SIEMENS

Insights Hub

Integrated Data Lake

System Manual

05/2024

Legal information

Warning notice system

This manual contains notices you have to observe in order to ensure your personal safety, as well as to prevent damage to property. The notices referring to your personal safety are highlighted in the manual by a safety alert symbol, notices referring only to property damage have no safety alert symbol. These notices shown below are graded according to the degree of danger.

⚠ DANGER
indicates that death or severe personal injury **will** result if proper precautions are not taken.

⚠ WARNING
indicates that death or severe personal injury **may** result if proper precautions are not taken.

⚠ CAUTION
indicates that minor personal injury can result if proper precautions are not taken.

NOTICE
indicates that property damage can result if proper precautions are not taken.

If more than one degree of danger is present, the warning notice representing the highest degree of danger will be used. A notice warning of injury to persons with a safety alert symbol may also include a warning relating to property damage.

Qualified Personnel

The product/system described in this documentation may be operated only by **personnel qualified** for the specific task in accordance with the relevant documentation, in particular its warning notices and safety instructions. Qualified personnel are those who, based on their training and experience, are capable of identifying risks and avoiding potential hazards when working with these products/systems.

Proper use of Siemens products

Note the following:

⚠ WARNING
Siemens products may only be used for the applications described in the catalog and in the relevant technical documentation. If products and components from other manufacturers are used, these must be recommended or approved by Siemens. Proper transport, storage, installation, assembly, commissioning, operation and maintenance are required to ensure that the products operate safely and without any problems. The permissible ambient conditions must be complied with. The information in the relevant documentation must be observed.

Trademarks

All names identified by ® are registered trademarks of Siemens AG. The remaining trademarks in this publication may be trademarks whose use by third parties for their own purposes could violate the rights of the owner.

Disclaimer of Liability

We have reviewed the contents of this publication to ensure consistency with the hardware and software described. Since variance cannot be precluded entirely, we cannot guarantee full consistency. However, the information in this publication is reviewed regularly and any necessary corrections are included in subsequent editions.

Siemens AG
Digital Industries
Postfach 48 48
90026 NÜRNBERG
GERMANY

5/2024 Subject to change

Table of contents

Introduction to Integrated Data Lake 1

1.1 Introduction

An Integrated Data Lake is a repository that allows to store structured and unstructured data or objects in its native format as needed. It handles large data pools for which the schema and data requirements are not defined until the data is queried. This offers more agility and flexibility than traditional data management systems. The stored data can be used for further analysis as per the customer's requirement.

Integrated Data Lake provides the following features:

- Import time series data or objects.
- Generate signed URLs to upload, update or download data or objects.
- Delete data or objects.
- Add, update and delete values for data or objects.
- Receive notifications.
- Data access using cross account access.
- Subtenancy support.

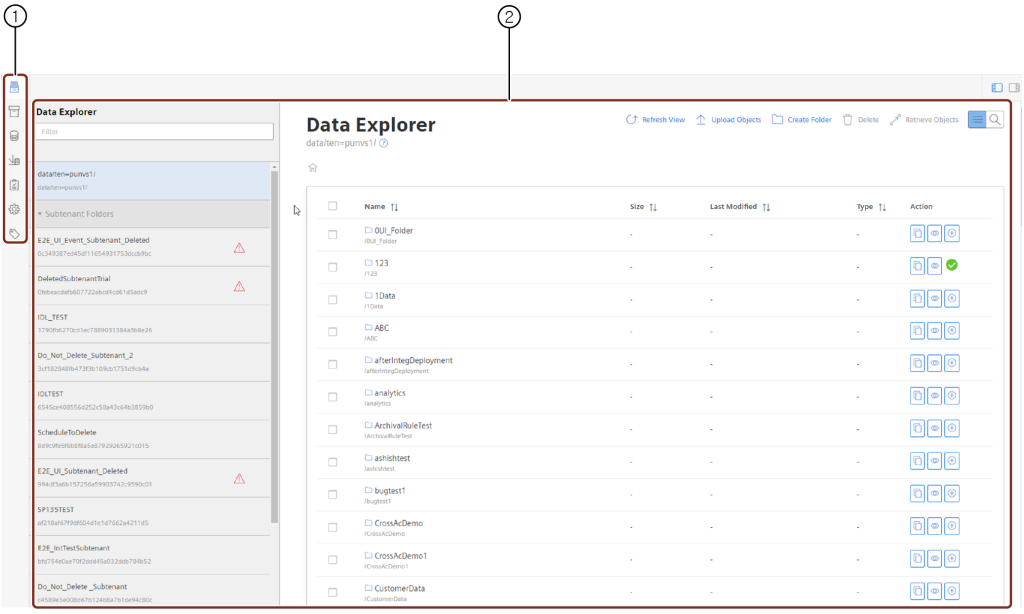
User interface Integrated Data Lake 2

2.1 User interface

Integrated Data Lake application can be accessed from the Insights Hub Launchpad by clicking the following icon:

Home screen

Integrated Data Lake displays the "Data Explorer" navigation tab. The following screenshot shows the different elements of the Integrated Data Lake user interface:

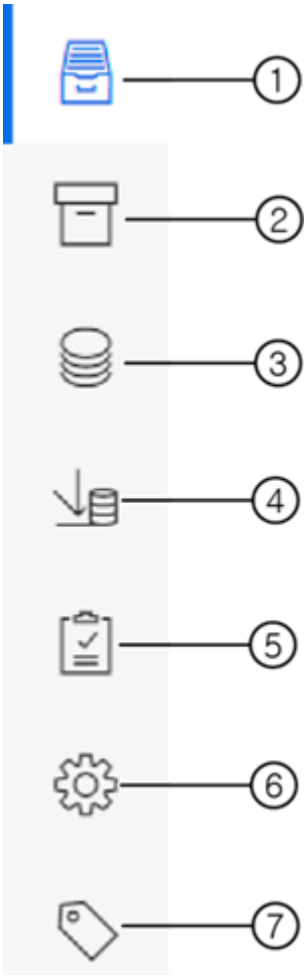


① Navigation tabs

② Work area

Navigation tabs

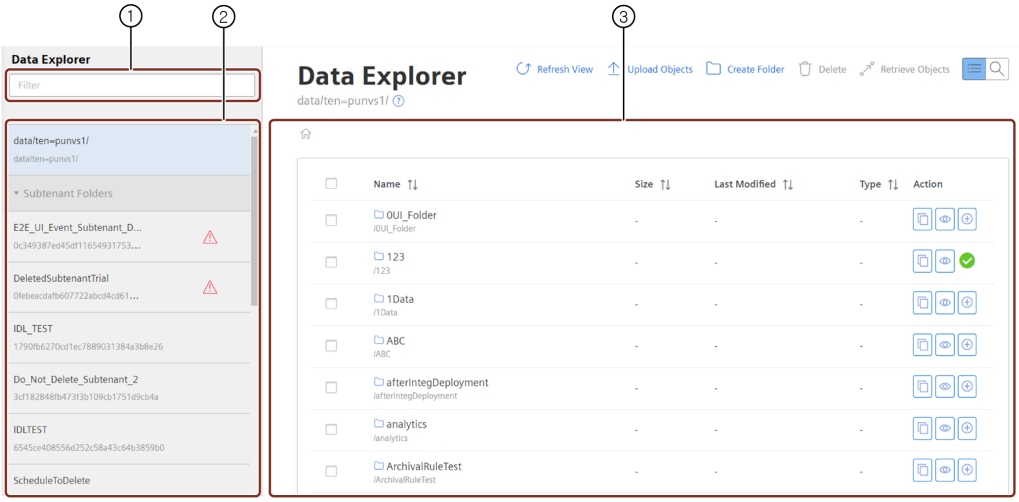
You can access the navigation tabs which are available in the left side of the work area.



① Data Explorer
② Archival and Retrieval
③ Data Catalog
④ Time Series Import
⑤ Event Subscription is to get notifications
⑥ Cross Account is applicable only for Europe 1
Service Principal is applicable for Virtual Private Cloud
⑦ Metadata Management is applicable only for Virtual Private Cloud

Data Explorer

The Data Explorer navigation tab is located on the left side of Integrated Data Lake UI.



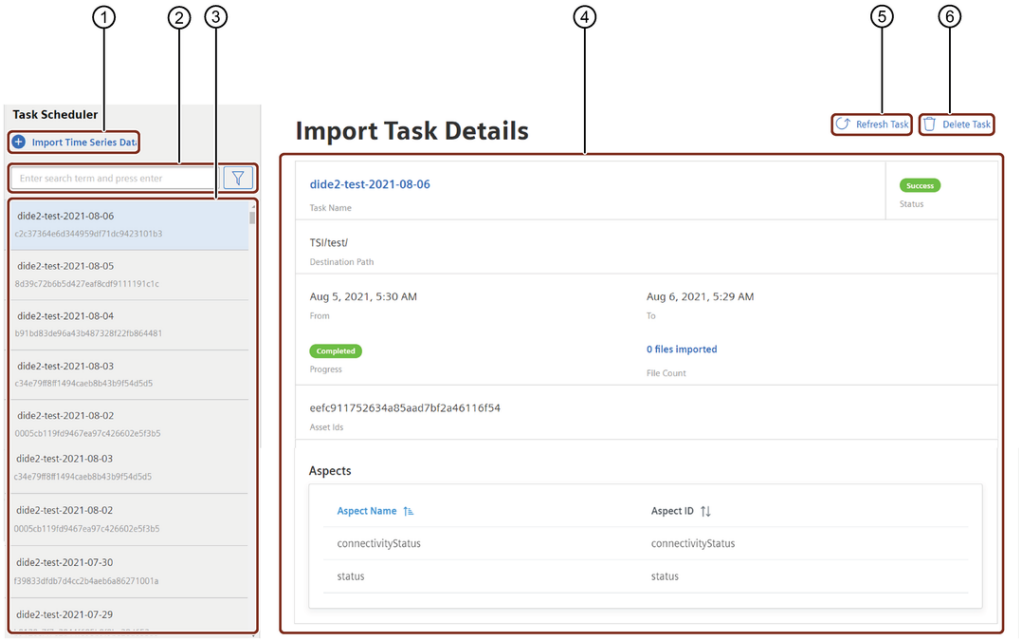
① Filter or search the folders

② List of the folders

③ Work area

Time Series Import

The Time Series Import navigation tab is located on the left side of Integrated Data Lake UI.



① The Time Series Import creates a job to import time series data to Integrated Data Lake

② Filter or search of the task created for import data

③ Task which are created for imported data

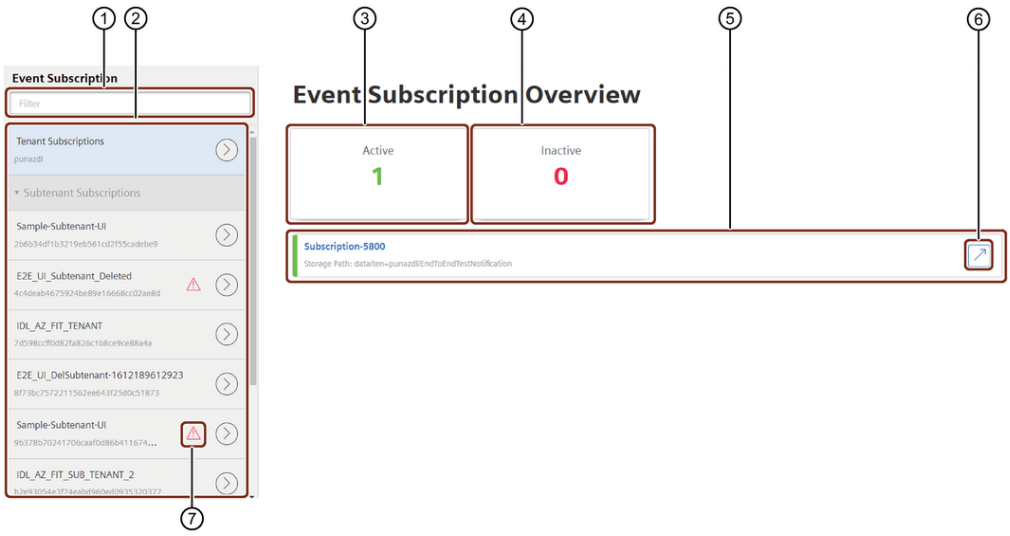
④ Details of the selected task

⑤ Updates the task detail

⑥ Deletes the selected task

Event Subscription

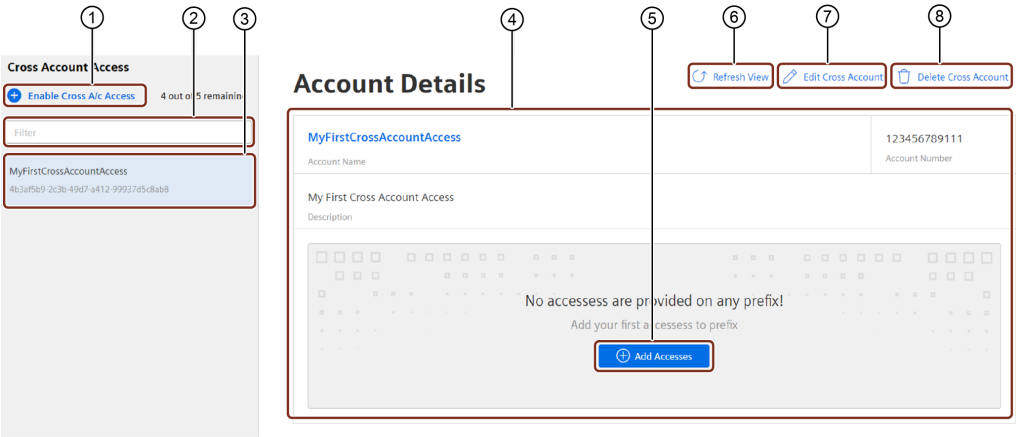
The Event Subscription navigation tab is located on the left side of Integrated Data Lake UI.



① Filter or search the event subscriptions
② Creates new Event Subscription
③ Active Event Subscription list
④ Inactive Event Subscription list
⑤ List of event subscriptions
⑥ Redirects to Data Explorer storage path
⑦ Deleted Subtenant

Cross Account

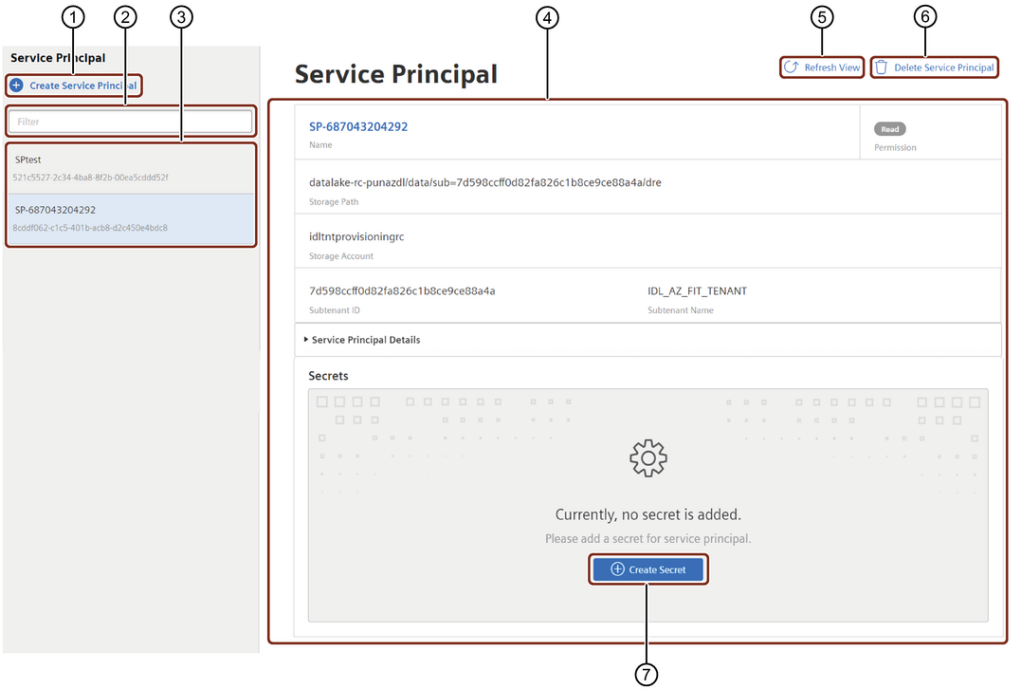
The Cross Account navigation tab is located on the left side of Integrated Data Lake UI. You can enable the Cross Account access only for the subscribers of Europe 1.



① Enable cross account access
② Filter or search the created cross account
③ Cross account drop-down list
④ Cross account details
⑤ Add Accesses
⑥ Updates the latest changes
⑦ Edit cross account
⑧ Delete cross account

Service Principal

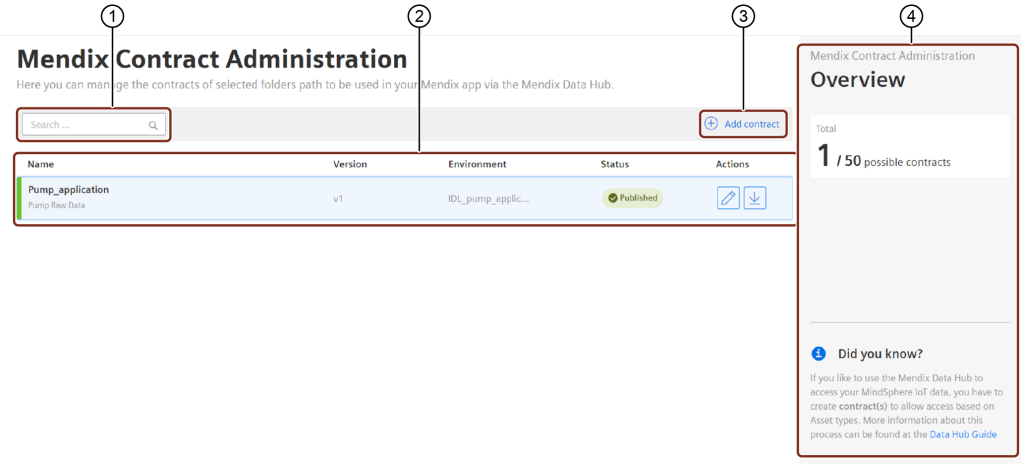
The Service Principal navigation tab is located on the left side of Integrated Data Lake UI. You can enable the Service Principal only for the subscribers of Virtual Private Cloud.



