modak

Modak Nabu™ 2.6

USER MANUAL

Contents

1 Introduction

1.1 Purpose

The purpose of this user manual is to provide usage instructions for all features in Modak Nabu. This includes a description of the system functionalities, capabilities, step-by-step procedures for access and usage.

1.2 Why Modak Nabu™?

Modak Nabu™ is an integrated data engineering product used for exploring, combining, cleaning, and transforming raw data into curated datasets. Modak Nabu converges data discovery, data ingestion, data preparation, data catalogue, data unification and data profiling into a single enterprise platform. With active metadata as primary driver, Modak Nabu automates repetitive tasks and thus enables 4X-10X acceleration in the data journey, both on-premises and on-cloud.

1.3 Features

Smart data discovery: Data Spiders automate the acquisition of technical metadata, for structured, semi-structured and unstructured data sources – whether on-prem or on-cloud. Examples of the captured metadata for structured data include details on schemas, tables, rows, and columns for each of the data sources. Examples for unstructured and semi-structured include file locations, file size and number.

Automated data ingestion: Using a metadata driven approach, ingestion pipelines are generated in an automated way for industry standard tools to ingest data into modern data platforms.

Data profiling: Automated profiling helps to quickly assess data quality and evaluate the suitability of using the data for projects. The profile of a table captures parameters like data types, range of values, number of null values, max and min values, and frequency distribution of values for each column in the table.

Data fingerprinting: The data fingerprint is a unique identifier for a set of data (the set of values in a column from a relational data source). The fingerprints are used to identify similarities between two or more data sets. Data fingerprinting is extensively utilized in Modak Nabu to help in discovery of similar columns, data unification, standardization of column names.

Active metadata catalog: Metadata captured is stored in Modak Nabu's rich data catalog. This helps to categorize enterprise data making it easily accessible to large number of users. The metadata stored also helps to automate and accelerate repetitive data engineering processes. Metadata captured includes technical, operational, business, and social metadata.

Search and data exploration: There are multiple ways for users to explore the data using Modak Nabu after the data that has been ingested, profiled, and indexed. The users can search for data, understand profile of a table and view knowledge graph to discover relationships.

Executive dashboard: The executive dashboard provides a view of key metrics related to crawling, ingestion, and profiling of data. The dashboard can be customized based on tags to provide custom views for different stakeholders.

Monitoring dashboard: Pipelines scheduled in Modak Nabu™ can be monitored through a unified interface that provides status, details about issues in pipelines and functionality to retry failed tables and files.

Multi-tenancy: Role-based and fine-grained access control for resources in Modak Nabu (e.g., pipelines, data domains, compute engines)

Enhanced security and governance: Governance on access to data domains, audit trail of user activities and use of secured vault for storing credentials.

1.4 Glossary of terms

Term	Description
Data Connection	Any existing repository of data in an enterprise. It can be a relational database, file system or native cloud data service.
Data domain	Logical grouping of data connections and selected schemas/tables/views.
Kosh	Metadata repository of Modak Nabu.
Crawl	Process of capturing metadata from tables/files in a data connection and storing them in Kosh.
Ingest	Process of moving data from a source data connection to a target data connection.
Profile	Profiling captures statistics including range of values in a column, frequency distribution, minimum and maximum values, cardinality, and number of null values.
Index	Indexing of metadata and data using Apache Solr makes it searchable through Nabu search interface
Entity	An entity is a collection of similar columns in a data domain. Data within these columns is searchable using Modak Nabu Search.
Facet	A collection of similar columns in a data domain. Difference between an entity and a facet is that the data in the columns grouped together under a facet is not searchable.
Filter	Filter is the parameter on which search results of an entity can be filtered.
Synonym	Alias for an entity. Search results of a synonym are the same as that of the entity for which it is an alias
Fieldstore	Fieldstore is an attribute of the entity which can be displayed in search results for an entity.

2 Getting Started

2.1 Logging in

On entering the Modak Nabu URL in browser, the user is taken to the login screen, where access credentials provided by Modak Nabu admin are entered. Modak Nabu integrates with an enterprise's LDAP or AD system, so the login credentials can be existing LDAP/AD ID. In case you are unable to login, please check with the Modak Nabu admin if you have been provided access.

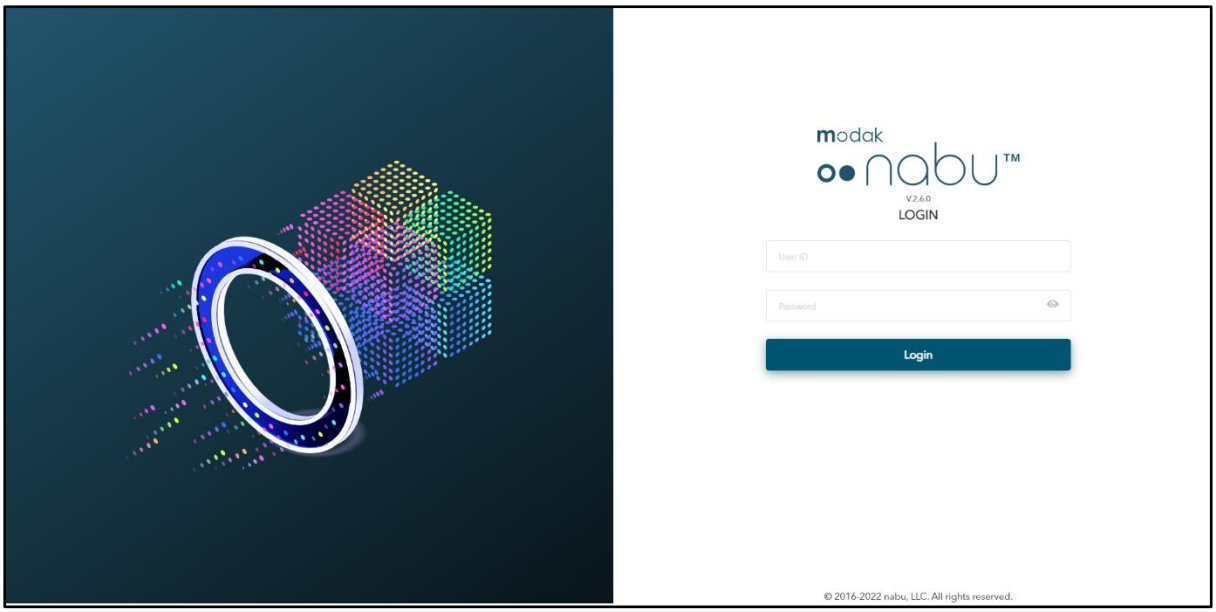


Figure 1: Login Page

2.2 Application Menu

On logging in, the user sees the Modak Nabu search screen, where search items can be entered in the search bar.

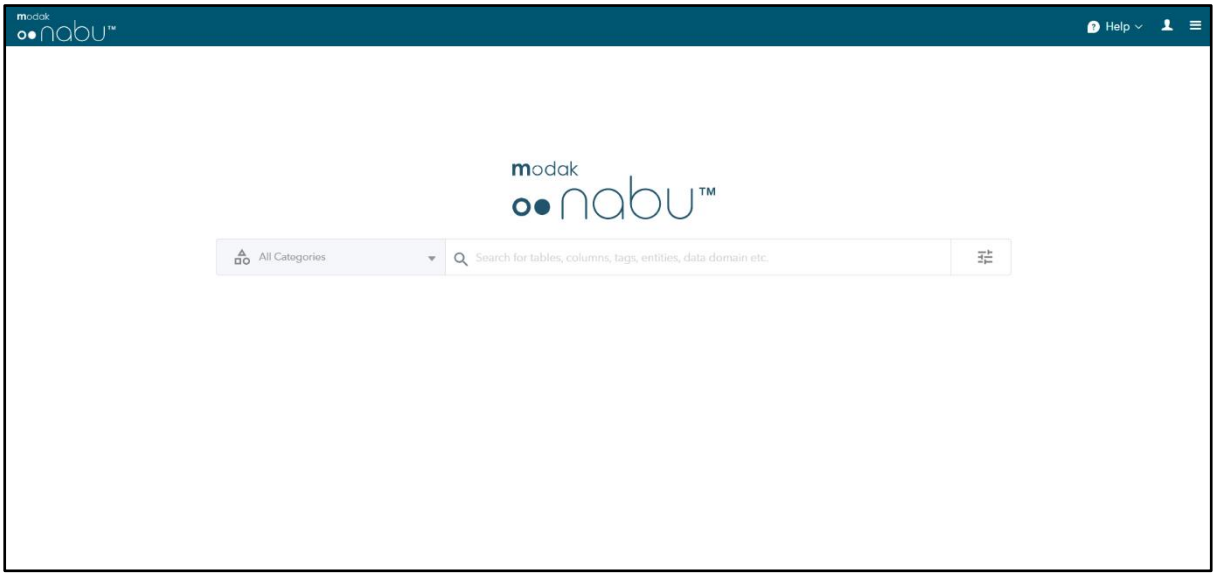


Figure 2: Modak Nabu Search Page

At the top right of the screen, on clicking the menu icon, the application menu is shown. The application menu lists the different modules of Modak Nabu.

1. Data Connections
2. Pipelines
3. Data domains
4. Data Catalogue
5. Dashboards
    a. Monitoring Dashboard

   b. Executive Dashboard
6. Access Management
   a. Roles
   b. Data Access
   c. Data Domain Group
7. Credentials
8. Compute Engines

The usage of these modules is described in detail in subsequent sections.

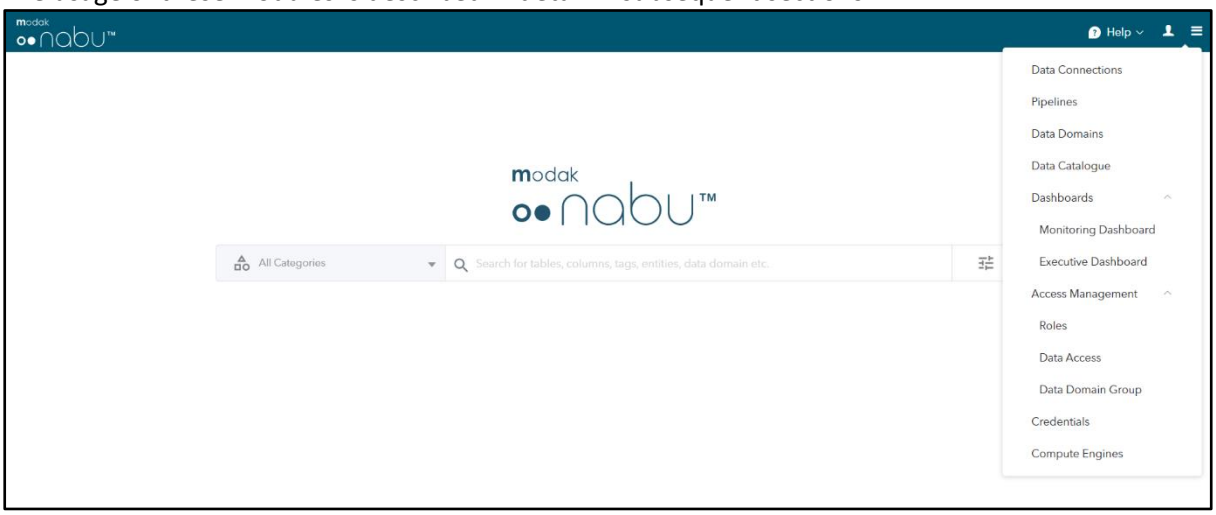


Figure 3: Modak Nabu Menu

3 Data Connections

You can access Data Connections by clicking on the menu button on the top right of the page.

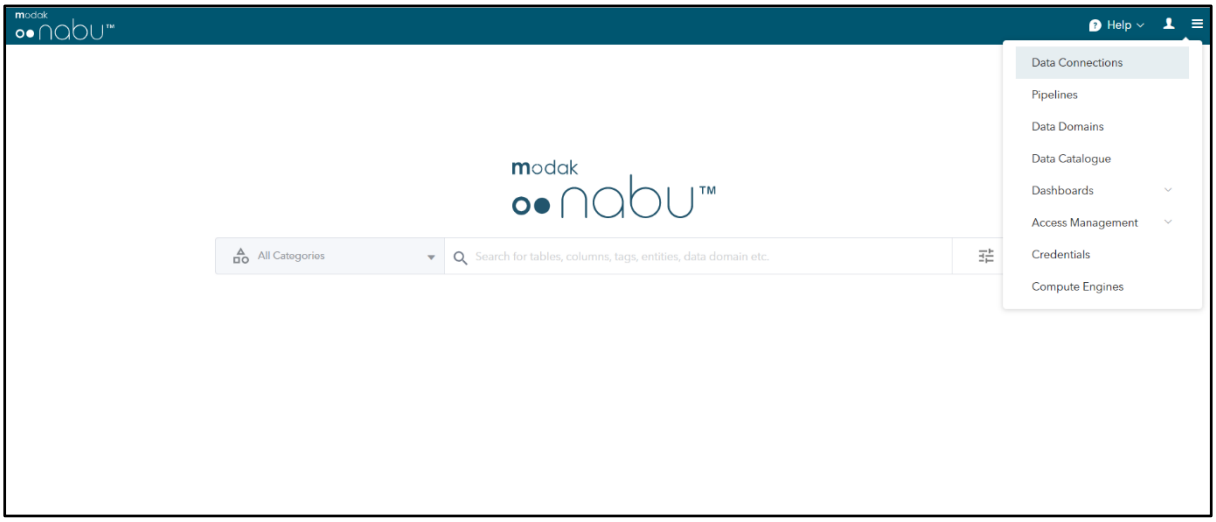


Figure 4: Data Connections

'Data Connection' in Modak Nabu refers to any repository of data in an enterprise. The data can be present in databases, files systems or cloud services.

The 'Data Connections' section is used to register das, add credentials to access those data connections, schedule ingestion from one data connection to another and monitor the ingestion between data connections.

3.1 Data Connections Landing Page

A data connection in Modak Nabu is any repository of data in an enterprise. These repositories can contain structured, semi-structured or unstructured data.

You can navigate to the 'Data Connections' module by clicking the menu icon " " at top right of the screen. Under 'Data Connections' option, select 'Data Connections'

The 'Data Connections' module is used to add/modify/delete data connections in Modak Nabu. Metadata from the added data connections is captured by 'Data Spiders' at a schedule specified by the user. This metadata is stored in the Modak Nabu's metadata repository called 'Kosh'.

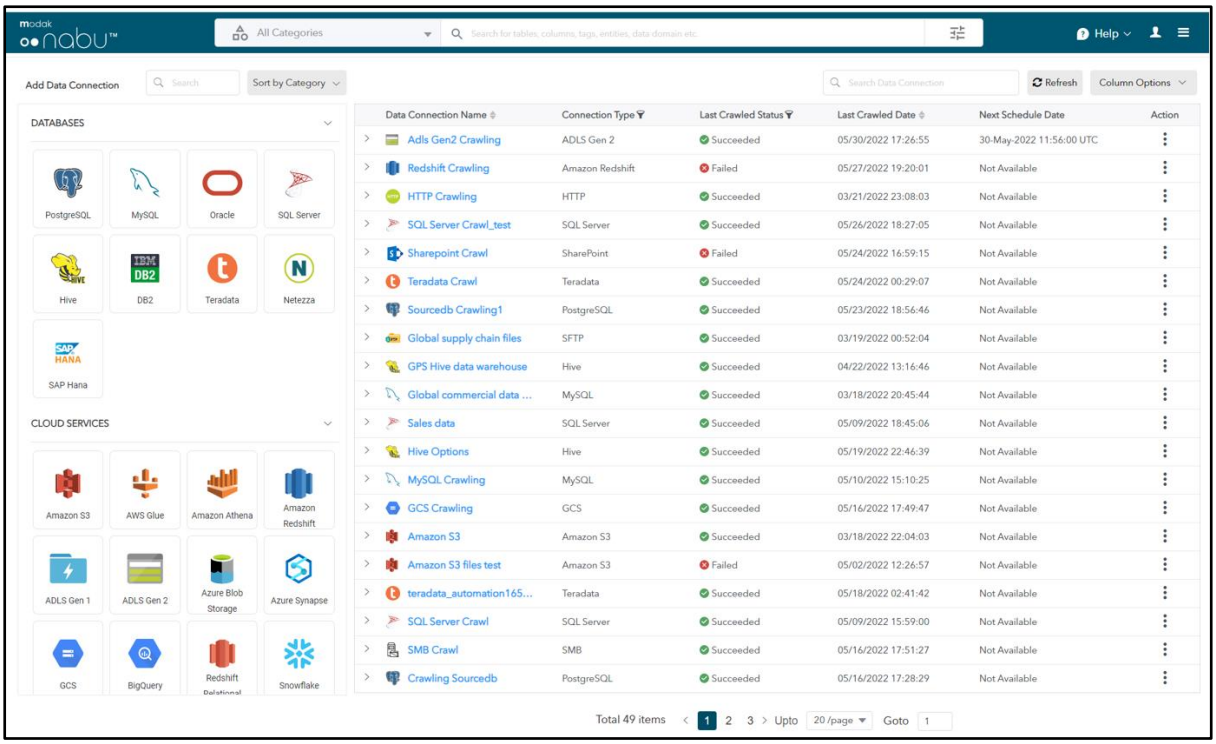


Figure 5: Data Connections Landing Page

The landing page of the data connection provides the user with following features:

In the left panel of the screen, it shows the icons for different data connection types that can be added to Modak Nabu. These include structured, semi-structured, and unstructured databases. This panel includes the following features.

Search for data connection type: This feature enables the user to search for the data connection type.

Sort by category/A to Z format: It helps the user to sort the data connections by category/A to Z.

On the right-hand side, the landing page shows a table which comprises the following list of columns.

Data Connection Name: This column displays the list of all the existing data connection names. This column is expandable and if user expanded it shows a crawling frequency information for the particular data connection This column is sortable.

Connection Type: This column shows the data connection type. The user can filter the results based on the connection type they want by using the filter option. Simply click on the filter icon, select the desired connection type from the menu, then click apply to view it and table results will get refreshed with the applied connection type.

Last Crawled Status: This column displays the outcomes of the data connections that have been crawled. They could be succeeded, running, or failed. Each state is indicated by a distinct colour and symbols, such as succeeded in green, running in blue, and failed in red, and the user can also check the needed status of the crawled data connection by using the filter option, simply click on the filter icon and select the required status from the menu and click on the apply button to view the filtered results.

Last Crawled Date: It shows the last crawled date and time for the data connection. This column is sortable.

Next schedule Date: It shows the next schedule date and time for crawling the data connection. This column is sortable.

Action: The action column provides the user to select any option (edit, duplicate, schedule, or delete) from the drop down. Based on the option selected the respective action will be applied.

Refresh: This indicator assists the user in refreshing the table. When user clicks on the refresh button, the results are refreshed and updated.

Column options: This option enables the user to select the required columns to display on the table. Some columns are selected by default, while others are not. For example: data connection name, connection type, last crawled status, last crawled date, next schedule date, and action are all selected by default. According to user preferences, the user can select/deselect the option from the menu.

Search data connection: This feature helps users to search for the required data connection. simply by entering the name of data connection in the search box.

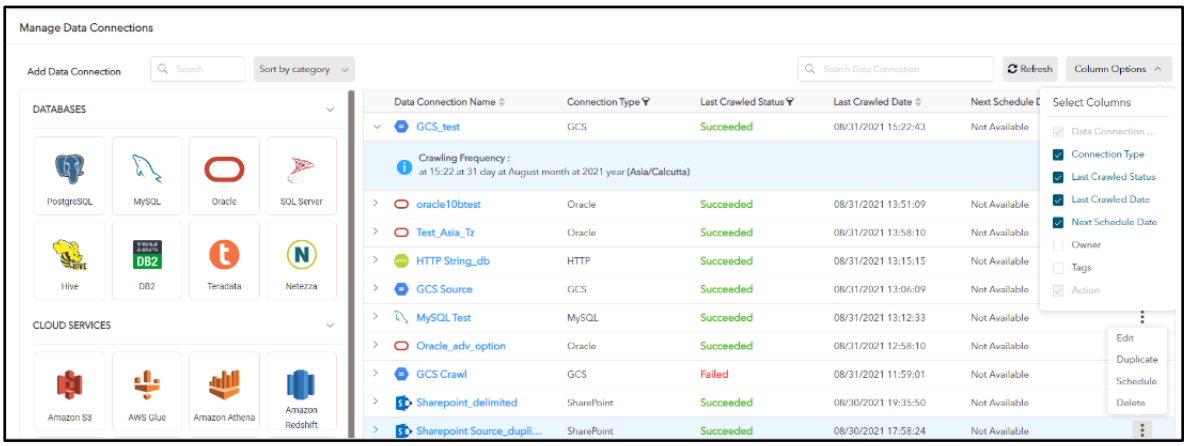


Figure 6 : Data Connections

3.1.1 DATABASES

A new field schema Name is added for the below data connections.

- Postgres, Oracle, SQL server, DB2, Netezza, AWS Glue, Amazon Redshift, Azure Synapse, Big query.

The below data connections are added as part of Nabu 2.6.

Snowflake, SAP Hana, Documentum, SharePoint Subsite, Redshift Relational

3.1.1.1 Add Postgres Data Connection

To add a Postgres data connection, click on the Postgres icon from the left pane of the Data Connections' screen. The Add Postgres Data Connection form opens, and the user is required to provide the below details.

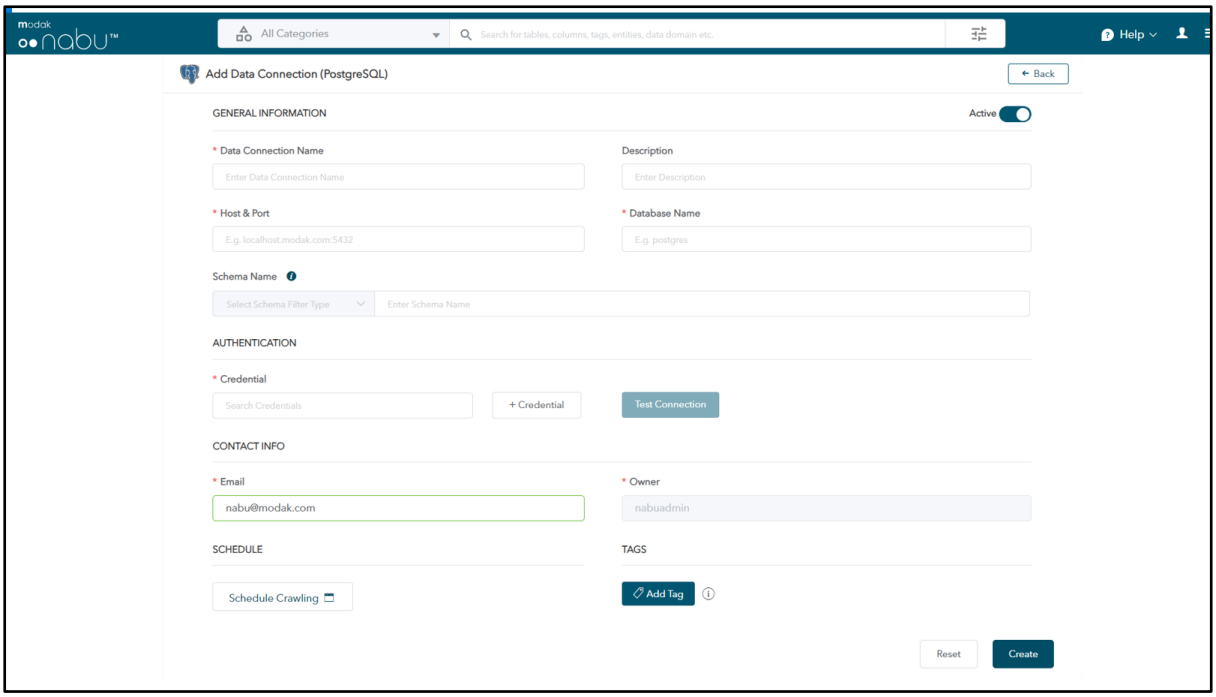


Figure 7 : PostgreSQL Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Postgres data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the Postgres database. Ex: localhost.modak.com:5432. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Postgres.
5	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for

		the data connection.
6	Credential	The user must select the credentials to connect to the Postgres database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port, and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.2 Add MySQL data connection

To add a MySQL data connection, click on the MySQL icon from the left pane of the 'Data connections' screen. The Add MySQL Data Connection form opens, and the user is required to provide the below details.

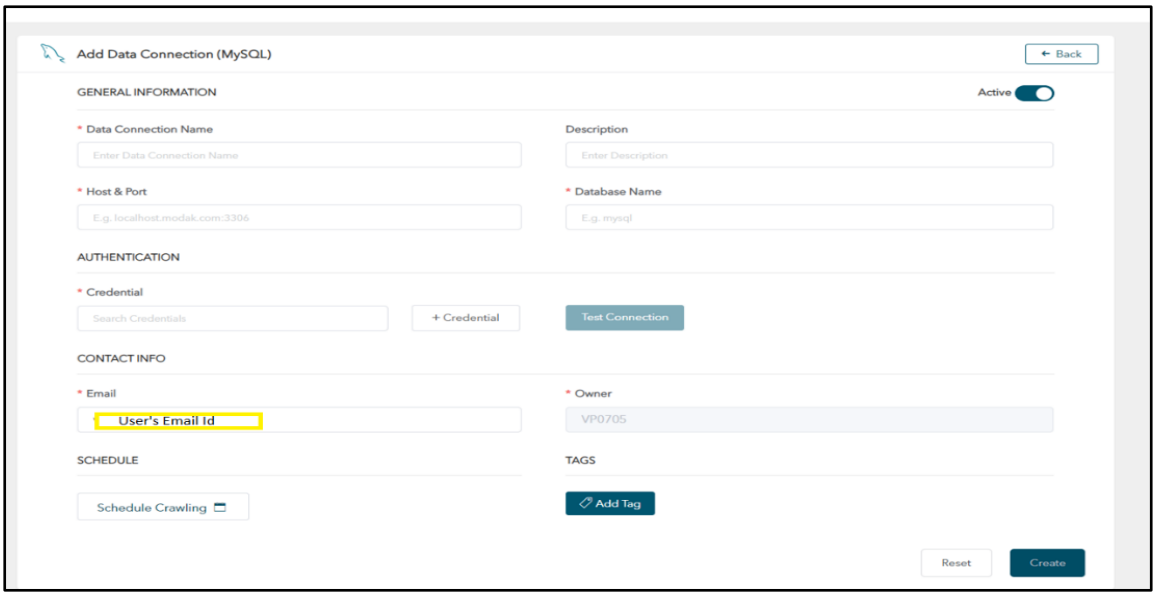


Figure 8: MySQL Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a MySQL data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the MySQL database. Ex: w3.devsql.modak.com:3306. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: MySQL.
5	Credential	The user must select the credentials to connect to the MySQL database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
6	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Modak Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port, and credentials.
7	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
8	Owner	This indicates the person who has created the data connection. The owner id is filled by default.

9	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9.
10	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10.
11	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.3 Adding Hive Data Connection

To add a Hive data connection, click on the Hive icon from the left pane of the 'Data connections' screen. The Add Hive data connection form opens, and the user is required to provide the below details.

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Hive data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
3	Email	This field is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
4	Crawl Frequency	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling.
5	Connection Mode	Choose between 'Cluster Mode' or 'JDBC Mode' for the Hive connection.
6	Description	This is an optional field where user can enter any information to add to the data connection.
7	Active slider	States if the data connection is active or not
8	Sensitive slider	States if the data connection has sensitive data or not
9	Reset button	Resets all the data entered in the fields.

Click on the **next** button to move on to the next section of the form

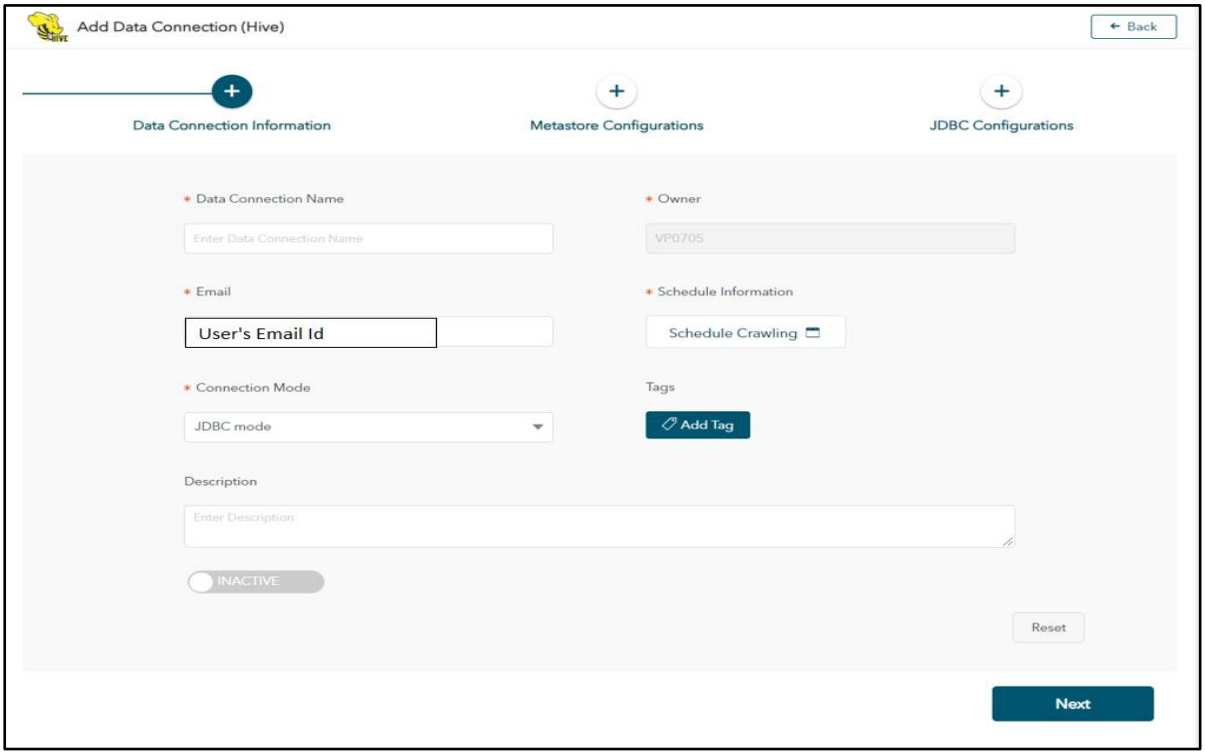


Figure 9: Hive Data Connection

Note: The process will differ depending on the connection mode you choose, if you choose 'cluster mode' you need to go through an extra step of 'File System Configuration' else you can skip one step.

JDBC Mode: When user selects JDBC mode, after the required details are filled above, click on Next".

S.no	Tag	Description
1	Metastore JDBC URL	The user must enter the URL for JDBC metastore. E.g., jdbc:mysql://host:port/databasename
2	Credentials	The user must select the credential from the credential's dropdown.

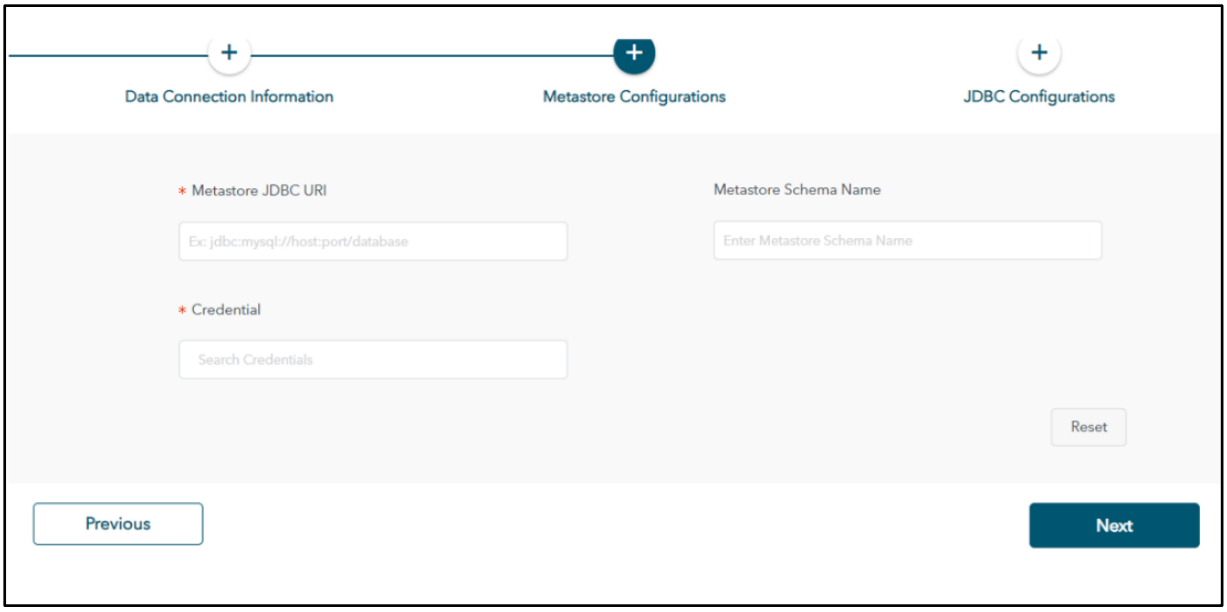


Figure 10: Hive Metastore Configuration

Click on **previous** button to go back to Data Connection Information section.

Click on the **next** button to move on to the next section of the form (Metastore Configurations)

User needs to fill the following details in the form

S.no	Tag	Description
1	Database name	The user needs to enter the database name where tables will be ingested
2	Hive JDBC URI	Host and Port for Hive JDBC
3	Credentials	The user must select the credential from the credential's dropdown.

Cluster Mode: when the user selects cluster configuration mode, on clicking 'Next' from Metastore configurations, you will be redirected to "File System Configurations" page. Fill the following attributes like

S.no	Tag	Description
1	File System type	States the file system type. User can choose between HDFS and S3A
2	Ingestion root path	Root path where objects are ingested
3	Credentials	The user must select the credential from the credential's dropdown.
4	Hadoop File System URI	Host and Port for Hadoop File System

	(If File system type is HDFS)	
5	Hadoop File System Config Directory (If File system type is HDFS)	Path for Hadoop File System Configuration

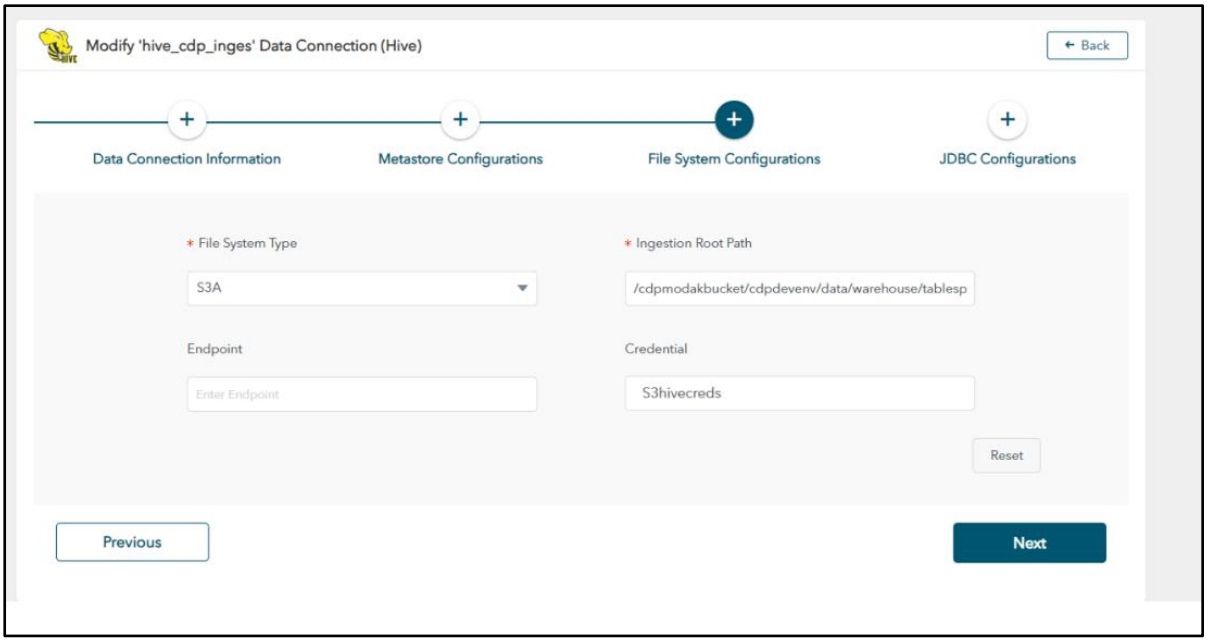


Figure 11: Data Connection File System config

Click on **previous** button to go back to Metastore Configurations.

Click on the **next** button to move on to the final section of the form (JDBC Configurations)

User needs to fill the following detail in the form

S.no	Tag	Description
1	Database name	This is a mandatory field where the user needs to enter the database name to connect to. Ex: Hive
2	Hive JDBC URI	Host and Port for Hive JDBC
3	Credentials	The user must select the credential from the credential's dropdown.

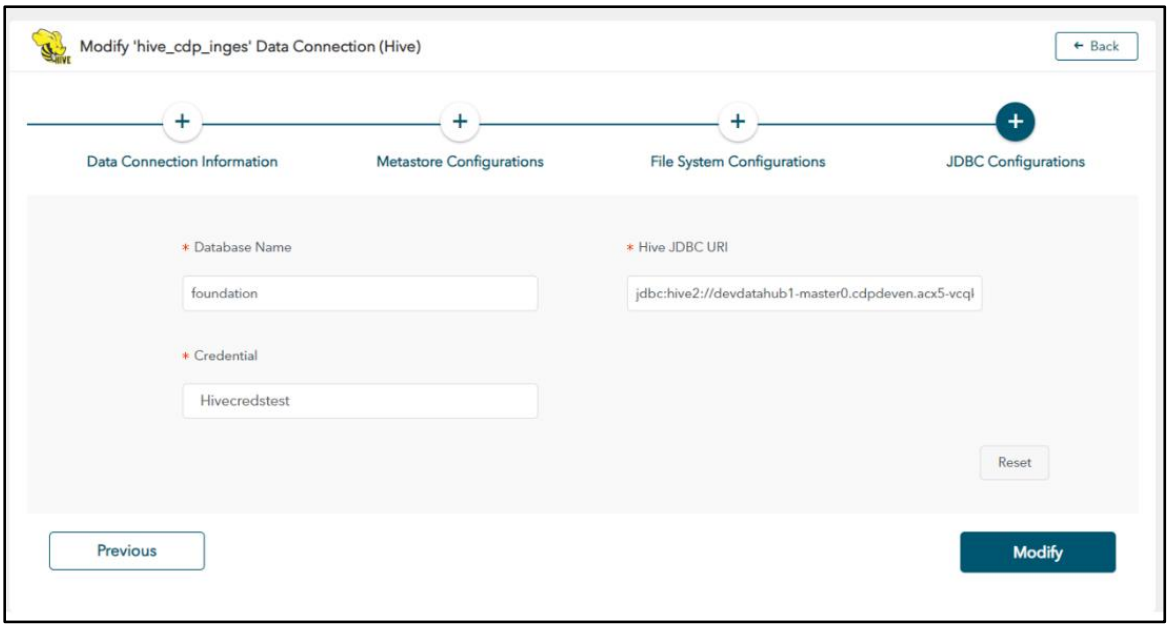


Figure 12: JDBC Configuration

Click on **previous** button to go back to File System Configurations.

Click on **Modify** button to make required changes to the existing data connection.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.4 Add Oracle Data Connection

To add an Oracle data connection, click on the Oracle icon from the left pane of the Data Connections' screen. The Add Oracle Data Connection form opens, and the user is required to provide the below details.

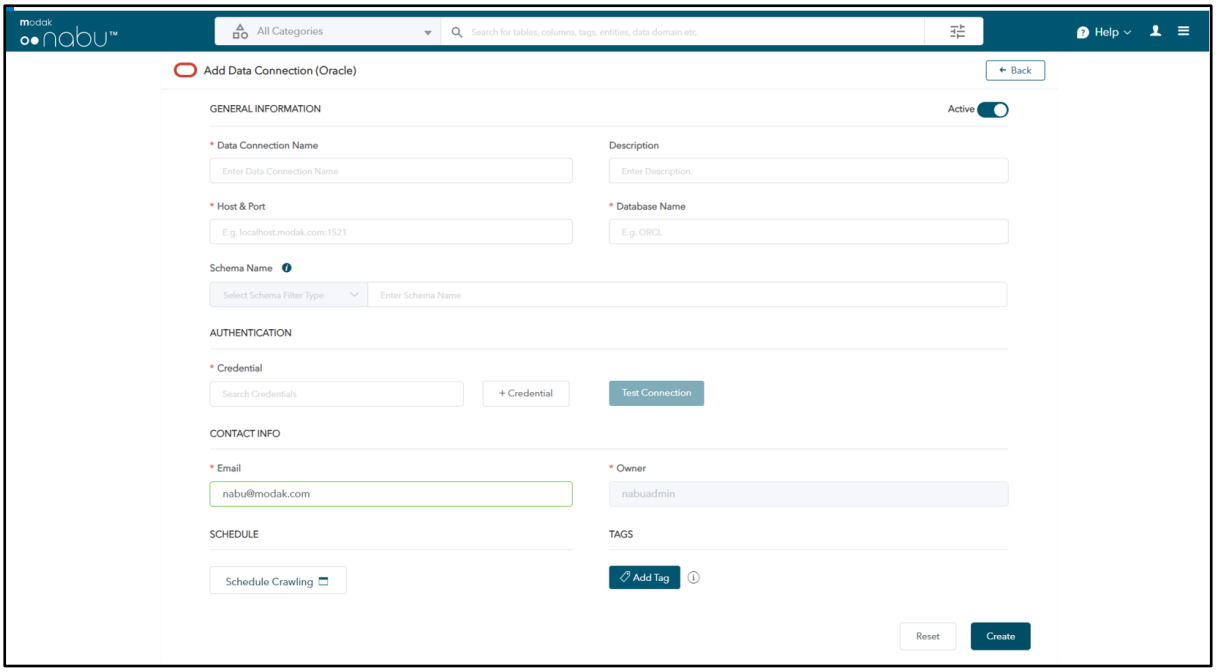


Figure 13: Oracle Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an Oracle data connection. A unique data connection name must be entered here. The Data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the Oracle database. Ex: w3.testorcl10b.modak.com:1521. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Oracle.
5	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
6	Credential	The user must select the credentials to connect to the oracle database. If the credential could not be fetched or would like to add any new credential, the user can click on **credential** button which will redirect to the Manage Credentials page where user can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port, and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9

11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.5 Add SQL server Data Connection

To add a SQL server data connection, click on the SQL server icon from the left pane of the 'Data connections' screen. The Add SQL server Data Connection form opens, and the user is required to provide the below details.

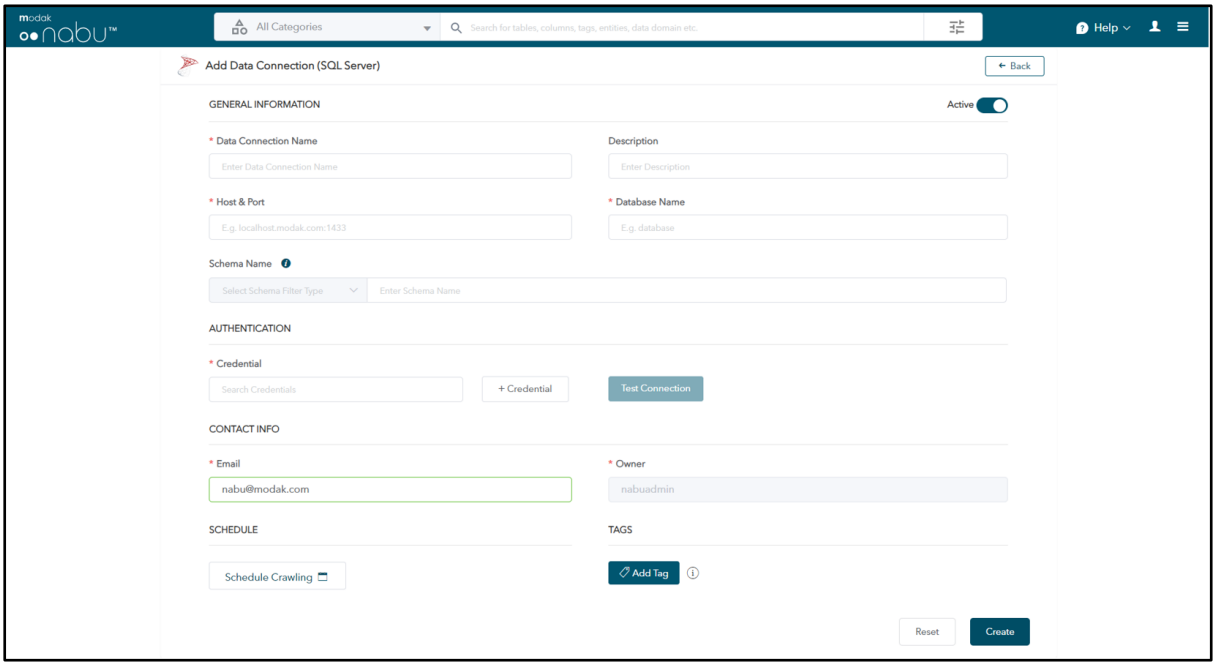


Figure 14: SQL server Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a SQL server data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the SQL server database. Ex: w3.devsql.modak.com:1433. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: SQL server
5	Schema Name	Select schema filter type from the below options and provide the schema name.

		• Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
6	Credential	The user must select the credentials to connect to the SQL server database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.6 Add DB2 Data Connection

To add a DB2 data connection, click on the DB2 icon from the left pane of the 'Data connections' screen. The Add DB2 Data Connection form opens, and the user is required to provide the below details.

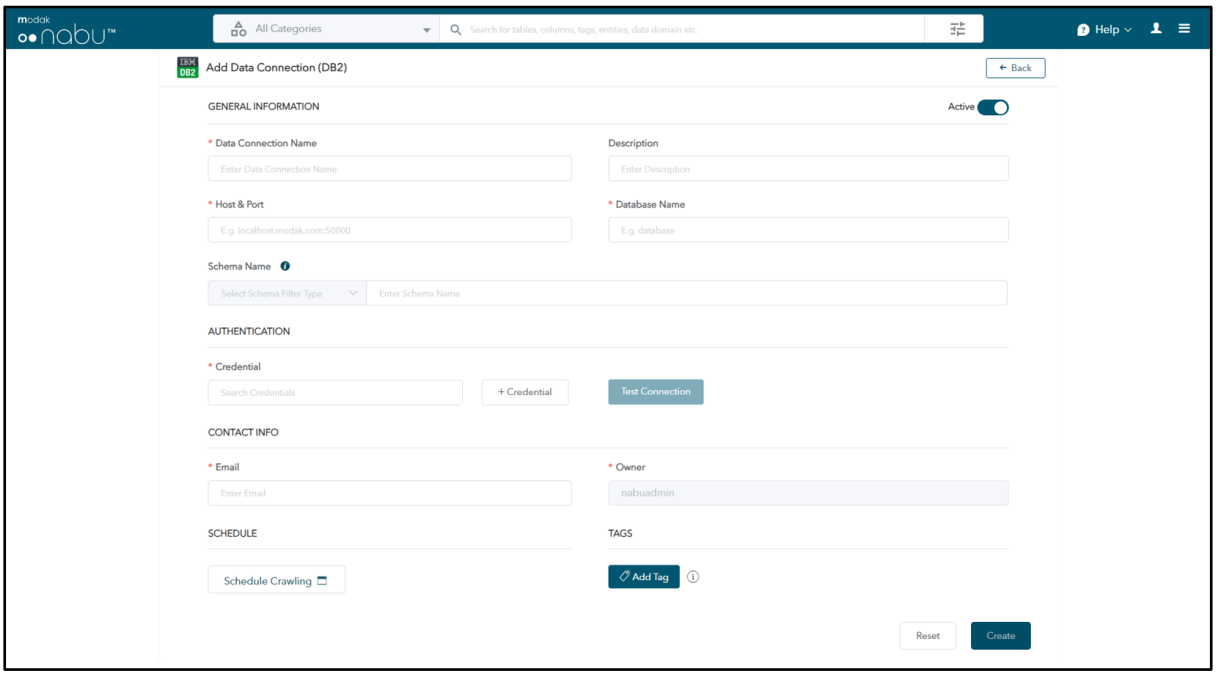


Figure 15 : DB2 Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a DB2 data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the DB2 database. Ex: w3.testsdc.modak.com:50000. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: DB2
5	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and

		the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
6	Credential	The user must select the credentials to connect to the DB2 database. If the credential could not be fetched or would like to add any new credential, the user can click on **credential** button which will redirect to the Manage Credentials page where user can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. In case the test connection fails, check the details – hostname, port and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.7 Add Teradata Data Connection

To add a Teradata data connection, click on the Teradata icon from the left pane of the 'Data connections' screen. The Add Teradata Data Connection form opens, and the user is required to provide the below details.

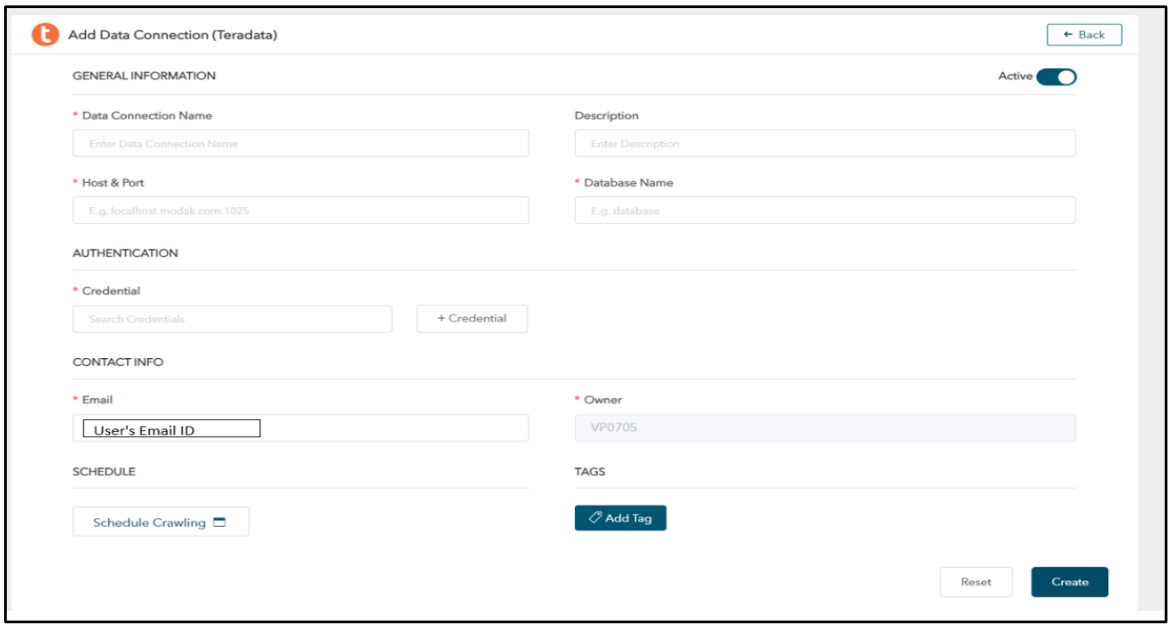


Figure 16: Teradata Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Teradata connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the Teradata database. Ex: localhost.modak.com:1025. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Teradata
5	Credential	The user must select the credentials to connect to the Teradata database. If the credential could not be fetched or would like to add any new credential, the user can click on **credential** button which will redirect to the Manage Credentials page where user can add a new credential.
6	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
7	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
8	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
9	Add Tag	The user can create a tag for the data connection by clicking on

		Add Tag. For Add Tag please refer to section:3.1.10
10	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.1.8 Add Netezza Data Connection

To add a Netezza data connection, click on the Netezza icon from the left pane of the 'Data connections' screen. The Add Netezza Data Connection form opens, and the user is required to provide the below details

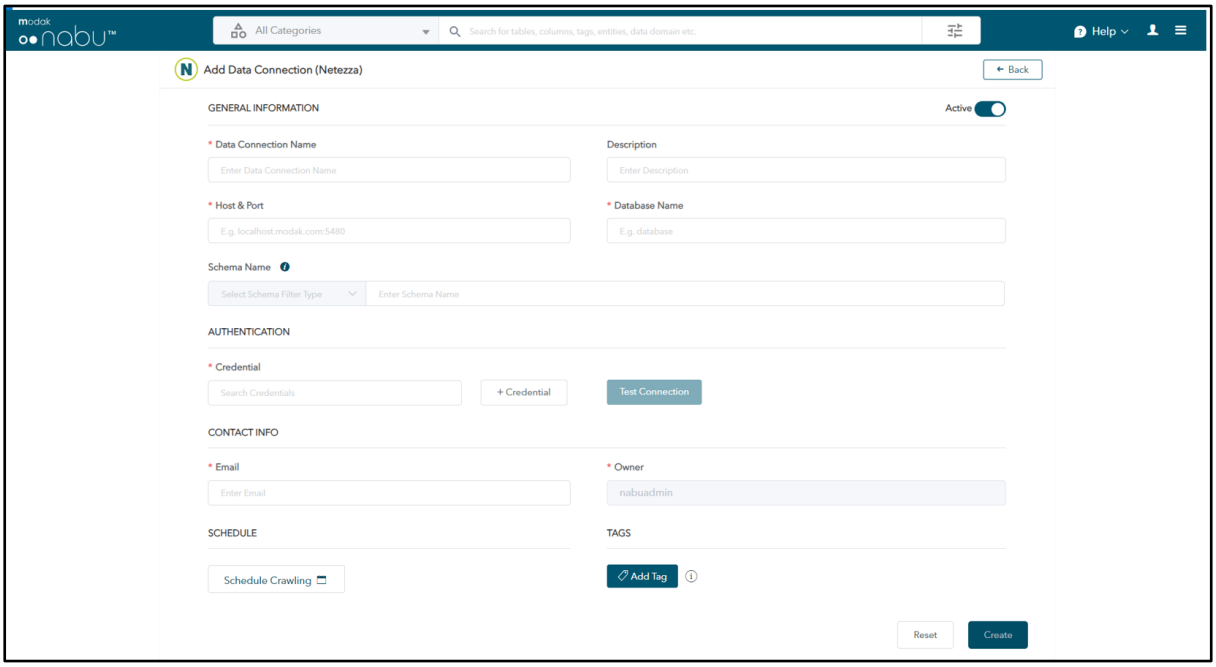


Figure 17: Netezza Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Netezza data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the Netezza database. Ex: localhost.modak.com:5480. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Netezza
5	Schema Name	Select schema filter type from the below options and provide the schema name.

		• Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
6	Credential	The user must select the credentials to connect to the Netezza database. If the credential could not be fetched or would like to add any new credential, the user can click on **credential** button which will redirect to the Manage Credentials page where user can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. In case the test connection fails, check the details – hostname, port, and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use back button which is on the top right corner of the window to go back to home page.

3.1.1.9 Add SAP Hana Data Connection

To add an SAP Hana data connection, click on the SAP Hana icon from the left pane of the 'Data connections' screen. The Add SAP Hana Data Connection form opens, and the user is required to provide the below details.

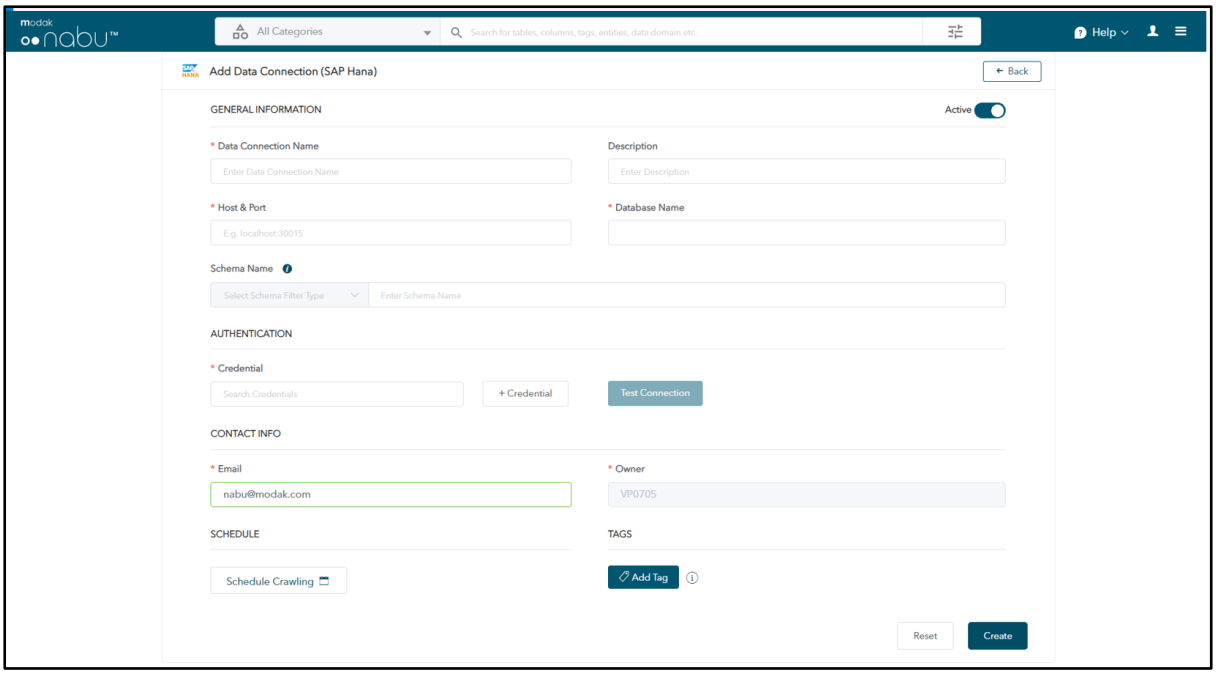


Figure 18 : Add SAP Hana Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an SAP Hana data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the SAP Hana database. Ex: hana.trial-us10.hanacloud.ondemand.com:443. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: H00.
5	Schema Name	This is an optional field where the user can select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for

		the data connection.
6	Credential	The user must select the credentials to connect to the SAP Hana database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port, and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page

3.1.2 CLOUD SERVICES

3.1.2.1 Add Amazon S3 data connection

To add an Amazon S3 data connection, click on the Amazon S3 icon from the left pane of the Data Connections' screen. The Add Amazon S3 Connection form opens, and the user is required to provide the below details.

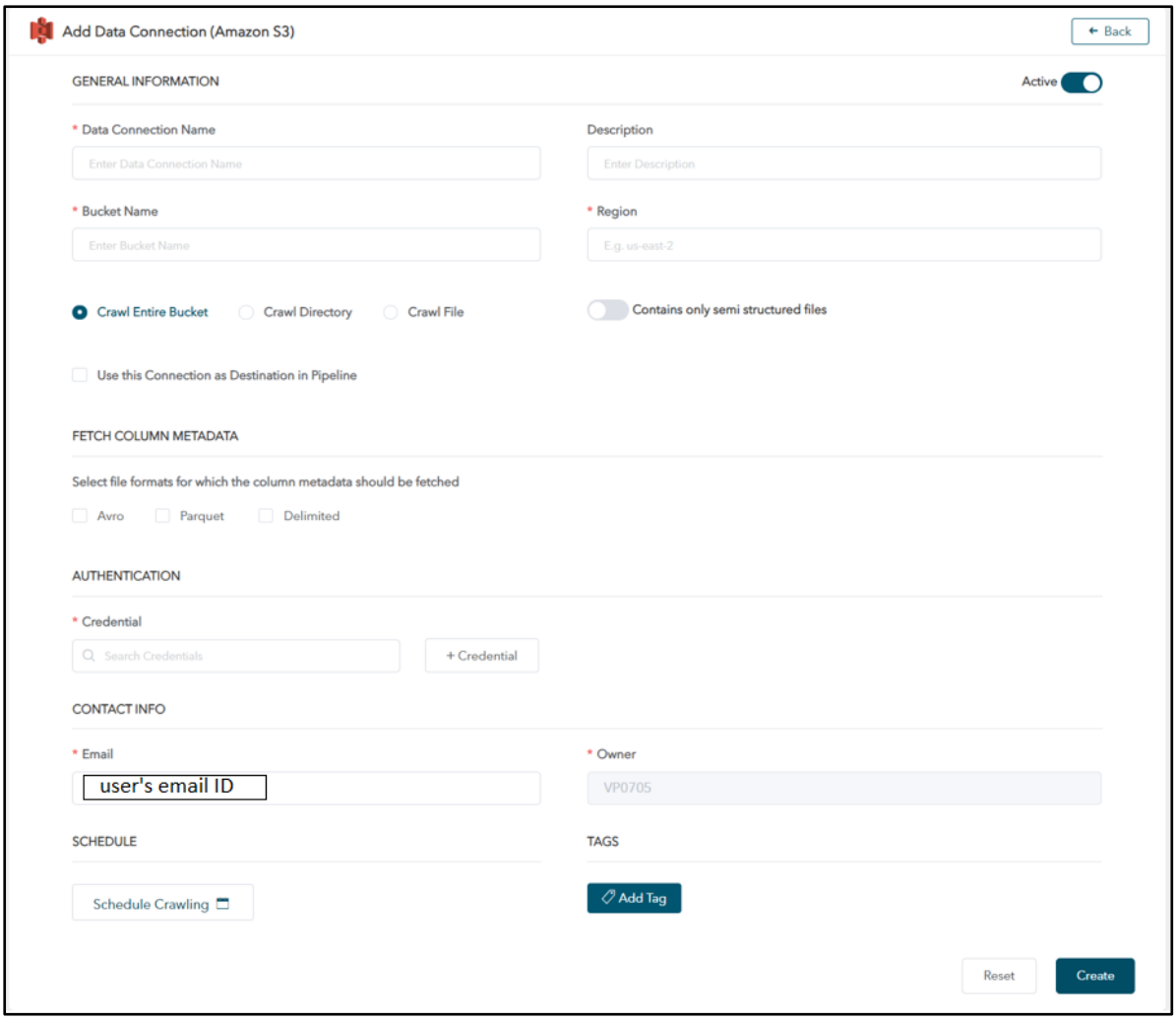


Figure 19: Amazon S3 Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an Amazon S3 data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Bucket Name	1. This is a mandatory field where the user needs to enter the bucket name to connect to the Amazon S3 data connection. An invalid bucket name will throw an error. This bucket name should start with an alphabet. It should contain at least 3 characters. The bucket name cannot exceed 63 characters. a) By default, the **crawl Entire Bucket** radio option is selected. As per the use case, the user can select any of the below.

		b) **Crawl Directory:** This option enables users to crawl the directory. When user selects crawl directory, the c) **Crawl File:** This option enables users to crawl the file. 2. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
4	Region	This is a mandatory field where the user needs to enter the region name for the amazon s3 bucket. Ex: us-east-2.
5	Directory Path	This is a mandatory field when user selects Crawl Directory radio option. The user should provide the directory path to crawl the directory. Ex: bucket/directory
6	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: bucket/file
7	Ingestion Root Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: bucket/directory. By default, this checkbox is unchecked.
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro, Parquet and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the Amazon S3 database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
10	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Modak Nabu is able to connect to the data connection. If the test connection fails, check the bucket details and credentials.
11	Email	This field is auto filled with the user's email when the user opens this form. This field is to contact the user for any issues related to this data connection.
12	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
13	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
14	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
15	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.2 Add AWS Glue data connection

To add an AWS Glue data connection, click on the AWS Glue icon from the left pane of the 'Data connections' screen. The Add AWS Glue Connection form opens, and the user is required to provide the below details.

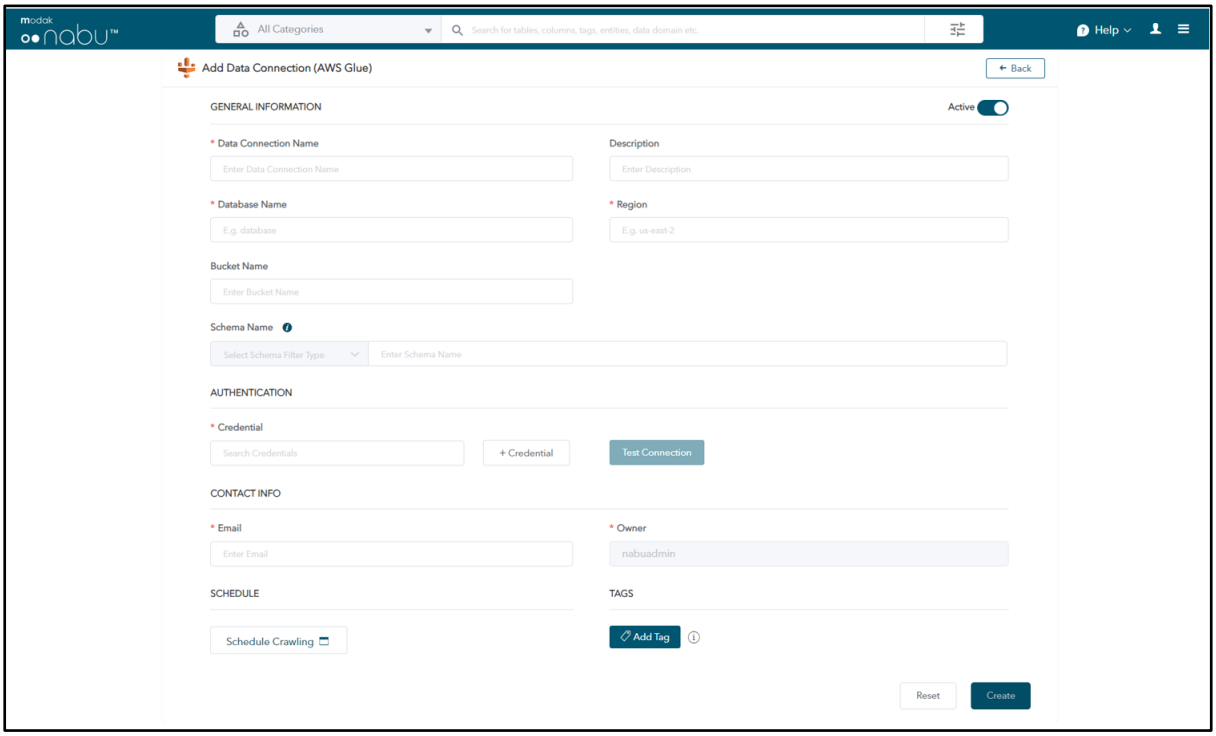


Figure 20: AWS Glue Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an AWS Glue data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet and should contain at least 3 characters. Special characters except underscore (_) are not allowed. The data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where the user can enter any information to add to the data connection.
3	Database Name	This is a mandatory field where the user needs to enter the database name to connect to. Ex: AWS Glue The Database name should start with an alphabet and should contain at least 3 characters and numbers are allowed. Special characters except underscore (_) are not allowed. The name cannot exceed more than 50 characters.
4	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like

		• Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
5	Bucket Name	This field is optional, and the user needs to enter the bucket name to connect to the AWS Glue data connection. An invalid bucket name will throw an error. This bucket name should start with an alphabet. It should contain at least 3 characters. Special characters except underscore (_) are not allowed. The name cannot exceed 63 characters.
6	Region	This is a mandatory field where the user needs to enter the region name for the AWS Glue bucket. Ex: us-east-2. Minimum characters are 9, maximum characters are 16
7	Credential	The user must select the credentials to connect to the AWS Glue database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
8	Email	This field is auto filled with the user's email when the user opens this form. This field is to contact the user for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.3 Add Amazon Athena data connection

To add an Amazon Athena data connection, click on the Amazon Athena icon from the left pane of the 'Data connections' screen. The Add Amazon Athena Connection form opens, and the user is required to provide the below details.

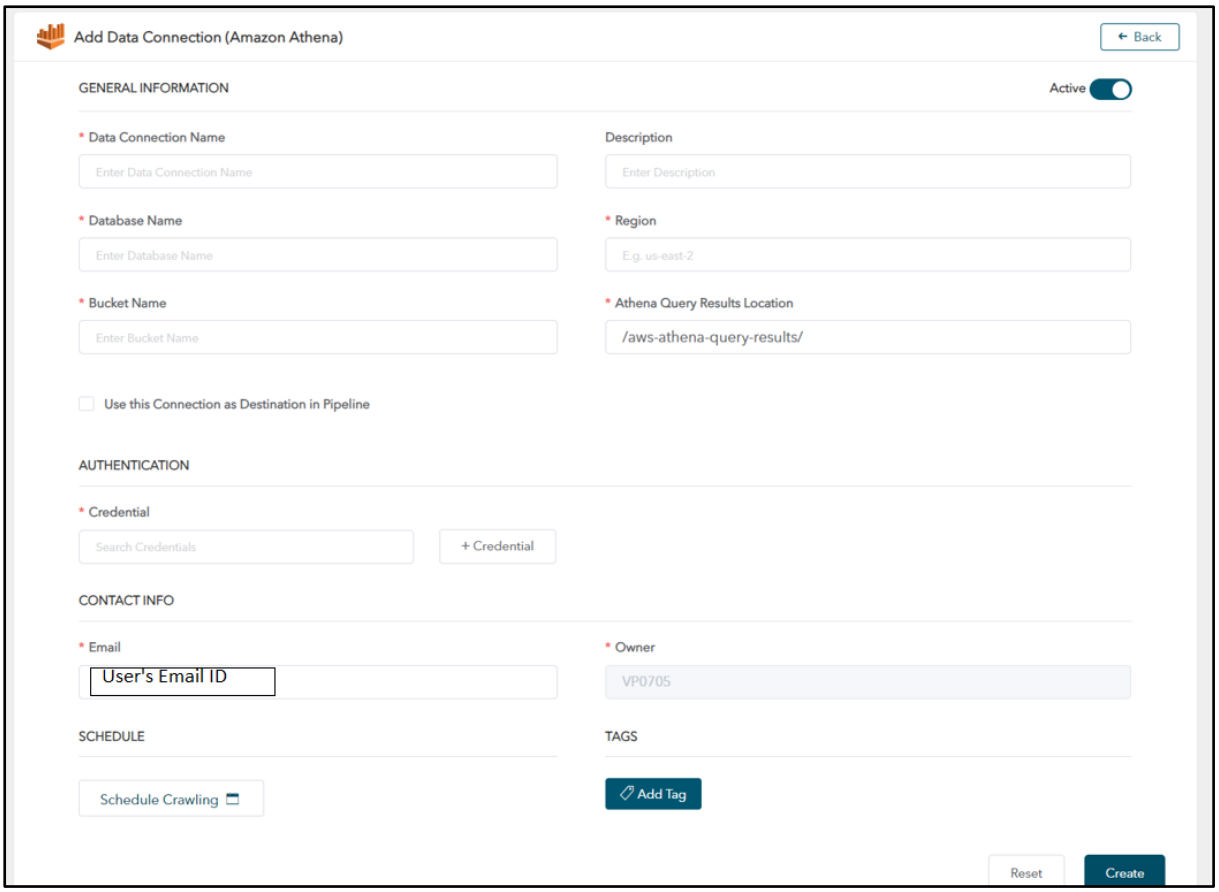


Figure 21: Amazon Athena Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an Amazon Athena data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet and should contain at least 3 characters. Special characters except underscore (_) are not allowed. The name cannot exceed more than 50 characters.
2	Description	This is an optional field where the user can enter any information to add to the data connection.
3	Database Name	This is a mandatory field where the user needs to enter the database name to connect to. Ex: Amazon Athena The Database name should start with an alphabet and should contain at least 3 characters. Special characters except underscore (_) are not allowed. The database name cannot exceed more than 50 characters.
4	Bucket Name	This is a mandatory field, and the user needs to enter the bucket name to connect to the Amazon Athena data connection. An invalid bucket name will throw an error. This bucket name should start with an alphabet. It should contain at least 3 characters. The bucket name cannot exceed 63 characters.
5	Region	This is a mandatory field where the user needs to enter the region for the Amazon Anthem bucket. Ex: us-east-2. Invalid

		region details will throw an error. Minimum characters are 9, maximum characters are 16.
6	Athena Query Results location	This is a field where the query results are stored/saved. The user needs to enter the query results location here. E.g., /aws-athena-query-results/
7	Ingestion Root Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: bucket/directory. By default, this checkbox is unchecked.
8	Credential	The user must select the credentials to connect to the Amazon Athena data connection. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
9	Email	This field is auto filled with the user's email when the user opens this form. This field is to contact the user for any issues related to this data connection.
10	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
11	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
12	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
13	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.4 Add Amazon Redshift data connection

To add an Amazon Redshift data connection, click on the Amazon Redshift icon from the left pane of the 'Data connections' screen. The Add Amazon Redshift Connection form opens, and the user is required to provide the below details.

Form contains two sections for which user needs to provide details

1. Data connection information
2. File system configurations

Data connections information section

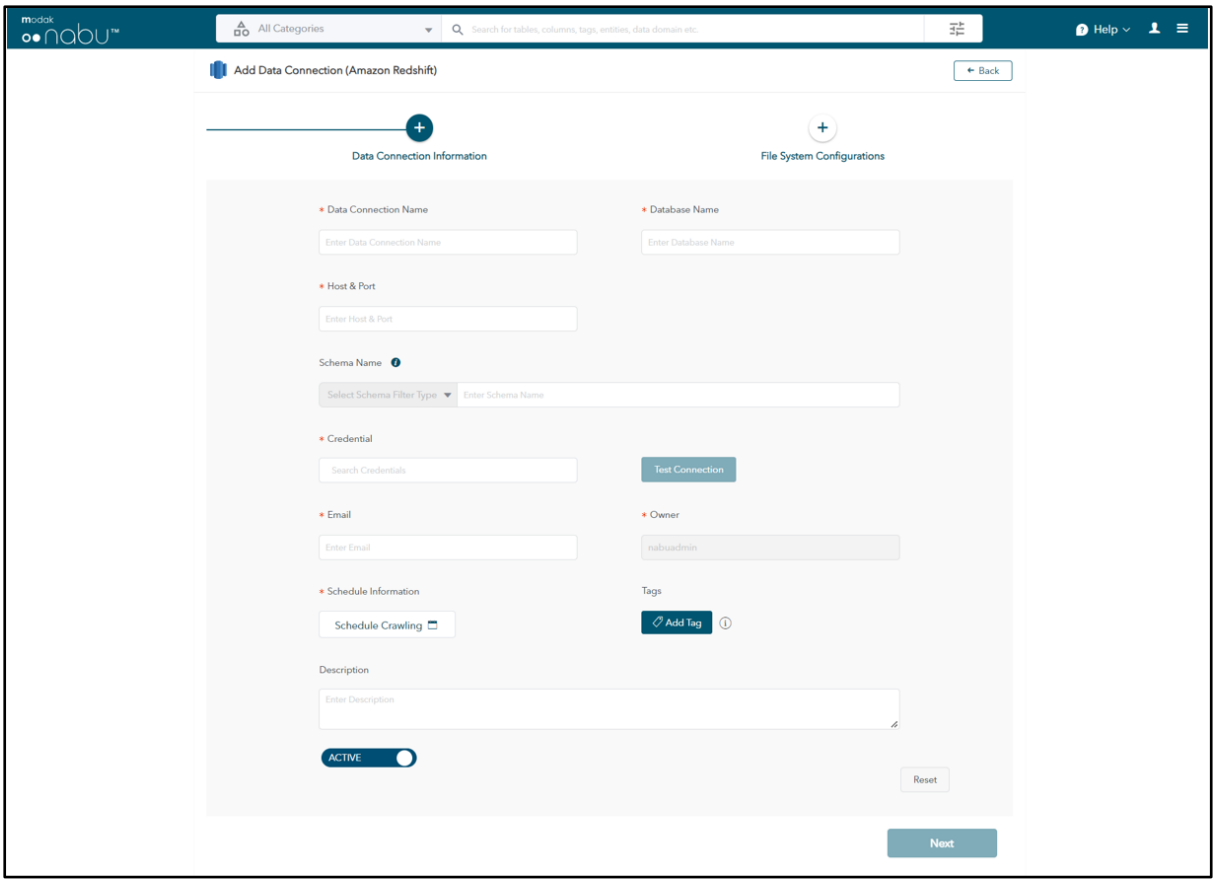


Figure 22: Amazon Redshift- Data Connection information

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an Amazon Redshift data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Host & Port	This is a mandatory field where the user needs to enter the host and port details to connect to the Amazon Redshift database. Invalid Host & Port details will throw an error.
3	Database Name	This is a mandatory field where the user needs to enter the database name to connect to. Ex: Amazon redshift. The database name should start with an alphabet and should contain at least 3 characters. The name cannot exceed more than 50 characters.
4	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause

		• Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
5	Email	This is a mandatory field where the user needs to enter the email to connect to. This field is to contact for any issues related to this data connection.
6	Credential	The user must select the credentials to connect to the Amazon Redshift database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
7	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
8	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
9	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
10	Description	This is an optional field where the user can enter any information to add to the data connection.
11	Reset	**Reset** This button resets the Add Data connection form.

Click on the **next** button to move on to the next section of the form

File system configurations section

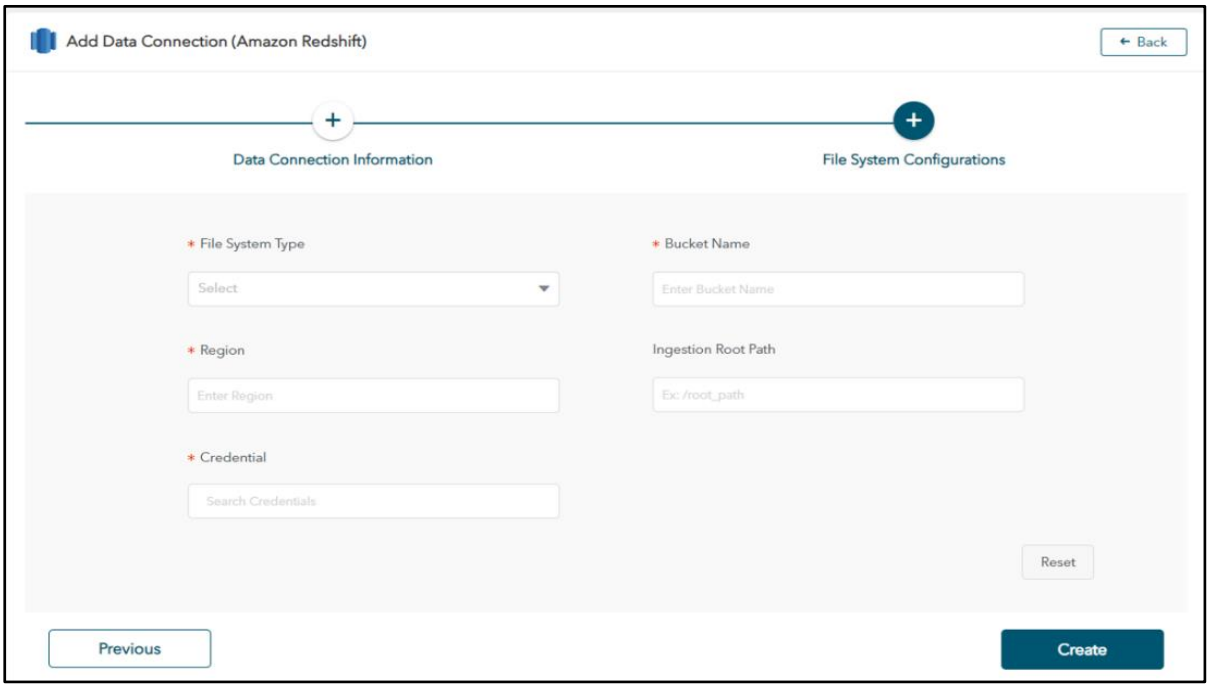


Figure 23: Amazon Redshift- File System Configurations

S.no	Tag	Description
1	File system type	This is a mandatory filed where user needs to select the file system type. Ex: S3A.
2	Bucket Name	The user needs to enter the bucket name to connect to the Amazon Redshift data connection. An invalid bucket name will throw an error. This bucket name should start with an alphabet. It should contain at least 3 characters. The bucket name cannot exceed 63 characters.
3	Region	This is a mandatory field where the user needs to enter the region name for the Amazon Redshift bucket. Ex: us-east-2.
4	Ingestion Root Path	This is an optional field where user needs to enter the ingestion root path. Ex: /root path
5	Credential	The user must select the credentials to connect to the Amazon Redshift database.
6	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Click on **previous** button to go back to data connection information section

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.5 Add ADLS Gen 1 data connection

To add an ADLS Gen 1 data connection, click on the ADLS Gen 1 icon from the left pane of the 'Data connections' screen. The Add ADLS Gen 1 Connection form opens, and the user is required to provide the below details.

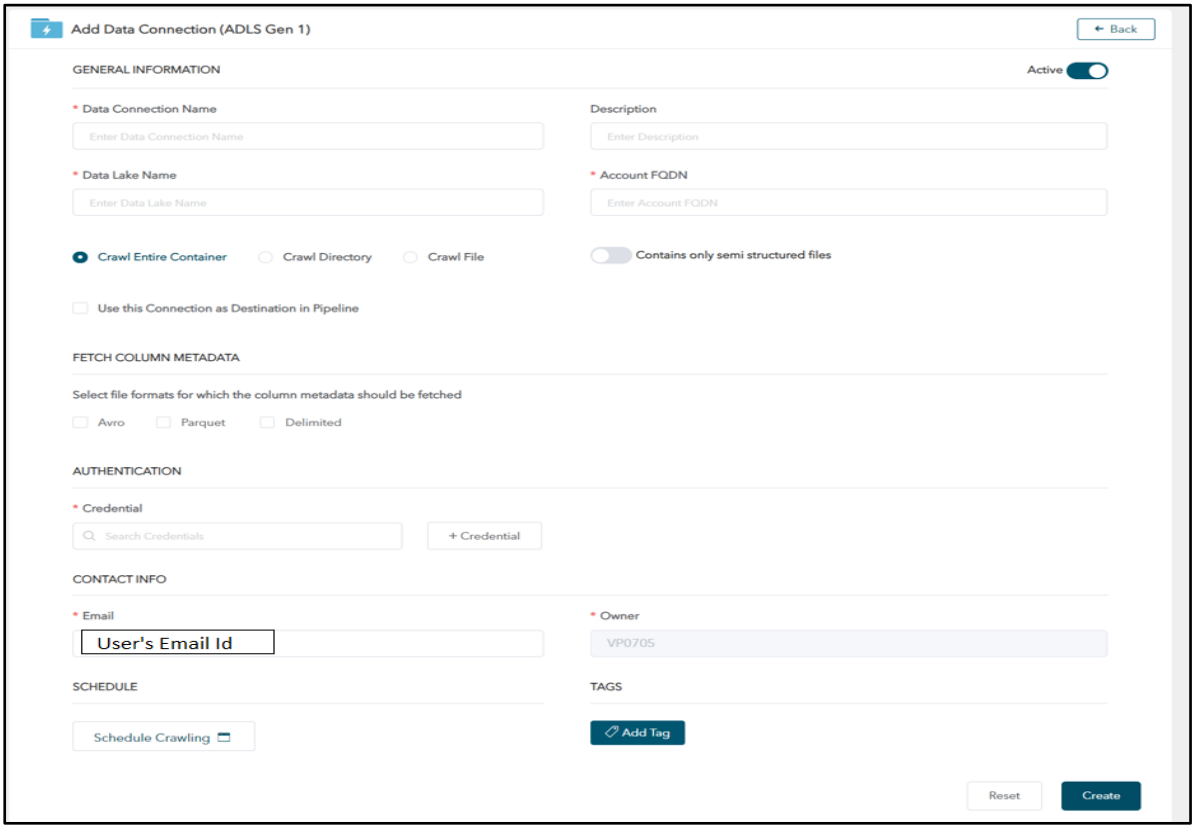


Figure 24: ADLS Gen 1 Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an ADLS Gen 1 data connection. A unique data connection name must be entered here. The Data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Data lake name	1. This is a mandatory field where the user needs to enter the data lake name to connect to the ADLS Gen 1 database. An invalid data lake name will throw an error. This data lake name should start with an alphabet. It should contain at least 3 characters. The data lake name cannot exceed 63 characters. a) By default, the **Crawl Entire Container** option is selected. As per the use case, the user can select any of the below. b) **Crawl Directory:** This option enables users to crawl the directory c) **Crawl File:** This option enables users to crawl the file. 2. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
4	Account FQDN	This is a mandatory field where user needs to enter the FQDN details to connect to ADLS Gen 1.
5	Directory Path	This is a mandatory field when user selects Crawl Directory radio option. The user should provide the directory path to crawl the directory. Ex: container/directory
6	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: container/file
7	Ingestion Root Directory Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: container/directory. By default, this checkbox is unchecked.
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro, Parquet and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the ADLS Gen 1 data connection. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
10	Email	This field is auto filled with the user's email when the user

		opens this form. This field is to contact the user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. The user can schedule a crawl for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.6 Add ADLS Gen 2 data connection

To add an ADLS Gen 2 data connection, click on the ADLS Gen 2 icon from the left pane of the 'Data connections' screen. The Add ADLS Gen 2 Connection form opens, and the user is required to provide the below details.

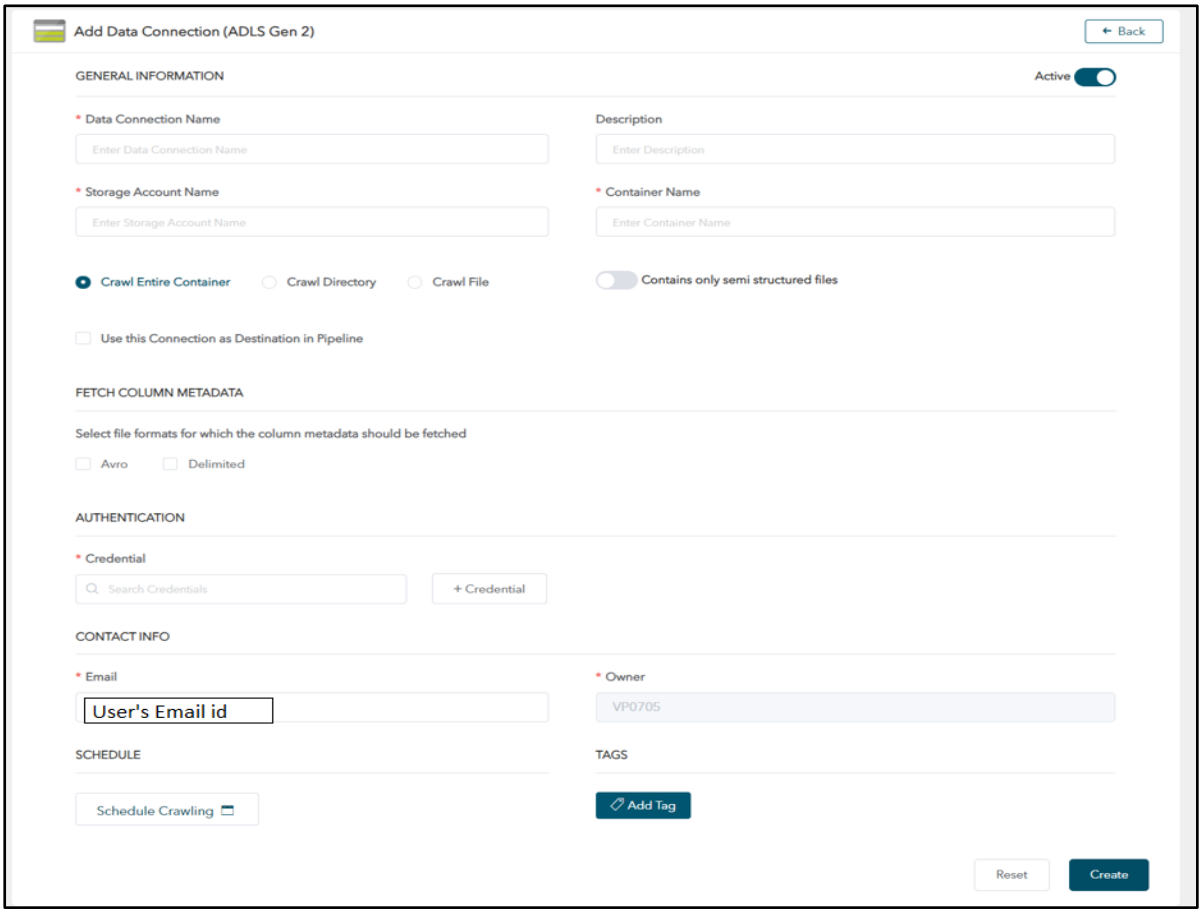


Figure 25: ADLS Gen 2 Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an ADLS Gen 2 data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Storage account name	1. This is a mandatory field where the user needs to enter the storage account name to connect to the ADLS Gen 2 database. An invalid storage account name will throw an error. It should contain at least 3 characters. Special characters are not allowed. The name cannot exceed 24 characters. a) By default, the **Crawl Entire Bucket** option is selected. As per the use case, the user can select any of the below. b) **Crawl Directory:** This option enables users to crawl the directory c) **Crawl File:** This option enables users to crawl the file. 2. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
4	Container name	This is a mandatory field where the user needs to enter the valid container name. It must start with an alphabet and should contain only letters and numbers.
5	Directory Path	This is a mandatory field when user selects Crawl Directory radio option. The user should provide the directory path to crawl the directory. Ex: container/directory
6	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: container/file
7	Ingestion Root Directory Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: container/directory. By default, this checkbox is unchecked.
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the ADLS Gen 2 data connection. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.

10	Email	This field is auto filled with the user's email when the user opens this form. This field is to contact the user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use back button which is on the top right corner of the window to go back to home page.

3.1.2.7 Add Azure Blob Storage

To add an Azure Blob Storage data connection, click on the Azure Blob Storage icon from the left pane of the 'Data connections' screen. The Add Azure Blob Storage Connection form opens, and the user is required to provide the below details.

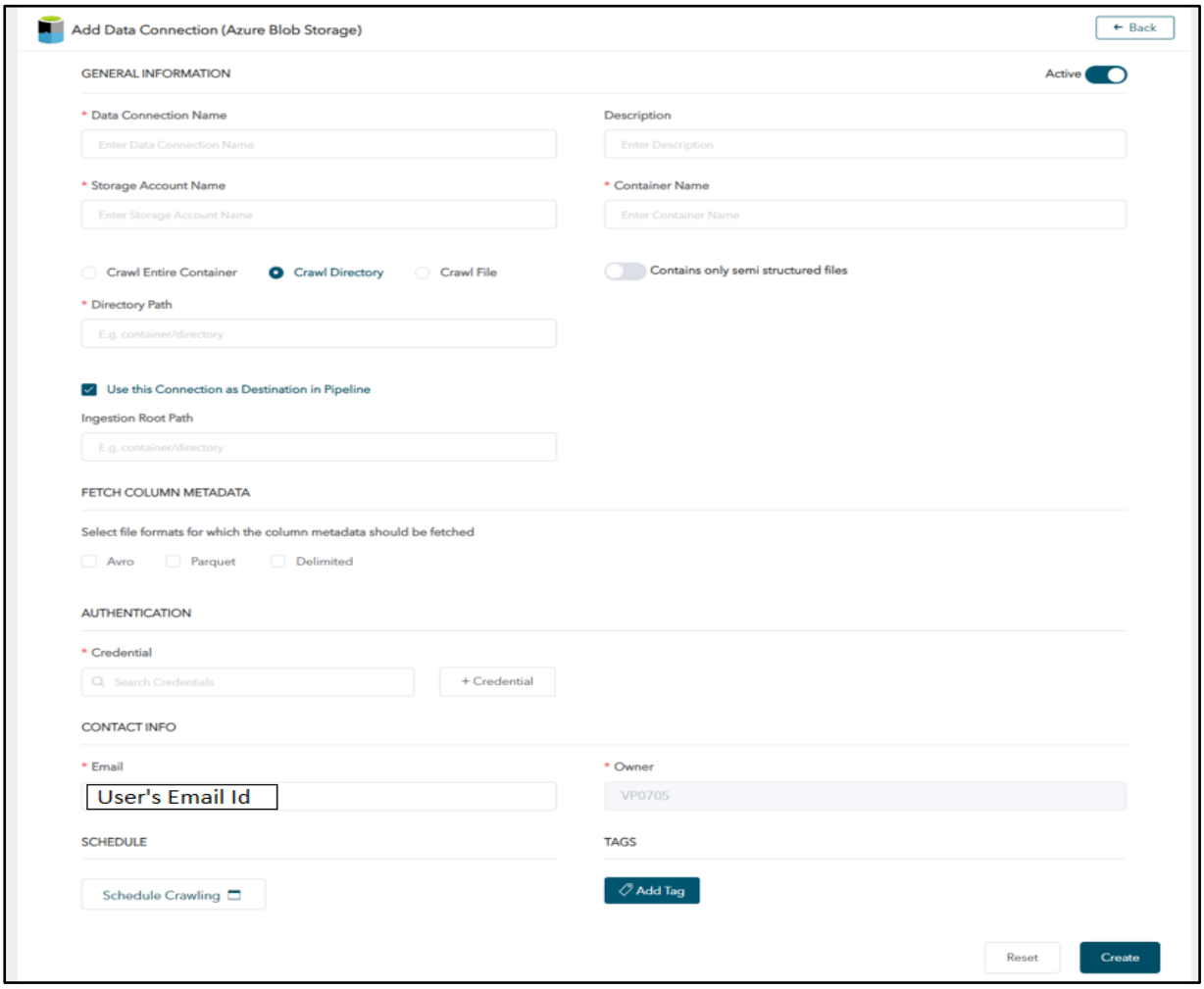


Figure 26: Azure Blob Storage Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an Azure Blob Storage data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Storage account name	3. This is a mandatory field where the user needs to enter the storage account name to connect to the Azure Blob Storage database. An invalid storage account name will throw an error. It should contain at least 3 characters. Special characters are not allowed. The name cannot exceed 24 characters. a) By default, the **Crawl Entire Container** option is selected. As per the use case, the user can select any of the below. b) **Crawl Directory:** This option enables users to crawl the directory c) **Crawl File:** This option enables users to crawl the file. 4. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
4	Container name	This is a mandatory field where the user needs to enter the valid container name. It must start with an alphabet and should contain only letters and numbers.
5	Directory Path	This is a mandatory field when user selects Crawl Directory radio option. The user should provide the directory path to crawl the directory. Ex: container/directory
6	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: container/file
7	Ingestion Root Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: container/directory. By default, this checkbox is unchecked.
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro, parquet and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the Azure Blob Storage data connection. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.

10	Email	This field is auto filled with the user's email when the user opens this form. This field is to contact the user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.8 Add Azure synapse data connection

To add an Azure synapse data connection, click on the Azure synapse icon from the left pane of the 'Data connections' screen. The Add Azure Synapse Connection form opens, and the user is required to provide the below details.

Form contains two section which user needs to fill

1. Data connection information
2. File system configurations

Data connections information section

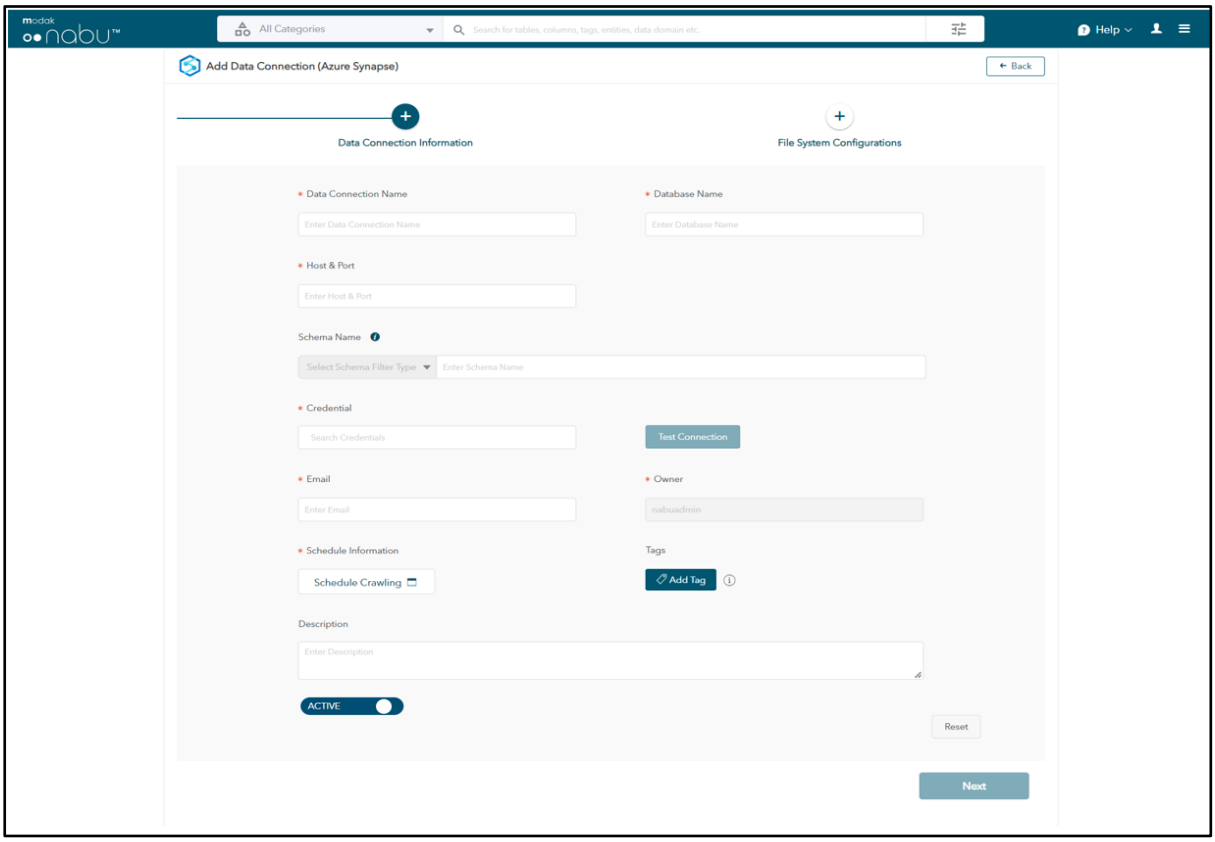


Figure 27: Azure Synapse- Data Connection Information

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an Azure Synapse data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Host & Port	This is a mandatory field where the user needs to enter the host and port details to connect to the Azure synapse database. Invalid Host & Port details will throw an error.
3	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Azure synapse
4	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
4	Email	This is a mandatory field where user needs to enter the email to connect to. This field is to contact for any issues related to this data connection.
5	Credential	The user must select the credentials to connect to the Azure Synapse database.
6	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
7	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
8	Add Tag	The user can create the tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
9	Description	This is an optional field where the user can enter any information to add to the data connection.
10	Reset	This button resets the Add Data connection form.

