Informatica®
10.1

User Guide

Informatica User Guide
10.1
August 2016

Publication Date: 2019-05-09

Table of Contents

Preface

The *Informatica Intelligent Data Lake User Guide* contains information for analysts about using the Intelligent Data Lake application to search for, discover, and prepare data for analysis. This book assumes that you know your data requirements and are familiar with applications such as Excel.

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.

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.
- 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

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 has the following features:

Search

- Find the data in the lake as well as in the other enterprise systems using smart search and inference-based results.
- Filter assets based on dynamic facets using system attributes and custom defined classifications.

Explore

- Get an overview of assets, including custom attributes, profiling statistics for data quality, data domains for business content, and usage information.
- Add business context information by crowd-sourcing metadata enrichment and tagging.
- Preview sample data to get a sense of the data asset based on user credentials.
- Get lineage of assets to understand where data is coming from and where it is going and to build trust in the data.
- Know how the data asset is related to other assets in the enterprise based on associations with other tables or views, users, reports and data domains.
- Progressively discover additional assets with lineage and relationship views.

Acquire

- Upload personal delimited files to the lake using a wizard-based interface.
  Hive tables are automatically created for the uploads in the most optimal format.
- Create, append to, or overwrite assets for uploaded data.

Collaborate

- Organize work by adding data assets to projects.
- Add collaborators to projects with different roles, such as co-owner, editor, or viewer, and with different privileges.

Recommendations

- Improve productivity by using recommendations based on the behavior and shared knowledge of other users.
- Get recommendations for alternate assets that can be used in a project.
- Get recommendations for additional assets that can be used a project.
- Recommendations change based on what is in the project.

Prepare

- Use excel-like environment to interactively specify transformation using sample data.
- See sheet-level and column-level overviews, including value distributions and numeric and date distributions.
- Add transformations in the form of recipe steps and see the results immediately on the sheets.
- Perform column-level data cleansing and data transformation using string, math, date, logical operations.
- Perform sheet-level operations to combine, merge, aggregate, or filter data.
- Refresh the sample in the worksheet if the data in the underlying tables change.
- Derive sheets from existing sheets and get alerts when parent sheets change.
- All transformation steps are stored in the recipe which can be played back interactively.

Publish

- Use the power of the underlying Hadoop system to run large-scale data transformation without coding or scripting.
- Run data preparation steps on actual large data sets in the lake to create new data assets.
- Publish the data in the lake as a Hive table in the desired database.
- Create, append, or overwrite assets for published data.

Data Asset Operations

- Export data from the lake to a CSV file.
- Copy data into another database or table.
- Delete the data asset if allowed by user credentials.

My Activities

- Keep track of upload activities and their status.
- Keep track of publications and their status.
- View log files in case of errors and share with IT administrators if needed.

IT Monitoring

- Keep track of user, data asset and project activities by building reports on top of the audit database.
- Find information such as the top active users, the top data assets by size, prior updates, most reused assets, and the most active projects.

IT Operationalization

- Operationalize the ad-hoc work done by analysts.
- User Informatica Developer to customize and optimize the Informatica Big Data Management mappings translated from the recipes that analysts create.
- Deploy, schedule, and monitor the Informatica Big Data Management mappings to ensure that data assets are delivered at the right time to the right destinations.
- Make sure that the entitlements for access to various databases and tables in the data lake are according to security policies.

Intelligent Data Lake Concepts

To successfully use the Intelligent Data Lake application, you must understand the concepts that are used in the tool.

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.

Use the Intelligent Data Lake application to search, discover, and prepare data that resides in the data lake. When you prepare the data, you can combine, cleanse, and transform the data to create new insights. You can add data to the data lake by uploading the data or publishing prepared data.

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

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

Data Asset

A data asset is data that you work with as a unit. 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.

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

After you find the data asset you are interested in, you can add the data asset to a project and then prepare the data for analysis.

Project

A project is a container stores data assets and worksheets. When you add a data asset to a project, Intelligent Data Lake uses the data asset as an input source and creates a corresponding worksheet that contains a sample of the data. When you publish the worksheet as a data asset to the data lake, the project displays the publication for the worksheet.

Data Preparation

The process of combining, cleansing, transforming, and structuring data from one or more data assets so that it is ready for analysis.

In Intelligent Data Lake you use worksheets in a project to create data preparation recipes.

Data Publication

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

You can publish a worksheet that contains prepared data. To publish a worksheet, you select the worksheet in the project. The published worksheet appears as a publication in the project.

When you 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.

Recipe

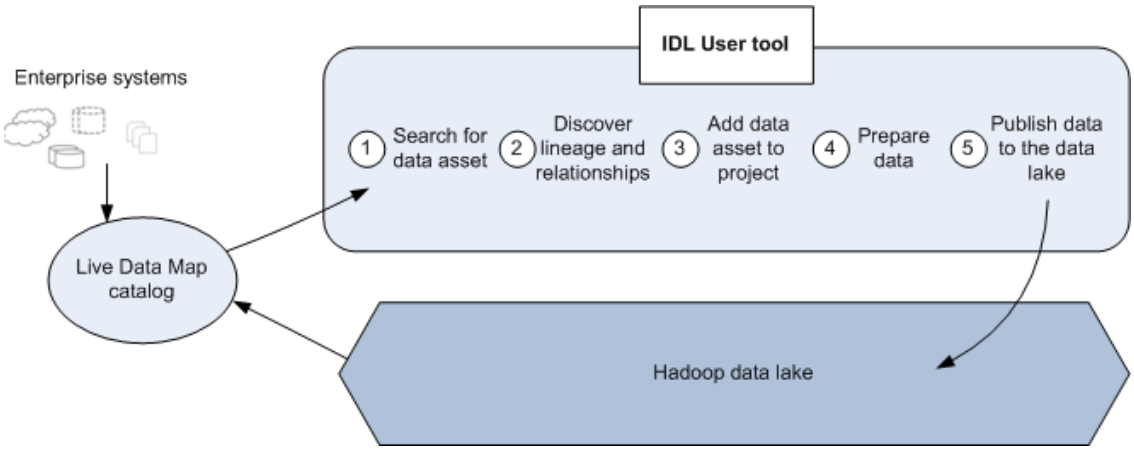
A recipe includes the list of input sources and the steps taken to prepare data in a worksheet. When you select a worksheet in a project to publish prepared data in the worksheet, Intelligent Data Lake applies the recipe to the input source.

Data Discovery and Analysis Process

Use the Intelligent Data Lake application to search for and discover data that resides in and outside the data lake. You can prepare data that resides in the data lake. You can also move data outside the data lake into the data lake to use and prepare the data for analysis.

You can use the Intelligent Data Lake application after an administrator uses the Informatica Live Data Map to create a catalog of assets that reside in the enterprise.

The following image shows the high-level tasks that you complete to prepare data for analysis using the Intelligent Data Lake application:



Complete the following high-level tasks in the Intelligent Data Lake application to prepare data:

1. Search for data assets.

   If you find data outside the data lake that you want to work with, ask your administrator to add the data to the data lake.
2. Discover the lineage and relationships between data.
3. When you find the data asset that you want to prepare, create a project and add the data asset to the project.
4. Prepare the data by combining, cleansing, transforming, and structuring the data so that it is ready for analysis.
5. Publish the prepared data to the data lake so that the data asset can be used in other projects..

   If you want to regularly load data into the data lake using your preparation steps, ask your administrator to operationalize the steps.

Example - Data Discovery and Analysis Process

Sarah is a data analyst in the marketing department of a national telecom company who needs to know how many customers are considering switching from her company to another provider. When she identifies the list of customers who might switch, she can run targeted marketing campaigns to retain these customers.

To get analyze data about customers who might switch to another provider, Sarah performs the following table tasks:

Search for data assets.

Sarah uses the Intelligent Data Lake application to search for data assets that contain carrier tracking numbers. She can use wildcards or search filters to search for relevant data assets. If Sarah finds relevant data that is not in the data lake, she can ask her administrator to add the data to the data lake.

Sarah finds an asset named customer_details_records in the data lake and views its details in the **Overview** view to identify the tables that contain the data. She notices from the columns that the table contains a log of all calls made by all customers who cancelled subscription in the previous year. A note in the table description states that because the data in this table was directly extracted from the enterprise data warehouse, it might not be clean and must be used with caution. Sarah then previews a sample of the data on the **Data Preview** view to verify the quality of the data in the customer_details_records table and determines that this is data that she might want to work with.

Discover the lineage and relationships between data.

Sarah views the **Lineage** view and notices that the data flow starts from the call_rec_agg table, which aggregates calls made by customer. She identifies the call_rec_agg table as another data asset for her analysis.

Sarah views the **Relationship** view and explores the relationships between the call_rec_agg table and other assets. She finds that the call_rec_table has a relationship with the CustomerID data domain. She further drills down into the CustomerID tables to discover that the customer_master table contains master data extracted from a CRM system on a daily basis. She decides to use the customer_master table for her analysis.

Add the data asset to the project.

Sarah decides to use the call_rec_agg table in her analysis. Before she prepares her data asset, she must add the data asset to a project. She creates a project named Customer_Churn_Analysis and adds the call_rec_agg table as a data asset to the project.

Prepare the data.

Sarah decides to prepare a trustworthy data asset by combing, cleaning, and structuring the worksheets in the Customer_Churn_Analysis project so that the data is ready for analysis.

In the project, she clicks **Prepare** > **Prepare** to edit the worksheets. She combines the worksheets into a new worksheet named Combination and works with recipes to add formulas with expression functions to the Combination worksheet. When she saves her work, the Combination worksheet appears in the Customer_Churn_Analysis project as a work in progress.

Publish the data to the data lake.

Sarah decides that she is really happy with the data in the Combination worksheet and decides to publish the worksheet. By publishing the worksheet, Sarah loads the data into the data lake using her data preparation steps.

To publish the Combination worksheet, she selects the worksheet and clicks **Publish**. After the worksheet is published, Sarah can view her publication history by selecting the **My Data Assets** view. Her recent publications appear on the **My Publications** view.

In the Customer_Churn_Analysis project, Sarah can select the published Combination worksheet and export the worksheet as a .csv file. Sarah can use third-party tools to generate business intelligence on this data asset. Since Sarah wants to regularly load data into the data lake using her data preparation steps, she asks her administrator to operationalize the steps so that the data asset is published to the data lake every month.

Logging In to the Intelligent Data Lake Application

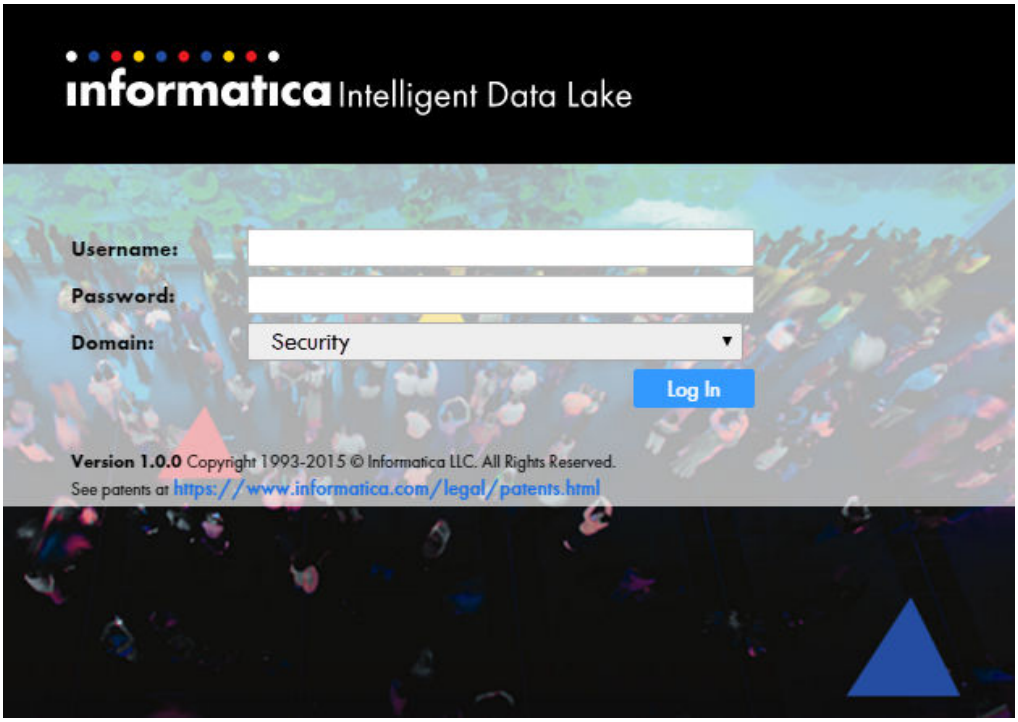
The Intelligent Data Lake application is a web application that becomes available after the system administrator creates the services that are required by Intelligent Data Lake. Contact your system administrator to get the URL for the Intelligent Data Lake application.

