



Informatica® Metadata Command Center
November 2025

Apache Hive Sources

© Copyright Informatica LLC 2023, 2025

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

Informatica, Informatica Cloud, Informatica Intelligent Cloud Services, PowerCenter, PowerExchange, and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at <https://www.informatica.com/trademarks.html>. Other company and product names may be trade names or trademarks of their respective owners.

Portions of this software and/or documentation are subject to copyright held by third parties. Required third party notices are included with the product.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at infa_documentation@informatica.com.

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Publication Date: 2025-11-20

Table of Contents

Preface	5
Chapter 1: Introduction to Apache Hive catalog sources.....	6
Extraction and view process.	7
About the Apache Hive catalog source	7
Extracted metadata.	8
Data profiling for Apache Hive objects.	8
Compatible connectors.	9
Chapter 2: Before you begin.....	10
Verify permissions.	10
Permissions to extract metadata.	10
Permissions to run data profiles.	10
Permissions to run data classification.	11
Permissions to run glossary association.	11
Configure Kerberos authentication.	11
Create a connection.	11
Create endpoint catalog sources for connection assignment.	13
Chapter 3: Create catalog sources in Metadata Command Center.....	14
Step 1. Register a catalog source.	14
Step 2. Configure capabilities.	15
Configure metadata extraction.	16
Configure lineage discovery.	17
Configure data profiling and quality.	18
Configure data classification.	22
Configure glossary association.	22
Step 3. Associate stakeholders and asset groups.	23
Step 4. Run or schedule the job.	25
Step 5. Assign reference catalog source connections to endpoint catalog source objects.	26
Chapter 4: View results in Data Governance and Catalog.....	28
View metadata extraction results.	28
View data lineage.	30
View lineage at the catalog source level.	31
View lineage at data set level.	31
View lineage at data element level.	31
View data profiling results.	32
View data observability results.	33
View classified data.	34

View glossary associations. 34

Preface

Read *Apache Hive Sources* to learn how to register and configure Apache Hive sources as catalog sources in Metadata Command Center. After you configure a catalog source, you extract metadata and then view the results in Data Governance and Catalog.

CHAPTER 1

Introduction to Apache Hive catalog sources

You can use Metadata Command Center to extract metadata from a source system.

A source system is any system that contains data or metadata. For example, Apache Hive is a source system from which you can extract metadata through an Apache Hive catalog source with Metadata Command Center. A catalog source is an object that represents and contains metadata from the source system.

Before you extract metadata from a source system, you first create and register a catalog source that represents the source system. Then you configure capabilities for the catalog source. A capability is a task that Metadata Command Center can perform, such as metadata extraction, lineage discovery, data profiling, data classification, or glossary association.

When Metadata Command Center extracts metadata, Data Governance and Catalog displays the extracted metadata and its attributes as technical assets. You can then perform tasks such as analyzing the assets, viewing lineage, and creating links between those assets and their business context.

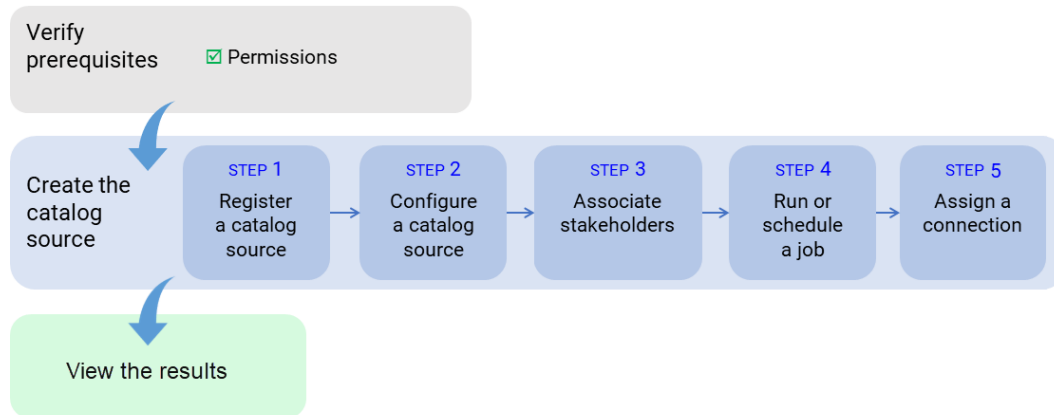
The following table describes the capabilities of the catalog source:

Capability	Description
Lineage Discovery	Builds the complete lineage of a catalog source by recommending endpoint catalog source objects to assign to reference catalog source connections. When you run the catalog source job, Metadata Command Center assigns the reference catalog source connections to CLAIRE recommended endpoint catalog source objects. You can then view the list of CLAIRE recommendations and accept or reject them.
Data Profiling and Quality	<ul style="list-style-type: none">- Data Profiling. Assesses source metadata and analyzes the collected statistics to discover content and structure, such as value distribution, patterns, and data types.- Data Quality. Measures the reliability of the data and enables data usage.- Data Observability. Identifies anomalies in the characteristics of the data.
Data Classification	Data classification is the process of identifying and organizing data into relevant categories based on the functional meaning of the data. Classifying data can help your organization manage risks, compliance, and data security.
Glossary Association	You can associate terms that are in the glossary with technical assets to provide user-friendly business names to technical assets. Glossary Association automatically associates glossary terms with technical assets or recommends glossary terms that you can manually associate with technical assets in Data Governance and Catalog.

Extraction and view process

To extract metadata from a source system, configure the catalog source and run the extraction job in Metadata Command Center. Then view the results in Data Governance and Catalog.

The following image shows the process to extract metadata from a source system:



After you verify prerequisites, perform the following tasks to extract metadata from Apache Hive:

1. Register a catalog source. Create a catalog source object, select Apache Hive, and then select and test the connection.
2. Configure the catalog source. Specify the runtime environment and configure parameters for metadata extraction. Optionally, add filters to include or exclude source system assets from metadata extraction. You can also configure other capabilities such as data profiling and quality, data classification, or glossary association.
3. Optionally, associate stakeholders. Associate users with technical assets, giving the users permission to perform actions determined by their roles.
4. Run or schedule the catalog source job.
5. Optionally, if the catalog source job generates referenced asset objects, you can assign a connection to referenced source system assets.
You can view the lineage with object references without performing connection assignment. After connection assignment, you can view the objects.

After you run the catalog source job, you view the results in Data Governance and Catalog.

About the Apache Hive catalog source

You can use the Apache Hive catalog source to extract metadata from an Apache Hive source system.

Apache Hive is a data warehouse software built on Apache Hadoop, used to query large datasets from various databases and file systems.

Extracted metadata

You can use the Apache Hive catalog source to extract metadata from an Apache Hive source.

Metadata Command Center extracts the following metadata from the Apache Hive source system:

- Database
- Schema
- Table
- Column
- External Table
- External Column
- View
- View Column
- Materialized View

Note: Objects of the Materialized View type appear as View in Data Governance and Catalog.

Data profiling for Apache Hive objects

Configure data profiling to run profiles on the metadata extracted from an Apache Hive source system. You can view the profiling statistics in Data Governance and Catalog.

You can run data profiles on the following objects:

- Views
- Tables
- External tables created in the following file formats:
 - AVRO
 - Parquet
 - Delimited
 - JSON

The data profiling task runs profiles on the following data types:

- bigint
- boolean
- char
- date
- decimal
- double
- float
- integer
- smallint
- string
- timestamp
- tinyint
- varchar

Compatible connectors

Before you configure an Apache Hive catalog source, you must connect to the Apache Hive source system.

Use Hive Connector to connect to the Apache Hive source system.

For information about configuring a connection, see *Connections* in the Administrator service.

CHAPTER 2

Before you begin

Before you can extract catalog source metadata, get information from the Apache Hive administrator.

Perform the following prerequisite tasks:

- Verify permissions.
- Configure authentication.
- Create a connection.

Verify permissions

To extract metadata and to configure other capabilities that a catalog source might include, you need account access and permissions on the source system. The permissions required might vary depending on the capability.

Permissions to extract metadata

To extract Apache Hive metadata, you need access to the Apache Hive source.

Verify that the Cloudera CDP, Amazon EMR, Google Cloud Dataproc, or Azure HDInsight cluster user has read permission on the source.