Click on the **next** button to move on to the next section of the form

File system configurations section

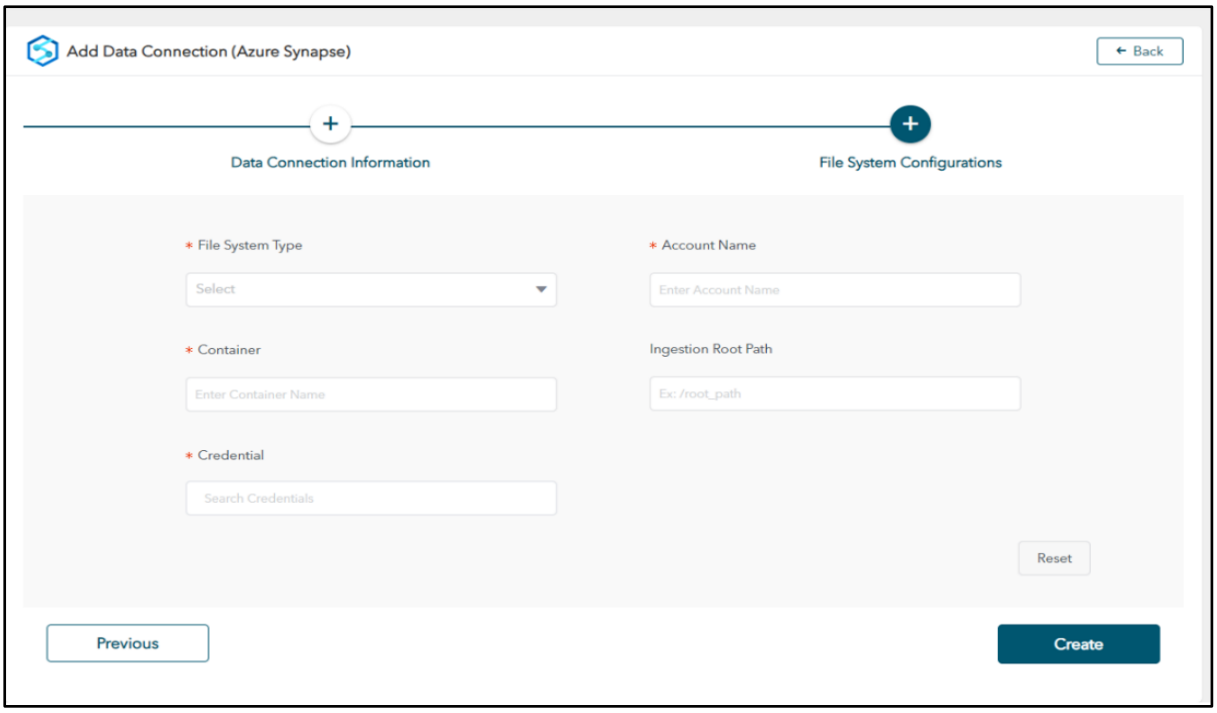


Figure 28: Azure Synapse- File System Configurations

S.no	Tag	Description
1	File system type	This is a mandatory filed where user needs to select the file system type.
2	Account Name	The is a mandatory field where the user needs to enter the account name to connect to the Azure Synapse. An invalid account name will throw an error. It should contain at least 3 characters. Special characters are not allowed.
3	container	This is a mandatory field where the user needs to enter the container name for the Azure Synapse.
4	Ingestion Root Path	This is an optional field where user needs to enter the ingestion root path. Ex: /root path.
5	Credential	The user must select the credentials to connect to the Azure Synapse database.
6	Create/ Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Click on **previous** button to go back to data connection information section.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.9 Add GCS data connection

To add an GCS data connection, click on the GCS icon from the left pane of the 'Data connections' screen. The Add GCS Connection form opens, and the user is required to provide the below details.

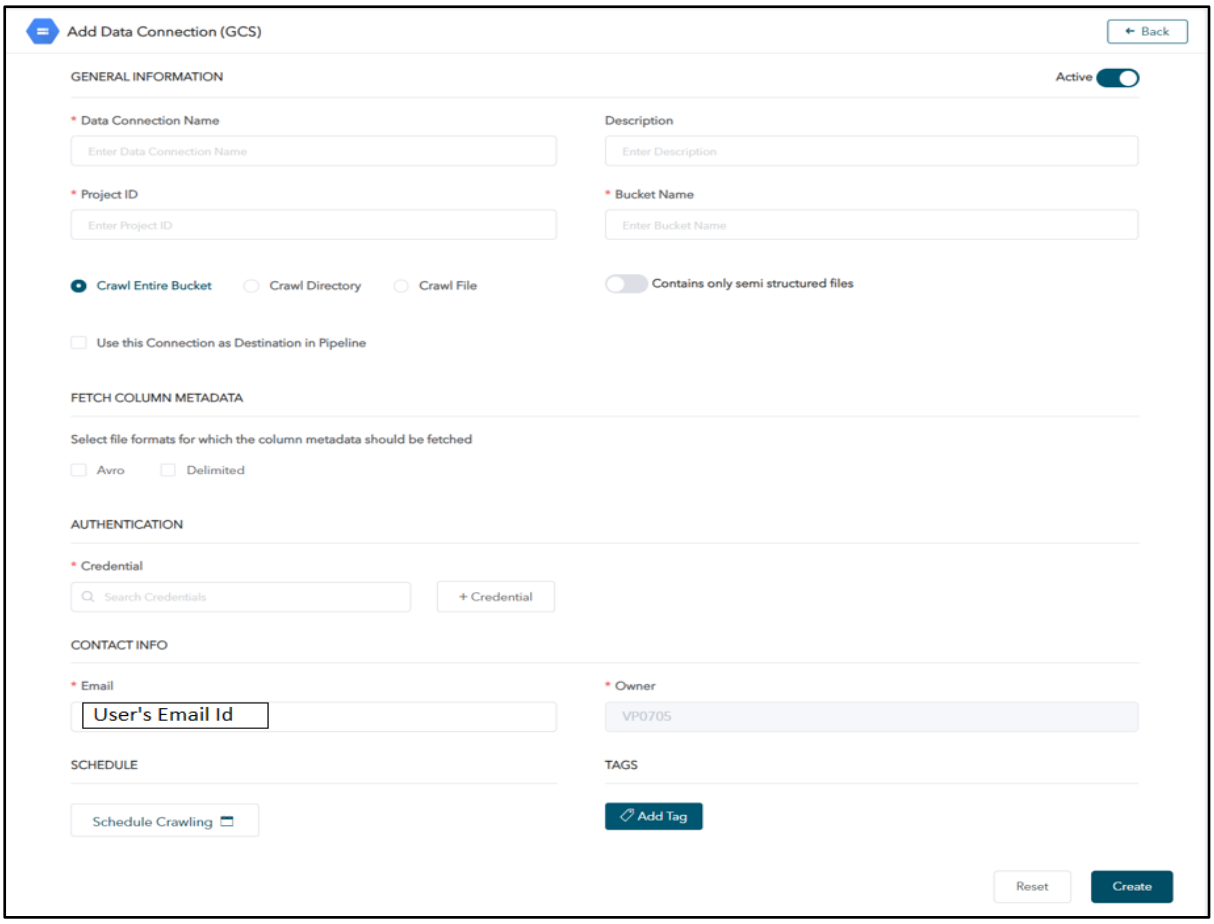


Figure 29: GCS Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an GCS data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Project ID	This is a mandatory field where the user needs to enter the Project ID to connect to the GCS database. An invalid project ID will throw an error. This Project ID should start with an alphabet. It should contain at least 6 characters and can contain numbers. The name cannot exceed 30 characters. 1. By default, the **Crawl Entire Bucket** option is selected. As per the use case, the user can select any of the below. 2. **Crawl Directory:** This option enables users to crawl the directory 3. **Crawl File:** This option enables users to crawl the file. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi

		structured files.
4	Bucket Name	The user needs to enter the bucket name to connect to the GCS. An invalid bucket name will throw an error. It should contain at least 3 characters. The bucket name cannot exceed 255 characters.
5	Directory path	This is a mandatory field when user selects Crawl Directory radio option. The user should provide the directory path to crawl the directory. Ex: bucket/directory
6	File path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: bucket/file
7	Ingestion Root Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: container/directory. By default, this checkbox is unchecked.
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the GCS database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
10	Email	This field is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create the tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.10 Add BigQuery data connection

To add a BigQuery data connection, click on the BigQuery icon from the left pane of the 'Data connections' screen. The Add BigQuery Connection form opens, and the user is required to provide the below details.

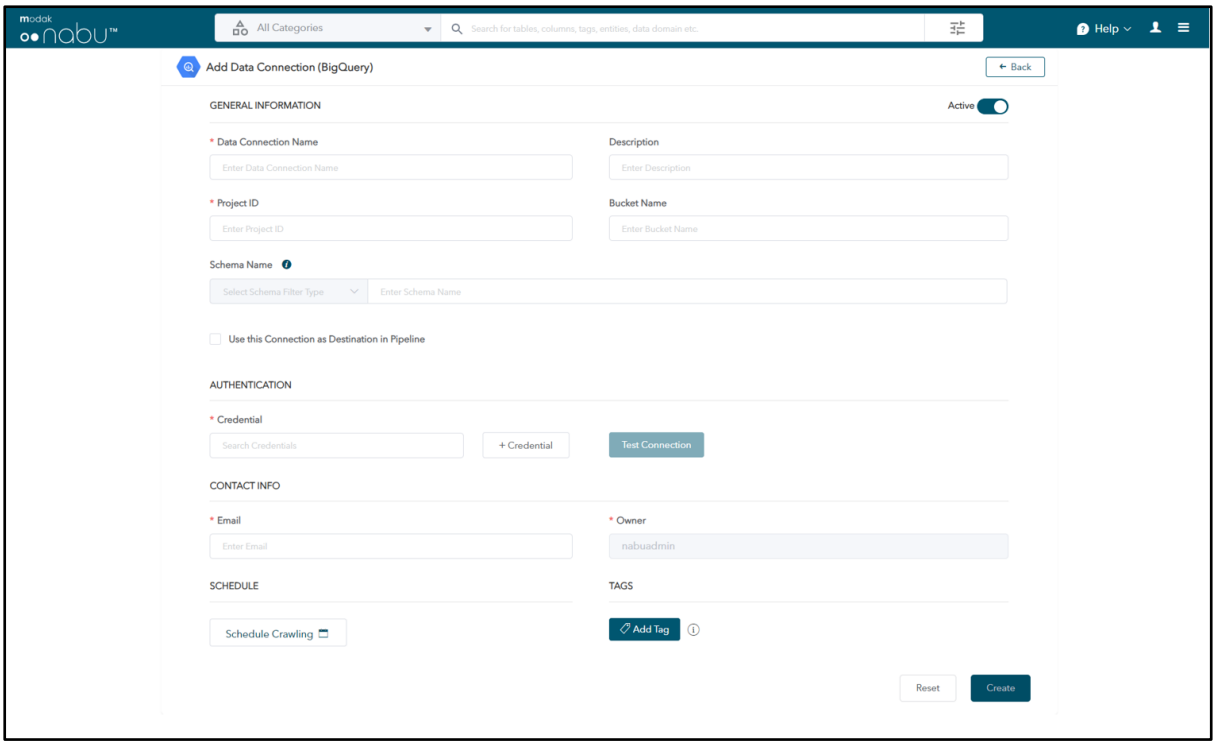


Figure 30: Big Query Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Big Query data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Project ID	This is a mandatory field where the user needs to enter the Project ID to connect to the Big Query database. An invalid data lake name will throw an error. It should contain at least 6 characters and numeric. The name cannot exceed 30 characters. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
4	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and

		the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
5	Bucket Name	The user needs to enter the bucket name to connect to the Big Query. An invalid bucket name will throw an error. It should contain at least 3 characters. The name cannot exceed 255 characters.
6	Ingestion Root Path	This is a mandatory field when user checks 'Use this Connection as Destination in Pipeline' checkbox. Ex: bucket/directory. By default, this checkbox is unchecked.
7	Credential	The user must select the credentials to connect to the Big Query database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule a crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create the tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.2.11 Add Redshift Relational Data Connection

To add a Redshift Relational data connection, click on the Redshift Relational icon from the left pane of the 'Data connections' screen. The Add Redshift Relational Data Connection form opens, and the user is required to provide the below details.

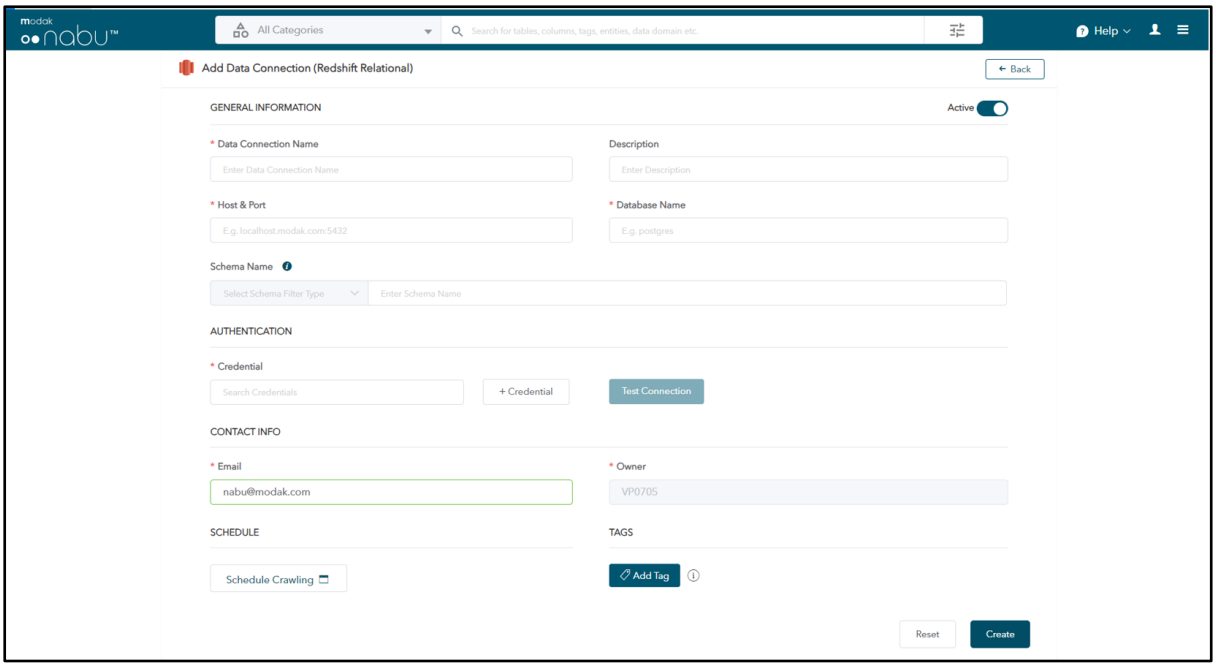


Figure 31 : Add Redshift Relational Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Redshift Relational data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to Redshift Relational. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Redshift.
5	Schema Name	This is an optional field where the user can select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.

6	Credential	The user must select the credentials to connect to Redshift relational. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port, and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page

3.1.2.12 Add Snowflake Data Connection

To add a Snowflake data connection, click on the snowflake icon from the left pane of the 'Data connections' screen. The Add Snowflake data connection form opens, and the user is required to provide the below details.

Data Connection Information:

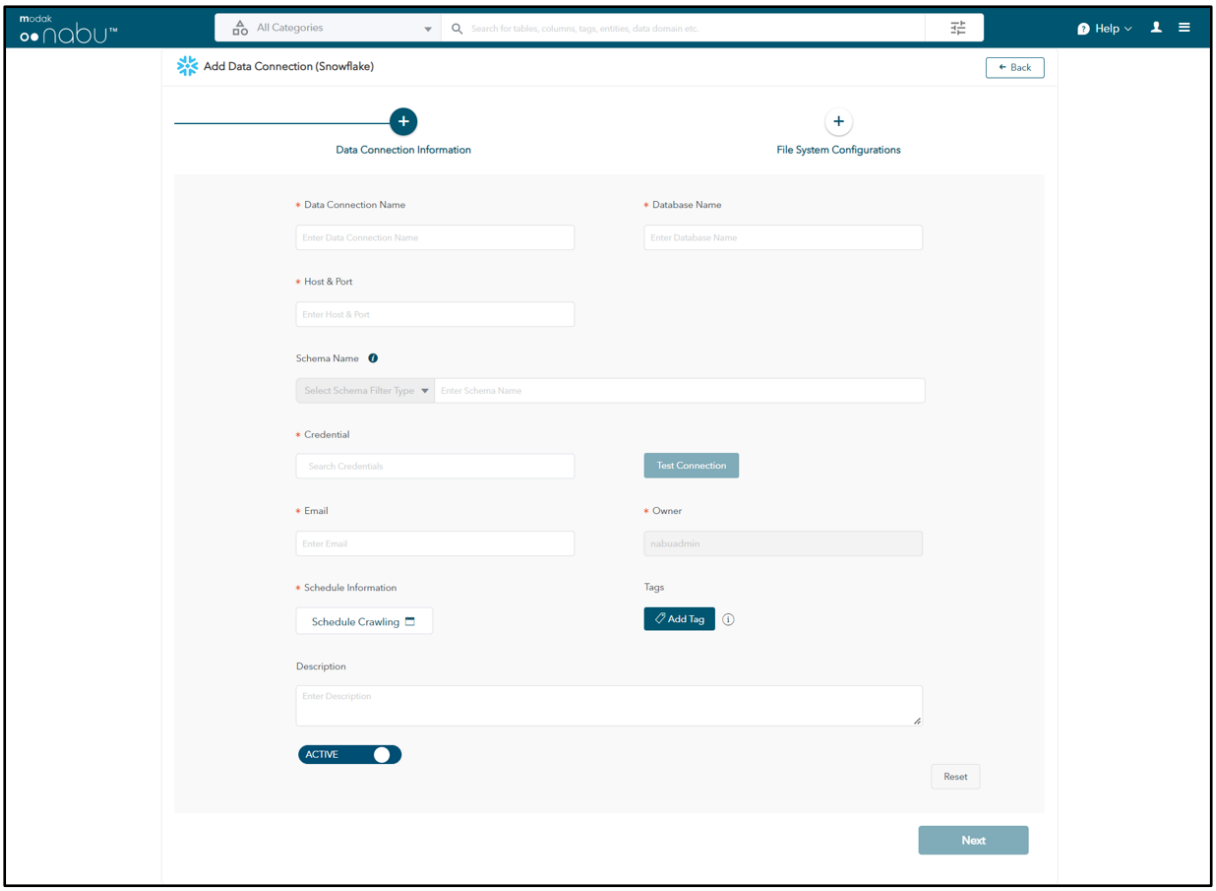


Figure 32 : Add Snowflake - Data Connection Information

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a snowflake data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
3	Email	This field is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
4	Host & Port	Host & Port details to connect to snowflake data connection.
5	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Snowflake
6	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause

		Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
7	Credential	The user can select the credential from the credential drop down
8	Crawl Frequency	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
9	Description	This is an optional field where user can enter any information to add to the data connection.
10	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – host and port, and credentials.

Click on the **next** button to move on to the next section of the form.

The Next Button is disabled by default and will be enable after successful test connection.

File System configuration:

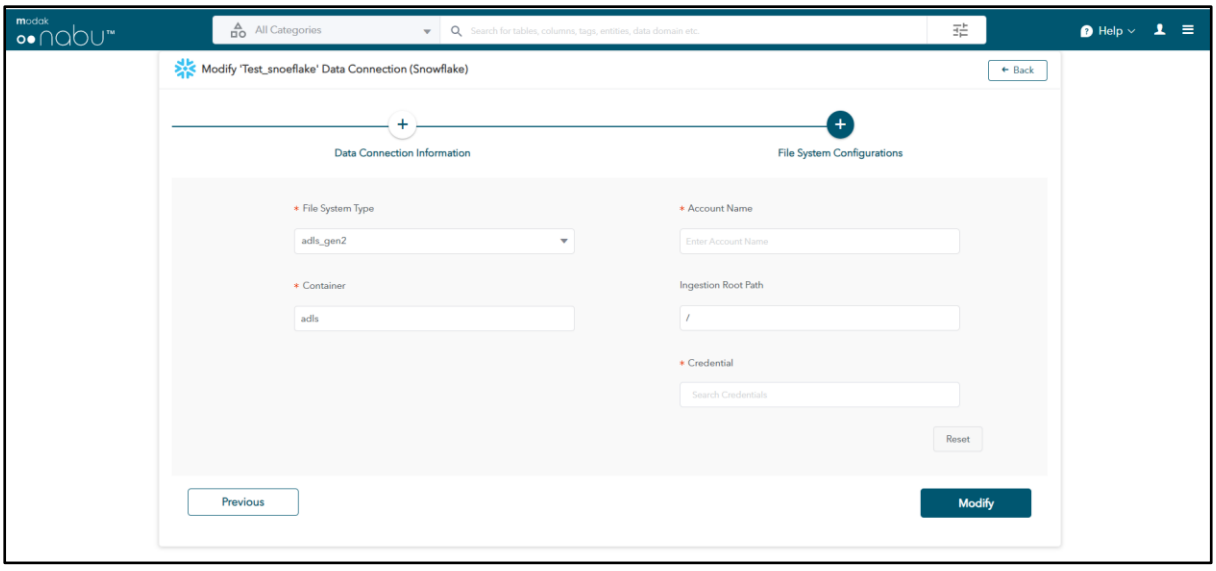


Figure 33 : Add Snowflake Data Connection – File system Configuration

S.no	Tag	Description
1	File System Type	States the file system type. User can choose the required file system type. Ex: S3, ADLS Gen2.
2	Credential	The user must select the credential from the credential's dropdown.
3	Bucket Name	This is a mandatory field when the user chooses S3 file system, and the user needs to enter the bucket name to connect to the Amazon S3 data connection. An invalid bucket name will throw an error. This bucket name should start with an alphabet. It

		should contain at least 3 characters. The bucket name cannot exceed 255 characters.
4	Container	This is a mandatory field where the user chooses ADLS File system type and the user needs to enter the container name for the snowflake
5	Account Name	The is a mandatory field where the user needs to enter the account name to connect to the Snowflake. An invalid account name will throw an error. It should contain at least 3 characters. Special characters are not allowed.
6	Ingestion Root Path	This is an optional field where objects are ingested, and the user needs to enter the ingestion root path. Ex: /root path
7	Create/Modify Button	The user can create the data connection with all the above details provided.

Click on the **Previous** button to move on to the previous section of the form.

Click on **Modify** button to make required changes to the existing data connection.

Use **back** button which is on the top right corner of the window to go back to home page

3.1.3 FILESHARES

3.1.3.1 Add File System data connection

To add File a System data connection, click on the File System icon from the left pane of the 'Data connections' screen. The Add File System Connection form opens, and the user is required to provide the below details.

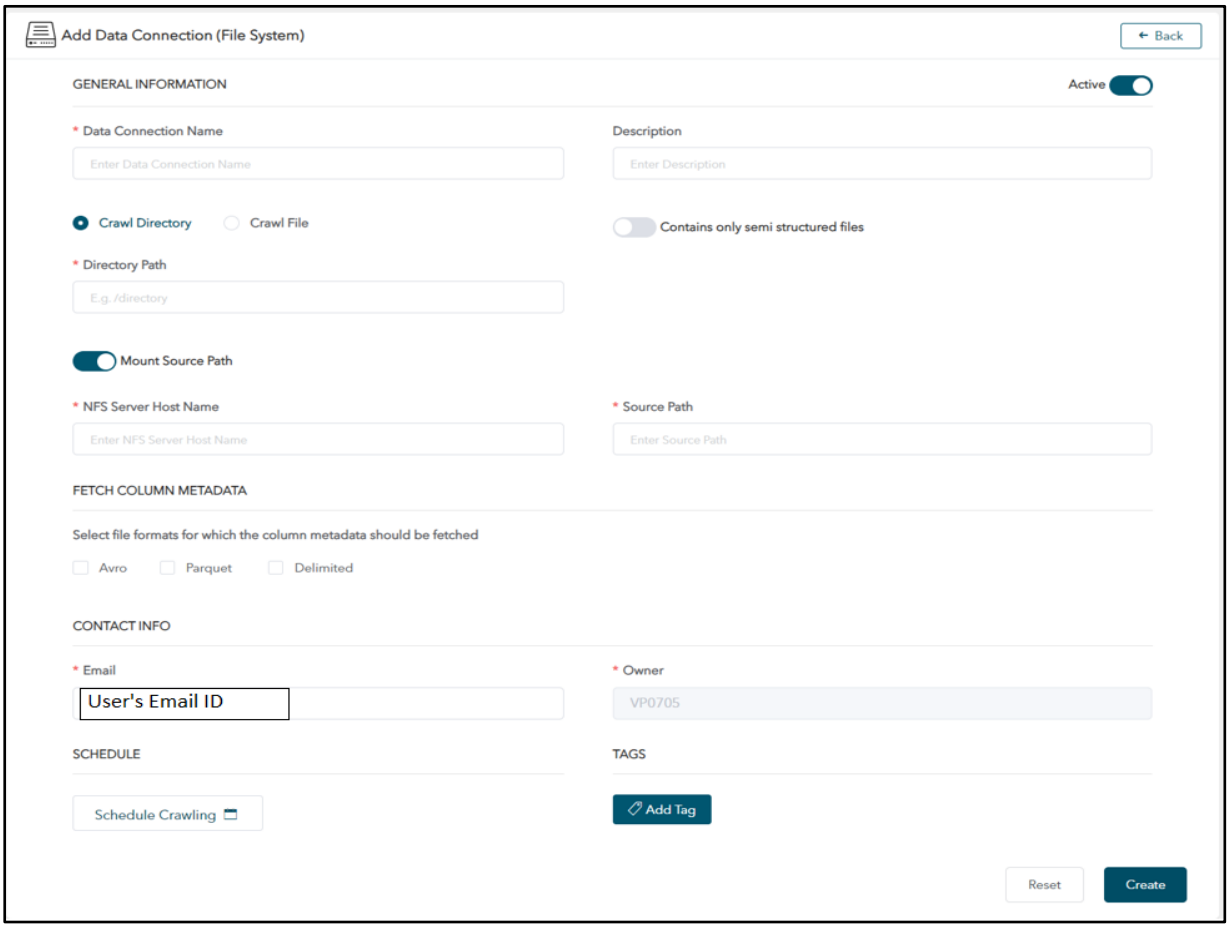


Figure 34: File System Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a File System data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Directory Path	1. This is a mandatory field where the user needs to enter the directory path to connect to the File System data connection. Ex:/directory. An invalid directory path will throw an error. The directory path does not accept few special characters. a. By default, the **Crawl Directory** option is selected. As per the use case, the user can select any of the below. b. **Crawl Directory:** This option enables users to crawl the directory c. **Crawl File:** This option enables users to crawl the file. 2. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi-structured files.
4	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: directory/file
5	Mount source path	By default, this switch is turned off. The user can turn on to mount the source path. To mount the source path, the user needs to provide the below details. NFS Server Host Name: The user needs to provide the valid host name. Source Path: The source path details need to be provided in this field.
6	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro, Parquet and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
7	Email	This field is auto filled with the user's email when the user opens this form. This field is to contact the user for any issues related to this data connection.
8	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
9	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the scheduled date and time for the crawl. For scheduling, please refer to section: 3.1.9
10	Add Tag	The user can create a tag for the data connection by clicking on

		Add Tag. For Add Tag please refer to section:3.1.10
11	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.3.2 Add SMB data connection

To add SMB data connection, click on the SMB icon from the left pane of the 'Data connections' screen. The Add SMB Connection form opens, and the user is required to provide the below details.

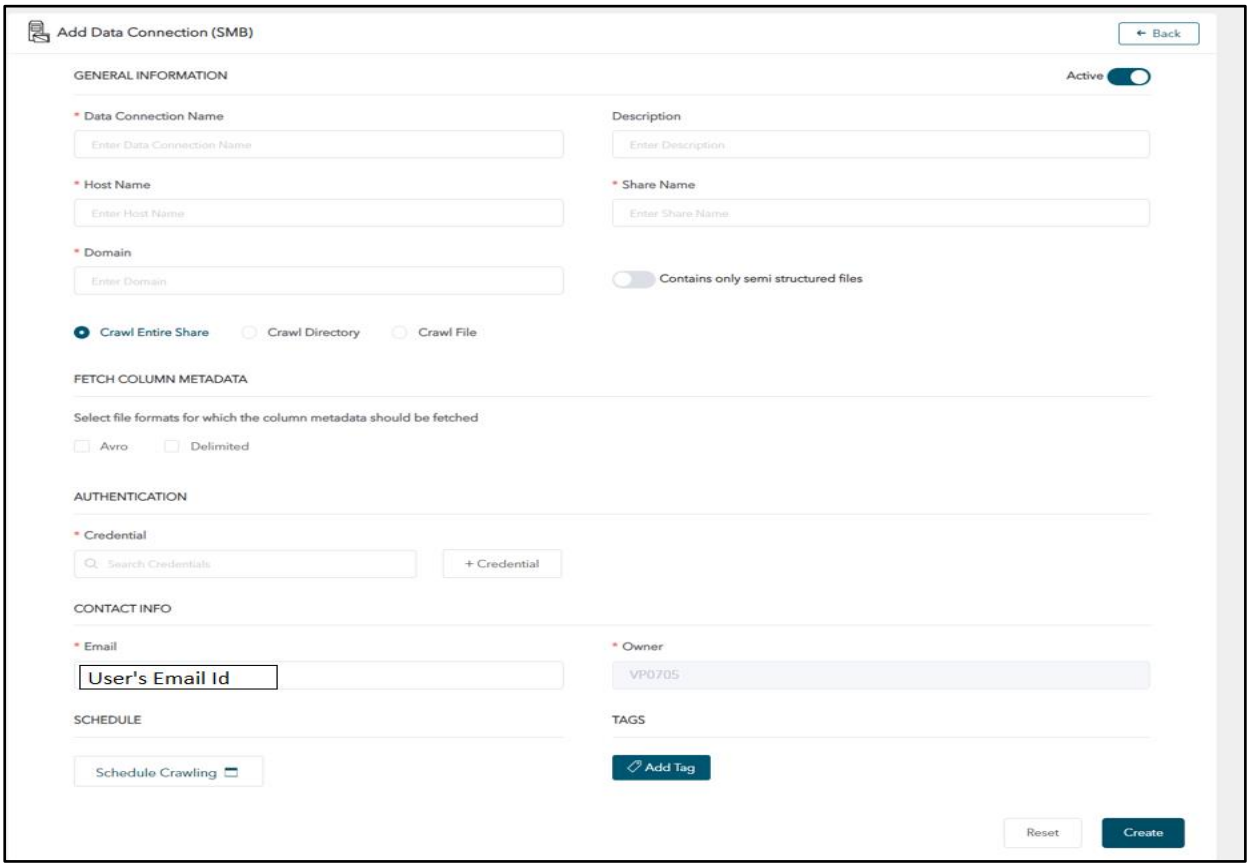


Figure 35: SMB Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a SMB Data connection. A unique Data connection name must be entered here. The Data Connection name should start with an alphabet, should contain at least 3 alphabets, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host Name	This is a mandatory field where the user needs to enter the host name to connect to the SMB data connection. Ex: w3.test.modak.com. An invalid host name will throw an error.
4	Share Name	This is a mandatory field where the user needs to enter the share name to connect to the SMB data connection. An invalid

		host name will throw an error. The share name should start with an alphabet, should contain at least 3 characters.
5	Domain	The user needs to enter the domain details in this field. Ex: modak.com. 1. By default, the **Crawl Entire Share** option is selected where the entire share will be crawled. As per the use case, the user can select any of the below. a. **Crawl Directory:** This option enables users to crawl the directory b. **Crawl File:** This option enables users to crawl the file. 2. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
6	Directory path	This is a mandatory field when user selects Crawl Directory radio option. The user should provide the directory path to crawl the directory. Ex: /directory
7	File path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: directory/file
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the SMB data connection. If the credential could not be fetched or if the user would like to add a new credential, the user can click on the **credential** button, which will redirect to the Manage Credentials page where the user can add a new credential.
10	Email	This field[ii] is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule the **crawl frequency** for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for the crawl. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.3.3 Add SFTP data connection

To add SFTP data connection, click on the SFTP icon from the left pane of the 'Data connections' screen. The Add SFTP Connection form opens, and the user is required to provide the below details.

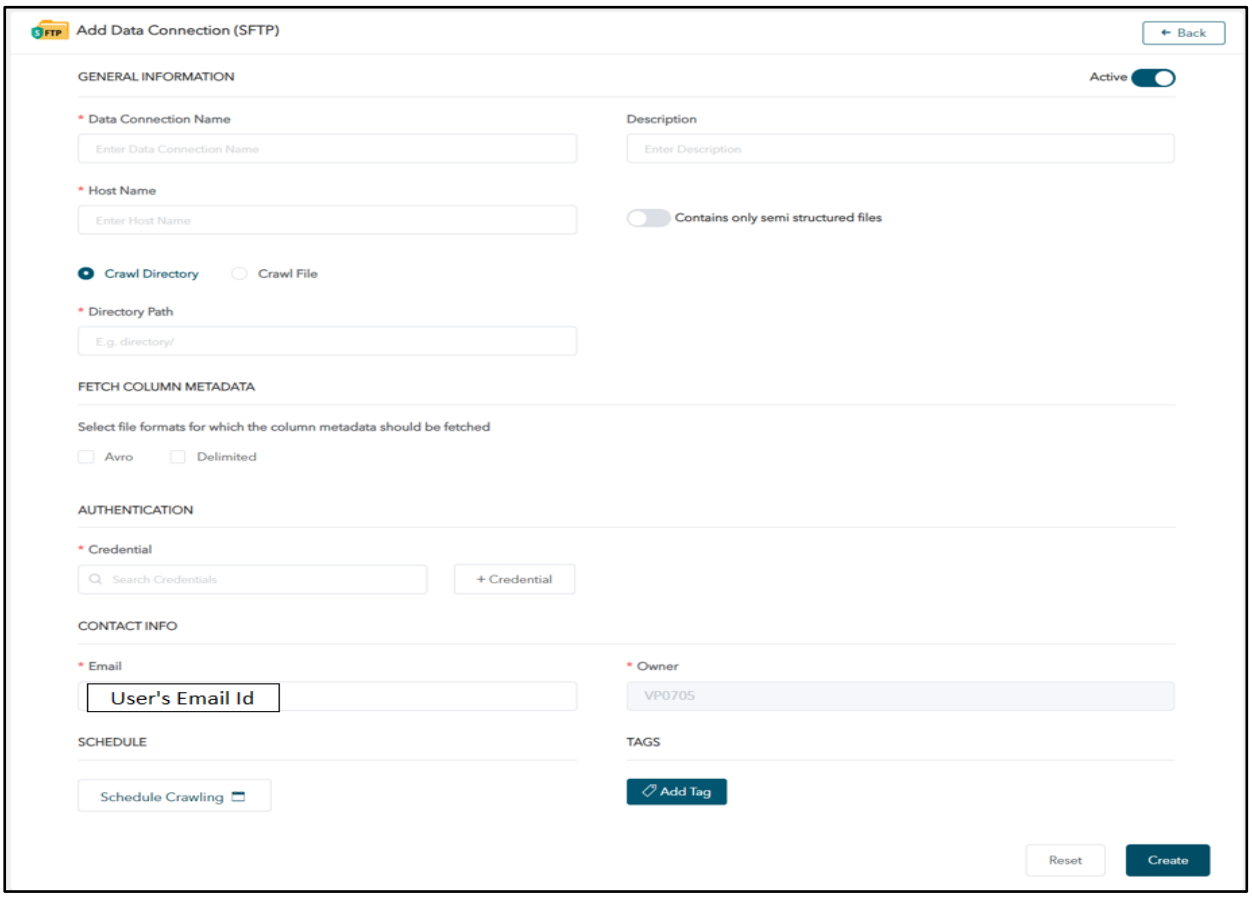


Figure 36: SFTP Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a SFTP data connection. A unique data connection name must be entered here. The Data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host Name	1. This is a mandatory field where the user needs to enter the host name to connect to the SFTP data connection. Ex: w3.test.modak.com. An invalid host name will throw an error. a. By default, the **Crawl Directory** option is selected. The user needs to provide the **directory path** for this option. Ex: directory/ b. **Crawl File:** This option enables users to crawl the file. The user needs to provide the file path for this option.

		2. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
4	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: directory/
5	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For Delimited option, it fetches for CSV, TSV. User can also specify the number of records to scan. By default, 100,000 is selected to read the records from the file and detect the data type of the columns. The user can select any value as per the preference.
6	Credential	The user must select the credentials to connect to the SFTP data connection. If the credential could not be fetched or if the user would like to add a new credential, the user can click on the **credential** button which will redirect to the Manage Credentials page where the user can add a new credential.
7	Email	This field[iii] is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
8	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
9	Schedule Crawling	The user can schedule a crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for the crawling. For scheduling, please refer to section: 3.1.9
10	Add Tag	The user can create the tag for the particular data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
11	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.3.4 Add FTP data connection

To add FTP data connection, click on the FTP icon from the left pane of the 'Data connections' screen. The Add FTP Connection form opens, and the user is required to provide the below details.

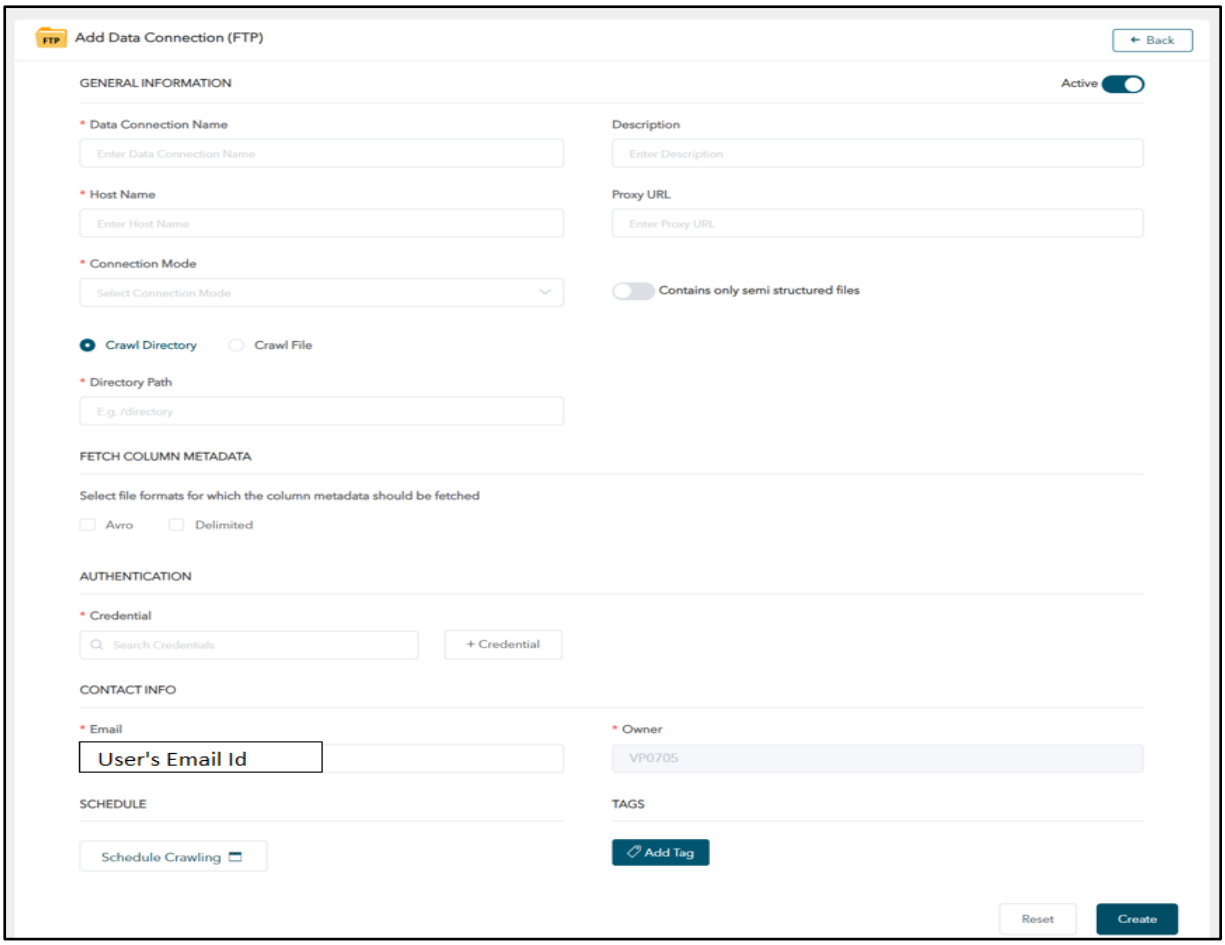


Figure 37: FTP Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to created/add an FTP Data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host Name	This is a mandatory field where the user needs to enter the host name to connect to the FTP data connection. Ex: w3.test.modak.com. An invalid host name will throw an error.
4	Proxy URL	This is an optional field where user can enter the proxy URL to connect to the FTP data connection.
5	Connection mode	This is a mandatory field where user need to provide the connection mode. Ex: active, passive 5. By default, the C**rawl Directory** option is selected. The user needs to provide the directory path for this option. Ex: directory/ 6. **Crawl File:** This option enables users to crawl the file. The user needs to provide the file path for this option.

		Contains only semi structured files This option, when switched on indicates that the directory/files selected contains only semi structured files.
6	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: directory/file
7	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For Delimited option, it fetches for CSV, TSV. User can also specify the number of records to scan. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the FTP data connection. If the credential could not be fetched or if the user would like to add a new credential, the user can click on the **credential** button which will redirect to the Manage Credentials page where the user can add a new credential.
10	Email	This field[iv]is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule a crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for the crawling. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create the tag for the particular data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.3.5 Add FTPS data connection

To add FTPS data connection, click on the FTPS icon from the left pane of the 'Data connections' screen. The Add FTPS Connection form opens, and the user is required to provide the below details.

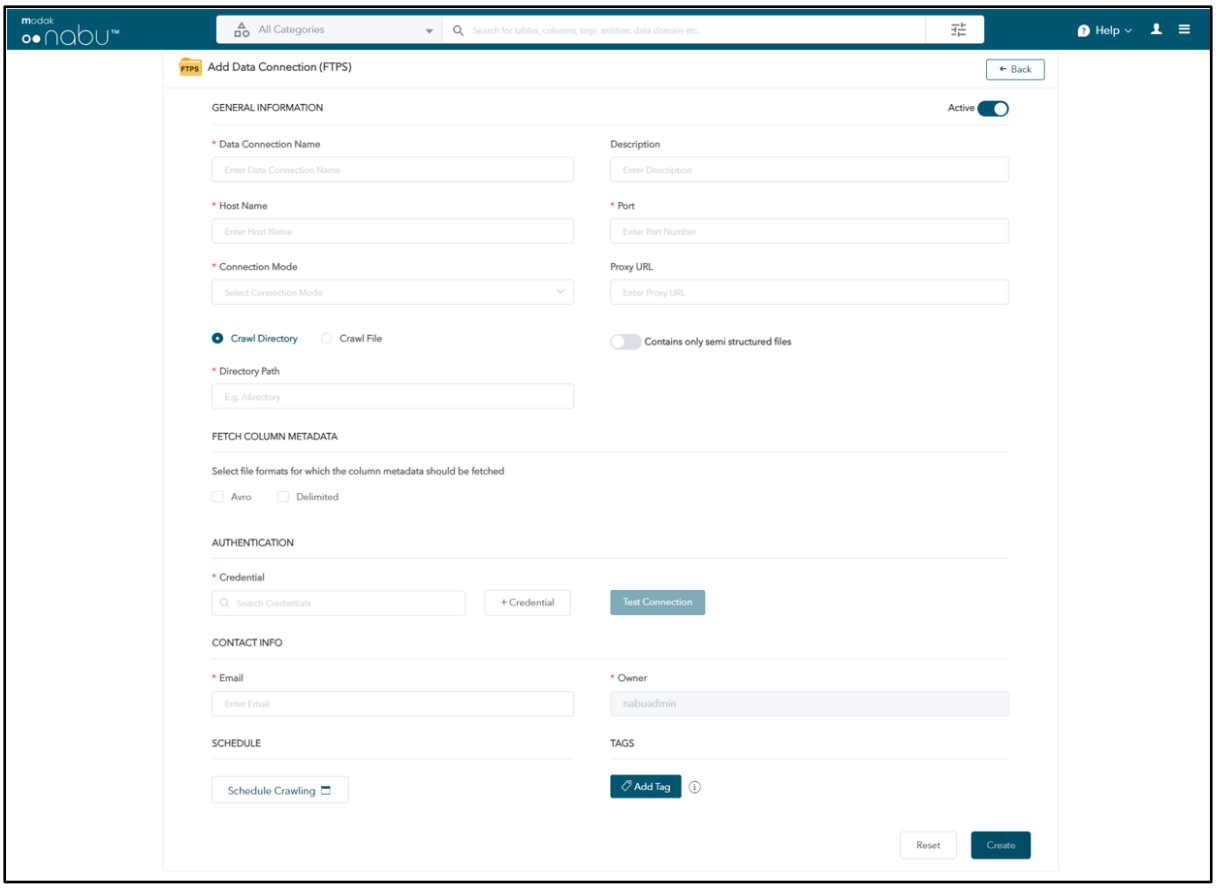


Figure 38: FTPS Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to created/add an FTPS Data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host Name	This is a mandatory field where the user needs to enter the host name to connect to the FTPS data connection. Ex: w3.test.modak.com. An invalid host name will throw an error.
4	Port	This is a mandatory field where user needs to enter the port details to connect to the FTPS
5	Proxy URL	This is an optional field where user can enter the proxy URL to connect to the FTPS data connection.
6	Connection mode	This is a mandatory field where user need to provide the connection mode. Ex: active, passive 7. By default, the C**rawl Directory** option is selected. The user needs to provide the directory path for this option. Ex: directory/

		8. **Crawl File:** This option enables users to crawl the file. The user needs to provide the file path for this option. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
7	File Path	This is a mandatory field when user selects Crawl File radio option. The user should provide the File path to crawl the file. Ex: directory/file
8	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For Delimited option, it fetches for CSV, TSV. User can also specify the number of records to scan. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
9	Credential	The user must select the credentials to connect to the FTP data connection. If the credential could not be fetched or if the user would like to add a new credential, the user can click on the **credential** button which will redirect to the Manage Credentials page where the user can add a new credential.
10	Email	This field[v]is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
11	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
12	Schedule Crawling	The user can schedule a crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for the crawling. For scheduling, please refer to section: 3.1.9
13	Add Tag	The user can create the tag for the particular data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
14	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page

3.1.4 OTHERS

3.1.4.1 Add SAS data connection

To add a SAS data connection, click on the SAS icon from the left pane of the 'Data connections' screen. The Add SAS Connection form opens, and the user is required to provide the below details.

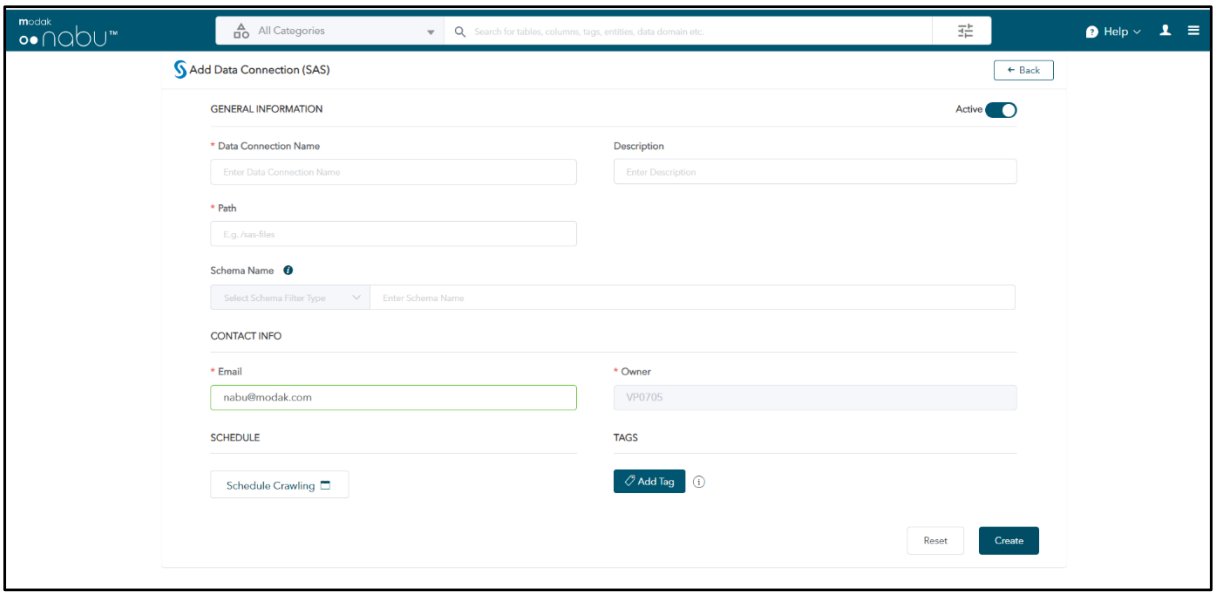


Figure 39: SAS Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add an SAS connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet and should contain at least 3 alphabets. Special characters except underscore (_) are not allowed. The name cannot exceed more than 50 characters.
2	Path	This is a mandatory field where the user needs to enter Path to connect to SAS. Ex., /sas-files
3	Credential	The user must select the credentials to connect to the SAS database. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button, which will redirect you to the Manage Credentials page where you can add a new credential.
4	Schema Name	Select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause • Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
5	Contact Info	This field is auto filled with the user's email when the user opens this form. This field is to contact user for any issues related to this data connection.

6	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
7	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
8	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
9	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.4.2 Add MongoDB data connection

To add a MongoDB data connection, click on the MongoDB icon from the left pane of the 'Data connections' screen. The Add MongoDB Connection form opens, and the user is required to provide the below details.

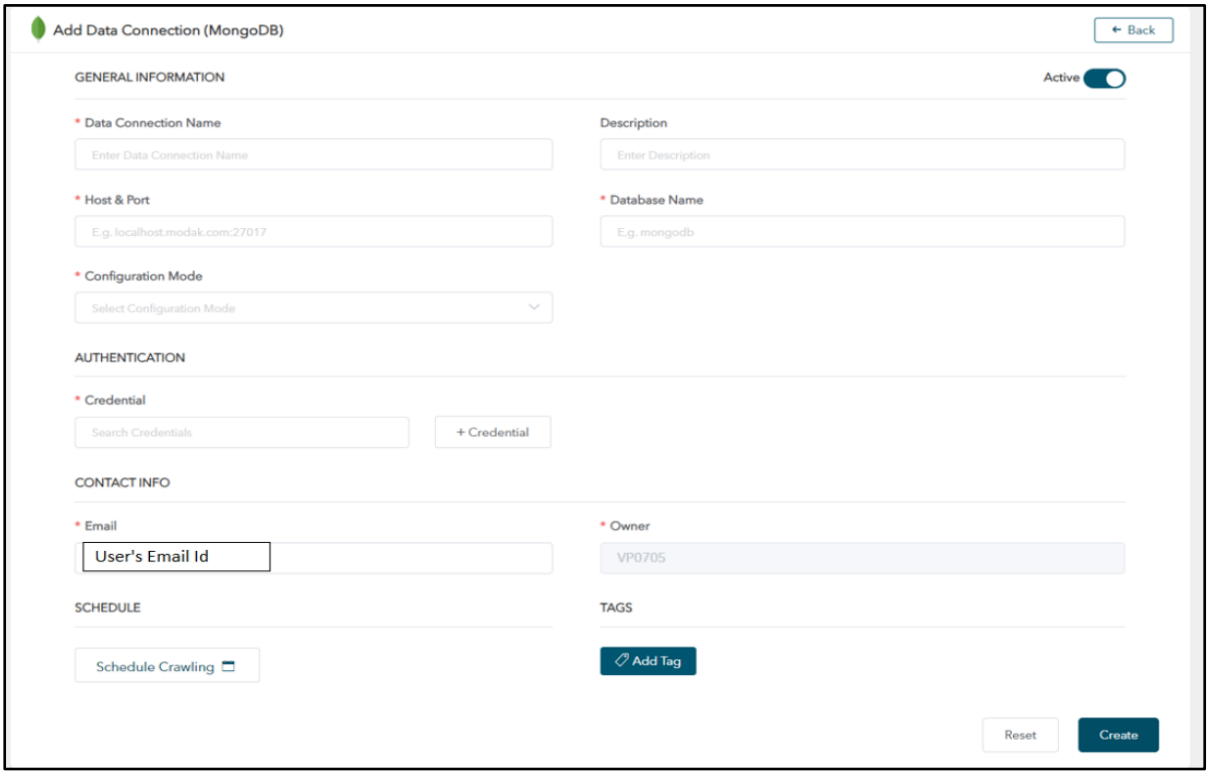


Figure 40: MongoDB Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to created/add a MongoDB connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.

2	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the MongoDB database. Ex: localhost.modak.com:27017. Invalid Host & Port details will throw an error.
3	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: MongoDB
4	Configuration Mode	This is a mandatory filed where it enables the user to select configuring mode from the dropdown.
5	Email	This is a mandatory field where user needs to enter the email to connect to. This field is to contact for any issues related to this data connection.
6	Credential	The user must select the credentials to connect to the MongoDB database. If the credential could not be fetched or would like to add any new credential, the user can click on **credential** button, which will redirect to the Manage Credentials page where user can add a new credential.
7	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
8	Schedule Crawling	The user can schedule a crawling for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
9	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
10	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.4.3 Add HTTP data connection

To add Http data connection, click on the Http icon from the left pane of the 'Data connections' screen. The Add Http Connection form opens, and the user is required to provide the below details.

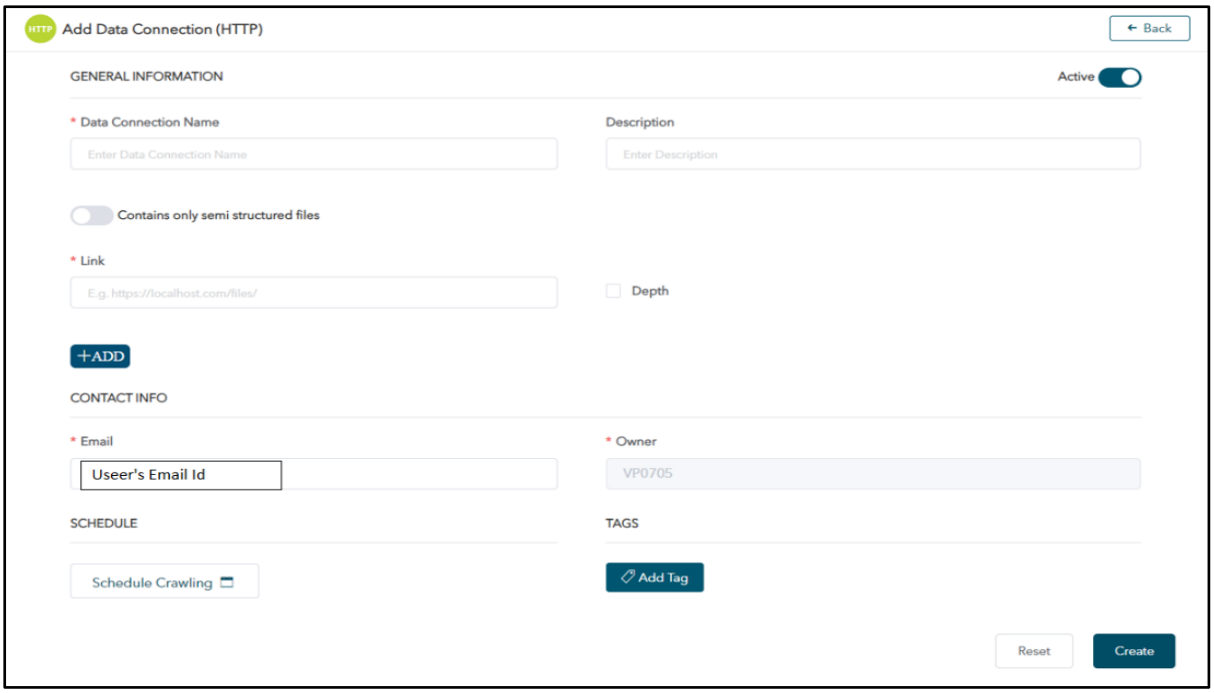


Figure 41: HTTP Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to created/add a Http Data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters., special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Link	This is a mandatory field where the user needs to enter the link to connect to the Http data connection. Ex: https://localhost.com/files. An invalid link will throw an error.
4	Depth	This is a checkbox provided to the user where the user can check/uncheck to include all the links (URLs) which are present in the above link.
5	Email	This field[vi]is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
6	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
7	Schedule Crawling	The user can schedule a crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for the crawling. For scheduling, please refer to section: 3.1.9
8	Add Tag	The user can create the tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10

9	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.4.4 Add Salesforce data connection

To add a Salesforce data connection, click on the Salesforce icon from the left pane of the 'Data connections' screen. The Add Salesforce Connection form opens, and the user is required to provide the below details.

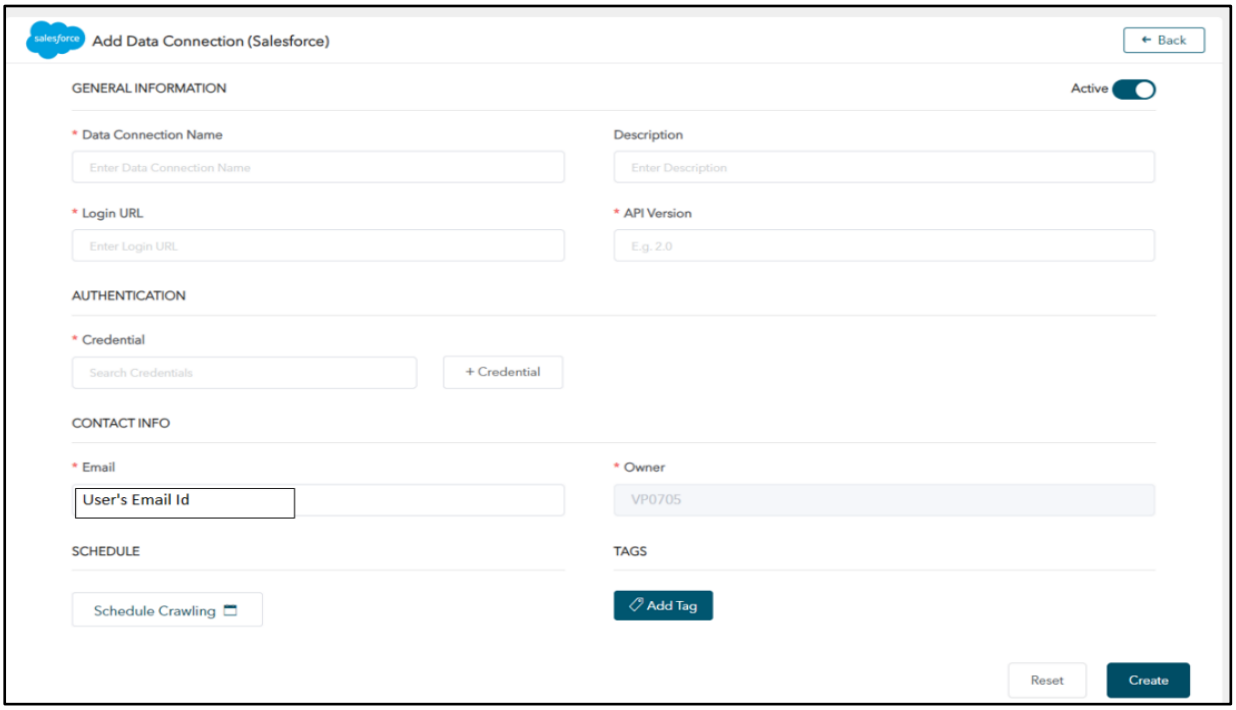


Figure 42: Salesforce Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a salesforce data connection. A unique data connection name must be entered here. The data connection name should start with an alphabet and should contain at least 3 characters. Special characters except underscore (_) are not allowed. The name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Login URL	This is a mandatory filed where the user needs to enter login URL e.g., https://login.salesforce.com
4	API Version	This is a mandatory field where user needs to enter version of salesforce API.
5	Credential	The user must select the credentials to connect to the salesforce database. If the credential could not be fetched or would like to add any new credential, the user can click on **credential** button which will redirect to the Manage Credentials page where user can add a new credential.

6	Email	This field[vii] is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
7	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
8	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
9	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
10	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.4.5 Add SharePoint data connection

To add SharePoint data connection, click on the SharePoint icon from the left pane of the 'Data connections' screen. The Add SharePoint data Connection form opens, and the user is required to provide the below details.

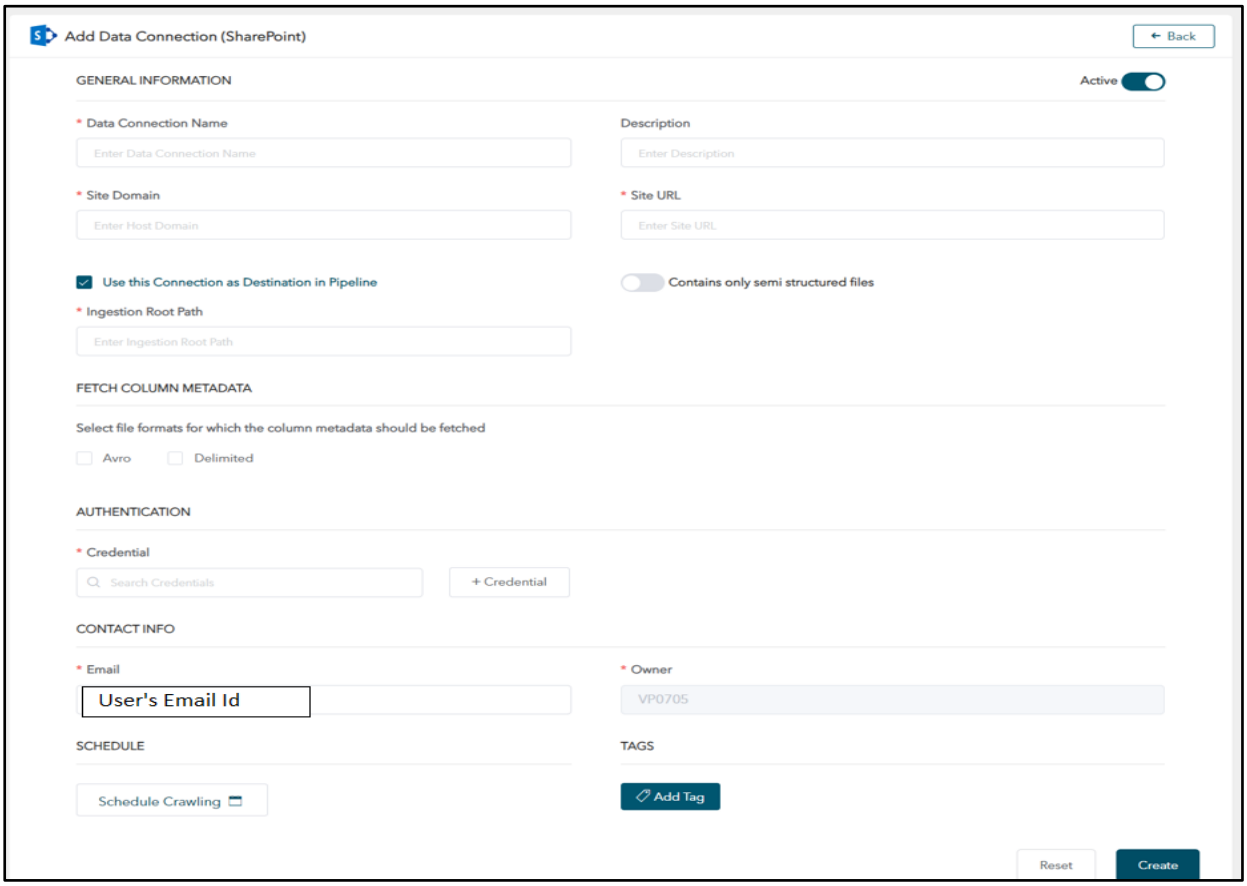


Figure 43: SharePoint Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a SharePoint data connection. A unique data connection name must be entered here. The Data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Site Domain	This is a mandatory field where the user needs to enter the site domain details to connect to the SharePoint data connection. Ex: modakanalytics0.sharepoint.com. An invalid site domain will throw an error.
4	Site URL	This is a mandatory field where user needs to enter the site URL for the domain specified above. An invalid site URL will throw an error “Invalid Site URL”.
5	Ingestion Root Path	This is a mandatory field where user need to provide the Ingestion root path when the checkbox ‘Use this Connection as Destination in Pipeline’ is checked. By default, this checkbox is unchecked. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
6	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default, 100,000 is selected to read the records from the file and detect the data type of the columns.
7	Credential	The user must select the credentials to connect to the SharePoint data connection. If the credential could not be fetched or if the user would like to add a new credential, the user can click on the **credential** button which will redirect to the Manage Credentials page where the user can add a new credential.
8	Email	This field[viii] is auto filled with the user’s email when user opens this form. This field is to contact user for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create / Reset	**Create** The user can click on create button to add the above data connection.

		Reset This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page.

3.1.4.6 Add Documentum Data Connection

To add a Documentum data connection, click on the Documentum icon from the left pane of the 'Data connections' screen. The Add Documentum Data Connection form opens, and the user is required to provide the below details.

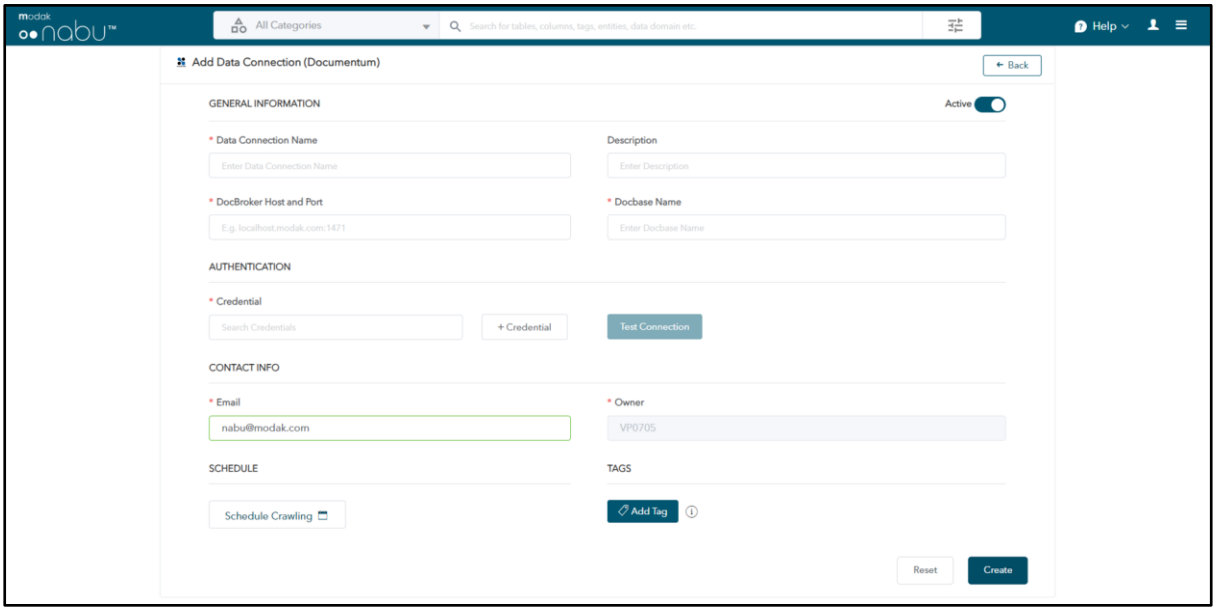


Figure 44 : Add Documentum Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a Documentum data connection. A unique data connection name must be entered here. The data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the data connection name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Host & Port	This is a mandatory field where user needs to enter the host and port details to connect to the Documentum database. Ex: localhost.modak.com:1471. Invalid Host & Port details will throw an error.
4	Database Name	This is a mandatory field where user needs to enter the database name to connect to. Ex: Documentum.
5	Schema Name	This is an optional field where the user can select schema filter type from the below options and provide the schema name. • Inclusive Regex • Exclusive Regex • Inclusive Like • Exclusive Like • Where IN Clause

		• Where NOT IN Clause Based on the selected filter type, the schemas are filtered, and the filtered schemas will be included as part of that data connection. Ex: Inclusive like-> schema Name: clinical The schemas inclusive of 'clinical' name will only be crawled for the data connection.
6	Credential	The user must select the credentials to connect to Documentum. If the credential could not be fetched or you would like to add a new credential, the user can click on the **credential** button which will redirect you to the Manage Credentials page where you can add a new credential.
7	Test Connection	This button is disabled by default and once the above details are entered, the button gets enabled, and the user can test the connection to ensure Nabu is able to connect to the data connection. If the test connection fails, check the details – hostname, port, and credentials.
8	Email	This field is auto filled with the user's email when user opens this form. This field is to contact for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create/Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page

3.1.4.7 Add SharePoint Subsite Data Connection

To add a SharePoint Subsite data connection, click on the SharePoint Subsite icon from the left pane of the 'Data connections' screen. The Add SharePoint Subsite Data Connection form opens, and the user is required to provide the below details.

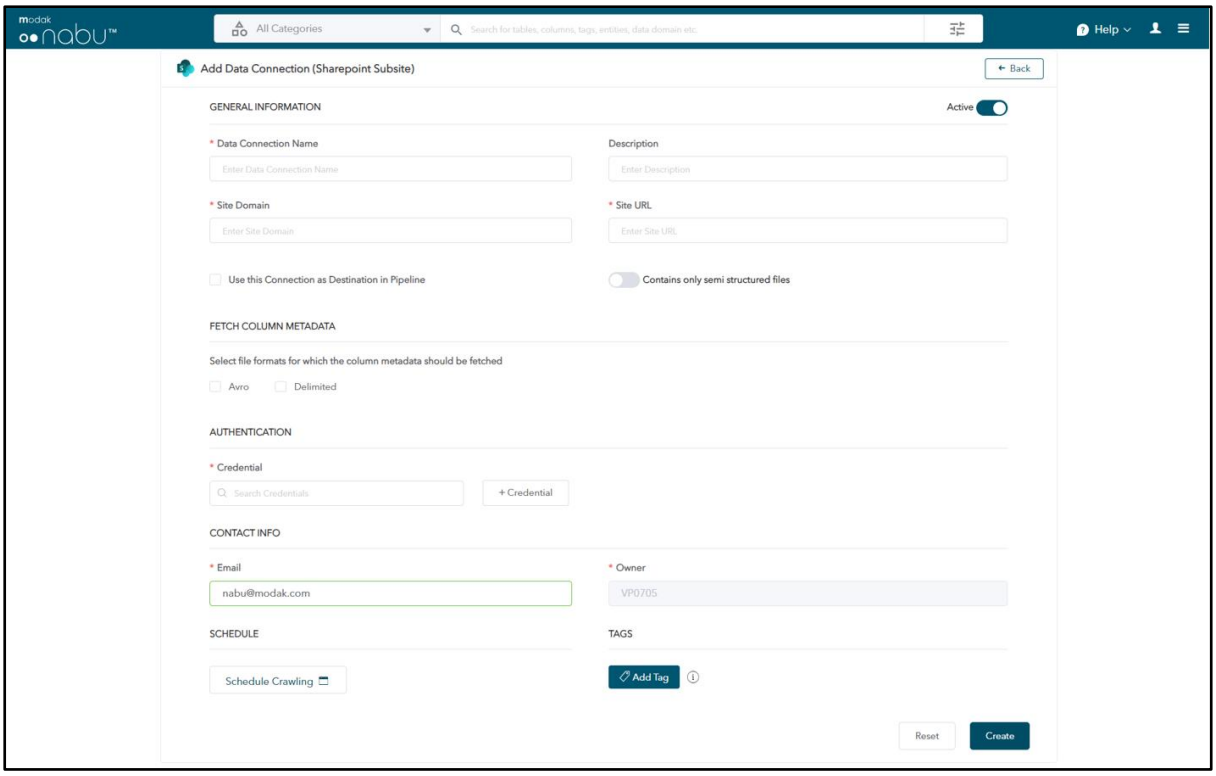


Figure 45 : Add SharePoint Subsite Data Connection

S.no	Field	Description
1	Data Connection Name	This is a mandatory field to create/add a SharePoint Subsite data connection. A unique data connection name must be entered here. The Data Connection name should start with an alphabet, should contain at least 3 characters, special characters except underscore (_) are not allowed, the name cannot exceed more than 50 characters.
2	Description	This is an optional field where user can enter any information to add to the data connection.
3	Site Domain	This is a mandatory field where the user needs to enter the site domain details to connect to the SharePoint Subsite data connection. An invalid site domain will throw an error.
4	Site URL	This is a mandatory field where user needs to enter the site URL for the domain specified above. An invalid site URL will throw an error “Invalid Site URL”.
5	Ingestion Root Path	This is a mandatory field where user need to provide the Ingestion root path when the checkbox ‘Use this Connection as Destination in Pipeline’ is checked. By default, this checkbox is unchecked. **Contains only semi structured files** This option, when switched on indicates that the directory/files selected contains only semi structured files.
6	Fetch Column Metadata	The user can select the file formats to fetch the column meta data. The options provided are Avro and delimited. For the Delimited option, it fetches CSV and TSV. The user can also specify the number of records to be scanned. By default,

		100,000 is selected to read the records from the file and detect the data type of the columns.
7	Credential	The user must select the credentials to connect to the SharePoint subsite data connection. If the credential could not be fetched or if the user would like to add a new credential, the user can click on the **credential** button which will redirect to the Manage Credentials page where the user can add a new credential.
8	Email	This field[ix]is auto filled with the user's email when user opens this form. This field is to contact user for any issues related to this data connection.
9	Owner	This indicates the person who has created the data connection. The owner id is filled by default.
10	Schedule Crawling	The user can schedule crawl frequency for this data connection. By clicking on this button, a popup modal is opened where the user can select the schedule date and time for crawling. For scheduling, please refer to section: 3.1.9
11	Add Tag	The user can create a tag for the data connection by clicking on Add Tag. For Add Tag please refer to section:3.1.10
12	Create / Reset	**Create** The user can click on create button to add the above data connection. **Reset** This button resets the Add Data connection form.

Use **back** button which is on the top right corner of the window to go back to home page

3.1.5 Edit Data Connection

The user can edit any of the above data connections from the Data connections landing page as below.

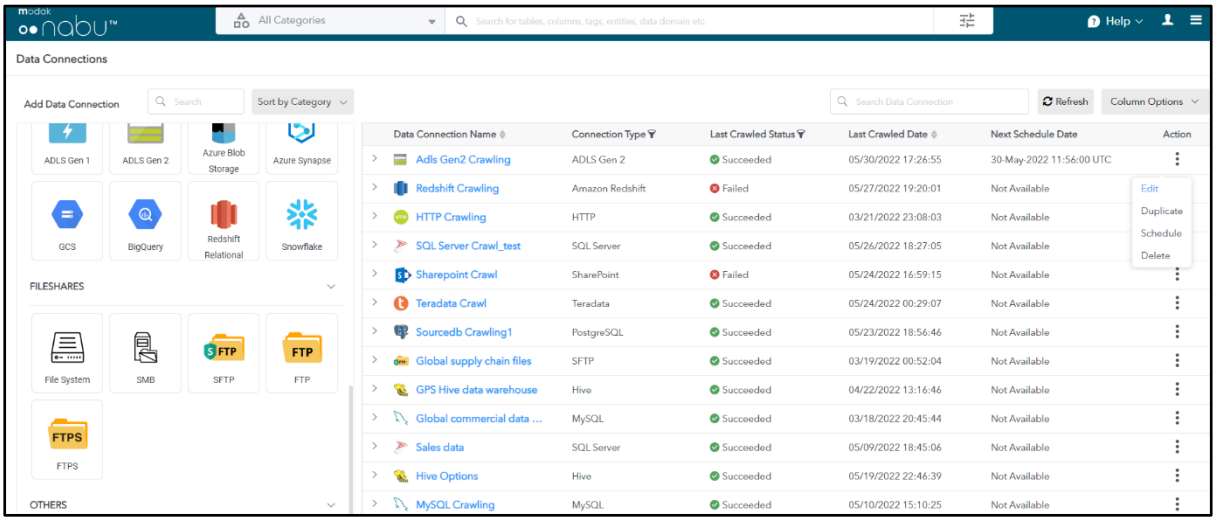


Figure 46: Edit Data Connection

The user can click on the Action column and select the Edit option or click on the Data connection name.

On selecting the edit option, the user will be redirected to the respective data connection with all the previous details. The user can edit the data connection as required and click on the modify button.

All the details entered/modified should be valid, else the user will not be able to modify the data connection.

3.1.6 Duplicate Data Connection

The user can duplicate any of the above data connections from the Data connection landing page as below.

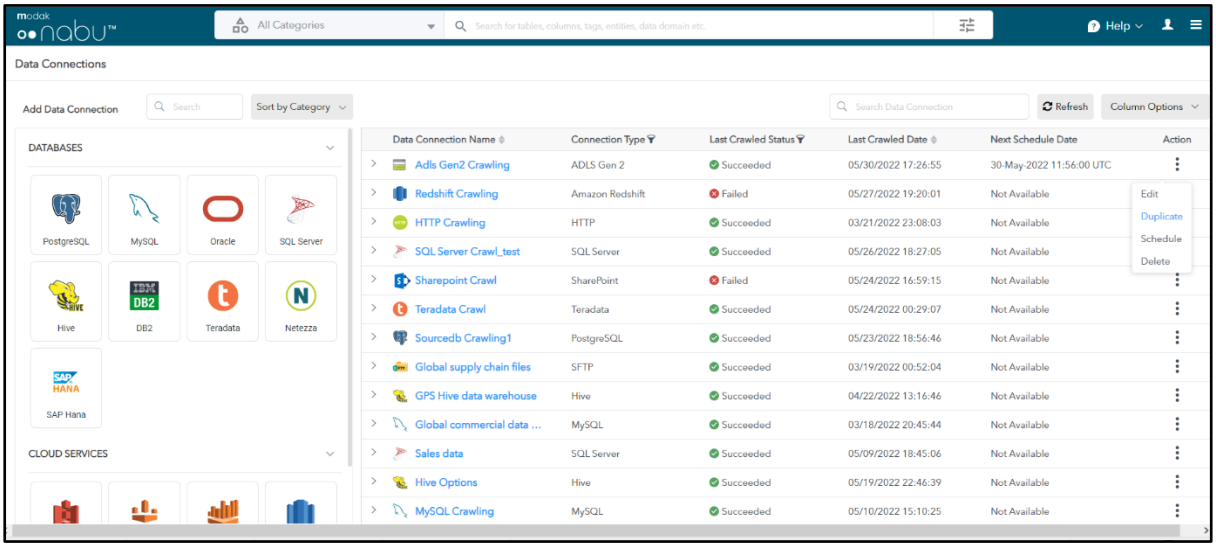


Figure 47: Duplicate Data Connection

The user can click on the Action column and select the duplicate option.

On selecting the duplicate option, the duplicate or copy of the selected data connection is created and the user will be redirected to the selected data connection with all the details. The user can take action as per their preference.

All the details entered should be valid, else the user will not be able to create the data connection.

3.1.7 Schedule Data Connection

The user can schedule crawling for any of the above data connections from the Data connections landing page as below.

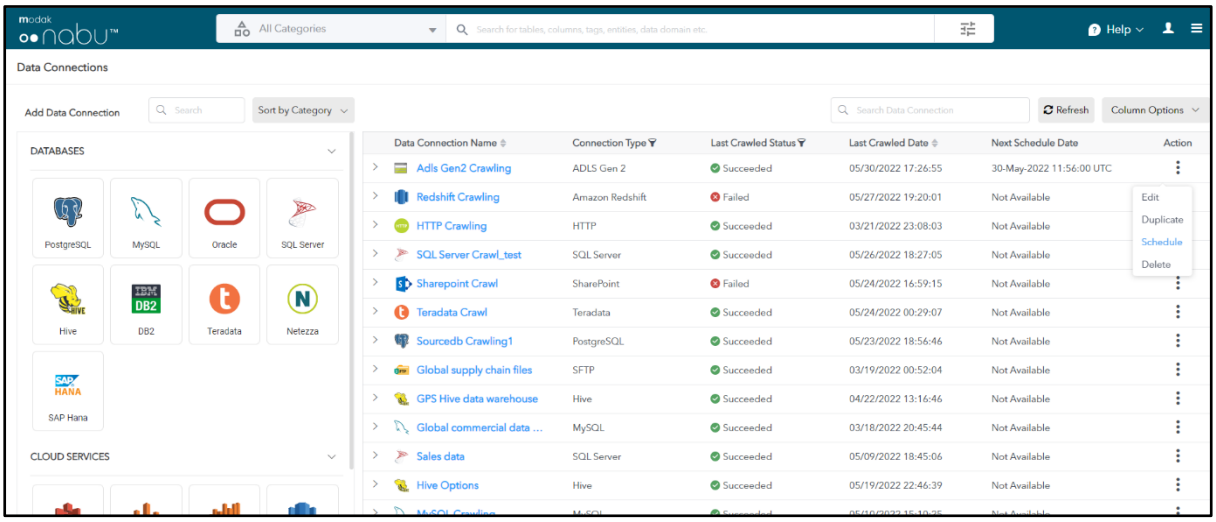


Figure 48: Schedule Data Connection

On clicking schedule, the user will be prompted with a popup as below where the user can select the crawling frequency. For scheduling, please refer to section: 3.2.10.

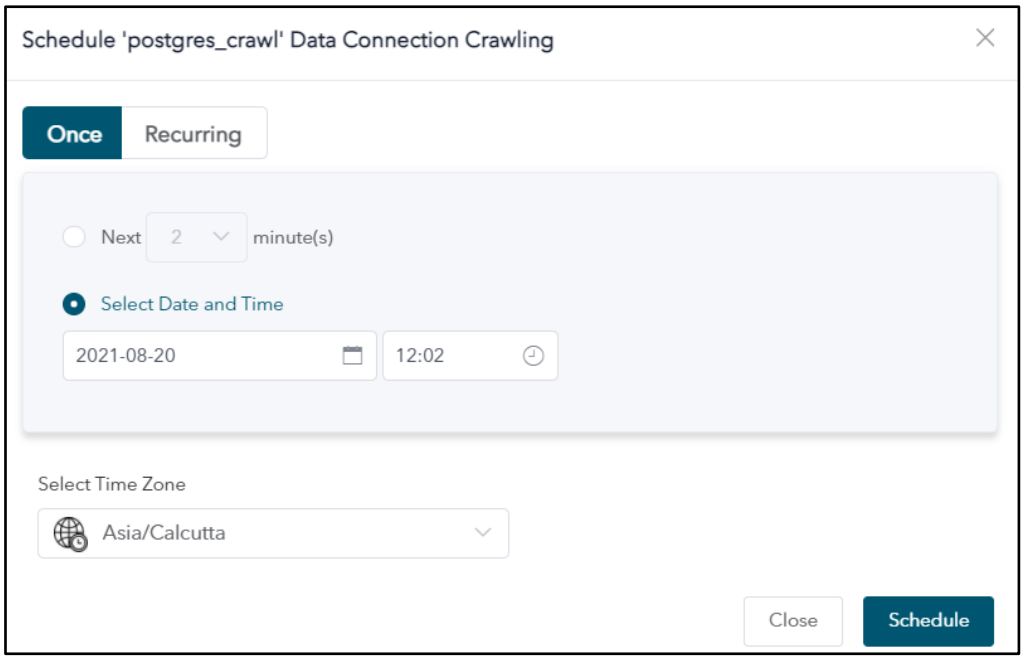


Figure 49: Schedule Data Connection Crawling

3.1.8 Delete Data connection

The user can delete any of the data connections from the data connection landing page as below.

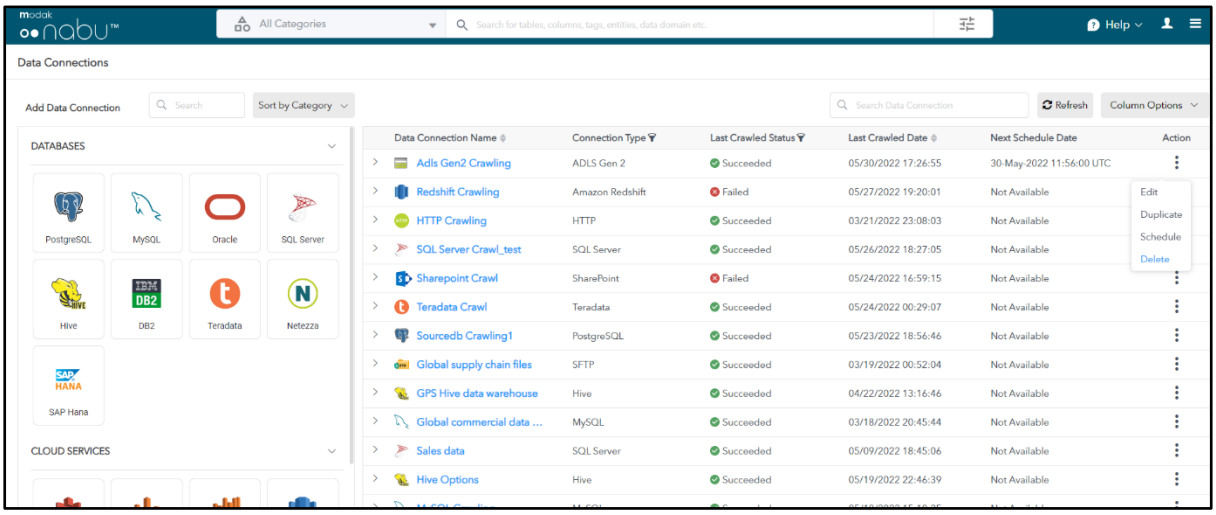


Figure 50: Delete Data Connection

The user can click on the Action column and select the Delete option.

On selecting the delete option, a popup showing the impacted data connection for the selected data connection will be shown as below.

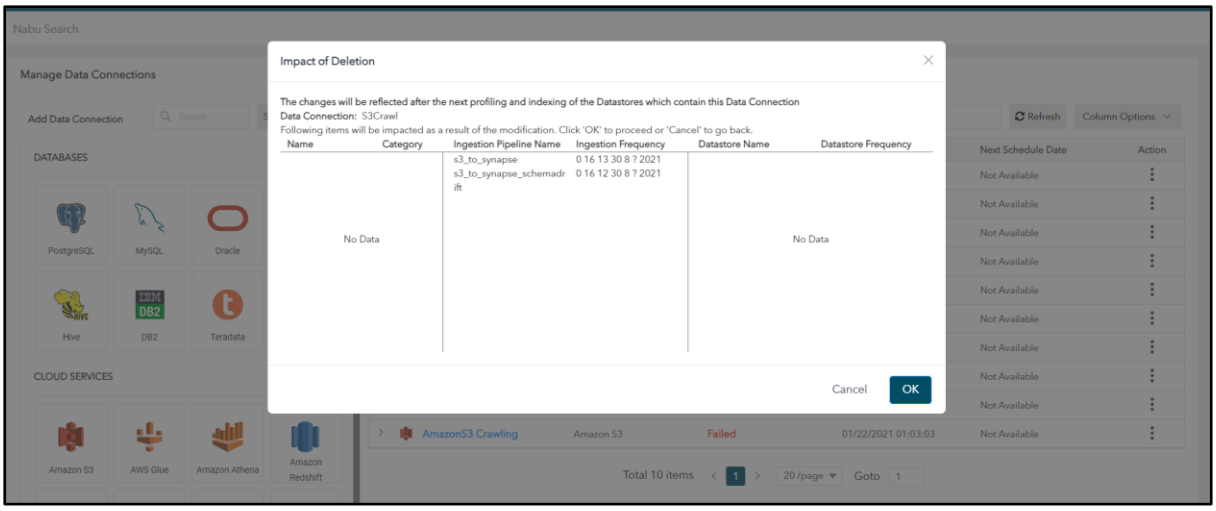


Figure 51: Impact of Deletion

If the user clicks ok, the selected data connection will be deleted

3.1.9 Scheduling

when user clicks on Schedule button, a popup modal opens where user can select the date and time for scheduling.

Once: This tab is selected by default. For One time scheduling, the user can select any of the below options.

Next ‘x’ minutes: The user can select ‘x’ minutes from the dropdown and click on save. The ‘Select Date and Time’ section is disabled when user selects this option.

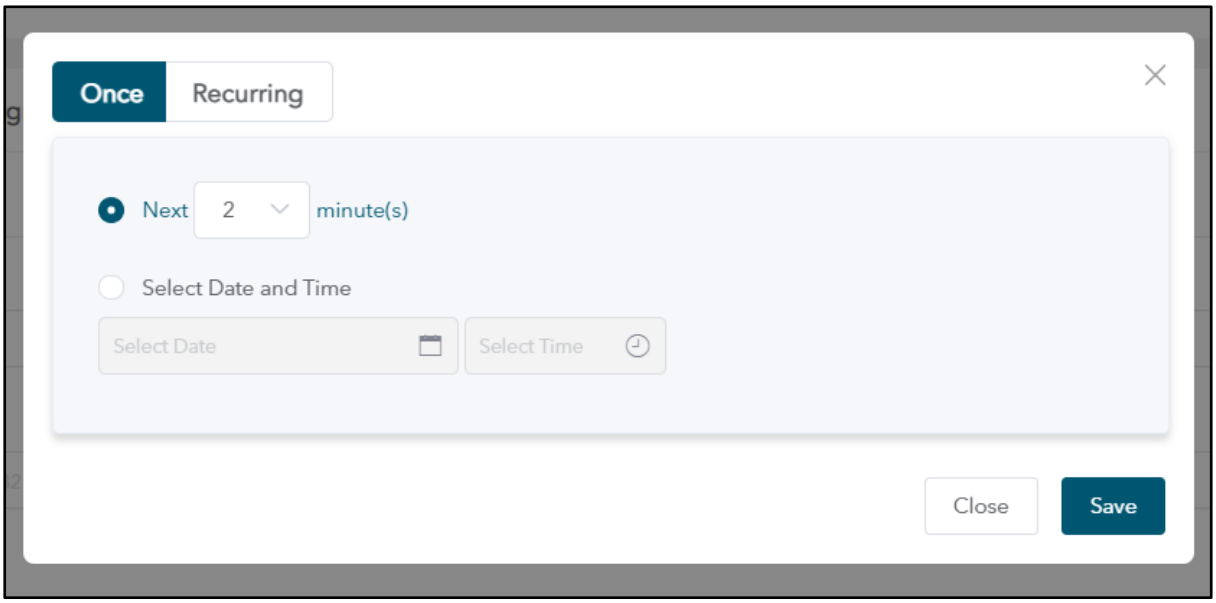


Figure 52: Scheduling

Select Date and Time: The user can select date and time from date and time pickers respectively for the scheduling and click on save. The next ‘x’ minutes section is disabled when user selects this option.

Select Time zone: By default, the dropdown shows the browser’s time zone, and the user can change the same as per their preference from the dropdown. The drop down shows the list of all the available time zones.

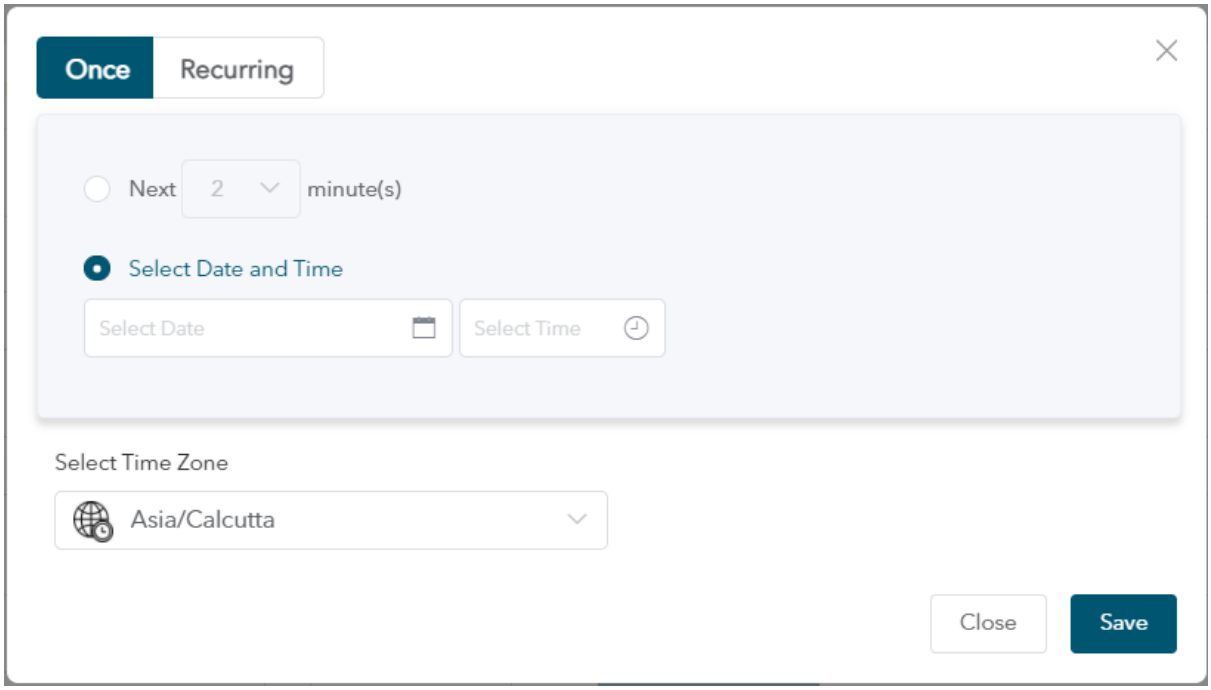


Figure 53: Scheduling Once Tab

Recurring: The user can schedule recurring jobs by switching to the Recurring tab. The recurring can be set to Daily/weekly/monthly/yearly. Daily tab is selected by default.
Daily: The user can select any of the options from the daily tab and select start and end dates and click on save.

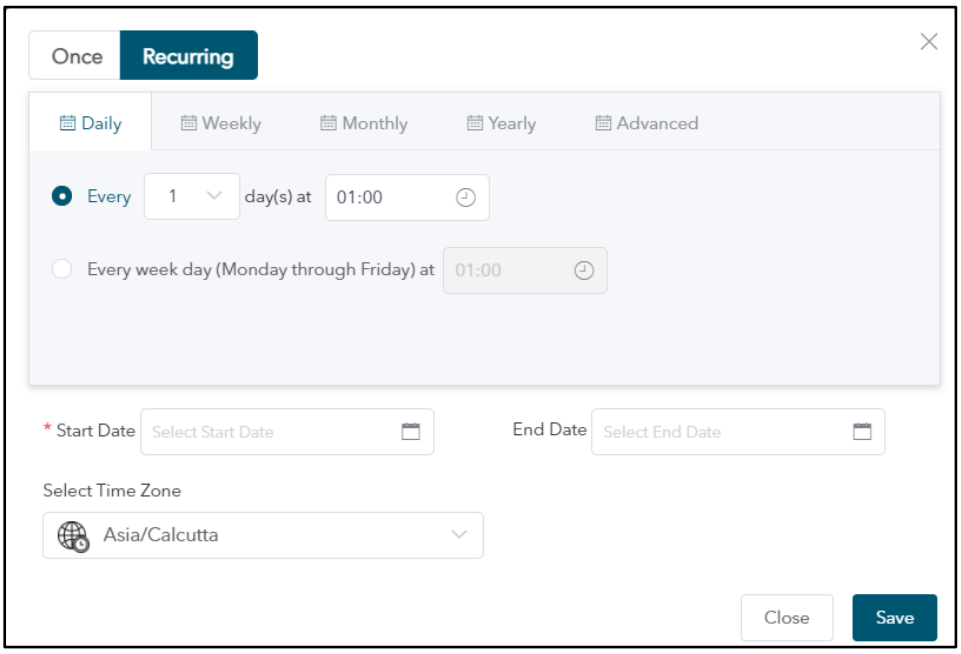


Figure 54: Recurring Daily

Weekly: The 'Monday' is checked by default and the user can check/uncheck the specific week and select the start time, start and end dates and click on save. At least one week name should be selected.

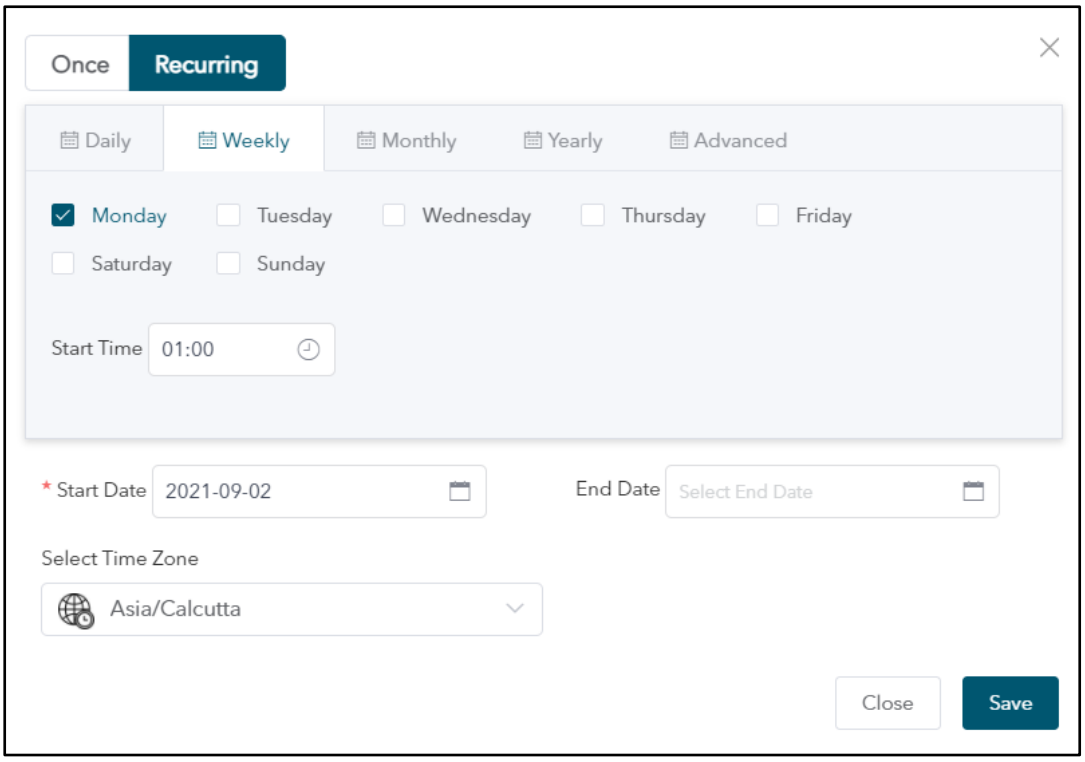


Figure 55: Recurring Weekly

Monthly: The user can select any of the available options and select start and end dates and click on save.

