



Informatica® Metadata Command Center
November 2025

Apache Atlas 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: 2025-11-19

Table of Contents

Preface.	4
Chapter 1: Introduction to Apache Atlas catalog sources.	5
Extraction and view process.	6
About the Apache Atlas catalog source.	6
Extracted metadata.	7
Chapter 2: Before you begin.	8
Verify authentication.	8
Verify permissions.	9
Import the SSL certificate to the Secure Agent machine.	9
Get Apache Atlas source information.	10
Chapter 3: Create catalog sources in Metadata Command Center.	11
Step 1. Register a catalog source.	11
Step 2. Configure capabilities.	13
Configure metadata extraction.	13
Configure lineage discovery.	16
Step 3. Associate stakeholders and asset groups.	17
Step 4. Run or schedule the job.	18
Step 5. Assign reference catalog source connections to endpoint catalog source objects.	19
Chapter 4: View results in Data Governance and Catalog.	21
View metadata extraction results.	21
View data lineage.	23
View lineage at the catalog source level	23
View lineage at the data set level.	23
View lineage at the data element level.	24

Preface

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

CHAPTER 1

Introduction to Apache Atlas 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 Atlas is a source system from which you can extract metadata through an Apache Atlas 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.

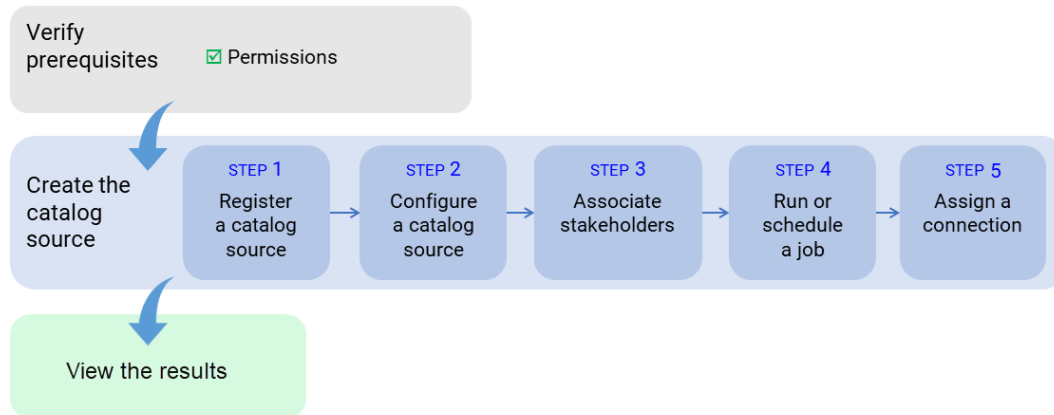
The following table describes the capabilities of the catalog source:

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.

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 an Apache Atlas source system:



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

1. Register a catalog source. Create a catalog source object, select the source system, and specify values for connection properties.
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 Atlas catalog source

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

Apache Atlas is the governance and metadata framework for Hadoop. Apache Atlas has a scalable and extensible architecture that can be plugged into many Hadoop components to manage their metadata in a central repository.

Extracted metadata

You can extract metadata from an Apache Atlas source system.

Objects extracted

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

- Atlas Server
- Hive Process
- Sqoop Process
- Calculation

Note: Calculation objects are extracted when there is column-level lineage from one asset to another in Hive and Sqoop processes.

- Spark Application
- Spark Process

The Apache Atlas catalog source extracts data lineage from the following data sources:

- Oracle
- MySQL
- PostgreSQL
- Apache Hive
- Hadoop Distributed File System (HDFS)
- Apache HBase

Note: Metadata Command Center skips extraction of Hive processes and the associated lineage links for the following operation types:

- CREATETABLE
- CREATEVIEW
- CREATE_MATERIALIZED_VIEW

Metadata Command Center extracts folders as reference objects from Hadoop Distributed File System.

