Informatica®

10.2

Intelligent Data Lake Administrator Guide

Informatica Intelligent Data Lake Administrator Guide
10.2
February 2018

DISCLAIMER: Informatica LLC provides this documentation "as is" without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of noninfringement, merchantability, or use for a particular purpose. Informatica LLC does not warrant that this software or documentation is error free. The information provided in this software or documentation may include technical inaccuracies or typographical errors. The information in this software and documentation is subject to change at any time without notice.

NOTICES

This Informatica product (the "Software") includes certain drivers (the "DataDirect Drivers") from DataDirect Technologies, an operating company of Progress Software Corporation ("DataDirect") which are subject to the following terms and conditions:

1. THE DATADIRECT DRIVERS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT.
2. IN NO EVENT WILL DATADIRECT OR ITS THIRD PARTY SUPPLIERS BE LIABLE TO THE END-USER CUSTOMER FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR OTHER DAMAGES ARISING OUT OF THE USE OF THE ODBC DRIVERS, WHETHER OR NOT INFORMED OF THE POSSIBILITIES OF DAMAGES IN ADVANCE. THESE LIMITATIONS APPLY TO ALL CAUSES OF ACTION, INCLUDING, WITHOUT LIMITATION, BREACH OF CONTRACT, BREACH OF WARRANTY, NEGLIGENCE, STRICT LIABILITY, MISREPRESENTATION AND OTHER TORTS.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, please report them to us in writing at Informatica LLC 2100 Seaport Blvd. Redwood City, CA 94063.

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: 2019-05-09

Table of Contents

Preface

The *Informatica Intelligent Data Lake™ Administrator Guide* contains information about managing access to datasets in the data lake and the users who access the datasets. This book assumes that you have knowledge of the Informatica domain and the Informatica Administrator. It also assumes that you are familiar with the data management processes in your enterprise.

Informatica Resources

Informatica Network

Informatica Network hosts Informatica Global Customer Support, the Informatica Knowledge Base, and other product resources. To access Informatica Network, visit https://network.informatica.com.

As a member, you can:

- Access all of your Informatica resources in one place.
- Search the Knowledge Base for product resources, including documentation, FAQs, and best practices.
- View product availability information.
- Review your support cases.
- Find your local Informatica User Group Network and collaborate with your peers.

Informatica Knowledge Base

Use the Informatica Knowledge Base to search Informatica Network for product resources such as documentation, how-to articles, best practices, and PAMs.

To access the Knowledge Base, visit https://kb.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Documentation

To get the latest documentation for your product, browse the Informatica Knowledge Base at https://kb.informatica.com/_layouts/ProductDocumentation/Page/ProductDocumentSearch.aspx.

If you have questions, comments, or ideas about this documentation, contact the Informatica Documentation team through email at infa_documentation@informatica.com.

Informatica Product Availability Matrixes

Product Availability Matrixes (PAMs) indicate the versions of operating systems, databases, and other types of data sources and targets that a product release supports. If you are an Informatica Network member, you can access PAMs at
https://network.informatica.com/community/informatica-network/product-availability-matrices.

Informatica Velocity

Informatica Velocity is a collection of tips and best practices developed by Informatica Professional Services. Developed from the real-world experience of hundreds of data management projects, Informatica Velocity represents the collective knowledge of our consultants who have worked with organizations from around the world to plan, develop, deploy, and maintain successful data management solutions.

If you are an Informatica Network member, you can access Informatica Velocity resources at http://velocity.informatica.com.

If you have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Marketplace

The Informatica Marketplace is a forum where you can find solutions that augment, extend, or enhance your Informatica implementations. By leveraging any of the hundreds of solutions from Informatica developers and partners, you can improve your productivity and speed up time to implementation on your projects. You can access Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

You can contact a Global Support Center by telephone or through Online Support on Informatica Network.

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
http://www.informatica.com/us/services-and-training/support-services/global-support-centers.

If you are an Informatica Network member, you can use Online Support at http://network.informatica.com.

CHAPTER 1

Introduction to Intelligent Data Lake Administration

This chapter includes the following topics:

Intelligent Data Lake Overview

With the advent of big data technologies, many organizations are adopting a new information storage model called data lake to solve data management challenges. The data lake model is being adopted for diverse use cases, such as business intelligence, analytics, regulatory compliance, and fraud detection.

A data lake is a shared repository of raw and enterprise data from a variety of sources. It is often built over a distributed Hadoop cluster, which provides an economical and scalable persistence and compute layer. Hadoop makes it possible to store large volumes of structured and unstructured data from various enterprise systems within and outside the organization. Data in the lake can include raw and refined data, master data and transactional data, log files, and machine data.

Intelligent Data Lake helps customers derive more value from their Hadoop-based data lake and make data available to all users in the organization.

Organizations are looking to provide ways for different kinds of users to access and work with all of the data in the enterprise, within the Hadoop data lake as well data outside the data lake. They want data analysts and data scientists to be able to use the data lake for ad-hoc self-service analytics to drive business innovation, without exposing the complexity of underlying technologies or the need for coding skills. IT and data governance staff want to monitor data related user activities in the enterprise. Without strong data management and governance foundation enabled by intelligence, data lakes can turn into data swamps.

Intelligent Data Lake is a collaborative self-service big data discovery and preparation solution for data analysts and data scientists. It enables analysts to rapidly discover and turn raw data into insight and allows IT to ensure quality, visibility, and governance. With Intelligent Data Lake, analysts to spend more time on analysis and less time on finding and preparing data.

Intelligent Data Lake provides the following benefits:

- Data analysts can quickly and easily find and explore trusted data assets within the data lake and outside the data lake using semantic search and smart recommendations.
- Data analysts can transform, cleanse, and enrich data in the data lake using an Excel-like spreadsheet interface in a self-service manner without the need for coding skills.
- Data analysts can publish data and share knowledge with the rest of the community and analyze the data using their choice of BI or analytic tools.
- IT and governance staff can monitor user activity related to data usage in the lake.
- IT can track data lineage to verify that data is coming from the right sources and going to the right targets.
- IT can enforce appropriate security and governance on the data lake
- IT can operationalize the work done by data analysts into a data delivery process that can be repeated and scheduled.

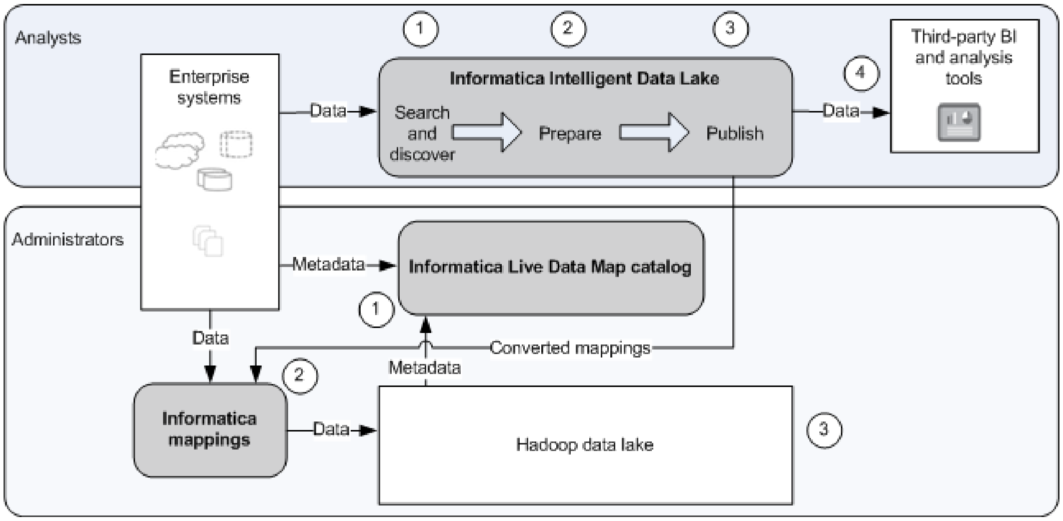
Intelligent Data Lake Users

Intelligent Data Lake users include analysts who search, discover, and prepare data for analysis and administrators who manage the catalog and monitor the data lake.

Analysts work with data that reside in and outside the data lake. Analysts search for, discover, prepare, and publish data back to the data lake so that it is ready for further analysis.

Administrators use Enterprise Data Catalog to extract metadata from information assets and load the metadata into the catalog. Administrators can run Informatica mappings to load data into the data lake. They also monitor data lake usage and work with Hadoop administrators to secure the data in the data lake.

The following image displays the high-level tasks that analysts and administrators complete in Intelligent Data Lake:



Analysts perform the following high-level tasks:

1. Search the catalog for data that resides in and outside the data lake. Discover the lineage and relationships between data that resides in different enterprise systems.

2. Create a project and prepare data.
3. Publish prepared data to the data lake to share and collaborate with other analysts.
4. Optionally use third-party business intelligence or advanced analytic tools to run reports to further analyze the published data.

Administrators perform the following high-level tasks:

1. Use Enterprise Data Catalog to create a catalog of the information assets that reside in and outside the data lake. Enterprise Information Catalog extracts metadata for each asset and indexes the assets in the catalog.
2. Operationalize the Informatica mappings created during the publication process to regularly load data with the new structure into the data lake.
   Optionally develop and run additional Informatica mappings to read data from enterprise systems and load the data to the data lake.
3. Secure the data in the data lake and monitor the usage of the data lake.

Intelligent Data Lake Concepts

To successfully administer Intelligent Data Lake, you must understand the concepts used in the product.

Catalog

A catalog is an indexed inventory of all the information assets in an enterprise. The assets can come from different types of enterprise systems. Assets can include such items as a database table, report, folder, user account, or business glossary definition.

The catalog provides a single comprehensive view of the data in the enterprise. The catalog contains metadata about each asset, including profile statistics, asset ratings, data domains, and data relationships. Metadata can come from scans of enterprise system repositories or from data enrichment by analysts.

You use Enterprise Data Catalog to create the catalog. When you manage the catalog, you create resources to represent data sources. Each resource contains the properties required to connect to the data source, extract metadata from the data source, and load the metadata to the catalog.

Intelligent Data Lake requires that you use Catalog Administrator to create the following resource types for the catalog:

Hive resource for the data lake

Create a Hive resource to represent the data lake. Create a schedule for the resource so that Enterprise Information Catalog regularly extracts metadata from the Hive tables in the data lake. Analysts can use the Intelligent Data Lake application to search for, discover, and prepare data stored in the Hive tables.

Resources for other enterprise systems

Create additional resources to represent other enterprise systems. You can create resources for data integration repositories, ETL and modeling tools and repositories, business glossaries, application databases, and other file storage and databases. Create a schedule for each resource so that Enterprise Information Catalog regularly extracts metadata from the enterprise system.

Analysts can use the Intelligent Data Lake application to search for and discover data that resides in or outside the data lake. Analysts cannot prepare data stored outside the lake. However, when the catalog contains metadata across disparate enterprise systems, analysts can discover lineage and relationships between the data. They use this information to help identify which data they want to prepare and to help identify data that should be added to the data lake.

Domain User resource

Use the DomainUsers resource in Enterprise Data Catalog to create a list of users who have Informatica user accounts and can log in to the Intelligent Data Lake application. Create a schedule for the resource so that Enterprise Data Catalog regularly extracts metadata from the Informatica user accounts.

When the catalog contains user metadata, the user information helps analysts discover the data. When analysts discover a data asset, the Intelligent Data Lake application lists other users who have used and prepared that data asset. This information helps analysts decide whether the data asset contains quality or trusted data.

Data Lake

A data lake is a centralized repository of large volumes of structured and unstructured data. A data lake can contain different types of data, including raw data, refined data, master data, transactional data, log file data, and machine data. In Intelligent Data Lake, the data lake is a Hadoop cluster.

Analysts use the Intelligent Data Lake application to search, discover, and prepare data that resides in the data lake. When analysts prepare the data, they combine, cleanse, and transform the data to create new insights.

Data can be added to the data lake in the following ways:

Analysts use the Intelligent Data Lake application to upload data.

Analysts can upload delimited text files to the data lake. When analysts upload data, Intelligent Data Lake writes the uploaded data to a Hive table in the data lake.

Analysts use the Intelligent Data Lake application to publish prepared data.

When analysts publish prepared data, Intelligent Data Lake writes the transformed input source to a Hive table in the data lake.

Administrators run Informatica mappings to populate the data lake.

As an administrator, you can run Informatica mappings to read data from enterprise systems and write the data to Hive tables in the data lake. You can develop mappings or you can operationalize the mappings created during the Intelligent Data Lake publication process.

Administrators and developers run third-party tools to load data into the data lake.

Administrators, developers, or analysts can use data movement tools from other vendors to load data into the data lake.

Data Asset

A data asset is data that you work with as a unit. A data asset is one of the assets described in the catalog. Data assets can include items such as a flat file, table, or view. A data asset can include data stored in or outside the data lake.

Analysts use the Intelligent Data Lake application to search for and discover any assets included in the catalog. However, analysts can only prepare data assets that are stored in the data lake as Hive tables.

After analysts find the data asset they are interested in, they add the data asset to a project and then prepare the data for analysis. Data preparation includes combining, cleansing, transforming, and structuring data in project worksheets.

Data Publication

Data publication is the process of making prepared data available in the data lake.

When analysts publish prepared data, Intelligent Data Lake writes the transformed input source to a Hive table in the data lake. Other analysts can add the published data to their projects and create new data assets. Or analysts can use a third-party business intelligence or advanced analytic tool to run reports to further analyze the published data.

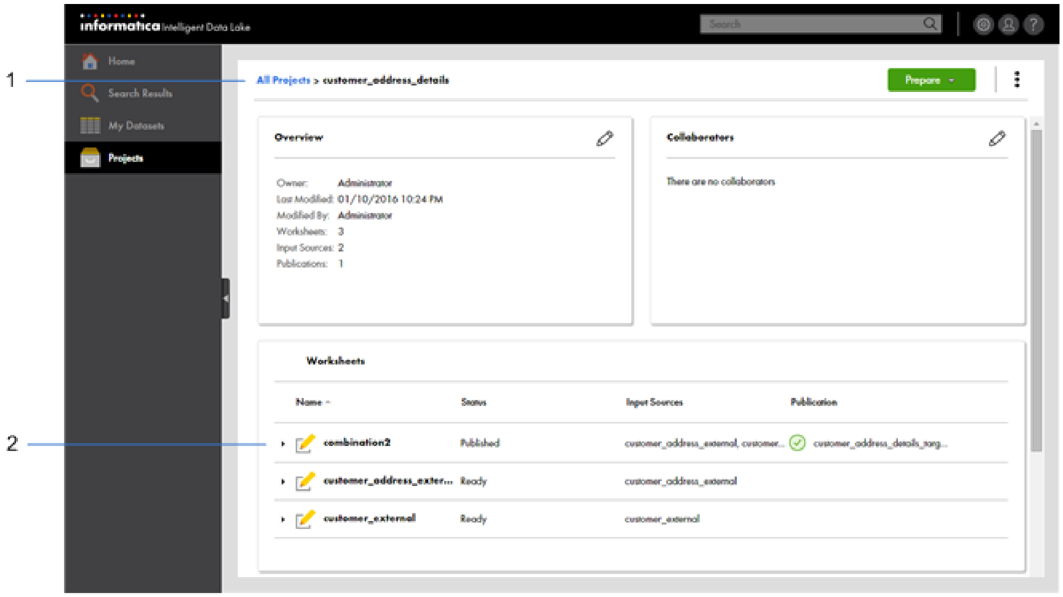
During the publication process, Enterprise Data Catalog scans the published data to immediately add the metadata to the catalog.

Recipes and Mappings

A recipe includes the list of input sources and the steps taken to prepare data in a worksheet. When analysts publish prepared data, Intelligent Data Lake applies the recipe to the data in the input source. Intelligent Data Lake converts the recipe into an Informatica mapping and stores the mapping in the Model repository.

During the publication process, Intelligent Data Lake uses a similar naming convention for the projects and mappings stored in the Model repository. The mappings are accessible from the Developer tool. You can use the Developer tool to view the mappings generated from the recipes. You can operationalize the mappings to regularly load data with the new structure into the data lake.

The following image displays a project named `customer_address_details` in the Intelligent Data Lake application. The project contains one worksheet with published data:



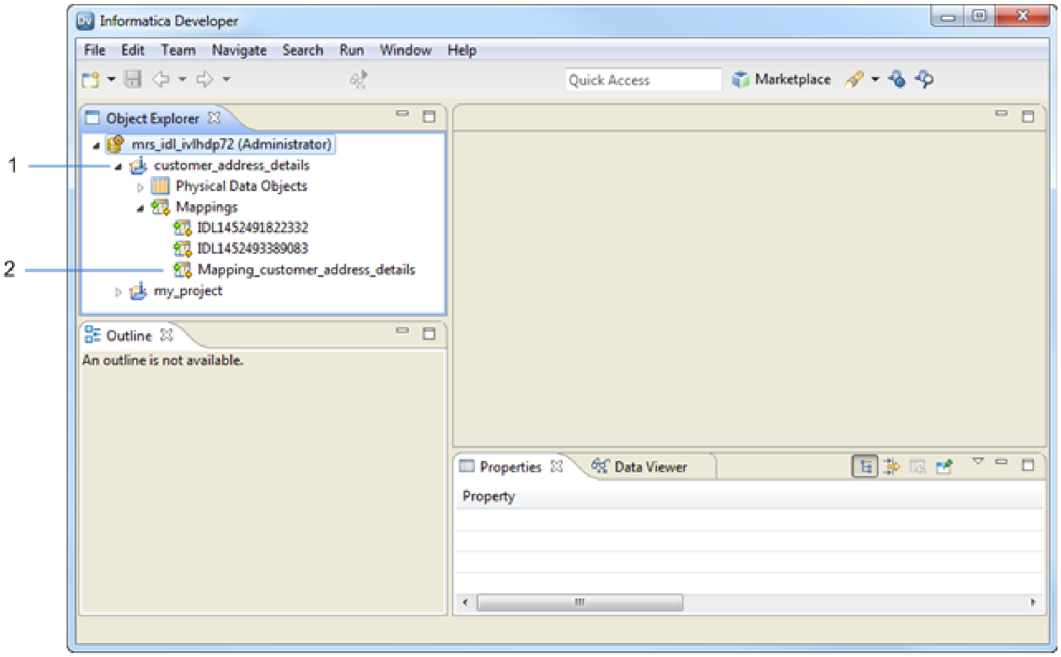
Owner: Administrator
Last Modified: 01/10/2016 10:24 PM
Modified By: Administrator
Worksheets: 3
Input Sources: 2
Publications: 1

There are no collaborators

Name	Status	Input Sources	Publication
combination2	Published	customer_address_external, customer...	✓ customer_address_details_targ...
customer_address_exter...	Ready	customer_address_external	
customer_external	Ready	customer_external	

1. Project in the Intelligent Data Lake application
2. Worksheet with published data

The following image displays the **Object Explorer** view in the Developer tool. The **Object Explorer** view displays the project and converted mapping stored in the Model repository during the publication of the prepared data in the `customer_address_details` project:



1. Project in the Developer tool
2. Converted mapping within the project

Architecture and Components

Intelligent Data Lake uses a number of components to search, discover, and prepare data.