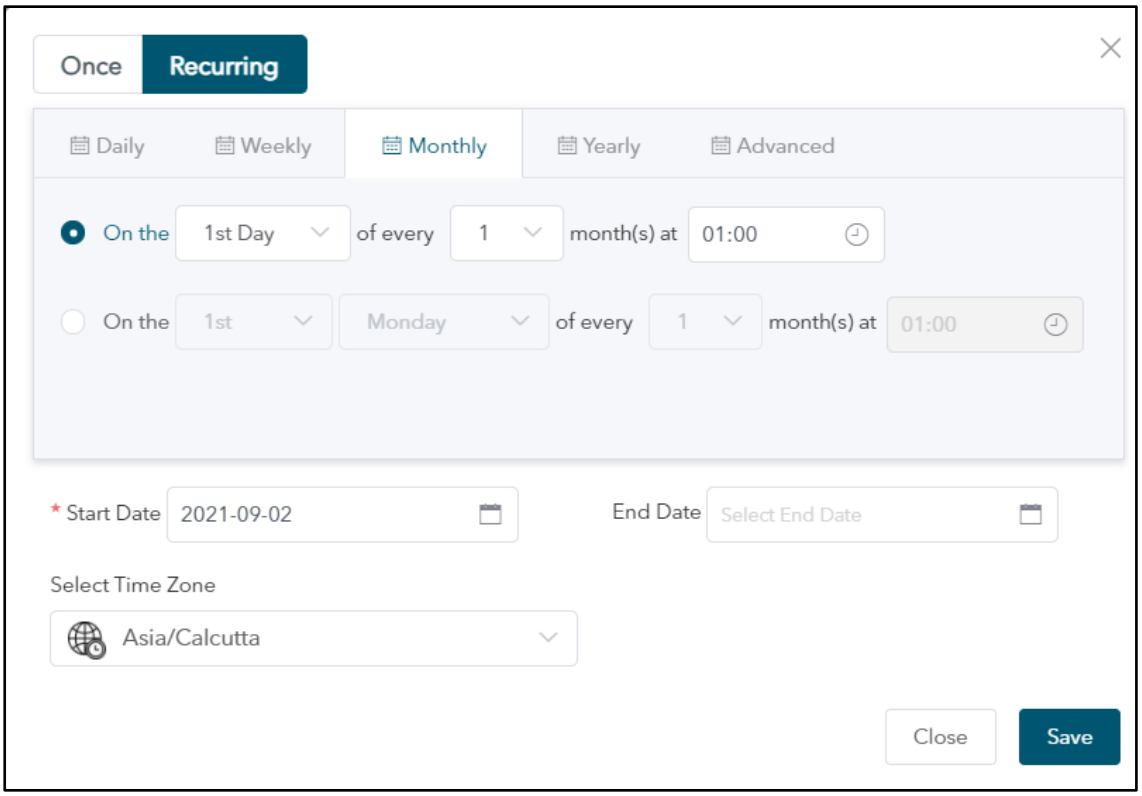


Figure 56: Recurring Monthly

Yearly: The user can select any of the available options from the tab and select start and end dates and click on save.

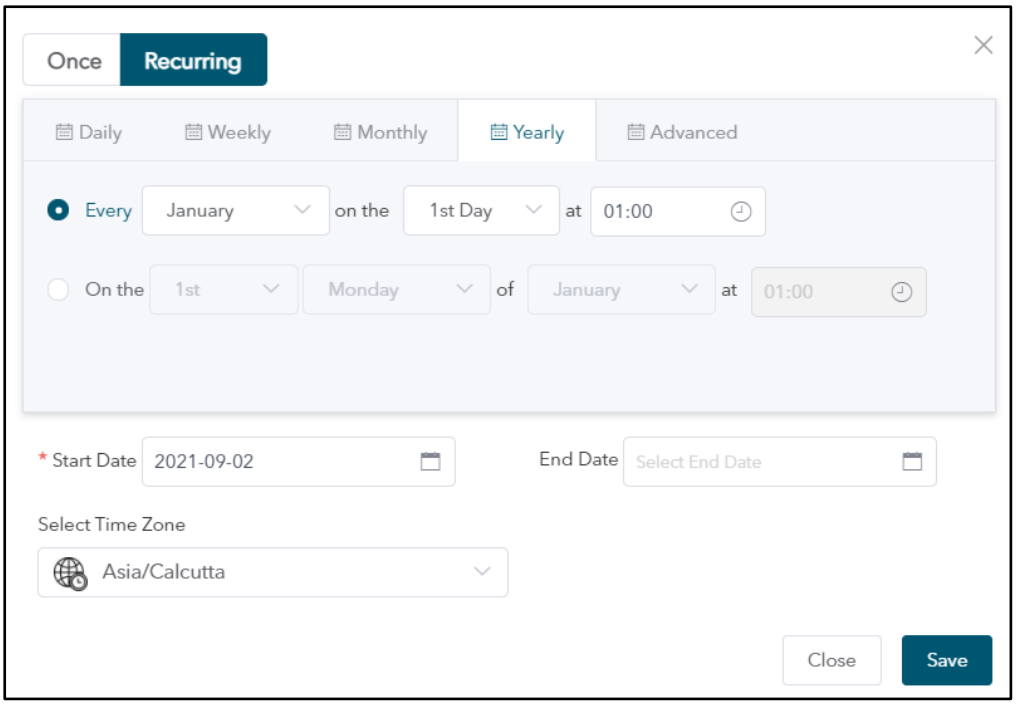


Figure 57: Recurring Yearly

Advanced: For Advanced tab, the user must enter the cron expression manually in the text box. The format to be followed to enter the cron expression is shown below the text box.

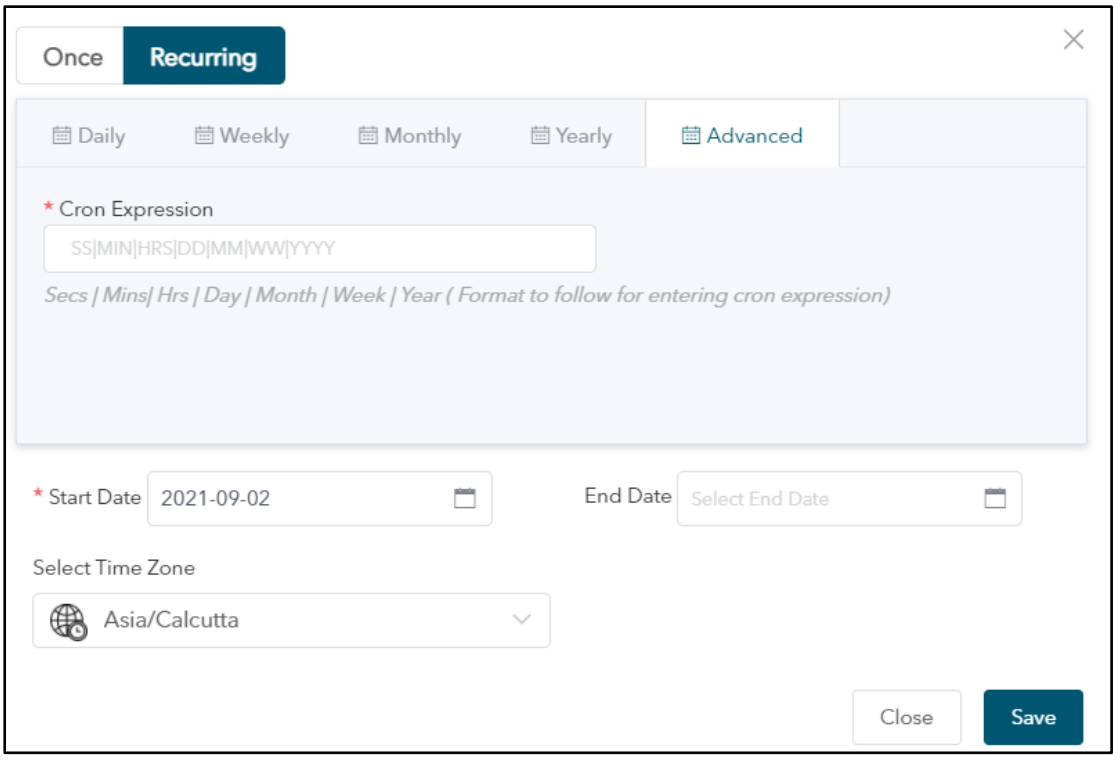


Figure 58: Advanced Recurring

1. Readable Format and Next scheduled Date: The user can view the readable format and next scheduled date for the valid cron expression entered as shown below. The error will be shown for any invalid cron expressions.

2. If the start date field is empty, it gets auto filled with the current date when user enters valid cron expression.

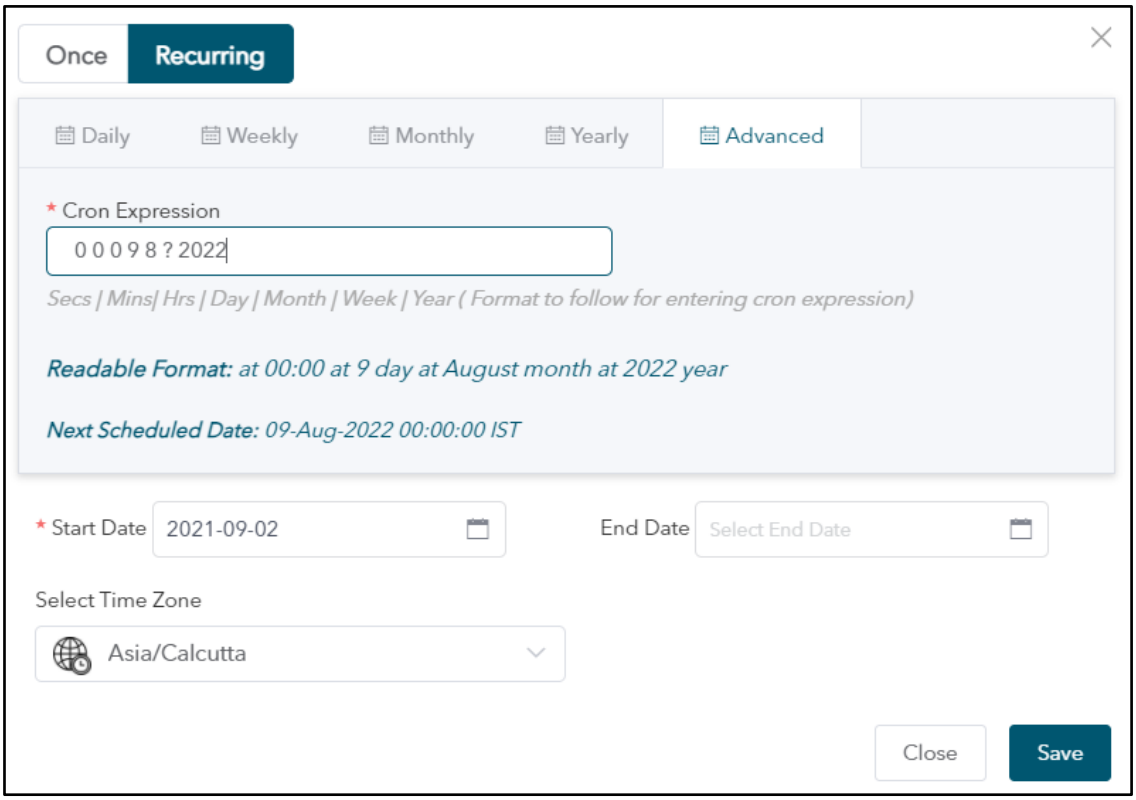


Figure 59: Advanced Recurring Tab

Start and End Dates: The start date is a mandatory field for the user to select to specify the schedule start date. Optionally, user can select the end date to specify the schedule end date. The validations for start and end dates are applicable.

Ex: If the start date is 2021-09-02, the time considered is 2021-09-02 00:00:00. If the end date is 2021-09-02 the time considered is 2021-09-02 23:59:59.

The Time zone can be selected as per user's preference.

Save/Close: For any of the above options selected, the user can click on save button for scheduling. The user cannot save/close the popup when there are any errors in the inputs and cannot save the schedule date /time. To close the popup anytime, you can click on the close button.

For the valid inputs, when user clicks on save, the popup will be closed, and you can see the next scheduled date on the respective page where the schedule button is.

3.1.10 Add Tag

To add a tag, click on the 'Add Tag'. Provide the tag category, and a tag value.

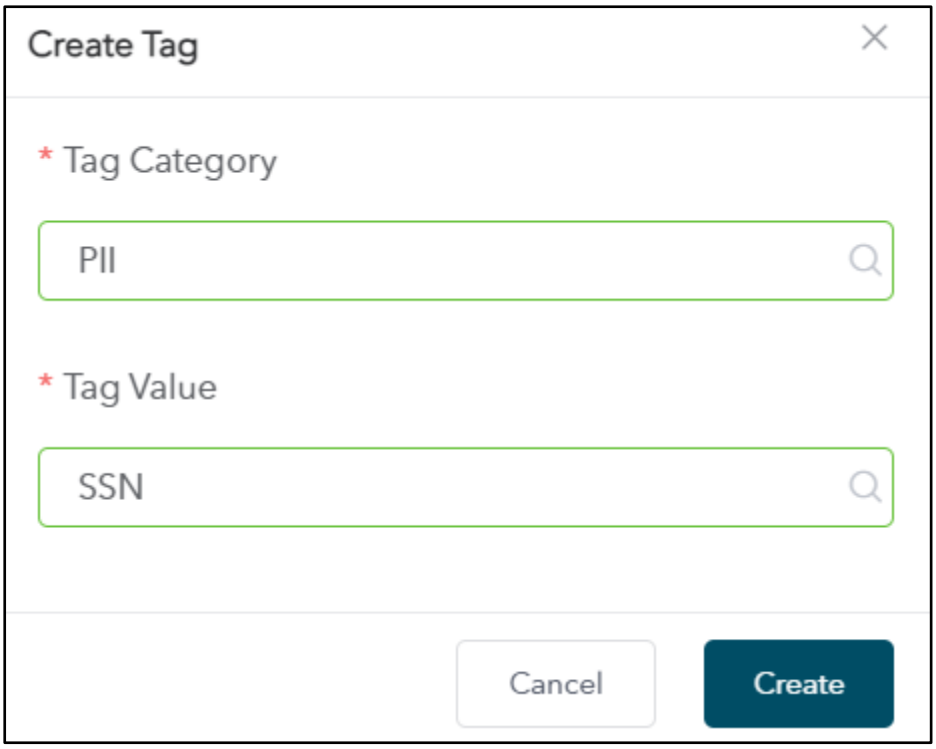


Figure 60: Add Tag

If a 'Tag Category' or a 'Tag Value' is already created, it will be shown as suggestion, when the user starts typing. Otherwise, the users will be asked to create the category.

The added tags will appear as <Tag Category>: <Tag Value> next to the table, or column for which they are added on the profile screen.

4 Pipelines

4.1 Pipelines

'Pipelines' section allows you to create pipelines from a source to destination. The source and destination first need to be added as data connections in Modak Nabu.

You can navigate to the 'Pipelines section by clicking the menu icon " ≡ " at top right of the screen. select 'Pipelines.

The 'Pipelines' module is used to add and modify the pipelines from source to destination. (figure)

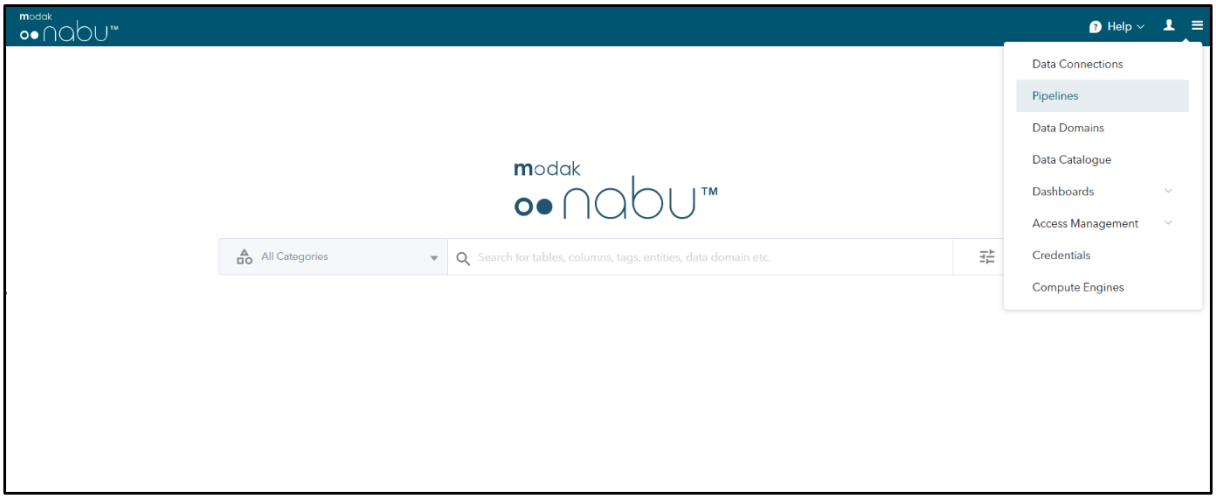


Figure 61: Pipelines Menu

The following features are available to the users on the Pipelines landing page.

In the left panel of the screen, it shows the icons for different data connection types that can be used as source and destination while creating pipelines. These include databases, file copy, and file ingestion and others for curations of pipelines. This panel includes the following features.

Search for data source type: This feature enables the user to search for the source type.

Sort by category/A to Z format: It helps the user to sort the source type by category/A to Z.

On the right-hand side, the landing page shows a table which comprises the following list of columns.

Pipeline Name: This column displays the list of all the existing pipeline names. To view details about a pipeline, click on the arrow next to the pipeline name to see pipeline schedule frequency.

Pipeline Type: This column shows the pipeline type. The user can filter the results based on the source type they want by using the filter option. Simply click on the filter icon, select the desired source type from the menu, then click apply to view it and table results will get refreshed with the applied source type. If the user wants to remove the applied filters, they can user clear button.

Last Run Status: This column displays the outcomes of the pipeline that have been scheduled. They could be waiting, succeeded, running, ignored, or failed. Each state is indicated by a distinct colour and symbols, such as waiting in grey, succeeded in green, running in blue, ignored in grey and failed in red, and the user can also check the needed status of the scheduled pipeline by using the filter option, simply click on the filter icon and select the required status from the menu and click on the apply button to view the filtered results. If the user wants to remove the applied filters, they can user clear button.

Last Run Date: It shows the last scheduled date and time for the pipeline. This column is sortable.

Next schedule Date: It shows the next schedule date and time for ingestion/curation of the pipelines. This column is sortable.

Owner: This column contains information about the person who created the pipeline.

Tags: This column displays tags linked to pipelines. It also gives filter choices, allowing users to apply the tag to the pipeline and filter the results according to applied tags.

Action: The action column provides the user to select any option (edit, duplicate, schedule, or delete) from the drop down. Based on the option selected the respective action will be applied.

Refresh: This indicator assists the user in refreshing the table. When user clicks on the refresh button, the results are refreshed and updated.

Column options: This option enables the user to select the required columns to display on the table. Some columns are selected by default, while others are not. For example: Pipeline name, pipeline type, last run status, last run date, next schedule date, and action are all selected by default. According to user preferences, the user can select/deselect the option from the menu, which includes the owner and tags.

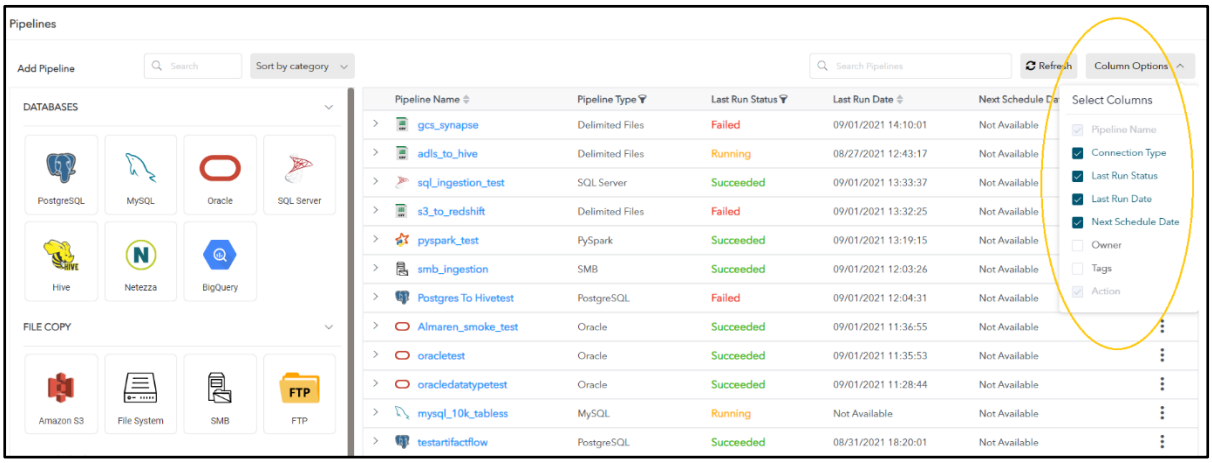


Figure 62: Column Option for Pipelines

Search Pipeline: This feature helps users to search for the required pipelines. simply by entering the name of pipeline in the search box.

4.1.1 Creation of pipelines

4.1.1.1 Creation of Database Pipeline

Following are the steps to execute the curation Pipeline.

9. Select type of data connection (i.e., Postgres, MySQL, Oracle, SQL Server, SAS) from the left pane on the Pipelines landing page.
10. You will be redirected to the “Add Pipeline” wizard for the selected source type. The wizard has 6 steps.
11. Pipeline information:

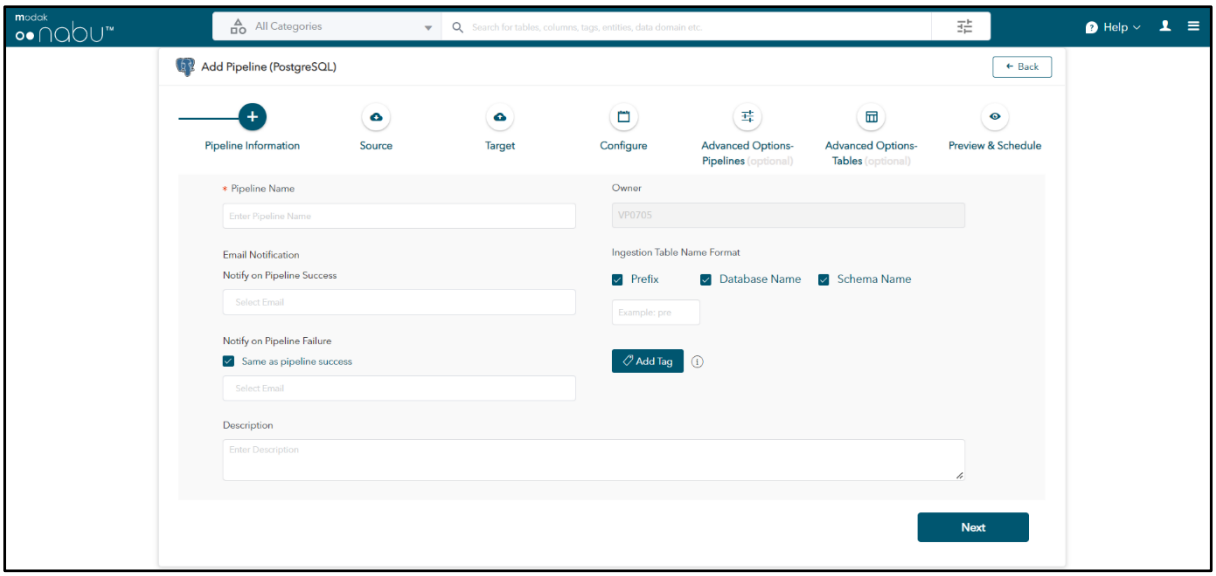


Figure 63: Database-Add Pipeline-Pipeline information

12. Pipeline name: Fill in the pipeline name which is a mandatory field. The pipeline name should start with an alphabet and should contain at least 3 characters, can contain numbers and special characters except underscore others are not allowed.
13. The owner field is auto filled with the user's ID and disabled.
14. Ingestion table name format: The ingestion table name format helps to define the naming format for the ingested table. Prefix, database name, and schema name are all checked by default. To name ingested table in the destination, the user must enter a prefix (maximum of 5 characters) in the prefix field. As a result, the user will see the ingested table name in the format specified under prefix field as an example
15. Add tag: This allows the user to add a tag to the pipeline. For Add Tag please refer to section:3.1.10.
16. Description: This field enables the user to enter description for the pipeline.
17. Email Notification: The user can select single/multiple email addresses in 'Notify on Pipeline Success' and 'Notify on Pipeline Failure' fields. The check box on 'same as pipeline success' will copy the same content from 'Notify on Pipeline Success' to 'Notify on Pipeline Failure' when checked.

Click Next to move to the next step.

18. Source:
19. Select a source data connection from left side of the pane, where you would like to move the data or user can search data connection by using search box.

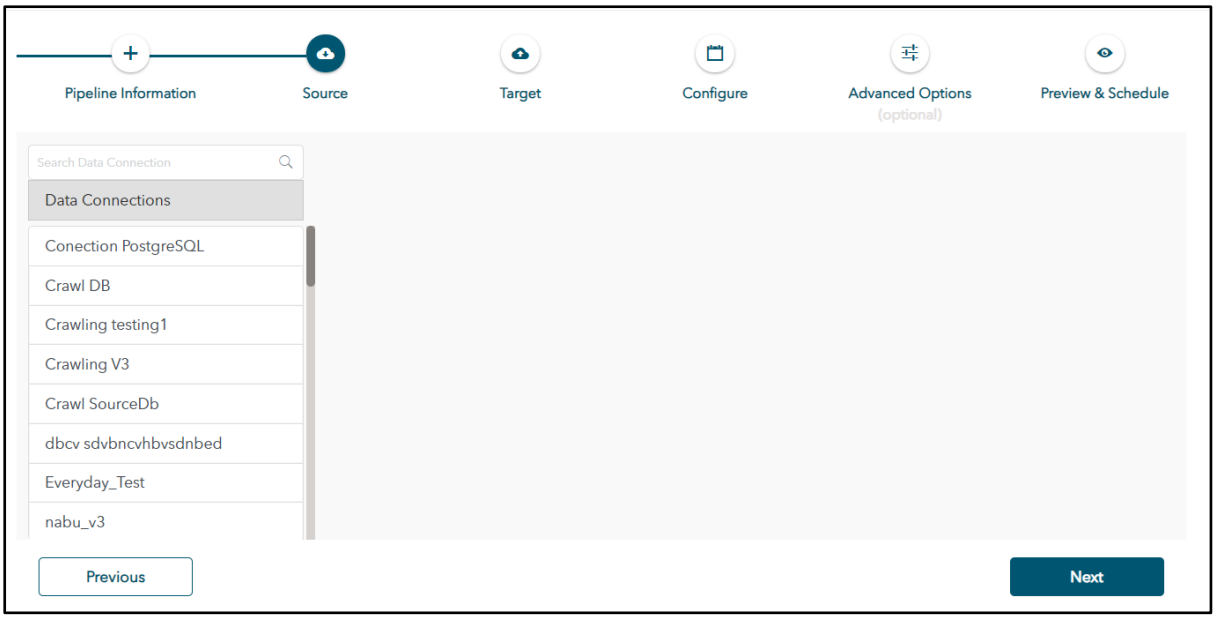


Figure 64: Database-Add Pipeline-Search for Data Connection

20. On the same screen, from the right side of the pane, select the schemas for the respective data connection (s) selected. Or use search box to search for the required schema.
21. By default, the tables/views button in Action column is disabled. The button gets enabled for the selected schema.

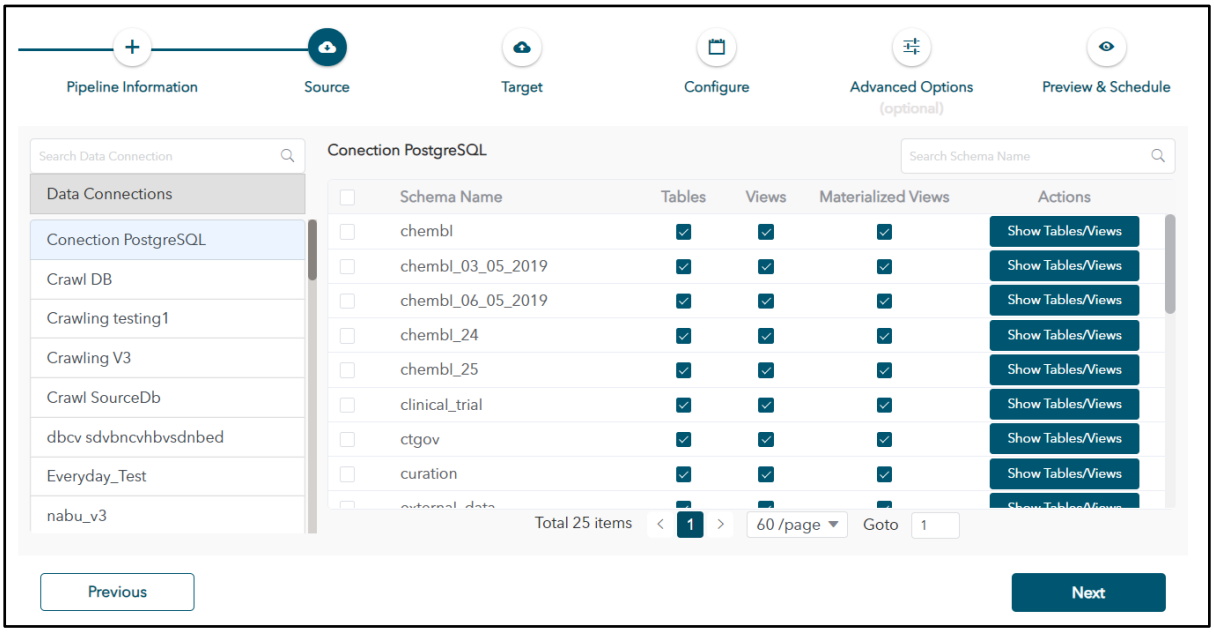


Figure 65: Database-Select/search for Schemas

22. Users can select the tables/views to be included as part of the data movement from the selected schemas by clicking on the ‘Show Tables/Views' button. All tables, views, and materialized views are checked by default. The boxes are checked/unchecked based on the user's preferences.
23. When the user clicks on show tables/views, a pop-up window appears where the user can select the tables/views one at a time or select all. The user can also look for tables/views using the search box.

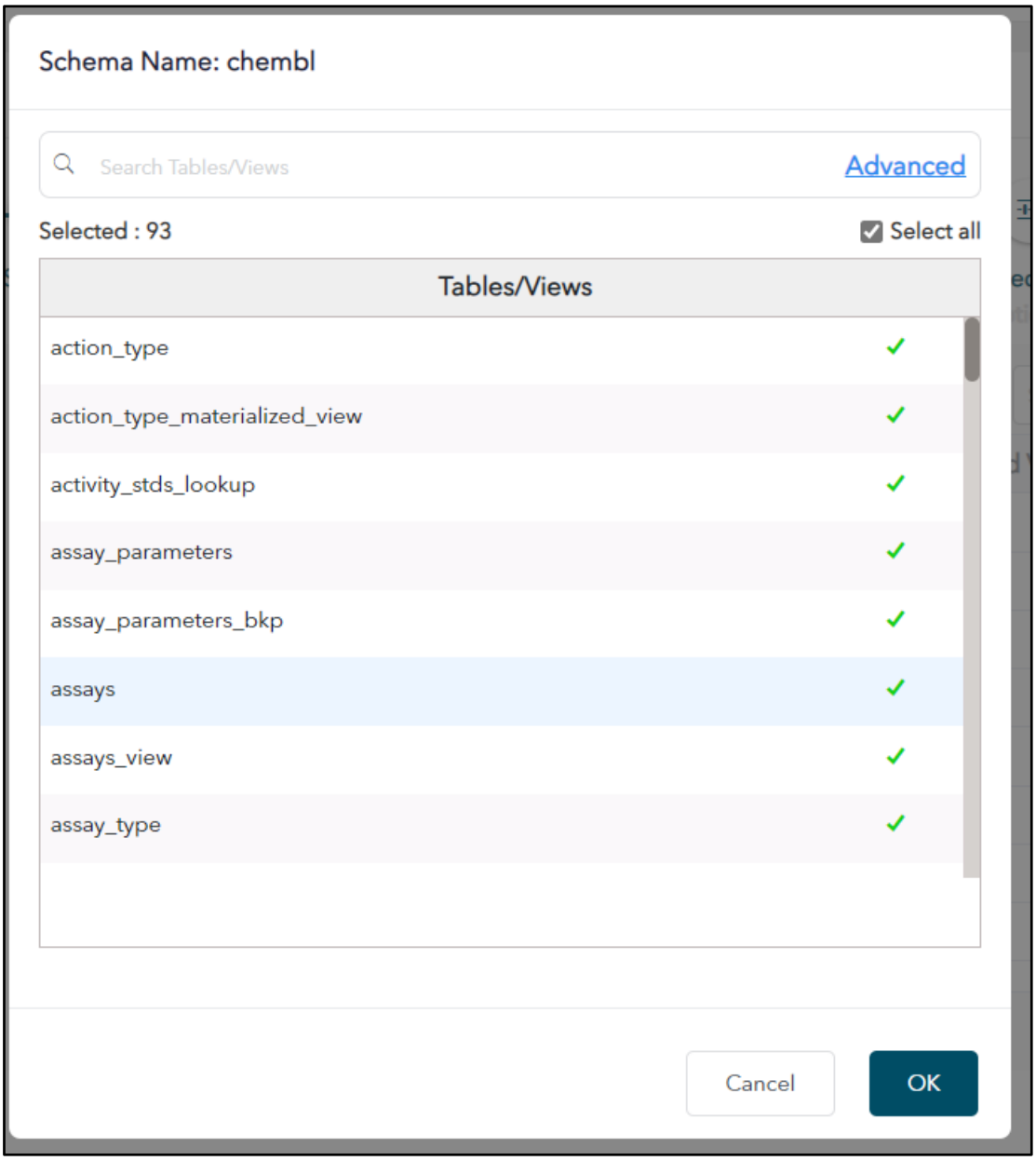


Figure 66: Database-Select/search for Tables/views

24. The advanced button allows the user to apply a filter. When you click the advanced button, an “Add filter” button is added to the popup. The user can select the filter and click on the add filter button. The results are updated with the applied filters, and the applied filter is visible as a tag in the popup's top left corner. Close the applied filter tag to remove the applied filters. The available filters are:
25. Inclusive Regex – Include all tables/views, whose name matches a regex pattern.
26. Exclusive Regex - Exclude all tables/views, whose name matches a regex pattern.
27. Inclusive Like – Include all tables/views whose name is like a certain name.
28. Exclusive Like – Exclude all tables/views whose name is like a certain name

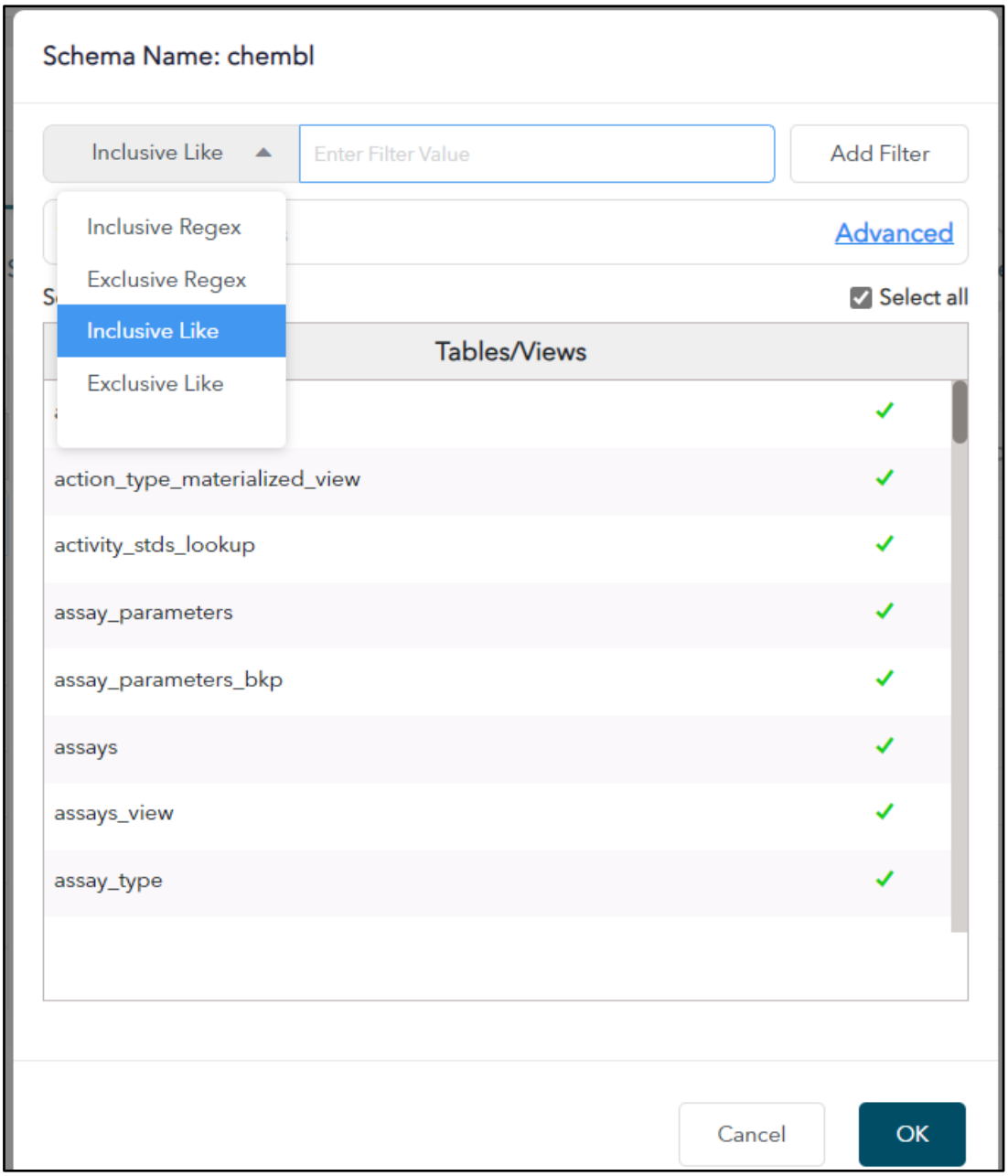


Figure 67: Database-Add Filter

29. Destination:
30. Select a destination data connection from the dropdown menu on the left side which shows the respective schemas/containers/buckets for the respective selected data connection.
31. For any destination data connection with no schemas/containers/buckets, the message 'No containers/schemas/buckets available. Please select any other destination' is shown.

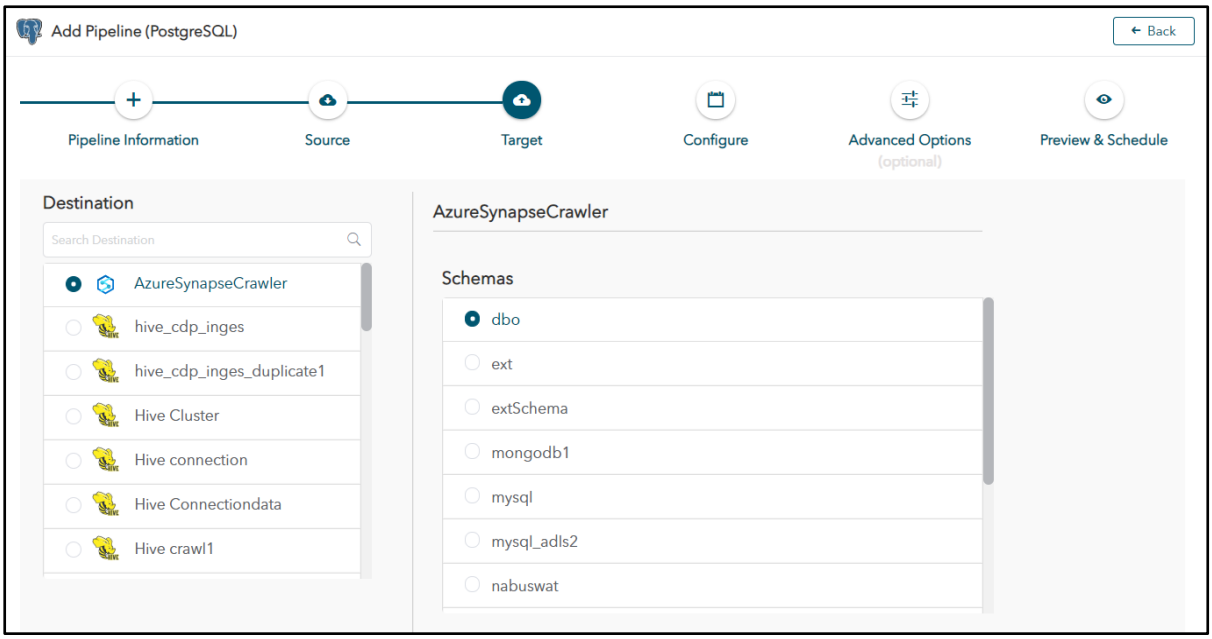


Figure 68: Database-Add Pipeline-Select Destination/Schemas

Click Previous to move to the previous step.

Click Next to move to the next step.

32. Configure:

    In this step, the user must first select the compute engine. On selection of compute engine, the user will be allowed to select the workflow engine.

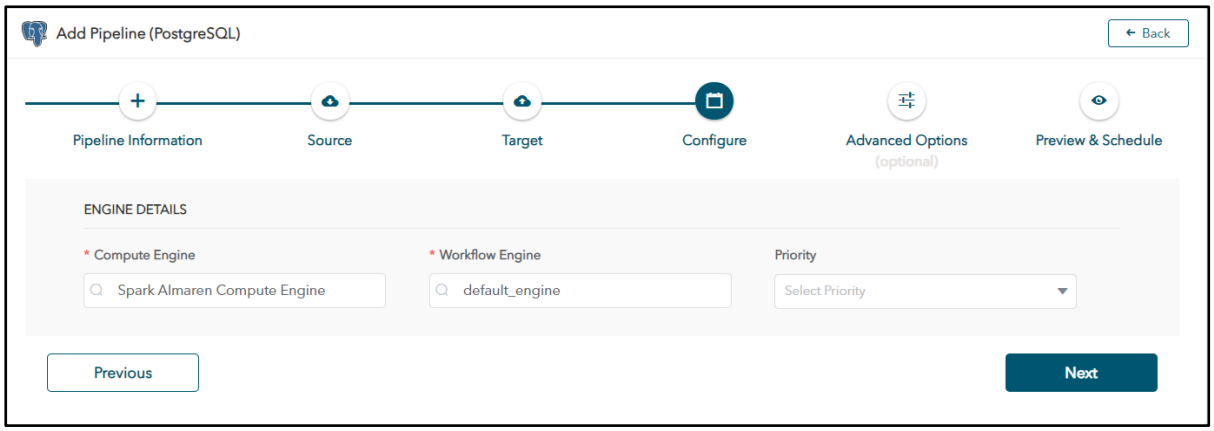


Figure 69: Database-Add Pipeline-Configure

    Click Previous to move to the previous step.

    Click Next to move to the next step.

33. Advance Options - Pipeline: please refer to Section 4.1.2

34. Advanced Options – Table: please refer to Section 4.1.3

35. Preview & Schedule:

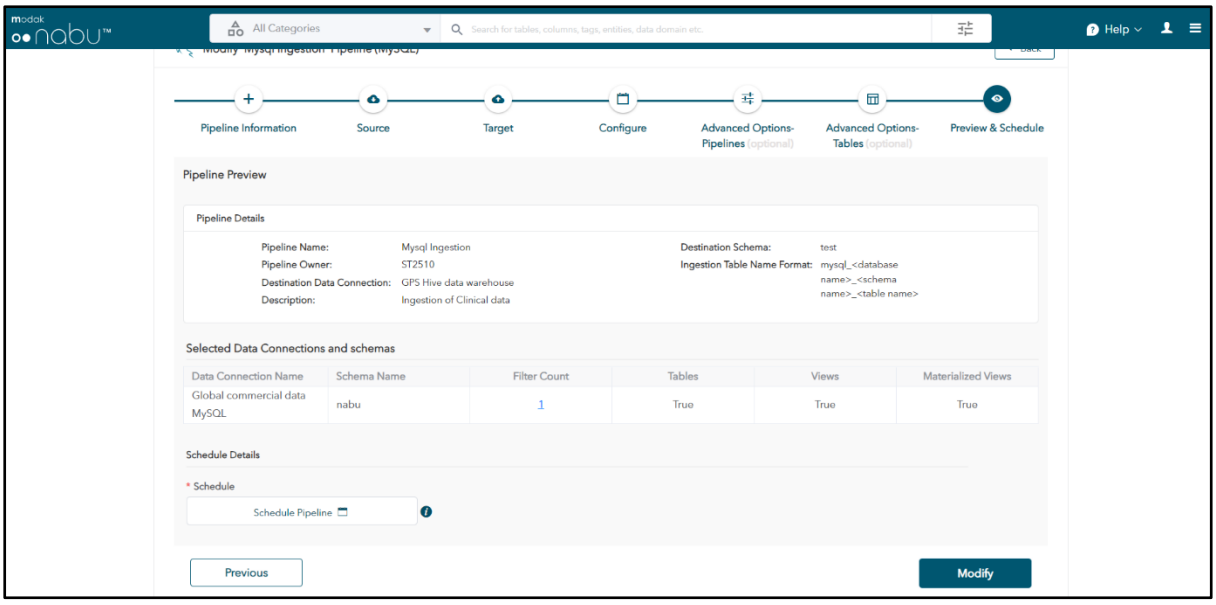


Figure 70: Database-Add Pipeline-Preview & Schedule

In this section, you can preview all the data entered in all the steps. You can schedule the pipeline by clicking on Schedule Pipeline button. For scheduling, please refer to section: 3.2.10.. The user can confirm the pipeline by clicking on "create" button.

Click Previous to move to the previous step.

4.1.1.2 Creation of file copy pipeline

Following are the steps to create a Pipeline.

1. Select type of data connection (i.e., Amazon S3, File System, SMB, FTP) from the left pane on the Pipelines landing page.
2. You will be redirected to the "Add Pipeline" wizard for the selected source type. The wizard has 6 steps.
3. Pipeline information:

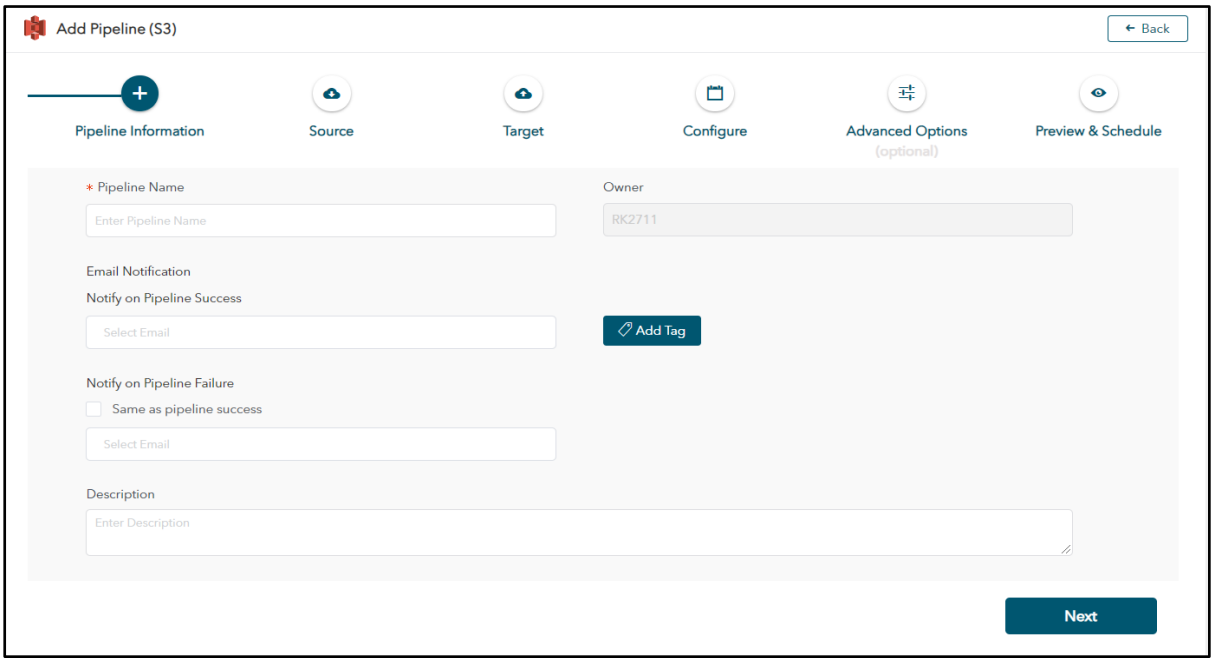


Figure 71: File Copy-Add Pipeline-Pipeline information

a. Pipeline name: Fill in the pipeline name which is a mandatory field. The pipeline name should start with an alphabet and should contain at least 3 characters, can contain numbers and special characters except underscore others are not allowed.
b. The owner field is auto filled with the user's ID and disabled.
c. Add tag: This allows the user to add a tag to the pipeline. For Add Tag please refer to section:3.1.10.
d. Description: This field enables the user to enter description for the pipeline.
e. Email Notification: The user can select single/multiple email addresses in 'Notify on Pipeline Success' and 'Notify on Pipeline Failure' fields. The check box on 'same as pipeline success' will copy the same content from 'Notify on Pipeline Success' to 'Notify on Pipeline Failure' when checked.

Click **Next** to move to the next step.

4. Source:
   a. Select a source data connection from left side of the pane, where you would like to move the data or user can search data connection by using search box.

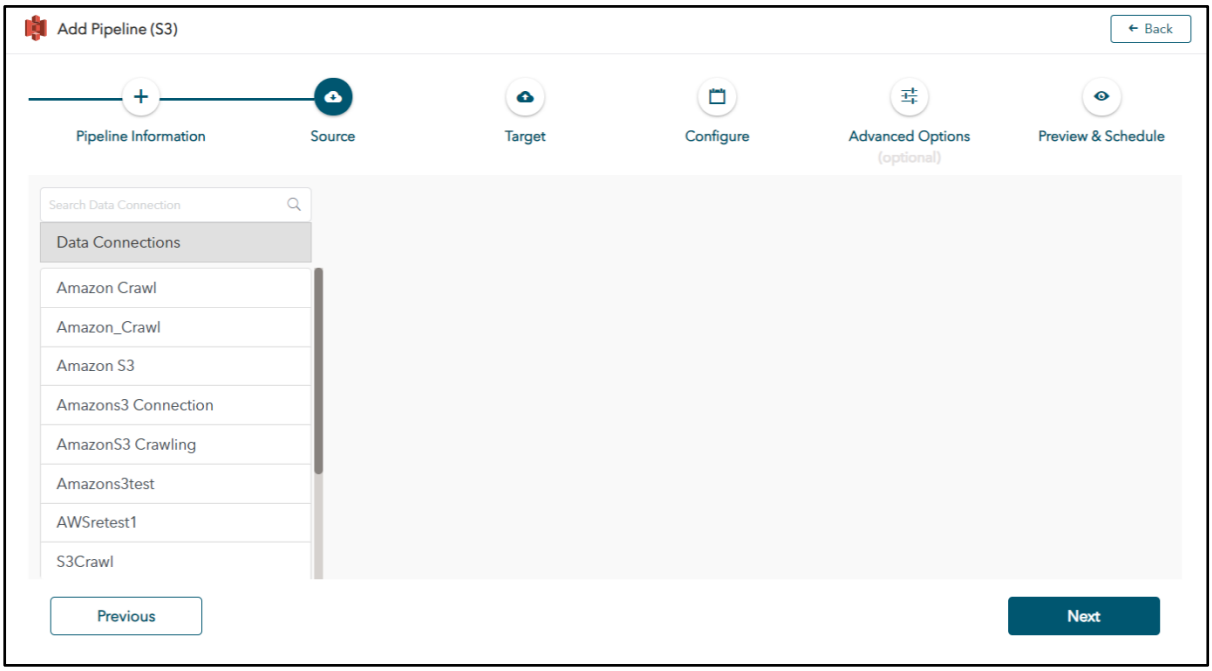


Figure 72: File Copy-Add Pipeline-Select/search for Source

   b. On the same screen, from the right side of the pane, select the bucket for the respective data connection (s) selected. Or use search box to search for the required bucket/schemas.

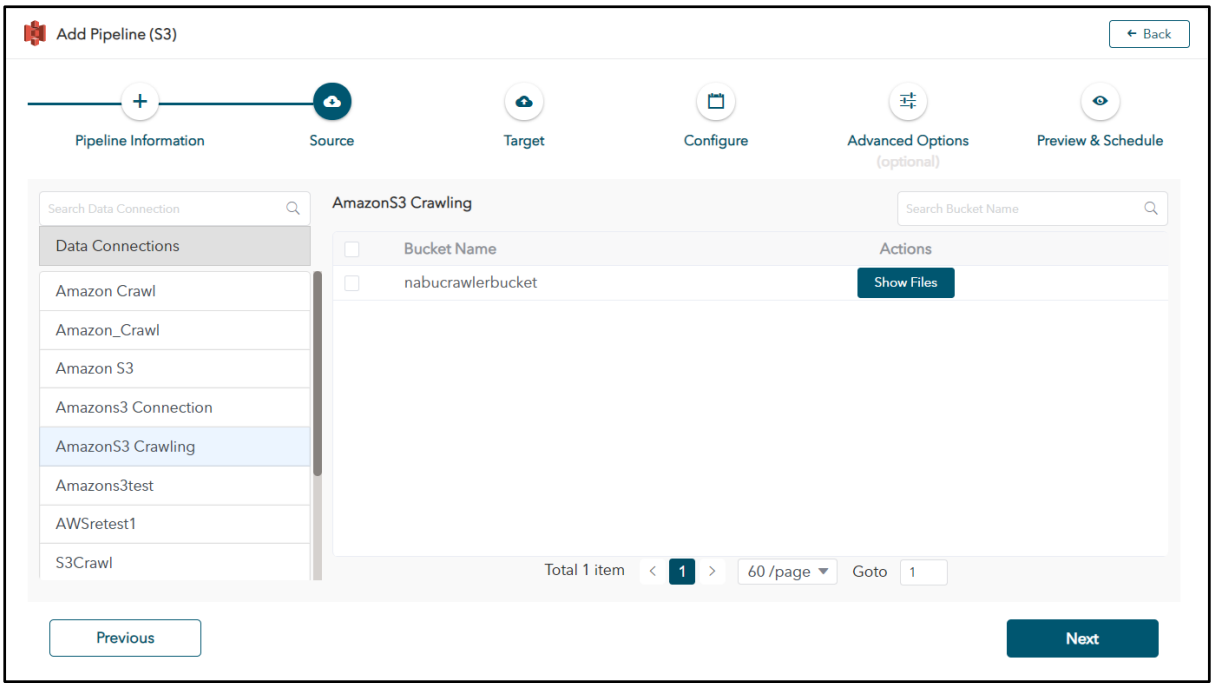


Figure 73: File Copy-select/search for Buckets

c. By default, the files button in Action column is disabled. The button gets enabled for the selected bucket/root path/share.
d. Users can select the files to be included as part of the data movement from the selected bucket by clicking on the ‘Show file' button.
e. When the user clicks on show files, a pop-up window appears where the user can select the files one at a time or select all. The user can also look for files using the search box.

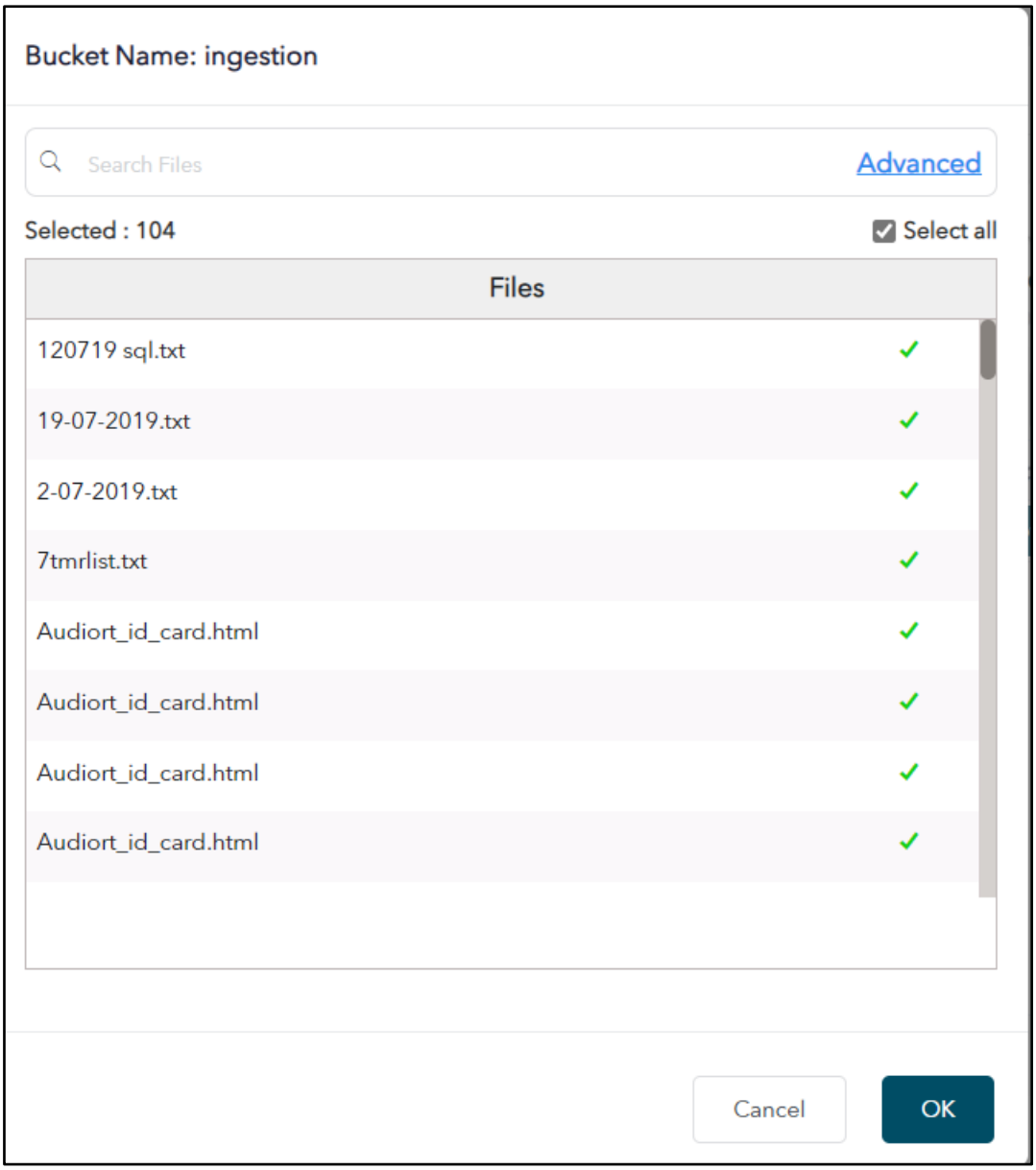


Figure 74: File Copy-Select/search for Files

f. The advanced button allows the user to apply a filter. When you click the advanced button, an “Add filter” button is added to the popup. The user can select the filter and click on the add filter button. The results are updated with the applied filters, and the applied filter is visible as a tag in the popup's top left corner. Close the applied filter tag to remove the applied filters. The available filters are:
    i. Inclusive Regex – Include all tables/views, whose name matches a regex pattern.
    ii. Exclusive Regex - Exclude all tables/views, whose name matches a regex pattern.
    iii. Inclusive Like – Include all tables/views whose name is like a certain name.
    iv. Exclusive Like – Exclude all tables/views whose name is like a certain name.

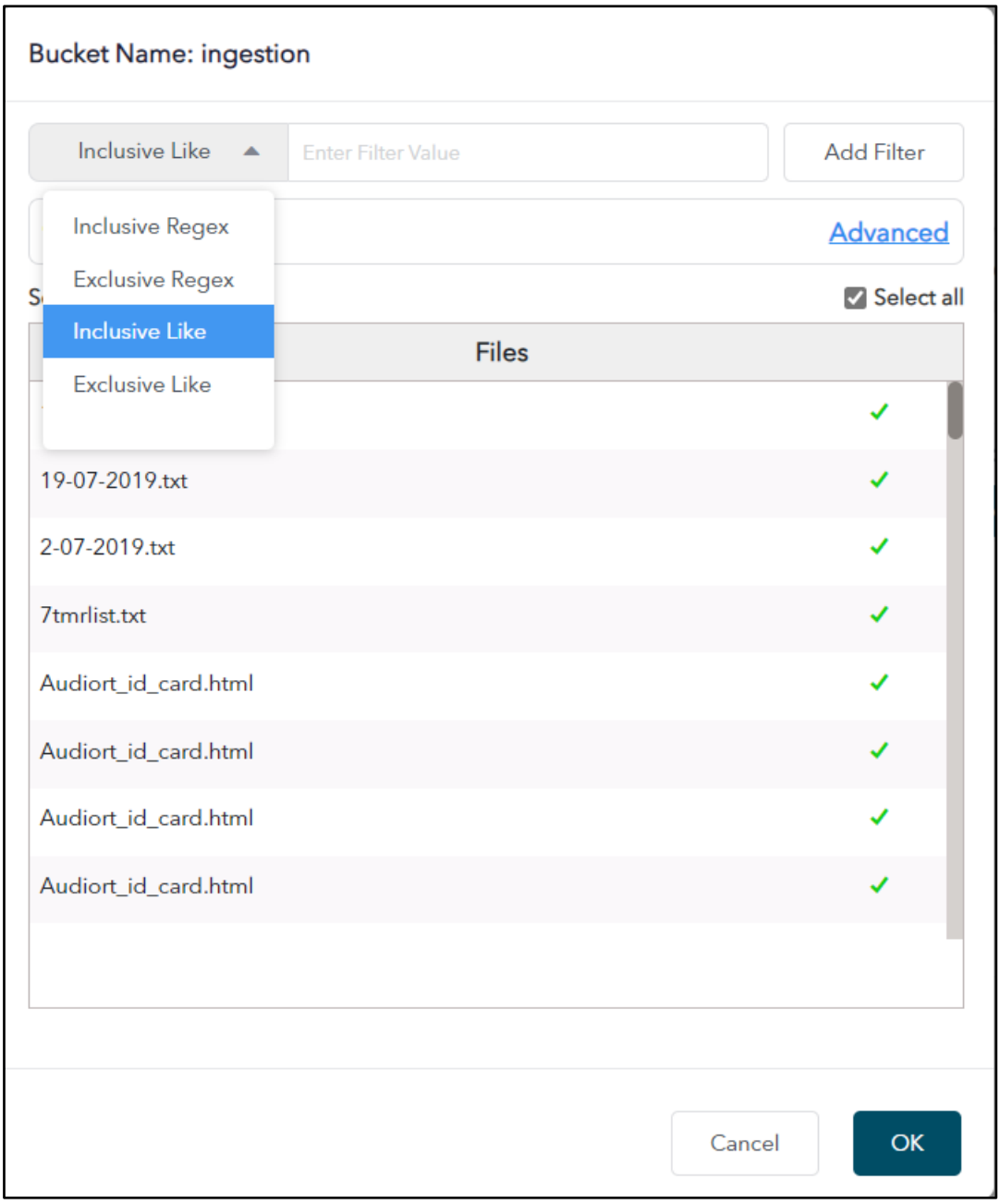


Figure 75: File Copy-Add Filter

Click **Previous** to move to the previous step.

Click **Next** to move to the next step.

5. Destination:

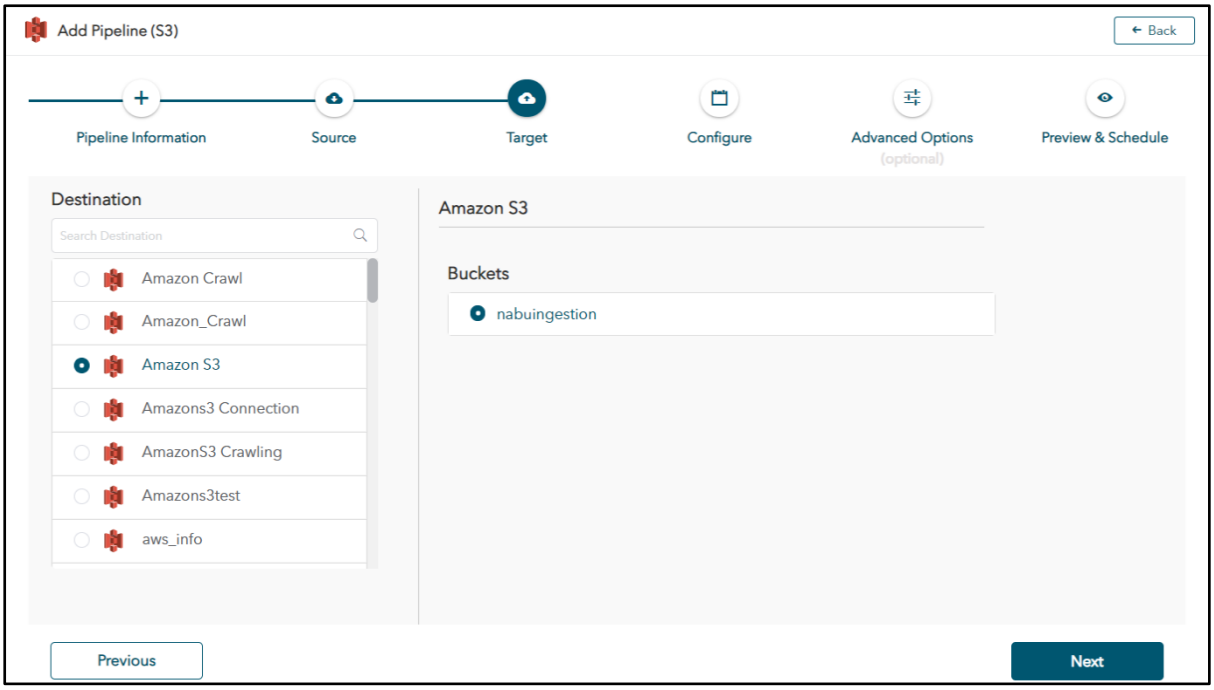


Figure 76: File Copy-Add Pipeline-Select/Search for Destination

a. Select a destination data connection from the dropdown menu on the left side which shows the respective bucket/root path/share for the respective selected data connection.
b. For any destination data connection with no bucket/root path/share, the message 'No containers/schemas/buckets available. Please select any other destination' is shown.

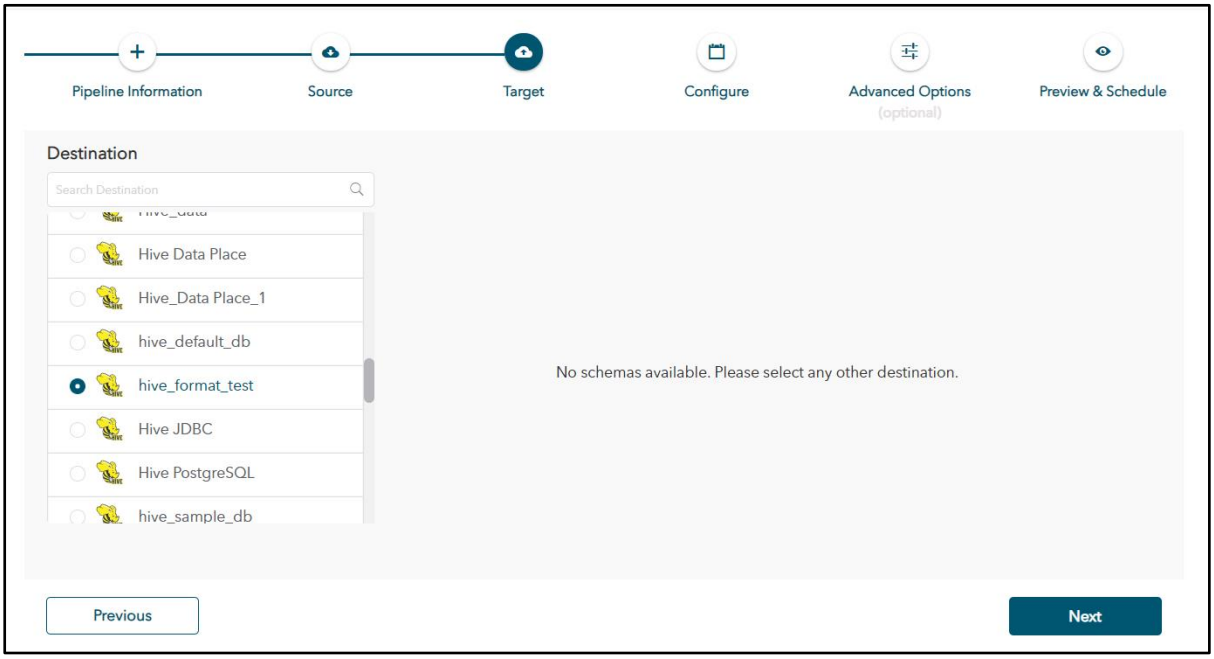


Figure 77: File Copy-No Schema Available

6. Configure:

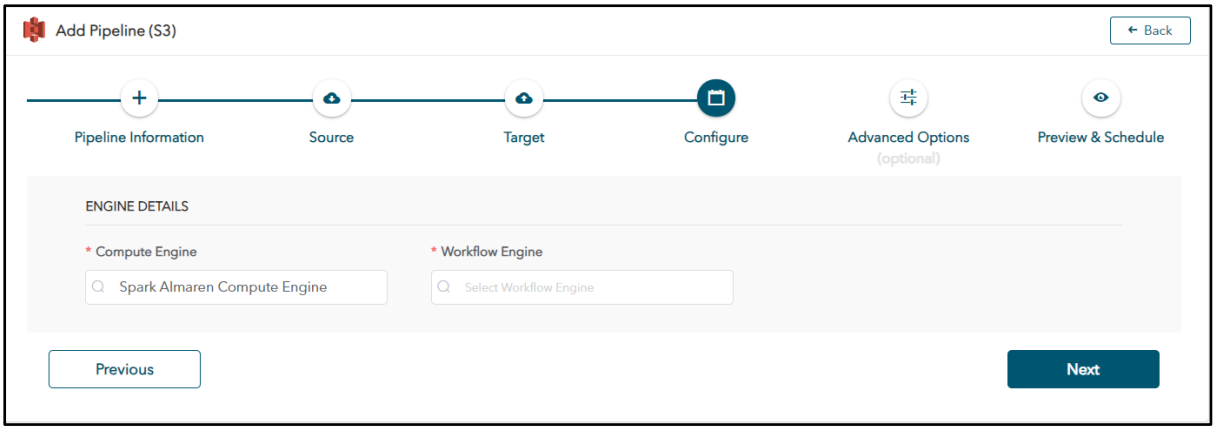


Figure 78: File Copy-Add Pipeline-Configure

In this step, the user must first select the compute engine. On selection of compute engine, the user will be allowed to select the workflow engine.

Click Previous to move to the previous step.

Click Next to move to the next step.

7. Advanced Options: please refer to Section 3.4.3
8. Preview & schedule:

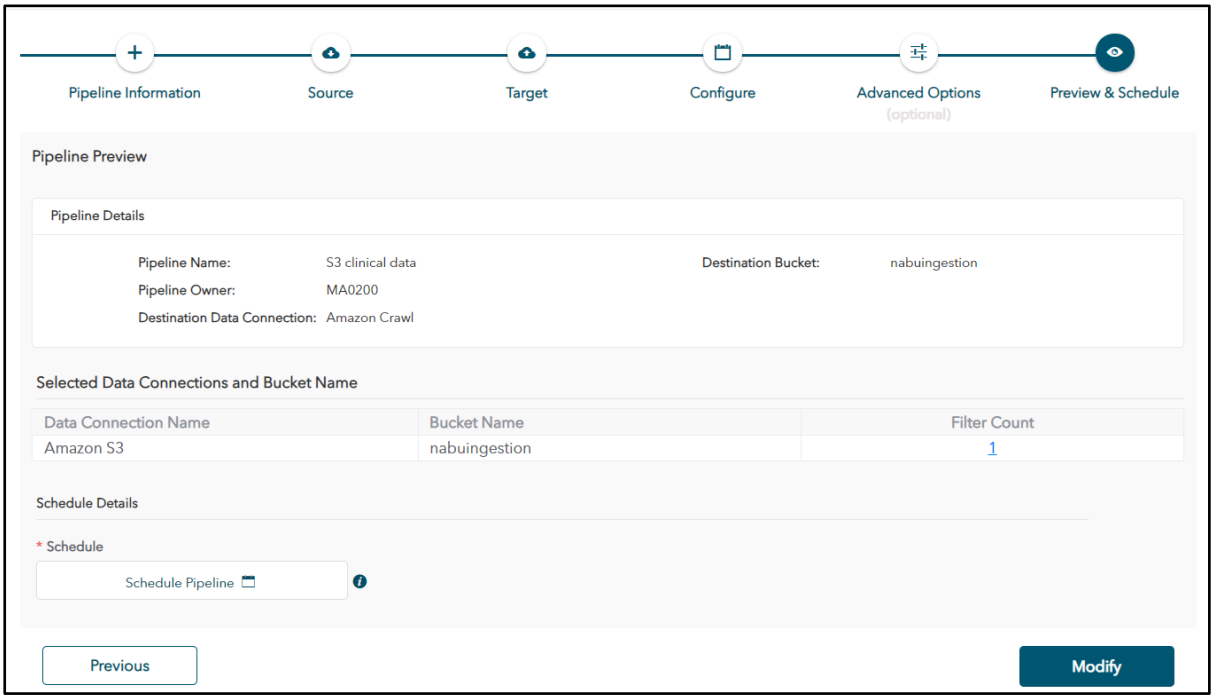


Figure 79: File Copy-Add Pipeline-Preview & Schedule

In this section, you can preview all the data entered in all the steps. You can schedule the pipeline by clicking on Schedule Pipeline button. For scheduling, please refer to section: 3.2.10.

The user can confirm the pipeline by clicking on “create” button.

Click **Previous** to move to the previous step.

4.1.1.3 Creation of File ingestion pipelines

Following are the steps to create a Pipeline.

1. Click on the Delimited files icon (File ingestion section) from the left pane on Pipelines landing page.
2. You will be redirected to the “Add Pipeline” wizard for the selected pipeline type. The wizard has 6 steps.
3. Pipeline information:

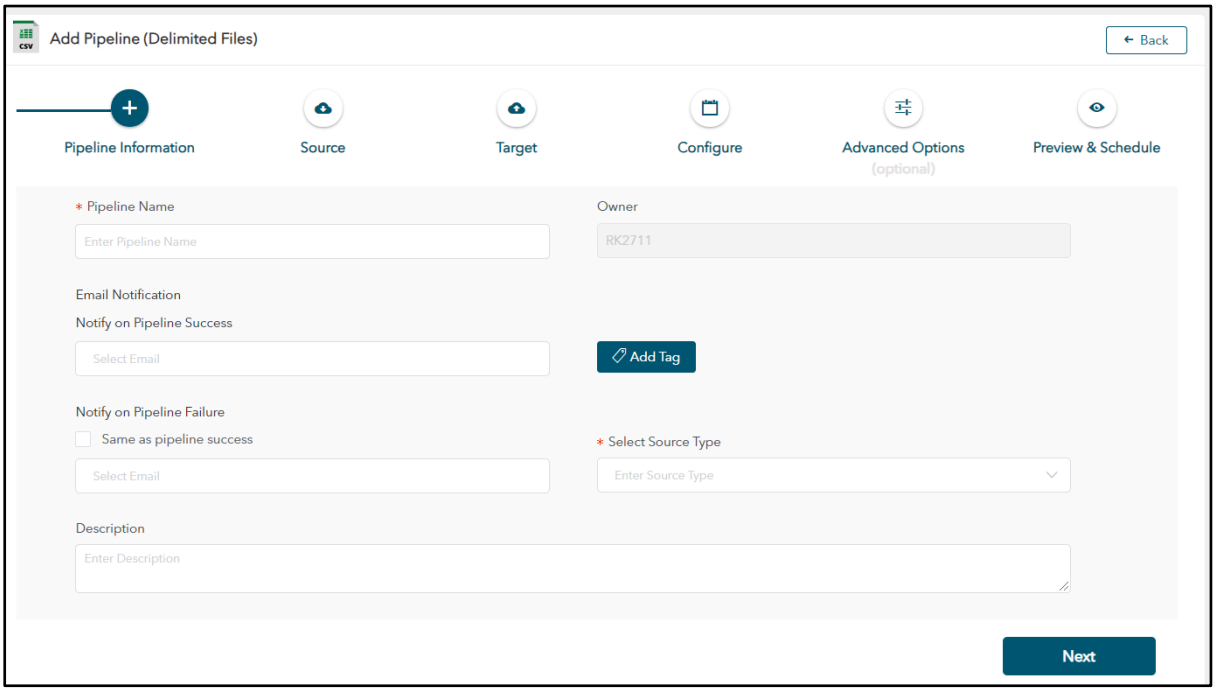


Figure 80: File Ingestion-Add Pipeline-Pipeline Information

   a. Fill in the pipeline name which is a mandatory field. The pipeline name should start with an alphabet and should contain at least 3 characters, can contain numbers and special characters except underscore are not allowed.
   b. The owner field is auto filled with the user’s ID and disabled.
   c. Email notification: This is an optional field where the user can select single/multiple email addresses in ‘Notify on Pipeline Success’ and ‘Notify on Pipeline Failure’ fields. The check box on ‘same as pipeline success’ will copy the same content from ‘Notify on Pipeline Success’ to ‘Notify on Pipeline Failure’ when checked.
   d. Select Source type: This is a mandatory field, and the user can select the source type for the pipeline. The data connections in the source step will be according to the selected source type.
   e. Add Tag: This allows the user to add a tag to the pipeline. For Add Tag please refer to section:3.1.10.
   f. Description: This field enables the user to enter description for the pipeline.

Click Next to move to the next step.

4. Source:
   a. Select a source data connection from left side of the pane, where you would like to move the data or user can search data connection by using search box.

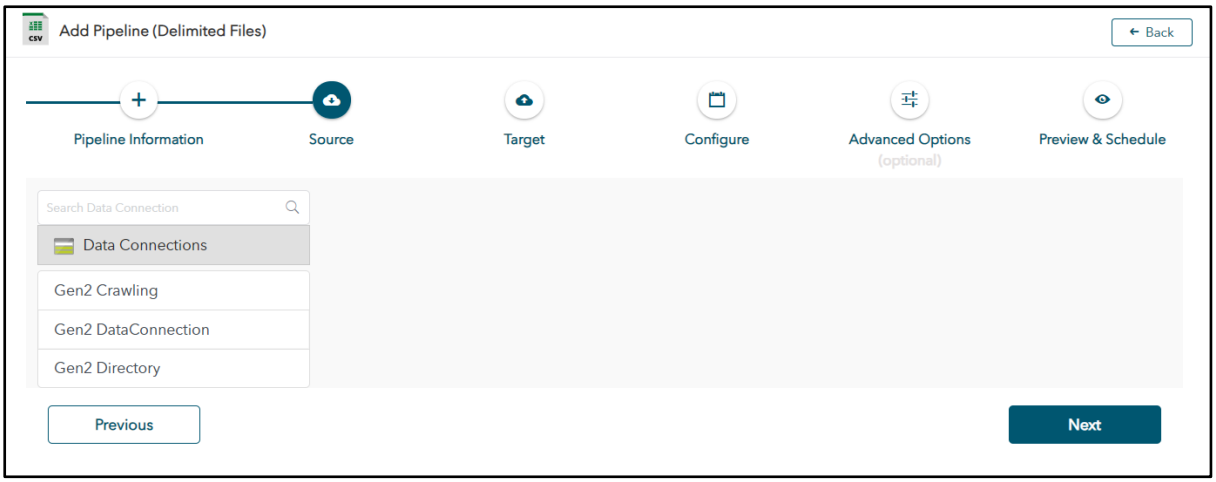


Figure 81: File Ingestion-Add Pipeline-Select/Search Data Connection

b. On the same screen, from the right side of the pane, select the schemas for the respective data connection (s) selected. Or use search box to search for the required schema.

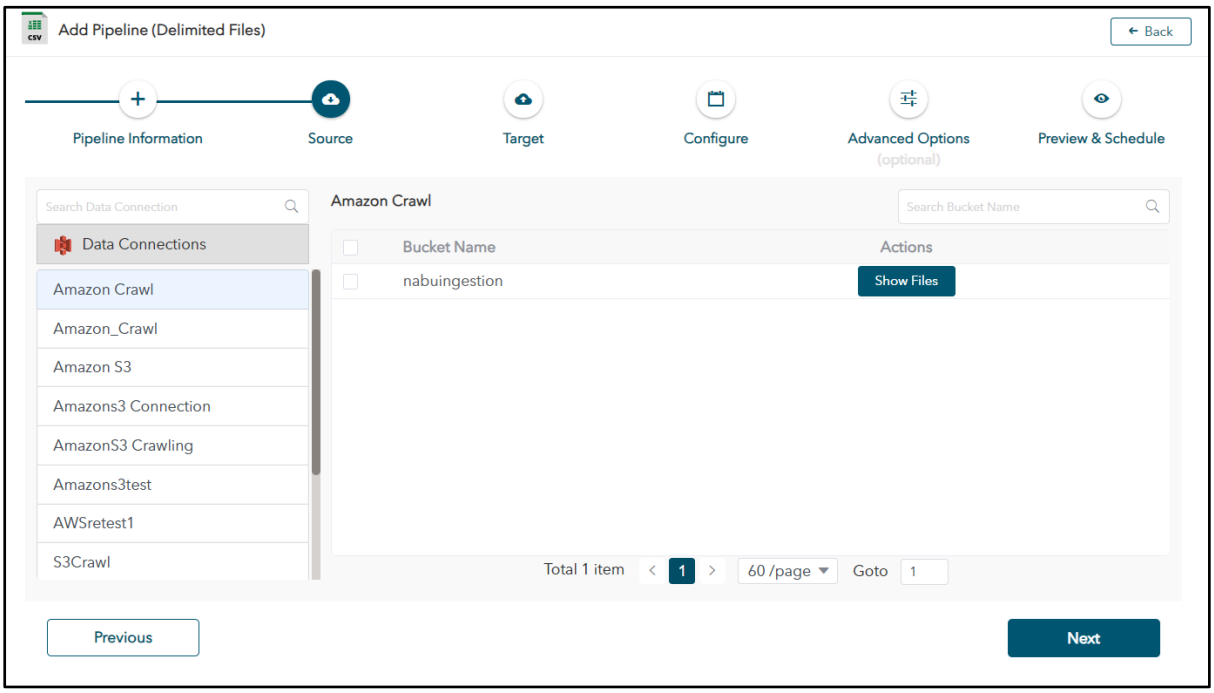


Figure 82: File Ingestion-Select/search Buckets

c. By default, the tables/views button in Action column is disabled. The button gets enabled for the selected schema.

d. Users can select the tables/views to be included as part of the data movement from the selected schemas by clicking on the ‘Show Tables/Views' button. All tables, views, and materialized views are checked by default. The boxes are checked/unchecked based on the user's preferences.

e. When the user clicks on show tables/views, a pop-up window appears where the user can select the tables/views one at a time or select all. The user can also look for tables/views using the search box.

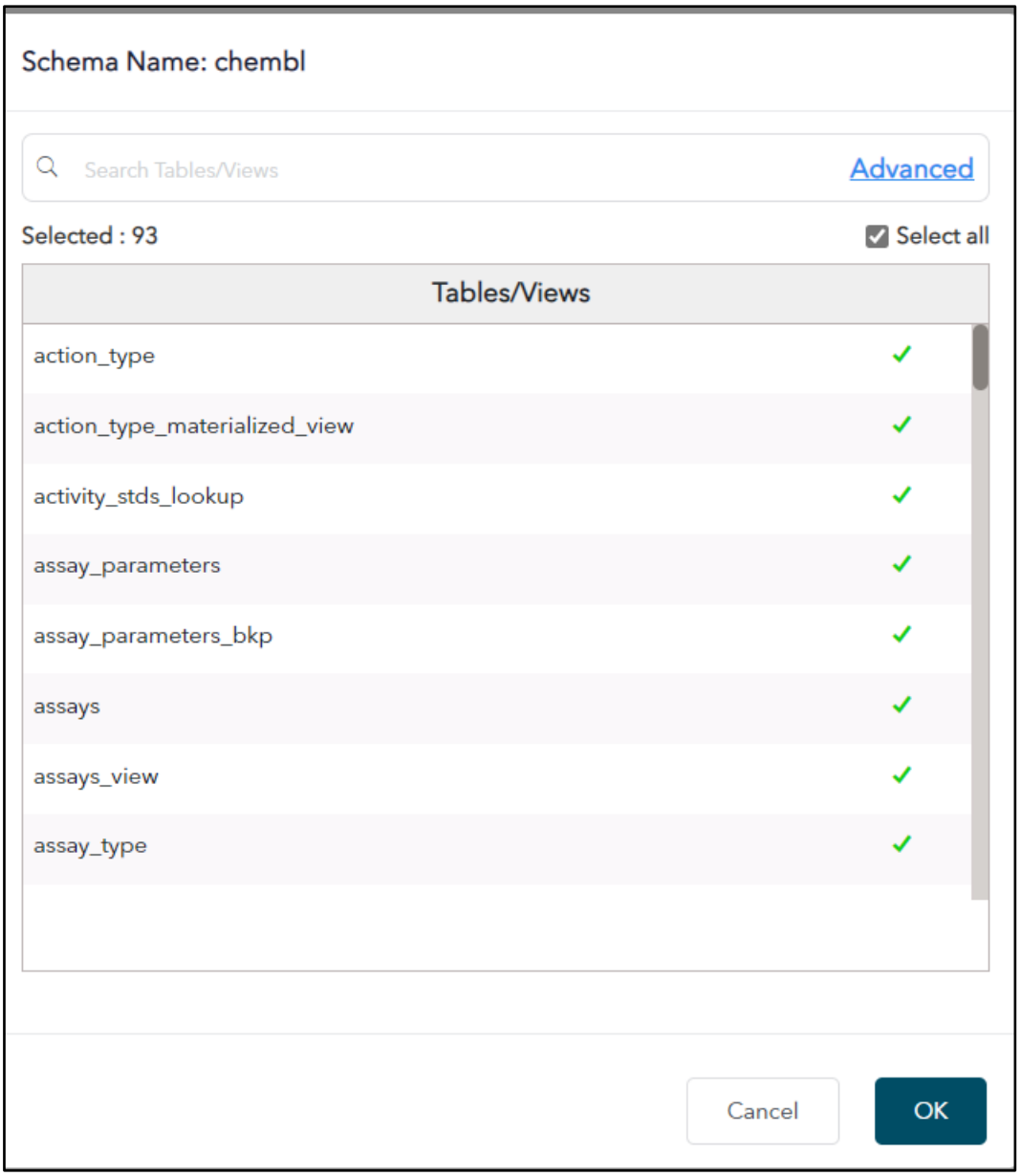


Figure 83: File Ingestion-Select/search for Tables/views

f. The advanced button allows the user to apply a filter. When you click the advanced button, an “Add filter” button is added to the popup. The user can select the filter and click on the add filter button. The results are updated with the applied filters, and the applied filter is visible as a tag in the popup's top left corner. Close the applied filter tag to remove the applied filters. The available filters are:
    i. Inclusive Regex – Include all tables/views, whose name matches a regex patter
    ii. Exclusive Regex - Exclude all tables/views, whose name matches a regex pattern
    iii. Inclusive Like – Include all tables/views whose name is like a certain name
    iv. Exclusive Like – Exclude all tables/views whose name is like a certain name

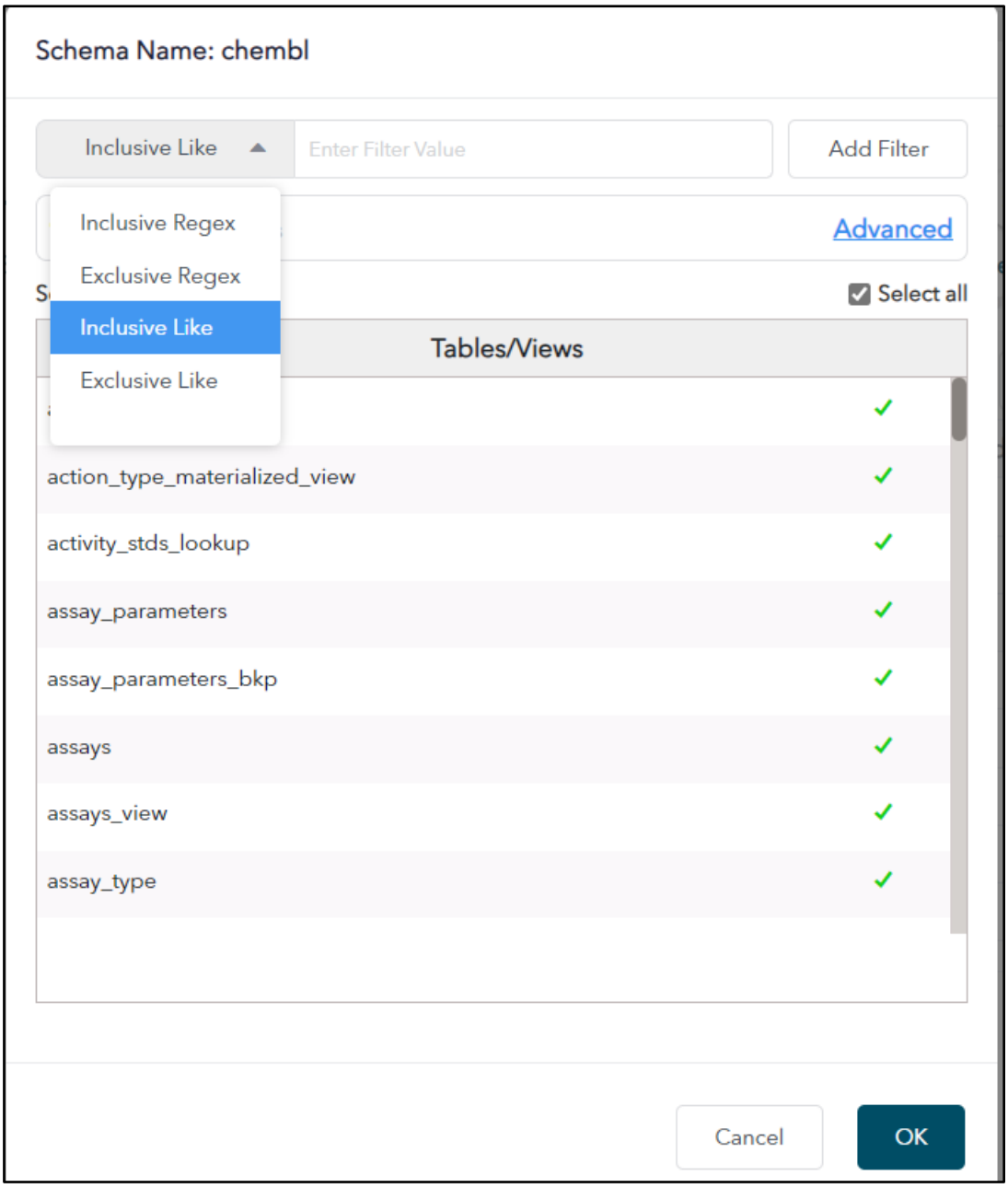


Figure 84: File Ingestion-Add Filter

5. Destination:
    a. Select a destination data connection from the dropdown menu on the left side which shows the respective schemas/containers/buckets for the respective selected data connection.

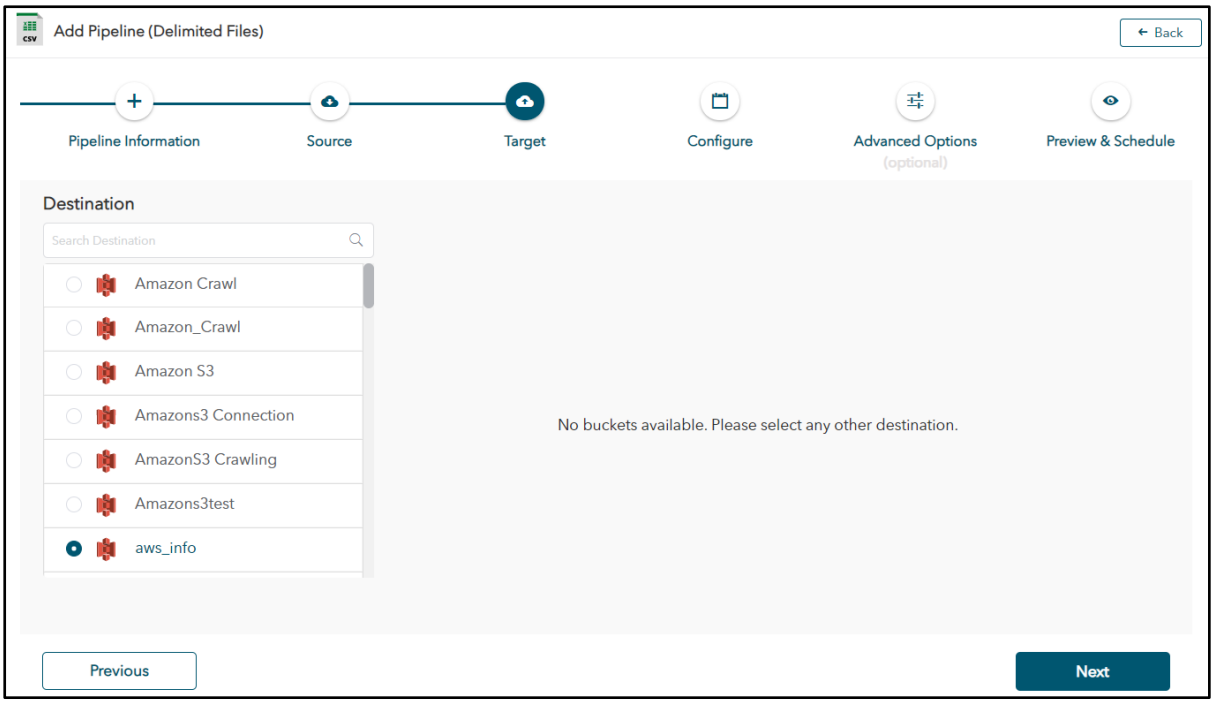


Figure 85: File Ingestion-Add Pipeline-Select/search for Destination

b. For any destination data connection with no schemas/containers/buckets, the message 'No containers/schemas/buckets available. Please select any other destination' is shown.

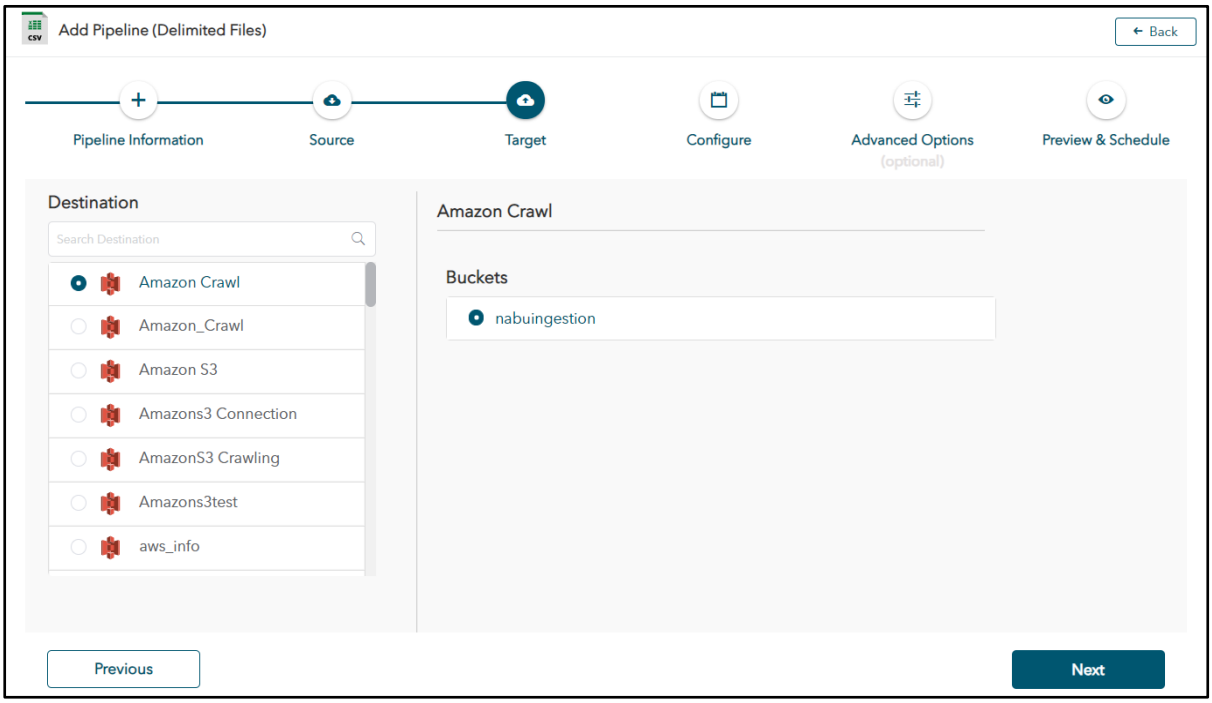


Figure 86: File Ingestion-Select/Search for Buckets

Click Previous to move to the previous step.

Click Next to move to the next step.

6. Configure:

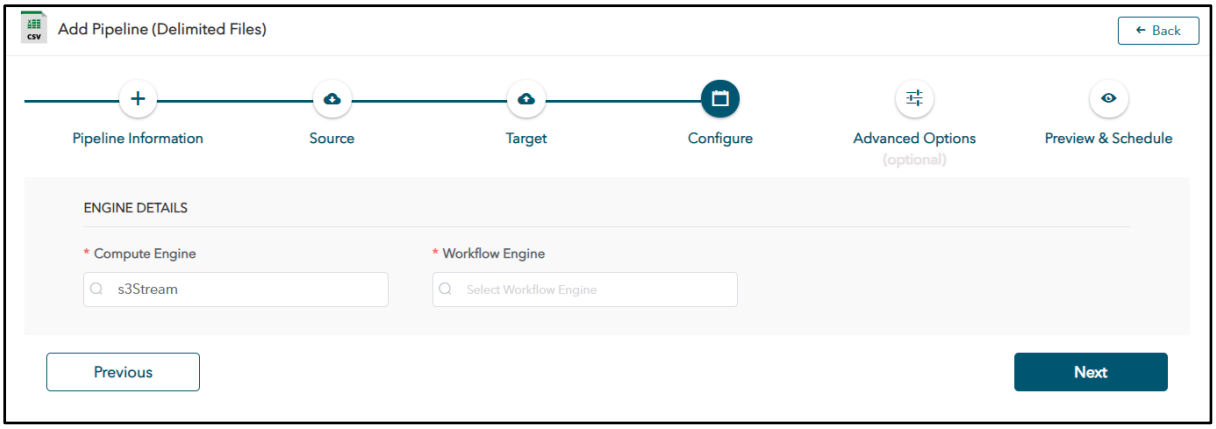


Figure 87: File Ingestion-Add Pipeline-Configure

In configuration, the user must first select the compute engine; this selection aids in the inclusion of the workflow menu option, allowing the user to select the workflow engine as well.

Click Previous to move to the previous step.

Click Next to move to the next step.

