



6. Click the action *Select a Connection* and then
 - a. From the list of available connections select *Data Warehouse*
 - b. Click *Next* to continue



7. Browse and select the Discovery Root. For purposes of this lab, we are only selecting a sub-set of the Schemas and Tables from the Database *BLUDB*
 - a. Expand the Database *BLUDB*
 - b. select the Database Schema *Auto_Insurance*



8. 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.

Automated discovery job

Connection *
Analytics Data Warehouse

Discovery root ⓘ
schema[BLUDB]AUTO_INSURANCE

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:
500

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
Seed
Example: 1234
Percentage
Example: 10

Workspace ⓘ
Bank Customers

Cancel Discover

IBM Cloud Pak for Data

AS

Search

Discovery job 1578339429757

Finished

General information

Start January 6, 2020, 9:37 PM

Started by admin

Discover options

Workspace Bank Customers

Discovery options used Column analysis, Term assignment, Data quality analysis

Source asset import All assets

Sampling options

Sample size 500

End January 6, 2020, 9:42 PM

Assets included in the discovery 1

Connection Analytics Data Warehouse

Discovery root schema[BLUDB]AUTO_INSURANCE

Analysis All assets

Sample type SEQUENTIAL

Discovered assets

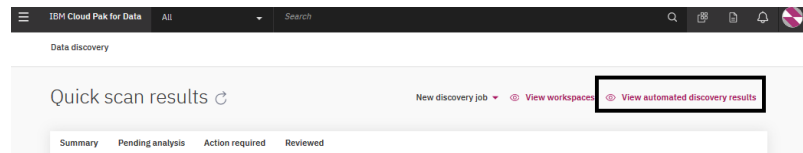
Number of schemas 1

Number of tables 3

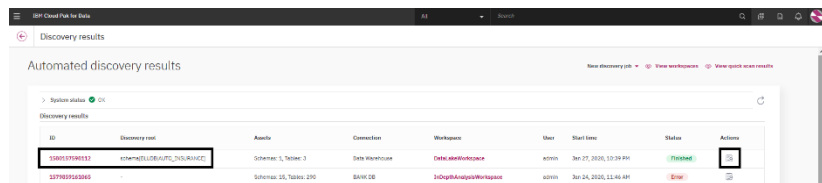
Asset name	Asset type	Tables	Status	Actions
AUTO_INSURANCE	Schema	3	<div><div>Phase Import</div><div>Phase Analyze</div><div>Start January 6, 2020, 9:37 PM</div><div>End January 6, 2020, 9:37 PM</div><div>Done 100%</div><div>Successful 100%</div><div>Cancelled 0%</div><div>Failed 0%</div></div>	<div><div></div><div></div></div>

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

11. From the menu, select the tab *View Automated Discovery Results*



12. From the list of previous results, identify the result for the Database Schema *Auto_Insurance* and click the *Discovery ID* to view its details

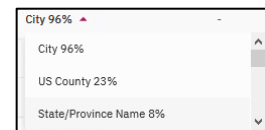


13. Click the action *Review Discovery Results* to view the results for the Schema *Auto_Insurance* and its included Tables.

Discovered assets		Number of tables: 3	
Asset name	Asset type	Tables	Status
AUTO_INSURANCE	Schema	3	<div> <div>Phase: Import</div> <div>Start: January 27, 2020, 10:40 PM</div> <div>End: January 27, 2020, 10:40 PM</div> </div> <div> <div>Phase: Analysis</div> <div>Start: January 27, 2020, 10:40 PM</div> <div>End: January 27, 2020, 10:44 PM</div> </div> <div> <div>Done: 100%</div> <div>Successful: 100%</div> <div>Cancelled: 0%</div> <div>Failed: 0%</div> </div>

14. Expand the Table *Customers* and view the Quality Score, Data Class and Assigned Term of its columns.

a. For the column *City*, click to expand the list of identified Data Classes. Review the list of suggested types and their relative score.



b. For the column *Email Address* review the Term Suggestion.

Discovered data sets		Audit		Refresh	
Name	Quality score	Data class	Assigned terms	Last analyzed	Actions
CLAIMS	99%	-	-	Jan 6, 2020, 9:40 PM	
CUSTOMERS	99%	-	Customer 100% X + 3 more	Jan 6, 2020, 9:42 PM	
CITY	96%	City 96%	Address 100% X + 5 more	Jan 6, 2020, 9:42 PM	
COUNTRY	100%	Country Code 100%	Country 100% X + 6 more	Jan 6, 2020, 9:42 PM	
CREDITCARD_OVV	100%	NoClassDetected 100%	Credit Card 57% ✓ + 2 more	Jan 6, 2020, 9:42 PM	
CREDITCARD_EXP	100%	Date 100%	Credit Card 57% ✓	Jan 6, 2020, 9:42 PM	
CREDITCARD_NUMBER	100%	Identifier 100%	Credit Card Number 100% X	Jan 6, 2020, 9:42 PM	
CREDITCARD_TYPE	99%	NoClassDetected 56%	Customer Ty... 100% X + 1 more	Jan 6, 2020, 9:42 PM	

15. 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.

<input checked="" type="checkbox"/>	Name	Quality score	Data class
-------------------------------------	------	---------------	------------

16. Click *Submit* to complete the publication process.

×


Publish selected

You are about to publish the selected elements to the catalog. Are you sure you want to publish them?