1. Start a Microsoft Internet Explorer or Google Chrome browser.
2. In the address field, enter the URL provided by the administrator.

   By default, the URL consists of the host name and port number and the /idl directory:

   ```
   http://<hostname>:<port>/idl/
   ```
3. Press **Enter**.

   The Intelligent Data Lake application login page appears.



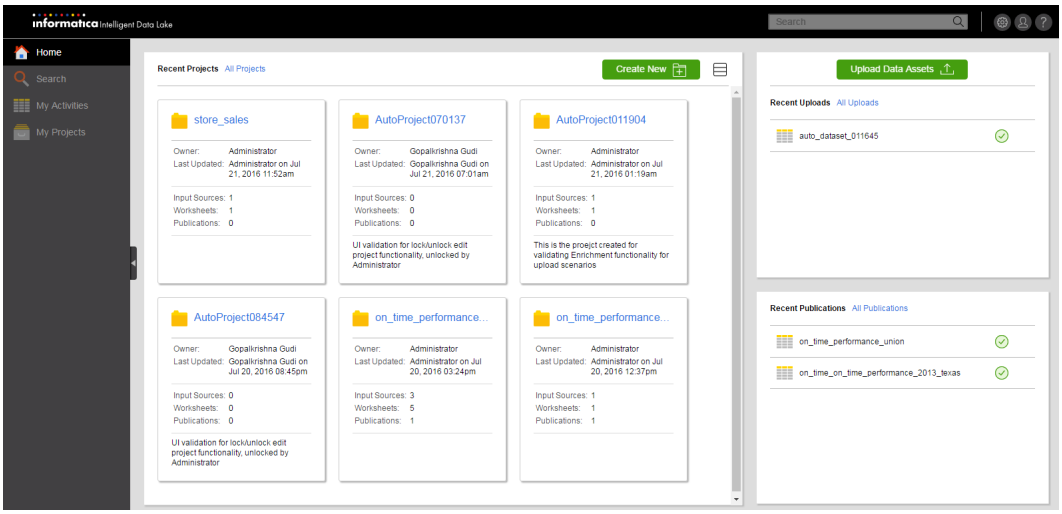
4. Enter your user name and password.
5. Select the security domain for your user account.
6. Click **Log In**.

   The **Home** view appears.

User Interface

The Intelligent Data Lake application is a web-based interface that enables you to search, discover, and prepare big data sets with minimal training or involvement from IT.

Let's take a look at the main components that appear in the Intelligent Data Lake application user interface:



Header

The header is a black bar at the top of the Intelligent Data Lake application. The header contains the search box and icons you can click to configure search filters, log out, or view online help.

Let's explore what you can do on the header:

- To search for data assets, enter a search query in the search box.
- To configure search filters, click the **Application Configuration** icon ( ).
- To log out of the Intelligent Data Lake application, click the **Log Out** icon ( ).
- To view online help, click the **About** icon ( ).

Home View

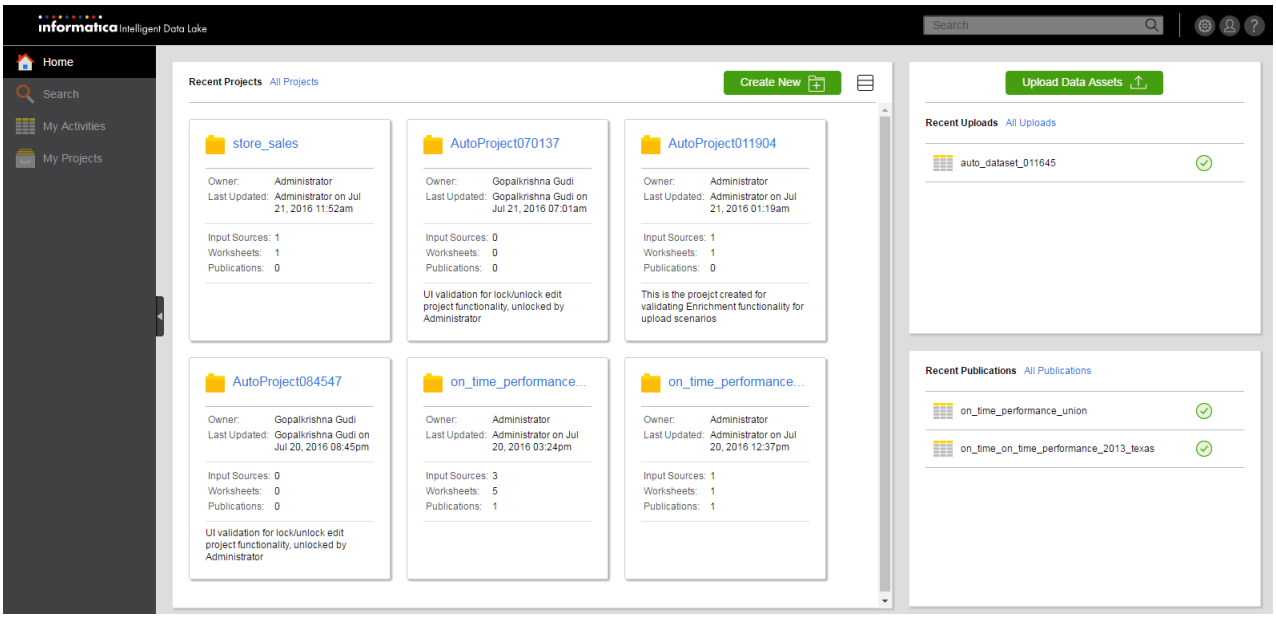
The **Home** view displays your recent projects, uploads, and publications. Use the **Home** view to get a quick picture of your activities in the Intelligent Data Lake application.

Let's explore what you can do in the **Home** view:

- View and create projects. The **Recent Projects** panel displays projects with the latest projects on top. To see a list of projects, click the **Switch to List View** icon ( ). If you want to see all projects, click **All Projects**. To create a new project, click **Create New**.
- View recent uploads. The **Recent Uploads** panel shows data assets that you recently added to the data lake. You can also click **Upload Data Assets** to upload a data asset into the data lake. Click **All Uploads** to get a complete list of all data assets that you added to the data lake.

- View recent publications. The **Recent Publications** panel shows publications that you recently published to the data lake. To view all published publications, click **All Publications**.

The following image shows the **Home** view:

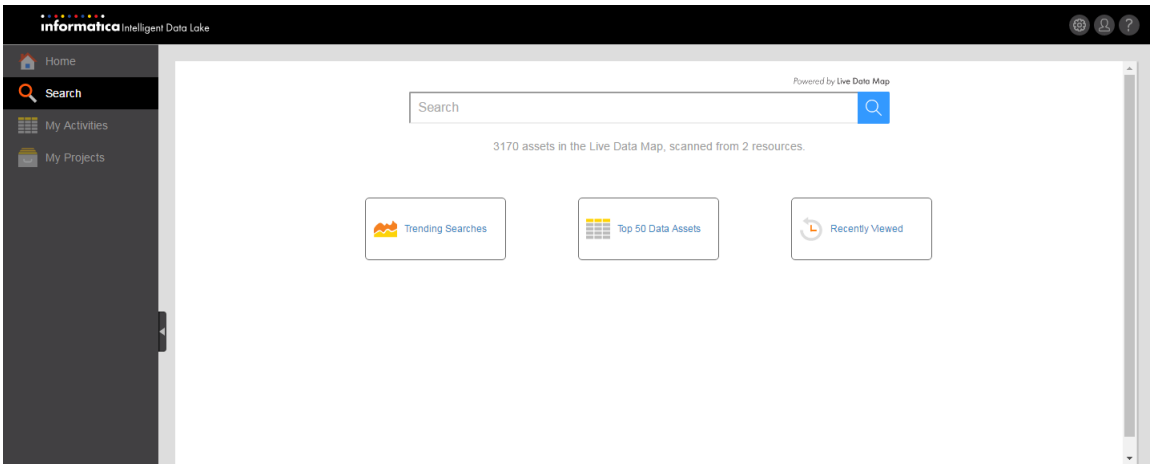


Search Results View

The **Search Results** view displays the Enterprise Information Catalog search.

The Enterprise Information Catalog displays the total number of data assets in the catalog along with the number of resources.

The following image shows the **Search Results View**:



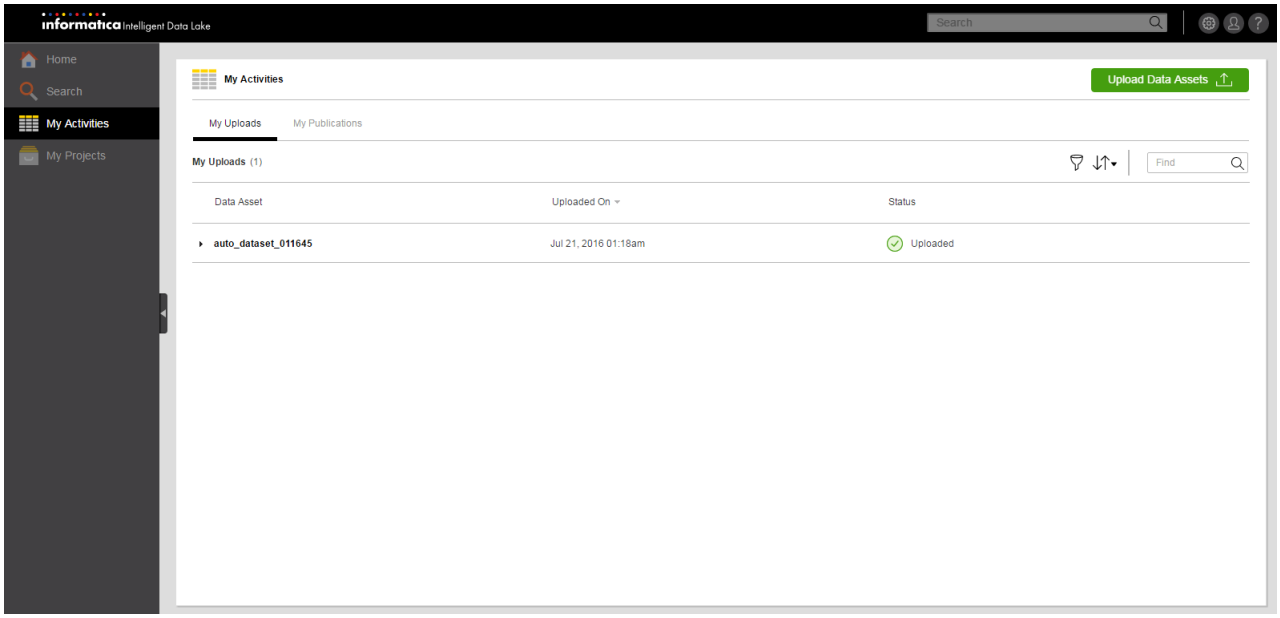
My Activities View

The **My Activities** view displays all activities that you have completed. Activities can include uploading and publishing data assets.