The following image shows the components that Intelligent Data Lake uses and how they interact:

Note: You must install and configure Enterprise Data Catalog before you install Intelligent Data Lake. Enterprise Information Catalog requires additional clients, Informatica services, repositories, and Hadoop services. For more information about Enterprise Data Catalog architecture, see the *Catalog Administrator Guide*.

Clients

Administrators and analysts use several clients to make data available for analysis in Intelligent Data Lake.

Intelligent Data Lake uses the following clients:

Informatica Catalog Administrator

Administrators use Informatica Catalog Administrator to administer the resources, scanners, schedules, attributes, and connections that are used to create the catalog. The catalog represents an indexed inventory of all the information assets in an enterprise.

Informatica Administrator

Administrators use Informatica Administrator (the Administrator tool) to manage the application services that Intelligent Data Lake requires. They also use the Administrator tool to administer the Informatica domain and security and to monitor the mappings run during the upload and publishing processes.

Intelligent Data Lake application

Analysts use the Intelligent Data Lake application to search, discover, and prepare data that resides in the data lake. Analysts combine, cleanse, transform, and structure the data to prepare the data for analysis. When analysts finish preparing the data, they publish the transformed data back to the data lake to make available to other analysts.

Informatica Developer

Administrators use Informatica Developer (the Developer tool) to view the mappings created when analysts publish prepared data in the Intelligent Data Lake application. They can operationalize the mappings so that data is regularly written to the data lake.

Application Services

Intelligent Data Lake requires application services to complete operations. Use the Administrator tool to create and manage the application services.

Intelligent Data Lake requires the following application services:

Intelligent Data Lake Service

The Intelligent Data Lake Service is an application service that runs the Intelligent Data Lake application in the Informatica domain. When an analyst publishes prepared data, the Intelligent Data Lake Service converts each recipe into a mapping.

When an analyst uploads data, the Intelligent Data Lake Service connects to the HDFS system in the Hadoop cluster to temporarily stage the data. When an analyst previews data, the Intelligent Data Lake Service connects to the Hadoop cluster to read from the Hive table.

As analysts complete actions in the Intelligent Data Lake application, the Intelligent Data Lake Service connects to HBase in the Hadoop cluster to store events that you can use to audit user activity.

Data Preparation Service

The Data Preparation Service is an application service that manages data preparation within the Intelligent Data Lake application. When an analyst prepares data in a project, the Data Preparation Service connects to the Data Preparation repository to store worksheet metadata. The service connects to the Hadoop cluster to read sample data or all data from the Hive table, depending on the size of the data. The service connects to the HDFS system in the Hadoop cluster to store the sample data being prepared in the worksheet.

When you create the Intelligent Data Lake Service, you must associate it with a Data Preparation Service.

Catalog Service

The Catalog Service is an application service that runs Enterprise Data Catalog in the Informatica domain. The Catalog Service manages the catalog of information assets in the Hadoop cluster.

When an analyst searches for assets in the Intelligent Data Lake application, the Intelligent Data Lake Service connects to the Catalog Service to return search results from the metadata stored in the catalog.

When you create the Intelligent Data Lake Service, you must associate it with a Catalog Service.

Model Repository Service

The Model Repository Service is an application service that manages the Model repository. When an analyst creates projects, the Intelligent Data Lake Service connects to the Model Repository Service to store the project metadata in the Model repository. When an analyst publishes prepared data, the Intelligent Data Lake Service connects to the Model Repository Service to store the converted mappings in the Model repository.

When you create the Intelligent Data Lake Service, you must associate it with a Model Repository Service.

Data Integration Service

The Data Integration Service is an application service that performs data integration tasks for Intelligent Data Lake. When an analyst uploads data or publishes prepared data, the Intelligent Data Lake Service connects to the Data Integration Service to write the data to a Hive table in the Hadoop cluster.

When you create the Intelligent Data Lake Service, you must associate it with a Data Integration Service.

Repositories

The Intelligent Data Lake Service connects to other application services in the Informatica domain that access data from repositories. The Intelligent Data Lake Service does not directly access any repositories.

Intelligent Data Lake requires the following repositories:

Data Preparation repository

When an analyst prepares data in a project, the Data Preparation Service stores worksheet metadata in the Data Preparation repository.

Model repository

When an analyst creates a project, the Intelligent Data Lake Service connects to the Model Repository Service to store the project metadata in the Model repository. When an analyst publishes prepared data, the Intelligent Data Lake Service converts each recipe to a mapping. The Intelligent Data Lake Service connects to the Model Repository Service to store the converted mappings in the Model repository.

Hadoop Services

Intelligent Data Lake connects to several Hadoop services on a Hadoop cluster to read from and write to Hive tables, to write events, and to store sample preparation data.

Intelligent Data Lake connects to the following services in the Hadoop cluster:

HBase

As analysts complete actions in the Intelligent Data Lake application, the Intelligent Data Lake Service writes events to HBase. You can view the events to audit user activity.

Hadoop Distributed File System (HDFS)

When an analyst uploads data to the data lake, the Intelligent Data Lake Service connects to the HDFS system to stage the data in HDFS files.

When an analyst prepares data, the Data Preparation Service connects to the HDFS system to store the sample data being prepared in worksheets to HDFS files.

When an analyst previews data, the Intelligent Data Lake Service connects to the Data Integration Service and reads the first 100 rows from the mapping using the JDBC driver.

When an analyst prepares data, the Data Preparation Service connects to HDFS system. Depending on the size of the data, the Data Preparation Service reads sample data or all data from the Hive table and displays the data in the worksheet.

When an analyst uploads data, the Intelligent Data Lake Service connects to the Data Integration Service to read the temporary data staged in the HDFS system and write the data to a Hive table. When an analyst publishes prepared data, the Intelligent Data Lake Service connects to the Data Integration Service to run the converted mappings in the Hadoop environment. The Data Integration Service pushes the processing to nodes in the Hadoop cluster. The service applies the mapping to the data in the input source and writes the transformed data to a Hive table.

CHAPTER 2

Administration Process

This chapter includes the following topics:

Overview of the Administration Process

Administrators use multiple tools to manage the catalog, application services, and data lake that make up Intelligent Data Lake.

Administrators complete the following tasks to manage Intelligent Data Lake:

1. Create the catalog of the information assets in the enterprise.
2. Configure and enable the Intelligent Data Lake Service and the Data Preparation Service.
3. Add users and grant them privileges to access the Intelligent Data Lake application.
4. Provide users access to the data stored in the data lake.
5. Operationalize Informatica mappings converted from recipes to regularly load data with the new structure into the data lake.
6. Regularly monitor data lake usage and activity.
7. Regularly monitor the catalog and the Enterprise Information Catalog tasks that extract metadata from the data lake and from enterprise systems outside the data lake.

Step 1. Configure Big Data Management

Use the cluster configuration to configure the Hadoop cluster.

A cluster configuration is an object in the domain that contains configuration information about the Hadoop cluster. The cluster configuration enables the Data Integration Service to push mapping logic to the Hadoop environment. The cluster configuration displays properties in configuration sets that are based on *-site.xml files on the cluster.

For more information about the integration tasks, see the *Informatica Big Data Management Hadoop Integration Guide*.

Tools to complete this step:

- Informatica Administrator

Step 2. Create the Catalog

Use the Catalog Administrator to create the catalog of information assets in the enterprise.

Complete the following steps to create the catalog:

1. Use the Catalog Administrator to create the following resources:
   - Hive resources for the data lake

     Create a Hive resource to represent the data lake. Analysts can use the Intelligent Data Lake application to search for, discover, and prepare data stored in the Hive tables. Include `"data lake"` in the resource name as an indicator to analysts that this resource represents the data lake. Choose to extract both source metadata and profiling metadata from the data lake.
   - Resources for other enterprise systems that are outside the data lake

     Create additional resources to represent other enterprise systems. You can create resources for data integration repositories, ETL and modeling tools and repositories, business glossaries, application databases, and other file storage and databases. Analysts can use the Intelligent Data Lake application to search for and discover data that resides outside the data lake. Analysts cannot prepare data stored outside the lake. Choose to extract both source metadata and profiling metadata from the enterprise systems.
   - Domain User resource

     Create a Domain User resource to represent users who have Informatica user accounts and log in to the Intelligent Data Lake application. When analysts discover a data asset, the Intelligent Data Lake application lists other users who have used and prepared that data asset.
2. Run a scan on the resources to load metadata into the catalog.
3. Create schedules for the resources so that Enterprise Data Catalog regularly scans the resources. As a best practice, schedule the resource scans to run during non-business hours.

Tools to complete this step:

- Informatica Catalog Administrator

Step 3. Configure the Intelligent Data Lake Application Services

Intelligent Data Lake requires the Intelligent Data Lake Service and the Data Preparation Service. You can create the application services during or after the Intelligent Data Lake installation.

If you created the services during the installation, use the Administrator tool to verify that the services are correctly configured and enabled.

If you chose to create the services after the installation, use the Administrator tool to create, configure, and enable the Intelligent Data Lake Service and the Data Preparation Service.

Tools to complete this step:

- Informatica Administrator

Related Topics:

- "Creating and Managing the Data Preparation Service" on page 33
- "Creating and Managing the Intelligent Data Lake Service" on page 47

Step 4. Add Users and Grant them Privileges

Set up user accounts in the Informatica domain and grant the users privileges to access the Intelligent Data Lake application. You can create user accounts in the Informatica domain or import LDAP user accounts from an LDAP directory service.

You can add the user accounts to a group to create a set of related users that have the same authorization. Assign the group the required privileges on the Intelligent Data Lake Service or assign the group a role that includes the required privileges on the Intelligent Data Lake Service.

Tools to complete this step:

- Informatica Administrator

Step 5. Provide Access to Data in the Data Lake

Intelligent Data Lake user accounts must be authorized to access the Hive tables in the Hadoop cluster designated as the data lake. Intelligent Data Lake user accounts access Hive tables in the Hadoop cluster when they preview data, upload data, and publish prepared data.

As an administrator, grant access to the data assets that an Intelligent Data Lake user.

When an Intelligent Data Lake user requests access to data assets, use established processes and follow best practices to grant permission to data assets. You must grant analysts the appropriate permissions to the data lake Hadoop cluster. You can set up user impersonation to run mappings in the Hadoop cluster when an analyst publishes prepared data.

Tools to complete this step:

- Third-party Hadoop tools

Step 6. Operationalize Mappings Converted from Recipes

When analysts publish data in the Intelligent Data Lake application, the Intelligent Data Lake Service converts the preparation recipes into Informatica mappings. You can operationalize the mappings to regularly load data with the new structure into the data lake.

Use the Developer tool to view and edit the converted mappings. The mappings use the same name as the Intelligent Data Lake worksheet with the published data. Verify that the converted mappings meet your requirements, and then deploy the mappings.

Use the Administrator tool or the infacmd command line program to run the mappings to load and transform the data. The Data Integration Service writes the data to Hive tables in the data lake. You can schedule the deployed mappings to regularly load data into the data lake. Use the Administrator tool to monitor each mapping run.

For more information about developing and running mappings that write to Hive, see the *Informatica Big Data Management User Guide*.

Tools to complete this step:

- Informatica Developer
- Informatica Administrator

Step 7. Monitor Data Lake Usage

As analysts use the Intelligent Data Lake application to prepare data in the data lake, you can regularly monitor the usage and activity on the data lake.

You can monitor the following activities:

Hadoop jobs that run in the Hadoop cluster when analysts upload and publish data

During the upload and publishing process, the Data Integration Service runs mappings in the Hadoop environment. The Data Integration Service pushes the processing to nodes in the Hadoop cluster. You can monitor the mappings that run during the upload and publishing processes on the **Monitor** tab of the Administrator tool. The mappings are listed as ad hoc jobs on the **Monitor** tab. For more information about monitoring mappings in a Hadoop environment, see the *Informatica Big Data Management Guide*.

Events that the Intelligent Data Lake Service writes to HBase in the Hadoop cluster

You can audit user activity on the data lake by viewing the events that the Intelligent Data Lake Service writes to HBase. The events include user activity in the Intelligent Data Lake application, such as when a user creates a project, adds data assets to a project, or publishes prepared data.

Tools to complete this step:

- Informatica Administrator
- Third-party reporting tools

Step 8. Monitor the Catalog

As analysts use the Intelligent Data Lake application to search for data in the catalog, you can regularly monitor the catalog and the Enterprise Information Catalog tasks that extract metadata from the data lake and from enterprise systems outside the data lake.

Use the Catalog Administrator to track the status and schedule of Enterprise Information Catalog tasks. You can monitor the duration of the tasks that are running. You can also monitor the resource distribution and track the number of resources for each resource type. To perform a detailed analysis of Enterprise Information Catalog performance, open the **Monitoring** workspace in Catalog Administrator.

For more information about monitoring the catalog, see the *Informatica Catalog Administrator Guide*.

Tools to complete this step:

- Catalog Administrator

Access the Administrator Tools

Use Informatica Administrator and Catalog Administrator to administer the application services and the catalog that make up Intelligent Data Lake.

Log In to Informatica Administrator

You must have a user account to log in to the Informatica Administrator web application.

If the Informatica domain runs on a network with Kerberos authentication, you must configure the browser to allow access to the Informatica web applications. In Microsoft Internet Explorer and Google Chrome, add the URL of the Informatica web application to the list of trusted sites. If you are using Chrome version 41 or later, you must also set the `AuthServerWhitelist` and `AuthNegotiateDelegateWhitelist` policies.

1. Start a Microsoft Internet Explorer or Google Chrome browser.
2. In the **Address** field, enter the URL for the Administrator tool:
   - If the Administrator tool is not configured to use a secure connection, enter the following URL:
     ```
     http://<fully qualified hostname>:<http port>/administrator/
     ```
   - If the Administrator tool is configured to use a secure connection, enter the following URL:
     ```
     https://<fully qualified hostname>:<http port>/administrator/
     ```

   Host name and port in the URL represent the host name and port number of the master gateway node. If you configured secure communication for the domain, you must use HTTPS in the URL to ensure that you can access the Administrator tool.

   If you use Kerberos authentication, the network uses single sign on. You do not need to log in to the Administrator tool with a user name and password.
3. If you do not use Kerberos authentication, enter the user name, password, and security domain for your user account, and then click **Login**.

   The **Security Domain** field appears when the Informatica domain contains an LDAP security domain. If you do not know the security domain that your user account belongs to, contact the Informatica domain administrator.

   Note: If this is the first time you log in with the user name and password provided by the domain administrator, change your password to maintain security.

Log In To Catalog Administrator

Use either Microsoft Internet Explorer or Google Chrome to log in to Catalog Administrator.

1. Start Microsoft Internet Explorer or Google Chrome.
2. In the Address field, enter the URL for the Catalog Administrator login page in the following format:

   ```
   http://<host>:<port>/catalogadmin
   ```

   The host is the gateway node host name. The port represents the port number configured for the Catalog Service.
3. In the Catalog Administrator login page, enter the user name and password.
4. Verify that the default domain option **Native** is selected. You can also select an LDAP domain.

   The Domain field appears when the Informatica domain contains an LDAP security domain.
5. Click **Log In**.

Example - Administration Process

You are an administrator in a multinational retail organization. Your organization has configured a data lake on a Hadoop cluster to store operational data, transactional data, log file data, and social media data from stores worldwide. You use Informatica Big Data Management to read data from the disparate enterprise systems and write the data to Hive in the Hadoop data lake.

The data analysts in the Marketing department need to discover the social media data stored in the data lake and prepare it for further analysis. The data analysts in the Sales department need to discover and prepare the transactional data. As the analysts search for data in the lake, they want to view the lineage of the data across the various enterprise systems. They also want to view how the data is related to other assets in the enterprise catalog. After finding the data they are interested in, the analysts want to combine and transform the data on their own, without requiring much involvement from the administrators on your team. They need to prepare the data so that it can be analyzed further using a more advanced third-party business intelligence tool.

As the administrator, you perform the following tasks to enable data analysts to discover data across all enterprise systems and to prepare data stored in the Hadoop data lake for further analysis:

- Use Catalog Administrator to create the following resources for the catalog:
  - Hive resource for the data lake
  - Domain User resource
  - Additional resources for enterprise systems that are outside the data lake

  Verify that Enterprise Information Catalog successfully extracts metadata from these resources so that the metadata exists in the catalog. Create schedules for the resources so that Enterprise Information Catalog regularly scans the resources.
- Use the Administrator tool to configure the Intelligent Data Lake Service and the Data Preparation Service.
- Use the Administrator tool to import LDAP users from an LDAP directory service. Assign the user accounts in the Marketing department to a Marketing group, and the user accounts in the Sales department to a Sales group. Assign each group the privilege to access the Intelligent Data Lake application.
- Use Hadoop tools to grant the user accounts in the Marketing and Sales groups access to the Hive tables in the data lake.

- Use the Administrator tool to regularly monitor the Hadoop jobs that run when analysts upload and publish data in the Intelligent Data Lake application.
- Use Catalog Administrator to regularly monitor the metadata extraction from the resources for the Hadoop data lake, the domain users, and the enterprise systems outside the data lake.
- When an analyst requests that you operationalize an Informatica mapping converted from the preparation recipe during the publication process, use the Developer tool to review and deploy the mapping. Use the Administrator tool to schedule and run the deployed mapping to regularly load data with the new structure into the data lake.

CHAPTER 3

User Account Setup

This chapter includes the following topics:

User Account Setup Overview

You must set up user accounts in the Informatica domain to access the Intelligent Data Lake application and services. The Intelligent Data Lake user accounts must also have access to the Hadoop clusters where the data lake is created.

Intelligent Data Lake uses impersonation to provide users access to the data lake. Set up impersonation for Intelligent Data Lake users based on the data access requirements of your organization

To set up Intelligent Data Lake user accounts, perform the following tasks:

1. Set up user accounts in the Informatica domain.
2. Assign Intelligent Data Lake privileges.
3. Assign connection privileges.
4. Set up user access to the data lake.
5. Set up user impersonation.

