Starburst + Collibra

Data Products Integration v1.0.0

(Python)

Name	Starburst to Collibra Data Products Integration
Category	Integration
Short Description	A Jupyter notebook that automatically brings business metadata for Starburst Data Products and Data Domains into Collibra and then links that metadata to the appropriate data sets (views and/or materialized views) in Collibra that have been ingested through the Starburst driver.
Use Case	Increase the visibility and usability of Starburst Data Products in Collibra.
Target Audience	• Data Architects • Data Engineers • Business Analysts • Data Analysts • Data Scientists
Who can set it up	Integration Engineer
Target Business Functions	Any
Target Industries	Any
License Requirements	• Collibra Platform • Collibra Catalog • Starburst Enterprise
Dependencies	• Collibra Platform v2021+ • Starburst Enterprise 380-e LTS+ • Python 3.9+ • Jupyter Notebook
Developer	Collibra, Starburst
License	Binary Code License Agreement

Overview

A Starburst data product is a schema that contains one or more data sets, which are represented as views and/or materialized views in the data source where they are located. The Starburst JDBC driver in the Collibra Marketplace will automatically ingest the following metadata for each Starburst data product:

- Schema name
- Data set name
    - Name of each column in the data set
    - Datatype for each column in the data set

The Starburst JDBC driver does not ingest the following metadata for Starburst data products:

- Data product name
- Data product summary
- Data product description
- Data domain
- Catalog (the data source where the data sets are located)
- Data product owners
- Tags
- SQL query used to create each data set

This integration will do the following:

- Create assets in Collibra for each published data product in Starburst
- Create assets for each domain the data products are associated with
- Extract the aforementioned data products metadata from Starburst
- Add the metadata to the corresponding assets in Collibra
- Link the data domain asset to the appropriate data product asset
- Link the data product asset to the appropriate data sets

Requirements

To use this integration, you must meet the following requirements:

- Valid license for Collibra Integration Cloud, Collibra Catalog and Starburst Enterprise
- Collibra Integration Cloud v2021+
- Starburst Enterprise 380-e LTS+
- Python 3.9+
- Jupyter Notebook
- Collibra CMA file containing the required operating model

Constraints

This integration has the following constraints:

Starburst

- The integration is only supported for Starburst Enterprise at this time and is not supported for Starburst Galaxy.
- Only **published** data products in Starburst Enterprise will be ingested into Collibra. Data products that are in draft status will not be ingested.
- Changes made to data domain and data product assets in Collibra will not be reflected in Starburst Enterprise. Any changes that need to be made on data domains or data products should be made in Starburst Enterprise and then brought into Collibra via this integration.

Collibra

- The integration is only supported for Collibra Integration Cloud.
- Requires Collibra Edge Catalog to ingest metadata using the certified Starburst JDBC driver found on the Collibra Marketplace.
- Create Data Domain and Data Product pulled from Starburst and then enrich the metadata created from the previous bullet.
- Associated Dataset of Starburst Data Products are treated as Views and Materialized Views are represented as Database View and Table, respectively, by Collibra.
- Collibra does not currently support spaces in tags, so any spaces in tags that exist in Starburst will be replaced with underscores (‘_’) when ingested into Collibra.

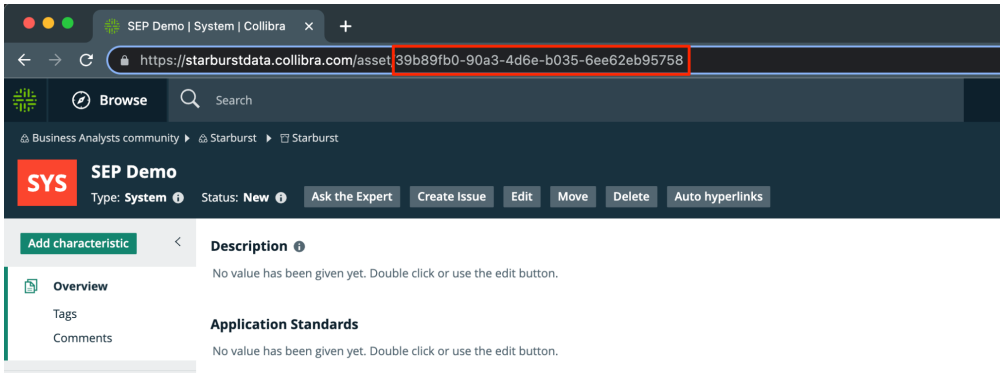
Installation

🛑 Python Version: Use Python 3.9+. Some dependencies in the Jupyter notebook might not be supported if the version is not met.

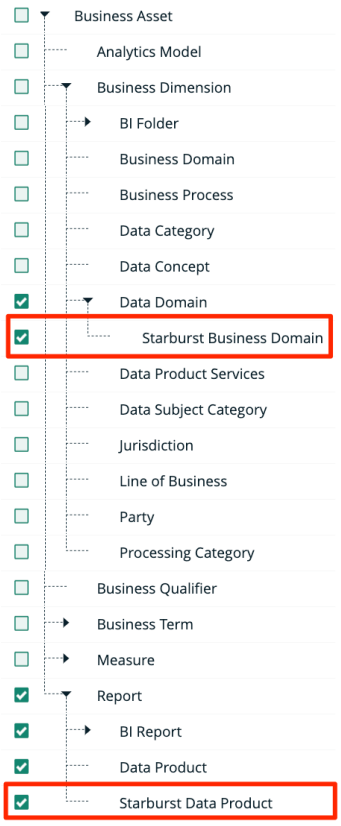
Prerequisites

- Install Python 3.9+
- Install PIP (Python packages installer)
- Download files from the starburst-collibra/data_products GitHub repo
    - requirements.txt
    - collibra_starburst_dataproducts.ipynb
    - helper.py
    - cma/starburst_collibra_integration-x.x.x.cma
- You have connected Collibra to your Starburst Enterprise cluster using the Starburst JDBC driver to pull metadata
- You have run metadata extraction for the data sources and schemas where your data products are located
- You have the URL for your Starburst Enterprise instance
- You have the URL for your Collibra Cloud instance
- Your Starburst Enterprise user must have read access to the data sources and schemas where the data products are located
    - If your Starburst Enterprise cluster has BIAC enabled, be sure to have the name of the role with the appropriate permissions
- Your Collibra Cloud user must have admin access or at least permission to create and update assets in a Domain

- The ID of the system asset you created for the connection to your Starburst Enterprise cluster. This ID can be found at the end of the URL on the catalog page for the system in the Collibra DGC.



- Add “Starburst Business Domain” and “Starburst Data Product” to the scope for the asset types that are searchable in the Data Marketplace in Collibra (Settings > Scopes > Data Marketplace).



Install Python on your preferred OS (Unix/Windows) or environment

Refer to this link to download Python for the operating system of your choice.