Let's explore what you can do in the **My Activities** view:

- Click the name of a data asset to view its details. The data asset opens in the **Open Asset** view.
- Upload a data asset. Click **Upload Data Assets** to upload a data asset to the data lake.
- View your uploaded data asset. The **My Uploads** tab displays the data asset that you uploaded into the data lake.
- Filter your list of uploaded data assets. On the **My Uploads** tab, click the filter icon ( ) to filter uploads by time frame and status. Select **Uploaded On** to filter by all, last 24 hours, last week, or last month. Select **Status** to filter by all, uploaded, in progress, or failed.
- Search for an uploaded data asset. On the **My Uploads** tab, use the search box to search for an uploaded data asset.
- Delete an uploaded data asset. On the **My Uploads** tab, click the mange data assets icon ( ) to delete an uploaded data asset.
- View your publications. The **My Publications** tab displays the data assets that you published.
- Filter your list of publications. On the **My Publications** tab, click the filter icon ( ) to filter publications by time frame and status. Select **Published On** to filter by all, last 24 hours, last week, or last month. Select **Status** to filter by all, uploaded, in progress, or failed.
- Search for a publication. On the **My Publications** tab, use the search box to search for a publication.
- Delete a publication. On the **My Publications** tab, click the manage data assets icon ( ) to delete a publication.

The following image shows the **My Activities** view:



Projects View

The **Projects** view displays all projects that you have created or that you are a collaborator on.

Let's explore what you can do on the **Projects** view:

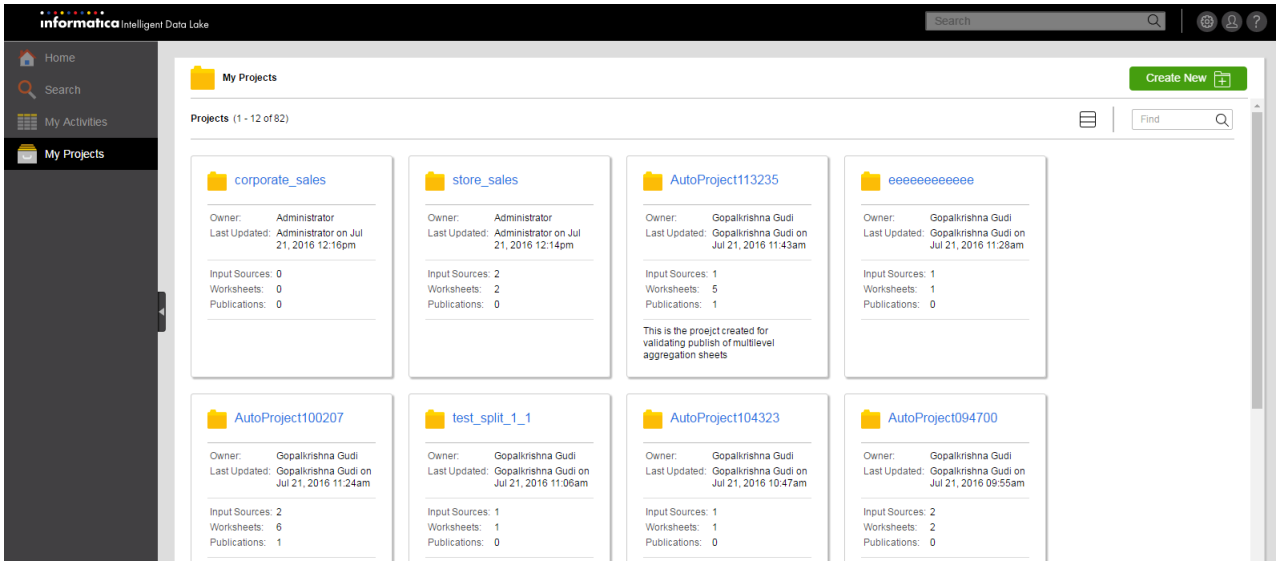
- Create a project. Click **Create New** to create a project.
- Switch the views on a project. By default, the Intelligent Data Lake application displays each project in a card view. Click the **Switch to List View** icon ( ) to switch to a list view. To return to card view, click the **Switch to Card View** icon ( ).
- Search for a project. Use the search box to search for a project.
- View all projects. To view all projects, click **All Pojects**.
- Click the name of a project to view details about the project. The project opens in a different view. You can perform the following tasks in this view:

  - Edit the name and description of the project. Click the edit icon ( ) in the **Overview** panel to edit these project properties.

  - Change the collaborators on the project. Click the edit icon ( ) in the **Collaborators** panel to change the users who can collaborate on the project.
  - View or edit a worksheet in the project. Each project displays a worksheet for each data asset in the project. In the **Worksheets** panel, click **Prepare** and choose to view a worksheet or open it for editing.

    You can also select a worksheet in a project and click the Publish icon ( ) to publish the worksheet.
  - View recommendations. View recommendations for data assets used in different projects by trusted people in the organization on the **Recommendations** panel.

The following image shows the **Projects** view:



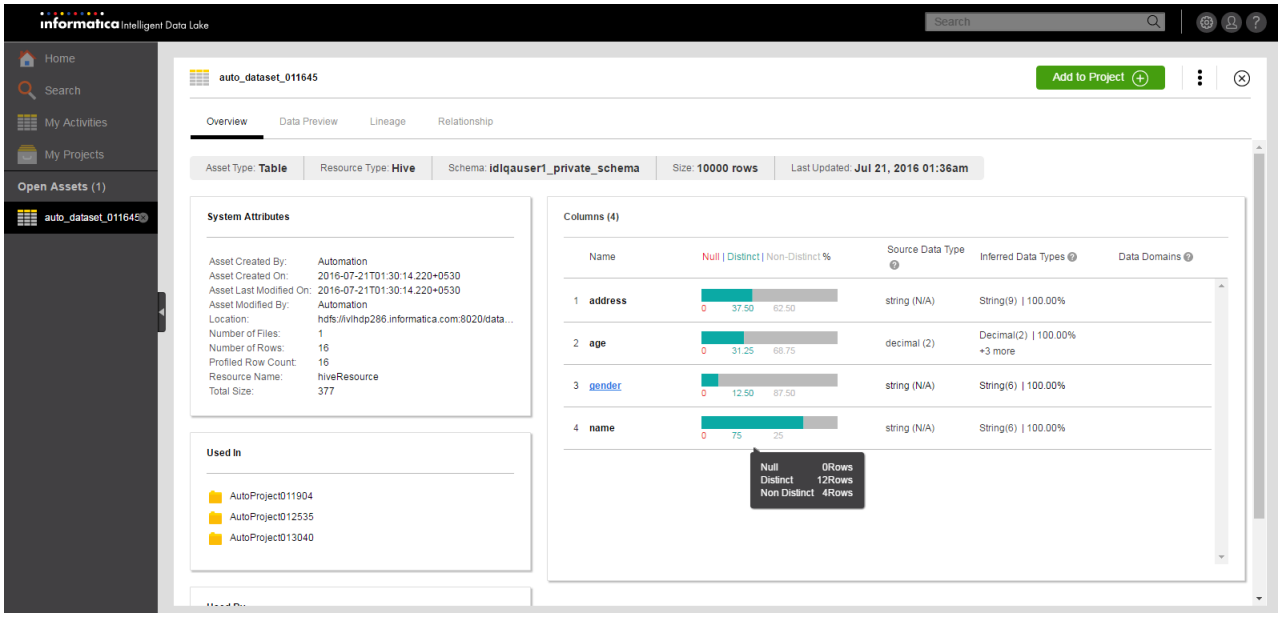
Open Assets

When you select a data asset from the **My Activities** or **Projects** view, each data asset opens in an asset view labeled with the data asset name.

Let's explore what you can do in the **Open Assets** view:

- Add the data asset to a project. Click **Add to Project** to add the data asset to a project.
- Manage the data asset. Click the Manage icon ( ) to edit data asset properties, publish the data asset, export the data asset as a CSV file or delete the data asset.
- View data asset details in the **Overview** tab. You can perform the following actions when you view asset details:
  - View properties such as the business or technical owner of the data asset on the **Custom Attributes** panel. Click the Edit icon ( ) to edit these properties.
  - View where the data asset was used in a project on the **Used In** panel.
  - View which users used the data asset on the **Used By** panel.
- Preview data on the **Data Preview** tab.
- Discover the lineage for the flow of data for the data asset on the **Lineage** tab.
- Discover the relationships between the open data asset with other data assets on the **Relationship** tab.

The following image shows the **Open Assets** view:



CHAPTER 2

Search

This chapter includes the following topics:

Overview

You can search for the required data assets in Intelligent Data Lake. You can search based on the name of the data asset or perform a generic search using the wildcard character asterisk (*) or the wildcard character question mark (?). Intelligent Data Lake also displays probable matches when you type the name of a required data asset. You can also search for assets such as users and data domains.

You can search for and discover any data asset described in the catalog. You can search for, discover, and prepare data assets that are stored in the data lake. Data assets can include items such as a flat file, table, or view.

From the search results displayed, you can sort the results based on the data asset name, the data source, or the system attributes or the custom attributes. You can use the search filters displayed to filter the search results and view additional details for the displayed data assets. After searching and finding the required data asset, you can annotate and enrich the required data assets with custom attributes.

After finding the required data asset, you can click the data asset to view the profiling details, the lineage, and the relationship of the data asset with other data assets. You can also add the data asset to a project.

Search Strings

You can use search strings to search for data assets. You can use wildcard characters in search strings. Intelligent Data Lake also lists search suggestions when you enter a search string.

Wildcard Search

You can use the wildcard characters asterisk (*) and the question mark (?) to perform a search to find matching data assets from the resources. If you specify * in the **Search** box and click the **Search** icon,

Intelligent Data Lake lists all the data assets in the catalog. You can use the ? wildcard character to substitute individual letters in the name of a data asset that you want to search. For example, if you know that the data asset that you want to search for, begins with the letters `HR`, followed by two numbers that denote the year, and ends with REPORT, you can specify the following string in the **Search** box: `HR??REPORT`. Intelligent Data Lake lists all the data assets that match the search criteria. For example, `HR12REPORT`, `HR13REPORT`, `HR14REPORT`.

You can also use asterisk along with parts of the data asset name. For example, if you want to search for all the data assets that begin with the word NAME in the catalog, you can specify `NAME*`.

Search Suggestions

Intelligent Data Lake lists matching data asset names when you type the first few letters of a data asset.

If you type the name of a data asset incorrectly, Intelligent Data Lake compares the typed letters with names of existing data assets in the catalog. The probable matches then appear as search suggestions. For example, if you typed `sela` to search for the data asset named `salary`, Intelligent Data Lake suggests `salary` as a probable match.

Note: The suggestions include all custom attributes that are of type string. Data assets or custom attribute names that include special characters do not get listed.