Intelligent Data Lake uses the Data Integration Service to upload and publish data and perform operations in the data lake. The Data Integration Service uses impersonation to identify the Hadoop user that can access data and perform operations in the data lake. To use impersonation, you must configure the Data Integration Service to use operating system profiles. Follow the guidelines to use operating system profiles in Intelligent Data Lake.

Step 1. Set up User Accounts

You set up Intelligent Data Lake user accounts in the same manner as other user accounts in the Informatica domain. You can create Native user accounts in the Informatica domain or import LDAP user accounts from an LDAP directory service into the domain.

You can set up the following types of user accounts:

Native user accounts

You can create native user accounts for Intelligent Data Lake. Native user accounts are created, stored, and managed in the Informatica domain.

LDAP user accounts

If you use LDAP or Kerberos authentication, you can import LDAP user accounts for Intelligent Data Lake. Use the Administrator tool to create an LDAP security domain and set up a connection to the LDAP server. Specify the users and groups from the LDAP directory service that can have access to Intelligent Data Lake. You can create a synchronization schedule to regularly synchronize the user accounts in the security domain and the LDAP server.

You can organize users into groups and assign privileges and data access based on the requirements of each group.

For more information about creating user accounts in the Informatica domain, see the *Informatica Security Guide*.

Step 2. Assign Intelligent Data Lake Privileges

To enable users to log in to the Intelligent Data Lake application, assign privileges on the Intelligent Data Lake service to the user accounts. If you organize users into groups, you can assign privileges to the group.

You can assign the specific privileges required by Intelligent Data Lake or a role that has the required privileges. Informatica recommends that you assign the Administrator role to Intelligent Data Lake users.

1. In the Administrator tool, click the **Security** tab.
2. In the Navigator, select a user or group.
3. Click the **Privileges** tab.
4. Click Edit.

   The **Edit Roles and Privileges** dialog box appears.
5. On the **Roles** tab, expand the Intelligent Data Lake service.

   The list of roles available for the Intelligent Data Lake service appears.
6. Select the Administrator role.
7. click **OK.**

Step 3. Assign Connection Privileges

Intelligent Data Lake uses connection objects to connect to the data lake. You must assign privileges on the Hadoop connections to Intelligent Data Lake users so they can access the data assets in the lake.

Assign Intelligent Data Lake users privileges on the following connections to the data lake:

- Hive connection
- HDFS connection
- JDBC connection with Sqoop

Use the same steps to assign privileges to Intelligent Data Lake users on each connection.

1. In the Administrator tool, click **Manage** > **Connections**.
2. In the Navigator, select the corresponding data lake connection.
3. In the contents panel, select the **Permissions** view.
4. Select **Groups** or **Users**.

   You can assign permissions based on how you organize Intelligent Data Lake users in the domain.
5. Click the **Actions** menu on the **Manage** tab.
6. Select **Assign Permission**.

   The **Assign Permissions** dialog box displays all users or groups that do not have permission on the connection. If you do not see the user or group to which you want to assign permissions, you can filter the list displayed on the dialog box by group or user name or by security domain.
7. Select a user or group, and click **Next**.
8. Select **Allow** for the permission type that you want to assign.

   Intelligent Data Lake users require read and write permissions on each connection to the data lake.
9. Click **Finish.**

Step 4. Set Up User Access to the Data Lake

To be able to perform operations in Intelligent Data Lake, users must have access to the data lake. Intelligent Data Lake users must have access to the Hive data and metadata in the data lake.

Intelligent Data Lake supports storage-based authorization. Access rights to Hive databases and tables are determined by the access rights to the underlying HDFS directories.

Assign users access to the data lake cluster based on the tasks that users perform. To preview, prepare, or export data, a user must have read access to data and metadata in the Hive database. To upload or publish data, a user must have write access to the Hive database. The user requires write access to create tables or to overwrite or append to existing tables.

The following table describes the user tasks and the permission required to perform the tasks:

Task	Permission
Data preview	Read and execute permission on the data asset.
Export	Read and execute permission on the data asset.

Task	Permission
Data upload	Read, write, and execute permission on the temporary HDFS location. Write permission on the Hive schema where the data is uploaded.
Data preparation	Read and execute permission on the data asset to be prepared.
Publish	Read and execute permission on the data asset to be published. Write permission on the Hive schema where the data asset is published.

Step 5. Set Up User Impersonation

When a user performs tasks in Intelligent Data Lake, the Data Integration Service connects to Hadoop services to access the data lake and perform the operation on behalf of the user. The Data Integration Service uses impersonation to pass the user account to Hadoop.

Configure user impersonation for Intelligent Data Lake based on the user access requirements of your organization.

You can configure user impersonation in the following ways:

Using the Intelligent Data Lake user account to connect to Hadoop services

You can configure the Data Integration Service to impersonate the user account logged in to the Intelligent Data Lake application. The user account logged in to Intelligent Data Lake must have user or group authorization to connect to the Hadoop services. To use this option, you must configure the Data Integration Service to use operating system profiles. You must assign the Intelligent Data Lake user account a default operating system profile with the **Use the logged in user as Hadoop Impersonation User** option selected.

Using an authorized user account to connect to Hadoop services

If the Intelligent Data Lake user accounts do not have logged authorization to connect Hadoop, you can configure the Data Integration Service to impersonate a specific authorized user. The impersonated user account must have authorization to connect to the data lake and the Hadoop services. To use this option, you must configure the Data Integration Service to use operating system profiles. The operating system profile must have the **Use the specified user as Hadoop Impersonation User** option selected and must specify the authorized user account to impersonate.

Using the Hive connection user account to connect to Hadoop services

If user access to the data lake does not require authorization or if all users have the same authorization, you do not need to set up operating system profiles to connect to the Hadoop services. The Data Integration Service connects to Hadoop services using the user account specified in the Hive connection object for Intelligent Data Lake.

To configure user impersonation for Intelligent Data Lake, perform the following steps:

- Configure user impersonation in the Hadoop cluster.
- Configure the user account in the Hive connection.

Configuring User Impersonation in the Hadoop Cluster

To enable the Data Integration Service to impersonate a user account on the Hadoop cluster, configure user impersonation in Hadoop. Set up the user account that runs the Data Integration Service as a proxy user in Hadoop.

The Hadoop configuration file core-site.xml defines the proxy user accounts that can impersonate other users. You can set the properties directly in the configuration file or use the Hadoop management tool for your Hadoop distribution to set the properties.

In the core-site.xml file, the following properties specify the impersonation settings:

hadoop.proxyuser.<user account>.groups

Defines the groups that a user account can impersonate. The user account can impersonate any member of the groups that you specify.

hadoop.proxyuser.<user account>.hosts

Defines the machines that a user account can connect from to impersonate the members of a group. The host machine you specify must be a machine where the Data Integration Service runs.

For more information about how to enable user impersonation in the Hadoop cluster, see the documentation for your Hadoop distribution.

Configuring the User Account in the Hive Connection

The Hive connection object defines the Data Integration Service user account that can impersonate Intelligent Data Lake users. Set the user name in the Hive connection object you use to connect to the data lake.

1. In the Administrator tool, click **ManageConnections**.
2. In the Navigator, select the Hive connection to the data lake.
3. In the properties view, edit the **Common Attribute to Both the Modes** section.

   The **Edit Common Attribute to Both the Modes** dialog box appears.
4. Set the user name and password for the Data Integration Service user account you want to use to impersonate Intelligent Data Lake users.

   The user account must be the user account specified for user impersonation in the Hadoop cluster.
5. Click **OK.**

For more information about connection objects and the Hive connection properties, see the *Informatica Administrator Guide*.

Using Operating System Profiles in Intelligent Data Lake

You can configure the Data Integration Service to use the permissions of the default operating system profile assigned to a user to perform operations initiated by the user.

You can assign a user a default operating system profile. If a user belongs to a group, the user inherits the default operating system profile of the group. If the user belongs to multiple groups with different default operating system profiles, the user can have multiple operating system profiles. The Data Integration Service cannot determine which profile to use for the user and cannot perform any operation initiated by the user.

Use the following guidelines when you assign an operating system profile to a user:

- A user must have only one default operating system profile.
- A user that is not assigned a default operating system profile cannot belong to multiple groups with multiple default operating system profiles.
- To avoid conflicts in group and user permissions, create a group that includes all users who require the same profile and assign a default operating system profile to the group.

For more information about creating an operating system profile, see the *Informatica Security Guide*.

For more information about configuring the Data Integration Service to use an operating system profile, see the *Informatica Application Service Guide*.

CHAPTER 4

Data Preparation Service

This chapter includes the following topics:

Data Preparation Service Overview

The Data Preparation Service is an application service that manages data preparation within the Intelligent Data Lake application.

When an analyst prepares data in a project, the Data Preparation Service connects to the Data Preparation repository to store worksheet metadata. The service connects to the Hadoop cluster to read sample data or all data from the Hive table, depending on the size of the data. The service connects to the HDFS system in the Hadoop cluster to store the sample data being prepared in the worksheet.

The Data Preparation Service uses an Oracle database or a MySQL database for the data preparation repository. You must configure a local storage location for data preparation file storage on the node on which the Data Preparation Service runs. The Data Preparation Service uses the Apache Solr indexing capabilities to provide recommendations of related data assets. This Solr instance does not run on the Hadoop cluster and is managed by the Data Preparation Service.

When you create the Intelligent Data Lake Service, you must associate it with a Data Preparation Service.

Before You Create the Data Preparation Service

Before you create the Data Preparation Service, complete the prerequisite tasks for the service.

Perform the following tasks before you create the Data Preparation Service, if Oracle is the database for the Data Preparation Service Repository:

- Set up the Oracle server database version 12cR2 or 11gR2 that the Data Preparation Service connects to. Ensure that the database is case insensitive.

- Set up the required user account for the Oracle database with create, drop, and alter permissions for tables and views.

Perform the following tasks before you create the Data Preparation Service, if MySQL is the database for the Data Preparation Service Repository:

- Set up the MySQL server database version 5.6.26 or above that the Data Preparation Service connects to. Ensure that the database is case insensitive. For MySQL version 5.6.26 and above, set lower_case_table_names=1 and for MySQL version 5.7 and above, set explicit_defaults_for_timestamp=1 in the my.cnf file.
- Set up the required user account for the MySQL database with create, drop, and alter permissions for tables and views.

If the domain is secure, you must secure the services that you create for use by Intelligent Data Lake.

- The following services in the domain and the YARN application must share the same common truststore file:
  - Data Integration Service
  - Model Repository Service
  - Catalog Service
  - Data Preparation Service
  - Intelligent Data Lake Service
- The Data Preparation Service and Intelligent Data Lake Service must also share the same keystore file.
- You can use different keystore files for the Data Integration Service, Model Repository Service, and Catalog Service. If you use different keystore files, you must add certificates corresponding to each of the keystores into a common truststore file.
- If you have configured Intelligent Data Lake with one primary node and one or more backup nodes, you must copy the truststore files to a common directory and specify the same directory path for all nodes assigned to Intelligent Data Lake.

Creating and Managing the Data Preparation Service

Use the Administrator tool to create and manage the Data Preparation service. When you change a service property, you must recycle the service or disable and then enable the service for the changes to take affect.

Creating the Data Preparation Service

Use the service creation wizard in the Administrator tool to create the service.

1. In the Administrator tool, click the **Manage** tab.
2. Click the **Services and Nodes** view.
3. In the Domain Navigator, select the domain.
4. Click **Actions** > **New** > **Data Preparation Service.**

5. On the New Data Preparation Service - Step 1 of 7 page, enter the following properties:

Property	Description
Name	Name of the Data Preparation service. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain spaces or the following special characters: ` ~ % ^ * + = { } \ ; : ' " / ? . , < > | ! ( )] [
Description	Description of the Data Preparation service. The description cannot exceed 765 characters.
Location	Location of the Data Preparation Service in the Informatica domain. You can create the service within a folder in the domain.
License	License object with the data lake option that allows the use of the Data Preparation Service.
Node Assignment	Type of node in the Informatica domain on which the Data Preparation Service runs. Select **Single Node** if a single service process runs on the node or **Primary and Backup Nodes** if a service process is enabled on each node for high availability. However, only a single process runs at any given time, and the other processes maintain standby status. The **Primary and Backup Nodes** option will be available for selection based on the license configuration. Select the **Grid** option to ensure horizontal scalability by using grid for the Data Preparation Service with multiple Data Preparation Service nodes. Improved scalability supports high performance, interactive data preparation during increased data volumes and number of users. Each user is assigned a node in the grid using round-robin method to distribute the load across the nodes. Default is Single Node.
Node	Name of the node on which the Data Preparation Service runs.
Backup Nodes	If your license includes high availability, nodes on which the service can run if the primary node is unavailable.
Grid	If you selected Grid for node assignment, select the grid that you want to use for the Data Preparation Service.

6. Click **Next**.
7. On the New Data Preparation Service - Step 2 of 7 page, enter the following properties:

Property	Description
Model Repository Service Name	Name of the Model Repository Service. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain spaces or the following special characters: ` ~ % ^ * + = { } \ ; : ' " / ? . , < > | ! ( )] [You cannot change the name of the service after you create it.
Model Repository Service User Name	User name to access the Model Repository Service.
Model Repository Service Password	Password to access the Model Repository Service.

Property	Description
Security Domain	Select the security domain to access the Model Repository Service.
Data Integration Service Name	Name of the Data Integration Service.

8. Click **Next**.
9. On the New Data Preparation Service - Step 3 of 7 page, enter the following properties:

Property	Description
HTTP Port	Port number for the HTTP connection to the Data Preparation Service.
Enable Secure Communication	Use a secure connection to connect to the Data Preparation Service. If you enable secure communication, you must set all required HTTPS properties, including the keystore and truststore properties.
HTTPS Port	Port number for the HTTPS connection to the Data Preparation Service.
Keystore File	Path and the file name of keystore file that contains key and certificates required for HTTPS communication.
Keystore Password	Password for the keystore file.
Truststore File	Path and the file name of truststore file that contains authentication certificates for the HTTPS connection.
Truststore Password	Password for the truststore file.

10. Click **Next**.
11. On the New Data Preparation Service - Step 4 of 7 page, enter the following properties:

Property	Description
Database Type	Type of database to use for the Data Preparation repository. Oracle and MySQL databases are supported.
Host Name	Host name of the machine that hosts the Data Preparation repository database. This field appears if you select MySQL as your Database Type.
Port Number	Port number for the database. This field appears if you select MySQL as your Database Type.
Connection String	Connection used to access the Oracle database. Use the following connection string: `jdbc:informatica:oracle://<machineName>:<PortNo>;ServiceName=<DBName>` This field appears if you select Oracle as your Database Type.

Property	Description
Secure JDBC Parameters	If the database is secured, information such as TrustStore and TrustStorePassword can be included in this field. It is saved in an encrypted format. Parameters usually configured include the following: *EncryptionMethod=<encryption method>;HostNameInCertificate=<hostname>;TrustStore=<truststore file with its location>;TrustStorePassword=<truststore password>;KeyStore==<keystore file with its location>;KeyStorePassword=<keystore password>;ValidateServerCertificate=<true /false>* This field appears if you select Oracle as your Database Type.
Database User Name	Database user account to use to connect to the Data Preparation repository.
Database User Password	Password for the Data Preparation repository database user account.
Schema Name	Schema or database name of the Data Preparation repository database. This field appears if you select MySQL as your Database Type.

12. Click **Next**.
13. On the New Data Preparation Service - Step 5 of 7 page, enter the following properties:

Property	Description
Solr Port	Port number for the Apache Solr server used to provide data preparation recommendations.

14. Click **Next**.
15. On the New Data Preparation Service - Step 6 of 7 page, enter the following properties:

Property	Description
Local Storage Location	Directory for data preparation file storage on the node on which the Data Preparation Service runs.
HDFS Connection	HDFS connection for data preparation file storage.
HDFS Storage Location	HDFS location for data preparation file storage. If the connection to the local storage fails, the Data Preparation Service recovers data preparation files from the HDFS location.
Hadoop Authentication Mode	Security mode of the Hadoop cluster for data preparation storage. If the Hadoop cluster uses Kerberos authentication, you must set the required Hadoop security properties for the cluster.
HDFS Service Principal Name	Service Principal Name (SPN) for the data preparation Hadoop cluster. Specify the service principal name in the following format: user/_HOST@REALM.

Property	Description
Hadoop Impersonation User Name	User name to use in Hadoop impersonation as set in core-site.xml.
SPN Keytab File for User Impersonation	Path and file name of the SPN keytab file for the user account to impersonate when connecting to the Hadoop cluster. The keytab file must be in a directory on the machine where the Data Preparation Service runs.

16. Click **Next**.
17. On the New Data Preparation Service - Step 7 of 7 page, enter the following properties:

Property	Description
Hadoop Distribution	Hadoop distribution used for the data lake Hadoop cluster. Select Cloudera or HortonWorks.

18. Click **Finish**.

Note: After you create the Data Preparation Service, ensure that you complete the following steps:

1. In the Administrator tool, click **Actions** > **Create Repository** to create the repository contents.
2. Click **Actions** > **Enable Service**.

Enabling, Disabling, and Recycling the Data Preparation Service

You can enable, disable, and recycle the service from the Administrator tool.

1. In the Administrator tool, click the **Manage tab** > **Services and Nodes view**.
2. In the Domain Navigator, select the service.
3. On the **Actions** tab, select one of the following options:
    a. **Enable Service** to enable the service.
    b. **Disable Service** to disable the service.

       Choose the mode to disable the service in. Optionally, you can choose to specify whether the action was planned or unplanned, and enter comments about the action. If you complete these options, the information appears in the Events and Command History panels in the Domain view on the Manage tab.
    c. **Recycle Service** to recycle the service.

Editing the Data Preparation Service

To edit the Data Preparation Service, select the service in the Domain Navigator and click the Properties view. You can change the properties while the service is running, but you must restart the service for the properties to take effect.

To edit the Data Preparation Service:

1. To edit specific properties, click the pencil icon in the selected properties area.
2. In the **Edit Properties** window, edit the required fields.
3. Click **OK**.

4. Click **Actions** > **Recycle Service**.
5. In the **Recycle Service** window, select the required options.
6. Click **OK** to restart the service.

Deleting the Data Preparation Service

Only users with ADMIN or WRITE permissions for the Data Preparation Service can delete the service.

To delete the Data Preparation Service:

1. On the **Manage** tab, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Data Preparation Service.
3. Disable the Data Preparation Service by clicking **Actions** > **Disable Service** .
4. To delete the Data Preparation Service, click **Actions** > **Delete**.