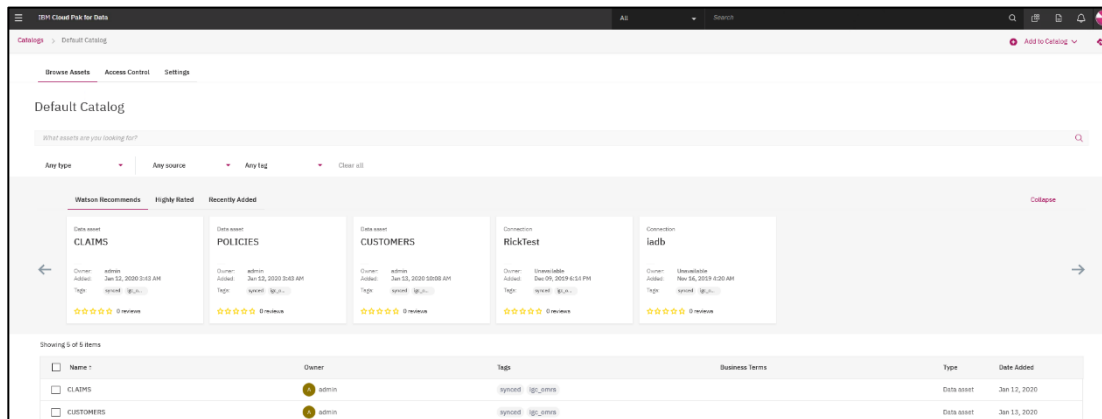
You will publish the following datasets:
CLAIMS
CUSTOMERS
POLICIES

Cancel

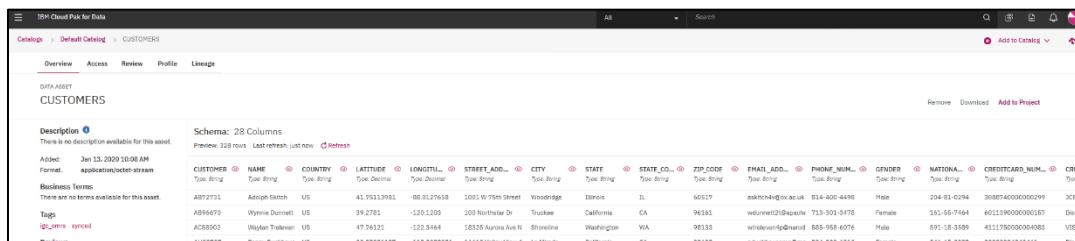
Submit

17. 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.

18. Select the catalog *Default Catalog*. The Catalog view will display.



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



CUSTOMER	NAME	COUNTRY	LATITUDE	LONGITUDE	STREET_ADDRESS	CITY	STATE	STATE_CODE	ZIP_CODE	EMAIL_ADDRESS	PHONE_NUMBER	GENDER	NATIONALITY	CREDITCARD_NUMBER	CREDITCARD_EXPIRATION_DATE
AB72731	Adriana Smith	US	41.75113981	-88.0127608	1001 W 75th Street	Woodbridge	Illinois	IL	60517	amsmith4@outlook	854-400-4440	Male	204-01-0294	3088740000000099	3CR
AB96470	Wynne Darnell	US	39.2781	-120.1200	100 Northstar Dr	Truckee	California	CA	96161	wdarnell2@outlook	713-301-5470	Female	161-55-7164	6011190000000087	DISC
AC68002	Walter Trueman	US	47.74121	-122.3444	18025 Aurora Ave N	Shoreline	Washington	WA	98133	wtrueman@shoreline	805-955-4070	Male	691-18-3389	4111750000000480	VISA
AD05577	Nancy Robinson	US	34.80061104	-110.0000000	100 S 10th Ave N	St. Cloud	Minnesota	MN	56340	nrobinson@stcloud	850-824-8514	Female	836-46-8439	8000000000000000	DISC

20. Preview the Columns and their Data Class and Term assignments.


21. Select the tab *Profile* to preview the Frequency Distribution for the Table and its columns.

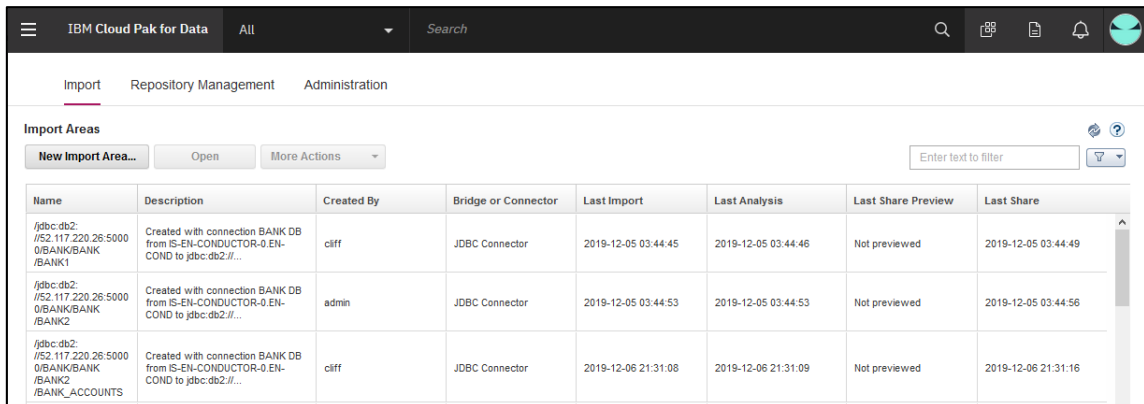
22. Optionally: Select the tab *Review* and set a Rating add a new Comment for the Table.

This completes Step 2 and the Data Discovery process

Step 3: Metadata Import

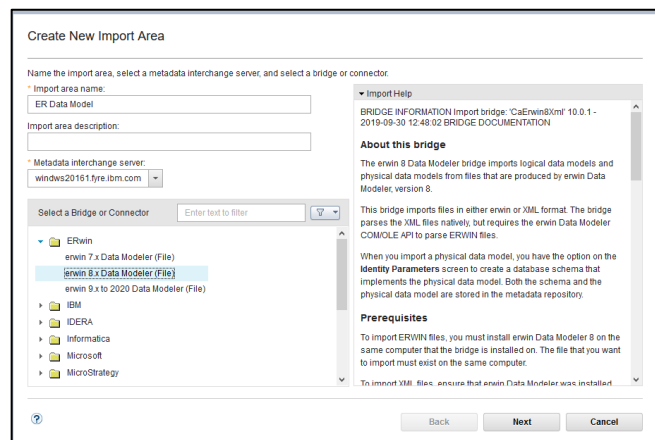
This step will allow the user to explore the process for importing and publishing Assets into the Catalog

1. Open the navigation menu by selecting the  action and expand the section *Organize* and further expand the section *Curation* selecting the item *Metadata Import*. This will open the *Metadata Asset Manager* management screen.