Metadata Command Center extracts the following objects as reference objects from Apache Hive:

- Schema
- Table
- View
- External Table
- Column

Field and column objects are extracted when there is column-level lineage from one asset to another in Apache Atlas.

CHAPTER 2

Before you begin

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

Perform the following prerequisite tasks:

- Verify authentication.
- Verify permissions.
- Import SSL certificate to the JRE folder of the Informatica Secure Agent.
- Get Apache Atlas source information.

Verify authentication

To extract Apache Atlas metadata, verify that you have the URL to access Apache Atlas and connect to the Atlas REST API.

You need to provide the Kerberos principal for authentication when you configure the Apache Atlas catalog source in Metadata Command Center.

Complete the following tasks:

- Add details of the Kerberos server to the host file located in the Secure Agent machine in the following format: <ip_address> <hostname>
On a Windows machine, the host file is available in the following path: C:\Windows\System32\drivers\etc\hosts
On a Linux machine, the host file is available in the following path: /etc/hosts
- Copy the Atlas Keytab file from the Hadoop cluster to any location on the Secure Agent machine.
- Enable Atlas hook from Hadoop Distributed File System and Apache Hive configurations so that Apache Atlas can read the metadata.
- Copy the Kerberos configuration file from the Hadoop cluster to any location on the Secure Agent machine. You can modify the Kerberos configuration file as per requirement.

The following code shows a sample Kerberos configuration file:

```
[libdefaults]
default_realm = *****
dns_lookup_kdc = false
dns_lookup_realm = false
ticket_lifetime = 86400
renew_lifetime = 604800
forwardable = true
default_tgs_enctypes = rc4-hmac
default_tkt_enctypes = rc4-hmac
```



```

permitted_encetypes = rc4-hmac
udp_preference_limit = 1
kdc_timeout = 3000
allow_weak_crypto=true
[realms]
<domain name> = {
  kdc = *****
  admin_server = *****
}
[domain_realm]

```

Note: If the Kerberos encryption algorithms are not compatible with Java Standard Edition version 11, you can add the `allow_weak_crypto=true` property in the Kerberos configuration file.

Verify permissions

Verify that you have the following account and permissions:

- A user account to access and extract metadata from the Apache Atlas source system.
- Read permission for the account to access the Apache Atlas source system.

Import the SSL certificate to the Secure Agent machine

If SSL is enabled on the Apache Atlas source system, import the SSL certificate to the JRE folder of the Informatica Secure Agent installation directory.

Complete the following steps to import the SSL certificate:

1. Download the SSL certificate from the Apache Atlas installation.
2. Copy the SSL certificate file to any location on the Informatica Secure Agent machine.
3. Identify the Java version for the Secure Agent.
You can identify the Java version from the following files:
 - `<Informatica Secure Agent installation directory>\apps\agentcore\agentcore.log`
Search for "AgentCore JRE version".
 - On Windows operating systems, open the `lcm-env.bat` file from one of the following locations:
 - `<Informatica Secure Agent installation directory>\apps\agentcore<latest version>\.lcm`
 - `<Informatica Secure Agent installation directory>\apps\DIS<latest version>\.lcm`
 - On Linux operating systems, open the `lcm-env.sh` file from one of the following locations:
 - `<Informatica Secure Agent installation directory>/apps/agentcore/<latest version>/.lcm`
 - `<Informatica Secure Agent installation directory>/apps/DIS/<latest version>/.lcm`
4. Open a command prompt from the following directory:
`<Informatica Secure Agent installation directory>\apps\jdk<latest version>\jre\bin`

5. Run the following command to import the SSL certificate:

```
keytool -import -alias <alias name> -keystore <path to cacert file> -file <absolute path to SSL certificate>
```

Note: The Java certificate file is named `cacerts` and is located in the following Java directory: `\jre\lib\security\cacerts`

For example, you can run the following command on Windows operating systems:

```
keytool -import -alias aliasname -keystore "C:\data\devprod\jdk\jre\lib\security\cacerts" -file "C:\data\devprod\filename.crt"
```

6. Restart the Secure Agent.

Get Apache Atlas source information

Before you configure the catalog source, ask the Apache Atlas administrator for connection information that you need to configure the catalog source.