Data Preparation Service Properties

To view the Data Preparation Service properties, select the service in the Domain Navigator and click the Properties View. You can edit the properties by clicking the pencil icon in the respective area, while the service is running, but you must restart the service for the properties to take effect. You can configure the following Data Preparation Service properties:

- General Properties
- Data Preparation Repository Options
- Data Preparation Storage Options
- Hive Security Options
- Hadoop Options
- Custom Properties

General Properties

General properties for the Data Preparation Service include the name, description, and the node in the Informatica domain that the service runs on.

To edit the general properties, click the pencil icon in the General Properties area. In the **Edit General Properties** window, edit the required fields.
The following table describes the general properties for the service:

Property	Description
Name	Name of the Data Preparation Service. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain spaces or the following special characters: ` ~ % ^ * + = { } \ ; : ' " / ? . , < > | ! ( )] [
Description	Description of the service. The description cannot exceed 765 characters.

Property	Description
License	License object with the data lake option that allows use of the service.
Node Assignment	Type of node in the Informatica domain on which the Data Preparation Service runs. Select **Single Node** if a single service process runs on the node or **Primary and Backup Nodes** if a service process is enabled on each node for high availability. However, only a single process runs at any given time, and the other processes maintain standby status. The **Primary and Backup Nodes** option will be available for selection based on the license configuration. Default is Single Node.
Node	Name of the node on which the Data Preparation Service runs.
Grid	If you selected Grid for node assignment, select the grid that you want to use for the Data Preparation Service. A grid ensures horizontal scalability. Improved scalability supports high performance, interactive data preparation during increased data volumes and number of users. Each user is assigned a node in the grid using round-robin method to distribute the load across the nodes.

Model Repository Options and Data Integration Service Options

To edit the model repository options, click the pencil icon in the Model Repository Options area. In the **Edit Model Repository Options** window, edit the required fields.

The following table describes the model repository options:

Property	Description
Model Repository Service Name	Name of the Model Repository Service. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain spaces or the following special characters: ` ~ % ^ * + = { } \ ; : ' " / ? . , < > | ! ( )] [You cannot change the name of the service after you create it.
Model Repository Service User Name	User name to access the Model Repository Service.
Model Repository Service Password	Password to access the Model Repository Service.
Security Domain	Select the security domain to access the Model Repository Service.

To edit the Data Integration Service options, click the pencil icon in the Data Integration Service Options area. In the **Edit Data Integration Service Options** window, edit the required fields.

The following table describes the Data Integration Service options:

Property	Description
Data Integration Service Name	Name of the Data Integration Service.

Data Preparation Repository Options

To edit the data preparation repository options, click the pencil icon in the Data Preparation Repository Options area. In the **Edit Data Preparation Repository Options** window, edit the required fields.

The following table describes the data preparation repository options:

Property	Description
Database Type	Type of database to use for the Data Preparation repository.
Host Name	Host name of the machine that hosts the Data Preparation repository database. This property appears only if your Database Type is MySQL.
Database Port Number	Port number for the database. This property appears only if your Database Type is MySQL.
Connection String	Connection used to access the Oracle database. Use the following connection string: `jdbc:informatica:oracle://<machineName>:<PortNo>;ServiceName=<DBName>` This field appears if you select Oracle as your Database Type.
Secure JDBC Parameters	If the database is secured, information such as TrustStore and TrustStorePassword can be included in this field. It is saved in an encrypted format. Parameters usually configured include the following: *EncryptionMethod=<encryption method>;HostNameInCertificate=<hostname>;TrustStore=<truststore file with its location>;TrustStorePassword=<truststore password>;KeyStore==<keystore file with its location>;KeyStorePassword=<keystore password>;ValidateServerCertificate=<true /false>* This field appears if you select Oracle as your Database Type.
Database User Name	Database user account to use to connect to the Data Preparation repository.
Database User Password	Password for the Data Preparation repository database user account.
Modify Database User Password	Select this checkbox to modify the database user password.
Schema Name	Schema or database name of the Data Preparation repository database. This field appears if you select MySQL as your Database Type.

Data Preparation Storage Options

Data preparation storage options enables you to specify the local storage and HDFS location for data persistence.

To edit the data preparation storage options, click the pencil icon in the Data Preparation Storage Options area. In the **Edit Data Preparation Storage Options** window, edit the required fields.

The following table describes the data preparation storage options:

Property	Description
Local Storage Location	Directory for data preparation file storage on the node on which the Data Preparation Service runs.
HDFS Connection	HDFS connection for data preparation file storage.
HDFS Storage Location	HDFS location for data preparation file storage. If the connection to the local storage fails, the Data Preparation Service recovers data preparation files from the HDFS location. If the Hadoop cluster uses Kerberos authentication, the impersonation user name must have read, write and execute permission on the HDFS storage location directory. The default location is: /datalake/dps_durable_storage.
Hadoop Distribution Directory	Location of the Hadoop distribution for data preparation storage.

Hive Security Options

You can specify the Hive security options for the Hadoop cluster.

To edit the Hive security options, click the pencil icon in the Hive Security Options area. In the **Edit Hive Security Options** window, edit the required fields.

The following table describes the Hive security options:

Property	Description
Hadoop Authentication Mode	Security mode of the Hadoop cluster for data preparation storage. If the Hadoop cluster uses Kerberos authentication, you must set the required Hadoop security properties for the cluster.
HDFS Principal Service Name	Service Principal Name (SPN) for the data preparation Hadoop cluster. Specify the service principal name in the following format: user/_HOST@REALM.
Hadoop Impersonation User Name	User name for the user account to impersonate when connecting to the data preparation Hadoop cluster.
SPN Keytab File for User Impersonation	Path and file name of the SPN keytab file for the impersonation user account. The keytab file must be in a directory on the machine where the Data Preparation Service runs.

Custom Properties

Configure custom properties that are unique to specific environments. You might need to apply custom properties in special cases.

When you define a custom property, enter the property name and an initial value. Define custom properties only at the request of Informatica Global Customer Support.
To view the custom properties, select the service in the Domain Navigator and click the Properties view. You can change the properties while the service is running, but you must restart the service for the properties to take effect.

To edit the custom properties, click the pencil icon in the Custom Properties area. In the **Edit Custom Properties** window, edit the required fields.

Data Preparation Service Process Properties

A service process represents a service running on a node.

To configure properties for the Data Preparation Service processes, click the **Processes** view. Select the node to configure properties specific to that node.

You can edit service process properties such as the HTTP configuration, Solr options, advanced options, and custom properties. You can change the properties while the Data Preparation Service process is running, but you must restart the process for the changed properties to take effect.

HTTP Configuration Options

The HTTP configuration options specify the HTTP or HTTPS port. The properties also specify the keystore file and truststore file to use when the Data Preparation Service process uses the HTTPS protocol.

To edit the HTTP configuration options, click the pencil icon in the HTTP Configuration Options area. In the **Edit HTTP Configuration Options** window, edit the required fields.

The following table describes the HTTP configuration options for a Data Preparation Service process:

Property	Description
HTTP Port	Port number for the HTTP connection to the Data Preparation Service.
Enable Secure Communication	Use a secure connection to the Data Preparation Service. If you enable secure communication, you must enter all required HTTPS options.
HTTPS Port	Port number for the HTTPS connection to the Data Preparation Service.
Keystore File	Path and the file name of the keystore file that contains key and certificates required for the HTTPS communication.
Keystore Password	Password for the keystore file.
Modify Keystore Password	Select this checkbox if you want to modify the keystore password.
Truststore File	Path and the file name of the truststore file that contains authentication certificates for the HTTPS connection.
Truststore Password	Password for the truststore file.
Modify Truststore Password	Select this checkbox if you want to modify the truststore password.

Solr Options

Intelligent Data Lake uses the Apache Solr indexing capabilities to provide recommendations of related data assets.

To edit the Solr options, click the pencil icon in the Solr Options area. In the **Edit Solr Options** window, edit the required fields.

The following table describes the Solr options:

Property	Description
Solr Port	Port number for the Apache Solr server used to provide data preparation recommendations.

Advanced Options

You can set the maximum heap size and Java Virtual Machine (JVM) options from the Advanced Options area.

To edit the advanced options, click the pencil icon in the Advanced Options area. In the **Edit Advanced Options** window, edit the required fields.

The following table describes the advanced options:

Property	Description
Maximum Heap Size	Maximum amount of RAM in megabytes to allocate to the Java Virtual Machine (JVM) that runs the Data Preparation Service.
JVM Command Line Options	JVM command line options for the Data Preparation Service processes.

Custom Properties

Configure custom properties that are unique to specific environments. You might need to apply custom properties in special cases.

When you define a custom property, enter the property name and an initial value. Define custom properties only at the request of Informatica Global Customer Support.
To view the custom properties, select the service in the Domain Navigator and click the Properties view. You can change the properties while the service is running, but you must restart the service for the properties to take effect.

To edit the custom properties, click the pencil icon in the Custom Properties area. In the **Edit Custom Properties** window, edit the required fields.

Configuring Data Preparation Service on Grid for Scalability

The Data Preparation Service requires most memory and CPU resources for in-memory database to support high performance interactive data preparation. When too many users try to prepare data simultaneously, performance of the interactive preparation can decline. The administrator might need to upgrade the hardware to improve the performance levels. To support increased preparation data volumes, the administrator can achieve horizontal scaling by creating a Data Preparation Service Grid with multiple Data Preparation Service nodes.

Each user is assigned a node in the grid using round-robin method to distribute the load across the nodes. Homogeneous combinations of nodes are allowed. You can combine nodes with the same operating system, same CPU, same memory, and security setup. This allows for seamless restoration of data after node failures, enabling the Data Preparation Service to be highly available.

1. Install the Intelligent Data Lake binaries on every node that is part of the grid.
2. Select Grid while configuring the Data Preparation Service.
3. Make sure all the folder locations mentioned in the configuration are present in all the nodes.

You can add or remove nodes dynamically from a grid. When a node is added into an active grid, the Data Preparation Service process will not start automatically. The Intelligent Data Lake administrator must enable the process in the **Processes** tab of the Data Preparation Service to start the process in the node.

Adding a New Node when the Data Preparation Service is Running

When you add a new node to the grid when the Data Preparation Service instance is running, the new node will be in the Disabled state.

1. Log in to the Administrator tool.
2. Click **Services**.
3. Select the Data Preparation Service from the list.
4. Click the Processes tab of the service.
5. Select the newly added node.
6. On the top right hand corner, click the **Enable** icon to start the process.

   A warning message appears.
7. Click **OK**.

Removing Nodes from the Grid

At least one node should be active for the Data Preparation Service to execute without being disabled.

When you shut down a node or a node goes down, it does not affect the Data Preparation Service as long as at least one node remains enabled in the grid. An active session will not be automatically recovered. An error will appear and the user must reconnect the session to continue. If all the nodes in a grid are removed or shut down, the Data Preparation Service will be disabled.

Monitoring Node States

You can troubleshoot by finding the state of the nodes at any given point in time.

To find the nodes of the service along with the state, connect to the Data Preparation Service repository and execute an SQL query on the metadata repository as follows:

```
select node_id, node_ip, state, created_ts, node_port, isp_node_name from dp_physical
node;
```

The state column shows the current state of the node service. It can be in any of the following states:

- ACTIVE: The node is ready to take new user sessions.
- SUSPECTED_UNREACHABLE: The node cannot accept new sessions as peer-check operation is failing on that node. The node might not be completely down as the server may recover after a brief period of high load.

To find the user to node assignment, connect to the Data Preparation Service repository and execute an SQL query on the metadata repository as follows:

```
select login_id, node_ip, a.node_id, isp_node_name from dp_physical_node a, dp_user u,
dp_user_to_node_map m where a.node_id = m.node_id and u.id = m.user_id;
```

CHAPTER 5

Intelligent Data Lake Service

This chapter includes the following topics:

Intelligent Data Lake Service Overview

Intelligent Data Lake requires the Intelligent Data Lake Service to complete operations. The Intelligent Data Lake Service is an application service that runs the Intelligent Data Lake application in the Informatica domain.

The Intelligent Data Lake application allows data analysts to create data preparation projects. Each step of the data preparation project is stored in a recipe that is translated into a mapping for execution on the Informatica platform.

When an analyst uploads data, the Intelligent Data Lake Service connects to the HDFS system in the Hadoop cluster to temporarily stage the data. When an analyst previews data, the Intelligent Data Lake Service connects to the Hadoop cluster to read from the Hive table.

As analysts complete actions in the Intelligent Data Lake application, the Intelligent Data Lake application connects to HBase in the Hadoop cluster to store events that you can use to audit user activity.

Before You Create the Intelligent Data Lake Service

Before you create the Intelligent Data Lake Service, complete the prerequisite tasks for the service.

Perform the following tasks before you create the Intelligent Data Lake Service:

- Verify that the Informatica domain has the following services which must be associated with the Intelligent Data Lake Service:
  - Data Integration Service
  - Model Repository Service

    Note: Ensure that the fined-grained-parsing and sync-search flags are set to **true** for the Model Repository Service.

    - Catalog Service
    - Content Management Service must be configured if you want to use the data domain discovery feature. Ensure that the database client is installed on all nodes of the Hadoop cluster with the correct settings in the hadoopEnv.properties file.
    - Data Preparation Service

      Note: AnIntelligent Data Lake Service and a Data Preparation Service must have a one to one association. Do not associate one Intelligent Data Lake Service with multiple Data Preparation Service instances, or one Data Preparation Service with multiple Intelligent Data Lake Service instances.

- If you use HTTPS to connect to the Intelligent Data Lake Service, verify the location and password of the keystore and truststore files.
  If the domain is secure, you must secure each Intelligent Data Lake Service that you create in Intelligent Data Lake. The Intelligent Data Lake Service instances must use the same keystore and truststore files that the domain uses. If you use separate security certificates, you must add the security certificates for the Intelligent Data Lake Service to the truststore and keystore file for the domain. You must use these keystore and truststore files for Intelligent Data Lake. You must import the certificate file to the truststore location for the Data Preparation Service and Intelligent Data Lake Service.

  You must also use the same truststore files for the following services:

  - Data Integration Service
  - Model Repository Service
  - Catalog Service
  - Data Preparation Service
  - Intelligent Data Lake Service

If the domain is secure, you must secure the services that you create in Intelligent Data Lake.

- The following services in the domain and the YARN application must share the same common truststore file:

  - Data Integration Service
  - Catalog Service
  - Data Preparation Service
  - Intelligent Data Lake Service

  Note: You can use different keystore files for the Data Integration Service, Model Repository Service, and Catalog Service. If you use different keystore files, you must add certificates corresponding to each of the keystores into a common truststore file.

- The Data Preparation Service and Intelligent Data Lake Service must also share the same keystore file.
- If you have configured Intelligent Data Lake with one primary node and one or more backup nodes, you must copy the truststore files to a common directory and specify the same directory path for all nodes assigned to Intelligent Data Lake.

Creating and Managing the Intelligent Data Lake Service

Use the Administrator tool to create and manage the Intelligent Data Lake Service. When you change a service property, you must recycle the service or disable and then enable the service for the changes to take affect.

Creating the Intelligent Data Lake Service

Use the service creation wizard in the Administrator tool to create the service.

1. In the Administrator tool, click the **Manage** tab.
2. Click the **Services and Nodes** view.
3. In the Domain Navigator, select the domain.
4. Click **Actions** > **New** > **Enterprise Data Lake Service.**
5. Enter the following properties:

Property	Description
Name	Name of the Intelligent Data Lake Service. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain spaces or the following special characters: ` ~ % ^ * + = { } \ ; : ' " / ? . , < > | ! ( )] [
Description	Description of theIntelligent Data Lake Service. The description cannot exceed 765 characters.
Location	Location of the Intelligent Data Lake Service in the Informatica domain. You can create the service within a folder in the domain.
License	License object that allows the use of the Intelligent Data Lake Service.
Node Assignment	Type of node in the Informatica domain on which the Intelligent Data Lake Service runs. Select **Single Node** if a single service process runs on the node or **Primary and Backup Nodes** if a service process is enabled on each node for high availability. However, only a single process runs at any given time, and the other processes maintain standby status. The **Primary and Backup Nodes** option is available based on the license configuration. Default is Single Node.
Node	Name of the node on which theIntelligent Data Lake Service runs.
Backup Nodes	If your license includes high availability, nodes on which the service can run if the primary node is unavailable.

6. Click **Next**.

7. Enter the following properties for the Model Repository Service:

Property	Description
Model Repository Service	Name of the Model Repository Service associated with the Intelligent Data Lake Service.
Model Repository Service User Name	User account to use to log in to the Model Repository Service.
Model Repository Service User Password	Password for the Model Repository Service user account.

8. Click **Next**.
9. Enter the following properties for the Data Preparation Service, Data Integration Service, and Catalog Service:

Property	Description
Data Preparation Service	Name of the Data Preparation Service associated with the Intelligent Data Lake Service.
Data Integration Service	Name of the Data Integration Service associated with the Intelligent Data Lake Service.
Catalog Service	Name of the Catalog Service associated with the Intelligent Data Lake Service.
Catalog Service User Name	User account to use to log in to the Catalog Service.
Catalog Service User Password	Password for the Catalog Service user account.
Data Lake Resource Name	Hive resource for the data lake. You configure the resource in Catalog Administrator.

10. Click **Next**.
11. Enter the following properties:

Property	Description
Hadoop Authentication Mode	Security mode of the Hadoop cluster for the data lake. If the Hadoop cluster uses Kerberos authentication, you must set the required Hadoop security properties for the cluster.
Principal Name for User Impersonation	Service principal name (SPN) of the user account to impersonate when connecting to the data lake Hadoop cluster. The user account for impersonation must be set in the core-site.xml file.

Property	Description
SPN Keytab File for User Impersonation	Path and file name of the SPN keytab file for the user account to impersonate when connecting to the Hadoop cluster. The keytab file must be in a directory on the machine where the Intelligent Data Lake Service runs.
HBase User Name	User name with permissions to access the HBase database.

12. Click **Next**.
13. Enter the following properties:

Property	Description
HDFS Connection	HDFS connection for the data lake.
HDFS Working Directory	HDFS directory where the Intelligent Data Lake Service copies temporary data and files necessary for the service to run.
Hadoop Distribution Directory	Directory that contains Hadoop distribution files on the machine where the Intelligent Data Lake Service runs.
Hive Connection	Hive connection for the data lake.
Hive Table Storage Format	Data storage format for the Hive tables. Select from the following options: - DefaultFormat - Parquet - ORC
Hadoop Connection	Hadoop connection for the data lake.