Name	Description	Created By	Bridge or Connector	Last Import	Last Analysis	Last Share Preview	Last Share
/jdbc:db2: /52.117.220.26:5000 0/BANK/BANK /BANK1	Created with connection BANK DB from IS-EN-CONDUCTOR-0.EN- COND to jdbc:db2://...	cliff	JDBC Connector	2019-12-05 03:44:45	2019-12-05 03:44:46	Not previewed	2019-12-05 03:44:49
/jdbc:db2: /52.117.220.26:5000 0/BANK/BANK /BANK2	Created with connection BANK DB from IS-EN-CONDUCTOR-0.EN- COND to jdbc:db2://...	admin	JDBC Connector	2019-12-05 03:44:53	2019-12-05 03:44:53	Not previewed	2019-12-05 03:44:56
/jdbc:db2: /52.117.220.26:5000 0/BANK/BANK /BANK2 /BANK_ACCOUNTS	Created with connection BANK DB from IS-EN-CONDUCTOR-0.EN- COND to jdbc:db2://...	cliff	JDBC Connector	2019-12-06 21:31:08	2019-12-06 21:31:09	Not previewed	2019-12-06 21:31:16

2. If not already selected, select the Import tab.
3. Click the action *New Import Area* to define the import parameters and initiate an import of a Data Model Asset. The Import wizard appears.
4. Step 1: Create New Import Area
 - a. Enter the name of the Import Area *ER Model <Your Name>*. The name uniquely identifies the import and its defined parameters.
 - b. Optionally enter a description for the Import Area *Import ERWin Data Model*
 - c. Select the Metadata Interchange Server *windowsvm*. The Bridges are hosted on an external MS Windows Server.
 - d. Expand the folder *Erwin* and select the import bridge *Erwin 8.x*
 - e. Click *Next* to continue



Create New Import Area

Name the import area, select a metadata interchange server, and select a bridge or connector.

* Import area name:
ER Data Model

Import area description:

* Metadata interchange server:
windows20161.fyre.ibm.com

Select a Bridge or Connector: Enter text to filter

- ERwin
 - erwin 7.x Data Modeler (File)
 - erwin 8.x Data Modeler (File)**
 - erwin 9.x to 2020 Data Modeler (File)
- IBM
 - IDERA
- Informatica
- Microsoft
- MicroStrategy

Import Help

BRIDGE INFORMATION import bridge: 'CaErwin8Xml' 10.0.1 - 2019-09-30 12:48:02 BRIDGE DOCUMENTATION

About this bridge

The erwin 8 Data Modeler bridge imports logical data models and physical data models from files that are produced by erwin Data Modeler, version 8.

This bridge imports files in either erwin or XML format. The bridge parses the XML files natively, but requires the erwin Data Modeler COMOLE API to parse ERWIN files.

When you import a physical data model, you have the option on the **Identity Parameters** screen to create a database schema that implements the physical data model. Both the schema and the physical data model are stored in the metadata repository.


Prerequisites

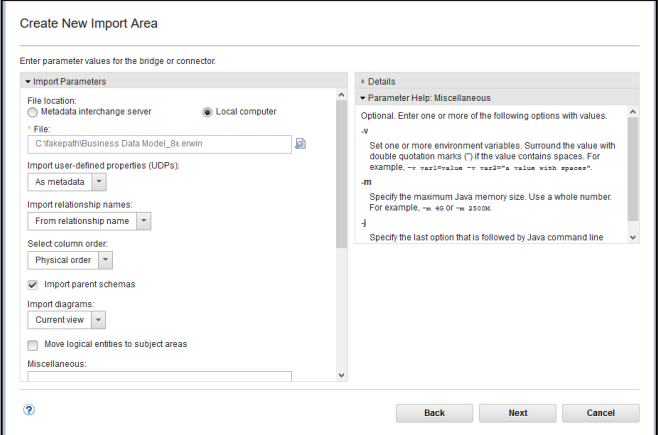
To import ERWIN files, you must install erwin Data Modeler 8 on the same computer that the bridge is installed on. The file that you want to import must exist on the same computer.

To import XML files, ensure that erwin Data Modeler was installed.

Back Next Cancel

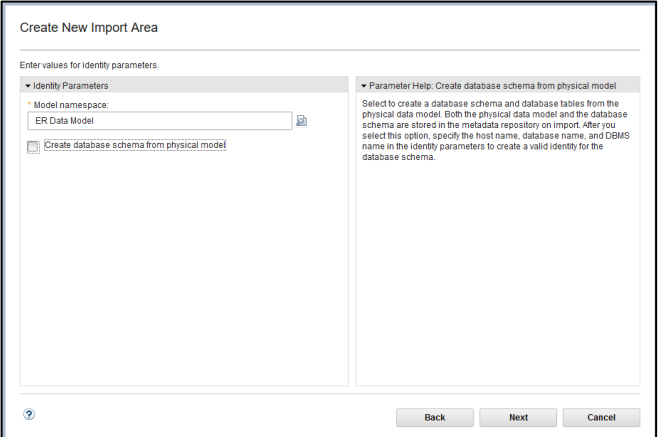
5. Step 2: Create New Import

- Browse to the provided Box Folder and download the file *Business Data Model.xml* to your machine
- Click the  action to browse the *File* location and select the downloaded file *Business Data Model.xml*. Alternately, if running the lab from the Skytap image directly, select the file from the directory *C:/FS*
- Click Next to continue



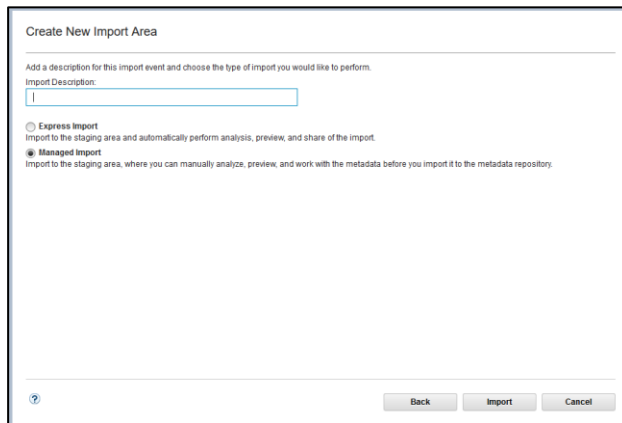
6. Step 3: Create Import Area

- Enter the Model Namespace *ER Model <Your Name>*
- Click Next to continue



7. Step 4: Create Import Area

- Optional enter an import description. The description is specific to this import, whereas the Import Area may be re-used for subsequent import.
- Select *Managed Import* to review the imported results, prior to sharing to the Catalog
- Click Import to complete the import process and view the results. The import may take a few minutes to complete.
- Click OK to close the Import Summary dialog