Grant permissions that allow you to perform the following operations:

- show schemas
- show tables
- show views
- show materialized views

Permissions to run data profiles

Ensure that you have the required permissions to run profiles.

Grant read permissions to the cluster user for all objects on which you want to run data profiles.

Also, grant write permission to the connection object used by the intermediate staging connection to write profiling results temporarily.

Permissions to run data classification

You can perform data classification with the permissions required to perform metadata extraction.

Permissions to run glossary association

You can perform glossary association with the permissions required to perform metadata extraction.

Configure Kerberos authentication

If you use Kerberos authentication, provide the Kerberos principal for authentication when you configure the Apache Hive catalog source in Metadata Command Center. Also, configure the Secure Agent machine to work with the Kerberos Key Distribution Center (KDC).

Verify that you have the URL to connect to Apache Hive.

1. Open the hosts file located on the Secure Agent machine.
On a Windows machine, the hosts file is available in the following path: `C:\Windows\System32\drivers\etc\hosts`
On a Linux machine, the hosts file is available in the following path: `/etc/hosts`
2. Add details of the Kerberos server to the hosts file in the following format: `<IP address> <Host name>`
3. Add the KDC server IP address to the file in the following format: `<KDC server IP address> <Fully qualified name of the KDC server> <Alias name>`
4. Save and close the file.
5. Verify that the Kerberos configuration file is available on the Secure Agent machine.
On a Windows machine, the `krb5.ini` configuration file is available in the following path: `C:\Windows`
On a Linux machine, the `krb5.conf` configuration file is available in the following path: `/etc`
6. Copy `hive.keytab`, `core-site.xml`, and `hive-site.xml` files from the Hadoop cluster node to a directory on the Secure Agent machine.
7. Download the Keytab file from the Kerberos administrator and copy it to a directory on the Secure Agent machine.

Create a connection

Create a Hive Connector connection object in Administrator.

Before you create a connection, configure the Hive Connector to download the Hive third-party libraries for Cloudera CDP, Amazon EMR, or the Azure HDInsight cluster. You can also connect to Apache Hive source systems hosted on the Google Cloud Dataproc cluster using non-Kerberos authentication. For more information about the Hive Connector, see Hive Connector in the *Data Integration Connectors* help.

1. In Administrator, select **Connections**.
2. Click **New Connection**.

3. In the **Connection Details** section, enter the following connection details:

Connection property	Description
Connection Name	Name of the connection. Each connection name must be unique within the organization. Connection names can contain alphanumeric characters, spaces, and the following special characters: _ . + -, Maximum length is 255 characters.
Description	Description of the connection. Maximum length is 4000 characters.

4. Select the Hive Connector connection type.
5. In the **Hive Connector Properties** section, select the runtime environment where you want to run the tasks.
6. In the **Connection** section, enter the following connection details:

Connection property	Description
Authentication Type	You can select one of the following authentication types: <ul style="list-style-type: none"> - Kerberos. Select Kerberos for a Kerberos cluster. - LDAP. Select LDAP for an LDAP-enabled cluster. <p>Note: LDAP is not applicable to mappings in advanced mode.</p> <ul style="list-style-type: none"> - None. Select None for a Hadoop cluster that is not secure or not LDAP-enabled.
JDBC URL *	The JDBC URL to connect to Hive. Specify the following format based on your requirement: <ul style="list-style-type: none"> - To view and import tables from a single database, use the following format: jdbc:hive2://<host>:<port>/<database name> - To view and import tables from multiple databases, do not enter the database name. Use the following JDBC URL format: jdbc:hive2://<host>:<port>/ <p>Note: After the port number, enter a slash.</p> <ul style="list-style-type: none"> - To access Hive on a Hadoop cluster enabled for TLS, specify the details in the JDBC URL in the following format: jdbc:hive2://<host>:<port>/<database name>;ssl=true;sslTrustStore=<TrustStore_path>;trustStorePassword=<TrustStore_password>, where the truststore path is the directory path of the truststore file that contains the TLS certificate on the agent machine.
JDBC Driver *	The JDBC driver class to connect to Hive.
Username	The user name to connect to Hive in LDAP or None mode.
Password	The password to connect to Hive in LDAP or None mode.
Principal Name	The principal name to connect to Hive through Kerberos authentication.

Connection property	Description
Keytab Location	The path and file name to the Keytab file for Kerberos login.
Configuration Files Path *	<p>The directory that contains the Hadoop configuration files for the client.</p> <p>Copy the site.xml files from the Hadoop cluster and add them to a folder in the Linux box. Specify the path in this field before you use the connection in a mapping to access Hive on a Hadoop cluster:</p> <ul style="list-style-type: none"> - For mappings, you require the core-site.xml, hdfs-site.xml, and hive-site.xml files. - For mappings in advanced mode, you require the core-site.xml, hdfs-site.xml, hive-site.xml, mapred-site.xml, and yarn-site.xml files.

* These fields are mandatory parameters.

7. Click **Test Connection**.
8. Click **Save**.

Create endpoint catalog sources for connection assignment

An endpoint catalog source represents a source system that the catalog source references. Before you perform connection assignment, create endpoint catalog sources and run the catalog source jobs.

You can then perform connection assignment to reference source systems to view complete lineage with source system objects.

CHAPTER 3

Create catalog sources in Metadata Command Center

Use Metadata Command Center to configure a catalog source for Apache Hive and extract metadata.

When you configure a catalog source, you define the source system that you want to extract metadata from. Optionally, configure filters to include or exclude source system metadata before you run the job.

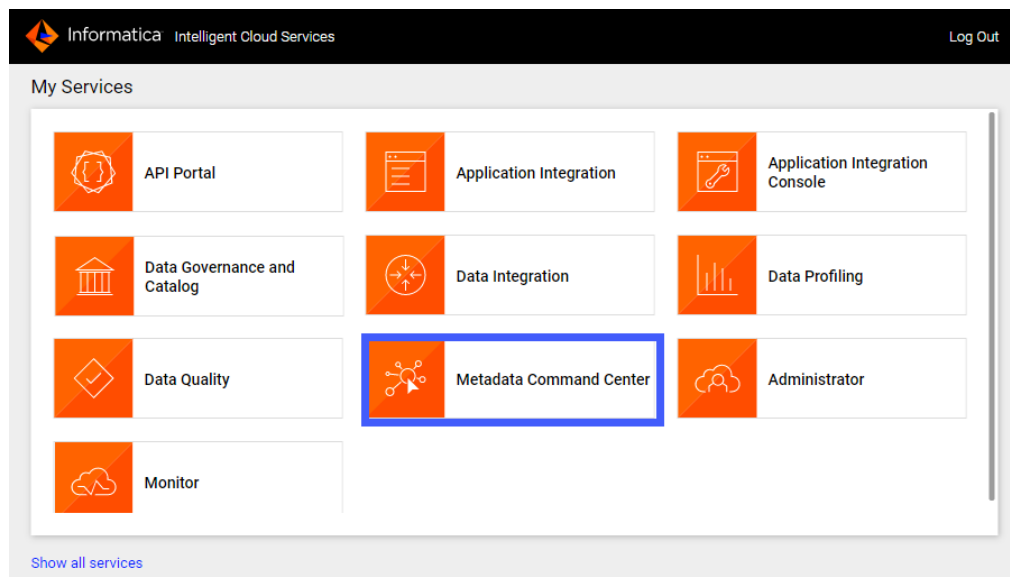
To provide stakeholders access to technical assets, you can assign access through roles. To view lineage for any system that the source system references, create a catalog source and a connection associated with the referenced source system after you run the job.

Step 1. Register a catalog source

When you register a catalog source, provide general information and connection values.