7. Advance Options: please refer to Section 3.4.3
8. Preview & Schedule:

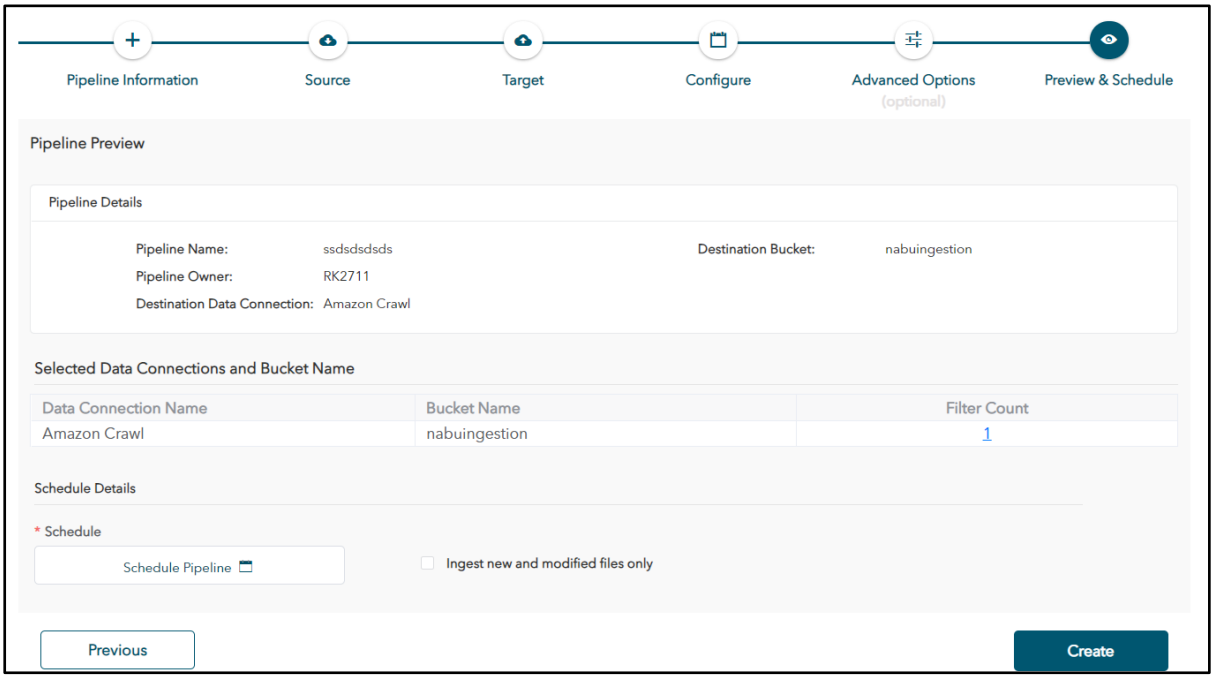


Figure 88: File Ingestion-Add Pipeline-Preview & Schedule

In this section, you can preview all the data entered in all the steps. You can schedule the pipeline by clicking on Schedule Pipeline button. For scheduling, please refer to section: 3.2.10.

The user can confirm the pipeline by clicking on "create" button.

Click Previous to move to the previous step.

4.1.1.4 Creation of Curation Pipelines

Following are the steps to execute the curation Pipeline.

1. Click on the Almaren/PySpark icon (Others section) from the left pane on Pipelines landing page.
2. You will be redirected to the “Add Pipeline” wizard for the selected pipeline type. The wizard has 4 steps.
3. Pipeline Information:

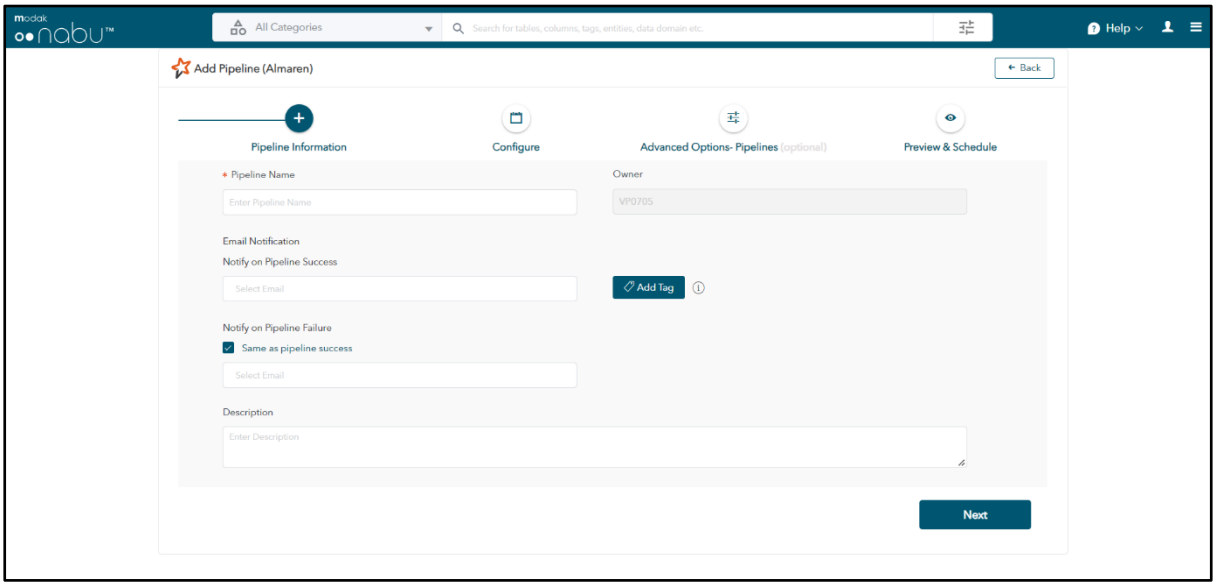


Figure 89: Curation-Add Pipeline-Pipeline Information

   a. Fill in the pipeline name which is a mandatory field. The pipeline name should start with an alphabet and should contain at least 3 characters, can contain numbers and special characters except underscore are not allowed.
   b. The owner field is auto filled with the user’s ID and disabled.
   c. Email notification: The user can select single/multiple email addresses in ‘Notify on Pipeline Success’ and ‘Notify on Pipeline Failure’ fields. The check box on ‘same as pipeline success’ will copy the same content from ‘Notify on Pipeline Success’ to ‘Notify on Pipeline Failure’ when checked.
   d. Add Tag: This allows the user to add a tag to the pipeline. For Add Tag please refer to section:3.1.10.
   e. Description: This field enables the user to enter description for the pipeline.

Click Next to move to the next step.

4. Configure:

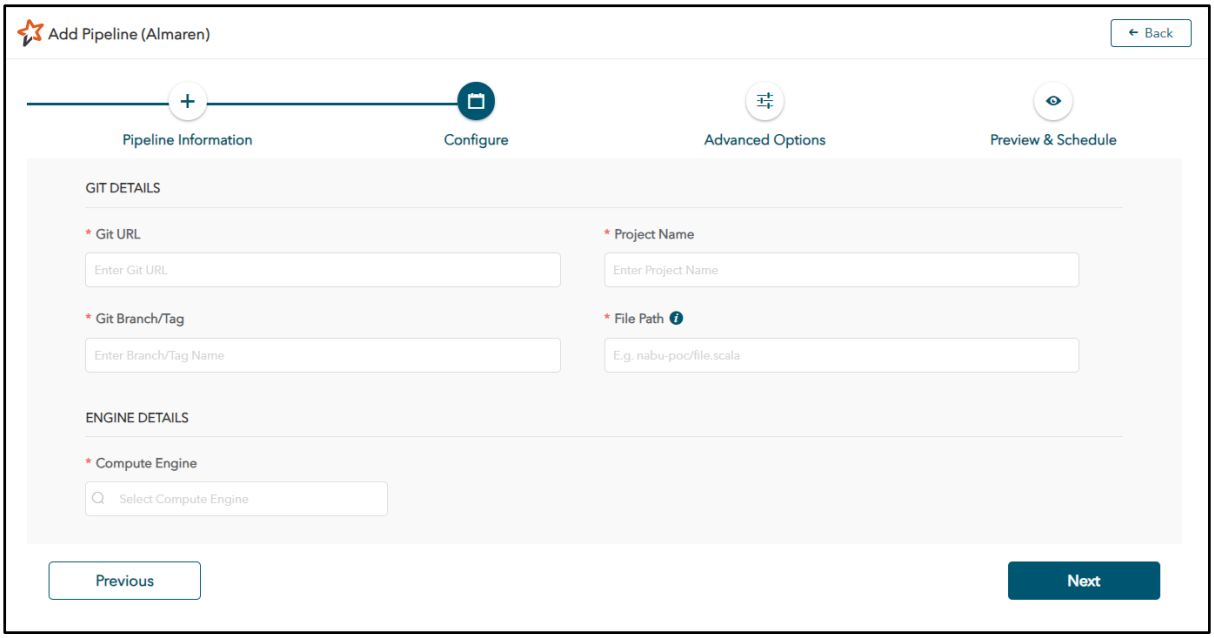


Figure 90: Curation-Add Pipeline-Configure

a. Git URL: The user can enter the Git URL to connect to.
b. Project Name: The user must enter the project name in this field.
c. Git Branch/Tag: The user must enter the Git Branch or tag for the above project.
d. File Path: The user must enter the **Scala file path for Almaren pipeline and Python file path for PySpark pipeline.**

Click Previous to move to the previous step.

Click Next to move to the next step.

5. Advanced Options:

The below are the advanced options available to the user for Almaren/PySpark pipelines.

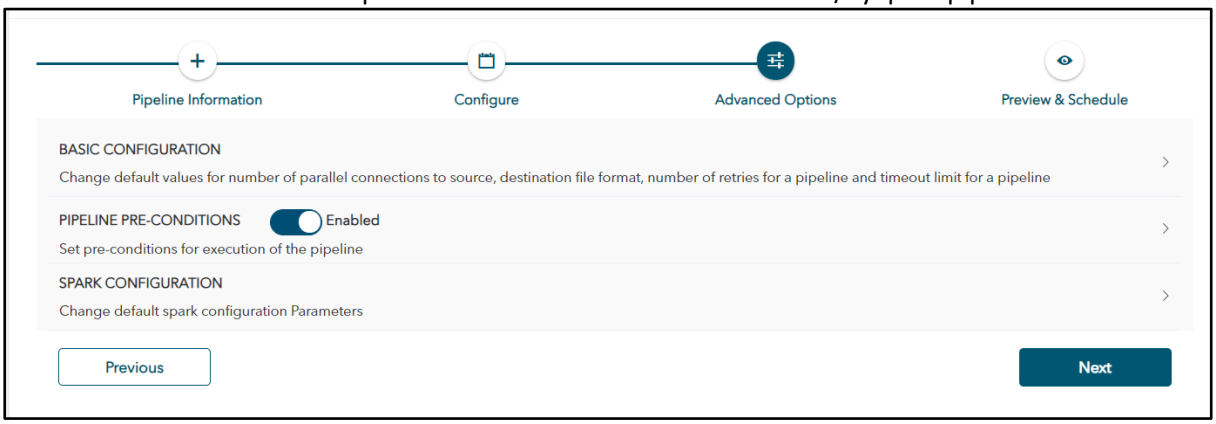


Figure 91: Curation-Add Pipeline-Advanced Options

Basic Configurations: The user can view/change the existing basic configuration for the pipeline.

1. Pipeline Retry Count: This option lets the user to specify how many times a failed object should be retried before returning an error. The default value is 3 and the maximum retry counts are 10. The default value can be changed by hovering on the textbox, the user can increment or decrement the value as required.

2. Pipeline pre-conditions: By default, the pipeline pre-conditions are disabled. The user can enable the pre-conditions by clicking on expand. Once enabled, the user can click on 'Add pipeline pre-condition' button to set pre-conditions for the pipeline.

To set the pre-conditions, user can enter the pipeline name and select the status check from the dropdown. The status can be success, error, completed etc.

The user can delete the configured pre-conditions anytime when not needed.

By clicking on the delete icon which is after the status check field.

3. Spark Configuration: The user can change the default Almaren/PySpark configurations here.

Click Previous to move to the previous step.

Click Next to move to the next step.

6. Preview & Schedule:

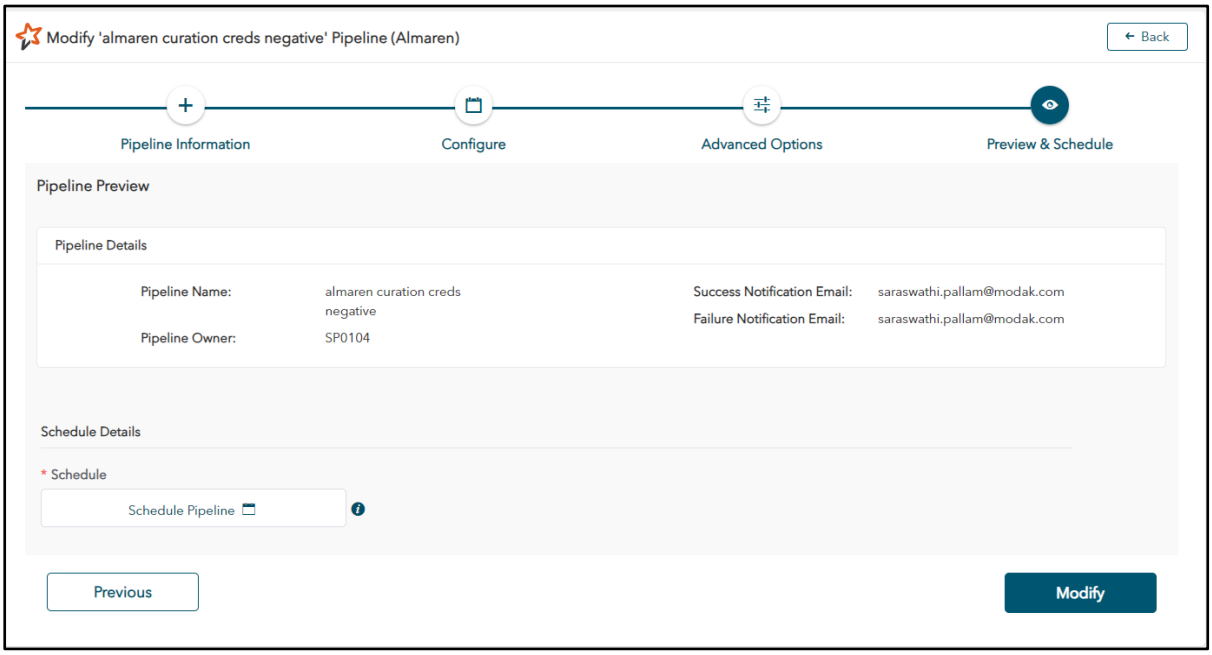


Figure 92: Curation-Add Pipeline-Preview & Schedule

In this section, you can preview all the data entered in all the steps. You can schedule the pipeline by clicking on Schedule Pipeline button. For scheduling, please refer to section: 3.2.10.

The user can confirm the pipeline by clicking on "create" button

Click Previous to move to the previous step.

Click Next to move to the next step.

4.1.2 Advanced Options - Pipeline

The Advanced options is an optional step for the create/modify Pipelines which is shown after Configure step where user selects compute engine and workflow engine.

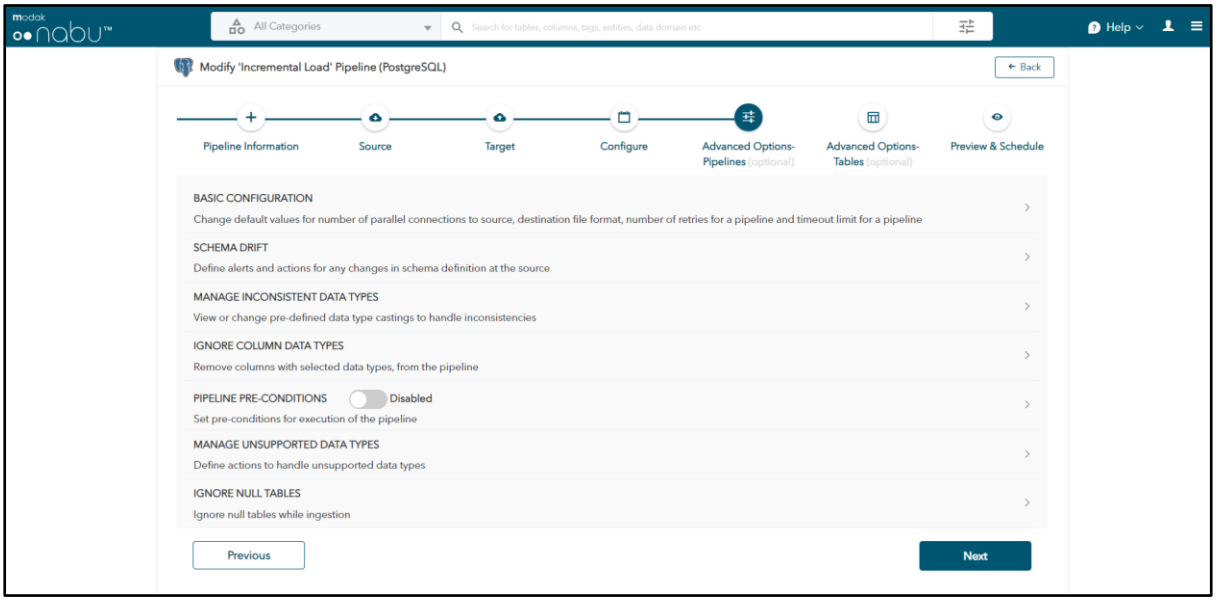


Figure 93: Advanced Options

The below are the advanced options available for the user.

Basic Configuration:

The user can view the existing basic configurations for a pipeline and can change the existing for the below options.

a. Pipeline flow timeout: Shows the maximum time a pipeline can run. The pipeline flow timeout has default value, and the default value can be changed by hovering on the textbox, the user can increment or decrement the value as required.
b. Destination file format shows the default destination file format in which the data will be written to the destination. The drop down shows other file formats, if any, and the user can change accordingly.
c. Pipelines retry count: This option lets the user to specify how many times a failed table or file should be retried before returning an error. The default value is 3 and the maximum retry counts are 10. The default value can be changed by hovering on the textbox, the user can increment or decrement the value as required.
d. Parallel source connection: This option allows the user to choose the number of parallel objects that will be processed by the pipeline. The default value is 10 and the default value can be changed by hovering on the textbox, the user can increment or decrement the source connections as required. The maximum parallel source connections that can be configured are 25.

Schema Drift Actions:

The user can define alerts or actions for any schema drift at the source. **Notify on schema drift**: By default, the 'Notify on Schema Drift' is disabled, and the user can switch to enable. On enable, the user will be shown the Email IDs field where the user can select the single/multiple email ids to notify the user.

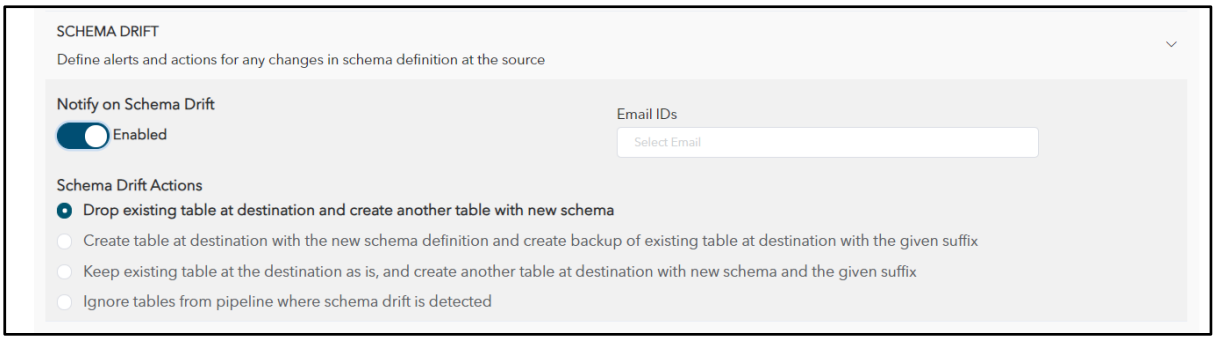


Figure 94: Notify on Schema Drift

Schema drift actions: The user can choose any of the below schema drift actions.

a. Drop existing table at destination and create another table with new schema.

b. Create a table at destination with the new schema definition and create a backup of an existing table at destination with the given suffix. The suffix which needs to be added can be provided in the 'Suffix for Back Table' field. The user can view/change the timestamp format by simply checking on the timestamp checkbox and selecting from the timestamp format drop down.

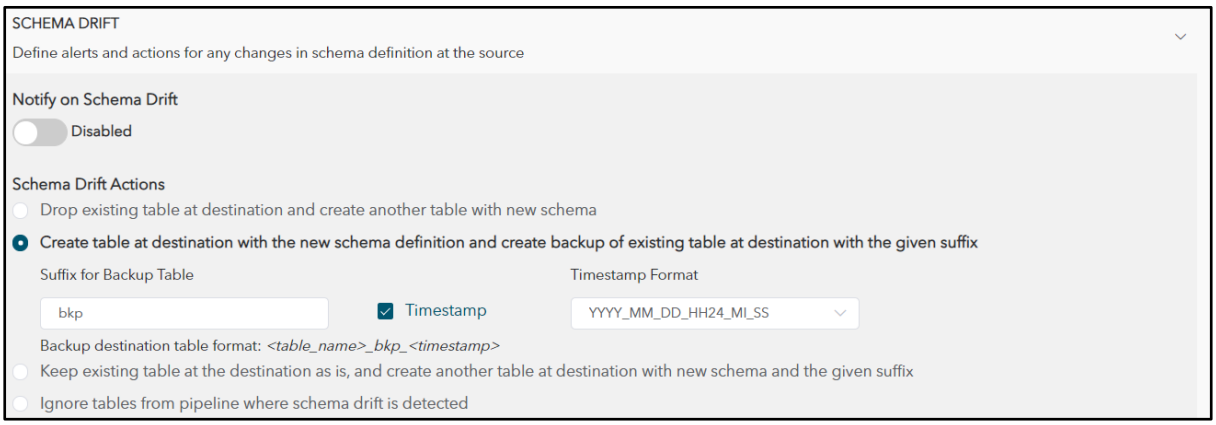


Figure 95: Creating Backup Table

c. Keep existing table at the destination as it is and create another table at destination with new schema and given suffix. The suffix for new table should be specified in 'Suffix for New Table' field. The user can view/change the timestamp format by simply checking on the timestamp checkbox and selecting from the timestamp format drop down.

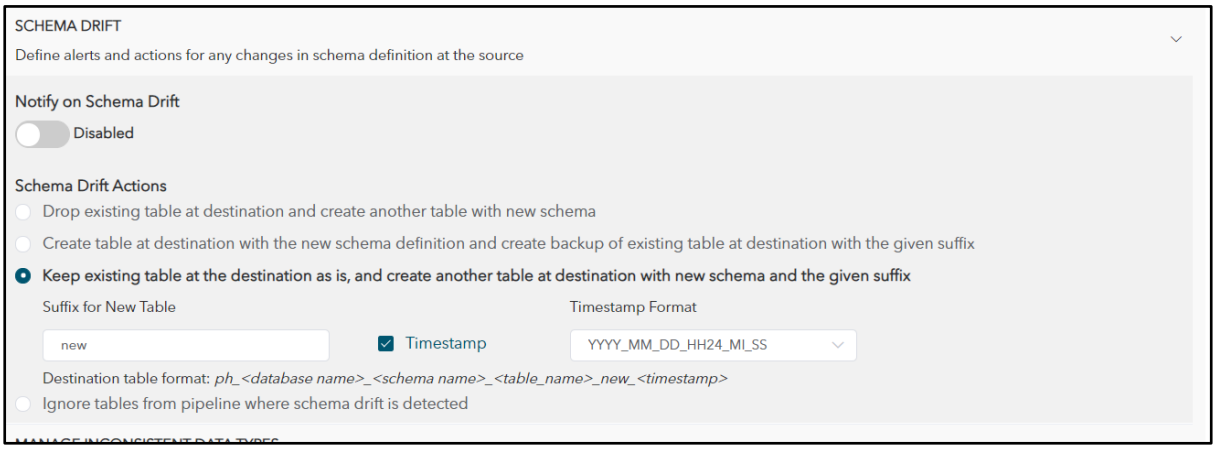


Figure 96: Create New Table

d. Ignore tables from the pipeline where schema drift is detected.

Manage Inconsistent Data Types: For any inconsistent data types from the source into the destination, the user can expand the **Manage Inconsistent Data Types** and can view the existing data type castings.

It shows the inconsistent data type mapping as below. The user can also change the destination data type for any of the source data types as required. The info icon after each source data type gives more details on the data type and mapping conversions.

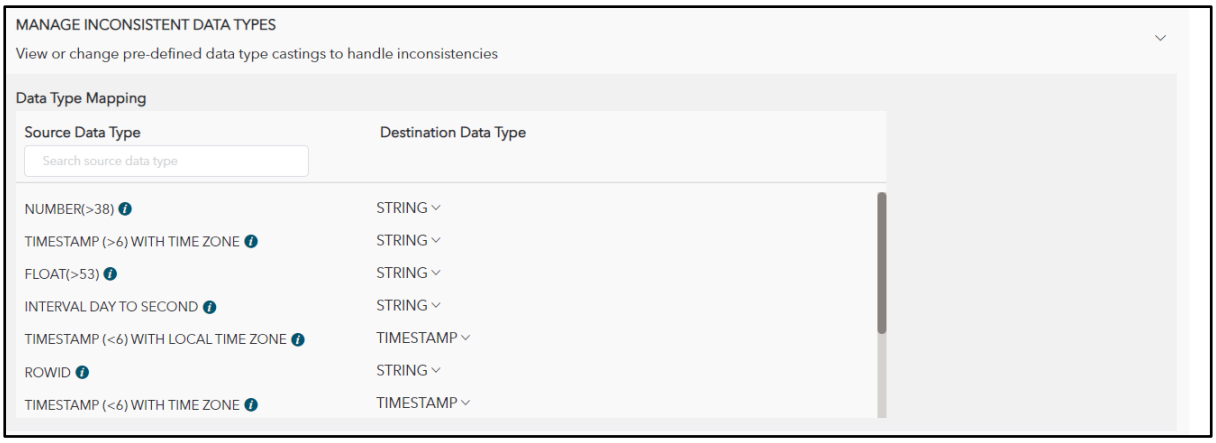


Figure 97: Manage Inconsistent Data Types

The destination data type can be selected from the available data types in the dropdown.

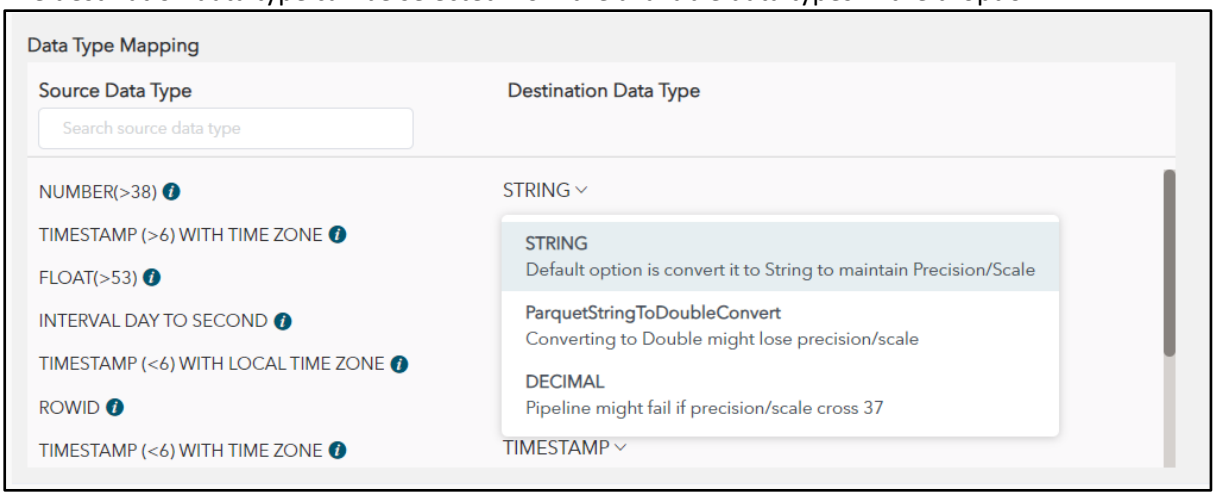


Figure 98: Data Type Mapping

Ignore Column Data Types: The user can select any specific column data types to ignore from the pipeline. The available source column data types will be displayed as below where user can check/uncheck the datatypes. Checked column data types will be ignored from the pipeline. **Search data types** allows the user to search for the required data type from the below list.

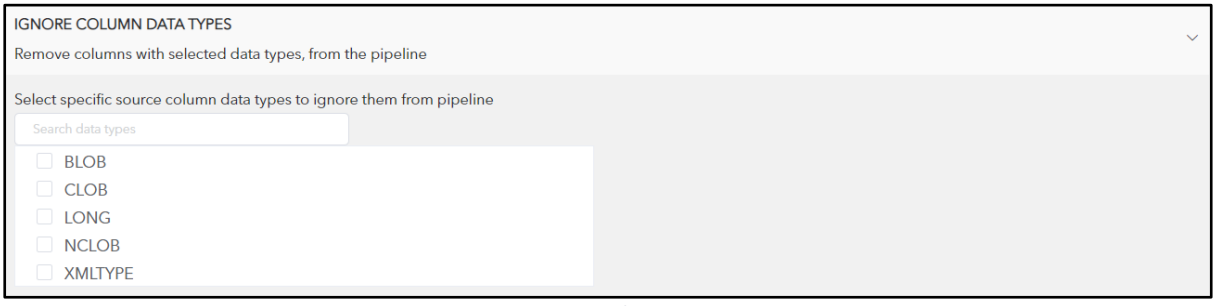


Figure 99: Ignore Column Data Types

When there no column datatypes available to ignore from the pipeline, the below information is shown as below.

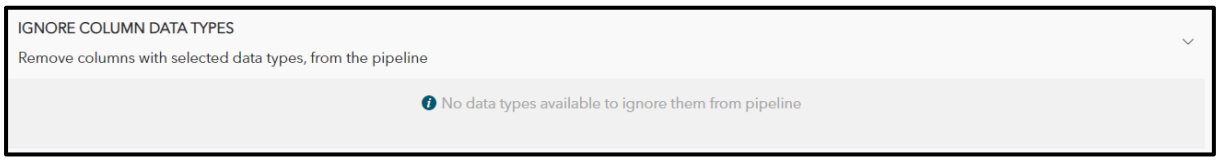


Figure 100: No Data Type Available to ignore

Pipeline pre-conditions: By default, the pipeline pre-conditions are disabled. The user can enable the pre-conditions by clicking on expand.

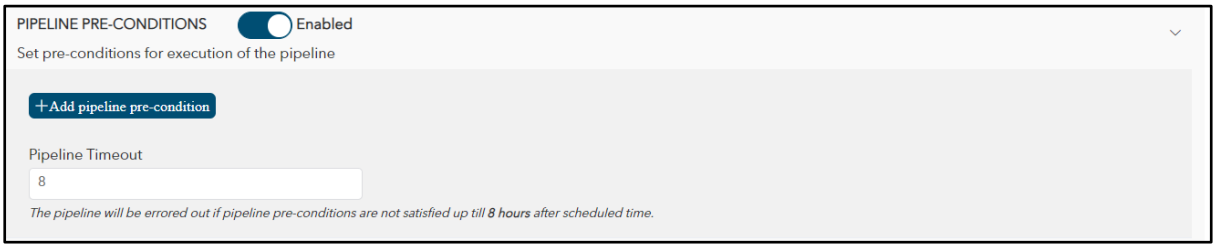


Figure 101: pipeline Pre-conditions

Once enabled, the user can click on 'Add pipeline pre-condition' button to set pre-conditions for the pipeline.

To set the pre-conditions, user can enter the pipeline name and select the status check from the dropdown. The status can be success, error, completed etc.

The user can delete the configured pre-conditions anytime when not needed by clicking on the delete icon which is after the status check field.

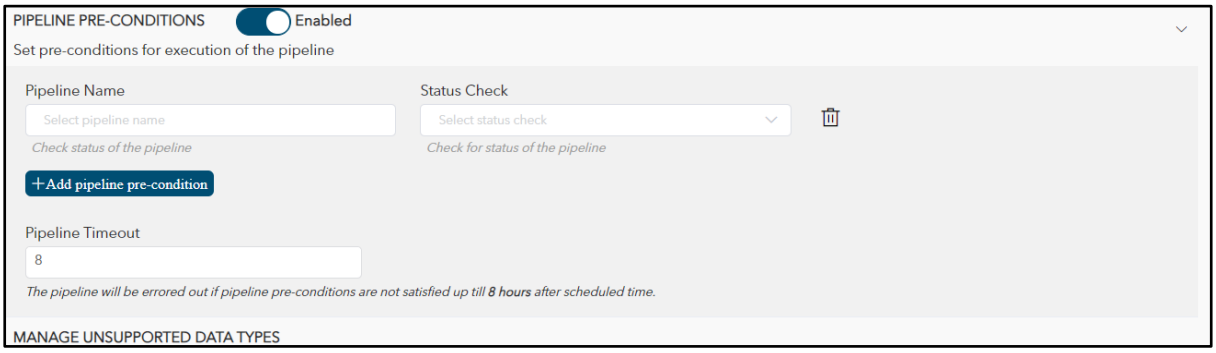


Figure 102: Add Pipeline Pre-Condition

Manage unsupported Data types: The user can define actions to handle when there are any unsupported data types for the destination in the pipeline.

For any unsupported data types, the user can choose any of the below options.

a. As Is: when user chose this option, it ingests the data into the destination in the same data type format as it exists in the source.
b. Null: It ingests the data into the destination as null.
c. Ignore: It removes the column for the selected data type from the pipeline.
d. Custom Text: It ingests the data to the destination as text.

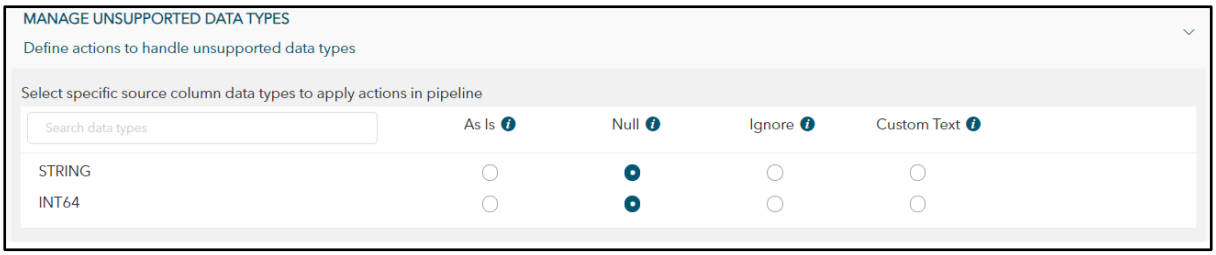


Figure 103: Manage Unsupported Data Types

When there are no unsupported data types available in the source datasets, the information is shown as below.

Figure 104: No Unsupported Data Types

To move to the next/previous section of the form, click the Next/Previous button.

Ignore Null Tables: To ignore tables that do not have any data, from the ingestion pipeline, users can enable 'Ignore Null Tables' option. By default, this option is disabled

4.1.3 Advanced Options – Table

The Advanced Options-Tables is an optional step for the creation or modification of pipelines, which is shown after the Advanced Options - Pipeline step, where the user can change the configurations for a table.

For each table(s) selected in the Source step for specific data connections and schemas, the user can configure advanced options.

Ex: If the user selected demographic_data tables in external schema for US Demographic DB Data connection. The user can configure advanced options for table demographic_data

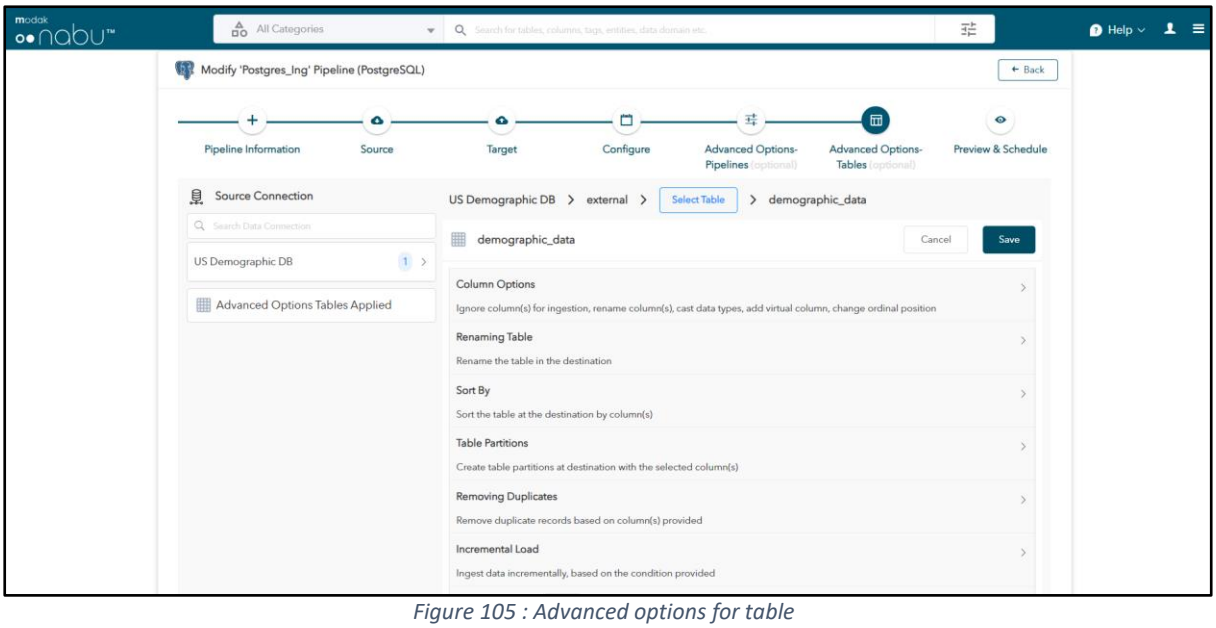


Figure 105 : Advanced options for table

The left pane shows the source data connection details. When the user clicks on the data connection, the schema and table details for the particular data connection is shown on the right.

The select table button can be clicked and it shows the list of tables that are selected in the source step.

The user can select the tables for which the configurations are required

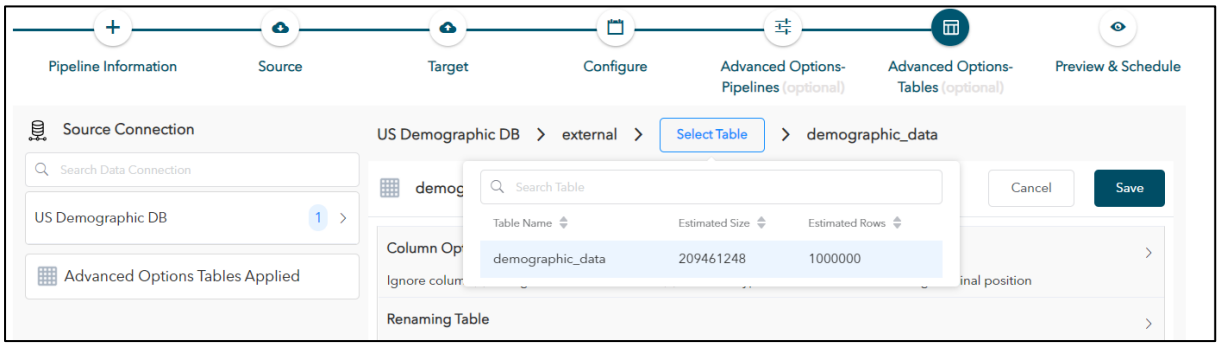


Figure 106: Advanced options for table – Select Table

The below are the advanced options for table available for the user.

Column options

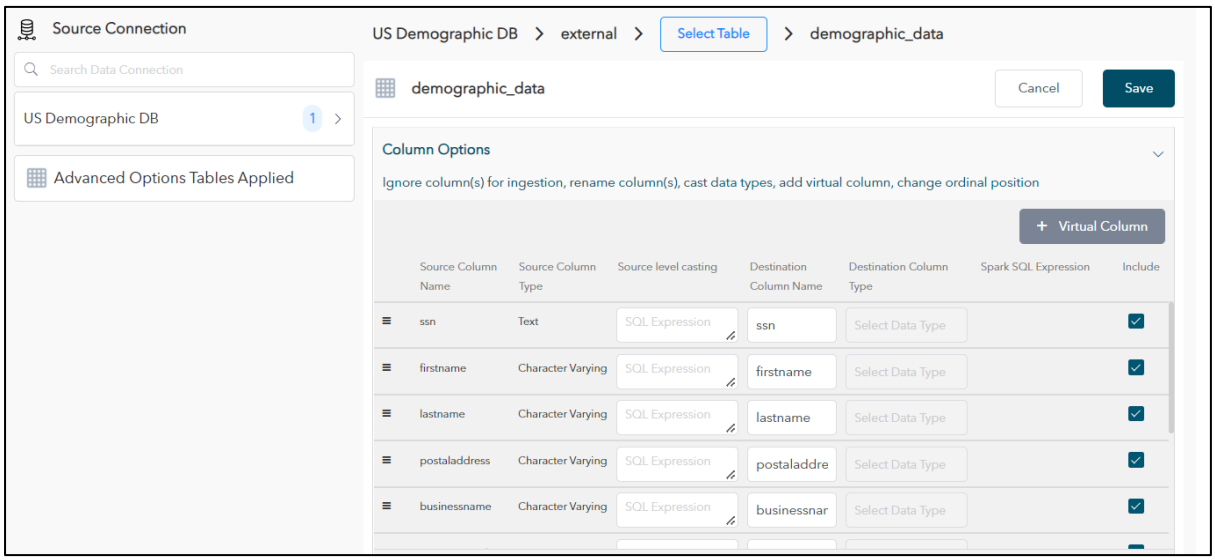


Figure 107: Advanced options for table – Column Options

When you click on "column options", it shows all the columns for the table. The below configurations can be done here.

- **Ignore columns for the ingestion:** This is an option to ignore any column from the list of columns for the ingestion. By default, all the columns are included as part of the ingestion. They include a checkbox that can be unchecked if it has to be ignored.

The ignored column will be removed at the destination.

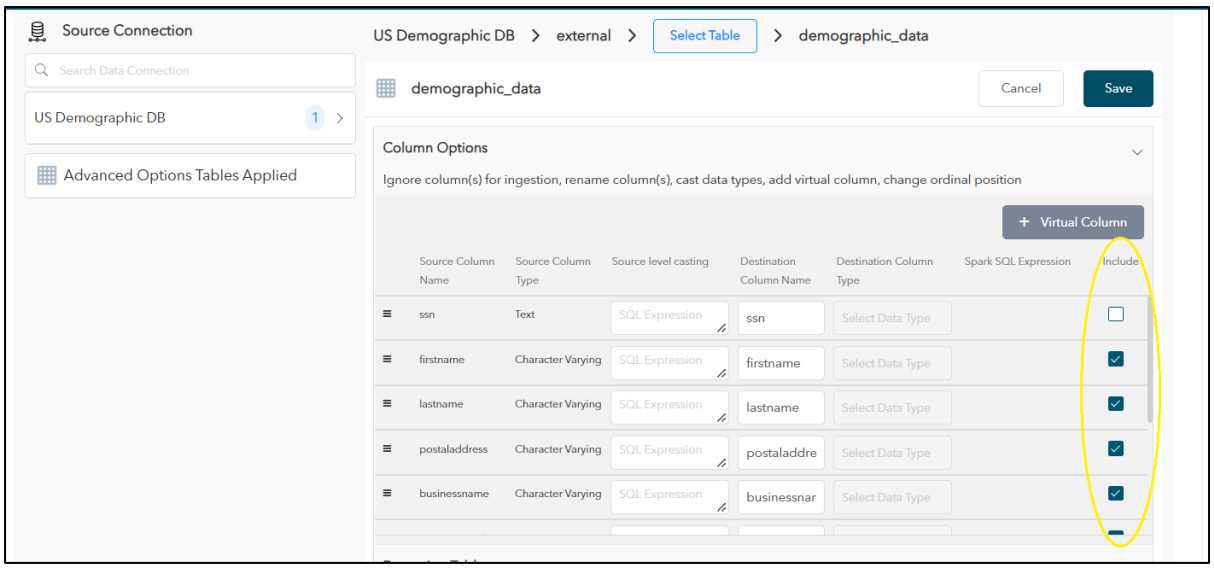


Figure 108: Advanced options for table – Include/Exclude columns

- **Rename the columns:** This option is to rename any column at the destination with the same data. For any source column, the destination column name field value can be changed.

Ex: If source column name is ssn, to rename the column as 'ssn_name' at the destination the destination column name can be given as ssn_name

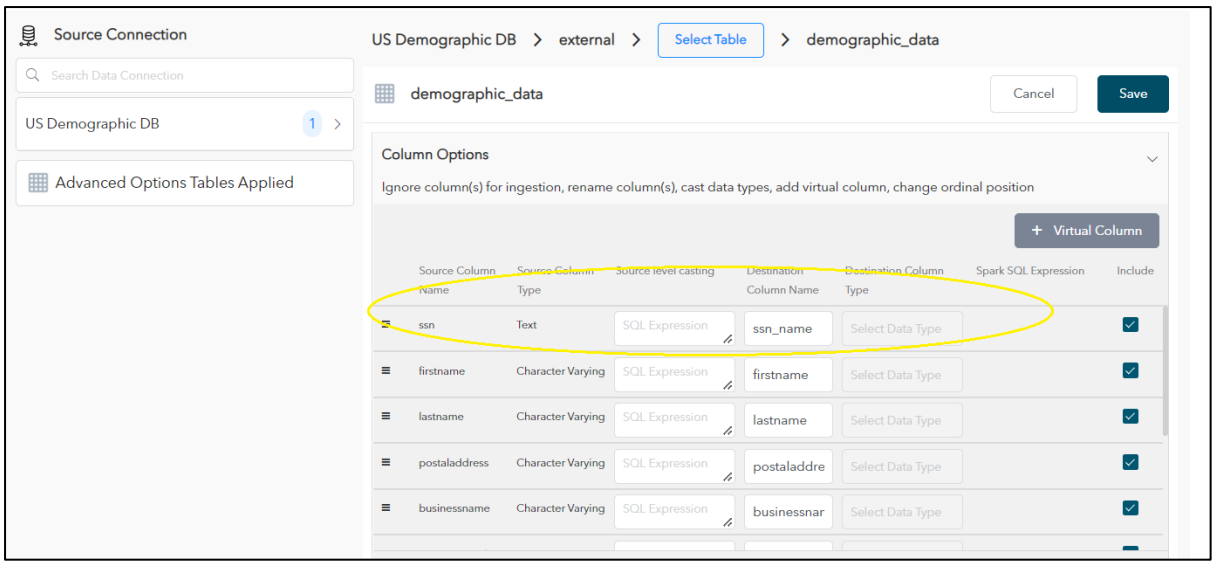


Figure 109: Advanced options for table – Rename columns

- **Add virtual column:** This is an option to create a virtual column with the selected spark data type at the destination.
- Click on +virtual column button and provide a column name and select data type. The spark SQL expression can be given to apply any function to the column and click on Add button

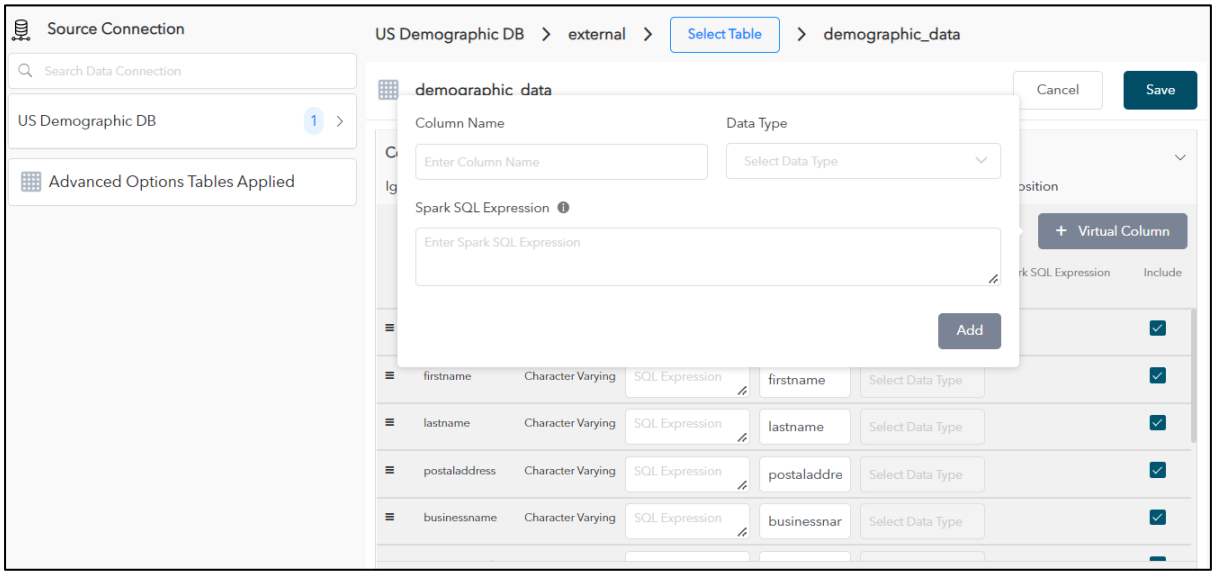


Figure 110: Advanced options for table – Add Virtual column

With the above configurations, the virtual column will be added at the destination. The info icon at **Spark SQL Expression** guides the user on applying the functions to the column.

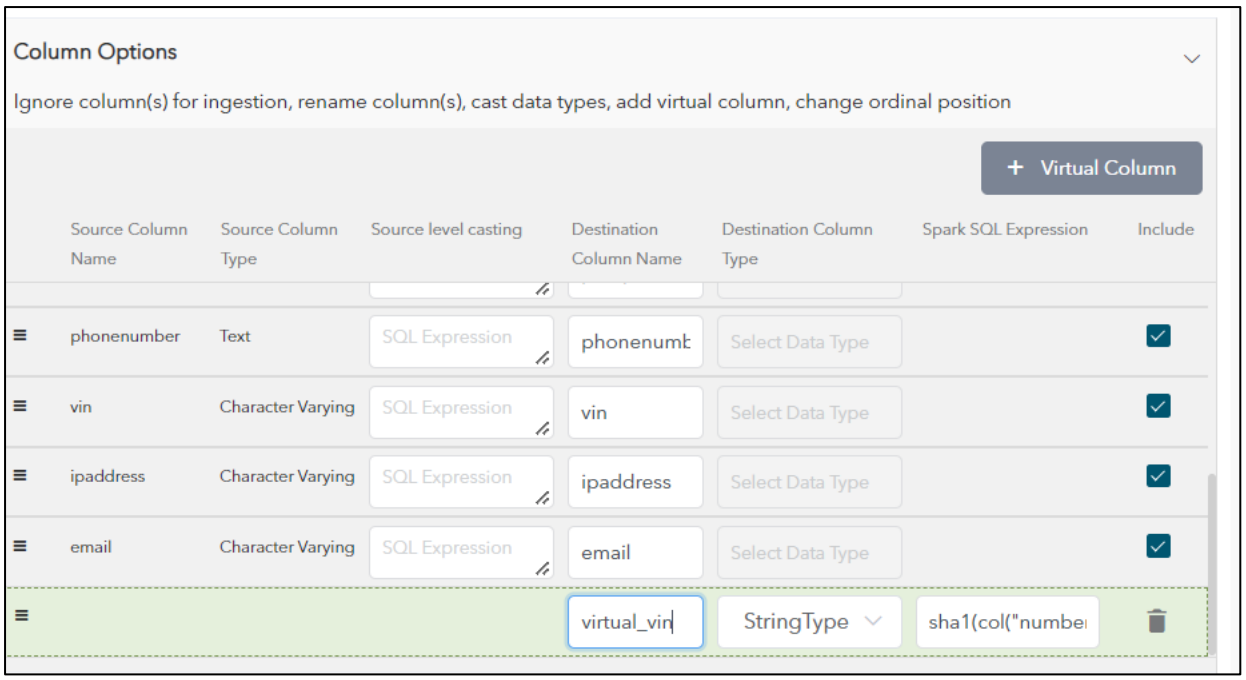


Figure 111: Advanced options for table – virtual column

Ex: 'vir_vin is added as a virtual column with new data type to show at the destination.

- **Change the ordinal position of the column:** This is an option to change or reorder the columns to show at the destination.

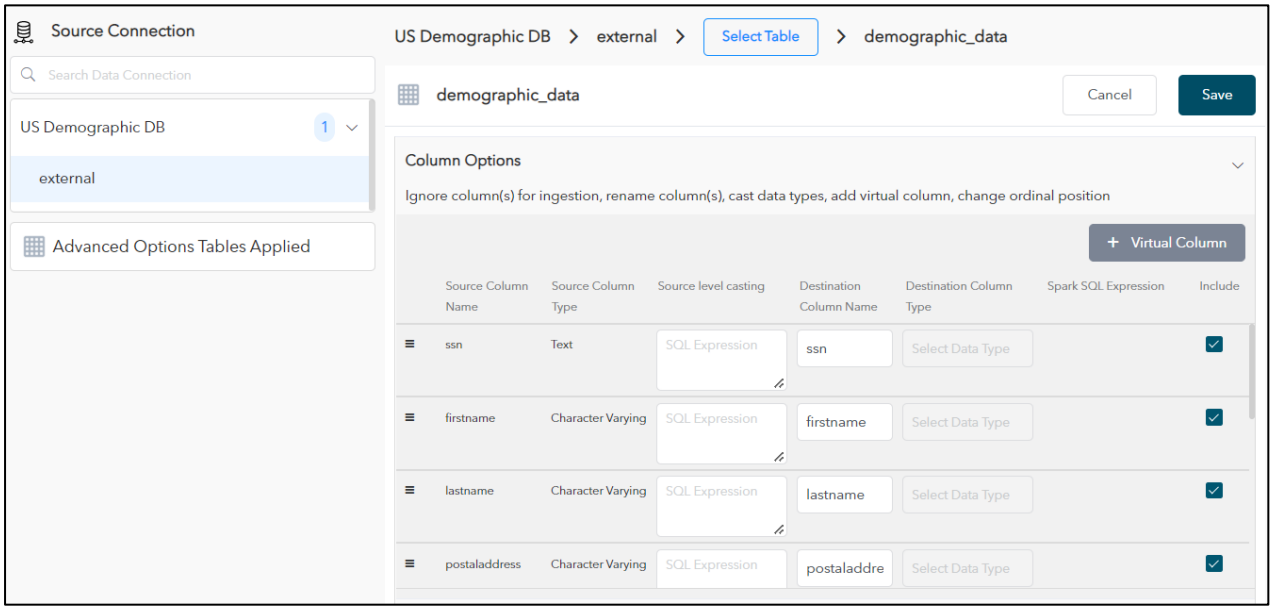


Figure 112: Advanced options for table – order columns

Ex: firstname can be reordered to be the first column at the destination. Drag the first name column and drop at the top.

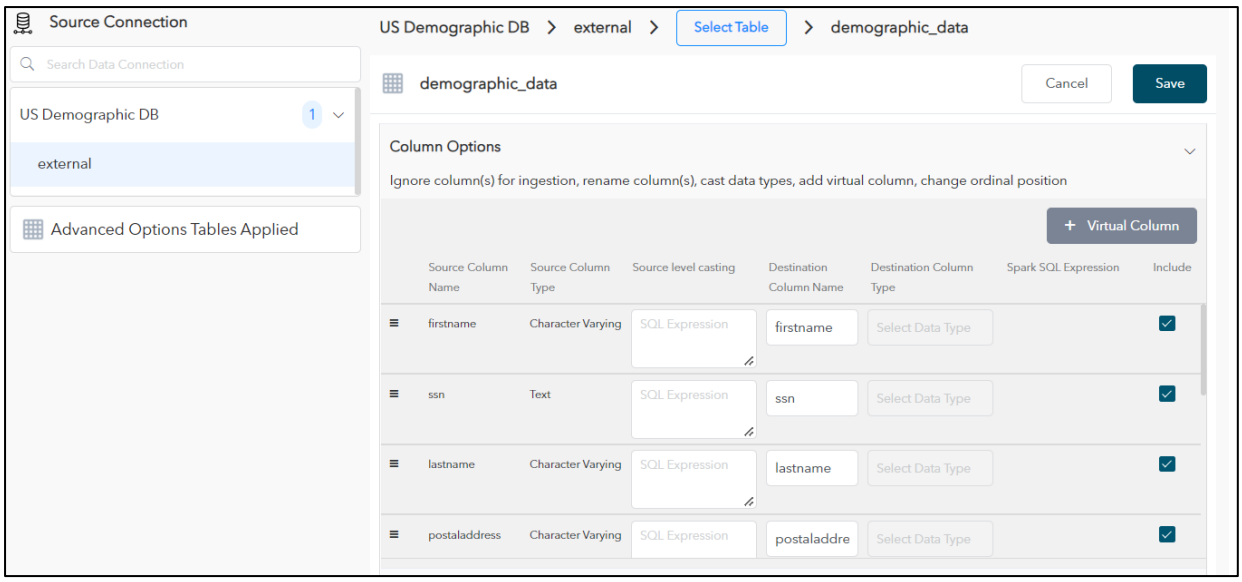


Figure 113: Advanced options for table – After columns reorder

- **Data type casting:** This is an option to make source level casting and change the destination column type accordingly.

For any destination column type casting, the user needs to perform source level casting and change the destination column type.

The destination column type is disabled by default and is enabled when the expression is provided in source level casting field.

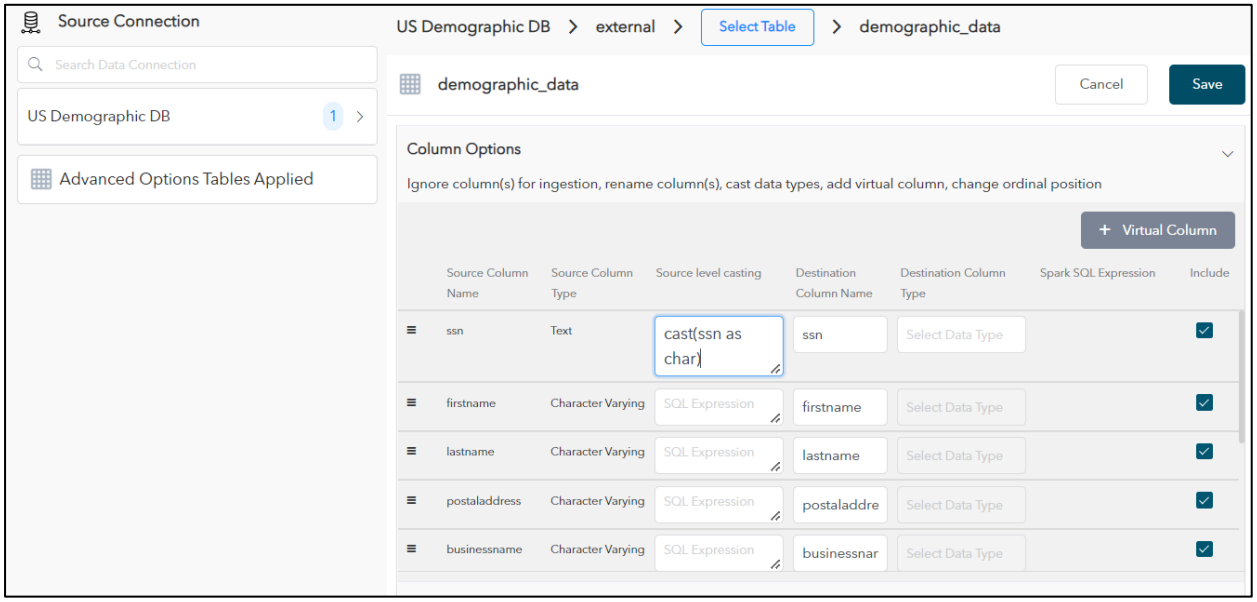


Figure 114: Advanced options for table – Data type casting

- **Spark Engine Sort:** This is an option where the user can give the names of the columns by which the ordering (ascending/descending) should be done at the Spark level. For any random data at the source, the user can sort by columns as per the preference.
    - Selecting any particular column and order by ascending or descending
    - The nulls position can also be ordered (Nulls first or Nulls last). If the column has null values, the nulls position can be ordered as per the option accordingly.

Multiple columns can be added to apply sorting. Click on '+' icon on the right to add a new row and delete icon to delete the existing row

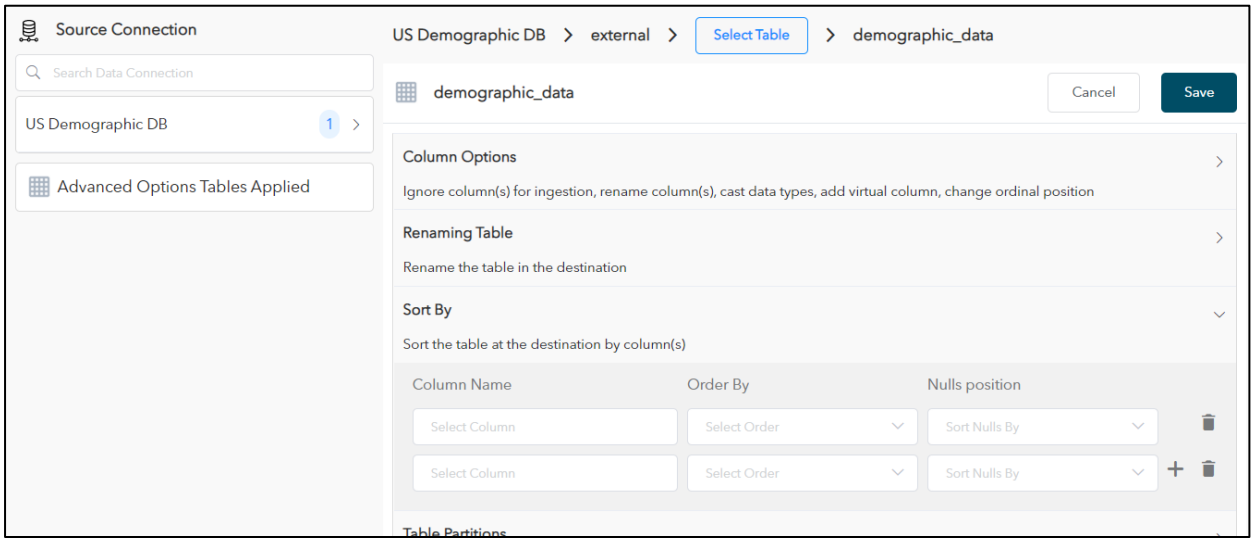


Figure 115: Advanced options for table – Sort By

- give name(s) of column(s) by which ordering should be done

Renaming table

This is an option to rename the table at the destination. The user can provide the preferred table name in the text box to show at the destination after ingestion.

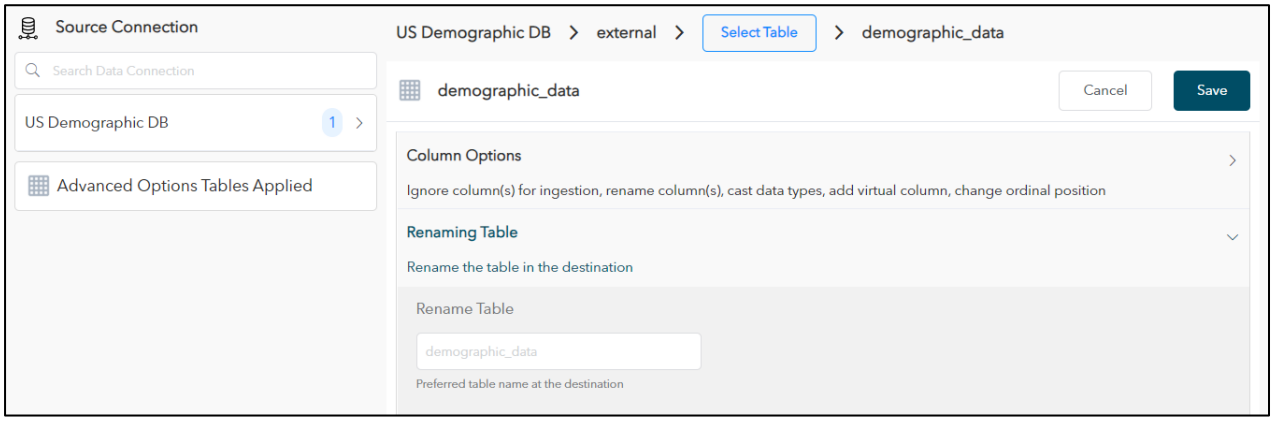


Figure 116: Advanced options for table – Rename Table

Removing duplicates

- This option is to remove duplicate records for the particular columns. The user can select the keys/columns to detect the duplicate records based on the column provided. Multiple keys and their sort can be selected to detect and remove duplicate records and write them into the destination.

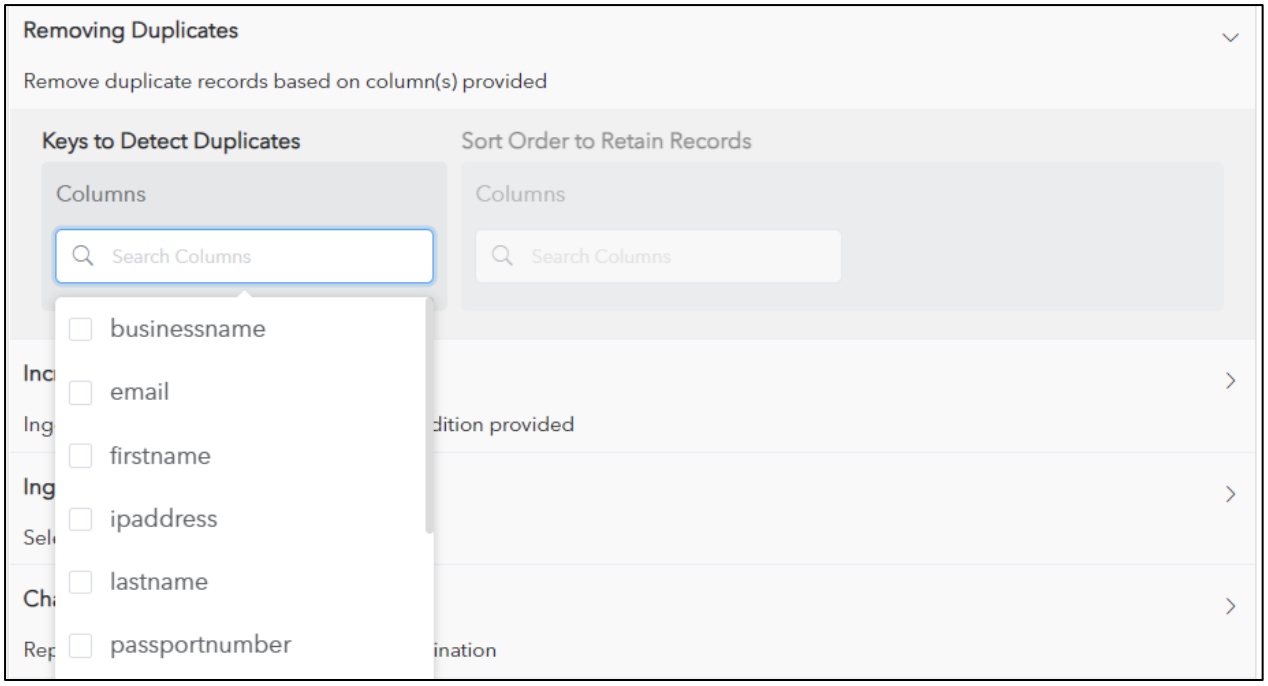


Figure 117 : Advanced options for table – keys to remove duplicates

- The sorting can also be applied for the records at the destination. The duplicate records will be removed, and the records can be sorted as provided below at the destination

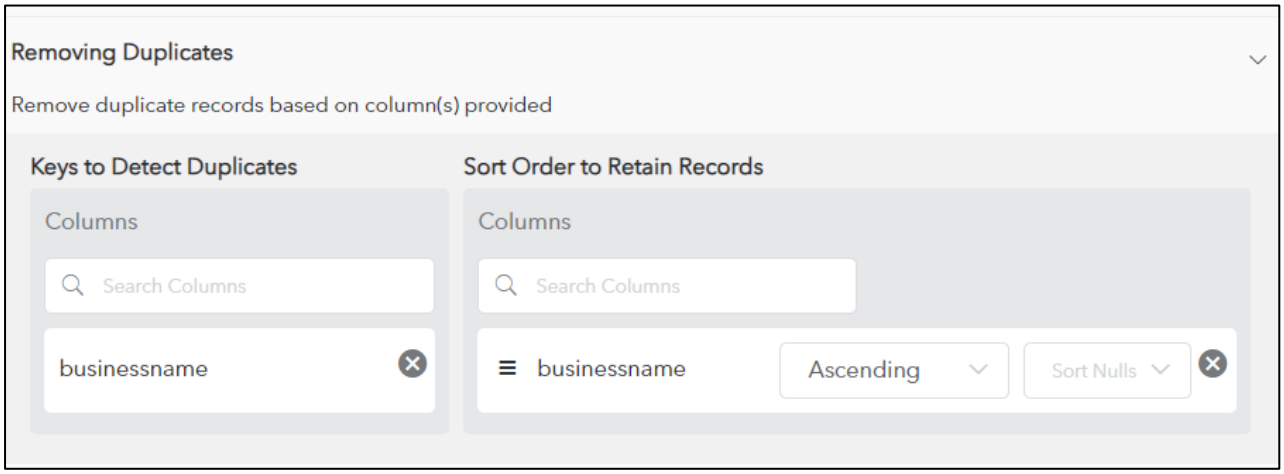


Figure 118: Advanced options for table – sort order after removing duplicates

Table partitions at destination

This is an option to create table partitions at the destination. For any vast data, after the ingestion, if the user prefers to create table partitions at the destination, the column can be selected as below.

- The selected can be removed by clicking on the close icon.

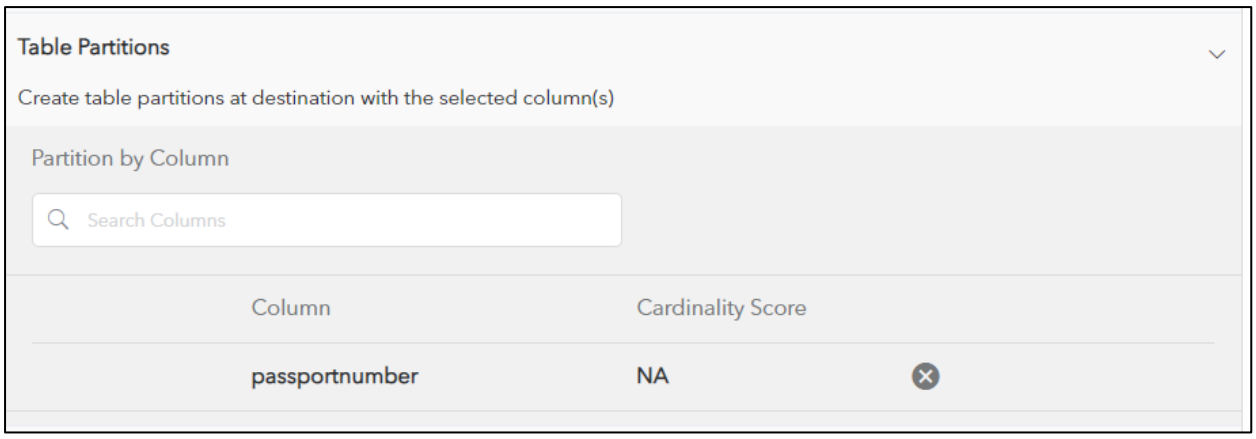
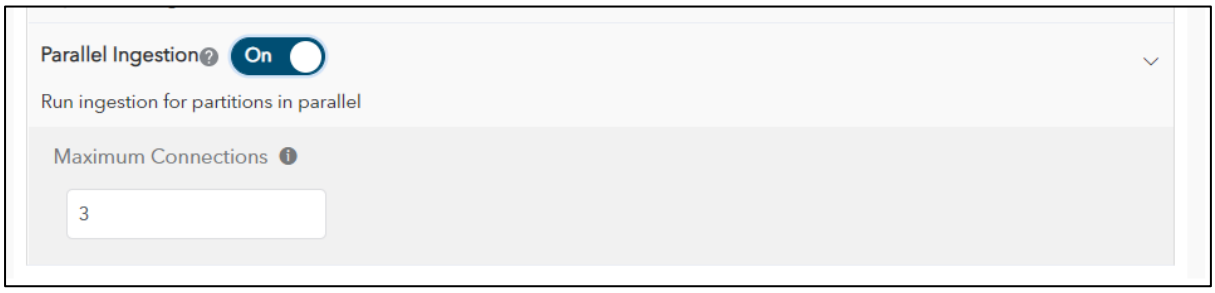


Figure 119: Advanced options for table – Table partitions

Parallel ingestion based on partition

This option is shown when there are table partitions at the source**.**

- The maximum number of connections should be provided for the partitions.
- The data is ingested in parallel based on partitions and the number of connections.



Incremental Load

This is an option to ingests data in small chunks based on the Where Clause condition or default values.

- Provide the condition for executing incremental load using placeholders or default values.
- Example: ssn > '<%ssn_name%>'
- Use placeholder by providing placeholder name that can be used in the where condition.
- Default value of placeholder will be replaced within the placeholder's name provided in the where condition.
- To change the default value from the next run, the max value column can be selected. Max value of this column will be considered for the next pipeline run.
- Multiple placeholders can be added by clicking on + icon
- Click on validate button to validate the condition

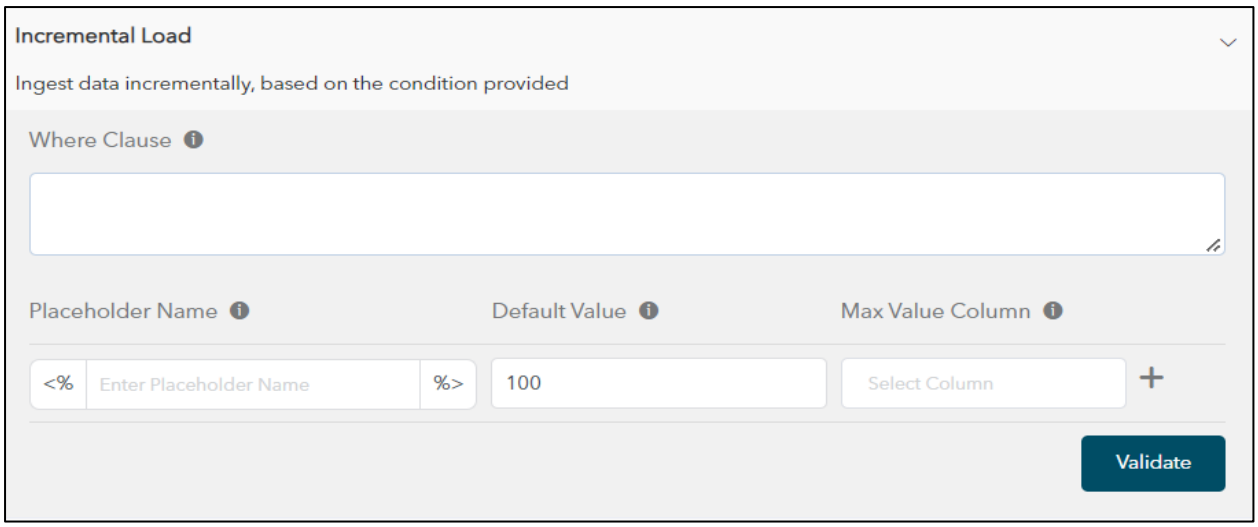


Figure 120: Advanced options for table – Incremental load

Slowly changing dimension (SCD)

- This option is to capture the changed data from the source to the destination instead of loading the entire table.
- The below are the SCD Process types:

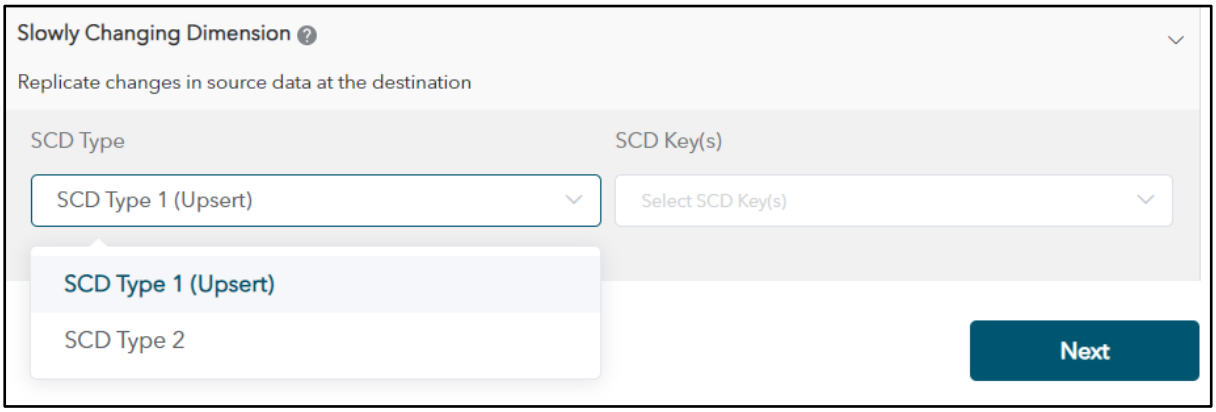


Figure 121 : SCD process types

- SCD Type 2: All the history data is maintained in the destination with the valid columns. For SCD Type 2, the process key needs to be selected from Unique key or Generate Hash key.
- For unique key selection, the user can select the unique keys from the columns field.

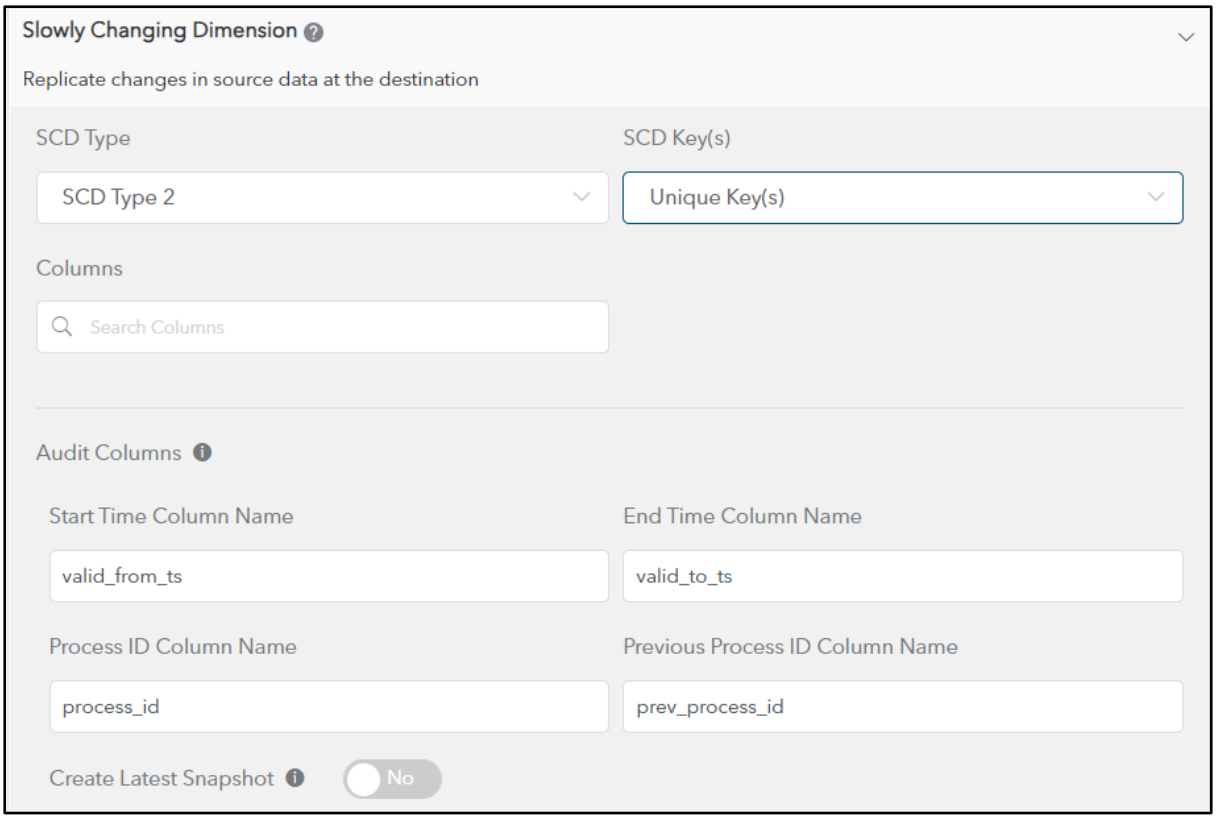


Figure 122 : SCD unique key

- For Generate Hash key selection, the user needs to select the hash type from the Hash type field as shown below.

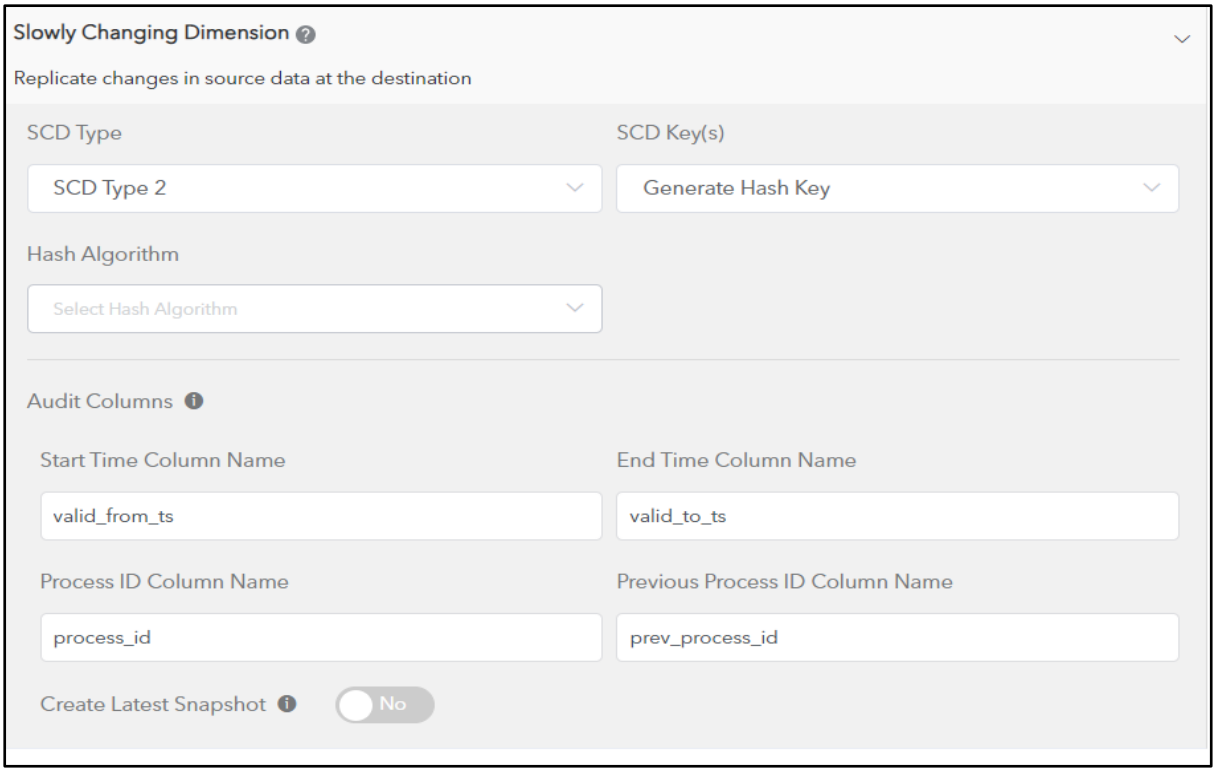


Figure 123 : SCD Generate Hash Key

- The Audit columns are mandatory for SCD Type 2 process type. The user can change the audit columns and they cannot be edited back again for the same pipeline.

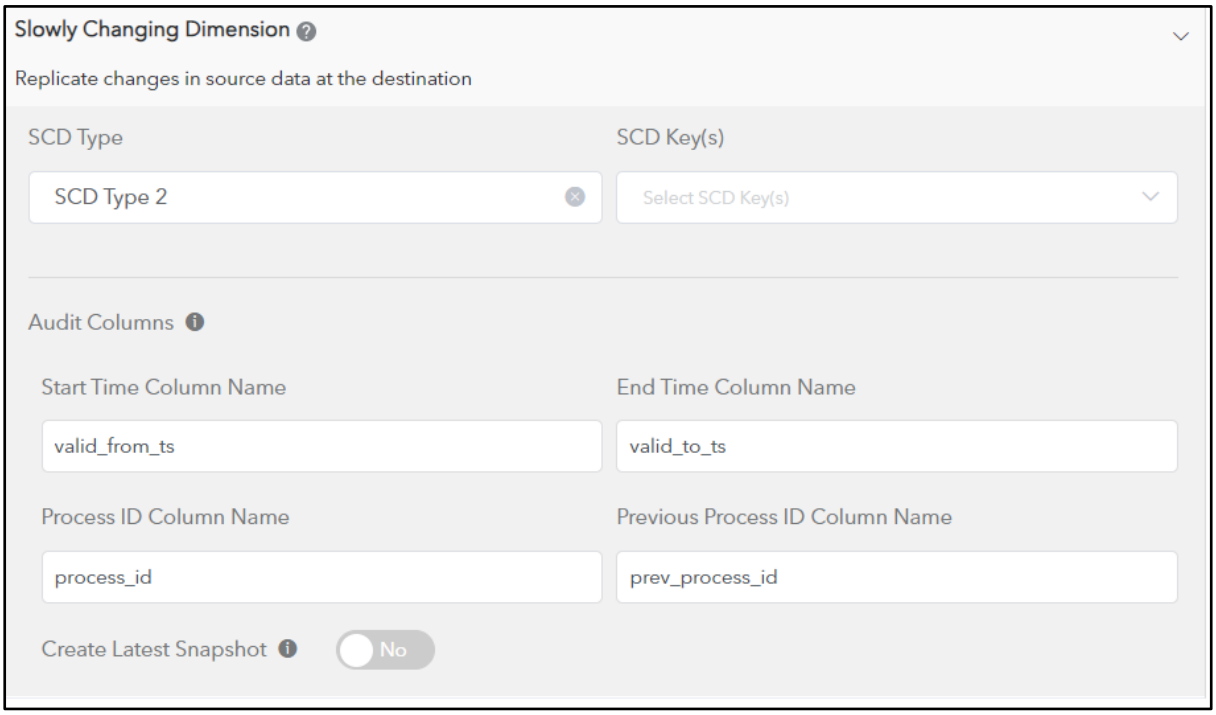


Figure 124 : SCD Audit Columns

- The create snapshot switch creates a view with the latest records. The user can turn the switch to enable.

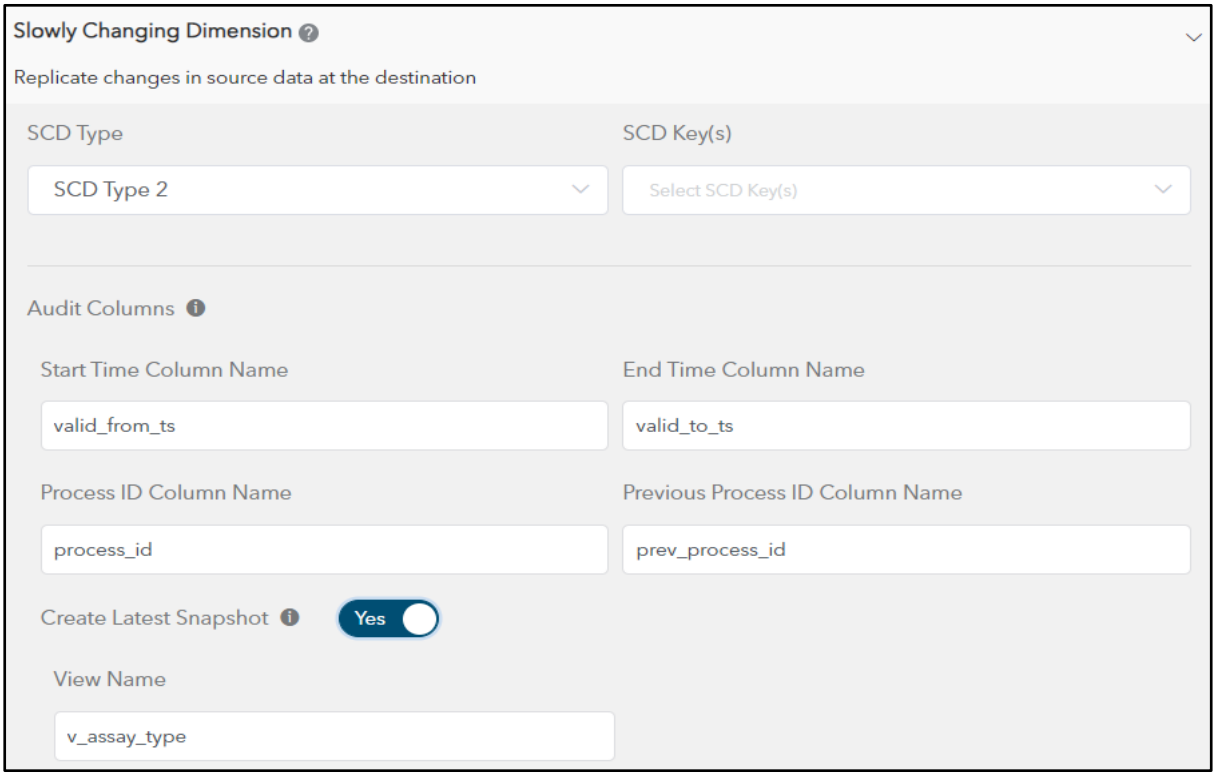


Figure 125 : SCD Create latest Snapshot

- Upsert: Will have all the deleted records without the valid columns and it performs only inserts and updates on destination.
- For upsert selection, the user needs to select the process key. For unique key, the user needs to select the columns for the unique key as shown below.

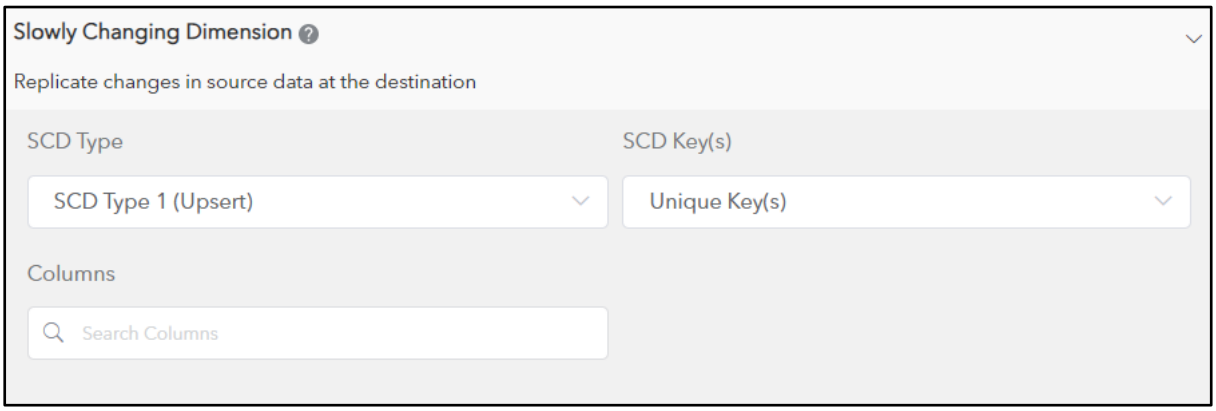


Figure 126 : SCD Upsert process type

Ingestion Mode (Drop & Recreate, Append, Overwrite):

- This option is to select mode to ingest the table. The ingestion mode is Off by default

Figure 127: Advanced options for table – Ingestion Mode Off

- Turn the ingestion mode switch to enable the modes.
  - **Append:** The ingested data will be appended to the existing data at the destination.
  - **Overwrite:** The ingested data will be overwritten at the destination.
  - **Ignore:** The Ignore mode ignores the ingesting data, if data already exists and the save operation is expected to not save the contents and to not change the existing data.
  - **Drop & Create:** The existing data will be dropped, and the ingested data is created at the destination.

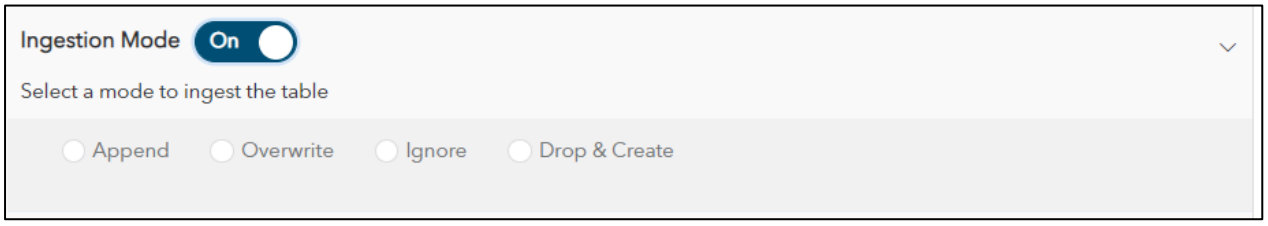


Figure 128: Advanced options for table – Ingestion Mode On

4.1.4 Edit Pipelines

The user can edit any of the above pipelines from the Pipeline landing page as below.

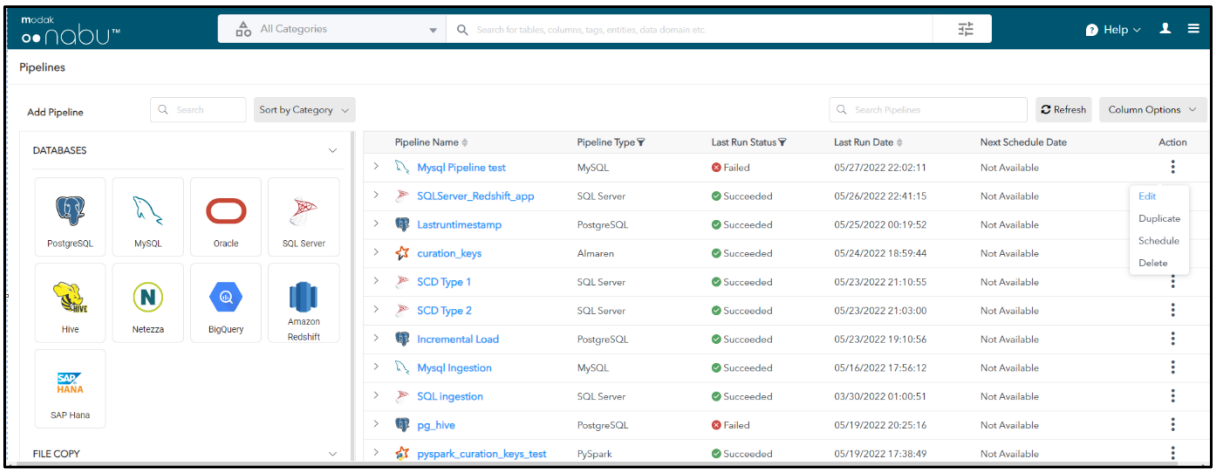


Figure 129: Edit Pipeline

The user can click on the Action column and select the Edit option or click on the pipeline name.

The user will be redirected to the respective pipeline with all the previous details. The user can edit the pipeline as required and click on the modify button.

All the details entered/modified should be valid, else the user will not be able to modify the data pipeline

4.1.5 Duplicate pipeline

The user can duplicate any of the above pipeline from the pipelines landing page as below.

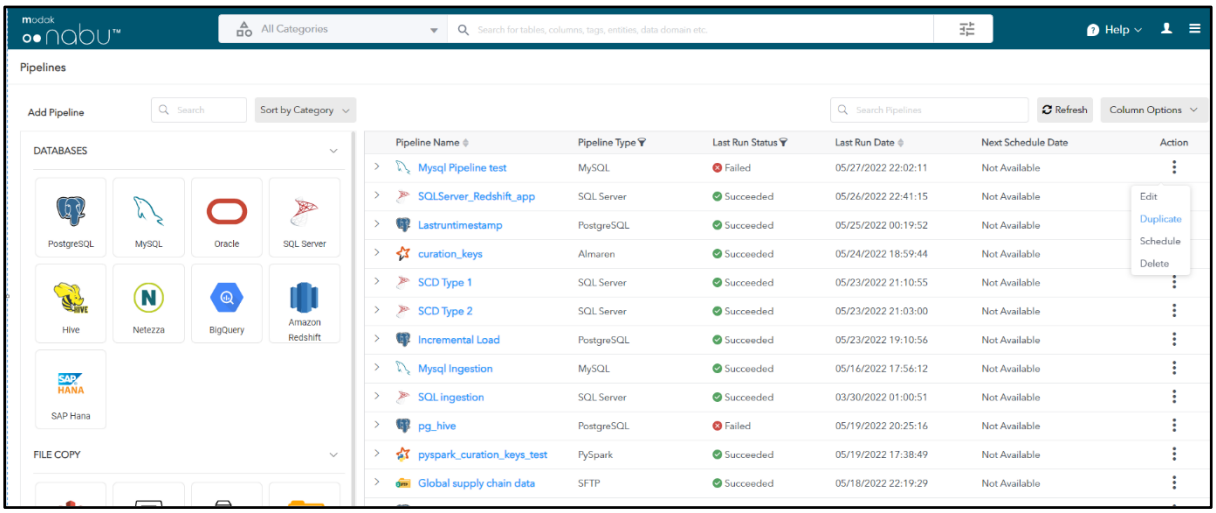


Figure 130: Duplicate Pipeline

The user can click on the Action column and select the duplicate option.

On selecting the duplicate option, the duplicate or copy of the selected pipeline is created and the user will be redirected to the selected pipeline with all the details. The user can take action as per their preference.

All the details entered should be valid, else the user will not be able to create the data pipeline.

4.1.6 Schedule Data pipeline

The user can schedule ingestion for any of the above pipeline from the pipelines landing page as below.

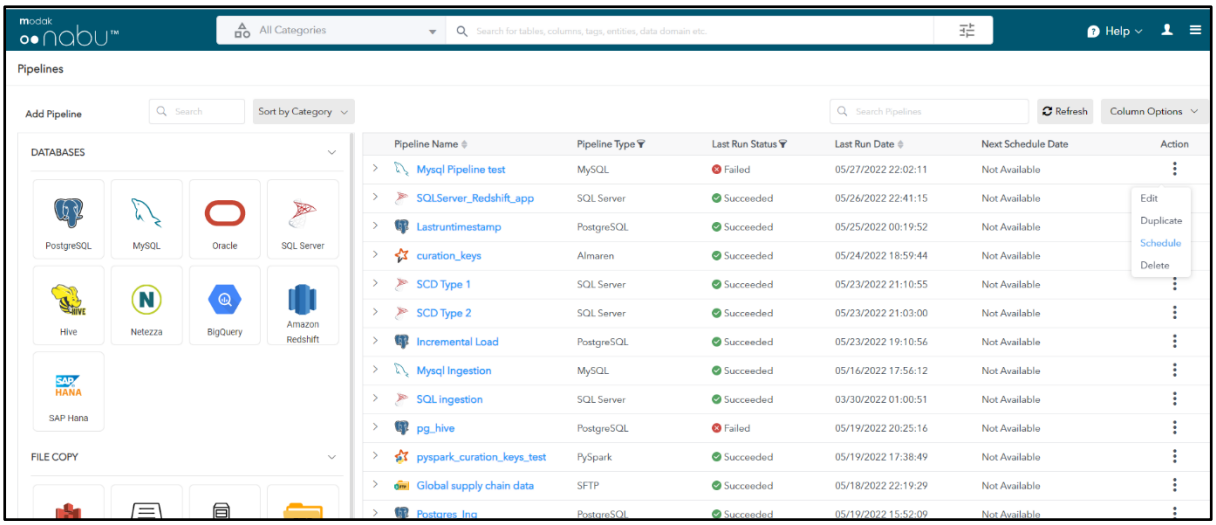


Figure 131: Schedule Pipeline

On clicking schedule, the user will be prompted with a popup as below where the user can select the schedule frequency. For scheduling, please refer to section: 3.1.9.