Create New Import Area

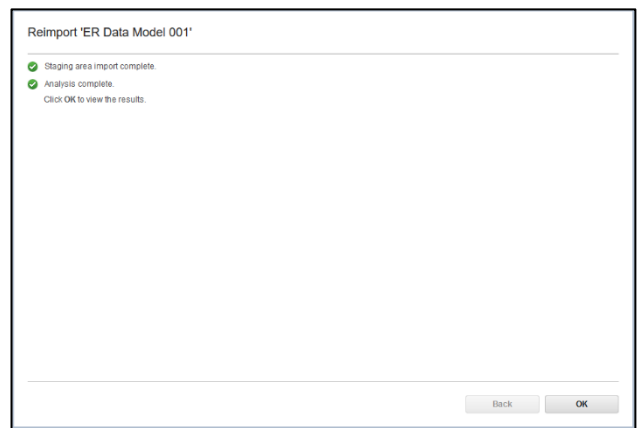
Add a description for this import event and choose the type of import you would like to perform.

Import Description:

☐ Express Import
Import to the staging area and automatically perform analysis, preview, and share of the import.

☒ Managed Import
Import to the staging area, where you can manually analyze, preview, and work with the metadata before you import it to the metadata repository.

[?](#) [Back](#) [Import](#) [Cancel](#)



Reimport 'ER Data Model 001'

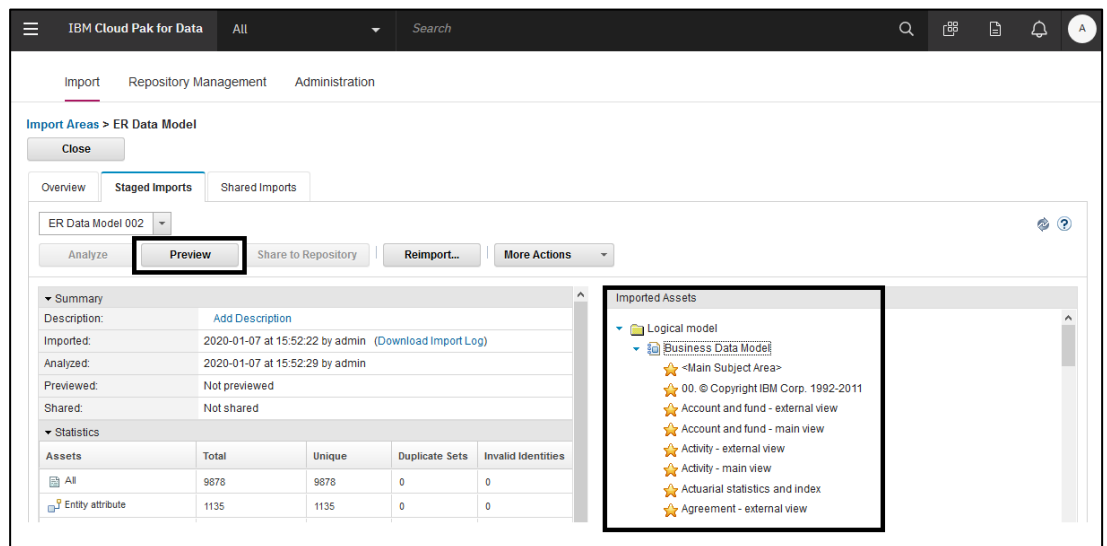
☒ Staging area import complete.

☒ Analysis complete.

Click OK to view the results.

[Back](#) [OK](#)

- View the Staged Import results, representing the Model and its Entities and Attributes captured from the import file. Click *Preview* to preview and analyze the content, prior to publishing and sharing with Catalog.



IBM Cloud Pak for Data

Import Repository Management Administration

Import Areas > ER Data Model

[Close](#)

Overview **Staged Imports** Shared Imports

ER Data Model 002

[Analyze](#) **[Preview](#)** [Share to Repository](#) [Reimport...](#) [More Actions](#)

Summary

Description: [Add Description](#)

Imported: 2020-01-07 at 15:52:22 by admin (Download Import Log)

Analyzed: 2020-01-07 at 15:52:29 by admin

Previewed: Not previewed

Shared: Not shared

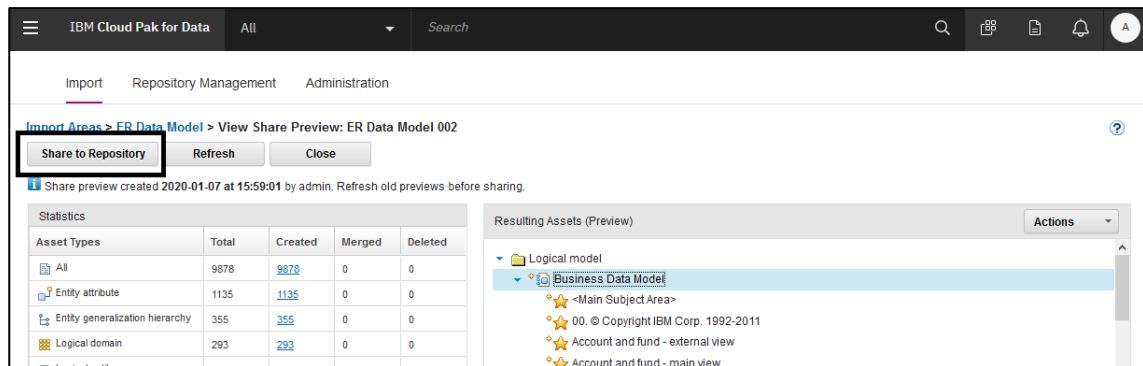
Statistics


Assets	Total	Unique	Duplicate Sets	Invalid Identities
All	9878	9878	0	0
Entity attribute	1135	1135	0	0