① Creates new Service Principal
② Filter or search the service principal
③ Service Principal drop-down list
④ Service Principal details
⑤ Updates service principal
⑥ Deletes service principal
⑦ Creates the secret for the service principal

oData Contract

The oData Contract navigation tab is located on the left side of Integrated Data Lake UI.



① Filter or search the required oData contract
② Displays the list of added oData Contracts

③ Creates oData Contract

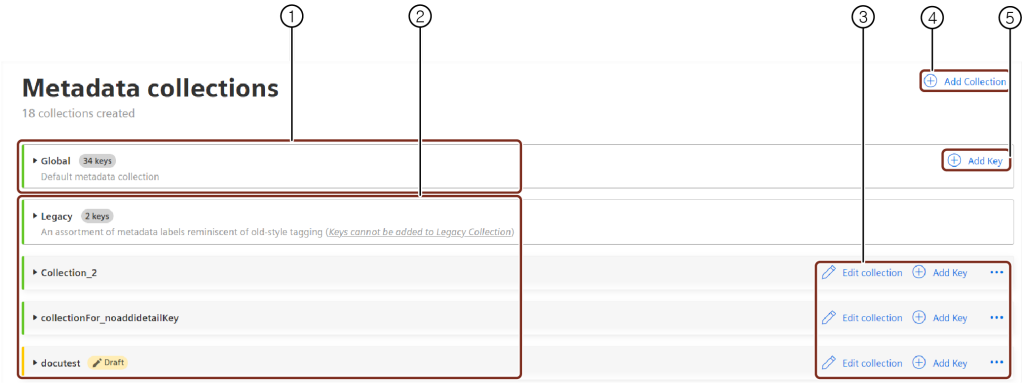
④ Displays the overview of selected oData Contract information

Region deviation

Feature "oData Contract" is supported only for Region Europe 1.

Metadata Management

The Metadata Management navigation tab is located on the left side of Integrated Data Lake UI.



① Global metadata collection.

② Custom metadata collections.

③ Actions to operate on custom metadata collection ie., edit collection, add key, delete collection.

④ Add a new custom metadata collection.

⑤ Add a new key in global collection.

User rights in Integrated Data Lake 3

3.1 User rights in Integrated Data Lake

The Integrated Data Lake application user rights are assigned from Settings application. After the activation of Integrated Data Lake, you will receive all data lake related read and write permissions from the tenant administrator.

User rights

The user rights depend on the following user roles:

- Administrator
- User
- Viewer

The following table shows the permissions:

Permission	Administrator Role	Standard user Role	Subtenant user Role
View all tenant's	✓	✓	✓
Time Series Import	✓	✓	✓
Cross Account	✓		
Event Subscription	✓		
Service Principal	✓		

Data Explorer

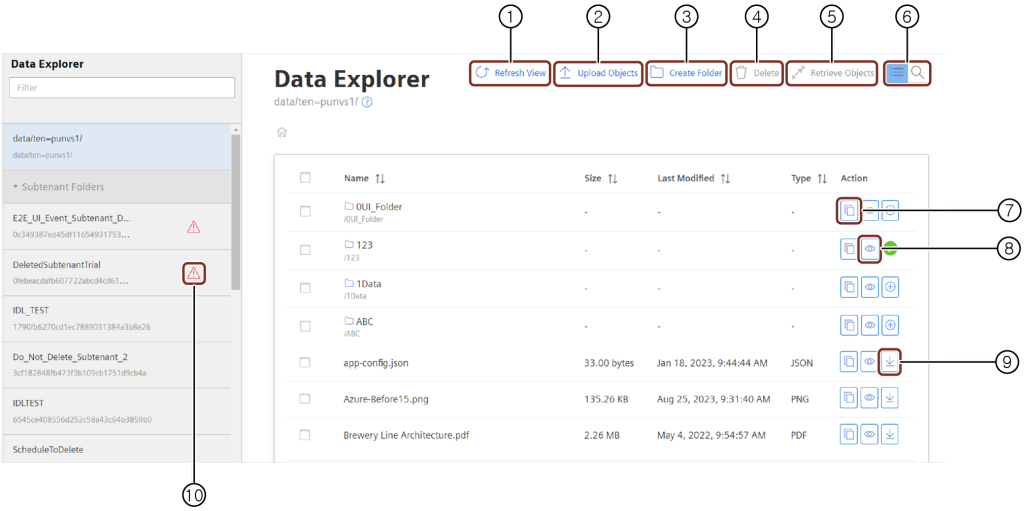
4

4.1 Exploring Data Explorer

Data Explorer feature enables you to explore the folders created by you and other subtenants. From the application, you can use the functionality of breadcrumb, upload, download, create the folder, sort and pagination.

You can view the data in the folders and objects available in Integrated Data Lake. The following screenshots display the folders and objects in Data Explorer screen:

Tenant admin view



① Refresh the page to get the latest data

② Upload the objects

③ Creates the new folder

④ Delete the objects

⑤ Retrieve the object

⑥ Search objects or files with object name, metadata or both

⑦ Copy the object path

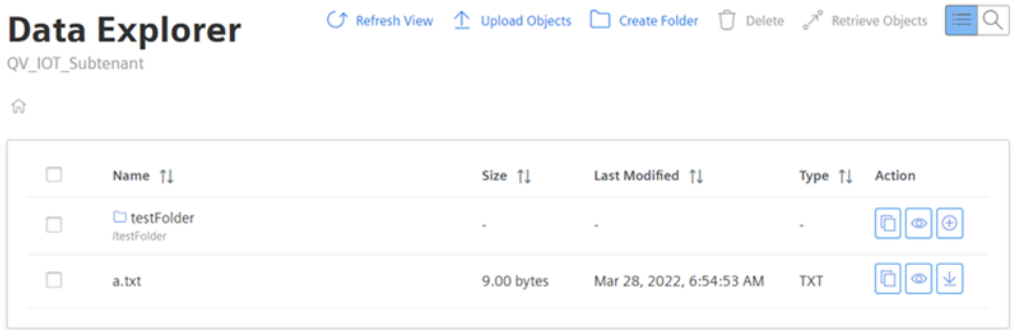
⑧ Displays Object View details and adds the metadata values to the folder and object

⑨ Download the object

⑩ Deleted subtenant symbol

- The subtenant users are permitted to access only those folders created by themselves, but not the folders created by other users.
- Tenant admin can see the folder or files created by subtenant user and access them from the Data Explorer folder pane.
- If the subtenant is deleted, uploading objects and adding metadata values are not allowed.

Subtenant admin view



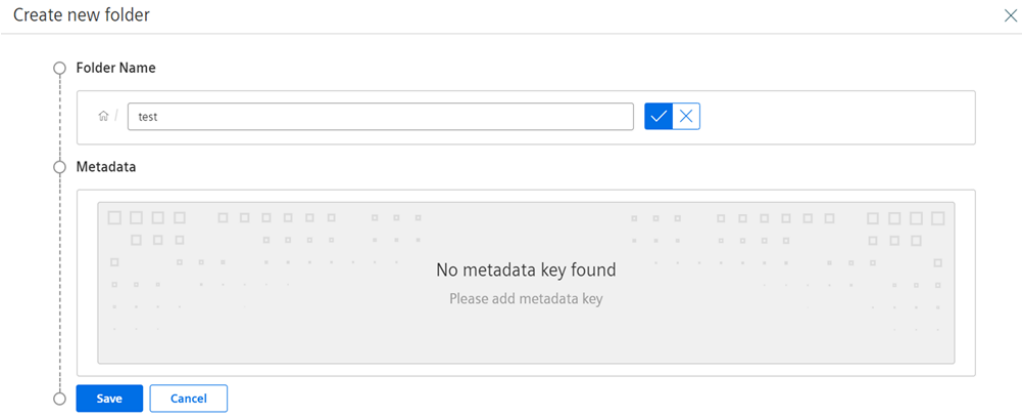
Data Explorer folder pane is not accessible for the subtenant user.

4.2 Creating a folder in Data Explorer

To create the folder with metadata values in Data Explorer, follow these steps:

1. In the left navigation, click "Data Explorer". The Data Explorer overview page is displayed.

2. In the Data Explorer overview page, click "Create Folder".



3. Enter a folder name and click to save the folder.

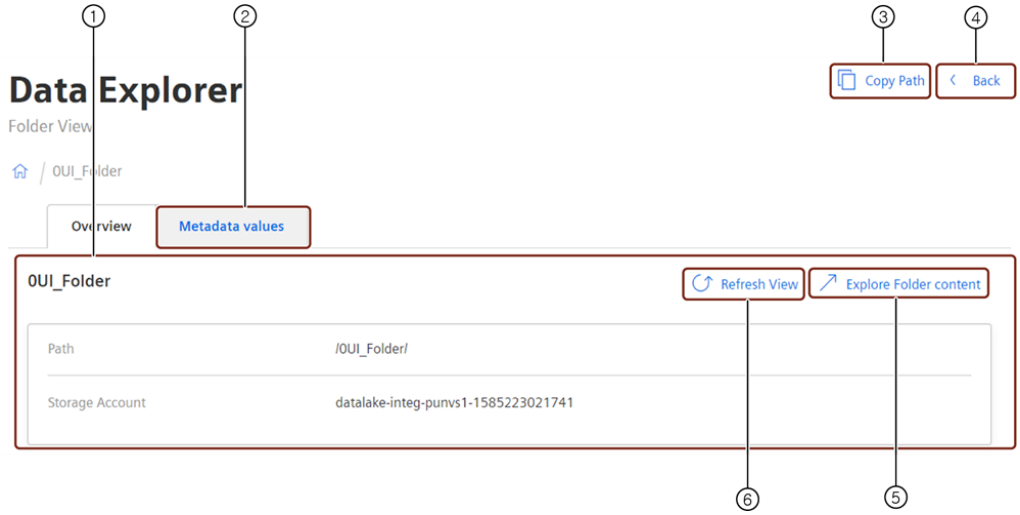
Enter the required details to add metadata values for the created folder.

4. Click toggle button to hide optional keys.

5. Click "Save".

- If the folder is added with metadata values, then deleting the folder is not possible.
- Folder name cannot contain '/'.
- Folder name can contain only alphanumeric with * ' & $ @ : + , < > ~ [ ] " # | ? { } ^ ` ( ) _ = - characters.

After creating a folder, you can view the folder details as shown below:



① Folder overview details
② Add the metadata values to the folder
③ Copy the folder path
④ Navigates to the parent folder or Data Explorer
⑤ Navigates to the folder
⑥ Refresh to update the folder details

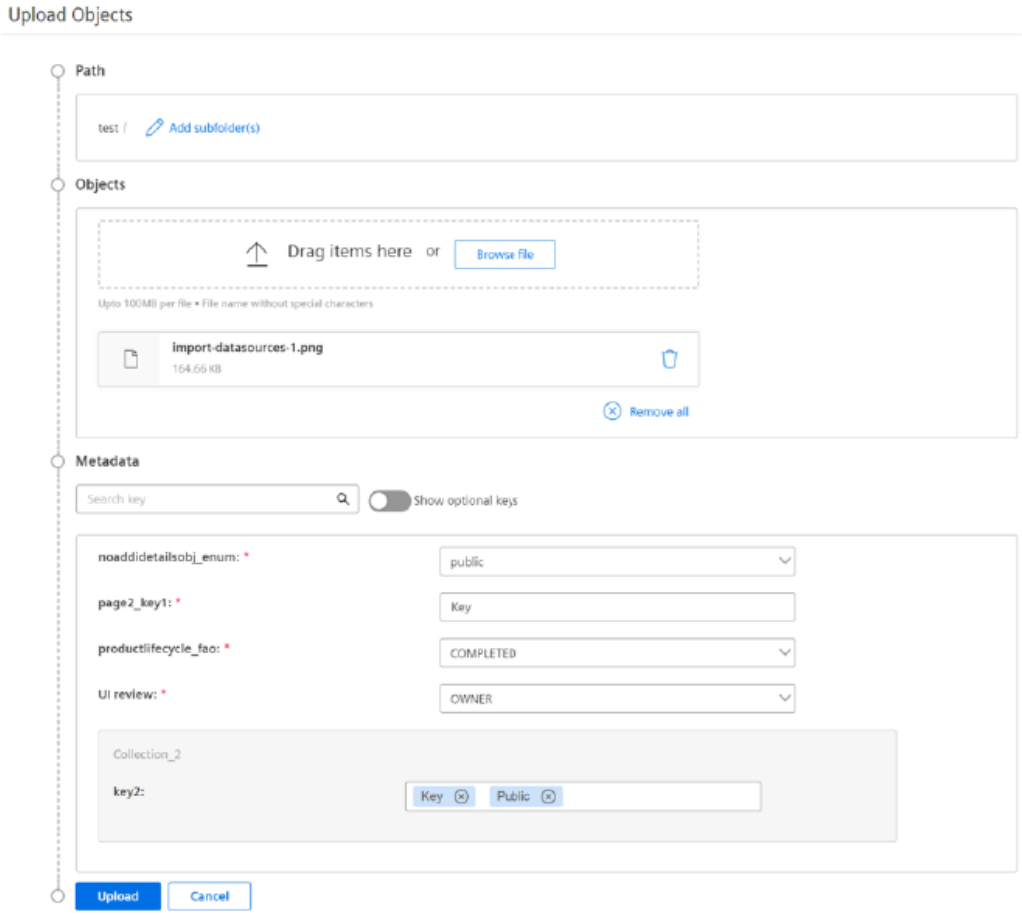
4.3 Uploading the files or objects to the folder or Data Explorer

You can upload the files/objects to the folder or "Data Explorer". It is possible to upload the files or objects with different formats.

Procedure

To upload the files or objects into Integrated Data Lake, follow these steps:

1. In the "Data Explorer" overview page, click "Upload Objects".



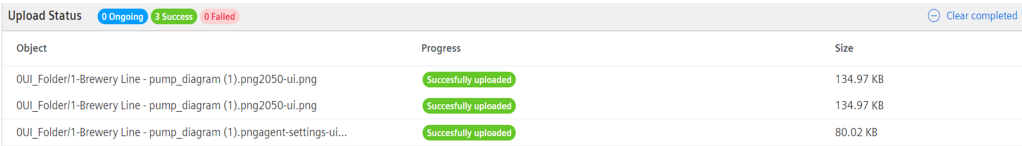
2. In the "Path" section,
   - Click Add subfolder(s).
   - Enter the subfolder name.
   - To save the subfolder name, click .
3. In the "Objects" section, select the files or objects to be uploaded.

   In the "Metadata" section, enter the details to add metadata values.
4. Click toggle button to show optional keys.
5. Click "Upload".

- A maximum of 30 files or objects with 5 GB per file can be uploaded.
- If duplicate files or objects are uploaded (with same name or different size with same name), the recently uploaded file or object will overwrite the existing file or object.
- If you refresh the page while ongoing upload operation, the browser will prompt the confirmation message.
- You will be notified with a warning message, when you try to access other pages or module during upload.
- The file size greater than 400mb is recommended to download through API.
- Click button to change the path.
- Click button to remove the file from the list.
- Uploading in "Time Series Import" folder will not be allowed in Integrated Data Lake.

Result

Uploading of files or objects is successful and the upload status can be viewed.



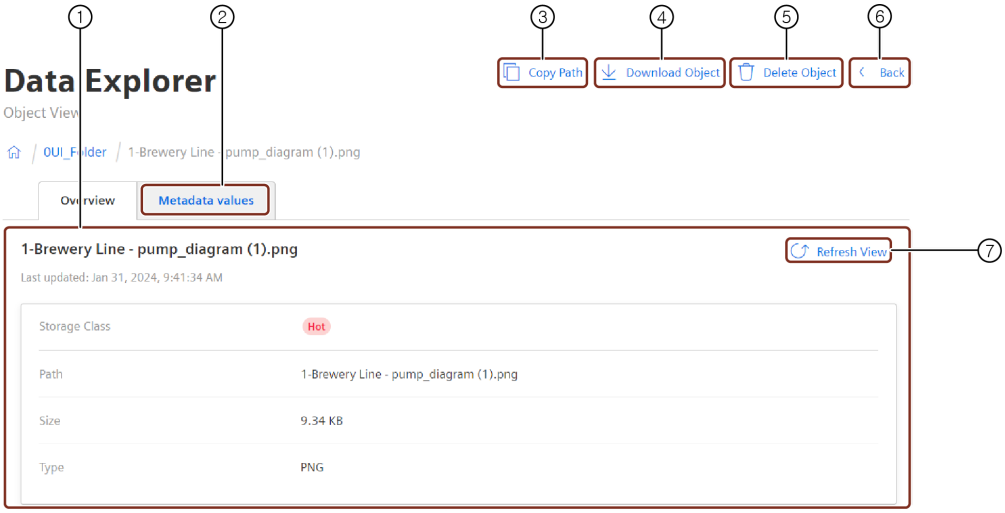
In Data Explorer, click "Refresh View" button to view the uploaded files or objects in the table.

4.4 Adding metadata values to the object

After creating a folder, upload the objects to the folder. Once the objects are uploaded, you can view the object details and add the metadata values.