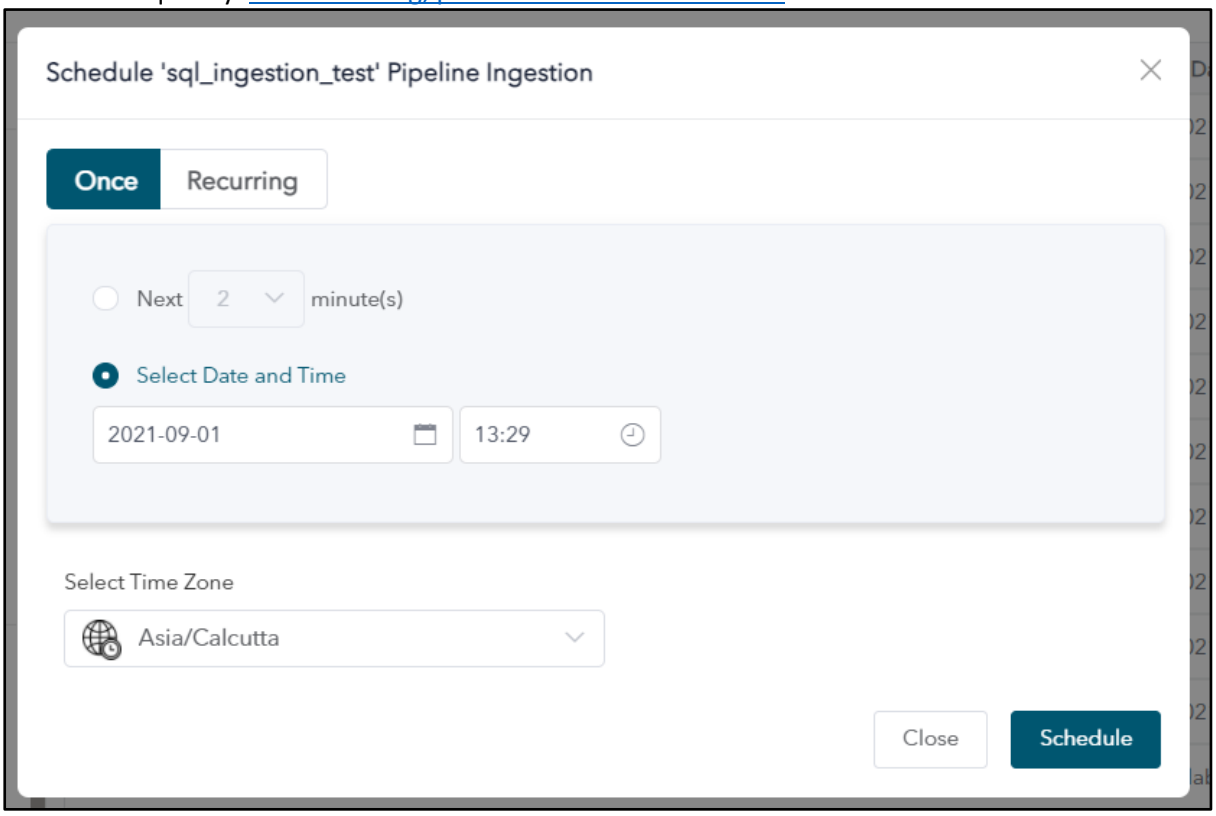


Figure 132: Pipeline Ingestion

4.1.7 Delete pipeline

The user can delete any of the pipeline from the pipeline landing page as below.

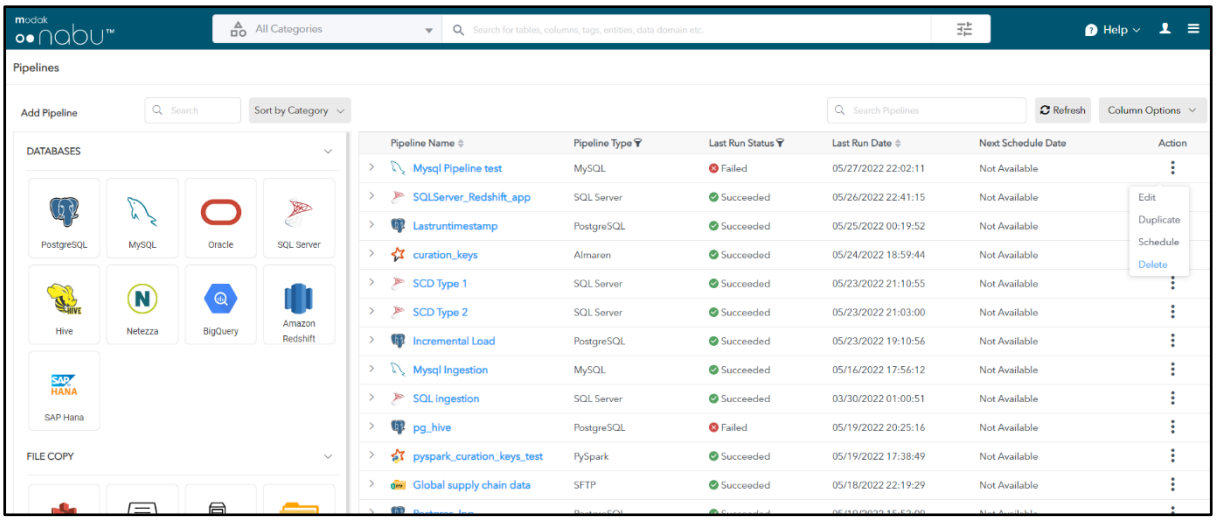


Figure 133: Delete Pipeline

The user can click on the Action column and select the Delete option.

On selecting delete option, a popup showing the impacted pipeline will be shown as below.

If the user clicks ok, the selected pipeline will be deleted.

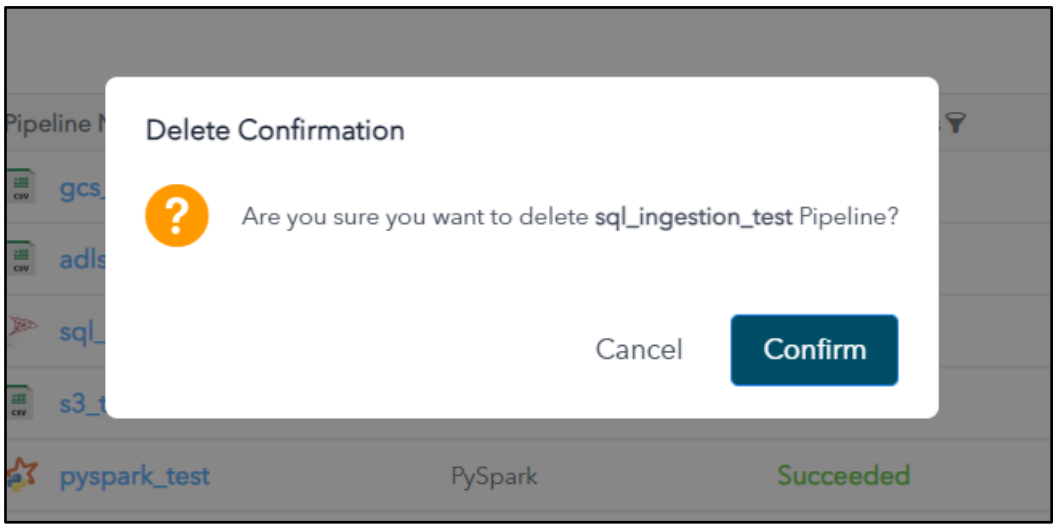


Figure 134: Delete Pipeline Confirmation Popup

5 Data Domains

Data domain is a group or combination of tables, directories, or files on which a user can perform profiling, indexing, and defining of entities. The available data domains are Postgres, MySQL, Oracle, SQL Server, Hive and SAS.

You can navigate to the 'Data domains' section by clicking the menu icon " ≡ " at top right of the screen. select 'Data domains'

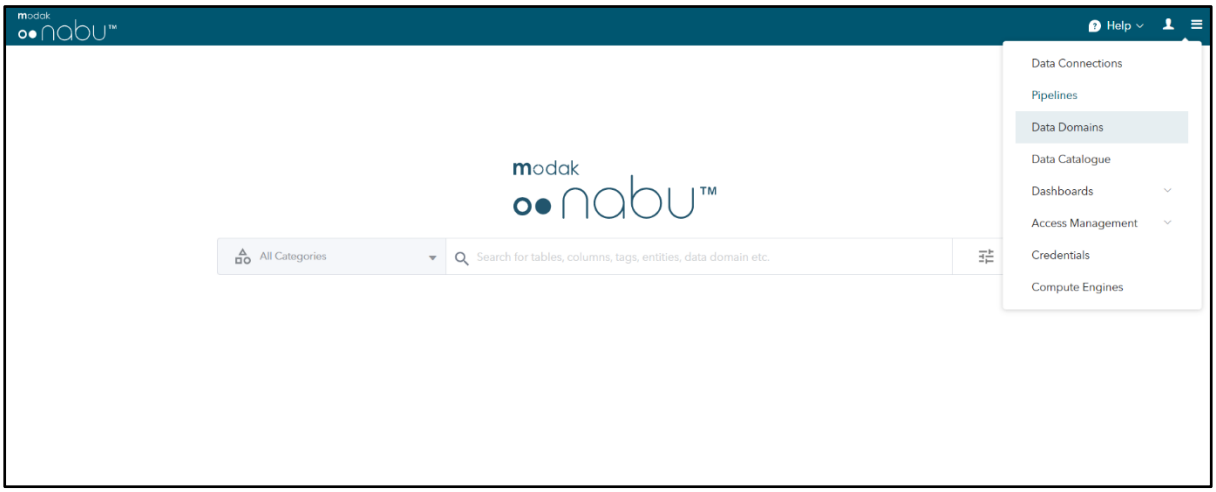


Figure 135: Data domains

Data domains landing page is as below.

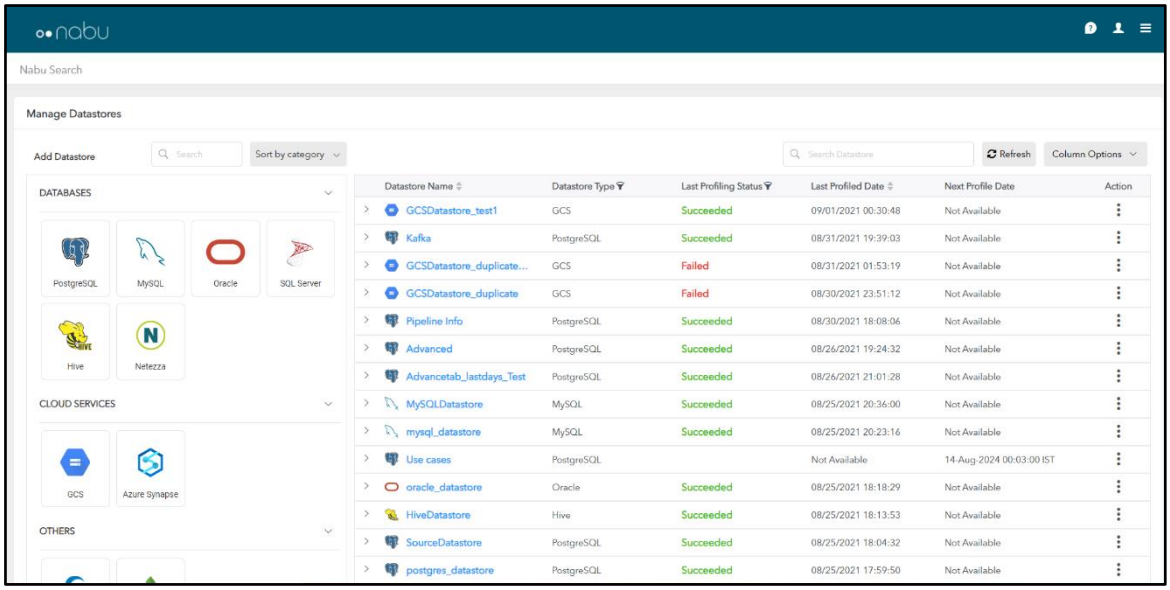


Figure 136: Data domains landing page

The following features are available to the users on the Data domains landing page.

In the left panel of the screen, it shows the icons for different data connection types. These include databases, cloud services, and others. This panel includes the following features.

Search for data domain type: This feature enables the user to search for the data connection type.

Sort by category/A to Z format: It helps the user to sort the connection type by category/A to Z.

On the right-hand side, the landing page shows a table which comprises the following list of columns.

Data domain Name: This column displays the list of all the existing data domains names. This column

is expandable and if user expands it shows a frequency detail for the data domains This column is sortable.

Data domain Type: This column shows the data connection type. The user can filter the results based on the connection type they want by using the filter option. Simply click on the filter icon, select the desired connection type from the menu, then click apply to view it and table results will get refreshed with the applied connection type.

Last Profiling Status: This column displays the outcomes of the data domains that have been profiled. They could be succeeded, running, or failed. Each state is indicated by a distinct colour and symbols, such as succeeded in green, running in blue, and failed in red, and the user can also check the needed status of the profiled data domains by using the filter option, simply click on the filter icon and select the required status from the menu and click on the apply button to view the filtered results.

Last Profiled Date: It shows the last profiled date and time for the data domain. This column is sortable.

Next profile Date: It shows the next profile date and time for profiling of the data domain. This column is sortable.

Action: The action column provides the user to select any option (edit, duplicate, schedule, or delete) from the drop down. Based on the option selected the respective action will be applied.

Refresh: This indicator assists the user in refreshing the table. When user clicks on the refresh button, the results are refreshed and updated.

Column options: This option enables the user to select the required columns to display on the table. Some columns are selected by default, while others are not. For example: data domain name, data domain type, last profiling status, last profiled date, next profiling date, and action are all selected by default. According to user preferences, the user can select/deselect the option from the menu. Which includes last indexed data, owner and tags

Last indexed data: it shows the last indexed date and time for the data domain. This column is sortable.

Owner: It shows the user id of the individual who created the data domain.

Tag: It shows the tags that are assosiated with the pipeline.

Search data domains: this feature helps users to search for the required data domains. simply by entering the name of data domain in the search box.

5.1.1 Creating Data domain

You can add a data domain by following below steps

Click on the preferred data domain, then you will be redirected to selected Add Data domain page where

1. You need to fill the required fields in the **Data domain info.** Following are the fields.

S.no	Field	Description
1	Data domain Name	Assigns a name to the current data domain
2	Add Tag	You can add a tag to categorize the current data domain
3	Email	Takes email address
4	Owner	Tells about the creator of this data domain
5	Description	A brief description can be added to understand why this data domain is created or used for.

6	Next button	Saves the values in current page and proceeds to next step

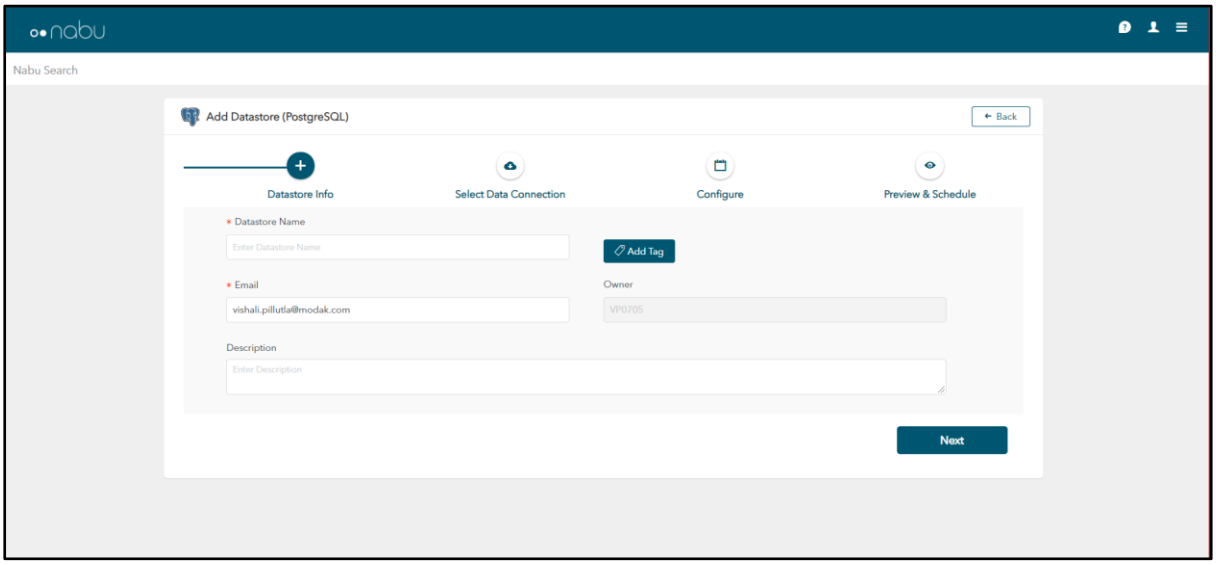


Figure 137: Data domain-Add Data domain info

Click Next to move to the next step.

2. Once you provide the basic information about the data domain, you will be asked to select a data connection.
3. Select a data connection from the available options in the list, else you can use the search box provided for easy search of the required data connection.

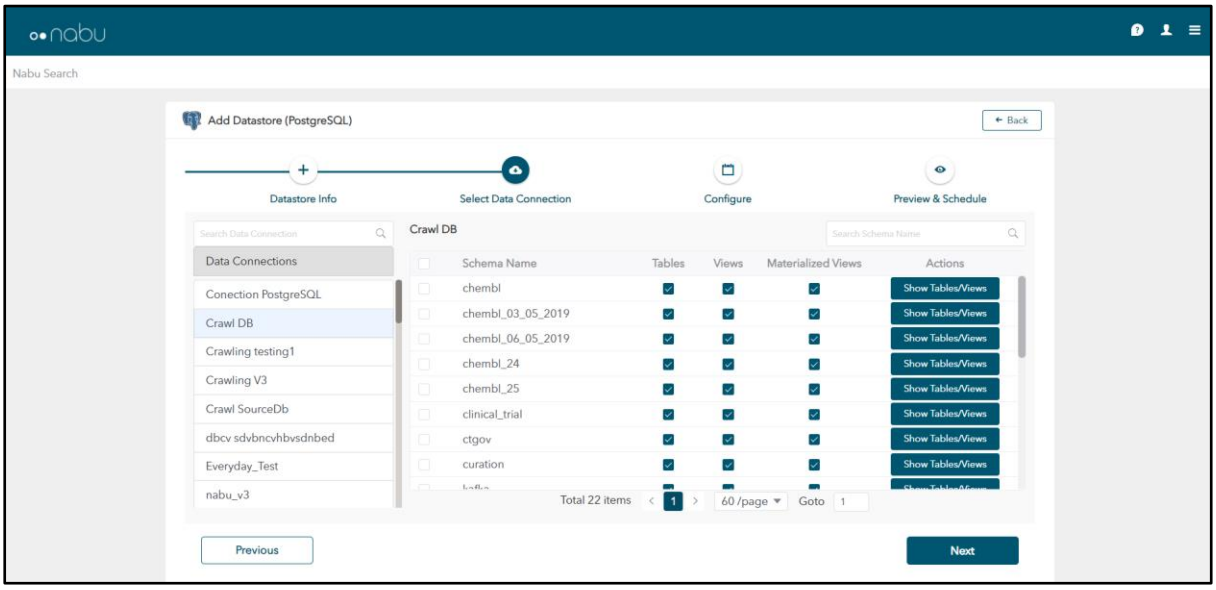


Figure 138: Data domain-Add Data domain- Select Data connection

Note: You can select multiple data connection, multiple tables, and views.

4. Once you select the required Data connection, schemas will be visible on where you can select multiple schemas.
5. For higher precision you can go to granular level and select a required table.

Note: You can select apply a filter to view data, we will see this in the upcoming topics.

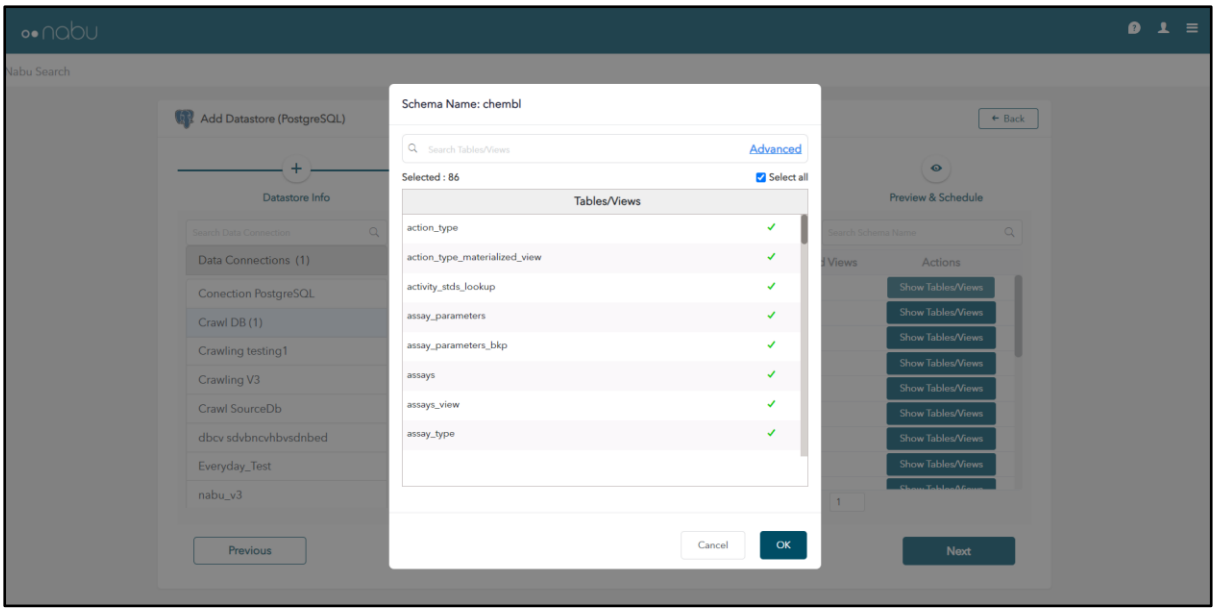


Figure 139:Data domain- Table popup

Click Previous to move to the previous step.

Click Next to move to the next step.

6. You will be redirected to Configure page where the user can enter Profiling, Indexing and Data domain refresh information.
    a. Profiling Information: The user can select compute engine, workflow engine which are the optional fields.
    b. Indexing Information: The user can select compute engine, workflow engine which are mandatory fields.
    c. Data domain refresh Information: The user can select compute engine which is mandatory field.

Click Previous to move to the previous step.

Click Next to move to the next step.

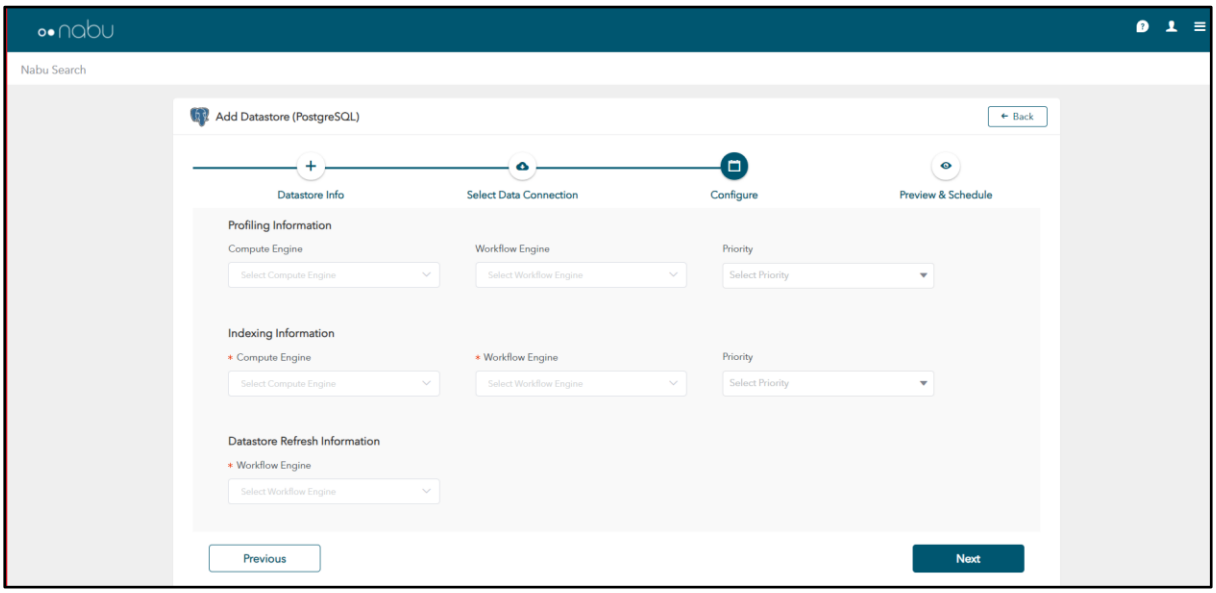


Figure 140: Data domain-Add Data domain-Configure

7. Preview & Schedule:

In this section, you can preview all the data entered in all the steps. You can schedule the profiling, indexing and data domain refresh as required by clicking on the respective Schedule button. More details on scheduling. For scheduling, please refer to section: 3.1.9.

Click Previous to move to the previous step.

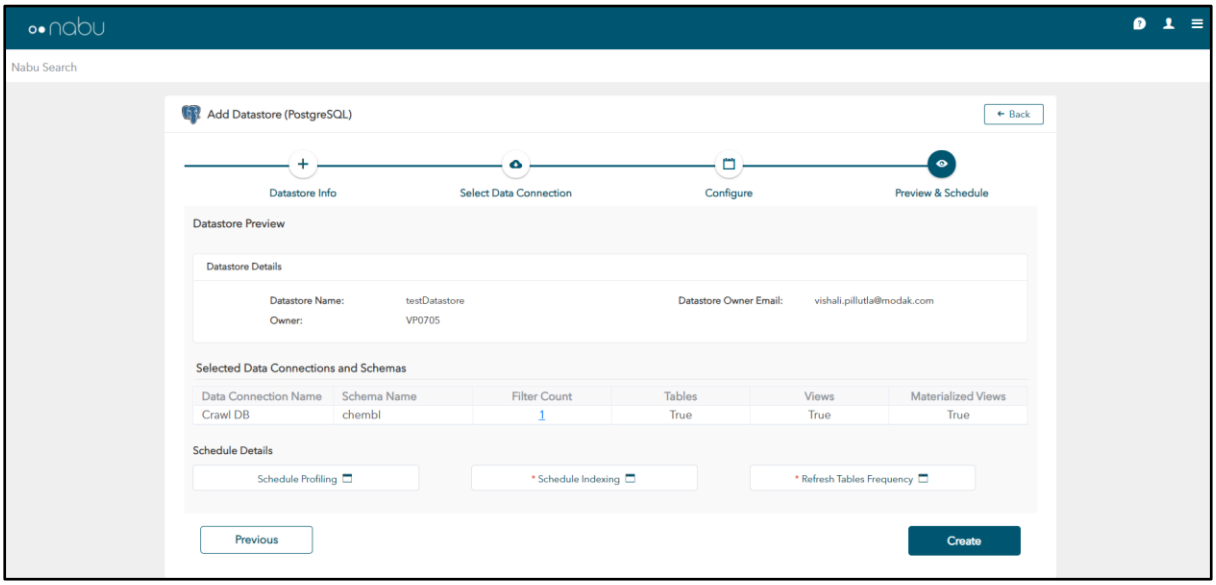


Figure 141: Data domain-Add Data domain-Preview & Schedule

8. Make any changes if required, click “Create” to create a Data domain.

5.1.2 Edit Data domain

The user can edit any of the data domain from the Data domain landing page as below.

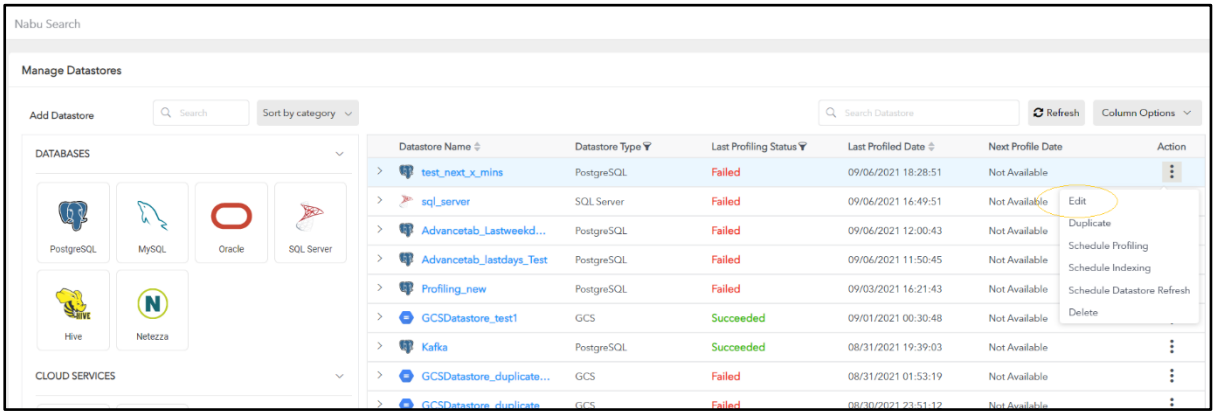


Figure 142: Edit Data domain

The user can click on the Action column and select the Edit option or click on the Data domain name.

On selecting the edit option, the user will be redirected to the respective data domain with all the previous details. The user can edit the data domain as required and click on the modify button.

All the details entered/modified should be valid, else the user will not be able to modify the data domain.

5.1.3 Duplicate Data domain

The user can duplicate any of the above data domain from the data domain landing page as below.

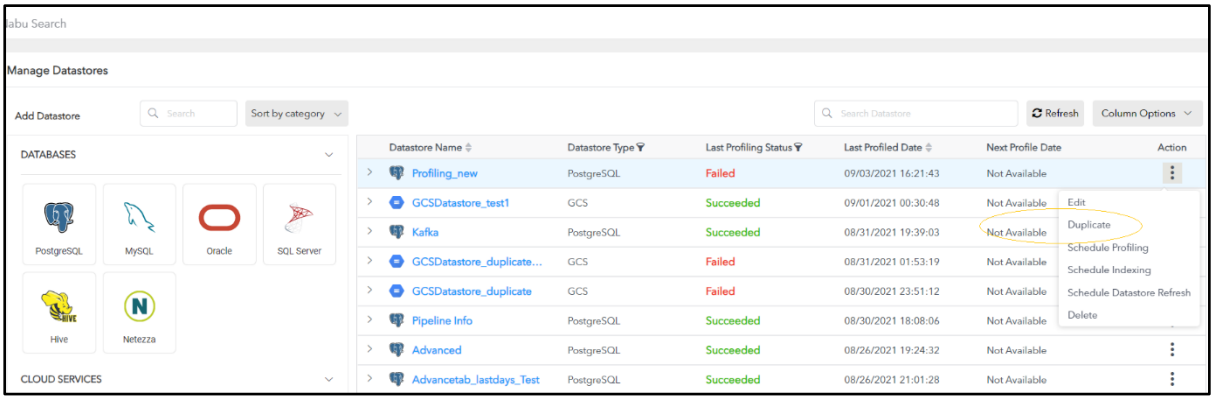


Figure 143: Duplicate Data domain

The user can click on the Action column and select the duplicate option.

On selecting the duplicate option, the duplicate or copy of the selected data domain is created and the user will be redirected to the selected data domain with all the details. The user can take action as per their preference.

All the details entered should be valid, else the user will not be able to create the data domain.

5.1.4 Schedule Data domain

The user can schedule profiling, indexing and data domain refresh for any of the data domains from the data domain landing page as below.

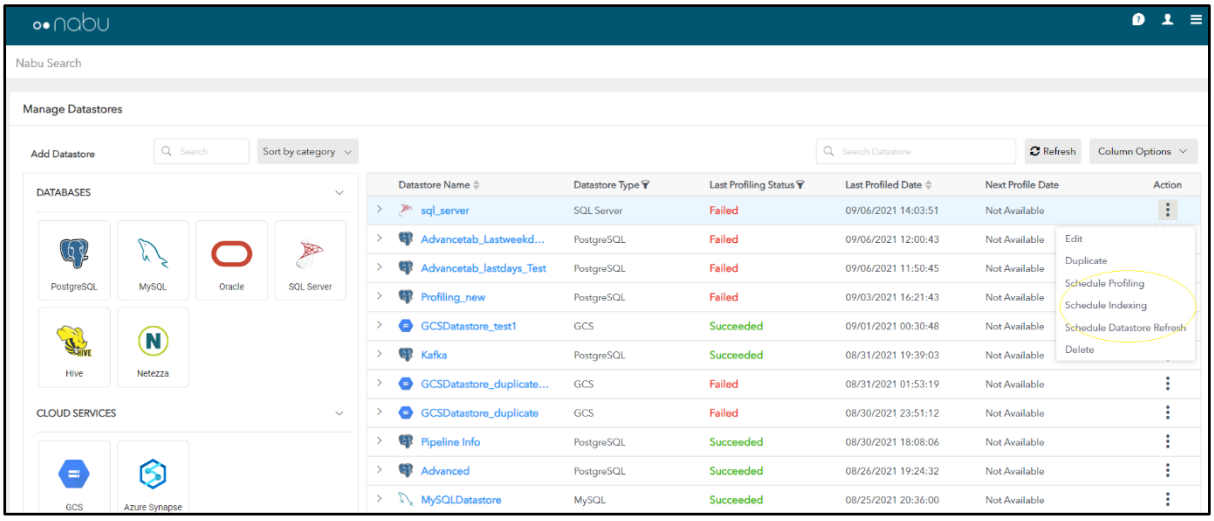


Figure 144: Schedule Data domain

On clicking schedule, the user will be prompted with a popup where the user can select the schedule frequency. [For scheduling, please refer to section: 3.1.9.]

5.1.5 Delete Data domain

The user can delete any of the data domain from the data domain landing page as below.

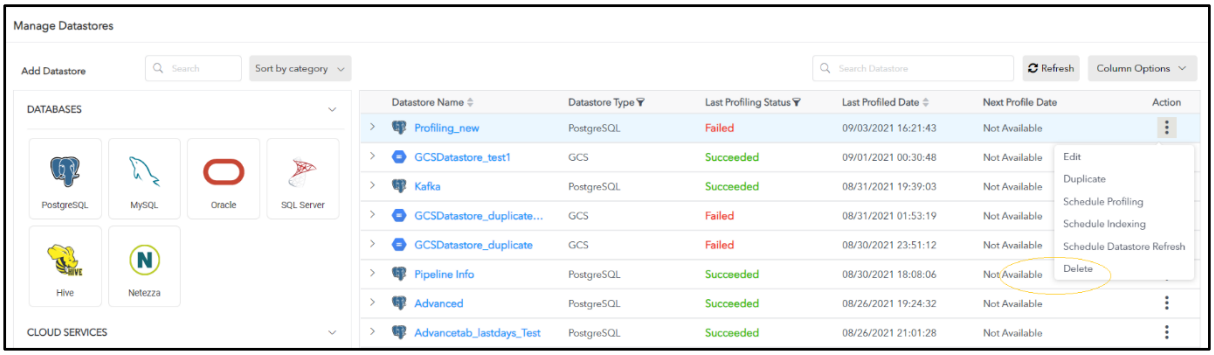


Figure 145: Delete Data domain

The user can click on the Action column and select the Delete option.

On selecting delete option, a popup showing the impacted data domains will be shown as below.

If the user clicks ok, the selected data domain will be deleted.

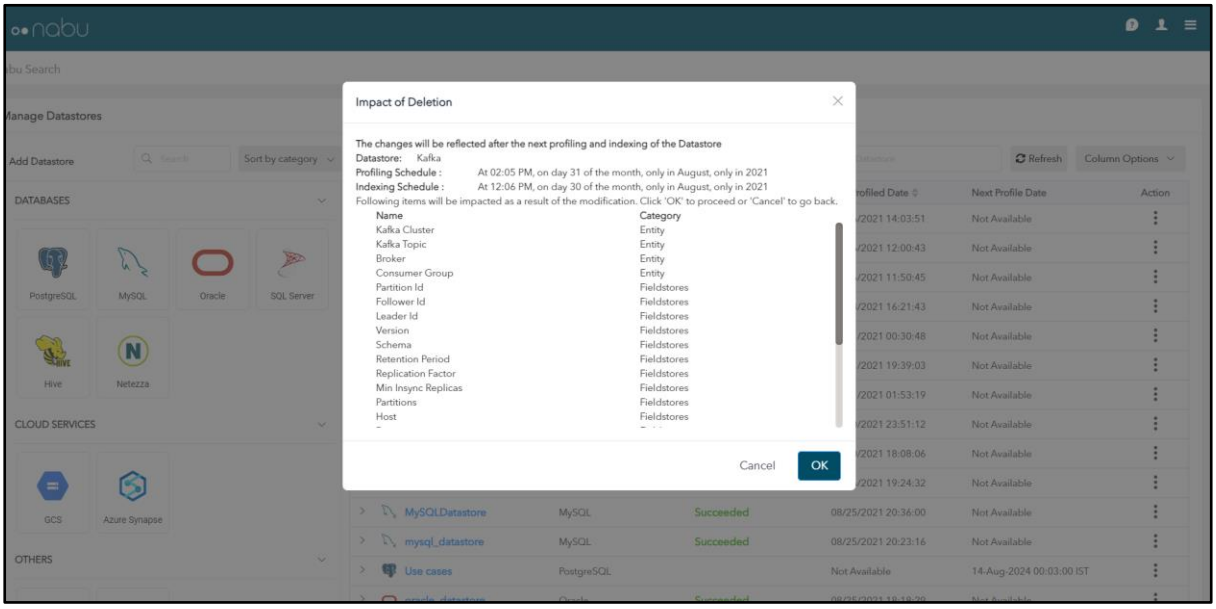


Figure 146: Impact of Deletion-Data domain

If the user clicks ok, the selected data domain will be deleted.

6 Data Catalogue

'Data Catalogue' section is used to create entries which enable quicker and more efficient discovery, exploration and understanding of data. The entries that can be defined in data catalogue are as under:

1 Facet – It is a group of columns which are similar. For e.g., a facet 'Age' could comprise a group of all columns that contain data for age.
2 Entity – It is a group of columns which are similar. Additionally, the users can search for values within a defined entity. For e.g., if an entity 'Study ID' is defined, then all columns which contain data for study ids are grouped together in the entity and specific values of 'Study Id' can be searched (e.g., Study Id of ABC10101) from Modak Nabu Search.
3 Synonym – It is an alias for an entity. For e.g., a 'Study Name' could serve as an alias for a 'Study Id'.
4 Filter – A filter is any attribute related to an entity on which the search results of the entity can be filtered.
5 Fieldstore – A fieldstore is an attribute which provides additional information/context for an entity. The fieldstore is displayed in the search results of the entity.

To create an entry in the data catalogue, navigate to the 'Data Catalogue' section from the menu, and click on 'Add' button at the top right to select the type of entry to be created.

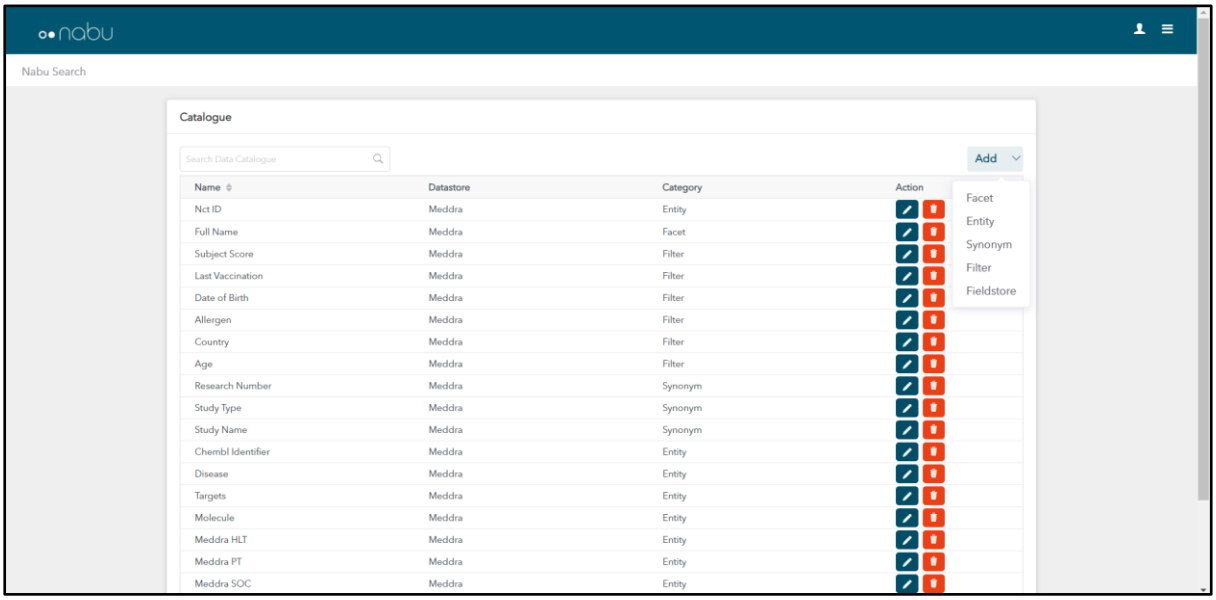


Figure 147: Catalogue Dashboard

6.1 Creating Facet

1. Select 'Facet' option from the 'Add' dropdown on the Catalogue dashboard screen. The following screen is shown.

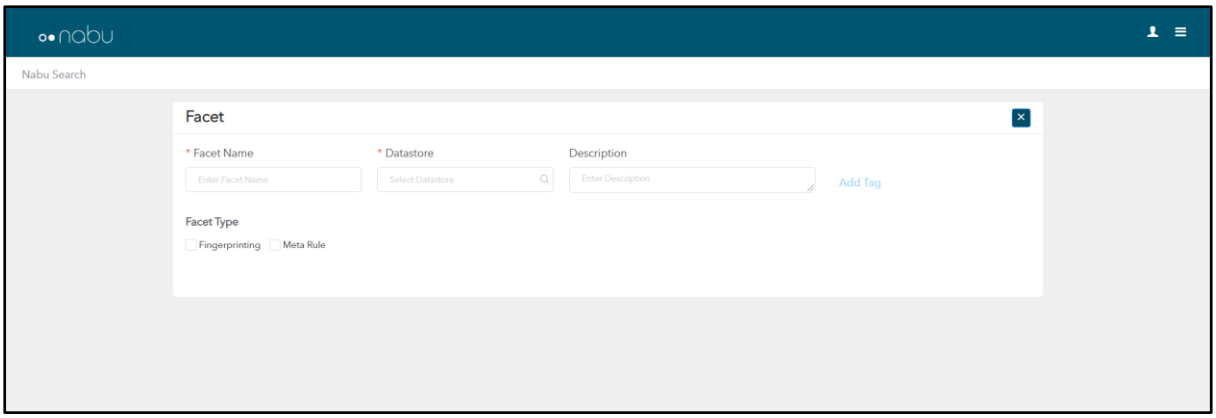


Figure 148: Add Facet

2. Enter the details for the fields as indicated below

S.no	Field	Description
1	Facet name	Name by which the facet will be identified
2	Data domain	Refers to the data domain on which the facet will be defined. Similar columns in a data domain are considered for defining a facet
3	Description	Optional description on the facet being created
4	Add Tag	An optional tag can be added while creating the facet. The tag can be used to search facets.
5	Facet Type	Select the method for defining the Facet Type. Fingerprinting or Meta Rule. Facets can be created using either Fingerprinting or Meta rule or both

3. The data domain selected should be profiled before creating facets.

6.1.1 Creating Facet with Fingerprinting

Fingerprinting identifies similar columns to a seed column which is given as a reference. It generates a score between 0 and 1 to denote the extent of match of a column to the reference column.

To define a facet using fingerprinting, follow the steps as below:

1 Select ‘Fingerprinting’ as the ‘Facet Type, as shown in the following

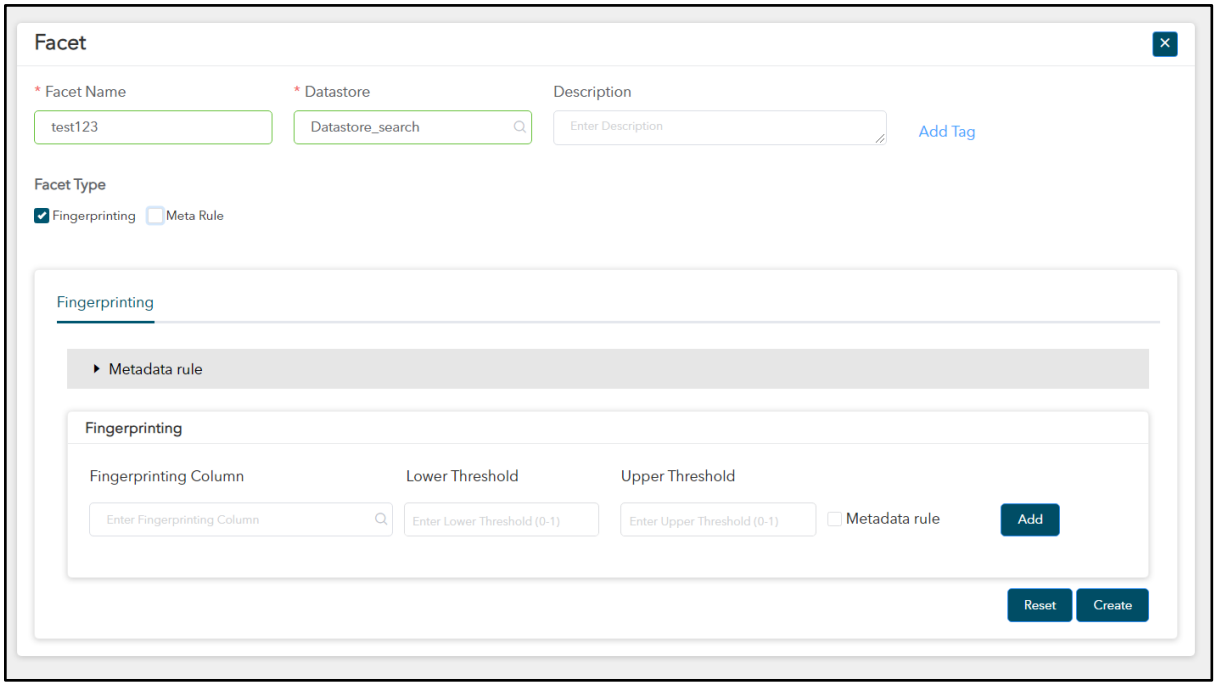


Figure 149: Facet Fingerprinting

2 For ‘Fingerprinting Column’ provide the name of the seed column which will be used as a reference column to look for other columns with similar data.

3 ‘Lower Threshold’ is a value between 0 and 1, which sets the lower limit for determining a match. If the match score calculated by data fingerprinting technique for a column is less than ‘Lower Threshold’, then it is not considered as part of the facet.

4 ‘Upper Threshold’ is a value between 0 and 1, which sets the upper bound for determining a match. If the match score calculated by data fingerprinting technique for a column is more than ‘Upper Threshold’, then it considered as part of the facet.

5 Optional metadata rules can be provided to help in narrowing the list of columns selected as part of facets. An overall metadata rule restricts the matches to those columns that satisfy the rule. For e.g., an overall metadata rule could be ‘meta.column_name like ‘age%’. This rule would ensure that only columns whose name starts with ‘age’ are shown as possible matches.

6 Metadata rule can also be provided for each ‘Fingerprinting Column’.

For e.g., in the figure below, for the fingerprinting column, ‘name(ctgov.facilities)’, a metadata rule ‘meta.table_name like ‘ctgov%’ is entered. This ensures that the list of matches for fingerprinting column ‘name(ctgov.facilities)’, include only those columns which are present in tables whose name starts with ‘ctgov’.

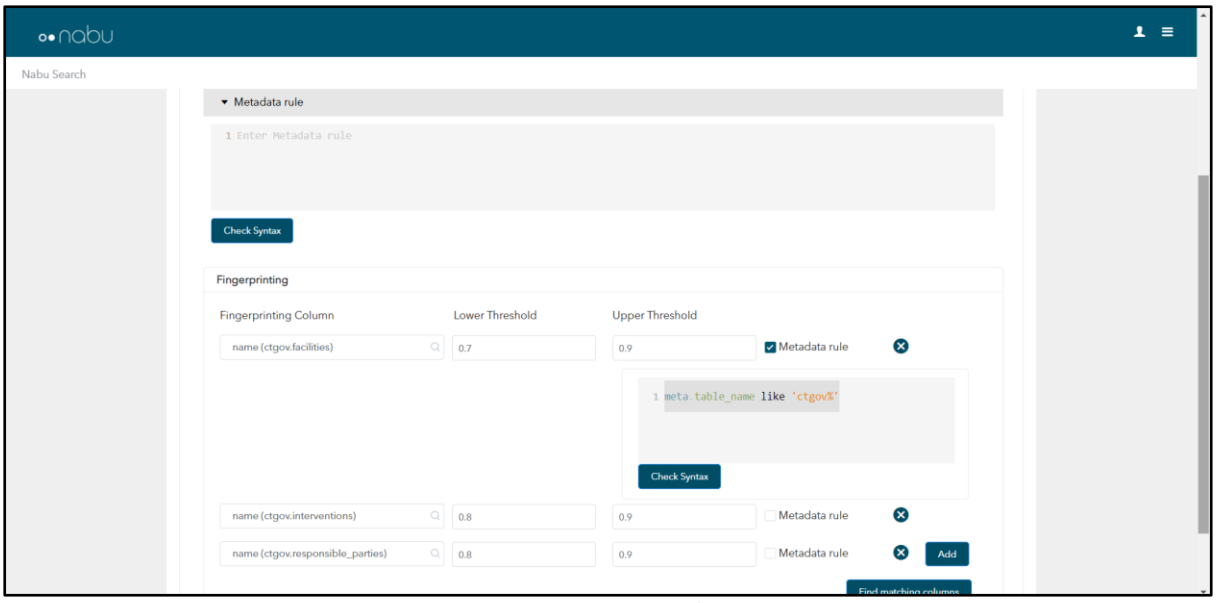


Figure 150: Meta Rule

7 More than one column can be given as 'Fingerprinting Column'. To add additional columns as fingerprinting columns, click on 'Add' button next to each fingerprinting rule.

8 Once all the fingerprinting columns are defined, click on 'Create' button.

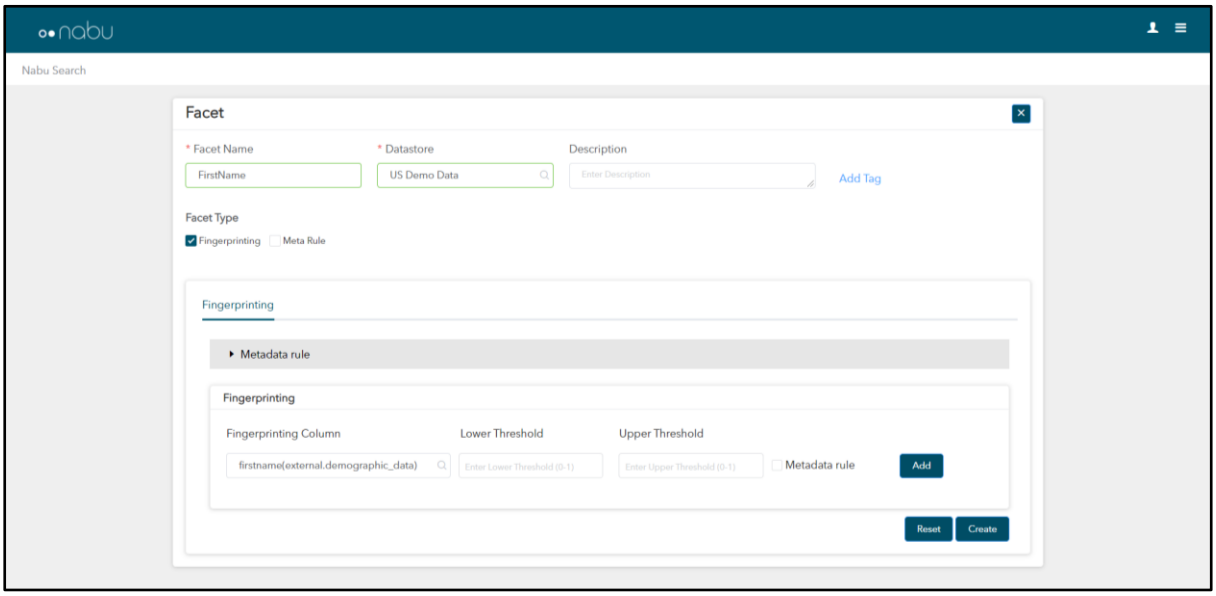


Figure 151: Create Facet

9 Schedule Profiling for the data domain from the 'Data domain' section. When the data domain is profiled, the fingerprinting for columns is also done. The fingerprinting will generate similarity score for each column in the data domain against the reference column(s) given in the Create Facet section.

10 Once profiling is completed, go to 'Data Catalogue' Section, and click on edit against the facet name created.

11 Click on 'Find Matching Columns' at the bottom of the screen.

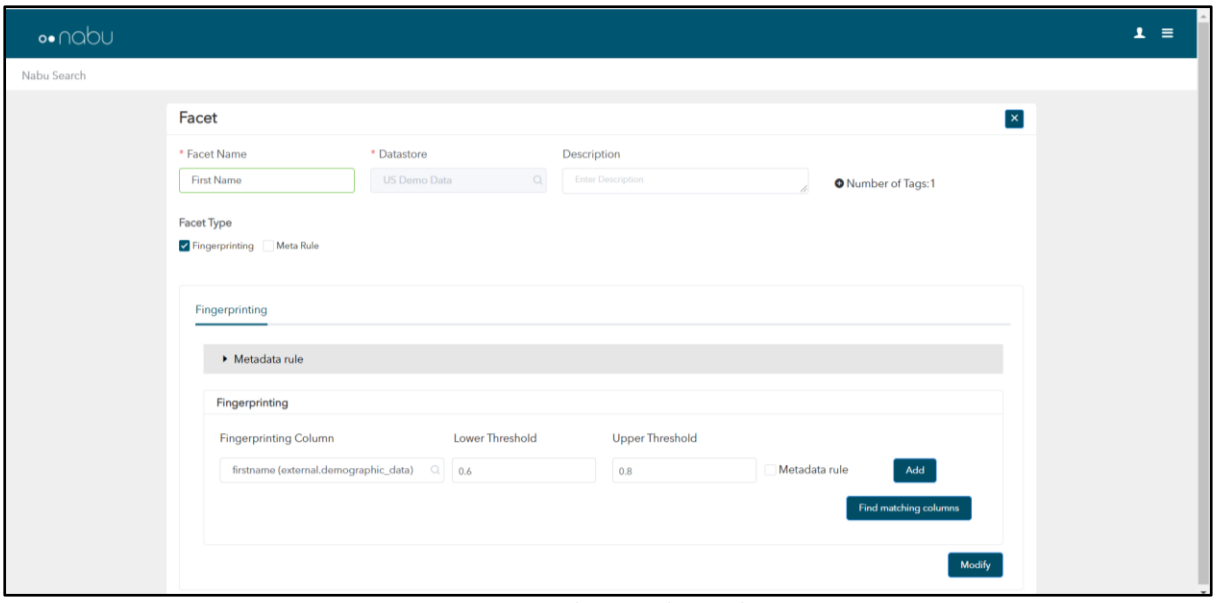


Figure 152: Finding Matching Column

12 Three tabs will be shown ‘Suggested’, ‘Accepted’ and ‘Rejected’. The ‘Suggested’ tab will have list of columns whose ‘Match score’ lies between ‘Lower Threshold’ and ‘Upper Threshold’. The user can choose to accept or reject the matched columns shown in ‘Suggested’ tab. The ‘Accepted’ tab shows columns whose ‘Match Score’ is higher than the ‘Upper Threshold’. Any columns that are rejected from ‘Suggested’ or ‘Accepted’ tabs can be seen under ‘Rejected’ tab.

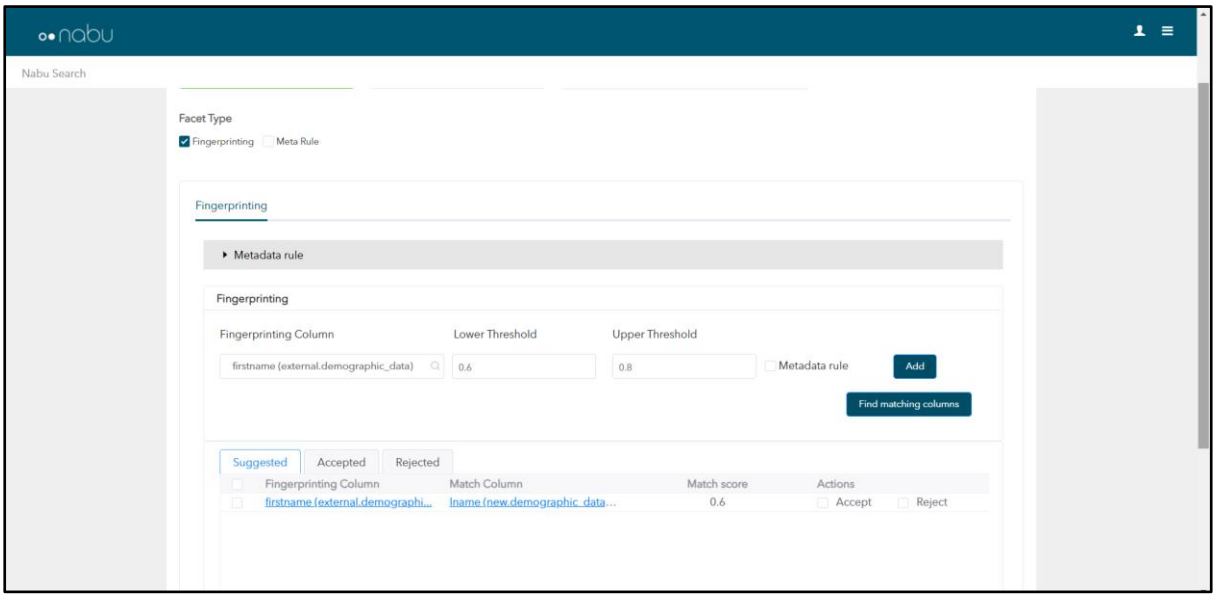


Figure 153 : Tabs to Find Matching Column

6.1.2 Creating Facet with Meta Rule

‘Meta Rule’ can be used to create a facet on its own or can be used along with ‘Fingerprinting’ option. ‘Meta Rule’ provides options to group columns based on column names or table names.

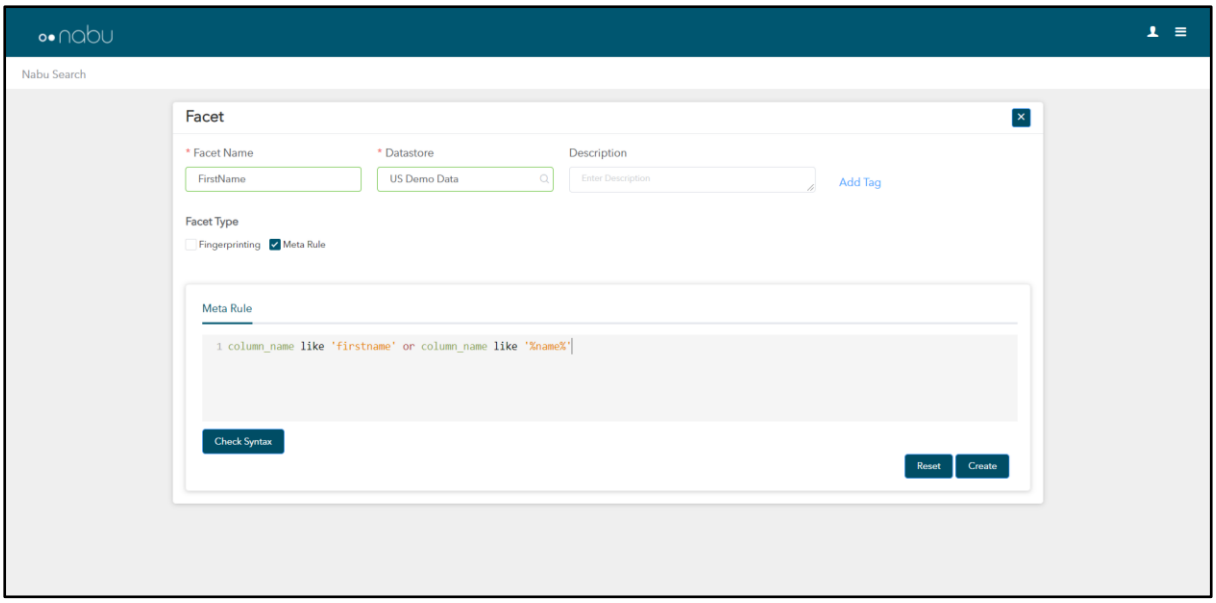


Figure 154: Facet Meta Rule

1. Enter the meta rule to identify similar columns. The syntax uses the following rules.
    a. 'column_name' or 'table_name' to compare column names or table names respectively.
    b. 'like', 'not like' to compare against a string.
    c. The string for comparison is given in single quotation marks. Use % for wildcard characters
    d. 'and' and 'or' are used to combine multiple conditions.
2. The syntax can be validated using the 'Check Syntax' button.
3. Click 'Create' button, to create facet using the 'Meta Rule' provided.

6.2 Creating Entity

36. Select 'Entity' option from the 'Add' dropdown on the Catalogue dashboard screen. The following screen is shown.

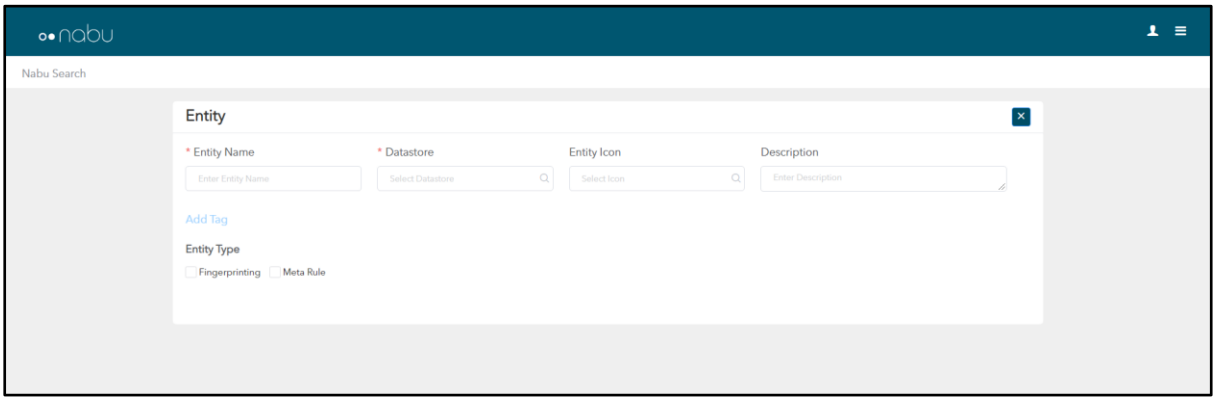


Figure 155: Add Entity

37. Enter the details for the required columns like

S.no	Field	Description
1	Entity name	Identifier for the entity being created.
2	Data domain	Refers to the data domain on which the entity will be created. Tables included in the data domain, will

		be considered for creation of entity.
3	Entity Icon	Optional. Select an icon to denote the entity. This icon will be used in the visualization of the knowledge graph for the entity.
4	Description	Optional. Description for the entity being created.
5	Add Tag	Optional. Tags are key value pairs that provide business context for an entity.

38. To create an entity, the data domain should have been profiled.

6.2.1 Creating Entity using Fingerprinting

Fingerprinting identifies similar columns to a seed column which is given as a reference. It generates a score between 0 and 1 to denote the extent of match of a column to the reference column.

To define a facet using fingerprinting, follow the steps as below:

1. Select 'Fingerprinting' as the 'Facet Type, as shown in the following

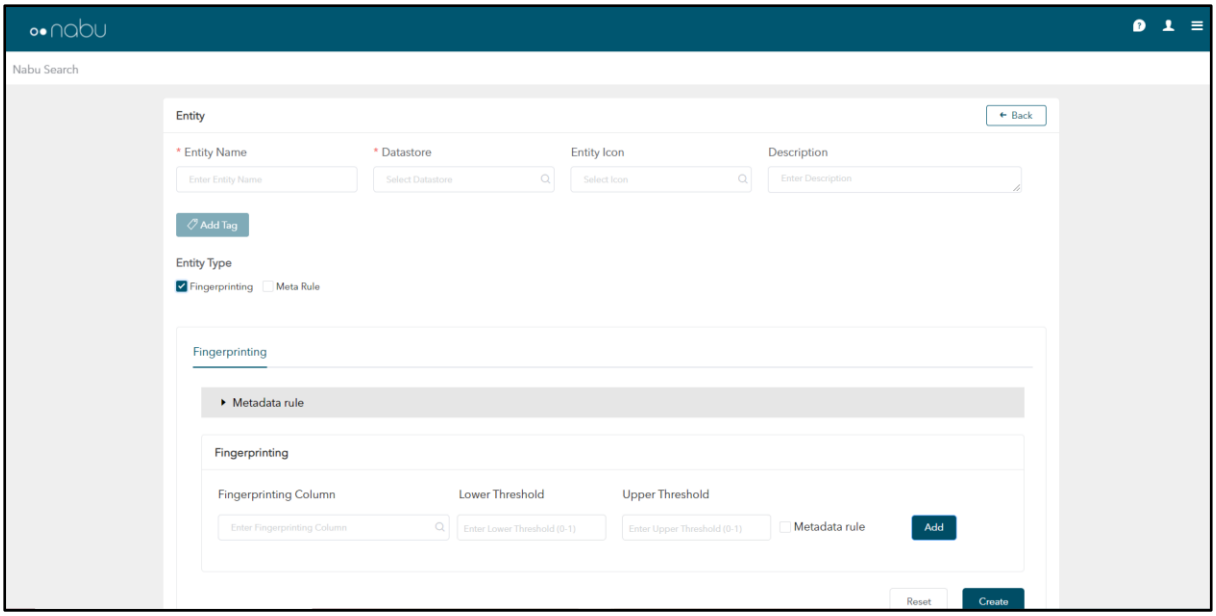


Figure 156: Entity-Fingerprinting

2. For 'Fingerprinting Column' provide the name of the seed column which will be used as a reference column to look for other columns with similar data.
3. 'Lower Threshold' is a value between 0 and 1, which sets the lower limit for determining a match. If the match score calculated by data fingerprinting technique for a column is less than 'Lower Threshold', then it is not considered as part of the facet.
4. 'Upper Threshold' is a value between 0 and 1, which sets the upper bound for determining a match. If the match score calculated by data fingerprinting technique for a column is more than 'Upper Threshold', then it considered as part of the facet.
5. Optional metadata rules can be provided to help in narrowing the list of columns selected as part of facets. An overall metadata rule restricts the matches to those columns that satisfy the rule. For e.g., an overall metadata rule could be 'meta.column_name like 'age%'. This rule would ensure that only columns whose name starts with 'age' are shown as possible matches.
6. Metadata rule can also be provided for each 'Fingerprinting Column'.

   For e.g., in the figure below, for the fingerprinting column, 'name(ctgov.facilities)', a metadata rule 'meta.table_name like 'ctgov%' is entered. This ensures that the list of matches for fingerprinting column 'name(ctgov.facilities)', include only those columns which are present in

tables whose name starts with ‘ctgov’.

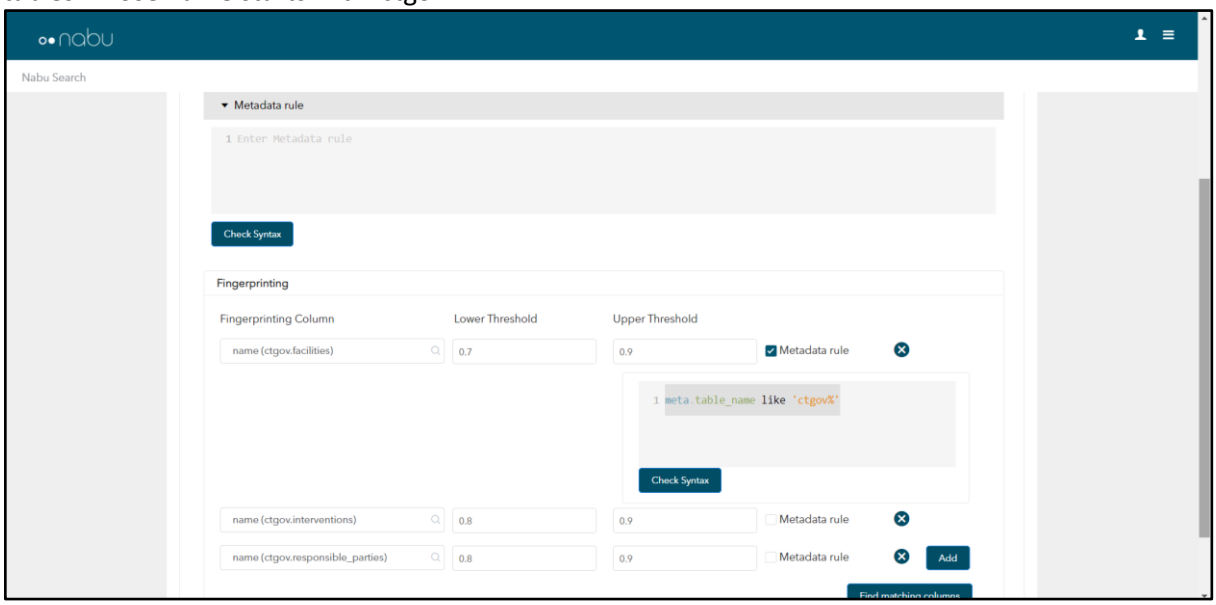


Figure 157: Entity-Metadata Rule

7. More than one column can be given as ‘Fingerprinting Column’. To add additional columns as fingerprinting columns, click on ‘Add’ button next to each fingerprinting rule.
8. Once all the fingerprinting columns are define, click on ‘Create’ button.

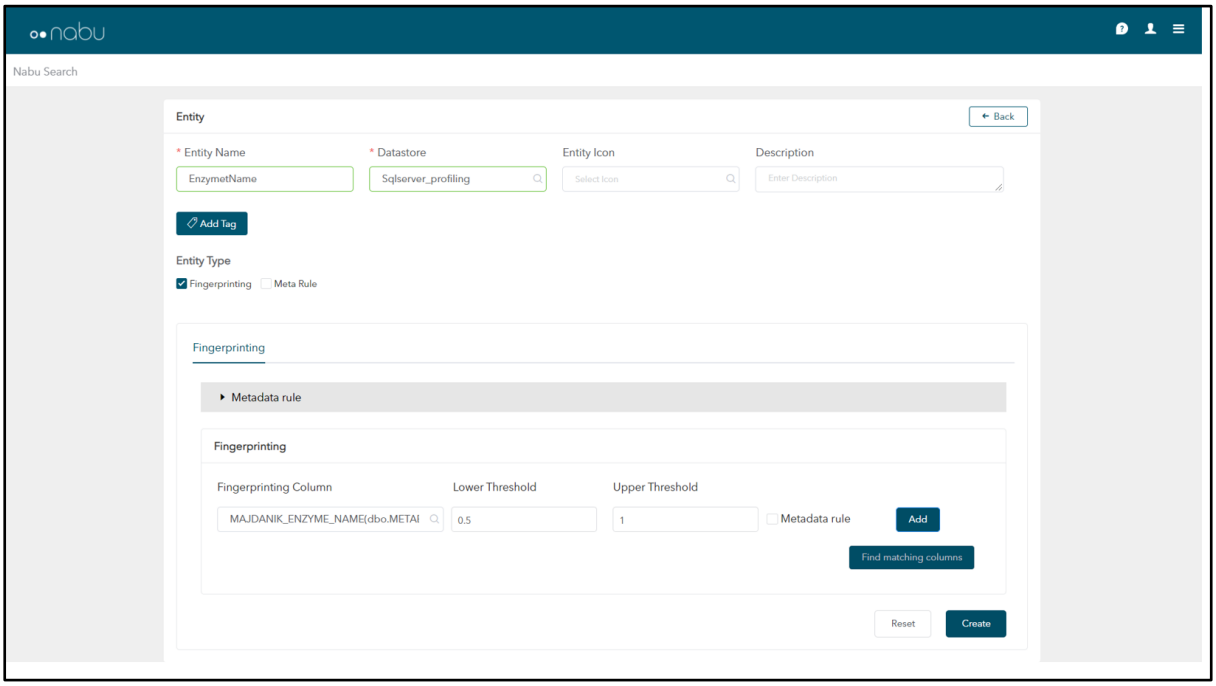


Figure 158: Entity Creation

9. Schedule Profiling for the data domain from the ‘Data domain’ section. When the data domain is profiled, the fingerprinting for columns is also done. The fingerprinting will generate similarity score for each column in the data domain against the reference column(s) given in the Create Facet section.
10. Once profiling is completed, go to ‘Data Catalogue’ Section, and click on edit against the facet name created.
11. Click on ‘Find Matching Columns’ at the bottom of the screen.

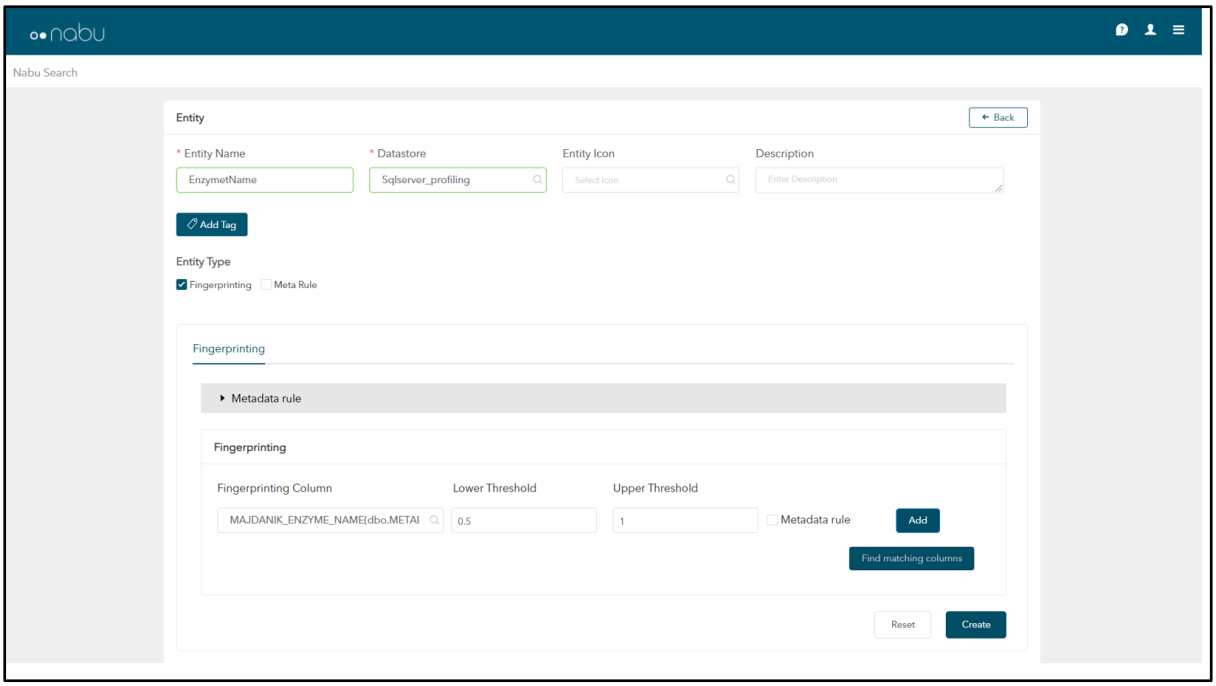


Figure 159: Find Matching Columns

12. Three tabs will be shown ‘Suggested’, ‘Accepted’ and ‘Rejected’. The ‘Suggested’ tab will have list of columns whose ‘Match score’ lies between ‘Lower Threshold’ and ‘Upper Threshold’. The user can choose to accept or reject the matched columns shown in ‘Suggested’ tab. The ‘Accepted’ tab shows columns whose ‘Match Score’ is higher than the ‘Upper Threshold’. Any columns that are rejected from ‘Suggested’ or ‘Accepted’ tabs can be seen under ‘Rejected’ tab.

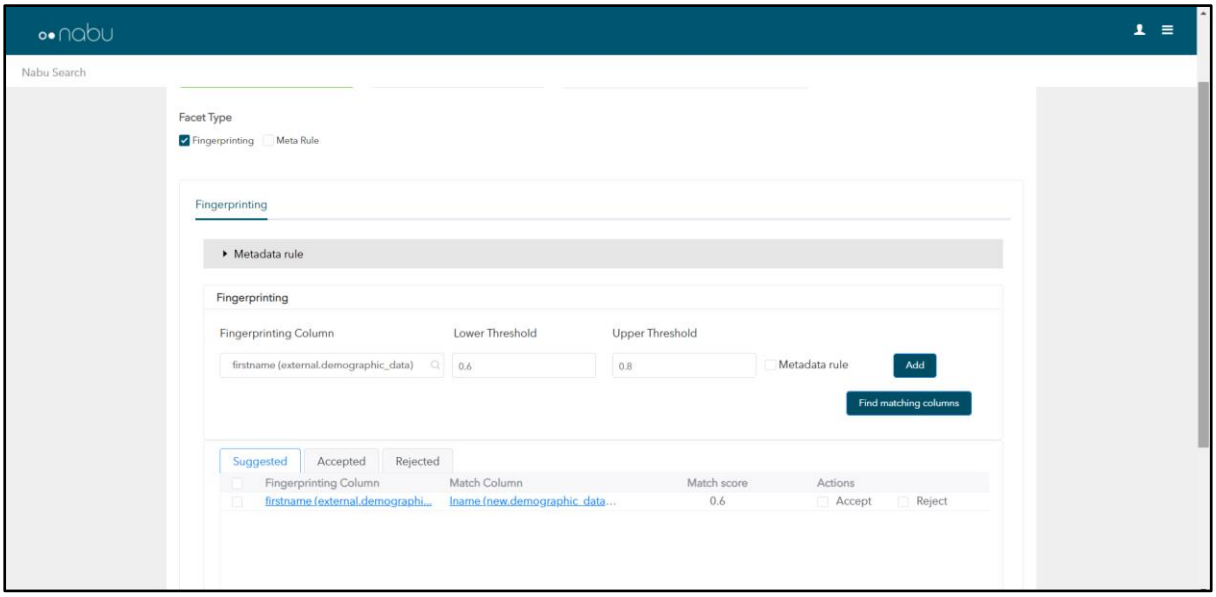


Figure 160: Find Matching Column Value

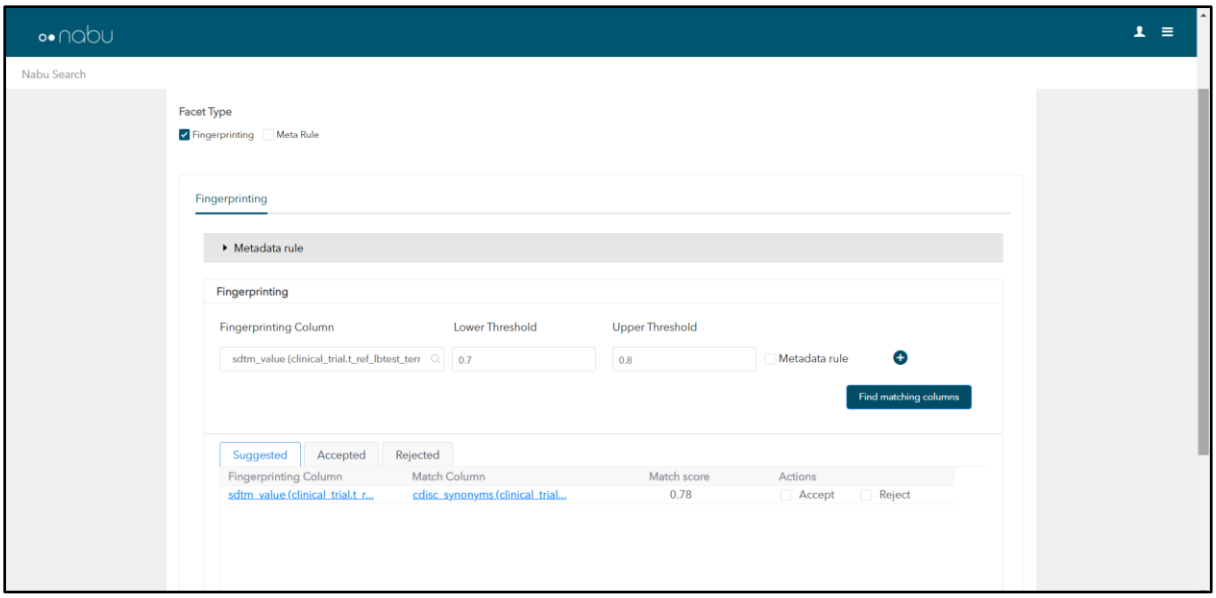


Figure 161: Suggested Tab for Matching Columns

6.2.2 Creating Facet with Meta Rule

'Meta Rule' can be used to create a facet on its own or can be used along with 'Fingerprinting' option. 'Meta Rule' provides options to group columns based on column names or table names.

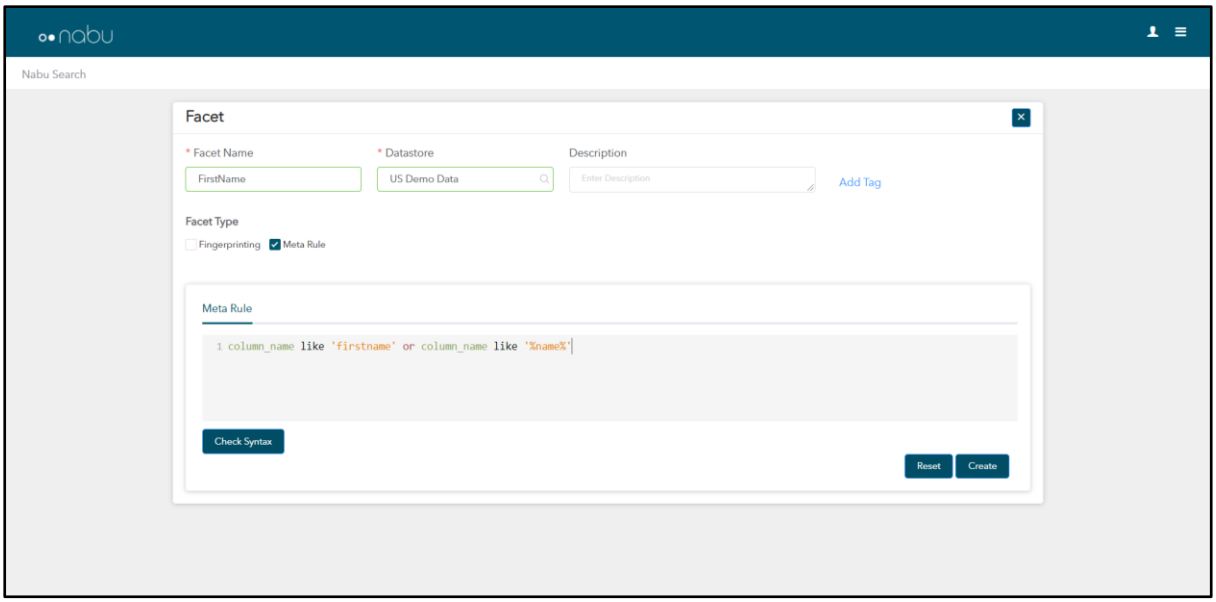


Figure 162: Meta Rule

1. Enter the meta rule to identify similar columns. The syntax uses the following rules.
    a. column_name' or 'table_name' to compare column names or table names respectively.
    b. 'like', 'not like' to compare against a string
    c. The string for comparison is given in single quotation marks. Use % for wildcard characters.
    d. 'and' and 'or' are used to combine multiple conditions.
2. The syntax can be validated using the 'Check Syntax' button.
3. Click 'Create' button, to create facet using the 'Meta Rule' provided.

6.3 Create Synonym, Filter and Fieldstore

The process for creating a Synonym, Filter and Fieldstore is same.

A “Synonym” acts as an alias of an entity, where a user can search the required value using synonym also.

A “Filter” is created on entities as it helps in narrowing down of search results for an entity value (E.g., you can apply a filter on age column where you can filter the results within a specific range)

A “Fieldstore” is an attribute of an entity which is displayed in search results of an entity

Provide the field details like:

S.no	Field	Description
1	Name	A reference name for either Synonym, Filter or Fieldstore
2	Data domain	The data domain where you want it to be created
3	Entity	Name of the related entity
4	Description	Brief explanation about the category being created.
5	Lookup Table	Table to select the required column for defining synonym/filter/fieldstore

Once all the fields are filled click on “Create” to save the category.

7 Dashboards

7.1 Monitoring Dashboard

The functionality of the monitoring dashboard is to provide a consistent interface for monitoring scheduled pipelines.

For any scheduled pipelines, the monitoring dashboard provides the following details to the user on the landing page.

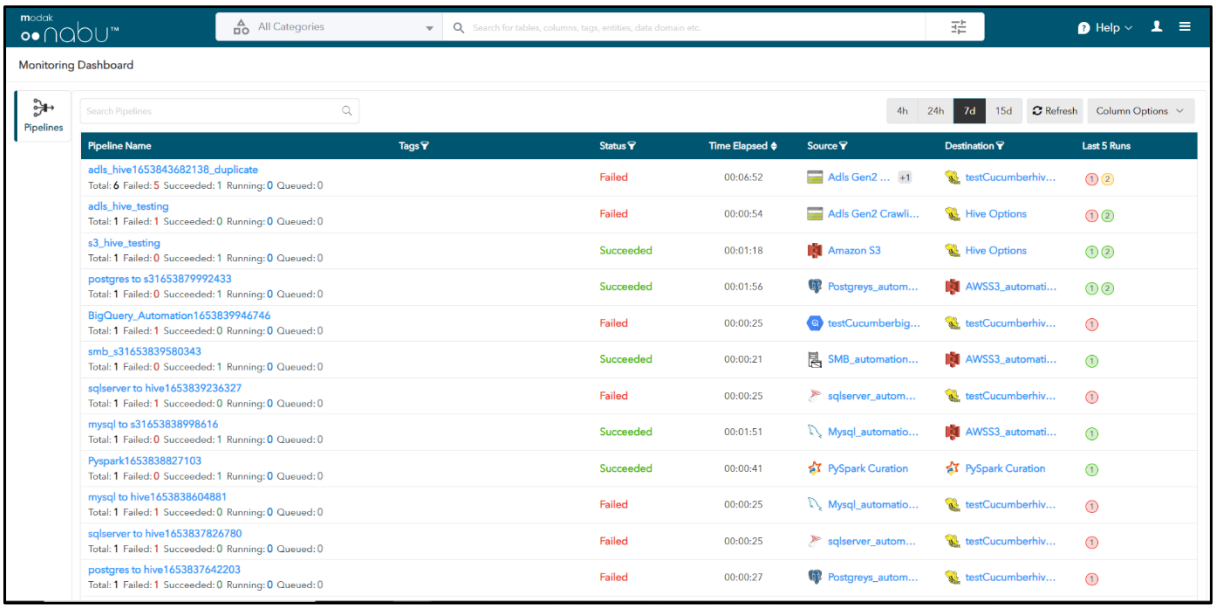


Figure 163: Monitoring Dashboard page

Pipeline name: This column shows the name of the pipeline, as well as the object details of the pipeline that are scheduled, failed, succeeded, running, and queued. Each result is represented by a distinct colour. Failed objects count is shown in red, succeeded objects count in green, running objects count in orange, and queued objects in black.

Search pipelines: This option aids the user to search for the specific pipeline. For the desired results, the user must enter the name of the pipeline in the search box.

Tags: This column displays tags linked to the pipelines. It also gives filter choices, allowing users to apply the tag to the pipeline and filter the results according to the applied tags.

Status: This column displays the outcomes of the pipelines that have been scheduled. They could be waiting, successful, running, ignored, or failed. Each state is indicated by a distinct colour and symbols, such as waiting in blue, successful in green, running in blue, ignored in grey, and failed in red, and the user can also check the needed status of the scheduled pipeline by using the filter option, simply click on the filter icon and select the required status from the menu and click on the apply button to view the filtered results.

Time Elapsed: Regardless of the outcome, this column provides the exact time taken by the scheduled pipeline to finish the ingestion process.

Source: This column gives the source details of the ingestion pipelines. The user can filter the results based on the sources they want by using the filter option. Simply click on the filter icon, select the desired source from the menu, then click apply to view it. The source details are clickable and when the user clicks, it displays the data connection name of that particular source, the path/schema based on the source type, the number of files/tables/views of the specific source type, as well as the estimated size of the file or table.

For file

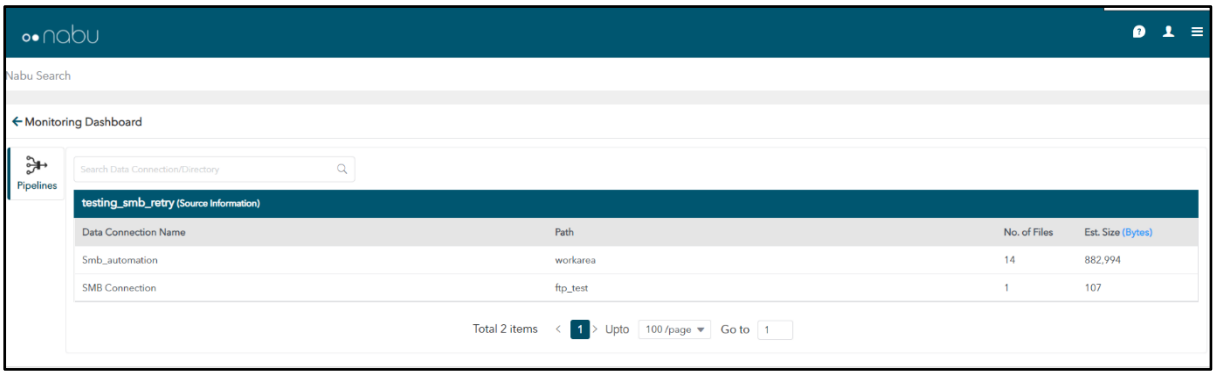


Figure 164: Monitoring Dashboards-Source Information for File

For schema

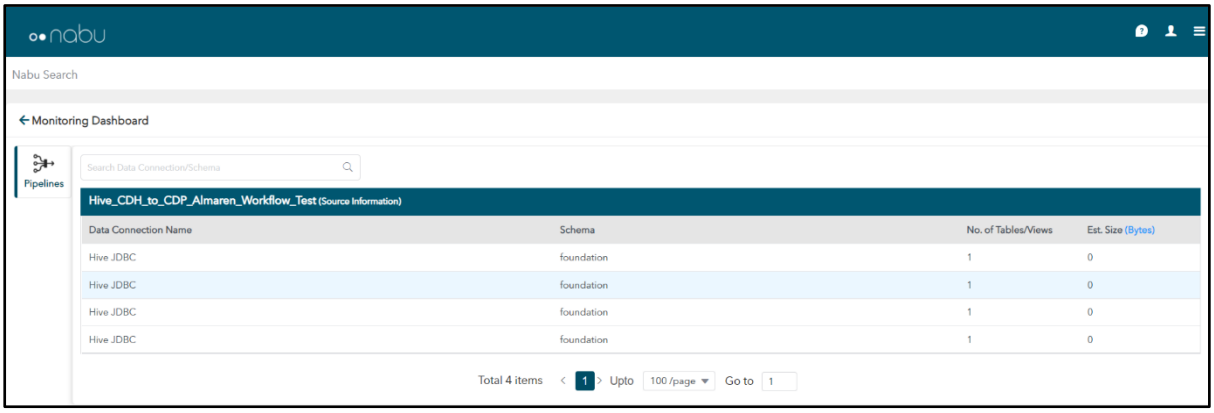


Figure 165: Monitoring Dashboards-Source Information for Schema

Use arrow which is on the top left corner of the window to go back to monitoring dashboard main page.

Destination: This column gives the destination details of the ingestion pipelines. The user can filter the results based on the destination they want by using the filter option. Simply click on the filter icon, select the desired destination from the menu, then click apply to view it. The destination details are clickable and when the user clicks, it displays the data connection name of that destination, the directory/schema name based on the destination type in the first column, and the destination path and file name if there is any file associated with that destination. If there is a file and path associated with the destination objects, both are displayed.

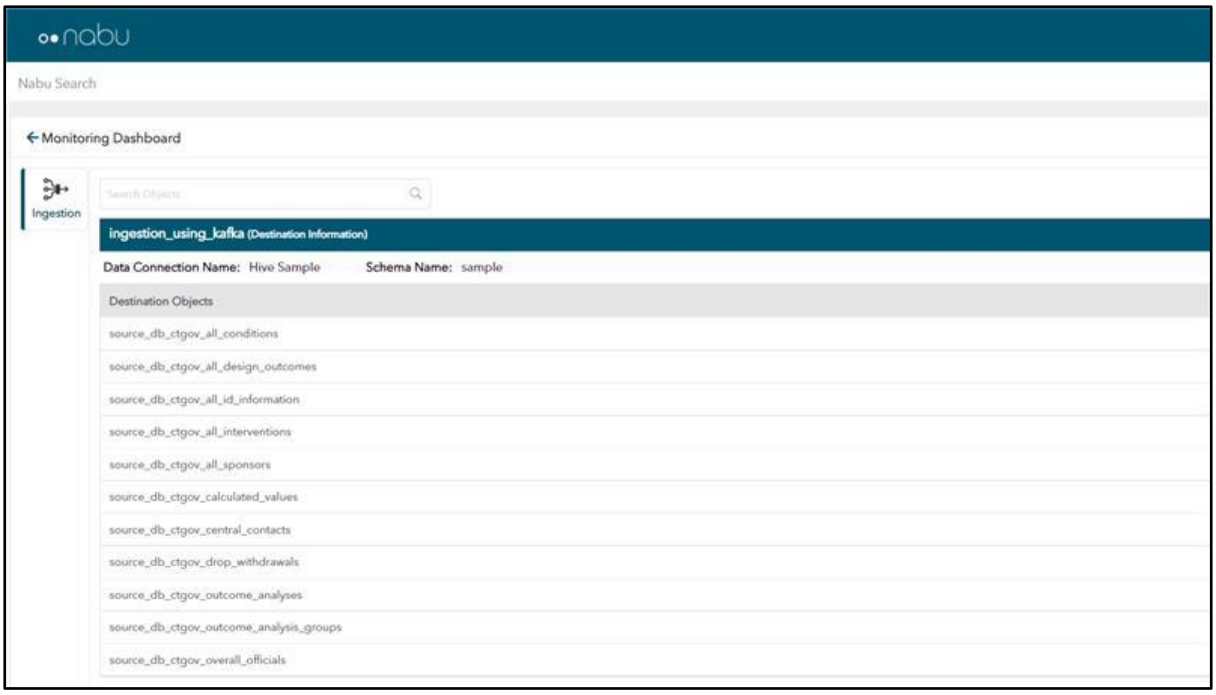


Figure 166: Monitoring Dashboards-Destination Information

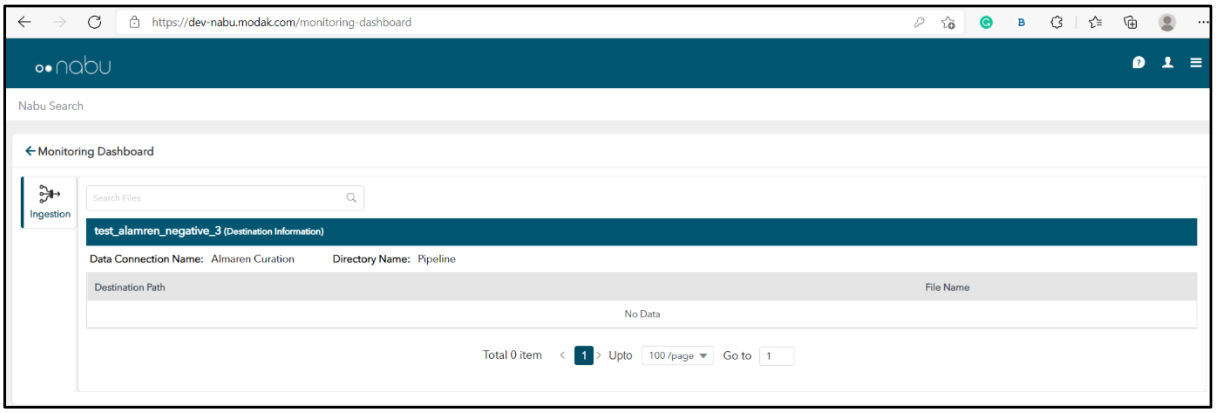


Figure 167: Monitoring Dashboards-Destination Information for File

Use arrow which is on the top left corner of the window to go back to monitoring dashboard main page.

Last 5 Runs: As named last 5 runs, it offers information of the last five scheduled pipeline runs. The last 5 run details are represented by numbers like (1,2,3,4, and 5). Those numbers may indicate any one or may be a combination of succeeded, failed, running, and queued status. Each is represented by a separate fixed colour, such as green for success, red for failure, orange for running, and black for queued.

Each run is clickable, and user can view more information about a specific run, they can do so by clicking on the run number.

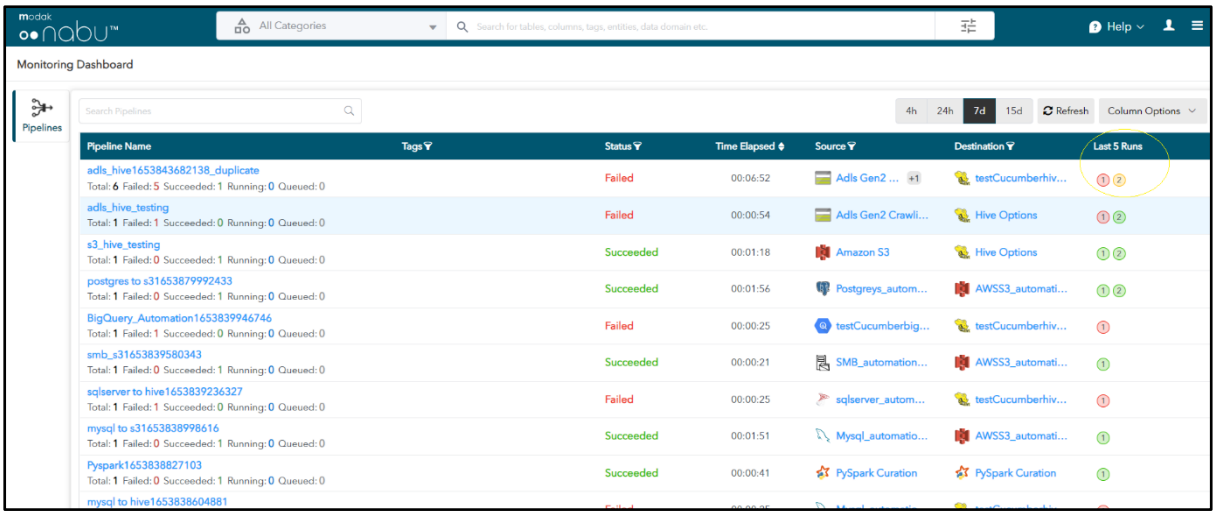


Figure 168: Monitoring Dashboards-Last 5 Runs

Use arrow which is on the top left corner of the window to go back to monitoring dashboard main page.