Note: You don't need to create a connection object for Apache Atlas. You provide this information when you configure the catalog source.

The following table describes the properties that you need:

Property	Description
Base URL	URL to access Apache Atlas and connect to the Atlas REST API.
Principal	The Kerberos principal used for authentication.
Keytab File Path	The absolute path to the Kerberos keytab file located on the Secure Agent machine used for authentication.
Configuration File Path	The absolute path to the Kerberos configuration file located on the Secure Agent machine used for authentication.

CHAPTER 3

Create catalog sources in Metadata Command Center

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

When you configure a catalog source, you define the source system that you want to extract metadata from. Configure filters to include or exclude source system metadata before you run the job. Optionally, configure other capabilities, such as lineage discovery, data profiling and quality, data classification, relationship discovery, and glossary association.

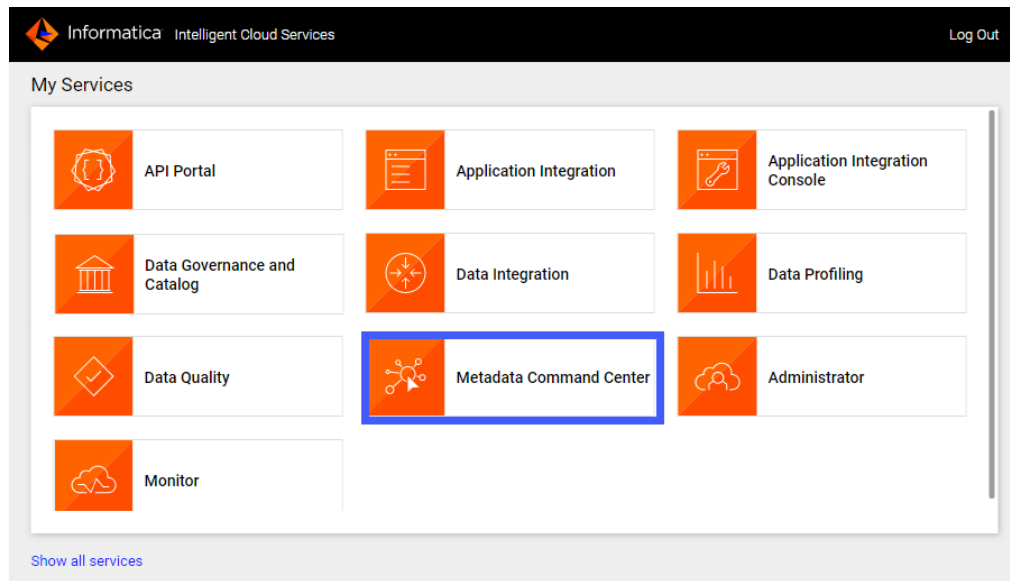
To provide stakeholders access to technical assets, you can assign access through stakeholder roles. You can also associate technical assets extracted from the catalog source to asset groups. If your catalog source references other source systems, you can create a connection assignment to the endpoint catalog source to view complete lineage.