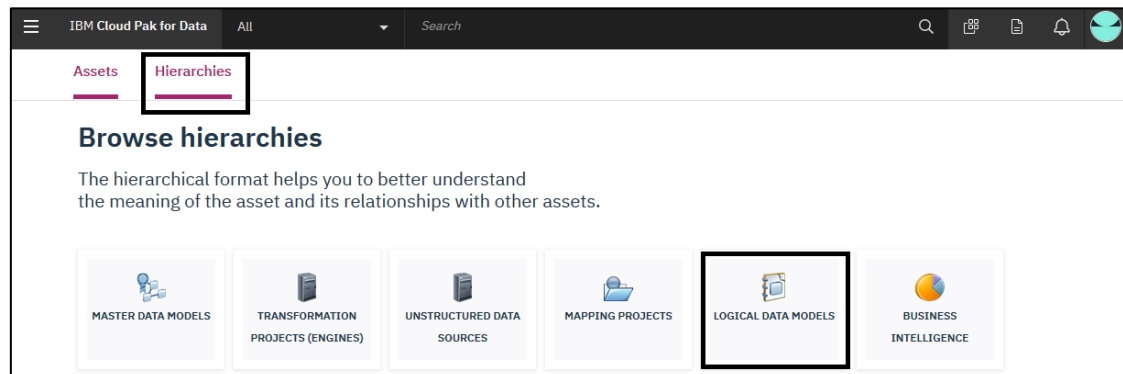
Imported Assets


- Logical model
 - Business Data Model
 - <Main Subject Area>
 - 00. © Copyright IBM Corp. 1992-2011
 - Account and fund - external view
 - Account and fund - main view
 - Activity - external view
 - Activity - main view
 - Actuarial statistics and index
 - Agreement - external view

9. View the Analysis results, the number of items to be created or merged into the Catalog. Click *Share to Repository* to continue and complete the import and share task. A confirmation dialog will ask you to confirm the action.



10. Click *Close* to close the Import Area display and return to the list of Import Areas
11. Open the navigation menu by selecting the  action and expand the section *Organize* and further expand the section *Information Assets*. This will open the *Information Asset* search screen.
12. Select *Hierarchies* and further select *Logical Data Models*




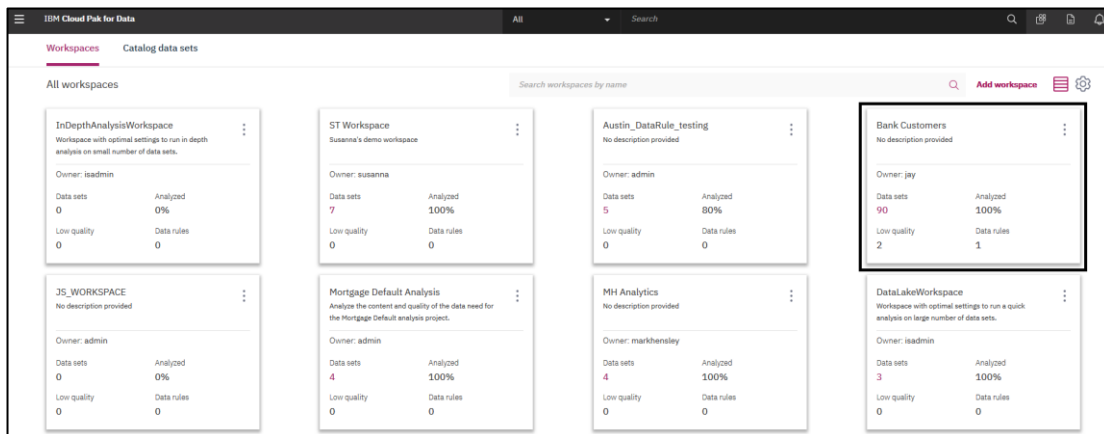
13. Expand the model *Business Data Model* and further expand the Subject Area *Main Subject Area* and select the Entity *Account*. The Entity details are displayed.
23. Optionally, expand the action  menu and select *Edit* to modify the following properties of the Entity:
 - a. Add a relationship to the Term *Customer*
 - b. Add a relationship to the Data Steward *CTP*
14. Optionally, expand the action menu and select *Explore Relationships* to view the usage or dependency relationships for the Entity

This completes Step 3 and the Metadata Import process

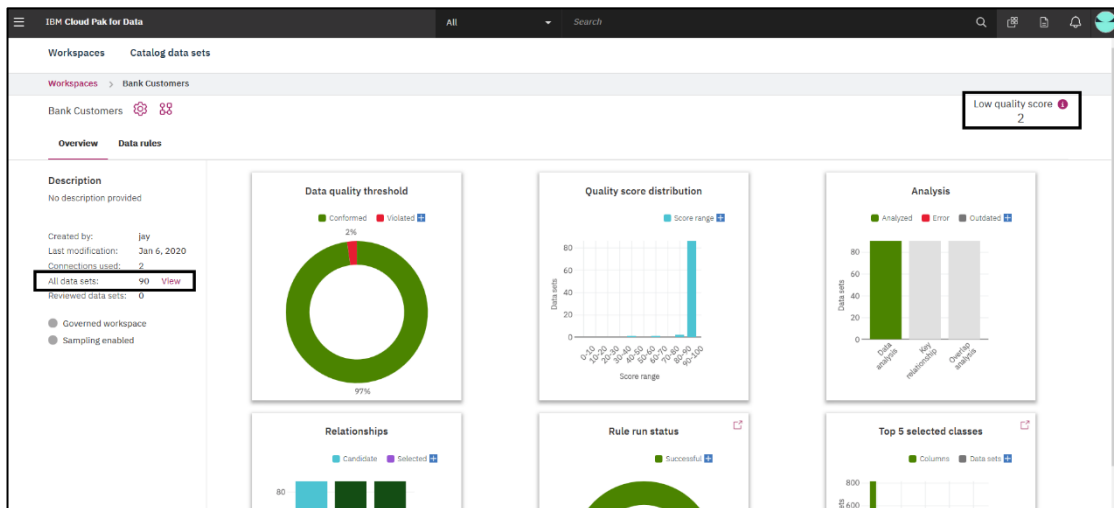
Step 4: Data Quality

This step will allow the user to explore the process for reviewing the Data Quality results

1. If not already open, logon to Watson Knowledge Catalog
2. Open the navigation menu by selecting the  action and expand the section *Organize* selecting the item *Data Quality*. This will open the *Data Quality* workspace screen.



