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

IBM **Cloud and Cognitive Software** Fast Start 2020

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, identity 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

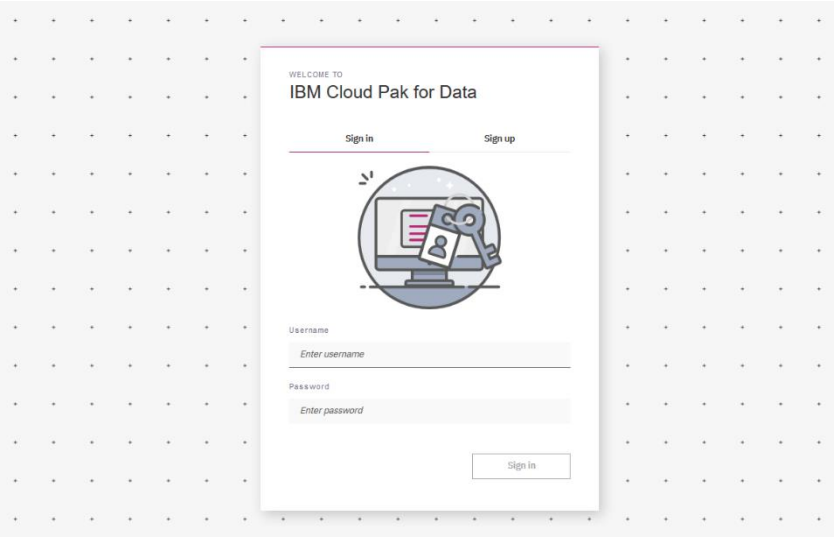
Step 1: Configure Metadata Import and Discovery

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

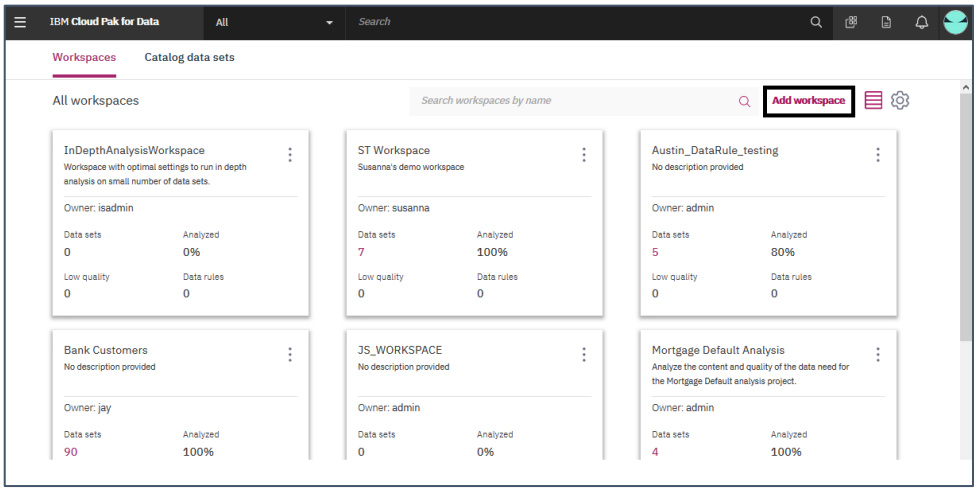
1. Launch Watson Knowledge Catalog from the desktop-shortcut, or enter the provided URL into a supported Internet Browser.

2. Enter the credentials for Watson Knowledge Catalog, as provided.



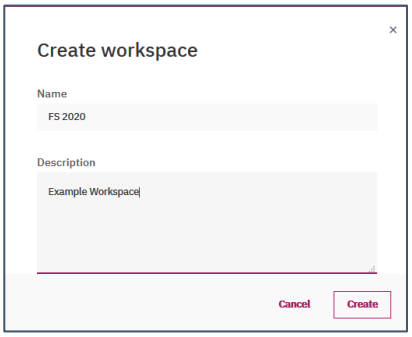
3. Open the navigation menu by selecting the action (from the upper-left corner of the navigation bar)

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.

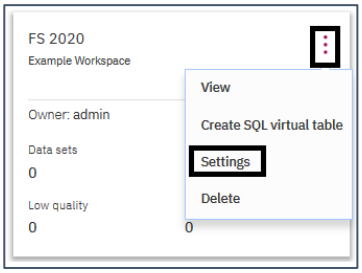


5. Click the action *Add Workspace* to create a new Quality Workspace. The create wizard appears.

   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

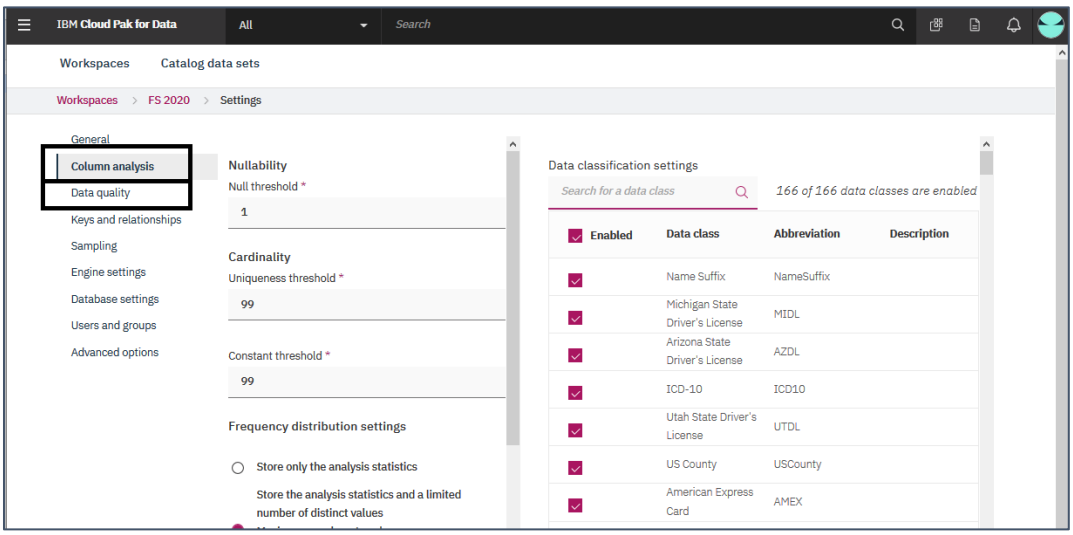


6. From the displayed list of existing Quality Workspaces, find the newly created Workspace *FS 2020 <Your Name>*. Expand the action menu and select *Settings* to review the Quality Dimension settings.



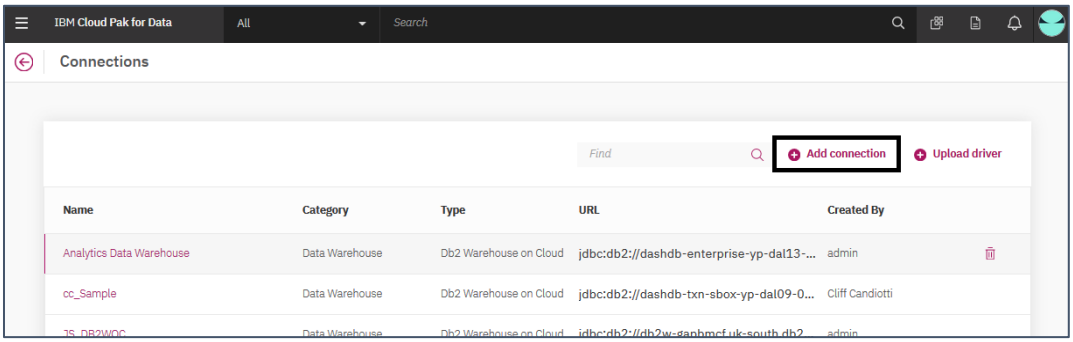
7. From the settings dialog, select the tab *Column Analysis*. Optionally:
    a. Enable or disable the Data Classes to be used when analyzing and evaluating data
    b. Adjust the threshold values for Cardinality and Nullability

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.