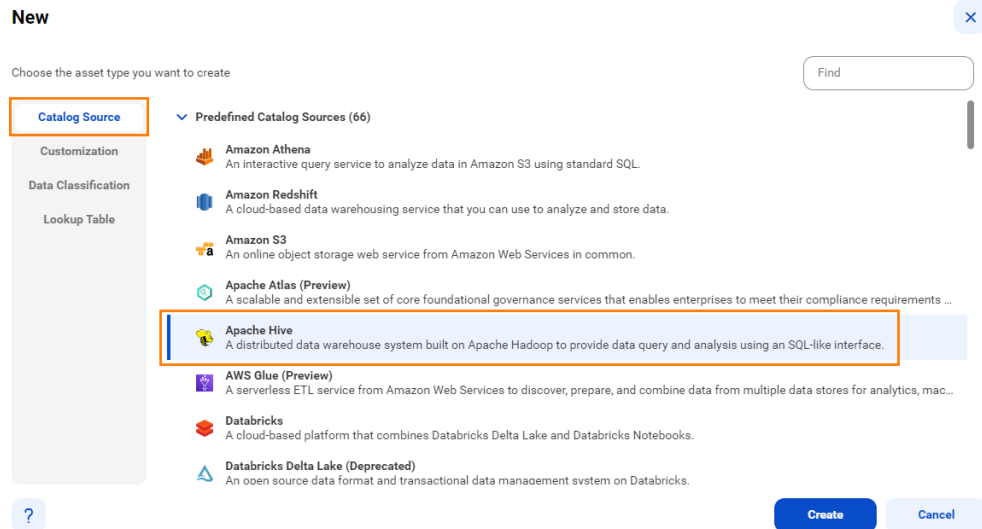
1. Log in to Informatica Intelligent Cloud Services.
The **My Services** page appears.
2. Click **Metadata Command Center**.

The following image shows the Metadata Command Center box on the **My Services** page:



The Metadata Command Center home page appears.

3. Click **New**.
4. Select **Catalog Source** from the list of asset types.
5. Select Apache Hive from the list of catalog source types.



6. Click **Create**.
7. The **New Catalog Source** page opens.
7. In the **General Information** section, enter a name and an optional description for the catalog source.
Note: You can rename a catalog source after you create it, but to apply the change to all associated objects you must rerun the metadata extraction job.

After you save the catalog source, you can update the description in Metadata Command Center and Data Governance and Catalog. The update appears only in the service in which you update it.

8. In the **Connection Information** area, select the connection that you created in Administrator.
Note: To create or edit a catalog source, you need permissions on the connection to the source system. Select a connection that you have access to, or ask the administrator to grant the necessary permissions to the connection that you want to use.
9. Click **Connection Properties** to expand and view the connection properties for the selected connection.
10. Click **Test Connection** to test your connection to the source system.
11. Click **Next**.

The **Configuration** page appears.

Step 2. Configure capabilities

When you configure the Apache Hive catalog source, you define the settings for the metadata extraction capability and other optional capabilities.

The metadata extraction capability extracts source metadata from external source systems. You can also configure other capabilities that the catalog source includes.

You can save the catalog source configuration at any point after you enter the connection information. After you save the catalog source, you can choose to run the catalog source job. To run the job once, click **Run**. To run metadata extraction and other capabilities on a recurring schedule, configure schedules on the **Schedule** tab.

Configure metadata extraction

When you configure the Apache Hive catalog source, you choose a runtime environment, define filters, and enter configuration parameters for metadata extraction.

1. In the **Connection and Runtime** area, choose a serverless runtime environment or the Secure Agent group where you want to run catalog source jobs.
Note: Serverless runtime environment options are available if the catalog source works with a serverless runtime environment.
2. Choose to retain, delete, or deprecate objects that are deleted from the source system in the catalog with the **Metadata Change Option**.

- **Retain.** Retains objects that are deleted from the source system in the catalog. If you update or add a filter, the catalog retains objects extracted from the previous job and extracts additional objects that match the current filter. Objects deleted from the source system are not deleted from the catalog. Enrichments added on deleted objects and relationships are retained.
- **Delete.** Deletes metadata from the catalog based on objects deleted from the source system and changes you make to the filter. Enrichments added on deleted objects and relationships are also permanently lost. Objects renamed in the source system are removed and recreated in the catalog.
- **Deprecate.** The lifecycle of objects imported into the catalog moves to Obsolete based on objects deleted from the source system and changes you make to the filter. This does not impact enrichments added on deprecated objects and relationships. Objects renamed in the source system are removed and recreated in the catalog. When you run the catalog source job again for other capabilities such as data classification, relationship discovery, or glossary association, the job doesn't consider obsolete objects. Obsolete objects remain in the catalog until they are purged when you run a **Purge Obsolete Objects** job on the **Explore** page.

Note: You can also change the configured metadata change option when you run a catalog source.

3. In the **Filters** area, define one or more filter conditions to apply for metadata extraction:
 - a. From the **Include or exclude metadata** list, choose to include or exclude metadata based on the filter parameters.
 - b. From the **Object type** list, select Tables, Views, or External Tables, depending on the object that you want to extract metadata from. Select All to extract metadata from all objects in the schema.
 - c. Enter the path to the object as the filter value.
Filters can contain the following wildcards:
 - Question mark. Represents a single character.
 - Asterisk. Represents multiple characters or empty text.

The following image shows the filter options:

Filters

Specify metadata filters: ☐ No ☒ Yes

> Show supported wildcards and examples

Include or exclude ...

Select the object ty...

Enter a value to specify the object location

+

🗑

- d. Optionally, to define an additional filter with an OR condition, click the **Add** icon.

The following image shows the filter conditions for an Apache Hive catalog source that includes metadata from all objects within the schema with names that start with EMPLOYEEDATA. The filter condition also excludes metadata from tables with names that start with department followed by a single character located in the schema EMPLOYEEDATA.

Filters

Specify metadata filters: ☐ No ☒ Yes

> Show supported wildcards and examples

Include Metadata	All	EMPLOYEEDATA *	+	🗑
Exclude Metadata	All	EMPLOYEEDATA .department?	+	🗑

4. Optionally, in the **Configuration Parameters** area, enter properties to override default context values and job parameters.

The following table describes the property that you enter for additional settings.

Note: The **Additional Settings** section appears when you click **Show Advanced**.

Property	Description
Expert Parameters	Enter additional configuration options to be passed at runtime. Required if you need to troubleshoot the catalog source job. Caution: Use expert parameters when it is recommended by Informatica Global Customer Support.

5. Configure additional capabilities for the catalog source by clicking on the tabs.

Configure lineage discovery

Enable the lineage discovery capability and use CLAIRE to build complete lineage by recommending endpoint catalog source objects to assign to reference catalog source connections.

1. Click the **Lineage Discovery** tab.
2. Select **Enable Lineage Discovery**.
3. In the **Filters** area, define one or more filter conditions to apply for lineage discovery.

To define filters, you can choose to select catalog source types, asset groups, or enter a catalog source name or search from a list of catalog sources.

- a. Select **Yes** to view filter options.
- b. From the Include/Exclude list, choose to include or exclude catalog sources for lineage discovery based on the filter parameters.
- c. From the filter type list, select catalog source type, catalog source name, or asset group.
- d. In the filter value field, select the required catalog source types, or click the Search button and select catalog sources or asset groups.

Filters can contain the asterisk wildcard to represent multiple characters or empty text.

The following image shows the filter condition options:

Enable Lineage Discovery: ☒

Filters

Specify lineage discovery filters: ☒ No ☒ Yes

> Show supported wildcards and examples

Include	Catalog Source Type	Select Catalog Source Types	+	-
Exclude	Catalog Source Name	Select Catalog Sources	+	-
Exclude	Asset Group	Select Asset Groups	+	-

Examples:

- To include or exclude all Oracle catalog sources, select **Catalog Source Type** as the filter type and select `Oracle` in the filter value field.
- To include or exclude the 'Oracle_Retail' catalog source, select **Catalog Source Name** as the filter type and search for the catalog source or enter `Oracle_Retail` in the filter value field.
- To include or exclude all catalog sources with names that start with 'Oracle', select **Catalog Source Name** as the filter type and search for the catalog source or enter `Oracle*` in the filter value field.
- To include or exclude all catalog sources with names that end with 'Retail', select **Catalog Source Name** as the filter type and search for the catalog source or enter `*Retail` in the filter value field.
- To include or exclude all catalog sources with names that contain 'Ret', select **Catalog Source Name** as the filter type and search for the catalog source or enter `*Ret*` in the filter value field.
- To include or exclude all catalog sources that are part of the 'Financial Group' asset group, select **Asset Group** as the filter type and search `Financial Group` in the filter value field.