3. Click the workspace *DataLakeWarehouse* to view its results and status. Explore the displayed dashboards, alerts for Data Set of low quality score and number of Data Sets contained by the



Workspace.

- From the left pane, note the count of Data Sets used within the Workspace. Click **View** to explore further the Data Sets.
- Filter the ensuing list and find and select the Data Set *Customers*.

Dataset name	Quality score	Location	Last data analysis	Columns	Rows	Key relationship found	Last key relationship analysis	Reviewed	Actions
CUSTOMERS	96.96%	jdbcd2://dashdb...	Jan 6, 2020, 21:42	28	328	0	--	<input checked="" type="checkbox"/>	[...]
POLICIES	98.71%	jdbcd2://dashdb...	Jan 6, 2020, 21:41	24	500	0	--	<input checked="" type="checkbox"/>	[...]
CLAIMS	99.50%	jdbcd2://dashdb...	Jan 6, 2020, 21:40	9	500	0	--	<input checked="" type="checkbox"/>	[...]
BANK_ACCOUNTS	97.06%	jdbcd2://s2.117...	Dec 4, 2019, 23:15	10	3532	0	--	<input checked="" type="checkbox"/>	[...]
BANK_CUSTOMERS	97.85%	jdbcd2://s2.117...	Dec 4, 2019, 23:28	16	3008	0	--	<input checked="" type="checkbox"/>	[...]

- Toggle the display to view the results within a grid by clicking the action icon

CUSTOMERS	
jdbcd2://dashdb-enterprise-yp-dal13-16.services.dal.ibm.net:50000/BLUDB...	
66	0
Last imported	Last data analysis
Jan 6, 2020, 21:37	Jan 6, 2020, 21:42
Last published	Threshold
Never	80%

- Select the Data Set to view its analysis results. The Table analysis view displays.

Column	Analysis status	Last analyzed	Data class	Term	Format	Nullability	Uniqueness	Minimum	Maximum	Distinct values
NAME	Completed	19 hours ago	Person Name	First Name	NA			Abigail Ingall	Yves Samways	328
COUNTRY	Completed	19 hours ago	Country Code	State	AA			US	US	1
LATITUDE	Completed	19 hours ago	Identifier		99.99999999			21.28201	61.19533942	328
LONGITUDE	Completed	19 hours ago	Identifier		-99.99999999			-157.890904	-70.887404	328
STREET_ADDRESS	Completed	19 hours ago	US Street Name	Street	NA			1 Adobe Court	22855 NE Park Lane	328
CITY	Completed	19 hours ago	City	City Name	NA			Shoreline	Yukon	260

8. Select the tab *Data Quality* to view the data quality violations for the Data Set.
 - a. Click the violation *Data Class Violations* to view which Data Classes have reported violations
 - b. Click the Data Class *State Code* to view the specific records that have violated the Data Class

Data class violations in column STATE_CODE

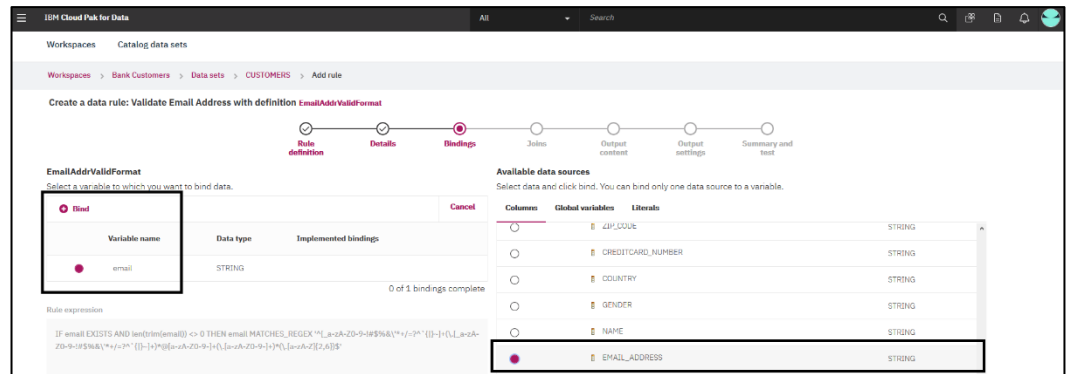
THE VA...	NUMBER OF COMPLAINTS	STATE_CODE	NAME	NUMBER_OF_COMMUNIC...	CITY	CREDITCARD_TYPE	PHONE_NUMBER	STREET_ADDRESS	LATITUDE	NUMBER_OF_CLOSE...
8		0000	Pat Werlock	6	Barrington	Diners Club	383-898-7295	180 County Road	41.7372687	1

- c. Click *OK* to close the preview dialog
9. Select the tab *Rules* and not there are no Data or Quality Rules defined. Rules evaluate the validity of data, ensuring it conforms to a standard, range or expectation.
10. Select the action item *Add Rule* and subsequently select the item *Data Rule* to view the available Data Rules that may be added to the workspace and evaluated in subsequent Discovery activities.
 - a. Expand the Folder *All* and select the Rule *EmailAddrValidFormat*. Click *Next* to continue

The screenshot shows the 'Add rule' dialog in the IBM Cloud Pak for Data interface. The 'Rule definition' tab is selected, and the 'EmailAddrValidFormat' rule is chosen from the 'All' folder. The 'Next' button is highlighted.

- b. Enter the name for the Data Rule *Validate Email Address*. Optionally enter a Description for the Rule. Click *Next* to continue
 - c. Define the binding for the Data Rule.
 - i. Select the Variable *Email* for the Email Address Format (left navigation pane)
 - ii. Scroll and select the column *Email_Address* from the *Auto Insurance // Customers* Table (right navigation pane)
 - iii. Click *Bind* (left navigation pane) to bind the variable to the column

iv. Click Next to continue.