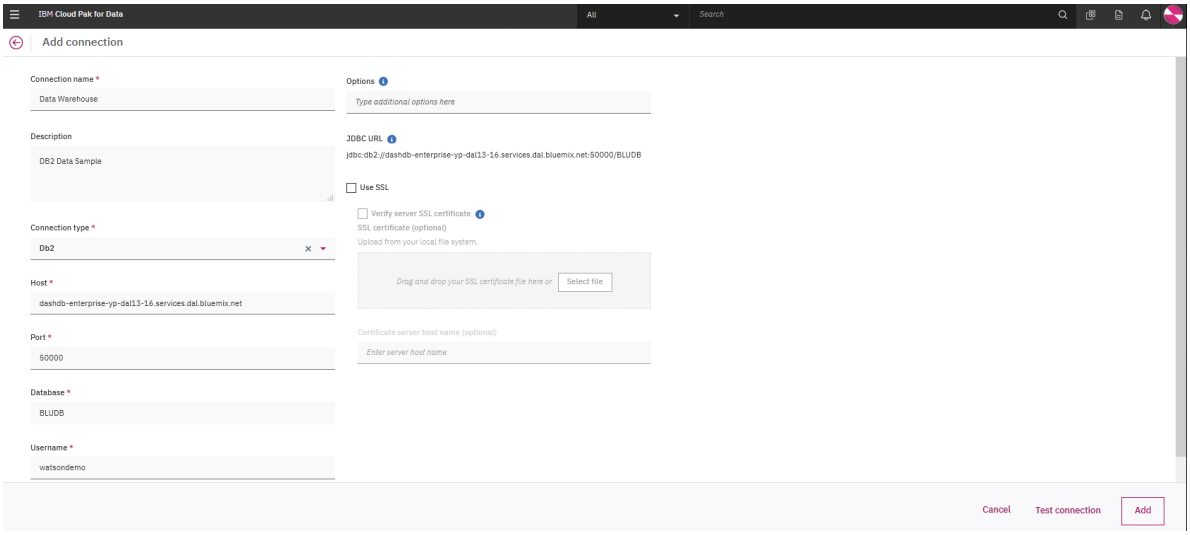
10. Open the navigation menu by selecting the action and select *Connections*. The Connection management dialog displays.



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.bluemix.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 *Wats0nDataandAI!*
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



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

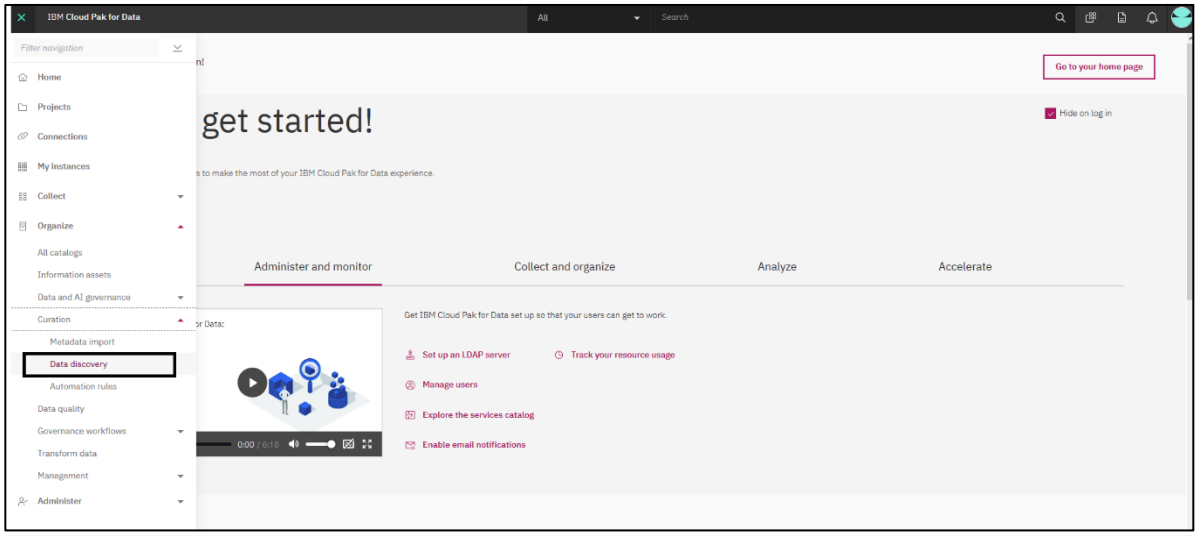
This step will allow the user to explore the process for initiating, reviewing and publishing the Data Discovery results

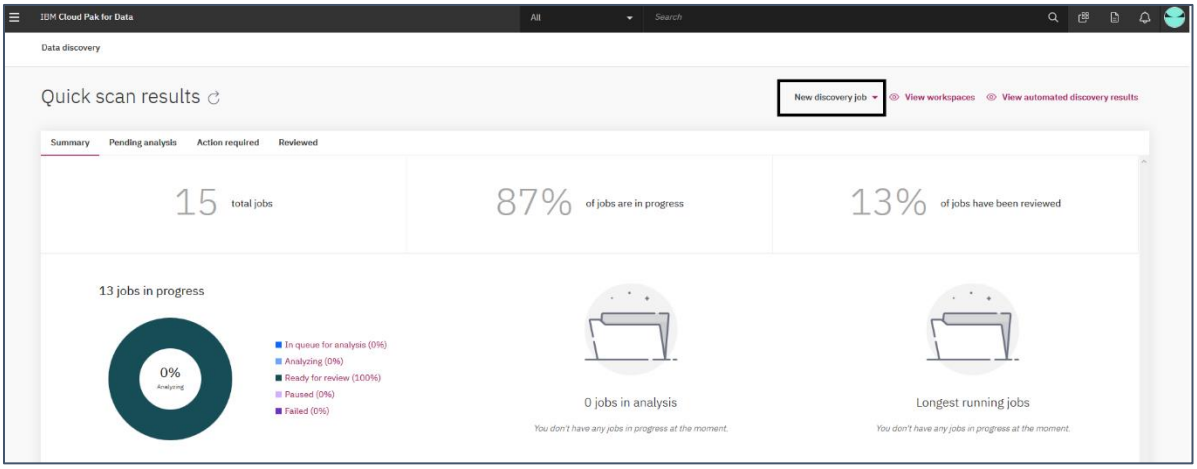
1. If not already open, launch Watson Knowledge Catalog from the desktop-shortcut, or enter the provided URL into a supported Internet Browser.

2. Enter the credentials for Watson Knowledge Catalog, as provided

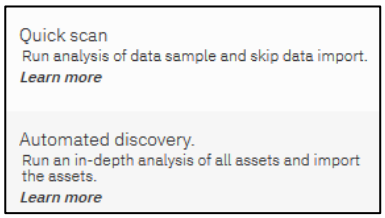
3. Open the navigation menu by selecting the ☰ action (from the upper-left corner of the navigation bar)

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.