View object details

The following image shows the object details window:



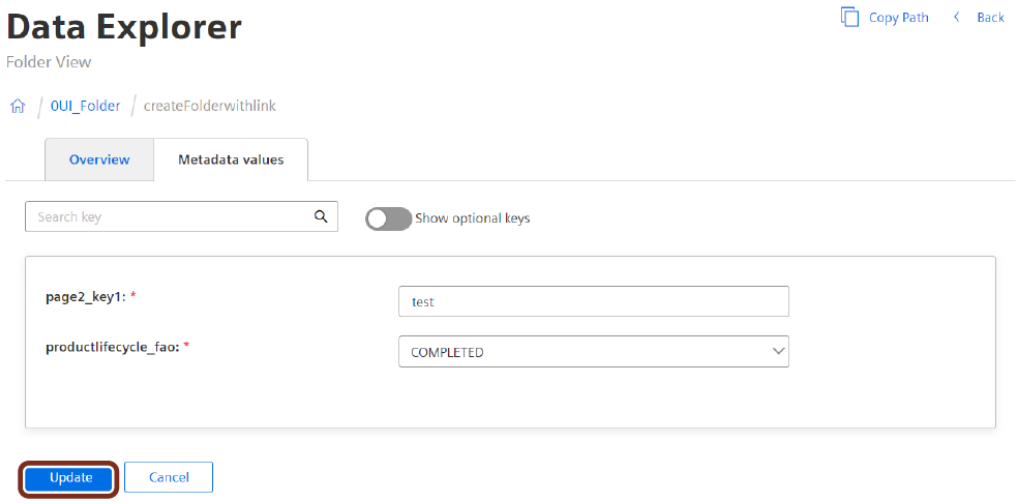
① Object overview details
② Add the metadata values to the object
③ Copy the object path
④ Download the object
⑤ Delete the object
⑥ Navigates to the parent folder or Data Explorer
⑦ Refresh to update the object details

Procedure

To add metadata values to the object, follow these steps:

1. In the Data Explorer overview page, select the folder and click to view the object details.
2. Click "Metadata values" tab.
3. Click "Add metadata".

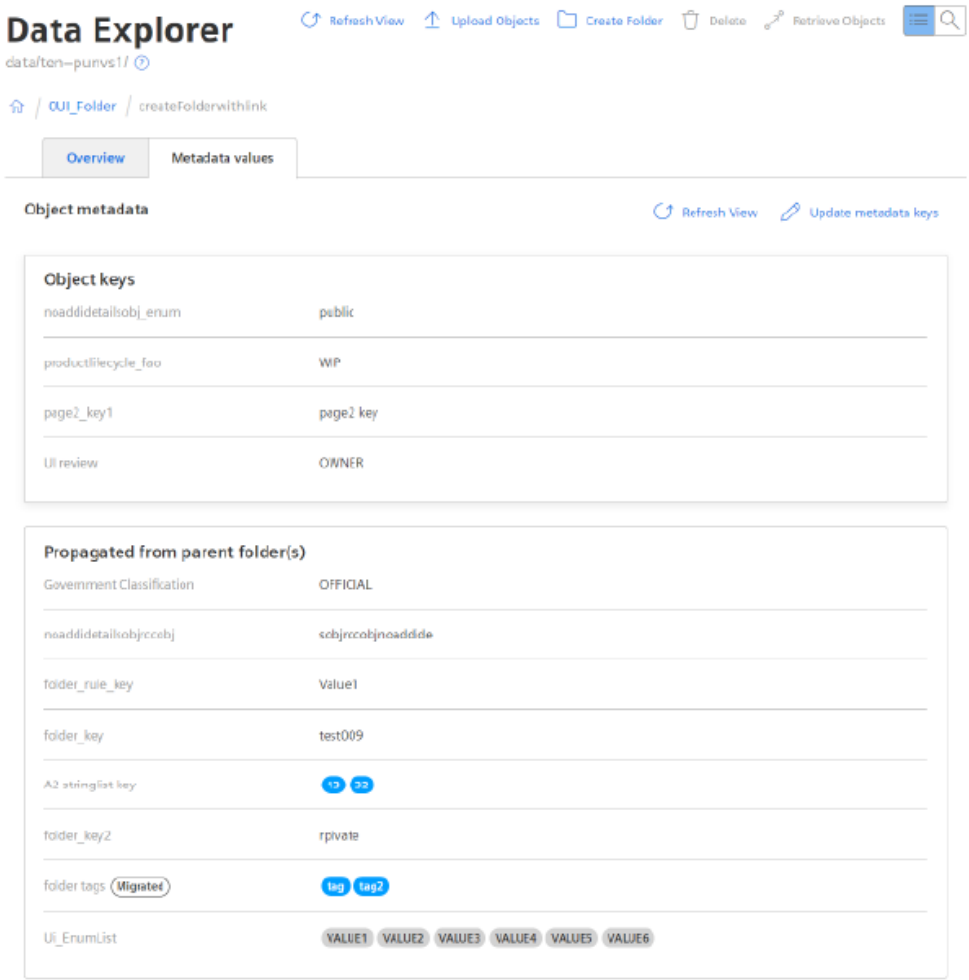
   Enter the required details.
4. Click toggle button to hide optional keys.
5. Click "Update".



- Maximum 8 multiple values can be added to the object.
- If metadata added to the file which contains special characters (including space), it will be considered in metadata search.
- If metadata values are added to folder then only first level files are tagged, second level or subfolder files will not be tagged with the metadata.

Result

The Metadata values are added successfully to the object.



4.5 Integrating Integrated Data Lake with Monitor

You can add Integrated Data Lake plugin to Monitor and tag the asset related files or objects using Data Explorer. The assetID of an asset in Monitor is used for tagging the files or objects

from Integrated Data Lake.
After tagging, these files or objects of an asset can be viewed in Monitor using Integrated Data Lake plugin extension.

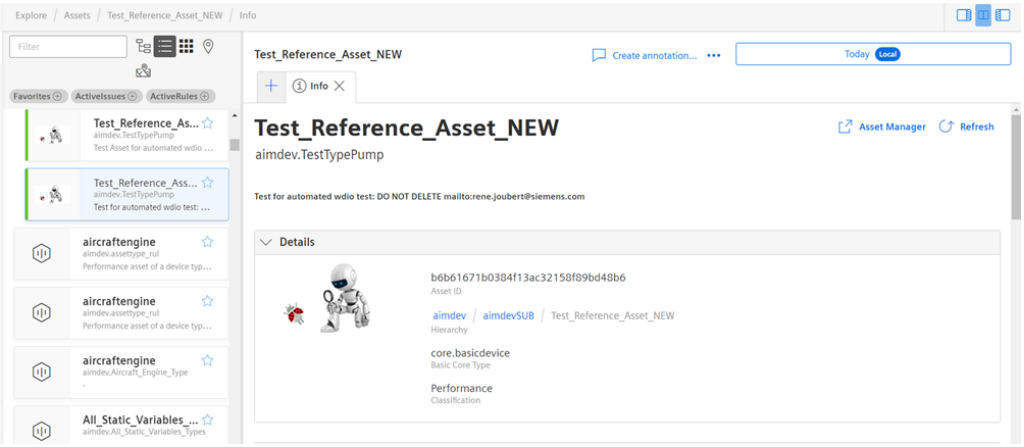
Procedure

To tag the files or objects from Integrated Data Lake to an asset in Monitor, follow these steps:

1. Open the Monitor from Insights Hub Launchpad.

2. Select an asset.

3. Click "Info" extension.



4. Copy Asset ID.

5. In Integrated Data Lake, open Data Explorer.

6. Click  of the file or object and click "Metadata Tags" tab.

7. Click "Add Tag", paste the Asset ID.

8. Click "Save".

For more information about how to add a plugin to Insights Hub Monitor, refer to [Add plugins to Insights Hub Monitor](){target="_blank"}.

4.6 Searching the files or objects in Data Explorer

The files or objects can be searched in Data Explorer by providing object name, metadata value or both. You can perform actions such as copying the path to the clipboard, view object details and download the object from the "Search view".

Metadata search is not possible, if metadata added to the file contains special characters (including space).

Procedure to search

To search the files or objects with Object name in Data Explorer, follow these steps:

1. In the "Data Explorer" overview page, click to open "Search view".
2. Select object name or metadata value with a condition or both parameters using an operator.
3. Enter file or object name or search text.
4. Click Search.

If the file or object with the appropriate option is not recognized, then search by selecting both Object name and Metadata value parameters with condition and operator.

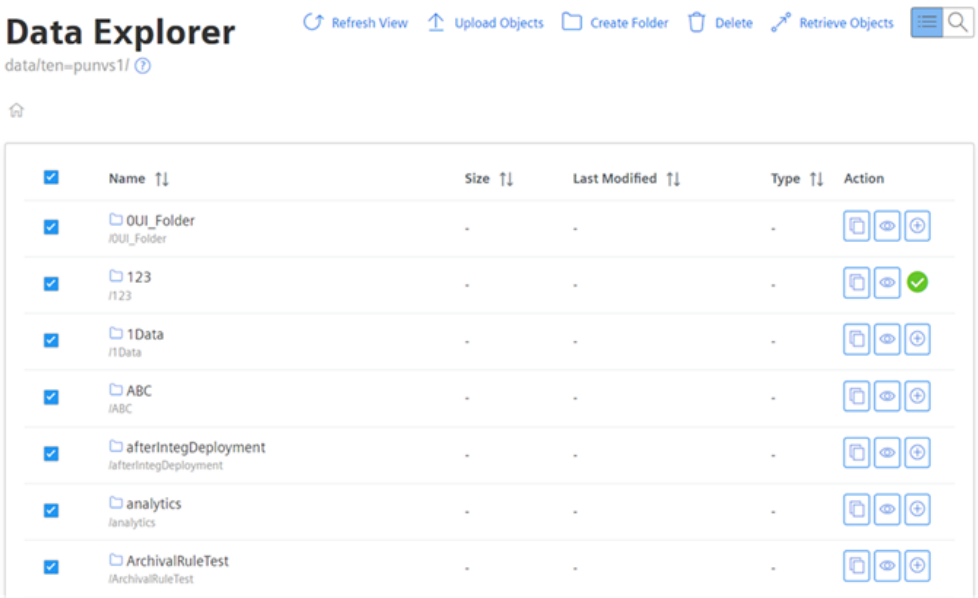
Result

The file or object details will be displayed and you can perform the actions from the Search view.

4.7 Deleting the folder or object in Data Explorer

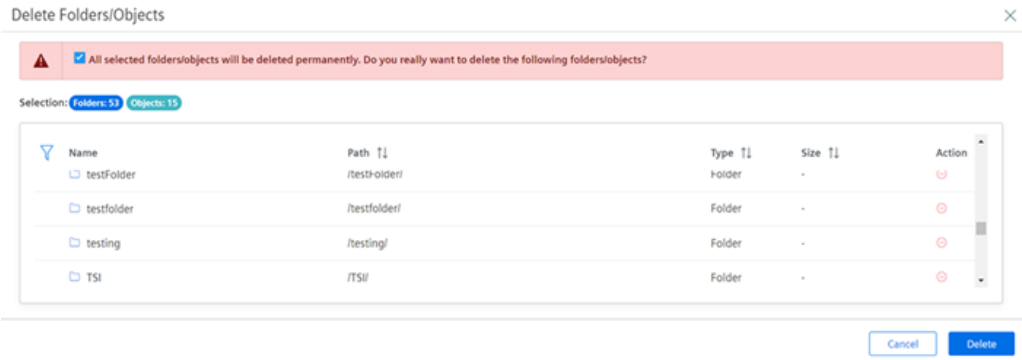
To delete the folder or object in Data Explorer, follow these steps:

1. In the Data Explorer overview page, select the folder or object.



2. Click .

3. Check to confirm.



4. Click "Delete".

- The folder will be deleted automatically, if you delete all files or objects from Data Explorer.
- Deleting a folder is allowed only if the folder is empty and no metadata value is added on the folder level.
- You can select single or multiple files.
- Click button to remove the file from the list.

Time Series Import

5

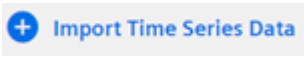
5.1 Importing data into Integrated Data Lake

The "Time Series Import" service allows the authorized users to import the historical IoT time series data to Integrated Data Lake. This enables an on-demand time series upload for analytics and machine learning tools.

Procedure

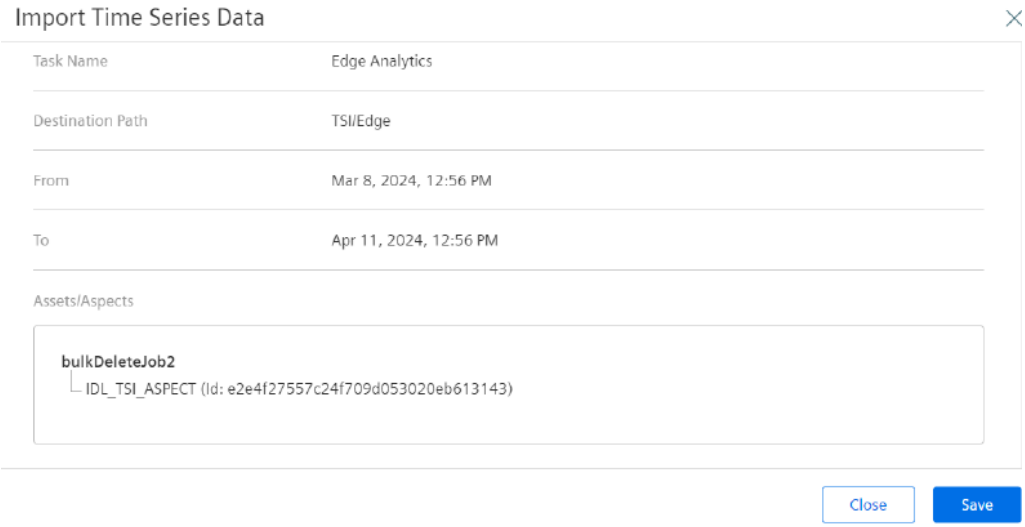
To import IoT time series data into Integrated Data Lake, follow these steps:

1. In the left navigation, select the "Time Series Import" tab and click "One Time Import". The "Import Task Details" screen is displayed.

   Click 

.

2. The "Import Time Series Data" screen appears.

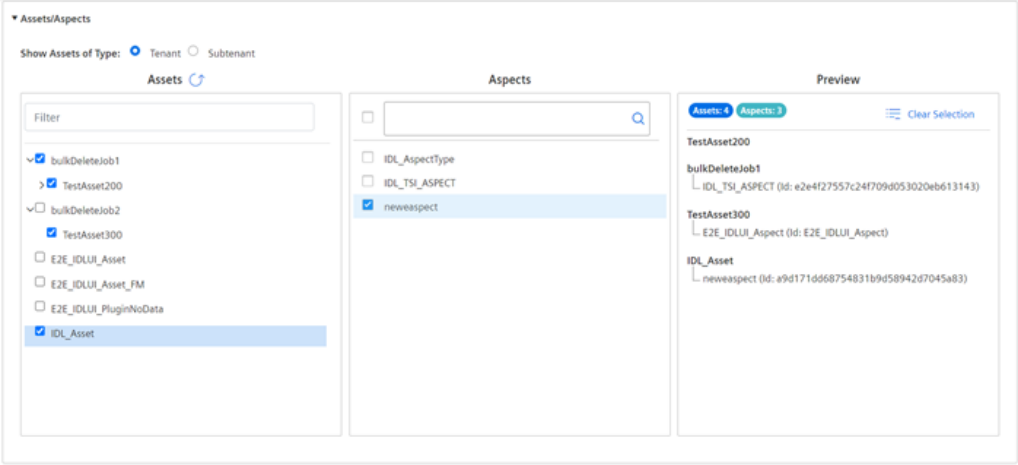


3. In the "Basic info" section, enter the "Destination Path" and the "Task Name".

4. Expand the "Date/time Range" option to set the date and time.

5. Click "Date/Time Range" to select the date and time range and click "Set".

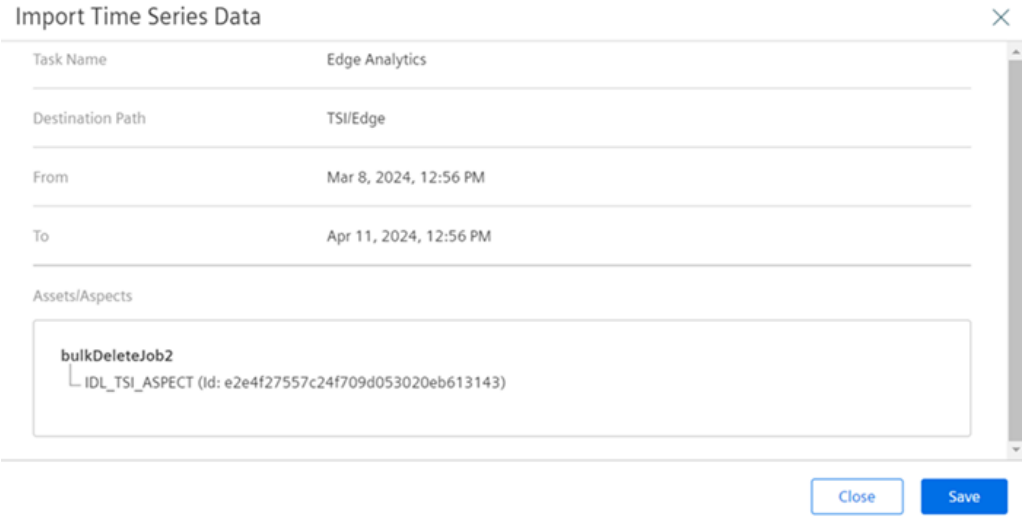
6. Expand the "Assets/Aspects" option to select the assets and corresponding aspects.