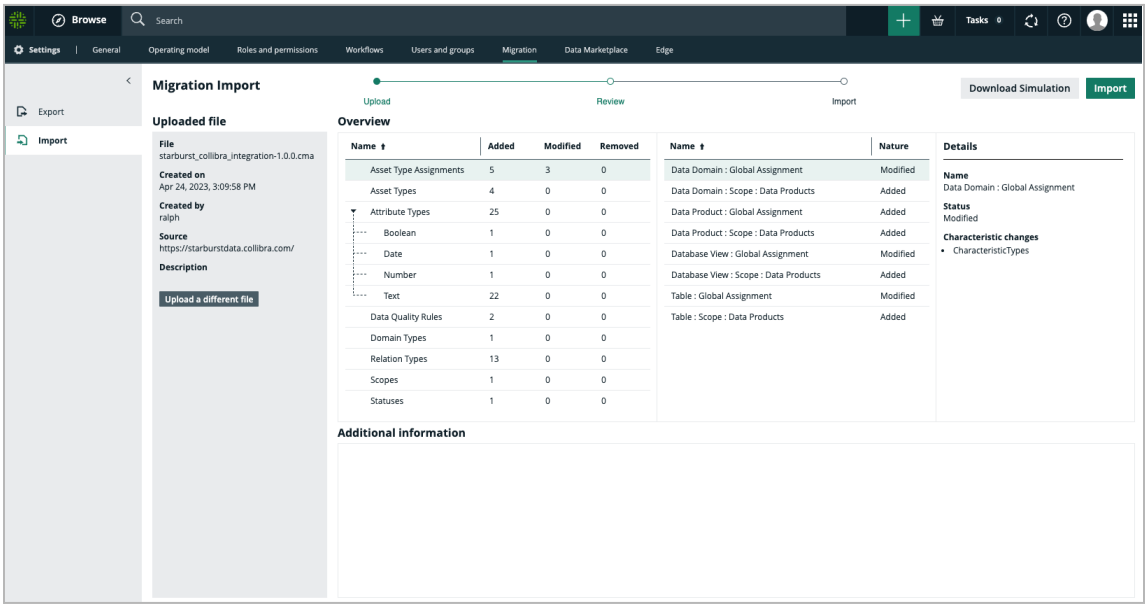
Install PIP - Python Packages Installer

PIP is a python packages installer that is required in order to install modules specified in the Jupyter notebook. Instructions for installing PIP in your preferred OS can be found here.

Upload the CMA Migration File to Install the Operating Model

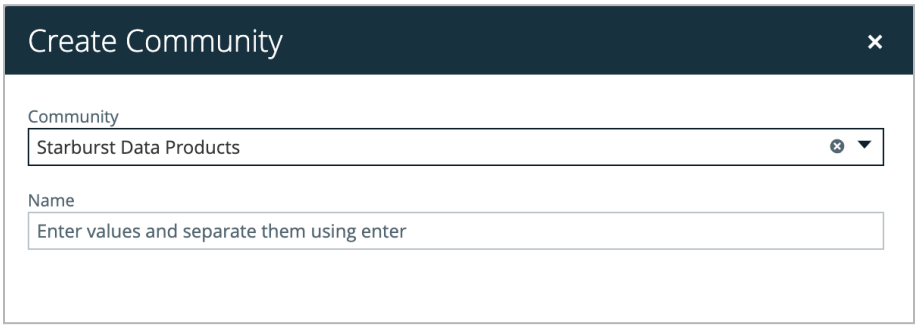
⚠ Remember to run the updated CMA file if you are upgrading from a previous version of the code. If there are any specific errors on the attribute and relation types associated with the Starburst asset types, try to delete it manually then import the CMA again.

To create the operating model dependencies, login to the Collibra DGC and then import the starburst_collibra_integration-1.0.0.cma file. Follow the Import model objects in the Collibra documentation to import the CMA file. After importing the CMA file, you should see the following summary in Collibra DGC.



Create a Starburst Data Products Community in Collibra

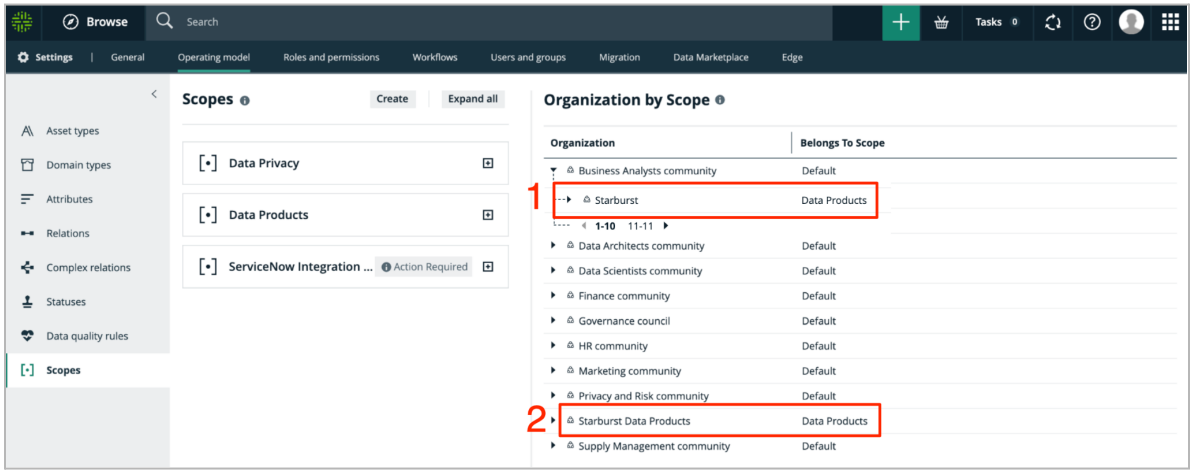
Before you can run the integration, you will need to create a community in Collibra that the Starburst data products will be associated with. Login to the Collibra DGC and follow the Create a community instructions in the Collibra documentation to create the new community. Be sure to remember the name of the community you created as this will be needed as an input to the Jupyter notebook. You can choose whatever name you want for the community.



