Watson Knowledge Catalog (WKC)

in Cloud Pak for Data V3.5

Hands-on lab

Robert Kern, IBM Cloud Pak for Data Client Experience Team

January 25, 2021

In this hands-on lab you will experience how to work with the three pillars of Watson Knowledge Catalog (WKC) in IBM Cloud Pak for Data: Data governance, data curation & quality, and data consumption. After completing some preparation steps in chapter **Error! Reference source not found.** and 2, you will establish a basic governance framework in chapter 3. It consists of a simple set of policies, rules, and a business glossary composed of a hierarchy of business terms. In chapter 4 you will get an impression of how to work with governed data in catalog and projects. You will learn how to profile data and how to create data protection rules to prevent unauthorized users from accessing sensitive data. In chapter 5 you will use data discovery to analyze data and automatically map business terms and data classes to data assets and fields. You will work with the discovered data in data quality projects and learn how to create quality rules and automation rules. Finally, you will explore the characteristics of Information assets view and the Default catalog in chapter 6.

*Source of introductory session header texts:* [*Official CPD V3.5 documentation*](https://www.ibm.com/support/knowledgecenter/SSQNUZ_3.5.0/cpd/overview/welcome.html)*.*

Table of Contents

[1 Introduction 4](#_Toc62428671)

[2 Preparation 5](#_Toc62428672)

[2.1 Creating your personal user account 5](#_Toc62428673)

[2.2 Adjusting the default catalog permissions 5](#_Toc62428674)

[2.3 Adjusting the platform connections permissions 6](#_Toc62428675)

[2.4 Downloading the workshop data 6](#_Toc62428676)

[3 Creating governance artifacts 7](#_Toc62428677)

[3.1 Governing artifacts with workflows 7](#_Toc62428678)

[3.1.1 Creating governance workflows 7](#_Toc62428679)

[3.1.2 Testing the governance workflow 8](#_Toc62428680)

[3.1.3 Deleting the newly created business term 8](#_Toc62428681)

[3.1.4 Disabling the workflow 9](#_Toc62428682)

[3.2 Importing a business glossary 9](#_Toc62428683)

[3.2.1 Importing categories 9](#_Toc62428684)

[3.2.2 Adjusting the category permissions 9](#_Toc62428685)

[3.2.3 Importing terms 10](#_Toc62428686)

[3.3 Governing & protecting data with policies & rules 11](#_Toc62428687)

[3.3.1 Creating a category 11](#_Toc62428688)

[3.3.2 Creating policies 11](#_Toc62428689)

[3.3.3 Creating governance rules 12](#_Toc62428690)

[3.3.4 Creating data protection rules 12](#_Toc62428691)

[4 Working with governed data assets in projects and catalogs 14](#_Toc62428692)

[4.1.1 Importing project data from a zip file 14](#_Toc62428693)

[4.1.2 Creating data profiles in projects 14](#_Toc62428694)

[4.1.3 Creating a governance catalog 16](#_Toc62428695)

[4.1.4 Adding a project data set to a catalog 16](#_Toc62428696)

[4.1.5 Validate effect of data protection rule 17](#_Toc62428697)

[5 Data curation & data quality 18](#_Toc62428698)

[5.1 Creating a data connection 18](#_Toc62428699)

[5.2 Discovering data: Automated Discovery 18](#_Toc62428700)

[5.2.1 Starting the Automated Discovery process 18](#_Toc62428701)

[5.2.2 Quick review of Automated Discovery results 19](#_Toc62428702)

[5.2.3 Detailed review of Automated Discovery results 20](#_Toc62428703)

[5.3 Discovering data: Quick scan 22](#_Toc62428704)

[5.3.1 Starting the QuickScan process 22](#_Toc62428705)

[5.3.2 Publishing Quick scan results 23](#_Toc62428706)

[5.4 Data quality related rules 24](#_Toc62428707)

[5.4.1 Creating rule definitions 24](#_Toc62428708)

[5.4.2 Creating quality rules 25](#_Toc62428709)

[5.4.3 Creating automation rules 26](#_Toc62428710)

[6 Working with the Information Assets view and the Default Catalog 28](#_Toc62428711)

[6.1 Browsing data sets in the Information assets view 28](#_Toc62428712)

[6.2 Browsing data sets in the Default catalog 29](#_Toc62428713)

# Introduction

Before starting with the actual hands-on lab, make yourself comfortable with the Cloud Pak for Data (CPD) knowledge center. It can be accessed under the following URL: <https://www.ibm.com/support/knowledgecenter/SSQNUZ_3.5.0/cpd/overview/welcome.html>

The ***Table of Contents*** tab on the left is the primary navigating tree that points to all sections of the document. The most relevant ones for the *Watson Knowledge Catalog* service are summarized in the following table:

|  |  |
| --- | --- |
| **Topics** | **Section** |
| Limitations and known issues | Overview -> Limitations and known issues -> Watson Knowledge Catalog |
| Overview, installation, migration, backup & restore, upgrade | Services and integrations -> Services in the catalog -> Data governance services -> Watson Knowledge Catalog |
| Creating and administering projects | Projects |
| Accessing and searching for data, connecting to data sources, supported data sources | Accessing data |
| Governance, catalogs, governance artifacts, data discovery, metadata import | Enterprise governance -> Watson Knowledge Catalog |
| Governance workflows | Enterprise governance -> Workflows |
| Developer resources & APIs | Developer resources -> Available APIs -> Watson Data API |
| Troubleshooting | Troubleshooting -> Watson Knowledge Catalog |

# Preparation

This chapter covers some preparation steps for running the WKC hands-on labs.

## Creating your personal user account

This section describes how to create and adjust your CPD user id to be used for the workshop. If you already have an account, verify that the correct roles and permissions are assigned to it.

1. Navigate to the URL of the CPD cluster using a recent version of the Mozilla Firefox, Google Chrome, or Microsoft Edge browser.
2. Logon to the CPD cluster using the ***admin*** user. If you don’t know the admin password, ask a colleague to create the account for you.
3. In the main menu, navigate to ***Administration-> User management****.*
4. In the *Users* tab, click on ***New user*** in the upper right corner.
5. On the *New user* panel enter
   1. ***User***: Your first and last name, e.g. “Robert Kern”
   2. ***Username***: Your user-id, e.g. “robkern”
   3. ***Email***: Your IBM email address, e.g. “robkern@de.ibm.com”
   4. ***Password*** / ***Re-enter password***: Choose a password.
6. Select all available roles and click ***Create***.