14. Click **Next**.
15. Enter the following properties:

Property	Description
Log User Activity Events	Indicates whether the Intelligent Data Lake Service logs the user activity events for auditing. The user activity logs are stored in an Hbase instance. If this option is enabled, select a HBase connection.
HBase Namespace	Namespace for the HBase tables.
Maximum File Size for Uploads (MB)	The maximum size of the files that can be uploaded.
Maximum Number of Rows to Download	Number of rows to export to a .csv file. You can specify a maximum of 2,000,000,000 rows. Enter a value of -1 to export all rows.
Number of Recommendations to Display	The number of recommended data assets to display on the Projects page. You can specify a maximum of 50 recommendations. A value of 0 means no recommendations will be displayed. You can use recommended alternate or additional data assets to improve productivity.

Property	Description
Maximum Data Preparation Sample Size	The maximum number of sample rows to fetch for data preparation. You can specify a maximum number of 1,000,000 rows.
Default Data Preparation Sample Size	Number of sample rows to fetch for data preparation. You can specify a maximum number of 1,000,000 rows and a minimum of 1,000 rows.

16. Click **Next**.
17. Enter the following properties:

Property	Description
Zeppelin URL	The URL to access the Zeppelin framework. The URL should be in this format: `http[s]://zepplinHost:PortNo`
Log Severity	Severity of messages to include in the logs. Select from one of the following values: - FATAL. Writes FATAL messages to the log. FATAL messages include nonrecoverable system failures that cause the service to shut down or become unavailable. - ERROR. Writes FATAL and ERROR code messages to the log. ERROR messages include connection failures, failures to save or retrieve metadata, service errors. - WARNING. Writes FATAL, WARNING, and ERROR messages to the log. WARNING errors include recoverable system failures or warnings. - INFO. Writes FATAL, INFO, WARNING, and ERROR messages to the log. INFO messages include system and service change messages. - TRACE. Write FATAL, TRACE, INFO, WARNING, and ERROR code messages to the log. TRACE messages log user request failures. - DEBUG. Write FATAL, DEBUG, TRACE, INFO, WARNING, and ERROR messages to the log. DEBUG messages are user request logs. Default value is INFO.
Log Directory	Location of the directory to save the log files.
Hive Execution Engine	The Hive execution engine for Intelligent Data Lake Service.
Local System Directory	Location of the local system directory.

18. Enter the following properties:

Property	Description
HTTP Port	Port number for the HTTP connection to the Intelligent Data Lake Service.
Enable Secure Communication	Use a secure connection to connect to the Intelligent Data Lake Service. If you enable secure communication, you must enter all required HTTPS options.
HTTPS Port	Port number for the HTTPS connection to the Intelligent Data Lake Service.
Keystore File	Path and the file name of keystore file that contains key and certificates required for the HTTPS connection.

Property	Description
Keystore Password	Password for the keystore file.
Truststore File	Path and the file name of the truststore file that contains authentication certificates for the HTTPS connection.
Truststore Password	Password for the truststore file.
Enable Service	Select this checkbox if you want to enable the service immediately after you create the service. If you want to enable the service at a later time, in the Domain Navigator, select the service and then select **Actions** > **Enable Service**.

19. Click **Finish**.

Enabling, Disabling and Recycling the Intelligent Data Lake Service

You can enable, disable, and recycle the service from the Administrator tool.

1. In the Administrator tool, click the **Manage tab** > **Services and Nodes view**.
2. In the Domain Navigator, select the service.
3. On the **Actions** tab, select one of the following options:
    a. **Enable Service** to enable the service.
    b. **Disable Service** to disable the service.

       Choose the mode to disable the service in. Optionally, you can choose to specify whether the action was planned or unplanned, and enter comments about the action. If you complete these options, the information appears in the Events and Command History panels in the Domain view on the Manage tab.
    c. **Recycle Service** to recycle the service.

Editing the Intelligent Data Lake Service

To edit the Intelligent Data Lake Service, select the service in the Domain Navigator and click the Properties view. You can change the properties while the service is running, but you must restart the service for the properties to take effect.

To edit the Intelligent Data Lake Service:

1. To edit specific properties, click the pencil icon in the selected properties area.
2. In the **Edit Properties** window, edit the required fields.
3. Click **OK**.
4. Click **Actions** > **Recycle Service**.
5. In the **Recycle Service** window, select the required options.
6. Click **OK** to restart the service.

Deleting the Intelligent Data Lake Service

Only users with ADMIN or WRITE permissions for the Intelligent Data Lake Service can delete the service.

To delete the Intelligent Data Lake Service:

1. On the **Manage** tab, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. Disable the Intelligent Data Lake Service by clicking **Actions** > **Disable Service** .
4. To delete the Intelligent Data Lake Service, click **Actions** > **Delete**.

Intelligent Data Lake Service Properties

To view the Intelligent Data Lake Service properties, select the service in the Domain Navigator and click the Properties View. You can edit the properties by clicking the pencil icon in the respective area, while the service is running, but you must restart the service for the properties to take effect. You can configure the following Intelligent Data Lake Service properties:

- General Properties
- Model Repository Service Options
- Data Preparation Service Options
- Data Integration Service Options
- Catalog Service Options
- Data Lake Security Options
- Data Lake Options
- Execution Options
- Event Logging Options
- Export Options
- Data Asset Recommendation Options
- Sampling Options
- Logging Options
- Custom Options

General Properties

General properties for the Intelligent Data Lake Service include the name, description, license, and the node in the Informatica domain that the Intelligent Data Lake Service runs on.

To edit the general properties, click the pencil icon in the general properties area. In the **Edit General Properties** window, edit the required fields.

The following table describes the general properties for the service:

Property	Description
Name	Name of the Intelligent Data Lake Service. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain spaces or the following special characters: ` ~ % ^ * + = { } \ ; : ' " / ? . , < > | ! ( )] [
Description	Description of the Intelligent Data Lake Service. The description cannot exceed 765 characters.
License	License object with the data lake option that allows the use of the Intelligent Data Lake Service.
Node Assignment	Type of node in the Informatica domain on which the Intelligent Data Lake Service runs. Select **Single Node** if a single service process runs on the node or **Primary and Backup Nodes** if a service process is enabled on each node for high availability. However, only a single process runs at any given time, and the other processes maintain standby status. The **Primary and Backup Nodes** option will be available for selection based on the license configuration. Default is Single Node.
Node	Name of the node on which the Intelligent Data Lake Service runs.
Backup Nodes	If your license includes high availability, nodes on which the service can run if the primary node is unavailable.

Model Repository Service Options

The Model Repository Service is an application service that manages the Model repository. When an analyst creates projects, the Model Repository Service connects to the Model repository to store the project metadata. When you create the Intelligent Data Lake Service, you must associate it with a Model Repository Service using the Model Repository Service Options properties.

To edit the Model Repository Service options, click the pencil icon. In the **Edit Model Repository Service Options**window, edit the required fields.
The following table describes the Model Repository Service options:

Property	Description
Model Repository Service	Name of the Model Repository Service associated with the Intelligent Data Lake Service.
Model Repository Service User Name	User account to use to log in to the Model Repository Service.
Model Repository Service Password	Password for the Model Repository Service user account.
Modify Repository Password	Select the checkbox to modify the Model Repository Service user password.
Security Domain	LDAP security domain for the Model repository user. The field appears when the domain contains an LDAP security domain.

Data Preparation Service Options

The Data Preparation Service is an application service that manages data preparation within the Intelligent Data Lake application. When you create the Intelligent Data Lake Service, you must associate it with a Data Preparation Service using the Data Preparation Service options.

To edit the Data Preparation Service options, click the pencil icon. In the **Edit Data Preparation Service Options**window, edit the required fields.
The following table describes the Data Preparation Service options:

Property	Description
Data Preparation Service	Name of the Data Preparation Service associated with the Intelligent Data Lake Service.

Data Integration Service Options

The Data Integration Service is an application service that performs data integration tasks for Intelligent Data Lake. When you create the Intelligent Data Lake Service, you must associate it with a Data Integration Service using the Data Integration Service options.

To edit the Data Integration Service options, click the pencil icon. In the **Edit Data Integration Service Options** window, edit the required fields.
The following table describes the Data Integration Service options:

Property	Description
Data Integration Service	Name of the Data Integration Service associated with the Intelligent Data Lake Service.

Catalog Service Options

The catalog represents an indexed inventory of all the configured assets in an enterprise. You can find metadata and statistical information, such as profile statistics, data asset ratings, data domains, and data relationships, in the catalog. The catalog options will be based on the Catalog Service configuration you have set up when you installed Enterprise Data Catalog.

To edit the catalog service options, click the pencil icon in the Catalog Service Options area. In the **Edit Catalog Service Options** window, edit the required fields.

The following table describes the catalog service options:

Property	Description
Catalog Service	Name of the Catalog Service associated with the Intelligent Data Lake Service.
Catalog Service User Name	User account to use to log in to the Catalog Service.
Catalog Service User Password	Password for the Catalog Service user account.
Modify Catalog Service User Password	Select this checkbox to modify the Catalog Service user password.

Property	Description
Data Lake Resource Names	Hive resources for the data lake. You configure the resources in Catalog Administrator.
Security Domain	LDAP security domain for the Catalog Service user. The field appears when the domain contains an LDAP security domain.

Data Lake Security Options

Select the security mode and specify the related details using the Data Lake Security Options.

To edit the data lake security options, click the pencil icon in the Data Lake Security Options area. In the **Edit Data Lake Security Options** window, edit the required fields.

The following table describes the data lake security options:

Property	Description
Hadoop Authentication Mode	Security mode of the Hadoop cluster for the data lake. If the Hadoop cluster uses Kerberos authentication, you must set the required Hadoop security properties for the cluster.
Principal Name for User Impersonation	Service principal name (SPN) of the user account to impersonate when connecting to the data lake Hadoop cluster. The user account for impersonation must be set in the core-site.xml file.
SPN Keytab File for User Impersonation	Path and file name of the SPN keytab file for the user account to impersonate when connecting to the Hadoop cluster. The keytab file must be in a directory on the machine on which the Intelligent Data Lake Service runs.
HBase User Name	User name with permissions to access the HBase database.

Data Lake Options

The data lake options include the Hive, HDFS, and Hadoop configuration details.

To edit the data lake options, click the pencil icon in the Data Lake Options area. In the **Edit Data Lake Options** window, edit the required fields.

The following table describes the data lake options:

Property	Description
HDFS Connection	HDFS connection for the data lake.
HDFS Working Directory	HDFS directory where the Intelligent Data Lake Service copies temporary data and files necessary for the service to run. This directory must have 777 permissions to enable users to upload data.
Hadoop Distribution Directory	Directory that contains Hadoop distribution files on the machine where Intelligent Data Lake Service runs. The directory must be within the Informatica directory. The default directory is <Informatica installation directory>/services/shared/hadoop/<hadoop distribution name>.

Property	Description
Hive Connection	Hive connection for the data lake.
Hive Table Storage Format	Data storage format for the Hive tables. Select from the following options: - DefaultFormat - Parquet - ORC
Hadoop Connection	Hadoop connection for the data lake.

Event Logging Options

Use the Event Logging Options area to configure user activity event logging options.

To edit the event logging options, click the pencil icon. In the **Edit Event Logging Options** window, edit the required fields.

Note: If the Hadoop cluster uses Kerberos authentication, the HBase administrator must create a namespace and provide 'Read, Write, Create' permissions to the Intelligent Data Lake user. The user details must be stored in the keytab provided while configuring the IIntelligent Data Lake Service.

The following table describes the event logging options:

Property	Description
Log User Activity Events	Indicates whether the Intelligent Data Lake Service logs the user activity events for auditing. The user activity logs are stored in an HBase instance.
HBase Connection	Connection for the HBase database.
HBase Namespace	Namespace for the HBase tables.

Upload and Download Options

After the project is published to the data lake, you can export the data to a .csv file and save it on the local drive. You can also upload flat files into the application.

To edit the upload and download options, click the pencil icon in the Upload and Download Options area. In the **Edit Upload and Download Options** window, edit the required fields.

The following table describes the upload and download options:

Property	Description
Maximum File Size for Uploads (MB)	Maximum size of the files that the users upload. The default value is 1 GB.
Download Rows Size	Number of rows to export to a .csv file. You can specify a maximum of 2,000,000,000 rows. Enter a value of -1 to export all rows.

Data Asset Recommendation Options

The recommendation options are used to define the number of recommended data assets that can be displayed on the Projects page in the Intelligent Data Lake application.

To edit the data asset recommendation options, click the pencil icon in the Data Asset Recommendation Options area. In the **Edit Data Asset Recommendation Options** window, edit the required fields.

The following table describes the data asset recommendation options:

Property	Description
Number of Recommendations to Display	The number of recommended data assets to display on the Projects page. You can specify a maximum of 50 recommendations. A value of 0 means no recommendations will be displayed. You can use recommended alternate or additional data assets to improve productivity.

Sampling Options

You can specify the sample size to be retrieved in the Intelligent Data Lake application in the Intelligent Data Lake Service properties.

To edit the sampling options, click the pencil icon in the Sampling Options area. In the **Edit Sampling Options** window, edit the required fields.

The following table describes the sampling options:

Property	Description
Maximum Data Preparation Sample Size	The maximum number of sample rows to fetch for data preparation. You can specify a maximum number of 1,000,000 rows.
Default Data Preparation Sample Size	The default number of sample rows to fetch for data preparation. You can specify a maximum number of 1,000,000 rows and a minimum of 1,000 rows.

Apache Zeppelin Options

You can specify the Apache Zeppelin URL in the Intelligent Data Lake properties.

To edit the Zeppelin options, click the pencil icon in the Zeppelin Options area. In the **Edit Zeppelin Options** window, edit the required fields.

The following table describes the Zeppelin options:

Property	Description
Zeppelin URL	The URL to access the Zeppelin framework. The URL should be formatted as: `http[s]://zepplinHost:PortNo`

Note that if Apache Zeppelin uses a Spark 1.x version, you must specify the Spark version in an environment variable named sparkVersion in theIntelligent Data Lake Service process properties. For more information, see "Apache Zeppelin Options" on page 60.

Logging Options

Logging options include properties for the severity level for service logs. Configure the Log Severity property to set the logging level.

To edit the logging options, click the pencil icon in the Logging Options area. In the **Edit Logging Options** window, edit the required fields.

The following table describes the logging options:

Property	Description
Log Severity	Severity of messages to include in the logs. Select from one of the following values: - FATAL. Writes FATAL messages to the log. FATAL messages include nonrecoverable system failures that cause the service to shut down or become unavailable. - ERROR. Writes FATAL and ERROR code messages to the log. ERROR messages include connection failures, failures to save or retrieve metadata, service errors. - WARNING. Writes FATAL, WARNING, and ERROR messages to the log. WARNING errors include recoverable system failures or warnings. - INFO. Writes FATAL, INFO, WARNING, and ERROR messages to the log. INFO messages include system and service change messages. - TRACE. Write FATAL, TRACE, INFO, WARNING, and ERROR code messages to the log. TRACE messages log user request failures. - DEBUG. Write FATAL, DEBUG, TRACE, INFO, WARNING, and ERROR messages to the log. DEBUG messages are user request logs.
Log Directory	Location of the directory of log files.

Execution Options

Execution options include properties for the execution engine and the local system directory.

To edit the execution options, click the pencil icon in the Execution Options area. In the **Edit Execution Options** window, edit the required fields.

The following table describes the execution options:

Property	Description
Hive Execution Engine	Engine for running the mappings in the Hadoop environment.
Local System Directory	Local directory that contains the files downloaded from Intelligent Data Lake application, such as .csv or .tde files

Custom Options

Configure custom properties that are unique to specific environments. You might need to apply custom properties in special cases.

When you define a custom property, enter the property name and an initial value. Define custom properties only at the request of Informatica Global Customer Support.
To view the custom options, select the service in the Domain Navigator and click the Properties view. You can change the properties while the service is running, but you must restart the service for the properties to take effect.

To edit the custom options, click the pencil icon in the Custom Options area. In the **Edit Custom Options** window, edit the required fields.

Intelligent Data Lake Service Process Properties

A service process is the physical representation of a service running on a node. When the Intelligent Data Lake Service runs on multiple nodes, a Intelligent Data Lake Service process can run on each node with the service role. You can configure the service process properties differently for each node.

To configure properties for the Intelligent Data Lake Service processes, click the **Processes** view. Select a node to configure properties specific to that node.

You can edit service process properties such as the HTTP port, advanced options, custom properties, and environment variables. You can change the properties while the Intelligent Data Lake Service process is running, but you must restart the process for the changed properties to take effect.

HTTP Configuration Options

The HTTP configuration options specify the keystore and truststore file to use when the Intelligent Data Lake Service uses the HTTPS protocol.

To edit the HTTP configuration options, click the pencil icon in the HTTP Configuration Options area. In the **Edit HTTP Configuration Options** window, edit the required fields. The following table describes the HTTP configuration options for a Intelligent Data Lake Service process:

Property	Description
HTTP Port	Port number for the HTTP connection to the Intelligent Data Lake Service.
Enable Secure Communication	Use a secure connection to connect to the Intelligent Data Lake Service. If you enable secure communication, you must enter all required HTTPS options.
HTTPS Port	Port number for the HTTPS connection to the Intelligent Data Lake Service.
Keystore File	Path and the file name of keystore file that contains key and certificates required for the HTTPS connection.
Keystore Password	Password for the keystore file.
Truststore File	Path and the file name of the truststore file that contains authentication certificates for the HTTPS connection.
Truststore Password	Password for the truststore file.

Advanced Options

You can set the maximum heap size and Java Virtual Machine (JVM) options from the Advanced Options area.

To edit the advanced options, click the pencil icon in the Advanced Options area. In the **Edit Advanced Options** window, edit the required fields.

The following table describes the advanced options:

Property	Description
Maximum Heap Size	Maximum amount of RAM in megabytes to allocate to the Java Virtual Machine (JVM) that runs the Intelligent Data Lake Service.
JVM Command Line Options	JVM command line options for the Intelligent Data Lake Service processes.

Custom Options

Configure custom properties that are unique to specific environments. You might need to apply custom properties in special cases.

When you define a custom property, enter the property name and an initial value. Define custom properties only at the request of Informatica Global Customer Support.
To view the custom options, select the service in the Domain Navigator and click the Properties view. You can change the properties while the service is running, but you must restart the service for the properties to take effect.

To edit the custom options, click the pencil icon in the Custom Options area. In the **Edit Custom Options** window, edit the required fields.

Environment Variables

You can configure environment variables for the Intelligent Data Lake Service process.

The following table describes the environment variables:

Property	Description
Environment variable	Enter a name and a value for the environment variable.

Apache Zeppelin Options

If Apache Zeppelin uses Spark 1.x version, you must specify the Spark version in an environment variable named sparkVersion in the Intelligent Data Lake Service process properties.