Note: You can't add more than one include or exclude filter for the same filter type.

- e. Optionally, to define an additional filter with an AND condition, click the **Add** icon.

For more information about lineage discovery, see *Lineage discovery* in the *Administration* help.

Configure data profiling and quality

Enable the data profiling capability to evaluate the quality of metadata extracted from the Apache Hive source system.

You can run data profiling and quality capabilities on Apache Hive on an elastic cluster.

1. Click the **Data Profiling and Quality** tab.
2. Expand **Data Profiling** and select **Enable Data Profiling**.

Note: Ensure that you have permissions on all the staging connections that you use in your data profiling configuration. You can't run the job if you don't have permissions on the connections that you use. Select connections that you have access to, or ask the administrator to grant the necessary permissions on the connections that you want to use.

3. Optional. In the **Filters** area, specify additional filters in addition to metadata filters:
 - a. Select **Yes** to view filter options.
 - b. From the Include/Exclude list, choose to include or exclude metadata based on the filter parameters.
 - c. From the object type list, select Tables, Views, or External Tables depending on the object that you want to extract metadata from. Select All to extract metadata from all objects in the schema.

- d. Enter the path to the object as the filter value.

Examples:

- You extracted metadata of all tables and views from a schema and now you want to profile a specific table or view from the schema. Select All/Tables/Views from the Object type option and then enter the Schema name followed by the table/view name in the input field. For example, `Schema_Name.TABLE_NAME` or `Schema_Name.VIEW_NAME`
- You extracted metadata from multiple schemas and now you want to run a profile on all the objects in a specific schema. Select All from the Object type option and then enter the Schema name in the input field.

To include or exclude multiple objects, click the **Add** icon to add filters with the OR condition.

4. In the **Parameters** area, configure the following parameters based on your requirements:

Parameter	Description
Modes of Run	Determines the type of data that you want the data profiling task to collect. Choose one of the following options: <ul style="list-style-type: none">• Keep signatures only. Collects only aggregate information such as data types, average, standard deviation, and patterns.• Keep signatures and values. Collects both signatures and data values.
Profiling Scope	Determines whether you want to run data profiling only on the changes made to the source system or on the entire source system. Choose one of the following options: <ul style="list-style-type: none">• Incremental. Includes only source metadata that is changed or updated since the last profile run.• Full. Includes the entire metadata that is extracted based on the filters applied for extraction.
Sampling Type	Determines the sample rows on which you want to run the data profiling task. Choose any of the following options: <ul style="list-style-type: none">• All rows. Runs data profiling on all rows in the metadata.• Limit N Rows. Runs data profiling on a limited number of rows.• Custom Query. Provides an SQL clause to select sample rows to run the data profiling task. For example, <code>where column1='X'; TABLESAMPLE (X ROWS); TABLESAMPLE (X PERCENT)</code>
No of rows to limit	Required if you select Limit N Rows in Sampling Type. Specify the number of rows that you want to run the profile on. Default is 1000.
Sampling Query	Required if you select Custom Query in Sampling Type. Specify an SQL clause to select sample rows to run the data profiling task.
Elastic Runtime Environment	Select an elastic runtime environment to run data profiling and quality capabilities on Apache Hive in the elastic cluster configuration. Note: To run a profile, use an elastic runtime environment configured with an advanced cluster. For more information about advanced clusters, see Advanced Clusters help.

Parameter	Description
Staging Connection	The staging connection where data profiling results are stored temporarily during the profile run. If the Apache Hive source system is hosted on a Cloudera CDP or an Amazon EMR cluster, configure Amazon S3 v2 as the staging connection. If the Apache Hive source system is hosted on an Azure HDInsight cluster, configure Microsoft Azure Data Lake Storage Gen2 as the staging connection.
Maximum Precision of String Fields	The maximum precision value to be used for profiling fields that include the string data type. Default is 50.
Text Qualifier	The character that defines string boundaries. If you select a quote character, profiling ignores delimiters within the quotes. Select a qualifier from the list. Default is Double Quote.

- Expand **Data Quality** and select **Enable Data Quality**.

Note: You can click **Use Data Profiling Parameters** to use the same parameters as in the **Data Profiling** section.

Note: Ensure that you have permissions on all the staging and flat file connections that you use in your data quality configuration. You can't run the job if you don't have permissions on the connections that you use. Select connections that you have access to, or ask the administrator to grant the necessary permissions on the connections that you want to use.

- In the **Parameters** area, configure the following parameters based on your requirements:

Parameter	Description
Data Quality Rule Automation	Enable the option to automatically create or update rule occurrences for data elements in the catalog source. Choose one of the following options: <ul style="list-style-type: none"> • Apply on Data Elements linked with Business Dataset. Creates rule occurrences for all data elements that are linked with business data sets in the catalog source. • Apply on all Data Elements. Creates rule occurrences for all data elements in the catalog source.
Data Quality Remediation	Enable the option to specify a flat file connection to store the list of failed rows so that users can remediate poor data quality scores. Choose one of the following options: <ul style="list-style-type: none"> • No. Doesn't enable the data quality failure ticket option. • Yes. Shows a list of flat file connections where you write failed rows to customer-managed locations.

Parameter	Description
Data Quality Failure Ticket	<p>Specify whether you want to create data quality failure tickets for poor data quality scores based on the threshold defined for the rule occurrence in Data Governance and Catalog.</p> <p>Choose one of the following options:</p> <ul style="list-style-type: none"> • No. Doesn't automatically create data quality failure tickets when the data quality scores are poor. • Yes. Automatically creates data quality failure tickets based on the data quality threshold values you define in Data Governance and Catalog, and notifies you when a data quality score is below the threshold. <p>Note: You must configure a workflow event for the data quality failure and enable the event in Metadata Command Center.</p>
Cache Result	<p>Select Agent Cache if you want to generate a cache file in the runtime environment and to preview the cached results faster in subsequent data preview runs. The results are cached for seven days by default after the first run in the runtime environment. Select No Cache if you don't want to cache the preview results and view the live results.</p>
Run Rule Occurrence Frequency	<p>Specify whether you want to run data quality rules based on the frequency defined for the rule occurrence in Data Governance and Catalog.</p>
Sampling Type	<p>Determines the sample rows on which you want to run the data quality task.</p> <p>Choose any of the following options:</p> <ul style="list-style-type: none"> • All rows. Runs data profiling on all rows in the metadata. • Limit N Rows. Runs data profiling on a limited number of rows. • Custom Query. Provides an SQL clause to select sample rows to run the data profiling task. <p>For example, <code>where column1='X'; TABLESAMPLE(X ROWS); TABLESAMPLE(X PERCENT)</code></p>
No of rows to limit	<p>Required if you select Limit N Rows in Sampling Type. Specify the number of rows that you want to run the profile on. Default is 1000.</p>
Sampling Query	<p>Required if you select Custom Query in Sampling Type. Specify an SQL clause to select sample rows to run the data profiling task.</p>
Elastic Runtime Environment	<p>Select an elastic runtime environment to run data profiling and quality capabilities on Apache Hive in the elastic cluster configuration.</p> <p>Note: To run a profile, use an elastic runtime environment configured with an advanced cluster. For more information about advanced clusters, see Advanced Clusters help.</p>
Staging Connection	<p>The staging connection where data quality results are stored temporarily during the profile run.</p> <p>If the Apache Hive source system is hosted on a Cloudera CDP or an Amazon EMR cluster, configure Amazon S3 v2 as the staging connection.</p> <p>If the Apache Hive source system is hosted on an Azure HDInsight cluster, configure Microsoft Azure Data Lake Storage Gen2 as the staging connection.</p>

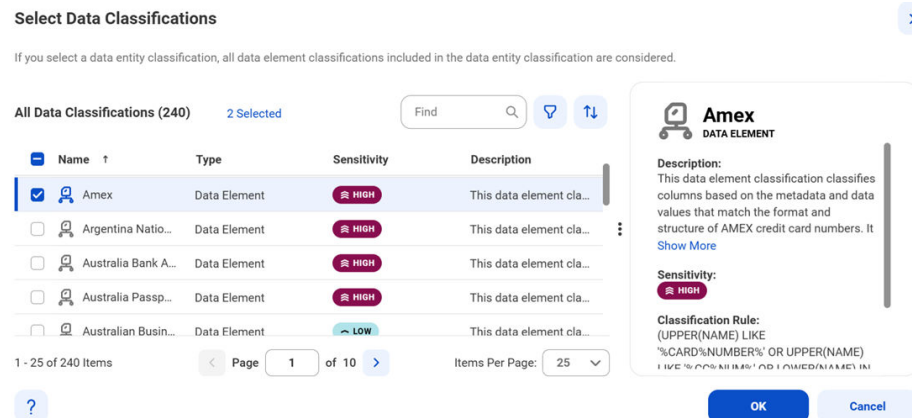
Parameter	Description
Maximum Precision of String Fields	The maximum precision value to be used for profiling fields that include the string data type. Default is 50.
Text Qualifier	The character that defines string boundaries. If you select a quote character, profiling ignores delimiters within the quotes. Select a qualifier from the list. Default is Double Quote.