To refresh the asset list, click .

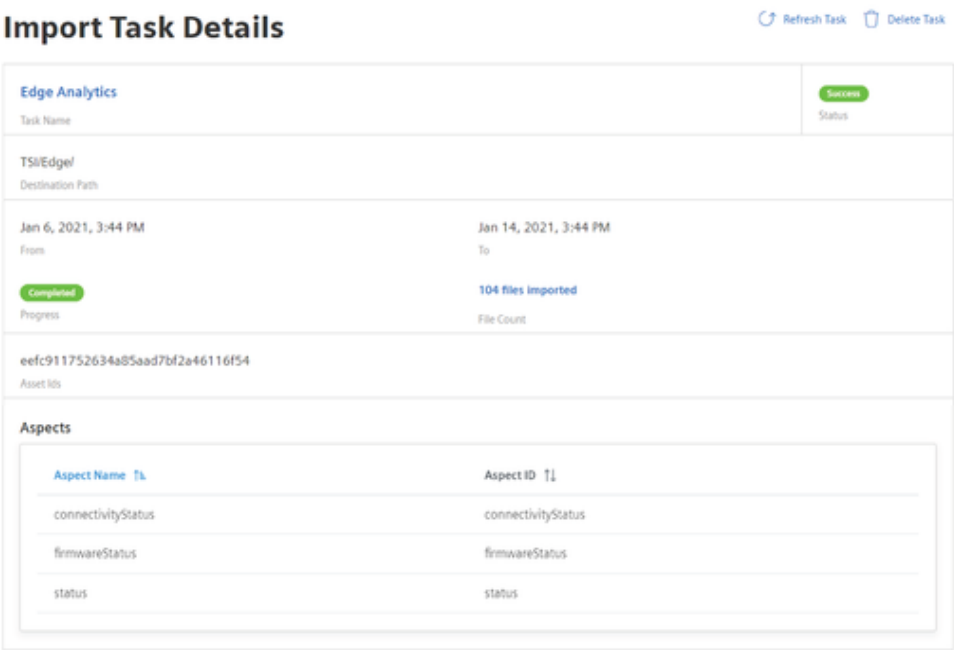
7.Select tenant or subtenant to get the respective assets from the "Assets" list and corresponding aspects from the "Aspects" list.

8.Click "Preview & Save".



9.Click "Save".

10.The "Import Task Details" status "Success" screen appears.



11.Click "Refresh Task" button to get the latest status of the import task.
Click "Delete Task" button to delete the task from the "Task Schedule" pane.

- Integrated Data Lake application will not allow the user to import the duplicate data based on "Date/Time Range", "Assets" and "Aspects" information. The application will provide the information about the existing bulk imports with the "Asset ID".
- A user is not allowed to select more than 3 months data to import.
- The Time Series data (parquet files) transferred to cold storage (typically within 7 days) and will be available in the Data Lake.
- There is a change in the namespace of the Time Series data imported into Integrated Data Lake.

Result

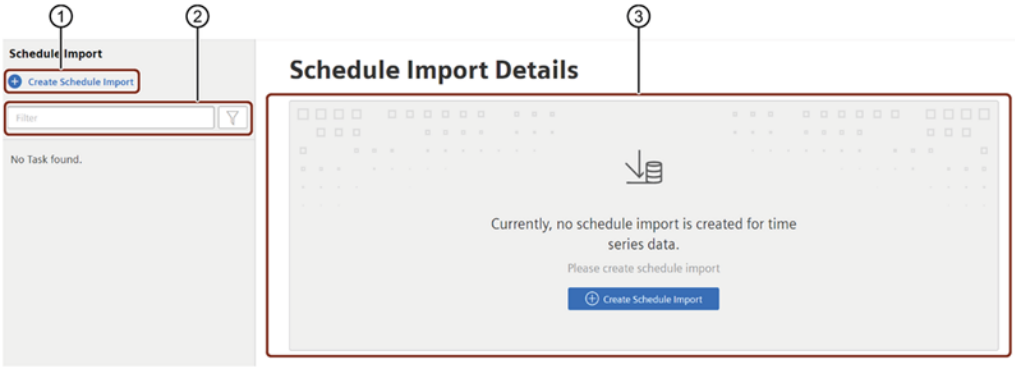
The historical IoT time series data will be imported to Integrated Data Lake and success status will be displayed with the number of file count.

5.2 Importing the scheduled Time Series data

In Time Series Import, you can schedule the import data. The import data can be scheduled with the frequency of "Daily", "Weekly" or "Monthly". Based on the frequency selection, your data will be imported as per the scheduled task.

User Interface

The "Schedule Import" is accessible from the Time Series Import tab.

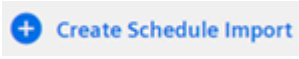


① Create Schedule Import jobs
② Filter or search the list of Schedule Import jobs
③ Displays the selected Schedule Import Details

Procedure

To create the scheduled IoT time series job, follow these steps:

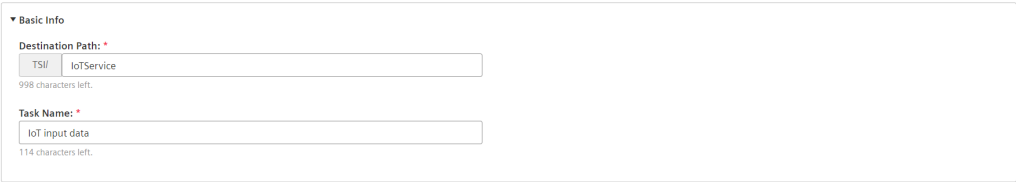
1. In the left navigation, select the "Time Series Import" tab and click "Schedule Import". The "Schedule Import Details" screen is displayed.

   Click 

.

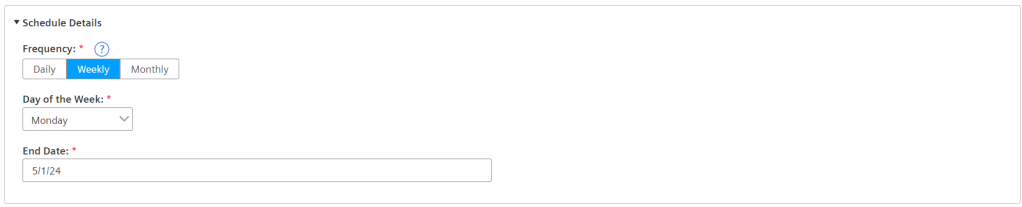
2. The "Import Time Series Data" screen is displayed.

3. In "Basic Info" section, enter the "Destination Path" and the "Task Name".

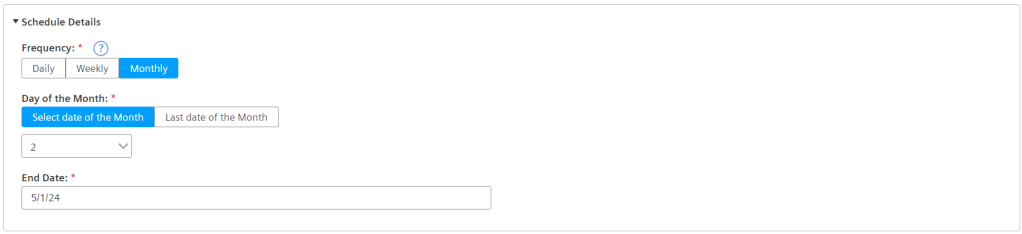


4. In "Schedule Details" section, choose any one "Frequency" to schedule the import:
   - **Daily**: This will schedule the job to run every day.

- **Weekly**: This will schedule the job to run once in a week on selected day of the week. Ensure the end date of the job is far enough to execute the at least 1 execution of the job.

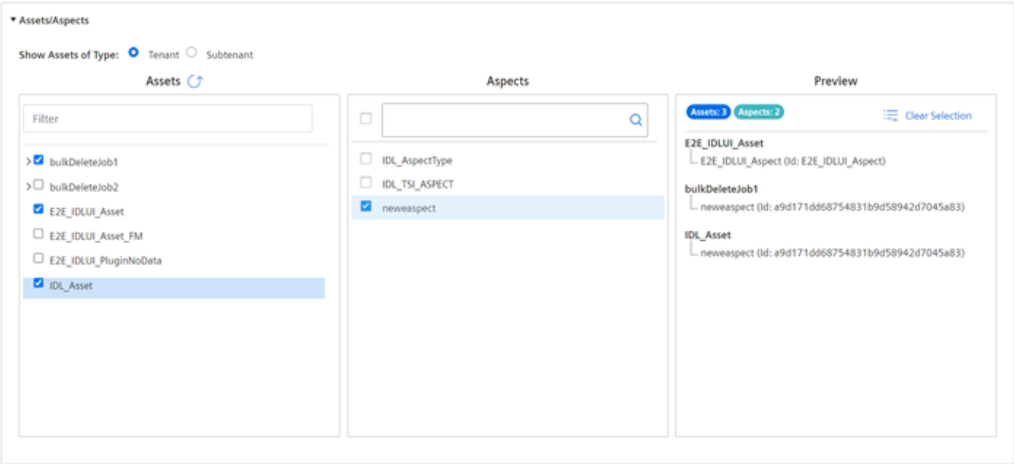


- **Monthly**: This will schedule the job to run once every month. Ensure the end date of the job is far enough to execute the at least 1 execution on the job. There are two option available in monthly schedule:
  - On the selected day of the month
  - On the last day of every month.



Job runs on the scheduled time and fetch the data one day prior to the execution date.

5.In "Assets/Aspects" section, select the Asset and Aspect.



Maximum 5 asset and aspect combination selection is allowed.

6.Click "Preview and Save".

7.Check the job preview and click "Save".

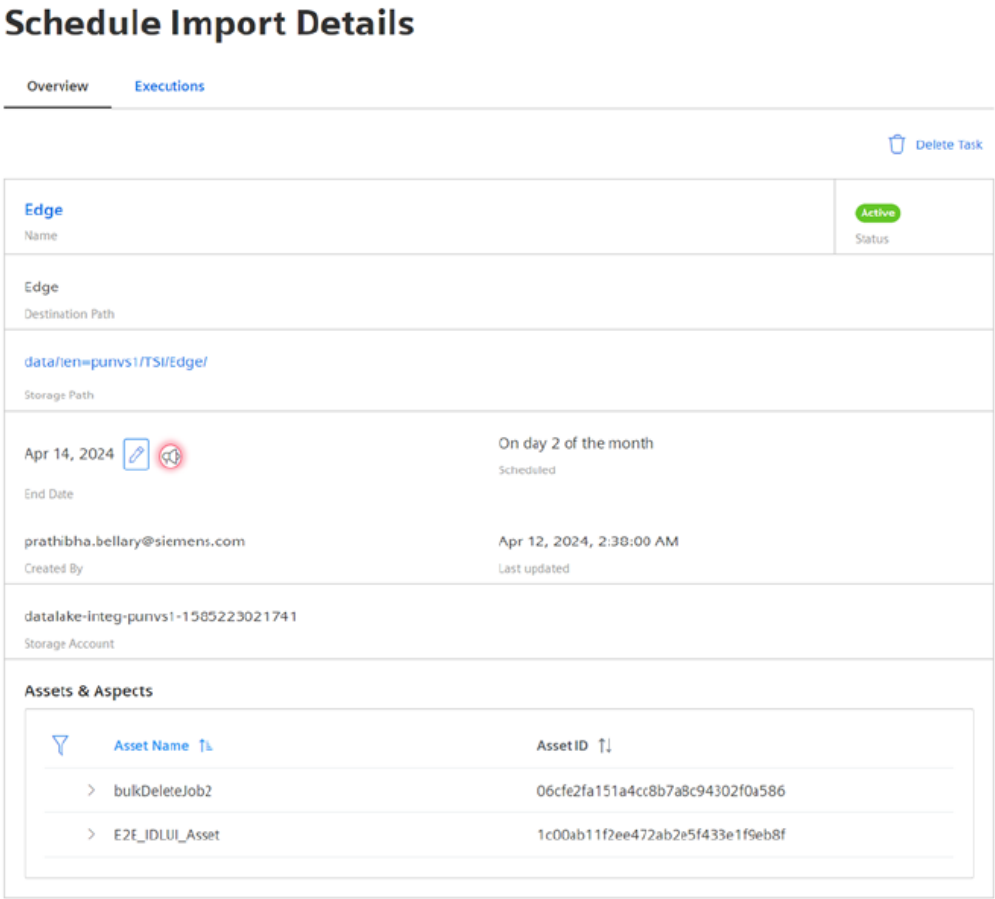
The status of the created job is described below:

- Active – It is the status of the job after creation. In "Active" status, you can extend "End Date" of the job.
- Inactive – It is the status of the job if any of "Asset/Aspect" from the job is deleted.
- Expired – It is the status of the job after "End Date" is met. You have to delete the job and create a new one after it expires.

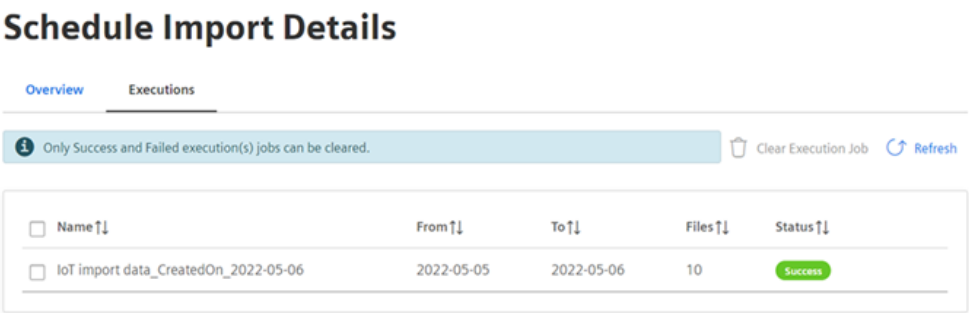
- Maximum limit to schedule the jobs is 10 for all users (tenant + subtenant).
- Only "Success" and "Failed execution" jobs can be cleared.

Result

The scheduled Time Series job is created successfully.



In "Schedule Import Details", click "Executions" tab for the outputs.


Schedule Import Details
Overview
Executions
Only Success and Failed execution(s) jobs can be cleared.
Clear Execution Job
Refresh
Name
From
To
Files
Status
IoT import data_CreatedOn_2022-05-06
2022-05-05
2022-05-06
10
Success

Cross Account Access

6

6.1 Enabling Cross Account access

Environment administrator can enable or revoke the cross account access for the AWS account. The administrator can also modify the relevant access to the specific prefixes for the folders in Integrated Data Lake. For more information, refer to User rights in Integrated Data Lake.

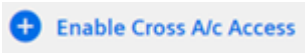
Region deviation

Cross account access is applicable only for Region Europe 1.

Procedure

To enable cross account access for the AWS account, follow these steps:

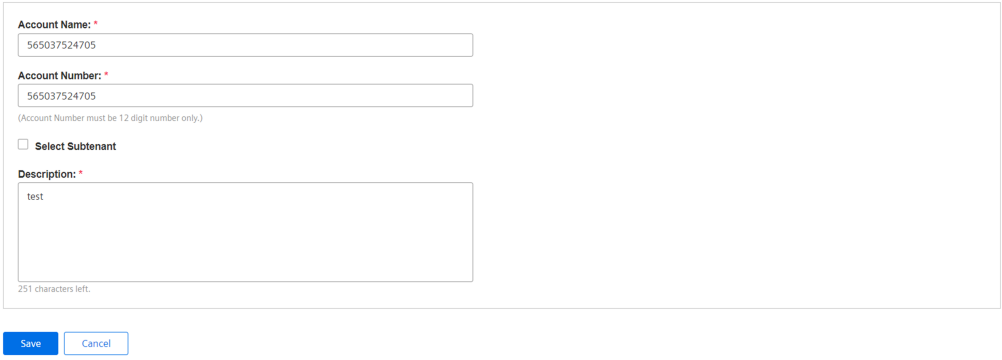
1. In the left navigation, click "Cross Account". The Account Details screen is displayed.

   Click 

.

2. The "Cross Account Access" screen appears.

Edit Cross Account Details

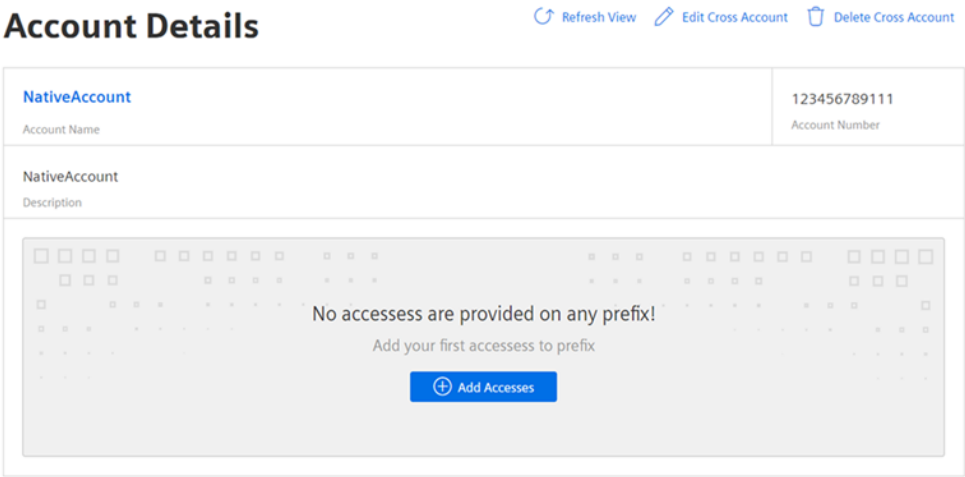


3. Enter the "Account Name".

4. Enter the valid AWS "Account Number" and "Description".

- Account name field is user defined.
- A user should have a valid 12 digit AWS account to enable the "Cross Account".
- Description field is mandatory to enable the cross account.