***Note:*** *In order to keep it simple, you will use a single user-id for running through the workshop exercises. You may restrict your user to certain roles later in order to see which subset of capabilities is available for the individual user roles.*

1. In the ***Roles*** tab, click on the ***Data Steward*** role and check whether the ***Access advanced governance capabilities*** permission was assigned to it. If it isn’t, click the ***Edit*** button on the right, select the permission, and click save.

***Important:*** *The* ***Access advanced governance capabilities*** *permission enables menu options* ***Import metadata****,* ***Define custom attributes****, and* ***View operational data lineage****. By default, it isn’t assigned to any user role, i.e. the corresponding menu options are hidden from all users.*

*By default, the Advanced*

## Adjusting the default catalog permissions

The Default catalog is automatically created when installing CPD. By default, user *admin* is the only user who can access it. Perform the following steps to add your user-id:

1. In the main menu, navigate to ***Catalogs-> All catalogs****.*
2. Open the ***Default catalog*** and navigate to the ***Access control*** tab.
3. Click ***Add collaborator*** on the right.
4. On the *Add collaborators as…* panel
   1. Select ***Role***: *Admin*,
   2. In the ***Collaborators*** section, search for your newly created user, select it, then click ***Add***.

## Adjusting the platform connections permissions

By default, new the admin user is the only user who can administrate the platform connections while all other users can only view them. Perform the following steps to add your newly created user as a collaborator to the platform connections:

1. In the main menu, navigate to ***Data-> Platform connections****.*
2. In the ***Access Control*** section click on ***Add collaborator*** on the right.
3. On the *Add collaborators as…* panel
   1. Select ***Role***: *Admin*,
   2. In the ***Collaborators*** section, search for your newly created user, select it, then click ***Add***.
4. Log out from CPD by clicking the icon in the upper right corner of the screen and selecting ***Log out***. Then confirm by clicking ***Log out*** again. Afterwards, login with your newly created user.

***Hint:*** *Under* ***Governance -> Data quality****, there are four projects (DataLakeWorkspace, InDepthAnalysisWorkspace, PIIWorkspace,* and *UGDefaultWorkspace) which created during the WKC installation and which are available to the admin user only by default. They won’t be needed for this workshop but if you want to use them, the admin user needs to add your user in the “Settings” tab of the corresponding project under “Users and groups”.*

## Downloading the workshop data

This hands-on lab is based on the ***IBM Data and AI Accelerators*** for Cloud Pak for Data.

Use the following steps to download the resources of the ***Utilities Payment Risk Prediction*** industry accelerator:

1. Navigate to the Industry Accelerator main page: <https://community.ibm.com/accelerators/>
2. Search for *“Utilities Payment”* using the search bar
3. Open ***Utilities Payment Risk Prediction*** in the search result
4. Click on button ***Get accelerator now***.
5. Download and unpack the *utilities-payment-risk-prediction-industry-accelerator.tar.gz* file for CPD V3.5.0.0 as described on the page.
6. In a subfolder, the archive contains the following files required in a later step:
   1. utilities-payment-risk-prediction-glossary-categories.csv
   2. utilities-payment-risk-prediction-glossary-terms.csv

# Creating governance artifacts

In this chapter some governance artifacts are created which are building the foundation for the data profiling and data quality analysis in the subsequent chapters.

***Note:*** *Governance artifacts like categories, business terms, data classes, classifications, reference data sets, policies, governance rules and also governance workflows, are global objects within CPD that may be shared by all users depending on the permissions of the category to which they belong to.*

## Governing artifacts with workflows

Workflows enforce a task-based process to control the creating, modifying, and deleting of governance artifacts like business terms, data classes, classifications, reference data, policies, and governance rules. Each type of artifact can be assigned to a workflow that defines the sequence of steps that must be completed before an artifact is published and available to use. The steps can include authoring, approving, reviewing, and publishing. A set of users are assigned to work on each step.

Initially, the ***default workflow configuration*** applies to all governance artifacts. It defines, that all users having the *author governance artifacts* permission are allowed to modify, delete, or publish new governance artifacts.

In this section you will create **your own instance** of a governance workflow and delete it afterwards in order to avoid conflicts with other workshop users.

### Creating governance workflows

Perform the following steps to create a governance workflow for business terms:

1. In the main menu, navigate to ***Administration -> Workflow management****.*
2. On the *Workflow Types* tab, click on the card ***Governance artifact management***.
3. Navigate to the ***Workflow configurations*** tab and click on click on ***New configuration*** on the right.
4. On the *New workflow configuration* panel, enter ***Name*** “<your-id>s review and approval of business terms”, where *user-id* is your username on the CPD cluster.
5. Under *Choose a workflow template*, select ***One approval step and one review step***.
6. Confirm by clicking on ***New workflow configuration***.
7. The workflow configuration has four tabs on the left, *Workflow conditions*, *Approve*, *Review*, and *Publish*.
8. On the ***Workflow conditions*** tab, click on ***Add condition***, leave the *All categories* toggle selected and in row ***Business term***, select all actions, ***Create***, ***Update***, ***Delete***, and ***Import***. This will disable the default workflow configuration and enable your newly created workflow configuration for all those actions performed on business terms.
9. Navigate to ***each*** of the tabs, *Approve*, *Review*, and *Publish* on the left. Select the ***Users*** checkbox and then click on the ***Add assignees*** button on the right.
10. On the *Add assignees* panel, select the ***admin*** user and click ***Add***. Feel free to assign additional users for the individual roles and use them for testing the individual workflow steps later.
11. Click ***Activate*** in the upper right corner of the screen to enable the workflow and confirm the upcoming dialog by clicking on the ***Activate*** button. If the action fais go back to step 8.
12. The workflow is now enabled and will be applied to the creation, import, update, or deletion of business terms.

### Testing the governance workflow

Given that your user id has all available permissions, you may run through all steps of the workflow including *authoring*, *review*, *approval*, and *publishing* of a new business term.

1. In the main menu, navigate to ***Governance -> Business terms****.*
2. Click on ***New business term -> Create new business term****.*
3. On the *Create new business term* panel, enter Business term name “<your id>s business term”.
4. Under Primary category, click on ***Change*** and select ***Industry Accelerators***, and confirm by clicking on ***Add***.
5. On the *Create new business term* panel, click ***Save as draft***.
6. A draft object is created that can be deleted or sent for approval:  
   
7. Click on ***Send for approval***, leave the comment field empty, and confirm by clicking on ***Send for approval*** once more.
8. The status of the business term changes to *Ready for approval*.  
   Graphical user interface

   Description automatically generated with low confidence
9. Under ***Take Action***, click on ***Approve***, leave the comment field empty, and confirm by clicking on ***Approve*** once more.
10. The status changes to *Ready for publishing*:  
    
11. Click on ***Publish***, leave the comment field empty, and confirm by clicking on ***Publish*** once more.
12. You now have successfully created a new business term.