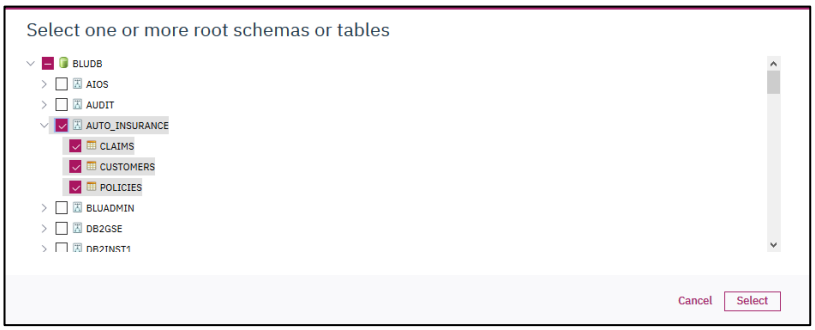


6. Click the action *Select a Connection* and then

   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*

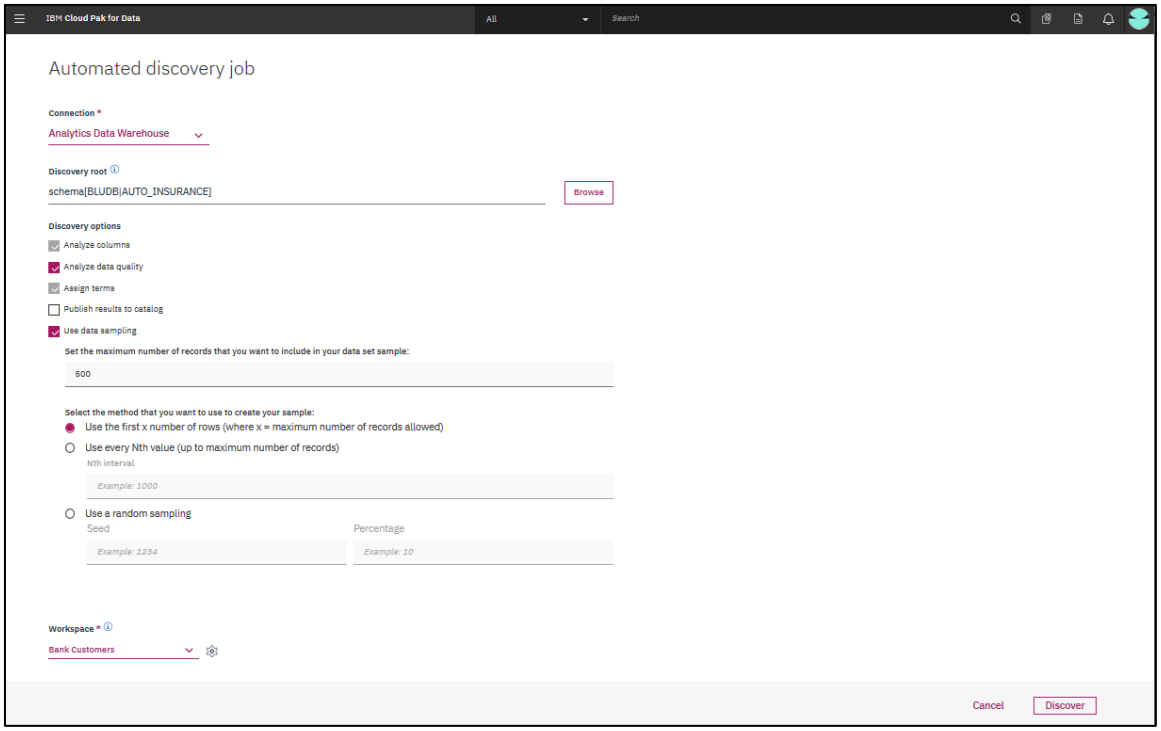


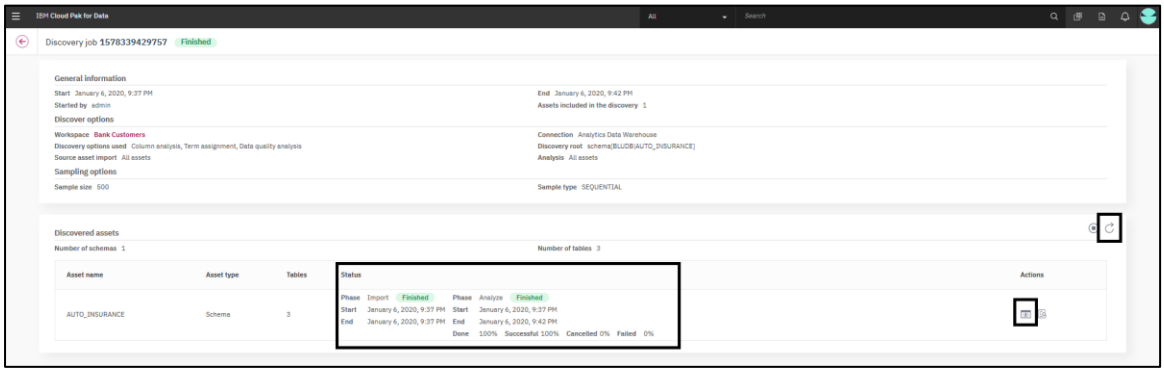
8. Select the following Discovery Options:

   a. Analyze Columns. Examine the characteristics and identify the matching Data Classes.
   b. Analyze Data Quality. Calculate the Quality Score based upon the Quality Dimensions.
   c. Assign Terms. Suggest and assign Business Terms.
   d. Use Data Sampling. Set the *Maximum Number of Records* to 500
   e. 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*

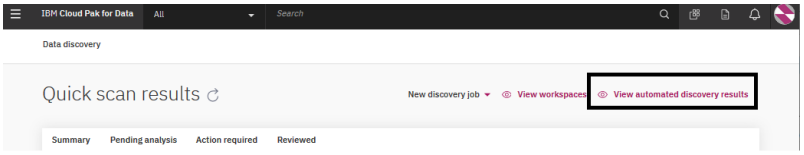
STOP. Due to constraints, you will not be completing the analysis process and viewing the calculated results. Rather, you will be directed to review previous analysis results and take the same steps to publish the results.



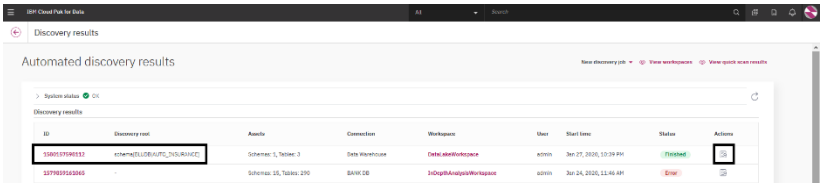


10. Click the action *Cancel* to cancel the process and return to the previous screen.

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



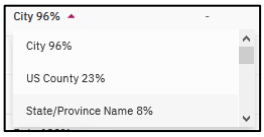
12. From the list of previous results, identity the result for the Database Schema Auto_Insurance and click the *Discovery ID* to view its details



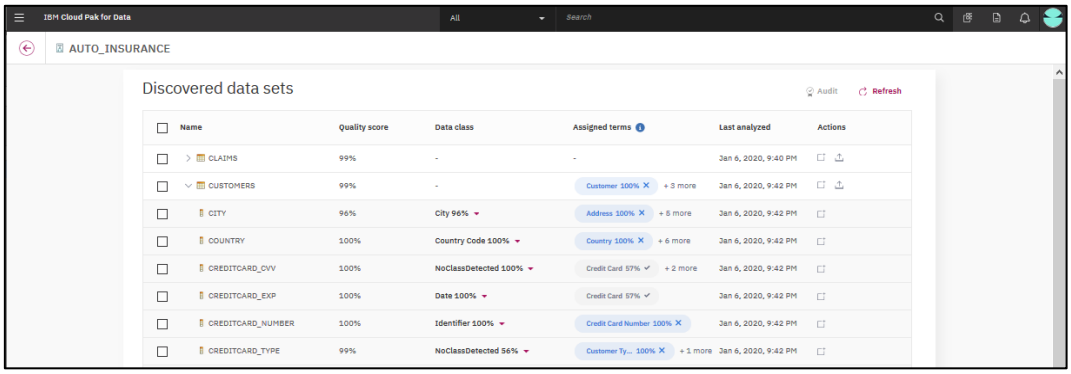
13. Click the action *Review Discovery Results* to view the results for the Schema *Auto_Insurance* and its included Tables.

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.