Step 1. Register a catalog source

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

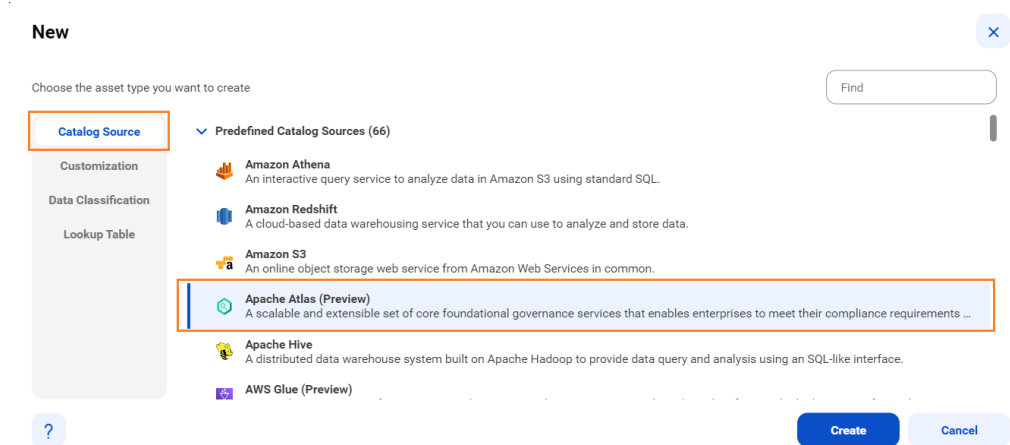
1. Log in to Informatica Intelligent Cloud Services.
The **My Services** page appears.
2. Click **Metadata Command Center**.

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 Atlas from the list of source systems.



6. Click **Create**.
- 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, enter the Apache Atlas connection information based on the connection values that you got from the administrator.

The following table describes the properties to configure:

Property	Description
Base URL	URL to access Apache Atlas and connect to the Atlas REST API.
Principal	The Kerberos principal used for authentication.
Keytab File Path	The absolute path to the Kerberos keytab file located on the Secure Agent machine used for authentication.
Configuration File Path	The absolute path to the Kerberos configuration file located on the Secure Agent machine used for authentication.

9. Click **Next**.

The **Configuration** page appears.

Step 2. Configure capabilities

When you configure the Apache Atlas catalog source, you define the settings for the metadata extraction capability.

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 Atlas 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 Hive Database, HDFS Path, or HBase Namespace.
 - c. Enter the filter values.

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 condition options:

The screenshot shows the 'Filters' section with the 'Specify metadata filters' toggle set to 'Yes'. Below this is a link 'Show supported wildcards and examples'. The main configuration area has three input fields: 'Include or exclude metadata...' (a dropdown menu), 'Select the object type' (a dropdown menu with 'Hive Database', 'HDFS Path', and 'HBase Namespace' options), and 'Enter a value to specify the object location' (a text input field). To the right of these fields are '+' and trash icons.

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

The following image shows that the filter includes metadata related to Hive tables in the HR database with names that start with EMP followed by a single character, includes metadata related to the table named HbaseTable located in the HbaseNS namespace, and excludes metadata related to all files in the hdfsfolder1 folder and its subfolders:

The screenshot shows the 'Filters' section with the 'Specify metadata filters' toggle set to 'Yes'. Below this is a link 'Show supported wildcards and examples'. The main configuration area displays three filter conditions, each with its own '+' and trash icons:

- Condition 1: 'Include Metadata' (dropdown), 'Hive Database' (dropdown), 'HR.EMP?' (text input).
- Condition 2: 'Include Metadata' (dropdown), 'HBase Namespace' (dropdown), 'HbaseNS.HbaseTable' (text input).
- Condition 3: 'Exclude Metadata' (dropdown), 'HDFS Path' (dropdown), '/hdfsfolder1/*' (text input).

Exclude filter conditions are considered if the assets in the include filter conditions are not related or linked through lineage to the excluded assets. For example, add a filter condition to include metadata related to all tables with the name EMP across all databases (*.EMP) and then add another filter condition to exclude metadata related to the EMP table located in the HR database (HR.EMP). Here, the exclude filter condition is considered as the assets are not related or linked through lineage.

Exclude filter conditions are not considered if the assets in the include filter conditions are related or linked through lineage to the excluded assets. For example, add a filter condition to include metadata related to EMP table in the HR database (HR.EMP) and then add another filter condition to exclude

metadata related to SAL table in the same HR database (HR.SAL). Here, the exclude filter condition is not considered due to the presence of lineage links between the EMP and SAL tables.