5.The subtenant option can be selected, if the AWS cross account access is for a subtenant.
6.From the "Select Subtenant" drop-down list, "Choose a subtenant" as per the requirements.
7.Click "Save" button.
The "Account Details" enabled screen appears.



8.Click "Edit Cross Account" button to edit the cross account.
9.Click "Delete Cross Account" button to delete the cross account.

Result

Cross Account access for a specific Europe 1 will be created. You can now create Cross Account Accesses to give access to specific folders.

6.2 Managing Cross Account accesses

You can manage the "Cross Account Accesses" by providing access to the specific folders based on the requirement and grant the permission to read, write or delete access as per the requirements.

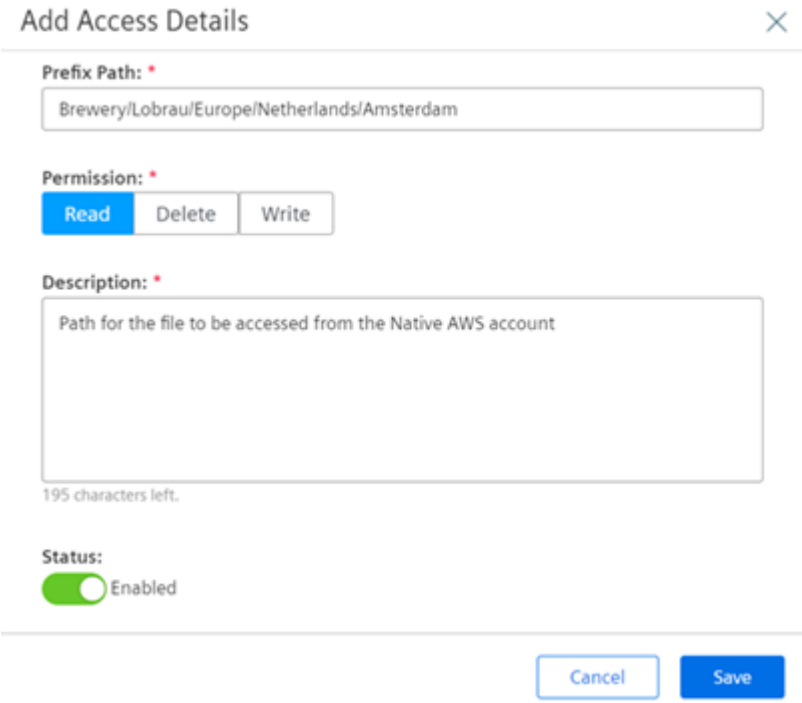
Procedure

To manage the cross accounts accesses, follow these steps:

1. In the left navigation, click "Cross Account". The Account Details screen is displayed.

In the "Account Details" screen, click the "Add Accesses" button.

2. The "Add Access Details" screen appears.



3. Enter the "Prefix Path" to specific account access path on which the accesses to be provided.

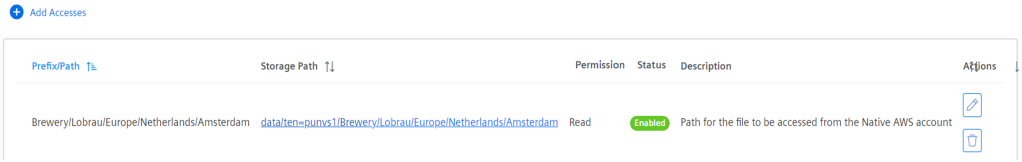
4. Select the "Access Permission" (Read, Delete or Write).

5. Enter the "Description" regarding the access details.

6. Click "Status" button to enable or disable the "Access Permission".

   Click "Save".

7. The account access details will added to the list.



8. Click the icon to edit the account access permission.

9. Click the icon to delete the account access.

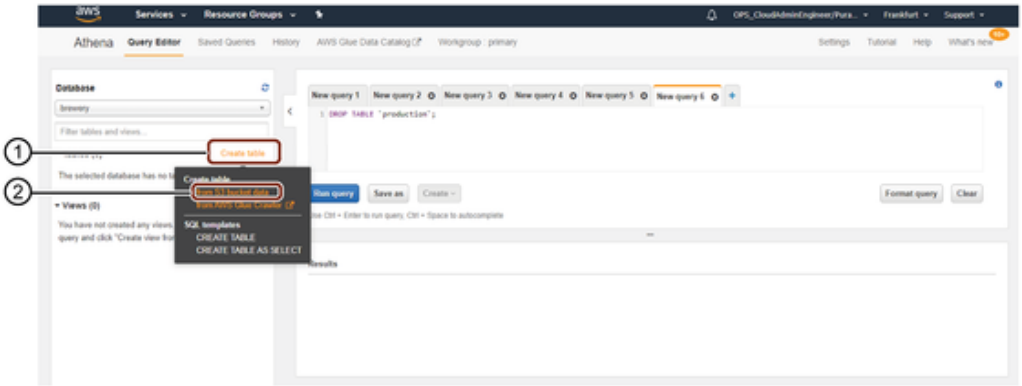
- A maximum of 5 prefixes can be enabled across one or multiple cross accounts.
- A maximum 10 cross account accesses can be created in a disabled state.
- Cross Account is set to be accessed from the native AWS tools.

6.3 Integrating AWS Athena tool with Integrated Data Lake

To integrate AWS Athena tool with Integrated Data Lake after enabling cross account access, follow these steps:

1. Open AWS Athena tool.

2. Create the table in Athena



① Click to create table
② Choose the option "from S3 bucket data"

3. Add the details related to database, table and input data.

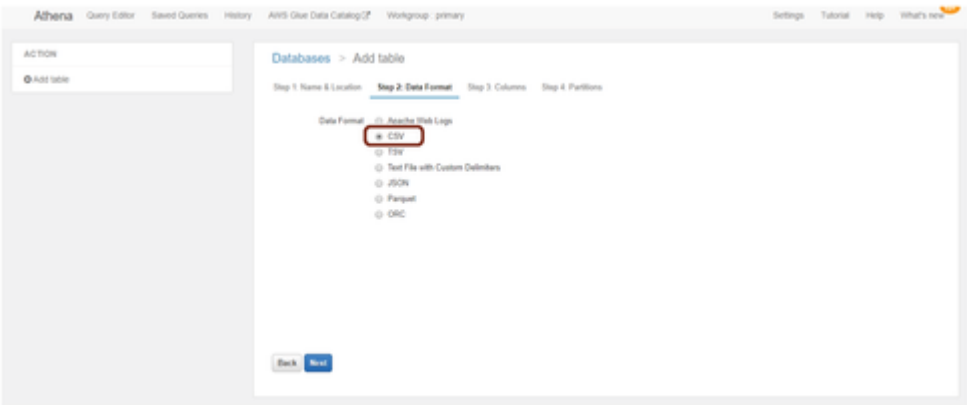
   - Database: Select the existing database or create a new by selecting "Create a new database".
   - Table Name: Enter the table name.
   - Location of Input Data Set: Path provided to access while configuring the cross account accesses.

Location of input data set would look like "`s3://+<<storage account>> + <<storage path>>` and path should end with '/'.
For example:

- Storage Account = datalake-prod-a-starter-576071197214
- Storage Path = data/ten=starter/Brewery/Lobrau/Europe/Netherlands/Amsterdam
- Location of Input Data Set = `s3://datalake-prod-a-starter-576071197214/data/ten=starter/Brewery/Lobrau/Europe/Netherlands/Amsterdam/`

4.Click "Next".

5.Select "Data Format".



File uploaded in the Data Explorer should be .csv file.

6.Add column name and column type corresponding to the data in the .csv file.
7.Click "Next".
8.Click "Create Table".

Result

The table is created successfully and now you can query the data in AWS Athena.

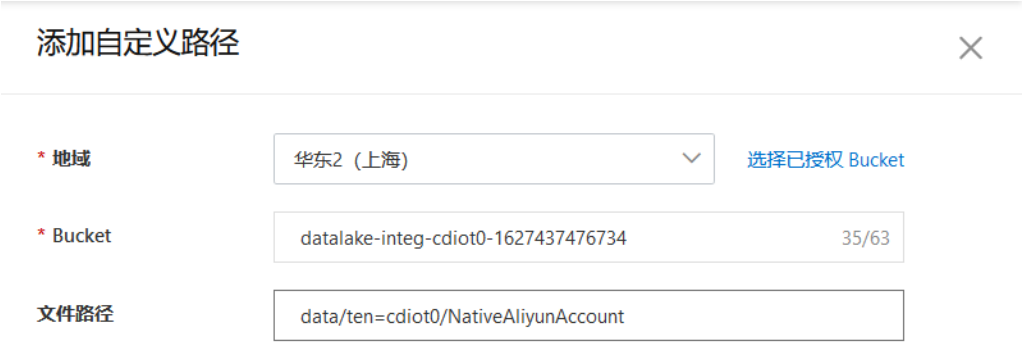
6.4 Integrating Aliyun OSS Web Console with Integrated Data Lake

To integrate Aliyun OSS Web Console with Integrated Data Lake after enabling cross account access, follow these steps:

1. Open Aliyun OSS Web Console.

2. Add My OSS Paths.

3. Click on the "+" to add the details related to access path and input data.
   - Region: The region of Integrate Data Lake Storage Account, default is `cn-shanghai`.
   - Bucket: Integrate Data Lake Storage Account name.
   - File Path: Storage Path provided to access while configuring the cross account accesses.



For example:

- Storage Account = `datalake-integ-cdiot0-1627437476734`
- Storage Path = `data/ten=cdiot0/NativeAliyunAccount`

4.Click "OK".

5.Select your added path.

There may some "Access Denied" pop up, as we only granted the Read Permission for this account.

Result

The data in Aliyun OSS Web Console is displayed and also user can access files under the Integrated Data Lake path using the account credentials through Aliyun OSS SDK or other Aliyun Native tools.

Service Principal

7

7.1 Creating Service Principal

You can create a Service Principal to configure the native tools of Microsoft Azure to access the files or objects in Integrated Data Lake. By using the Service Principal, you can call Azure Data Lake Storage (ADLS) APIs directly. You can generate Service Principal for each environment with the specific access limited to your own data.
The Service Principal's ClientId or Secret can be managed as per the guidelines.

Region deviation

Service Principal is available for Virtual Private Cloud.

Procedure

To configure the native tools of Microsoft Azure with Service Principal, follow these steps:

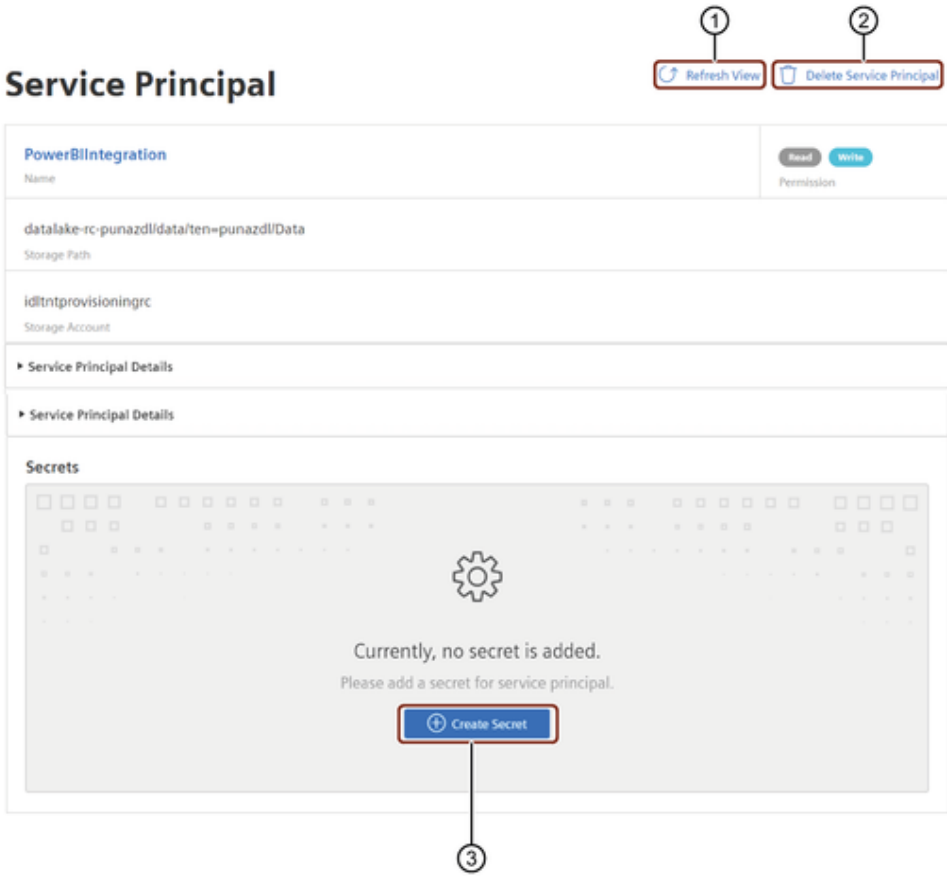
1. In the left navigation, click "Service Principal". The Service Principal screen is displayed.

2. In the "Service Principal" screen, click Create Service Principal.

- Enter the "Name".
- Enter the path for "Service Principal".
- Specify the access permission (Read, Write or Read and Write)

The subtenant option can be selected, if the Service Principal is for a subtenant.

3.Click "Create Secret".



Maximum of 5 Service Principals can be created in Integrated Data Lake.

A secret is created for the Service Principal to access the files or objects from the specified path. To create a secret, refer to Managing Service Principal.

7.2 Managing Service Principal

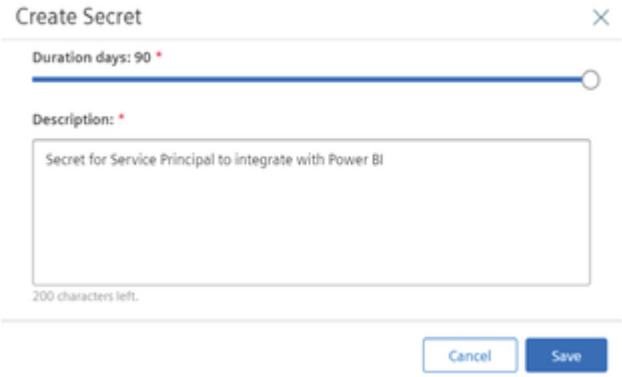
You can manage the "Service Principal" by adding a secret to configure with Microsoft Azure Native tools.

Procedure

To manage the service principal, follow these steps:

1. In the left navigation, click "Service Principal". The Service Principal screen is displayed.

2. In the "Service Principal" screen, click "Add Secret".



   - Select the duration from minimum 1 to maximum 90 day.
   - Enter the description.

3. Click Save.
4. Click to copy the generated secret and confirm the secret is copied.
5. Click Close.

- Maximum 2 Secrets can be added for each Service Principal in Integrated Data Lake.
- If the Secret is added, the Service Principal cannot be deleted.
- The duration of the secret is left with 30 to 10 days, then symbol will appear.
- The duration of the secret is left with 10 to 1 days, then symbol will appear.

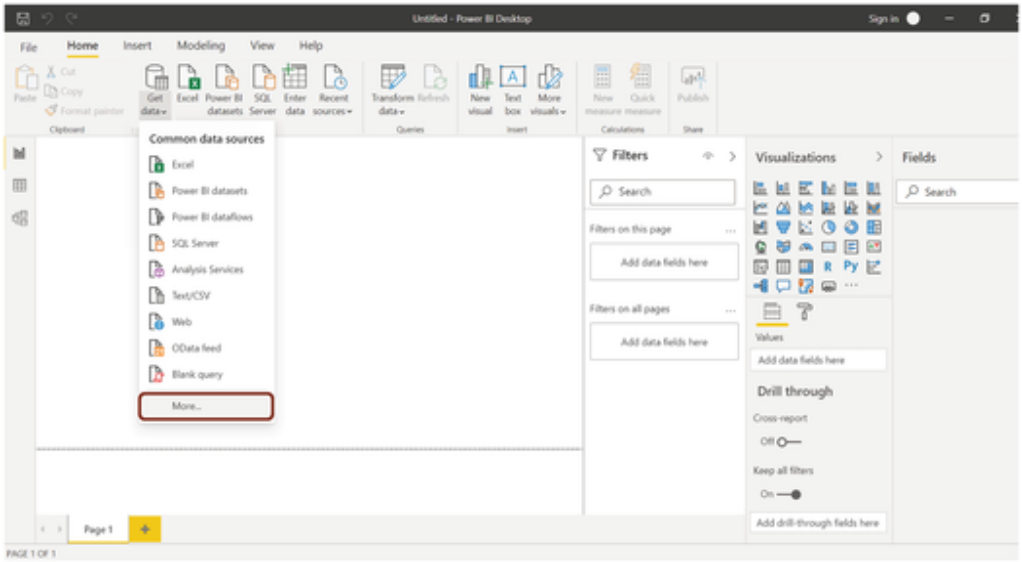
Result

The secret for the Service Principal is successfully generated. The Service Principal can now be configured with Microsoft Azure Native tools.