- d. Click Next, there is no Database Join information to define
- e. Click Next, there is no Output Content to define
- f. Click Next, there is no Output Settings to define (the default output table will be used)
- g. Click Save to complete the process. the Data Rules included for the Data Set are displayed

11. Click **⋮**, the action menu, for the new Data Rule, and click the action *Run* to immediately invoke the Rule. Confirm the Rule Execution. It may take several minutes for the Rule execution to complete.

Rule details	Rule status	Execution status	Bindings	Last run time	Number met	Percentage met	Number not met	Percentage not met
> Validate Email Address	Candidate	Not started	EMAIL_ADDRESS	--	--	--	--	--

Run

View in catalog

View relationships

Edit

12. Refresh the display of Watson Knowledge Catalog until the Rule execution completes.

Rule details	Rule status	Execution status	Bindings	Last run time	Number met	Percentage met	Number not met	Percentage not met
> Validate Email Address	Candidate	Successful	EMAIL_ADDRESS	1/7/2020, 8:14 PM	328	100%	0	0%

13. Click on the tab *Columns* to view the columns of the Data Set and their analysis summary.

14. Browse and select the column *State* to further analyze it. View the Quality Score, Assigned Term and Suggested Term of the column.

IBM Cloud Pak for Data

Workspaces Catalog data sets

Workspaces > Bank Customers > Data sets > CUSTOMERS > Columns > STATE

STATE

Quality score 99%

Selected data class US State Name

Selected data type VarChar(14)

Analysis status Completed

Last column analysis 1/6/2020, 9:40 PM

Last DQ analysis 1/6/2020, 9:42 PM

Find a column

Name	Quality score
CUSTOMER	100
NAME	97%
COUNTRY	100
STATE	99%

Assigned terms

State State State Code State Name

Suggested terms

Province 79% Country 78% Country 78% Country 78% Country 78% Country 78%

Country Name 78% US State 70% Postal Address State 67% City 50%

Notes

No notes


Add note

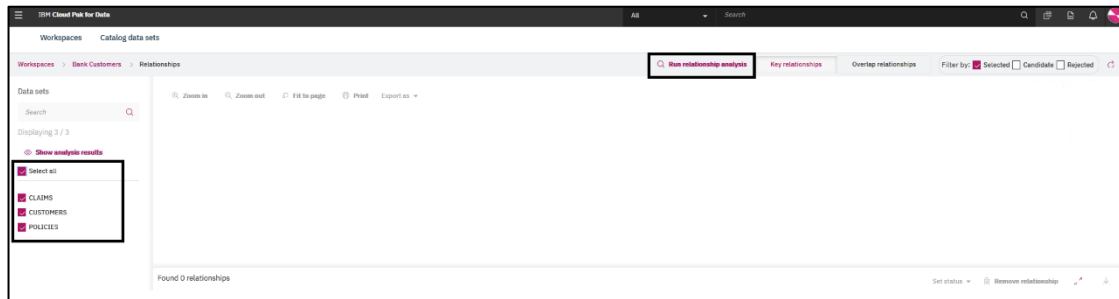
Long description


No description

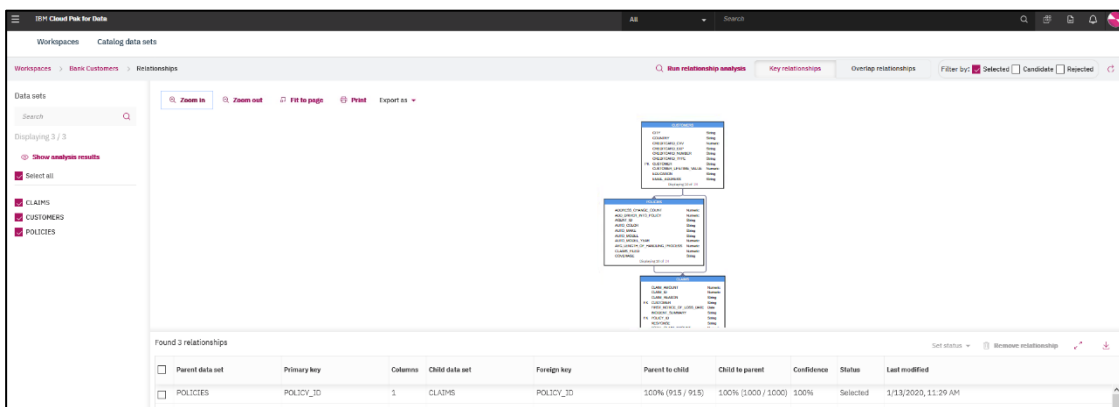
15. Click on the tab Data Quality for the Data Set *Customers* to view the resulting quality dimension details. Those details determined the Quality Score. Expand the violation *Inconsistent capitalization* and explore each violation.
16. Click on the tab Data Classes to view the suggested and selected Data Class. The Data Class of the column can help determine the sensitivity of the information and can be used to set and enforce Data Protection Rules that mask or otherwise hide the data from preview.
17. Click on the tab Frequency Distribution to view the distinct State Values. View the results graphically by clicking the action icon.
18. Return to the *DataLakeWarehouse* Workspace view by clicking *DataLakeWarehouse* from the breadcrumb menu

Workspaces > Bank Customers > Data sets > CUSTOMERS > Columns > STATE

19. From the Workspace view, click the action  (from the left pane) to view or initiate an analysis of Key Relationships. The Analysis screen displays.




- Select all Tables to be included in the Key Relationship analysis
 - Select the action *Run Relationship Analysis* to invoke the action
 - From the ensuing dialog select both *Key Relationship* and *Overlap*, as the type of relationships to be analyzed for, and click *Analyze* to continue
20. Click the action  to refresh the display of the Key Relationships to view the results of the analysis
21. Review the suggested Relationships and explore each further.