If you add a filter condition to include metadata from a table deleted from the Apache Atlas source system, Metadata Command Center ignores the filter condition.

If the value of the HDFS Path filter contains special characters, replace the special characters with an asterisk wildcard character. For example, replace `/Test$~^!()**<>_Folder` with `/Test*Folder`.

4. In the **Configuration Parameters** area, enter configuration properties.

Note: Click **Show Advanced** to view all configuration parameters.

The following table describes the properties that you can enter:

Property	Description
Lineage Direction	The direction of data flow between assets that you extract from Apache Atlas with the direction parameter of the LineageRESTAPI. Select one of the following options: <ul style="list-style-type: none">- BOTH. Extracts both input and output data flow between assets.- INPUT. Extracts only input data flow between assets.- OUTPUT. Extracts only output data flow between assets.
Lineage Depth	The number of lineage hops to extract from Apache Atlas for filtered assets with the depth parameter of the LineageRESTAPI. Default is 3.
Page Result Limit	Advanced parameter. The maximum number of search result entries per page from a fetch using the limit parameter of the DiscoveryRESTAPI. Default is 1000.
Entity Bulk Fetch Count	Advanced parameter. The maximum number of entities to include in a bulk fetch when you use the BulkEntityRESTAPI. Default is 100.
Connection Timeout	Advanced parameter. The maximum amount of time, in milliseconds, that the Secure Agent waits to set up an HTTP connection to communicate and get a response from the Apache Atlas server. Default is -1 which means timeout is disabled.
Parallel Lineage Fetch Count	Advanced parameter. The maximum number of LineageRESTAPI calls that can run simultaneously to retrieve lineage data. Default is 5.

5. Optional. In the **Configuration Parameters** area, enter additional settings.

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.

6. 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.

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

Users User Groups

All Users (1)

Find 🔍 ↕

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

? OK Cancel

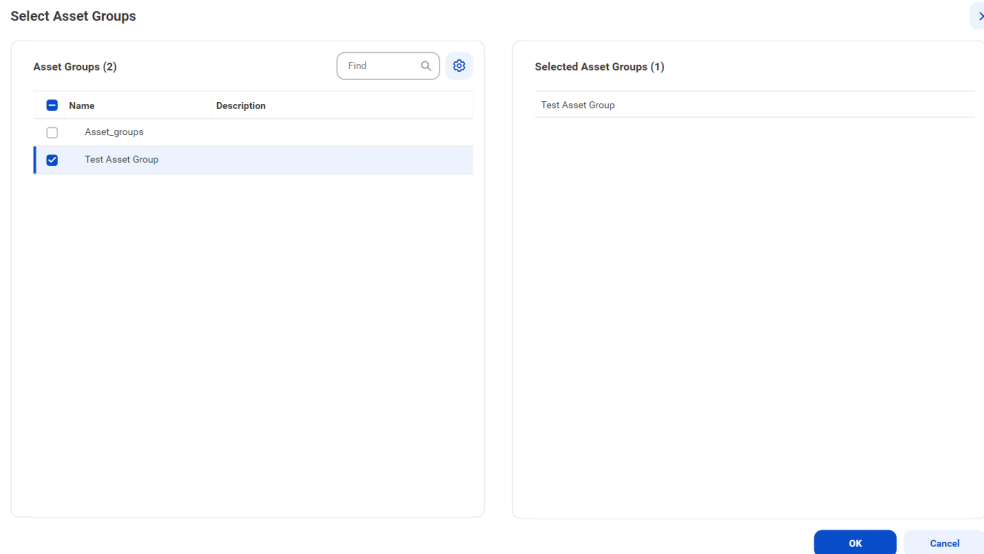
- e. Select one or more users or user groups to assign as stakeholders for the technical assets, and click **OK**.

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 file system such as Hadoop Distributed File System or relational databases such as Apache Hive, Oracle, MySQL, and PostgreSQL. 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.