- To enable the data observability capability, expand **Data Observability** and select **Enable Data Observability**.

Configure data classification

Enable the data classification capability to identify and organize data into relevant categories based on the functional meaning of the data.

- Click the **Data Classification** tab.
- Select **Enable Data Classification**.
- Choose one or both of the following options:
 - Generated Data Classifications.** CLAIRE automatically generates data classifications for the data elements.
 - Data Classification Rules.** Choose from predefined or custom data classifications.
 - Click **Add Data Classification**. The following image shows the **Select Data Classifications** dialog box:



- Select the data classifications that you want to use.
- Click **OK**.

Configure glossary association

Enable the glossary association capability to associate glossary terms with technical assets, or to get recommendations for glossary terms that you can manually associate with technical assets in Data Governance and Catalog.

Metadata Command Center considers all published business terms in the glossary while making recommendations to associate your technical assets.

- Click the **Glossary Association** tab.

2. Select **Enable Glossary Association**.
3. Select **Enable auto-acceptance** to automatically accept glossary association recommendations.
4. Specify the **Confidence Score Threshold for Auto-Acceptance** to set a threshold limit based on which the glossary association capability automatically accepts the recommended glossary terms.
Note: Specify a percentage from 80 to 100. If the score is higher than the specified limit, the glossary association capability automatically assigns a matching glossary term to the data element.
5. Select **Enable Below-threshold Recommendations** to receive glossary association recommendations below the auto-acceptance threshold. If you enable auto-acceptance, you can enable below-threshold recommendations to receive glossary recommendations below the auto-acceptance threshold.
6. Specify the **Confidence Score Threshold for Recommendations** to set a threshold based on which the glossary association capability makes recommendations
 If you enable auto-acceptance, specify a percentage from 80 to the selected auto-acceptance threshold. You can accept or reject the recommended glossary terms that fall within this range in Data Governance and Catalog.
 If you disable auto-acceptance, specify a percentage from 80 to 100 inclusive.
7. Choose to automatically assign business names and descriptions to technical assets. You can then choose to retain existing assignments and only assign business names and descriptions to assets that don't have assignments, or allow overwrite of existing assignments.
 By default, existing assignments are retained.
8. Optional. Choose to ignore specific parts of data elements when making recommendations. Select **Yes** and enter prefix and suffix keyword values as needed.
 Click **Select** to enter a keyword. You can enter multiple unique prefix and suffix keywords. Keyword values are case insensitive.
9. Optional. Choose specific top-level business glossary assets to associate with technical assets. Selecting a top-level asset selects its child assets as well. Select **Top-level Glossary Assets** and specify the assets on the **Select Assets** page.
10. Optional. Choose to use abbreviations and synonym definitions from lookup tables for accurate glossary association. Select **Yes** to enable, and then click **Select** to upload a lookup table.
11. Click **Next**.
 The **Associations** page appears.

Step 3. Associate stakeholders and asset groups

Associate users or user groups within a stakeholder role as stakeholders for technical assets in Data Governance and Catalog. Also, you can choose to assign technical assets extracted from the catalog source to asset groups. You can then use access policies to control permissions on assets that are assigned to asset groups.

Verify that the administrator assigned users and user groups to the stakeholder role that you want to associate with technical assets.

1. To associate users or user groups as stakeholders with technical assets extracted from the catalog source, perform the following steps:
 - a. On the **Associations** page, click **Stakeholders**.

- b. Select **Assign Stakeholders**.
- c. Select a stakeholder role.
- d. Click **Select** to add users and user groups from the stakeholder role as stakeholders for the technical assets.

The **Add Users & User Groups** dialog box displays a list of users and user groups assigned to the selected stakeholder role.

Add Users & User Groups [X]

Users User Groups

All Users (1) Find [magnifying glass] [sort]

Full Name	Email	User Name	Status
<input type="checkbox"/> gov owner_09	[blurred]	[blurred]	Active

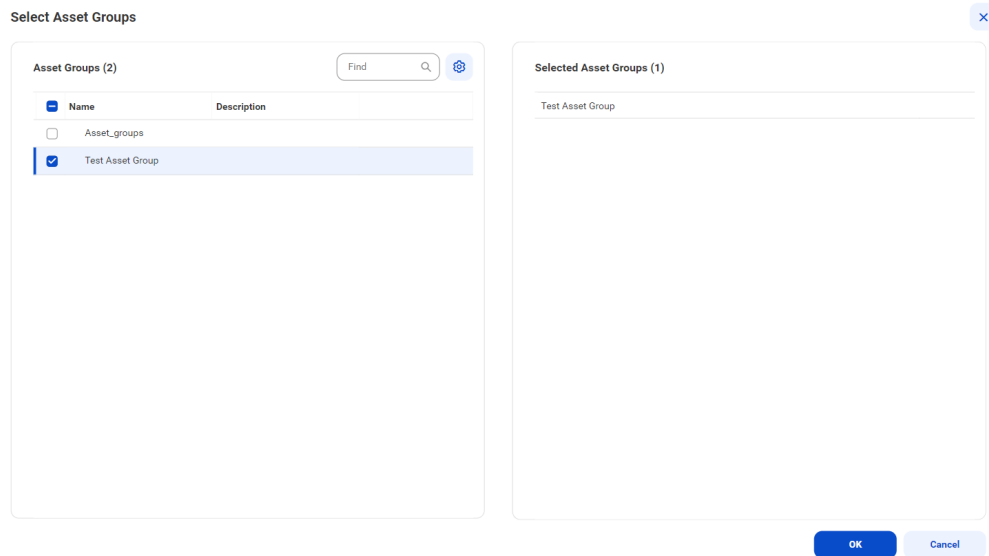
[?] [OK] [Cancel]

- e. Select one or more users or user groups to assign as stakeholders for the technical assets, and click **OK**.
Only the selected users and user groups belonging to the specified stakeholder role are granted the permissions to technical assets.
 - f. To assign users or user groups from another stakeholder role, click **Add** and then repeat the steps.
2. To assign asset groups to technical assets extracted from the catalog source, perform the following steps:
 - a. On the **Associations** page, click **Asset Groups**.
 - b. Select **Assign Asset Groups**.
 - c. Click **Select**.

The **Select Asset Groups** dialog box displays the list of asset groups.

If you enabled an access policy that includes an asset group, you can only view assets that belong to that asset group.

3. Select the asset groups to which you want to assign technical assets extracted from the catalog source, and click **OK**.



4. Choose to save and run the job or to schedule a recurring job.
 - To save and run the job, click **Save** and then **Run**.
 - To schedule a recurring job, click **Next** to open the **Schedule** page.

Step 4. Run or schedule the job

Choose to run a catalog source job manually, or configure it to run on schedule.

Note: You can't run multiple jobs simultaneously.

You can choose to perform a full or an incremental metadata extraction. A full metadata extraction extracts all objects from the source to the catalog. An incremental metadata extraction extracts only the changed and new objects since the last successful catalog source job run. Incremental metadata extraction doesn't remove deleted objects from the catalog and doesn't extract metadata of code-based objects if applicable.

When you run an incremental metadata extraction job with a filter to include metadata from objects, the job extracts only the objects that have the latest timestamp since the last successful job.

Note: The incremental extraction option appears if it is available for the catalog source.