Assign Data Products Scope to Collibra Communities

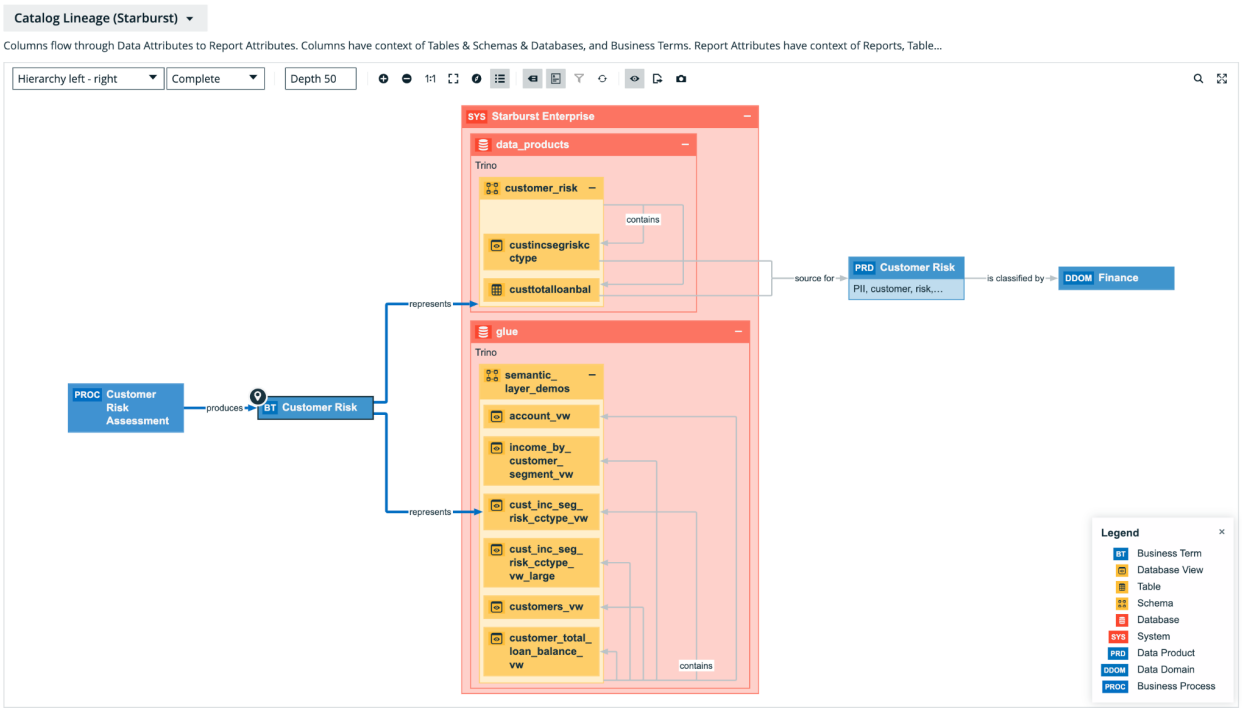
The CMA file you import will create a “Data Products” scope, which is to be used for the Starburst data products you ingest into Collibra. To ensure this scope only applies to Starburst data sources, follow the Edit a scope instructions in the Collibra documentation to assign the “Data Products” scope to the following communities:

1. The community associated with the data source connection(s) you created for your Starburst Enterprise cluster.
2. The Starburst Data Products community you created in the previous section of this guide.



Add Starburst Catalog Lineage View (optional)

A lineage view has been created that shows the physical data assets as well as the domain, business term(s) and business process(es) that are associated with a Starburst data product. The JSON for this view is available in the GitHub repository (starburst_lineage_view.json) While this lineage view is not required to run the integration, it does provide a simplified way to see how the relationships that exist around a data product. Below is an example of this lineage view.



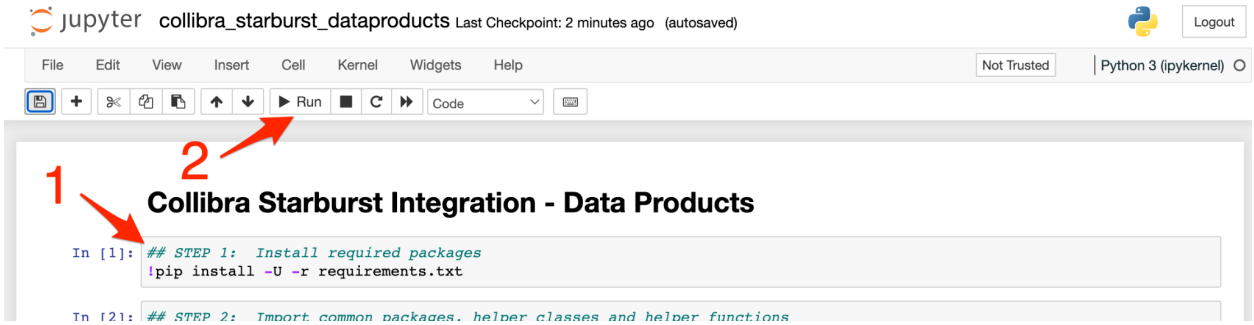
Configuration

After satisfying the prerequisites and completing the installation instructions, you are ready to run the integration. The Jupyter notebook is self contained and imports all of the required packages for you. No additional configuration is required.

Usage

The following instructions will walk you through how to run the integration.

- Open the Jupyter notebook
- Select the first cell (1) and click **run** (2).



- After running the cell, you will automatically be moved to the next cell.
- Click **run** for each cell until you reach the end.
    - If you encounter an error when running a cell, correct the error and run the cell again.

Below are guidelines for cells that require input from you or that produce output.

- **## STEP 3 - This cell will prompt you for your Starburst connection information:**
    - Enter the full **URL** of your Starburst Enterprise environment, including the HTTP protocol (example: https://mycluster.mydomain.com/)
    - Enter the username to log into Starburst Enterprise with
    - Enter the name of a role your username has access to
        - This is only required if BIAC is enabled in your Starburst Enterprise cluster
        - Leave this blank if BIAC is not enabled in your cluster
    - Enter the password associated with the username you provided

- **## STEP 4 - This cell will prompt you for your Collibra connection information:**
    - Enter the full **URL** of your Collibra Cloud environment, including the HTTP protocol (example: https://mycatalog.collibra.com/)

    - Enter the **username** to log into Collibra Cloud with
    - Enter the **password** associated with the username you provided

- **## STEP 5 - This cell will prompt you for the following information:**
    - A temporary directory where files will be created for the JSON payloads that are needed as input to the Collibra API calls
    - The name of the **Collibra community** where the data domains and data products will be loaded
    - The **system ID** of the data source schema where the data sets for the Starburst data products are located

- **## STEP 6 - This cell is used to run the integration. If everything works correctly, you should see the following output:**

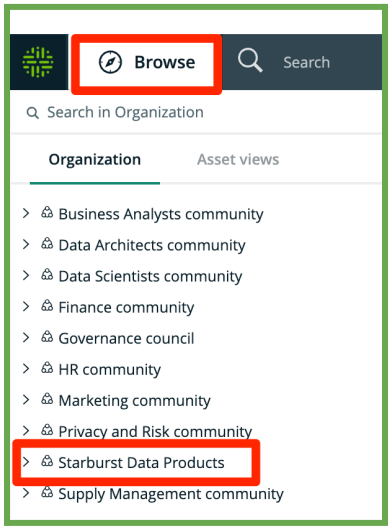
```
About to pull/ingest all the data domains from Starbur to
Collibra.

About to pull/ingest all the published data products from
Starburst to Collibra.

Completed.
```

After running all of the cells in the Jupyter notebook, you can use the starburst_collibra_integrations.xlsx file in the GitHub repository as a guideline to confirm all of the expected assets have been added to your Collibra environment.