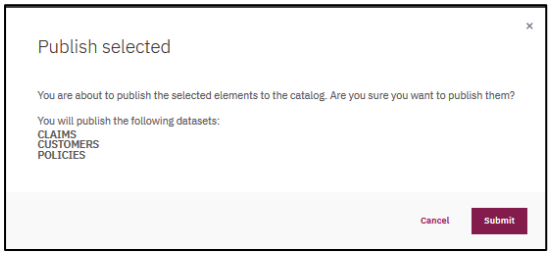


   b. For the column *Email Address* review the Term Suggestion.



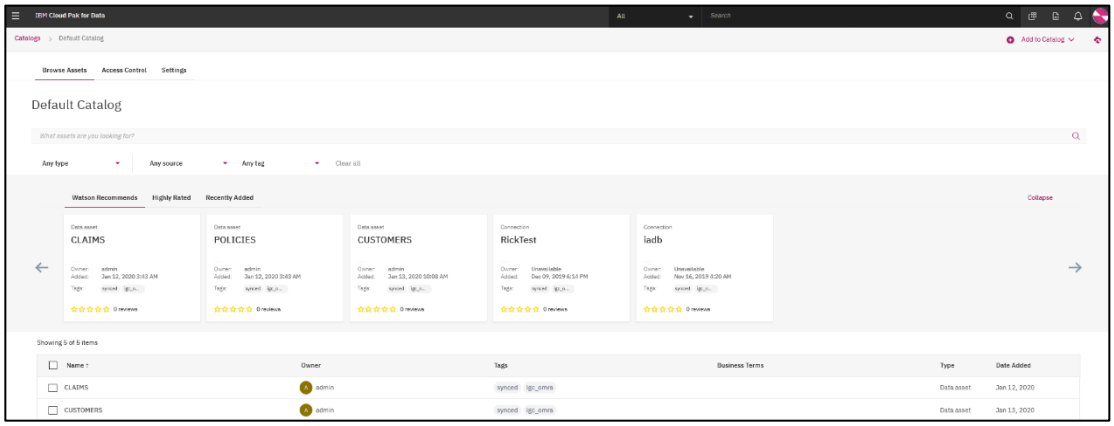
15. Select all Tables and Columns by clicking the checkbox within the header section and click the action *Publish* to publish the Table and their Quality Score and assigned Data Class and Term to the Catalog. The Publish summary dialog appears.

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

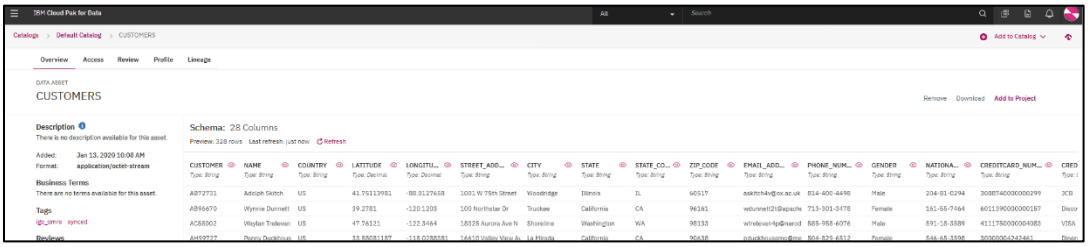


17. optionally, 0pen the navigation menu by selecting the ☰ action (from the upper-left corner of the navigation bar) and expand the section *Organize* and select the item *All Catalogs*. The list of Catalogs will display. Catalogs include a sub-set of Data Sets which specific users have been granted 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.



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.

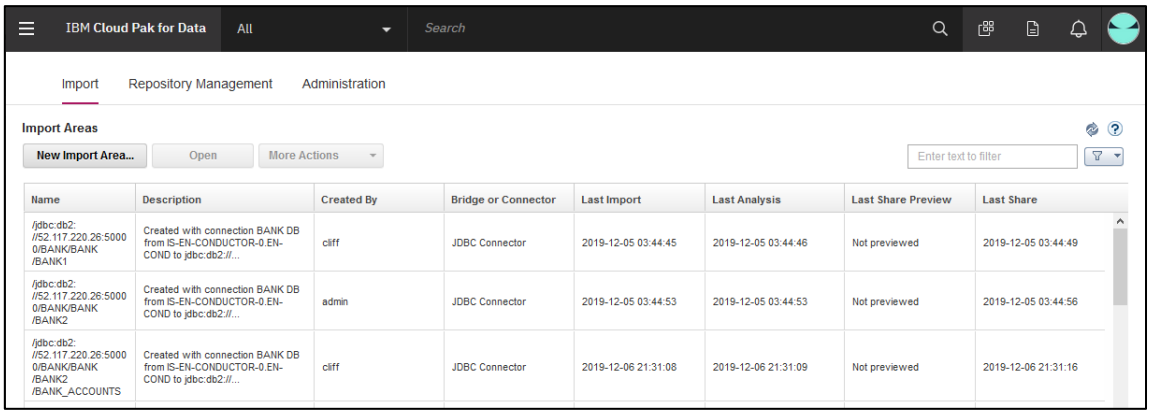
22. Optionally: Select the tab *Review* and set a Rating add a new Comment for the Table.

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.



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

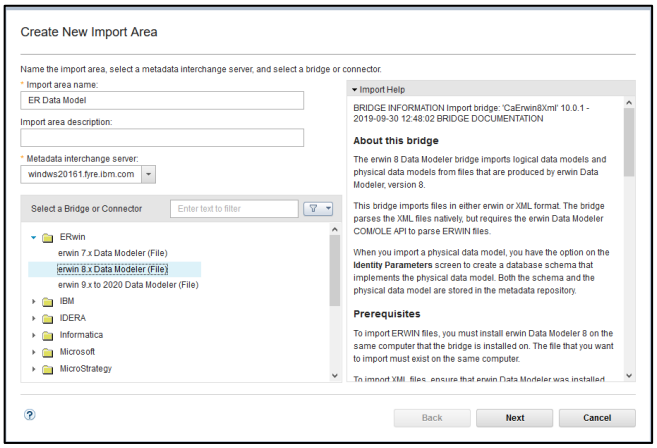
   a. Enter the name of the Import Area *ER Model <Your Name>*. The name uniquely identifies the import and its defined parameters.

   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

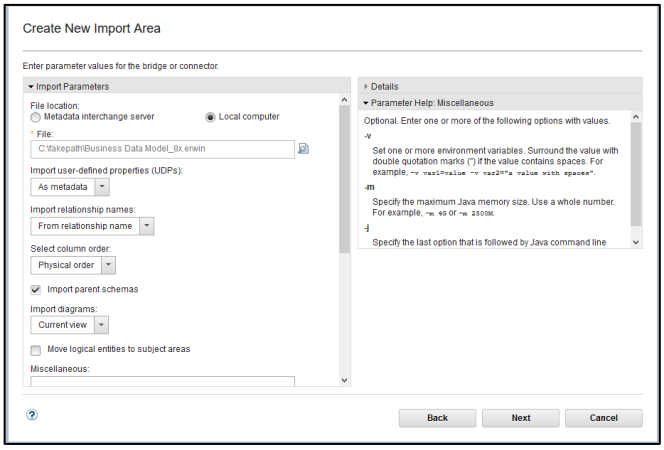


5. Step 2: Create New Import

   a. Browse to the provided Box Folder and download the file *Business Data Model.xml* to your machine