You do not need to add the environment variable if Zeppelin uses a Spark 2.x version.

To add the environment variable, click the pencil icon in the Environment Variables area. The following table describes the sparkVersion environment variable:

Property	Description
sparkVersion	The Spark 1.x version used by Apache Zeppelin.

CHAPTER 6

Application Configuration

This chapter includes the following topics:

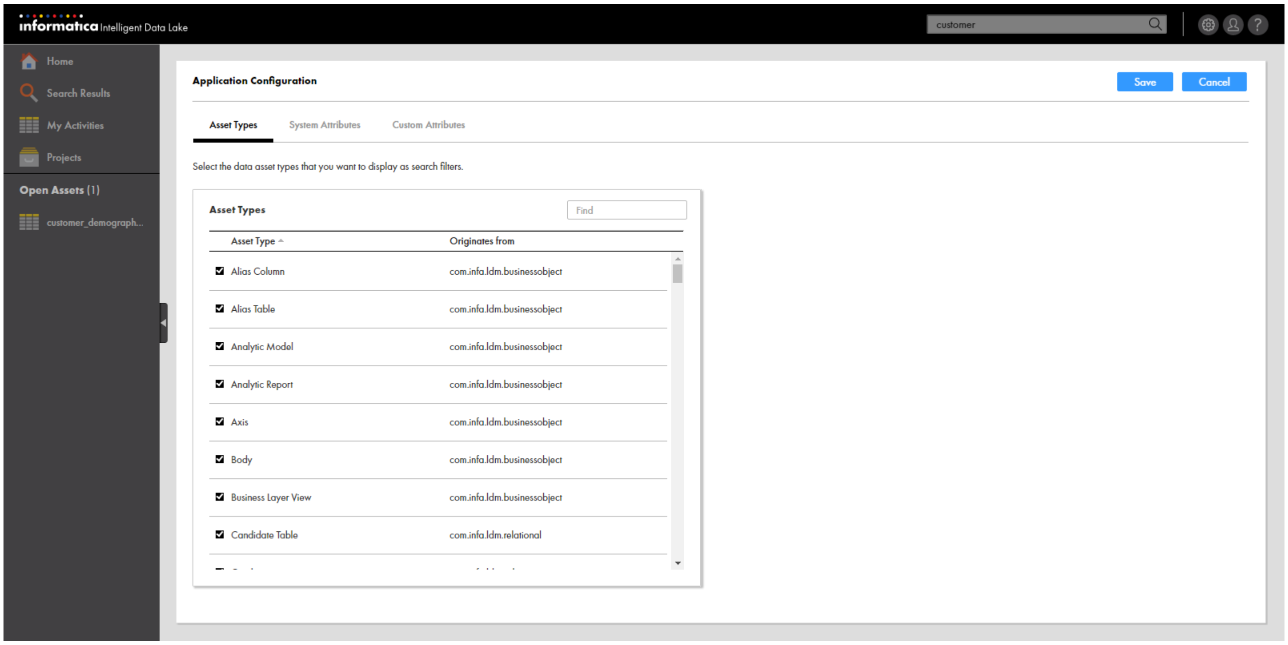
Application Configuration Overview

You can configure the filter options and set the attributes that display in the search view. The data asset types and attributes that you set determine the filters that analysts can create on search results.

Configuring Asset Types

You can configure the data asset types that display in the search filters. Use the **Application Configuration** icon to configure asset types.

The following image shows the settings that you can configure for data asset types:



Select the data asset types that you want to display as search filters.

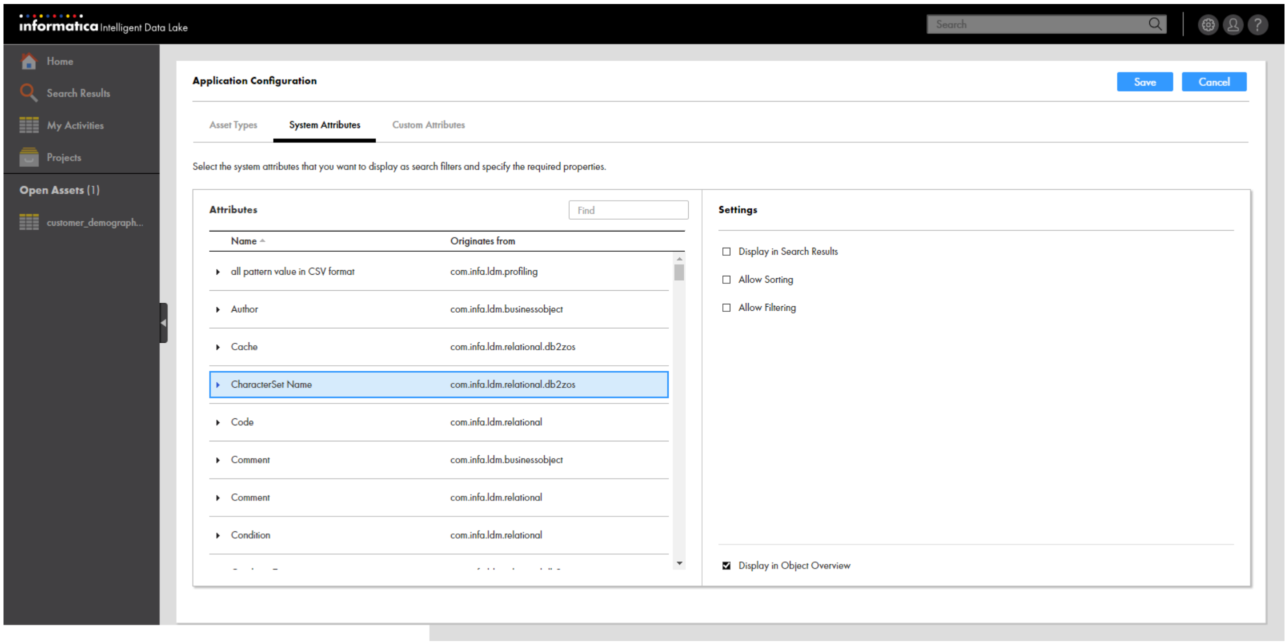
Asset Type	Originates from
☑ Alias Column	com.infa.ldm.businessobject
☑ Alias Table	com.infa.ldm.businessobject
☑ Analytic Model	com.infa.ldm.businessobject
☑ Analytic Report	com.infa.ldm.businessobject
☑ Axis	com.infa.ldm.businessobject
☑ Body	com.infa.ldm.businessobject
☑ Business Layer View	com.infa.ldm.businessobject
☑ Candidate Table	com.infa.ldm.relational

1. On the Intelligent Data Lake application header, click the **Application Configuration** icon.
2. Select the **Asset Types** view.
3. Search for an asset type or select an asset type from the list of asset types.
4. Click **Save**.

Configuring System Attributes

You can configure the system attributes that display in the search filters. Use the **Application Configuration** icon to configure system attributes.

The following image shows the settings that you can configure for **System Attributes**:



Application Configuration

Select the system attributes that you want to display as search filters and specify the required properties.

Attributes

Name	Originates from
all pattern value in CSV format	com.infa.ldm.profiling
Author	com.infa.ldm.businessobject
Cache	com.infa.ldm.relational.db2zos
CharacterSet Name	com.infa.ldm.relational.db2zos
Code	com.infa.ldm.relational
Comment	com.infa.ldm.businessobject
Comment	com.infa.ldm.relational
Condition	com.infa.ldm.relational

Settings

☐ Display in Search Results

☐ Allow Sorting

☐ Allow Filtering

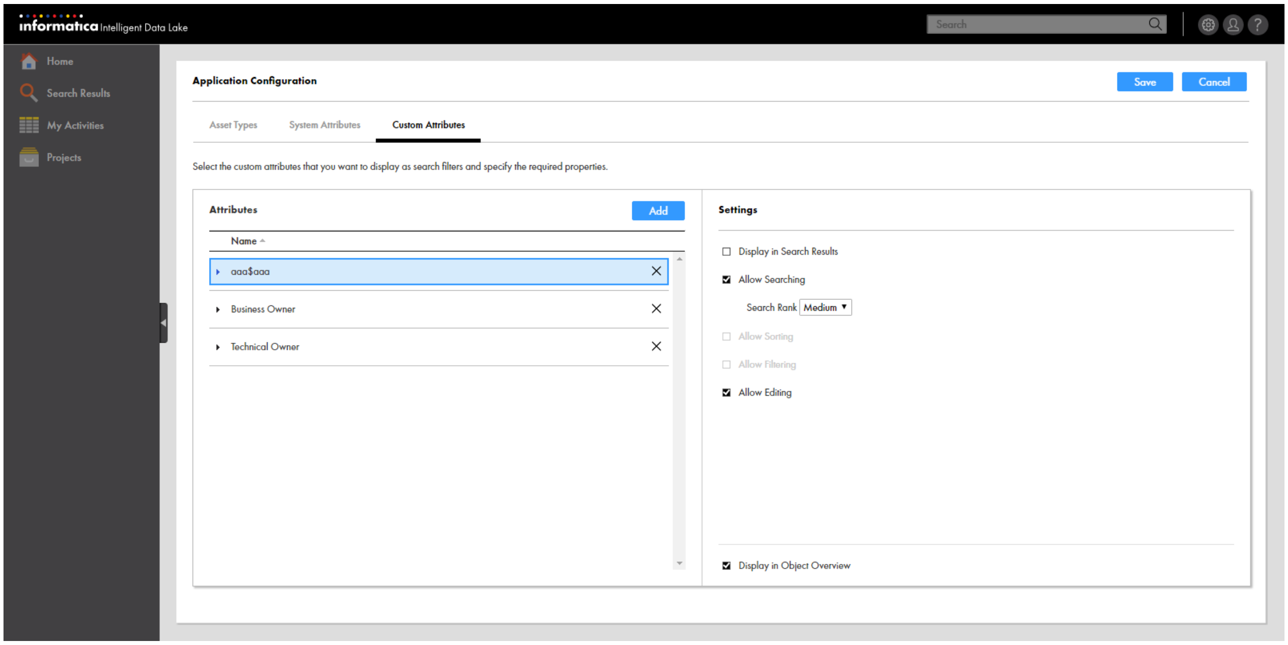
☑ Display in Object Overview

1. On the Intelligent Data Lake application header, click the **Application Configuration** icon.
2. Select the **System Attributes** view.
3. On the **Attributes** panel, choose to search for a system attribute or scroll to select a system attribute.
4. On the **Settings** panel, choose a filter setting for the system attribute that you want to display as search filters.
   - To display the system attribute in the search filter, select **Display in Search Results**.
   - To enable the system attribute to be sorted with the search results, select **Allow Sorting**.
   - To enable the system attribute to be filtered with the search results, select **Allow Filtering**.
   - To enable the system attribute to be included in the Overview view of the data asset, select **Display in Object Overview**.
5. Click **Save**.

Configuring Custom Attributes

You can configure the custom attributes that display in the search filters. Use the **Application Configuration** icon to configure custom attributes.

The following image shows the settings that you can configure for **Custom Attributes**:



Select the custom attributes that you want to display as search filters and specify the required properties.

- aaa$aaa
- Business Owner
- Technical Owner

□ Display in Search Results

☑ Allow Searching

Search Rank: Medium

□ Allow Sorting

□ Allow Filtering

☑ Allow Editing

☑ Display in Object Overview

1. On the Intelligent Data Lake application header, click the **Application Configuration** icon.
2. Select the **Custom Attributes** view.
3. On the **Attributes** panel, choose to select the custom attribute that you want to display as a search filter or add a custom attribute.
   - To select a custom attribute, select a custom attribute under the **Attributes** panel.
   - To add a custom attribute, click **Add**, select an attribute and click **OK**.
4. On the **Settings** panel, choose a filter setting for the system attribute that you want to display as search filters.
   - To display the custom attribute in the search filter, select **Display in Search Results**.
   - To enable the custom attribute to appear in the search results, select **Allow Searching**, and choose a search rank from low, medium, or high.
   - To enable the custom attribute to be sorted with the search results, select **Allow Sorting**.
   - To enable the custom attribute to be filtered with the search results, select **Allow Filtering**.
   - To enable the system attribute to be included in the Overview view of the data asset, select **Display in Object Overview**.
5. Click **Save**.

CHAPTER 7

Roles, Privileges, and Profiles

This chapter includes the following topics:

Intelligent Data Lake User Management Overview

You can assign privileges, roles, and permissions to users or groups of users to manage the level of access users and groups can have and the scope of the actions that users and groups can perform in the domain.

To access the application services and objects in the Informatica domain and to use the application clients, you must have a user account.

During Enterprise Data Catalog installation, a default administrator user account is created. Use the default administrator account to log in to the Informatica domain and manage application services, domain objects, and other user accounts.

Groups, privileges, and roles determine the tasks that a user can perform. You can add a user account to a group to create a set of users that have the same authorization. Assign the group or user accounts the privileges on the role that includes the required privileges on the Intelligent Data Lake Service.

For more information about Intelligent Data Lake user management , refer to the PDF documentation on Informatica Network. To access Informatica Network, visit https://network.informatica.com.

Intelligent Data Lake Users and Groups

A group is a collection of users and groups that can have the same privileges, roles, and permissions.

If you add a new Intelligent Data Lake user to the Informatica domain, you **must** run the user resource in Catalog Administrator. This is required to associate the user with the catalog and ensure that the latest changes to the catalog are available to the Intelligent Data Lake user. For more information, see the *Catalog Administrator Guide*.

Enterprise Data Catalog Permissions

Use the Catalog Administrator tool to set permissions for users to access data from the catalog.

Set the following permissions for the users and groups to access data assets and perform operations on the data assets.

Operation	Permissions
Search	Read access to the resource to view the data in the search results. Rest API Privilege is also required for the catalog service.
Import	Read and Write access to the relational source, Read and Write access to the Hive resource in the data lake.
Export	Read access from the Hive resource in the data lake, Read and Write access to the relational target.
HiveScanner	Read and Write access to the Hive resource in the data lake.

For more information on giving permissions to users and groups, see the *Catalog Administrator Guide*.

Intelligent Data Lake Service Privileges

The Intelligent Data Lake Service privileges determine actions that users can perform using the Intelligent Data Lake application.

The following table lists the required permissions and the actions that users can perform with the privilege in the Data Discovery and Preparation privilege group:

Privilege Name	Permission On	Description
Data Discovery and Preparation > Access Intelligent Data Lake	Intelligent Data Lake Service	User is able to perform the following actions: - Log in to the Intelligent Data Lake application. - Search data assets. - Upload data assets. - Prepare data assets. - Publish data assets. **Note:** You can access the Application Configuration details only if you are logged in as an Administrator.

Intelligent Data Lake Administrator

The Intelligent Data Lake administrator has full permissions and privileges for managing the Intelligent Data Lake application.

Follow these steps to create an Intelligent Data Lake administrator:

1. In the Administrator tool, click the **Security** tab.

2. In the Navigator, select a user or group.
3. Click the **Privileges** tab.
4. Click Edit. The **Edit Roles and Privileges** dialog box appears.
5. On the **Roles** tab, expand the Intelligent Data Lake service. The list of roles available for the Intelligent Data Lake service appears.
6. Select the Administrator role.
7. Click **OK.**
8. Assign permissions for the Hive and HDFS connections created for the Intelligent Data Lake Service.

For more information about creating users, assigning permissions, and security settings, see the *Informatica Application Security Guide*.

Operating System Profiles

An operating system profile is a type of security that the Data Integration Service uses to run mappings, workflows, and profiling jobs. Use operating system profiles to increase security and to isolate the run-time environment for users.

If the Data Integration Service runs on UNIX or Linux, create operating system profiles and configure the Data Integration Service to use operating system profiles. By default, the Data Integration Service process runs all jobs, mappings, and workflows using the permissions of the operating system user that starts Informatica services. For more information about Data Integration Service operating system profiles, see the *Informatica Application Service Guide.*

When you create the Intelligent Data Lake Service, you must associate it with a Data Integration Service. The operating system profile configuration for the associated Data Integration Service will be applicable to the Intelligent Data Lake Service.

Note: If you plan to use the operating system profiles option for the Data Integration Service, ensure that you create and associate a different Data Integration Service for Enterprise Data Catalog and Intelligent Data Lake. Enterprise Data Catalog does not support operating system profiles.

Creating an Operating System Profile for Intelligent Data Lake

Create an operating system profile to provide a level of security to users and groups in the run-time environment. The Data Integration Service you have associated with the Intelligent Data Lake Service uses the operating system profile of the user to run workflows or jobs. Ensure that the assigned license allows you to use the operating system profiles feature.

1. In the Administrator tool, click the Security tab.
2. On the Security Actions menu, click **Create Operating System Profile**.

   The **Create Operating System Profile** dialog box appears.

3. Enter the following general properties for the operating system profile:

Property	Description
Name	Name of the operating system profile. The name is not case sensitive and must be unique within the domain. It cannot exceed 128 characters or begin with @. It also cannot contain the following special characters: % * + \ / . ? < > The name can contain an ASCII space character except for the first and last character. All other space characters are not allowed.
System User Name	Name of an operating system user that exists on the machines where the Data Integration Service runs. The Data Integration Service runs workflows or jobs using the system access of the system user defined for the operating system profile. **Note:** When you create operating system profiles, you cannot specify the system user name as root or use a non-root user with uid==0.

4. Click **Next**.

   The **Configure Operating System Profile** dialog box appears.
5. Select the Data Integration Service checkbox and configure the operating system profile properties.
6. Configure service process variables in the operating system profile to specify different output file locations based on the operating system profile that is assigned to the user or group. The Data Integration Service writes output files to a single shared location specified in the $DISRootDir service process variable.
7. Select **Enable Hadoop Impersonation Properties**.
8. Choose to use the logged in user or specify a Hadoop impersonation user to run Hadoop jobs. For a secure Hadoop cluster, the logged in user or the user specified as the Hadoop impersonation user should be valid and have the required permissions in the cluster nodes.
9. Optionally, configure the environment variables.
10. Click **Next**.

    The **Assign Groups and Users to Operating System Profile** dialog box appears.
11. In the Groups tab, select the groups that you want to assign the operating system profile.

    A list of all the groups with permission on the operating system profile appears.
12. In the Users tab, select the users that you want to assign the operating system profile.

    A list of all the users with permission on the operating system profile appears.
13. Click **Finish**.

    After you create the operating system profile, the details panel displays the properties of the operating system profile and the groups and users that the profile is assigned to.

Assigning a Default Operating System Profile to an Intelligent Data Lake User or Group

For Intelligent Data Lake, each user must have only one default operating system profile. If a user inherits multiple operating system profiles based on permissions assigned to groups or direct assignment, ensure that the user is assigned only one default operating system profile.