7.3 Configuring Azure Power BI tool with Integrated Data Lake

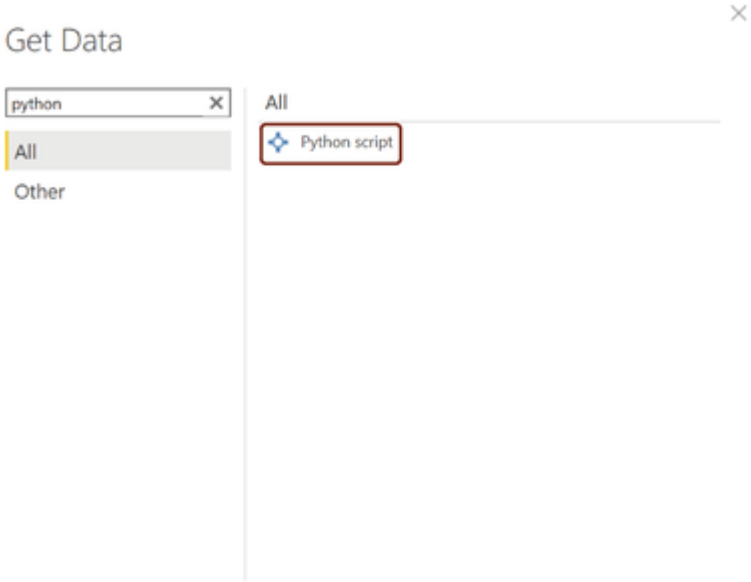
To configure Azure Power BI tool with Integrated Data Lake, follow these steps:

1. Open Power BI tool.



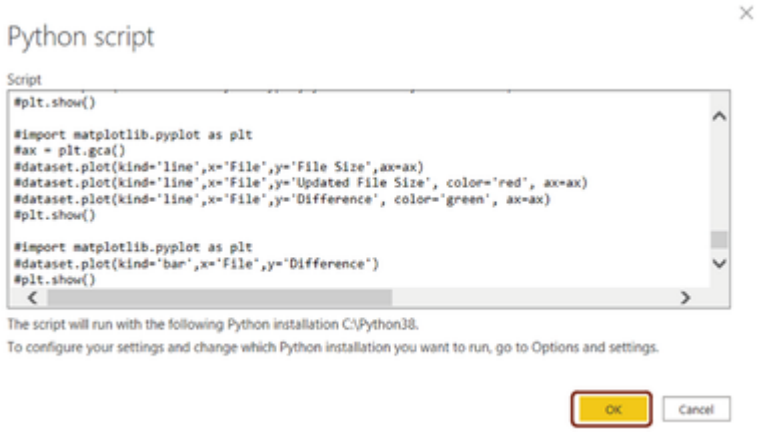
2. In Power BI tool, open "Get data" and select "More".

3. In Get Data screen, search for "python" and choose "Python script".



4. Click Connect.

5. Configure the python script connector



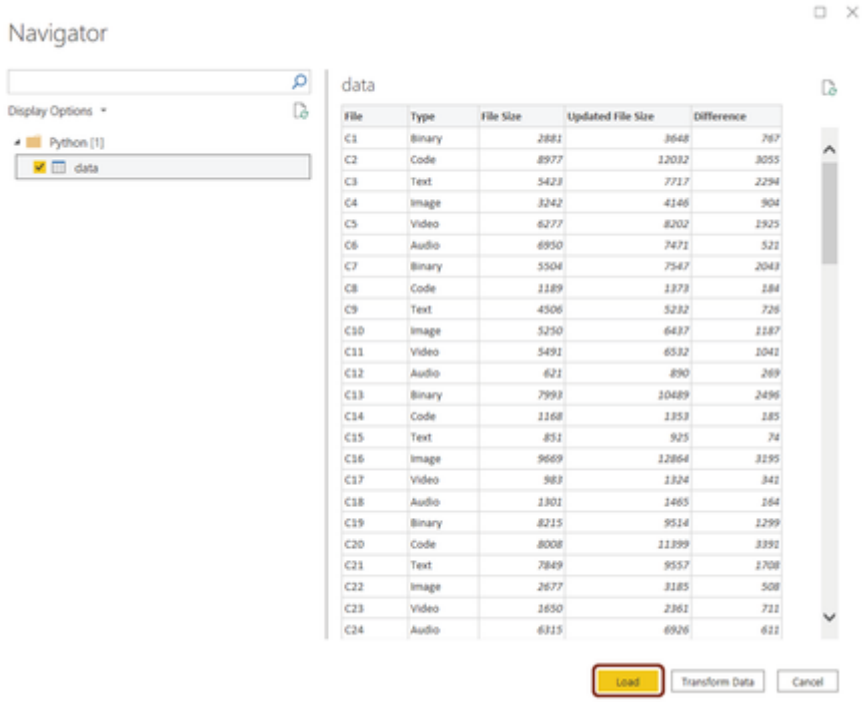
- Sample Python Script to be used as given below:

??? SamplePythonScript

```python
from pprint import pprint
import os, uuid, sys
from azure.identity import ClientSecretCredential
from azure.storage.filedatalake import DataLakeServiceClient
import pandas as pd
import matplotlib.pyplot as plt
from io import StringIO
def load_file_in_powerbi(file_system_client):
    try:
        directory_client = file_system_client.get_directory_clien
t(sampledata_directory)
        file_client = directory_client.get_file_client(sampledata
_filename)
        download = file_client.download_file()
        downloaded_bytes = download.readall()
        bytes_string = str(downloaded_bytes, 'utf-8')
        data_stringio = StringIO(bytes_string)
        global data
        data = pd.read_csv(data_stringio)
        print(data)
    except Exception as e:
        print(e)
def initialize_storage_account_ad(storage_account_name, client_i
d, client_secret, tenant_id):
    try:
        global datalake_service_client
        credential = ClientSecretCredential(tenant_id, client_id,
client_secret)
```

```
            datalake_service_client = DataLakeServiceClient(account_u
rl="{}://{}.dfs.core.windows.net".format("https",
            storage_account_name), credential=credential)
        except Exception as e:
            print(e)
    def get_file_system_client(file_system_name):
        try:
            file_system_client = datalake_service_client.get_file_sys
tem_client(file_system_name)
            return file_system_client;
        except Exception as e:
            print(e)
    def get_directory_client(file_system_client, directory_path):
        try:
            directory_client = file_system_client.get_directory_clien
t(directory_path)
            return directory_client;
        except Exception as e:
            print(e)
    def upload_file_to_directory(directory_client, src_file_path,file
_name):
        try:
            file_client = directory_client.create_file(file_name)
            local_file = open(src_file_path, 'r')
            file_contents = local_file.read()
            file_client.append_data(data=file_contents, offset=0, len
gth=len(file_contents))
            file_client.flush_data(len(file_contents))
        except Exception as e:
            print(e)
    def list_directory_contents(file_system_client, directory_path):
        try:
            paths = file_system_client.get_paths(directory_path)
            return paths
        except Exception as e:
            print(e)
    def download_file_from_directory(directory_client, file_name):
        try:
            local_file = open(file_name, 'wb')
            file_client = directory_client.get_file_client(file_name)
            download = file_client.download_file()
            downloaded_bytes = download.readall()
            # bytes_string = str(downloaded_bytes,'utf-8')
```

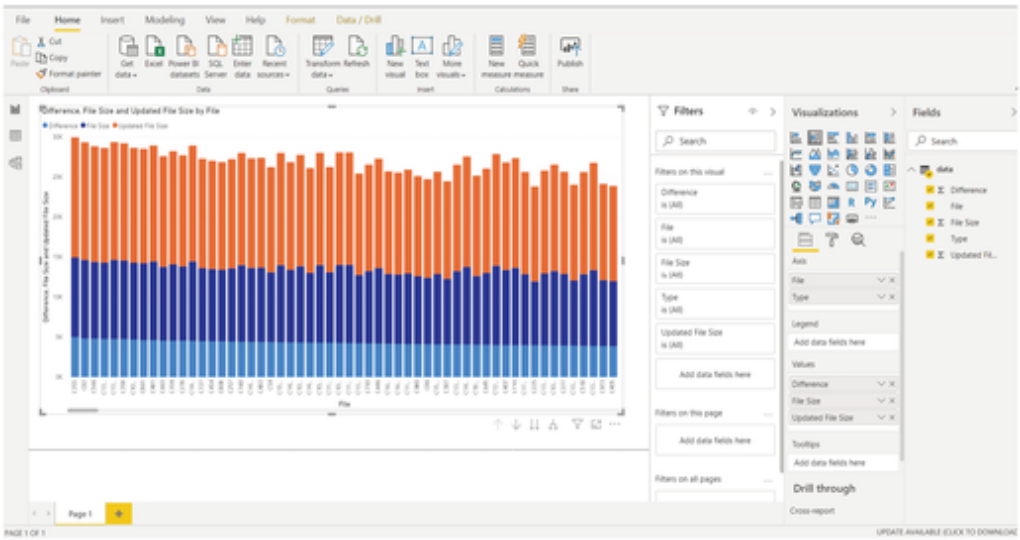
```
            # data_stringio = StringIO(bytes_string)
            # data = pd.read_csv(data_stringio)
            # completedData = data.fillna(method='backfill', inplace=
False)
            # data["completedValues"] = completedData["SMI missing va
lues"]
            # pprint(data)
            # pprint(vars(data))
            # pprint(data.head())
            # data.plot(kind='bar', x='Day', y='completedValues', col
or='red')
            # plt.show()
            local_file.write(downloaded_bytes)
            local_file.close()
        except Exception as e:
            print(e)
    def download_hierarchy_directory(file_system_client, directory_pa
th):
        try:
            directory_client = get_directory_client(file_system_clien
t, directory_path)
            paths_list = list_directory_contents(file_system_client,
directory_path)
            for path in paths_list:
                if path.is_directory == True:
                    download_hierarchy_directory(file_system_client,
path.name)
                else:
                    # print(path.name + '\n')
                    # pprint(vars(path))
                    path_split = path.name.rsplit("/", 2)
                    if path_split[1] == directory_path.rsplit("/", 1)
[1]:
                        file_name = path_split[2]
                        print(file_name + '\n')
                        download_file_from_directory(directory_clien
t, file_name);
        except Exception as e:
            print(e)
    `if __name__ == "__main__":`
        tenant_id = "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx"
        client_id = "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx"
        client_secret = "xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx"
```

```
        storage_account_name = "idltntprovisioningrc"
        file_system_name = "datalake-rc-punazdl"
        directory_path = "data/ten=punazdl/powerbi"
        sampledata_directory = "data/ten=punazdl/powerbi/sampledata"
        sampledata_filename = "file-size-upload.csv"
        initialize_storage_account_ad(storage_account_name, client_i
d, client_secret, tenant_id)
        file_system_client = get_file_system_client(file_system_name)
        paths = list_directory_contents(file_system_client, directory
_path)
        directory_client = get_directory_client(file_system_client, d
irectory_path)
        load_file_in_powerbi(file_system_client)
 ```
```

| Application ID | Location |
|---|---|
| tenant_id | Tenant ID will be available on the Service Principle page in Data Lake. |
| client_id | This is the application ID that can be copied from the Service Principle page in Data Lake. |
| client_secret | This is the Service Principle secret that was generated and copied. |
| sampledata_directory | Path on which Service Principle is generated. |
| sampledata_filename | File to be pulled into Azure Power BI. |

6.Click OK to load data into Azure Power BI.


```

7.Click "Load".

Result

The data is successfully loaded in Azure Power BI. You can now visualize your data in Azure Power BI.



Event Subscription

8

8.1 Creating an Event Subscription

Event Subscription allows you to subscribe the events to get notifications. You can register the destination Service Bus for Azure and Simple Notification Service for AWS to get notifications, which will be published by Integrated Data Lake service. These notifications include object events like create, update or delete in the environment prefix. You can add, view, edit and delete event subscriptions.
If the permission to send notification to the Service Bus or Simple Notification Service topic is removed or if there is any misconfiguration, the tenantAdmin is notified via email to check and respond. Even the status of the subscription will be changed to "Inactive" on the user interface. You can change the status to "Active" by resolving the configuration in the user interface.

Subscription name is not mandatory. Incase the name is not provided, default name for the subscription will be chosen as `(subscription_<<unique_identifier>>)`.

Procedure

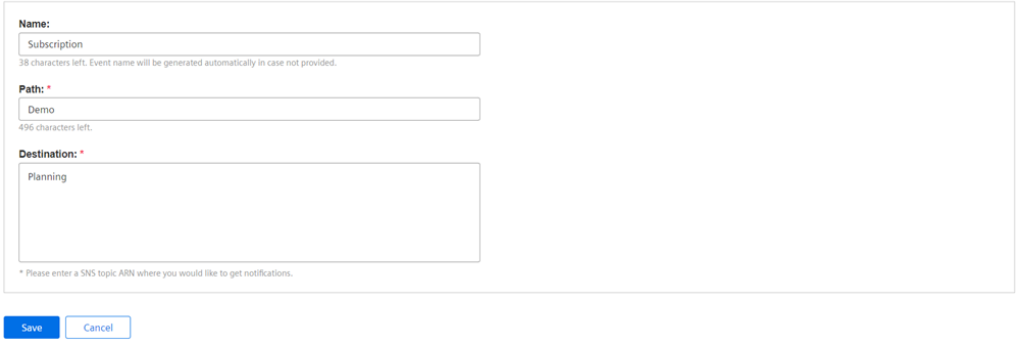
To subscribe the events with the notifications, follow these steps:

1. In the left navigation, click "Event Subscription". The Event Subscription screen is displayed.

2. In the "Event Subscription" screen, select the Tenant Subscriptions and click .

3. Click Create Event Subscription.

Create Event Subscription



- Enter the "Name".
- Enter the path for storage.
- Enter the path for destination.

4. Click "Save".

- You can only change the status from "Inactive" to "Active".
- Maximum 15 Event Subscriptions can be created for each environment.
- If the subtenant is deleted, you can only view its existing event subscriptions and delete them if not required.

oData Contract

9

9.1 Adding and managing oData Contract

The oData Contract allows you to add and manage the contracts of selected folders path to be used in the Mendix app via the Mendix Data Hub.
To access the data using Mendix Data Hub, you need to create oData contract(s) to allow access based on selected folder. For more information about this process, refer to Data Hub Guide{target="_blank"}.

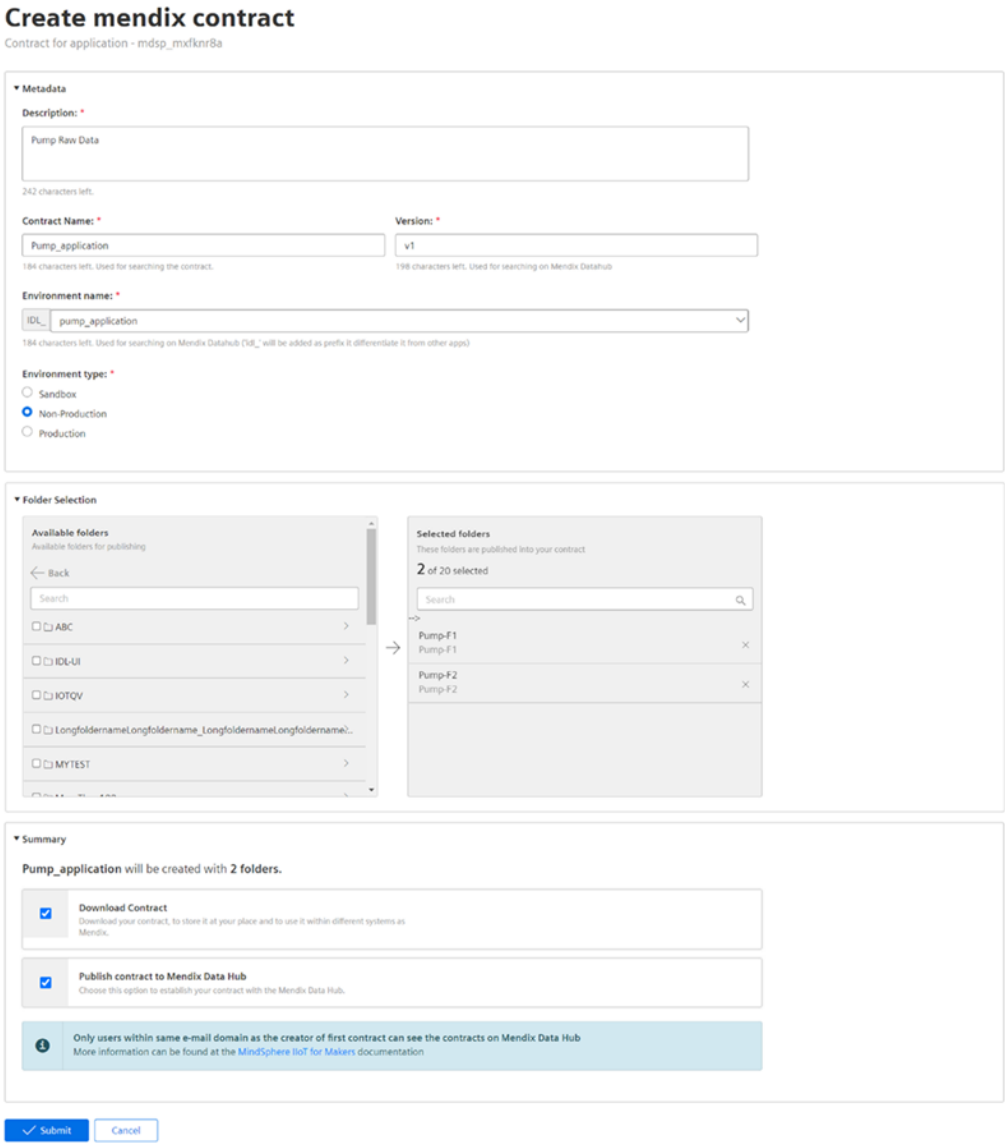
oData Contract Status

Status	Description
Published	Contract is published to Mendix Data hub
Publishing	Contract is not published due to network issue while Publishing to Mendix Data hub
Not published	Contract is ready to be published to Mendix Data hub

Procedure

To add and manage the oData contract, follow these steps:

1. In oData Contract, click "Add contract".



2. In "Create mendix contract" screen, configure the following sections:
   - In "Metadata" section, enter the contract details for the application.
   - In "Folder Selection" section, select the required folders for the contract.