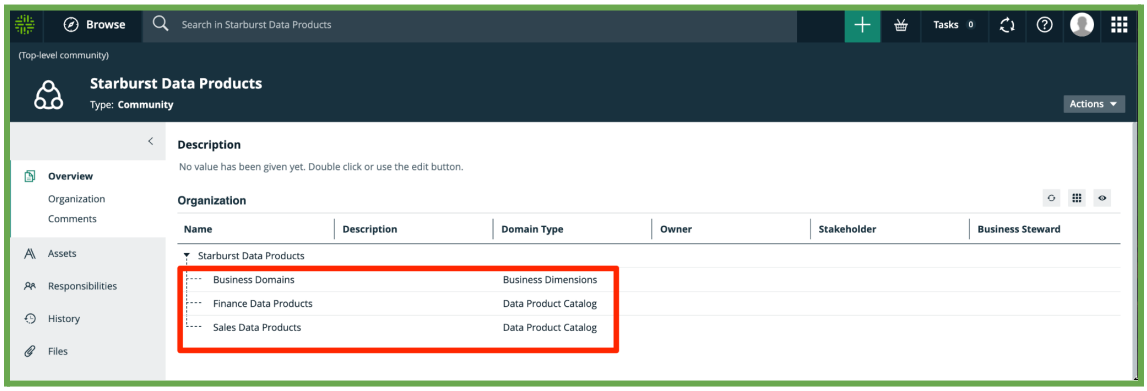
Delete Integration Artifacts

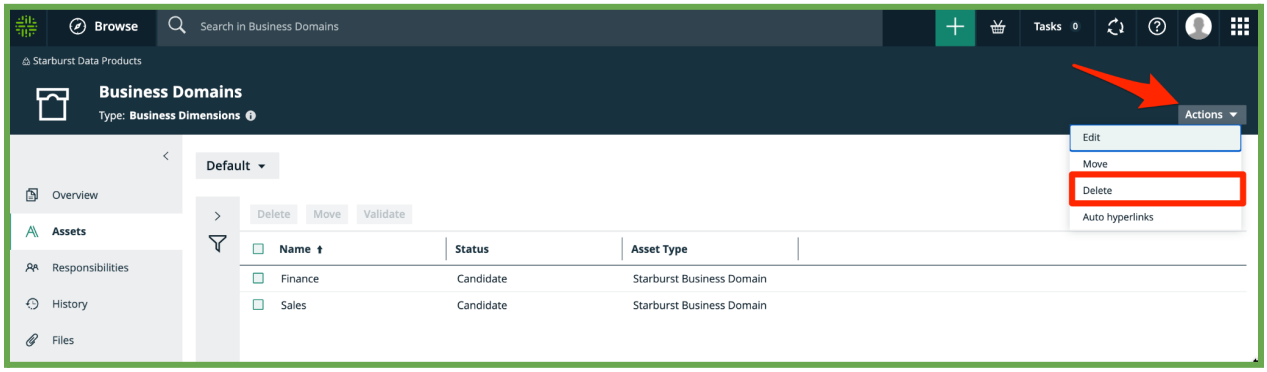
Follow the steps below to remove the artifacts that are created by the integration in Collibra.

1. In your Collibra DGC environment, go to **Browse** and then click on the name of the community you created for your Starburst Data products.



2. Delete “Business Domains” as well as any other items in the community by clicking on the name of the item and then selecting **Delete** from the **Actions** menu.

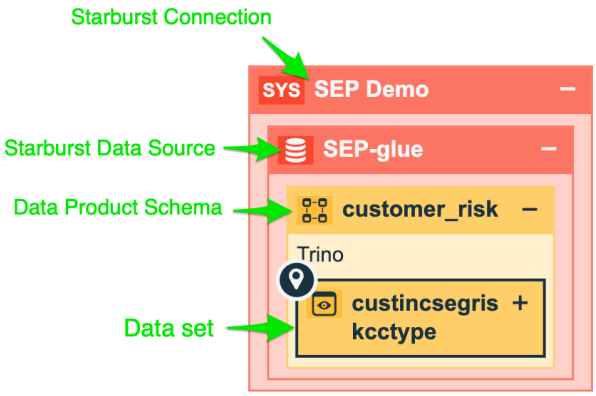




3. All artifacts of the integration have now been cleared out.

Functional Design

Before you run the integration, you will need to have already ingested metadata into Collibra from Starburst for the data sources (“catalogs”) and schemas where your data products are located. For example, the image below depicts a data set (view) and data product schema that were automatically ingested using the certified Starburst JDBC driver.



When you run the integration, you will iterate through each cell in the Jupyter notebook starting at the top of the notebook and working your way to the bottom. Provided below is a high-level overview of the operations that are performed by running the Jupyter notebook.

1. Prompt user for Starburst and Collibra connection information (URL, username, password)

2. Prompt user for the type of pull they want to perform
    a. *Valid values*: data domains, data products

3. Prompt user for the Collibra community and system where the data products and data domains will be added

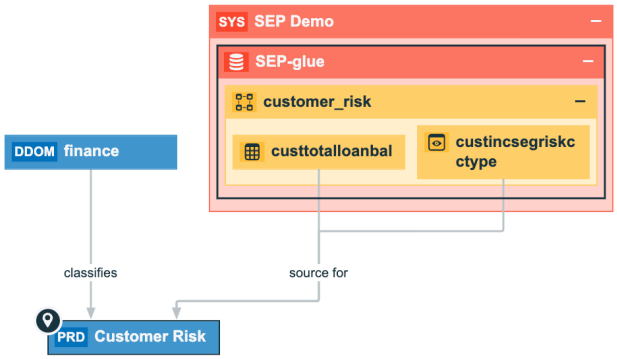
4. Retrieve all data domains and their metadata from Starburst and create a domain asset for each one in Collibra

5. Retrieve all “published” data products and their from Starburst and create a data product asset for each one in Collibra

6. Update the views/materialized views associated with each data product in Collibra with the definition SQL queries from Starburst

7. Link the data domain asset and data set(s) (views/materialized views) to the appropriate data product asset in Collibra

After running the integration, you will see a data product asset in Collibra for each published data product that exists in your Starburst Enterprise environment. You will also see a data domain asset in Collibra for each domain that was extracted from the published data products in Starburst Enterprise. Each data product asset in Collibra will be linked to the appropriate data domain asset and the data set(s) that are part of the data product. The example below shows the result of ingesting the **Customer Risk** data product from Starburst Enterprise into Collibra using this integration.