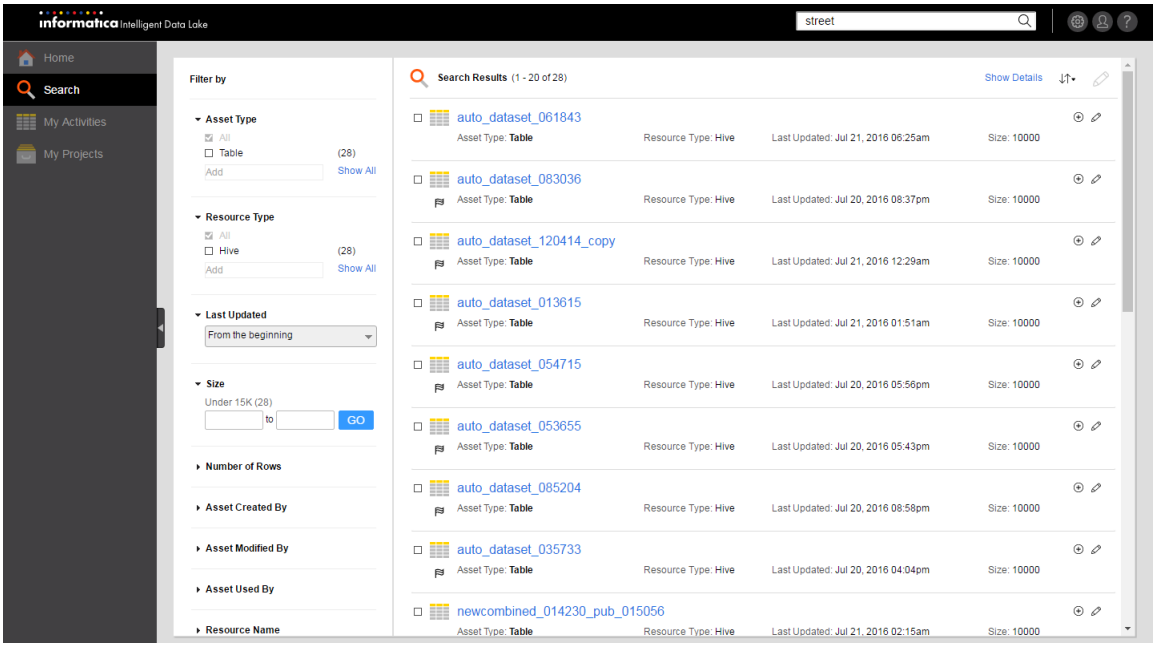
Search Results

Search results can include data assets stored in and outside the data lake. If you select a data asset to add to a project, the data asset must be stored in the data lake. Search results can also include a user asset that lists the owner of the found data asset.

When you type the data asset that you want to search for in the search box, the matching results appear. You can interpret the search results in the following ways:

- The search results show a summary of the details of the data asset.
- The number in parentheses under the **Filter by** panel sections indicates the number of matching results found for the searched data asset under those categories.

The following image shows the search results:



Sort Search Results

You can sort the displayed data assets based on the name of the listed data assets or the relevance of the data asset.

- To sort data assets based on the name, click ↓↑ and select **Name**.
- To sort data assets based on the relevance to the search, click ↓↑ and select **Relevance**.

Note: If you sort the data assets based on name, the sort options list displays an up arrow (↑) or a down arrow (↓) before the sort criteria selected. An up arrow indicates that the data assets are sorted in the ascending order based on the selected criteria and a down arrow indicates that the data assets are sorted in the descending order.

Search Filters

After you search for a data asset, you can use the following search filters to refine the search results. These filters help you filter the data assets based on your requirements.

You can use the following search filters:

- **Asset Type**: The type of data asset.
- **Resource Type**: The resource type from where the data asset details were collected.
- **Last Updated**: The last time the details of the data asset were updated in the catalog.

- **Size**: The size of the data asset.
- The system attributes or custom attributes.

Refine Search by Data Asset Type

You can refine your search based on the type of data asset. You can use the **Asset Type** filter for refining the search.

The Asset Type Filter

You can select **All** from the **Asset Type** filter to specify that the search is applicable to all types of data assets. You can click a specific type of data asset or click multiple data assets based on your search requirement. For example, you can use the following data types:

- Table
- Resource
- View

For example, when you select **Table** from the **Asset Type** list, Intelligent Data Lake lists all the data assets of type Table.

Adding Data Asset Types

By default, Intelligent Data Lake does not list all the data asset types in the **Asset Type** list. You can add data asset types that are not displayed to the **Asset Type** list. You can then use these data asset types to refine your search.

To add data asset types, perform the following steps:

1. Click **Add**. A box appears where you can type the name of the asset to be added.
2. Start typing the name of the data asset in the box. A list of matching data asset types appears.
3. Select the data asset type that you want. Intelligent Data Lake refreshes the search results based on the type of data asset selected.

Viewing all the Data Asset Types

You can view all the data asset types that are not listed in the **Asset Type** list. You can then add the required data asset types to the **Asset Type** list and refine your search.

Perform the following steps to add data asset types:

1. Click **show all**. The **Select Asset Type** dialog box appears.

   Note: The **show all** option appears if the number of data assets is more than five in the catalog.
2. Select the types of data assets from the **Name** column as required.
3. Click **OK**. Intelligent Data Lake refreshes the search results based on the attribute selected.

Refine Search by Resource Type

You can search for data assets based on specific data sources.

The Resource Type Filter

You can select **All** from the **Resource Type** list to specify that the search is applicable to all the data source types.

Adding Resource Types

By default, Intelligent Data Lake does not display all the data sources in the **Resource Type** list. You can add data sources that are not displayed to the **Source** list. You can then use these sources to refine your search. To add data sources to the **Resource Type** list, perform the following steps:

1. Click **Add**. A box appears where you can type the name of the data source to be added.
2. Start typing the name of the data source in the box. A list of matching data sources appears.
3. Select the data source that you want. Intelligent Data Lake refreshes the search results based on the type of data source selected.

Viewing all the Resource Types

You can view all the data source types that are not listed in the **Resource Type** list. You can then add the required data source types to the **Resource Type** list and refine your search. Perform the following steps to view all the data source types:

1. Click **show all**. The **Select Resource Type** dialog box appears.

   Note: The **show all** option appears if the number of resource types is more than five in the catalog.
2. Select the required data sources from the **Name** column.
3. Click **OK**. Intelligent Data Lake refreshes the search results based on the type of data source selected.

Refine Search by Last Updated Time

You can search for data assets based on when the information about the data asset was last updated in the catalog.

You can select one of the following options from the **Last Updated** list:

Option	Description
Today	Displays the search results for data asset details updated on the present day.
Yesterday	Displays the search results for data asset details updated on the previous day.
Last 7 days	Displays the search results for data asset details updated during the last seven days.
Last 1 month	Displays the search results for data asset details updated during the last one month.
Last 6 months	Displays the search results for data asset details updated during the last six months.
Last 1 year	Displays the search results for data asset details updated during the last year.
From the beginning	Displays the search results for data asset details present in the catalog from the time the catalog was created.
Custom	Displays the search results based on a range of dates specified by you. To specify the date ranges, perform the following steps: 1. Select the **Custom** option from the **Last Updated** list. 2. Click the boxes to launch the calendar and select the date ranges. 3. Click **GO** to display the search results.

Refine Search by Asset Created By

You can search for data assets based on the users who created the data assets.

The Asset Created By Filter

You can select **All** from the **Asset Created By** list to specify that the search is applicable to all the users.

Adding Users

By default, Intelligent Data Lake does not display all the users in the **Asset Created By** list. You can add users who are not displayed to the list. You can then use these users to refine your search. To add users to the **Asset Created By** list, perform the following steps:

1. Click **Add**. A dialog box appears where you can type the name of the user to be added.
2. Start typing the name of the user in the text box. A list of matching user names appears.
3. Select the user that you want. Intelligent Data Lake refreshes the search results based on the users selected.

Viewing all the Users

You can view all the users that are not listed in the **Asset Created By** list. You can then add the required users to the **Asset Created By** list and refine your search. Perform the following steps to view all the users:

1. Click **show all**. The **Select Asset Created By** dialog box appears.
2. Select the required users from the **Name** column.
3. Click **OK**. Intelligent Data Lake refreshes the search results based on the users selected.

Refine Search by Asset Modified By

You can search for data assets based on the users who modified the data assets.

The Asset Modified By Filter

You can select **All** from the **Asset Modified By** list to specify that the search is applicable to all the users.

Adding Users

By default, Intelligent Data Lake does not display all the users in the **Asset Modified By** list. You can add users who are not displayed to the list. You can then use these users to refine your search. To add users to the **Asset Modified By** list, perform the following steps:

1. Click **Add**. A dialog box appears where you can type the name of the user to be added.
2. Start typing the name of the user in the text box. A list of matching user names appears.
3. Select the user that you want. Intelligent Data Lake refreshes the search results based on the users selected.

Viewing all the Users

You can view all the users that are not listed in the **Asset Modified By** list. You can then add the required users to the **Asset Modified By** list and refine your search. Perform the following steps to view all the users:

1. Click **show all**. The **Select Asset Modified By** dialog box appears.
2. Select the required users from the **Name** column.
3. Click **OK**. Intelligent Data Lake refreshes the search results based on the users selected.

Refine Search by Asset Used By

You can search for data assets based on the users who used the data assets.

The Asset Used By Filter

You can select **All** from the **Asset Used By** list to specify that the search is applicable to all the users.

Adding Users

By default, Intelligent Data Lake does not display all the users in the **Asset Used By** list. You can add users who are not displayed to the list. You can then use these users to refine your search. To add users to the **Asset Used By** list, perform the following steps:

1. Click **Add**. A dialog box appears where you can type the name of the user to be added.
2. Start typing the name of the user in the text box. A list of matching user names appears.
3. Select the user that you want. Intelligent Data Lake refreshes the search results based on the users selected.

Viewing all the Users

You can view all the users that are not listed in the **Asset Used By** list. You can then add the required users to the **Asset Used By** list and refine your search. Perform the following steps to view all the users:

1. Click **show all**. The **Select Asset Used By** dialog box appears.
2. Select the required users from the **Name** column.
3. Click **OK**. Intelligent Data Lake refreshes the search results based on the users selected.

Refine Search by Data Asset Size

You can search for data assets based on the size of the data asset. Intelligent Data Lake displays multiple ranges for the sizes based on the data assets in the resources.

You can use the following options to search for a data asset based on size:

- Select **All** to specify that data assets of all sizes must be considered for a search.
- Select a specific range listed.
- Specify a range for the data asset sizes if you do not find the range of data asset in the displayed list. To specify a range of sizes for the searched data assets, type the ranges in the boxes provided and click **GO**.

Refine Search by System Attributes or Custom Attributes

You can refine your search based on the system attributes or the custom attributes listed.

Adding Attributes

1. Click **Add**.
2. Start typing the name of the attribute in the box. A list of matching attributes appears.
3. Select the attribute that you want. Intelligent Data Lake refreshes the search results based on the attribute selected.

Viewing all the Attributes

1. Click **show all**. The attribute dialog box appears.

   Note: The **show all** option appears if the number of attributes defined is more than five.

2. Select the required attributes and the included attributes from the **Name** column.
3. Click **OK**.

Editing Data Asset Properties to Assign Custom Attributes

You can edit the properties of a data asset to assign custom attributes to that data asset. Assigning custom attributes to a data asset helps you refine your search and find the required data asset faster. You can search for the required data assets using the custom attribute attached to the data assets.

1. Select the data asset from the list of search results.
2. Click the **Edit Properties** () icon.

   The **Edit Properties** dialog box appears. If you had select multiple data assets, the **Edit Properties** dialog box displays an additional column named **Assigned To**. This column indicates the number of objects in the selected data asset that are assigned the specific custom attribute.
3. Select the required custom attributes from the **Properties** section.

   You can also type the required attribute in the **Find** box. The list of matching attributes appears.
4. Select the required attributes from the **Name** section.
5. Click **OK**.