When you assign an operating system profile as the default profile to a user or group, the Data Integration Service uses the default operating system profile to run jobs and workflows. You can assign only an operating system profile with direct permission as the default profile to a user or group.

1. On the Security tab, select the **Users** or **Groups** view.
2. In the Navigator, select the user or group.
3. In the content panel, select the **Permissions** view.
4. Click the **Operating System Profiles** tab.
5. Click the **Assign or Change the Default Operating System Profile** button.

   The **Assign or Change the Default Operating System Profile** dialog box appears.
6. Select a profile from the **Default Operating System Profile** list.
7. Click **OK**.

   In the details panel, the **Default Profile** column displays **Yes (Direct)** for the operating system profile.

CHAPTER 8

Data Asset Access and Publication Management

This chapter includes the following topics:

Data Asset Access Overview

Intelligent Data Lake user accounts must have access to the Hive tables in the data lake when they preview data, upload data, and publish prepared data.

The administrator must follow the standards and best practices of their organization for providing data access to their users. The administrator must follow the established policies and procedures to provide the right level of access to the data asset a user requires for a project.

For more information about data asset access and publication management, refer to the PDF documentation on Informatica Network. To access Informatica Network, visit https://network.informatica.com.

Providing and Managing Access to Data

Intelligent Data Lake user accounts must be authorized to access the Hive tables in the Hadoop cluster designated as the data lake. Intelligent Data Lake user accounts access Hive tables in the Hadoop cluster when they preview data, upload data, and publish prepared data.

HDFS permissions

Grant each user account the appropriate HDFS permissions in the Hadoop cluster. HDFS permissions determine what a user can do to files and directories stored in HDFS. To access a file or directory, a user must have permission or belong to a group that has permission to the file or directory.

A Hive database corresponds to a directory in HDFS. Each Hive table created in the database corresponds to a subdirectory. You grant Intelligent Data Lake user accounts permission on the appropriate directory, based on whether you want to provide permission on a Hive database or on a specific Hive table in the database.

Note: As a best practice, you can set up private/shared/public databases (or schemas) in a single Hive resource and grant users appropriate permissions on those corresponding HDFS directories.

Data asset preview, import, and export

To access and preview non-lake data assets, grant users permission to connections including Oracle and Teradata Import.

For more information about Sqoop connectivity through JDBC for external databases, see the chapters on Sqoop sources and targets in the *Informatica Big Data Management User Guide*.

Set the custom flag `DoNotUseOwnerNameForSqoop` to true if the database does not require the owner name while connecting through Sqoop. For more information, see the *Informatica Big Data Management User Guide*.

User impersonation

User impersonation allows different user accounts to run mappings in a Hadoop cluster that uses Kerberos authentication. When users upload and publish prepared data in the Intelligent Data Lake application, the Data Integration Service runs mappings in the Hadoop environment. The Data Integration Service pushes the processing to nodes in the Hadoop cluster. The Data Integration Service uses the credentials you have specified to impersonate the user accounts that publish and upload the data. Create a user account in the Hadoop cluster for each Intelligent Data Lake user account.

When the Data Integration Service impersonates a user account to submit a mapping, the mapping can only access Hadoop resources that the impersonated user has permissions on. Without user impersonation, the Data Integration Service uses its credentials to submit a mapping to the Hadoop cluster. Restricted Hadoop resources might be accessible.

Data Asset Publication

Data publication is the process of making prepared data available in the data lake.

When you publish prepared data, Intelligent Data Lake applies the recipe to the data in the input source. Intelligent Data Lake writes the transformed input source to a Hive table in the data lake. These mappings are stored in the Model repository and can be accessed using the Developer tool.

Use the Developer tool to modify the Informatica mappings created when you published a project in the Intelligent Data Lake application. Use the Administrator tool to run and monitor the mappings. For more information, see the *Informatica Big Data Management User Guide* and the *Informatica Administrator Guide*.

You can use a third-party business intelligence or advanced analytic tool to run reports to further analyze the published data. Other analysts can add the published data to their projects and create new data assets.

Operationalizing the Loading of Data Assets

When analysts publish data in the Intelligent Data Lake application, the Intelligent Data Lake Service converts the preparation recipes into Informatica mappings. You can operationalize the mappings to regularly load data with the new structure into the data lake.

Use the Developer tool to view and edit the converted mappings. The mappings use the same name as the Intelligent Data Lake worksheet with the published data. Verify that the converted mappings meet your requirements, and then deploy the mappings.

Use the Administrator tool or the infacmd command line program to run the mappings to load and transform the data. The Data Integration Service writes the data to Hive tables in the data lake. You can schedule the deployed mappings to regularly load data into the data lake. Use the Administrator tool to monitor each mapping run.

For more information about developing and running mappings that write to Hive, see the *Informatica Big Data Management User Guide*.

Enabling Hive Statistics

The Intelligent Data Lake application displays upload and publication statistics in the My Activity page using the Hive statistics stored in the database metastore. Hive is configured to use Derby as the default metastore and the statistics information is stored in the Derby metastore. However, Derby cannot be used for concurrent execution in Hive. To display the Intelligent Data Lake upload and publishing statistics, you must enable Hive statistics.

To change the default metastore to MySQL and enable Hive statistics:

1. Create a mysql table using the command: CREATE TABLE <db>.PARTITION_STATS_V2 (TS TIMESTAMP DEFAULT CURRENT_TIMESTAMP, ID VARCHAR(255) PRIMARY KEY , ROW_COUNT BIGINT , RAW_DATA_SIZE BIGINT);
2. On the machine where the Data Integration Service is configured, edit the hive.xml file as follows:

```
<property>
<name>hive.stats.dbclass</name>
<value>jdbc:mysql</value>
<description>The default database that stores temporary hive statistics.</
description>
</property>
<property>
<name>hive.stats.autogather</name>
<value>true</value>
<description>A flag to gather statistics automatically during the INSERT OVERWRITE
command.</description>
</property>
<property>
<name>hive.stats.jdbcdriver</name>
<value>com.mysql.jdbc.Driver</value>
<description>The JDBC driver for the database that stores temporary hive
statistics.</description>
</property>
<property>
<name>hive.stats.dbconnectionstring</name>
<value>jdbc:mysql://<host>:3306/<db>?
useUnicode=true&characterEncoding=UTF-8&user=<user>&password=<pwd></value>
<description>The default connection string for the database that stores temporary
hive statistics.</description>
</property>
```

3. Copy the mysql-connector.jar file to the following location: `INFA_HOME/services/shared/hadoop/cloudera_cdh5u4/lib`.
4. In the `INFA_HOME/services/shared/hadoop/cloudera_cdh5u4/infaConf/hadoopEnv.properties` file, add or update the following line: `infapdo.aux.jars.path=file://$DIS_HADOOP_DIST/lib/mysql-connector-java-5.1.23.jar`.

CHAPTER 9

Monitoring Intelligent Data Lake

This chapter includes the following topics:

Monitoring Intelligent Data Lake Overview

Monitoring Intelligent Data Lake involves auditing the log that stores the events and their details.

Intelligent Data Lake saves the details of all events performed in the HBase database. You can find answers to questions like the following by running a query on the audit database:

- Who are the top 10 most active users (in last week, last month, last year)?
- What are the top 10/50.100 recently published data assets (new data assets and updated data assets)?
- What are the largest data assets published and what are their sizes?
- What are the largest data assets uploaded and what are their sizes?
- What are the largest data assets exported and what are their sizes?
- Who exported data assets and when?
- Who created/updated/deleted/shared/viewed which project and when?
- Who added which data asset to which project and when?
- Who created/updated which recipe for a worksheet and when?

Events Logged In The Audit Database

Based on the type of activity, you can find the details of the following events saved in the audit database. You can use the event name and run a query to get the required details.

Application Access Events

LOGIN

LOGOUT

TIMEOUT

Project Management Events

CREATE

VIEW

SHARE

CHANGEOWNER

UPDATENAME

UPDATEDESC

DELETE

ADDWORKSHEET

Data Asset Related Events

UPLOAD

ADDDATASET

PUBLICATION

EXPORT

DOWNLOAD

INGESTION

COPY

DELETEDATASET

Data Preparation Related Events

PREPINEDITMODE

PREPINVIEWMODE

UPDATERECIPE

RENAMEWORKSHEET

REFRESH

CHAPTER 10

Backing Up and Restoring Intelligent Data Lake

This chapter includes the following topics:

Back Up and Restore Overview

Informatica recommends that you periodically back up the service metatdata, repositories, and storage locations associated with Intelligent Data Lake. Backing up Intelligent Data Lake enables you to recover from data loss due to hardware or software failures. You should also back up Intelligent Data Lake before you upgrade Intelligent Data Lake.

Back up the following application services, repositories, and storage locations associated with Intelligent Data Lake:

- Back up the Model repository associated with the Intelligent Data Lake Service.
- Document the configuration details for the Hive scanners and HDFS scanners configured in the Enterprise Data Catalog used by the Intelligent Data Lake Service.
- Back up the MySQL database or Oracle database used as the Data Preparation Service repository.
- Back up the Data Preparation Service durable storage location on HDFS.

You might also need to back up the following:

- If rules are configured for the Data Preparation Service, back up the rules metadata.
- If the Intelligent Data Lake Service logs user activity events, back up the HBase table that contains user activity log event metadata.

Back Up Intelligent Data Lake

Back up the application service metatdata, repositories, and storage locations associated with Intelligent Data Lake.

You do not need to complete the back up procedures in sequence. However, you should disable the Intelligent Data Lake Service and the associated application services before performing a back up procedure.

Step 1. Disable the Application Services

Disable the Intelligent Data Lake Service and all associated application services.

You use Informatica Administrator (the Administrator tool) to disable the application services. You can also use the Administrator tool to identify the application services used by the Intelligent Data Lake Service.

Disable the application services in the following order:

- Intelligent Data Lake Service
- Data Preparation Service
- Content Management Service
- Data Integration Service
- Model Repository Service

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. Click the **Disable the Service** button, and then disable the application service.
4. In the **Properties** view, click each application service associated with the Intelligent Data Lake Service, and then disable the service.

Step 2. Back Up the Model Repository Content

Back up the content of the Model repository associated with the Intelligent Data Lake Service. The Model repository stores project definitions and transient publication mappings associated with projects created by analysts in Intelligent Data Lake.

Use the Administrator tool to identify the Model Repository Service associated with the Intelligent Data Lake Service. You also use the Administrator tool to back up the Model repository. When you back up a Model repository, the Model Repository Service writes the repository contents to a backup file in the service backup directory. The service backup directory is a subdirectory of the node backup directory.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. In the **Properties** view, expand **Model Repository Service Options**.
4. Select the Model Repository Service.
5. On the **Manage** tab **Actions** menu, select **Repository Contents** > **Back Up**.

   The **Back Up Repository Contents** dialog box appears.
6. Enter the options required to back up the Model Repository Service.

The following table describes the options required to back up the Model Repository Service:

Option	Description
Username	User name of any user in the domain.
Password	Password of the domain user.
SecurityDomain	Domain to which the domain user belongs. Default is native.
Output File Name	Name of the backup file.
Description	Description of the contents of the backup file.

7. Click **Overwrite** to overwrite a file with the same name.
8. Click **OK**.

   The Model Repository Service writes the backup file to the service backup directory.

Step 3. Document the Configuration Properties for the Scanner Resources

Document the configuration details for each Hive scanner resource and each HDFS scanner resource configured in the Enterprise Data Catalog used by the Intelligent Data Lake Service.

Use the Administrator tool to identify the Catalog Service associated with the Intelligent Data Lake Service. Use the Catalog Administrator to find the configuration properties for each scanner resource.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. In the **Properties** view, expand **Catalog Service Options**.
4. Select the Catalog Service.
5. Click the Catalog Administrator URL.

   The URL has the following format:

   ```
   <Catalog Service host name>:<Catalog Service port>/ldmadmin
   ```
6. In Catalog Administrator, click **Open**.

   The Library view opens.
7. Select a scanner resource.
8. Document the properties shown on each tab in Catalog Administrator.

Step 4. Back Up the Intelligent Data Lake Connections

Back up the connection objects for the HDFS, Hive, Hadoop, and HBase connections used by the Intelligent Data Lake Service.

Use the infacmd isp ExportDomainObjects command to export the connection object for each connection to an .xml file.

You can create an export control file to filter the objects to export from the domain. An export control file is an .xml file that you use with the infacmd isp ExportDomainObjects command. Specify the name of each connection object you want to export in an `<objectList>` element within the export control file.

The following example shows an export control file that filters the connection objects for the HDFS, Hive, Hadoop, and HBase connections used by the Intelligent Data Lake Service:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<exportParams xmlns="http://www.informatica.com/oie/exportControl/9">
  <objectList type="connection" >
    <object name="HDFS_CCO" />
  </objectList>
  <objectList type="connection" >
    <object name="HIVE_CCO" />
  </objectList>
  <objectList type="connection" >
    <object name="HADOOP_CCO" />
  </objectList>
  <objectList type="connection" >
    <object name="HBASE_CCO" />
  </objectList>
</exportParams>
```

If you use an export control file, use the Administrator tool to find the name of each connection used by the Intelligent Data Lake Service. If you do not use an export control file, the infacmd isp ExportDomainObjects command exports all native users, native groups, roles, connections, and cluster configurations from the Informatica domain.

For more information about creating export control files and using the infacmd isp ExportDomainObjects command, see the *Informatica Command Reference*.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. In the **Properties** view, expand **Data Lake Options**.
4. Copy the values for the **HDFS Connection**, **Hive Connection** and **Hadoop Connection** properties.
5. Expand **Event Logging Options**.
6. Copy the value for the **HBase Connection** property.
7. Create an export control file to filter the objects that the infacmd isp ExportDomainObjects command exports.

   Specify the name of each connection object you want to export in an `<objectList>` element within the file.
8. Run the following command:

   ```
   infacmd isp ExportDomainObjects -dn <domain name> -un <user name> -pd <password>
   -sdn <security domain> -fp <export file> -cp <export control file> -rp <true|false>
   ```

   The following table describes the infacmd isp ExportDomainObjects command options and arguments you might specify to export connection objects for the HDFS, Hive, Hadoop, and HBase connections:

Option	Argument	Description
-DomainName -dn	domain_name	Required. Name of the Informatica domain.
-UserName -un	user_name	Required if the domain uses Native or LDAP authentication. User name to connect to the domain. Optional if the domain uses Kerberos authentication. To run the command with single sign-on, do not set the user name. If you set the user name, the command runs without single sign-on.

Option	Argument	Description
-Password -pd	password	Required if you specify the user name.
-SecurityDomain -sdn	security_domain	Required if the domain uses LDAP authentication. Optional if the domain uses native authentication or Kerberos authentication. Name of the security domain to which the domain user belongs. If the domain uses native or LDAP authentication, the default is Native. If the domain uses Kerberos authentication, the default is the LDAP security domain created during installation.
-ExportFile -fp	export_file_name	Required. Path and file name of the export file. If you do not specify the file path, infacmd creates the file in the directory where you run infacmd.
-ExportControlFile -cp	export_control_file	Optional. Name and path of the export control file that filters the objects to export.
-RetainPassword -rp	true|false	Optional. Set to true to retain encrypted passwords for users and connections in the exported file. When set to false, user and connection passwords are exported as empty strings. Default is false.

The following example exports the connection objects for the HDFS, Hive, Hadoop, and HBase connections defined in the specified export control file:

```
infacmd isp ExportDomainObjects -dn InfaDomain -un Administrator -pd password -fp
"/backup/idl_connections.xml" -cp "/backup/idl_exportControl.xml" -rp true
```

Step 5. Back Up the Data Preparation Service Repository

Back up the MySQL database or the Oracle database used as the Data Preparation Service repository.

Use the Administrator tool to find the database type and schema name values configured for the repository database. Follow the instructions in the MySQL documentation or the Oracle documentation to back up the repository database.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Data Preparation Service.
3. In the **Properties** view, expand **Data Preparation Repository Options**.
4. Find the values for the **Database Type** and **Schema Names** properties configured for the repository database.
5. Back up the Data Preparation Service repository database.

   For more information, see the MySQL documentation or the Oracle documentation.

Step 6. Back Up the Data Preparation Service Durable Storage

Back up the HDFS location where the Data Preparation Service stores data.

Use the Administrator tool to find the durable storage configuration properties. You can use the Hadoop DistCp tool to export the storage metadata to a backup location.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Data Preparation Service.

3. In the **Properties** view, expand **Data Preparation Storage Options**, and then document the durable storage configuration properties.
4. On the HDFS cluster, run the following command:

   ```
   hadoop distcp hdfs://<durable storage location> <local file system location>
   ```

   The following table describes the DistCp command arguments:

Argument	Description
Durable Storage Location	The HDFS storage location to back up. Specify the value of the HDFS Storage Location property displayed in the Administrator tool.
Backup Location	The path to the backup directory to which to write the storage metadata.

Step 7. Back Up the User Activity Event Auditing Data

If the Intelligent Data Lake Service logs user activity events for auditing, you must back up the HBase table that the contains user activity event metadata. The Intelligent Data Lake Service logs user activity events if the **Log User Activity Events** property under **Event Logging Options** is set to true in the Administrator tool.

You can use the HBase Driver export command to back up the HBase table that the contains user activity event metadata and write it to a backup location.

On the HDFS cluster, run the following command:

```
org.apache.hadoop.hbase.mapreduce.Driver export  <HBase table name> <output directory>
```

The following table describes the command arguments:

Argument	Description
HBase Table Name	The HBase table that the contains user activity event metadata. The default name of the HBase table is USER_ACTIVITIES.
Output Directory	The path to the backup directory to which the metadata is written.

Step 8. Back Up the Rules Metadata

If rules are configured for the Data Preparation Service, back up the rules metadata stored in the Model repository used by the Data Preparation Service.

The rules metadata is stored in projects in the Model repository used by the Data Preparation Service.

Use the Administrator tool to find the Model Repository Service used by the Data Preparation Service. Use the infacmd oie ExportObjects command to export the rule objects, the dictionary content, and the NER and Classifier model files from a project in the Model repository.

The infacmd oie ExportObjects command creates an .xml file that contains the rule objects. The command also creates a .zip file containing the dictionary content as .dic extension flat files and the NER and Classifier models as .ner and .classifer files.

Run the infacmd oie ExportObjects command to export rules metadata from each project that contains rules metadata. When you run the command, you specify the name of a project that contains the rule objects. You also specify the name of the Model Repository Service that manages the Model repository that contains the project. The Model Repository Service must be running when you run the command.

You can create an export control file to filter the objects to export from the Model repository. An export control file is an .xml file that you use with the infacmd oie ExportObjects command. Specify the name of each rule object you want to export in an `<objectList>` element within the export control file.

The following example shows an export control file that filters the rule objects to export from the Model repository:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<exportParams xmlns="http://www.informatica.com/oie/exportControl/9">
  <folders>
    <folder recursive="false" select="all">
      <objectList type="Rule" >
        <object name="validate_address" />
      </objectList>
      <objectList type="Rule" >
        <object name="validate_USA_zip_code" />
      </objectList>
      <objectList type="Rule" >
        <object name="validate_Canadian_postal_code" />
      </objectList>
    </folder>
  </folders>
</exportParams>
```

If you do not use an export control file, the infacmd oie ExportObjects command exports all objects in the specified project.