     Note:

     Maximum 20 possible folders can be selected for a single contract.
   - In "Summary" section, select "Download" and "Publish contract to Mendix Data Hub".

     Note:

     Downloading a contract is optional.
3. Click "Submit".
4. In Mendix "Authorize access to account" screen, confirm and accept the agreement.

Maximum 50 possible contracts can be created in oData Contract.

Result

The oData Contract is successfully added and you can also view the contract in Mendix data hub.

Metadata Management 10

10.1 Exploring Metadata Management

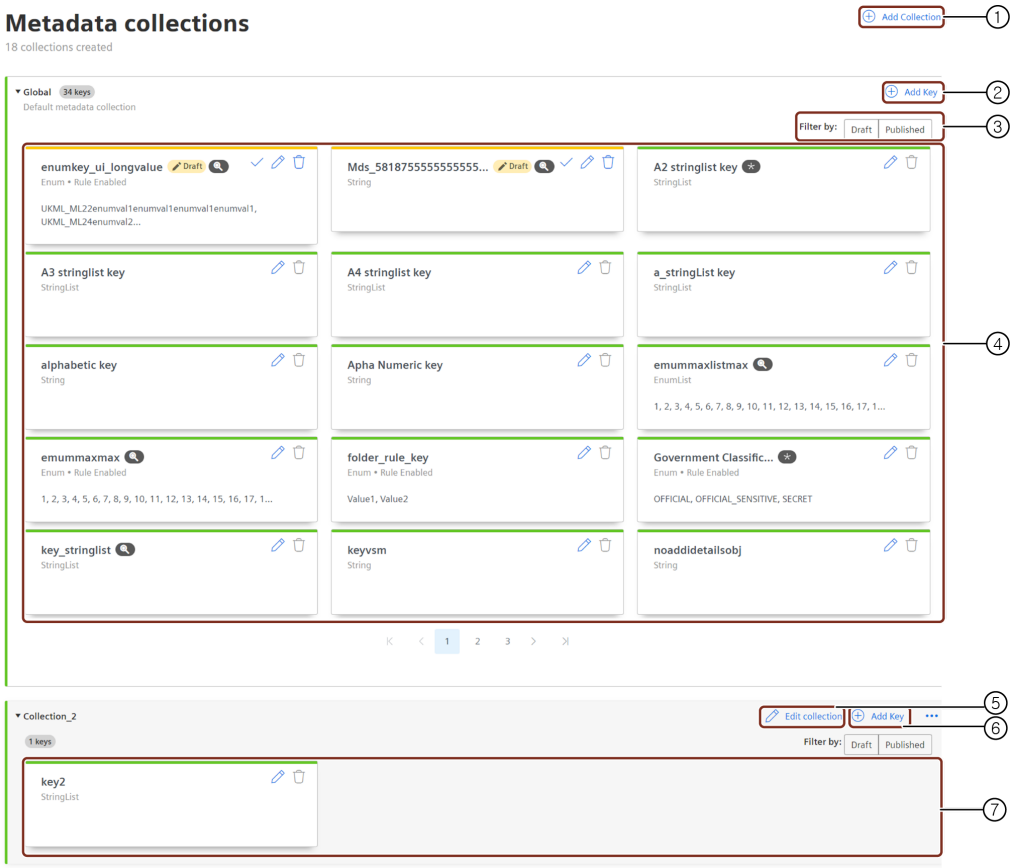
Metadata Management is a key element of the data management in the Integrated Data Lake application. It involves the systematic organization, storage, retrieval and maintenance of metadata. Metadata Management includes information on the characteristics, origin, usage and connections of the actual data stored in the Integrated Data Lake.
In Integrated Data Lake, metadata is available in the form of values and can be provided in folders or objects. Metadata values provides a way to categorize content, focus on key concepts or themes. Integrated Data Lake introduces the structured approach of metadata by utilizing key-value pairs. Users can configure metadata keys, metadata collections, and define rules that control the behavior of these collections.
Metadata Management is a well-organized and functional behavior Integrated Data Lake, which serves as a fundamental framework and plays an important role in enhancing user decision, data quality and contributing to the overall success of data-driven initiatives.

Metadata Management is available only for Virtual Private Cloud.

Metadata Management offers the following benefits:

- Metadata Management enables efficient data exploration by providing detailed information about the contents of the Integrated Data Lake. User can easily search and locate relevant datasets based on the attributes such as file type, creation dates or keywords.
- Metadata Management enhances raw data by providing insights into its structure, format and meaning. In the large and complex Integrated Data Lake environments, it is essential for users to understand the context of the data to ensure precise interpretation and effective utilization.
- Metadata Management is required for enforcing access control policies and ensuring restricted access. It plays an important role in providing conditional access permissions within the access control policies, which are essential for data security.
- Metadata Management enables connecting various consumer applications to the data storage, enabling Integrated Data Lake as a single source of data storage.

The following image shows the user interface of the metadata collection screen:



① Adds a new custom collection.

② Adds a new key in global collection.

③ Filters the keys available in global collection by draft or published status.

④ Displays the list of keys available in global collection.

⑤ Edits a custom collection.

⑥ Adds a new key in custom collection.

⑦ Displays the list of keys available in custom collection.

10.2 Overview

Metadata collection is the systematic process of collecting metadata keys that manage all the relevant metadata information associated with the specific dataset or system. In custom metadata collection, metadata keys acts as identifiers for a specific attributes, characteristics or properties of the data by providing a structured overview of the dataset. This approach enables the user for efficient retrieval and management of the metadata, contributing to the enhanced data understanding and overall system functionality.

Metadata collections are classified into three categories.

Global collection: Global collection is a default collection created by the system when the initial key is defined in the metadata collection. By default, the global collection is always in published status. Users can directly use this collection to define Metadata keys.

Custom collection: Custom collections are user created collections to define metadata keys. Admin can create up to 50 custom collections in the system. By default, any custom collection

created by users are in draft status and it is changed to published once the collection is published.

In the metadata collection screen, select a custom collection and click ••• to view the following options.

① View the details of the collection.

② Adding a rule to the collection.

③ Publish the collection.

④ Delete the collection.

- When the collection is in draft status, "View details", "Add Rule", "Publish" and "Delete" options are enabled.
- When the collection is in published status, only "View details" and "Add Rule" are enabled.

Reserved collection: Reserved collections are created, managed and are used by the system for the metadata like file type, file size, created date etc. This collection is always in published status.

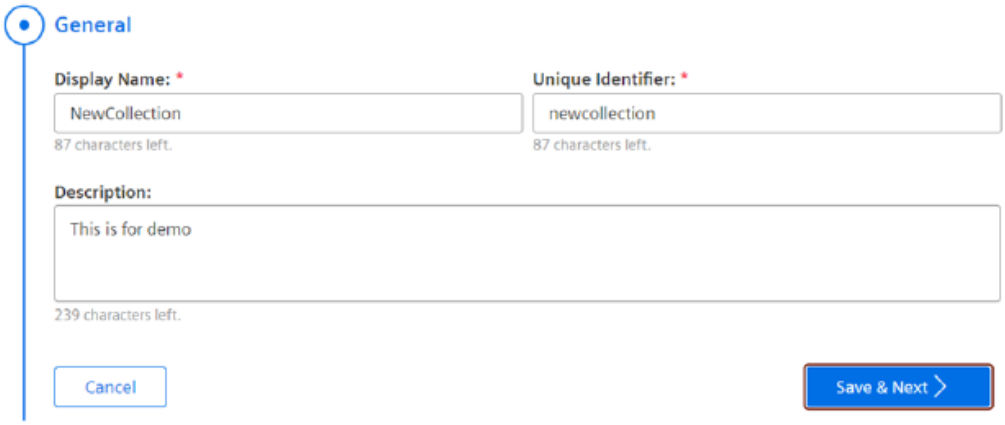
For more information on adding a new collection, refer to Adding a collection.

10.3 Adding a collection

It is only possible for the user to create custom collections to define the keys. To add a new custom collection, follow these steps:

1. In the left navigation, click "Metadata Management". The Metadata collection overview page is displayed.
2. Click "Add Collection".
3. In the "General" section,
   - Enter the "Display Name" and "Description".
   - By default, "Unique Identifier" is automatically filled with the same name as "Display Name".

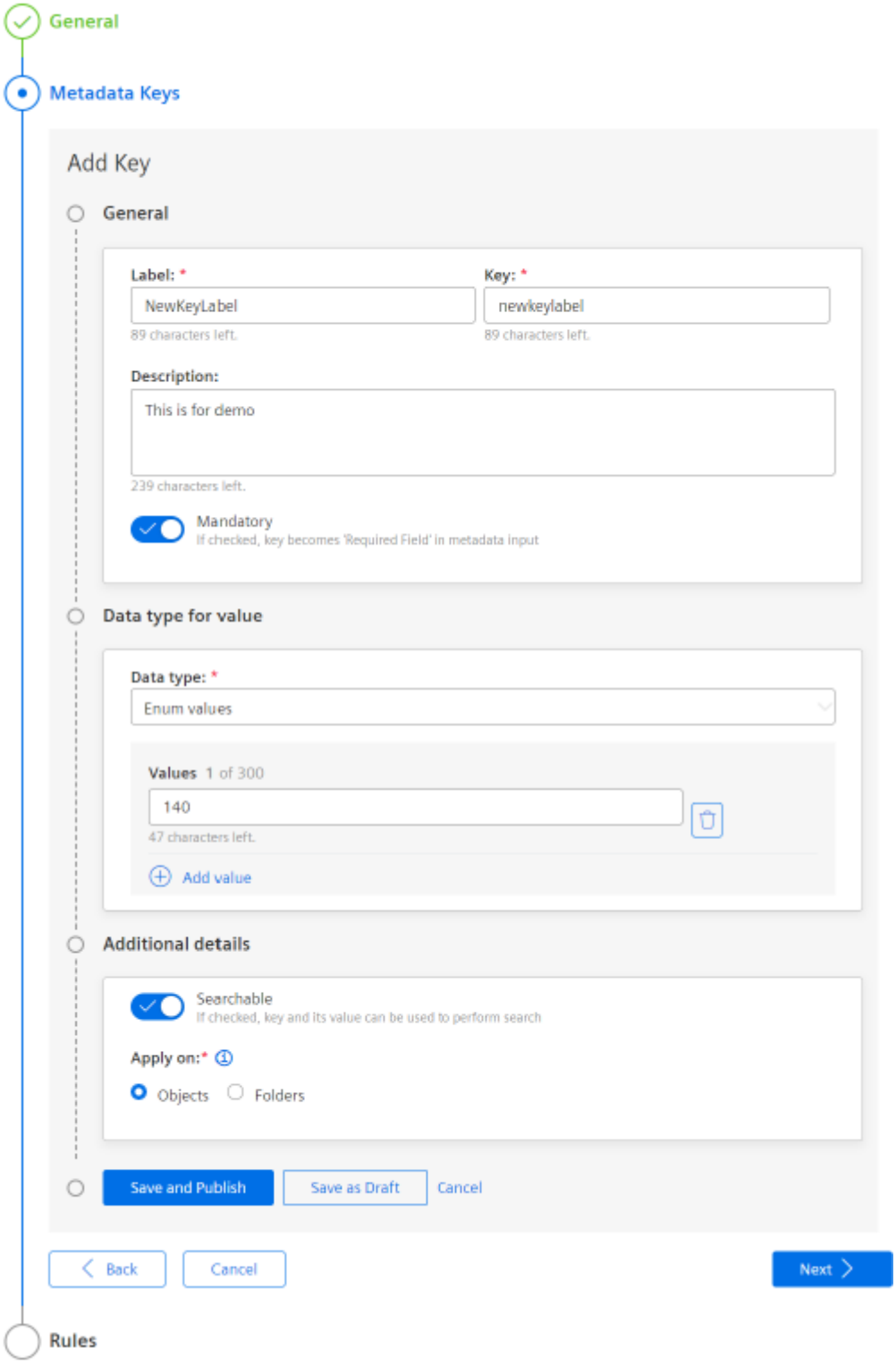
4. Click "Save & Next".



5. In the "Metadata Keys" section, click "Add Key". For more information to add a key, refer to [Adding a Key](#).

Click "Next".

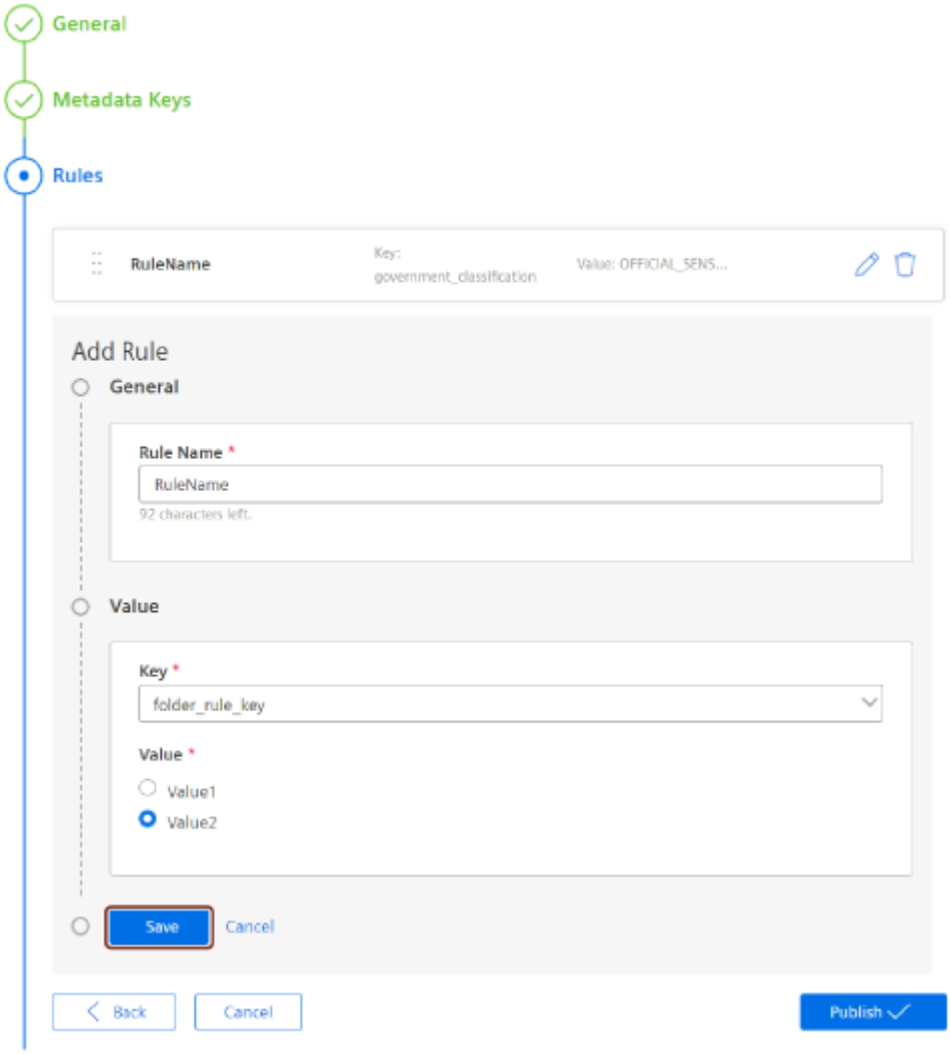
6. To return back to the "General" section, click "Back".



7. In the "Rules" section, click "Add Rule".
   - In the "General" section, enter the "Rule Name". Rule name should be unique, the same rule name cannot be re-used.
   - In the "Value" section, select the "Key" and the corresponding "Value" of your choice.

- In the "Key" dropdown, only the keys defined in the global collection with Enum data types are listed.
- The rule name defined with the key and its value can be used only once.

8.Click "Save".



9.Click "Publish".

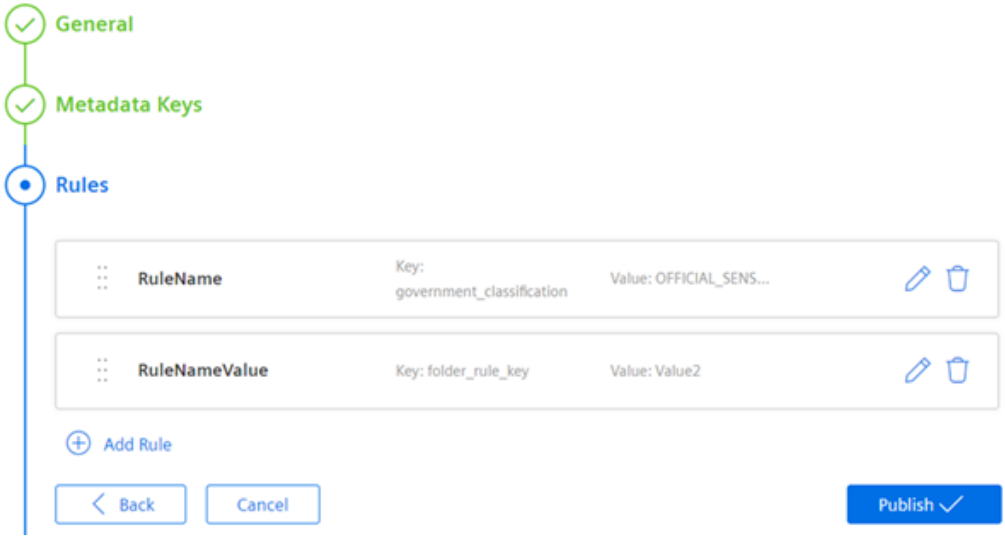
- To publish a collection, it should have at least one key and an associated rule.
- When a metadata collection is in draft status, the defined "Data Type" in "Metadata Keys" section cannot be modified.
- When a metadata collection is in published status, the defined "Rules" and the defined "Data Type" in "Metadata Keys" section cannot be modified.