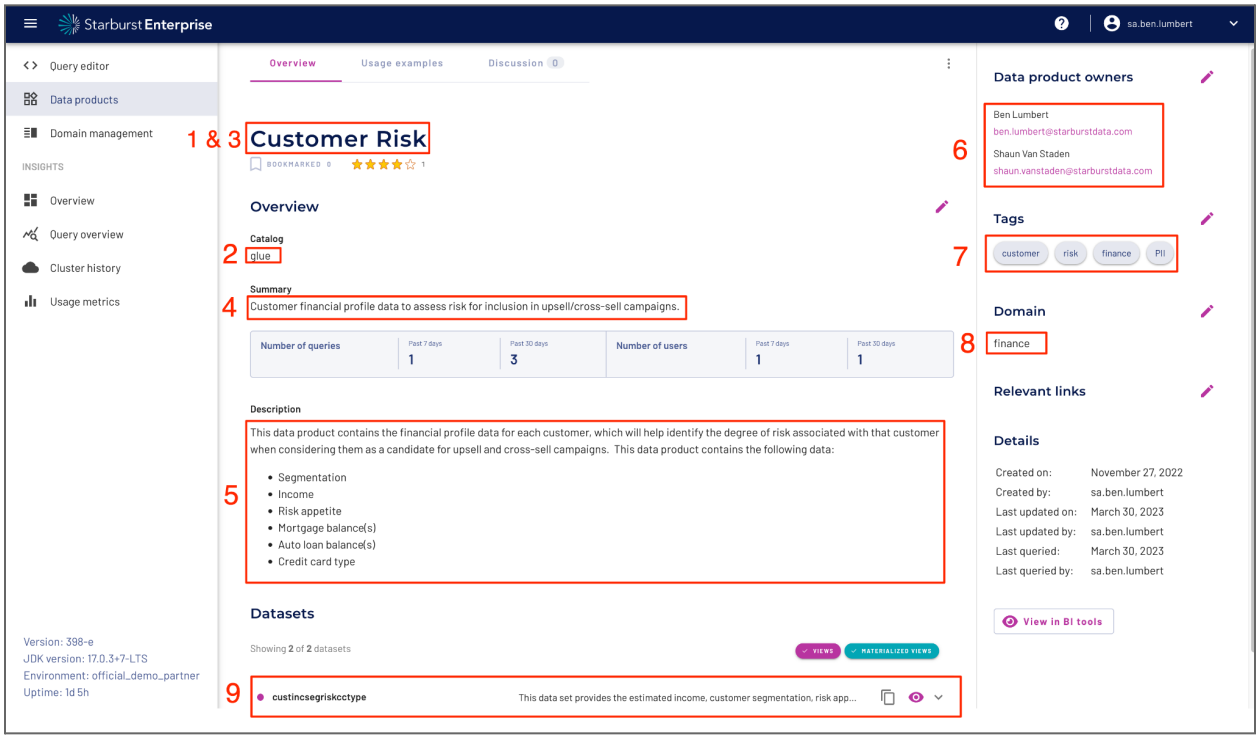
- These assets were created through the Starburst JDBC driver and were not created through this integration:
    - SEP Demo (*system*)
    - SEP-glue **(***data source*)
    - customer_risk (*schema*)
    - custtotalloanbal (*materialized view, represented as a table*)
    - custincsegrickcctype (*view*)

- These assets were created after running this integration:
    - finance (data domain)
    - Customer Risk (data product)

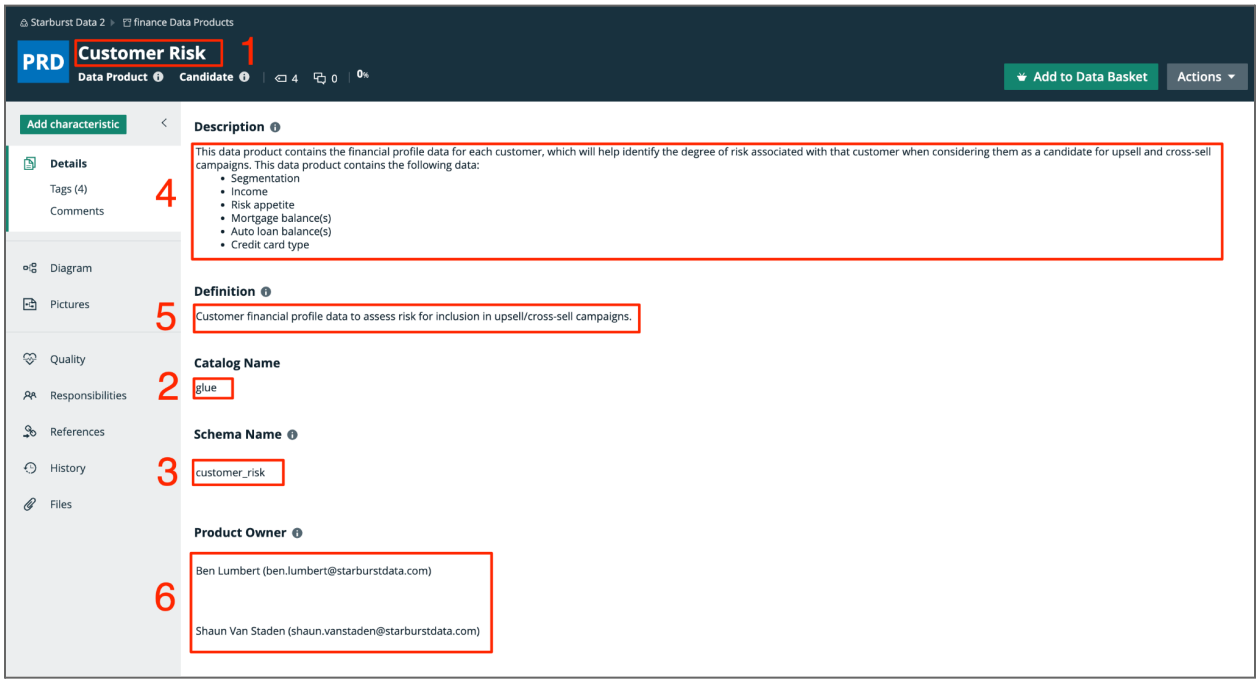
The table below shows the mapping between the metadata fields for Starburst Enterprise data products and the corresponding fields on the catalog page for the data product in Collibra.

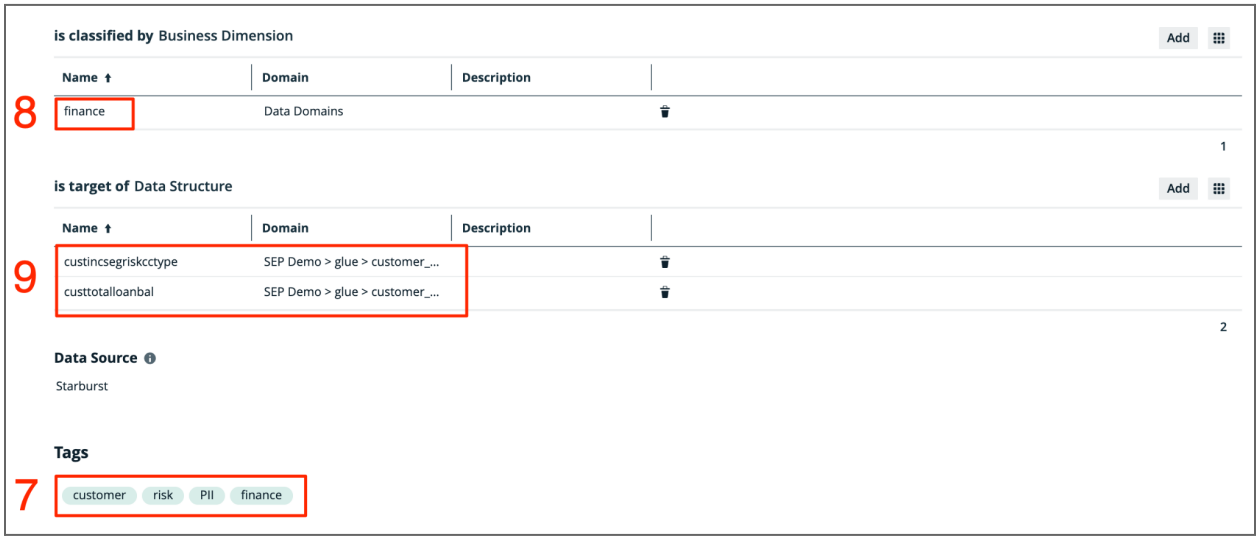
#	Starburst Enterprise	Collibra
1	Name	Name
2	Catalog	Catalog Name
3	*The schema name is automatically generated from the name of the data product.*	Schema Name
4	Summary	Definition
5	Description	Description
6	Data product owners	Product Owner
7	Tags	Tags
8	Domain	is classified by Business Dimension
9	Datasets	Is target of Data Structure

Starburst Enterprise



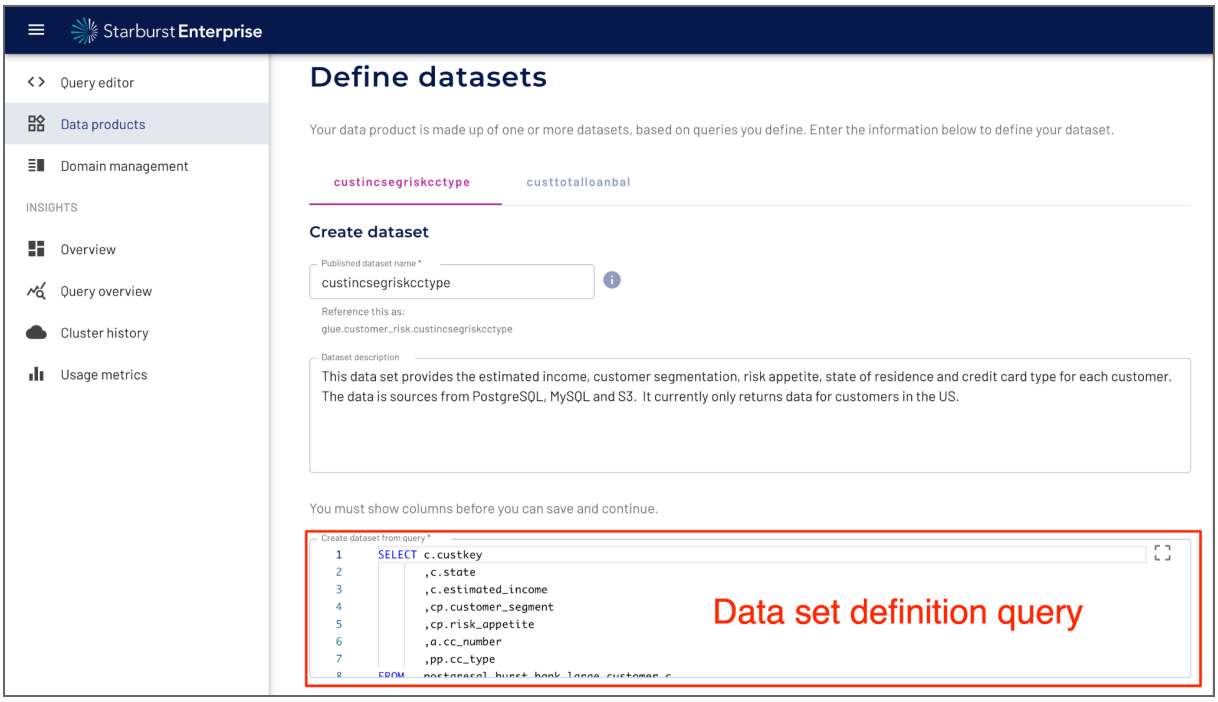
Collibra Cloud



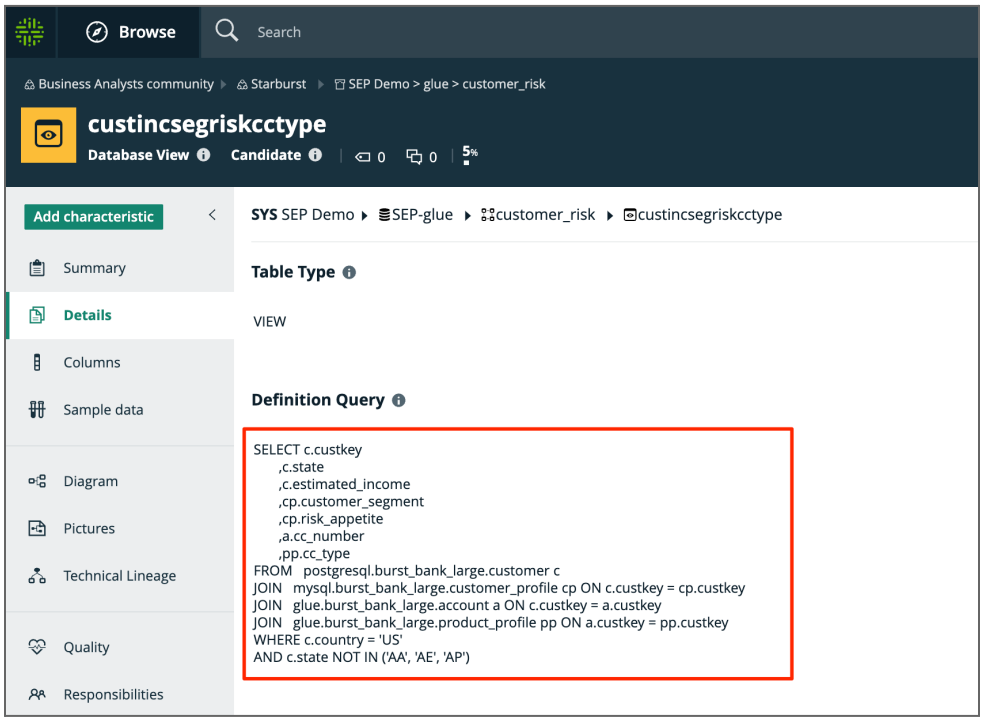


Additionally, the definition query for each data set in the data product will be extracted from Starburst Enterprise and added to the **Definition Query** field on the **Details** page for the corresponding view/materialized view in Collibra Cloud.