   b. 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*

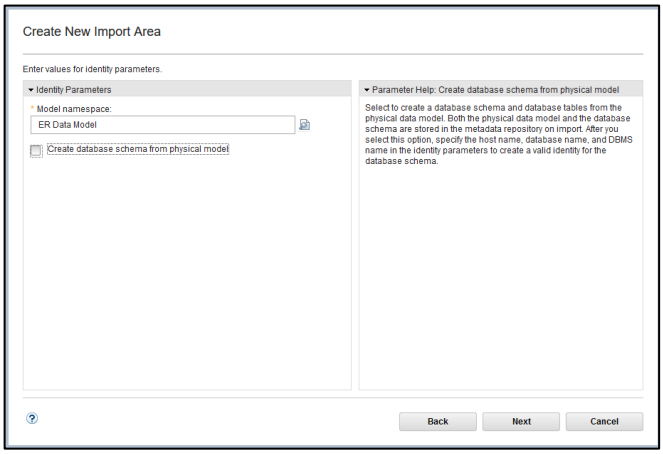
   c. Click Next to continue



6. Step 3: Create Import Area

   a. Enter the Model Namespace *ER Model <Your Name>*

   b. Click Next to continue



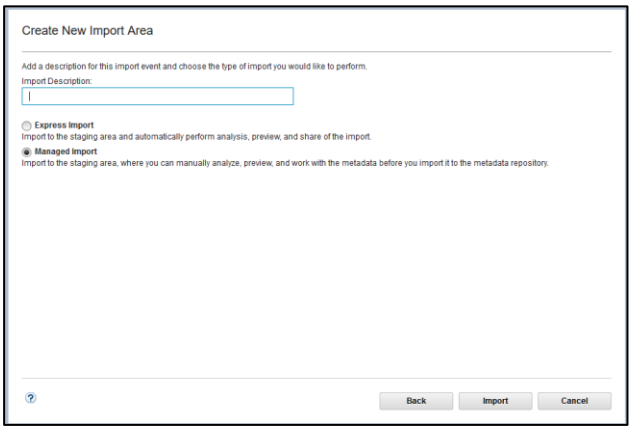
7. Step 4: Create Import Area

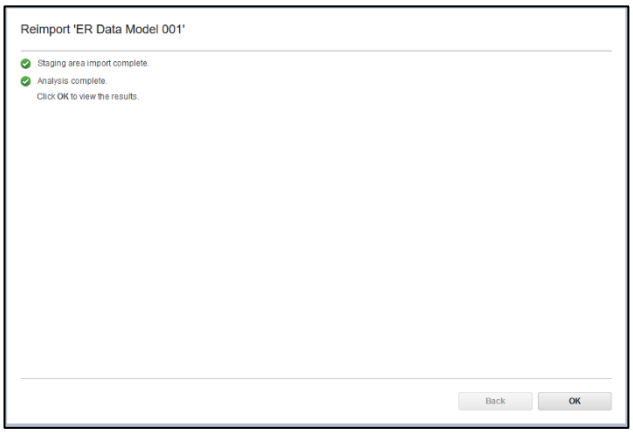
   a. Optional enter an import description. The description is specific to this import, whereas the Import Area may be re-used for subsequent import.

   b. Select *Managed Import* to review the imported results, prior to sharing to the Catalog

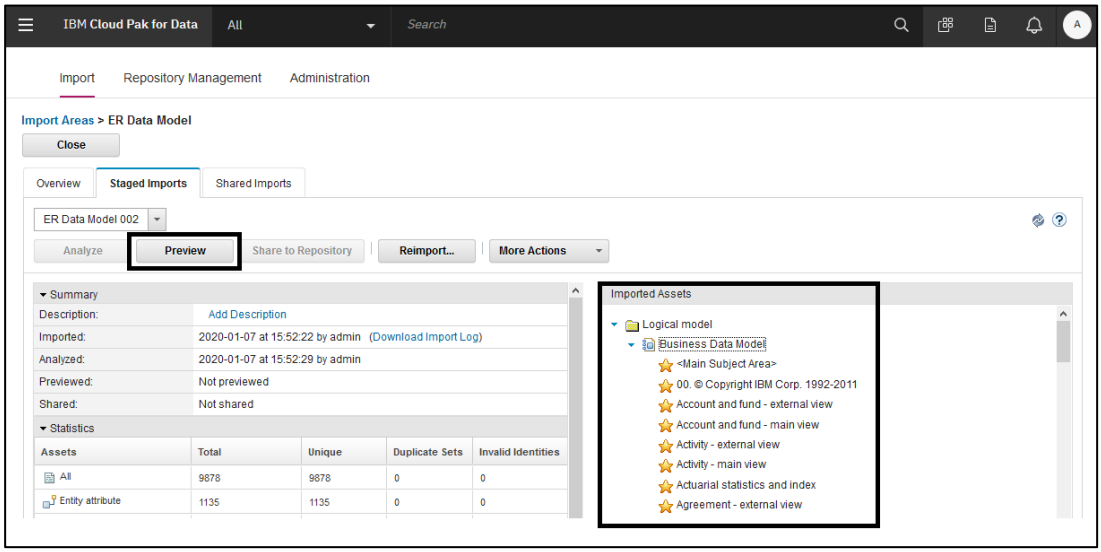
   c. Click Import to complete the import process and view the results. The import may take a few minutes to complete.

   d. Click OK to close the Import Summary dialog

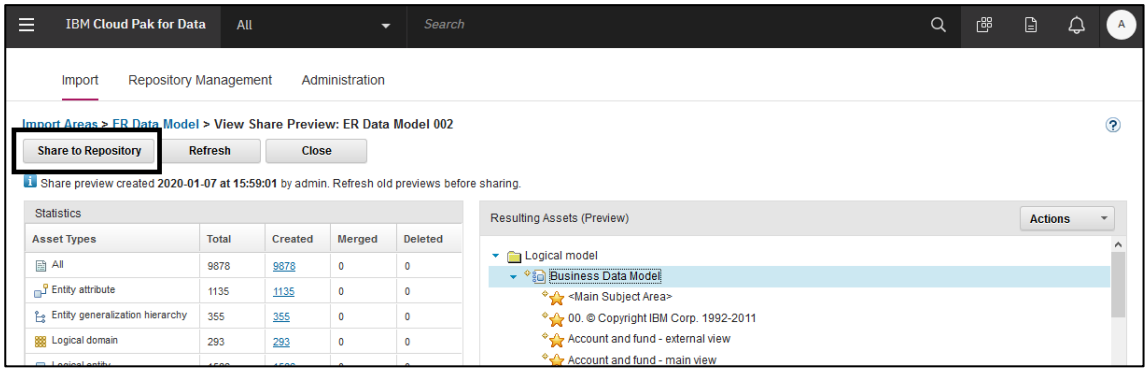




8. View the Staged Import results, representing the Model and its Entities and Attributes captured from the import file. Click *Preview* to preview and analyze the content, prior to publishing and sharing with Catalog.



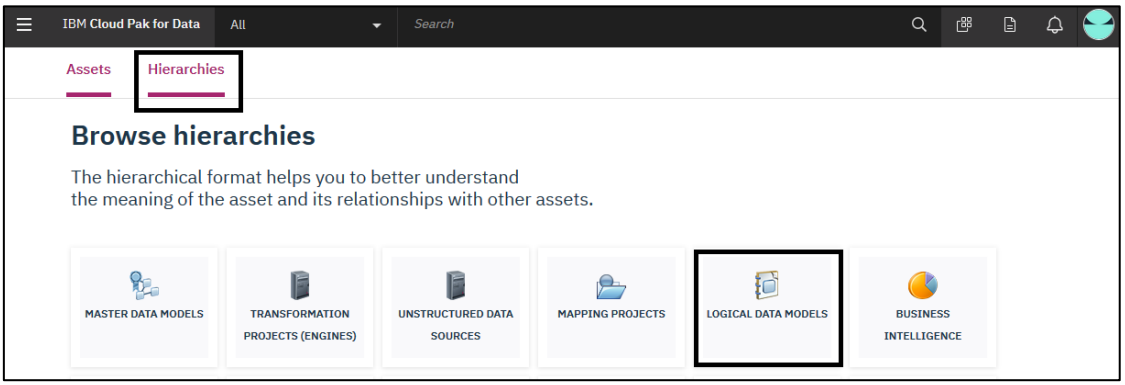
9. View the Analysis results, the number of items to be created or merged into the Catalog. Click *Share to Repository* to continue and complete the import and share task. A confirmation dialog will ask you to confirm the action.