### Deleting the newly created business term

The governance workflow applies to the deletion of business terms as well. Perform the following steps, in order to delete the business term created in the previous section:

1. In the main menu, navigate to ***Governance -> Business terms****.*
2. On the *Business terms* panel, search for “<your id>” your business term under the published business terms.
3. Click on your business term to open the details page.
4. Click on the ***three-dot-menu*** () next to the business term name and select ***Mark for deletion*** in order to trigger the deletion workflow, then confirm by again clicking on ***Mark for deletion****.*
5. The status of the business term changes to *Marked for deletion*:  
   
6. Under ***Take Action***, click on ***Approve***, leave the comment field empty, and confirm by clicking on ***Approve*** once more.
7. The status changes to *Ready for publishing* which means that the *deletion* *request* is ready to be published, not the business term:  
   
8. Click on ***Delete*** in order to publish the deletion request and to delete the business term. Then confirm by clicking on ***Delete*** once more.

### Disabling the workflow

In order to avoid conflicts with other workshop users, delete the governance workflow by performing the following steps:

1. In the main menu, navigate to ***Administration -> Workflow management****.*
2. On the *Workflow Types* tab, click on the card ***Governance artifact management***.
3. Navigate to the ***Workflow configurations*** tab.
4. Click to the workflow configuration “<your id>s review and approval of business terms” in order to open the details page.
5. Click on the ***Deactivate*** button on the upper right of the screen*.*
6. Confirm by clicking ***Deactivate*** again.
7. Your workflow has now been deactivated and can be re-activated at a later time if needed.

## Importing a business glossary

In this section you will import the categories and business terms of the industry accelerator downloaded above. You may re-import the business terms even if they are already available. Make sure to use option “*Replace only empty values*” as indicated below in order to avoid conflicts with other workshop users.

### Importing categories

Perform the following steps to import the categories:

1. In the main menu, navigate to ***Governance -> Categories****.*
2. Click on ***Add category -> Import from file*** in the upper right corner.
3. Click ***Add file*** and navigate to the category csv file *utilities-payment-risk-prediction-glossary-categories.csv* downloaded in section 2.4.
4. Click ***Next***.
5. Select option ***Replace only empty values*** and click ***Import****.*
6. If they didn’t exist before, two new categories will be added to the business glossary: *Industry Accelerators* and subcategory *Utilities Payment Risk Prediction*.
7. Click ***Close***.
8. Refresh the browser screen in order to see the newly created category.

### Adjusting the category permissions

If the newly imported categories have not existed before, you need to add the other users as collaborators in order to make the categories accessible to them.  
  
In order to adjust the permissions of the category,

***Important:*** *Access to categories and the artifacts stored within them can be provided to users by adding them as collaborators of the category. However, even if a user does not have access to a specific category the user cannot create a category with the same name within the same place of the category hierarchy. Even if the user cannot see the category, trying to create a category with the same name results in a conflict. For the workshop, we are therefore sharing the same category tree among all users rather than using an individual category tree for each user.*

1. Click on the ***Industry Accelerators*** category in order to open it.
2. Navigate to the ***Access control*** tab.
3. Click on ***Add collaborators*** on the right.
4. If not done by another workshop participant yet, check the ***All users*** group in the *Collaborators* list.
5. On the ***Roles*** drop-down of the group, check the ***Owner*** role and make sure that all other roles are unchecked.
6. Click ***Add***.

### Importing terms

Perform the following steps to import the business terms

1. In the main menu, navigate to ***Governance -> Business terms****.*
2. Click on ***New business term -> Import from file*** in the upper right corner.
3. Click on ***Add file*** and navigate to the term csv file *utilities-payment-risk-prediction-glossary-terms.csv* downloaded above.
4. Click ***Next***.
5. Select option ***Replace only empty values*** and click ***Import****.*
6. If they didn’t exist before, *drafts* for 152 new business terms will be added to the *Utilities Payment Risk Prediction* category of the business glossary.  
   Chart

   Description automatically generated with medium confidence
7. In the ***Task inbox***, review the 152 business terms to be published.
8. Click on ***Publish*** in the right in order to publish the business terms to the glossary. (Note: If a governance workflow has been setup on your CPD cluster you might need to *review* and *approve* the task like in section 3.1.2).

A screenshot of a cell phone

Description automatically generated

## Governing & protecting data with policies & rules

In order to avoid duplicate entries, before creating any of the objects in this section, please check whether it may have been already created by another workshop user.

### Creating a category

Perform the following steps to create an additional category:

1. In the main menu, navigate to ***Governance -> Categories***.
2. Click on ***Add category -> New category*** on the right.
3. On the *New category* panel, specify *Category name* “<your id>s data protection”.
4. Click ***Save***.

### Creating policies

Policies are used to describe and document your organization’s guidelines, regulations, standards or procedures to ensure data and information assets are properly managed and used. A policy is a natural-language description of a governance subject area. Policies describe how to control data. A policy consists of one or more rules.

Perform the following steps to create a set of policies for data protection:

1. In the main menu, navigate to ***Governance -> Policies***.
2. Click on ***New policy*** on the right and select ***Create new policy***.
3. On the *Create new Policy* panel
   1. Enter ***Policy name*** “Sensitive information must be protected”
   2. Under *Primary category*, click on ***Change*** and select the newly created category “<your id>s data protection” and click ***Add***.
   3. Click ***Save as draft***.
4. A draft of the new policy will be created which will need to be published in order for the changes to take effect. The following options will be displayed: (Note: If a governance workflow has been setup on your CPD cluster you might need to *review* and *approve* the task like in section 3.1.2)  
   
5. Click ***Publish***, leave the comment field empty and confirm with ***Publish***.
6. Follow steps 1-4 again in order to create additional policies to be used as subpolicies:
   1. ***Policy name****: “*Protect confidential information”
   2. ***Policy name****:* “Protect personal identifiable information (PII)”
   3. ***Policy name****:* “Protect sensitive personal information (SPI)”
7. In the main menu, navigate to ***Governance -> Policies***.
8. Open policy *“Sensitive information must be protected”.*
9. In the *Subpolicies* section, click ***Add subpolicies***, select the three newly created policies and click ***Add***.
10. ***Publish*** the modified policy according to steps 4 and 5 above. The policy should have the following subpolicies now:  
    A screenshot of a cell phone

    Description automatically generated

### Creating governance rules

Governance rules provide the business description of the required behavior or actions to be taken to implement a given governance policy. These are business descriptive rules and can’t be enforced, unlike other rules that you can define in Watson Knowledge Catalog such as data protection rules, data rules, data quality rules or automation rules.

Perform the following steps to create a governance rule:

1. In the main menu, navigate to ***Governance -> Rules***.
2. Click on ***New rule*** on the right and select ***Create new rule***.
3. Click on the ***Governance rule*** icon.
4. On the *Create new Governance rule* panel
   1. Enter ***Governance rule name*** “Hide personal identifiable information”
   2. Under *Primary category*, click on ***Change*** and select category “<your id>s data protection” and click ***Add***.
   3. Click ***Save as draft***.
5. In the ***Parent policies*** section of the policy draft screen, click ***Add policy***, select policy “Protect personal identifiable information (PII)” and click ***Add***.
6. ***Publish*** the modified policy according to steps described in section 3.3.2.

### Creating data protection rules

Data protection rules apply to all governed catalogs and all assets within these catalogs. Data protection rules are automatically enforced when a catalog member attempts to view or act on a data asset in a governed catalog to prevent unauthorized users from accessing sensitive data. However, if the user who is trying to access the asset is the owner of the asset (by default, the user who created the asset), then access is always granted.

***Important:*** *Data masking is limited to certain data formats (relational data, CSV, Avro, partitioned data, and Parquet files) and tools (projects, catalogs, data refinery, data virtualization). Data is* ***not being masked*** *by tools that access the original data source (e.g. notebooks, model builder dashboards).*

In this section you will create a data protection rule that will partially implement the policy *Protect personal identifiable information (PII)* by masking email addresses in data sets.

Perform the following steps to create the data protection rule:

1. In the main menu, navigate to ***Governance -> Rules***.
2. Click on ***New rule*** on the right and select ***Create new rule***.
3. Click on the ***Data protection rule*** icon.
4. Under *Rule details*
   1. Enter ***Name*** “Mask email addresses - <your id>”.
   2. Select ***Type*** “Access” which is the default.
   3. Enter ***Business definition*** “Replace e-mail addresses by X characters”.
5. In the *Criteria* section of the *Rule builder*
   1. Select ***Criteria*** “Data class” from the if condition selection list.
   2. In the *Search for data class* field, type “Email” and click on the “Email address” data class once it appears
6. In the *Action* section of the *Rule builder*
   1. Select ***mask data*** and **in *columns containing***.
   2. In the *Search for data class* field, type “Email” and click on the ***Email address*** data class once it appears.
   3. In section *Select how to mask data*, keep ***Redact*** selected which is the default.
7. The final *Rule details* should look like this:  
   A screenshot of a cell phone

   Description automatically generated
8. Click ***Create rule*** to create and enable the rule. The rule will be immediately enabled. Note that data protection rules are processed in the order of their creation. The first data protection rule whose conditions result in a deny action blocks access to the asset and stops further data protection rule processing.

***Important:*** *The data protection rule relies on the proper assignment of data class* Email address*. If the data class is not assigned to a column, its data will not be masked even if it contains sensitive email addresses.*

1. For documentation purposes, add the data protection rule to the *Protect personal identifiable information (PII)* policy.
   1. In the main menu, navigate to ***Governance -> Policies***.
   2. Click on the *Protect personal identifiable information (PII)* policy to open its details page.
   3. In the *Data protection rules* section click on the ***Add data protection rules*** link on the right.
   4. Select the newly created data protection rule and confirm by clicking ***Add***.
2. Note that the data protection rule will remain active even if the policy is deleted.

# Working with governed data assets in projects and catalogs

In this chapter you will create a project in CPD, add data assets to it, and analyze the data using the implicit profiling capability. The data is then being shared with a data catalog and it is demonstrated how a data protection rule can automatically hide sensitive information from a user.

### Importing project data from a zip file

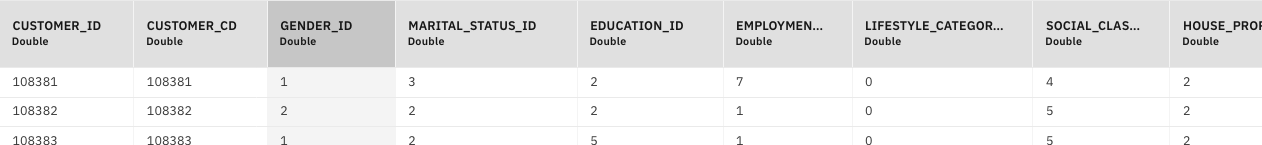
Perform the following steps to import the project zip file delivered with the industry accelerator downloaded above:

1. In the main menu, navigate to ***Projects -> All projects****.*
2. Click on ***New project*** in the upper right corner.
3. Select option ***Analytics project***and click ***Next***.
4. Click ***Create a project from a file***.
5. Browse for the *utilities-payment-risk-prediction-analytics-project.zip* file downloaded in section 2.4 and select it.
6. In the *Name* field, enter “*Utilities Payment Risk Prediction – <your id>”* where your id stands for your mail id, e.g. *robkern*.
7. Click ***Create***.
8. Wait for the import to be completed and click ***View new project***.
9. Navigate to the ***Assets***tab of the project. You will see the imported CSV files and notebooks.

### Creating data profiles in projects

1. In the main menu, navigate to ***Projects -> All projects***.
2. Open the “*Utilities Payment Risk Prediction – <your id>”* project you created above.
3. In the *Data assets* section, click on ***View all***.
4. Click on ***Show more*** at the bottom of the screen to see all data assets.
5. Navigate to the CUSTOMER.csv file.

***Hint:*** *A simpler way of navigating to a specific asset is to type part of its name into the search field at the top of the screen which will show a subset of matching assets.*

1. Navigate to the ***Preview*** tab which will display a preview of the data. The technical data types are displayed underneath the column names.  
   

***Important:*** *Initially, there are no data classes being displayed for the columns. After completing data profiling, an additional data class row will appear underneath the header row.*

1. Navigate to the ***Profile*** tab of the CUSTOMER.csv file.
2. You will see two buttons, ***Select data classes*** and ***Create profile***.

***Hint:*** *By default, all available data classes will be considered when running automatic data class matching. After clicking on the “Select data classes” button you can reduce the matching process to the relevant data classes which may considerably improve the profiling performance.*

1. Click ***Create profile*** to start the profiling process which may take **several minutes**. The process will run in the background so you can navigate away from the screen and check the progress later. Otherwise, click on the ***Refresh*** button from time to time in order to refresh the screen.

***Important:*** *The profiling tab uses a different architecture from the data discovery process. Even though the same Java libraries are used in the backend, the results may be slightly different because different parameters are being used for the analysis.*

1. After the profiling process has finished, the data profile will be displayed which shows the automatically identified data classes, the frequency distribution and some statistics about the data:  
   Graphical user interface, application

   Description automatically generated
2. Navigate back to the ***Preview*** tab. As mentioned above, there is an additional row now underneath the column headers that indicates the data classes of the individual rows:  
   Table

   Description automatically generated

***Note:*** *Within analytics projects, no business terms are being displayed for data assets or columns. After sharing the data asset with a data catalog, an additional eye-icon will appear next to each column name for managing business terms.*

### Creating a governance catalog

1. In the main menu, navigate to ***Catalogs -> All catalogs****.*
2. Click on ***Create Catalog*** in the upper right corner.
3. In the ***Name*** field, enter “<your id>s catalog”, e.g. “robkerns catalog”.
4. Select checkbox ***Enforce data protection rules*** and confirm the warning that will pop up.

***Important:*** *If you do not enable this check box during catalog creation, data protection rules won’t be enforced within the catalog, i.e. sensitive information might be disclosed to the users even if data protection rules apply.*

1. Click ***Create***.
2. Add *admin* user to the catalog as a collaborator (needed for validating the data protection rule later).
   1. Navigate to the ***Access Control*** tab
   2. Click on the ***Add collaborator*** link on the right.
   3. Select the ***Role*** toggle *Admin*.
   4. In the ***Collaborator*** field, type “admin” to find the *admin* user,
   5. Select *admin* user and click ***Add***.
3. The *admin* user is now an additional Admin of the catalog.

### Adding a project data set to a catalog

Perform the following steps to add a project data set to the newly created governance catalog:

1. In the main menu, navigate to ***Projects – All projects***.
2. Open the “*Utilities Payment Risk Prediction – <your id>”* project you created above.
3. Navigate to the ***Assets*** tab.
4. Type “customer” into the search field underneath the project tabs and press enter.
5. Locate the *customers.csv* file in the *Data assets* section of the search result, click on the ***three-dot-menu*** icon () on the right of the row and select ***Publish to catalog***.
6. On the *Publish to catalog* screen, select your catalog (“<your id>s catalog”) from the *Target* selection list and click ***Publish*** in the lower right of the screen.
7. In the main menu, navigate to ***Catalogs -> All catalogs*** and click on your catalog (“<your id>s catalog”) in order to open it.
8. Navigate to or search for the newly added *customer.csv* file and click on it.
9. Compared to the asset browser within your analytics project, there is also an additional *Access* tab for managing access to the data set on the asset level:  
   
10. In the ***Overview*** tab, you will a *Business terms* section on the left where you may assign business terms to the data set itself. In the ***Asset*** tab, there are eye-icons ( ) near each of the column names for assigning business terms to the individual columns:  
    Table

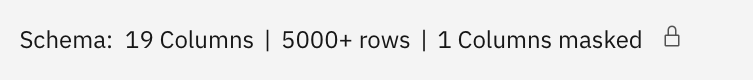
    Description automatically generated
11. Scroll to the right to see the EMAIL column. It turns out that the data is not masked despite of the data protection rule created in section 3.3.4:  
    Graphical user interface, table

    Description automatically generated

***Important:*** *If the user who is trying to access the asset is the owner of the asset (by default, the user who created the asset), then access is always granted. In Data Virtualization, the asset owner is not exempt from enforcement of data protection rules.*

### Validate effect of data protection rule

In order to validate the data protection rule, we need to logon using a different catalog user who is not the owner of the data assets.

1. Logoff from CPD by clicking on the user icon at in the upper right corner of the screen and selecting ***Log out*** and confirming with ***Log out*** again.
2. Sign in again using the *admin* user.
3. In the main menu, navigate to ***Catalogs-> All catalogs****.*
4. Open the “*<your id>s catalog”* created above.
5. Type “customer” into the search field underneath the project tabs and press enter.
6. Click on the *customers.csv* file in the search result.
7. Navigate to the ***Asset*** tab.
8. It the header, there should be an indication that at least one column is masked (in this case, two columns are masked because there is an additional data protection rule in place):  
   
9. The data masking might still be in progress and might take several minutes to complete. If you see an error, try to refresh the screen.
10. Once masking has finished, the data in the EMAIL column should be masked by XXXX characters and there should be a lock () indicating that the data is masked:  
    Table

    Description automatically generated
11. **Logoff** from the admin user again and logon with your personal user id.

# Data curation & data quality

This chapter shows how data assets can be discovered from a connection using *Automated discovery* or *QuickScan*. The results are then being further investigated in a data quality project before they are published to the metadata catalog. Finally, the use of quality rules and automation rules is being demonstrated.

## Creating a data connection

For this workshop, the sample data of the *Utilities Payment Risk Prediction* industry accelerator has been uploaded to DB2 in order to use it for data discovery. Use the following steps in order to setup a data connection to the DB2 database:

1. In the main menu, navigate to ***Platform connections***and click ***New connection*** in the upper right corner.
2. In section ***IBM***, select ***DB2***.
3. Specify the following parameters:
   * 1. ***Name***: “WKC workshop <your id>”
     2. ***Database***: CPD
     3. ***Hostname or IP address***: dataquack.ddns.net
     4. ***Port***: 50001
     5. ***Username***: db2inst1
     6. ***Password***: Rob123!
4. Clicking the ***Test*** button should show a message “*The test connection was successful”*.
5. Click ***Create***.

## Discovering data: Automated Discovery

Automated discovery provides detailed analysis results of all assets from data sources. Unlike quick scan, with automated discovery the metadata and analysis results are automatically imported into the default catalog. Also, a broader set of analysis results is available for viewing in a project, including data quality score, automatically assigned data classes and business terms, data types, formats, frequency distributions, and more.

### Starting the Automated Discovery process

Perform the following steps to run *Automated Discovery*:

1. In the main menu, navigate to ***Governance -> Data discovery****.*
2. Under ***New discovery job*** *select* ***Automated discovery***.
3. Under ***Connection***, select ***Find or add connection***.
4. On the ***Add existing connection*** panel, select your newly created connection, and click ***Add***.  
   If you have used the connection for discovery before, it will appear in the dropdown list underneath the ***Find or add connection***option*.*

***Important:*** *This will create a copy of the global connection in IMAM. Because of an open issue, any updates performed to the global connection afterwards will not be applied to the copy. With cpd-3.5.1-wkc-patch-2, at least passwords changes will be applied.*

1. For *Discovery root*, click on the ***Browse*** button, select schema DB2INST1 from the list and click ***Select***. This will populate the Discovery root parameter with “schema[CPD|DB2INST1]”.

***Hint:***You may run Automated Discovery on individual data sets as well which is not possible with QuickScan.

1. In the *Project* field, select ***Add a project***, enter ***Name*** “Automated Discovery project - <your id>”, e.g. “Automated discovery project - robkern” and click ***Create***.
2. In the *Discovery options* section, select check boxes ***Analyze columns***, ***Analyze data quality****.* Option ***Assign terms*** should be automatically checked.
3. Make sure that the checkbox ***Use data sampling*** is selected and ***Set the maximum number of records that you want to include in your data set sample*** is set to 1000.
4. Finally, click on the ***Discover*** button at the top right.
5. The UI indicates the status of the discovery process. You may click on the refresh button () to update the view. The *Import status* refers to the import of the technical metadata, while the *Analyze status* refers to the actual analysis of the data. Use the refresh icon () in order to refresh the screen. Once the *Analyze status* is *Running,* the schema name (DB2INST1) on the left will turn in to a link. When you click on it, you will see the analysis status for all tables of the schema. Use the browser back button to get back to the overall status.  
   Graphical user interface

   Description automatically generated with low confidence
6. The analysis can be cancelled either for an individual schema or for all schemas by clicking on the corresponding cancel () button.

***Note:*** *While the analysis step can be cancelled this is not possible for the import of the technical metadata. When you cancel the analysis for all schemas, their technical metadata will still be imported.*

1. The discovery process may run several minutes. There is no need to wait for analysis to finish. The following section describes how you may browse the results at any time later.

### Quick review of Automated Discovery results

Perform the following steps to review the discovery results:

1. Previous Automated Discovery jobs can be found by navigating to ***Governance -> Data discovery***and clicking on ***View automated discovery results*** at the top of the screen.
2. Click on the ID of the **first job** in the *Discovery results* table to see the results of the latest job.
3. Click on the schema name on the left to see the discovery results for all tables of a schema.  
   Graphical user interface, text, application, email

   Description automatically generated
4. Review and adjust the term assignments using the ***Manage term assignments*** icon (). Click on the angle bracket () in front of the table to see the column details.
5. Using the *Publish analysis results to catalog* icon () you may publish the results of individual tables. In order to publish the results of multiple tables at once you may select the corresponding tables and click the *Publish* button at the top of the table. Do not publish any results yet.  
   Graphical user interface, application

   Description automatically generated

### Detailed review of Automated Discovery results

The discovery results can be reviewed in a more comprehensive way in the data quality project:

1. In the main menu, navigate to ***Governance -> Data quality****.*
2. Open the “Automated Discovery project - <your id>” project you have used for storing the Automated discovery results.  
   Graphical user interface, application, Teams

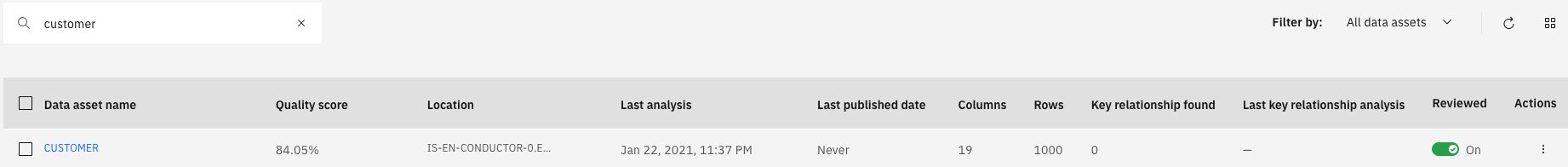
   Description automatically generated

***Hint:***There are two menu options for navigating to data quality projects: ***Governance -> Data quality*** and ***Projects -> All projects***. The latter also provides access to other types including *Analytics projects* and *Data transformation projects*.

.

1. The ***Dashboard*** tab of the project shows a summary of the quality characteristics:
   1. Data quality threshold: 100%
   2. Data quality score distribution: All data sets above 80%
   3. Analysis: Data analysis done for all 12 data sets
   4. Rules run status: no rules created yet
   5. Relationships: Candidates for primary key relationships: 12
   6. Top 5 data classes: Code, Boolean, Identifier, Quantity, Indicator
2. The *Data rules* and *Relationships* tabs address optional analysis capabilities that will be partially covered later in the workshop.
3. Click on the ***Data assets*** tab to see basic information about the datasets: Quality score, number of columns and rows per data set, etc.
4. Search for “CUSTOMER” in the ***Find a data asset*** field and click on the ***CUSTOMER*** data set in the search result, in order to see the detailed column analysis, data class matching, and business term assignment results for this data set:  
   Graphical user interface, application

   Description automatically generated
5. Navigate to the individual tabs to see details on business terms (***Governance*** tab), ***Data quality***, and detected ***Data classes***, ***Data types***, quality and data rule violations (***Rules***), as well as primary and foreign keys (***Keys***).
6. Click on one of the ***LAST\_NAME*** column to see details on the column level, again organized by tabs: *Column properties*, *Data quality*, *Rules*, *Data classes*, business terms (*Governance* tab), *Data types*, *Formats*, and *Frequency distribution*.  
   Graphical user interface, application

   Description automatically generated
7. Switch to the edit mode by clicking on the ***Edit*** button () in the upper right corner of the screen in order to modify the suggested results in the various tabs according to your needs (feel free to play with the available options).
8. When done, click on the ***Done*** button () to switch back to the *Preview mode*.
9. When you are ready to publish the results to the catalog, go back to the ***Data assets*** page. You may directly navigate to that page by clicking on ***Data assets*** in the object hierarchy displayed at the upper left of the screen: 
10. By default, you will see the card view i.e. each data asset is displayed as a large card on the screen. Click on the ***Show list*** icon () on the right to display the data assets as a list.
11. Use the ***Reviewed*** toggle on the right to indicate which data set’s results have been validated.  
    
12. Select the data sets to be published and click the ***Publish*** button () at the top.

***Important:*** *The business term assignments are published to the Default catalog as well as to the Information assets view. All other analysis results are* ***only*** *published to the Information assets view. The profiling tab of the Default catalog is* ***not*** *being updated except for the business terms. (See details in chapter 6)*

## Discovering data: Quick scan

*Quick scan* is best suited to get a fast initial analysis of large numbers of tables and files from data sources that you might not be familiar with. Quick scan analyzes a sample of each table or file to quickly provide analysis results, including data quality score, automatically assigned data classes and business terms. The source assets are not imported into any catalog during the discovery. However, you can easily add selected data assets to one or multiple catalogs later.

### Starting the QuickScan process

Perform the following steps to run *Quick scan* ona database schema:

1. In the main menu, navigate to ***Governance -> Data discovery****.*
2. At the top of the screen, select ***New discovery job -> Quick scan*** in the upper right corner of the screen.
3. Under *Connection*, click on ***Select a connection*** and select your connection ***“WKC workshop <your id>”***. (If you have not used this connection for data discovery before you need to add it first by selecting *Find or add connection*. Refer to the corresponding information in section 5.2.1)
4. For *Discovery root*, click on the ***Browse*** button, expand the CPD database by clicking on the *Open* icon () in front of the database name
5. Select schema ***DB2INST1*** from the list and click ***Select***. This will populate the Discovery root parameter with “schema[CPD|DB2INST1]”.

***Note:*** *You cannot run QuickScan on individual data sets but only on schemas.*

1. In the *Project* field, select ***Add a project***, enter *Name* ***“Quick scan project - <your id>”***, e.g. “Quick scan project - robkern”, and confirm by clicking ***Create***.

***Important:*** *This project serves as a working area but the discovered assets are not visible in the selected project and cannot be accessed there. They be accessed from the quick scan results only. The aim of the project is to manage access to the Quick scan results and to define their lifecycle. Only collaborators in the project have access to the quick scan job and the results. When a project is deleted, all Quick scan jobs tied it are also deleted.*

1. In the *Discovery options* section, keep the defaults and click **Discover.**

***Note:*** *For the sake of performance, Quick scan always operates on a sample of the data. The sample size can be specified in field “The maximum number of records included in the data set sample“.*

### Publishing Quick scan results

Once you have started the Quick scan process, you will be forwarded to the *Quick scan results* page where you can see the current discovery job in the *Pending analysis* tab.

1. Use the ***Refresh*** button (A close up of a mans face

   Description automatically generated) at the top of the screen in order to refresh the screen.
2. Once the job disappears from the *Pending analysis* tab navigate to the ***Action required*** tab.
3. Click on the ***Job ID*** in order to view the results, e.g. .
4. You will see summary of the discovery results covering *Discovery insights* and *Data quality insights*. At the bottom of the screen, you see the suggested *business terms* and *data classes*.  
   Chart, bar chart

   Description automatically generated
5. Click on the ***Explore assets*** button at the in the upper right corner in order to see the discovery details.  
   Graphical user interface, application

   Description automatically generated
6. Before adding the results to the data quality project, you may review and update the assigned business terms. The other results may be updated later in the data quality project (including the business terms again). In order to review and update the business to be assigned to the columns, make sure that *Asset type* ***Column*** is selected on the left, then click on the pencil icon () right from the corresponding column in the list.
7. As an example, find the ***CUSTOMER\_TYPE*** column by searching for it in the ***Find column*** field on the right. Then click on the pencil icon () on the right.
8. There are two *Assigned terms* and two *Suggested terms* for this column. You may click on the *check mark* () at the right of a business term in order to move it from the suggested to the assigned terms section or the X-icon () in order to move it back. Then click ***Update***.  
   Graphical user interface, application

   Description automatically generated
9. In order to publish the results, select *Asset type* toggle***Table*** on the left, select all tables by checking the check box near the *Table name* column at the top of the result table, and click ***Publish assets***.
10. On the ***Publish assets*** panel, select a catalog to which the assets and the discovery results are to be published to. Then click ***Publish***.

***Hint:***You can publish assets to as many catalogs as you want. If the respective connection does not yet exist in the catalog, the connection is also published.

1. The table metadata will be processed one after the other and the status changes from Loading to Approved as you refresh the screen using the refresh button ().  
   Graphical user interface, application

   Description automatically generated
2. Navigate to the ***CUSTOMER*** table and click on the ***View details*** link. Similar as described for Automated Discovery in section 5.2.3 this will show the detailed discovery results for the table. However, for Quick scan, the results are read-only. If you want to change the assigned data classes or business terms later, you need to do this within the catalog to which you have published the data asset.

## Data quality related rules

There are three types of data quality related rules, *Quality rules*, *Data rules* and *Automation rules*. Quality rules and Automation rules will be discussed in this section.

### Creating rule definitions

Data rule definitions are the basis for both, *Quality rules* and *Data rules*. Data rule definitions are used to develop rule logic to analyze data. You can set up data rule definitions so that a variable, such as a word or term, is evaluated based on a specific condition or type of check. After you create data rule definitions, you can associate physical data sources to the definition.

1. In the main menu, navigate to ***Governance -> Data quality****.*
2. Open the ***“Automated Discovery project - <your id>”*** project you have used for storing the Automated discovery results.
3. Navigate to the project’s ***Data rules*** tab.
4. There is a wide range of ready to use data rule definitions delivered with CPD.

In the available *Data rule definition* section tree, navigate to ***Published rules -> 01 Personal identity -> Email address*** and locate the ***EmailAddrValidFormat*** rule definition by expanding the folders using the expand button ().

1. Click on the three-dot-menu () right to the rule name and select ***Manage in project***. This will enable you to use the rule when creating an *automation rule* in step 5.4.3.
2. You will see a confirmation and the rule will be copied into the ***All*** folder above the ***Published rules*** folder.   
   Graphical user interface, application

   Description automatically generated
3. Click on the three-dot-menu () right to the rule name and select ***View***.
4. Select the ***Run logic*** tab.
5. The rule is being displayed in two ways:
   1. A visual representation:  
      Graphical user interface, website

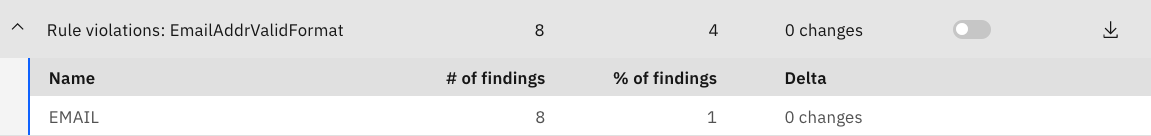
      Description automatically generated
   2. A textual representation:  
      IF email EXISTS AND len(trim(email)) <> 0 THEN email MATCHES\_REGEX '^[\_a-zA-Z0-9-!#$%&\'\*+/=?^`{|}~]+(\.[\_a-zA-Z0-9-!#$%&\'\*+/=?^`{|}~]+)\*@[a-zA-Z0-9-]+(\.[a-zA-Z0-9-]+)\*(\.[a-zA-Z]{2,6})$'
6. New rule definitions can be defined in either of the two ways. In this workshop, existing rule definitions are being used.

### Creating quality rules

Quality rules are run implicitly as part of the data quality analysis. The output of a quality rule is displayed as a rule violation in the data quality score analysis results.

1. In the main menu, navigate to ***Governance -> Data quality****.*
2. Open the ***“Automated Discovery project - <your id>”*** project.
3. Navigate to the ***Data assets*** tab and click on the **CUSTOMER** table card or link.
4. In the *Columns* tab, click on the ***EMAIL*** column.
5. On the ***Rules*** tab**,** click on the ***Create rule*** link in the in the upper right corner.
6. The Create data rule wizard comes up that leads you through the steps of the rule creation.
7. In the Rule definition step, select *Rule type* ***Quality rule***.
8. In the available *Data rule definition* section tree, navigate to ***Published rules -> 01 Personal identity -> Email address*** and select the ***EmailAddrValidFormat*** rule definition.
9. A panel will pop up asking whether you want to add the data rule definition to the project. Confirm by clicking the ***Add*** button.
10. In the ***All*** folder, select the newly added data rule definition *EmailAddrValidFormat* and click ***Next***.
11. On the left, select ***email*** variable, on the right, select the ***EMAIL***column of the ***CUSTOMER*** table. Then click the ***Bind***.  
    Graphical user interface, application

    Description automatically generated
12. Click ***Next***.
13. On the *Summary and test* tab, click on the ***Test*** button in order to validate the new Quality rule.
14. The *Test results* show that 99 percent of the records meet the rule. The detailed records that did meet and did not meet the text are displayed in separate tabs underneath.  
    A picture containing logo

    Description automatically generated
15. Click ***Save***.
16. The quality rule will be run next time data quality analysis is being performed on the data set. To re-analyze the dataset, click on the *Analyze* button at the top of the screen, select the ***Analyze data quality*** *check* box, and click ***Analyze***.
17. Once the analysis has finished, the results are available in the *Data quality* tab in section *Data quality dimension results*.  
    

### Creating automation rules

Automation rules help you automate tasks like binding rule definitions to assets, using quality dimensions, or setting the quality score. For example, instead of manually adding a quality rule to all columns containing Email addresses, an automation rule can be defined that does this automatically. The following example demonstrate how to create an automation rule.

1. In the main menu, navigate to ***Governance -> Automation rules***.
2. Click on the ***Create automation rule***button in the upper right corner.
3. In the ***Name*** field, enter *“Automatically bind email verification rule to columns containing emails – <your user-id>”*
4. In the Rule logic section, **drag and drop** the following components to the rule:
   1. ***AND*** operator to the ***if*** section of the rule.
   2. *“the asset is a column”* condition to the ***and*** operator.
   3. *“the asset is classified by the data class Email address”* also to the *and* operator (search for “Email address” by clicking on the magnifier icon).
   4. *“bind the data rule definition EmailAddrValidFormat”* action to the *then* section of the rule (search for “EmailAddrValidFormat” by clicking on the magnifier icon).
   5. The final rule should look like this:  
      Graphical user interface, application

      Description automatically generated
5. Leave the status on CANDIDATE and click ***Save***. Then confirm the warning by clicking ***Save*** again.
6. After changing its status is changed to ACCEPTED, the rule would automatically bind the *EmailAddrValidFormat* rule to all columns which are classified by data class Email address when running data quality analysis. In order to not affect other workshop users, we will not do that right now.

# Working with the Information Assets view and the Default Catalog

In this chapter, you will browse the metadata and discovery results you have published to the Information assets view and the Default catalog in chapter 5.

## Browsing data sets in the Information assets view

The Information assets view contains metadata of assets that have been discovered with *Automated Discovery* or imported with the *Metadata import* capability. Use the following steps to browse information about a data asset and its associate quality characteristics:

1. From the main menu, navigate to ***Catalogs -> Information assets***.
2. Type “CUSTOMER” in the search field and press *enter* or click the *Search* icon (). You may want to limit the search result to *Asset type* ***Database table*** on the left.
3. Locate the following search result and click on it. Note that the search result might have been proposed underneath the search bar even before pressing enter.  
   
4. The Database Table details screen will present summary information about the data set including corresponding *database schema*, *quality score*, assigned *business terms*, number of *database columns*, number of suggested *term assignments*.  
   Graphical user interface, application, Teams

   Description automatically generated

***Important:*** *This is the place where quality analysis results become available when “publishing” them after running Automated Discovery or quality analysis. The results will* ***not*** *be available in any other data catalog (except for the business term assignments which can be found in the Default catalog as well)*

1. Click on the ***Database Columns*** link on the left to open the *Database Columns* list.
2. Locate the ***EMAIL*** column from the list and click on it.
3. In the *Database Columns details* view, click on the ***General Information*** link. You will see that the EMAIL Address data class was assigned to the column:  
   Graphical user interface, application

   Description automatically generated
4. Click on the *three-dot-menu* icon () in the upper right corner and select ***Explore relationships***.
5. After expanding some of the nodes you will see how the EMAIL column relates to the *CUSTOMER* table and the *EMAIL Address* data class, and how the data class relates to the automation rule you created above:  
   Diagram

   Description automatically generated

## Browsing data sets in the Default catalog

Some asset metadata is automatically synchronized between the *Information assets* view and the *Default catalog*. This includes database tables and files as well as their related columns. A subset of the data connections is synchronized as well.

Use the following steps to browse the synchronized *CUSTOMER* table:

1. From the main menu, navigate to ***Catalogs -> All catalogs***.
2. Click on the ***Default catalog*** in order to open it.
3. Type *CUSTOMER* into the search field.
4. Click on the ***Any source*** selection list to limit the result set to source “WKC workshop <your id>”.
5. Click on the ***CUSTOMER*** table in the search result. The data asset overview page is being displayed. The *Tag* ***info\_asset*** on the left indicates that the asset was synchronized from the *Information assets* view.Graphical user interface, application, Teams

   Description automatically generated
6. Click on the eye icon () near in the header CUSTOMER\_ID column (first column) in the **Asset** section.
7. The panel shows that there are two business terms assigned to the column which have been successfully synchronized from the Information assets view. However, the assigned business terms cannot be edited here but only in the *Information assets* view.  
   Graphical user interface, text, application

   Description automatically generated
8. Switch to the ***Profile*** tab to see that the analysis results have not been synchronized from the *Information assets* view. You would have to profile the data once more in order to see statistics about the columns.

*This concludes the hands-on lab for WKC in CPD V3.5.*