CHAPTER 3

Data Discovery

This chapter includes the following topics:

Data Discovery Overview

After you search for data assets in the catalog, click a data asset to discover the data. During data discovery, you can view details about the data asset, lineage for the data asset, and relationships between the selected data asset and other data assets in the catalog.

You can discover all data assets in the catalog. However, you can only prepare data assets that are stored in the data lake.

Data Asset Views

Use the data asset views to get more information about the data asset by viewing the data asset details, previewing the data, and looking at the data relationship and lineage diagrams.

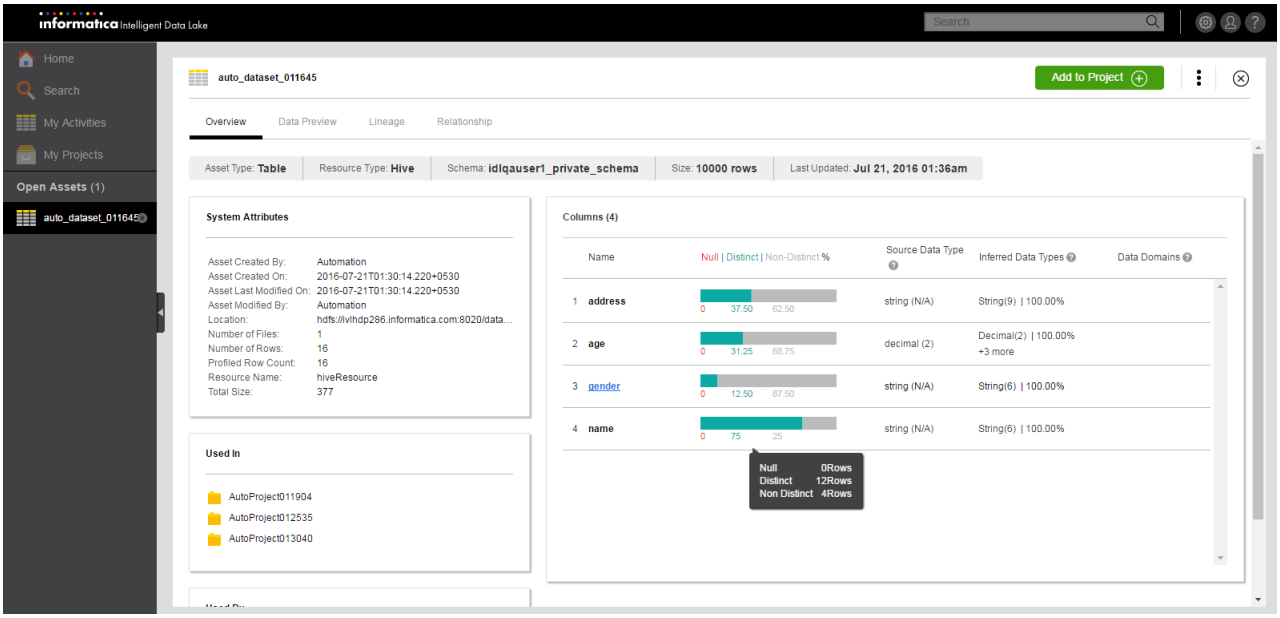
The Intelligent Data Lake application displays additional details about the selected data asset in the following views:

- Overview: Displays the details about the data asset. Details include the type of the data asset, the last time the data asset was updated, and the data source, along with other details.
- Data Preview: Displays a sample of the data in the data source.
- Lineage: Displays the lineage for the data asset selected. The lineage diagram shows the end-to-end data flow for the selected data asset.
- Relationship: Displays the relationships between the selected data asset and other data assets in the catalog.

Overview View

The **Overview** view displays details of the selected data asset.

The following image shows the **Overview** view:



Based on your data asset type, you can view the following details in the **Overview** view.

- Asset Type: The type of the data asset.
- Resource Type: The source type for the data set.
- The parent data asset for the selected data asset. For example, for a column, this is the name of the table that includes the selected column.
- Size: The size of the data asset. For example, for an Oracle table, this is the number of rows in the table.
- Last Updated: The time and date at which the details of the data asset was last updated in the catalog.
- System Attributes: Displays the system attributes associated with the selected data asset along with the values for the attributes.
- Custom Attributes: Displays the custom attributes associated with the selected data asset along with the values configured for the custom attributes.

Based on the type of the data asset selected, the Intelligent Data Lake application displays additional details for the data asset.

The following table contains additional details displayed for some of the common types of data assets:

Data Asset	Description

Table	- Columns: Displays the profiling information about the data in the data asset if your selected data asset is a table:
	  - Name: The name of the column in the table. Click a column to view the details for that column.
	  - Null Unique Non Unique: The number of null, unique, and non unique values in the columns.
	  - Source Data Type: The basic type of data, namely number, character, or date.
	  - Inferred Data Types: The type of data stored in the database that is derived from the basic type of data defined in **Source Data Type**. For example, string for character type data, decimals for numeric data, or time for date data type.
	  - Data Domains: Data that matches patterns defined earlier. For example, the format of the Social Security Number, bank account numbers, and credit card numbers.
	Note: You can sort the listed details in the **Columns** section based on any of the details by selecting the respective header for that detail. For example, selecting the **Name** header sorts the listed data alphabetically in the ascending or the descending order.

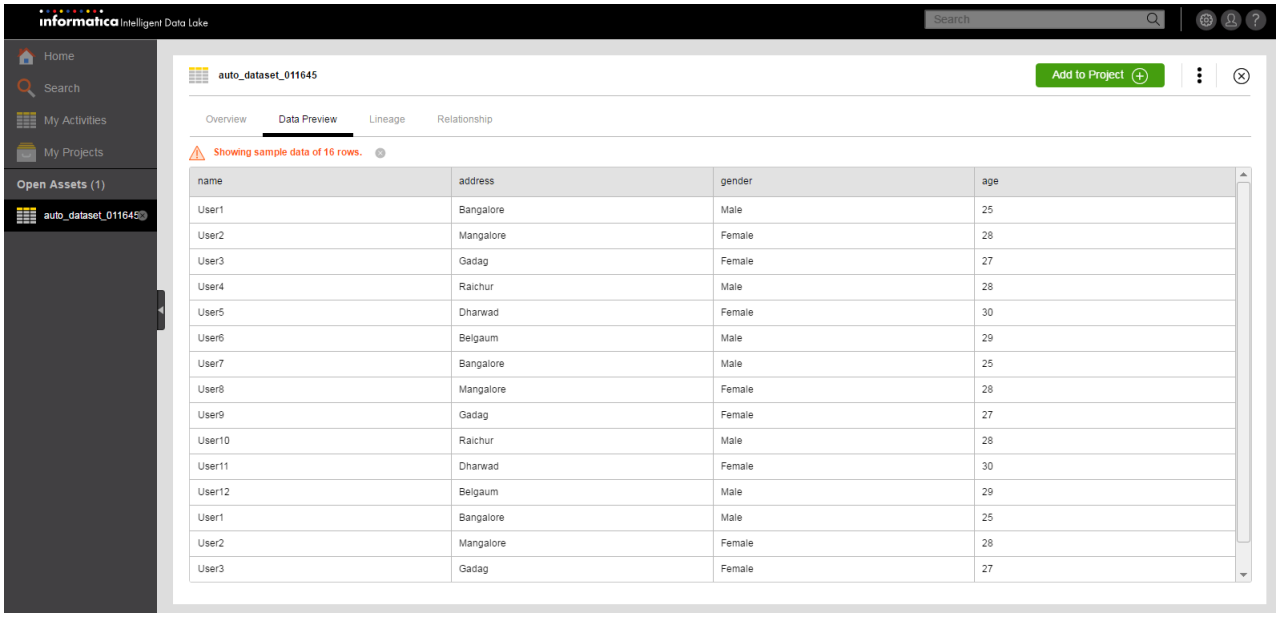
Column	- Value Distribution: Displays a color-coded representation of the null, unique, and non unique data values in the columns. - Pattern: Displays the percentage of data values grouped into ranges. - Inferred Data Types: The percentage of the type of data derived from the basic data type defined. - Data Domains: Data that matches patterns defined earlier. For example, the format of the Social Security Number, bank account numbers, and credit card numbers.
Resource	- A list of records included in the resource with the name and the type of record. - Resource Contains: Displays the list of assets in the resource. - A panel that displays the time when the resource was last scanned along with the status of the last scan.
View	Same as the additional details displayed for a table data asset.

Data Preview View

The **Data Preview** view displays the first 500 rows of a data asset stored in the data lake. The **Data Preview** is not available for data assets that are not in the data lake.

You can view data from tables and views on the Data Preview view. Each page shows 20 rows of data. You can use pagination options to navigate if a table or view contains more than 20 rows of data.

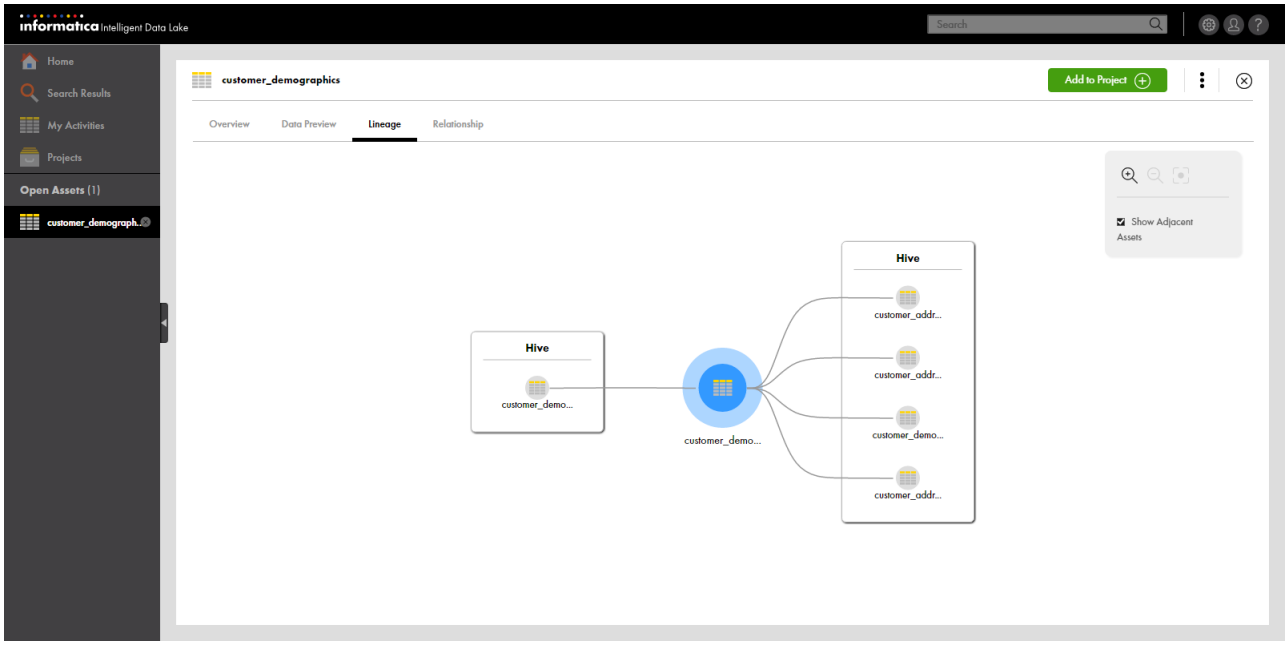
The following image shows the **Data Preview** view:



Lineage View

The **Lineage** view shows the start point of the data, describes the path, and shows how the data asset arrives at the end point. The lineage diagram shows the end-to-end data flow for the selected data asset.

The following image shows the **Lineage** view:



You can use the following icons to view the lineage diagram based on your requirements:

- Use the **Zoom in** ( ) icon to increase the magnification and the **Zoom out** icon ( ) to decrease the magnification.