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

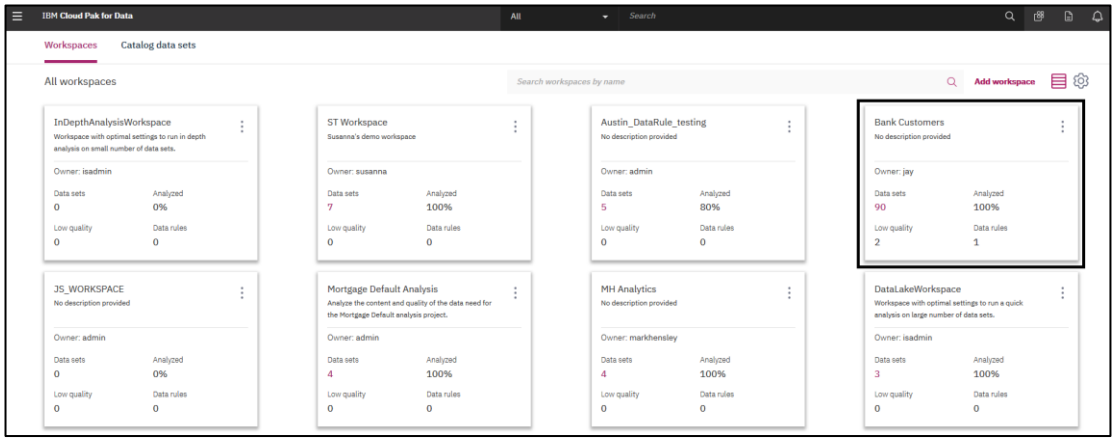
This completes Step 3 and the Metadata Import process

Step 4: Data Quality

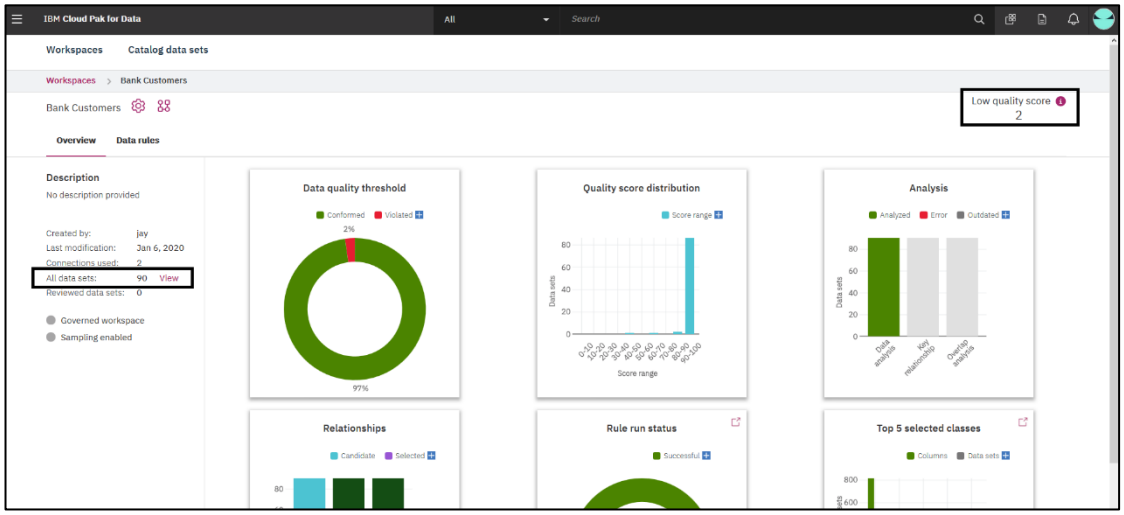
This step will allow the user to explore the process for reviewing the Data Quality results

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.



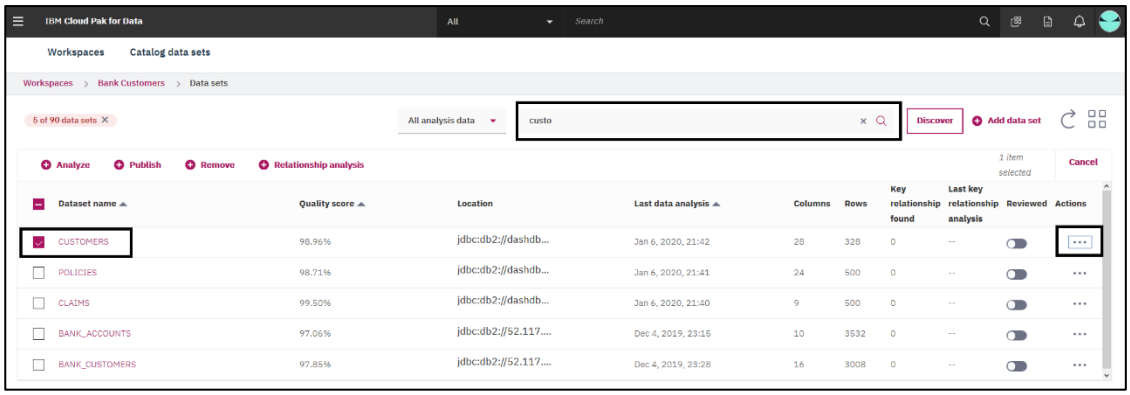
3. Click the workspace *DataLakeWarehouse* to view its results and status. Explore the displayed dashboards, alerts for Data Set of low quality score and number of Data Sets contained by the



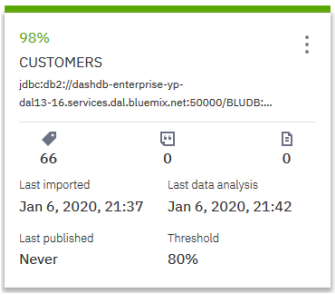
Workspace.

4. From the left pane, note the count of Data Sets used within the Workspace. Click *View* to explore further the Data Sets.

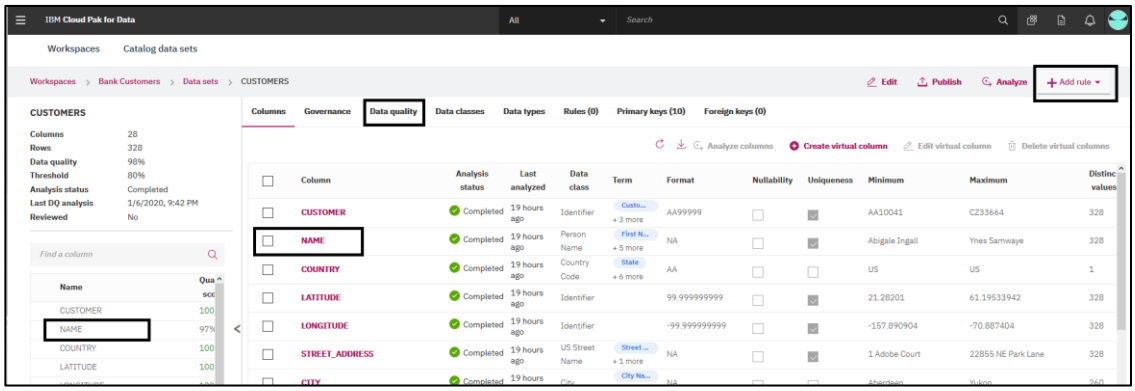
5. Filter the ensuing list and find and select the Data Set *Customers*.