For more information about creating export control files and using the infacmd oie ExportObjects command, see the *Informatica Command Reference*.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. In the **Properties** view, expand **Data Preparation Service Properties**.
4. Select the Data Preparation Service shown.
5. In the **Properties** view, expand **Model Repository Service Properties** to identify the Model Repository Service used by the Data Preparation Service.

   You must identify each project within the Model repository that contains rules metadata.
6. If the Model Repository Service is not running, recycle the Model Repository Service.

   For more information, see "Step 7. Recycle the Application Services" on page 87.
7. Run the following command:

   ```
   infacmd oie ExportObjects -dn <domain name> -un <user name> -pd <password>
   -sdn <security domain> -pn <project name> -rs <Model Repository Service name>
   -fp <export file> -ow <true|false> -cp <export control file> -oo <options>
   ```

   The following table describes infacmd oie ExportObjects options and arguments you might specify to back up the Content Management Service metadata:

Option	Argument	Description
-DomainName -dn	domain_name	Name of the Informatica domain.
-UserName -un	user_name	User name to connect to the domain. Optional if the domain uses Kerberos authentication.

Option	Argument	Description
-Password -pd	password	Password for the user name.
-SecurityDomain -sdn	security_domain	Name of the security domain to which the domain user belongs.
-ProjectName -pn	project_name	Name of the project in the Model repository from which to export the rules metadata.
-RepositoryService -rs	service_name	Model Repository Service name.
-ExportFilePath -fp	export_file_path	Path and XML file name of the export file that contains the rules metadata. You can specify an absolute path or a relative path to the file name.
-OverwriteExportFile -ow	true|false	Set to true to overwrite an existing export file. Default is false.
-ControlFilePath -cp	control_file_path	Path and file name of the export control file that filters the exported objects.
-OtherOptions -oo	options	Additional options to export the dictionary content and the NER and Classifier models to a zip file. Enter options using the following format: `rtm:<option_name>=<value>,<option_name>=<value>` Option names include: - codePage. Code page of the reference data. - refDataFile. Path and file name of the zip file where you want to export the reference table data.

The following example exports the rules objects defined in the specified export control file from a project named TEST_PROJECT to the dq_rules_metadata.xml file. The example also exports the dictionary content and the NER and Classifier models from the project named TEST_PROJECT to a dq_rules_artifacts.zip file.

```
infacmd oie exportObjects -dn InfaDomain -un Administrator -pd password -sdn native
-pn TEST_PROJECT -rs MRS_dev2 -fp "/export/dq_rules_metadata.xml" -ow true
-cp "/dq_rules_export.xml" -oo "rtm:codePage=UTF-8,refDataFile="/output/
dq_rules_artifacts.zip""
```

Restore Intelligent Data Lake

Restore the application service metatdata, repositories, and storage locations associated with Intelligent Data Lake.

You do not need to complete the restore procedures in sequence. However, you should disable the Intelligent Data Lake Service and the associated application services before performing a restore procedure.

Step 1. Restore the Model Repository Content

Restore the Model repository content from the repository backup file.

Verify that the Model repository is empty. If the Model repository contains content, the restore option is disabled. You must delete the content from the Model repository before restoring the repository content from the backup file.

1. In the Administrator tool, select the select the **Services and Nodes** view.
2. In the Domain Navigator, select the Model Repository Service associated with the Intelligent Data Lake Service.
3. If the Model repository contains content, delete the content. From the **Actions** menu, select **Repository Contents** > **Delete**.
4. From the **Actions** menu, select **Repository Contents** > **Restore**.

   The **Restore Repository Contents** dialog box appears.
5. Select the Model repository backup file to restore.
6. Enter the properties required to restore the Model repository.

   The following table describes the properties:

Property	Description
Username	User name of any user in the domain.
Password	Password of the domain user.
Security Domain	Security domain to which the domain user belongs. Default is native.

7. Click **OK**.

Step 2. Create the Hive and HDFS Resources in the Catalog

Create the Hive and HDFS scanner resources in the catalog used by the Intelligent Data Lake Service.

For each scanner resource, enter the property values you copied during the backup phase into each tab in the Resource view in Catalog Administrator.

For more information about creating resources, see the *Informatica Catalog Administrator Guide*.

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator, select the Intelligent Data Lake Service.
3. In the **Properties** view, expand **Catalog Service Options**.
4. Select the Catalog Service.
5. Click the Catalog Administrator URL.

   The URL has the following format:

   ```
   <Catalog Service host name>:<Catalog Service port>/ldmadmin
   ```
6. In Catalog Administrator, click **Resource**, and then select **New > Resource**.

   The create resource view opens.
7. Create each Hive and HDFS scanner resource.

   Enter the property values you copied on each tab in Catalog Administrator.

Step 3. Restore the Intelligent Data Lake Connections

Import the connection objects for the HDFS, Hive, Hadoop, and HBase connections used by the Intelligent Data Lake Service from the .xml file to which you exported the connection objects.

Use the infacmd isp ImportDomainObjects command to import the connection object for each connection from the .xml file.

For more information about using the infacmd isp ImportDomainObjects command, see the *Informatica Command Reference*.

Run the following command:

```
infacmd isp ImportDomainObjects -dn <domain name> -un <user name> -pd <password>
-sdn <security domain> -fp <import file>
```

The following table describes the infacmd isp ImportDomainObjects command options and arguments you might specify to import the connection objects for the HDFS, Hive, Hadoop, and HBase connections:

Option	Argument	Description
-DomainName -dn	domain_name	Required. Name of the Informatica domain.
-UserName -un	user_name	Required if the domain uses Native or LDAP authentication. User name to connect to the domain. Optional if the domain uses Kerberos authentication. To run the command with single sign-on, do not set the user name. If you set the user name, the command runs without single sign-on.
-Password -pd	password	Required if you specify the user name.
-SecurityDomain -sdn	security_domain	Required if the domain uses LDAP authentication. Optional if the domain uses native authentication or Kerberos authentication. Name of the security domain to which the domain user belongs. If the domain uses native or LDAP authentication, the default is Native. If the domain uses Kerberos authentication, the default is the LDAP security domain created during installation.
-ImportFilePath -fp	import_file_path	Required. Path and file name of the XML file from which you import the objects.

The following example imports the connection objects for the HDFS, Hive, Hadoop, and HBase connections from the specified .xml file:

```
infacmd isp ImportDomainObjects -dn InfaDomain -un Administrator -pd password
-fp "/backup/edl_connections.xml"
```

Step 4. Restore the Data Preparation Service Repository

Restore the Data Preparation Service repository, and then use the Administrator tool to configure the Data Preparation Service to use the repository schema.

1. Restore the MySQL database or Oracle database used as the Data Preparation Service repository.

   For more information, see the MySQL documentation or the Oracle documentation.
2. In the Administrator tool, select the **Services and Nodes** view.
3. In the Domain Navigator, select the Data Preparation Service.
4. In the **Properties** view, expand **Data Preparation Repository Options**.
5. Click the **Edit** icon.
6. Enter the properties required to connect to the Data Preparation Service repository.

   The following table describes the Data Preparation Service repository properties:

Property	Description
Database Type	Type of database to use for the Data Preparation repository.
Host Name	Host name of the machine that hosts the Data Preparation repository database. This property appears only if your Database Type is MySQL.
Database Port Number	Port number for the database. This property appears only if your Database Type is MySQL.
Connection String	Connection used to access the Oracle database. Use the following connection string: `jdbc:informatica:oracle://<host name>:<port>;ServiceName=<database name>` This field appears if you select Oracle as your Database Type.
Secure JDBC Parameters	If the database is secured, information such as TrustStore and TrustStorePassword can be included in this field. It is saved in an encrypted format. Parameters usually configured include the following: *EncryptionMethod=<encryption method>;HostNameInCertificate=<hostname>;TrustStore=<truststore file with its location>;TrustStorePassword=<truststore password>;KeyStore==<keystore file with its location>;KeyStorePassword=<keystore password>;ValidateServerCertificate=<true /false>* This field appears if you select Oracle as your Database Type.
Database User Name	Database user account to use to connect to the Data Preparation repository.
Database User Password	Password for the Data Preparation repository database user account.
Modify Database User Password	Select this checkbox to modify the database user password.
Schema Name	Schema or database name of the Data Preparation repository database. This field appears if you select MySQL as your Database Type.

Step 5. Restore the Data Preparation Service Durable Storage

Import the Data Preparation Service durable storage metadata into a durable storage location in the HDFS. After you import the durable storage metadata, use the Administrator tool to configure the Data Preparation Service to use the durable storage location.

You can use the Hadoop DistCp tool to import the storage metadata from the backup location to the HDFS location.

1. On HDFS, run the following command:

   ```
   hadoop distcp <backup location> hdfs://<durable storage location>
   ```

   The following table describes the DistCp command arguments:

Argument	Description
Durable Storage Location	The HDFS storage location to which to copy the durable storage metadata.
Backup Location	The path to the backup directory from which to copy the durable storage metadata.

2. In the Administrator tool, select the **Services and Nodes** view.
3. In the Domain Navigator, select the Data Preparation Service.
4. In the **Properties** view, expand **Data Preparation Storage Options**, and then click the **Edit** icon.
5. Enter the following properties:

Property	Description
Local Storage Location	Directory for data preparation file storage on the node on which the Data Preparation Service runs.
HDFS Connection	Connection to the durable storage location in HDFS.
HDFS Storage Location	Durable storage location in HDFS.
Hadoop Distribution Directory	Location of the Hadoop distribution on HDFS data nodes.

Step 6. Restore the User Activity Event Auditing Data

Import the user activity event metadata from the backup location into an HBase table, and then set the HBase table namespace in the Intelligent Data Lake Service properties in the Administrator tool.

You must create the HBase table before you import the user activity event metadata from the backup location. The default name of the HBase table is USER_ACTIVITIES.

You can use the HBase Driver import command to import the user activity event metadata from the backup location to the HBase table.

1. Run the following command:

   ```
   org.apache.hadoop.hbase.mapreduce.Driver import  <backup location> <HBase table name>
   ```

The following table describes the command arguments:

Argument	Description
Backup Location	Path to the backup directory containing the user activity event metadata.
HBase Table Name	The HBase table to which to write the user activity event metadata. The default table name is USER_ACTIVITIES.

2. In the Administrator tool, select the Intelligent Data Lake Service, and then select the **Properties** view.
3. In the **Properties** view, expand **Event Logging Options**, and then click the **Edit** icon.
4. Enter the following properties:

Property	Description
HBase Connection	Connection for the HBase database.
HBase Namespace	Namespace for the HBase table.

Step 7. Recycle the Application Services

Start the application services used by Intelligent Data Lake.

Start the application services in the following order:

- Model Repository Service
- Data Integration Service
- Content Management Service
- Data Preparation Service
- Intelligent Data Lake Service

1. In the Administrator tool, select the **Services and Nodes** view.
2. In the Domain Navigator of the Administrator tool, select an application service.
3. Click the **Recycle the Service** button.
4. Click **OK**.

Step 8. Restore the Rules Metadata

If rules are configured for the Data Preparation Service, restore the rules metadata to the Model repository used by the Data Preparation Service.

Use the infacmd oie ImportObjects command to import the rules metadata, the dictionary content, and the NER and Classifier model files into a project in the Model repository associated with the Intelligent Data Lake Service.

The Model Repository Service that manages the Model repository used by the Data Preparation Service must be running when you run the command.

For more information about using the infacmd oie ImportObjects command, see the *Informatica Command Reference*.

Run the following command:

```
 infacmd oie ImportObjects -dn <domain name> -un <user name> -pd <password>
-sdn <security domain> -tn <target project name> -rs <Model Repository Service name>
-fp <import file> -sp <source project name> -tf <target folder name> -oo <options>
```

The following table describes the infacmd oie ImportObjects options and arguments:

Option	Argument	Description
-DomainName -dn	domain_name	Name of the Informatica domain.
-UserName -un	user_name	User name to connect to the domain. Optional if the domain uses Kerberos authentication.
-Password -pd	password	Password for the user name.
-SecurityDomain -sdn	security_domain	Required if the domain uses LDAP authentication. Optional if the domain uses native authentication or Kerberos authentication. Name of the security domain to which the domain user belongs. If the domain uses native or LDAP authentication, the default is Native. If the domain uses Kerberos authentication, the default is the LDAP security domain created during installation.
-TargetProject -tp	target_project	Name of the project in the Model repository into which you want to import the objects. The project must exist in the Model repository before you import the objects. The option is ignored if you use an import control file.
-RepositoryService - rs	service_name	Model Repository Service name.
-ImportFilePath -fp	file_path	Path and file name of the XML file from which to import the rules metadata. You can specify an absolute path or a relative path to the file name.
-SourceProject -sp	project_name	Source project name in the file to import.

Option	Argument	Description
-TargetFolder -tf	folder_name	Target folder within the project into which you want to import the rules objects. If you do not specify a target folder, the objects are imported into the target project. The folder must exist in the repository before you import the objects.
-OtherOptions -oo	options	Additional options to import the dictionary content and the NER and Classifier model files from the specified zip file. Enter options using the following format: `rtm:<option_name>=<value>,<option_name>=<value>` Required option names include: - codePage. Code page of the reference data. - refDataFile. Path and file name of the zip file from where you want to import the reference table data.

The following example imports rules metadata from a dq_rules_metadata.xml file into a folder named dq_rules within a project named NEW_PROJECT. The example also imports the dictionary content and the NER and Classifier models from a dq_rules_artifacts.zip file into the project.

```
infacmd oie importObjects -dn InfaDomain -un Administrator -pd password -sdn native
-rs MRS_dev3 -tp NEW_PROJECT -tf dq_rules -fp "/export/dq_rules_metadata.xml" -ow true
-oo "rtm:codePage=UTF-8,refDataFile="/output/dq_rules_artifacts.zip""
```

CHAPTER 11

Data Type Reference

This chapter includes the following topics:

Data Type Reference Overview

Intelligent Data Lake supports most of the data types for data discovery, data preparation, data export, and other activities. Depending on the type of activity, some of the data types are not supported.

Data Type Support for Data Preview

When Intelligent Data Lake reads source data for an activity or event, it converts the native data types to the comparable transformation data types wherever possible. The following tables show the data type support for the data preview activity.

Data Preview from Hive Sources

The following table shows the data type support for the data preview activity from Hive sources.

Data Type	Support
Binary	Supported
Date/Time	Supported

Data Type	Support
Double	Supported
Integer	Supported
Decimal	There is a data loss for the decimal values.

Data Preview from Oracle Sources

The following table shows the data type support for the data preview activity from Oracle sources.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Number	Not Supported
Array	Not Supported
Struct	Not Supported
Union	Not Supported
Binary	Not Supported

When you try to preview any data of the unsupported data types, no data appears in the tables.

Data Preview from Teradata Sources

The following table shows the data type support for the data preview activity from Teradata sources.

Data Type	Support
Binary	Supported
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Interval	Not Supported

When you try to preview any data of the unsupported data types, no data appears in the tables.

Data Type Support for Import Operation

When Intelligent Data Lake reads source data for an activity or event, it converts the native data types to the comparable transformation data types wherever possible. The following tables show the data type support for the import activity.

Data Import from Oracle Sources

The following table shows the data type support for the data import activity from Oracle sources.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Number	Not Supported
Negative Scale Number	Not Supported
Struct	Not Supported
Union	Not Supported
Binary	Not Supported

When you try to import any data of the unsupported data types, no data appears in those columns.

Data Import from SQL Server Sources

The following table shows the data type support for the data import activity from SQL Server sources.

Data Type	Support
Binary	Supported
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Text	Not Supported
Float	Not Supported

When you try to import any data of the unsupported data types, no data appears in those columns.

Data Import from Azure Database Sources

The following table shows the data type support for the data import activity from Azure database sources.

Data Type	Support
Binary	Supported
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Text	Supported
Tinyint	Not Supported

When you try to import any data of the unsupported data types, no data appears in those columns.

Data Type Support for Export Operation

When Intelligent Data Lake exports data, it converts the native data types to the comparable transformation data types. The following tables show the data type support for the export operation.

Data Export to Oracle

The following table shows the data type support for the data export activity to Oracle.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Number	Not Supported
Double	Not Supported
Float	Not Supported
Timestamp	Not Supported

Data Export to Azure Database

The following table shows the data type support for the data export activity to an Azure database.

Data Type	Support
Binary	Supported
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Text	Supported
Varchar	Not Supported
Float	Not supported when the rows have an infinity value in the column

Data Type Support for Copy Operation

When Intelligent Data Lake reads source data for an activity or event, it converts the native data types to the comparable transformation data types wherever possible.

The following table shows the data type support for the activity of copying projects and worksheets.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Float	Supported
Map	Not Supported
Array	Not Supported
Struct	Not Supported
Union	Not Supported
Binary	Not Supported

Data Type Support for Publish Operation

When Intelligent Data Lake reads source data for an activity or event, it converts the native data types to the comparable transformation data types wherever possible.

The following table shows the data type support for the activity of publishing a worksheet.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Float	Supported
Map	Not Supported
Array	Not Supported
Struct	Not Supported
Union	Not Supported
Binary	Not Supported

Data Type Support for Upload Operation

When Intelligent Data Lake reads source data for an activity or event, it converts the native data types to the comparable transformation data types wherever possible.

The following table shows the data type support when uploading a file toIntelligent Data Lake.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Float	Supported
Map	Not Supported
Array	Not Supported

Data Type	Support
Struct	Not Supported
Union	Not Supported
Binary	Not Supported

Data Type Support for Download Operation

When Intelligent Data Lake reads source data for an activity or event, it converts the native data types to the comparable transformation data types wherever possible.

The following table shows the data type support for the activity of downloading a data asset as CSV or TDE file.

Data Type	Support
Date/Time	Supported
Double	Supported
Integer	Supported
Decimal	Supported
Float	Supported
Map	Not Supported
Array	Not Supported
Struct	Not Supported
Union	Not Supported
Binary	Not Supported

INDEX