Run the job manually

Click **Save** to save the catalog source and click **Run**. On the **Run Catalog Source Job** window, click **Run** to run the job.

You can override the capabilities that you selected while configuring your catalog source on the **Configuration** page. The first time you run the catalog source job, the metadata extraction capability is mandatory. From the second run onwards, you can choose to override the configured metadata change option. You can retain, delete, or deprecate objects that are deleted from the source in the catalog. For subsequent runs of the catalog source job, the metadata extraction capability is optional.

Note: You can choose incremental metadata extraction for subsequent runs only after one full metadata extraction job completes successfully. Incremental metadata extraction jobs run with the **Retain** metadata change option even if you set the option to **Delete** or **Deprecate** in the catalog source.

Note: To run a catalog source job, you need permissions on the connection to the source system. To run a catalog source job for catalog sources that reference other source systems, you need permissions on the connections for all the reference source systems.

Run the job on a schedule

You can choose to run metadata extraction and other capabilities on a recurring schedule. You can't choose incremental metadata extraction and full metadata extraction in the same schedule. To create a schedule for incremental metadata extraction, you must have completed at least one full metadata extraction job successfully. If not, first create a schedule for a full metadata extraction.

If an incremental metadata extraction is scheduled to run when the last run details aren't available, the job first performs a full metadata extraction, followed by incremental metadata extraction on subsequent runs.

For example, this can happen in the following scenarios:

- You create schedules for both incremental metadata extraction and full metadata extraction, but schedule the incremental extraction to run before the first full metadata extraction job.
- You create schedules for both incremental metadata extraction and full metadata extraction, but delete the full metadata extraction schedule before its first run.

1. On the **Schedule** tab, select **Run on Schedule**.
The **Schedule** configuration page opens.
2. Click the checkbox corresponding to each capability that you want to include in the schedule.
3. Enter the start date, time zone, and the interval at which you want to run the job.
4. You can manage additional schedules using the following options:
 - To create a new schedule, click the **Add** button.
 - To delete a schedule, click the **Delete** button.
 - To enable or disable a schedule, click the **Enable Schedule** toggle button.

Note: You can create a maximum of one schedule per capability that you enable. If you purged a catalog source or did not run the metadata extraction job, the catalog source job runs metadata extraction before running other scheduled capabilities.

Note: To create a schedule, you need permissions on the connection to the source system. If you lose permissions on the connection after you create a schedule, the scheduled jobs continue to run.

5. Click **Save** to save the schedule.

Monitor job status

After the job runs, you can monitor the status of the job on the **Overview** page of the job.

For more information about job monitoring, see *Administration*.

Step 5. Assign reference catalog source connections to endpoint catalog source objects

When you run the catalog source job, if the catalog source references another source system, a reference catalog source and connection get created that point to the reference source system. To view the complete lineage for your catalog source, you can perform connection assignment from the reference catalog source connection to the objects in the reference source system. A reference source system might be a database,

such as Oracle. You must first create and run an endpoint catalog source that connects to the reference source system.

Before you assign a connection, ensure that you have created and run an endpoint catalog source for each reference source system.

Note: If the source schema contains case-sensitive tables or if the reference objects contain multiple objects with the same name in different cases, perform case-sensitive connection assignment to get correct lineage.

If you enabled the lineage discovery capability for your catalog source, you can either curate the CLAIRE recommended endpoint objects on the **Related Catalog Sources** tab or assign connections manually.

For more information about related catalog sources and lineage discovery, see *Lineage discovery* in the *Administration* help.

1. On the **Configure** page, select the **Lineage** tab, and then select the **Lineage Discovery** tab. On the **Catalog Sources** panel, select the required catalog source and click the **Assign Connections** tab.
The **Assign Connections** tab displays a list of assigned and unassigned connections along with details for each connection. Use filters to view the connections based on the connection names. Click the **Add Filter** menu to add filters.
2. Select the connection to the reference source system and click **Assign**.
The connection name appears prefixed to the reference catalog source name on the **Hierarchy** tab of your catalog source in Data Governance and Catalog.
The **Assign Connection** dialog box appears with a list of recommended objects from the endpoint catalog sources. Click **All** to view all endpoint catalog source objects.
3. In the **Assign Connections** dialog box, select one or more endpoint objects to assign to the selected connection and click **Assign**.
You can filter the list in the **Assign Connection** dialog box by name, type, or endpoint. You can assign an Hadoop Distributed File System source system as an endpoint catalog source. The objects must be the Schema class type.

When you click **Assign**, Metadata Command Center creates links between matching objects in the connected catalog sources, and it calculates the percentage of matched and unmatched objects. The higher the percentage of matched objects, the more accurate the lineage that you view in Data Governance and Catalog.

CHAPTER 4

View results in Data Governance and Catalog

After Metadata Command Center runs a job, you can view the results in Data Governance and Catalog where the catalog source and its elements are called technical assets. You can view a catalog source as a hierarchy. Expand each technical asset to see its components.

When referenced source systems are connected to a catalog source, you can expand the hierarchy to see details about the technical asset's component elements.

You can view the data lineage of an asset contained within a catalog source to see individual elements such as data sources, calculations, and filters. When you view data lineage, you can see the individual upstream elements that contribute data or expressions to each component of a data flow or catalog source.

View metadata extraction results

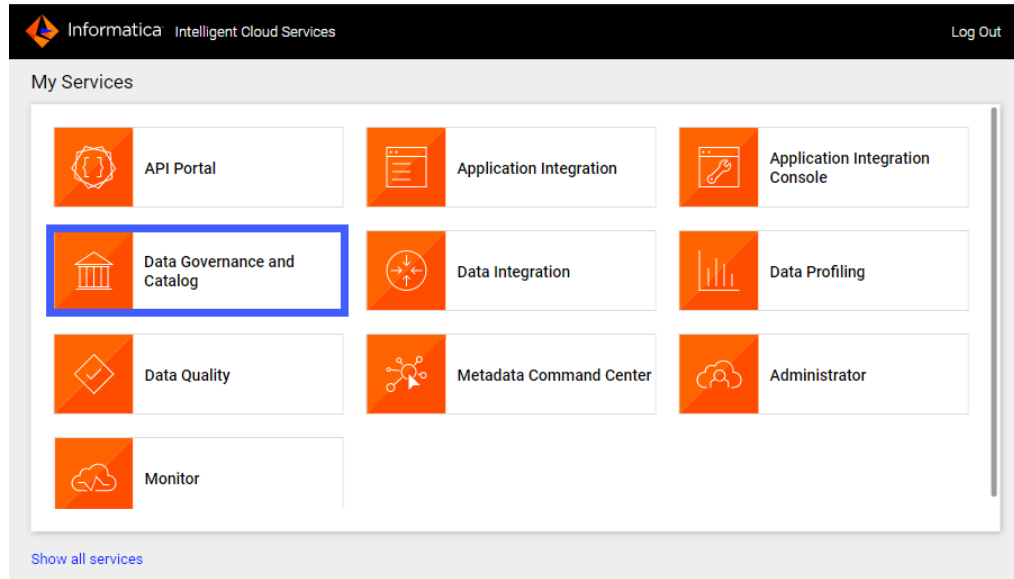
After a job runs in Metadata Command Center, view the results in Data Governance and Catalog. You can view details about source system contents as hierarchical displays and trace data lineage.

1. Log in to Informatica Intelligent Cloud Services.

The **My Services** page appears.