Note: You can view the lineage with reference objects without creating a connection assignment. After connection assignment, you can view the actual objects.

Apache Atlas uses Sqoop queries to import data from a reference source system to a Hive database. If the Sqoop query contains double quotes, replace the double quotes with backticks (`) in the Apache Atlas source system to view the reference objects correctly in Data Governance and Catalog.

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 Connections** dialog box by name, type, or endpoint.

You can connect to the following source system:

- Apache Hive. The target endpoint object must belong to the **Database** class type.
- Hadoop Distributed File System. The target endpoint object must belong to the **File System** class type.
- Oracle. The target endpoint object must belong to the **Database** class type.
- MySQL. The target endpoint object must belong to the **Database** class type.
- PostgreSQL. The target endpoint object must belong to the **Database** class type.

Note: You can assign connections to Oracle, MySQL, and PostgreSQL catalog sources only when Metadata Command Center extracts Sqoop processes from an Apache Atlas source system.

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

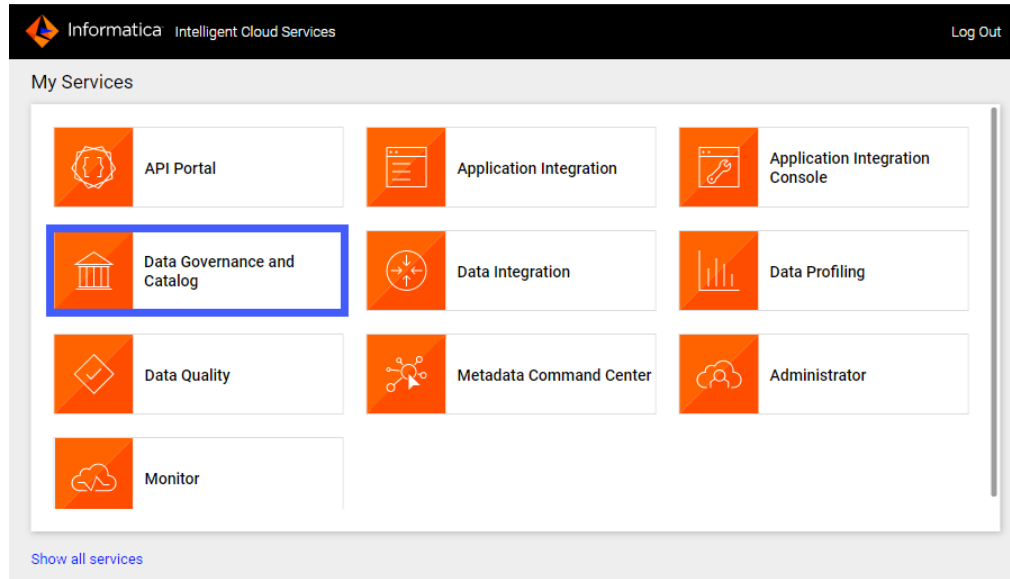
After a job runs in Metadata Command Center, view the results in Data Governance and Catalog. You can view details about source system contents in a hierarchical structure and trace data lineage.