- Click **View Adjacent Assets** to toggle between displaying and hiding the immediate neighboring data assets.
- Click the **Reset** icon ( ) to reset the lineage diagram back to the initial view size.

Expanded Lineage View

The **Expand Lineage Path** ( ) icon in the lineage diagram indicates that there are more objects in the data flow. You can place your pointer on any data asset toward the origin or end point in the data flow to view the expanded path.

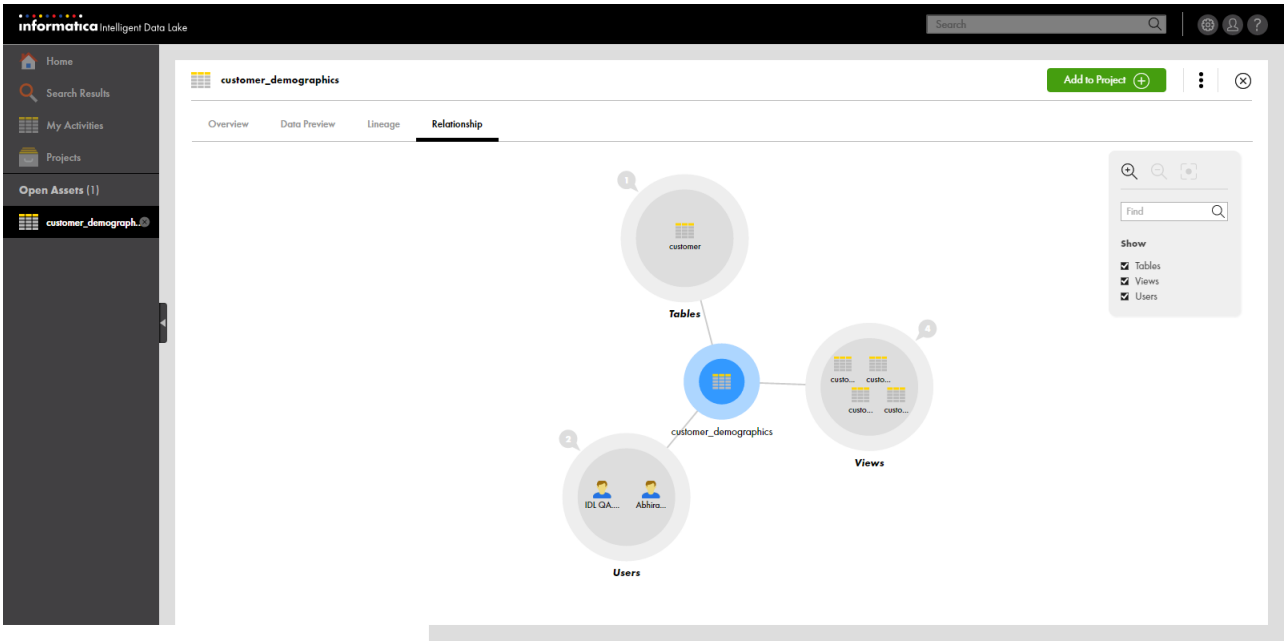
Click the number with a plus sign (+) in the origin or the end point to view all the data assets included.

Click any of the data assets listed on the lineage diagram to view a summary of the details for that data asset.

Relationship View

The **Relationship** view displays the relationship in a diagram that shows how the selected data asset is related to the other assets. You can view the relationships between the selected data asset and other data assets in the catalog.

The following image shows the **Relationship** view:



Based on the type of data asset, you can use the following icons to view the relationship between the selected data asset and other data assets:

- Use the **Zoom in** ( ) icon to increase the magnification and the **Zoom out** icon ( ) to decrease the magnification.
- Click the **Show** option to toggle between displaying and hiding the related data assets.
  This option changes based on the data asset that you selected. For example, if you selected a view that has related tables, you get the option **Tables** to show the tables related to that view. If you selected a

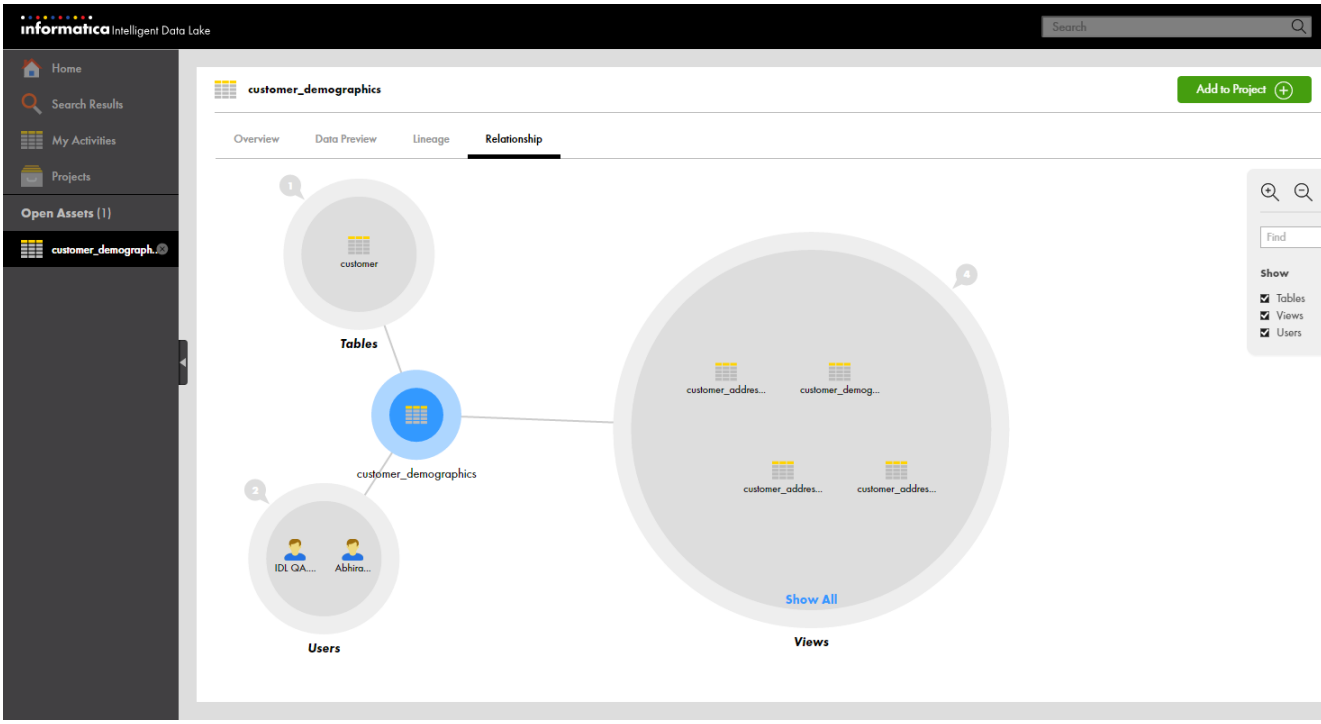
table that has related views, you get the option **Views** to view all the views associated with the selected table. Or, if you selected a table that has related users, you get the option **Users** to view all the users for the selected table.

- Click the Reset icon ( ) to reset the relationship diagram back to the initial view size.

The number at the top of a related asset in the diagram represents the number of related asset types included in the selected data asset. You can click this number to display all the related asset types included in the data asset. For example, in the image the number 4 on top of the related asset views represents the number of views included in the selected data asset.

You can click a related data asset to view a summary of the details for that data asset.

You can click the number of related data assets to display all the related data assets included in the table. For example, the following image shows that when you click on number 4 on top of the related views, the four views related to the data asset are expanded.



Click **Show All** to view all the related data assets included in the selected data asset.

Copying a Data Asset

You can copy a data asset stored in the lake to another Hive table in the lake.

1. On any data asset view, click the **Manage Data Assets** icon ( ) and then click **Copy**.
   The **Copy** dialog box appears.

2. Use one of the following options to copy the data asset:
   - New: Selected by default. Use this option if the table is completely new for the selected schema.
   - Append: Use this option if the entered table name is same as the existing table name in the schema and if the schema of this table is same as the existing table schema.
   - Overwrite: Use this option if the entered table name is same as the existing table name in the schema and if the schema of this table is different than the existing table schema.
3. Select the Hive database.

   The Hive database menu shows the data lake name.
4. Optionally, enter a name for the Hive table

   By default, Intelligent Data Lake appends the string "copy" to the name of the data asset.
5. Click **OK**.

Deleting a Data Asset

You can delete a data asset stored in the lake. You need to check before deleting data assets as they can be used by other users, projects, and publications.

1. On any data asset view, click the **Manage Data Assets** icon () and then click **Delete**.

   The **Delete** dialog box appears.
2. If the data asset is not part of any project, you can click **OK** to delete the project. If the data asset is in a project, the following message appears:

   This data asset is being used in 1 projects. It is recommended to go to Object Overview page to see details of which project it is used in. Click OK to delete anyway.
3. Click **Cancel** if you want to view the details of the project. If the project needs to be deleted, go to My Projects page and delete the project.

   Note: You need to check before deleting data assets and projects. Data assets can be shared by different users, projects, and publications. Check with the collaborators before deleting a data asset in a shared project.
4. Click **OK** to delete the project.

Access to Data

Based on your organization's security policies, you might not have access to a data asset stored in the data lake.

If you find a data asset in the search results but find that you cannot view the data asset details, you might not have the permission to view the data asset. You must contact your administrator to request access to the data asset

CHAPTER 4

Projects

This chapter includes the following topics:

Overview

A project is a container for organizing the worksheets and input sources that you use to prepare data. You can add data assets that are stored in the data lake as Hive tables to a project. After you publish a data asset to the data lake, the publication appears in the project.

Intelligent Data Lake adds data assets to projects as worksheets. By definition, projects are private and can only be viewed by the user who creates the project.

You can share projects with other analysts to collaborate on activities related to the project. You can also edit a project to change the project properties or delete a redundant project. When you edit a shared project, the project is locked for editing by the user who edits the project. In a project, you can also get recommendations for data assets based on the data assets added to the project.

You can access projects from the Home page or My Projects page. The projects in the **Projects Overview** section appear as cards by default. The project card shows the following details:

- Owner: Displays the name of the owner of the project.
- Last Updated: Displays the name of the user who updated the project recently and the date and time at which the project has been updated.
- Input Sources: Displays the number of input sources for this project.
- Worksheets: Displays the number of worksheets in this project.
- Publications: Displays the number of publications in this project.

Click the list icon () to view the projects in list view. Click the cards icon () to view the projects as cards.

Worksheets

When you add a data asset to a project, the data asset becomes a worksheet within the project.

When you open a project, you can perform the following tasks for a worksheet in the project:

- To view a worksheet, select the worksheet and click **Prepare** > **View**.
- To edit a worksheet, select a worksheet and click **Prepare** > **Prepare**.

The worksheet opens in a spreadsheet interface.

The worksheets list displays the following details about each worksheet:

- Name: Displays the name of the worksheet.
- Status: Displays the status of the worksheet.
  - New: The worksheet is newly added to the project.
  - Ready: The worksheet is open for data preparation.
  - Work in progress: The data in the worksheet has been prepared and the worksheet is ready to be published.
  - Published: The worksheet is prepared and published.
  - Prepare Error: The data preparation failed due to errors such as data type mismatch and so on.
  - Publish Error: The publishing failed due to errors such as division by zero.
- Input Sources: Displays the input sources of the worksheet. If the worksheet is prepared from other worksheets or operations, this column provides the details.
- Publication: Displays the name of the publication if the worksheet is already published.

The publishing icon () is enabled after you edit the worksheet and prepare the data. Click the () icon or () icon to publish the worksheet. You can also use this () icon to view logs and export a worksheet that is published earlier.