Starburst Enterprise



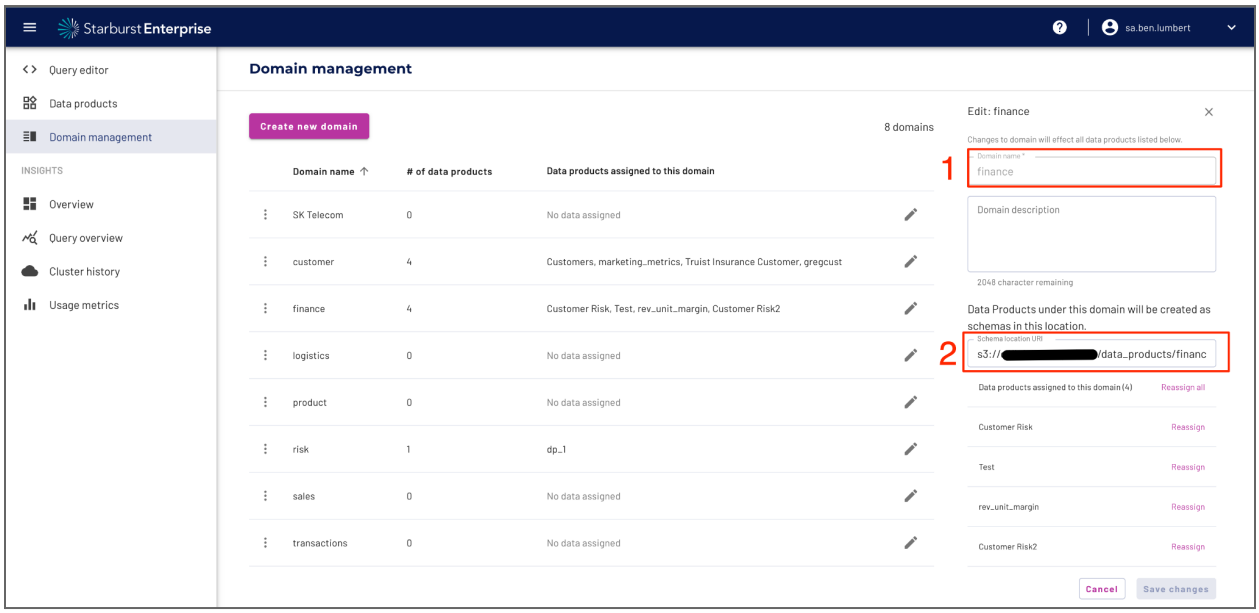
Collibra Cloud



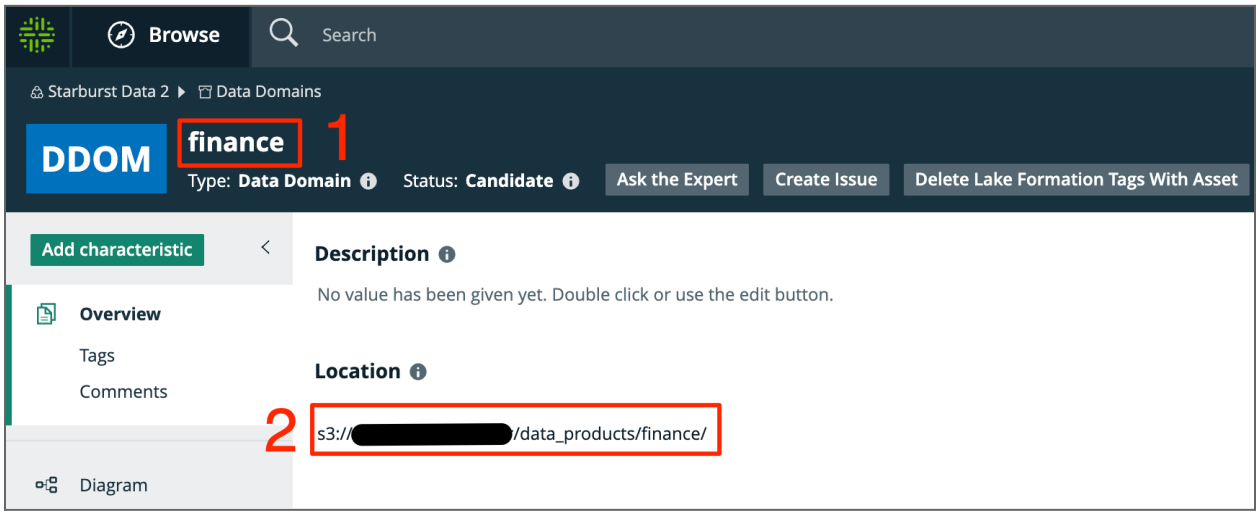
The table below shows the mapping between the metadata fields for Starburst Enterprise domains and the corresponding fields on the catalog page for the data domain in Collibra.