6. Toggle the display to view the results within a grid by clicking the action icon ⁞⁞.



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



8. Select the tab *Data Quality* to view the data quality violations for the Data Set.

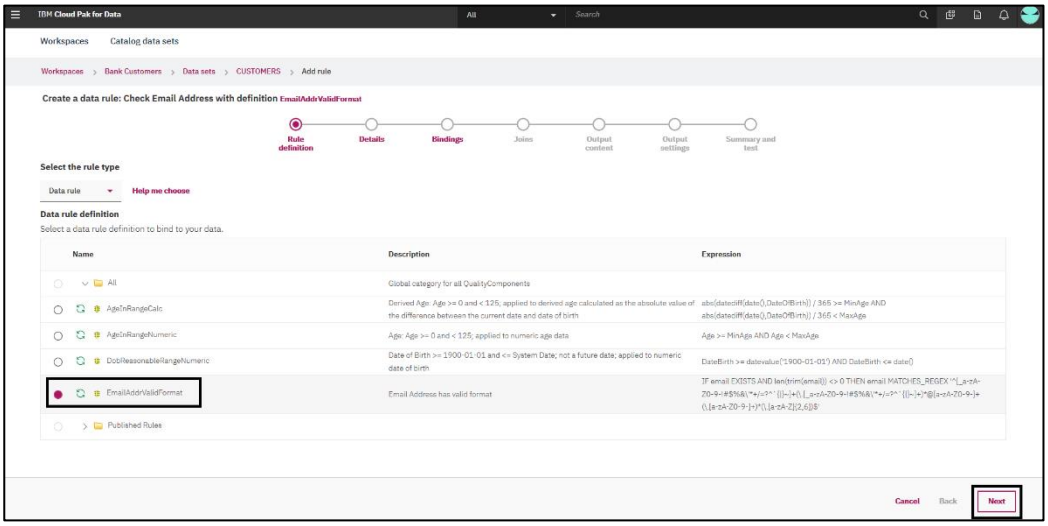
    a. Click the violation *Data Class Violations* to view which Data Classes have reported violations
    b. Click the Data Class *State Code* to view the specific records that have violated the Data Class

    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

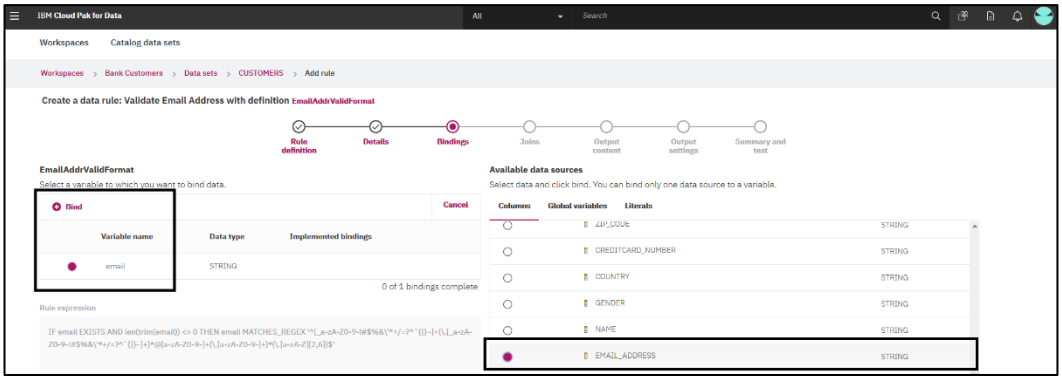


    b. Enter the name for the Data Rule *Validate Email Address*. Optionally enter a Description for the Rule. Click Next to continue

    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

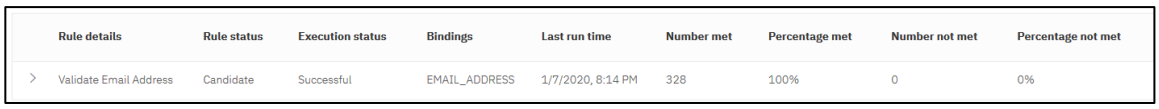
iv. Click Next to continue.



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

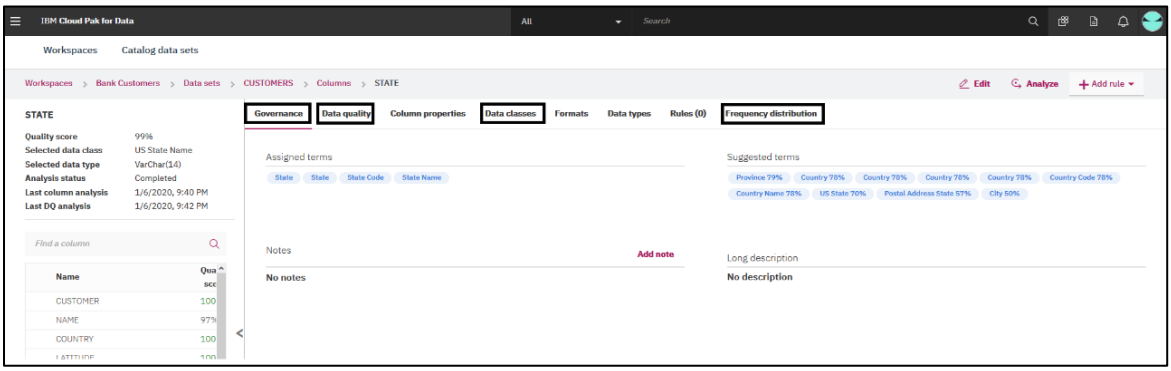
11. Click ⋯, the action menu, for the new Data Rule, and click the action *Run* to immediately invoke the Rule. Confirm the Rule Execution. It may take several minutes for the Rule execution to complete.

12. <u>Refresh the display of Watson Knowledge Catalog</u> until the Rule execution completes.



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.

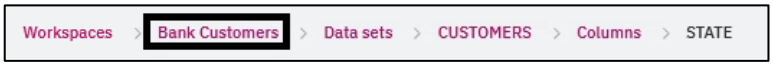


15. Click on the tab Data Quality for the Data Set *Customers* to view the resulting quality dimension details. Those details determined the Quality Score. Expand the violation *Inconsistent capitalization* and explore each violation.

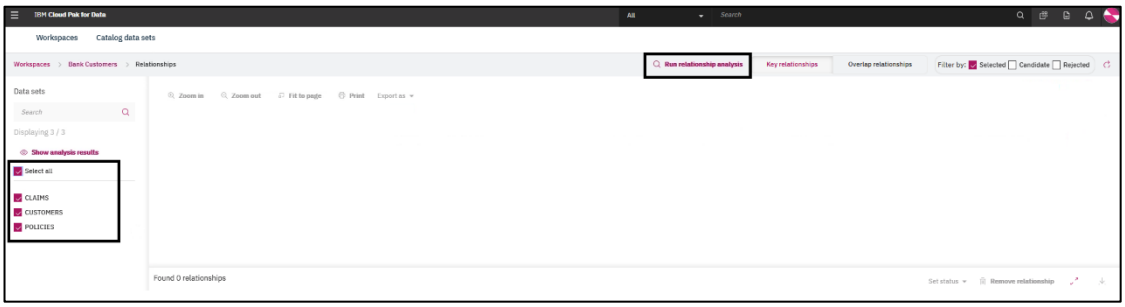
16. Click on the tab Data Classes to view the suggested and selected Data Class. The Data Class of the column can help determine the sensitivity of the information and can be used to set and enforce Data Protection Rules that mask or otherwise hide the data from preview.