Creating a Project

Create a project to organize the worksheets and input sources that you use to prepare data. You can add data assets that are stored in the data lake to any project.

1. You can create a project in the following ways:
   - On the **Home** page, click **Create New**.
   - On the **My Projects** page, click **Create New**.
   - To create a project from the **Search Results** or **Open Assets** view, select a data asset and click **Add to Project**. Then click **Create New**.

The **Add to Project** dialog box appears.

2. Enter a name and an optional description.
3. Click **Finish**.

Adding a Data Asset to a Project

Add a data asset to a project to prepare the data asset. Intelligent Data Lake adds the data asset as a worksheet to the project.

When you view the details of a data asset, you can add the data asset to a project. On the **Search Results** view or the **Open Assets** view, click the name of a data asset.

1. On the data asset view, click the **Add to Project** icon ( ).

   The **Add to Project** dialog box appears.
2. Add the data asset to an existing project or to a new project.
   - To add to an existing project, select a project under **Recent Projects** or **All Projects**.
   - To add to a new project, click **Create New** and enter a name and optional description for the project.
3. Click **OK**.

   The data asset appears as a worksheet in the selected project.

Recommendations

Intelligent Data Lake makes recommendations based on the data assets added to the project and enables you to determine how best to make use of the recommendations.

You can view recommendations for data assets used in different projects by trusted people in your organization or for data assets that share similar columns. Intelligent Data Lake makes the following types of recommendations:

- Additional recommendations. Data assets that are used in different projects by different users.
- Alternate recommendations. Data assets that share similar columns.

Editing a Project

Edit a project to change the name and description of the project.

To edit a project, you must have edit permissions.

1. In the **Projects** view, select a project.
   - Click the manage projects icon ( ) and select **Edit Properties**.

- Click the edit project icon () on the Overview tab.

The **Edit Project** dialog box appears.

2. Edit the name or description.
3. Click **Save**.

Sharing a Project

Share a project with other users to collaborate on the data assets in the project. When you share a project with a user, you can also assign a role to the user.

1. In the **Projects** view, select a project.
2. On the Project Details page, you can perform one of the following operations:
   - On the Collaborators panel, click the add icon ().
   - Click the manage projects icon () and select **Share Project**.

   The **Share Project** dialog box appears.
3. Enter the full name of the user that you want to add as a collaborator on the project. If you enter the login name, the search will not yield any results.

   Intelligent Data Lake displays a list of names that match the name you enter.
4. Select the user name from the list.
5. In the list, choose to assign a permission to the user to enable the user to view, edit, or become a project co-owner:
   - Assign the **View** permission to enable the user to view the project. The user can view the project and cannot edit, share, publish, or prepare the worksheet.
   - Assign the **Edit** permission to enable the user to view and edit the project. The user can edit the properties, view data assets, prepare data assets, and publish data assets.
   - Assign the **Co-owner** permission to enable the user to view, edit, and add data assets or work on the worksheets in the project. The user can also edit properties, share the project, prepare, and publish. Co-owners can share projects with other users and give view and edit permissions to other users. Co-owners cannot assign the co-owner permission to other users.

   Collaborators cannot change the owner of the project, add another co-owner to the project, or delete the project. The owner of the project can perform all these operations. However, the owner cannot unlock a project. Domain administrators and users with IDL administrator privileges can perform all the operations in projects including unlocking the projects.
6. Click **OK**.

Changing the Project Owner

Change the project ownership from one user to another user. Project owners, administrators, and users with administrator privileges can change the project ownership.

1. In the **Projects** view, select a project.
2. Click the manage projects icon () and select **Change Owner**.

   The **Change Owner** dialog box appears.
3. Enter the full name of the new owner. If you enter the login name, the search will not yield any results.

   Intelligent Data Lake displays a list of names that match the name you enter.
4. Select the user name from the list.
5. Click **OK**.

Object Missing

This object is not available in the repository.

Deleting a Project

You can delete projects. Project owners and administrators have the permission to delete projects.

1. In the **Projects** view, select a project.
2. Click the manage projects icon () and select **Delete**.

   The **Delete Project** dialog box appears.
3. Click **Yes**.

CHAPTER 5

Data Preparation

This chapter includes the following topics:

Overview

Data preparation is the process of combining, cleansing, transforming, and structuring data from one or more data assets so that it is ready for analysis.

When you add a data asset to a project, Intelligent Data Lake creates a corresponding worksheet in the project. A worksheet has an interactive data-driven spreadsheet interface. The worksheet contains the data that you prepare and a recipe that tracks the changes you make to the data as you prepare it. Intelligent Data Lake loads sample data into the worksheet depending on the pre-configured sample size. You can view the total number of actual rows of the data asset in the worksheet metadata.

When you prepare data, you use the sample data loaded in the worksheet and all operations are performed on this sample data. You do not directly change the data in the input source. When you publish the prepared data, Intelligent Data Lake applies the recipe to the data in the input source and creates a new data asset.

Suggestions and Previews

When you use a worksheet to prepare your data, you can get more information relevant to the data asset with suggestions and preview capabilities of the worksheet interface.

Intelligent Data Lake application infers String/Text data types for columns depending the content in the sample data. If all cells in a column have numeric values, the application infers this column as numbers and allows you to perform various numeric operations on this column. You can choose to change the data type of the selected column in the application if you find the inferred data type to be incorrect.

When you click a column in the worksheet, Intelligent Data Lake analyzes the data and makes suggestions on how you can improve or manipulate the data. Intelligent Data Lake displays the data analysis and suggestions in panels at the bottom of the worksheet.

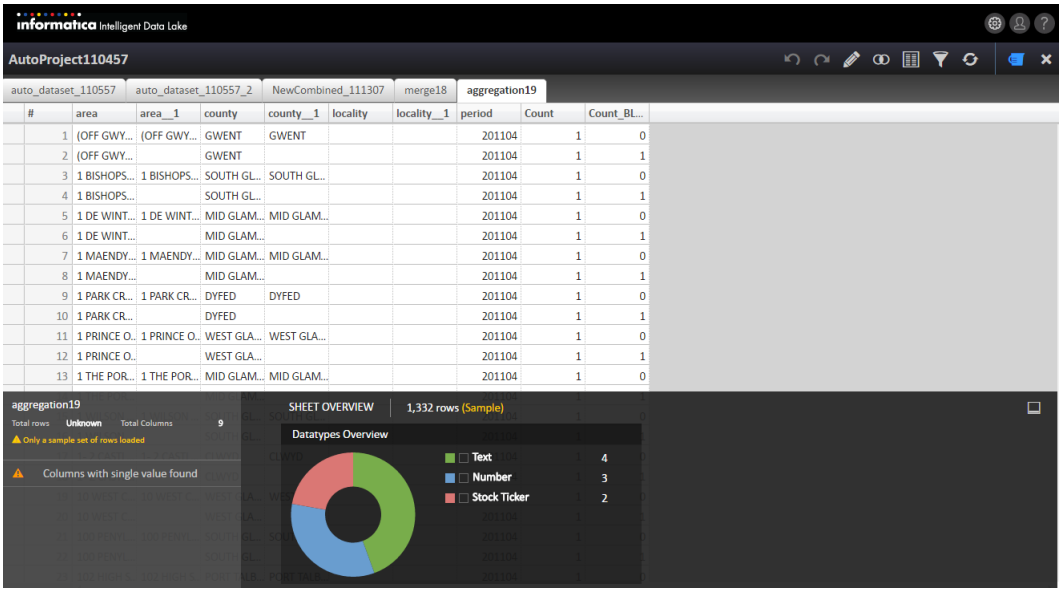
For example, Sarah, a data analyst wants to prepare a data asset that contains an email blast list that she can use to send emails to customers. She decides to click on a column in the worksheet. Intelligent Data Lake displays information about this column in an **Overview** panel that displays the column details. The **Overview** panel shows that the column contains email addresses and displays column details such as blank or duplicate values. Intelligent Data Lake also displays suggestions about how to manipulate the data in a **Suggestions** panel. For example, the **Suggestions** panel contains a suggestion to extract the domain name from an email. When Sarah hovers over a suggestion, Intelligent Data Lake shows her a preview of the data manipulation. Since Sarah would like to know where these emails go so she chooses the suggestion to extract the domain name from an email. Now, she can see how many of her customers use each domain.

Note: Intelligent Data Lake performs all operations and provides suggestions based only on the sample data included in the worksheeet.

Recipes and General Features

When you choose to view or prepare a worksheet in a project, the worksheet opens in a spreadsheet interface. The tasks to manipulate the data get added to a recipe as data preparation steps. The ingredients in a recipe are the list of input sources for the worksheet. You can go back to a previous step in a recipe to start over.

The following image shows the worksheet interface:



Let's explore the tasks that you can perform in a worksheet:

Edit the worksheet.

Use the edit icon ( ) to edit the columns in the worksheet. You can rename the worksheet. You can choose to delete blank columns or columns with one value, or hide or unhide columns in the worksheet. You can also right-click the header of a worksheet to rename, delete, or copy the worksheet. You can

perform operations to change the case of the column data, trim the column data, and split the column data. You can also right-click the columns to perform these operations on column data.

Combine worksheets.

Use the blend icon () to combine worksheets. You can perform operations such as inner join, left join, and right join on the worksheets. You can merge worksheets into one worksheet. Use this operation to unite worksheets and columns with matching data type. You can also view the data sources in the worksheet.

Aggregate the data in the worksheet.

Use the summarize icon () to aggregate the data by columns in the worksheet. You can aggregate the data on one or more keys and perform mathematical operations like sum or average on aggregated columns.

Filter the data in the worksheet.

Use the filter icon () to filter the data on any column in the worksheet. You can choose to add a filter or modify a filter condition. You can also clear all filters. When you filter your data, you select a column to filter and choose to search for a value in a column or create a custom filter and select your filter conditions. Different filters are available for String, Date, and Number columns depending on the column selected. You can filter using the top menu or you can use the filter panel available at the bottom or the window.

Apply or create formulas for data in the worksheet.

Use the formula icon () to create a column and apply your own formula using the Text, Numeric, Date, or other functions available.

Refresh the data in the worksheet.

Use the refresh icon () to update the data in the worksheet after you perform your data preparation tasks or import a new data asset . If you refresh a worksheet, the data from the Hive table or storage is refreshed. If you refresh a joined, merged, or aggregated worksheet, the data is refreshed from the parent worksheets if there is any change in the parent worksheets.

- Use the revert icon to revert the data to its previous state before making the refresh.
- Click **Highlight Changes** to view the changes that are made during the refresh.
- Click **Refresh Summary** to view the description of the changes made as part of the refresh in the form of instructions.

View your data preparation steps.

Use the recipe icon () to view your data preparation steps as you perform tasks on the data in the worksheet.

Go back to a previous step or move forward a step in a recipe.

Use the undo change recipe icon () to go back to a previous step, or use the redo icon () to perform a step again.

Sort the data.

These changes are for visualization purposes and will not persist during publishing.

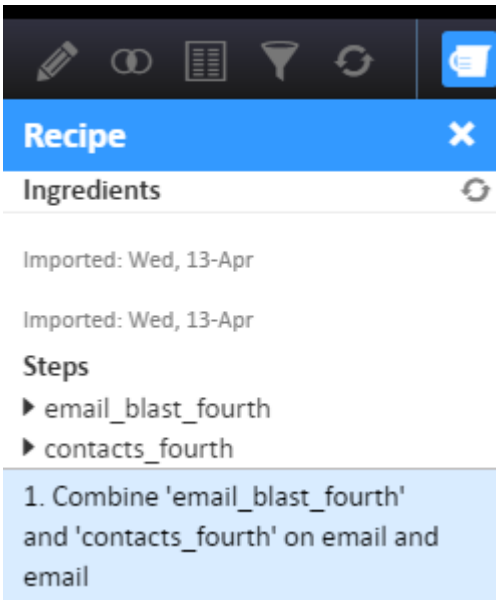
Close the worksheet.