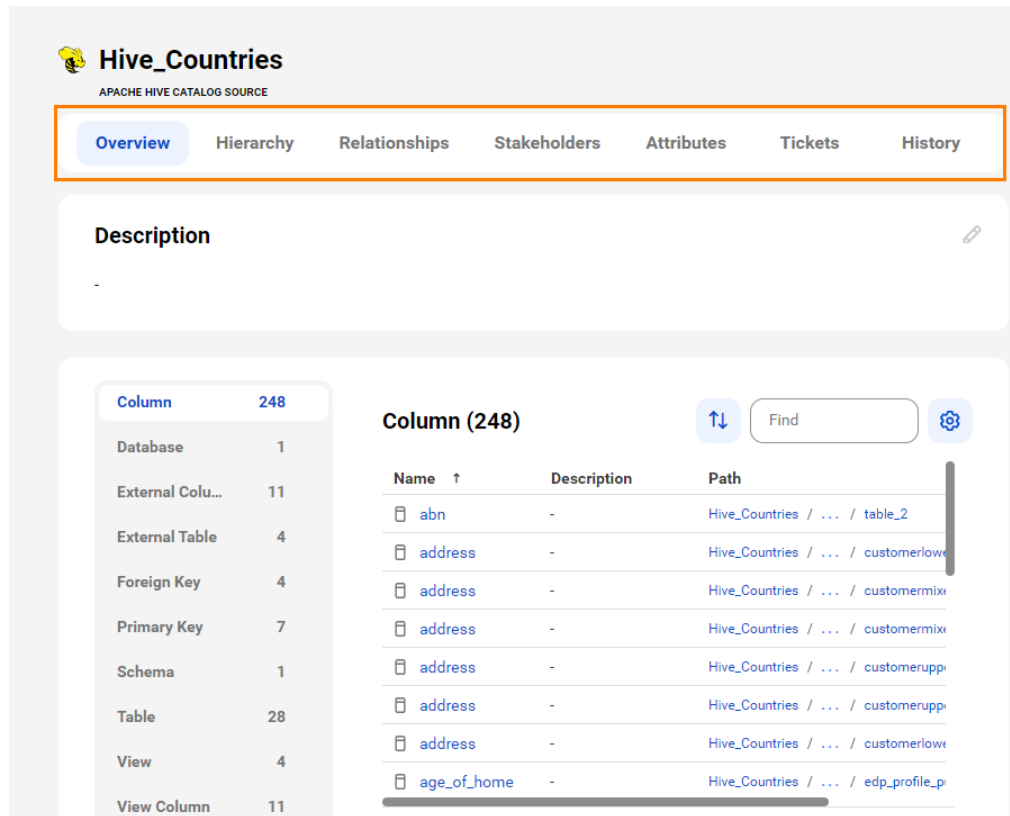
2. Click Data Governance and Catalog.

The following image shows the Data Governance and Catalog box on the **My Services** page:



3. On the Data Governance and Catalog home page, click the number in the **Technical Assets** panel.
The **Technical Assets** page opens.
4. Select **Catalog Source** in the **Filter** list.
The list of catalog sources opens.
5. Search for the catalog source from which you extracted metadata, and click the name.
The **Overview** tab of the asset opens.

The following image shows a sample asset page:



6. View the asset from different perspectives by clicking on the tabs.

For more information about working with assets, see *Working with Assets* in the Data Governance and Catalog help.

View data lineage

Data lineage is a visual representation of the flow of data across the systems in your organization. Lineage depicts how the data flows from the system of its origin to the system of its destination.

Data lineage views are available for technical assets in the catalog source. You can view lineage at the catalog source, data set, or data element level.

The lineage at the catalog source level shows how data flows from one catalog source to another. The lineage at the data set and the data element levels show how other technical assets such as files or tables contribute to the selected asset.

If linking catalog sources is available for your catalog source, you can use Metadata Command Center to generate data lineage based on rules or by generating automated lineage with CLAIRE. You can choose source and target catalog sources and objects to link and generate lineage.

To determine whether linking catalog sources is available for your catalog source, navigate to the **Configuration** tab of the **Link Catalog Sources** page. The catalog source must appear in the list of source and target catalog sources.

For information about linking catalog sources, see *Link catalog sources* in the Administration help.

View lineage at the catalog source level

The catalog source level shows how data flows from one catalog source to another with the lineage aggregating data from the data set and data element levels.

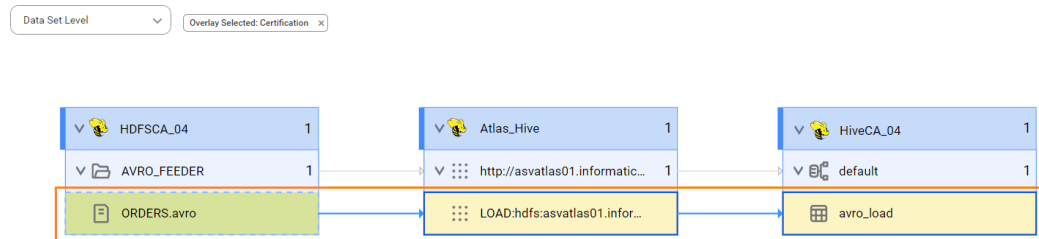
To view data lineage at the catalog source level, open a technical asset, click the **Lineage** tab, and then verify that the level is set to **Catalog Source Level**.

View lineage at data set level

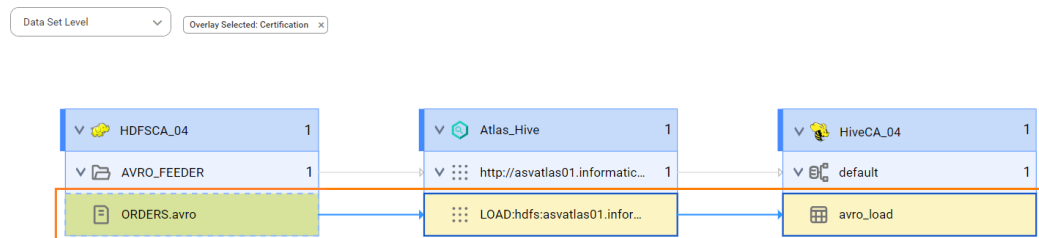
The data set level is a view that shows individual sets of data in the data flow.

To view lineage at the data set level, open a technical asset, click the **Lineage** tab, and then verify that the level is set to **Data Set Level**.

The following image shows how the avro_load referenced table in Apache Hive gets data from the ORDERS.avro referenced file in the HDFS source system before connection assignment:



The following image shows how the avro_load actual table in Apache Hive gets data from the ORDERS.avro actual file in the HDFS source system after connection assignment:

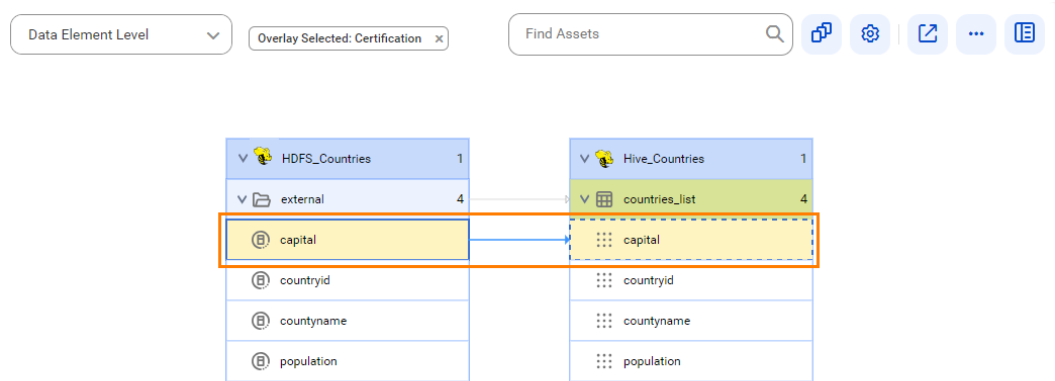


View lineage at data element level

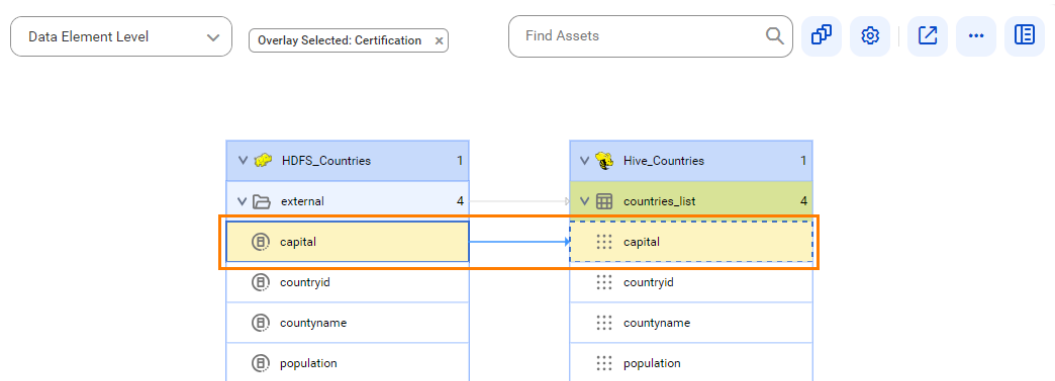
The data element level displays details of the data set level. At the data element level, you can see the input sources for expressions or commands and calculations or transformations on the data.