#	Starburst Enterprise	Collibra
1	Domain Name	Name
2	Schema location URI	Location

Starburst Enterprise

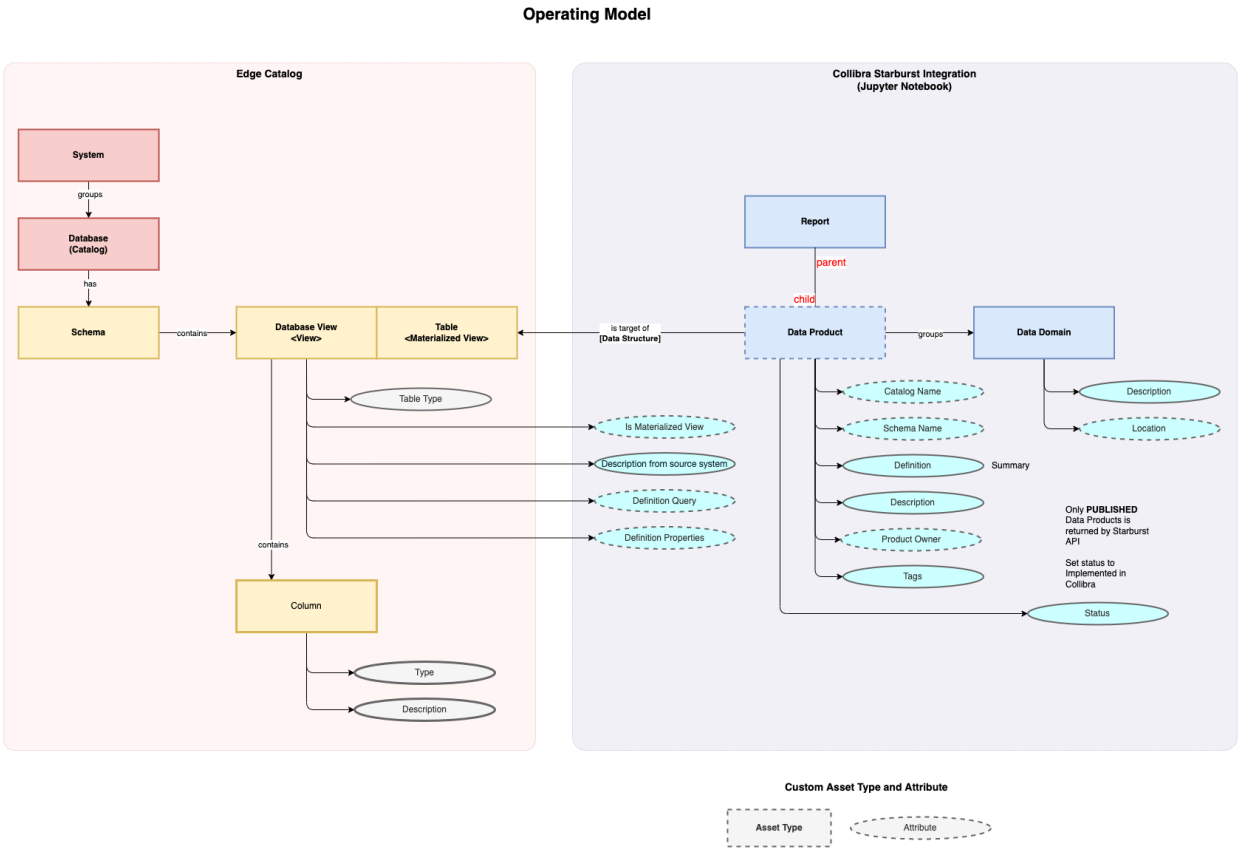


Collibra Cloud

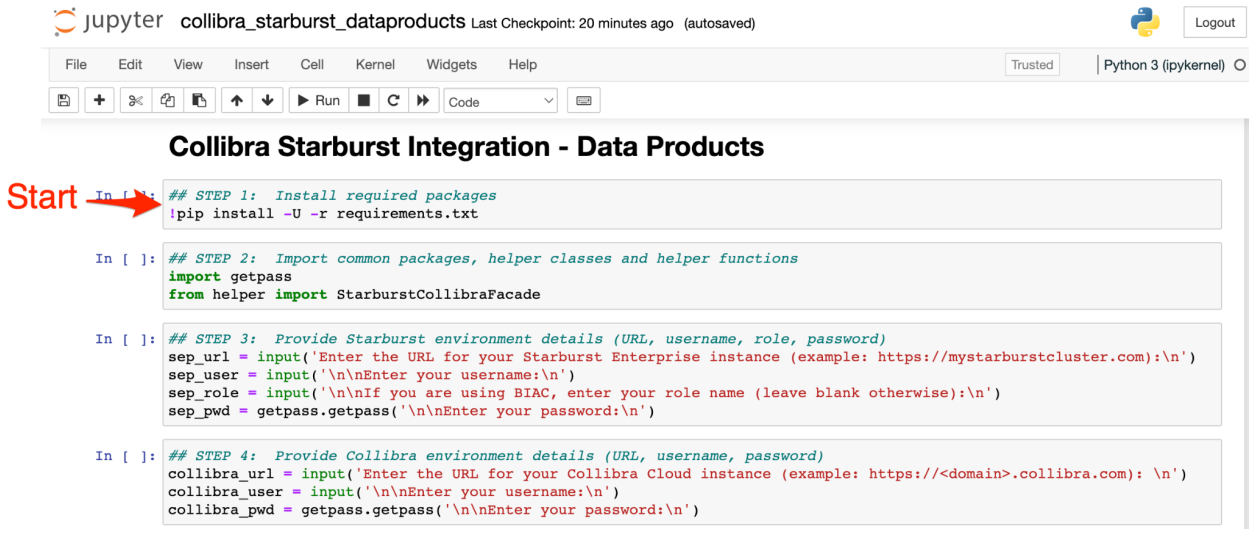


Technical Design

Below is the operating model for this integration. On the left are the asset types and attributes that are ingested by the Edge Catalog. On the right are the metadata that are ingested and linked to the Edge Catalog metadata. To be able to create the custom asset types, attribute types and the Starburst scope on the environment, make sure that you imported the provided CMA file.



To use the integration, you will run each cell in the Jupyter notebook in sequence, starting at the first cell at the top and working your way down.



Provided below is an overview of the operations that are performed in each of the cells you see in the notebook.

STEP 1: Install required packages

- Installs common Python packages and Collibra Python packages in your local environment

STEP 2: Import common packages, helper classes and helper functions

- Imports common Python packages
- Imports Collibra Python packages
- Imports helper classes for Starburst and Collibra
- Imports helper functions

STEP 3: Provide Starburst environment details (URL, username, role, password)

- Prompts user for Starburst information
    - Starburst Enterprise URL
    - Starburst username
    - Starburst role (only required if BIAC is enabled)
    - Starburst password

STEP 4: Provide Collibra environment details (URL, username, password)

- Prompts user for Collibra information
    - Collibra Cloud URL
    - Collibra username
    - Collibra password

STEP 5: Provide temp directory, Collibra community name and system ID

- *Temp directory* - This will be used by the notebook to generate JSON payloads for the Collibra API calls
    - This directory will automatically be created for you in the location where you run the Jupyter notebook from
- *Collibra Community* - The name of the Collibra community where the data products and data domains from Starburst will be added
- *System ID* - The ID of the Starburst Enterprise system in Collibra where the data sets for the data products in Starburst are cataloged

STEP 6: Import Starburst data domains and data products into Collibra

- Instantiates the facade class in order to use the classes and functions defined in helper.py
- Calls the function to extract all data domains from Starburst and import them into Collibra
- Calls the function to extract all “published” data products from Starburst and import them into Collibra

The Jupyter notebook requires 2 additional files for it to operate correctly. These files need to be located in the same directory as the Jupyter notebook.

requirements.txt

This file contains the Python packages that need to be installed in your local environment in order for the Jupyter notebook to operate correctly. These packages will installed for you when you run the first cell (“##STEP 1: Install required packages”) in the notebook.

helper.py

This file ensures the appropriate packages are imported into your local environment and defines the helper classes and functions required for the Jupyter notebook to operate correctly. The packages will be imported to your local environment and the helper classes and functions will be defined for you when you run the second cell (“## STEP 2: Import common packages, helper classes and helper functions”) in the notebook.

Provided below is an overview of what is contained in this file:

- Import common Python packages
- Import Collibra Python packages
- Define a reusable *CollibraService* class with functions that:
    - Import assets
    - Delete files
    - Read files
    - Write files
- Defines a reusable *ImportCommand* class to generate Import API requests in JSON format
- Defines a reusable *StarburstService* class with functions that:

    - Execute an API call to Starburst to retrieve all data domains
    - Execute an API call to Starburst to retrieve all tags for a data product
    - Execute an API call to Starburst to retrieve all of the metadata for a data product
    - Execute an API call to Starburst to retrieve the list of all ‘published’ data products in Starburst
- Defines a reusable *StarburstCollibraFacade* class
    - This class is used in the Jupyter notebook to call the other classes defined in this file
    - It also defines the following functions:
        - Query and import data domains
            - Pull all domains from Starburst via an API call
            - Create *Starburst Domain* Import Command and then send to Collibra Import API to create as assets
        - Update data product views
            - Link the data products to their respective views and/or materialized views in Collibra
        - Query and import data products
            - Export all “published” data products from Starburst via an API call
            - Create *Starburst Data Product* Import Command and then send to Collibra Import API to create as assets
        - Update data product view columns
            - Update descriptions in Collibra for columns in the views/materialized views that are associated with the data products

Release History

- Version 1.0.0 - Initial release of this integration.

What's Next?

The initial release of this integration focuses on extracting data domains and data products from Starburst Enterprise, then importing them as assets into Collibra and connecting them to the appropriate data sets (views and/or materialized views). The next phase of this integration will focus on creating a bi-directional sharing of data product and data domain metadata between Starburst Enterprise and Collibra. In that phase, Collibra users will be able to initiate the creation of data products from within Collibra Cloud.

A Collibra user will start the process by providing the following information:

- Select the view(s) and/or materialized view(s) in Collibra that will be used as data sets for the data product.
- Provide metadata (including the data domain) about the data product
- Provide notes regarding any operations (transformations, aggregations, etc) that need to be applied to the data sets

A designated admin in Collibra will be notified when the request has been made. Some of the information above will be used to create a draft data product in Starburst Enterprise. The admin will then use the provided notes to finish creating the data product in Starburst. Once that has been done, the metadata for the newly created data product will be automatically exported from Starburst and imported into Collibra using the functionality provided by the initial release of this integration.