Close the worksheet to return to the **Projects** view. Use the close icon ( ) to close the worksheet and return to the **Projects** view.

Ingredients

In the Intelligent Data Lake application, the ingredients are the list of input sources in a worksheet.

The following image shows two ingredients in a recipe, which contains one step to combine the two ingredients:



Steps

In a recipe, you can go back to the previous step and start over. Use the icons on the worksheet to perform these steps.

To go back to a previous step in a recipe, use the **Undo last step in recipe** ( ).

To perform a step in the recipe again, use the **redo** icon ( ).

To remove a step from a recipe, click the remove icon ( ) in the recipe steps.

Data Blending

You can blend the data by combining worksheets, merging worksheets into one worksheet, or viewing the data sources in the worksheet. Use the blend icon ( ) to blend the data in your worksheet. As you perform

data blending, Intelligent Data Lake displays suggestions and previews relevant to the data asset in a panel at the bottom of the worksheet.

For example, Sarah, a data analyst, imports multiple data assets to an email blast list project. The email blast list data asset does not contain contact information. Therefore, she wants to overlay it with the contacts list data asset so that all the information is in one email blast list data asset. Sarah uses the blend icon to combine the email blast list data asset with the contacts list data asset.

In the preview of the combined worksheet, Sarah can view the columns from the email blast list data asset, columns from the contacts list data asset, and a key column that Intelligent Data Lake uses to combine the worksheets and generate a new worksheet.

Sarah can choose to join worksheets using the Intelligent Data Lake suggested columns or change the join key to use some other column to join. She can use any column with matching data type for joining purposes. Sarah can edit the text box at the left side of the panel to rename the worksheet with a desired name.

In a panel at the bottom of the worksheet, Sarah can also view details about the columns that Intelligent Data Lake uses to make the key from key columns, how much overlap there is between key columns, and which rows overlap. When Sarah determines that the preview of the combined worksheet looks correct and clicks **Combine**, Intelligent Data Lake creates a combined worksheet. Sarah can perform other data blending operations such as viewing the data sources in the combined worksheet.

Data Aggregation

You can aggregate data by grouping data by column content or reporting data about each grouped by column in your worksheet. Use the summarize icon () to aggregate the data by columns in the worksheet. As you perform the data aggregation, Intelligent Data Lake displays suggestions and previews relevant to the data asset in a panel at the bottom of the worksheet.

For example, Sarah, a data analyst wants to aggregate data in an email blast list data asset that contains a list of users that have a subscription with her telecom company. She wants to know how the users ages vary across geographic regions. Sarah uses the summarize icon to aggregate the worksheet by geographic regions. In order to group and condense the rows by geographic region, she selects the geographic region column as the column to aggregate by. Since there are 32 geographic regions, the aggregation returns 32 rows.

In a panel at the bottom of the worksheet, Sarah can also view suggestions for columns to use with the aggregation to get more relevant information about the email blast list data asset. For example, Sarah can view how many rows the data aggregation will have if she breaks down each geographic region by a suggested column named country. When she performs an aggregation of each geographic region by gender, she notices that the aggregation has twice as many rows because each geographic region now has a male row and a female row.

Sarah can select one or more columns to aggregate the values. She can also perform operations such as Sum and Average on the values in the data asset.

Intelligent Data Lake also shows the count of the number of rows that have been condensed into each category. She notices that there are 600 females and 619 males in the geographic regions. When Sarah determines that the preview of the data aggregation looks correct, she clicks **Aggregate** to get a report of users in a geographical region by gender.

Formulas

When you use a worksheet to prepare your data, you can use formulas that are included in the data-driven spreadsheet interface or create your own formulas to manipulate your data. As you apply or create your formulas, Intelligent Data Lake displays suggestions and previews relevant to the data asset in a panel at the bottom of the worksheet.

You can apply several common formulas to your worksheet. For example, Sarah, a data analyst wants to extract the area code from phone numbers in the email blast list data asset. Sarah selects the phone number column, right-clicks the column, and selects **Extract** to extract a formula. She can choose to extract a column that gets added to the left, middle, or right of the phone number column. Intelligent Data Lake applies a formula based on the contents of the phone number column. Sarah enters `Area Code` as the name of the column that contains the extracted formula and clicks **Done**.

You can also create your own formulas. Select a column, and click the formula icon () to insert a formula for the column. Intelligent Data Lake provides suggestions at the bottom of the worksheet to pick a formula and fill in the blanks. Click **Done** to finish adding your formula.

If you make a mistake, you can always go back to rectify it. Click the recipe icon (), find your formula and remove it from the recipe.

CHAPTER 6

Data Publication

This chapter includes the following topics:

Data Publication Overview

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

You can prepare data after adding the data assets to a project. After addition, the data assets in a project are called worksheets. The inputs to these worksheets can be Hive tables or views.

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. 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.

Publishing Prepared Data

After you prepare your data in the worksheet, you can publish the data asset in the project to apply the recipe to the input source and write it as a Hive table to the data lake.

1. In the **Projects** view, select the worksheet that you want to publish.
2. Select the () icon to publish the worksheet and then select **Publish**.

   The **Publish** dialog appears.
3. Use one of the following options to publish the data asset:
   - New: Selected by default. Use this option if the table is completely new for the selected schema.
   - Append: Use this option if the entered table name is same as the existing table name in the schema and if the schema of this table is same as the existing table schema.

- Overwrite: Use this option if the entered table name is same as the existing table name in the schema and if the schema of this table is different than the existing table schema.

4. Select the Hive schema in the data lake in which to publish the data.
5. Enter a name for the Hive table.
6. Click **OK**.

Exporting a Publication

Export a publication to make the data asset available as a CSV file. You can use the file with a third-party application to run reports on the data.

1. In the **Projects** view, select the worksheet that you want to publish.
2. You can choose to export using one of the following options:

   - Select the () icon to publish the worksheet and then select **Export Publication**.
   - Click the name of the data asset under the **Input Sources** column. On the Data Assets page, select the () icon and click **Export**.

   Intelligent Data Lake exports the publication as a CSV file that you can download.

Operationalize Mappings

If you want your prepared data to be uploaded into the data lake on a regular basis, contact your administrator to operationalize your data preparation steps.

Before you ask your administrator to operationalize your data preparation steps, you must publish your prepared data to the data lake. Intelligent Data Lake uploads the prepared data as a Hive table to the data lake. Now, you can ask your administrator to create an operationalized mapping based on this Hive table.

Your administrator can import the Hive table into a mapping in Informatica Developer. The administrator can then run the mapping in the Hadoop environment to load the prepared data back into the data lake on a regular basis.

CHAPTER 7

Data Upload

This chapter includes the following topics:

Data Upload Overview

Data upload is the process of adding your own data to the data lake from a local drive. When you upload data, you create a new data asset that you can add to a project and prepare for analysis.

You can upload delimited text files to the data lake. Intelligent Data Lake stores the uploaded data as a Hive table in the data lake.

Uploading Data

You can upload a delimited text file from your local drive to the data lake.

Note: You cannot upload numeric data with a precision of more than 15 digits.

1. In the **Projects** view, click **Upload Data Assets**.

   The **Upload Data Assets** dialog box appears.
2. Browse for a file or drag a file into the dialog box.

   The name of the selected file appears on the window.
3. Click **Next**.
4. On the **Upload Data Assets (Step 2)** page, specify that code page and format of the file.

The following table describes the file properties that you specify:

Property	Description
Code Page	Character encoding used in the file. The default code page is UTF-8.
Delimiters	Character used to separate columns of data in the file. You can use multiple delimiters in the file. The default delimiter is a comma.
Text Qualifier	Character used to specify the start and end of a text string. If you select single quotes, the Intelligent Data Lake Service ignores delimiters within pairs of single quotes. If you select double quotes, the service ignores delimiters within pairs of double quotes.
Start import at row	Row number at which Intelligent Data Lake Service starts the import. The default starting import row is 1.
Import field names from first row	Select this option to use the values in the first row as names of the imported columns.

The page also displays a preview of the rows and columns to be created from the file. The page displays 100 records even if the number of records exceeds 100.

5. Click **Next**.
6. On the **Upload Data Assets (Step 3)** page, optionally click a column to configure column attributes.

   The following table describes the column properties that you specify for the columns:

Property	Description
Column Name	Name of the column. The column name cannot exceed 120 characters.
Column Type	Data type of the column.
Precision	If the column has a numeric data type, set the precision. The value of the precision must be larger than the value of the scale.
Scale	If the column has a numeric data type, set the scale. The value of the scale must be smaller than the value of the precision.
Format	If the column has a data type of DateTime, set the format.

7. Click **Next**.
8. On the **Upload Data Assets (Step 4)** page, specify Hive schema and table properties.

The following table describes the Hive schema and table properties that you specify:

Property	Description
Hive Schema	Name of the hive schema in which to upload the file.
Hive Table	Name of the table in the Hive schema in which to upload the file. The name of the table must be in lower case. The name can include letters, numbers, and the underscore character. The first character of the name cannot be an underscore or a number.
Action on Table	Action that the Intelligent Data Lake Service can perform during the import process. - To create a Hive table, select **Create**. - To append to an existing Hive table, select **Append**. You cannot append to a partitioned table. - To overwrite an existing Hive table, select **Overwrite**. The Intelligent Data Lake Service drops the existing table and creates a new table for the uploaded data. You cannot overwrite a partitioned table.

9. Click **Upload**.

APPENDIX A

Glossary

asset

An information object that is described in the catalog. Assets can include items such as a database table, report, folder, user account, or business glossary definition.

catalog

An indexed inventory of the assets in an enterprise. The assets can come from different types of enterprise systems. The catalog contains metadata about each asset, including profile statistics, asset ratings, data domains, and data relationships.

data asset

Data that you work with as a unit. A data asset is one type of asset 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. You can add data assets that are stored in the data lake as Hive tables to a project.

data lake

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 most cases, a data lake is a Hadoop cluster used for big data initiatives. When different types of data are stored in one repository, data analysts can more easily combine and transform the data to create new insights.

data preparation

The process of combining, cleansing, transforming, and structuring data from one or more data assets so that it is ready for analysis.

In Intelligent Data Lake, you use worksheets in a project to create data preparation recipes.

data publication

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. 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.

data upload

The process of adding data to the data lake from a local drive. When you upload data, you create a new data asset that you can add to a project and prepare for analysis.

You can upload delimited text files to the data lake. Intelligent Data Lake writes the uploaded data to a Hive table in the data lake.

input source

The data source for a worksheet in a project. An input source can be a Hive table in the data lake.

project

A container used to organize the worksheets and input sources that you use to prepare data.

You can add data assets that are stored in the data lake as Hive tables to a project. By definition, projects are private and can only be viewed by the user who creates the project. You can share projects with other analysts to collaborate on activities related to the project.

recipe

The list of input sources and the steps taken to prepare data in a worksheet.

When you 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.

worksheet

The component within a project where you prepare data. When you add a data asset to a project, Intelligent Data Lake creates a corresponding worksheet in the project.

A worksheet has an interactive data-driven spreadsheet interface. The worksheet contains the data that you prepare and a recipe that tracks the changes you make to the data as you prepare it. Depending on the size of the input source, Intelligent Data Lake loads sample data or all data into the worksheet. When you prepare data, you use the data loaded in the worksheet. You do not directly change the data in the input source. When you publish the prepared data, Intelligent Data Lake applies the recipe to the data in the input source and creates a new data asset.

INDEX