1. Log in to Informatica Intelligent Cloud Services.

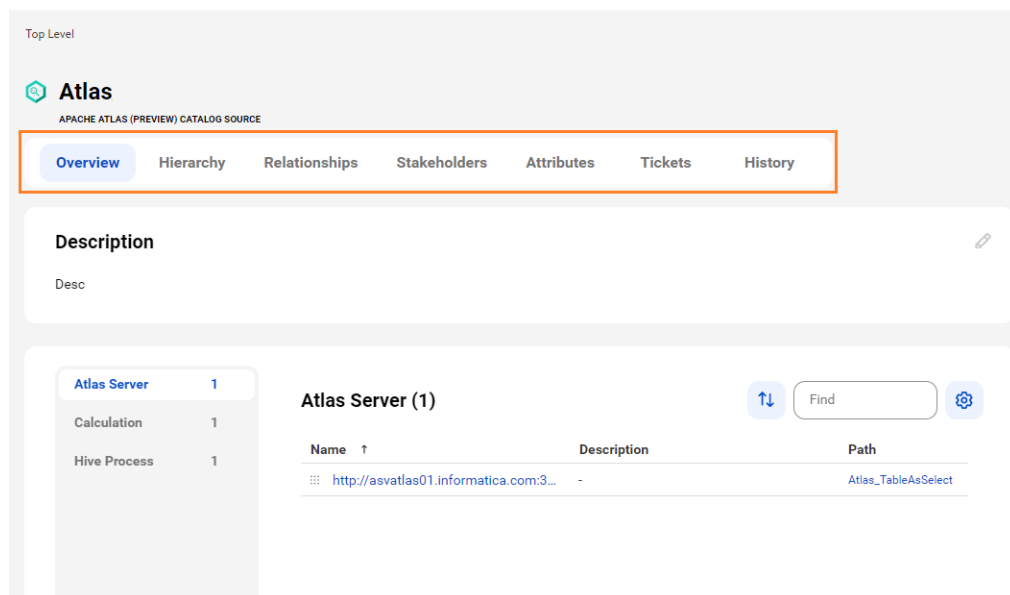
The **My Services** page appears.

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.

Note: You can view the calculation properties such as expression, control conditions, and calculation complexity in the **Overview** tab of a calculation asset.

If a table or column is deleted from the source system, the **Technical Description** field in the **Overview** tab and the **Comment** field in the **System Attributes** tab display the value as **Deleted**.

For more information about working with assets, see *Working with Assets* in *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.

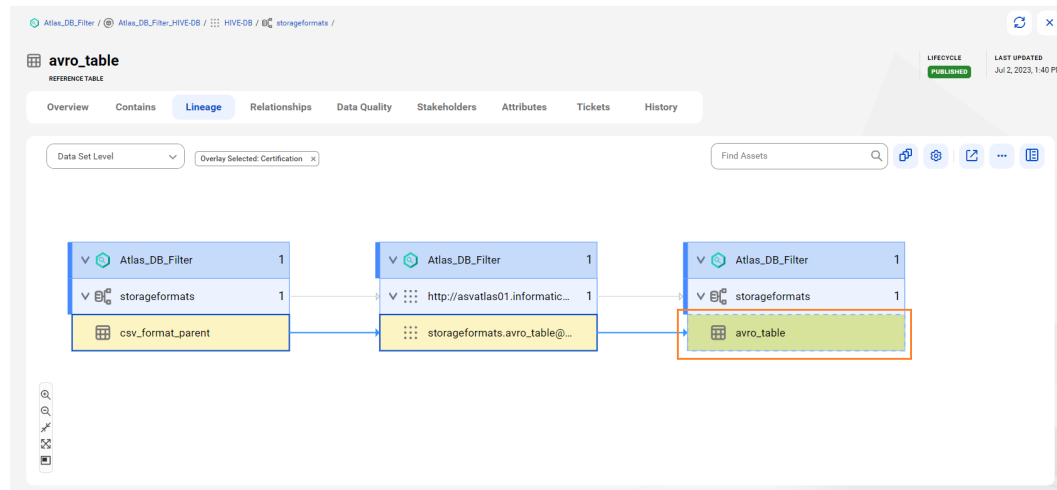
To view data lineage at the catalog source level, open a technical asset, click the **Lineage** tab, and then verify that the level is set to **Catalog Source Level**.