The user can view all the aforementioned features by using any of the criteria listed below:

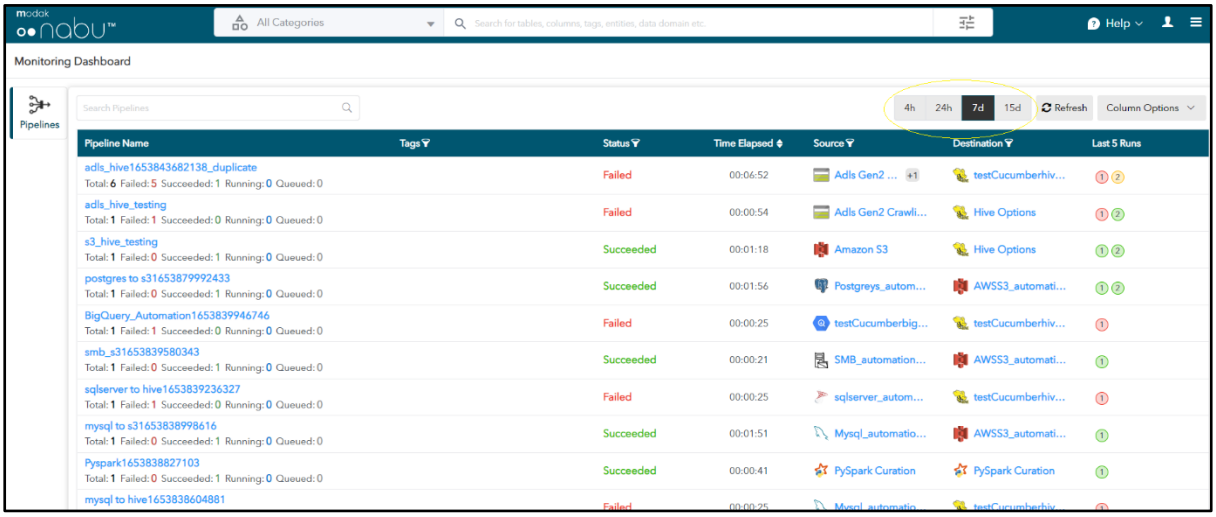


Figure 169: Monitoring Dashboards-Pipeline information for 7 Days

1. 4hours: By selecting this condition, the user can examine all the scheduled pipeline data for the last 4 hours.
2. 24hours: By selecting this condition, the user can examine all the scheduled pipeline data for the last 24 hours.
3. 7 days: By selecting this condition, the user can examine all the scheduled pipeline data for the 7 days.
4. 15days: By selecting this condition, the user can examine all the scheduled pipeline data for the last 15 days.

Refresh: This indicator assists the user in refreshing the table. When you click on the refresh button, the results are refreshed and updated.

Column options: This option enables the user to select the required columns to display on the monitoring dashboard. Some columns are selected by default, while others are not. For example, pipeline name, tags, status, time elapsed, source, destination, and the last 5 runs are all selected by default. According to user preferences, the user can deselect the option from the menu.

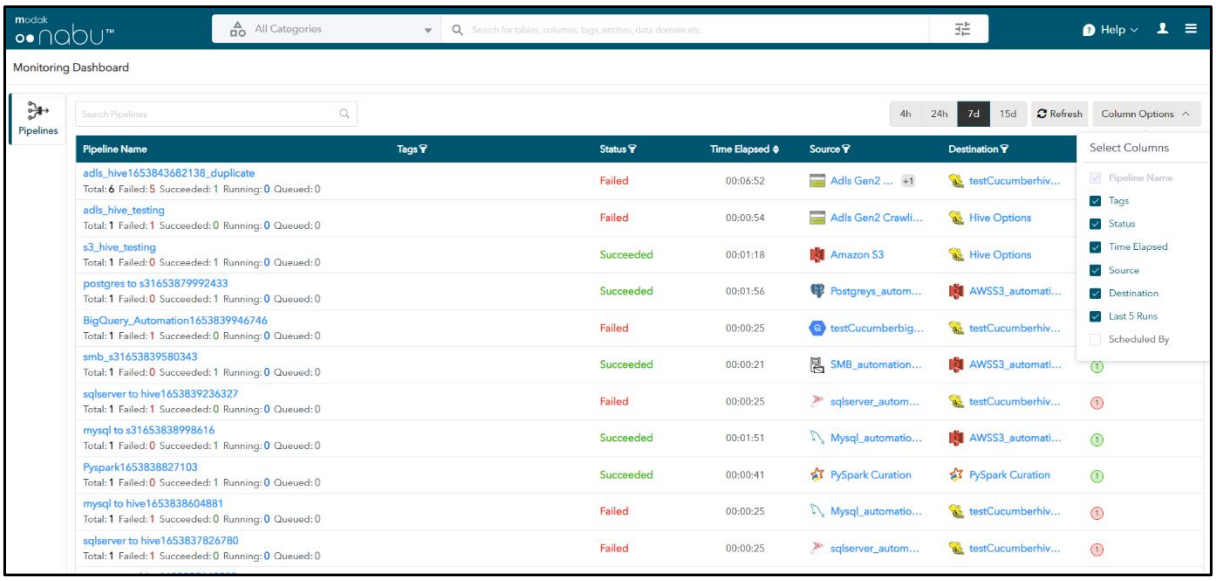


Figure 170: Monitoring Dashboards-Column Options

Users can obtain detailed information about the pipeline by clicking on the name of the pipeline from the list. It takes viewers to a page where they can see the features listed below in the first table. As shown in the below image.

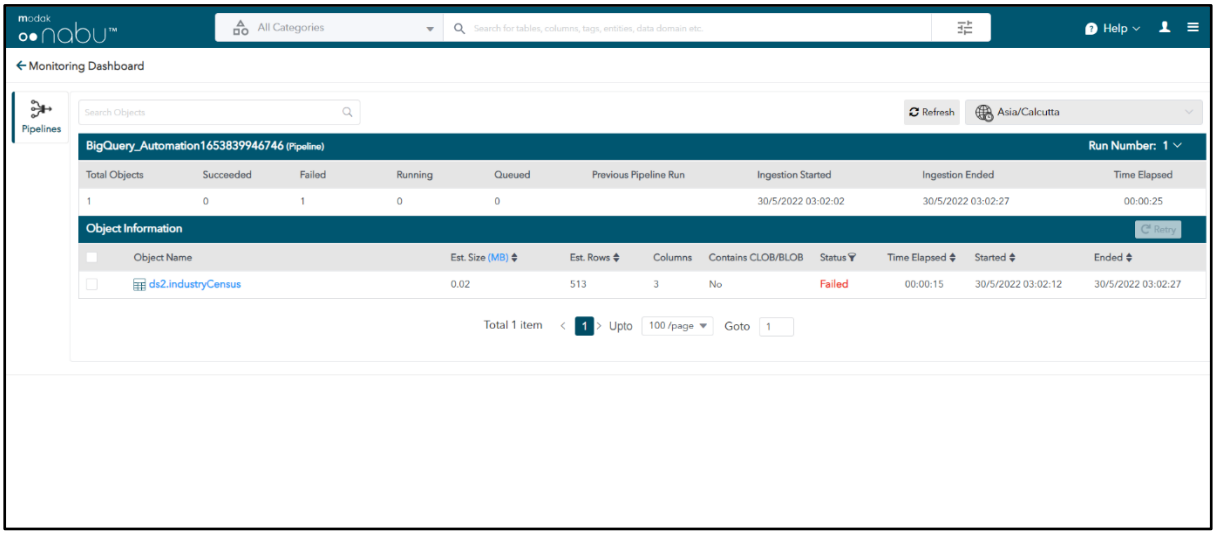


Figure 171: Monitoring Dashboards-Pipeline Information

1. Users will see the name of the selected pipeline and the run number in the header. This is a rundown of the pipeline's last five runs; this option allows users to view any run from the list as per their preference.
2. Displays precise information about the overall number of objects, how many succeeded, failed, are running, and are queued.
3. Shows the preceding pipeline run's information, such as the date and time of the last pipeline run. The date and time format (DD/MM/YYYY) can be adjusted according to user's choice from the time zone dropdown on the top right. By default, the time zone selected is user's browser time zone.
4. It specifies when pipeline ingestion started and ended, as well as the exact time and data. The date and time format (DD/MM/YYYY) can be adjusted according to user's choice as mentioned above.
5. The time elapsed, which reflects the real time required for the ingestion procedure, or the time difference between the start and end periods of the ingestion process.

Refresh: This button assists the user in refreshing the table. When you click on the refresh button, the results are refreshed and updated.

The Object information section shows the detailed information about the objects for that selected pipeline. It includes the below

Object name: showcases the name of the object for the pipeline, the object may be a table, file, View or Materialized view. Each represent by a distinct symbol. Once user clicks on the object name it shows a popup with the below information.

a. The object's name: the object could be a table, file, view, or materialized view.
b. Retry attempts indicates how many times it has been rescheduled/retried for ingestion.
c. Application ID for the object
d. BOTs process ID of the ingestion process.
e. The bot's process context displays the stages of the process, status of the stages, and time elapsed it took to complete the ingestion process.

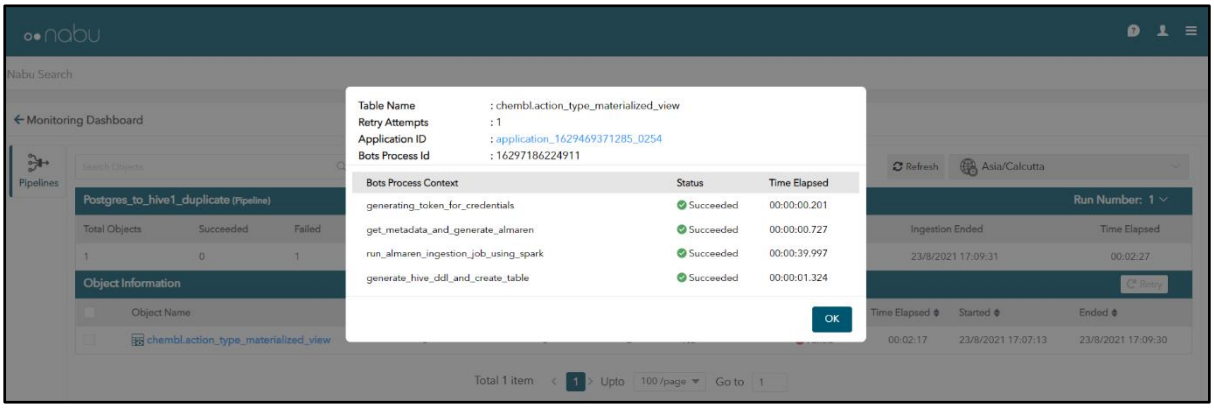


Figure 172: Object Name Popup

Est size: It shows the object's estimated size. The user can select the size as per preference. It could be bytes, KB, MB, GB, or TB. Users can sort the object's size according to their preferences.

Est Rows: It displays the number of rows in the table if the object is the table. Users can sort the rows of the table according to their preferences.

Columns: It shows the number of columns in the table if the object is the table.

Contains CLOB/BLOB: If that object includes any CLOB/BLOB datatype, it displays as yes, else no.

Status: It displays the state of the scheduled object, such as failed, succeeded, running, or ignored. Each state is denoted by a unique symbol.

Time Elapsed: It shows the exact amount of time taken to complete the ingestion process for the objects. Users can sort the time elapse of the objects according to their preferences.

Started: Displays the data and the time of when the ingestion process of the object began. Users can sort the started time and date of the objects according to their preferences.

Ended: Displays the data and the time of when the ingestion process of the object was completed. Users can sort the end time and date of the objects according to their preferences.

Retry: This option allows users to reschedule the ingestion process for failed objects. The user can reschedule only the failed objects from the pipeline, not the succeeded ones. To reschedule, simply check the box next to the object's name. This check box is initially disabled. It is enabled when the user selects the failed objects. The user has the option of selecting single or multiple objects as they see fit. After selecting, the user is taken to a popup window where they can select reschedule options

for the failed objects.

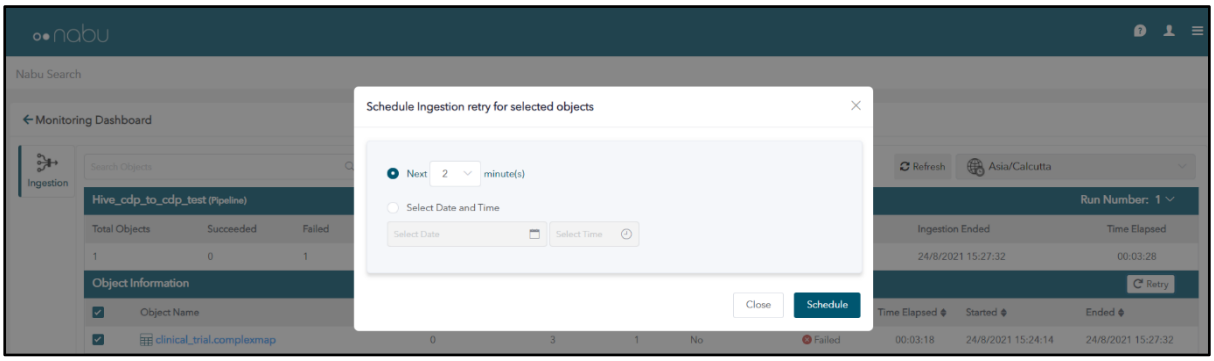


Figure 173: Monitoring Dashboards-Retry Schedule Popup

7.2 Executive Dashboard

The executive dashboard's functionality is to provide an interface for users to present the data, performance, progress, and metrics of various jobs for Data crawling, Pipelines and Data profiling

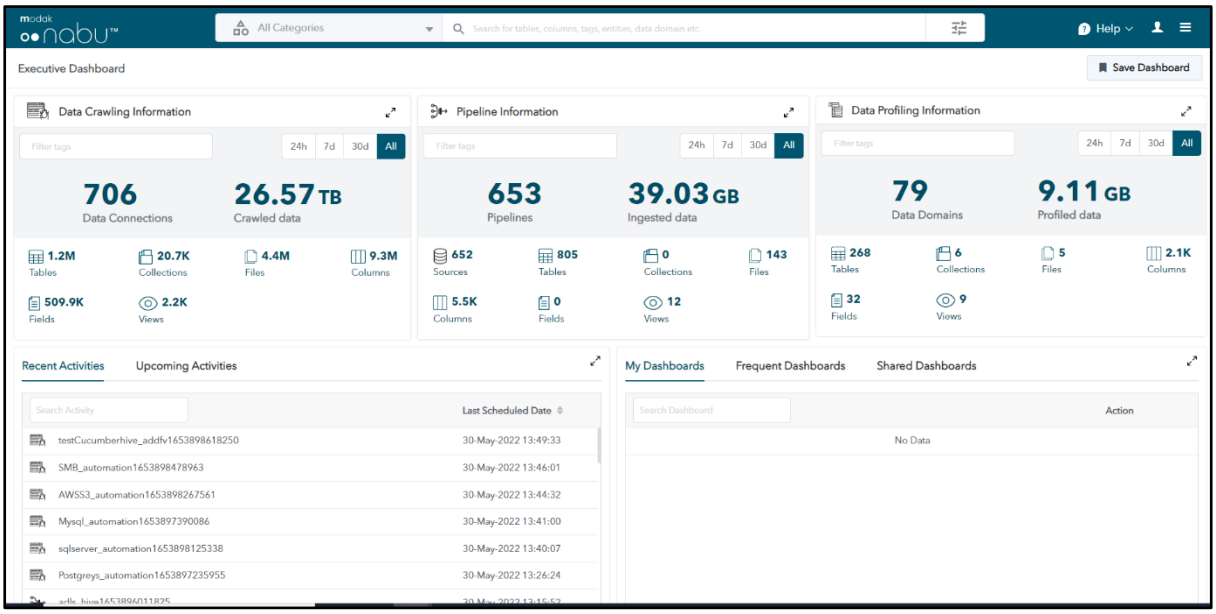


Figure 174: Executive Dashboard page

7.2.1 Data crawling information

This shows the detailed information related to the existing data connections and the crawled data. The Data crawling section shows the metrics of the below.

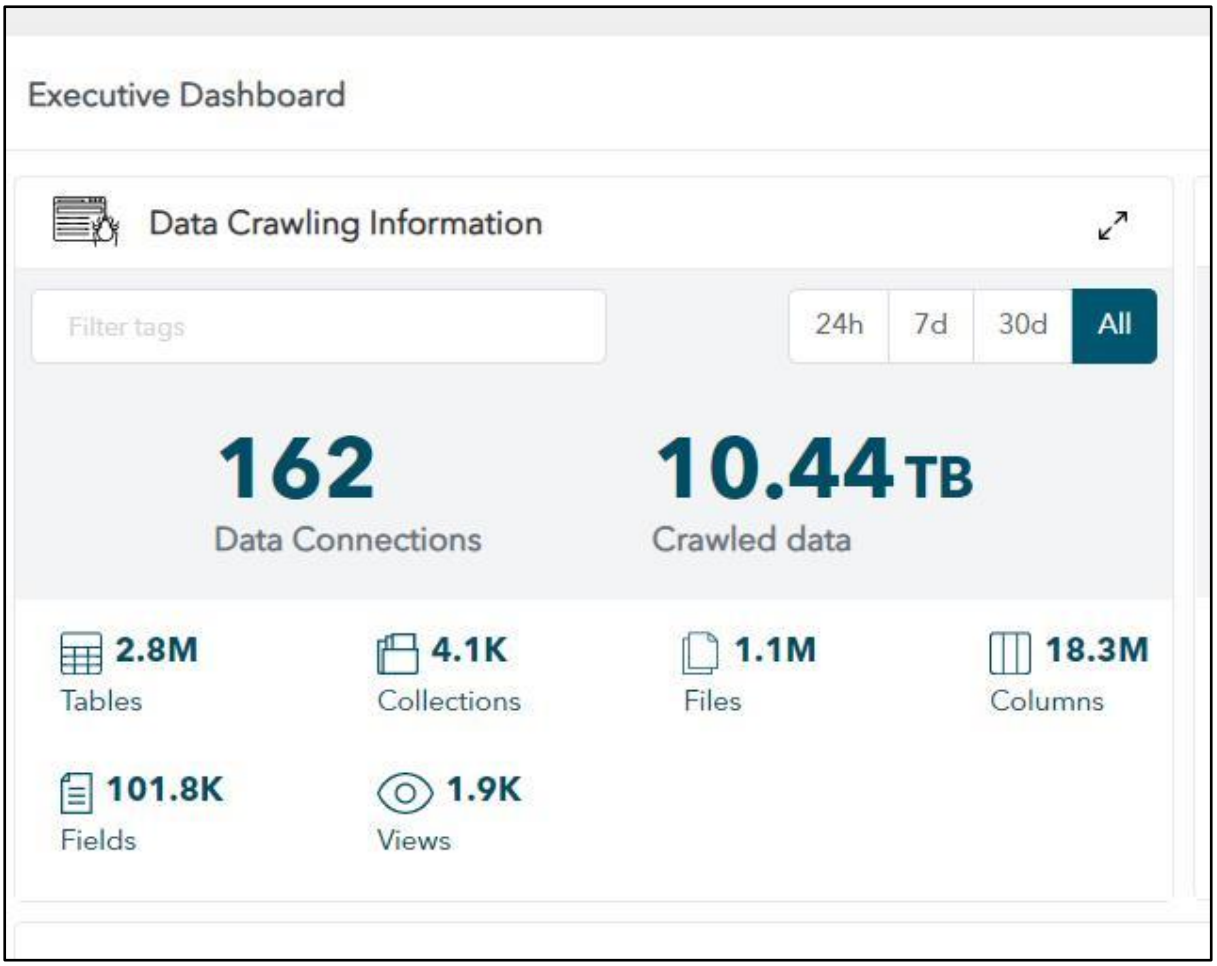


Figure 175: Executive Dashboard- Data Crawling Information

1. The total number of existing data connections.
2. The information on the size of the data that is crawled. (Ex: GB, MB, TB, PB etc.)
3. Filter tags: After selecting a tag from the drop-down, it assists the user to filter the data connection information and metrics for the tag applied. The data crawling information widget is refreshed and updated with new results. The user can apply single or multiple tags. The user can view applied filters on the expanded/collapsed view of the executive dashboard. Ex: If the user applies filter on the landing page, the expanded view also shows the same tag that is applied.
4. Number of tables/collections/files crawled for the existing data connections depending on the data connection type.
5. Number of columns/fields for the associated tables/collections.
6. Number of views crawled for the existing data connection.

The user can view all the above data/metrics by selecting any of the below options:

1. **24 hours:** The user can view all the above information for the last 24 hours.
2. **7 days:** The user can view all the above information for the last 7 days.
3. **30 days:** The user can view all the above information for the last 30 days.
4. **All:** This option shows the total Data crawling information. By default, this option is selected on the landing page.

The data crawling information widget can be expanded further to view detailed information of data crawling. Simply by clicking on the icon in the upper right corner of the data crawling information widget. The following features are included in the expanded widget.

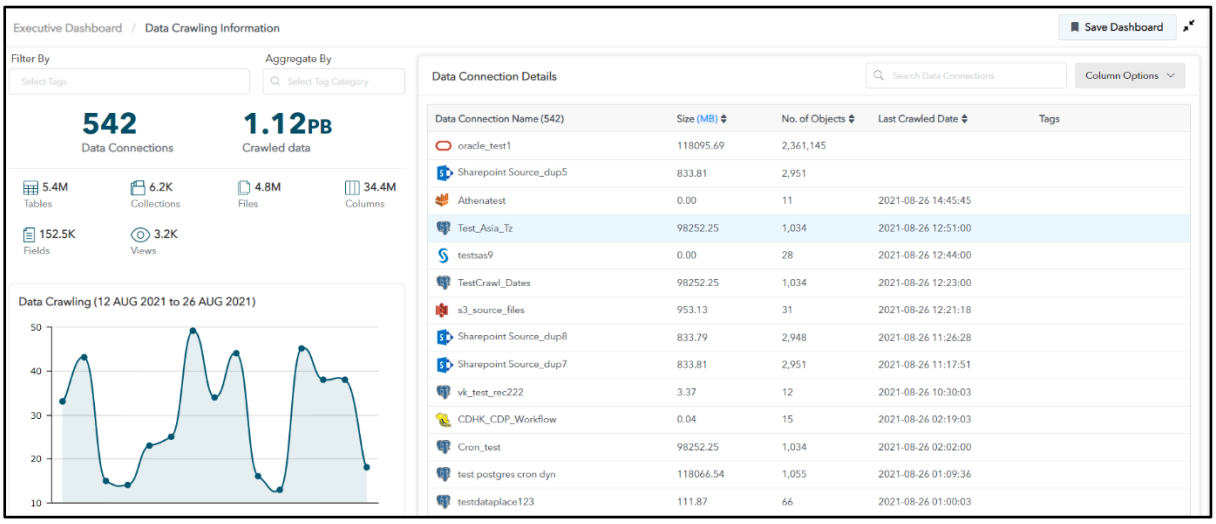


Figure 176: Executive Dashboard-Expanded View of Data Crawling Information

By default, after expanding the widget, the user can see detailed information about the entire data connections in the form of a table on the right side of the window with a title as **Data connection details**, the table includes the below columns.

1. **Data Connections Name:** this column displays the list of all the existing data connections names.
2. **Size:** This column shows the size of the data crawled for the data connection. The user can select the size type from the drop-down list as per preference. It could be bytes, KB, MB, GB, or TB. This column is sortable.
3. **No. of objects:** this column displays the total number of objects for the data connection. This column is sortable.
4. **Last crawled date:** It shows the last crawled date for the data connection. This column is sortable.
5. **Tags:** This column shows tags that are part of the data connection.

Column options: This option enables the user to select the required columns to display on the **data connection table**. Some columns are selected by default, while others are not.

Search data connections: This is a search box which enables the user to find the required data connection by simply typing the name of the data connection into the search box.

Filter By: This dropdown allows the user to filter the data connections results based on a specific tag. The user can select the tag from the drop down and the **Data connections details table** will get refreshed for the tag applied.

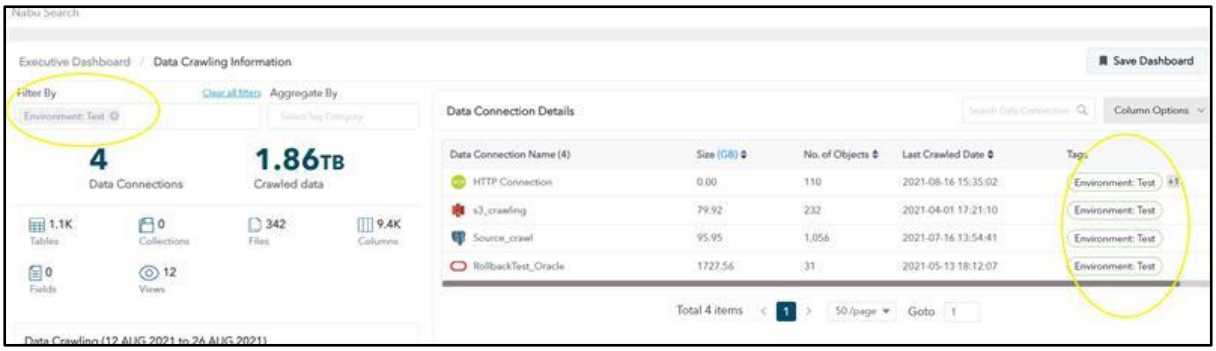


Figure 177: Executive Dashboard-Data Crawling Information-Filter By

Aggregate By: This dropdown allows the user to filter the data connections results based on a specific tag category. The user can view the refreshed results on the right side of the window with a title as **Aggregate results on Selected tag category.** The user can access specific information about the data connections linked with a selected tag category. This table can be expanded by clicking on the icon on the top right side of the table. The table comprises of the following columns.

1. A list of the tag value names for the selected tag category.
2. Shows the number of data connections associated with the tag category.
3. Shows the size of the data crawled for the data connection. The user can select the size type from the drop-down list as per preference. It could be bytes, KB, MB, GB, or TB. This column is sortable.
4. Number of tables/collections/files/views crawled for the existing data connections depending on the data connection type.
5. Number of columns/fields for the associated tables/collections.

The user has a clickable **clear all filters** link above the filter by drop down. When the user clicks that link, all the applied filters will be removed, and the data gets refreshed.

Export: It enables the users to export the results displayed in the window. The user must give the file a specific name and choose a format to store it in from the menu, which includes.png,.jpg, and.pdf.

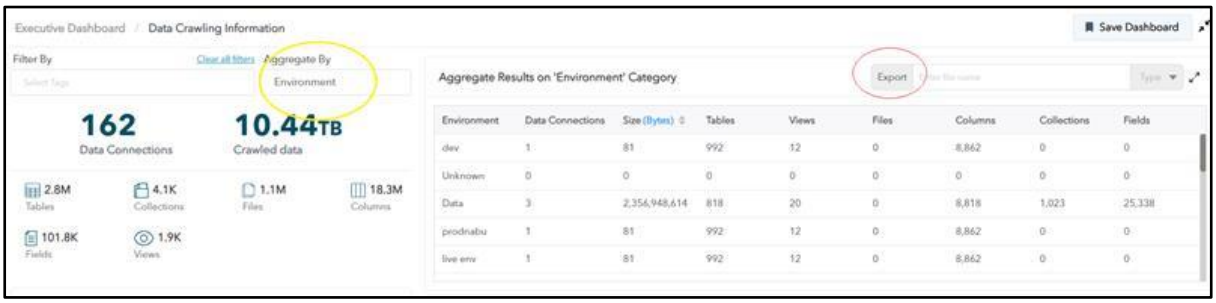


Figure 178: Executive Dashboard-Data Crawling Information-Aggregate By

Data crawling chart: It is the visual representation for the user to show completed crawling jobs for a specific period. Users will be able to observe how many jobs were successfully completed on that day on each graph plot. E.g., This chart shows the jobs completed each day for a period of 15 days.

The x-axis shows the days, and the y-axis shows the jobs count.

Figure 179: Executive Dashboard-Data Crawling Chart

7.2.2 Pipeline information

This shows the detailed information related to the existing data pipelines and the ingested data. The Data pipelines information section shows the metrics of the below.

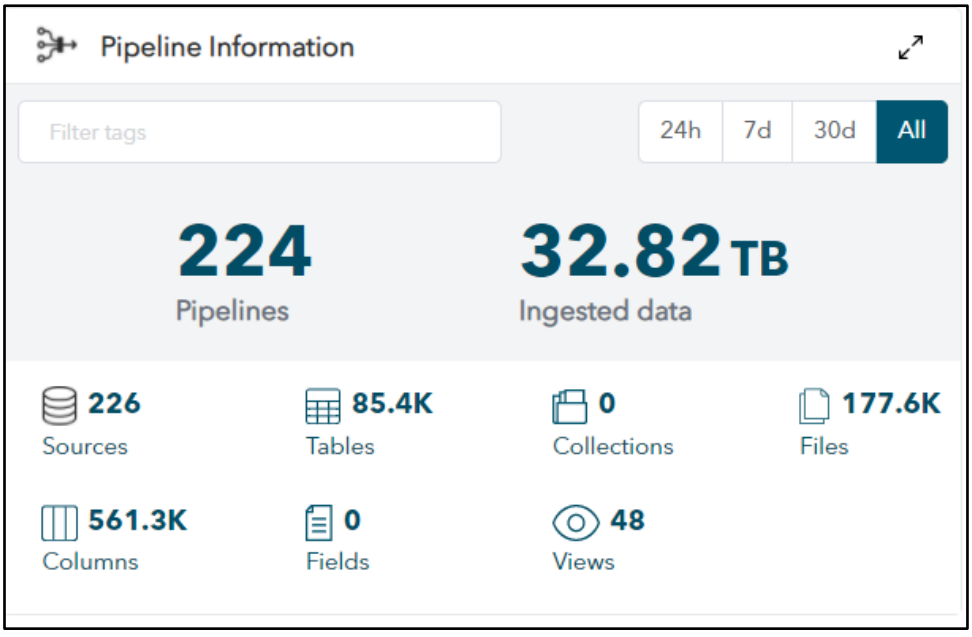


Figure 180: Executive Dashboard-Pipeline Information

1. The total number of existing data pipelines.
2. The information on the size of the data that is ingested. (Ex: GB, MB, ZB, PB etc.)
3. Filter tags: After selecting a tag from the drop-down, it assists the user in searching for a data pipeline that relates to the specified tag. The data pipeline information widget is refreshed and updated with new results. The user can apply single or multiple tags. The user can view applied filters on the expanded/collapsed view of the executive dashboard. Ex: If

the user applies filter on the landing page, the expanded view also shows the same tag that is applied.

4. The number of source/tables/collections/files scheduled for the existing data pipelines depends on the source type.
5. Number of columns/fields for the associated tables/collections.
6. Number of views scheduled for the existing data pipelines.

The user can view all the above data/metrics by selecting any of the below options:

1. **24 hours:** The user can view all the above information for the last 24 hours.
2. **days:** The user can view all the above information for the last 7 days.
3. **30 days:** The user can view all the above information for the last 30 days.
4. **All:** This option shows the total Data pipeline information. By default, this option is selected on the landing page.

The data pipeline information widget can be expanded further to view detailed information of data ingested. Simply by clicking on the icon in the upper right corner of the data pipeline information widget. The following features are included in the expanded widget.

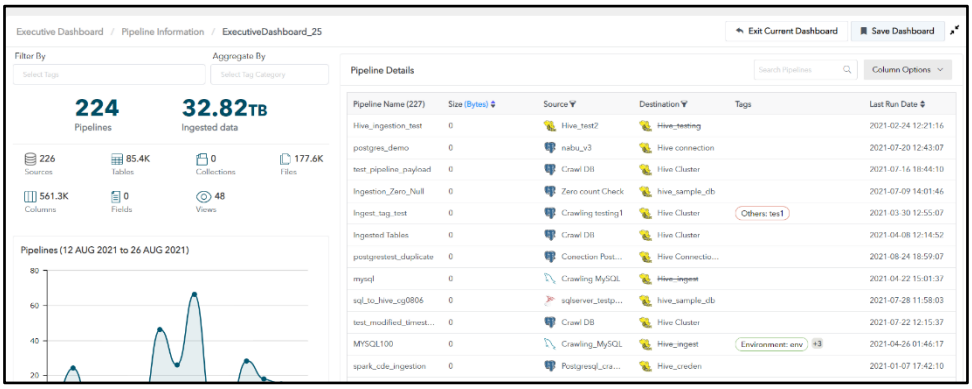


Figure 181: Executive Dashboard-Expanded view of Pipeline Information

By default, after expanding the widget, the user can see detailed information about the entire data pipelines in the form of a table on the right side of the window with a title as **data pipeline details**, the table includes the below columns.

Pipeline Name: this column displays the list of all the existing data pipeline names.

Size: This column shows the size of the data that is ingested for the data pipeline. The user can select the size type from the drop-down list as per preference. It could be bytes, KB, MB, GB, or TB. This column is sortable.

Source: This column gives the source details of the ingestion pipelines. The user can filter the results based on the sources they want by using the filter option. Simply click on the filter icon, select the desired source from the menu, then click apply to view it.

Destination: This column gives the destination details of the ingested pipelines. The user can filter the results based on the destination they want by using the filter option. Simply click on the filter icon, select the desired destination from the menu, then click apply to view it.

Tags: This column shows tags that are part of the ingestion pipelines.

Last run date: It shows the last scheduled or ingested date for the data pipelines. This column is sortable.

Column options: This option enables the user to select the required columns to display on the **pipeline**

details table. Some columns are selected by default, while others are not.

Search data connections: This is a search box which enables the user to find the required pipeline name from the table by simply typing the name of the pipeline in the search box.

Filter By: This dropdown allows the user to filter the pipeline results based on a specific tag. The user can select the tag from the drop down and the **pipeline details table** will get refreshed for the tag applied.

Aggregate By: This dropdown allows the user to filter the pipeline results based on a specific tag category. The user can view the refreshed results on the right side of the window with a title as **Aggregate result Selected tag category.** The user can access specific information about the pipelines linked with a selected tag category. This table can be expanded by clicking on the icon on the top right side of the table. The table comprises of the following columns.

1. A list of the tag value names for the selected tag category.
2. It shows the number of pipelines associated with that specific tag category.
3. shows the number of data connections associated with a specific tag category.
4. Shows the size of the data ingested for the pipelines. The user can select the size type from the drop-down list as per preference. It could be bytes, KB, MB, GB, or TB. This column is sortable.
5. Number of tables/collections/files/views crawled for the existing pipelines depending on the source type.
6. Number of columns/fields for the associated tables/collections.

The user has a clickable **clear all filters** link above the filter by drop down. When the user clicks that link, all the applied filters will be removed, and the data gets refreshed.

Export: It enables users to export the results displayed in the window. The user must give the file a specific name and choose a format to store it in from the menu, which includes.png,.jpg, and.pdf.

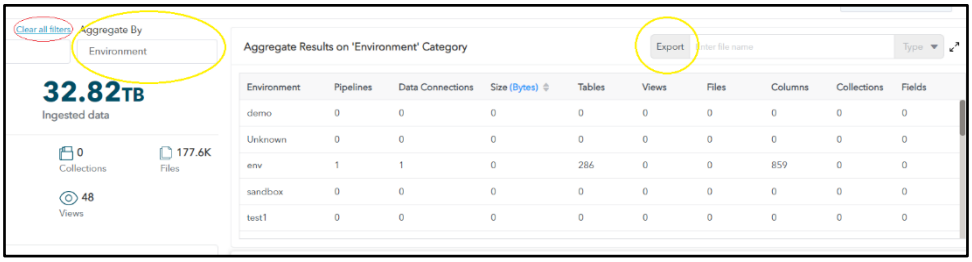


Figure 182: Executive Dashboards-Export, Aggregate and Clear all Filters by Options

Pipeline chart: It is the visual representation for the user to show completed pipeline jobs for a specific period. Users will be able to observe how many jobs were successfully completed on that day on each graph plot. E.g., This chart shows the jobs completed each day for a period of 15 days.

The x-axis shows the days, and the y-axis shows the jobs count.

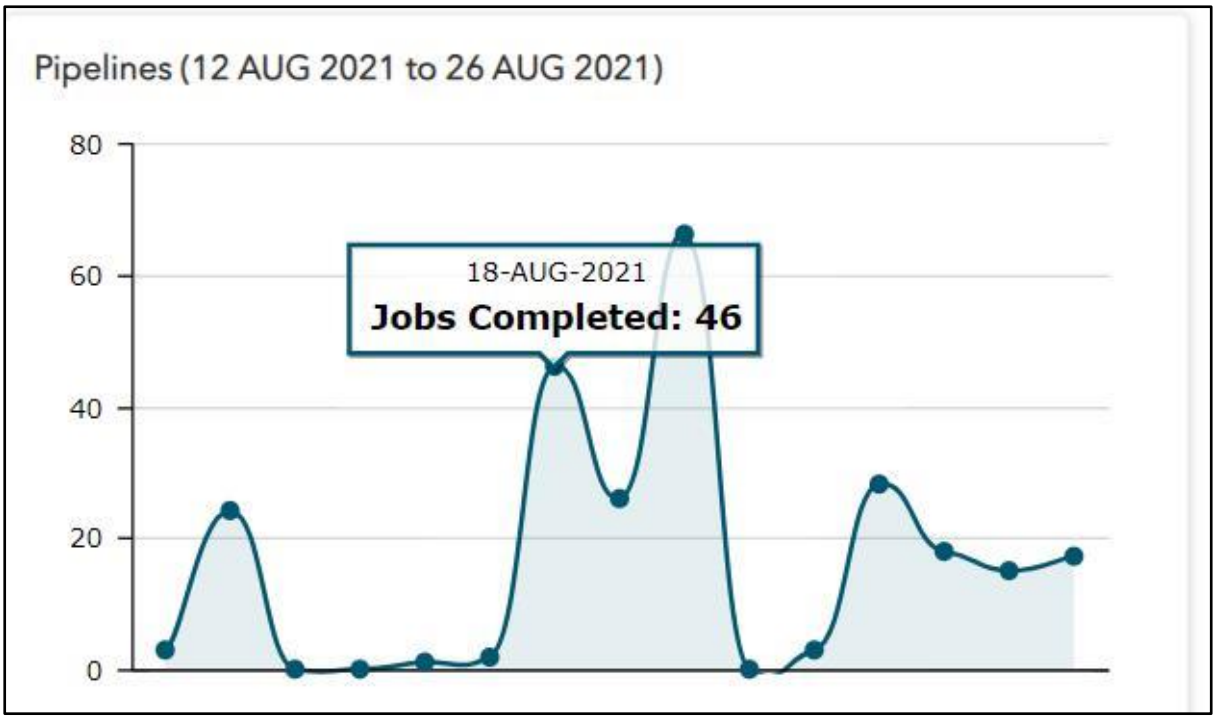


Figure 183: Executive Dashboards-Pipeline Chart

7.2.3 Data profiling information

This shows the detailed information related to the existing data domains and the profiled data. The Data profiling section shows the metrics of the below.

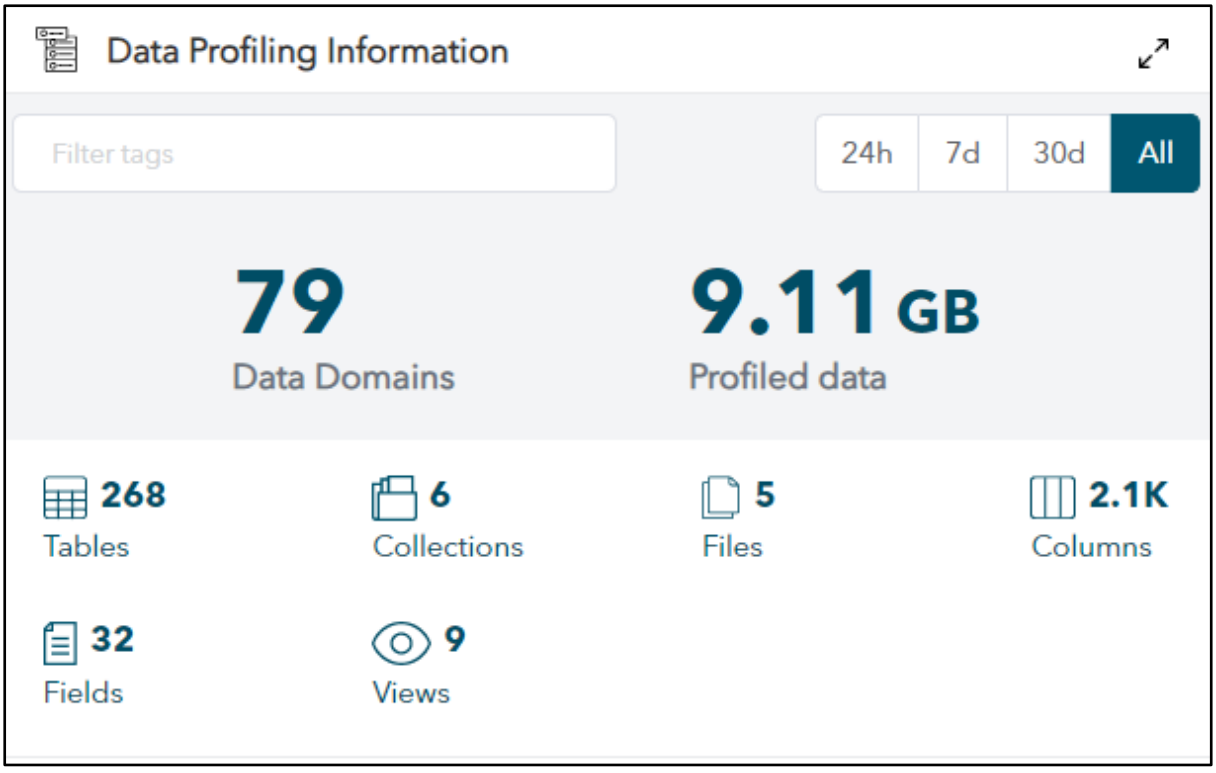


Figure 184: Executive Dashboard-Data Profiling Information

1. The total number of existing data domains.
2. The information on the size of the data that is profiled (Ex: GB, MB, TB, PB etc.)

3. Filter tags: After selecting a tag from the drop-down, it assists the user to filter the data profiling information and the metrics for the tag applied. The data profiling information widget is refreshed and updated with new results. The user can apply single or multiple tags. The user can view applied filters on the expanded/collapsed view of the executive dashboard. Ex: If the user applies filter on the landing page, the expanded view also shows the same tag that is applied.
4. Number of tables/collections/files profiled for the existing data domains depending on the data domain type.
5. Number of columns/fields for the associated tables/collections.
6. Number of views profiled for the existing data domains.

The user can view all the above data/metrics by selecting any of the below options:

1. **24 hours:** The user can view all the above information for the last 24 hours.
2. **7 days:** The user can view all the above information for the last 7 days.
3. **30 days:** The user can view all the above information for the last 30 days.
4. **All:** This option shows the total Data profiling information. By default, this option is selected on the landing page.

The data profiling information widget can be expanded further to view detailed information of data profiled and data domains. Simply by clicking on the icon in the upper right corner of the data profiling information widget. The following features are included in the expanded widget.

By default, after expanding the widget, the user can see detailed information of the data domains in the form of a table on the right side of the window with a title as **Data domains details**, the table includes the below columns.

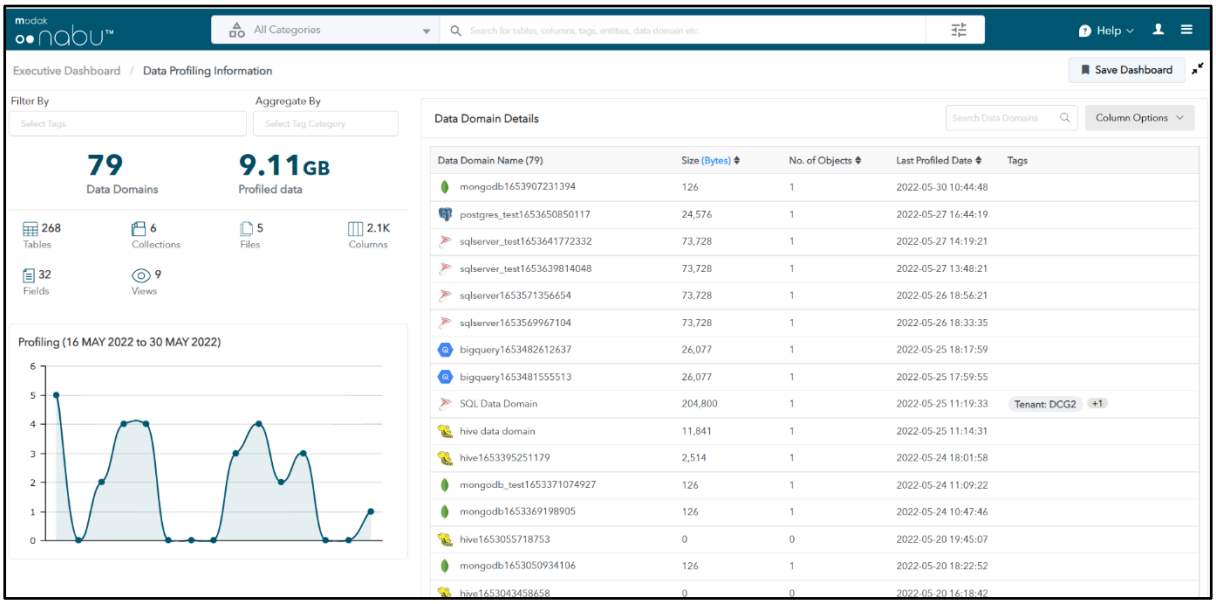


Figure 185: Executive Dashboard-Expanded view of Data Profiling Information

Data domain name: This column displays the list of all the existing data domains names.

Size: This column shows the size of the data profiled for the data domains. The user can select the size type from the drop-down list as per preference. It could be bytes, KB, MB, GB, or TB. This column is sortable.

No. of objects: this column displays the total number of objects for the data domain. This column is sortable.

Last profiled date: It shows the last profiled date for the data connection. This column is sortable.

Tags: This column shows tags that are part of the data domain

Column options: This option enables the user to select the required columns to display on the **data domain details table**. Some columns are selected by default, while others are not.

Search data domain: This is a search box which enables the user to find the required data domain by simply typing the name of the data domain into the search box.

Filter By: This dropdown allows the user to filter the data domains results based on a specific tag. The user can select the tag from the drop down and the **Data domain details table** will get refreshed for the tag applied.

Aggregate By: This dropdown allows the user to filter the data domains results based on a specific tag category. The user can view the refreshed results on the right side of the window with a title as **Aggregate results on Selected tag category.** The user can access specific information about the data domains linked with a selected tag category. This table can be expanded by clicking on the icon on the top right side of the table. The table comprises of the following columns.

1. A list of the tag value names for the selected tag category.
2. Shows the number of data domains associated with the tag category.
3. Shows the size of the data profiled for the data domains. The user can select the size type from the drop-down list as per preference. It could be bytes, KB, MB, GB, or TB. This column is sortable.
4. Number of tables/collections/files/views crawled for the existing data domain depending on the data domain type.
5. Number of columns/fields for the associated tables/collections.

The user has a clickable **clear all filters** link above the filter by drop down. When the user clicks that link, all the applied filters will be removed, and the data gets refreshed.

Export: It enables users to export the results displayed in the window. The user must give the file a specific name and choose a format to store it in from the menu, which includes.png,.jpg, and.pdf.

The below figure shows the Data profiling expanded view with the tag, tag category applied.

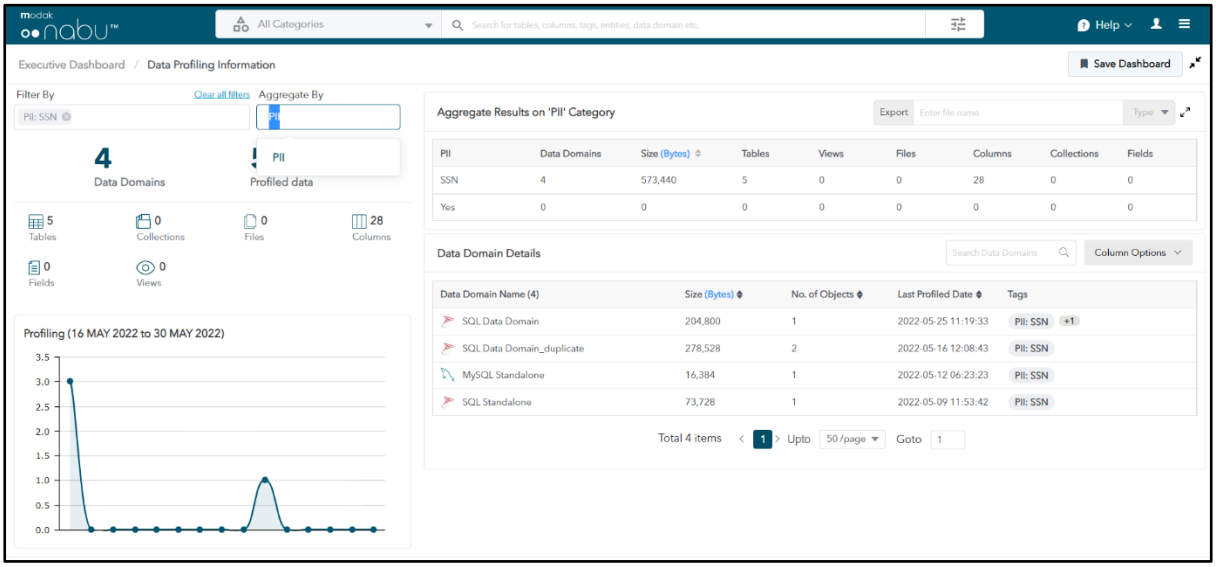


Figure 186: Executive Dashboard-Export Aggregate by Results

Profiling chart: It is the visual representation for the user to show completed profiling jobs for a specific period. Users will be able to observe how many jobs were successfully completed on that day on each graph plot. E.g., This chart shows the jobs completed each day for a period of 15 days. The x-

axis shows the days, and the y-axis shows the jobs count.

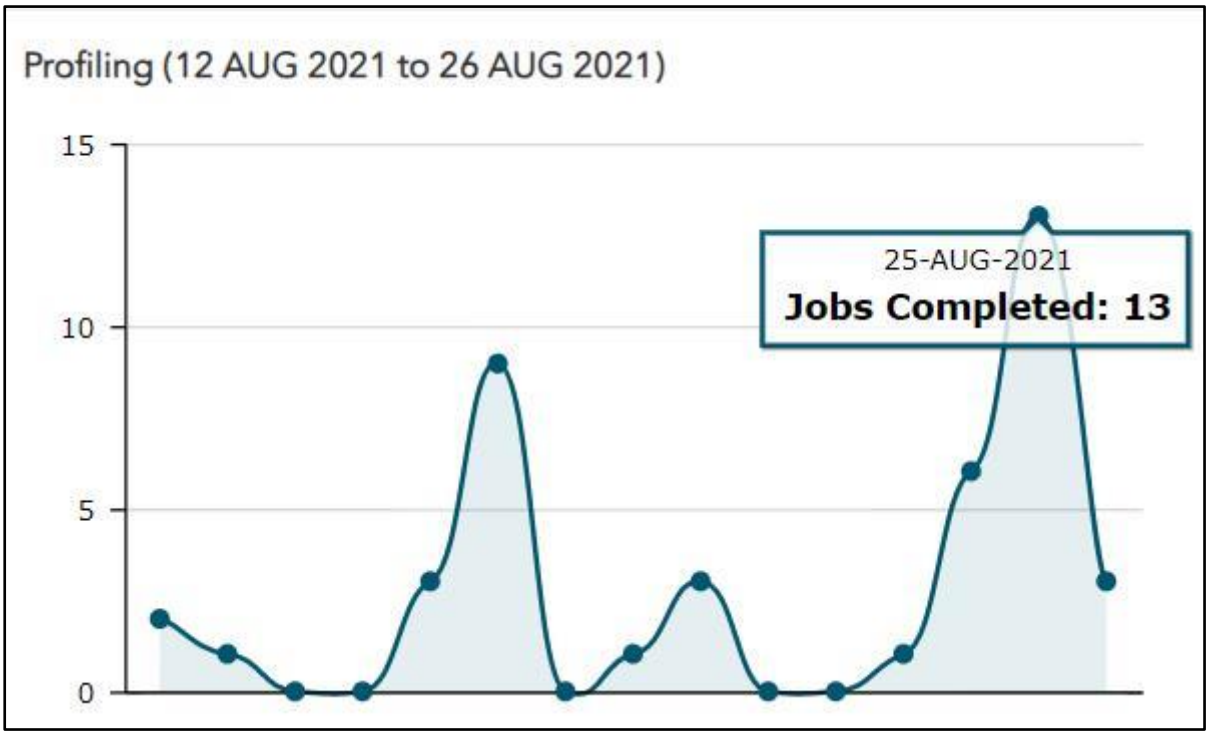


Figure 187: Executive Dashboard-Data Profiling Chart

Save and Share Dashboard: The executive dashboard provides the feature to save/share the dashboard from the landing page or from the expanded view of Data crawling, Pipeline Information and Data profiling Information.

1. The user can click on save dashboard button which is at the top right corner of the executive dashboard landing page or the expanded view. On clicking the save dashboard button, the user will be prompted with a popup as below. By default, the owner details will be populated as below, and the user can provide a unique name and click on the save button in the popup to save the dashboard.

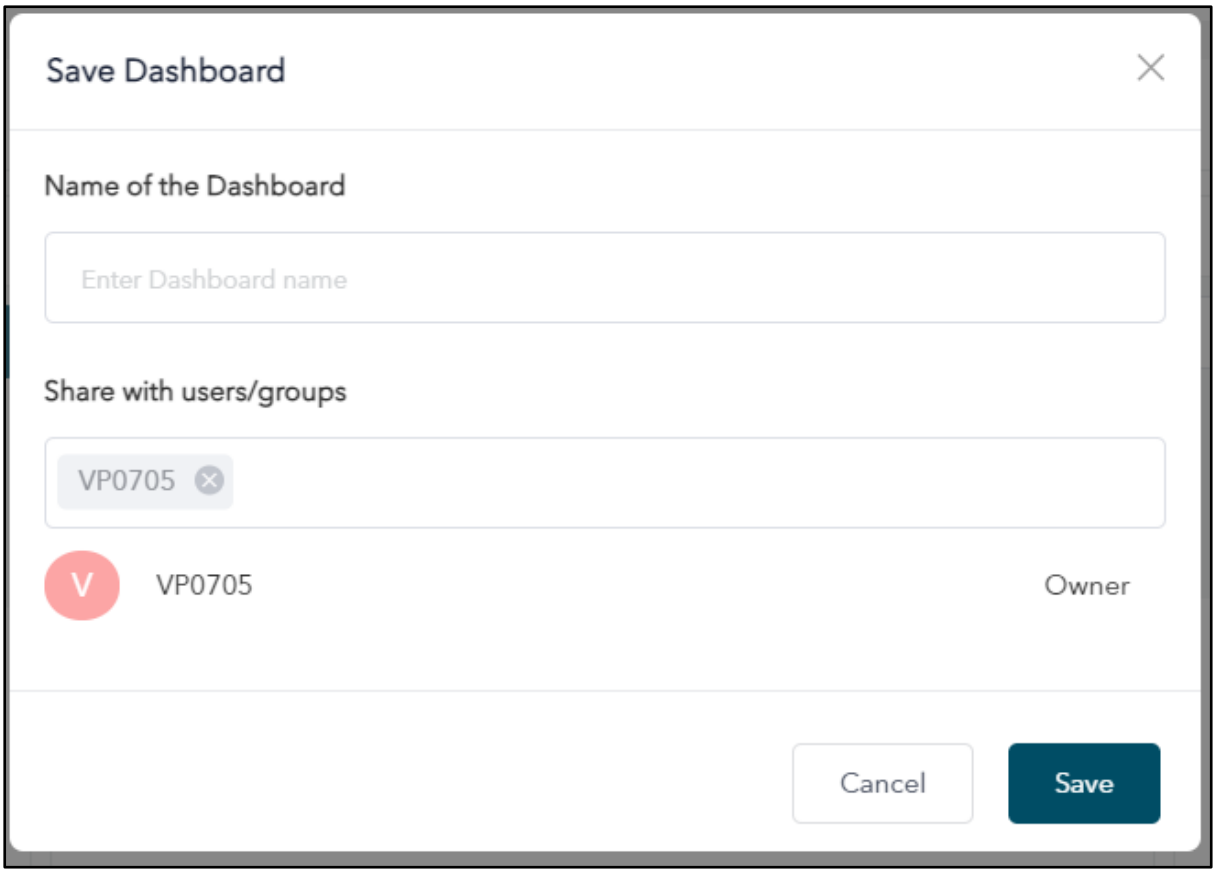


Figure 188: Save Dashboard Popup

Apart from saving the dashboard, the user can also share the dashboard with any other user or groups. The user can select the user/groups from the dropdown as shown below.

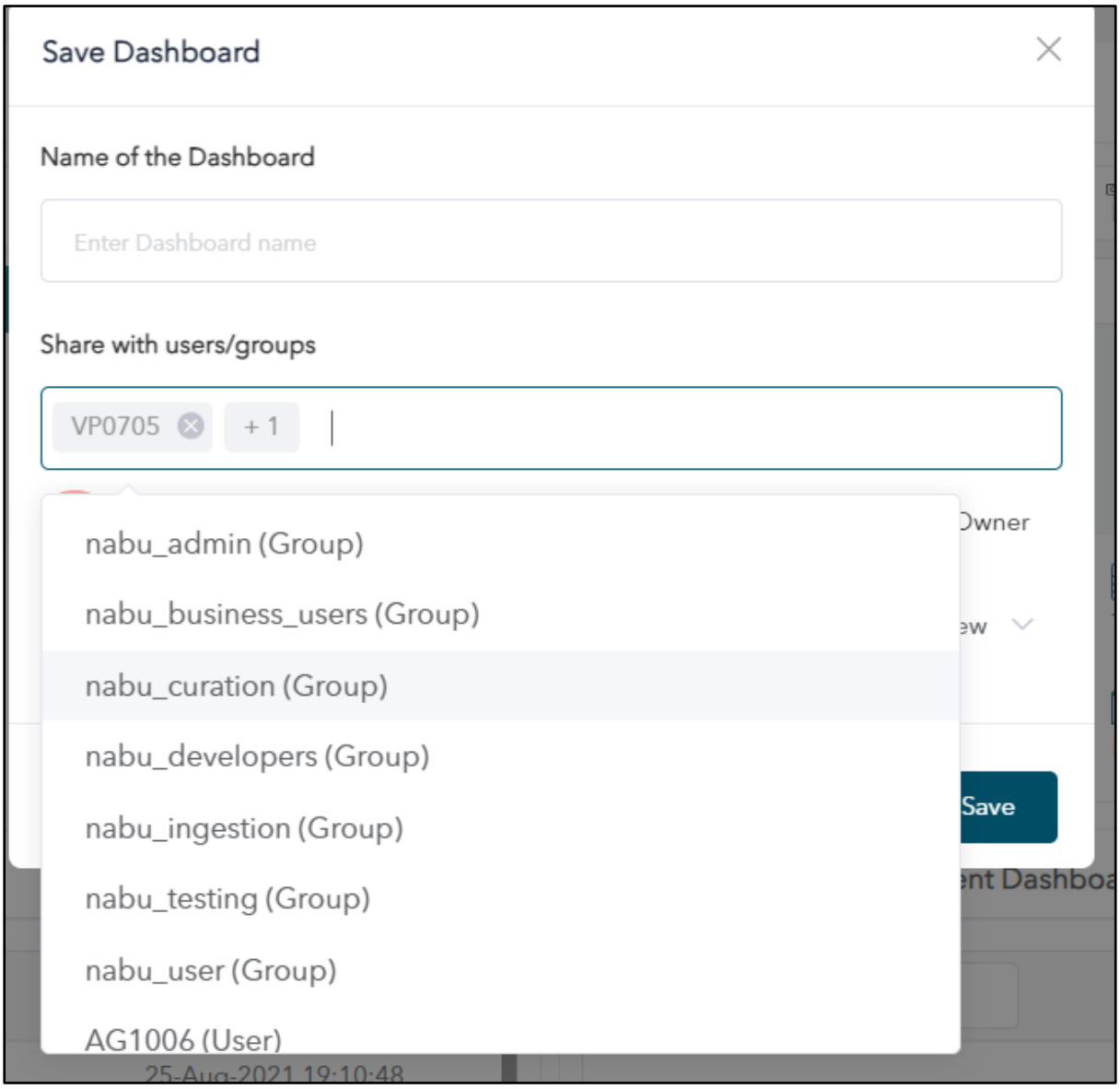


Figure 189: Search for Users/Groups

The user/group selected will be added to the list and the user can select the permissions to share the dashboard.

a. If the user grants **can view** access, the shared user can just view the dashboard and cannot perform any edit operations. The shared user can also share the respective dashboard to any other user/group with only view access.
b. If the user grants **can edit** access, the shared user can view and edit the dashboard. The shared user can also share the respective dashboard to any other user/group with both view and edit access.

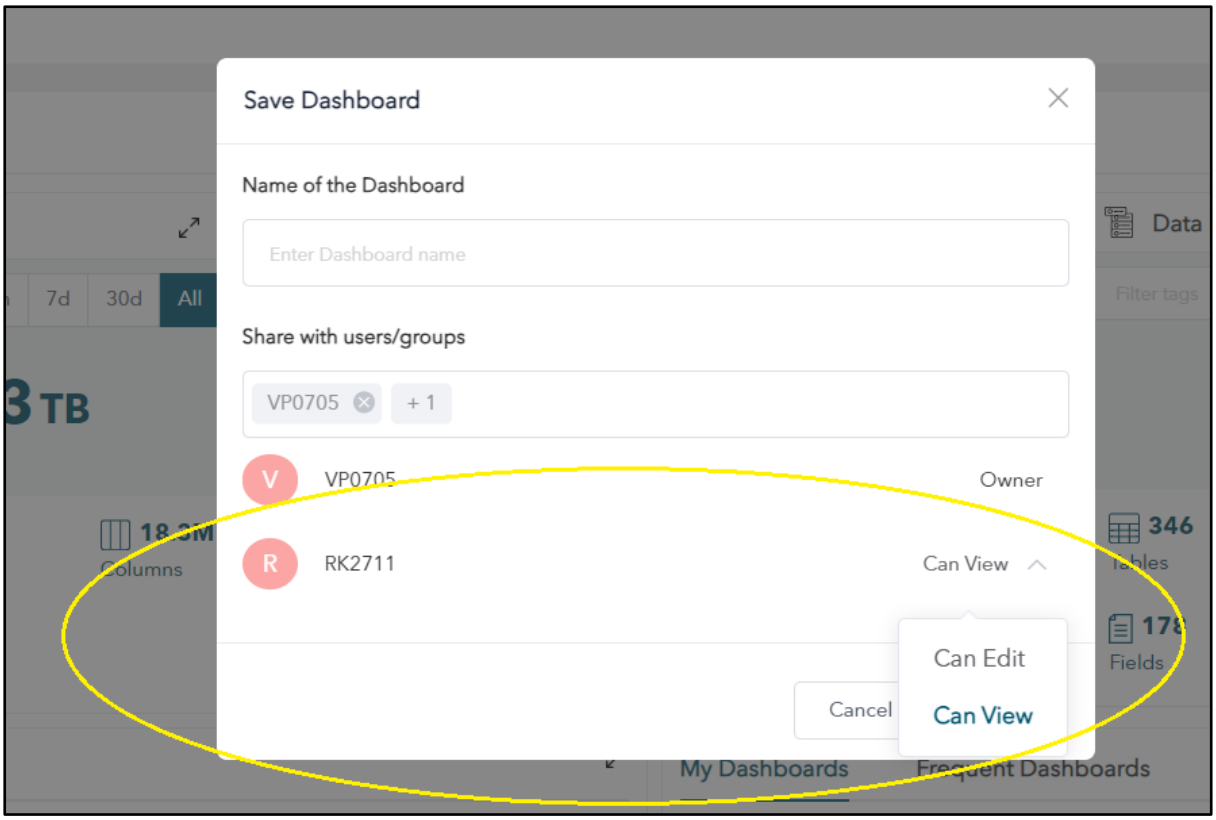


Figure 190: Edit/View Dashboard

As mentioned above, the dashboards which are saved by the user can be viewed/edited in the My Dashboards section which is at the bottom on the landing page. Three tabs (My Dashboards, Frequent Dashboards, Shared Dashboards) are shown as below.

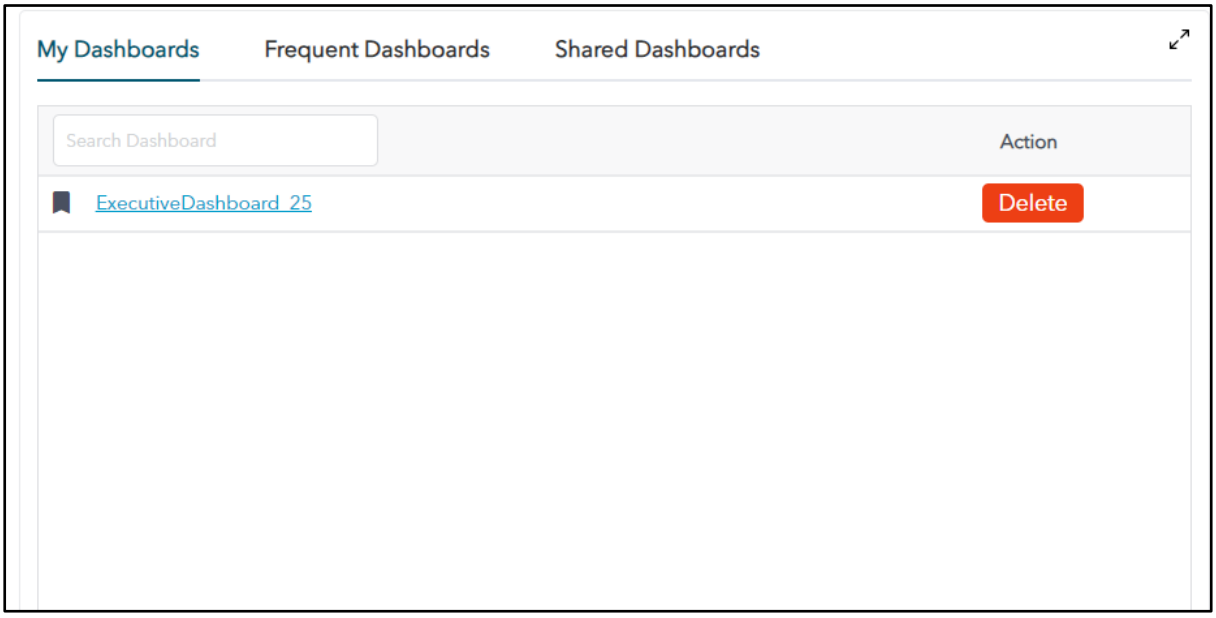


Figure 191: My Dashboard

My Dashboards tab is selected by default and the user can switch between different tabs as per their preferences.

My Dashboards: This section shows the dashboards which are created by the user. The user is the owner for these dashboards.

1. It shows the name of the dashboard that is saved with and when user clicks on the dashboard name, it shows the respective saved dashboard details. If the dashboard saved is in the expanded view, it redirects to the dashboard details in the expanded view or the vice versa.
2. The user can view or edit the dashboard as preferred from this tab. If the user edits the dashboard, the user can again save the edited dashboard. After edit, when user clicks on save dashboard button, the dashboard name will be populated as below, and the user can save the dashboard.

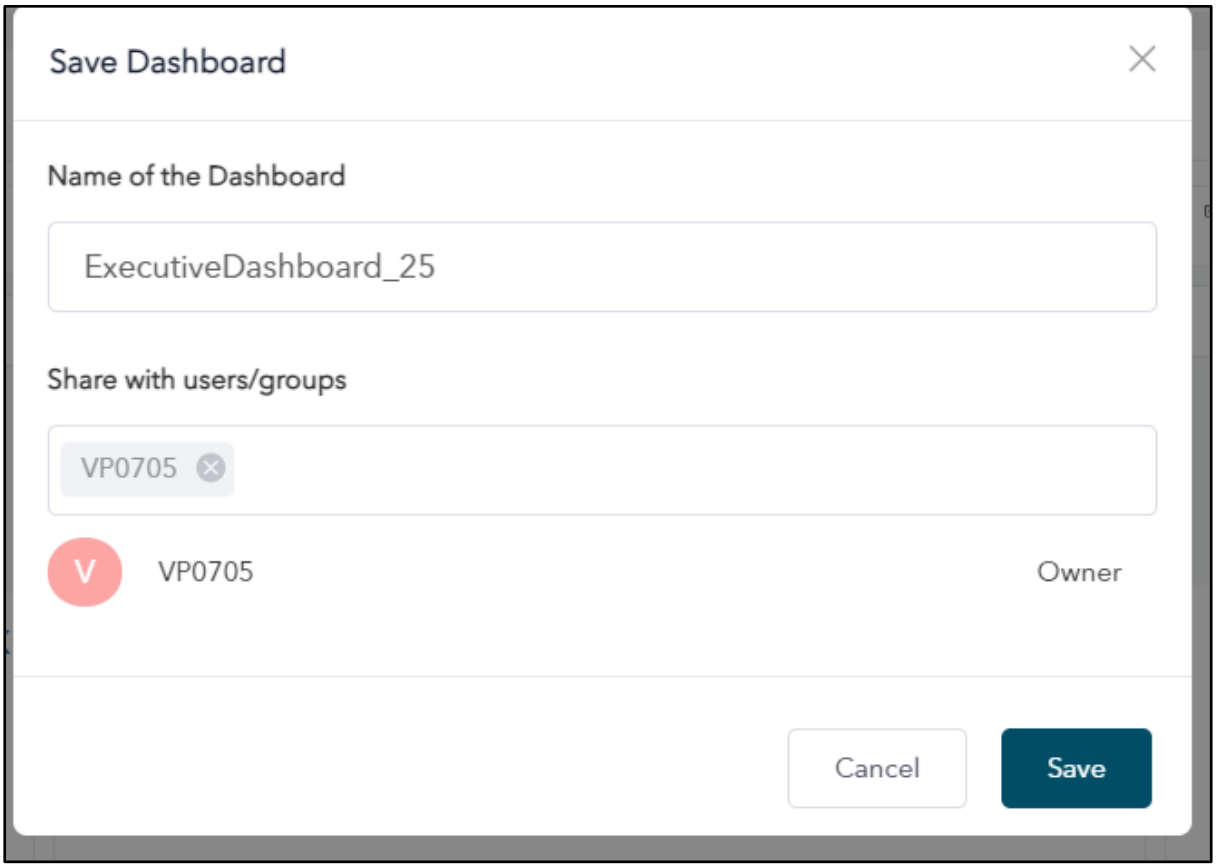


Figure 192: Edit Saved Dashboard

3. The user can also delete the dashboard by clicking on the delete button. On clicking the delete button, the user will be prompted for the confirmation to delete the dashboard as below. The dashboard will be deleted once the user clicks on confirm.

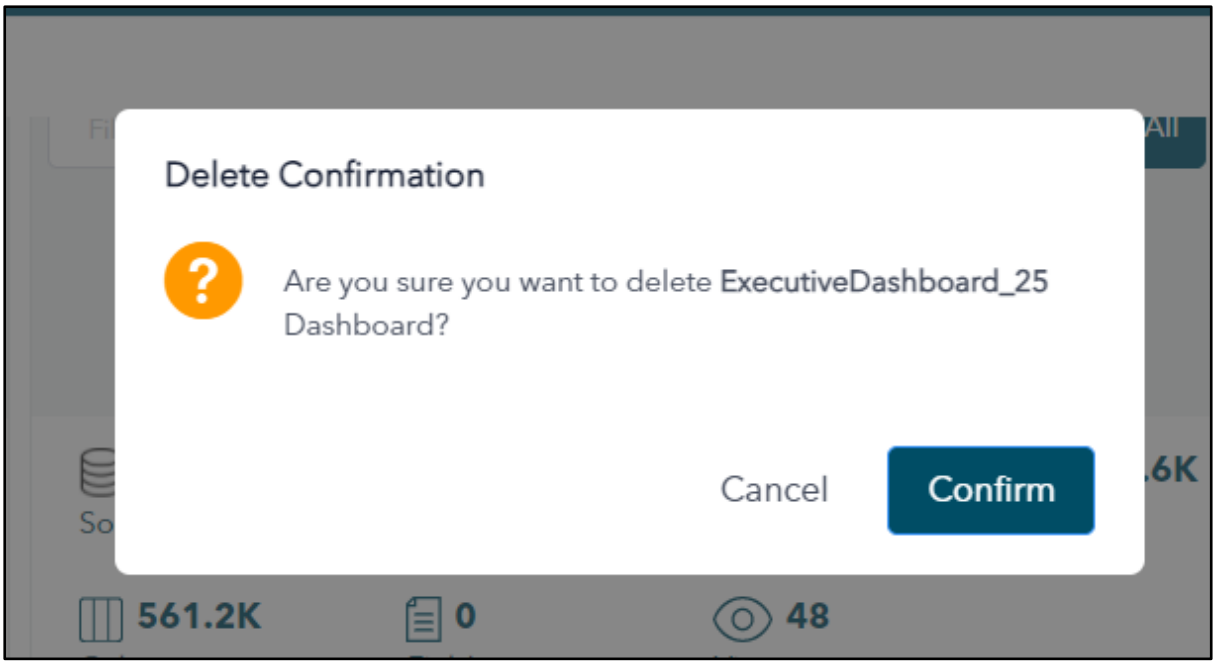


Figure 193: Executive Dashboard-Delete Confirmation Popup

Frequent Dashboards: This tab shows all the frequent dashboards that are accessible to the user.

1. It shows the name of the dashboard and when user clicks on the dashboard name, it shows the respective dashboard details.
2. The user can view/edit these dashboards according to the permission that is granted for the respective dashboard.
3. The user cannot delete the dashboards from this section

Shared Dashboards: This tab shows all the dashboards that are shared to the user by any other user/group.

1. It shows the name of the dashboard and when user clicks on the dashboard name, it shows the respective dashboard details.
2. The user can view/edit these dashboards according to the permission that is granted for the respective dashboard.
3. The user cannot delete the dashboards from this section.

The user can expand/collapse these tabs section by clicking on the expand/collapse icon to the top right corner of the tabs. On expand, the tabs section will be shown as below.

Figure 194: Dashboard Expanded View

Exit Current dashboard: For any of the activities above, if the user clicks on Exit current Dashboard button, the previous details that are on the dashboard gets refreshed.

Ex: If the user views/edits the dashboard by clicking any of the tabs above, the user can refresh the dashboard by clicking on exit current dashboard button.

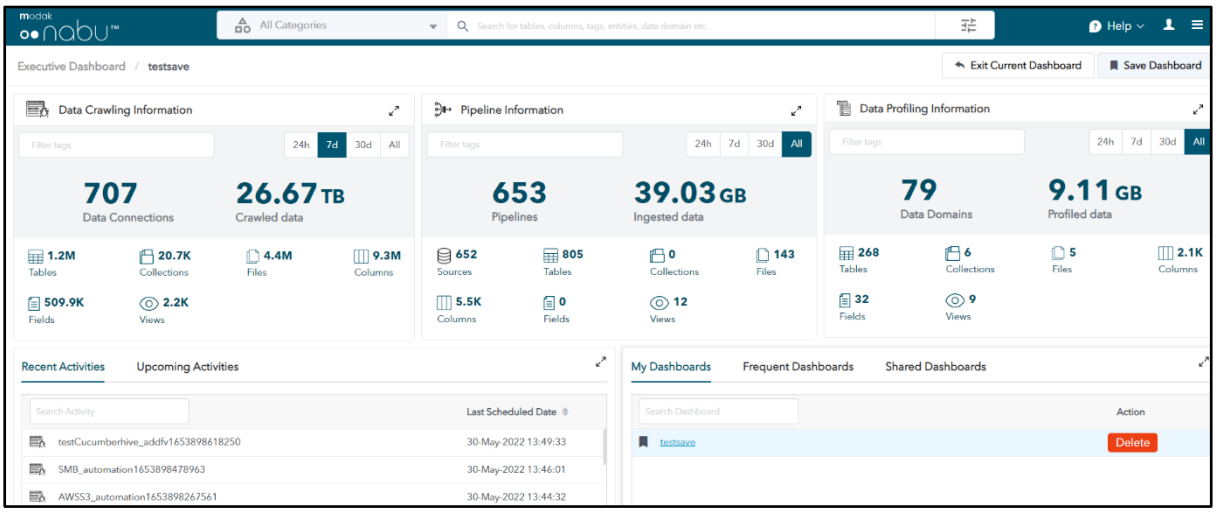


Figure 195: Executive Dashboard- Exit Current Dashboard

Recent and Upcoming Activities:

The user can view the Data crawling/Pipeline/Data Profiling activities in this section which is to the bottom left of the executive dashboard landing page.

Recent Activities: The below details are shown for the recent activities.

The name of the recent or active data connections/data domains/pipelines.

1. The last scheduled date of the above data connections/data domains/pipelines.

Upcoming Activities: The below details are shown for the upcoming activities.

1. The name of the upcoming data connections/data domains/pipelines. The upcoming activities are the activities that are scheduled for the future dates.
2. The next scheduled date of the above data connections/data domains/pipelines.

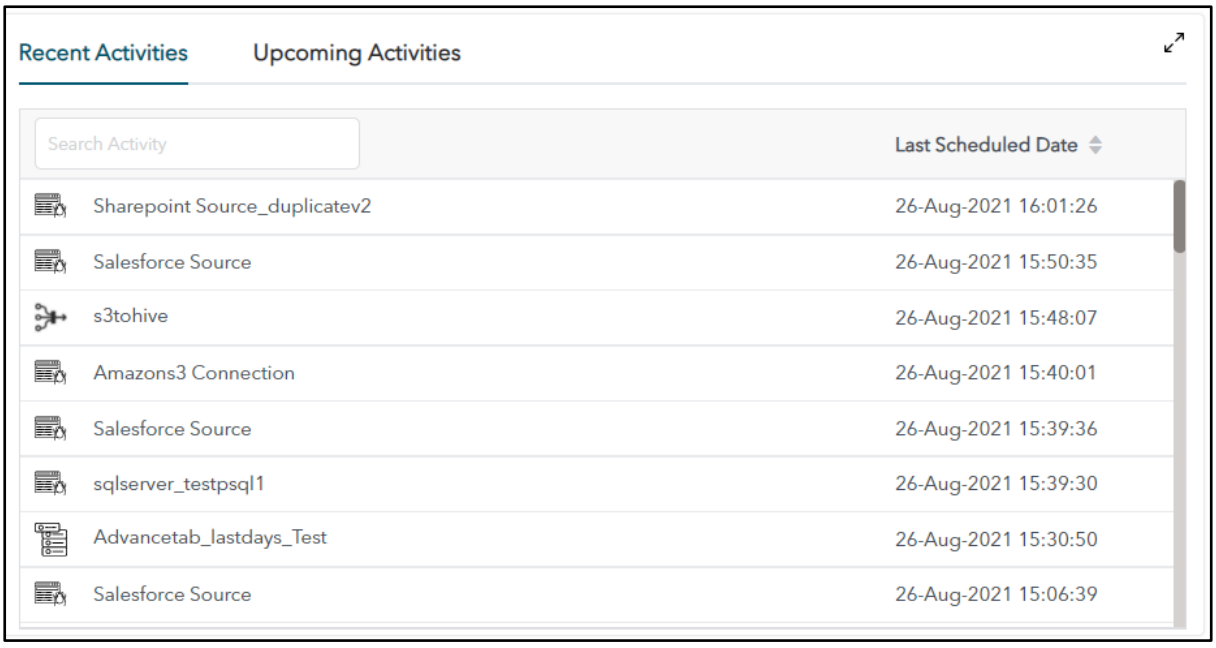


Figure 196: Recent Activities

This can be expanded by clicking on the icon on the top corner. On expanding, the recent and upcoming activities details can be viewed separately as below.

8 Access Management

The 'Access Management' feature in Modak Nabu 2.6 enables fine grain, access control over various functionalities, resources, data domains in Modak Nabu.

The 'Roles' functionality allows creation of roles, which define the access for UI functionalities and resources in Modak Nabu, for users added to the role.

The access to data domains is governed through 'Data Access' functionality.

8.1 Roles

Roles are created to define access to Modak Nabu functionalities (e.g., pipelines, data domains, compute engines etc.) and resources (specific pipelines, data domains, compute engines etc., that are created in Modak Nabu) in Modak Nabu. User ids, from an enterprise's LDAP or AD ID, which are added to a role, will have the level of access that is defined for the role.

For UI functionalities, the levels of access available are:

1. View – users can view the screens for that functionality on the Modak Nabu UI but cannot create/edit any resource.
2. Modify – users can modify resources. This is applicable for data catalogue entries (i.e., entity, facet, fieldstore, synonym, filter) and access management.
3. Create – users can create a resource using that functionality. For e.g., users with 'Create' access for pipelines, can create a new pipeline.

For resources created in Modak Nabu (i.e., credentials, compute engines, data connections, pipelines, data domains), the level of access available are as under:

1. View – Users can view details of the resource
2. Usage – In addition to 'View' permissions, Users can use the resource, but cannot modify any details. For e.g., a user with 'usage' permission on a pipeline can run the pipeline, but cannot modify the details of the pipelines
3. Modify – In addition to 'Usage' permissions, users have full access on the resource and can edit any details of the resource.

8.1.1 Create Roles

Roles can be created as required as defined by the requirement in an organization and assign members.

You can create a role by following the below steps.

1. Click on menu button "≡"at the top right of the screen, go to "Access management" and click on "Roles". You will be seeing the below screen.

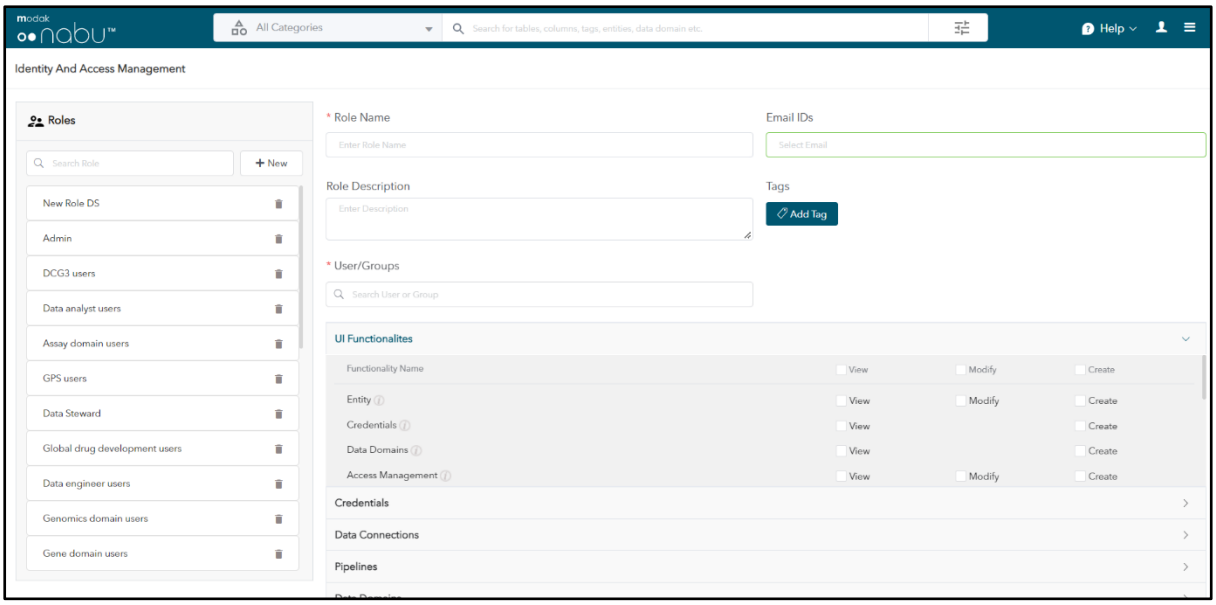


Figure 197: Create Role

2. Add ‘Role Name’, optional ‘Role Description’, creator ‘Email IDs’, and tags.
3. Add ‘Users/Groups’ that should be part of the role. These user ids and group names are obtained from the integration of enterprise’s LDAP/Azure AD.
4. Provide access to UI functionalities. Please refer section 8.1.1.1
5. Provide access to Modak Nabu resource. Please refer section 8.1.1.2
6. Click on “Create” button to create the role.

8.1.1.1 UI Functionalities

Under UI functionalities for a role, select the level of access to be provided to a role for each UI functionality.

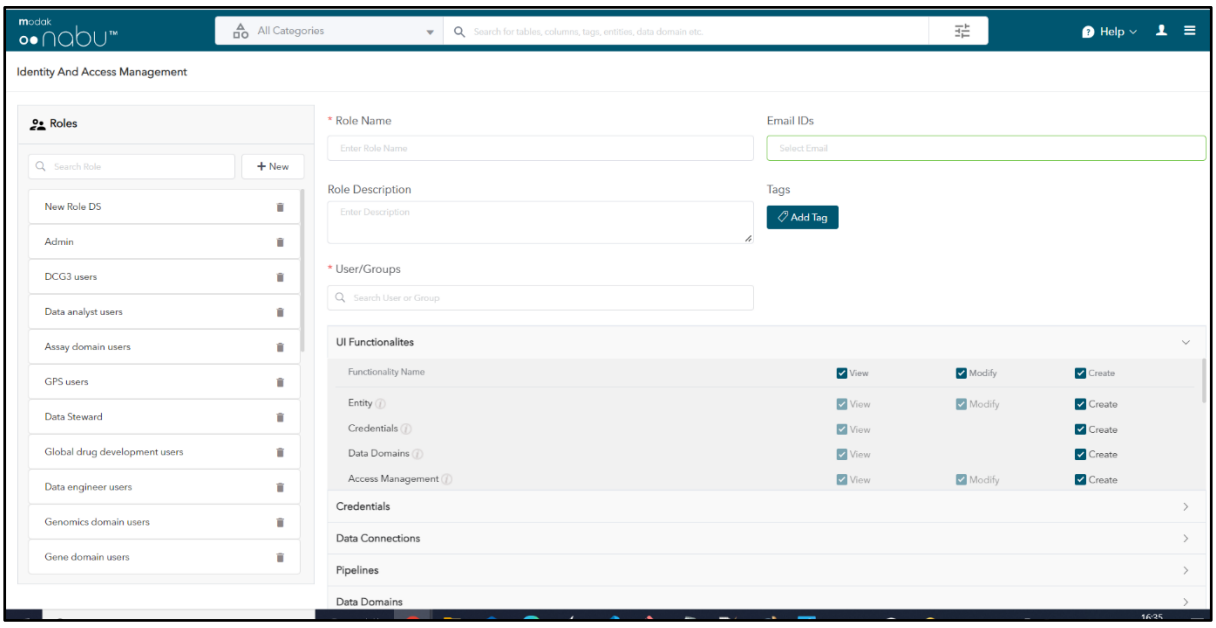


Figure 198: Roles Functionality

For UI functionalities, the levels of access available are:

- View – users can view the screens for that functionality on the Modak Nabu UI but cannot create/edit any resource.
- Modify – users can modify resources. This is applicable for data catalogue entries (i.e., entity, facet, fieldstore, synonym, filter) and access management.
- Create – users can create a resource using that functionality. For e.g., users with 'Create' access for pipelines, can create a new pipeline.

8.1.1.2 Access to resources in Modak Nabu

Access to different resources in Modak Nabu is provided as explained in the sections below.

Access for for Credentials: Includes access to specific credentials that are added in Modak Nabu to access a data source. The credentials are stored in a secure vault and are referred by their name in Modak Nabu.

The access to credentials is provided to a role, using:

a. Credential Name – Level of access defined on specific credential names. For e.g., for a credential name 'Oracle credentials', a particular role can be granted 'Usage' access. The user ids, who are assigned this role, will then be able to use these credentials while adding a data connection. These credentials are used by Nabu's data spiders to connect to the data source and crawl metadata from it.

b. Credential Tag – Level of access defined using tags attached to credentials. For e.g., for a credential tag 'Tenant: DCG3', a particular role can be granted 'Modify' access. In that case, user ids that are assigned this role can modify the details of the credential as well as use it for adding data connections

Figure 199 : IAM for credentials

The condition type can be Credential Name or Credential Tag.

Condition Type -> Credential Name:

- Credential name condition type with IS operator.

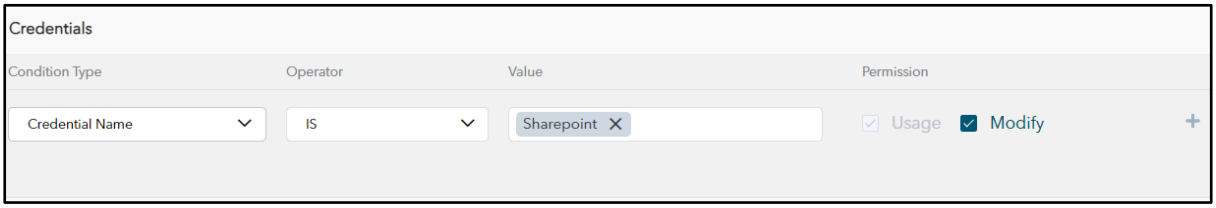


Figure 200 : IAM Credentials – Credential Name – IS Operator

- Condition type **credential name** with **IS operator**.
- Value is the credential name. Multiple credential names can be provided here.
- Permission can be Usage/Modify.

- Usage: The user with that role will have usage access to the credential names provided in the value field and the user can use the credentials for adding a data connection.
- Modify: The user with that role will have modify access to the credential names provided in the value field and can be used by the user to edit any details of credentials.

- Credential name condition type with like operator

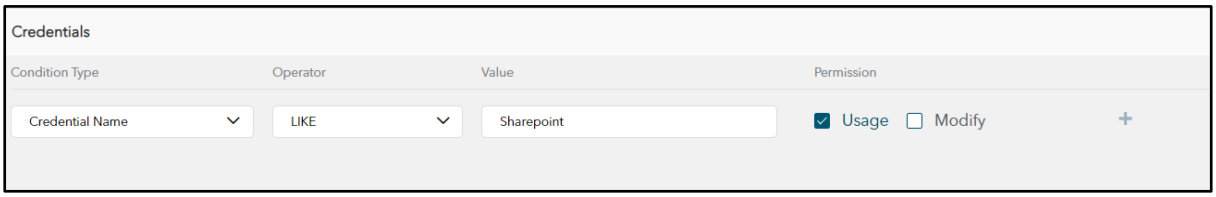


Figure 201 : IAM Credentials – Credential Name – LIKE Operator

- Condition type **credential name** with **LIKE operator**.
- Value is the specified pattern value that matches in the credential name. single pattern should be provided here.
- Permission can be Usage/Modify.
- Usage: The user with that role will have usage access to the credential pattern provided in the value field and the user can use the credentials for adding a data connection.
- Modify: The user with that role will have modify access to the credential pattern provided in the value field and can be used by the user to edit any details of those credentials.

Condition Type -> Credential Tag:

- Credential Tag condition type has only IS operator.

Figure 202 : IAM Credentials – Credential Tag – IS Operator

- Condition type **credential Tag** with **IS operator** and tag Environment: environment and permission **Modify**.
- Value is the tag with pattern tag category: tag value. Multiple Tags can be selected here.
- Permission can be Usage/Modify.
- Usage: The user with that role will have usage access to the credential with the provided tag in the value field and the user can use the credentials for adding a data connection.
- Modify: The user with that role will have modify access to the credential with the provided tag in the value field and the same credential can be used by the user to edit any details of those credentials

Access for Data Connections: Includes access to specific data connections that are added into Modak Nabu.

Access to data connections is provided to a role, using:

a. Names of data connections – Level of access defined on specific data connection names. For e.g., for a data connection name 'Clinical Trial 2022', a particular role can

be granted 'Modify' access. The user ids who are assigned this role, will then be able to change details of the data connection.

b. Tags attached to data connections – Level of access defined using tags attached to data connections. For e.g., for a data connection tag 'Tenant: GPS', a particular role can be granted 'View' access. In that case user ids who are assigned this role will have 'View' access on all data connections that have the tag 'Tenant: GPS'.

The condition type can be Data Connection Name or Data Connection Tag.

Figure 203 : IAM for Data Connection

Condition Type -> Data Connection Name:

- Data Connection Name condition type with IS operator.

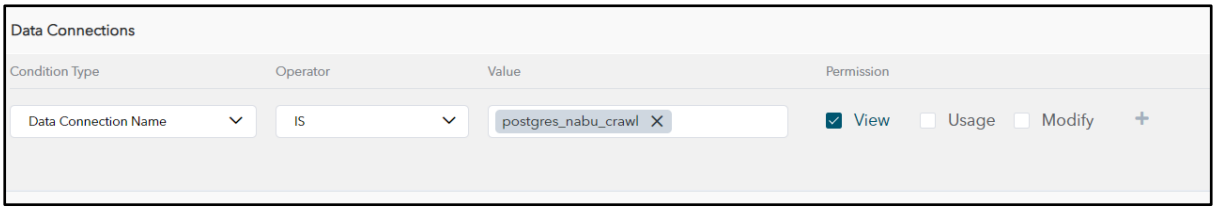


Figure 204 : IAM for Data Connection – Data Connection Name - IS Operator

  - Condition type Data Connection Name with **IS operator**.
  - Value is the Data connection name. Multiple Data Connection Names can be selected here.
  - Permission can be View/Usage/Modify.
  - **View**: If the permission is View, the user with that role can only view those data connection(s) provided in the value field and cannot perform any schedule/modify operations.
  - **Usage**: If the permission is Usage, the user with that role can view and schedule crawling of metadata from that data connection(s) provided in the value field and cannot perform any modify operations.
  - **Modify**: If the permission is Modify, the user with that role can view, modify and schedule crawling of metadata from that data connection(s) provided in the value field.

- Data Connection Name condition type with LIKE operator.

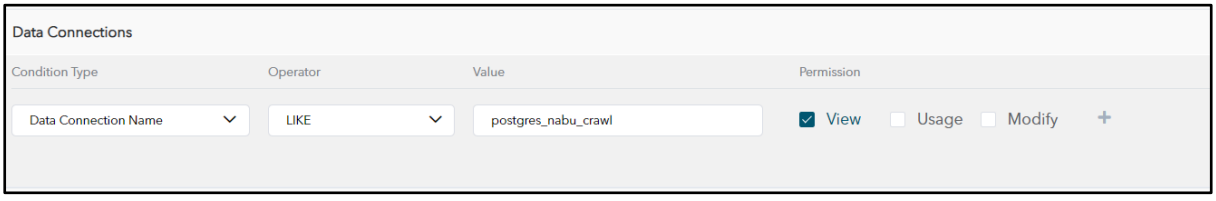


Figure 205 : IAM for Data Connection – Data Connection Name - LIKE Operator

  - Condition type Data Connection Name with **LIKE operator**.

- Value is the specified pattern value that matches in the Data Connection Name. single pattern should be provided here.
- Permission can be View/Usage/Modify.
- **View**: If the permission is View, the user with that role can only view that data connection(s) provided in the value field and cannot perform any schedule/modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and schedule crawling of meta data from that matched data connection(s) provided in the value field.
- **Modify**: If the permission is Modify, the user with that role can view, modify and schedule crawling of metadata from that matched data connection(s) provided in the value field.

Condition Type -> Data Connection Tag:

Data Connection Tag condition type has only IS operator.

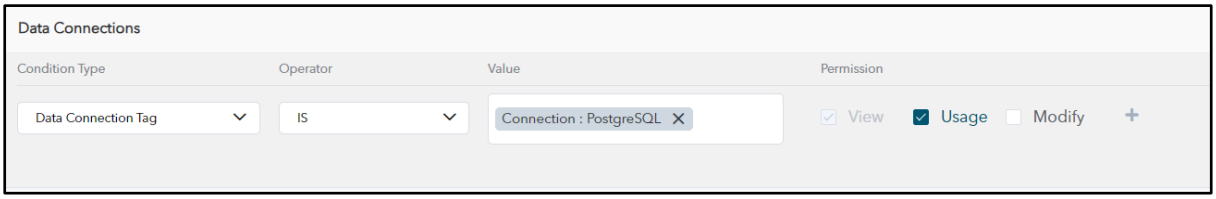


Figure 206 : IAM for Data Connection – Data Connection Tag - IS Operator

- Condition type **Data Connection Tag** with **IS operator**.
- Value is the tag with pattern tag category: tag value. Multiple Tags can be selected here.
- Permission can be View/Usage/Modify.
- **View**: If the permission is View, the user with that role can only view that data connection(s) with the specified tags and cannot perform any schedule/modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and schedule crawling of metadata from that data connection(s) with the specified tags and cannot perform any create/modify operations.
- **Modify**: If the permission is Modify, the user with that role can view, create, modify, and schedule crawling of metadata from that data connection(s) with the specified tags.

Access for Pipelines: The role can be granted or limit access at each Pipeline level.

Access to pipelines is provided to a role, using:

a. **Names of pipelines** – Level of access defined on specific pipeline names. For e.g., for a pipeline name 'Clinical Trial S3', a particular role can be granted 'Modify' access. The user ids who are assigned this role will be able to change details of the pipeline.

b. ***Tags attached to pipelines*** – Level of access defined using tags attached to pipelines. For e.g., for a pipeline tag 'Tenant: DCG', a particular role can be granted 'View' access. In that case user ids who are assigned this role will have 'View' access on all pipelines that have the tag 'Tenant: GPS'

The tags to a pipeline can be attached while creating the pipeline or can be added later by any user who is part of a role that has 'Modify' access on a pipeline.

The condition type can be Pipeline Name or Pipeline Tag.

Figure 207 : IAM for Pipeline

Condition Type -> Pipeline Name:

- Pipeline Name condition type with IS operator.

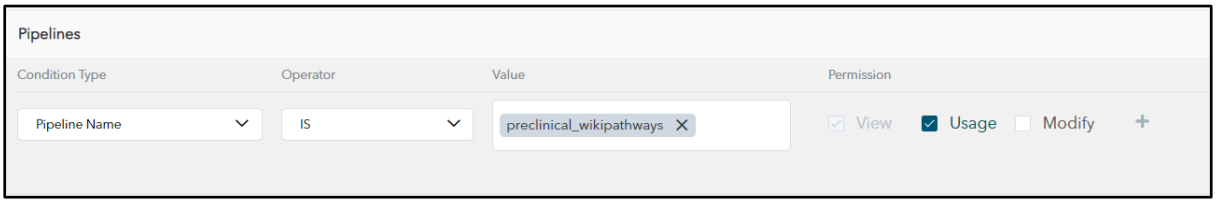


Figure 208 : IAM for Pipeline – Pipeline Name – IS Operator

  - Condition type Pipeline Name with **IS operator**.
  - Value is the Pipeline Name. Multiple Pipeline Names can be selected here.
  - Permission can be View/Usage/Modify
  - **View**: If the permission is View, the user with that role can only view those pipeline(s) provided in the value field and cannot perform any schedule/modify operations.
  - **Usage**: If the permission is Usage, the user with that role can view and schedule that pipeline(s) provided in the value field and cannot perform any modify operations.
  - **Modify**: If the permission is Modify, the user with that role can view, modify and schedule that pipeline(s) provided in the value field.

- Pipeline Name condition type with LIKE operator.

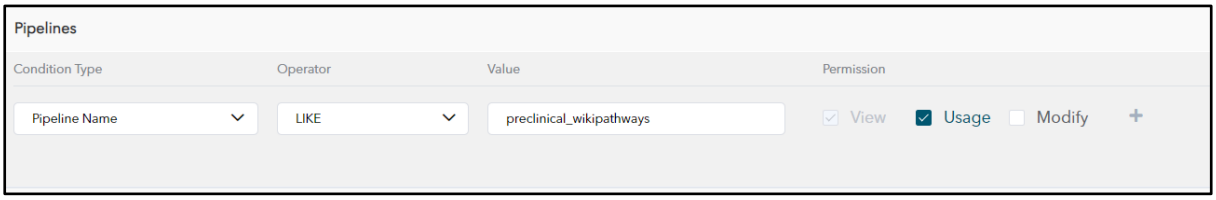


Figure 209 : IAM for Pipeline – Pipeline Name – LIKE Operator

  - Condition type Pipeline Name with **LIKE operator**.
  - Value is the specified pattern value that matches in the Pipeline Name. single pattern should be provided here.
  - Permission can be View/Usage/Modify.
  - **View**: If the permission is View, the user with that role can only view that pipeline(s) provided in the value field and cannot perform any schedule/modify operations.
  - **Usage**: If the permission is Usage, the user with that role can view and schedule that matched pipeline(s) provided in the value field and cannot perform any schedule/modify operations.
  - **Modify**: If the permission is Modify, the user with that role can view, modify and schedule that matched pipeline(s) provided in the value field.

Condition Type -> Pipeline Tag:

Pipeline Tag condition type has only IS operator.

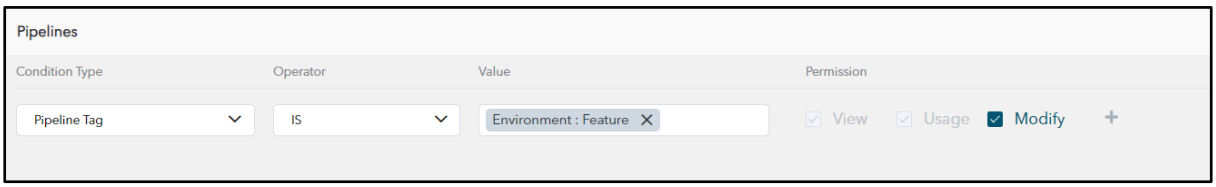


Figure 210 : IAM for Pipeline – Pipeline Tag – IS Operator

- Condition type **Pipeline Tag** with **IS operator**.
- Value is the tag with pattern tag category: tag value. Multiple Tags can be selected here.
- Permission can be View/Usage/Modify.
- **View**: If the permission is View, the user with that role can only view that pipeline(s) with the specified tags and cannot perform any schedule/modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and schedule that pipeline(s) with the specified tags and cannot perform any create/modify operations.
- **Modify**: If the permission is Modify, the user with that role can view, modify and schedule that pipeline(s) with the specified tags.

Access for Data Domains: The role can be granted or limit access at each Data Domain level.

The access to data domains is provided to a role, using:

a. ***Names of data domains*** – Level of access defined on specific data domain names. For e.g., for a data domain name 'Real World Data', a particular role can be granted 'Usage' access. The ids, who are assigned this role, will then be able to view the details of the data domain and schedule profile and indexing on this data domain.

b. ***Tags attached to data domains*** – Level of access defined using tags attached to data domains. For e.g., for a data domain tag 'Tenant: DCG1', a particular role can be granted 'View' access. In that case, user ids that are assigned this role will have 'View' access on all data domains that have the tag 'Tenant: DCG1'

The tags to a data domain can be attached while creating the data domain or can be added later by any user who is part of a role that has 'Modify' access on a data domain.

The condition type can be Data Domain Name or Data Domain Tag.

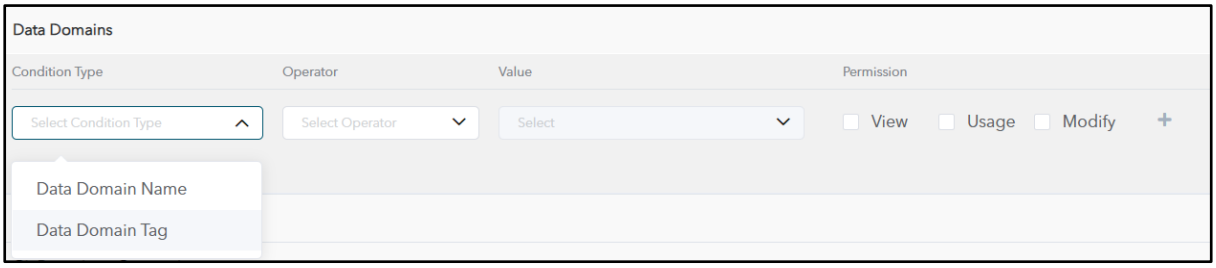


Figure 211 : IAM for Data Domain

Condition Type -> Data Domain Name:

- Data Domain Name condition type with IS operator.

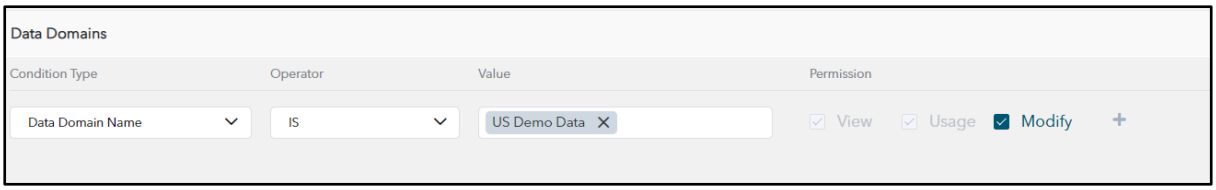


Figure 212 : IAM for Data Domain – Data Domain Name – IS Operator

- Condition type Data Domain Name with **IS operator**.
- Value is the Data Domain Name. Multiple Data Domain Names can be selected here.
- Permission can be View/Usage/Modify.
- **View**: If the permission is View, the user with that role can only view those data domain(s) provided in the value field and cannot perform any schedule/modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and schedule profiling and indexing for that data domain(s) provided in the value field and cannot perform any create/modify operations.
- **Modify**: If the permission is Modify, the user with that role can view, modify and schedule profiling and indexing for that data domain(s) provided in the value field.

- Data Domain Name condition type with LIKE operator.

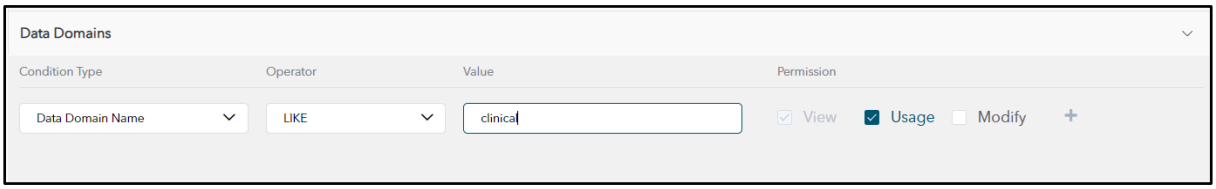


Figure 213 : IAM for Data Domain – Data Domain Name – LIKE Operator

- Condition type Data Domain Name with **LIKE operator**.
- Value is the specified pattern value that matches in the Data Domain Name. single pattern should be provided here.
- Permission can be View/Usage/Modify
- **View**: If the permission is View, the user with that role can only view that data domain(s) provided in the value field and cannot perform any create/modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and schedule that matched data domain(s) provided in the value field and cannot perform any create/modify operations.
- **Modify**: If the permission is Modify, the user with that role can view, create, modify and schedule that matched data domain(s) provided in the value field.

Condition Type -> Data Domain Tag:

Data Domain Tag condition type has only IS operator.

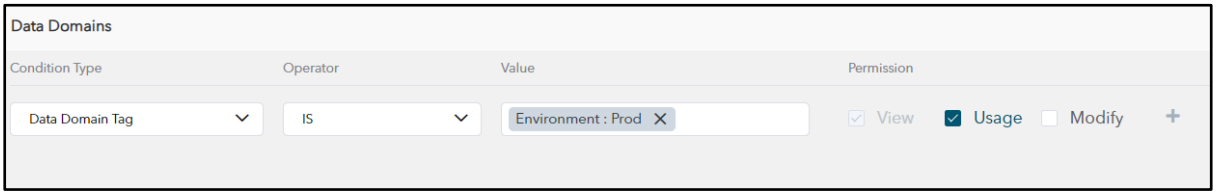


Figure 214 : IAM for Data Domain – Data Domain Tag – IS Operator

- Condition type **Data Domain** Tag with **IS operator**.

- Value is the tag with pattern tag category: tag value. Multiple Tags can be selected here.
- Permission can be View/Usage/Modify.
- **View**: If the permission is View, the user with that role can only view that data domain(s) with the specified tags and cannot perform any create/modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and schedule that data domain(s) with the specified tags and cannot perform any create/modify operations.
- **Modify**: If the permission is Modify, the user with that role can view, create, modify and schedule that data domain(s) with the specified tags.

Access for Compute Engines: The role can be granted or limit access at each compute engine level.

Access to data domains is provided to a role, using:

a. ***Names of compute engines*** – Level of access defined on specific compute engine names. For e.g., for a compute engine name 'CDE Compute Engine', a particular role can be granted 'Usage' access. The user ids, who are assigned this role, will then be able to view the details of the compute engine and use it for running pipelines, profiling, or indexing.

b. ***Tags attached to compute engines*** – Level of access defined using tags attached to compute engines. For e.g., for a compute engine tag 'Tenant: DCG2', a particular role can be granted 'Modify' access. In that case, user ids that are assigned this role can modify the details of the compute engine as well as use it for running pipelines, or profiling/indexing a data domain.

The tags for a compute engine can be attached while creating the compute engine or can be added later by any user who is part of a role that has 'Modify' access on a compute engine.

The condition type can be Compute Engine Name or Compute Engine Tag.

Figure 215 : IAM for Compute Engine

Condition Type -> Compute Engine Name:

- Compute Engine Name condition type with IS operator.

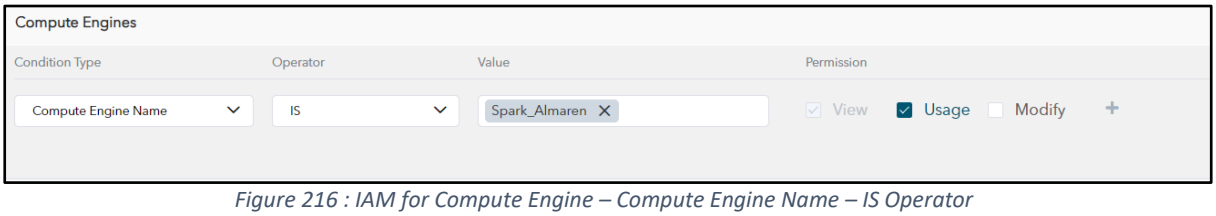


Figure 216 : IAM for Compute Engine – Compute Engine Name – IS Operator

- Condition type Compute Engine Name with **IS operator**.
- Value is the Compute Engine Name. Multiple Compute Engine Names can be selected here.
- Permission can be View/Usage/Modify

- **View**: If the permission is View, the user with that role can only view those compute engine(s) provided in the value field and cannot perform any modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and can configure the compute engine provided in the value field for running a pipeline or for schedule profiling, indexing for a data domain and cannot modify any details for the compute engine.
- **Modify**: If the permission is Modify, the user with that role can view, modify the details of compute engine(s) provided in the value field and can configure that compute engine to run pipeline or profiling or indexing schedule that.

- Compute Engine Name condition type with LIKE operator.

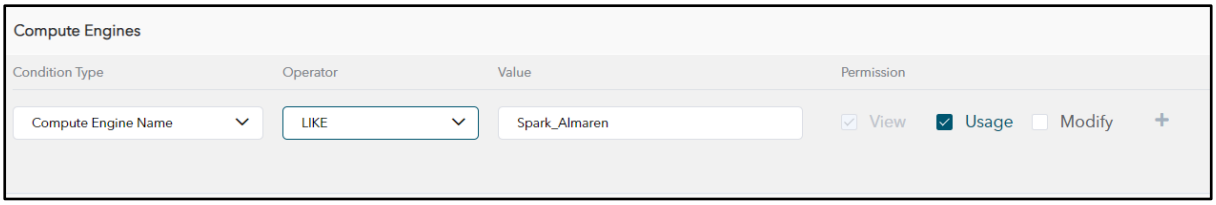


Figure 217 : IAM for Compute Engine – Compute Engine Name – LIKE Operator

- Condition type Compute Engine Name with **LIKE operator**.
- Value is the specified pattern value that matches in the Compute Engine Name. single pattern should be provided here.
- Permission can be View/Usage/Modify
- **View**: If the permission is View, the user with that role can only view that compute engine(s) provided in the value field and cannot perform any modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and can configure the compute engine provided in the value field for running a pipeline or for running profiling, indexing for a data domain and cannot modify details of compute engine.
- **Modify**: If the permission is Modify, the user with that role can view, modify the details of compute engine(s) provided in the value field and can configure the compute engine for a pipeline or for running profiling, indexing for a data domain.

Condition Type -> Compute Engine Tag:

Compute Engine Tag condition type has only IS operator.

Figure 218 : IAM for Compute Engine – Compute Engine Tag – IS Operator

- Condition type **Compute Engine Tag** with **IS operator**.
- Value is the tag with pattern tag category: tag value. Multiple Tags can be selected here.
- Permission can be View/Usage/Modify.
- **View**: If the permission is View, the user with that role can only view that compute engine(s) with the specified tags and cannot perform any modify operations.
- **Usage**: If the permission is Usage, the user with that role can view and can configure the compute engine provided with the specified tags for a pipeline or for running profiling, indexing for a data domain and cannot perform any modify operations.

- **Modify**: If the permission is Modify, the user with that role can view, modify the details of compute engine(s) provided with the specified tags and can configure the compute engine for running a pipeline or for running profiling, indexing for a data domain.

8.1.2 Edit Role

1. Click on menu button "≡"at the top right of the screen, go to "Access management" and click on "Roles". You will be redirect to the below screen.

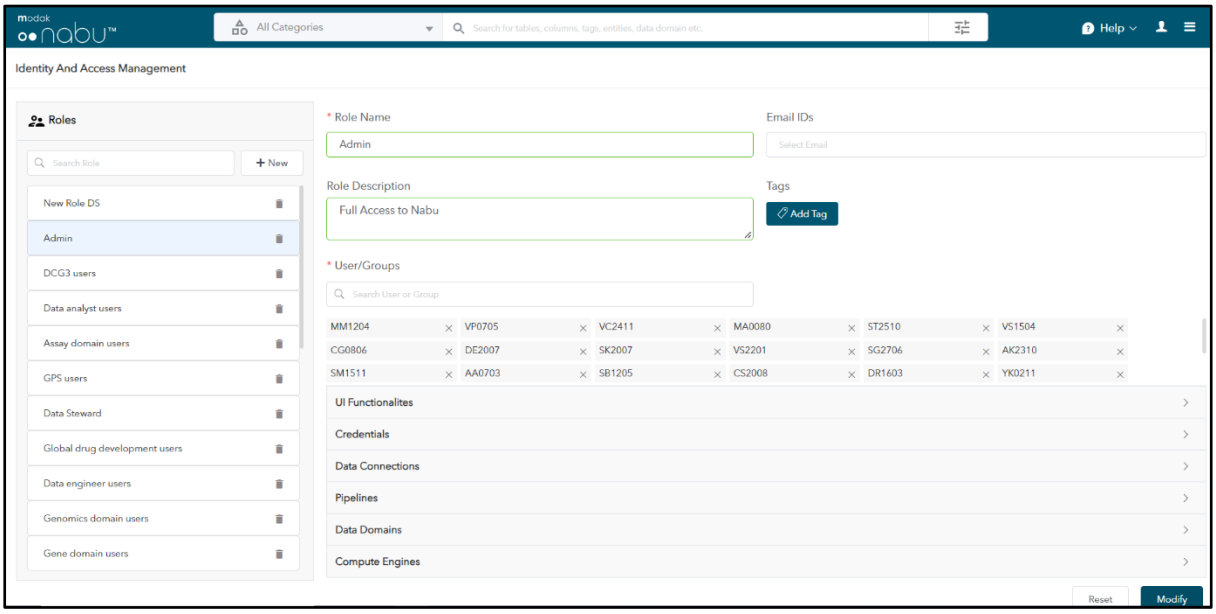


Figure 219: Edit Role Page

2. Select the required role from the left pane or search with the name in the search box for faster results.
3. Once you click on the required role, you can edit it on the right pane where all the information about the role is displayed.

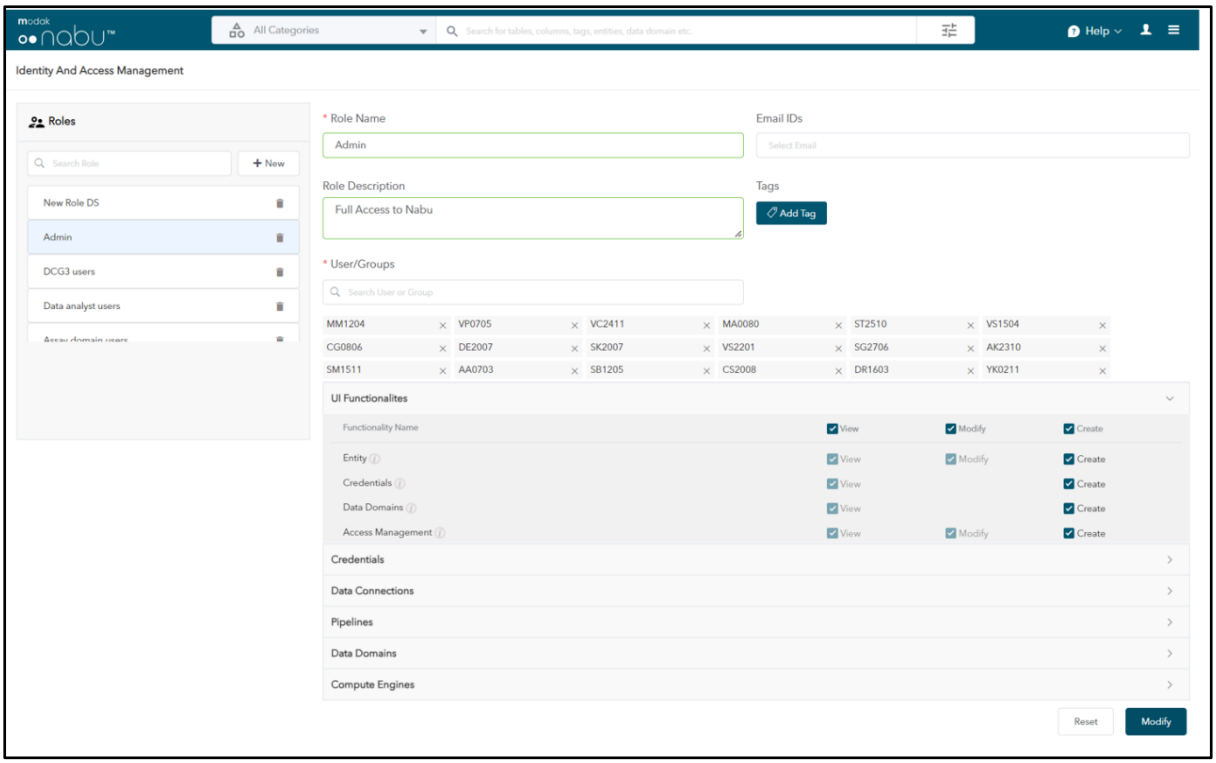


Figure 220: Edit Roles

4. Once all the required changes are made click on “modify” button to fix the changes.

<u>Navigate to Access management -> Roles</u>

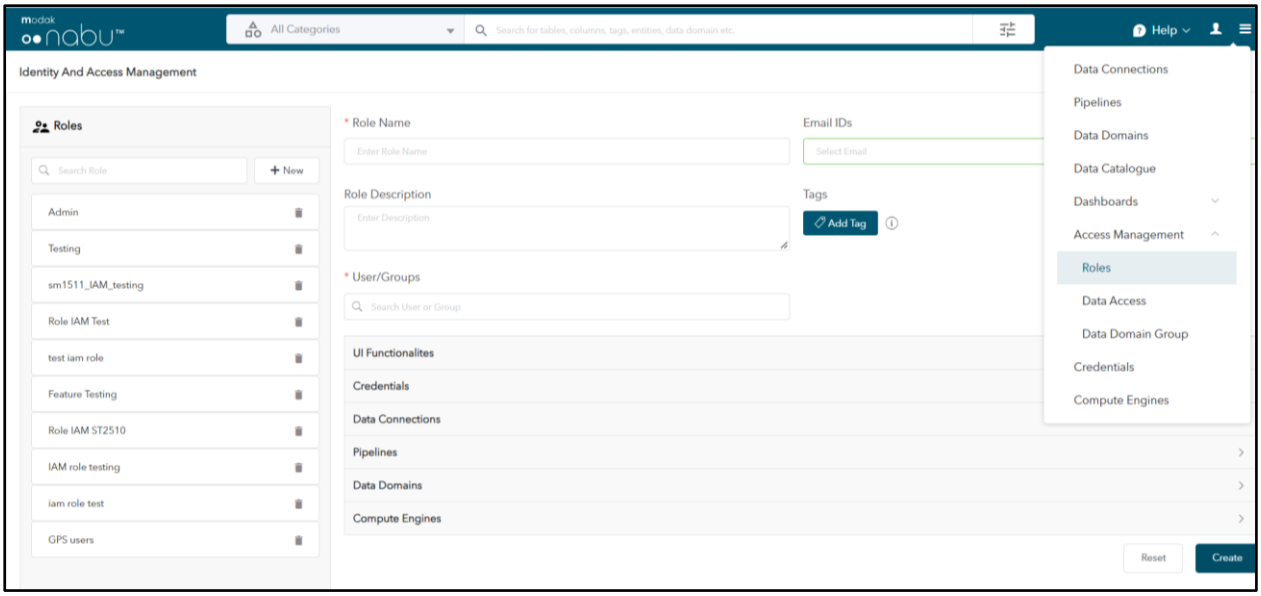


Figure 221 : Navigation of Roles

The user can perform any operations on Nabu based on the type of access that role has for that functionality/module.

<u>Ex: If the role has only view access for any functionality, the user can only view and cannot perform any add/modify operations for that functionality</u>.

The permission for the role can be controlled or granted for each module as below.

8.1.3 Impact of Roles Based Access Management

Access Management on Data Connections:

The below are the changes on Data Connections dashboard and Data connection forms with IAM.

- The user can view only those data connections for which the user has at least **view permission.**
    - The schedule, delete buttons will be disabled on dashboard and the user cannot perform any of those operations.

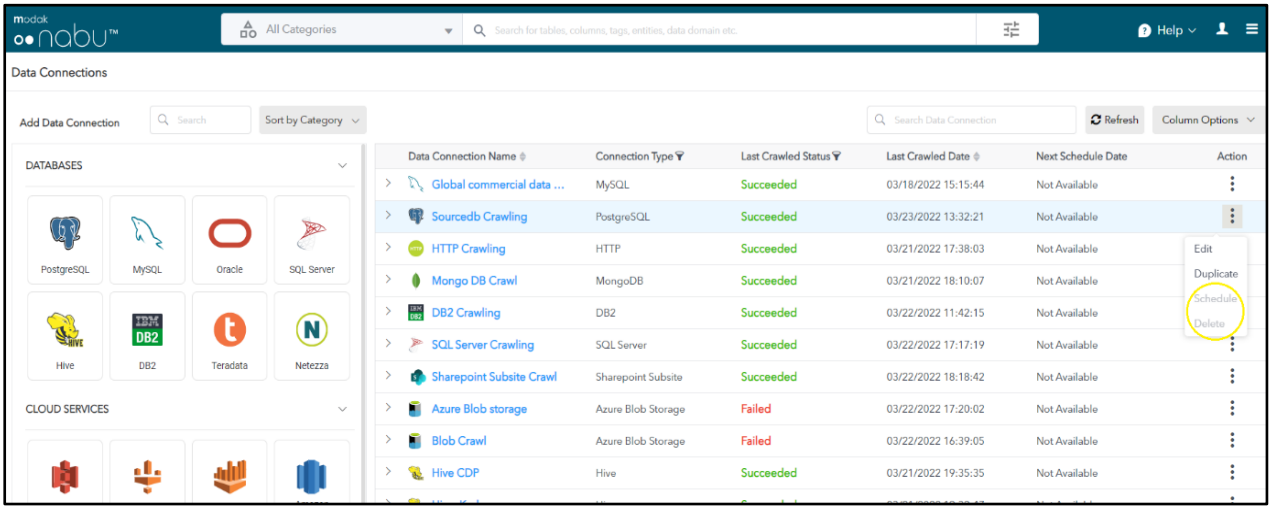


Figure 222 : Data Connection Dashboard – View Permission

The user can duplicate the data connection with view access.

- When user clicks on edit data connection, the user can only view the details and the modify button will be disabled in the data connection form as below

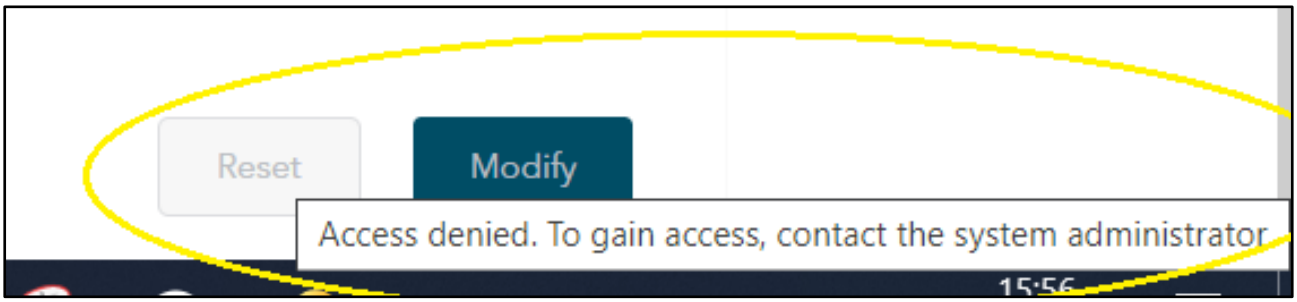


Figure 223 : Access denied for Modify Data Connection – View permission

- The user can view, duplicate and schedule crawling of metadata from those data connection(s) from the Data connections dashboard page for which the user has **Usage Permission**.
    - The delete button will be disabled and the user cannot delete the data connection.
    - The user needs to raise and gain access for those data connections

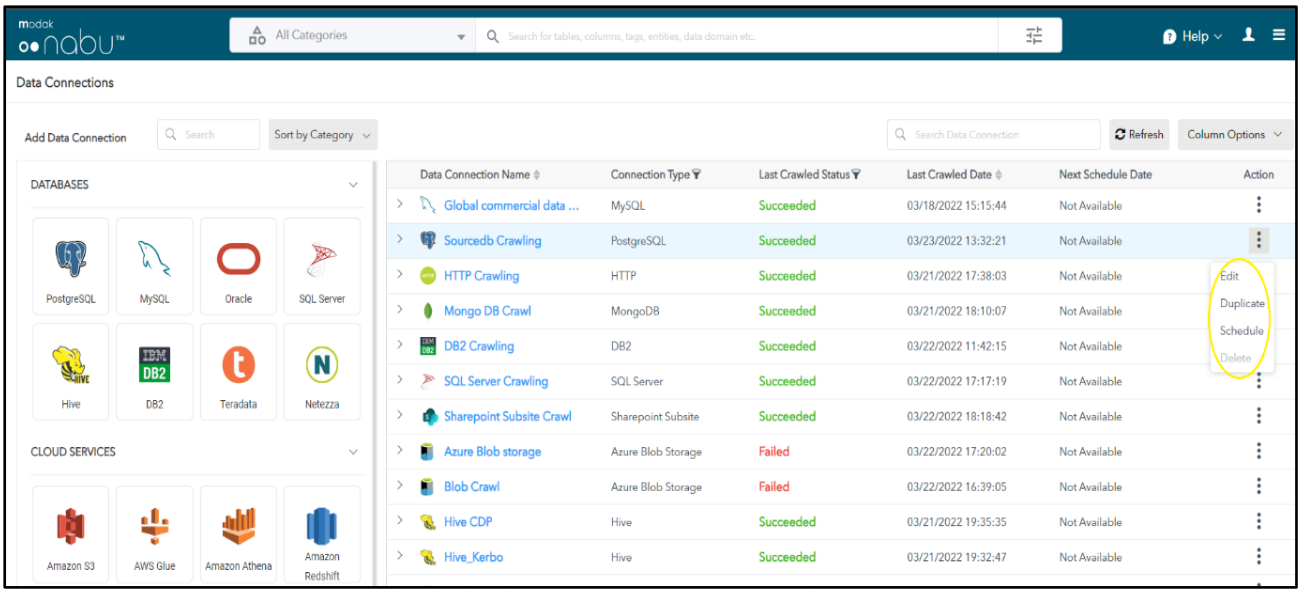


Figure 224 : Data Connection Dashboard – Usage Permission

- The user can view, duplicate, edit and schedule crawling of metadata from those data connection(s) on the Data connections dashboard page for which the user has **Modify Permission**.
- All the buttons are enabled for modify permission. The user can perform all the operations for those data connections.

NOTE:

- The credentials dropdown on the data connections page will have only those credentials for which the user has Usage access.
- For only view access, the user will be on 'View Mode' for that data.

Access Management on Pipelines:

The below are the changes on **Pipelines dashboard and Pipelines** pages.

- The user can view and duplicate those Pipelines for which the user has at least **view permission**.
  - The schedule, delete buttons on the dashboard will be disabled and the user cannot perform any of those operations.
  - The user needs to raise and gain access for those pipelines.

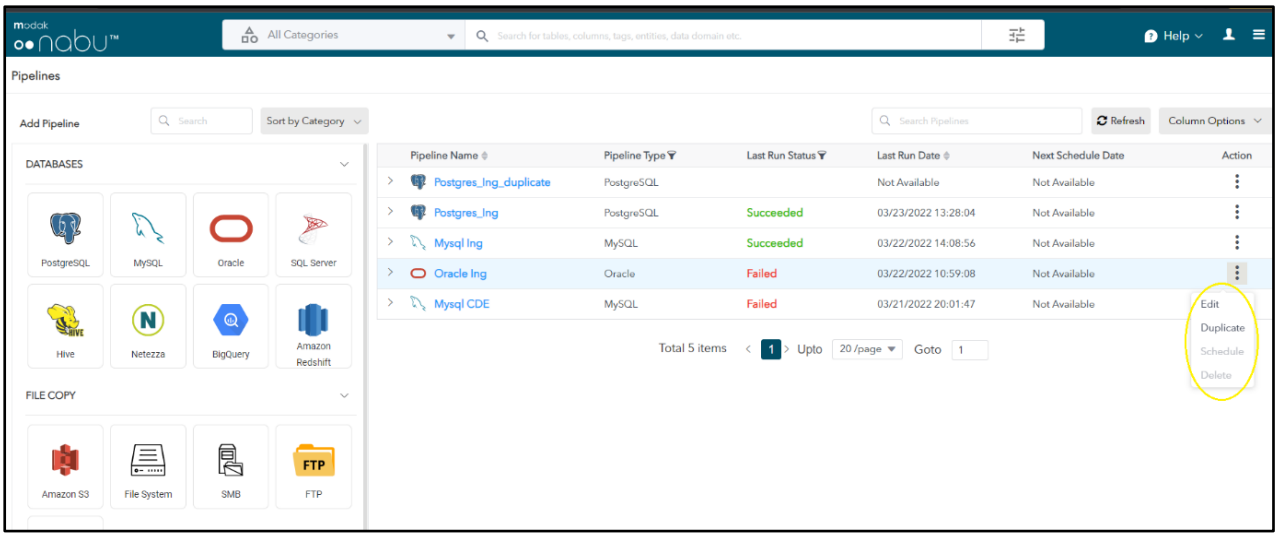


Figure 225 : Pipelines Dashboard – View Permission

- The user can view, duplicate and schedule those pipeline(s) from the Pipelines dashboard page or the preview section of pipelines for which the user has **Usage Permission**.
    - The delete button will be disabled and the user cannot delete the pipeline.
    - The user needs to raise and gain access for those pipelines

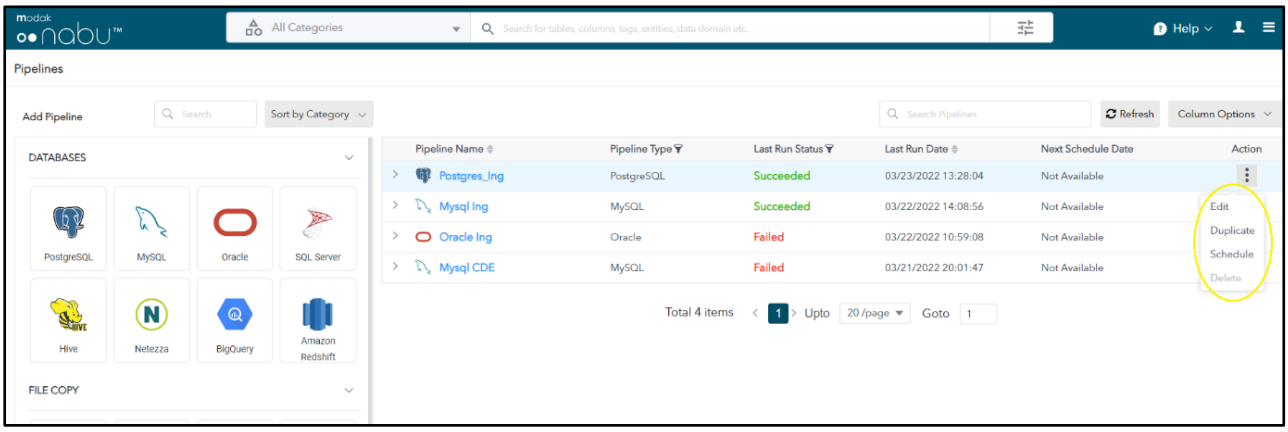


Figure 226 : Pipelines Dashboard – Usage Permission

- The user can view, edit, duplicate and schedule those pipeline(s) from the pipeline dashboard page or in edit mode of pipeline for which the user has **Modify Permission**. The user can also create new pipeline.
    - All the buttons are enabled for modify permission. The user can perform all the operations for those pipelines.

NOTE:

- For only view access, the user will be on 'View Mode' for that pipeline and the user cannot edit or schedule those pipeline(s).

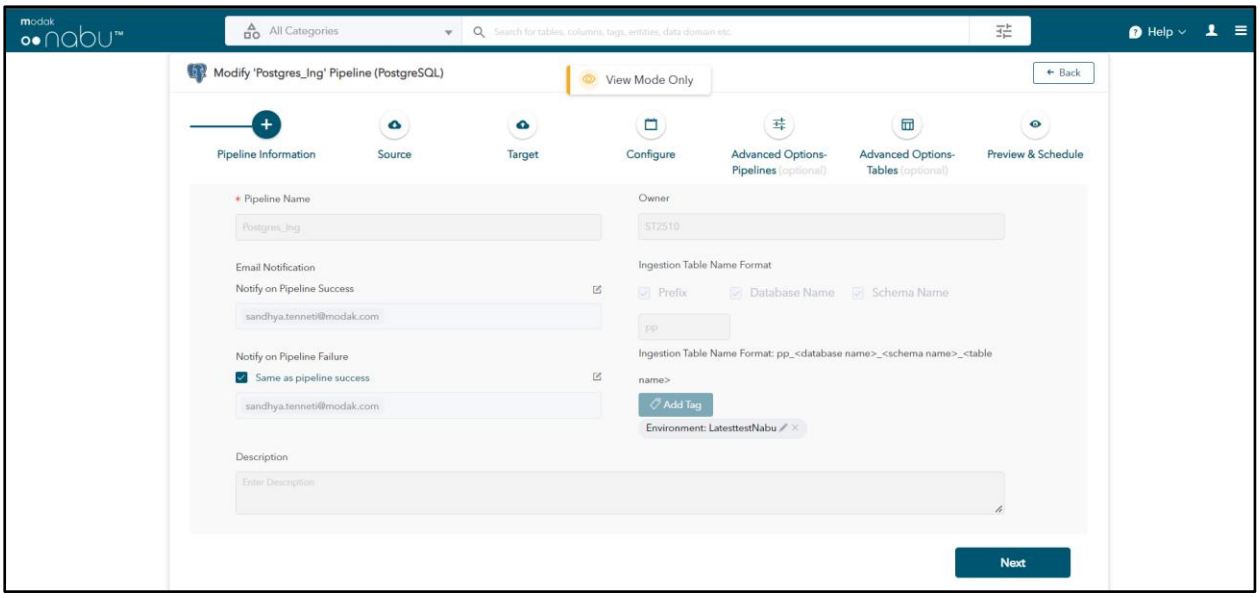


Figure 227 : Modify Pipeline – View Mode

- The source and destination data connections list will have those for which the user has Usage access.
- With View access, the user can duplicate a pipeline.
- The compute engine list will have those compute engines for which the user has Usage access.
- **source/destination data connection deleted**: For any deleted source or destination data connections, the user cannot perform any operations except delete. The user will be shown a message 'All sources are deleted for this pipeline' for source deleted scenario. The user cannot even view those pipelines.
- **Compute engine deleted**: For any pipeline where the compute engine is deleted, the user will be shown a message that 'compute engine has been deleted for this pipeline'

Access Management on Data Domains:

The below are the changes on Data Domains dashboard and Data Domain pages.

- The user can view and duplicate those data domains for which the user has at least view permission.
  - The schedule, delete buttons will be disabled and the user cannot perform any of those operations.
  - The user needs to raise and gain access for those data domains.
- The user can view, duplicate and schedule profiling, indexing for those data domain(s) from the Data Domains dashboard page or the preview section of data domains for which the user has Usage Permission
- The delete button will be disabled and the user cannot delete the data domain.
  - The user needs to raise and gain access for those data domains

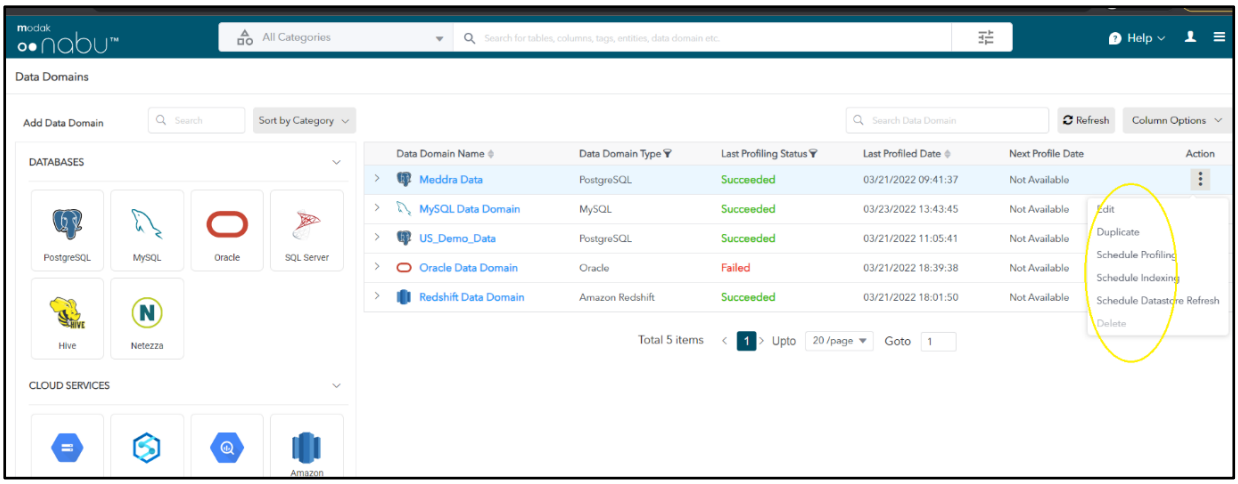


Figure 228 : Data Domain Dashboard – View Permission

- The user can view, edit, duplicate and schedule profiling, indexing those data domain(s) from the data domains dashboard page or in edit mode of data domain for which the user has Modify Permission. The user can also create new data domain.
    - All the buttons are enabled for modify permission. The user can perform all the operations for those data domains.

NOTE:

- For only view access, the user will be on 'View Mode' for that data domain and the user cannot edit or schedule the particular data domain

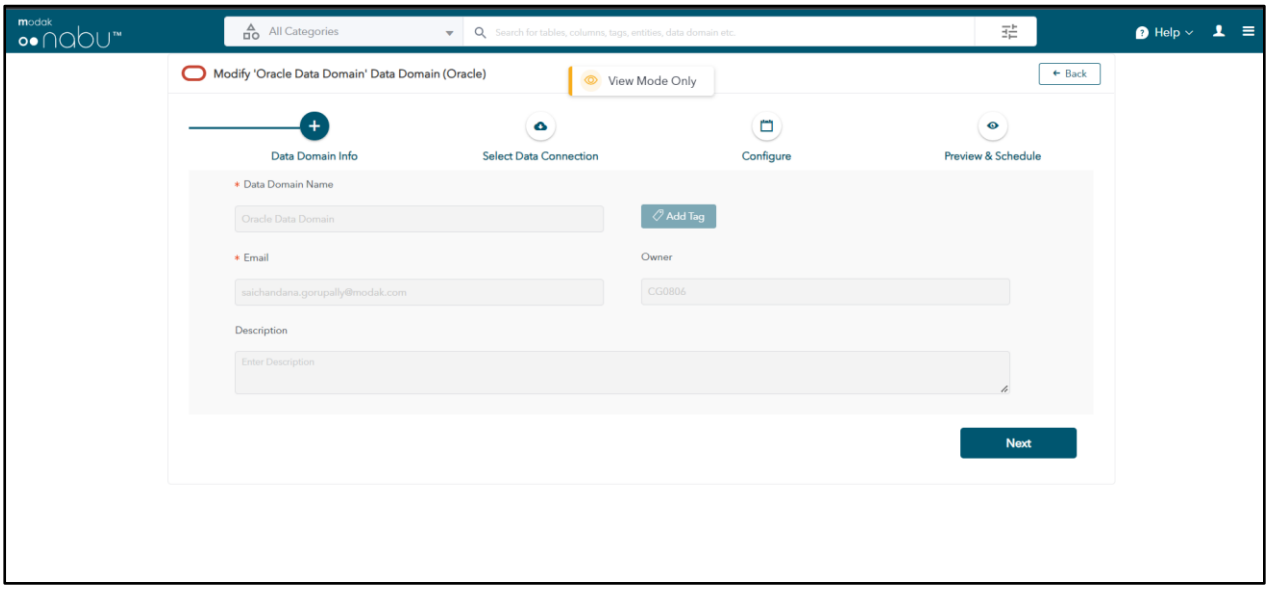


Figure 229 : Modify Data Domain – View Mode

- The source data connections list will have those for which the user has Usage access.
- The compute engine list will have those compute engines for which the user has Usage access.

Impact of Access Management on Monitoring Dashboard:

With IAM, the below are the changes on **Monitoring Dashboard** page.

- The user can view only those pipelines for which the user has at least view permission. The user cannot retry or reschedule any pipeline with View only access.

- For any failed objects, the user can retry or reschedule the pipeline only if the user has Usage access

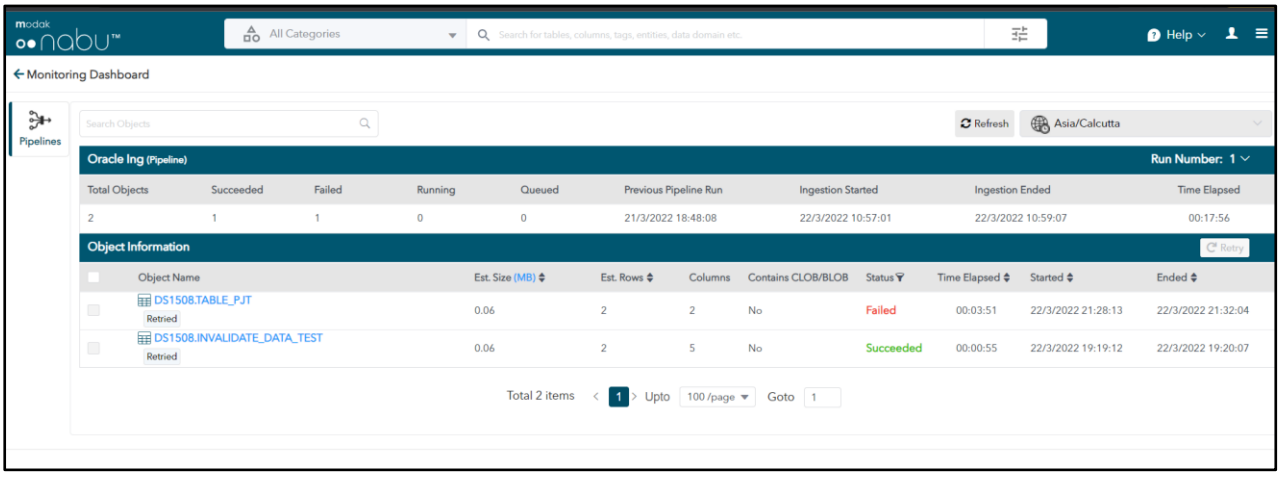


Figure 230 : IAM for monitoring Dashboard

Access Management on Credentials Page:

With IAM, the below are the changes on Credentials Page.

- **View Only Access:** The credentials list on the left pane will have only those credentials for which the user has at least view access.
  - The delete icon will be disabled for those credentials.
  - When clicked on the credential with view only access, the user can view the details but cannot modify the credential. The modify and reset buttons will be disabled

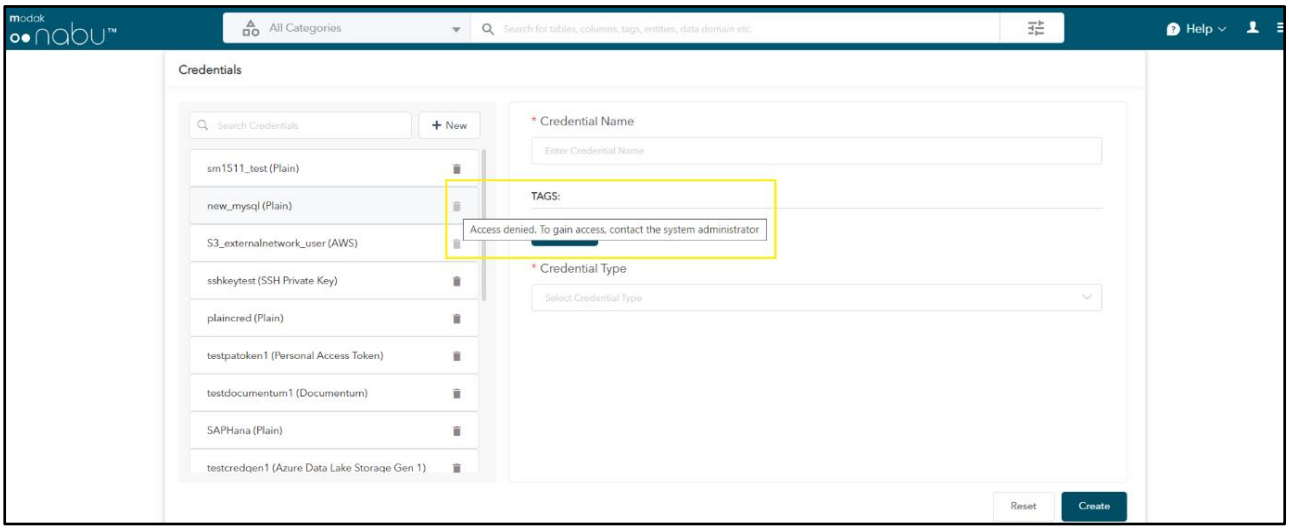


Figure 231 : IAM For Credentials – View Permission

- **Usage Access:** The credentials with the usage access will be part of the credentials dropdown for the data connections. The user can schedule any data connection with the usage access credentials.

NOTE: The user should also have Usage access for the data connection.

- **Modify Access:**
  - The delete icon will be enabled for those credentials.

- When clicked on the credential with modify access, the user can view the details and modify the credential. The modify and reset buttons will be enabled

Access Management for Compute Engines:

With IAM, the below are the changes on Compute Engines Page.

- **View Only Access:** The compute engines list on the left pane will have only those engines for which the user has at least view access.
  - The delete icon will be disabled for those compute engines.
  - When clicked on the compute engine with view only access, the user can view the details but cannot modify the compute engine. The modify and reset buttons will be disabled.

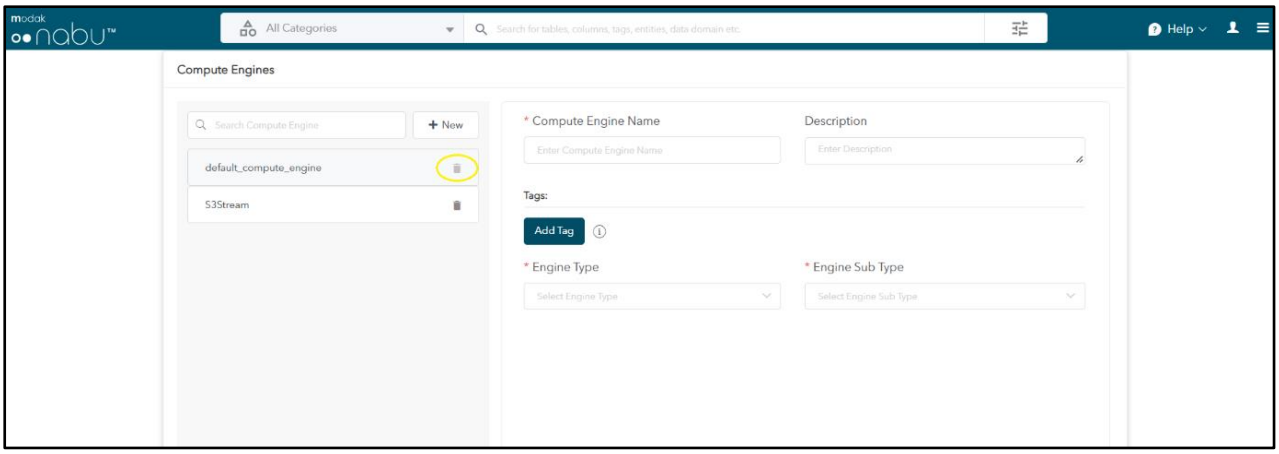


Figure 232 : IAM For Compute Engine – View Permission

- **Usage Access:** The compute engines with the usage access will be part of the compute engine list for the pipelines and data domains. The user can schedule any pipeline or data domain with the usage access compute engine.
- **Modify Access:**
  - The delete icon will be enabled for those compute engines.

When clicked on the compute engine with modify access, the user can view the details and modify the compute engine. The modify and reset buttons will be enabled

8.2 Data Access

Data access helps in keeping the data available for the required members of an organization, this makes sure only a specific member is performing operations on the permitted data.

Data Access can be given to Data domains (which are created in “Data domain”) and Data Domain Group (which are created in the “Access Management -> Data domain group”). Follow below steps to manage data access.

1. Click on menu button “≡”at the top right of the screen, go to “Access management” and click on “Data Access”. You will be redirect to the below screen.

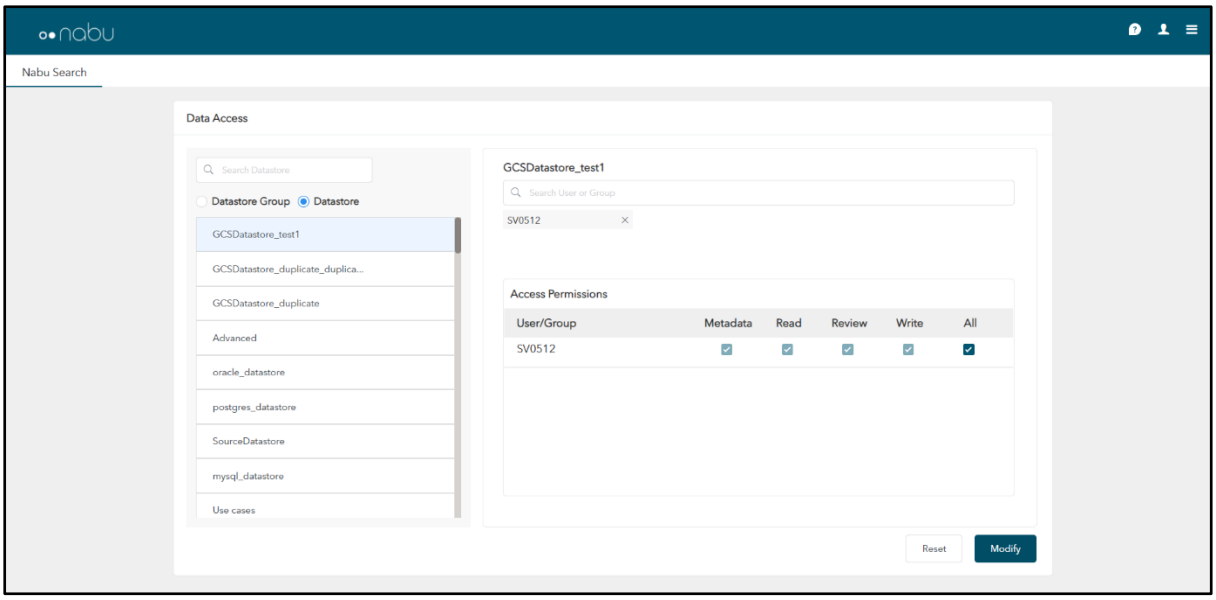


Figure 233: Data Access

2. Select a Data domain or group with the help of radio button.
3. Select the required Data domain or Data domain Group from the left pane and you will see the information about it on the right pane.
4. For the selected data domain/data domain group, users/groups can be added for access. These users/groups are available through Modak Nabu's integration with enterprise LDAP/AD ID.
5. For each user/group, the level of access provide to a data domain can be varied. The levels of access available are as under:
    a. 'Metadata' - only metadata (such as table names, column names) from the data domain can be viewed through the search functionality.
    b. 'Read' – Sample data of a dataset can be viewed by the user.
    c. 'Review' – Users can add metadata to a dataset, in the form of 'comments', 'tags', 'descriptions'.
    d. 'Write' – Users can edit the definition of the data domain.
    e. 'All' – Users have all the privileges on a data domain.
6. Make the necessary changes in access permissions or access and click on "modify" to apply changes to a Data domain or Data domain Group.

8.3 Data Domain Group

Data domain Groups help in providing accesses for a group of data domains at once to the required members. This makes an easy way to organise the permissions given to a member or a team.

8.3.1 Create Data Domain Group

To create a Data Domain Group, follow the given steps.

1. Click on menu button “ ”at the top right of the screen, go to “Access Management” and click on “Data Domain Group”. You will be redirect to the below screen.

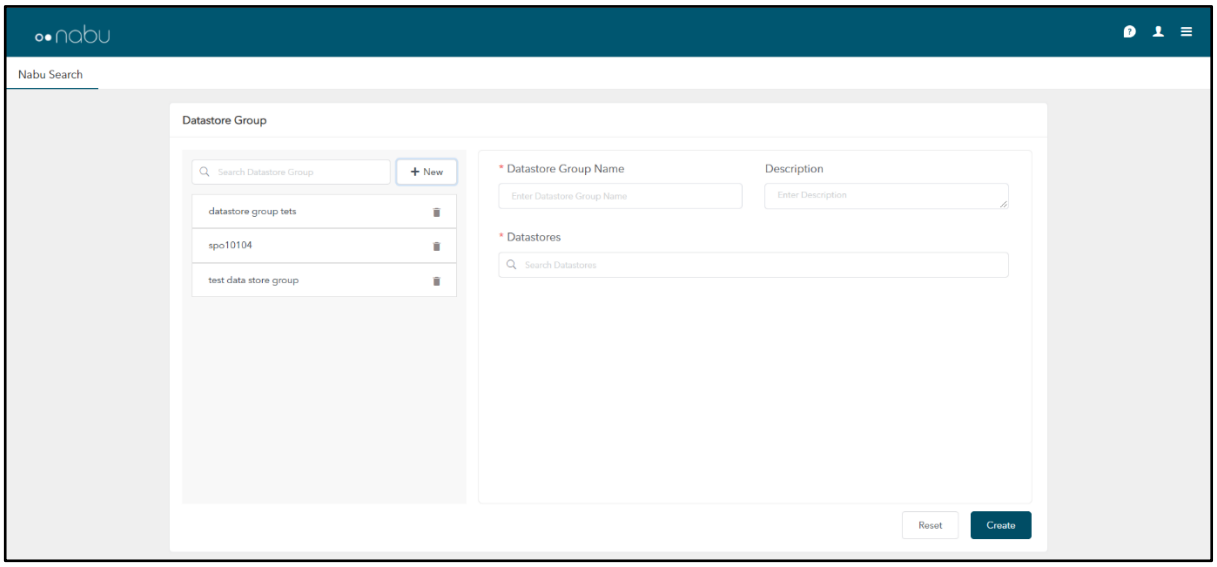


Figure 234: Data domain Group

2. Enter the fields like Data domain Group Name, Description of Group and Select the required Data domains from the dropdown.
3. Click on “Create” to create a Data domain group with the required data domains in it.

8.3.2 Edit Data domain Group

To edit a Data domain Group, follow below steps.

1. Click on menu button “ ”at the top right of the screen, go to “Access management” and click on “Data domain Group”. You will be redirect to the below screen.

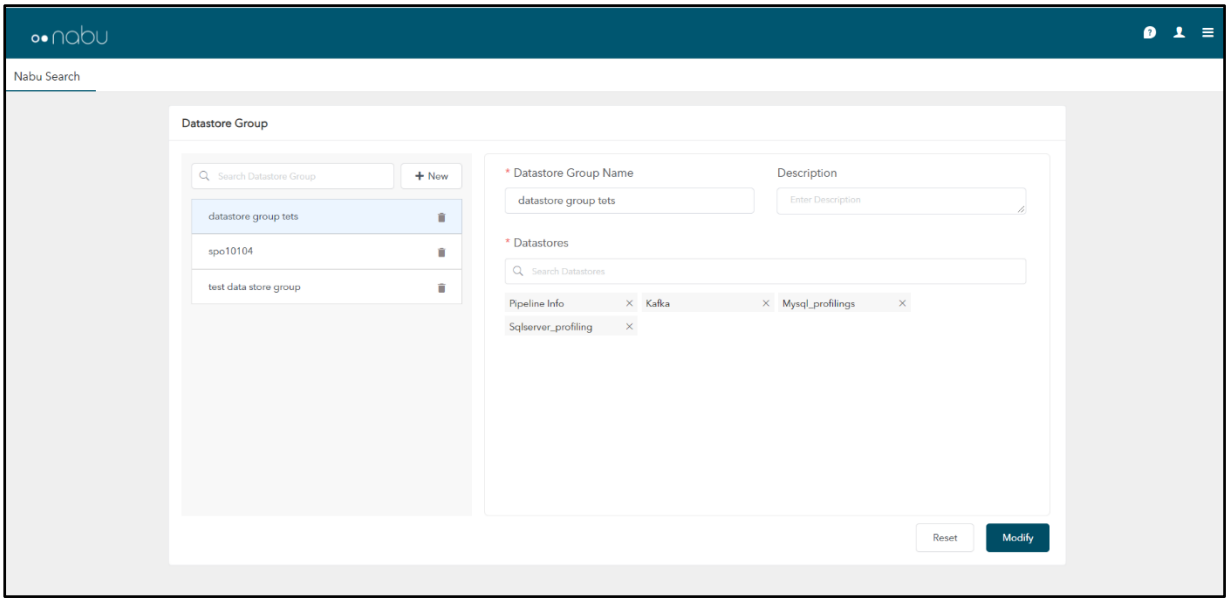


Figure 235: Edit Data domain Group

2. Select the required group from the left pane, then all the information about the group is displayed at the right pane.

3. Make the required changes in the selected group and click on “Modify” this will make sure your changes are applied.

9 Credentials

You can access Credentials by clicking on the menu button on the top right of the page.

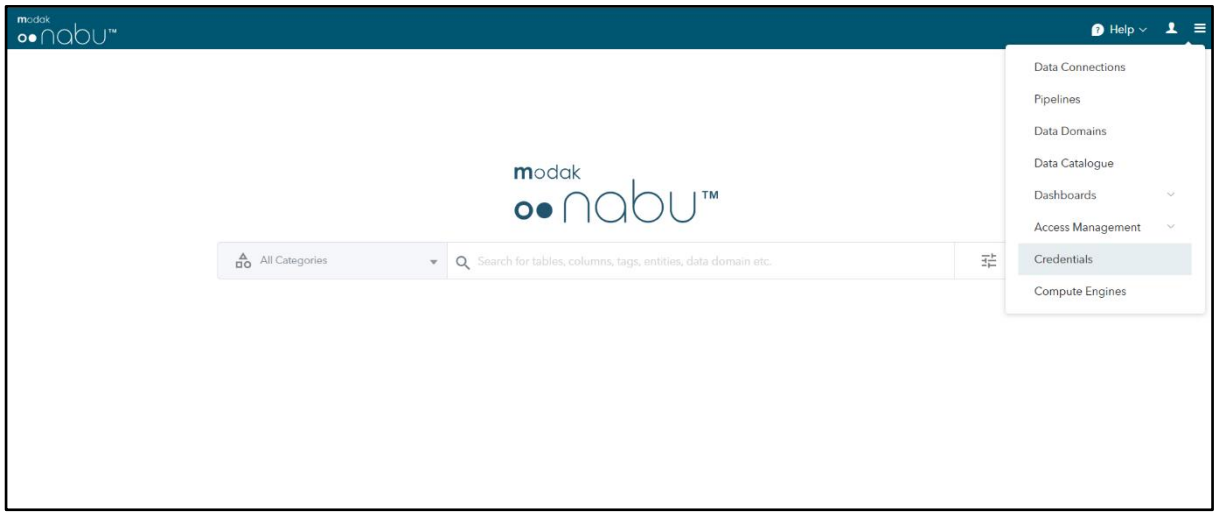


Figure 236: Credentials Menu

The credentials to connect to data connections are defined in this screen. The credential types supported include LDAP, Kerberos, AWS, GCP etc.

9.1 Adding new Credential

There are two basic fields which you need to fill to create a credentials i.e.

Credential Name: Where you can any name through which you can easily identify a credential

Credential Type: You need to select a type of credential from options like Plain, Aws, Kerberos, LDAP, GCP, Kerberos Truststore, Azure Blob Storage, Azure Data Lake Storage Gen 1, Azure Data Lake Storage Gen 2.

Depending on the type of credential you choose the following fields will be displayed.

Credential Type	Required Credentials to create
Plain	Username and Password
AWS	Secret Access Key and Access Key ID
Kerberos	Principle and Keytab
LDAP	Username and Password
GCP	Type, Project ID, Private Key ID, Private Key, Client Email, Client ID, Auth URI, Token URI, Auth Provider and Client Certificate.
Kerberos Truststore	Principal, Keytab, Truststore path, Truststore password
Azure Blob Storage	SAS (Shared Access Storage) Token, Connection String, Endpoint
Azure Data Lake Storage Gen 1	Client ID, Client Secret, Token
Azure Data Lake Storage Gen 2.	Account Key, Account Name

Once all the fields are filed click on the “Create Button” to create specified credentials.

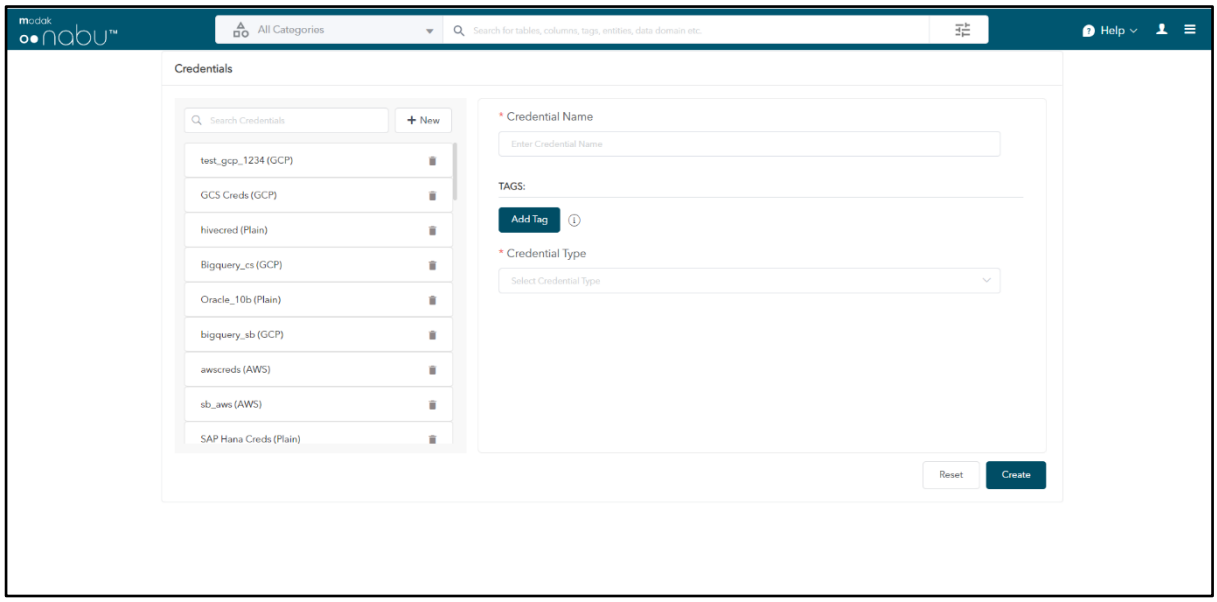


Figure 237: Credentials page

9.2 Editing of Credential

To edit an existing credential, click on the required credential from the left side of the panel, else search for the required credential by typing the name in the search box provided.

Once you select the required credential, textboxes are highlighted where you can edit the required field.

To edit password, click on “Change Password” this resets the password and you can change it easily.

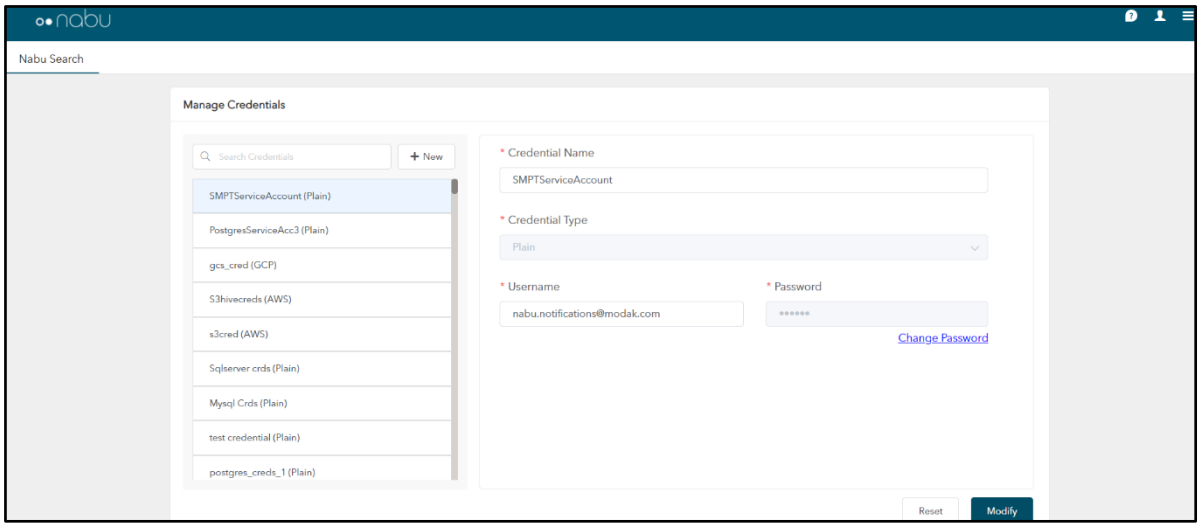


Figure 238: Edit Credential

Once all the required changes are done click on “Modify button” to freeze the changes and save it.

10 Compute Engines

To manage the compute engines, the user can navigate to ‘Manage Compute Engines’ from the menu.

Figure

The user can view all the available/existing compute engines on the left pane. To search for a specific compute engine, type compute engine name in the search box.

10.1 Create Compute Engine

The user must enter the below fields to create the compute engine.

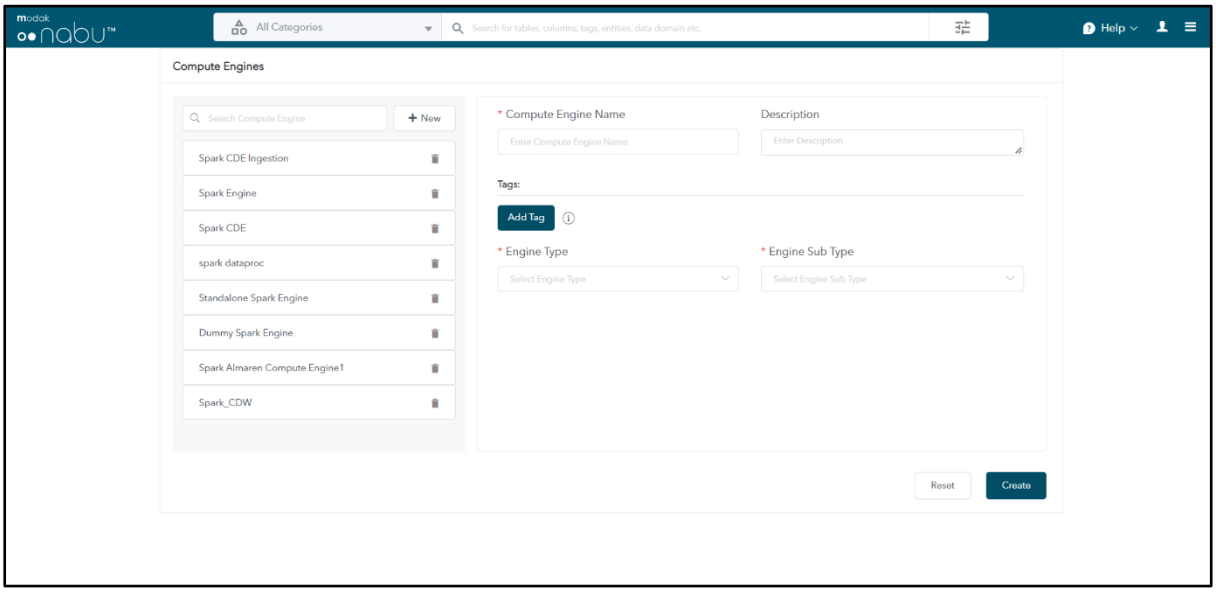


Figure 239: Compute Engine Page

1. Compute Engine Name: This is a mandatory field, and the compute engine name should start with an alphabet, should contain at least 3 characters, and can contain numbers, special characters except underscore are not allowed.
2. Engine Type: This is a mandatory field, and you must select the engine type from the dropdown.
3. Description: This is an optional field where user can add some description to the compute engine.
4. Engine Sub Type: This is a mandatory field and depends on the engine type value. This must be selected after selecting the engine type.
5. Create/Reset: The user can create the compute engine by clicking on the “create” button. The form should be valid for creating the compute engine.
6. “Reset” button resets all values in the form.
7. create success message will be shown on successful creation.

10.2 Modify Compute Engine

Select the compute engine from the left pane and the respective details will be shown on the right. The user can view the details or modify the same and click on “modify” button. Modify success message will be shown on successful modify. Any time, the user can switch back to "New” compute engine by clicking on "New” button.

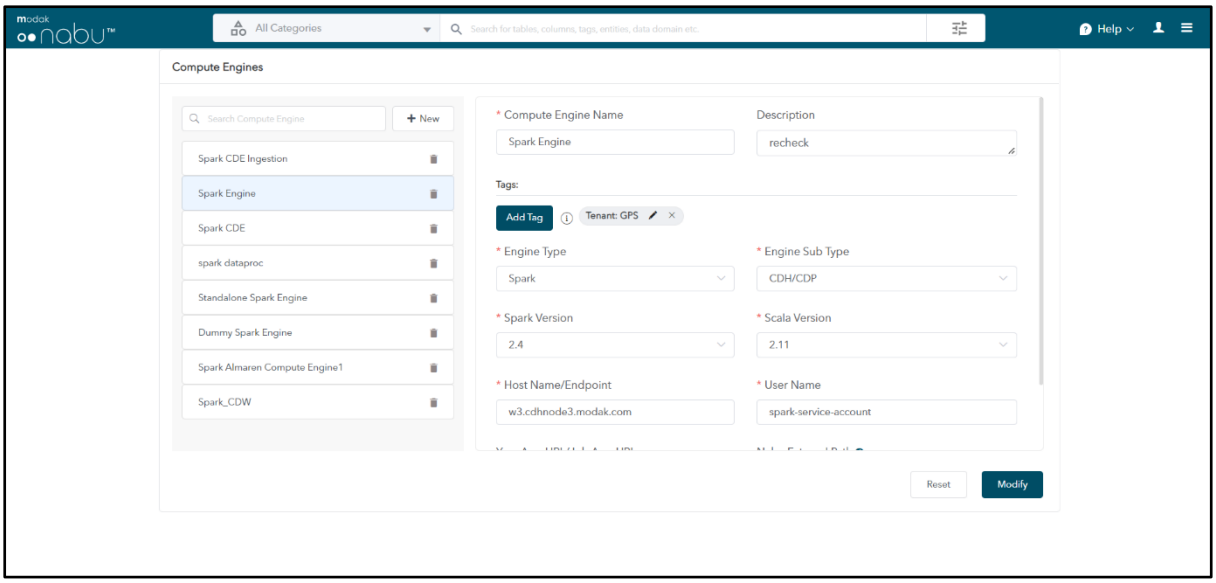


Figure 240: Modify Compute Engine

10.3 Delete Compute Engine

To delete a particular compute engine, the user can click on the delete icon on the right of the compute engine name.

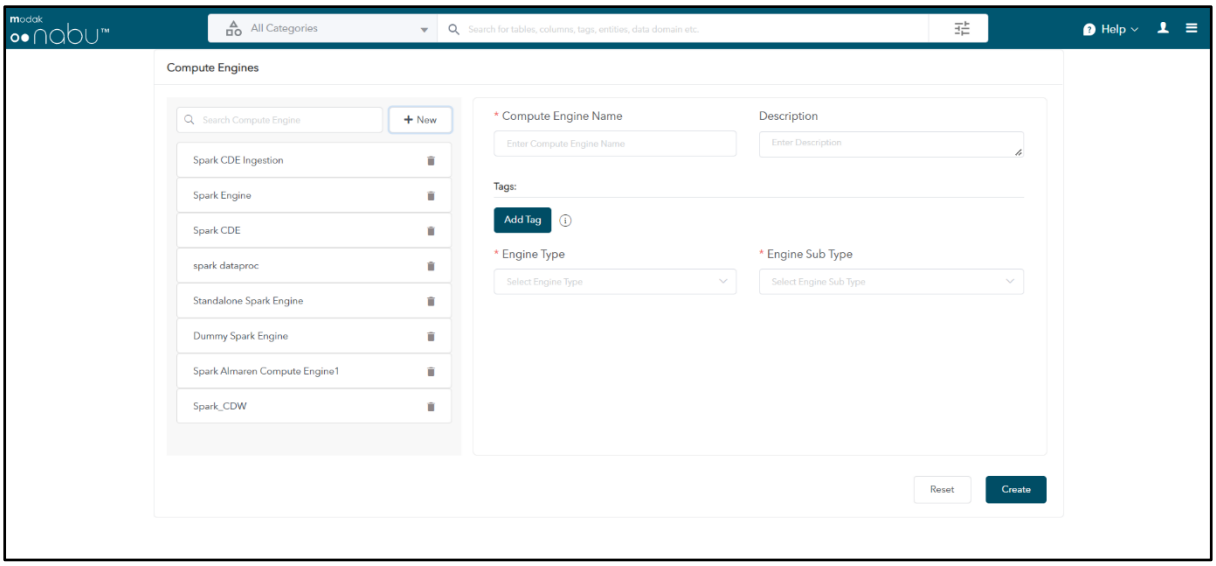


Figure 241: Delete Compute Engine

The delete confirmation will be prompted to the user to confirm the deletion. On confirm, the selected compute engine will be deleted.

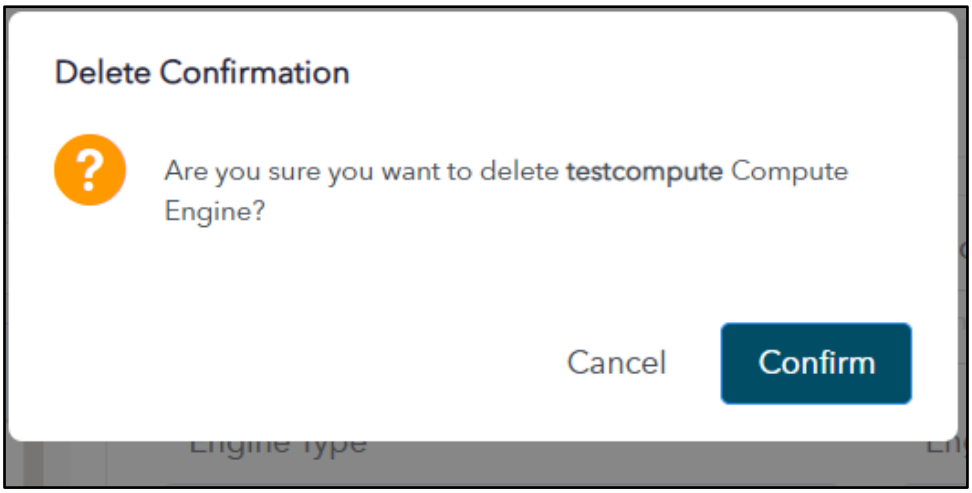


Figure 242: Compute Engine-Delete Confirmation Popup

11 Nabu Search

This functionality helps in discovering and exploring data in Nabu. The search bar helps to display the relevant results depending on the keyword typed. 'Nabu Search' has auto-complete feature that helps in recommending the search terms based on the typed characters. This helps the user in exploring data as well as avoid typos in search terms.

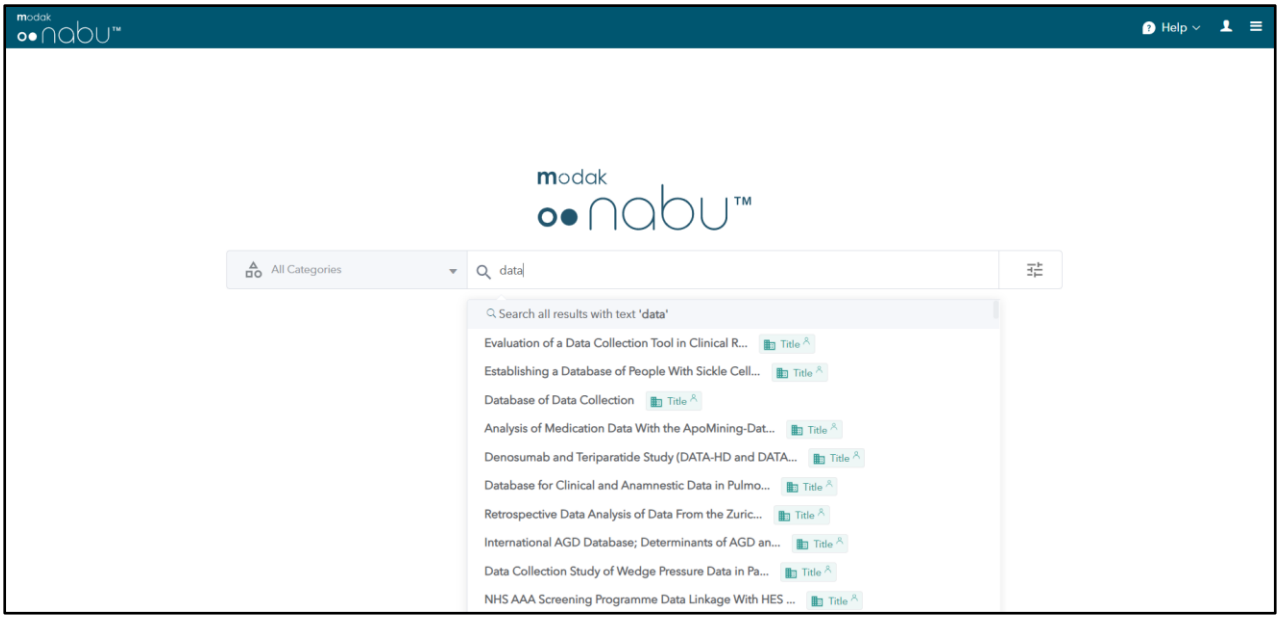


Figure 243: Nabu Search

The search bar has select category option to refine the data from the search. By default, "**All categories"** is enabled with all categories selected as shown below.

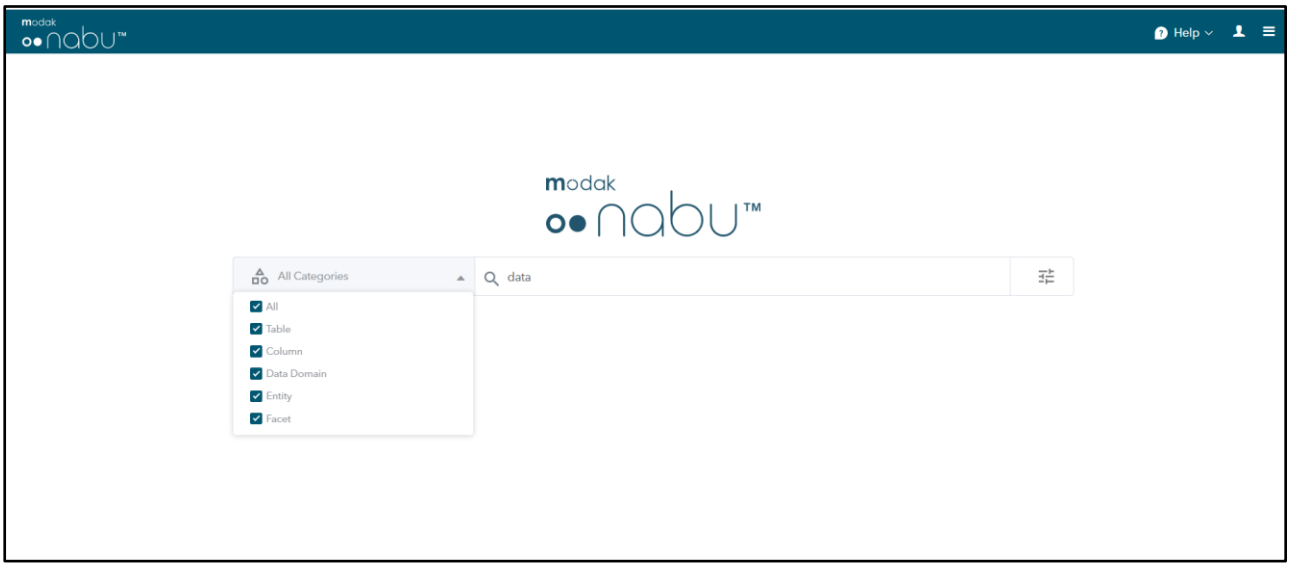


Figure 244: Nabu search - categories

The user can select/deselect the category as preferred. The selected categories are shown as tags.

Ex: If table and column are selected as categories, the results relevant to only table and data are displayed for the search term.

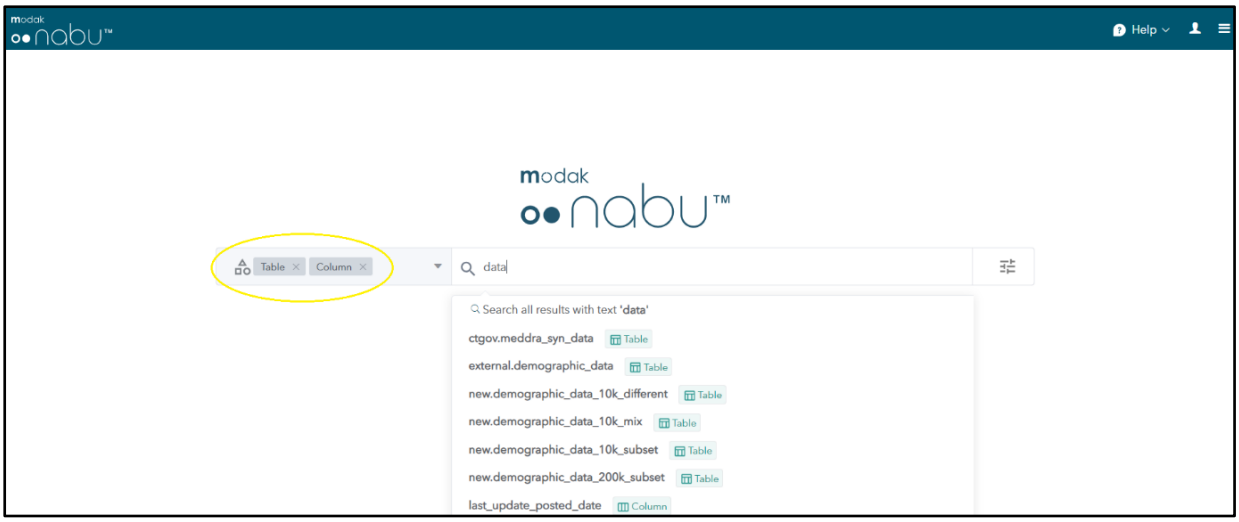


Figure 245: Nabu Search – applied categories

In the search box, along with the autocomplete suggestions, tags are shown which provide additional information about the search term. For e.g., the tags identify whether the search term is a data domain, table, column etc.

11.1 Searching data

Users can search data, using Nabu search, by entering a search term in the search box. As the user types few characters (3 or more), autocomplete suggestions are shown of the terms which match the typed characters.

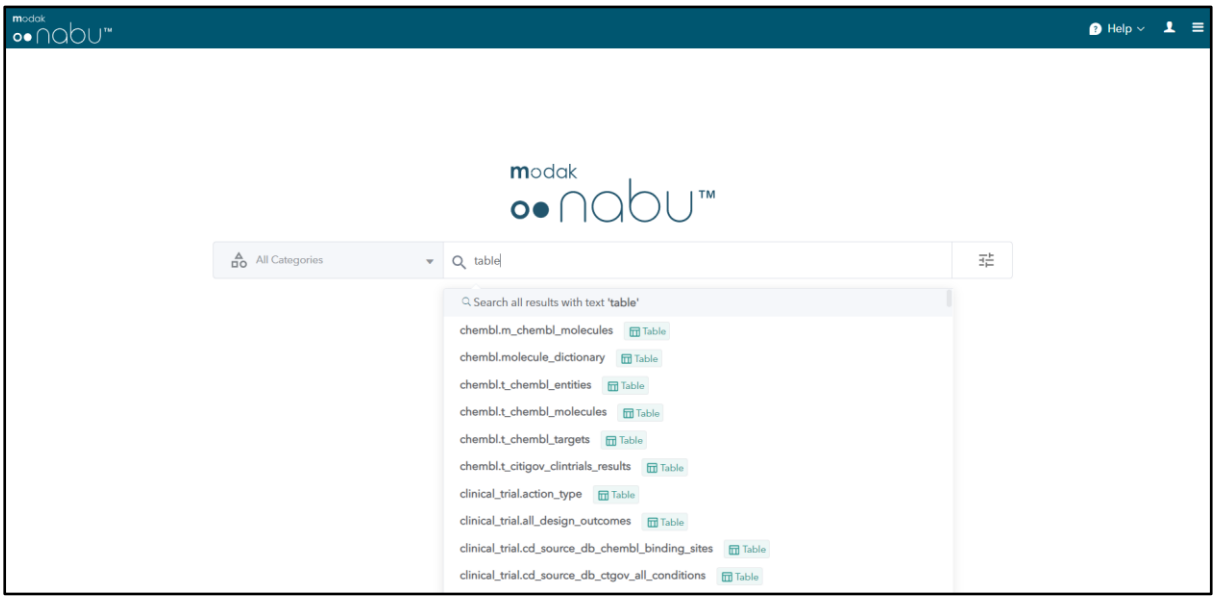


Figure 246: Search for data

Select any of the suggested results from the autocomplete search box, by clicking on it.

Alternatively, to view the complete set of results that match the typed term, press the enter key after providing the term.

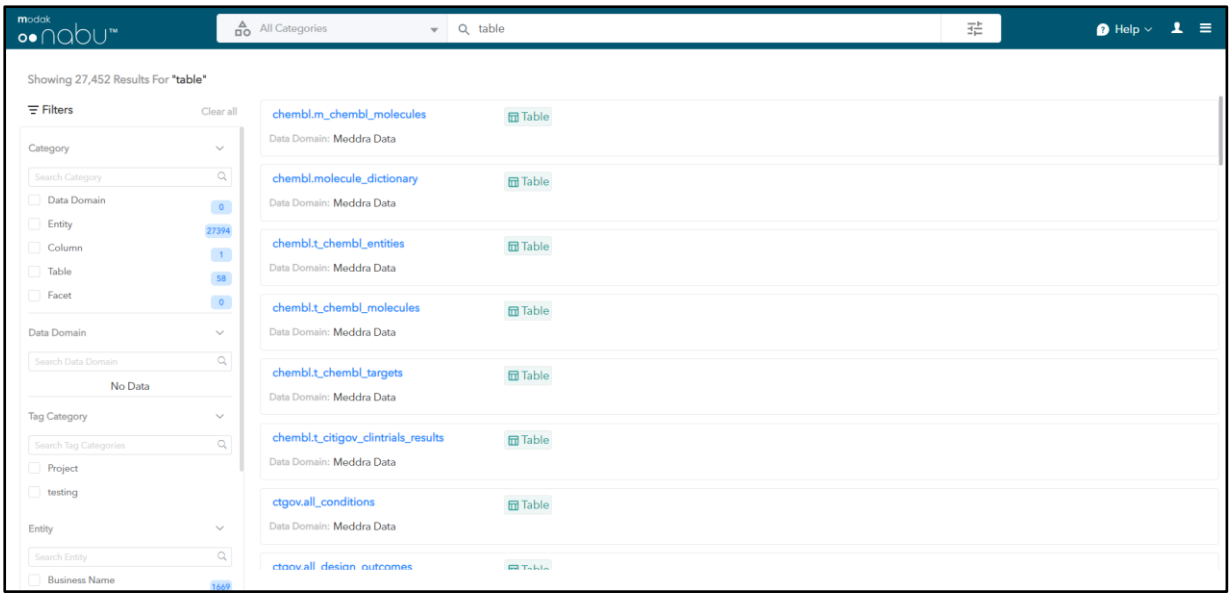


Figure 247: Nabu search - search results

The results page above shows the complete set of results with the results count for the typed term. The left side shows all the categories and the count that matched the typed term along with the category.

*Ex: The search term **table** has total 27,452 results, and each category like Entity, Data Domain, Columns etc has separate counts for the matched term.*

The user can select any category by checking the checkbox and the relevant results will be displayed on the right. By unchecking the category, the complete results will be displayed again.

For any huge data, the user also has option to search for any particular category like search Data Domain, search Entity etc.

Any selection of the left side will be applied as tags and the relevant refined results will be displayed on the right.

Note: The Clear all will clear any applied tags.

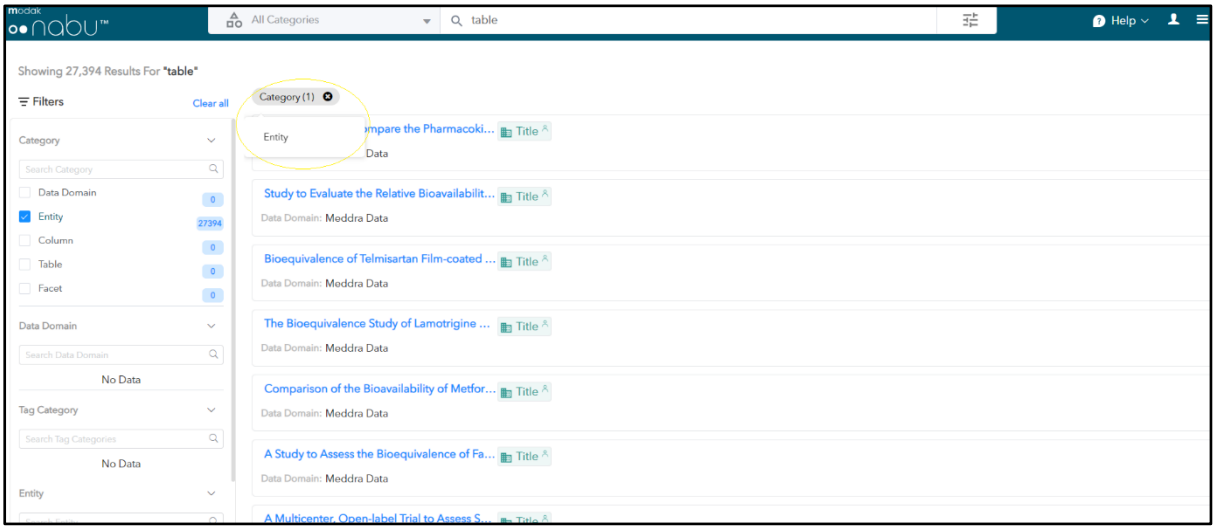


Figure 248: Nabu Search – search results – applied filter

Multiple filters can be applied which will refine the search results. For category entity, the user can again select any particular entity as preferred.

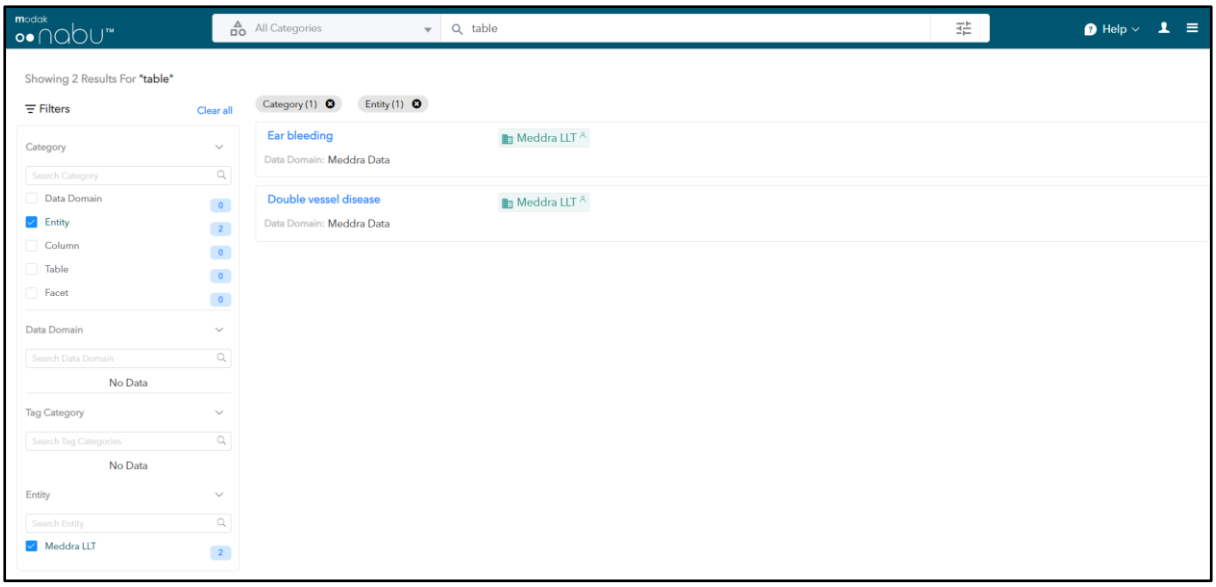


Figure 249: Nabu Search – results for applied filter

The user can navigate further by clicking on the table name on the right.

11.2 Search for Table

To search for a table, type few characters from its name, and then select one of the results in the autocomplete search box or select any table.

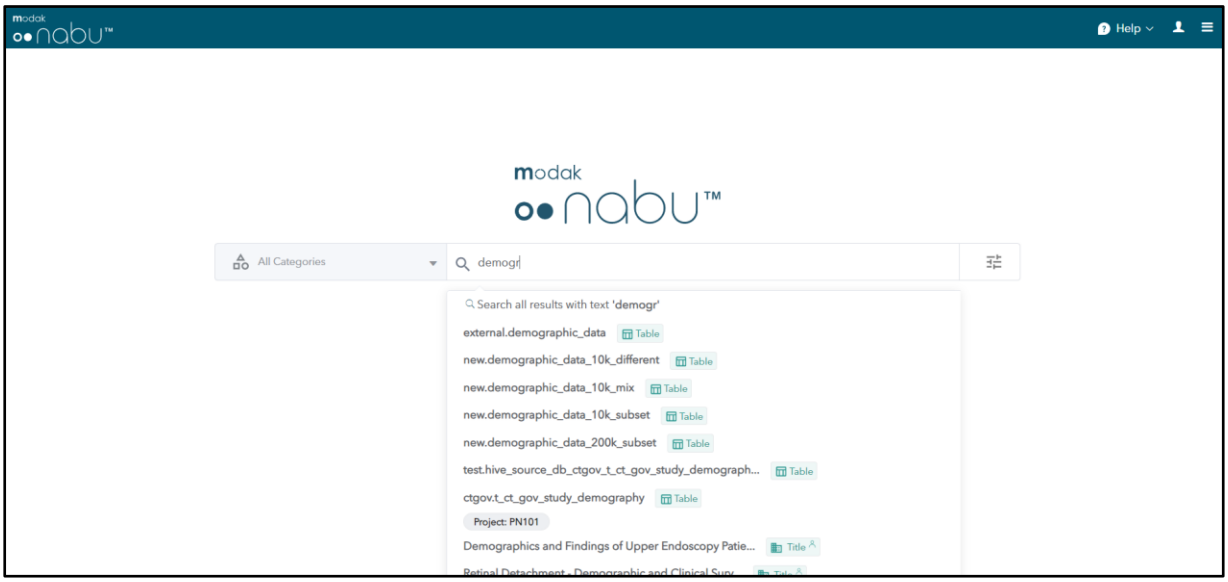


Figure 250: Nabu Search – Search for Table

When user clicks on the table name, the further table details like table overview, profile of a table, Data, knowledge graph and Data Lineage will be shown as below. The Overview option is shown by default as the first tab.