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



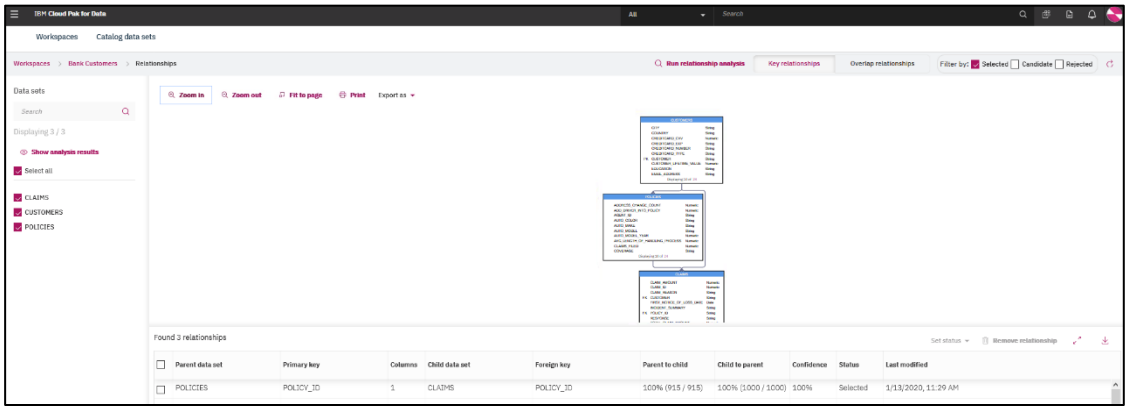
19. From the Workspace view, click the action (from the left pane) to view or initiate an analysis of Key Relationships. The Analysis screen displays.



   a. Select all Tables to be included in the Key Relationship analysis
   b. Select the action *Run Relationship Analysis* to invoke the action
   c. 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.



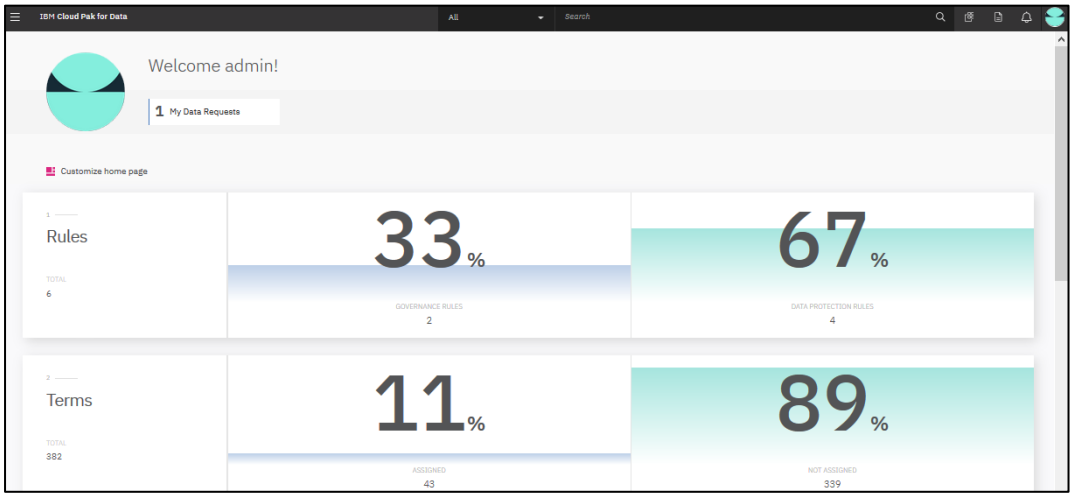
This completes Step 4 and the Data Quality review

Step 5: Search and Explore Business Ready Information

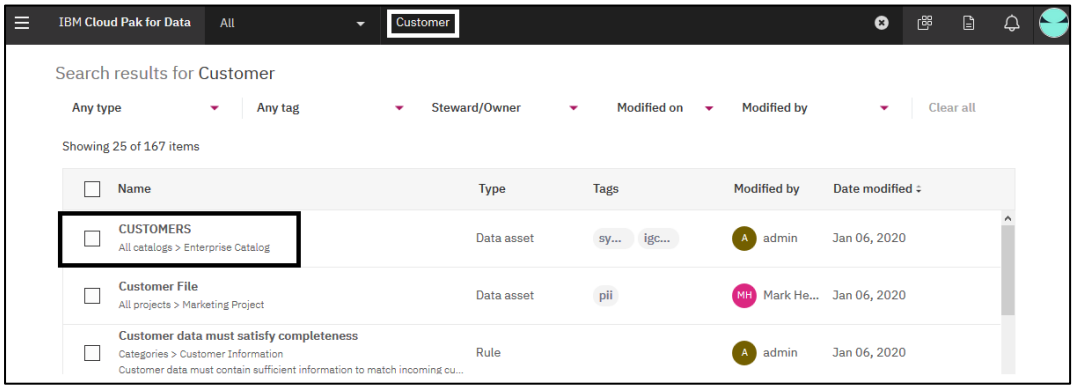
This step will allow the user to search and explore the Catalog and find relevant information, leveraging the automated Discovery services that have applied meaning and given quality to the information

1. If not already open, logon to Watson Knowledge Catalog.

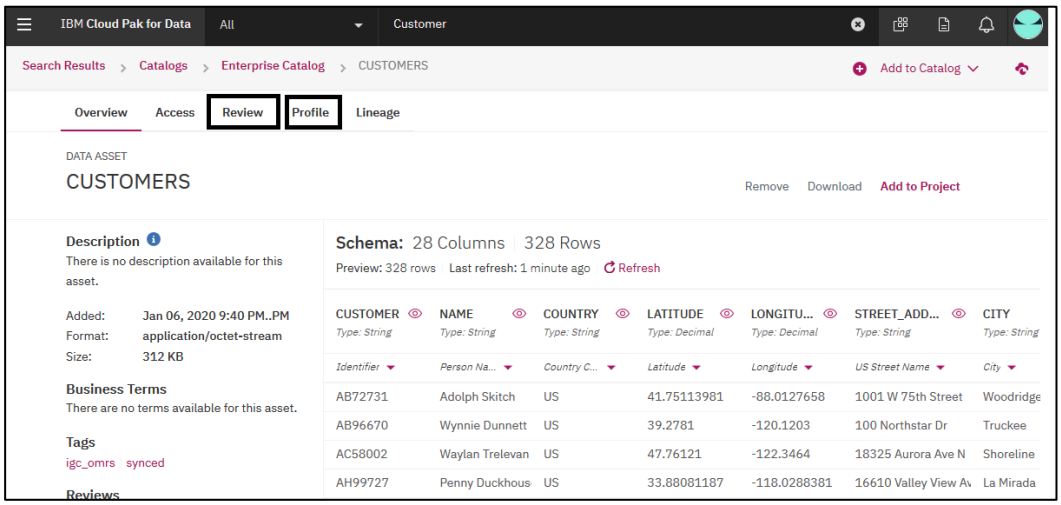
2. If the Homepage is not displayed, navigate to the Homepage by opening the navigation menu by selecting the ☰ action and select *Homepage*.



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



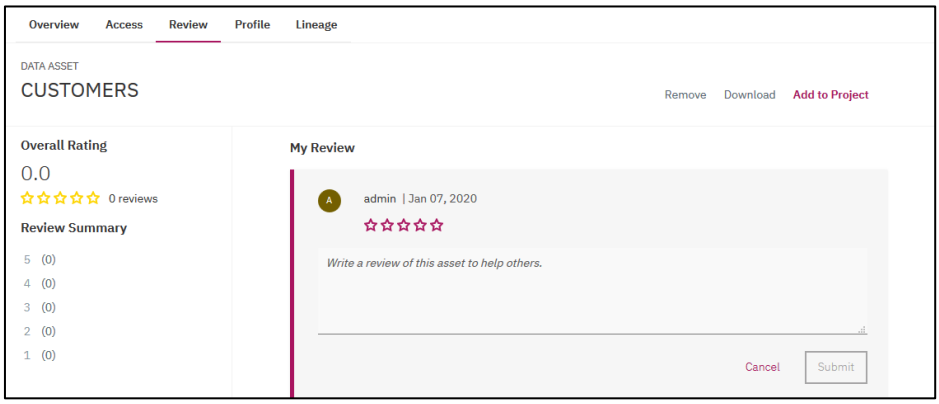
4. Select the table *Customers* to view its details from the Catalog



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.



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