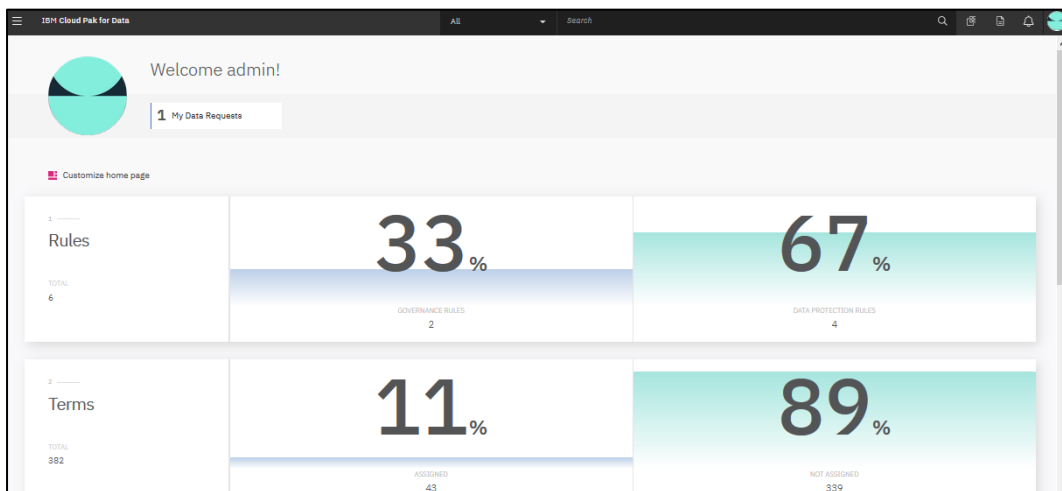


This completes Step 4 and the Data Quality review

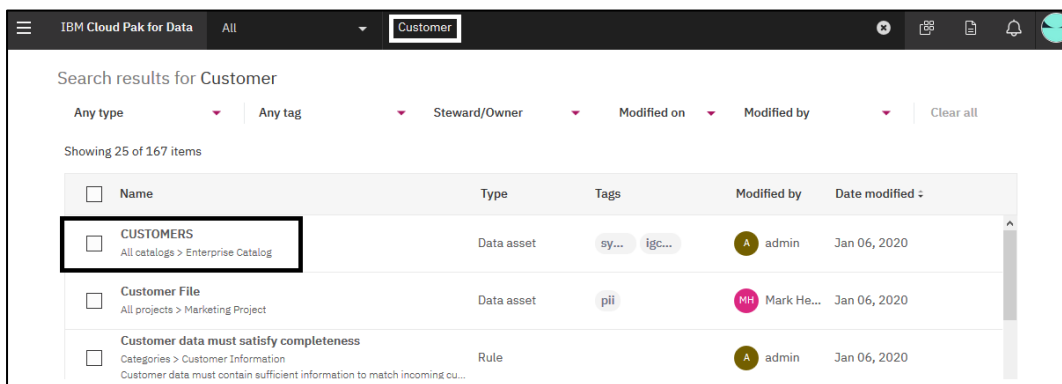
Step 5: Search and Explore Business Ready Information

This step will allow the user to search and explore the Catalog and find relevant information, leveraging the automated Discovery services that have applied meaning and given quality to the information

1. If not already open, logon to Watson Knowledge Catalog.
2. If the Homepage is not displayed, navigate to the Homepage by opening the navigation menu by selecting the  action and select *Homepage*.



3. From the Search widget, enter the search string *Customers* to find the Table of the same name



The screenshot shows the search results for 'Customer' in the IBM Cloud Pak for Data. The search bar at the top contains the text 'Customer'. Below the search bar, there are filters for 'Any type', 'Any tag', 'Steward/Owner', 'Modified on', and 'Modified by'. The results are displayed in a table with columns: Name, Type, Tags, Modified by, and Date modified. The first result, 'CUSTOMERS', is highlighted with a red box.

Name	Type	Tags	Modified by	Date modified
<input type="checkbox"/> CUSTOMERS All catalogs > Enterprise Catalog	Data asset	sy... igc...	A admin	Jan 06, 2020
<input type="checkbox"/> Customer File All projects > Marketing Project	Data asset	pii	MH Mark He...	Jan 06, 2020
<input type="checkbox"/> Customer data must satisfy completeness Categories > Customer Information Customer data must contain sufficient information to match incoming cu...	Rule		A admin	Jan 06, 2020

4. Select the table *Customers* to view its details from the Catalog

IBM Cloud Pak for Data | All | Customer

Search Results > Catalogs > Enterprise Catalog > CUSTOMERS

Overview Access **Review** Profile Lineage

DATA ASSET
CUSTOMERS

Remove Download Add to Project

Description
There is no description available for this asset.

Added: Jan 06, 2020 9:40 PM..PM
Format: application/octet-stream
Size: 312 KB

Business Terms
There are no terms available for this asset.

Tags
igc_omrs synced

Schema: 28 Columns 328 Rows
Preview: 328 rows Last refresh: 1 minute ago Refresh

CUSTOMER	NAME	COUNTRY	LATITUDE	LONGITUDE	STREET_ADD...	CITY
Type: String	Type: String	Type: String	Type: Decimal	Type: Decimal	Type: String	Type: String
Identifier	Person Na...	Country C...	Latitude	Longitude	US Street Name	City
AB72731	Adolph Skitch	US	41.75113981	-88.0127658	1001 W 75th Street	Woodridge
AB96670	Wynnie Dunnnett	US	39.2781	-120.1203	100 Northstar Dr	Truckee
AC58002	Waylan Trelevan	US	47.76121	-122.3464	18325 Aurora Ave N	Shoreline
AH99727	Penny Duckhous	US	33.88081187	-118.0288381	16610 Valley View Av	La Mirada

5. Review the table and column information, including the detected and assigned Term Assignments and Data Class Assignment.
6. Select the tab *Profile* to view the frequency distribution results
7. Select the tab *Review* and add a new Comment and Rating for the asset.

Overview Access **Review** Profile Lineage

DATA ASSET
CUSTOMERS

Remove Download Add to Project

Overall Rating
0.0
☆☆☆☆☆ 0 reviews

Review Summary

5	(0)
4	(0)
3	(0)
2	(0)
1	(0)

My Review

admin | Jan 07, 2020
☆☆☆☆☆

Write a review of this asset to help others.

Cancel Submit

This completes Step 5 and the review of Business Ready information

Thank You

Marc Haber, Senior Offering Manager - Data Governance and Catalog
Rick Buglio, Digital Technical Engagement - IBM Watson Data and AI