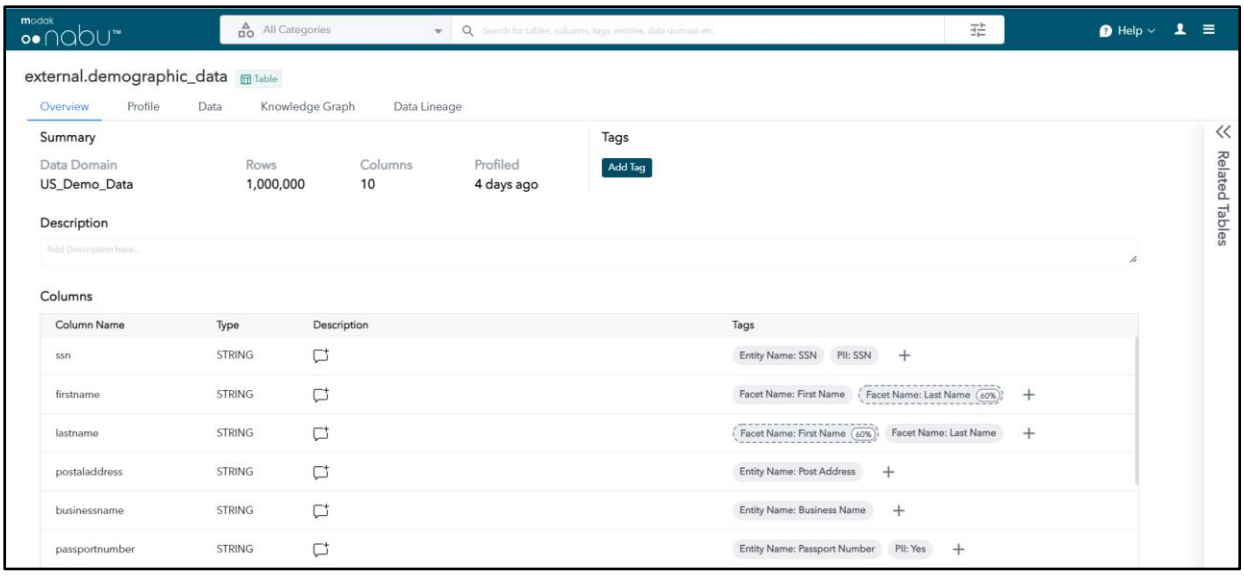


Figure 251: Nabu Search – Table results

11.2.1 Overview of a table

The overview shows the summary of the table like number of rows in the table, number of columns, Data Domain of the table, the details on when it was profiled, description of the table.

It also shows the column details of the table like column name, type, description for each column and tags which are added for each column.

Users can also add reviews and ratings about the table. The reviews and ratings can provide a quick view of the quality and reliability of the data.

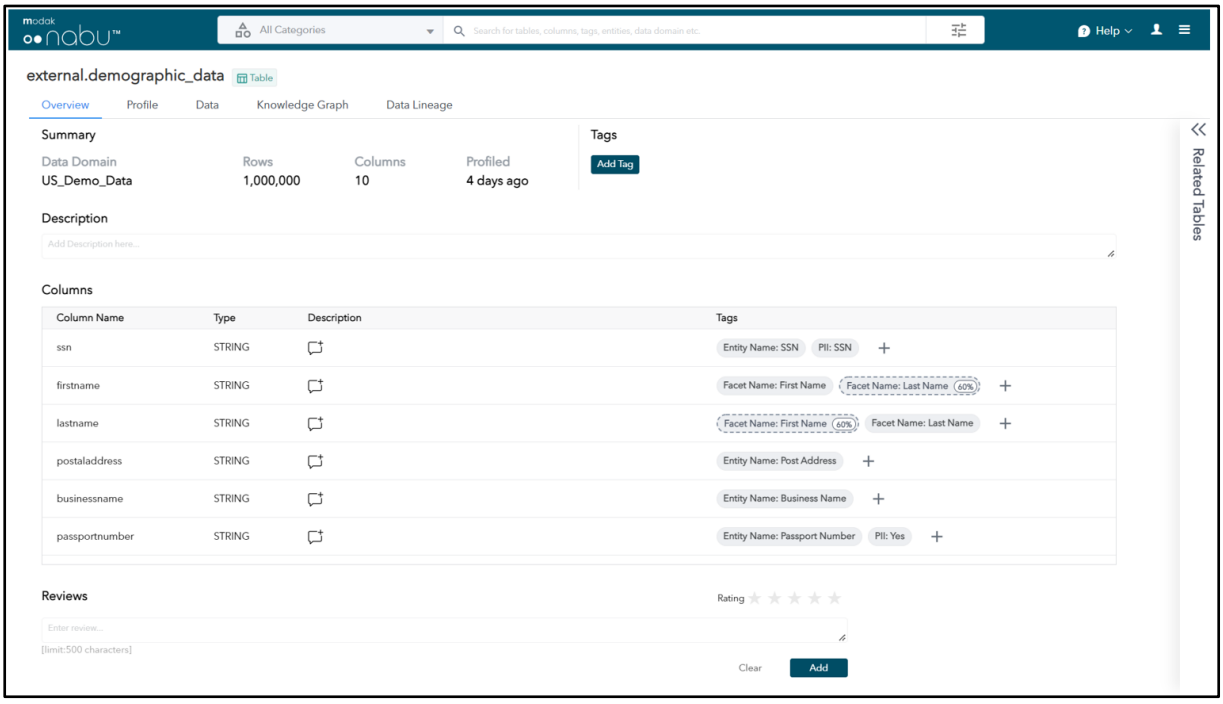


Figure 252: Nabu Search – overview of table

Table description: The overview tab also enables tribal knowledge about a data to be captured. For a table, description can be added to provide any information about the data in the table.

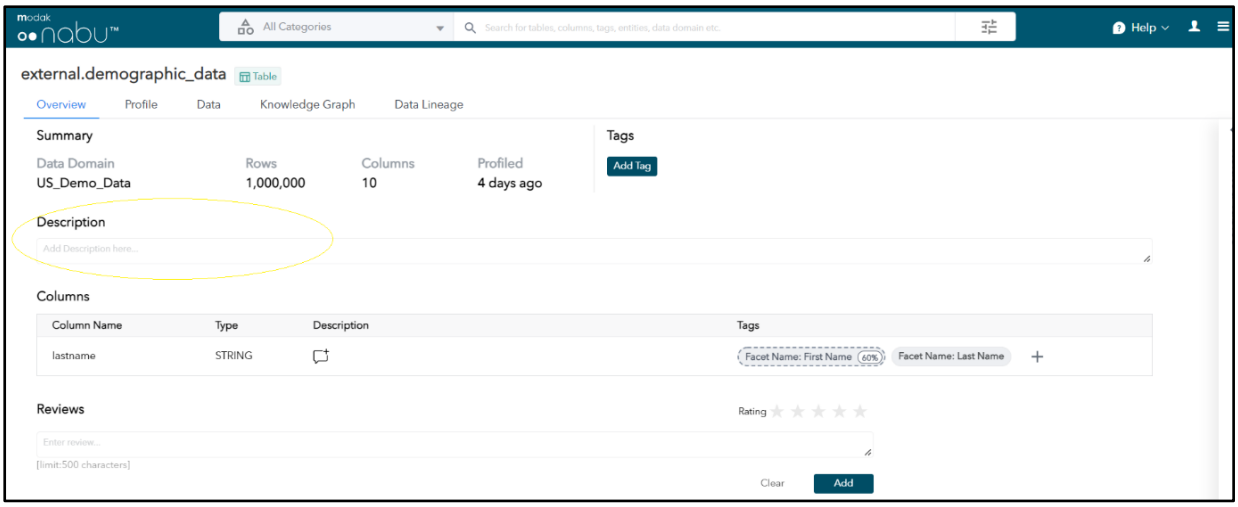


Figure 253: Nabu Search – Table description

Column description: Description can be added for each column by clicking on the icon in the description column and provide the description in the field.

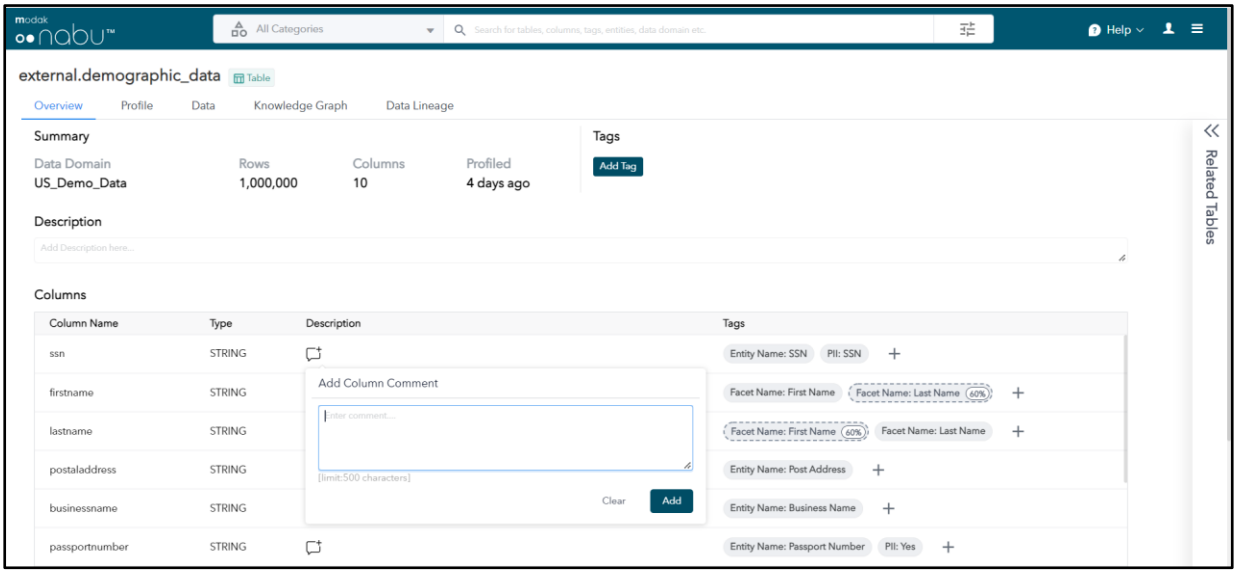


Figure 254: Nabu Search – column description for table

Tags: Users can also add tags to tables and columns on the overview tab. These tags are searchable and can help users to discover similar tables and columns.

To add a tag, click on the 'Add Tag' button for a table and + icon for a column. Provide the tag category and a tag value.

The user can add any new tag for the table by clicking on Add Tag button which will open the Add Tag modal to add tags.

If a 'Tag Category' or a 'Tag Value' is already created, it will be shown as suggestion, when the user starts typing. Otherwise, the users will be asked to create the category.

The added tags will appear as <Tag Category>: <Tag Value> for the table or column.

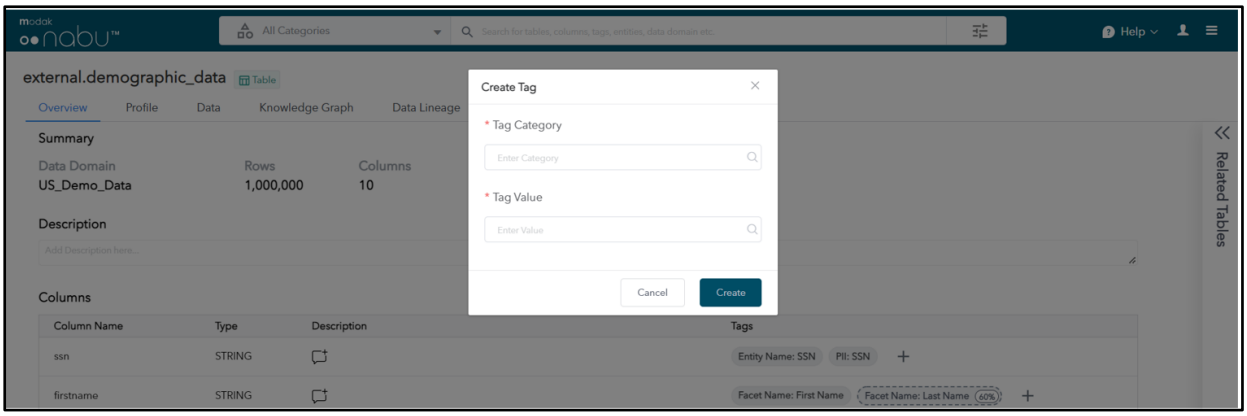


Figure 255: Create Tag Modal

Fading Tags

The tags of an entity, facet are shown as fading tags for the suggested columns which are matched with the fingerprinting column.

Ex: The column 'firstname' is defined as fingerprinting column for Facet 'First Name' and it has suggested column 'name' with match percentage of 60% with the column 'firstname'.

The tag is shown as faded till they are approved by a data steward.

As shown below, the tag **Facet Name: Last Name** is fading tag with match percentage of 60. Once the suggested column is approved, the tag will be shown as a normal tag

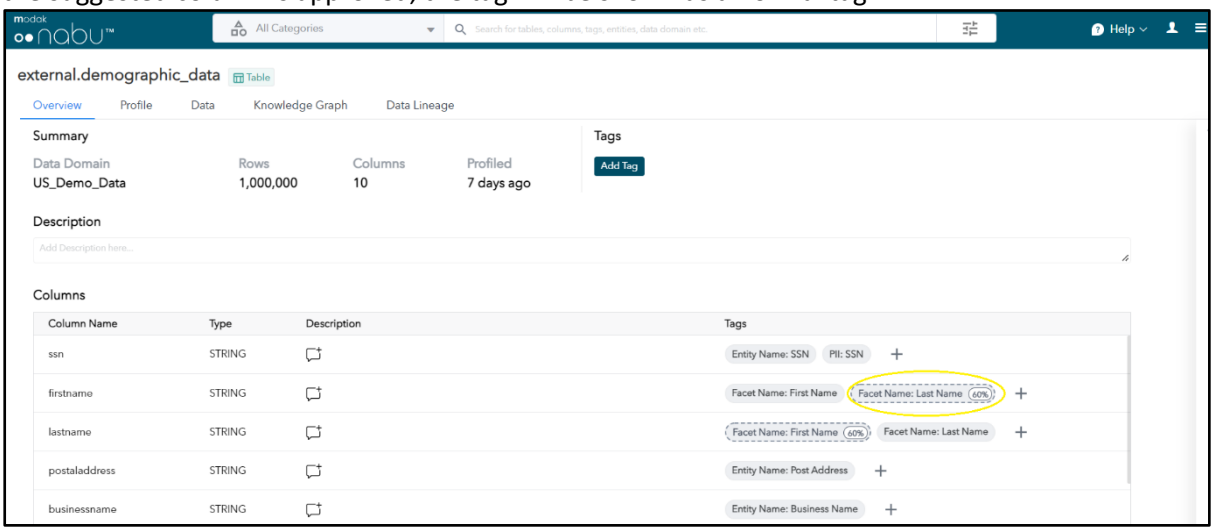


Figure 256

Related tables: To find the related tables with the table searched, click on the double arrows icon in the overview tab as below.

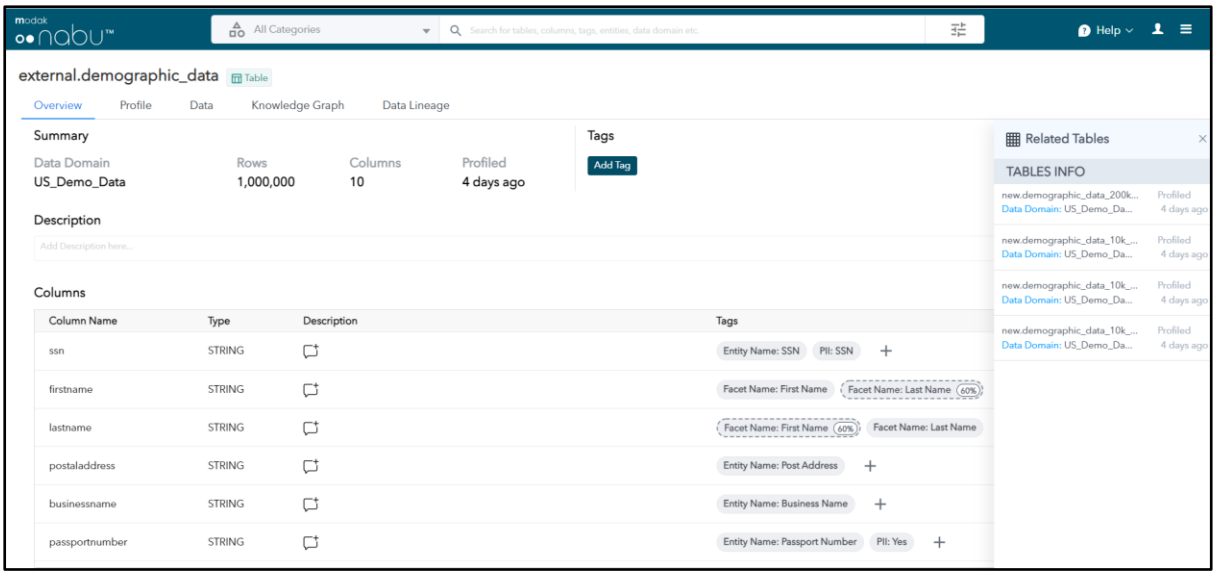


Figure 257: Nabu Search – Related Tables

11.2.2 Profile of a table

Click on the profile tab to view the profile information about a table. It provides a summary of the data present in the table, which helps to assess the quality of the data. Profile shows list of columns, their data types, the distribution of values for each column, number of distinct values, null count etc.

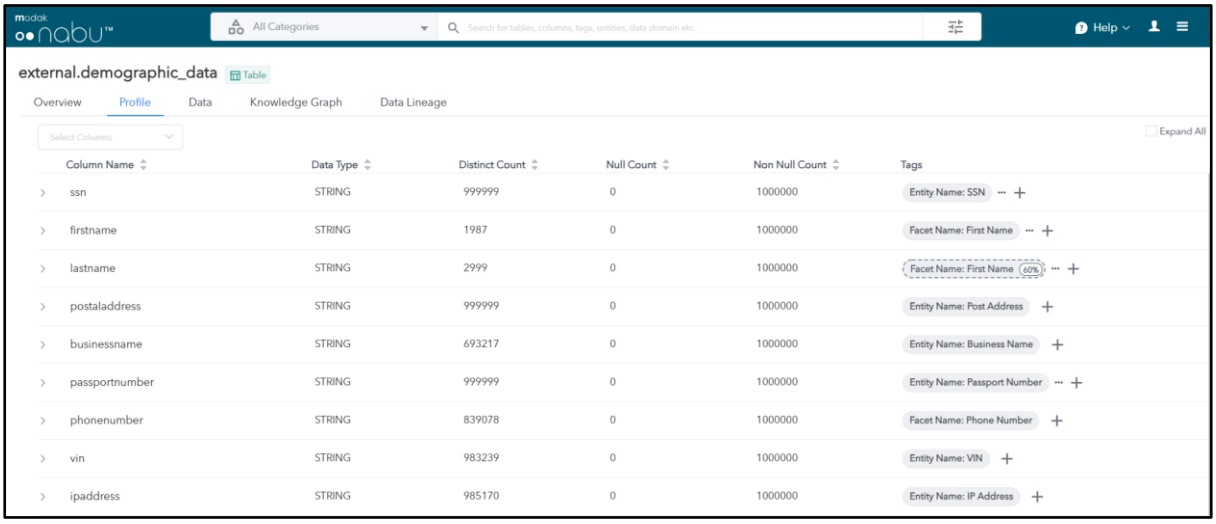


Figure 258: Nabu Search – Profile of a table

The profiling page provides a list of columns in the table along with data type, number of null values, number of distinct values for each column.

To view further details about a column, click on the arrow '>' which is at the end of the column name. To view details about all the columns, check the 'Expand All' checkbox.

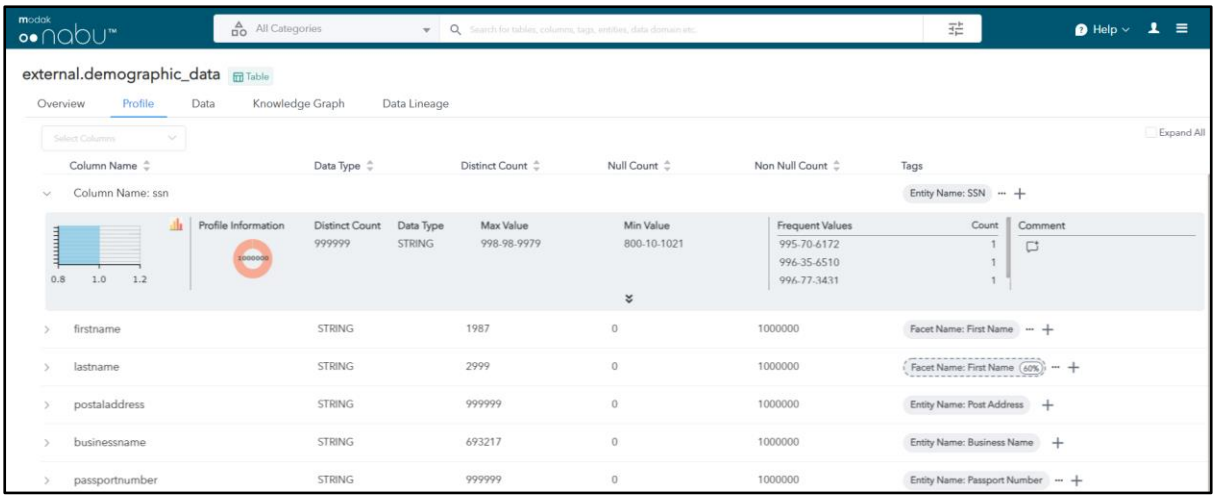


Figure 259: Nabu Search – column information in profile tab

The column details shown include frequency distribution of values in a column, data type, maximum and minimum values, most frequently occurring values.

Comments can be added to a column to share additional information which would be shown to other users as well.

The user can also search for any particular columns to show by selecting on search columns as below.

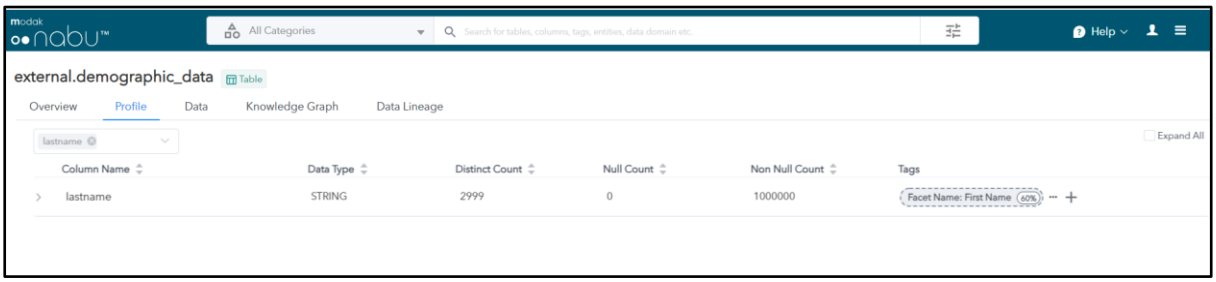


Figure 260: Nabu Search – Profile tab – search for column

11.2.3 Data tab

On clicking the "Data" tab, the data in a table will be displayed.

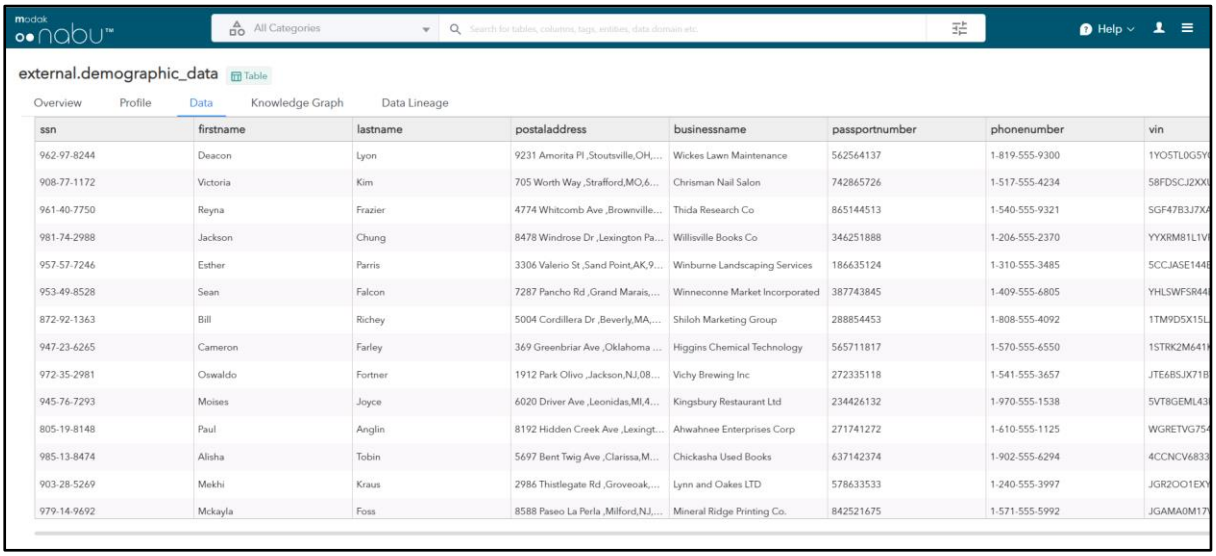


Figure 261: Nabu Search – Data Tab for table

11.2.4 Knowledge graph

Knowledge graph for a table, displays how the table is related to other tables, through similar columns. The similar columns are discovered by Nabu's fingerprinting process.

To view a knowledge graph for a table, search for the table name and click on Knowledge graph tab.

The knowledge graph screen opens up, with names of columns that are similar to other columns. To view tables that have similar columns, double click on the table icon 

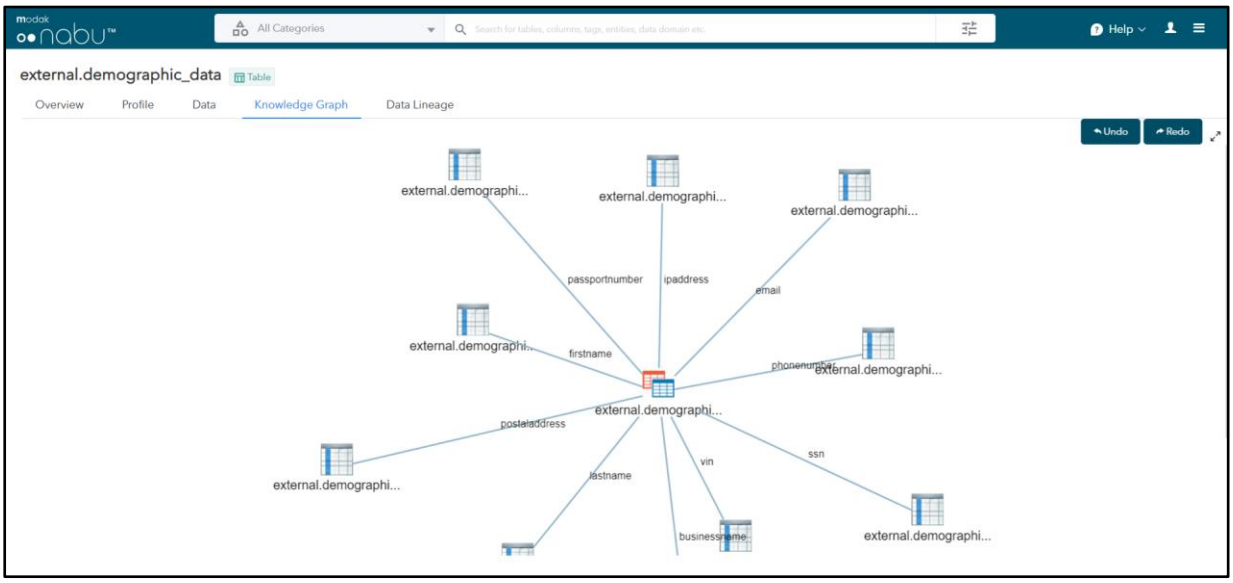


Figure 262: Nabu Search – Knowledge graph of a table

The lines linking the searched table to other tables, shows the names of columns from those tables, that are similar. The percentage of similarity, as calculated by fingerprinting is shown next to the column name.

The below is the expanded view of knowledge graph for tables.

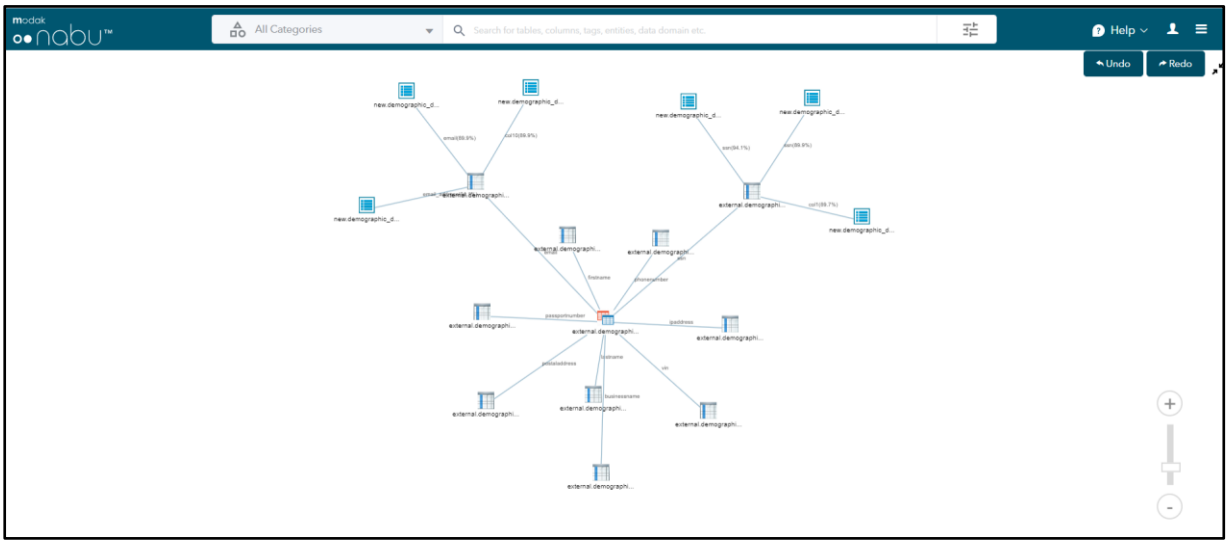


Figure 263: Nabu Search – knowledge graph expanded view

11.2.5 Data Lineage

To view lineage of a table, search for a table and then click on ‘Data Lineage’ tab.

The ‘Data Lineage’ screen shows the name of the source from where the table was ingested. It also shows the source type (PostgreSQL, SQL server etc.), the destination type and name of the data domain that the table currently belongs to.

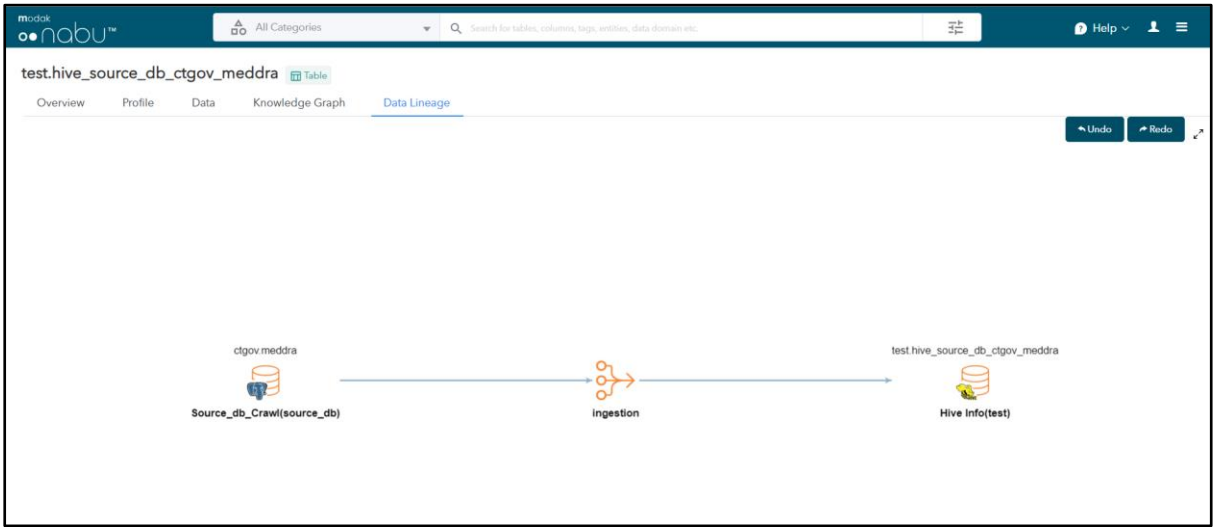


Figure 264: Nabu Search – Data Lineage

11.3 Search for Entity

An entity in Nabu is defined in the ‘Data Catalogue’ section. An entity is a collection of similar columns in a data domain. Similar columns are identified using a meta rule or using data fingerprinting. Data within these columns is searchable using Nabu Search.

To search for an entity, type its name in the search box. The autocomplete suggestion box shows all items that match the typed term. An entity can be identified with the ‘Entity’ tag next to it.

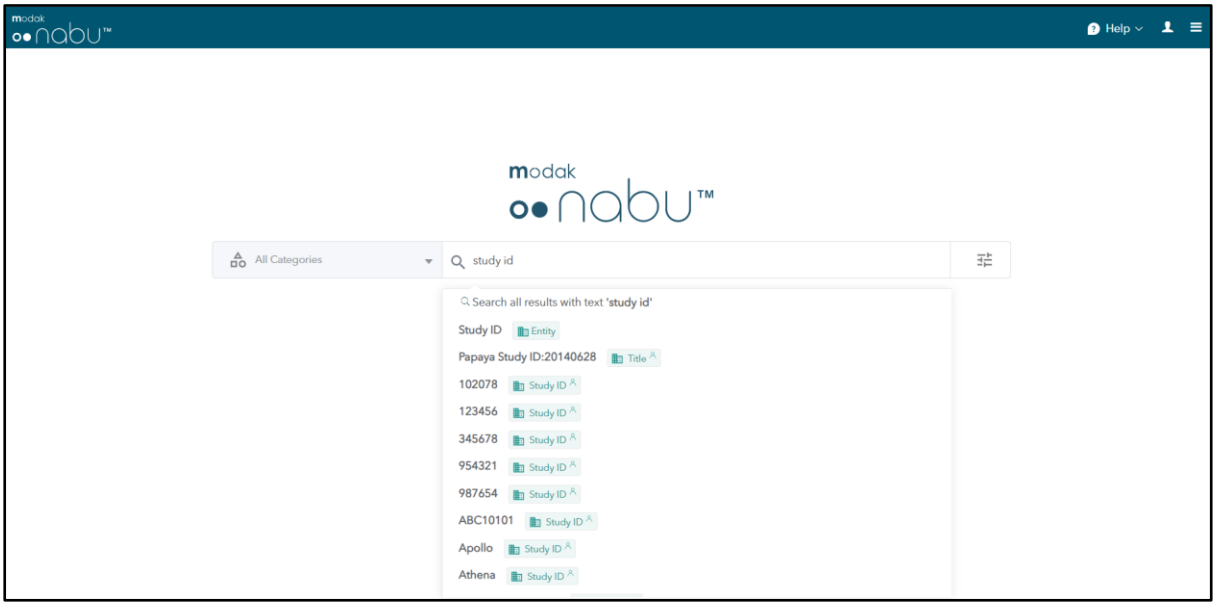


Figure 265: Nabu Search – Entity search

The search results for an entity show the number of columns that are grouped as part of the entity, and the number of tables where the columns are found. The name of the data domain on which the entity is defined is also displayed.

To see the list of tables or columns click on ‘Tables’ or ‘Columns’ tabs respectively. To view further details about the table, click on the table name.

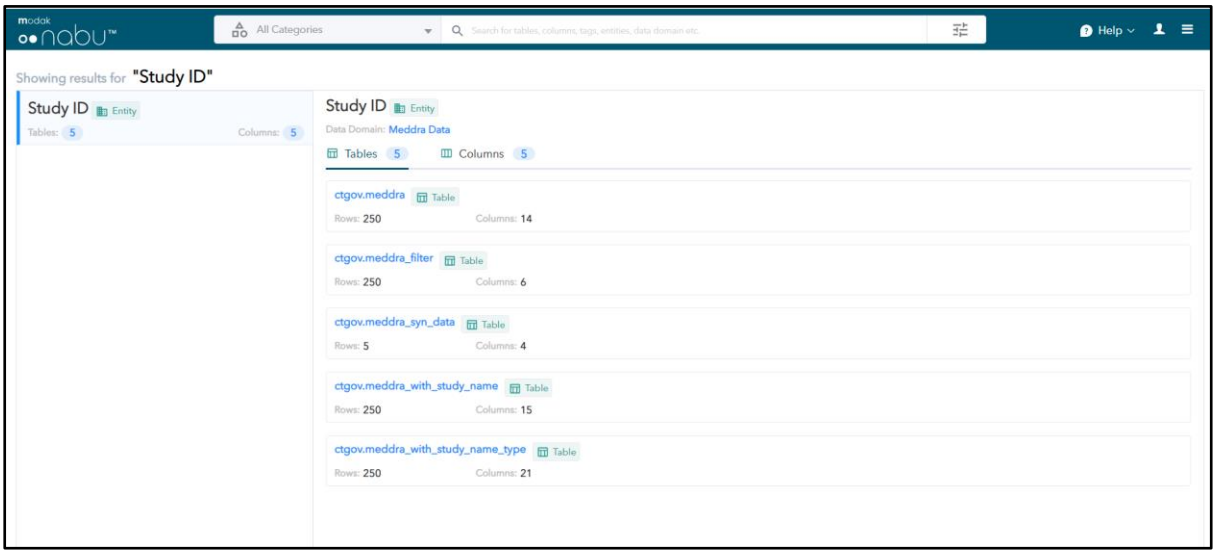


Figure 266: Nabu Search – Entity search results - tables

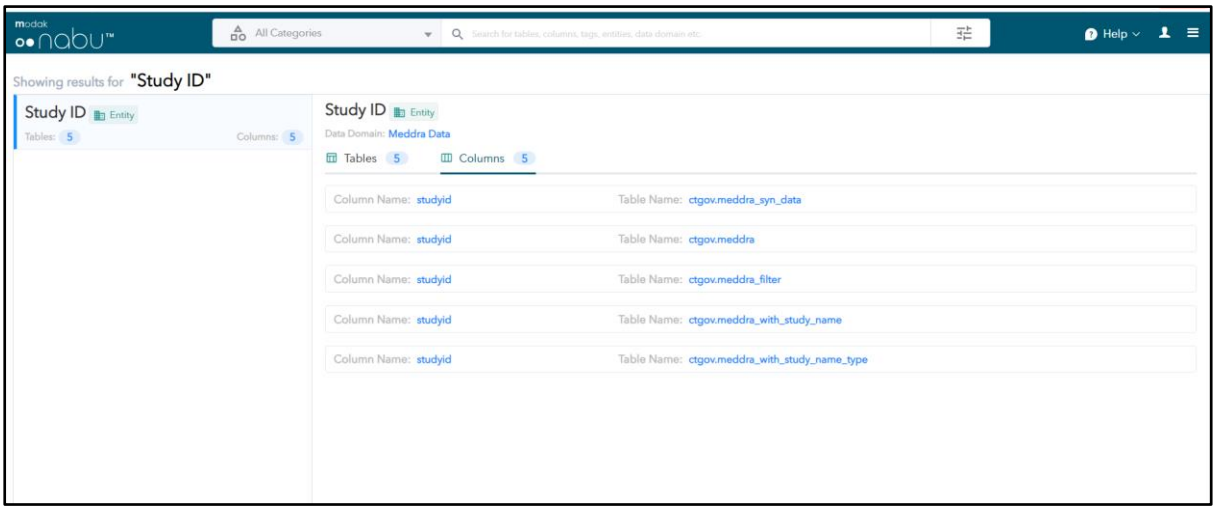


Figure 267: Nabu Search – Entity search results - columns

To view the Data Domain details, click on the Data Domain.

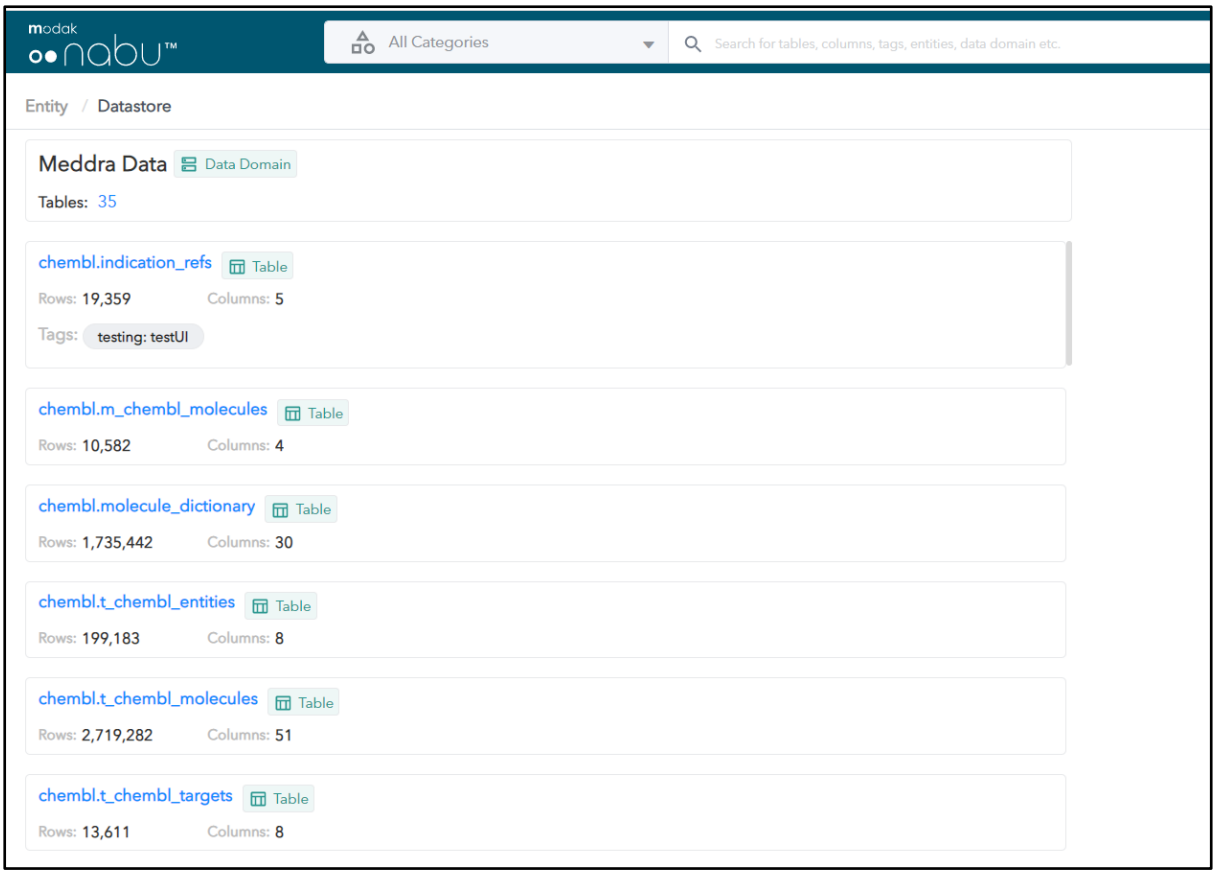


Figure 268: Nabu Search – Data Domain details

11.4 Search for Entity Value

To search for a value in a column defined as an entity, type few characters of the value. The values are shown. The tag next to the value shows the name of the entity.

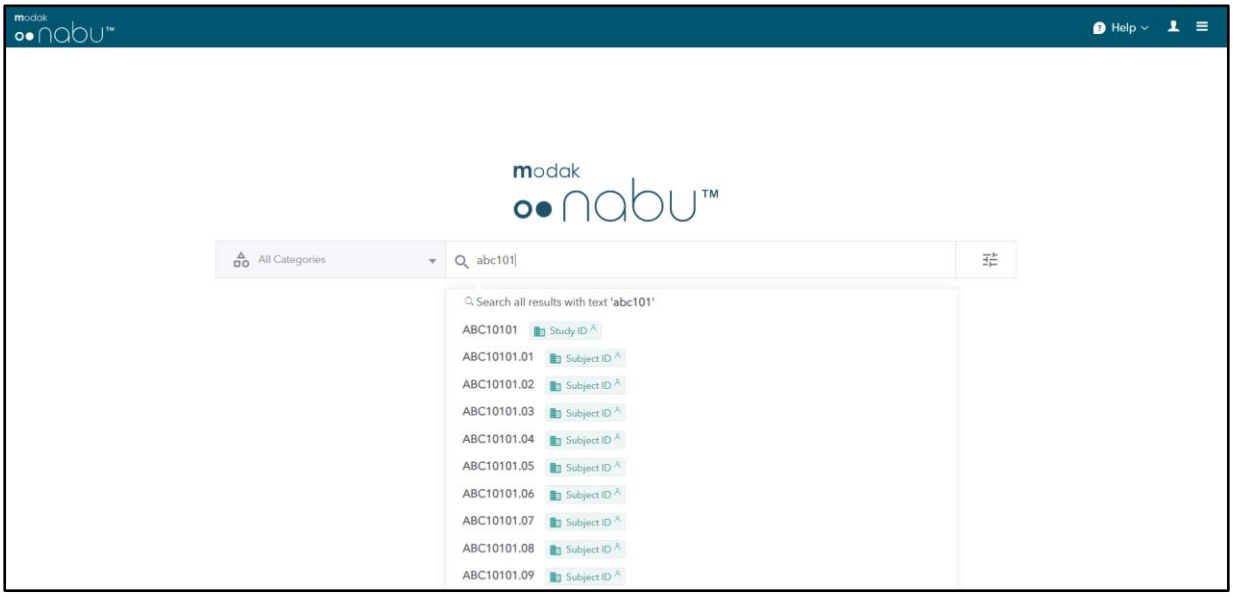


Figure 269 : Nabu Search – Search for entity value

Searching for an entity value gives the list of tables in which that value is found. Additionally, the list of related entities is also shown. For e.g., in the screenshot below, entity value ABC10101 is searched for. The summary shows that the number of tables (5) where the entity value is located. It also shows there are 50 values of 'Subject ID' entity and 7 values of 'Meddra SOC' entity that are related to Study

ID ABC10101.

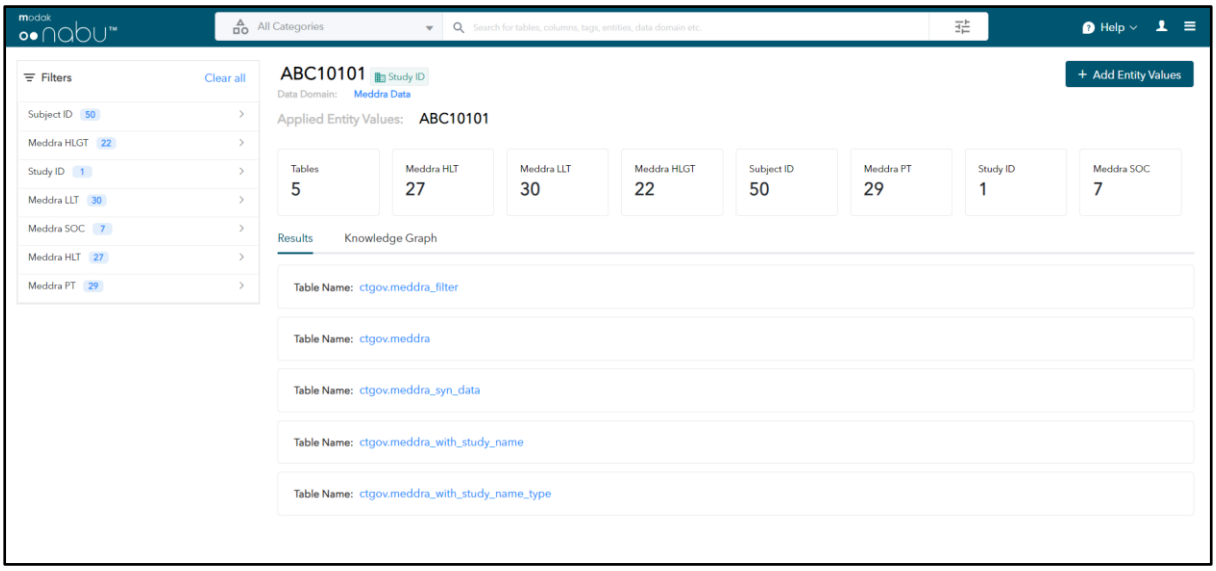


Figure 270: Nabu Search – Entity value results

To view results for 2 or more entity values together, the 'OR' search functionality can be used.

First search for an entity value. For e.g., in the screen below the entity value 'ABC10101' is searched for. Then, click on +Add Entity Values button which is at the right. In the resulting pop-up, type other entity values that need to be included in the 'Values' field. As few characters are typed, suggestions matching the entered characters are shown.

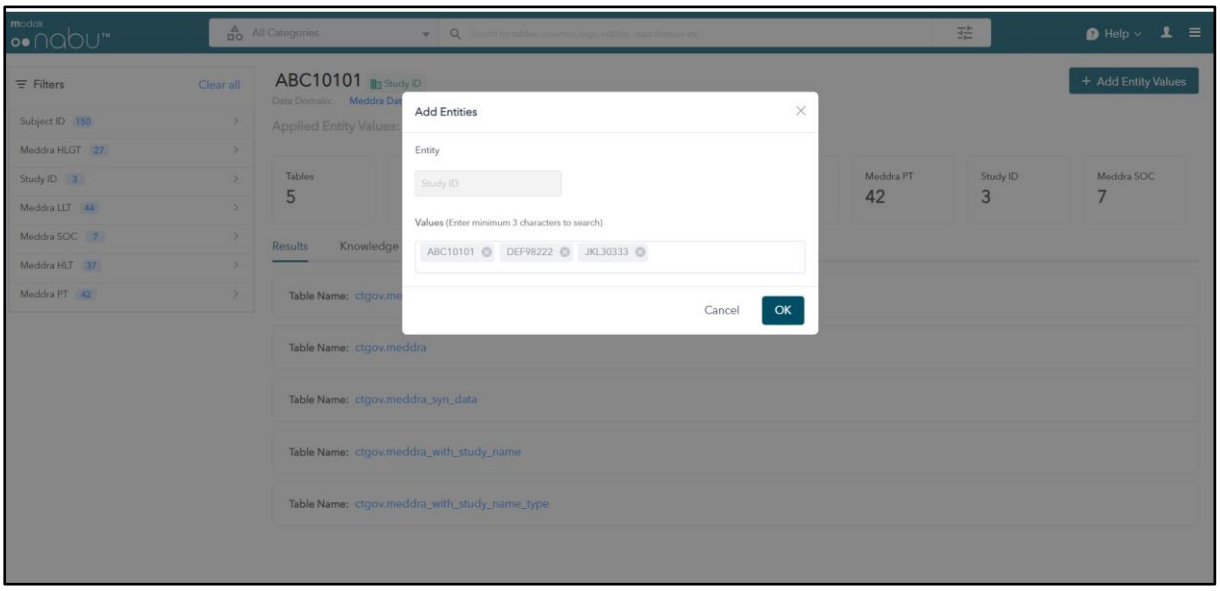


Figure 271: Nabu Search – Add entity values

Click 'OK' when required entity values have been entered. The search results now show results for both entity values together and added as 'Applied Entity values'.

The applied entity values can be anytime removed by hovering on the entity value and close the entity as shown below.

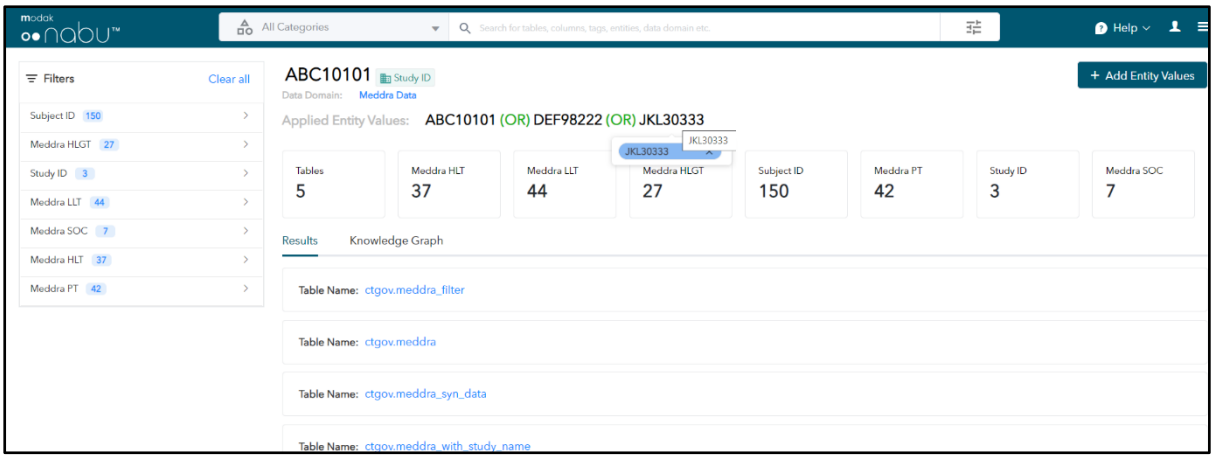


Figure 272: Nabu Search – remove entity value

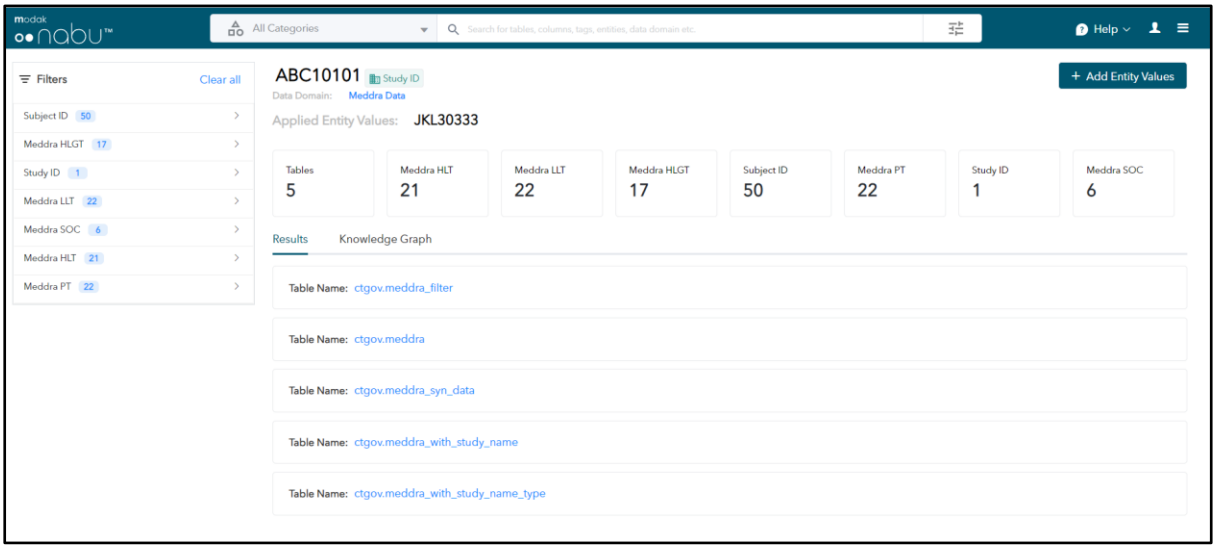


Figure 273: Nabu Search – remove entity value results

11.4.1 Knowledge graph

The relations between different entities can be visualised using a knowledge graph. To visually explore relationships, click on 'Knowledge Graph' tab.

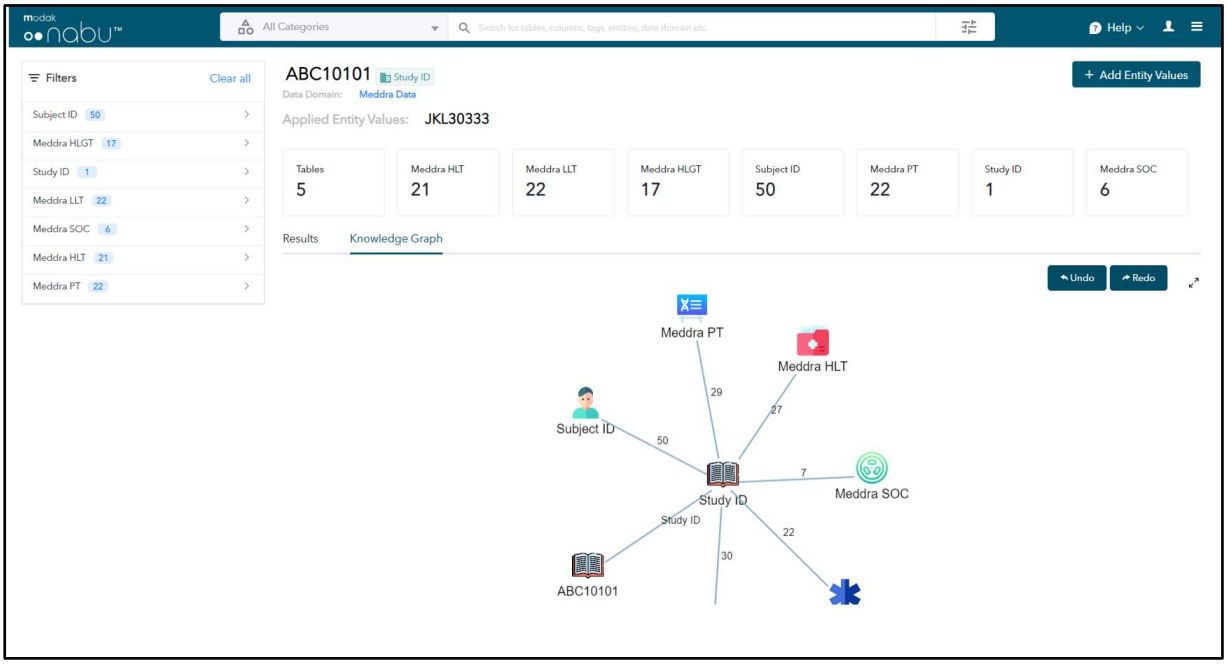


Figure 274: Nabu Search – related entities knowledge graph

The knowledge graph shows the searched entity value (in this case Study ID: ABC10101), linked to other entities. The numbers on the links denote, the number of distinct values of the related entity.

To view the distinct entity values, click on the node of the knowledge graph corresponding to the entity. For e.g., to view number of different values of 'Meddra SOC' in the above screen, click on 'Meddra SOC' node.

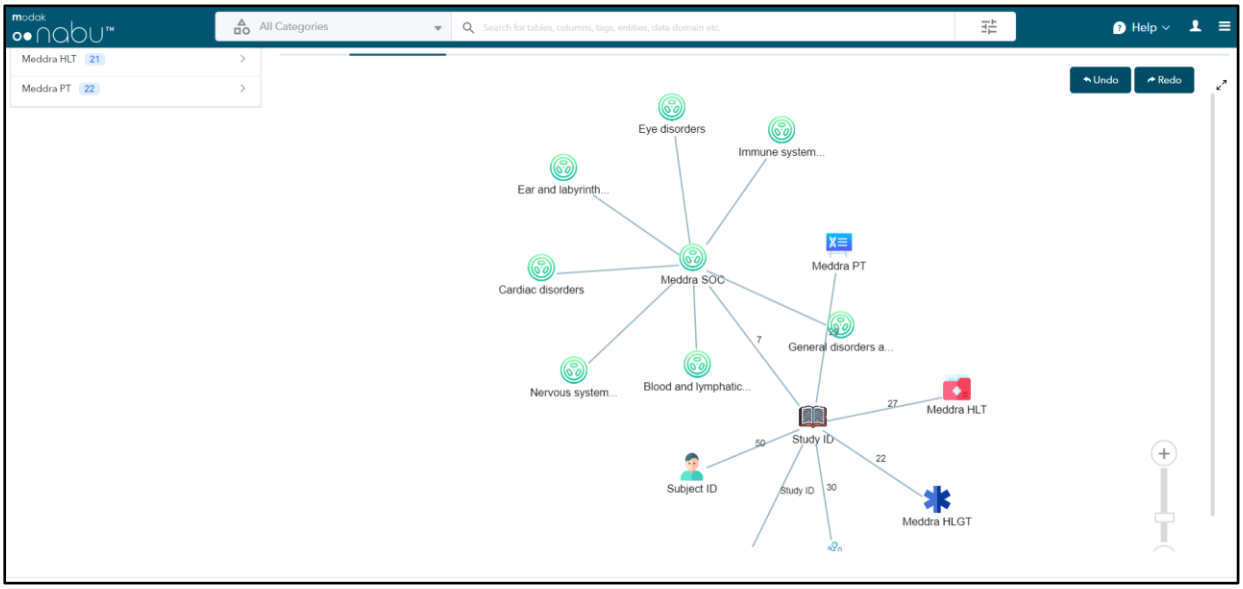


Figure 275: Nabu Search – related entities knowledge graph – node expanded

To progressively explore the relationship of a node with other entities, double click on that node. For e.g., to see other entities related to 'Cardiac Disorders' in the screen above, double click on the node 'Cardiac Disorders'

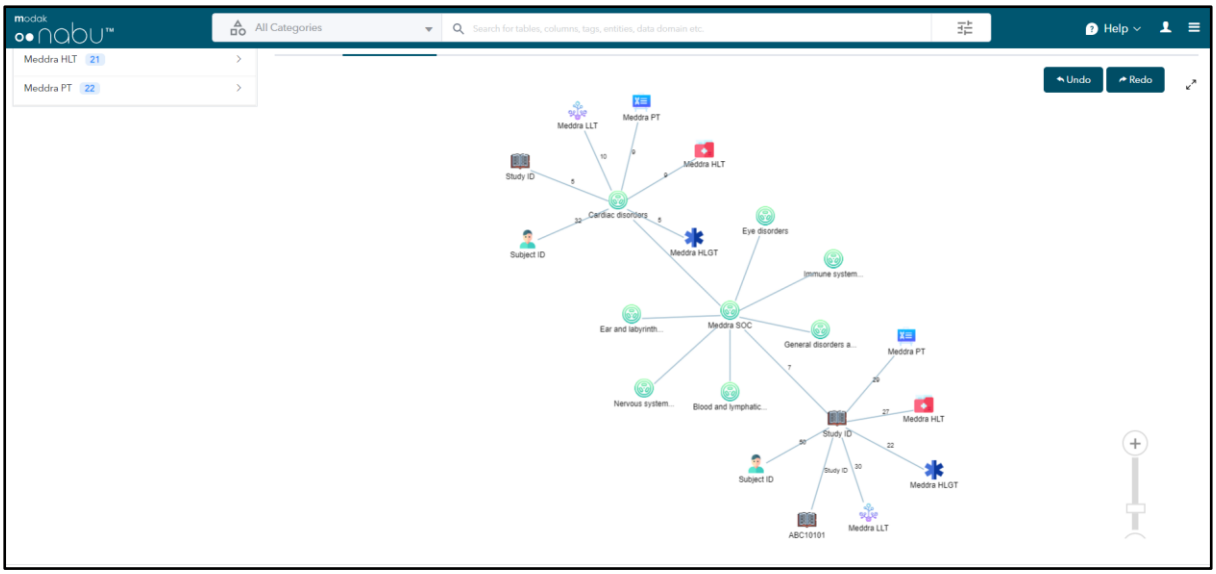


Figure 276: Nabu Search - related entities knowledge graph – another node expanded

11.4.2 Filters

Search results for an entity value can be refined by filtering them. The available filters are grouped under entities on the left side of the screen. These filters are defined in the 'Data Catalogue' section.

To filter on values of other entities, hover on the entity name on the left side.

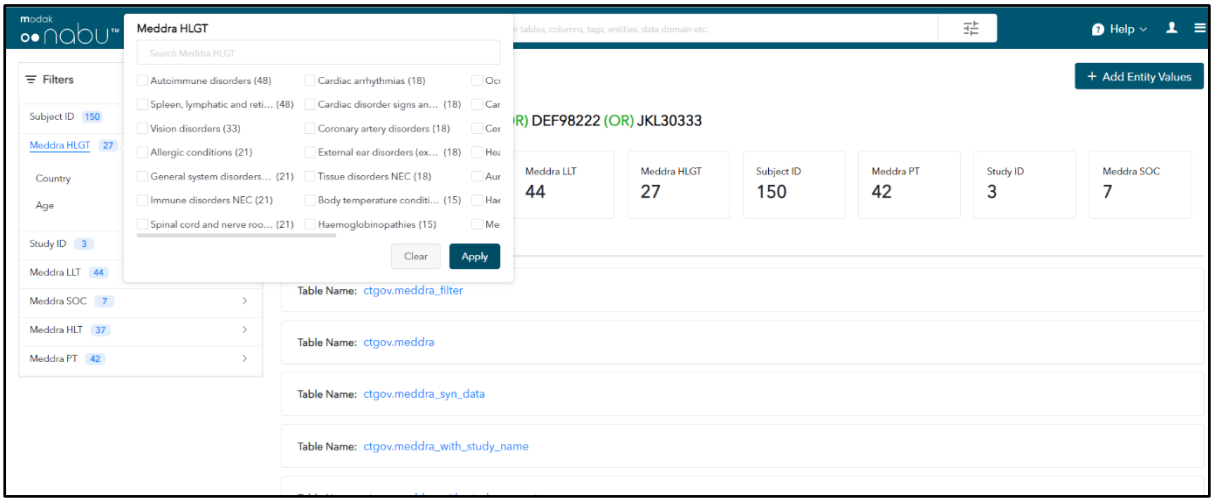


Figure 277: Nabu Search - entity filters

select the values of the entity to filter on and then click on 'Apply' button. The value will be added as filter and the results will be refreshed for the applied filter.

The applied filters can be seen at the top of related entities.

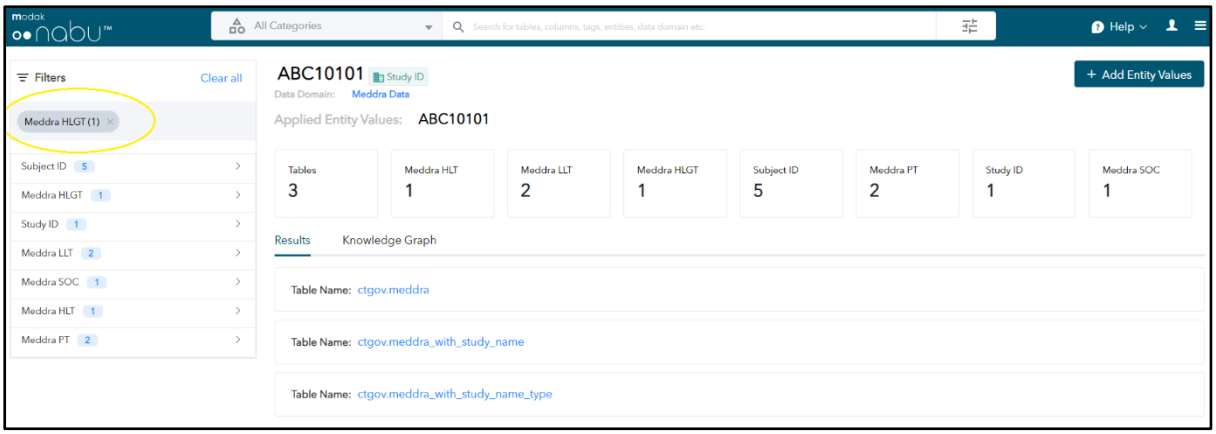


Figure 278: Nabu Search - entity filters applied

Filters defined for the searched entity can be seen by clicking on the arrow icon next to the filter name, on the left panel. For e.g., to view filters applicable for ABC10101, which is a value of an entity 'Study ID', click on the arrow icon next to the 'Study ID' on the left panel. The filters defined for 'Study ID' entity, like 'Age', 'Country' are shown in the screen below.

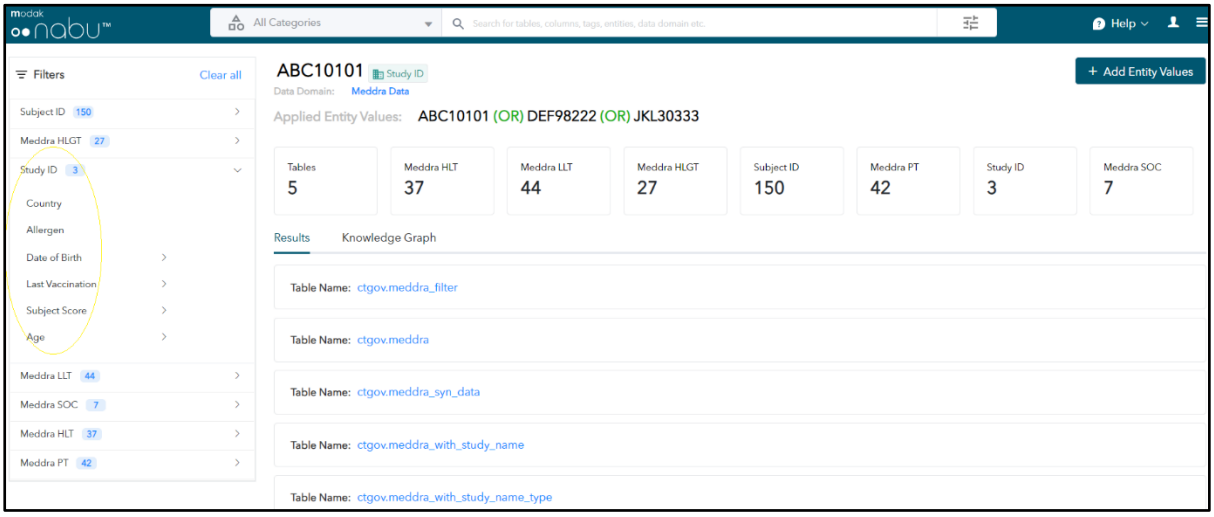


Figure 279: Nabu Search – Filters for entities

Depending on data type, cardinality the filters are of different types, such as range boxes, slider, multi-select etc.

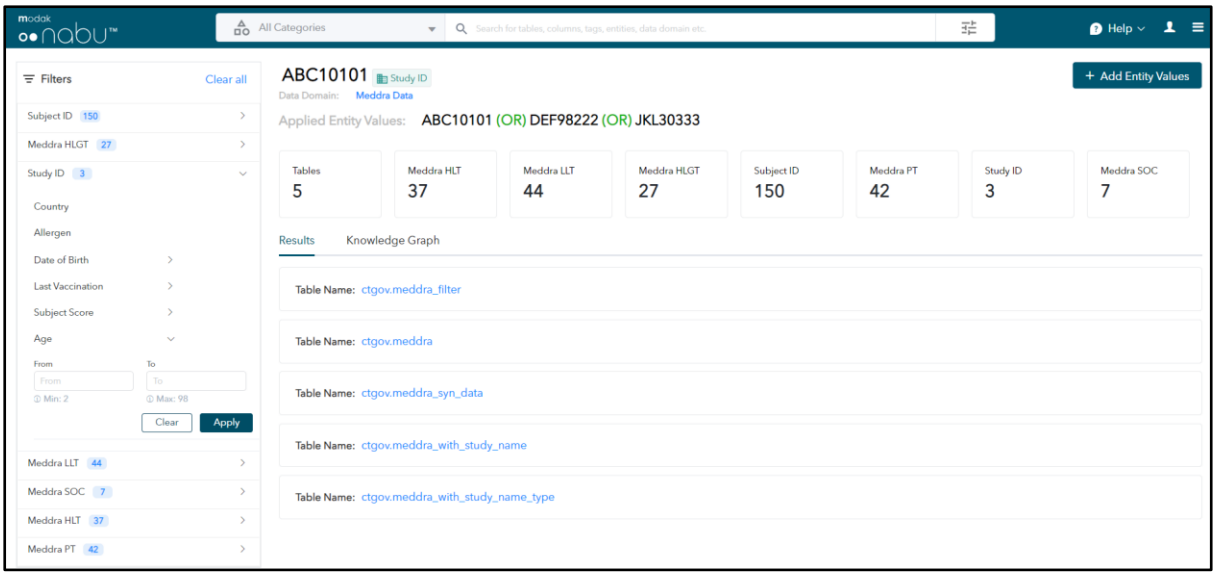


Figure 280: Nabu Search – Range box filter

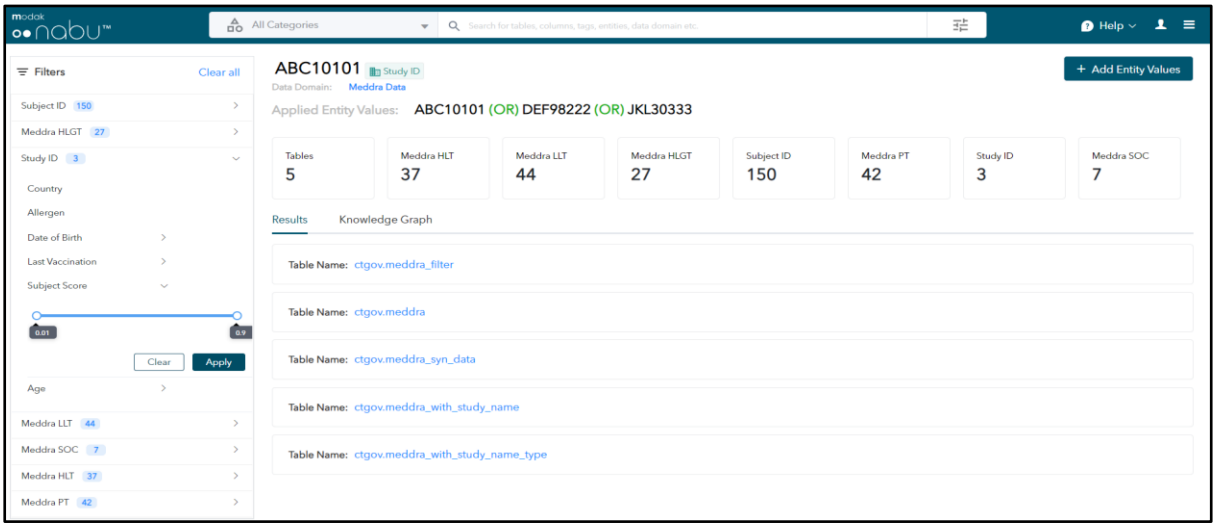


Figure 281: Nabu Search – slider filter

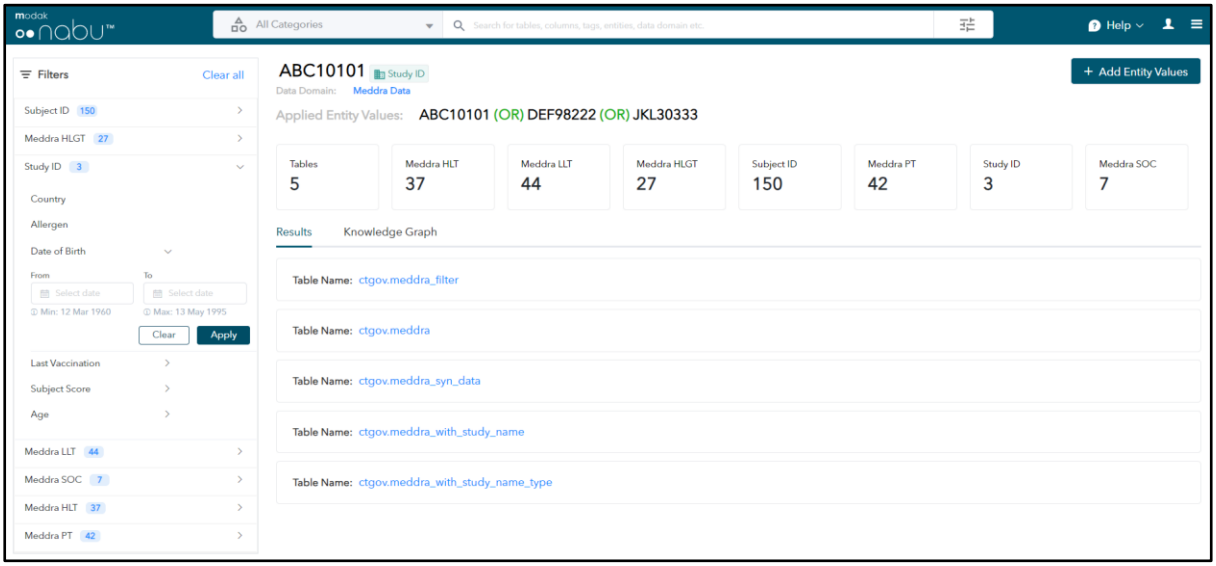


Figure 282: Nabu Search - select filter

The applied filters are shown on the left side in the filters section as below.

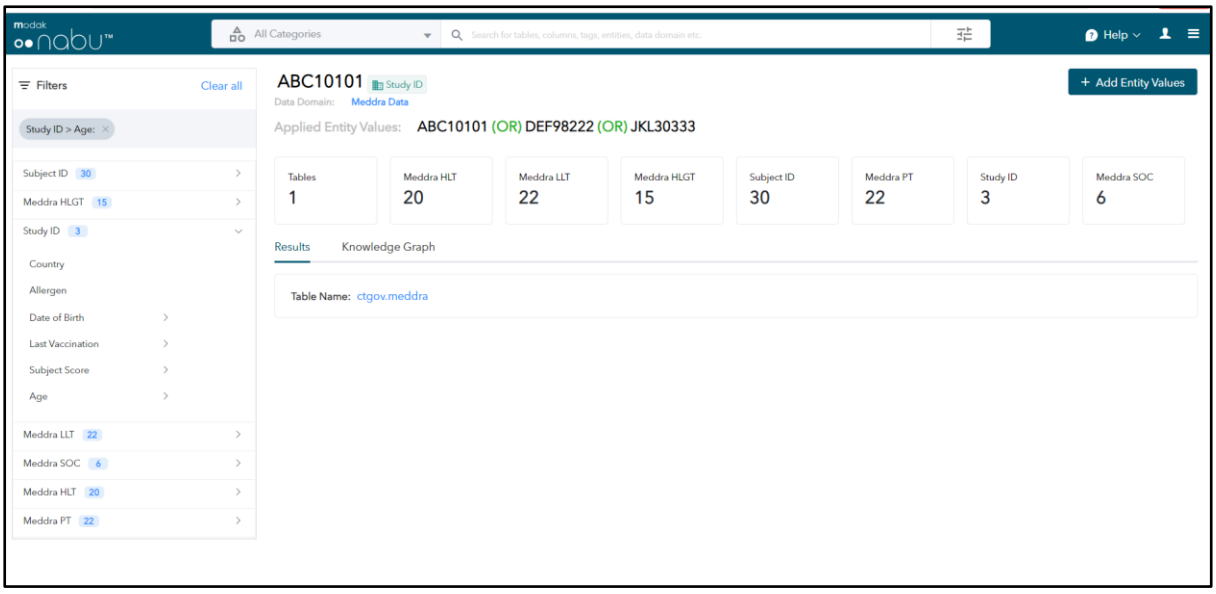


Figure 283: Nabu Search – Applied filters

11.5 Search for Synonym

'Synonym' in Nabu is an alias for an entity. Search results of a synonym are the same as that of the entity for which it is an alias. Synonyms are defined in the 'Data Catalogue' section.

To search for a synonym, type few characters from its name. The suggestions in the autocomplete search box show the synonyms that match the characters typed. The tags next to the suggested synonym show the entity name and entity value for the synonym.

For e.g., in the screen below, as the user types 'Athena', the first result in the autocomplete box shows that it is a synonym of Study Id - ABC10101.

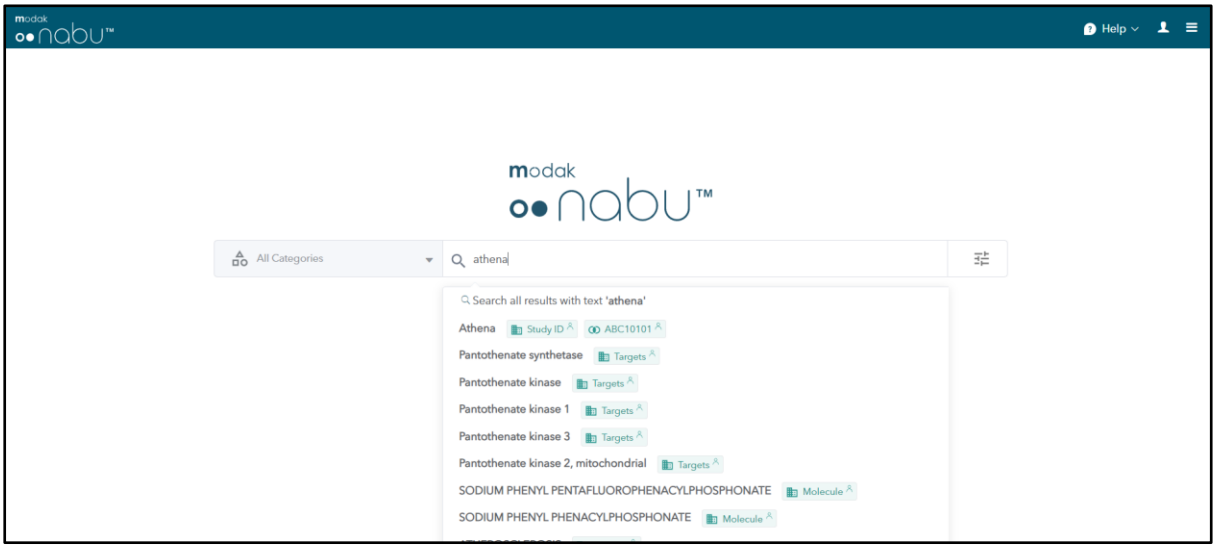


Figure 284: Nabu Search – Synonym search

Clicking on the suggested result, shows the same results as that for entity value 'ABC10101'

11.6 Search for Facet

Facet is a group of similar columns in a data domain. Facets are defined in the 'Data Catalogue' section.

To search for a facet, type few characters from its name. Any facet that matches the entered characters will be shown in the autocomplete search box.

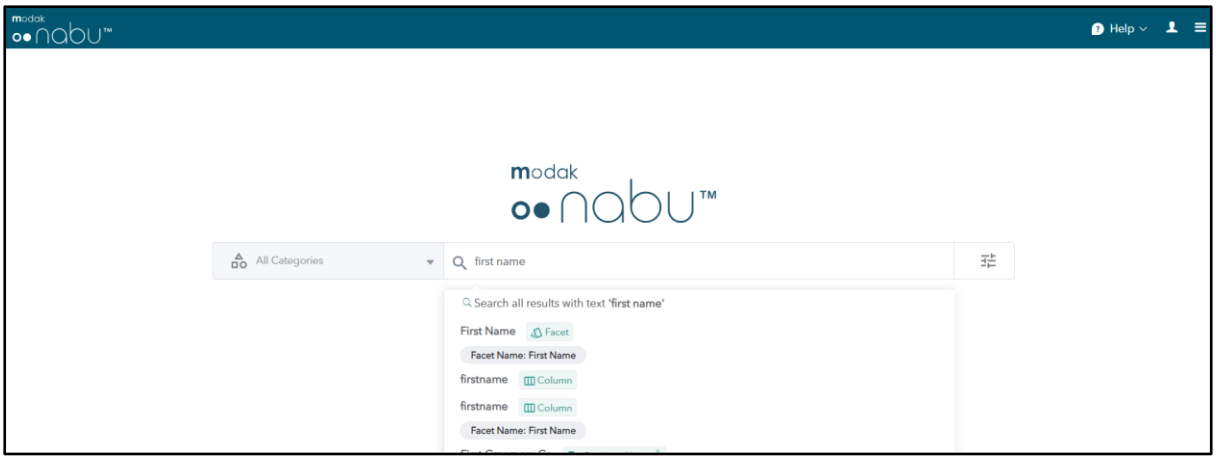


Figure 285: Nabu search – Facet search

Clicking on the facet name, shows number of columns which are grouped under the facet, number of tables those columns are in, any tags applied to the facets and the name of the data domain on which the facet is defined.

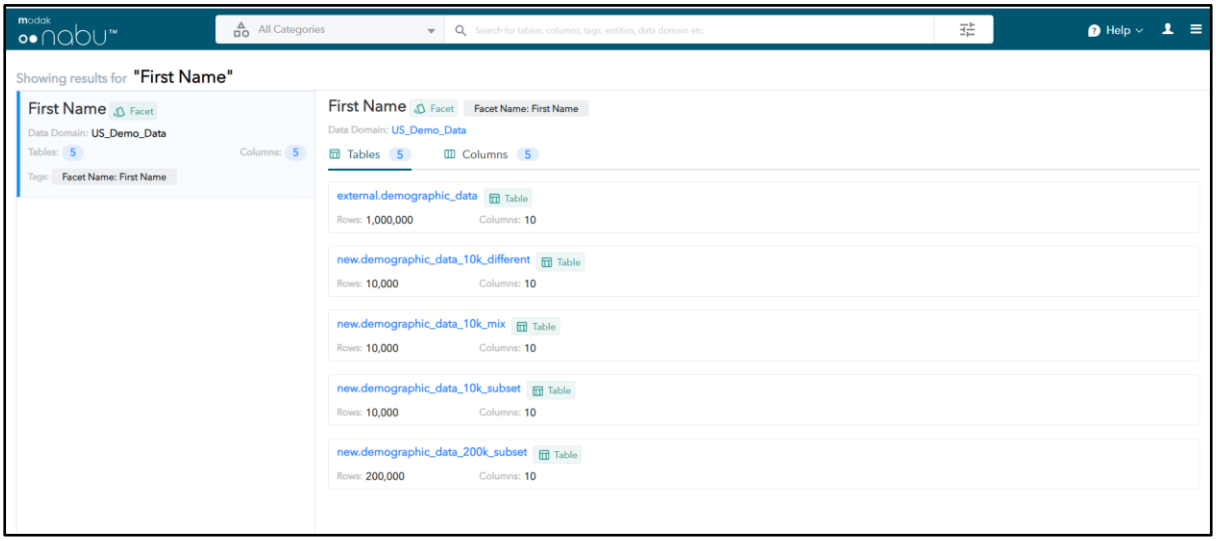


Figure 286 : Facet Search results

To view the list of tables, columns, click on 'Tables' or 'columns' tabs.

11.7 Search for Data Domain

A Data Domain in Nabu is a logical grouping of one or more data connections, and selected tables/files within them. To search for a data domain, type few characters from its name in the search box. The suggestions shown includes data domain whose name matches with the entered characters. A 'Data Domain tag next to the suggestion identifies that the result is for a data domain.

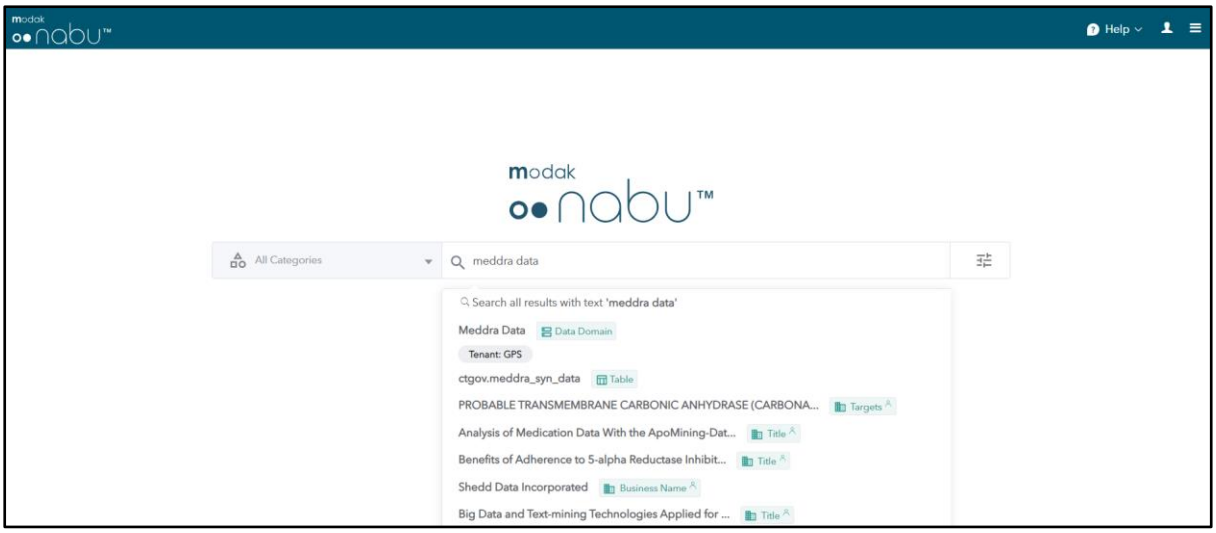


Figure 287: Nabu Search – Data Domain Search

Clicking on the data domain result shows the number and the list of tables as below.

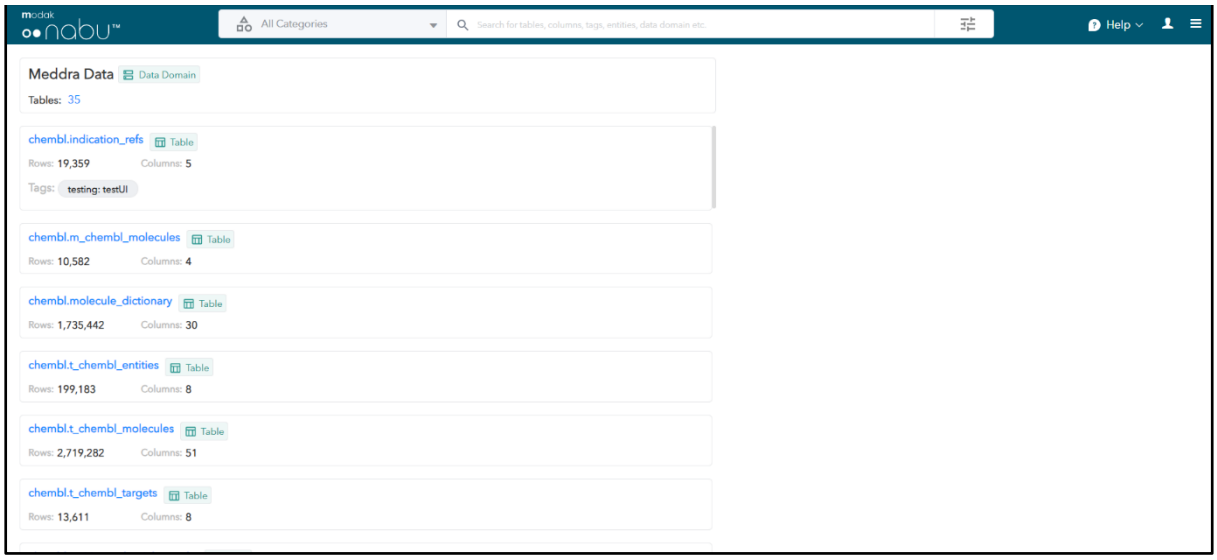


Figure 288: Nabu Search – Data Domain Search results

Search for Column

To search for a column in a table, type few characters from its name in the search box. The suggestions shown includes columns whose name matches with the entered characters. A 'Column' tag next to the suggested results, identifies that the result is for a column. The Tags that are defined for the column will also be shown as below.

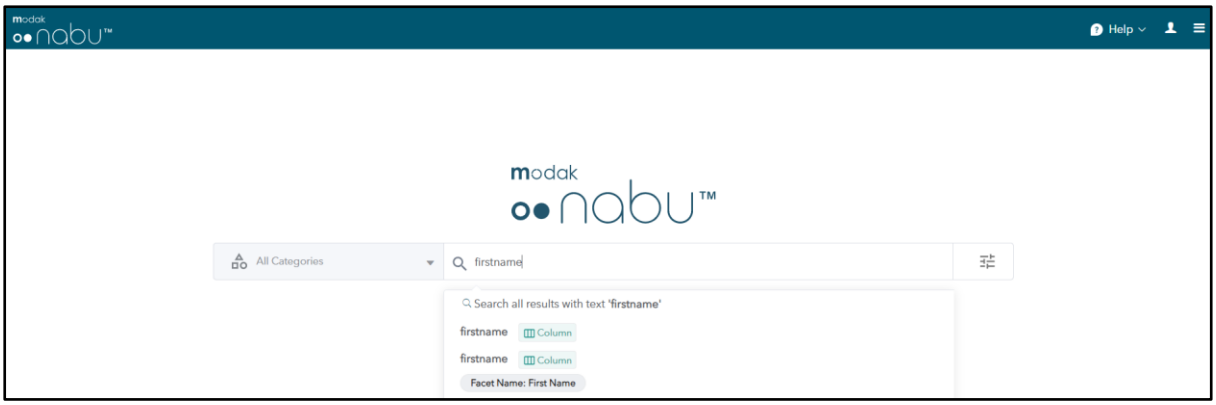


Figure 289: Nabu Search – Column Search

Clicking on the column result, shows the list of tables, number of rows and columns for each table. Any tags applied to the column are also shown.

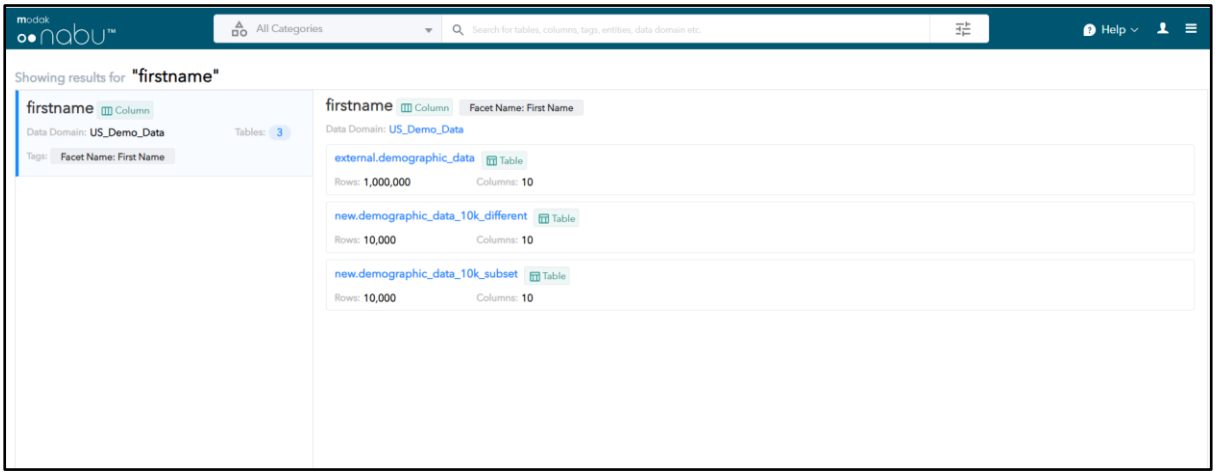


Figure 290: Nabu Search – Column Search results

11.8 Filter Tags

Tags are applied on a table, column, data domain, entity or facet and provide business context. Tags are in the form of a key value pair, where the key is tag category and value is tag value. The category and value are defined by the users. For e.g., a table containing sensitive data can have the tag 'PII: Sensitive'. Here the tag category is 'PII', and tag value is 'Sensitive'.

To find the tag defined for a table, column or data domain, click on the filter on the right of the search bar. It shows the options to enter the tag category, tag value and search on the tag value or category in the search as below. User cannot directly search for the tags but can filter the search results based on tags.

To filter the entities based on tags, the user can provide the entities in the entities field and search.

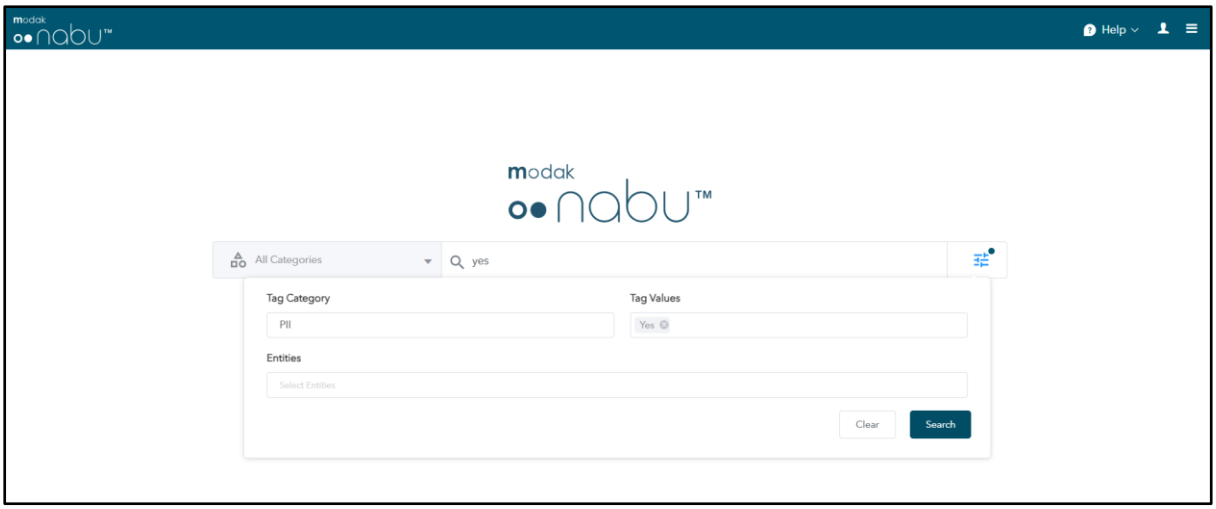


Figure 291: Nabu Search – Filter with tags

It shows the options as below.

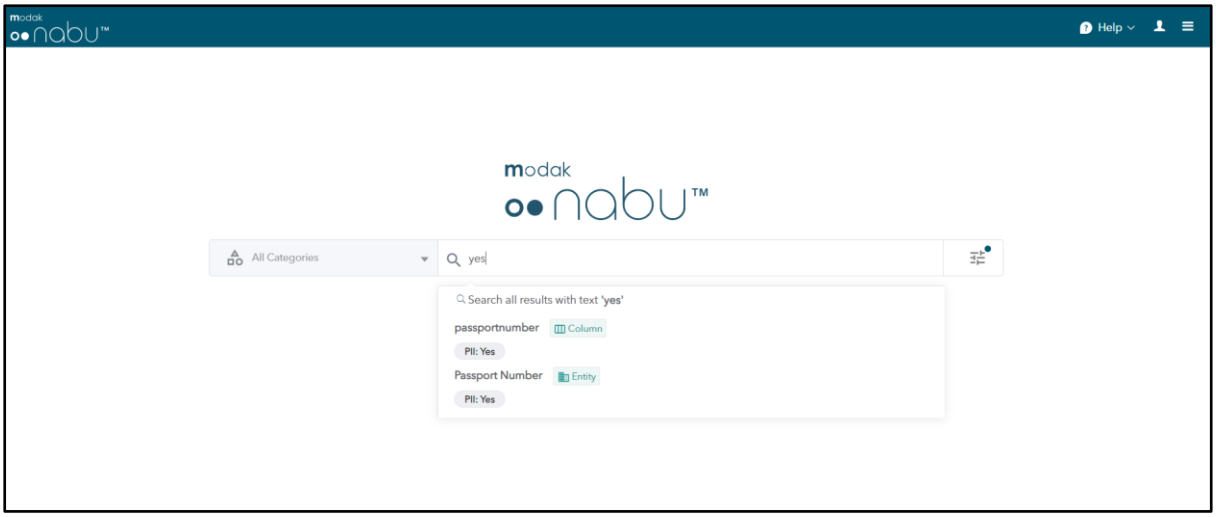


Figure 292: Nabu Search – Filter with tag value