By default, the collection created is automatically saved in draft status until the collection is published.

10.4 Editing the collection

The user can edit the metadata collection in both draft and published status. To edit the metadata collection, follow these steps:

1. In the left navigation, click "Metadata Management". The Metadata collection overview page is displayed.
2. Select the metadata collection to edit.
3. Click and modify the details. For more information to edit the details, refer to Adding a collection.



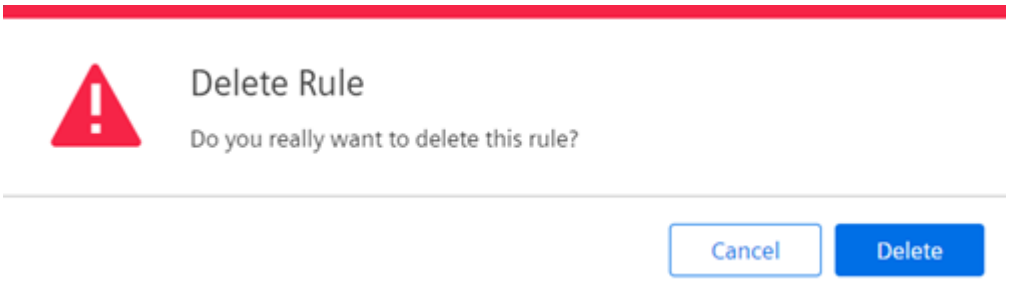
4. Click "Publish".

10.5 Deleting the collection

It is possible to delete the metadata collection only if it is in the draft status. To delete the metadata collection, follow these steps:

1. In the left navigation, click "Metadata Management". The Metadata collection overview page is displayed.
2. Select the metadata collection to be deleted.

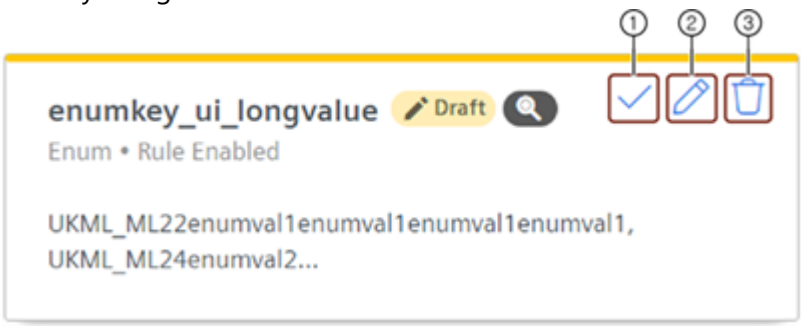
3. Click .



4. Click "Delete".

10.6 Overview

Metadata keys are the unique identifiers used to represent key-value pair structure, representing specific attributes or characteristics linked to data. They are a part of a global or custom collection and are used to create rules in a custom collection. Metadata key is a reference point for a specific metadata information to categorize and describe the specific metadata information to various aspects of the data.
For global collection, the Integrated Data Lake admin is not required to create the collection to add the metadata key and can directly use the global collection ID. For custom collection, the Integrated Data Lake admin should create a custom collection and then add a metadata key to the collection. To include metadata in Integrated Data Lake resources, users should define metadata keys either in the global or custom collection.
The key in a global or custom collection is shown below:



① Publish the key.
② Edit the key.
③ Delete the key.

- When the selected key is in draft status, publish, edit and delete options are enabled.
- When the selected key is in published status, only edit option is enabled.

Key parameters

Consider the following parameters when adding a key to metadata collection.

Data type values

Metadata key should be defined with a data value type as mentioned in the below table:

Data type	Field Name	Field Values
Enum values	options	The values accepted for the "Enum values" data type are in capitals and separated by commas.
Enum list values	options	The values accepted for the "Enum list values" data type are in capitals and separated by commas.
	Minimum Size	The minimum number of values that are selected from the enum list should be greater than 0 and less than the total number of options.
	Maximum Size	The maximum number of values that are selected from the enum list should be greater than minimum size and less than or equal to number of options.
String values	options	The values accepted for the "String values" data type are in capitals and separated by commas.
	Minimum length	The minimum length of character set in the list that is accepted as value should to be greater than 0.
	Maximum Length	The maximum length of character set in the list that is accepted as value should to be greater than or equal to minimum length.
List of String	options	The values accepted for the "List of String" data type are in capitals and separated by commas.
	Minimum Size	The minimum number of values are selected from the list of string has to be greater than 0 and less than number of options.
	Maximum Size	The maximum number of values are selected from the list of string should to be greater than minimum size and less than or equal to number of options.

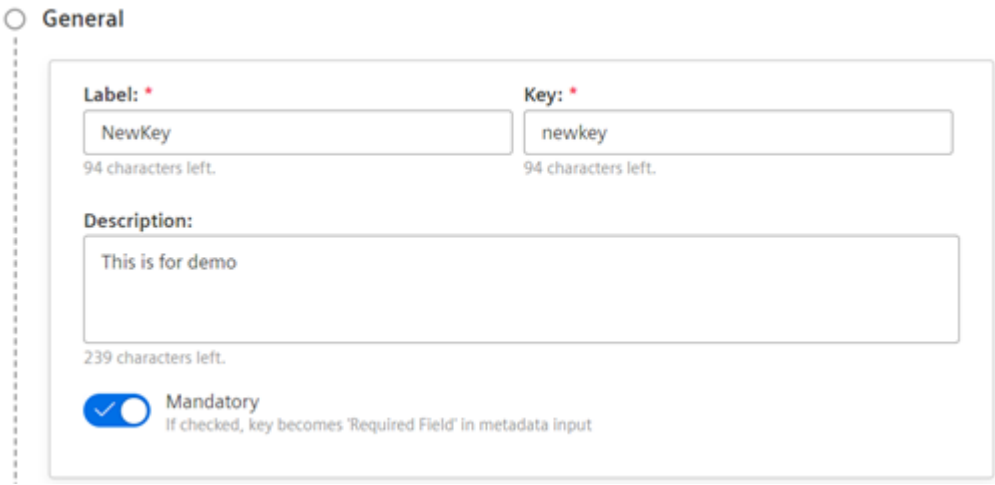
10.7 Adding a key

The user can add keys in global and custom collections. To add a new key, follow these steps:

1. In the left navigation, click "Metadata Management". The Metadata collection overview page is displayed.

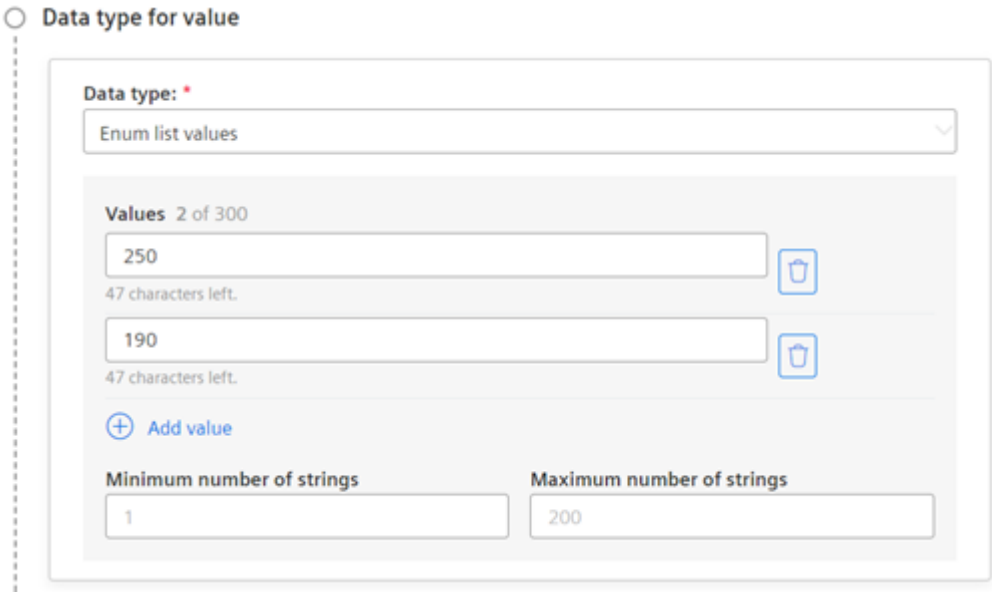
2. In the Metadata collection overview page, choose the collection of your choice, click "Add Key".

3. In the "General" section,
   - Enter the "Label" and "Description".
   - By default, "Key" field is automatically filled with the same name as "Label". Label name should be unique, the same label name cannot be re-used.
   - Enable "Mandatory" toggle button to define a key as "Required Field" in metadata input.



In the "Data type for value" section, select the "Data type" and enter the values. For more information on data types, refer to Data type values.

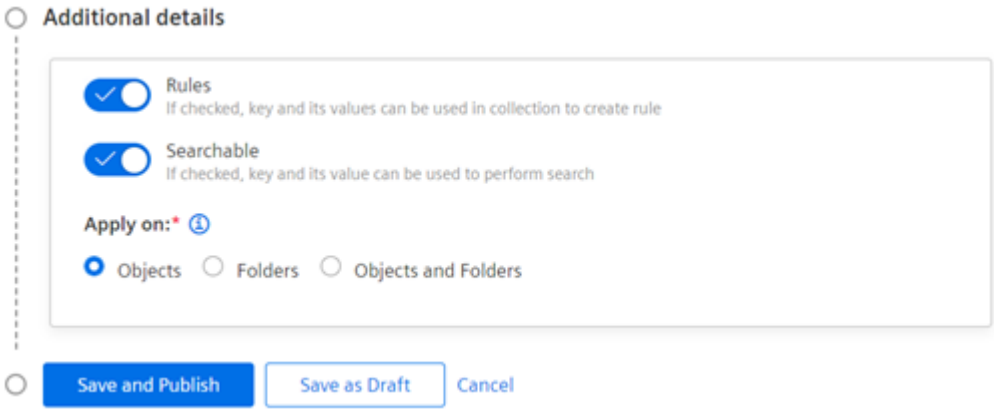
4. For example, to select "Enum values", follow these steps:
   - Select "Enum values" from the "Data type" dropdown.
   - Click "Add values".



5. In the Additional details,
   - Enable "Rules" toggle button to include this key and its values in the "Rules" section while creating a custom collection.

!!! note This toggle button is available only while creating a key for global collection, and can be enabled only for Enum values data type.

- Enable "Searchable" toggle button to perform search with authorized set of keys.
- Select "Objects", "Folders" or "Folders and Objects" radio button to apply on defined keys.



!!! note - **Objects**: Metadata is applied on objects.
- **Folders**: Metadata is applied on folders.
- **Folders and Objects**: Metadata is applied on both folders and objects. This radio button is available for only global collection.

6. Click "Save and Publish" to publish the key or click "Save as Draft" to save the key as draft.

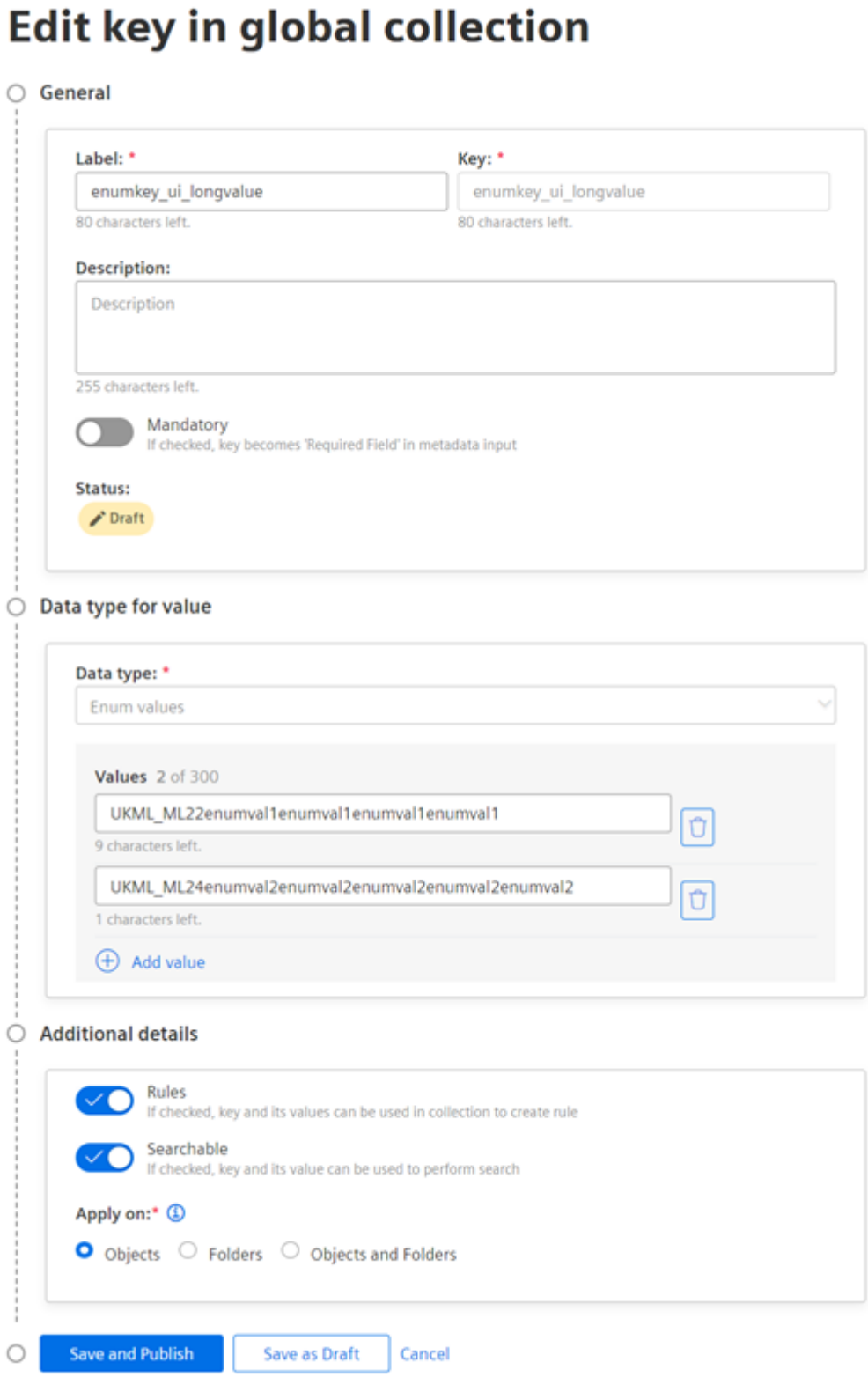
- When a key is in draft status, the defined "Data type" cannot be modified.
- When a key is in published status, the defined "Data type" and their "Values" cannot be modified.

10.8 Editing the key

The user can edit the key in both draft and published status. To edit the key, follow these steps:

1. In the left navigation, click "Metadata Management". The Metadata collection overview page is displayed.
2. Select the key to edit from global or custom collection.

3. Click and modify the details. For more information to edit the details, refer to Adding a key.

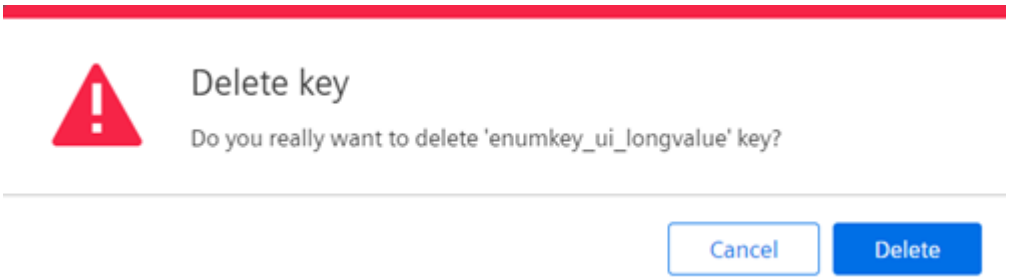


4. Click "Save and Publish" to publish the key or click "Save as Draft" to save the key as draft.

10.9 Deleting the key

It is possible to delete the key only if it is in the draft status. To delete the key, follow these steps:

1. In the left navigation, click "Metadata Management". The Metadata collection overview page is displayed.
2. Select the key to be deleted from the desired metadata collection.
3. Click .



4. Click "Delete".

10.10 Metadata Rule

Metadata Rule is a set of predefined conditions that determines the metadata associated with a custom collection that applies for the Integrated Data Lake resources. A metadata collection rule is associated with two functions:

- **Rule key**: The rule key is characterized as a metadata key with an enum data type in the global collection, facilitates the implementation of custom collection attributes in the system. By choosing a value for the rule key, users can activate specific keys in the custom collection to their unique requirements.
- **Custom collection attributes**: Custom collection attributes are only accessible when a particular value for the rule key(Enum values) is selected. This selection ensures the availability of accurate metadata information for Integrated Data Lake resources.

Rules Status

Rules are created to trigger custom metadata collection and provide metadata for the keys defined within the collection. They are applicable only when the metadata collection is defined with a rule which should be in published status. Following are the rules scenarios:

Scenarios	Create a collection	Adding Keys to collection	Keys in draft status	Keys in published status	Rule creation	Result
Scenario 1	✔	✔	✘	✔	✔	Collection is created.

Scenarios	Create a collection	Adding Keys to collection	Keys in draft status	Keys in published status	Rule creation	Result
Scenario 2	✓	✗	✗	✗	✓	Collection is saved only in draft status as there are no keys added to it.
Scenario 3	✓	✓	✓	✗	✓	Collection is saved only in draft status as none of the keys are published.
Scenario 4	✓	✓	✓	✓	✗	Collection is saved only in draft status as rule is not created.
Scenario 5	✓	✗	✗	✗	✗	Collection is saved only in draft status as keys are not published and rules are not created.