To view data lineage at the data element level, open a technical asset, click the **Lineage** tab, and then verify that the level is set to **Data Element Level**.

The following image shows the lineage where the capital referenced column in Apache Hive gets data from the capital referenced data element of the HDFS source system before connection assignment:



The following image shows the lineage where the capital actual column in Apache Hive gets data from the capital actual data element of the HDFS source system after connection assignment:

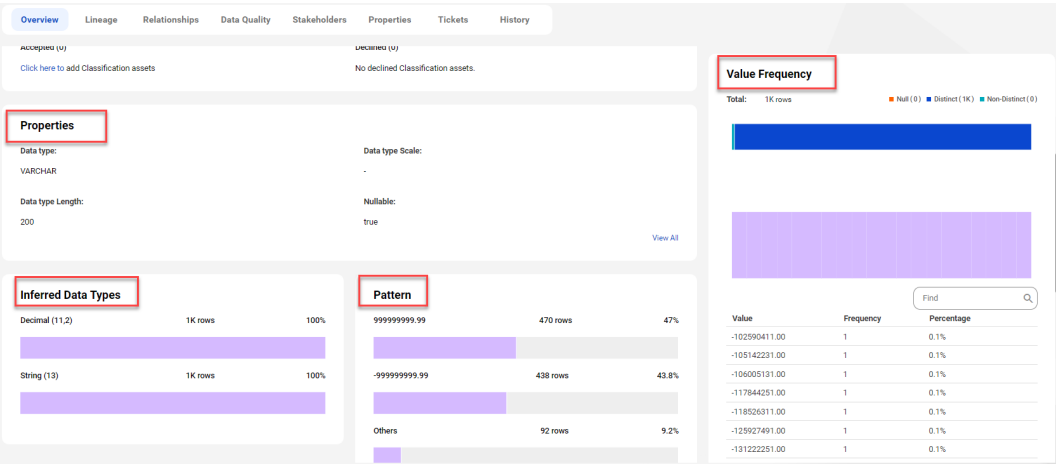


View data profiling results

When you enable the data profiling task for a catalog source in Metadata Command Center, the system runs a profile to evaluate the quality of the metadata extracted from the source system. The profiling statistics appear in Data Governance and Catalog when you open the technical assets.

The scope of profiling statistics that Data Governance and Catalog displays depends on the data profiling configuration parameters that you set when you configured the catalog source in Metadata Command Center.

The following image shows the data profiling statistics that appear on a column asset page in Data Governance and Catalog:



For more information about data profiling results, see *Asset Details* in the Data Governance and Catalog help.

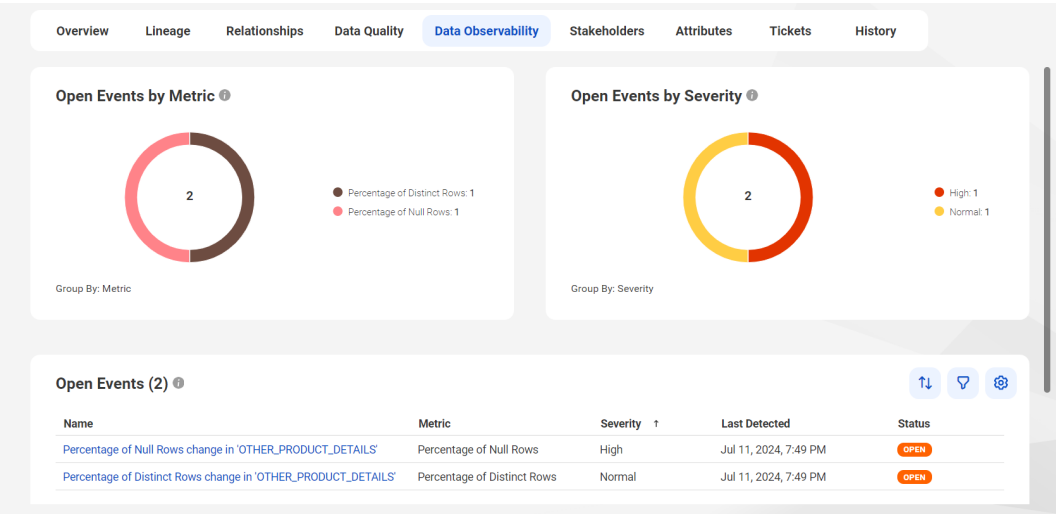
View data observability results

When you enable data observability for a catalog source in Metadata Command Center, you can view and evaluate the events that it generates in Data Governance and Catalog. These events indicate anomalies identified in the characteristics of the profiled data in your source system.

You can view the events that data observability generates for anomalies identified for catalog sources, technical data sets, and data elements. You can then take appropriate actions for the generated events.

Note: The administrator of the catalog source might have applied filters to the data to narrow down the data elements that are applicable for business users in Data Governance and Catalog. The data for which users receive anomaly notifications depend on the filters that are configured for the catalog source.

The following image shows the open events for a column asset in Data Governance and Catalog:

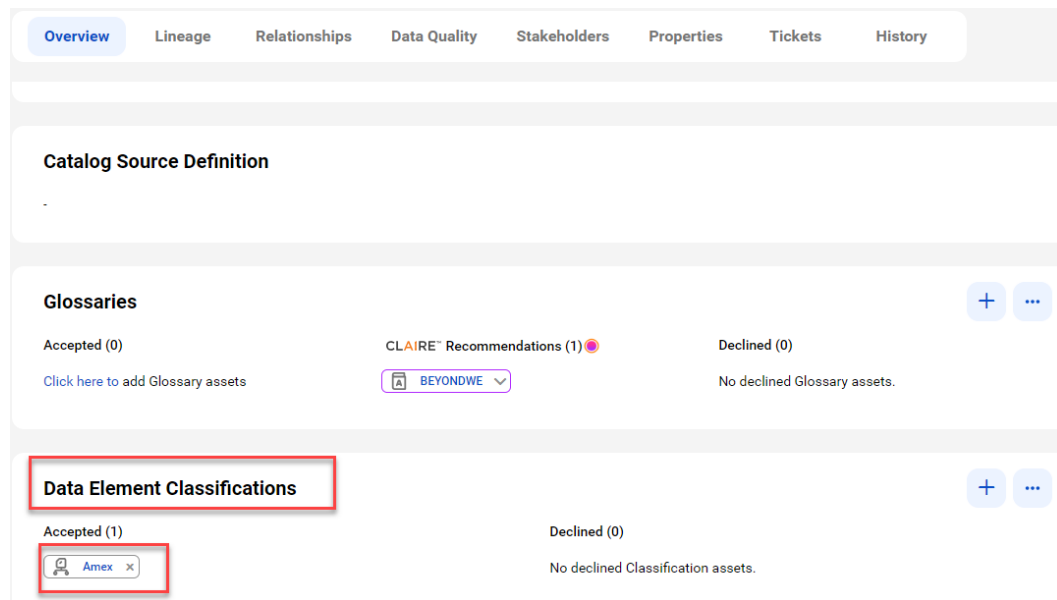


For more information about data observability results, see *Working With Assets* in the Data Governance and Catalog help.

View classified data

When you add data classification rules to a catalog source in Metadata Command Center, the system identifies the columns and tables that match the rules and displays one or more matched data classifications on the column or table asset pages in Data Governance and Catalog.

The following image shows a column asset page with the inferred data element classifications that match the column data and metadata:



For more information about data classification assets, see *Asset Details* in the Data Governance and Catalog help.

View glossary associations

When you enable the glossary association capability for a catalog source in Metadata Command Center, you can view the accepted glossary assets in Data Governance and Catalog.

The **Overview** tab for a technical asset in the catalog source displays glossary assets in the Accepted and CLAIRE Recommendations sections.

The **Glossaries** panel shows the automatically accepted and CLAIRE® recommended terms.

The following image shows a sample asset page:

