Technical preview: Watson Knowledge Catalog in Data Virtualization

Contents

Overview

Data Virtualization integrates data sources across multiple types and locations and turns it into one logical data view. Data Virtualization uses policies, data protection rules, and business terms that are provided by Watson Knowledge Catalog to govern your virtual data.

Watson Knowledge Catalog is a secure enterprise data catalog management platform that is supported by a data governance framework. A catalog connects data and knowledge with the people who need to use it. The data governance framework ensures that data access is compliant with your business rules. Governance is the process of curating, enriching, and controlling your data. You govern your data with governance artifacts.

You can govern your virtual data by:

- Creating data protection rules. You create data protection rules to identify the data to control and to specify the method of control.
- Creating policies to include and enforce your data protection rules. You create policies to describe how to govern data in catalogs.
- Using business terms to standardize your virtual data. You use business terms to standardize definitions of business concepts so that your data is described in a uniform and easily understood way across your enterprise. Data Virtualization can automatically use term assignment on assets while virtualizing your data.

When you use policies to enforce data protection rules, you must consider the following items:

1. Only tables and views that were approved to be published to the governed catalog are impacted by policy enforcement.
2. All policies and data protection rules are enforced at the table and view level. This means that if you have a policy that denies access to a single column, access to the entire table is denied.
3. Policies work in addition to Data Virtualization authorizations. These policies can only further restrict access in addition to the authorizations. Thus, all authorizations that are defined in the service must be always respected and access can be further restricted by the policies only. For more information, see Managing access to virtual objects.

Related information

Data Virtualization documentation

Watson Knowledge Catalog documentation

Connecting Data Virtualization to Watson Knowledge Catalog

Watson Knowledge Catalog includes a connection type that is specific to Data Virtualization.

About this task

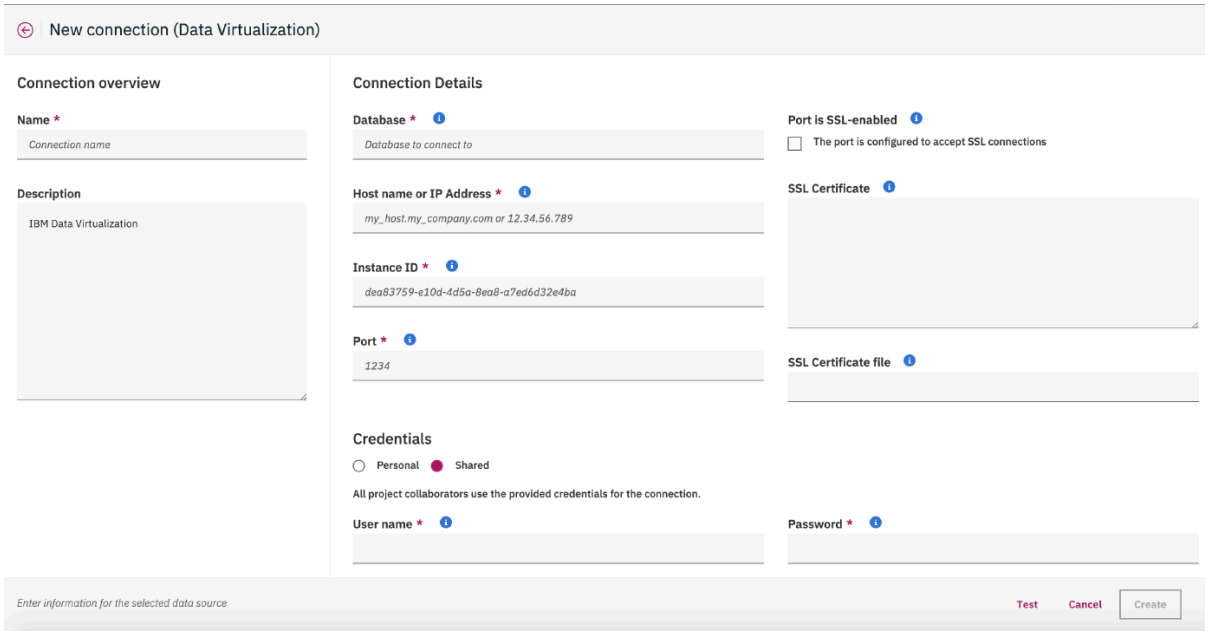
The Data Virtualization connection to Watson Knowledge Catalog is created automatically when a request to publish a Data Virtualization table or view is approved for the first time. This connection is listed in **Connections** page with the Data Virtualization connection name. It is a shared connection that uses the credentials of the Data Steward who approves the publish request. Alternatively, you can connect Data Virtualization to Watson Knowledge Catalog manually. To do so, you must use the Data Virtualization connection type.

Procedure

To manually connect Data Virtualization to Watson Knowledge Catalog:

1. Click **Add to catalog** > **Connection**.
2. Choose the Data Virtualization connection type.
3. Enter the connection information required for your service.

   Typically, you need to provide information like the host, port number, service instance ID, username, and password. You can find the service instance ID in the **Collect** > **Data Virtualization** > **Connection details** page.



4. Click **Create**.

   The connection appears on the **Assets** page. You can edit the connection by clicking the connection name on the **Assets** page. For more information see, Adding connections to projects.

Creating policies and data protection rules

Policies describe how to control data. A policy consists of one or more data protection rules. You can add data protection rules to create policies that specify types of data to restrict.

About this task

Policies are artifacts that you can create, view, edit, rename, publish, delete, or import. For more information, see Policies.

You create data protection rules to specify what data to control and how to control it. Data protection rules apply to all governed catalogs and all assets within these catalogs. Data protection rules are published and active after creation. They are not subject to workflow. You can add data protection rules to policies, however, data protection rules are enforced regardless of whether they are included in any published policies. For more information, see Data protection rules.

Procedure

- To create a policy:
  a) Go to **Organize** > **Data and AI governance** > **Policy**.
  b) Click **Create policy** .
  By creating a policy, you can group and organize multiple data protection rules. For more information, see Managing policies
- To create a data protection rule:
  a) Go to **Organize** > **Data and AI governance** > **Rules**.
  b) Click **Create rule** and select **Data protection rule**.
  For more information, see Managing data protection rules.

Enabling policy enforcement in Data Virtualization

You can determine whether Watson Knowledge Catalog policies are enabled in Data Virtualization.

Before you begin
To ensure that Watson Knowledge Catalog is installed, go to **Organize** > **All Catalogs**. If the **Default catalog** is available, Watson Knowledge Catalog is installed correctly.

About this task

Policies describe how to control data. A policy consists of one or more data protection rules. Strictly speaking, policies are not enforced. You enforce data protection rules that are part of a policy. For more information about catalog policies, see Policies.

Procedure

1. Log in to IBM® Cloud Pak for Data as an Administrator.
2. Check the status of the policy setting by using the following API endpoint:

   ```
   /icp4data-databases/dv/zen/dvapiserver/v1/dv/security/policy/status
   ```

   You can get the following responses:

   ```
   {
   "status: "disabled"
   }
   ```

   Or

   ```
   {
   "status: "enabled"
   }
   ```

   If you get this response, you can skip the rest of the steps in this procedure.
3. In your browser, add the following information to the URL of your Cloud Pak for Data cluster:

   ```
   /icp4data-databases/dv/zen/dvapiserver/v1/dv/security/policy/on
   ```

   You get the following response:

   ```
   {
   "status: "enabled"
   }
   ```

4. Check the status of the policy setting by using the following API endpoint:

   ```
   /icp4data-databases/dv/zen/dvapiserver/v1/dv/security/policy/on
   ```

   Wait a few minutes for the policy status to change.
5. To validate policy in your Data Virtualization service, follow these steps:
   a) Virtualize and publish some of your tables.
   b) Ensure that your virtualized tables are available in Watson Knowledge Catalog.
   c) Assign tags or business terms that match your governance rules to the asset or asset columns.
   d) Perform a SELECT operation or create views by using the SQL editor or a JDBC connection.
      Your policies and rules are enforced, and access is denied as defined by your rules and policies.

Results
The Data Virtualization service requires a restart after you enable policy enforcement. Wait until the restart completes to use the service.

SQL statements that involve data access are governed by data protection rules.

Managing business terms

You can map virtual table and column names to business terms in Watson Knowledge Catalog assets. When you create virtual tables from catalog data assets that have term assignments, these tables can use the assigned terms to rename table and column names. Therefore, you can create views of the asset that are better adapted to your organization.

Before you begin
Ensure that you have a catalog available to publish your virtual objects. See Setting up your default catalog for more information.

About this task

Business terms are governance artifacts that are available in all catalogs. You can assign business terms to table and column names to form a logical structure of your virtual data. A common language increases trust and confidence in the information of your organization.

When using strict mode, Data Virtualization can use terms that are assigned on data assets to name virtual tables and their columns with these business terms. Thus, new virtual tables that are shared in Watson Knowledge Catalog can be searched by business terms.

Procedure

- Import your business terms:
  a) Click **Organize** > **Data and AI governance** > **Business terms** > **Import**.
  b) Select CSV file that contains your business terms.
  c) Select merge option:
     – If you select **Replace all values**, the imported values in the CSV file will replace existing values in the catalog.
     – If you select **Replace with defined values**, the imported CSV values that are not empty will replace existing values in the catalog.
     – If you select **Replace empty values**, the imported values in the CSV file will replace only empty values in the catalog.
  d) Click **Import**.

- Create business terms:
  a) Click **Create business term**.
  b) Enter name and abbreviation for the new business term.
  c) Specify the primary category. To change the category, click **Change**, select a new category, and click **Add**.
  d) Enter description and click **Save as draft**.
  You can then decide to:
  – Save the business term as a draft on the **Draft** tab.
  – Send the business term for approval. See Workflows for governance artifacts
- Extract business terms from a document:
  a) Click **Extract**.
  b) Upload CSV file that contains the business terms.
  c) Click **Extract terms**.
- To discover data, click **Organize** > **Metadata curation** > **Data discovery** > **New discovery job**:
  - Choose **Quick scan** if you want to quickly get a general overview of the quality of your data based on the analysis of data sample. The import of source assets is skipped. Quick scan is suitable for large data sources.
    1. Select a connection. To add a connection, click **Add a connection** and follow instructions in Add data sources.
    2. Specify a discovery option. You can discover a full database, or individual schemas and database tables. For JDBC connection types, the database name to use might differ from the actual database name, depending on the JDBC driver.
    3. Specify the discovery options.
    4. Specify the workspace. You can add a workspace by clicking **Add a new workspace**. You must specify a name and description for the new workspace.
       When you approve assets, they are added to the selected workspace, along with the analysis results.
    5. Click **Discover**.
  - Choose **Automated discovery** if you want to see the details about the quality of your data based on an in-depth analysis of all assets. The source assets are imported. Automated discovery is suitable for smaller data sources, or selected components of larger data sources.
    1. Select a connection. To add a connection, click **Add a connection** and follow instructions in Add data sources.
    2. Specify a discovery option. You can discover a full database, or individual schemas and database tables. For JDBC connection types, the database name to use might differ from the actual database name, depending on the JDBC driver.
    3. Specify the discovery options.
    4. Specify the workspace. You can add a workspace by clicking **Add a new workspace**. You must specify a name and description for the new workspace.
       When you approve assets, they are added to the selected workspace, along with the analysis results.
    5. Click **Discover**.
- Find or view a published business term on the **Published** tab. Published business terms are active and ready for use. They are inactive if the specified effective dates are expired.
- Find, view, or work on a business term on the **Draft** tab. Draft business terms are inactive, preliminary versions of business terms that you can save as drafts or send for approval according to your workflow process. By default the business term is published if you send it for approval.

- View the **Related content** page for details on relationships defined by other types of artifacts. You can then select one of the listed artifacts to view its content or edit it, for example, to remove its relationship to this business term.

What to do next

- Create a virtualized table.
- You can publish your virtual table to the catalog. Once in the catalog, you can assign business terms to the table and its columns.
- If Data Virtualization is in strict mode, you cannot edit column names. However, you can exclude a column when you virtualize data.

Related information

Virtualizing data

Enabling the strict virtualization mode

Data Virtualization administrators and stewards can determine the virtualization mode of the service.

About this task

The virtualization mode determines whether to enforce use of business terms in virtual table and column names. Use the **Strict** virtualization mode if you want virtual objects to map to existing business terms only. In this virtualization mode, users can virtualize only assets in the catalog, and their associated columns, that have assigned business terms. In the Strict virtualization mode, you can use only data sources that you add in **Organize** > **Discover assets**. You must use the Data Virtualization API to enable the strict virtualization mode.

Procedure

To enable the strict virtualization mode:

1. Log in to IBM Cloud Pak for Data.

   You must have the Cloud Pak for Data data steward role and the Data Virtualization Admin or Data Virtualization Steward role.
2. Enter the following GET request on your browser:

   ```
   https://OpenShift_URL:port/icp4data-databases/dv/Project/dvapiserver/v1/dv/security/strict/on
   ```

   Replace the following values:

Required information	Description
OpenShift_URL:port	The URL and port number to use when logging in to your Red Hat OpenShift cluster. Ensure that you have the appropriate credentials to log in to the cluster by using `oc login`. Your cluster administrator should tell you whether your cluster is connected to the internet or is air-gapped.
Project	The project (namespace) where the IBM Cloud Pak for Data control plane is installed.

Disabling the strict virtualization mode

Disable the strict virtualization mode to enable users to discover data sources and to virtualize objects without automatically using business terms for table and column names.

Procedure

To disable the strict virtualization mode:

1. Log in to IBM Cloud Pak for Data.

   You must have the Cloud Pak for Data data steward role and the Data Virtualization Admin or Data Virtualization Steward role.
2. Enter the following GET request on your browser:

   ```
   https://OpenShift_URL:port/icp4data-databases/dv/Project/dvapiserver/v1/dv/security/strict/
   off
   ```

   Replace the following values:

Required information	Description
OpenShift_URL:port	The URL and port number to use when logging in to your Red Hat OpenShift cluster. Ensure that you have the appropriate credentials to log in to the cluster by using `oc login`. Your cluster administrator should tell you whether your cluster is connected to the internet or is air-gapped.
Project	The project (namespace) where the IBM Cloud Pak for Data control plane is installed.

Limitations and known issues in policy enforcement

The following limitations and known issues apply to the policy enforcement feature in Data Virtualization.

Do not use transformation or masking policies.
If you define one transformation policy, assets have blocked access. Thus, if a transformation policy applies to an asset, even to a single column, Data Virtualization denies access to the entire table or view.

Rules based on data classes
Creating a rule based on data class blocks access to all Data Virtualization assets in the catalog that are not profiled. Profiling needs to be run so that proper data classes are assigned to columns before the assets can be evaluated for the rules which are based on data classes.

Do not use duplicate assets for the same table
The policy service is unable to decide which of the duplicated assets to use for policy enforcement and does not aggregate the rules. If you do happen to have duplicates, the observed behavior has been that policies are evaluated against the oldest asset.

Policy enforcement for assets added by using the Watson Knowledge Catalog user interface
Policy enforcement is not supported for assets that are added by using the Watson Knowledge Catalog user interface. Only virtual assets that are approved to be published to the catalog are impacted by policy enforcement.

Cannot profile data assets
You cannot profile Data Virtualization data assets in Watson Knowledge Catalog. To solve this issue, select one of the following workarounds:

- Disable the **Restrict visibility of virtual objects list to authorized users** option in the Data Virtualization **Service details** page or
- Grant the Data Virtualization Steward role to the `ICP4D-DEV` service ID by running the following command:

  ```
  GRANT ROLE DV_STEWARD TO USER "ICP4D-DEV"
  ```

  You must have the Administrator role to run this command.