View lineage at the data set level

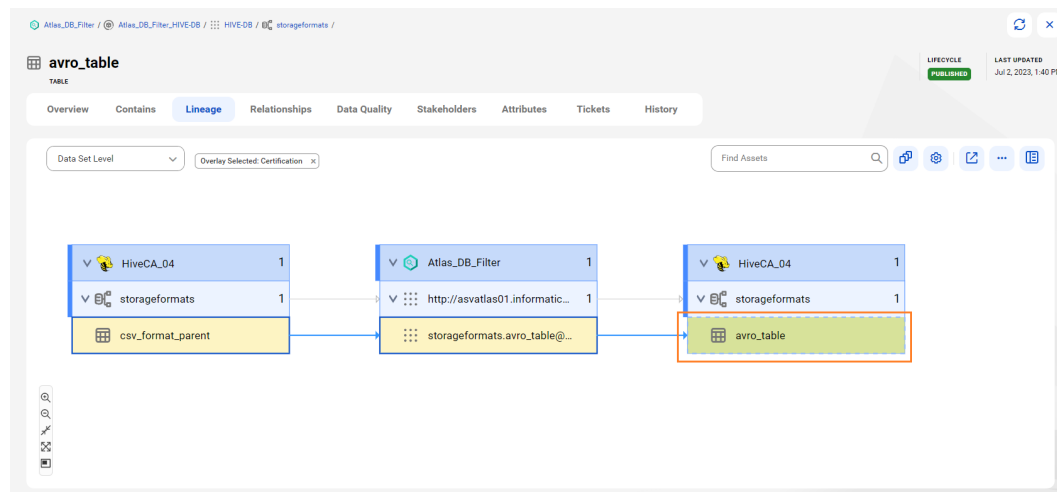
The data set level is a view that shows individual sets of data in the data flow.

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 table-level lineage where the avro_table referenced table gets data from the csv_format_parent referenced table after data transformation using the storageformats.avro Hive process before connection assignment:



The following image shows table-level lineage where the avro_table actual table gets data from the csv_format_parent actual table after data transformation using the storageformats.avro Hive process after connection assignment:

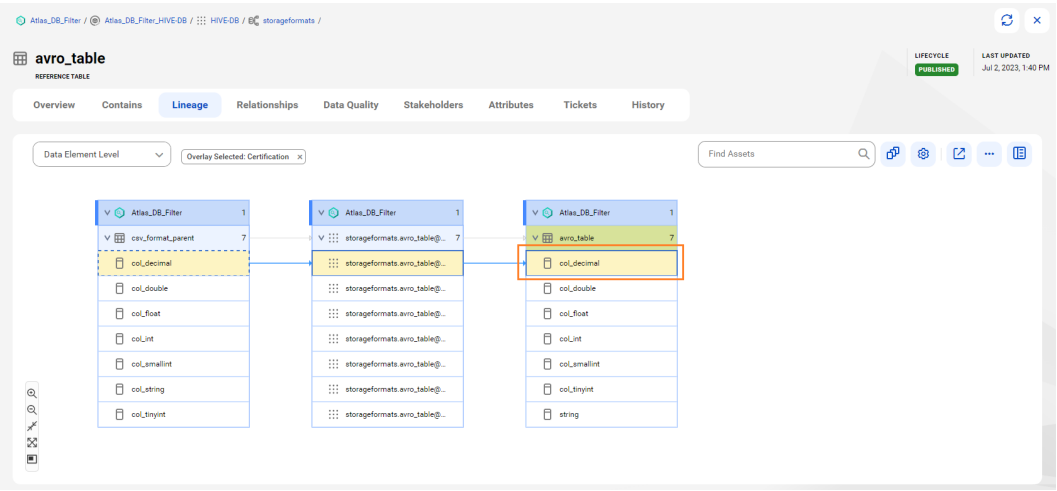


View lineage at the data element level

The data element level displays detailed information. 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 column-level lineage where the col_decimal referenced column of the avro_table gets data from the col_decimal referenced column of the csv_format_parent table after data transformation using the storageformats.avro Hive process before connection assignment:



The following image shows column-level lineage where the col_decimal actual column of the avro_table gets data from the col_decimal actual column of the csv_format_parent table after data transformation using the storageformats.avro Hive process after connection assignment:

