Informatica®
10.2.2

Enterprise Data Lake User Guide

Informatica Enterprise Data Lake User Guide
10.2.2
February 2019

Publication Date: 2019-09-03

Table of Contents

Preface

The *Enterprise Data Lake User Guide* contains information for analysts about using the Enterprise 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 provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Network

The Informatica Network is the gateway to many resources, including the Informatica Knowledge Base and Informatica Global Customer Support. To enter the Informatica Network, visit https://network.informatica.com.

As an Informatica Network member, you have the following options:

- Search the Knowledge Base for product resources.
- View product availability information.
- Create and 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 find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit https://search.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

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit https://docs.informatica.com.

Informatica maintains documentation for many products on the Informatica Knowledge Base in addition to the Documentation Portal. If you cannot find documentation for your product or product version on the Documentation Portal, search the Knowledge Base at https://search.informatica.com.

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

Informatica Product Availability Matrices

Product Availability Matrices (PAMs) indicate the versions of the operating systems, databases, and types of data sources and targets that a product release supports. You can browse the Informatica 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 and based on real-world experiences from hundreds of data management projects. Informatica Velocity represents the collective knowledge of Informatica consultants who work with organizations around the world to plan, develop, deploy, and maintain successful data management solutions.

You can find 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 extend and enhance your Informatica implementations. Leverage any of the hundreds of solutions from Informatica developers and partners on the Marketplace to improve your productivity and speed up time to implementation on your projects. You can find the Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

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

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
https://www.informatica.com/services-and-training/customer-success-services/contact-us.html.

To find online support resources on the Informatica Network, visit https://network.informatica.com and select the eSupport option.

CHAPTER 1

Introduction to Enterprise Data Lake

This chapter includes the following topics:

Enterprise 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.

Enterprise Data Lake helps you derive more value from your 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.

Enterprise 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 Enterprise Data Lake, analysts to spend more time on analysis and less time on finding and preparing data.

Enterprise 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.

Enterprise Data Lake Concepts

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

Data Lake

The data lake used by Enterprise 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.

The data lake can be deployed on-premise or in the cloud. The data lake utilizes Hive, the Spark engine, and a Hadoop-compatible file system. The data lake must be collocated with the Informatica application services associated with Enterprise Data Lake. For example, if the data lake is deployed in Amazon EMR, you must also deploy Enterprise Data Lake, Enterprise Data Catalog, and the Informatica services in Amazon EMR.

Enterprise Data Lake works with the security mechanism used by the cluster, including Kerberos, Apache Ranger, and Apache Sentry, to securely access data in the data lake.

You can upload assets such as comma-separated value files as Hive tables in the data lake using Enterprise Data Lake.

You can also ingest data from external sources outside the data lake using Hadoop tools or Informatica Mass Ingestion.

You can prepare a Hive or file-based asset that exists in Enterprise Data Catalog. When you a prepare file-based asset, Enterprise Data Lake creates an external temporary table in the Hive schema specified in the Enterprise Data Lake Service.

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

Data Asset

A data asset is source data that you work with as a unit. A data asset can include items such as a flat file, a table, or a table view. A data asset can include data stored in or outside the data lake, such as in external databases.

You use Enterprise Data Lake to search for assets described in the Enterprise Data Catalog, which serves as a centralized repository that stores the metadata for assets extracted from various external sources. Enterprise Data Catalog contains metadata such as profile statistics, asset ratings, data domains, and data relationships for each asset. When you search for assets in Enterprise Data Lake, you actually search the metadata in Enterprise Data Catalog.

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

Projects, Worksheets and Recipes

A project is a container that stores data assets and worksheets.

When you add a data asset to a project, Enterprise Data Lake creates a corresponding worksheet that contains a sample of the data in a spreadsheet-like presentation. You use the worksheet to preview the data and to prepare it for publication. You do not directly change the data in the input source.

When you publish the prepared data, Enterprise Data Lake applies the recipe to the data in the input source and creates a new data asset.

Enterprise Data Lake captures each step you complete in the worksheet in the recipe for the worksheet. When you publish the prepared data, Enterprise Data Lake applies the recipe to transform the data.

Data Publication

Data publication is the process of making data that you prepared in a worksheet available in the data lake.

When you publish prepared data, Enterprise 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 use a third-party business intelligence or advanced analytic tool to run reports to further analyze the published data.

Data Visualization

Data visualization is the process of assessing and validating published data using ad-hoc queries on the data to generate charts, tables, and other visual formats.

You can run queries and use dynamic forms to view the data in the form of graphs, charts, and other visual formats to assess its relevance for your business purposes. You can only run visualization on worksheets that are prepared and published. Collaborators on the project can access the notebook of the worksheet. You can share the notebook so that all users with the required permissions can view the visualization of the worksheet.

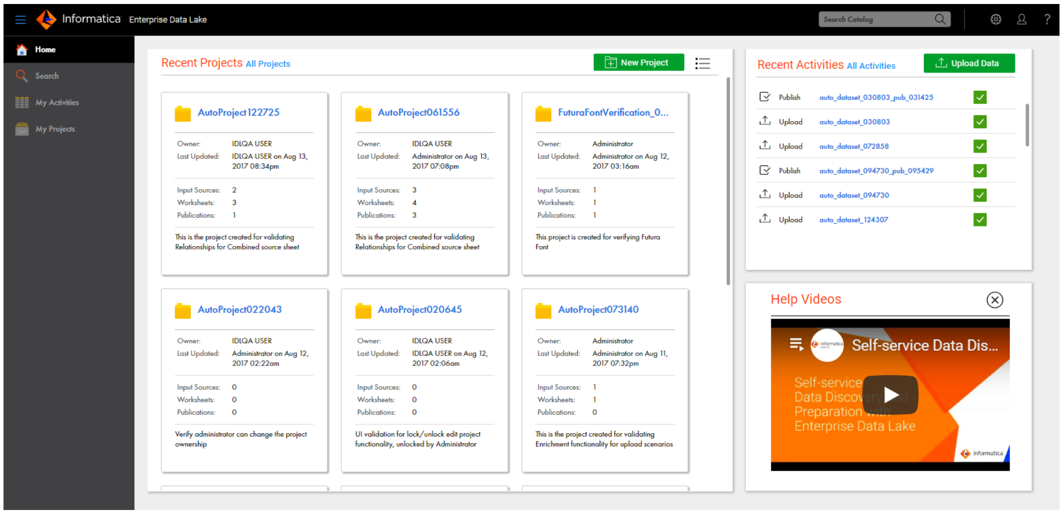
User Interface

The Enterprise Data Lake application provides you with a web-based user interface that you use to search for, discover, and prepare data assets that you publish to the data lake.

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 Enterprise Data Lake application.

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



AutoProject122725
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 08:34pm
Input Sources: 2
Worksheets: 3
Publications: 1
This is the project created for validating Relationships for Combined source sheet

AutoProject061556
Owner: IDLQA USER
Last Updated: Administrator on Aug 13, 2017 07:08pm
Input Sources: 3
Worksheets: 4
Publications: 3
This is the project created for validating Relationships for Combined source sheet

FuturaFontVerification_0...
Owner: Administrator
Last Updated: Administrator on Aug 12, 2017 03:16am
Input Sources: 1
Worksheets: 1
Publications: 1
This project is created for verifying Futura Font

AutoProject022043
Owner: IDLQA USER
Last Updated: Administrator on Aug 12, 2017 02:22am
Input Sources: 0
Worksheets: 0
Publications: 0
Verify administrator can change the project ownership

AutoProject020645
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 12, 2017 02:06am
Input Sources: 0
Worksheets: 0
Publications: 0
UI validation for lock/unlock edit project functionality, unlocked by Administrator

AutoProject073140
Owner: Administrator
Last Updated: Administrator on Aug 11, 2017 07:32pm
Input Sources: 1
Worksheets: 1
Publications: 0
This is the project created for validating Enrichment functionality for upload scenarios

Publish	auto_dataset_030803_pub_031425	✓
Upload	auto_dataset_030803	✓
Upload	auto_dataset_072858	✓
Publish	auto_dataset_094730_pub_095429	✓
Upload	auto_dataset_094730	✓
Upload	auto_dataset_124307	✓

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

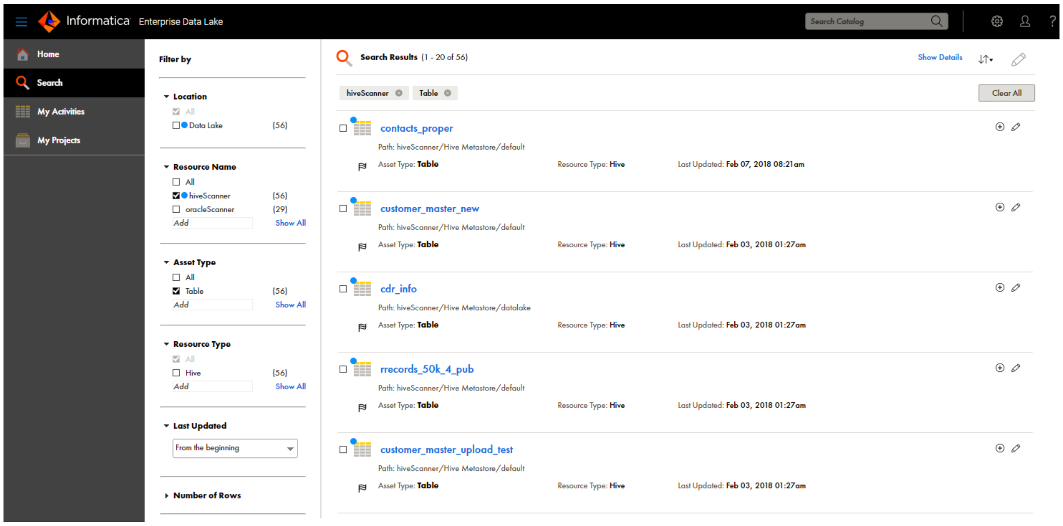
- View and create projects. The **Recent Projects** panel displays projects with the latest projects on top. To see a list of projects, click the **Switch to List View** icon. If you want to see all projects, click **All Projects**. To create a new project, click **New Project**.
- View recent activities.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

Search Results View

The **Search Results** view displays the results of your Enterprise Data Catalog search.

The application displays the total number of data assets in the catalog along with the number of resources that match the search.

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



Search Results (1 - 20 of 56)

contacts_proper
Path: hiveScanner/Hive Metastore/default
Asset Type: **Table** Resource Type: **Hive** Last Updated: **Feb 07, 2018 08:21am**

customer_master_new
Path: hiveScanner/Hive Metastore/default
Asset Type: **Table** Resource Type: **Hive** Last Updated: **Feb 03, 2018 01:27am**

cdr_info
Path: hiveScanner/Hive Metastore/datalake
Asset Type: **Table** Resource Type: **Hive** Last Updated: **Feb 03, 2018 01:27am**

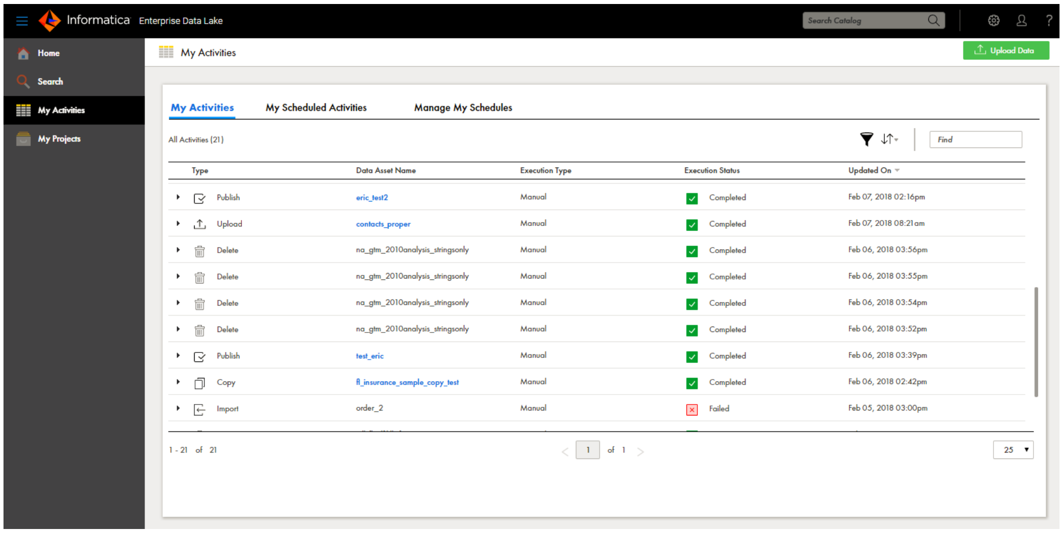
rrecords_50k_4_pub
Path: hiveScanner/Hive Metastore/default
Asset Type: **Table** Resource Type: **Hive** Last Updated: **Feb 03, 2018 01:27am**

customer_master_upload_test
Path: hiveScanner/Hive Metastore/default
Asset Type: **Table** Resource Type: **Hive** Last Updated: **Feb 03, 2018 01:27am**

My Activities View

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

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



All Activities (21)

Type	Data Asset Name	Execution Type	Execution Status	Updated On
Publish	eric_test2	Manual	✓ Completed	Feb 07, 2018 02:16pm
Upload	contacts_proper	Manual	✓ Completed	Feb 07, 2018 08:21am
Delete	na_gtm_2010analysis_stringsonly	Manual	✓ Completed	Feb 06, 2018 03:56pm
Delete	na_gtm_2010analysis_stringsonly	Manual	✓ Completed	Feb 06, 2018 03:55pm
Delete	na_gtm_2010analysis_stringsonly	Manual	✓ Completed	Feb 06, 2018 03:54pm
Delete	na_gtm_2010analysis_stringsonly	Manual	✓ Completed	Feb 06, 2018 03:52pm
Publish	test_eric	Manual	✓ Completed	Feb 06, 2018 03:39pm
Copy	fl_insurance_sample_copy_test	Manual	✓ Completed	Feb 06, 2018 02:42pm
Import	order_2	Manual	☒ Failed	Feb 05, 2018 03:00pm

1 - 21 of 21 1 of 1 25

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** to upload a data asset to the data lake.
- View your data assets on which an activity have been performed. The **My Activities** view displays all the data assets on which you performed any of the following activities - copying, deleting, downloading, exporting, importing, publishing, and uploading.

- Filter your list of data assets. Click the **filter** icon to filter data assets by type of activity, time frame and status.
  - Select **Type** to filter by all, copy, delete, download, export, import, publish, and upload types of activities.
  - Select **Updated On** to filter by all, last 24 hours, last week, or last month. You can also select a custom date range.
  - Select **Status** to filter by all, completed, completed (with warnings), in progress, or failed.
- Sort your list of data assets. Click the **sort** icon to sort data assets.
  - Select **Type** to sort by all, copy, delete, download, export, import, publish, and upload types of activities.
  - Select **Updated On** to sort by the time when they were last updated.
  - Select **Name** to sort the data asset names alphabetically.
- Search for a data asset. Use the search box to search for a data asset on which any of these activities have been performed.

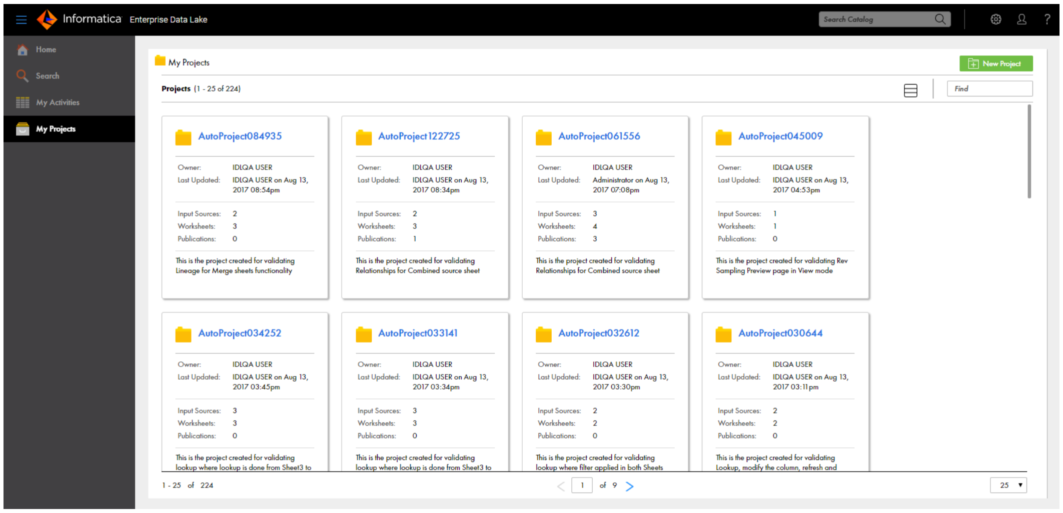
You can manage the schedules used to import, export, or publish data assets on a recurring basis from the **My Activities** view. For more information about scheduling import, export, or publish activities, see Chapter 7, “Publish Data” on page 91.

- Click **My Scheduled Activities** page to delete or suspend a scheduled activity, or to select a new schedule to use for an activity.
- Click **Manage Schedules** to edit, delete or suspend schedules that you own. You can also create a new schedule from the page.

Projects View

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

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



Projects (1 - 25 of 224)

AutoProject084935
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 08:54pm
Input Sources: 2
Worksheets: 3
Publications: 0
This is the project created for validating Lineage for Merge sheets functionality

AutoProject122725
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 08:34pm
Input Sources: 2
Worksheets: 3
Publications: 1
This is the project created for validating Relationships for Combined source sheet

AutoProject061556
Owner: IDLQA USER
Last Updated: Administrator on Aug 13, 2017 07:08pm
Input Sources: 3
Worksheets: 4
Publications: 3
This is the project created for validating Relationships for Combined source sheet

AutoProject045009
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 04:53pm
Input Sources: 1
Worksheets: 1
Publications: 0
This is the project created for validating Rev Sampling Preview page in View mode

AutoProject034252
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 03:45pm
Input Sources: 3
Worksheets: 3
Publications: 0
This is the project created for validating lookup where lookup is done from Sheet3 to

AutoProject033141
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 03:34pm
Input Sources: 3
Worksheets: 3
Publications: 0
This is the project created for validating lookup where lookup is done from Sheet3 to

AutoProject032612
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 03:30pm
Input Sources: 2
Worksheets: 2
Publications: 0
This is the project created for validating lookup where filter applied in both Sheets

AutoProject030644
Owner: IDLQA USER
Last Updated: IDLQA USER on Aug 13, 2017 03:11pm
Input Sources: 2
Worksheets: 2
Publications: 0
This is the project created for validating Lookup, modify the column, refresh and

1 - 25 of 224
1 of 9
25

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

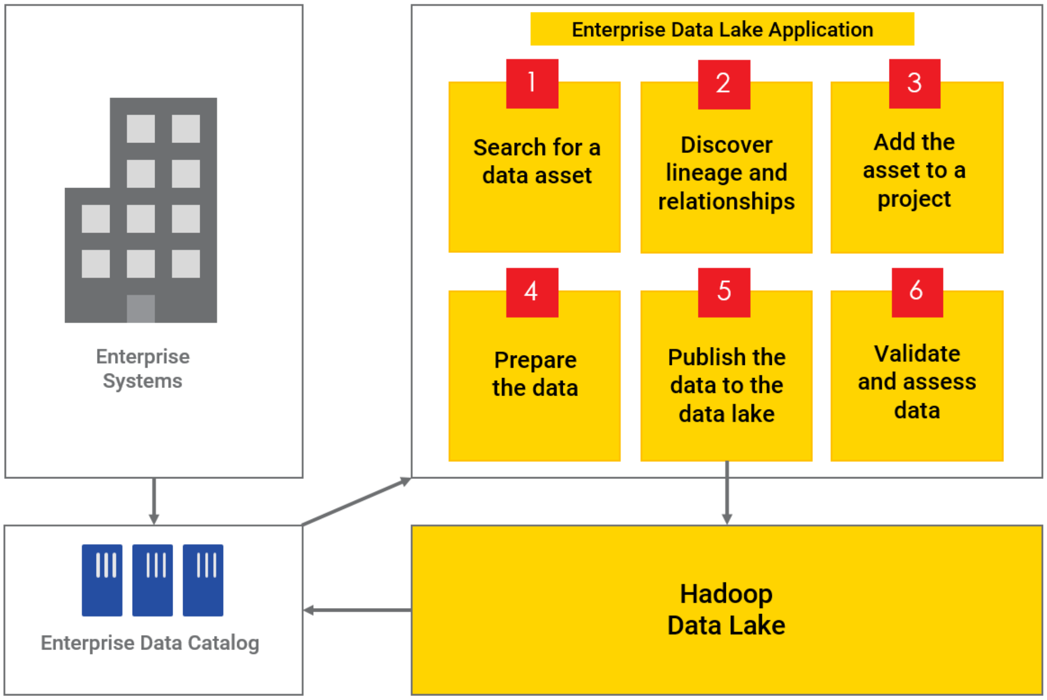
- Create a project. Click **New Project** to create a project.
- Switch the views on a project. By default, the Enterprise Data Lake application displays each project in a card view. Click the **Switch to List View** icon to switch to a list view. To return to card view, click the **Switch to Card View** 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 Properties** icon in the **Overview** panel to edit these project properties.
  - Change the collaborators on the project. Click **Add a Collaborator** 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** to load all worksheets in the project You can also select a worksheet in a project and click the **Publish** icon to publish the worksheet. You can also view logs by clicking the **Publish** icon.
  - View recommendations. View recommendations for data assets used in different projects by trusted people in the organization on the **Recommendations** panel.
- Upload a data asset. Click **Upload Data** to upload a data asset to the data lake.

Data Discovery and Analysis Process

Use the Enterprise Data Lake application to discover and prepare data.

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



Complete the following high-level activities in the Enterprise Data Lake application to prepare data.

Search for data assets.

Search Enterprise Data Catalog and the data lake for data assets you can use. Data assets can include items such as delimited files, tables and views, JSON Lines files, and more.

For more information, see Chapter 2, “Search for Data Assets” on page 16.

You can also move data from outside sources such as external databases into the data lake. 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.

For more information, see “Importing a Data Asset into the Data Lake” on page 26.

Examine each asset and understand how it relates to other assets.

After you discover an asset in Enterprise Data Catalog, view details about the data within the asset, the end-to-end data flow for the asset, and relationships between the asset and other data assets in the catalog.

For more information, see Chapter 3, “Discover Data” on page 19.

Add the data asset to a project.

Add the discovered data assets to a project, which is a container for organizing the worksheets and input sources that you use to prepare the data.

For more information, see Chapter 5, “Create and Manage Projects” on page 30.

Prepare the data for analysis.

Prepare the data by combining, cleansing, transforming, and structuring the data so that it is ready for analysis. Enterprise Data Lake provides you with numerous options for preparing data.

For more information, see Chapter 6, “Prepare Data” on page 36.

Publish the prepared data to the data lake.

After you finish preparing the data in the asset, you publish the prepared data to the data lake, where analysts and other users can access it.

For more information, see Chapter 7, “Publish Data” on page 91.

You can schedule publication of the prepared data to regularly update the asset in the data lake on a recurring basis.

For more information, see Chapter 9, “Schedule Export, Import and Publish Activities” on page 97.

Assess the quality of the published data.

After you publish the data to the data lake, you can visualize the data to ensure the content and quality are sufficient for use by analysts and other users. If you determine that the data is not ready for use, you can modify the data preparation steps and republish the data.

For more information, see Chapter 8, “Visualize and Assess Published Data” on page 94.

Note: Activities and operations you perform in Enterprise Data Lake do not support high precision mode.

CHAPTER 2

Search for Data Assets

This chapter includes the following topics:

Overview

You can search for and discover any data asset described in Enterprise Data Catalog. Data assets can include items such as a flat file, table, or view.

You can search for a data asset by name, use a wildcard character in your search, and Enterprise Data Lake displays probable matches as you type the name of a required data asset. You can also search for assets such as users and data domains.

You can sort search results, filter search results, and see more information about data assets. After finding the data asset that you need, you can annotate and enrich it with custom attributes.

After you find the data asset that you're looking for, you can view the profiling details and the lineage of the data asset. You can see the relationship between the data asset and 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. Enterprise Data Lake also lists search suggestions when you enter a search string.

Wildcard Search

You can use the asterisk character (*) and the question mark character (?) to perform a wildcard search.

- Asterisk character (*). If you enter `*` in the **Search** box and click the **Search** icon, Enterprise Data Lake lists all the data assets in the catalog. You can also use the asterisk with parts of a data asset name. For example, to search for all the data assets that end with the string "_name," enter *_name in the **Search** box. Enterprise Data Lake lists all of the data assets that match the search criteria, for example, `Associate_Name`, `party_name`, and `cust_name`.
- Question mark character (?). Use the `?` character to substitute individual characters in the name of an asset. For example, if the data asset that you want to search for begins with the letters HR, is followed by two numbers that denote the year, and ends with REPORT, enter the following string in the **Search** box: `HR??REPORT`. Enterprise Data Lake lists all of the data assets that match the search criteria, for example, `HR12REPORT`, `HR13REPORT`, and `HR14REPORT`.

If you know the name of a column in a table, you can enter the column name and search for the table. The search results display all tables that have data or columns with the name as your search string.

Search Suggestions

When you type the first few letters of an asset name into the Search box, Enterprise Data Lake suggests matching data asset names.

If you type the name of a data asset incorrectly, Enterprise 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`, Enterprise 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.

You can interpret the search results in the following ways:

- The search results show a summary of the details of the data asset.
- The search results show assets that are in the lake and assets that are not yet in the lake. You can identify the assets within the lake with presence of a blue dot at the top left corner of the asset icon.
- 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.
- Data assets within the lake appear with a blue dot in the icon. Thus, you can distinguish between data assets in the lake and data assets that are not in the lake.

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 the sort icon and select **Name**.
- To sort data assets based on the relevance to the search, click the sort icon 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.

CHAPTER 3

Discover Data

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 also import data assets into the lake and export data assets from the lake.

You can add an asset that is in the data lake to a project. For more information on adding a data asset to a project, see "Adding a Data Asset to a Project" on page 31.

To prepare a data asset that is not stored in the data lake, you must import the asset into the data lake. For information about importing data assets into the data lake, see "Importing a Data Asset into the Data Lake" on page 26.

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

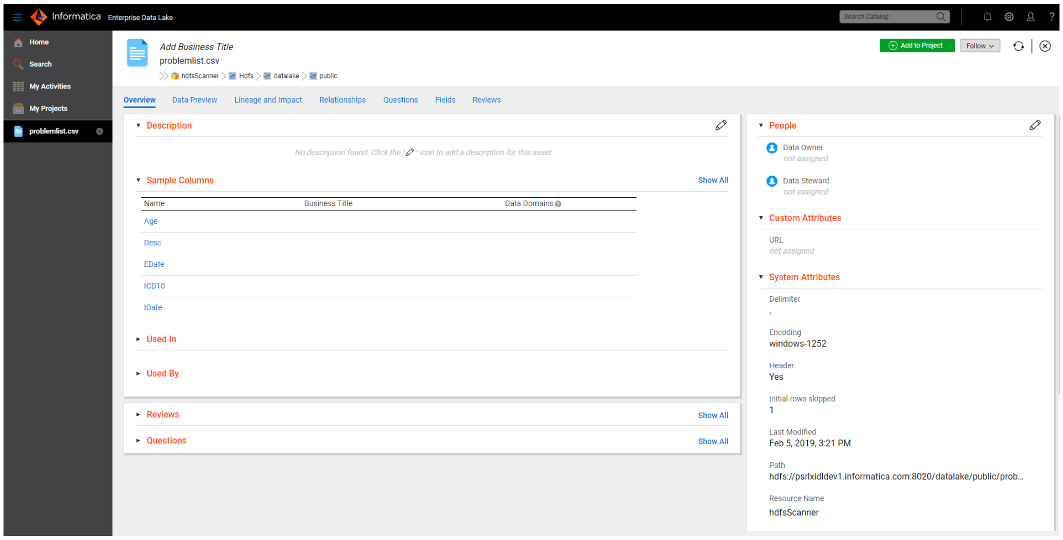
- Overview. Displays the details about the data asset. Details include the type of the data asset, the last time the data asset was updated, and the data source, along with other details.
- Data Preview. Displays a sample of the data in the data source.
- Lineage and Impact. Displays the lineage for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

- Relationships. Displays the relationships between the selected data asset and other data assets in the catalog.
- Columns or Fields. Displays details about the columns in the data asset, including the column data types and the percentage on distinct and non-distinct values. The tab displayed depends on the data asset type.

Asset View

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.

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



No description found. Click the " " icon to add a description for this asset.

Name	Business Title	Data Domains
Age		
Desc		
EDate		
ICD10		
IDate		

Data Owner
not assigned

Data Steward
not assigned

URL
not assigned

Delimiter
,

Encoding
windows-1252

Header
Yes

Initial rows skipped
1

Last Modified
Feb 5, 2019, 3:21 PM

Path
hdfs://psrlxidldev1.informatica.com:8020/datalake/public/prob...

Resource Name
hdfsScanner

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

- Click **Add to Project** to add the data asset to a project.
- Click the **Manage Data Assets** icon to edit data asset properties, copy the data asset, export the data asset, download the data asset as a CSV file or TDE file, or delete the data asset.
- View data asset details in the **Overview** tab.
- Preview data on the **Data Preview** tab.
- Discover the lineage for the flow of data for the data asset on the **Lineage and Impact** tab.
- Discover the relationships between the open data asset with other data assets on the **Relationships** tab.
- Add questions you have about the asset on the **Questions** tab.
- View details on the data in the asset on the **Columns** or **Fields** tab. The tab displayed depends on the data asset type.
- Read reviews of the asset from other users or add your own review on the **Reviews** tab.
- For assets that are published to the data lake, you can click the **Visualize** tab to visualize the data before you operationalize the recipes created during publishing.

Associating a Business Term with an Asset

You can associate a business term with any asset in the catalog that is not a business glossary asset. You can associate one business term with a data asset.

When you associate a business term with a data asset, Enterprise Data Lake displays the business term with the asset in the following places:

- In the search results next to the asset name.
- In the Asset Details view next to the asset name.
- In the Lineage and Impact diagram above the asset name when you show business terms in the diagram.

Complete the following steps to associate a business term with a data asset.

1. In the Asset Details view for the asset, click **Associate a Business Term**.

   Alternatively, you can click the Edit Properties icon for the asset in the search results.
2. In the Edit Properties dialog box, click **Business Terms**.
3. The options to associate a term vary based on whether the asset has inferred terms:
   - If one or more data domains with an associated business term are assigned to the asset, the inferred terms appear in the **Recommended Business Terms** section. To accept an inferred term, click the check mark icon next to the term.
   - If the asset has no inferred terms or you want to associate a different term with the asset, select the business glossary from the **Glossary** list. Then, select the term that you want to associate from the list of terms.

     If the list of terms is long, you can search for a term in the list by entering the first few characters of the term name in the search field.
4. Click **OK**.

Data Preview View

The **Data Preview** view displays the first 500 rows of a data asset. The **Data Preview** is also 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.

To preview data assets that are not in the lake, you must connect to the data source. Contact your Administrator for the connection details to the required source.

If the data asset preview does not show any data, click **Select Connection**. The Connections dialog box appears. Select the connection from the list. Click **OK**.

You can also filter the data during data preview for better assessment of data assets. You can add filters for multiple fields and apply combinations of such filters. Filter conditions depend on the data types. If available, you can view column value frequencies found during profiling for string values. If no value frequencies are available, you can use the free form text filtering options. Click the filter icon and select a column to filter the data asset. The filter results show the data from the entire data asset and not just the 500 rows displayed in the preview.

If Enterprise Data Lake does not support the data type of a column of the data source, the column will appear as a blank column.

Lineage View

The **Lineage and Impact** 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.

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.

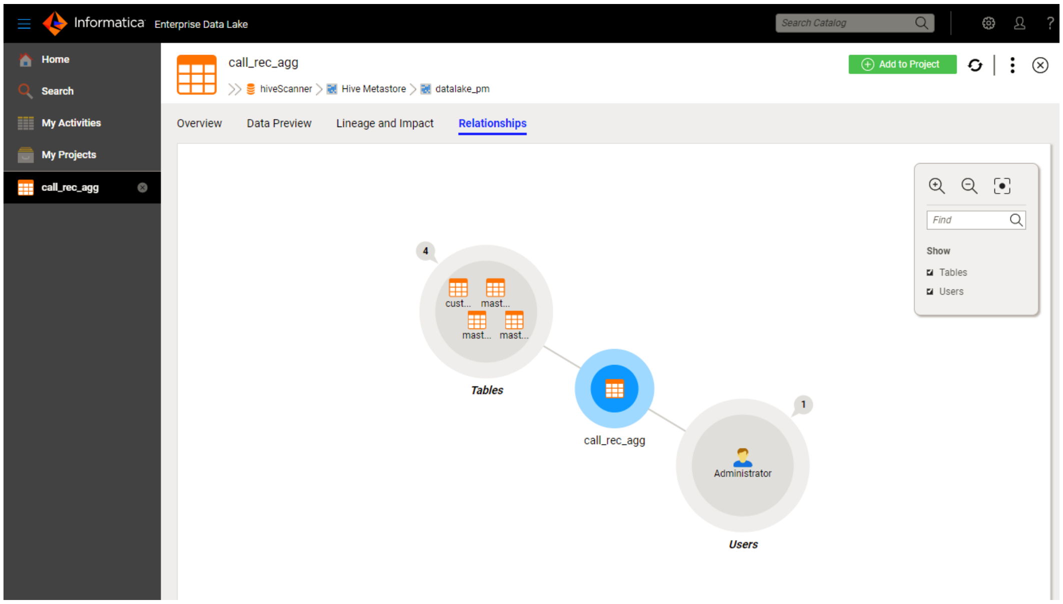
Click any of the data assets listed on the lineage diagram to view a summary of the details for that data asset.

You can also view lineage of individual columns in a table. Activities including copy, import, export, publication, and upload will add column-level lineage information in the Catalog. You can also view column-level lineage details for data preparation operations that you perform such as transformations, join, lookup, union, and aggregate. Expand the data asset and select a column to view its lineage details.

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:

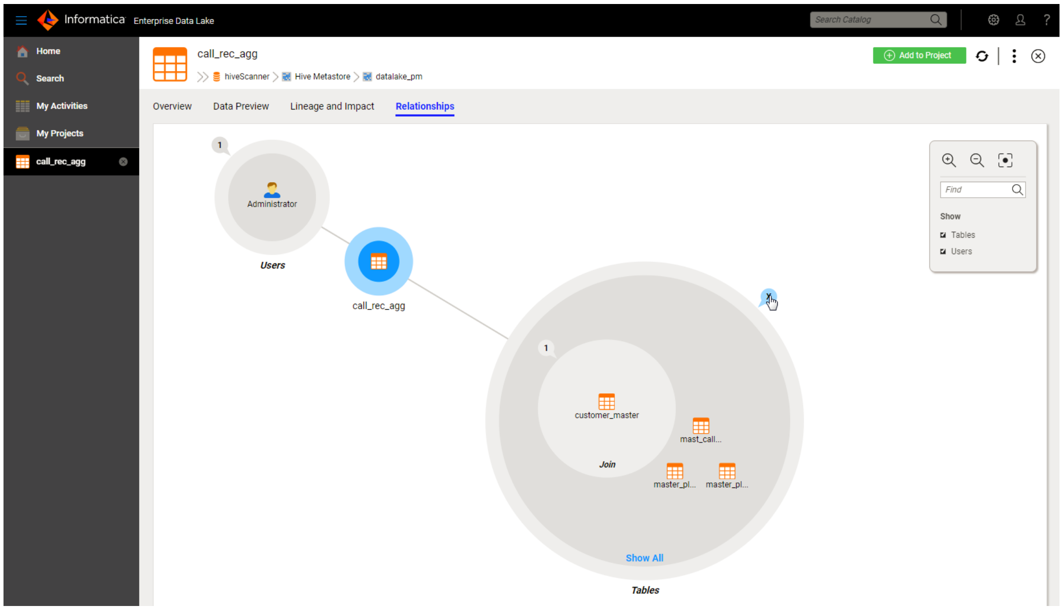


Based on the type of data asset, you can use the following icons to view the relationship between the selected data asset and other data assets:

- 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.

Click a related data asset to view a summary of the details for that data asset.

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. Click this number to display all the related asset types included in the data asset. The following image shows the impact of clicking the number 4 above Tables:



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 data lake to another Hive table in the data 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 and if the table structure is the same.
- Overwrite. Use this option if the entered table name is same as the existing table name. It drops the table and re-creates the table with the data from the copying table.

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, Enterprise Data Lake appends the string "copy" to the name of the data asset.
5. Click **OK**. Check the progress of the copy activity on My Activities page. You can see if the copy activity is in progress, completed, or if it has failed.

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. Check the progress of the delete activity on My Activities page. You can see if the deletion is in progress, completed, or if it has failed.

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

Import, Export, and Upload Data

This chapter includes the following topics:

Overview

You can import, export, and upload data assets.

You can import data from tables in external sources outside the data lake. A source from which you import data must already be cataloged in Enterprise Data Catalog.

You can import data from the following sources:

- Amazon Redshift
- Microsoft Azure SQL Database
- Microsoft SQL Server
- Oracle
- Teradata

When you import data, Enterprise Data Lake creates an import activity. Once the import activity is completed successfully, a Hive table is created and registered in the catalog. You can also check the lineage of the table.

If the asset already exists in the data lake, you can append the updated data to the asset, or overwrite the asset with the new data.

You can also export a data asset to external targets outside of the data lake. Once the export activity completes, you can verify the exported data on the external target system.

If the target already exists, you can append the updated data to the target, or overwrite the target with the new data.

You can choose to schedule import and export activities. Scheduling an activity enables you to import or export updated data assets on a recurring basis. For more information about scheduling activities, see Chapter 9, “Schedule Export, Import and Publish Activities” on page 97.

You can also upload a delimited text file from your local drive to the data lake. Enterprise Data Lake stores the uploaded data as a Hive table in the data lake.

Importing a Data Asset into the Data Lake

You can import a data asset from an external source such as an Oracle database or a Teradata database into the data lake.

You can choose to import an asset immediately, or you can schedule the import activity. Scheduling an import activity enables you to import updated data assets on a recurring basis.

For more information about scheduling an import activity by creating a schedule or using an existing schedule, see Chapter 9, "Schedule Export, Import and Publish Activities" on page 97.

1. Click **Search**, and then search for an asset you want to import into the data lake.
2. Select the asset in the search results.
3. On the Overview page for the selected asset, click **Import to Lake**.

   The **Import** dialog box appears.
4. Select the connection to use to connect to the external source, and then click **Next**.

   You must have **Read** and **Execute** permissions for the database connection you use to import a data asset.
5. Select the schema into which to import the asset from the drop down list.
6. Enter the name of the table to import. 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.
7. If a table with the same name already exists in the data lake, choose whether to append the data to the table, or to overwrite the table.
   - Select **Append** to append the data to the existing table. Make sure the column names, the data types of the columns, the precision and scale of the columns, and the sequence of the columns are identical in both the imported table and the existing table.
   - Select **Overwrite** to overwrite the existing table. Enterprise Data Lake drops the table and re-creates the table with the data from the imported table.
8. Choose whether to import the asset now, or to schedule the import activity.
   - Select **Import Now** to start the import activity.
   - Select **Schedule Import** to schedule the import activity.
     If you decide to schedule the activity, specify what to do if the asset already exists in the data lake.
     - Select **Append** to append the data to an existing table.
     - Select **Overwrite** to overwrite the existing table.
9. Click **OK**.

   You can monitor the progress of the import activity on the **My Activities** page in Enterprise Data Lake.

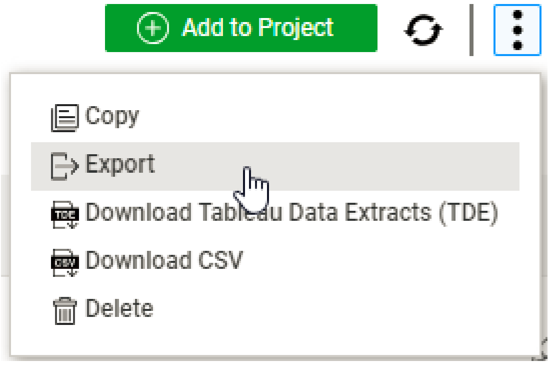
Exporting a Data Asset from the Data Lake

You can export a data asset or a publication from the data lake to external targets.

You can choose to export the asset immediately, or you can schedule the export activity. Scheduling an export activity enables you to export updated data assets on a recurring basis. For more information about scheduling activities, see Chapter 9, "Schedule Export, Import and Publish Activities" on page 97.

1. Click **Search**, and then search for the asset you want to export.
2. Click the asset in the search results.
3. On the **Overview** page for the data asset, click the **Manage Data Assets** icon, and then click **Export**.

   The following image shows the **Export** option selected from the **Manage Data Assets** icon:

   The **Export** dialog box appears.
4. Select the connection to use to connect to the target, and then click **Next**.

   You must have **Read** and **Execute** permissions for the connection you use to export a data asset.
5. Enter the name of the schema to which to export the asset.
6. Enter the name of the table to create in the target. The name can include letters, numbers, and the underscore character. The first character of the name cannot be an underscore or a number.
7. If a table with the same name already exists in the target, choose whether to append the data to the table, or to overwrite the table.
   - Select **Append** to append the data to the existing table.
   - Select **Overwrite** to overwrite the existing table.

     Choose to overwrite the table if the source table has a different schema structure than the exported table. Enterprise Data Lake drops the table and re-creates the table with the data from the exported table.
8. Choose whether to export the asset now, or to schedule the export activity.
   - Select **Export Now** to start the export activity.
   - Select **Schedule Export** to schedule the export activity.
     If you decide to schedule the activity, specify what to do if the asset already exists in the target.

     - Select **Append** to append the data to an existing table.

     - Select **Overwrite** to overwrite the existing table.
9. Click **OK**.

   You can monitor the progress of the export activity on the **My Activities** page in Enterprise Data Lake.

Uploading Data to the Data Lake

You can upload a delimited text file from your local drive to the data lake. Enterprise Data Lake stores the uploaded data as a Hive table in 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**.

   The **Upload Data** 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 (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 Enterprise 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 Enterprise 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 (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.

Property	Description
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 (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 Enterprise 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 Enterprise 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**.

CHAPTER 5

Create and Manage 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.

You can import a data asset from external sources into the data lake such as Oracle and Teradata databases. Once the import activity is completed successfully, a Hive table is created in the data lake that you can add to a project. After you publish a data asset to the data lake, the publication appears in the project.

Enterprise Data Lake adds data assets to projects as worksheets. Projects are private and can only be viewed by the user who creates the project. However, 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** view or the **My Projects** view. The projects in the **Projects Overview** section appear as cards by default.

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 **New Project**.
   - On the **My Projects** page, click **New Project**.
   - To create a project from the **Search Results** or **Open Assets** view, select a data asset and click **Add to Project**. Then click **New Project**.

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

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. You can also add a data asset by clicking **Search for asset** on the project page.

   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

Enterprise Data Lake provides recommendations on additional data assets to add to a project, based on the data assets added to the project.

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. Enterprise Data Lake makes the following types of recommendations:

Additional recommendations.

Data assets that are used in similar projects by other users.

Alternate recommendations.

Data assets that share similar columns with assets in the project.

Hover over the **Why?** text in a recommendation card for details on why Enterprise Data Lake recommends that you add the data asset to the project.

Editing a Project

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

To edit a project, you must have edit permissions.

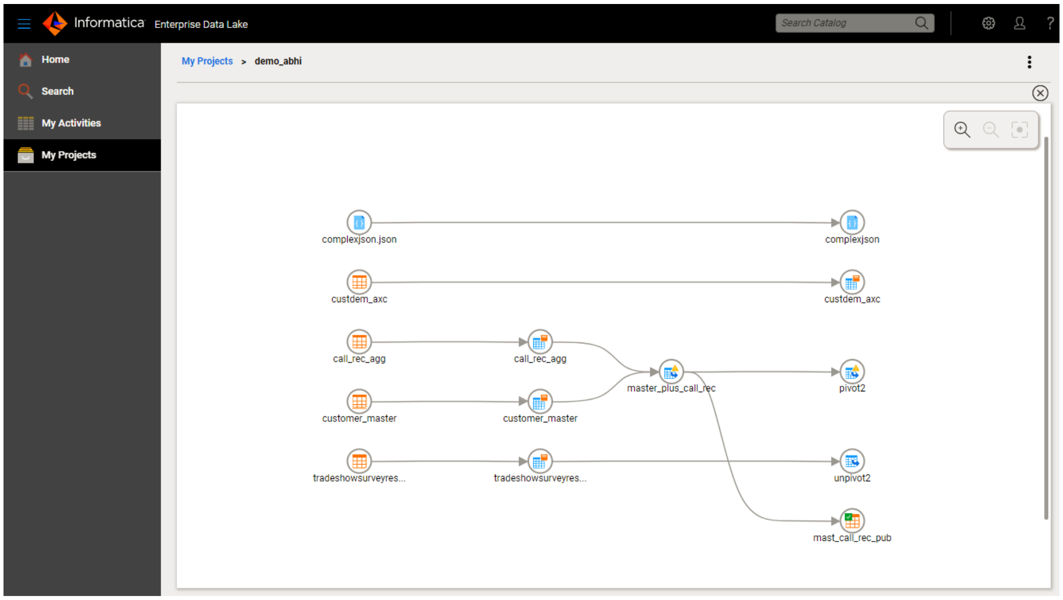
1. In the **Projects** view, select a project.
2. Click the **Manage Projects** icon and select **Edit Properties**.

   The **Edit Project** dialog box appears.
3. Edit the name or description.
4. Click **Save**.

Viewing Project Flows

You can view a flow diagram that shows you how worksheets in a project are related and how they are derived. The diagram is especially useful when you work on a complex project that contains numerous worksheets and includes numerous assets.

The following image shows the flow diagram for a project:



1. In the **Projects** view, select a project.
2. Click the **View Project Flow** icon.

   The project flow view displays the assets in the project and the relationships between assets.
3. Click an icon in the diagram to view details for the corresponding asset.

Viewing Project History

You can review the activities performed within a project to gain insights into the project, or to perform root cause analysis to uncover issues with a project.

You can review the following activities:

- All activities within the project.
- Activities performed on worksheets within the project.
- Publication activities for assets within the project.

Perform the following steps to view the history for a project:

1. In the **Projects** view, select a project.
2. Select **View Project History** from the **Manage Projects** menu.

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.

   Enterprise 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 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 then 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.

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

Locked Projects

When you edit shared projects, the projects are locked automatically for editing. Locks on shared projects are automatically released when you sign out of Enterprise Data Lake.

A project is locked when a user performs any of the following tasks:

- Edits a project.
- Shares a project.
- Prepares a project.
- Publishes a project.
- Changes the project owner.
- Deletes the project.

When other users try to access this project, a project locked icon appears. They cannot perform any of these tasks when a project is locked. To view the name of the user who is working on a project, hover the mouse over the locked icon.

If a user is performing tasks other than preparing and publishing, other users can access the project and view the data assets in the worksheet. However, if a user has a worksheet open in view mode, other users cannot access the worksheet.

Administrators and users with administrator privileges can unlock projects.

Deleting a Worksheet

You can delete a worksheet from a project that you own. You cannot delete a worksheet if it has dependent worksheets.

1. In the **Projects** view, select a project.
2. In the project page, select a worksheet.

3. Click the **Manage Worksheets** icon, and then select **Delete Worksheet**.

   The **Delete Worksheet** dialog box appears.
4. Click **Yes**.

Deleting a Project

You can delete projects that you own from Enterprise Data Lake.

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 6

Prepare Data

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, Enterprise Data Lake creates a corresponding worksheet in the project. A worksheet has an interactive data-driven spreadsheet like interface. The worksheet contains the data and a recipe that tracks the changes you make to the data as you prepare it. Enterprise Data Lake loads sample data into the worksheet depending on your sampling and filtering selection when you open a worksheet.

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

You can prepare data from the following sources:

- Amazon S3
- MapR-FS
- Microsoft Azure Data Lake Storage
- Microsoft Windows Azure Storage Blob

Data Preparation Features

Enterprise Data Lake provides a number of features you can use to prepare data in a worksheet.

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

Filter data

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.

Edit a worksheet and its columns

Use the **Edit** menu to edit the columns in a worksheet. You can choose to delete blank columns or columns with one value, or hide or unhide columns in the worksheet.

You can also delete duplicate rows within a selected column. When delete duplicate rows, you can choose to ignore case as well as leading or trailing whitespace in rows containing Text values. To delete duplicate rows, select a column, and then select **Delete** > **Duplicate rows** from the **Edit** menu.

Note: You can only delete duplicate rows from a worksheet that contains a maximum of 100 columns.

You can right-click the worksheet tab 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.

You can hide and unhide columns in a worksheet. You can also easily find and navigate to specific columns using the **Columns** panel, which also shows you the percentage of unique versus duplicate values a column contains. These features are especially useful when you prepare large data sets with dozens of columns.

View suggestions for improving the data

Enterprise Data Lake provides an overview of the data in the worksheet and of the values in each column. The application also offers suggestions on how you can manipulate or improve the data.

Refresh the data in a 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.

Blend data

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, or you can add columns from another sheet into your current sheet based on a common lookup key.

You can also group similar values in a column to into categories to make analysis easier.

Summarize data

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

You can pivot columns to reshape the data into a pivot table format for analysis. You can also unpivot columns in a worksheet into rows containing the column data in key value format.

Apply formulas, rules and window functions to data

Use the **Functions** icon to formulas, rules and window functions to worksheets.

You can apply formulas to columns using the Text, Numeric, Date, and other functions available. You can also create your own formulas.

You can apply rules to data in a worksheet to help you cleanse, transform, or validate the data. You can apply passive rules, which operate on columns in the current worksheet; or active rules, which operate on one more columns in one or more worksheets within the project.

You can use window functions in Enterprise Data Lake to perform operations on a small subset of a larger data set. A window function operates on a group of rows in a worksheet, and calculates a return value for every input row.

View and manage your data preparation steps in the worksheet recipe

Use the recipe icon to view and manage the steps you take to prepare the data in the worksheet.

You can edit a step in a recipe, or insert a new step. You can also reuse recipe steps. After you modify a recipe, you can use the back in time mode to review the impact of each recipe step on the data in the worksheet.

Data Masking

Enterprise Data Lake integrates with Informatica Dynamic Data Masking, a data security product, to enable masking of sensitive data in data assets. When you view or perform operations on columns containing masked data, the actual data is fully or partially obfuscated based on the masking rules applied.

Data masking can be applied to data in the following scenarios:

When previewing data.

When you preview a data asset that is in the data lake or in an external database, data is masked according to the masking rules defined on the source data.

When importing data from an external database.

When you import data from an external database, data masking rules are applied to the data in the Hive table created for the asset.

When preparing data.

During data preparation, columns in worksheets are populated with masked data according to the masking policies defined on the source data. When you perform a blending operation such as a join or lookup, data in masked columns cannot be blended with unmasked columns.

When exporting or downloading data to an external database or file.

When you export data from the data lake to an external database, or downloads data to a file, data is written to the database or file in masked format.

When publishing data.

When you publish prepared data, data in columns that were masked during data preview, data sampling, and data preparation is written to the Hive table in masked format.

Sampling Data

You must sample the data for each asset you add to a project before you can begin preparing the data. You perform data preparation operations using the sample data in the worksheet created for the asset.

The sampling preview page in the Enterprise Data Lake application displays the sample data to load based on the sampling selection settings specified for the asset. You can edit the sampling settings to meet your needs.

Sampling Table Data

You must sample the data in each database table you add to your project as the first step in data preparation. Enterprise Data Lake presents the data in a worksheet that you use to sample the data.

The sampling preview page displays the sample data to load based on the current sampling selection settings. You can edit the settings to meet your needs. If you do not need to change the settings, click **Load** to load the sample data.

1. On the project page, click **Prepare**.

   The Sampling Preview panel appears along with the worksheet.
2. To edit the current sampling selection settings, click the **edit** icon in the Current Sampling Selection panel.
3. In the **Columns** tab, select the columns that you want to work with. To search for a column, enter the column name in the Search text box. You can search for multiple columns by entering a comma-separated list.

   Note that if columns are added to the source data after you complete sampling the data, and you want to sample the data again, you must select the new columns in the **Columns** tab.
4. Click the **Filters** tab.
   a. Click **New Filter**.
   b. In the **Create filter with** text box, enter the name of the column to use for filtering.
   c. From the **select** drop down list, click one of the following options for string columns:
      - **Equals**. Select to filter and display rows with the column value equal to this entered value. You can enter multiple values in this field, separated by a comma. If you enter multiple values, rows that match each value are filtered. Only the first value is considered and the other entries are ignored.
      - **Contains**. Select to filter and display data containing the entered value. You can enter multiple values in this field, separated by a comma. If you enter multiple values, rows that match each value are filtered.
      - **Starts with**. Select to filter and display rows with the column value that start with the entered value. You can enter multiple values in this field, separated by a comma. If you enter multiple values, rows that match each value are filtered.
      - **Ends with**. Select to filter and display rows with the column value that end with the entered value. You can enter multiple values in this field, separated by a comma. If you enter multiple values, rows that match each value are filtered.

      For date or numeric columns, select **Equals**, **Less than**, or **Greater than**.
   d. Enter the value to use to filter the column.
   e. Click **Save**.

   Note: Filters are not available when you try to configure sampling for worksheets created for HDFS files.

5. Click the **Sampling** tab.
6. Select the Sampling type:
    - **Random Rows**. Select to view the rows at random from the entire data asset. The number of rows displayed may be less than or equal to the sample size.
    - **First Rows**. Select to view consecutive rows starting with the first row in the file.

    Note: Random sampling is not available when you try to configure sampling for worksheets created for HDFS files.
7. Enter the number of rows to sample. The Informatica administrator can set the maximum number of rows in the Administrator tool.
8. Click **Save** to save your changes.
9. Click **Load** to load the sample data.

Sampling Delimited Files

You must sample the data in each delimited file you add to your project as the first step in data preparation. Enterprise Data Lake presents the data in a worksheet that you use to sample the data.

The sampling preview page displays the sample data to load based on the current sampling selection settings. You can edit the settings to meet your needs. If you do not need to change the settings, click **Load** to load the sample data.

Note that if columns are added to the source data after you complete sampling the data, and you want to sample the data again, you must select the new columns in the **Columns** tab. Columns that have changed in the source data but are used in a recipe step are included in the sample data by default.

1. On the project page, click **Prepare**.

    The Sampling Preview panel appears along with the worksheets.
2. To edit the current sampling selection settings, click the **edit** icon in the Current Sampling Selection panel.
3. In the **Format** tab, enter the code page, delimiter and text qualifier format details for the flat file.
    a. Start at row. Enter the row number from which you want to start the sampling.
    b. Column Names. Select **Get column names from the specified Start at Row** if the first row contains column names.
4. In the **Columns** tab, select the columns that you want to work with. To search for a column, enter the column name in the Search text box. You can search for multiple columns by entering a comma-separated list.
5. In the **Sampling** tab, enter the number of rows to sample.

    The Informatica administrator can set the maximum number of rows in the Administrator tool.
6. Click **Save** to save your changes.
7. Click **Load** to load the sample data.

Sampling JSON Data

You can sample files in the JSON Lines (JSONL) format, which is also called newline-delimited JSON. Enterprise Data Lake flattens the data in the JSON Lines file into a flat structure and presents the data in a worksheet that you use to sample the data.

The Sampling Preview panel displays the JSON file structure as a tree. Each node in the tree represents a key that exists in one or more lines within the JSON file.

Enterprise Data Lake creates columns in the worksheet for the keys you include in the sample data. The values corresponding to the keys within each line in the file are displayed in rows in the worksheet.

The nodes for top-level keys, such as a string, number, or boolean, are selected in the tree by default. To select a nested key, you can either select the key to include all of the keys below it, or expand the key and select each key to include.

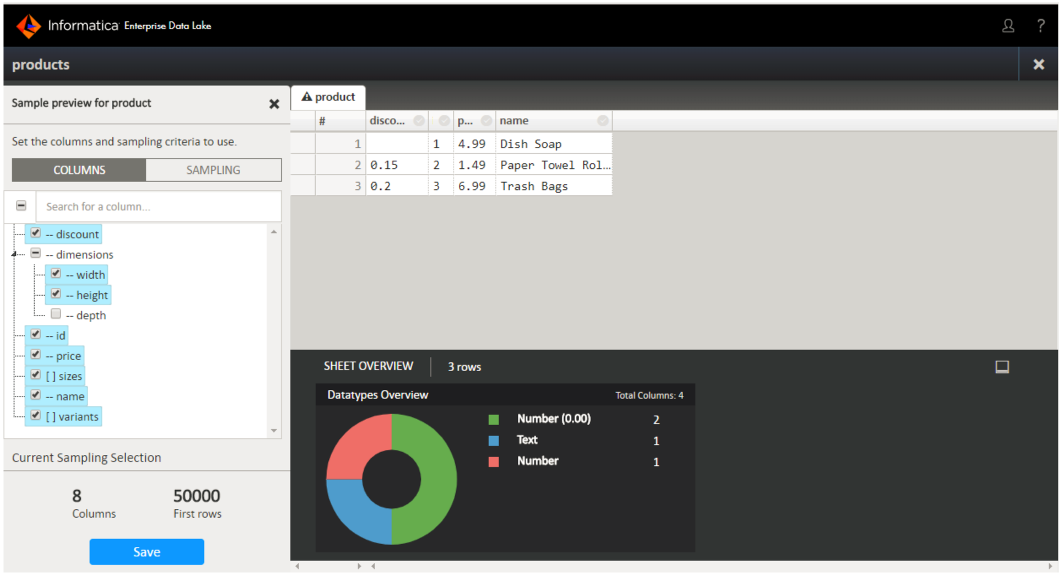
When you select key that is an array, Enterprise Data Lake explodes the values in the array into rows in the worksheet. Rows are created for each array element in each line in the JSON file. A file that contains numerous arrays might result in a large number of rows added to the worksheet, up to the maximum number of rows specified in the sampling settings.

If a line in the file does not contain a particular key or array, the cell in the corresponding column is empty.

The following image shows the contents of a JSON Lines file. The file contains a key with nested keys and two arrays.

```
{ "id": 1, "name": "Dish Soap", "price": 4.99, "variants": [ "Lemon", "Pine", "Lavender" ], "sizes" : ["Single", "Four Pack"] }
{ "id": 2, "name": "Paper Towel Roll", "price": 1.49, "discount": 0.15, "variants": [ "Regular", "Floral Print" ], "sizes" :
  ["Single", "Six Pack"], "dimensions": { "height": 10, "width": 6, "depth": 6 } }
{ "id": 3, "name": "Trash Bags", "price": 6.99, "discount": 0.2, "variants": [ "Draw String", "Tie" ], "sizes" : ["48 Count"],
  "dimensions": { "height": 12, "width": 8, "depth": 6 } }
```

The following image shows the keys to sample selected in the tree. Note that the width and height keys within the dimensions key are selected.



Set the columns and sampling criteria to use.

- ☑ -- discount
- ⊟ -- dimensions
  - ☑ -- width
  - ☑ -- height
  - ☐ -- depth
- ☑ -- id
- ☑ -- price
- ☑ [ ] sizes
- ☑ -- name
- ☑ [ ] variants

Current Sampling Selection

8 Columns

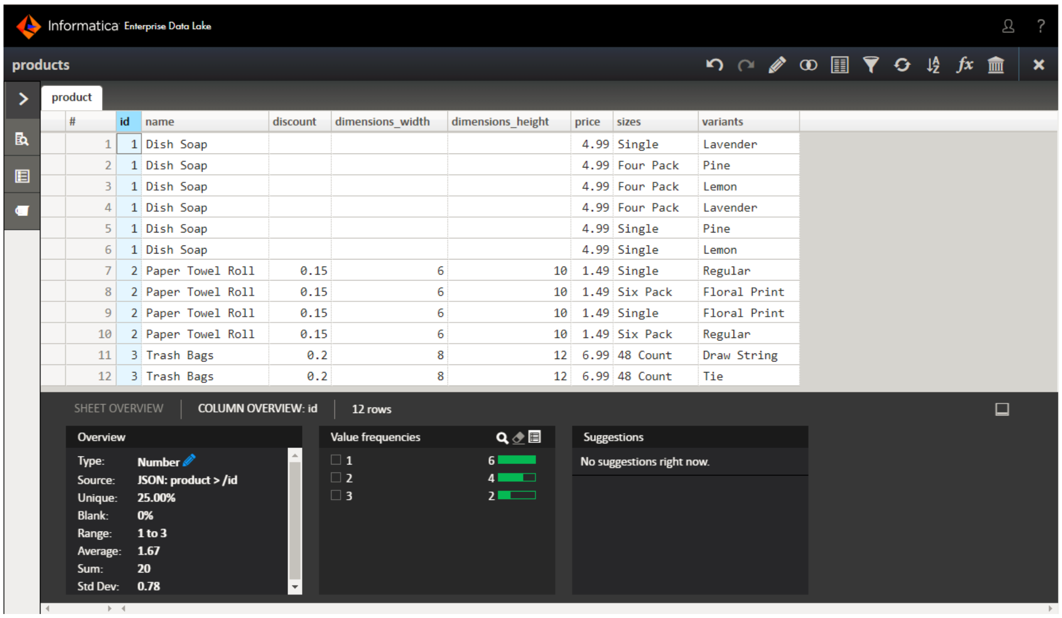
50000 First rows

#	disco...		p...	name
1		1	4.99	Dish Soap
2	0.15	2	1.49	Paper Towel Rol...
3	0.2	3	6.99	Trash Bags

3 rows

🟩	Number (0.00)	2
🟦	Text	1
🟥	Number	1

The following image shows the worksheet created for the JSON file. The width and depth keys within the dimensions key are added as columns to the worksheet. The sizes and variants arrays are exploded into rows.



#	id	name	discount	dimensions_width	dimensions_height	price	sizes	variants
1	1	Dish Soap				4.99	Single	Lavender
2	1	Dish Soap				4.99	Four Pack	Pine
3	1	Dish Soap				4.99	Four Pack	Lemon
4	1	Dish Soap				4.99	Four Pack	Lavender
5	1	Dish Soap				4.99	Single	Pine
6	1	Dish Soap				4.99	Single	Lemon
7	2	Paper Towel Roll	0.15	6	10	1.49	Single	Regular
8	2	Paper Towel Roll	0.15	6	10	1.49	Six Pack	Floral Print
9	2	Paper Towel Roll	0.15	6	10	1.49	Single	Floral Print
10	2	Paper Towel Roll	0.15	6	10	1.49	Six Pack	Regular
11	3	Trash Bags	0.2	8	12	6.99	48 Count	Draw String
12	3	Trash Bags	0.2	8	12	6.99	48 Count	Tie

Type: Number
Source: JSON: product > /id
Unique: 25.00%
Blank: 0%
Range: 1 to 3
Average: 1.67
Sum: 20
Std Dev: 0.78

Value	Frequency
☐ 1	6
☐ 2	4
☐ 3	2

No suggestions right now.

Because line 1 in the JSON file does not contain the discount or dimensions keys, the cells in the corresponding rows are empty.

Sampling a JSON File

You must sample the hierarchal data in each JavaScript key Notation Lines (JSONL) file you add to your project as the first step in data preparation.

The sampling preview page displays the sample data to load based on the current sampling selection settings. You can edit the settings to meet your needs. If you do not need to change the settings, click **Load** to load the sample data.

Note that if columns are added to the source data after you complete sampling the data, and you want to sample the data again, you must select the new columns in the **Columns** tab. Columns that have changed in the source data but are used in a recipe step are included in the sample data by default.

1. On the project page, click **Prepare**.

   The Sampling Preview panel displays the JSON key hierarchy in a tree structure.

2. To edit the current sampling selection settings, click the **edit** icon in the Current Sampling Selection panel.

3. In the **Columns** tab, select or deselect each node to add the corresponding key as a column in the worksheet.

   The nodes for primitive keys, such as strings and numbers, are selected and displayed as columns in the worksheet by default.

   To select an array or a key that contains other primitive keys, you can either select the node to include all of the keys below, or expand the node and then select each key to include. Note that if a key within a top level key contains a control character, whitespace, or any punctuation other than an underscore, you cannot select the key.

4. In the **Sampling** tab, enter the number of rows to sample.

The Informatica administrator can set the maximum number of rows in the Administrator tool.

5. Click **Save** to save your changes.
6. Click **Load** to load the sample data.

JSON Data Preparation Errors

You might encounter an error when you attempt to load sample data from a JSONL file. The most common causes of errors that might occur include the following:

- The file is not in JSON Lines format.
- The file and the file schema do not match.
- The file is corrupted.
- A column name in the file exceeds 128 characters.

Data preview of a JSON file that contains a column name that includes double quotes fails. For example, the following column causes data preview to fail:

```
{"id": 2,"\"first_name\"": "Bonnie","\"last_name\"":
"Cruz","Salary": 8489}
```

Limitations and Constraints

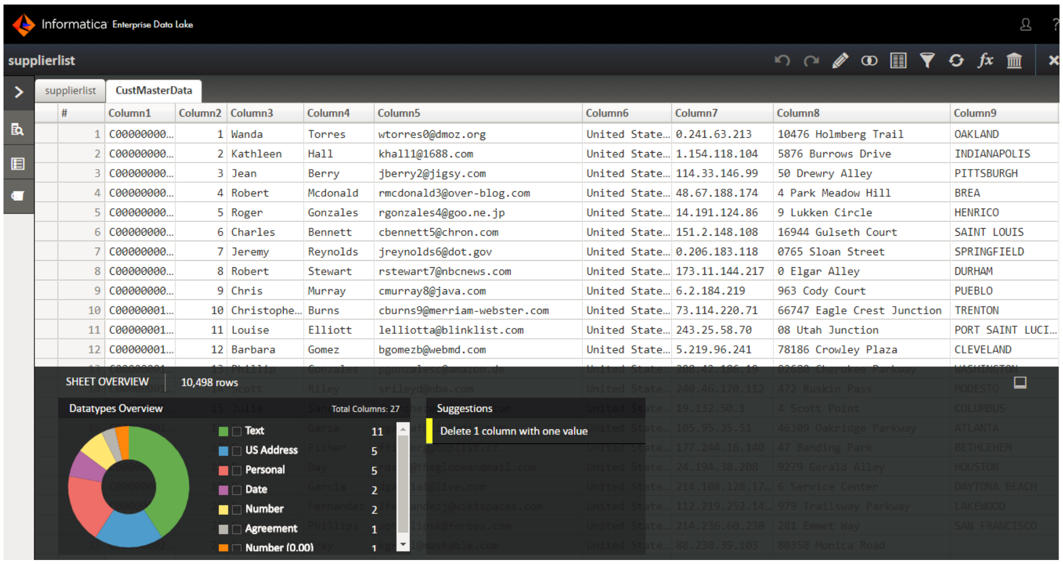
When there is a change in the source data, note the following constraints before you try to edit the sampling settings.

- Missing columns used in the recipe. You can't edit the sample if one or more columns used in the recipe are missing in the source. Restore the missing columns, or remove the columns from the sample.
- Missing columns used in sampling filter criteria. If one or more columns used to filter the sample are missing in the source, the filters will be automatically removed when you edit the sample. Restore the missing columns, or remove the columns from the sample.
- Missing columns selected for sampling. If you selected all columns during sampling, a warning will appear if some of the columns are missing in the source. Restore the missing columns, or remove the columns from the sample.
- Data type of a column has changed. If the data type of one or more columns used to filter the sample changes in the source, the filters are automatically removed when you edit the sample. Revert the data types for these columns before editing, or select different filters.

Working with Worksheets

When you use a worksheet to prepare your data, Enterprise Data Lake provides an overview of the data in the worksheet and of the values in each column. The application also offers suggestions on how you can manipulate or improve the data.

When you open a worksheet, Enterprise Data Lake provides you with an overview of the data within the worksheet, as shown in the following image.



#	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
1	C00000000...	1	Wanda	Torres	wtorres0@dmoz.org	United State...	0.241.63.213	10476 Holmberg Trail	OAKLAND
2	C00000000...	2	Kathleen	Hall	khall1@1688.com	United State...	1.154.118.104	5876 Burrows Drive	INDIANAPOLIS
3	C00000000...	3	Jean	Berry	jberry2@jigsy.com	United State...	114.33.146.99	50 Drewry Alley	PITTSBURGH
4	C00000000...	4	Robert	Mcdonald	rmcdonald3@over-blog.com	United State...	48.67.188.174	4 Park Meadow Hill	BREA
5	C00000000...	5	Roger	Gonzales	rgonzales4@goo.ne.jp	United State...	14.191.124.86	9 Lukken Circle	HENRICO
6	C00000000...	6	Charles	Bennett	cbennett5@chron.com	United State...	151.2.148.108	16944 Gulseth Court	SAINT LOUIS
7	C00000000...	7	Jeremy	Reynolds	jreynolds6@dot.gov	United State...	0.206.183.118	0765 Sloan Street	SPRINGFIELD
8	C00000000...	8	Robert	Stewart	rstewart7@nbcnews.com	United State...	173.11.144.217	0 Elgar Alley	DURHAM
9	C00000000...	9	Chris	Murray	cmurray8@java.com	United State...	6.2.184.219	963 Cody Court	PUEBLO
10	C00000001...	10	Christophe...	Burns	cburns9@merriam-webster.com	United State...	73.114.220.71	66747 Eagle Crest Junction	TRENTON
11	C00000001...	11	Louise	Elliott	lelliotta@blinklist.com	United State...	243.25.58.70	08 Utah Junction	PORT SAINT LUCI...
12	C00000001...	12	Barbara	Gomez	bgomezb@webmd.com	United State...	5.219.96.241	78186 Crowley Plaza	CLEVELAND

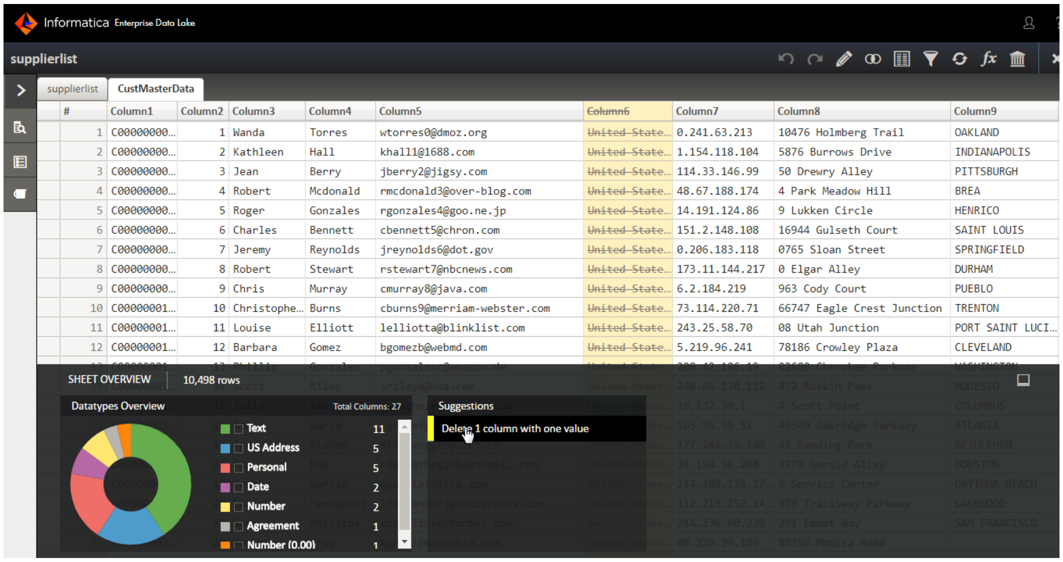
Datatypes Overview	
🟩 ☐ Text	11
🟦 ☐ US Address	5
🟥 ☐ Personal	5
🟪 ☐ Date	2
🟨 ☐ Number	2
⬜ ☐ Agreement	1
🟧 ☐ Number (0.00)	1

Suggestions: Delete 1 column with one value

The **Datatypes Overview** panel shows you how many columns in the worksheet are of each type or data domain. You can select one or more types to filter in the worksheet.

The **Suggestions** panel offers suggestions on how you might improve the worksheet. When you hover over a suggestion, the application updates the worksheet with a preview of the impact of accepting the suggestion.

The following image shows the impact of accepting the suggestion to delete a column that contains only one distinct value has on the worksheet, and may not add any value when you publish the asset to the data lake:

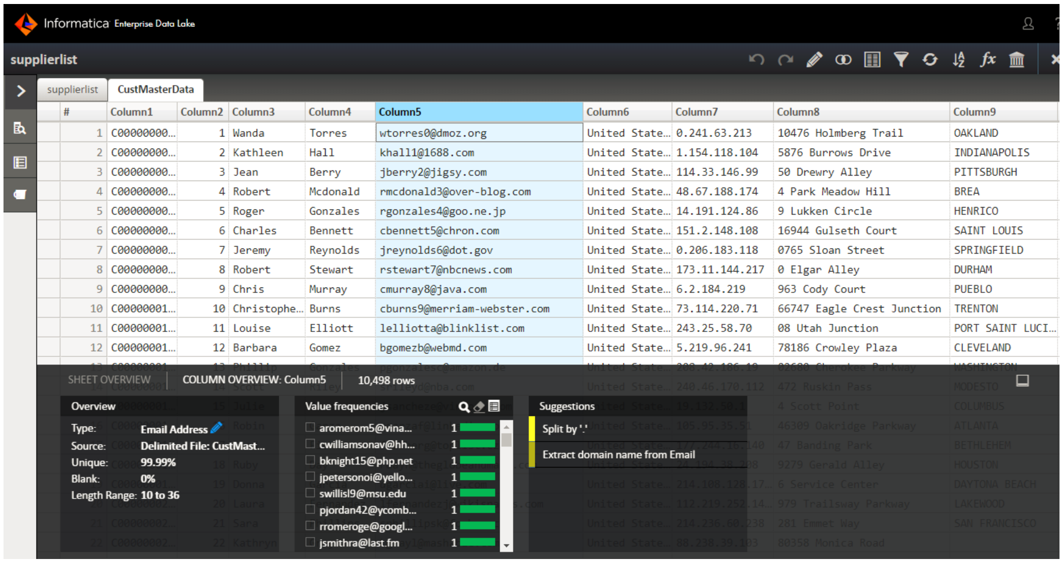


#	Column1	Column2	Column3	Column4	Column5	~~Column6~~	Column7	Column8	Column9
1	C00000000...	1	Wanda	Torres	wtorres0@dmoz.org	~~United State...~~	0.241.63.213	10476 Holmberg Trail	OAKLAND
2	C00000000...	2	Kathleen	Hall	khall1@1688.com	~~United State...~~	1.154.118.104	5876 Burrows Drive	INDIANAPOLIS
3	C00000000...	3	Jean	Berry	jberry2@jigsy.com	~~United State...~~	114.33.146.99	50 Drewry Alley	PITTSBURGH
4	C00000000...	4	Robert	Mcdonald	rmcdonald3@over-blog.com	~~United State...~~	48.67.188.174	4 Park Meadow Hill	BREA
5	C00000000...	5	Roger	Gonzales	rgonzales4@goo.ne.jp	~~United State...~~	14.191.124.86	9 Lukken Circle	HENRICO
6	C00000000...	6	Charles	Bennett	cbennett5@chron.com	~~United State...~~	151.2.148.108	16944 Gulseth Court	SAINT LOUIS
7	C00000000...	7	Jeremy	Reynolds	jreynolds6@dot.gov	~~United State...~~	0.206.183.118	0765 Sloan Street	SPRINGFIELD
8	C00000000...	8	Robert	Stewart	rstewart7@nbcnews.com	~~United State...~~	173.11.144.217	0 Elgar Alley	DURHAM
9	C00000000...	9	Chris	Murray	cmurray8@java.com	~~United State...~~	6.2.184.219	963 Cody Court	PUEBLO
10	C00000001...	10	Christophe...	Burns	cburns9@merriam-webster.com	~~United State...~~	73.114.220.71	66747 Eagle Crest Junction	TRENTON
11	C00000001...	11	Louise	Elliott	lelliotta@blinklist.com	~~United State...~~	243.25.58.70	08 Utah Junction	PORT SAINT LUCI...
12	C00000001...	12	Barbara	Gomez	bgomezb@webmd.com	~~United State...~~	5.219.96.241	78186 Crowley Plaza	CLEVELAND

Datatypes Overview	
🟩 ☐ Text	11
🟦 ☐ US Address	5
🟥 ☐ Personal	5
🟪 ☐ Date	2
🟨 ☐ Number	2
⬜ ☐ Agreement	1
🟧 ☐ Number (0.00)	1

Suggestions: Delete 1 column with one value

When you select a column in the worksheet, the application analyzes the data and makes suggestions on how you can improve or manipulate the data. Enterprise Data Lake displays the data analysis and suggestions in panels at the bottom of the worksheet, as shown in the following image:



#	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
1	C00000000...	1	Wanda	Torres	wtorres0@dmoz.org	United State...	0.241.63.213	10476 Holmberg Trail	OAKLAND
2	C00000000...	2	Kathleen	Hall	khall1@1688.com	United State...	1.154.118.104	5876 Burrows Drive	INDIANAPOLIS
3	C00000000...	3	Jean	Berry	jberry2@jigsy.com	United State...	114.33.146.99	50 Drewry Alley	PITTSBURGH
4	C00000000...	4	Robert	Mcdonald	rmcdonald3@over-blog.com	United State...	48.67.188.174	4 Park Meadow Hill	BREA
5	C00000000...	5	Roger	Gonzales	rgonzales4@goo.ne.jp	United State...	14.191.124.86	9 Lukken Circle	HENRICO
6	C00000000...	6	Charles	Bennett	cbennett5@chron.com	United State...	151.2.148.108	16944 Gulseth Court	SAINT LOUIS
7	C00000000...	7	Jeremy	Reynolds	jreynolds6@dot.gov	United State...	0.206.183.118	0765 Sloan Street	SPRINGFIELD
8	C00000000...	8	Robert	Stewart	rstewart7@nbcnews.com	United State...	173.11.144.217	0 Elgar Alley	DURHAM
9	C00000000...	9	Chris	Murray	cmurray8@java.com	United State...	6.2.184.219	963 Cody Court	PUEBLO
10	C00000001...	10	Christophe...	Burns	cburns9@merriam-webster.com	United State...	73.114.220.71	66747 Eagle Crest Junction	TRENTON
11	C00000001...	11	Louise	Elliott	lelliotta@blinklist.com	United State...	243.25.58.70	08 Utah Junction	PORT SAINT LUCI...
12	C00000001...	12	Barbara	Gomez	bgomezb@webmd.com	United State...	5.219.96.241	78186 Crowley Plaza	CLEVELAND

Type: Email Address
Source: Delimited File: CustMast...
Unique: 99.99%
Blank: 0%
Length Range: 10 to 36

☐ aromerom5@vina...	1
☐ cwilliamsonav@hh...	1
☐ bknight15@php.net	1
☐ jpetersonoi@yello...	1
☐ swillisl9@msu.edu	1
☐ pjordan42@ycomb...	1
☐ rromeroge@googl...	1
☐ jsmithra@last.fm	1

The **Overview** panel provides a summary of the column data, including column type and the source of the data.

The Type property indicates the column type. When you prepare a worksheet, Enterprise Data Lake infers types or data domains for string values in each column based on the data types or data domains assigned to data in similar assets. For example, if all of the values in a column contain digits, Enterprise Data Lake infers that the column is a Number type. If values in the column contain digits in the format of U.S. postal codes, Enterprise Data Lake might infer that the column values belong to the US 9 Digit Zip Code data domain.

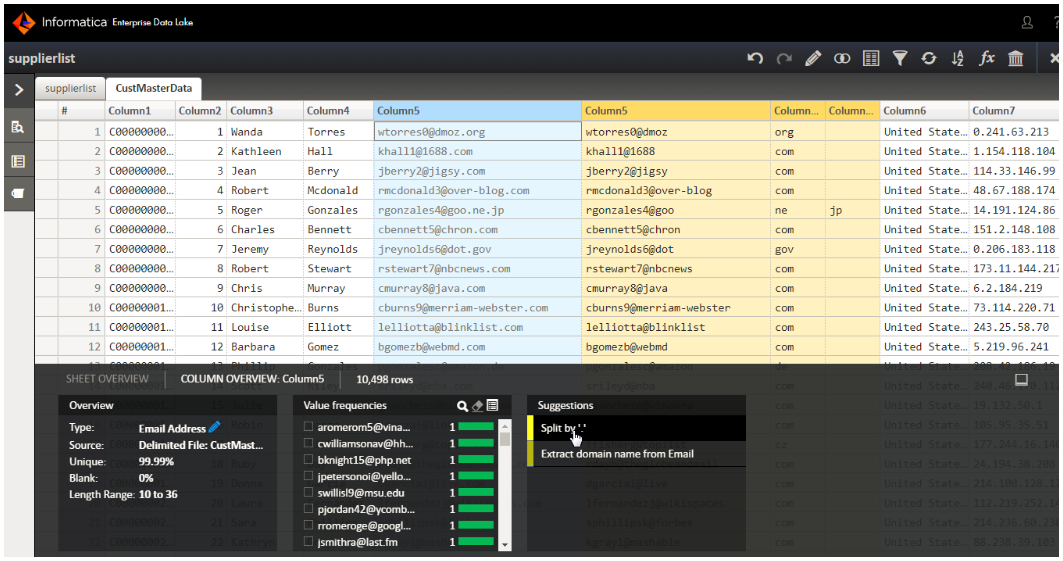
You can choose to accept the inferred type or data domain for a column, or you can revert the column back to the source data type. You can also convert some types to another type, such as Date to Text. For more information, see "Reverting an Inferred Data Type" on page 48.

The Source property indicates the source of the data displayed in the column. For more information, see "Viewing Column Data Sources" on page 48.

The **Value frequencies** panel displays the occurrence of each distinct value in the column. You can select one or more values to filter in the worksheet.

Suggestions panel offers suggestions on how you might manipulate the column data. When you hover over a suggestion, the application shows you a preview of the impact of accepting the suggestion.

The following image shows a preview of the columns that selecting the Split by suggestion adds to the worksheet:



#	Column1	Column2	Column3	Column4	Column5	Column5	Column...	Column...	Column6	Column7
1	C00000000...	1	Wanda	Torres	wtorres0@dmoz.org	wtorres0@dmoz	org		United State...	0.241.63.213
2	C00000000...	2	Kathleen	Hall	khall1@1688.com	khall1@1688	com		United State...	1.154.118.104
3	C00000000...	3	Jean	Berry	jberry2@jigsy.com	jberry2@jigsy	com		United State...	114.33.146.99
4	C00000000...	4	Robert	Mcdonald	rmcdonald3@over-blog.com	rmcdonald3@over-blog	com		United State...	48.67.188.174
5	C00000000...	5	Roger	Gonzales	rgonzales4@goo.ne.jp	rgonzales4@goo	ne	jp	United State...	14.191.124.86
6	C00000000...	6	Charles	Bennett	cbennett5@chron.com	cbennett5@chron	com		United State...	151.2.148.108
7	C00000000...	7	Jeremy	Reynolds	jreynolds6@dot.gov	jreynolds6@dot	gov		United State...	0.206.183.118
8	C00000000...	8	Robert	Stewart	rstewart7@nbcnews.com	rstewart7@nbcnews	com		United State...	173.11.144.217
9	C00000000...	9	Chris	Murray	cmurray8@java.com	cmurray8@java	com		United State...	6.2.184.219
10	C00000001...	10	Christophe...	Burns	cburns9@merriam-webster.com	cburns9@merriam-webster	com		United State...	73.114.220.71
11	C00000001...	11	Louise	Elliott	lelliotta@blinklist.com	lelliotta@blinklist	com		United State...	243.25.58.70
12	C00000001...	12	Barbara	Gomez	bgomezb@webmd.com	bgomezb@webmd	com		United State...	5.219.96.241

SHEET OVERVIEW | COLUMN OVERVIEW: Column5 | 10,498 rows

Overview

Type:	Email Address
Source:	Delimited File: CustMast...
Unique:	99.99%
Blank:	0%
Length Range:	10 to 36

Value frequencies

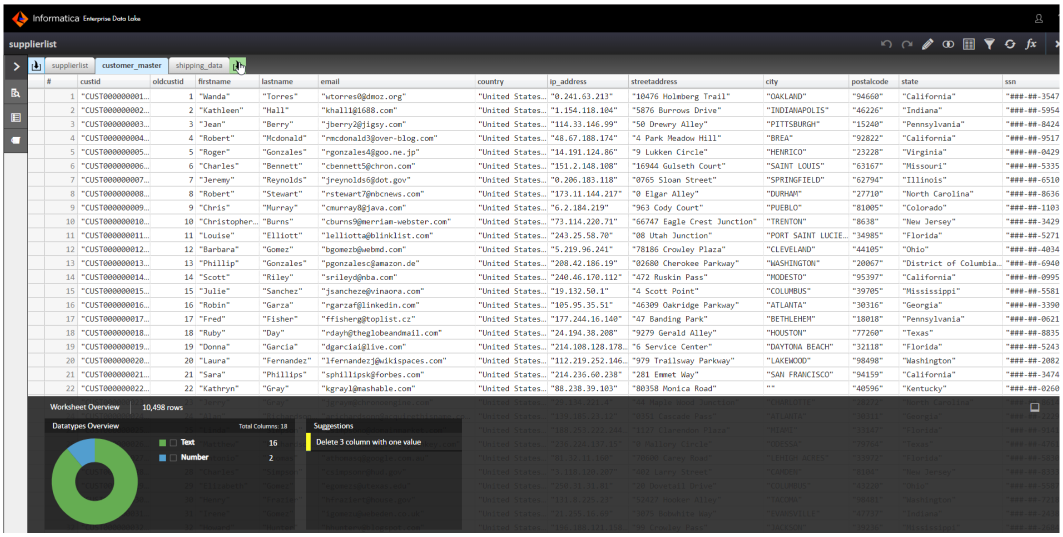
☐ aromerom5@vina...	1
☐ cwilliamsonav@hh...	1
☐ bknight15@php.net	1
☐ jpetersonoi@yello...	1
☐ swillisl9@msu.edu	1
☐ pjordan42@ycomb...	1
☐ rromeroge@googl...	1
☐ jsmithra@last.fm	1

Suggestions

- Split by '.'
- Extract domain name from Email

If a project contains multiple worksheets, you can reorder the worksheets in the data preparation page by moving a worksheet to a different position. For example, you might want to move the second worksheet displayed in the page to the third position.

To move a worksheet, select the worksheet tab, and then select Move Worksheet. You can then click the icon indicating the position you want to move the worksheet to, as shown in the following image:



#	custid	oldcustid	firstname	lastname	email	country	ip_address	streetaddress	city	postalcode	state	ssn
1	"CUST000000001...	1	"Wanda"	"Torres"	"wtorres0@dmoz.org"	"United States...	"0.241.63.213"	"10476 Holmberg Trail"	"OAKLAND"	"94660"	"California"	"###-##-3547
2	"CUST000000002...	2	"Kathleen"	"Hall"	"khall1@1688.com"	"United States...	"1.154.118.104"	"5876 Burrows Drive"	"INDIANAPOLIS"	"46226"	"Indiana"	"###-##-5954
3	"CUST000000003...	3	"Jean"	"Berry"	"jberry2@jigsy.com"	"United States...	"114.33.146.99"	"50 Drewry Alley"	"PITTSBURGH"	"15240"	"Pennsylvania"	"###-##-8424
4	"CUST000000004...	4	"Robert"	"Mcdonald"	"rmcdonald3@over-blog.com"	"United States...	"48.67.188.174"	"4 Park Meadow Hill"	"BREA"	"92822"	"California"	"###-##-9517
5	"CUST000000005...	5	"Roger"	"Gonzales"	"rgonzales4@goo.ne.jp"	"United States...	"14.191.124.86"	"9 Lukken Circle"	"HENRICO"	"23228"	"Virginia"	"###-##-0429
6	"CUST000000006...	6	"Charles"	"Bennett"	"cbennett5@chron.com"	"United States...	"151.2.148.108"	"16944 Gulseth Court"	"SAINT LOUIS"	"63167"	"Missouri"	"###-##-5335
7	"CUST000000007...	7	"Jeremy"	"Reynolds"	"jreynolds6@dot.gov"	"United States...	"0.206.183.118"	"0765 Sloan Street"	"SPRINGFIELD"	"62794"	"Illinois"	"###-##-6510
8	"CUST000000008...	8	"Robert"	"Stewart"	"rstewart7@nbcnews.com"	"United States...	"173.11.144.217"	"0 Elgar Alley"	"DURHAM"	"27710"	"North Carolina"	"###-##-8636
9	"CUST000000009...	9	"Chris"	"Murray"	"cmurray8@java.com"	"United States...	"6.2.184.219"	"963 Cody Court"	"PUEBLO"	"81005"	"Colorado"	"###-##-1103
10	"CUST000000010...	10	"Christopher...	"Burns"	"cburns9@merriam-webster.com"	"United States...	"73.114.220.71"	"66747 Eagle Crest Junction"	"TRENTON"	"8638"	"New Jersey"	"###-##-3429
11	"CUST000000011...	11	"Louise"	"Elliott"	"lelliotta@blinklist.com"	"United States...	"243.25.58.70"	"08 Utah Junction"	"PORT SAINT LUCIE...	"34985"	"Florida"	"###-##-5271
12	"CUST000000012...	12	"Barbara"	"Gomez"	"bgomezb@webmd.com"	"United States...	"5.219.96.241"	"78186 Crowley Plaza"	"CLEVELAND"	"44105"	"Ohio"	"###-##-4034
13	"CUST000000013...	13	"Phillip"	"Gonzales"	"pgonzalesc@amazon.de"	"United States...	"208.42.186.19"	"02680 Cherokee Parkway"	"WASHINGTON"	"20067"	"District of Columbia...	"###-##-6940
14	"CUST000000014...	14	"Scott"	"Riley"	"srileyd@nba.com"	"United States...	"240.46.170.112"	"472 Ruskin Pass"	"MODESTO"	"95397"	"California"	"###-##-0995
15	"CUST000000015...	15	"Julie"	"Sanchez"	"jsancheze@vinaora.com"	"United States...	"19.132.50.1"	"4 Scott Point"	"COLUMBUS"	"39705"	"Mississippi"	"###-##-5581
16	"CUST000000016...	16	"Robin"	"Garza"	"rgarzaf@linkedin.com"	"United States...	"105.95.35.51"	"46309 Oakridge Parkway"	"ATLANTA"	"30316"	"Georgia"	"###-##-3390
17	"CUST000000017...	17	"Fred"	"Fisher"	"ffisherg@toplist.cz"	"United States...	"177.244.16.140"	"47 Banding Park"	"BETHLEHEM"	"18018"	"Pennsylvania"	"###-##-0621
18	"CUST000000018...	18	"Ruby"	"Day"	"rdayh@theglobeandmail.com"	"United States...	"24.194.38.208"	"9279 Gerald Alley"	"HOUSTON"	"77260"	"Texas"	"###-##-8835
19	"CUST000000019...	19	"Donna"	"Garcia"	"dgarciai@live.com"	"United States...	"214.108.128.178...	"6 Service Center"	"DAYTONA BEACH"	"32118"	"Florida"	"###-##-5243
20	"CUST000000020...	20	"Laura"	"Fernandez"	"lfernandezj@wikispaces.com"	"United States...	"112.219.252.146...	"979 Trailsway Parkway"	"LAKEWOOD"	"98498"	"Washington"	"###-##-2082
21	"CUST000000021...	21	"Sara"	"Phillips"	"sphillipsk@forbes.com"	"United States...	"214.236.60.238"	"281 Emmet Way"	"SAN FRANCISCO"	"94159"	"California"	"###-##-3474
22	"CUST000000022...	22	"Kathryn"	"Gray"	"kgrayl@mashable.com"	"United States...	"88.238.39.103"	"80358 Monica Road"	""	"40596"	"Kentucky"	"###-##-0260

Worksheet Overview | 10,498 rows

Datatypes Overview — Total Columns: 18

🟩 ☐ Text	16
🟦 ☐ Number	2

Suggestions

- Delete 3 column with one value

Converting a Column Data Type

You can convert column data to a different type.

You can convert the following data types:

- Number to Text
- Date to Text

- Text to Number
- Text to Date

The conversion options available depend on the column data type.

The application adds a new column containing the converted data to the worksheet. You can specify

1. Select the column containing the data to convert.
2. Right-click the column, and then select the operation to perform:
    - Select **Convert to Text** to convert values in a Date or Number column to Text format.
    - Select **Convert to Number** to convert values in a Text column to numerical values.
      Enter a name for the column that contains the converted data, and then select the separator to use for thousand or decimal values.
    - Select **Convert to Date** to convert values in a Text column to Date format.
      Enter a name for the column that contains the converted data, and then select the Date format to apply to the data.

      You can also specify a custom format. Custom formats are not case sensitive. The following table describes supported custom Date formats:

Time period	Format
Year	- YYYY: four digit year - YY: last two digits of year
Month	- MONTH: full name of month - MON: abbreviated name of month (Jan-Dec) - MM: numeric designation of the month (01-12)
Day	- DD: day of month (1-31) - DY: abbreviated name of day (Sun-Sat) - DAY: full name of day
Hour	- HH24: hour of day (0-23) - HH12: hour of day (1-12) - HH: hour of day (1-12)
Minute	MI
Second	SS
Millisecond	MS
A.M./P.M.	AM, PM

You can use the following delimiters with custom Date formats:

comma, dash, forward slash, space

You can also use a colon to separate formats in a timestamp, such as in the following example:

HH24:MI:SS

Reverting an Inferred Data Type

You can revert inferred types and data domains back to the source type. You might want to revert an inferred type or data domain back to the source type if you want to use the column data in a formula.

String values that appear to be text, numbers or dates might be inferred as being of the Text, Number, or Date type. You can convert the inferred type to another type. For example, you might want to convert a string inferred as the Date type to the Text type.

String column values that appear to match a data domain might be inferred as being of that data domain type. You can revert the inferred data domain type to the source type.

When you create a derived worksheet through an aggregate, join, or union operation, the types and data domains inferred for columns in the source worksheet are applied to the corresponding columns in the derived worksheet. If you change a data type in a source worksheet, you must refresh all of the corresponding derived worksheets.

You cannot revert the type or data domain assigned to a column if the column is included in a step in a recipe.

After you revert or convert a column type, you can revert the column back to the inferred type if needed.

1. On the project page, click **Prepare**.

   The Sampling Preview panel appears along with the worksheet.
2. Select a string column in a worksheet.

   The inferred type or data domain for the column appears in the Type attribute in the Overview panel.
3. Hover over the **Change type** icon, and then select the operation to perform:
   - Select **Revert column data to <type>** to revert a column inferred as belonging to a data domain to the source type.
     The column reverts to the source type.
   - Select **Convert column data to <type>** to convert an inferred type to another type.
     The column converts to the selected type, and the action is added as a step to the worksheet recipe.
   - Select **Re-infer column type** to revert the column back to the inferred type.
     The column reverts back to the inferred type, and the action is added as a step to the worksheet recipe.

Viewing Column Data Sources

You can view the source of the data for a selected column in the Source property in the **Overview** panel. You might want to view the source of the data in a column to help you troubleshoot an issue.

Columns added by importing source data into a worksheet.

Displays the source type, the source name, and the name of the column from which the column data is derived.

Columns added through operations such as join, lookup, or aggregate.

Displays the names of the worksheet and column from which the column data is derived.

Columns added by a union operation.

Displays the names of each of the worksheets and columns from which the concatenated column data is derived, separated by a comma.

Columns added by operations such as split, concatenate, or extract.

Displays the step in the recipe that generated the column data. Open the Recipe panel and mouse over the Source property to highlight the step in the recipe.

Columns added by applying a formula or a rule to a column.

Displays the step in the recipe that generated the column data. Because the formula or rule modifies the source data, Enterprise Data Lake cannot display the actual source. Open the Recipe panel and mouse over the Source property to highlight the step in the recipe.

Using Recipes

A recipe defines the steps Enterprise Data Lake follows to combine, cleanse, transform, and structure the data in a worksheet. Each operation you perform in a worksheet during data preparation is added as a step to the recipe.

You can edit or delete a step in a recipe. You can also apply a filter to the columns used in a step.

You can insert a new step below an existing step in a recipe. You can insert a step below any step except the last step in the recipe. If filters are applied on the step above the inserted step, Enterprise Data Lake applies the filters to the inserted step. You cannot perform some operations when you insert a step in a recipe, including aggregate, join, or union operations.

To insert a step, select **Insert Step** from the menu in the existing step. The application adds a placeholder step below the selected step. Select the column you want to perform an operation on, and then select the operation to perform.

You can add or edit comment in a recipe step. Use comments to improve collaboration with other users, or to provide details to comply with auditing requirements.

Reviewing and Troubleshooting Recipe Steps

You can use the back in time mode to review the impact of each recipe step in the worksheet. The back in time mode is useful for enabling collaborators on a project to see what has been done thus far, or if you are troubleshooting issues discovered in a worksheet.

To view a step's impact, select the step in the recipe, and then select **Review steps back in time** from the menu. The worksheet updates based on the impact of the selected step.

You can edit or delete a step in back in time mode. Enterprise Data Lake recalculates the subsequent steps in the recipe based on your changes. You can also apply a temporary filter to a step to preview the impact of adding a filter, although you must create and apply the filter in the regular editing mode.

Reusing Recipes

You can reuse recipe steps created in a worksheet, including steps that contain complex formulas or rule definitions. You can reuse recipe steps within the same worksheet or in a different worksheet, including a worksheet in another project. You can copy and reuse selected steps from a recipe, or you can reuse the entire recipe.

When you copy a recipe step, Enterprise Data Lake includes all input columns used in the step, including the column name and type. When you insert a copied step into the recipe in the target worksheet, the application prompts you to map each column in the step to a column of the same type in the target worksheet. The target worksheet must contain columns of the same type as the columns included in each copied recipe step.

If a column with the same name and type exists in the worksheet, Enterprise Data Lake populates the column name. The application notes the number of columns modified in or added to the worksheet as a result of inserting the copied recipe steps.

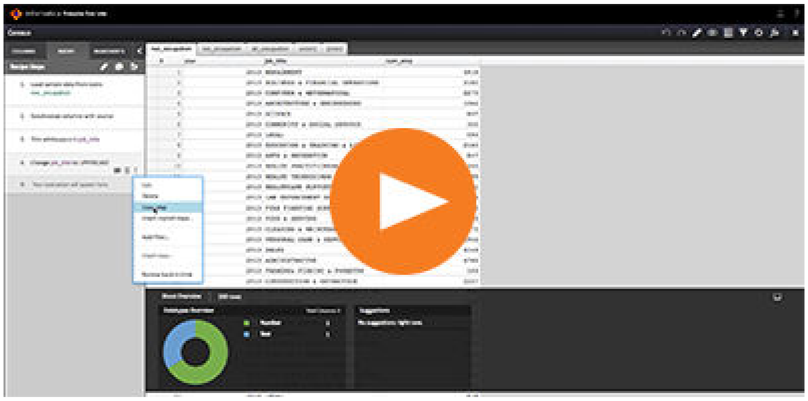
Filters applied in the source worksheet to columns used in copied steps are also copied with the steps. You can choose to apply or ignore the filters when you insert the copied steps into the worksheet.

Input parameters used in formulas or rules are included with the copied recipe steps. You can modify the inputs inserted with a formula or rule after the copied steps are saved in the worksheet.

You can insert copied steps below any step in a recipe, except above any step that is followed by a Synchronize Columns step. The copied steps remain available for insertion into a recipe until you copy any other content in Enterprise Data Lake or your session ends.

You can reuse recipe steps in a worksheet in a different project. If you copy a lookup step into a worksheet in another project, you must select a worksheet in the project that contains a lookup key and added columns with the same names and types as those in the worksheet used in the lookup operation. If a matching worksheet does not exist in the project, you might need to add the data asset used in the lookup operation as a worksheet to the project.

Click the image below to view a video about copying recipe steps:



Copying a Single Recipe Step

You can copy and reuse a single step from a recipe. Filters applied to columns in the step are copied by default.

1. On the project page, click **Prepare**.

   The Sampling Preview panel appears.
2. Select the worksheet containing the recipe step you want to reuse.
3. Click the **Recipe** icon.
4. Select the recipe step you want to copy.
5. Select **Actions** > **Copy Step**.
6. Open the worksheet in which you want to reuse the step.
7. Select the recipe step below which you want to add the copied step.
8. Select **Actions** > **Insert Copied Steps**.
9. To use filters included with the copied step, click **Apply filters from <source sheet>**.
10. Map the column in the step to a corresponding column of the same type in the worksheet.

11. Click **Done**.

    Enterprise Data Lake updates the worksheet based on the step, and displays the number of columns modified in or added to the worksheet as a result of inserting the copied recipe steps.

Copying Multiple Recipe Steps

You can copy and reuse selected steps from a recipe, or you can reuse the entire recipe.

You must select contiguous steps to copy. For example, if you want to copy the second and fourth steps in the recipe, you must select the second, third, and fourth steps. Filters applied to columns in a step are copied by default.

1. On the project page, click **Prepare**.

   The Sampling Preview panel appears.
2. Select the worksheet containing the recipe step you want to reuse.
3. Click the **Recipe** icon.
4. Select the steps you want to copy.
   - To copy all of steps in the recipe, click the **Copy Recipe Steps** icon and select **Copy All Steps**.
   - To copy multiple contiguous steps in the recipe, use **Shift + Click** to select the steps, and then click the **Copy Recipe Steps** icon and select **Copy Selected Steps**.
5. Open the worksheet in which you want to reuse the steps.
6. Select the recipe step below which you want to insert the copied steps.
7. Select **Actions** > **Insert Copied Steps**.
8. To use filters included with the copied steps, click **Apply filters from <source sheet>**.
9. Map each column in the step to a column of the same type in the worksheet.
10. Click **Done**.

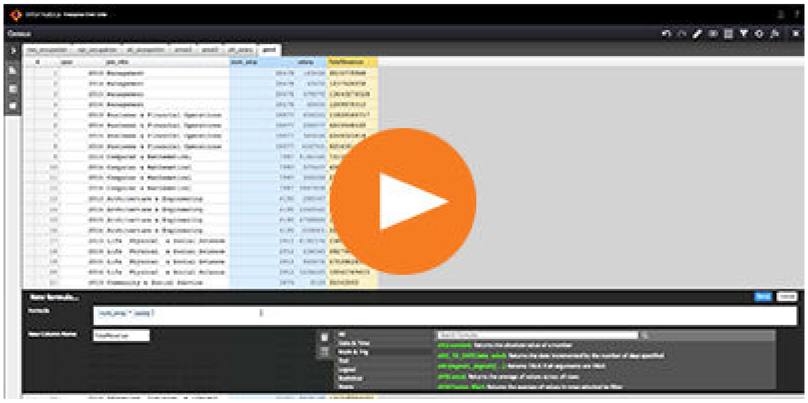
    Enterprise Data Lake updates the worksheet based on the step, and displays the number of columns modified in or added to the worksheet as a result of inserting the copied recipe steps.

Using Formulas

You can use formulas to perform calculations and return values in a worksheet. Enterprise Data Lake adds a column containing the formula results to the worksheet.

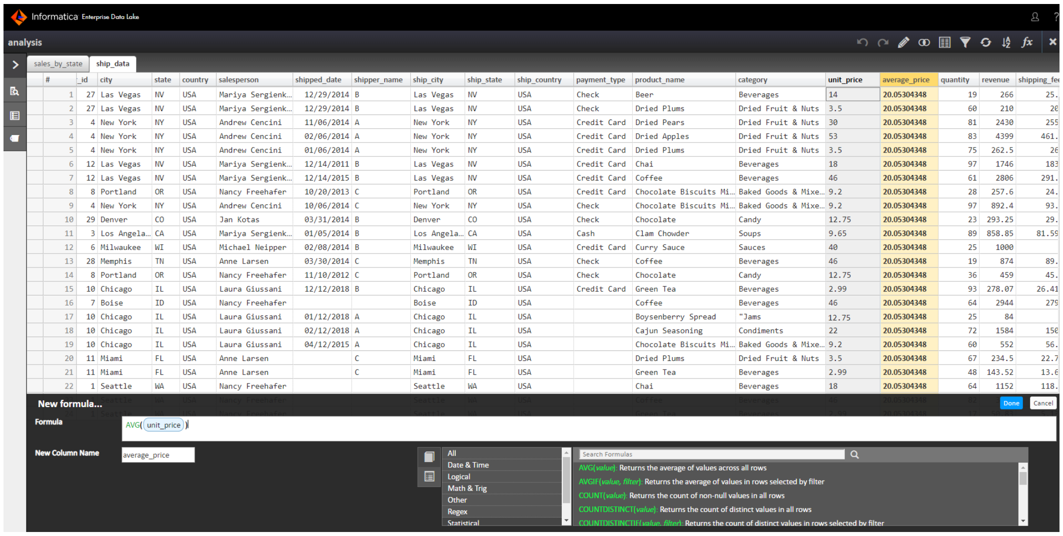
The application provides a variety of functions that you can use in formulas. To help you find the function you need, functions are organized into groups, such as Date and Time functions.

Click the image below to view a video about using formulas:



Example

The following example uses the AVG aggregate function to calculate the average of the values in the unit_price column. The results are displayed in the average column, as shown in the following image:



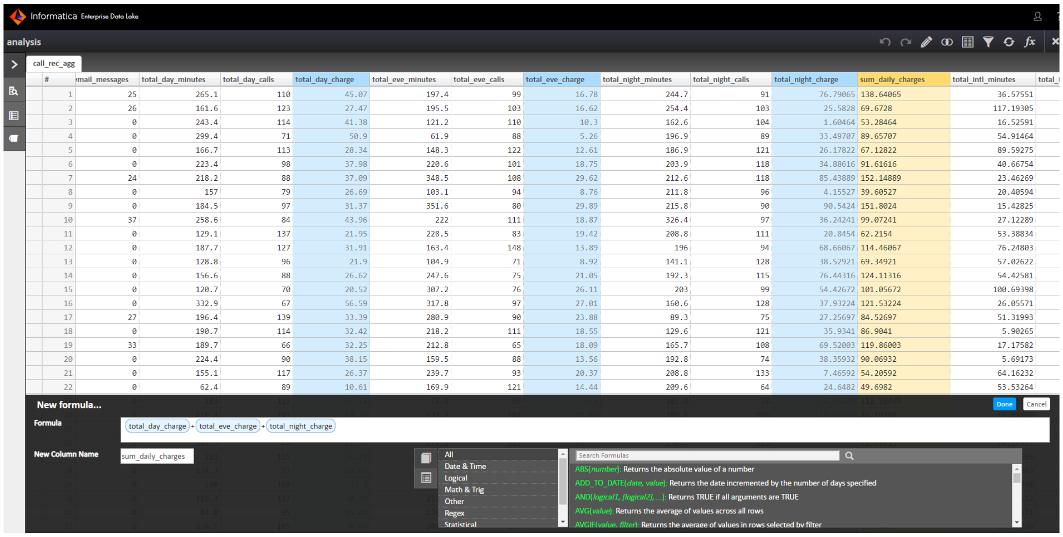
#	_id	city	state	country	salesperson	shipped_date	shipper_name	ship_city	ship_state	ship_country	payment_type	product_name	category	unit_price	average_price	quantity	revenue	shipping_fee
1	27	Las Vegas	NV	USA	Mariya Sergienk...	12/29/2014	B	Las Vegas	NV	USA	Check	Beer	Beverages	14	20.05304348	19	266	25.
2	27	Las Vegas	NV	USA	Mariya Sergienk...	12/29/2014	B	Las Vegas	NV	USA	Check	Dried Plums	Dried Fruit & Nuts	3.5	20.05304348	60	210	2
3	4	New York	NY	USA	Andrew Cencini	11/06/2014	A	New York	NY	USA	Credit Card	Dried Pears	Dried Fruit & Nuts	30	20.05304348	81	2430	255
4	4	New York	NY	USA	Andrew Cencini	02/06/2014	A	New York	NY	USA	Credit Card	Dried Apples	Dried Fruit & Nuts	53	20.05304348	83	4399	461.
5	4	New York	NY	USA	Andrew Cencini	01/06/2014	A	New York	NY	USA	Credit Card	Dried Plums	Dried Fruit & Nuts	3.5	20.05304348	75	262.5	2
6	12	Las Vegas	NV	USA	Mariya Sergienk...	12/14/2011	B	Las Vegas	NV	USA	Credit Card	Chai	Beverages	18	20.05304348	97	1746	18
7	12	Las Vegas	NV	USA	Mariya Sergienk...	12/14/2015	B	Las Vegas	NV	USA	Credit Card	Coffee	Beverages	46	20.05304348	61	2806	291.
8	8	Portland	OR	USA	Nancy Freehafer	10/20/2013	C	Portland	OR	USA	Credit Card	Chocolate Biscuits Mi...	Baked Goods & Mixe...	9.2	20.05304348	28	257.6	24.
9	4	New York	NY	USA	Andrew Cencini	10/06/2014	C	New York	NY	USA	Check	Chocolate Biscuits Mi...	Baked Goods & Mixe...	9.2	20.05304348	97	892.4	93.
10	29	Denver	CO	USA	Jan Kotas	03/31/2014	B	Denver	CO	USA	Check	Chocolate	Candy	12.75	20.05304348	23	293.25	29.
11	3	Los Angela...	CA	USA	Mariya Sergienk...	01/05/2014	B	Los Angela...	CA	USA	Cash	Clam Chowder	Soups	9.65	20.05304348	89	858.85	81.5
12	6	Milwaukee	WI	USA	Michael Neipper	02/08/2014	B	Milwaukee	WI	USA	Credit Card	Curry Sauce	Sauces	40	20.05304348	25	1000	
13	28	Memphis	TN	USA	Anne Larsen	03/30/2014	C	Memphis	TN	USA	Check	Coffee	Beverages	46	20.05304348	19	874	89.
14	8	Portland	OR	USA	Nancy Freehafer	11/10/2012	C	Portland	OR	USA	Check	Chocolate	Candy	12.75	20.05304348	36	459	45.
15	10	Chicago	IL	USA	Laura Giussani	12/12/2018	B	Chicago	IL	USA	Credit Card	Green Tea	Beverages	2.99	20.05304348	93	278.07	26.41
16	7	Boise	ID	USA	Nancy Freehafer			Boise	ID	USA		Coffee	Beverages	46	20.05304348	64	2944	27
17	10	Chicago	IL	USA	Laura Giussani	01/12/2018	A	Chicago	IL	USA		Boysenberry Spread	"Jams	12.75	20.05304348	25	84	
18	10	Chicago	IL	USA	Laura Giussani	02/12/2018	A	Chicago	IL	USA		Cajun Seasoning	Condiments	22	20.05304348	72	1584	15
19	10	Chicago	IL	USA	Laura Giussani	04/12/2015	A	Chicago	IL	USA		Chocolate Biscuits Mi...	Baked Goods & Mixe...	9.2	20.05304348	60	552	56.
20	11	Miami	FL	USA	Anne Larsen		C	Miami	FL	USA		Dried Plums	Dried Fruit & Nuts	3.5	20.05304348	67	234.5	22.7
21	11	Miami	FL	USA	Anne Larsen		C	Miami	FL	USA		Green Tea	Beverages	2.99	20.05304348	48	143.52	13.
22	1	Seattle	WA	USA	Nancy Freehafer			Seattle	WA	USA		Chai	Beverages	18	20.05304348	64	1152	118.

Formula AVG(unit_price)

New Column Name average_price

AVG(value): Returns the average of values across all rows
AVGIF(value, filter): Returns the average of values in rows selected by filter
COUNT(value): Returns the count of non-null values in all rows
COUNTDISTINCT(value): Returns the count of distinct values in all rows
COUNTDISTINCTIF(value, filter): Returns the count of distinct values in rows selected by filter

You can also create your own formulas. The following example adds the total day, evening and night call charges and provides the results in the sum_daily_charges column:



#	mail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge	total_night_minutes	total_night_calls	total_night_charge	sum_daily_charges	total_intl_minutes
1	25	265.1	110	45.07	197.4	99	16.78	244.7	91	76.79065	138.64065	36.57551
2	26	161.6	123	27.47	195.5	103	16.62	254.4	103	25.5828	69.6728	117.19305
3	0	243.4	114	41.38	121.2	110	10.3	162.6	104	1.60464	53.28464	16.52591
4	0	299.4	71	50.9	61.9	88	5.26	196.9	89	33.49707	89.65707	54.91464
5	0	166.7	113	28.34	148.3	122	12.61	186.9	121	26.17822	67.12822	89.59275
6	0	223.4	98	37.98	220.6	101	18.75	203.9	118	34.88616	91.61616	40.66754
7	24	218.2	88	37.09	348.5	108	29.62	212.6	118	85.43889	152.14889	23.46269
8	0	157	79	26.69	103.1	94	8.76	211.8	96	4.15527	39.60527	20.40594
9	0	184.5	97	31.37	351.6	80	29.89	215.8	90	90.5424	151.8024	15.42825
10	37	258.6	84	43.96	222	111	18.87	326.4	97	36.24241	99.07241	27.12289
11	0	129.1	137	21.95	228.5	83	19.42	208.8	111	20.8454	62.2154	53.38834
12	0	187.7	127	31.91	163.4	148	13.89	196	94	68.66067	114.46067	76.24803
13	0	128.8	96	21.9	104.9	71	8.92	141.1	128	38.52921	69.34921	57.02622
14	0	156.6	88	26.62	247.6	75	21.05	192.3	115	76.44316	124.11316	54.42581
15	0	120.7	70	20.52	307.2	76	26.11	203	99	54.42672	101.05672	100.69398
16	0	332.9	67	56.59	317.8	97	27.01	160.6	128	37.93224	121.53224	26.05571
17	27	196.4	139	33.39	280.9	90	23.88	89.3	75	27.25697	84.52697	51.31993
18	0	190.7	114	32.42	218.2	111	18.55	129.6	121	35.9341	86.9041	5.90265
19	33	189.7	66	32.25	212.8	65	18.09	165.7	108	69.52003	119.86003	17.17582
20	0	224.4	90	38.15	159.5	88	13.56	192.8	74	38.35932	90.06932	5.69173
21	0	155.1	117	26.37	239.7	93	20.37	208.8	133	7.46592	54.20592	64.16232
22	0	62.4	89	10.61	169.9	121	14.44	209.6	64	24.6482	49.6982	53.53264

New formula...

Formula total_day_charge + total_eve_charge + total_night_charge

New Column Name sum_daily_charges

ABS(*number*): Returns the absolute value of a number

ADD_TO_DATE(*date, value*): Returns the date incremented by the number of days specified

AND(*logical1, [logical2], ...*): Returns TRUE if all arguments are TRUE

AVG(*value*): Returns the average of values across all rows

AVGIF(*value, filter*): Returns the average of values in rows selected by filter

Applying a Formula

Select one or more functions to use in a formula, or create your own formula.

1. On the project page, click **Prepare**.
2. Select the worksheet to which you want to apply the formula.
3. Click the **Functions** icon, and then select **New formula**.
4. Enter the name of the column containing the formula results to add to the worksheet.
5. Select a function to use in the formula from the list, or type a custom formula in the **Formula** field.

   To add a column to the formula, start typing the name of the column, and then select the column name from the list.
6. Click **Done** to apply the formula.

Limitations and Constraints

Enterprise Data Lake uses lenient parsing of input values in the DATE and DATETIME formulas. If you specify a value for a DATE or DATETIME formula that contains a value that is out of range, the application returns a value within the preceding or following unit of time.

For example, if you specify DATE(2018, 13, 30), the application returns 01/30/2019, because 13 is outside of the valid range for the month argument.

In addition, you cannot use nested aggregate functions in a worksheet. For example, the following nested aggregate function is not valid:

SUM(AVG(quantity) + AVG(unit_price))

Using Window Functions

You can apply window functions in Enterprise Data Lake to groups of rows within a worksheet.

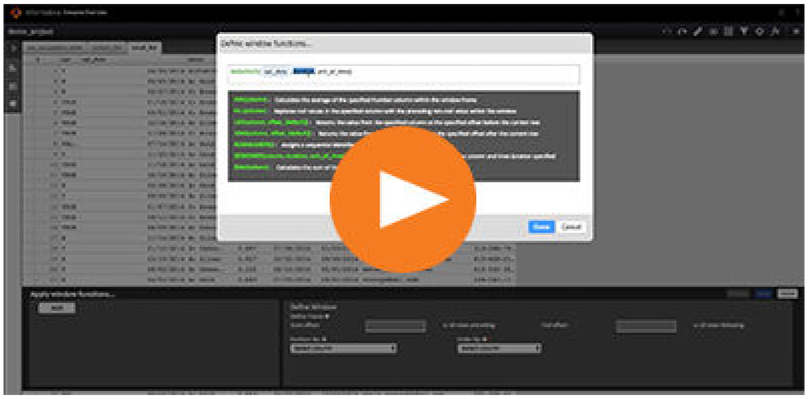
The group of rows on which a function operates is called a window. A window function calculates a return value for every input row within a window.

You can use window functions to perform the following tasks:

- Retrieve data from a previous row or a subsequent row.
- Calculate a sum or an average based on a group of rows.
- Assign a sequential row number to each row in a group of rows.
- Replace null values in rows with the preceding non-null value within a group of rows.
- Generate session identifiers you can use to group rows based on a specific time period, such as web site visits recorded in a log file.

You can apply multiple window functions to a worksheet. For example, you might apply a function to calculate the sum of values for each row preceding or following the current row within a window, and apply another function to calculate the average of the same values. Enterprise Data Lake adds a column containing the results of each function you apply to the worksheet.

Click the image below to view a video about applying window functions:



Window Configuration

When you use a window function, you configure the windowing properties associated with the function. Windowing properties define the frame, partitioning, and ordering boundaries associated with a particular input row.

Windowing properties define the following:

- A frame that defines the boundaries associated with a particular input row. You can define a frame if you apply a SUM or AVG function to a worksheet.
- Partition by columns containing the categorical values to use to group rows into windows.
- Order by columns that define the order of rows in each window.

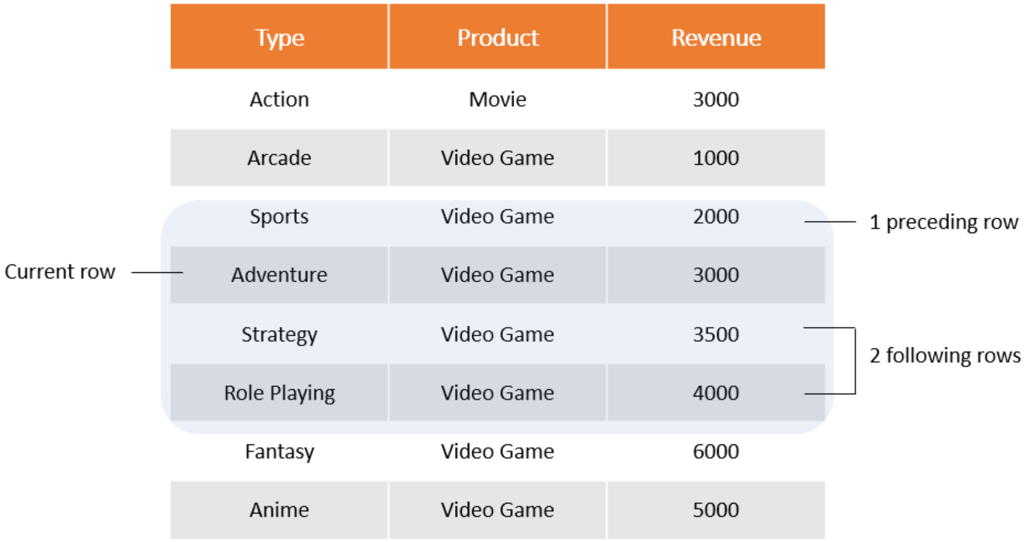
Frame

The frame definition determines which rows are included in the calculation for the current input row, based on the relative position of the rows to the current row.

You can configure frames when you use the SUM and AVG functions. The application ignores frame definitions when you use all other functions.

The start offset defines the number of rows before or after the current row at which to begin the calculation. The end offset describes the number of rows after the current input row at which to end the calculation. For example, a start offset of -1 and an end offset of 2 describes a frame including the row before the current row, the current input row, and the two rows after the current row.

The following image shows a frame with a start offset of -1 and an end offset of 2:



Type	Product	Revenue
Action	Movie	3000
Arcade	Video Game	1000
Sports	Video Game	2000
Adventure	Video Game	3000
Strategy	Video Game	3500
Role Playing	Video Game	4000
Fantasy	Video Game	6000
Anime	Video Game	5000

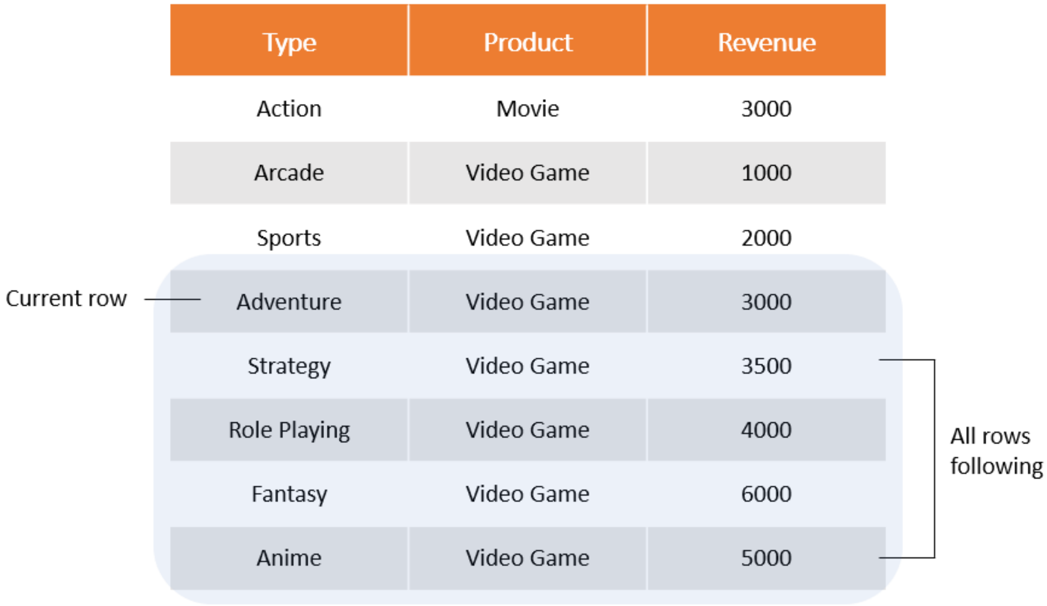
1 preceding row
Current row
2 following rows

You can also specify a frame that does not include the current input row. For example, a start offset of 2 and an end offset of 4 describes a frame that includes three rows, from the second to the fourth row after the current row.

Note: The start offset must be less than or equal to the end offset.

Offsets of **All Rows Preceding** and **All Rows Following** represent the first row of the partition and the last row of the partition. For example, if the start offset is All Rows Preceding and the end offset is -1, the frame includes one row before the current row, and all rows before it.

The following image shows a frame with a start offset of 0, indicating that the start offset is the current row, and an end offset of All Rows Following:



Type	Product	Revenue
Action	Movie	3000
Arcade	Video Game	1000
Sports	Video Game	2000
Adventure	Video Game	3000
Strategy	Video Game	3500
Role Playing	Video Game	4000
Fantasy	Video Game	6000
Anime	Video Game	5000

Current row
All rows following

Partition By and Order By Columns

You use partition by columns to group rows into windows, and use order columns to define the order or sequence of rows within each window.

Partition by columns

Partition by columns define window boundaries, based on categorical values within the columns. A window function operates across the rows that fall into the same window as the current row.

Partition by columns are optional for all functions. If you do not specify partition by columns, the function treats the entire data set as a window. You must select columns containing categorical values as partition by columns.

You can select multiple partition by columns to apply to a window function. The application creates a window based on each unique combination of categorical values in the selected columns.

Order by columns

Order by columns determine the order of the rows within a window.

Order by columns are required for all functions except AVG and SUM. For the AVG and SUM functions, specifying order by columns is optional when you select **All Rows Preceding** and **All Rows Following** as the offsets in the frame definition.

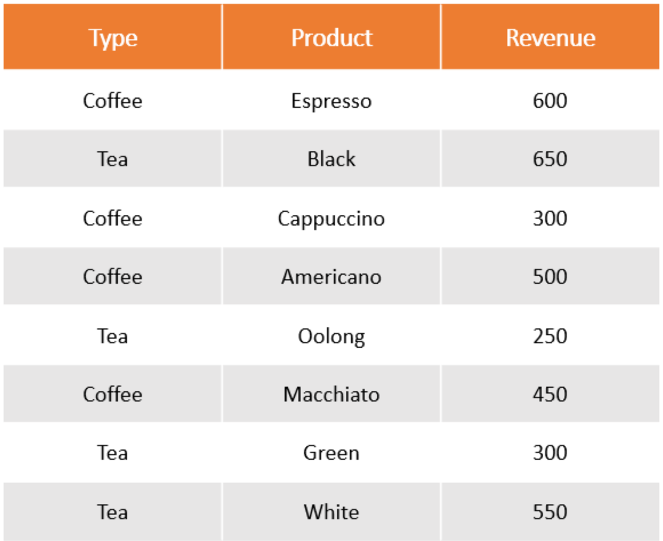
You can select multiple order by columns to apply to a window function. The application orders rows based on precedence of the order by columns selected.

When you apply an order by column to a function, you choose to arrange the data in ascending or descending order.

Example

In this example, you want to calculate the revenue for coffee and tea products.

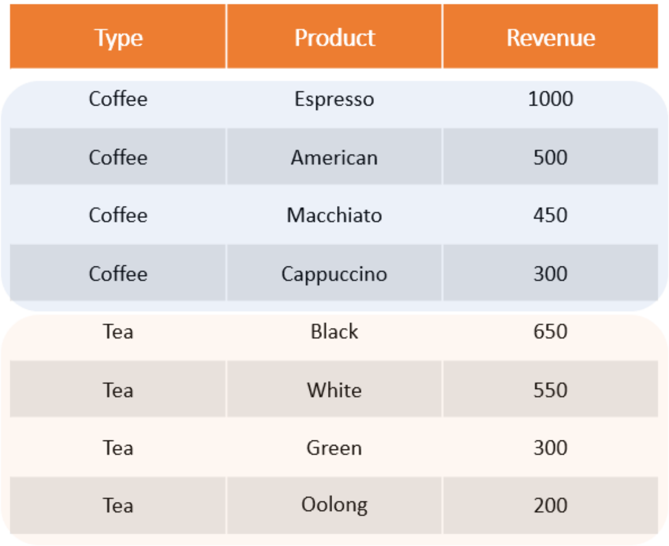
The following image shows a table listing the product types, the corresponding product categories, and the revenue from each product:



Type	Product	Revenue
Coffee	Espresso	600
Tea	Black	650
Coffee	Cappuccino	300
Coffee	Americano	500
Tea	Oolong	250
Coffee	Macchiato	450
Tea	Green	300
Tea	White	550

Select the Type column, which contains categorical values, as the partition by column. You select the Revenue column as the order by column, and choose to order the data by descending revenue.

The following image shows the data grouped into two windows based on the values in the Type column. Within each window, the revenue data is displayed in descending order:



Type	Product	Revenue
Coffee	Espresso	1000
Coffee	American	500
Coffee	Macchiato	450
Coffee	Cappuccino	300
Tea	Black	650
Tea	White	550
Tea	Green	300
Tea	Oolong	200

Supported Window Functions

AVG

The AVG function calculates the average of the values preceding or following the current row within the specified window.

You can apply the function to Number columns in a worksheet. The function adds a column containing the calculation results to the worksheet.

You must specify the Start Offset and End Offset properties in a window frame definition to define the rows preceding or following the current row to include in the calculation. The function iterates through each row within the window, performing the calculation based on the offset values specified.

For example, if you set Start Offset to 2 and End Offset to 4, the function calculates the average of the values in the second, third, and fourth rows after the current row.

To set the current row as the start offset, enter 0 as the value for the Start Offset property. To set a row before the current row as the start offset, enter a negative number.

For example, if you set Start Offset to -2 and End Offset to 4, the function calculates the average of the values in the current row, the two rows before the current row, and the four rows after the current row.

Partition by columns are optional. Specifying order by columns is optional when you select **All Rows Preceding** and **All Rows Following** as the offsets in the frame definition; otherwise, you must specify order by columns.

Syntax

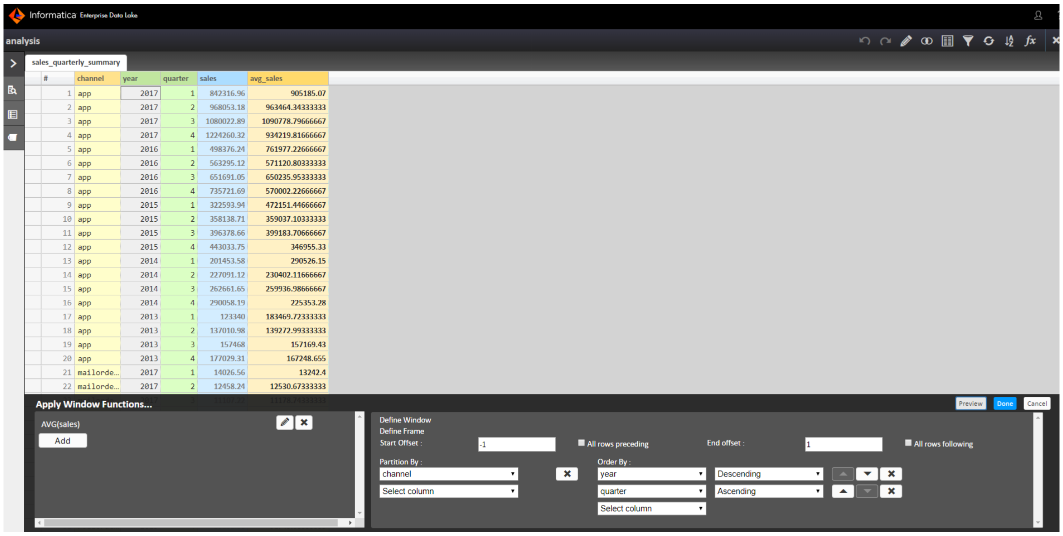
```
AVG (Column)
```

The following table describes the function arguments:

Argument	Required/Optional	Description
Column	Required	The name of the column containing the values to include in the calculation. The column must be of the Number type.

Example

The following example calculates the average sales per channel for the current, previous, and following quarters. To include the rows for the previous and following quarters, set the offset values to -1 and 1.



#	channel	year	quarter	sales	avg_sales
1	app	2017	1	842316.96	905185.07
2	app	2017	2	968053.18	963464.34333333
3	app	2017	3	1080022.89	1090778.79666667
4	app	2017	4	1224260.32	934219.81666667
5	app	2016	1	498376.24	761977.22666667
6	app	2016	2	563295.12	571120.80333333
7	app	2016	3	651691.05	650235.95333333
8	app	2016	4	735721.69	570002.22666667
9	app	2015	1	322593.94	472151.44666667
10	app	2015	2	358138.71	359037.10333333
11	app	2015	3	396378.66	399183.70666667
12	app	2015	4	443033.75	346955.33
13	app	2014	1	201453.58	290526.15
14	app	2014	2	227091.12	230402.11666667
15	app	2014	3	262661.65	259936.98666667
16	app	2014	4	290058.19	225353.28
17	app	2013	1	123340	183469.72333333
18	app	2013	2	137010.98	139272.99333333
19	app	2013	3	157468	157169.43
20	app	2013	4	177029.31	167248.655
21	mailorde...	2017	1	14026.56	13242.4
22	mailorde...	2017	2	12458.24	12530.67333333

Define Window
Define Frame
Start Offset : -1
☐ All rows preceding
End offset : 1
☐ All rows following
Partition By : channel
Select column
Order By : year Descending
quarter Ascending
Select column

Enter the following function arguments:

- Column: sales

Enter the following window properties:

- Start Offset: -1
- End Offset: 1

- Partition By column: sales
- Order By columns: year, quarter

FILL

The FILL function replaces null values in rows in the specified column with the preceding non-null value within a window, based on the partition by and order by columns specified. You must specify at least one order by column.

Syntax

```
FILL (Column)
```

The following table describes the function arguments:

Argument	Required/Optional	Description
Column	Required	The name of the column name within which to replace null values.

LAG

The LAG function returns the value from the specified column at the specified offset before the current row.

Use the LAG function to compare the values in the current row with the values in a preceding row. You must specify at least one order by column.

Use the offset parameter to specify the number of rows before the current row for which to return the value. For example, set the offset parameter to 3 to return the value of the row that is three rows before the current row.

Syntax

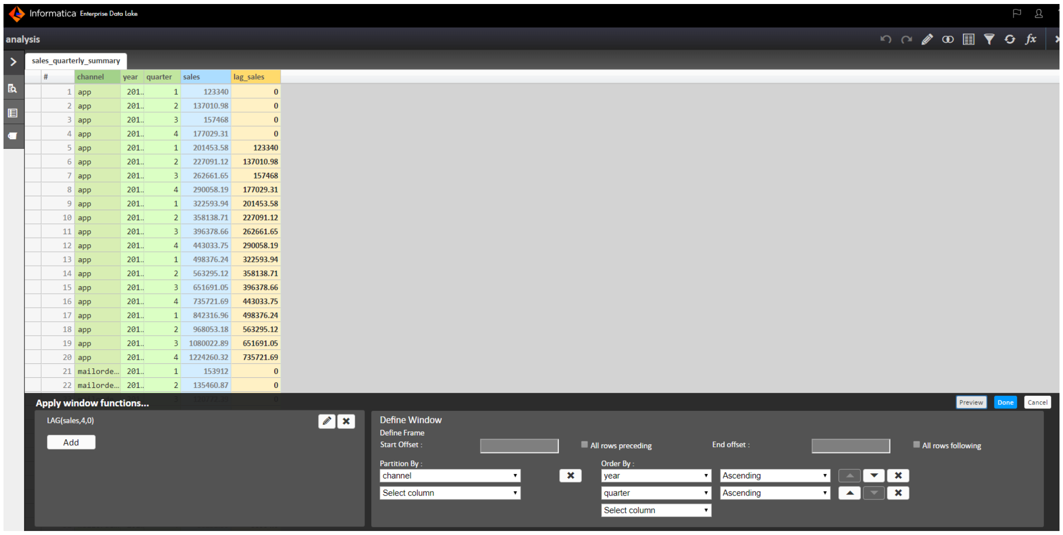
```
LAG (Column, Offset, [Default])
```

The following table describes the function arguments:

Argument	Required/ Optional	Description
Column	Required	The name of the column name containing the value to retrieve.
Offset	Required	The number of rows preceding the current row from which to retrieve the data. The value must be a positive number.
Default	Optional	The default value to return if a row does not exist at the specified offset value. If you apply the function to a Date column, enter the parameter value in the in the following formats: 'mm/dd/yyyy' or 'mm/dd/yyyy hh24:mi:ss' If you apply the value to a Text column, the value must be enclosed in quotes. For example: 'Annual' If you do not specify a value, the default value is null. Remove this argument from the function if it is not needed.

Example

The following example compares sales per product channel for each quarter with sales for the same quarter in the previous year. Enter 0 as the default value to return for rows that are not within the specified offset value.



#	channel	year	quarter	sales	lag_sales
1	app	201..	1	123340	0
2	app	201..	2	137010.98	0
3	app	201..	3	157468	0
4	app	201..	4	177029.31	0
5	app	201..	1	201453.58	123340
6	app	201..	2	227091.12	137010.98
7	app	201..	3	262661.65	157468
8	app	201..	4	290058.19	177029.31
9	app	201..	1	322593.94	201453.58
10	app	201..	2	358138.71	227091.12
11	app	201..	3	396378.66	262661.65
12	app	201..	4	443033.75	290058.19
13	app	201..	1	498376.24	322593.94
14	app	201..	2	563295.12	358138.71
15	app	201..	3	651691.05	396378.66
16	app	201..	4	735721.69	443033.75
17	app	201..	1	842316.96	498376.24
18	app	201..	2	968053.18	563295.12
19	app	201..	3	1080022.89	651691.05
20	app	201..	4	1224260.32	735721.69
21	mailorde…	201..	1	153912	0
22	mailorde…	201..	2	135460.87	0

LAG(sales,4,0)

Start Offset :
☐ All rows preceding
End offset :
☐ All rows following

Partition By : channel
Order By : year Ascending
Order By : quarter Ascending

Enter the following function arguments:

- Column: sales
- Offset: 4
- Default: 0

Enter the following window properties:

- Partition By column: sales
- Order By columns: year, quarter

LEAD

The LEAD function returns the value from the specified column at the specified offset after the current row.

Use the LEAD function to compare the values in the current row with the values in a following row. You must specify at least one order by column.

Use the offset parameter to specify the number of rows after the current row for which to return the value. For example, set the offset parameter to 3 to return the value of the row that is three rows after the current row.

Syntax

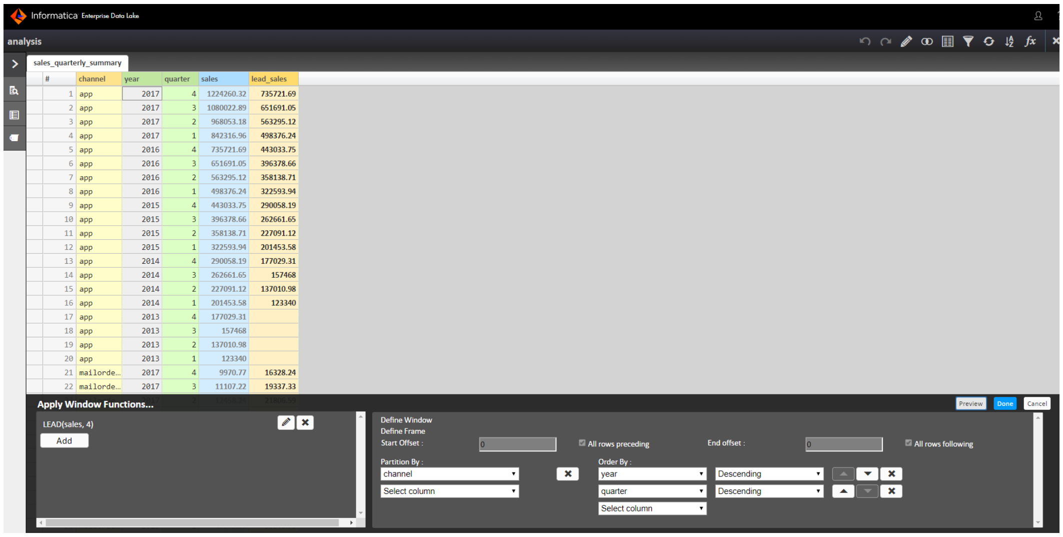
```
LEAD (Column, Offset, [Default])
```

The following table describes the function arguments:

Argument	Required/ Optional	Description
Column	Required	The name of the column name containing the value to retrieve.
Offset	Required	The number of rows preceding the current row from which to retrieve the data. The value must be a positive number.
Default	Optional	The default value to return if a row does not exist at the specified offset value. If you apply the function to a Date column, enter the parameter value in the in the following formats: 'mm/dd/yyyy' or 'mm/dd/yyyy hh24:mi:ss' If you apply the value to a Text column, the value must be enclosed in quotes. For example: 'Annual' If you do not specify a value, the default value is null. Remove this argument from the function if it is not needed.

Example

The following example compares sales per product channel for each quarter with sales for the same quarter in the following year. Remove the default argument from the function to return null values for rows that are not within the specified offset value.



#	channel	year	quarter	sales	lead_sales
1	app	2017	4	1224260.32	735721.69
2	app	2017	3	1080022.89	651691.05
3	app	2017	2	968053.18	563295.12
4	app	2017	1	842316.96	498376.24
5	app	2016	4	735721.69	443033.75
6	app	2016	3	651691.05	396378.66
7	app	2016	2	563295.12	358138.71
8	app	2016	1	498376.24	322593.94
9	app	2015	4	443033.75	290058.19
10	app	2015	3	396378.66	262661.65
11	app	2015	2	358138.71	227091.12
12	app	2015	1	322593.94	201453.58
13	app	2014	4	290058.19	177029.31
14	app	2014	3	262661.65	157468
15	app	2014	2	227091.12	137010.98
16	app	2014	1	201453.58	123340
17	app	2013	4	177029.31	
18	app	2013	3	157468	
19	app	2013	2	137010.98	
20	app	2013	1	123340	
21	mailorde…	2017	4	9970.77	16328.24
22	mailorde…	2017	3	11107.22	19337.33

Apply Window Functions...

LEAD(sales, 4)

Define Window
Define Frame
Start Offset : 0 ☑ All rows preceding
End offset : 0 ☑ All rows following
Partition By : channel
Order By : year Descending
quarter Descending

Enter the following function arguments:

- Column: sales
- Offset: 4
- Default: remove this argument from the function

Enter the following window properties:

- Partition By column: sales
- Order By columns: year, quarter

ROWNUMBER

The ROWNUMBER function assigns a sequential row number, starting at 1, to each row within a window.

The function adds a column containing the row identifiers to the worksheet.

When the function reaches the end of a window, it assigns identifiers to the rows in the next window, starting at 1.

Syntax

```
ROWNUMBER ()
```

The function does not take any arguments.

SESSIONIZE

The SESSIONIZE function generates session identifiers based on the Date column and the period of time specified. You can use the generated session identifiers to group rows based on a specific time period, such as web site visits recorded in a log file.

You can apply the function to Date columns in a worksheet. The function adds a column containing the session identifiers to the worksheet.

When you configure the function properties, you must select the name of the Date column you use in the function as the first order by column. You must also configure the order by column to list rows in ascending order.

Syntax

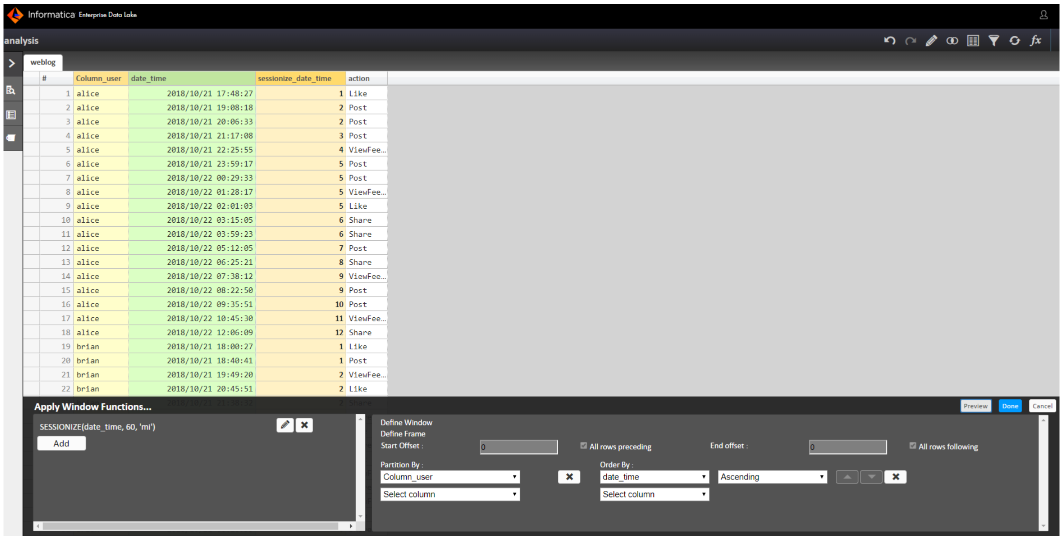
```
SESSIONIZE (Column name, Duration, Unit of time)
```

The following table describes the function arguments:

Argument	Required/Optional	Description
Column	Required	The column to which to apply the function. The column must be of the Date type. You must select the same column as the value of the Order By property.
Duration	Required	The time period by which to group rows included in a session. The value must be greater than 0. You can specify fractional values, such as 1.5.
Unit of time	Required	The unit of time to apply to the time interval value. Valid values are: - 'ms': milliseconds - 'ss': seconds - 'mi': minutes - 'hh': hours - 'dd': days Enclose values in quotes.

Example

The following example groups user visits to a particular web site recorded in a log file based on the date and time noted in the date_visited column. The function generates a new session identifier when the date_visited column value is greater than the previous value plus the specified time period.



#	Column_user	date_time	sessionize_date_time	action
1	alice	2018/10/21 17:48:27	1	Like
2	alice	2018/10/21 19:08:18	2	Post
3	alice	2018/10/21 20:06:33	2	Post
4	alice	2018/10/21 21:17:08	3	Post
5	alice	2018/10/21 22:25:55	4	ViewFee...
6	alice	2018/10/21 23:59:17	5	Post
7	alice	2018/10/22 00:29:33	5	Post
8	alice	2018/10/22 01:28:17	5	ViewFee...
9	alice	2018/10/22 02:01:03	5	Like
10	alice	2018/10/22 03:15:05	6	Share
11	alice	2018/10/22 03:59:23	6	Share
12	alice	2018/10/22 05:12:05	7	Post
13	alice	2018/10/22 06:25:21	8	Share
14	alice	2018/10/22 07:38:12	9	ViewFee...
15	alice	2018/10/22 08:22:50	9	Post
16	alice	2018/10/22 09:35:51	10	Post
17	alice	2018/10/22 10:45:30	11	ViewFee...
18	alice	2018/10/22 12:06:09	12	Share
19	brian	2018/10/21 18:00:27	1	Like
20	brian	2018/10/21 18:40:41	1	Post
21	brian	2018/10/21 19:49:20	2	ViewFee...
22	brian	2018/10/21 20:45:51	2	Like

SESSIONIZE(date_time, 60, 'mi')

Define Window
Define Frame
Start Offset : 0 ☑ All rows preceding
End offset : 0 ☑ All rows following
Partition By : Column_user
Select column
Order By : date_time Ascending
Select column

Enter the following function arguments:

- Column: date_time
- Time Interval: 60
- Unit of time: 'mi'

Enter the following window properties. Note that you must set the value specified in the Column argument as the value for the first Order By column. You must also set the Order By column to list results in ascending order:

- Partition By column: column_user
- Order By columns: date_time

SUM

The SUM function calculates the sum of values preceding or following the current row within the specified window.

You can apply the function to Number columns in a worksheet. The function adds a column containing the calculation results to the worksheet.

You must specify the Start Offset and End Offset properties in the window frame definition to define the rows preceding or following the current row to include in the calculation. The function iterates through each row within the window, performing the calculation based on the offset values specified.

For example, if you set Start Offset to 2 and End Offset to 4, the function calculates the sum of the values in the second, third, and fourth rows after the current row.

To set the current row as the start offset, enter 0 as the value for the Start Offset property. To set a row before the current row as the start offset, enter a negative number.

For example, if you set Start Offset to -2 and End Offset to 4, the function calculates the sum of the values in the current row, the two rows before the current row, and the four rows after the current row.

Partition by columns are optional. Specifying order by columns is optional when you select **All Rows Preceding** and **All Rows Following** as the offsets in the frame definition; otherwise, you must specify order by columns.

Syntax

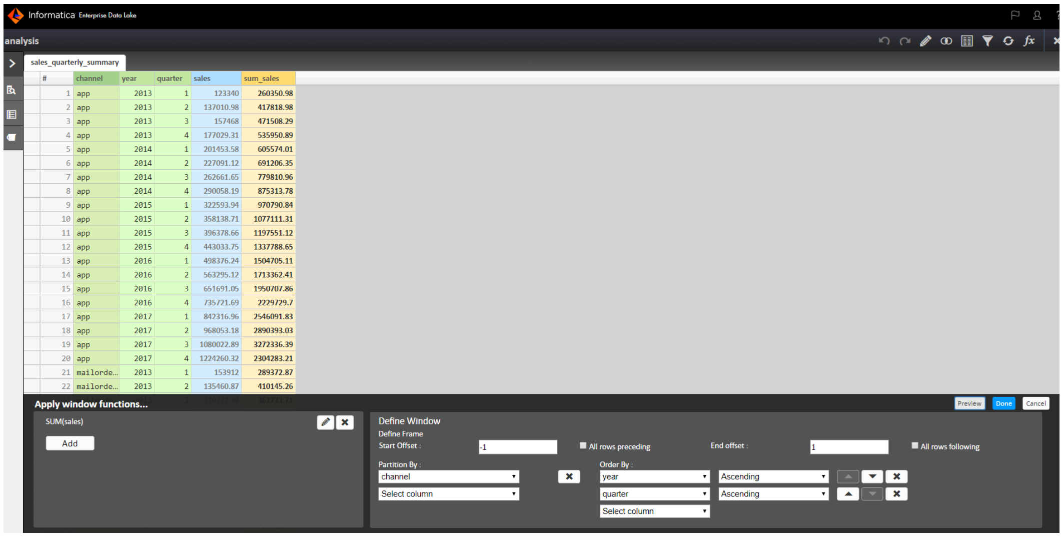
```
SUM (Column)
```

The following table describes the function arguments:

Argument	Required/Optional	Description
Column	Required	The name of the column containing the values to include in the calculation. The column must be of the Number type.

Example

The following example calculates the total sales per channel for the current, previous, and following quarters. To include the rows for the previous and following quarters, set the offset values to -1 and 1.



#	channel	year	quarter	sales	sum_sales
1	app	2013	1	123340	260350.98
2	app	2013	2	137010.98	417818.98
3	app	2013	3	157468	471508.29
4	app	2013	4	177029.31	535950.89
5	app	2014	1	201453.58	605574.01
6	app	2014	2	227091.12	691206.35
7	app	2014	3	262661.65	779810.96
8	app	2014	4	290058.19	875313.78
9	app	2015	1	322593.94	970790.84
10	app	2015	2	358138.71	1077111.31
11	app	2015	3	396378.66	1197551.12
12	app	2015	4	443033.75	1337788.65
13	app	2016	1	498376.24	1504705.11
14	app	2016	2	563295.12	1713362.41
15	app	2016	3	651691.05	1950707.86
16	app	2016	4	735721.69	2229729.7
17	app	2017	1	842316.96	2546091.83
18	app	2017	2	968053.18	2890393.03
19	app	2017	3	1080022.89	3272336.39
20	app	2017	4	1224260.32	2304283.21
21	mailorde...	2013	1	153912	289372.87
22	mailorde...	2013	2	135460.87	410145.26

SUM(sales)

Define Window
Define Frame
Start Offset : -1
☐ All rows preceding
End offset : 1
☐ All rows following
Partition By : channel
Order By : year Ascending
quarter Ascending

Enter the following function arguments:

- Column: sales

Enter the following window properties:

- Start Offset: -1
- End Offset: 1
- Partition By column: sales
- Order By columns: year, quarter

Applying Window Functions

You can apply one or more window functions to a worksheet.

For each function, optionally select one or more partition by columns to create windows from groups of rows. Select one or more order by columns to specify the order or sequence of rows within each window.

If you apply an AVG function or a SUM function, you must specify the rows preceding and following the current row to include in the frame within which to perform the calculation. The application ignores frame definitions for all other window functions.

If you select **All Rows Preceding** and **All Rows Following** in the frame definition for an AVG function or a SUM function, you do not need to select an order by column.

1. On the project page, click **Prepare**.
2. Select the worksheet to which you want to apply the functions.
3. Click the **Functions** icon, and then select **Apply window functions**.
4. Click **Add**.
5. Click or type the name of the function you want to apply.
6. Click each function argument to enter a value. To specify a column value, start typing the name of the column, and then select the column name from the list.
7. Click **Done**.

   Enter the window properties in the Define Window panel.
8. If you apply an AVG function or a SUM function, enter the number of rows before or after the current row at which to begin the calculation from the **Start Offset** menu.

   To begin the calculation at a row before the current row, enter a negative number.

   To include all rows before the current row in the calculation, select **All Rows Preceding**.
9. If you apply an AVG function or a SUM function, select the number of rows following the current row to include in the calculation from the **End Offset** menu.

   To include all rows following the current row in the calculation, select **All Rows Following**.
10. Select a column to use as a partition key from the **Partition By** menu.

    The window function operates across the rows that fall into the same group as the current row.
11. Select a column to use to specify how rows in the group or partition are ordered in the function results from the **Order By** menu.

    If you do not select an order by column, the rows have no particular order.
12. Select whether to order rows in the output column in ascending or descending order.
13. Click **Preview** to preview the function results in the worksheet.
14. Click **Done** to apply the function.

Using Rules

You can apply rules during data preparation to help you cleanse, transform, or validate data. Rules provide you with consistent, complex logic for transforming data that can be reused across projects.

Enterprise Data Lake provides hundreds of rules that you can apply to data. Rules can also be created by developers using the Administrator tool, or by analysts using the Analyst tool. To use rules, your Informatica administrator must deploy rules-related services and content in your Informatica installation. Your Informatica administrator must also give you privileges to access rules.

You can apply two types of rules: passive rules and active rules.

Passive Rules

A passive rule operates on columns in the current worksheet. The application adds columns containing the rule output to the current worksheet.

Active Rules

An active rule uses all rows within a data set as input. You can select multiple worksheets containing data to use as inputs to the rule. The application adds a worksheet containing the rule output to the project.

An active rule contains one or more input groups. For each input group, you select a worksheet that contains the data that you want to use in the rule. You can select the same worksheet for multiple input groups, or you can select a different worksheet for each group.

You map inputs in the input group to columns in the worksheet that contain the data you want to pass to the rule. You can also specify a constant value as an input, instead of a column name.

An active rule might also include parameters that are passed to the rule at runtime. Each parameter has a default value that you can modify.

An active rule contains one or more output groups. Each output group contains the names of columns containing the rule results that are added to the generated worksheet. You can select a single output group to use in an active rule.

Using Passive Rules

A passive rule operates on columns in the current worksheet. The application adds columns containing the rule output to the current worksheet.

When you apply a rule, Enterprise Data Lake adds the rule as a step to the worksheet recipe. You can edit the rule in the recipe, or remove it from the recipe.

1. On the project page, click **Prepare**.
2. Select the worksheet to which you want to apply the rule.
3. Click the **Functions** icon, and then select **Apply rule**.
4. Select a rule from the menu.

   To find the rule you want to use, start typing the name of the rule.
5. Start typing the name of a column to which you want to apply the rule.
6. Select the name of the column to add to the worksheet containing the rule output.
7. Click **Apply**.

Using Active Rules

An active rule uses all of the rows within a data set as input. You can select multiple worksheets to use as input sources. Enterprise Data Lake adds a worksheet containing the output columns chosen from the output group to the project.

When you apply a rule, the application adds the rule as a step to the worksheet recipe. You can edit the step created for the rule in the recipe, or you can remove the step from the recipe.

You can configure and run one active rule at a time.

1. On the project page, click **Prepare**.
2. Click the **Functions** icon, and then select **Apply rule**.
3. Select an active rule from the list.

To find the rule you want to use, start typing the name of the rule.

4. Select a worksheet to use as an input source.

   The input group displays the columns in the worksheet that contain values you can use as inputs to the rule.

5. Provide the inputs to the rule. You can select the names of columns in the worksheet, or enter constant values.

   - To use a column in the worksheet as an input for a field, select the column name from the menu. Select at least one column in each input group.
   - To use a constant value as an input for a field, select **Enter a Constant Value** from the menu, and then enter the value. For more information about valid values, see "Valid Rule Inputs" on page 67.

6. If the rule contains parameters, the data type and default value of each parameter applied in the rule definition appear. To change a parameter value, click the **edit** icon next to the parameter.

7. Expand an output group.

   The output group displays the columns that you can include in the generated worksheet containing the rule results. If the rule contains multiple output groups, select only one output group.

8. Select the output column associated with each input. You must select an output column for each input you include in the rule.

9. Click **Run**.

   Enterprise Data Lake runs the rule and adds a worksheet containing the rule output to the project.

Valid Rule Inputs

You can use the name of a column in the worksheet as a rule input in a worksheet. You can also use strings, dates and numbers as constant value inputs.

The following table describes the valid formats for values you can use as rule inputs:

Value Type	Valid Format
Number	Up to 19 numeric characters and a single decimal point. For example, 3.15 is valid; 3.1.5 is not valid.
String	No restrictions.

Blending Data

You can create blended data sets by combining, importing, or merging data in worksheets within a project. Enterprise Data Lake provides you with several methods for blending data.

You can use the join operation to combine data from two different worksheets based on a common column containing identical distinct values. Similarly, you can perform a lookup to import selected columns from another worksheet into your current worksheet based on a common column in both worksheets that contains identical distinct values.

You can use the union operation to merge data from one worksheet into another worksheet. When you perform a union, you match the columns in both worksheets that contain similar data.

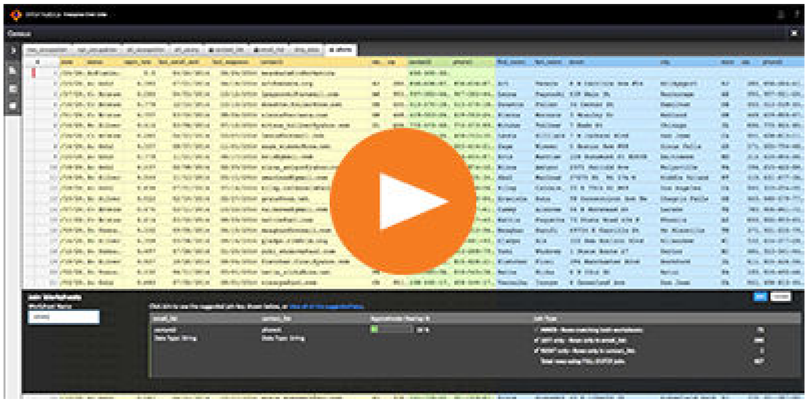
You can categorize or group similar values in a column to make analysis easier. For example, you might want to group values in a column containing the color of various item according to the corresponding primary color, such as red, blue or yellow.

Joining Worksheets

You can use the join operation to combine data from two different worksheets within a project. The operation adds a new worksheet containing the combined data to the project.

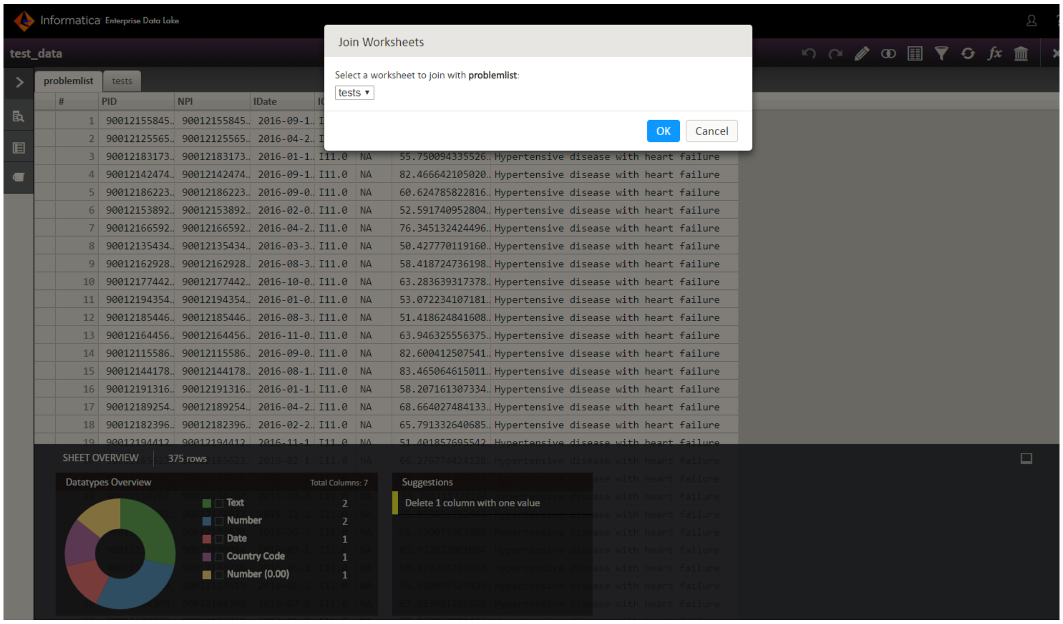
When you perform a join, you specify columns to use in each worksheet that contain the same distinct values. Both worksheets must contain at least one column with identical distinct values. For example, if both worksheets contain a column with a customer identifier, you can use these columns to combine columns containing data associated with each customer identifier in a single worksheet.

Click the image below to view a video about the join operation:



Example

The following image shows a worksheet selected to use in a join:



Join Worksheets

Select a worksheet to join with **problemlist**:

tests ▾

OK Cancel

#	PID	NPI	IDate	I			
1	90012155845..	90012155845..	2016-09-1..	I			
2	90012125565..	90012125565..	2016-04-2..	I			
3	90012183173..	90012183173..	2016-01-1..	I11.0	NA	55.750094335526..	Hypertensive disease with heart failure
4	90012142474..	90012142474..	2016-09-1..	I11.0	NA	82.466642105020..	Hypertensive disease with heart failure
5	90012186223..	90012186223..	2016-09-0..	I11.0	NA	60.624785822816..	Hypertensive disease with heart failure
6	90012153892..	90012153892..	2016-02-0..	I11.0	NA	52.591740952804..	Hypertensive disease with heart failure
7	90012166592..	90012166592..	2016-04-2..	I11.0	NA	76.345132424496..	Hypertensive disease with heart failure
8	90012135434..	90012135434..	2016-03-3..	I11.0	NA	50.427770119160..	Hypertensive disease with heart failure
9	90012162928..	90012162928..	2016-08-3..	I11.0	NA	58.418724736198..	Hypertensive disease with heart failure
10	90012177442..	90012177442..	2016-10-0..	I11.0	NA	63.283639317378..	Hypertensive disease with heart failure
11	90012194354..	90012194354..	2016-01-0..	I11.0	NA	53.072234107181..	Hypertensive disease with heart failure
12	90012185446..	90012185446..	2016-08-3..	I11.0	NA	51.418624841608..	Hypertensive disease with heart failure
13	90012164456..	90012164456..	2016-11-0..	I11.0	NA	63.946325556375..	Hypertensive disease with heart failure
14	90012115586..	90012115586..	2016-09-0..	I11.0	NA	82.600412507541..	Hypertensive disease with heart failure
15	90012144178..	90012144178..	2016-08-1..	I11.0	NA	83.465064615011..	Hypertensive disease with heart failure
16	90012191316..	90012191316..	2016-01-1..	I11.0	NA	58.207161307334..	Hypertensive disease with heart failure
17	90012189254..	90012189254..	2016-04-2..	I11.0	NA	68.664027484133..	Hypertensive disease with heart failure
18	90012182396..	90012182396..	2016-02-2..	I11.0	NA	65.791332640685..	Hypertensive disease with heart failure
19	90012194412..	90012194412..	2016-11-1..	I11.0	NA	51.401857695542..	Hypertensive disease with heart failure

SHEET OVERVIEW 375 rows

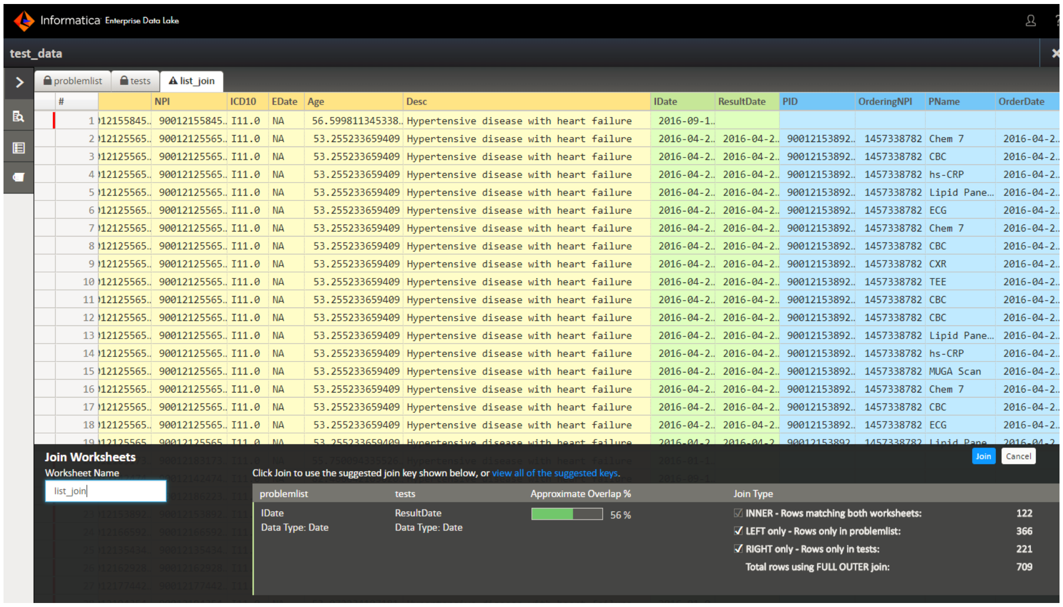
Datatypes Overview Total Columns: 7

🟩 ☐	Text	2
🟦 ☐	Number	2
🟥 ☐	Date	1
🟪 ☐	Country Code	1
🟨 ☐	Number (0.00)	1

Suggestions

Delete 1 column with one value

Enterprise Data Lake suggests columns to use as join keys. The application adds a worksheet containing the combined data as a preview, as shown in the following image:



#		NPI	ICD10	EDate	Age	Desc	IDate	ResultDate	PID	OrderingNPI	PName	OrderDate
1	12155845..	90012155845..	I11.0	NA	56.599811345338..	Hypertensive disease with heart failure	2016-09-1..					
2	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	Chem 7	2016-04-2..
3	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	CBC	2016-04-2..
4	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	hs-CRP	2016-04-2..
5	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	Lipid Pane...	2016-04-2..
6	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	ECG	2016-04-2..
7	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	Chem 7	2016-04-2..
8	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	CBC	2016-04-2..
9	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	CXR	2016-04-2..
10	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	TEE	2016-04-2..
11	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	CBC	2016-04-2..
12	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	CBC	2016-04-2..
13	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	Lipid Pane...	2016-04-2..
14	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	hs-CRP	2016-04-2..
15	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	MUGA Scan	2016-04-2..
16	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	Chem 7	2016-04-2..
17	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	CBC	2016-04-2..
18	12125565..	90012125565..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure	2016-04-2..	2016-04-2..	90012153892..	1457338782	ECG	2016-04-2..

Join Worksheets

Worksheet Name: list_join

Click Join to use the suggested join key shown below, or view all of the suggested keys.

problemlist	tests	Approximate Overlap %
IDate Data Type: Date	ResultDate Data Type: Date	56 %

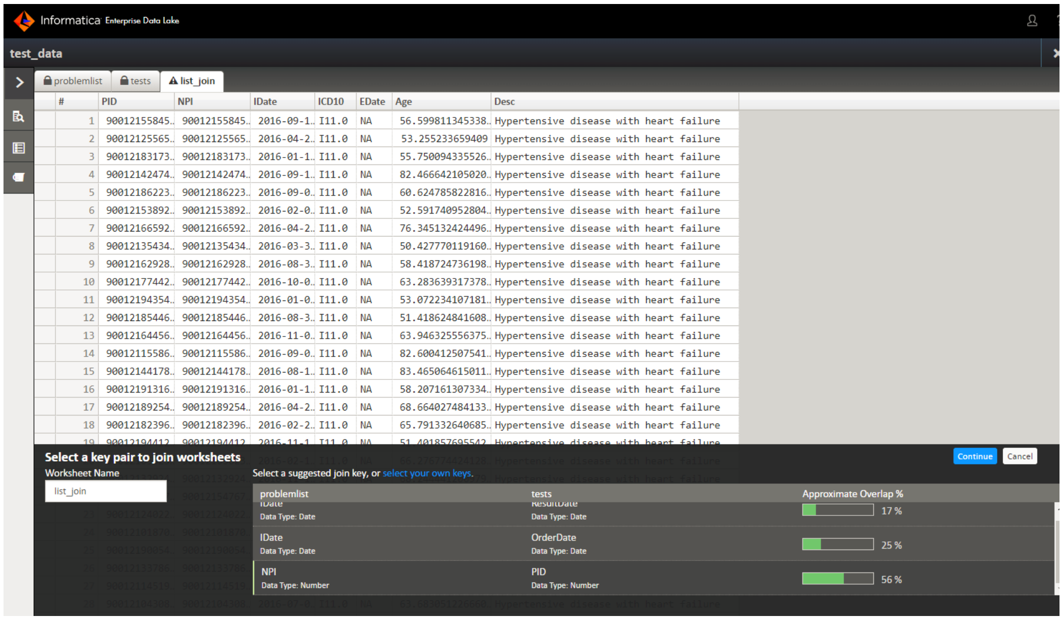
Join Type

☑ INNER - Rows matching both worksheets: 122
✓ LEFT only - Rows only in problemlist: 366
✓ RIGHT only - Rows only in tests: 221
Total rows using FULL OUTER join: 709

Enterprise Data Lake performs a FULL OUTER JOIN by default, which combines all rows containing matching data in the selected columns in both worksheets. You can change the selections in the Join Type column to determine which join type best meets your needs. You can also see the number of rows each join type adds to the worksheet.

You can click **Join** to perform the join using the selected key pair, or you can view all of the suggested key pairs.

The following image shows all of the suggested key pairs:



#	PID	NPI	IDate	ICD10	EDate	Age	Desc
1	90012155845..	90012155845..	2016-09-1..	I11.0	NA	56.599811345338..	Hypertensive disease with heart failure
2	90012125565..	90012125565..	2016-04-2..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure
3	90012183173..	90012183173..	2016-01-1..	I11.0	NA	55.750094335526..	Hypertensive disease with heart failure
4	90012142474..	90012142474..	2016-09-1..	I11.0	NA	82.466642105020..	Hypertensive disease with heart failure
5	90012186223..	90012186223..	2016-09-0..	I11.0	NA	60.624785822816..	Hypertensive disease with heart failure
6	90012153892..	90012153892..	2016-02-0..	I11.0	NA	52.591740952804..	Hypertensive disease with heart failure
7	90012166592..	90012166592..	2016-04-2..	I11.0	NA	76.345132424496..	Hypertensive disease with heart failure
8	90012135434..	90012135434..	2016-03-3..	I11.0	NA	50.427770119160..	Hypertensive disease with heart failure
9	90012162928..	90012162928..	2016-08-3..	I11.0	NA	58.418724736198..	Hypertensive disease with heart failure
10	90012177442..	90012177442..	2016-10-0..	I11.0	NA	63.283639317378..	Hypertensive disease with heart failure
11	90012194354..	90012194354..	2016-01-0..	I11.0	NA	53.072234107181..	Hypertensive disease with heart failure
12	90012185446..	90012185446..	2016-08-3..	I11.0	NA	51.418624841608..	Hypertensive disease with heart failure
13	90012164456..	90012164456..	2016-11-0..	I11.0	NA	63.946325556375..	Hypertensive disease with heart failure
14	90012115586..	90012115586..	2016-09-0..	I11.0	NA	82.600412507541..	Hypertensive disease with heart failure
15	90012144178..	90012144178..	2016-08-1..	I11.0	NA	83.465064615011..	Hypertensive disease with heart failure
16	90012191316..	90012191316..	2016-01-1..	I11.0	NA	58.207161307334..	Hypertensive disease with heart failure
17	90012189254..	90012189254..	2016-04-2..	I11.0	NA	68.664027484133..	Hypertensive disease with heart failure
18	90012182396..	90012182396..	2016-02-2..	I11.0	NA	65.791332640685..	Hypertensive disease with heart failure

Select a key pair to join worksheets

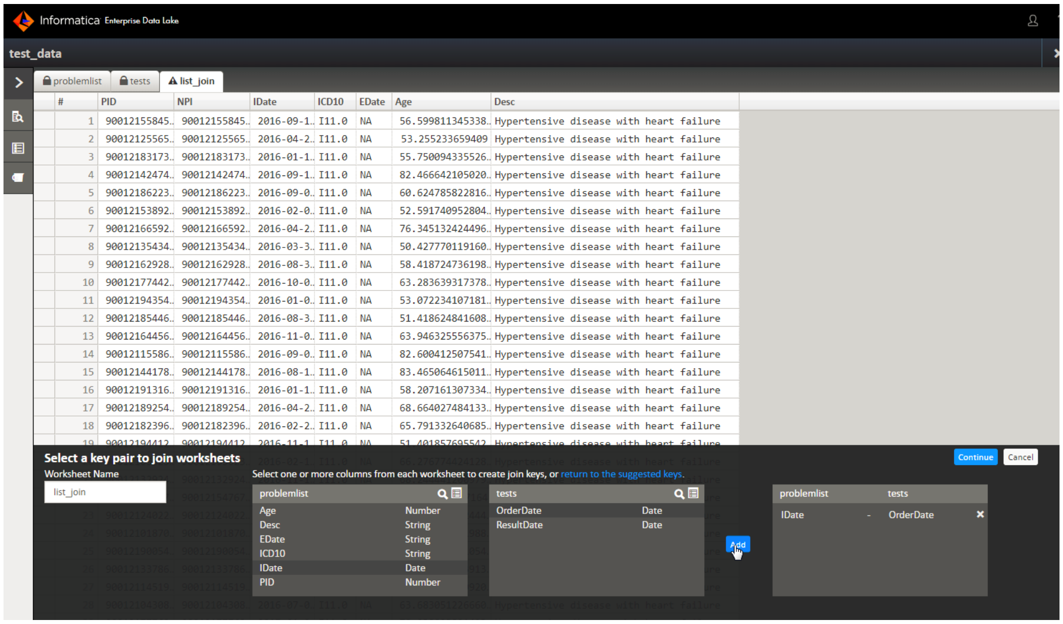
Worksheet Name: list_join

Select a suggested join key, or select your own keys.

problemlist	tests	Approximate Overlap %
IDate Data Type: Date	ResultDate Data Type: Date	17 %
IDate Data Type: Date	OrderDate Data Type: Date	25 %
NPI Data Type: Number	PID Data Type: Number	56 %

If you know that the worksheets contain other columns that also contain matching data, you can choose your own join key pairs from the list of columns in both worksheets. Select a column in each worksheet, and then click **Add**. You must select columns of the same type.

The following image shows how you can select your own join keys:



#	PID	NPI	IDate	ICD10	EDate	Age	Desc
1	90012155845..	90012155845..	2016-09-1..	I11.0	NA	56.599811345338..	Hypertensive disease with heart failure
2	90012125565..	90012125565..	2016-04-2..	I11.0	NA	53.255233659409	Hypertensive disease with heart failure
3	90012183173..	90012183173..	2016-01-1..	I11.0	NA	55.750094335526..	Hypertensive disease with heart failure
4	90012142474..	90012142474..	2016-09-1..	I11.0	NA	82.466642105020..	Hypertensive disease with heart failure
5	90012186223..	90012186223..	2016-09-0..	I11.0	NA	60.624785822816..	Hypertensive disease with heart failure
6	90012153892..	90012153892..	2016-02-0..	I11.0	NA	52.591740952804..	Hypertensive disease with heart failure
7	90012166592..	90012166592..	2016-04-2..	I11.0	NA	76.345132424496..	Hypertensive disease with heart failure
8	90012135434..	90012135434..	2016-03-3..	I11.0	NA	50.427770119160..	Hypertensive disease with heart failure
9	90012162928..	90012162928..	2016-08-3..	I11.0	NA	58.418724736198..	Hypertensive disease with heart failure
10	90012177442..	90012177442..	2016-10-0..	I11.0	NA	63.283639317378..	Hypertensive disease with heart failure
11	90012194354..	90012194354..	2016-01-0..	I11.0	NA	53.072234107181..	Hypertensive disease with heart failure
12	90012185446..	90012185446..	2016-08-3..	I11.0	NA	51.418624841608..	Hypertensive disease with heart failure
13	90012164456..	90012164456..	2016-11-0..	I11.0	NA	63.946325556375..	Hypertensive disease with heart failure
14	90012115586..	90012115586..	2016-09-0..	I11.0	NA	82.600412507541..	Hypertensive disease with heart failure
15	90012144178..	90012144178..	2016-08-1..	I11.0	NA	83.465064615011..	Hypertensive disease with heart failure
16	90012191316..	90012191316..	2016-01-1..	I11.0	NA	58.207161307334..	Hypertensive disease with heart failure
17	90012189254..	90012189254..	2016-04-2..	I11.0	NA	68.664027484133..	Hypertensive disease with heart failure
18	90012182396..	90012182396..	2016-02-2..	I11.0	NA	65.791332640685..	Hypertensive disease with heart failure

Select a key pair to join worksheets

Worksheet Name: list_join

Select one or more columns from each worksheet to create join keys, or return to the suggested keys.

problemlist	
Age	Number
Desc	String
EDate	String
ICD10	String
IDate	Date
PID	Number

tests	
OrderDate	Date
ResultDate	Date

problemlist		tests
IDate	-	OrderDate

Joining Data in Worksheets

A project must contain at least two worksheets to perform a join operation. You can modify the suggested columns to use as join keys.

1. On the project page, click **Prepare**.
2. Click the **Blending** icon, and then select **Join worksheets**.
3. In the **Join Worksheets** dialog box, select the worksheet that you want to join with the current worksheet.
4. Select the worksheet in the project containing the data you want to join with the current worksheet.

   The application adds a worksheet containing the joined columns to the project. Matching columns that occur in both worksheets are displayed in green. Columns that occur in the current worksheet only are shown in yellow. Columns that occur only in the worksheet selected for the join are shown in blue.
5. Enter a name for the worksheet.
6. Review the suggested key pair. You can view the occurrence of identical distinct values in both columns in the Approximate Overlap % column.

   Decide which keys to use to join the worksheets:
   - Use the suggested keys.
   - Click **View all of the suggested keys** to choose from the complete list of selected join keys. Select a row containing the keys to use, and then click **Continue**.
   - Click **Select your own keys** to choose one or more join key pairs from all possible keys in the worksheets. Click a key in each worksheet, and then click **Add**. You must select columns of the same type to use as join keys. Click **Continue** when you finish selecting the columns.
7. Select the type of join to perform in the Join Type column. Enterprise Data Lake updates the worksheet based on your selection.
   - Select **INNER** to join the matching rows in both the worksheets.
   - Select **LEFT** to join all of the rows from the first worksheet and the rows with matching values from the second worksheet.

- Select **RIGHT** to join all of the rows from the second worksheet and the with matching values from the first worksheet.
- Select **RIGHT** and **LEFT** to perform a FULL OUTER JOIN of all rows containing matching distinct values in the selected columns in both worksheets.

8. Click **Join** to join the worksheets using the join keys you select.

   Enterprise Data Lake adds the join worksheet to the project.

Editing Joins

On a joined worksheet, you can edit the join and choose different join keys. You cannot edit joins if any non-key columns used in the recipe are missing from the source sheets. You cannot edit joins created in an earlier version of Enterprise Data Lake.

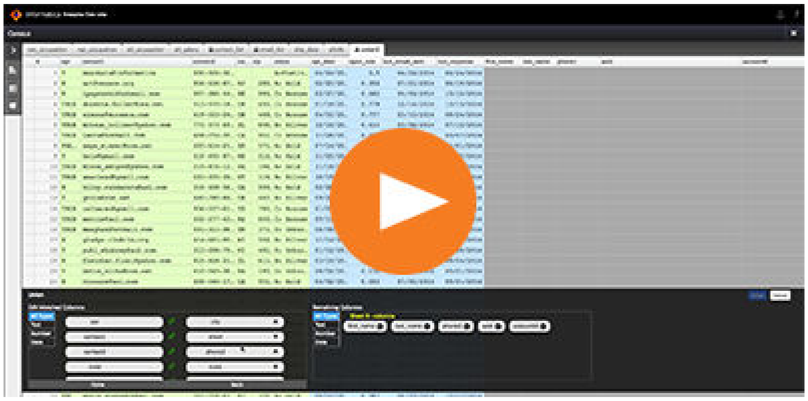
1. Select the joined worksheet.
2. Select the join step in the recipe.
3. Click the **Edit** icon to edit the join.
4. Review the suggested selected key. You can view the approximate overlap percentage and results.
5. Click **Change Key** to select another join key.

   Note: If the join key is not present, you can select a new join key.
6. From the Results displayed, select or deselect to modify the type of join: inner join, left join, or right join.
7. Click **Done**. To undo the edits to the join, click **Cancel**.

Merging Worksheets

You can use the union operation to merge data from another worksheet into the current worksheet. The operation adds a new worksheet containing the merged data to the project.

Click the image below to view a video about the union operation:

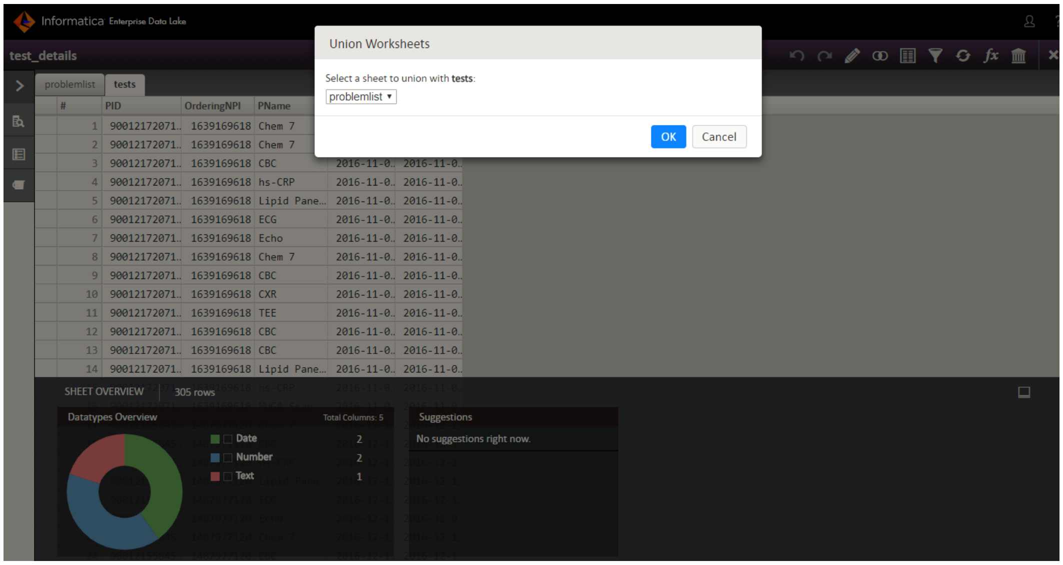


When you perform a union, you match columns in both worksheets that contain similar data. Enterprise Data Lake suggests matching columns of the same type to include in the union. Enterprise Data Lake merges the data in a matched column in the other worksheet into the corresponding matched column in the current worksheet.

You can modify the suggested matching columns by removing a column from a matched pair. You can also select columns in the other worksheet to match with columns in the current worksheet. The columns you select must be of the same type.

Example

To perform a union, you select the worksheet you want to merge with the current worksheet, as shown in the following image:



Union Worksheets

Select a sheet to union with **tests**:

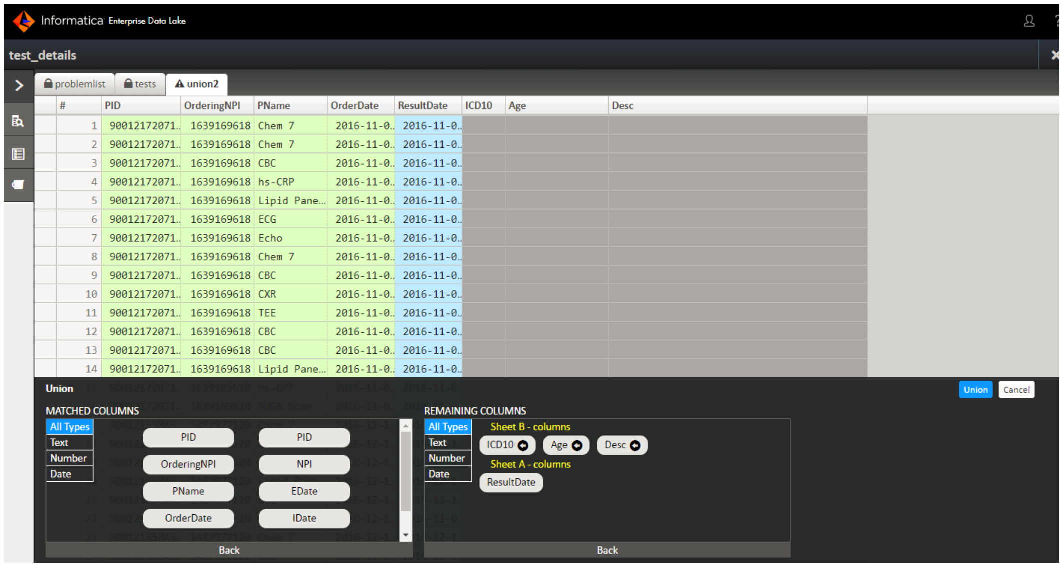
problemlist ▾

OK Cancel

#	PID	OrderingNPI	PName		
1	90012172071..	1639169618	Chem 7		
2	90012172071..	1639169618	Chem 7		
3	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..
4	90012172071..	1639169618	hs-CRP	2016-11-0..	2016-11-0..
5	90012172071..	1639169618	Lipid Pane..	2016-11-0..	2016-11-0..
6	90012172071..	1639169618	ECG	2016-11-0..	2016-11-0..
7	90012172071..	1639169618	Echo	2016-11-0..	2016-11-0..
8	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..
9	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..
10	90012172071..	1639169618	CXR	2016-11-0..	2016-11-0..
11	90012172071..	1639169618	TEE	2016-11-0..	2016-11-0..
12	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..
13	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..
14	90012172071..	1639169618	Lipid Pane..	2016-11-0..	2016-11-0..

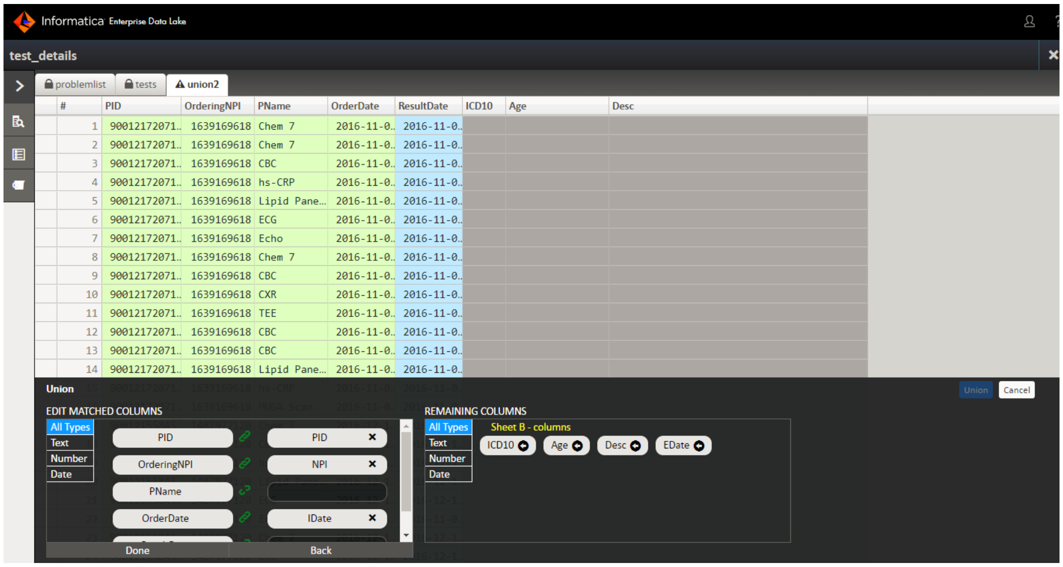
Enterprise Data Lake updates the worksheet with a preview of the union operation results. Matched columns are displayed in green. Columns in the tests worksheet that Enterprise Data Lake does not match with columns in the problemlist worksheet are shown in blue.

Click **View** in each panel to review the matched column pairs, as shown in the following image:



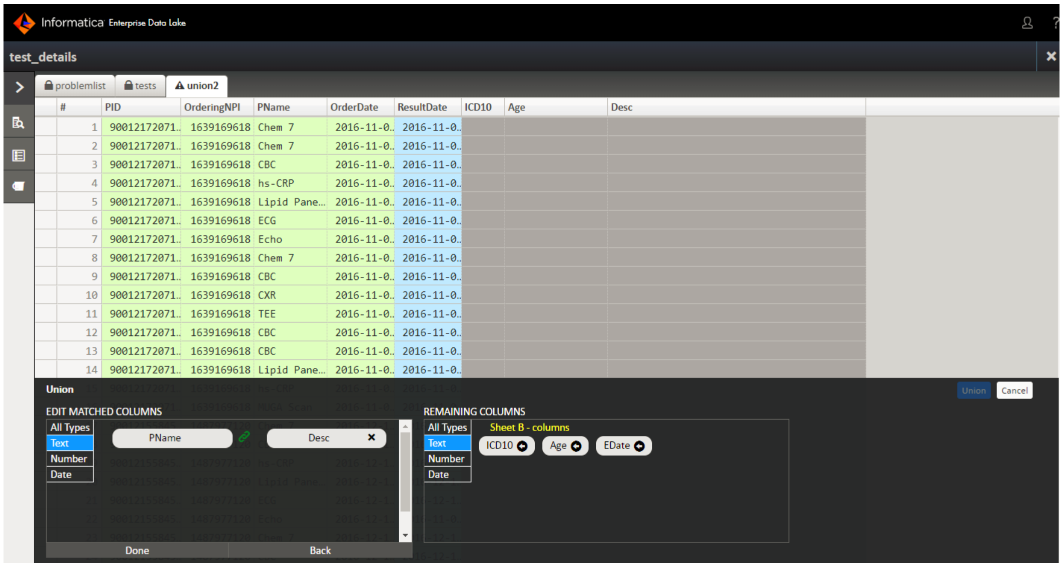
#	PID	OrderingNPI	PName	OrderDate	ResultDate	ICD10	Age	Desc
1	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
2	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
3	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
4	90012172071..	1639169618	hs-CRP	2016-11-0..	2016-11-0..			
5	90012172071..	1639169618	Lipid Pane..	2016-11-0..	2016-11-0..			
6	90012172071..	1639169618	ECG	2016-11-0..	2016-11-0..			
7	90012172071..	1639169618	Echo	2016-11-0..	2016-11-0..			
8	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
9	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
10	90012172071..	1639169618	CXR	2016-11-0..	2016-11-0..			
11	90012172071..	1639169618	TEE	2016-11-0..	2016-11-0..			
12	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
13	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
14	90012172071..	1639169618	Lipid Pane..	2016-11-0..	2016-11-0..			

You realize that the data in the PName column in the tests worksheet does not match the EDate column in the problemlist worksheet. Click the **Change** button in the **Matched Columns** panel, and then click EDate to remove it from the panel, as shown in the following image:



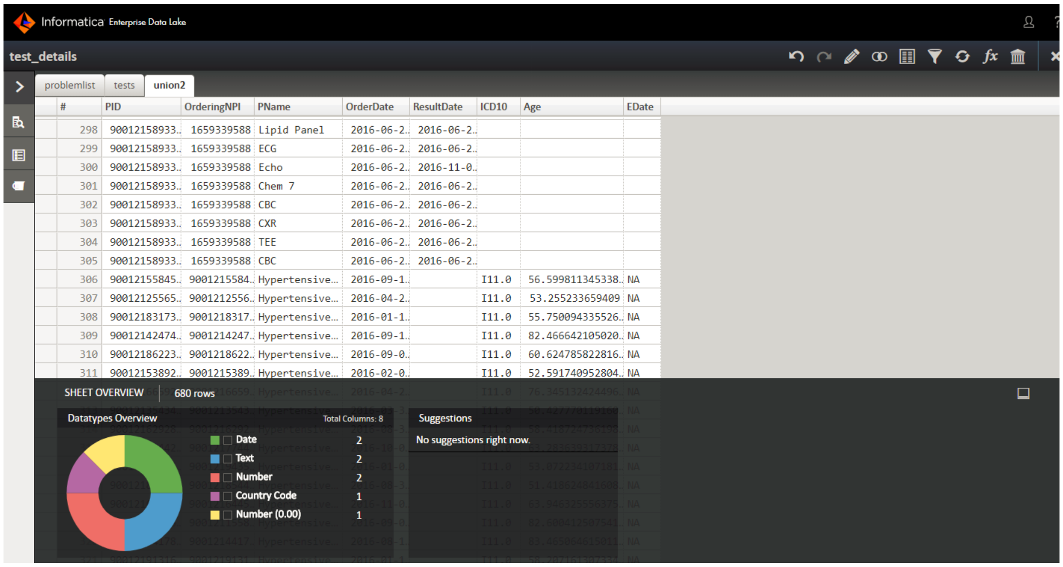
#	PID	OrderingNPI	PName	OrderDate	ResultDate	ICD10	Age	Desc
1	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
2	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
3	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
4	90012172071..	1639169618	hs-CRP	2016-11-0..	2016-11-0..			
5	90012172071..	1639169618	Lipid Pane...	2016-11-0..	2016-11-0..			
6	90012172071..	1639169618	ECG	2016-11-0..	2016-11-0..			
7	90012172071..	1639169618	Echo	2016-11-0..	2016-11-0..			
8	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
9	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
10	90012172071..	1639169618	CXR	2016-11-0..	2016-11-0..			
11	90012172071..	1639169618	TEE	2016-11-0..	2016-11-0..			
12	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
13	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
14	90012172071..	1639169618	Lipid Pane...	2016-11-0..	2016-11-0..			

Select **Text** in each panel to display only values of the same type. You know that the data in the Desc column in the problemlist worksheet is a good match with the PName column. Click the empty field next to PName, and then click Desc in the **Remaining Columns** panel to select it as the column to match with the PName column, as shown in the following image:



#	PID	OrderingNPI	PName	OrderDate	ResultDate	ICD10	Age	Desc
1	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
2	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
3	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
4	90012172071..	1639169618	hs-CRP	2016-11-0..	2016-11-0..			
5	90012172071..	1639169618	Lipid Pane...	2016-11-0..	2016-11-0..			
6	90012172071..	1639169618	ECG	2016-11-0..	2016-11-0..			
7	90012172071..	1639169618	Echo	2016-11-0..	2016-11-0..			
8	90012172071..	1639169618	Chem 7	2016-11-0..	2016-11-0..			
9	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
10	90012172071..	1639169618	CXR	2016-11-0..	2016-11-0..			
11	90012172071..	1639169618	TEE	2016-11-0..	2016-11-0..			
12	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
13	90012172071..	1639169618	CBC	2016-11-0..	2016-11-0..			
14	90012172071..	1639169618	Lipid Pane...	2016-11-0..	2016-11-0..			

When you are satisfied that the matched column pairs are correct, click **Done** in the **Matched Columns** panel, and then click **Union** to complete the union. Enterprise Data Lake adds the union worksheet to the project, as shown in the following image:



#	PID	OrderingNPI	PName	OrderDate	ResultDate	ICD10	Age	EDate
298	90012158933..	1659339588	Lipid Panel	2016-06-2..	2016-06-2..			
299	90012158933..	1659339588	ECG	2016-06-2..	2016-06-2..			
300	90012158933..	1659339588	Echo	2016-06-2..	2016-11-0..			
301	90012158933..	1659339588	Chem 7	2016-06-2..	2016-06-2..			
302	90012158933..	1659339588	CBC	2016-06-2..	2016-06-2..			
303	90012158933..	1659339588	CXR	2016-06-2..	2016-06-2..			
304	90012158933..	1659339588	TEE	2016-06-2..	2016-06-2..			
305	90012158933..	1659339588	CBC	2016-06-2..	2016-06-2..			
306	90012155845..	9001215584..	Hypertensive...	2016-09-1..		I11.0	56.599811345338..	NA
307	90012125565..	9001212556..	Hypertensive...	2016-04-2..		I11.0	53.255233659409	NA
308	90012183173..	9001218317..	Hypertensive...	2016-01-1..		I11.0	55.750094335526..	NA
309	90012142474..	9001214247..	Hypertensive...	2016-09-1..		I11.0	82.466642105020..	NA
310	90012186223..	9001218622..	Hypertensive...	2016-09-0..		I11.0	60.624785822816..	NA
311	90012153892..	9001215389..	Hypertensive...	2016-02-0..		I11.0	52.591740952804..	NA

Merging Columns in Worksheets

A project must contain at least two worksheets to perform a union operation. You can modify the suggested matched columns. Columns matched together must be of the same type.

1. Open the worksheet you want to union with another worksheet.
2. Click the **Blend** icon, and then select **Union worksheets**.
3. Select the worksheet in the project containing the data you want to merge into the current worksheet.

   The application adds a worksheet containing the merged data to the project. Matching columns that occur in both worksheets are displayed in green. Columns that occur in the current worksheet only are shown in blue. Columns that occur only in the worksheet selected for the union are shown in yellow.
4. The **Matched Columns** panel contains the matching column pairs found in both worksheets. Columns are matched based on type. The **Remaining Columns** panel contains columns for which a matching column cannot be found.

   Click **View** in each panel to view the columns. Click **Back** when finished.
5. Click **Change** in the **Matched Columns** panel to modify the matched column pairs.
   a. If a matched column pair is not correct in the **Matched Columns** panel, click the column to the right to remove it from the panel.
   b. To match a column in the **Remaining Columns** panel with a column in the **Matched Columns** panel, click the entry field for the corresponding column in the **Matched Columns** panel, then click a column of the same type in the **Remaining Columns** panel.
6. Click **Done** when you finish making changes to the matching columns.
7. Click **Union**.

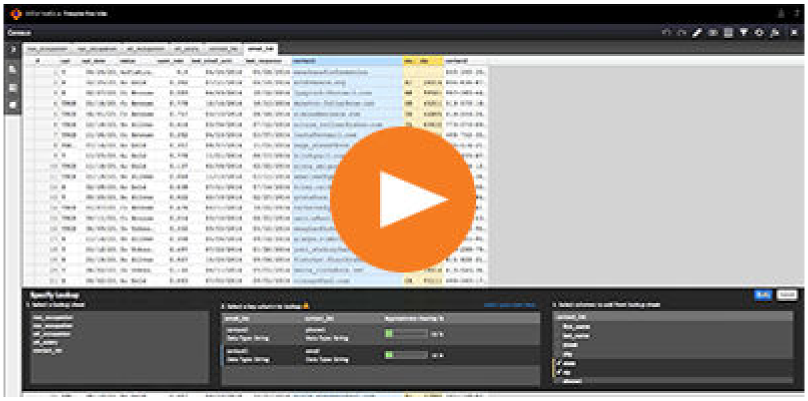
   Enterprise Data Lake adds the union worksheet to the project.

Using Lookup

You can use the lookup operation to import selected columns from another worksheet into the current worksheet.

When you perform a lookup, you select a column in the current worksheet and a column in another worksheet that contain similar distinct values to use as lookup keys. For example, if both worksheets contain a column with a customer e-mail address, you can select these columns to use as lookup keys. You then select columns in the other worksheet that contain data associated with each customer email address, and import the columns into the current worksheet.

Click the image below to view a video about the lookup operation:

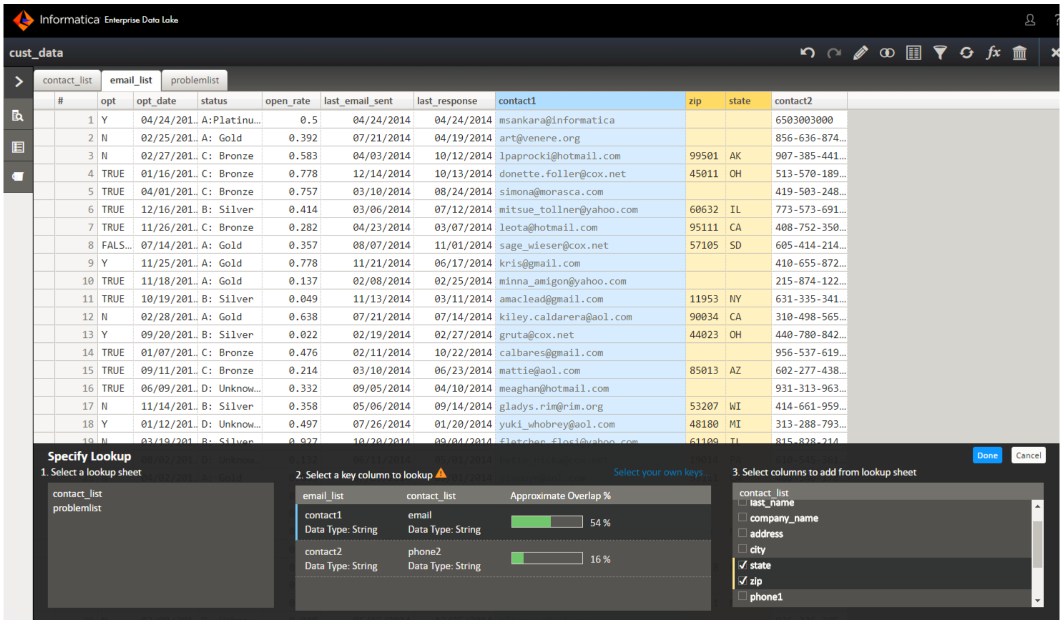


Example

After you select the lookup worksheet, which is the worksheet containing the data to import, you select the columns to use as lookup keys in both the current worksheet and in the lookup worksheet. You then select columns in the lookup worksheet that you want to import into the current worksheet.

Enterprise Data Lake suggests columns that contain a high frequency of identical distinct values to use as lookup keys, as indicated in the Approximate Overlap % property.

Enterprise Data Lake adds the imported columns to the current worksheet as a preview, as shown in the following image:



#	opt	opt_date	status	open_rate	last_email_sent	last_response	contact1	zip	state	contact2
1	Y	04/24/201..	A:Platinu...	0.5	04/24/2014	04/24/2014	msankara@informatica			6503003000
2	N	02/25/201..	A: Gold	0.392	07/21/2014	04/19/2014	art@venere.org			856-636-874...
3	N	02/27/201..	C: Bronze	0.583	04/03/2014	10/12/2014	lpaprocki@hotmail.com	99501	AK	907-385-441...
4	TRUE	01/16/201..	C: Bronze	0.778	12/14/2014	10/13/2014	donette.foller@cox.net	45011	OH	513-570-189...
5	TRUE	04/01/201..	C: Bronze	0.757	03/10/2014	08/24/2014	simona@morasca.com			419-503-248...
6	TRUE	12/16/201..	B: Silver	0.414	03/06/2014	07/12/2014	mitsue_tollner@yahoo.com	60632	IL	773-573-691...
7	TRUE	11/26/201..	C: Bronze	0.282	04/23/2014	03/07/2014	leota@hotmail.com	95111	CA	408-752-350...
8	FALS...	07/14/201..	A: Gold	0.357	08/07/2014	11/01/2014	sage_wieser@cox.net	57105	SD	605-414-214...
9	Y	11/25/201..	A: Gold	0.778	11/21/2014	06/17/2014	kris@gmail.com			410-655-872...
10	TRUE	11/18/201..	A: Gold	0.137	02/08/2014	02/25/2014	minna_amigon@yahoo.com			215-874-122...
11	TRUE	10/19/201..	B: Silver	0.049	11/13/2014	03/11/2014	amaclead@gmail.com	11953	NY	631-335-341...
12	N	02/28/201..	A: Gold	0.638	07/21/2014	07/14/2014	kiley.caldarera@aol.com	90034	CA	310-498-565...
13	Y	09/20/201..	B: Silver	0.022	02/19/2014	02/27/2014	gruta@cox.net	44023	OH	440-780-842...
14	TRUE	01/07/201..	C: Bronze	0.476	02/11/2014	10/22/2014	calbares@gmail.com			956-537-619...
15	TRUE	09/11/201..	C: Bronze	0.214	03/10/2014	06/23/2014	mattie@aol.com	85013	AZ	602-277-438...
16	TRUE	06/09/201..	D: Unknow...	0.332	09/05/2014	04/10/2014	meaghan@hotmail.com			931-313-963...
17	N	11/14/201..	B: Silver	0.358	05/06/2014	09/14/2014	gladys.rim@rim.org	53207	WI	414-661-959...
18	Y	01/12/201..	D: Unknow...	0.497	07/26/2014	01/20/2014	yuki_whobrey@aol.com	48180	MI	313-288-793...

Specify Lookup

1. Select a lookup sheet
   - contact_list
   - problemlist
2. Select a key column to lookup ⚠ Select your own keys...

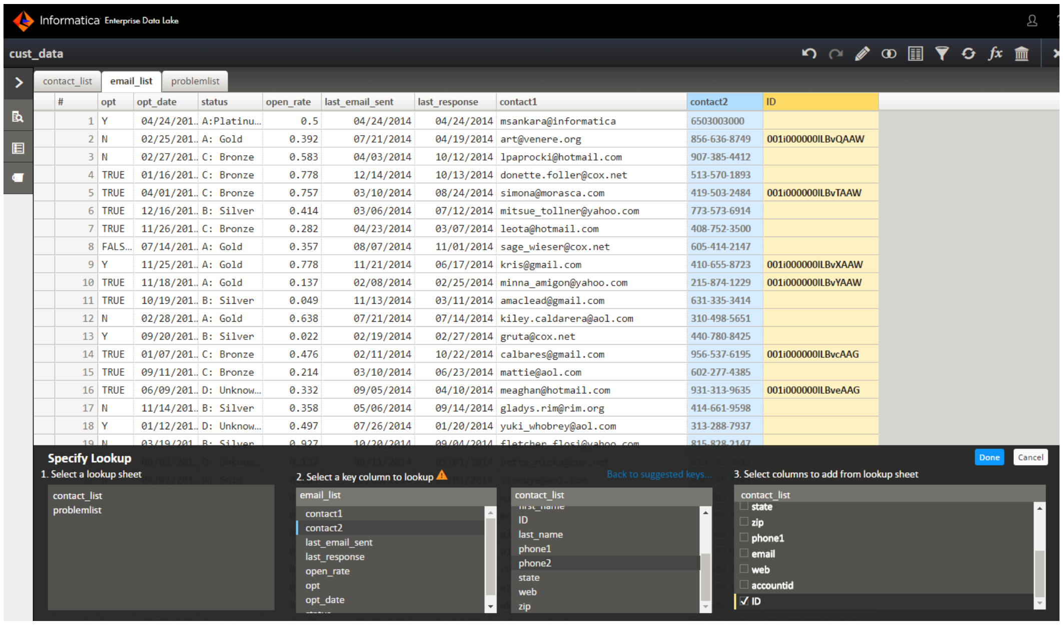
email_list	contact_list	Approximate Overlap %
contact1 Data Type: String	email Data Type: String	54 %
contact2 Data Type: String	phone2 Data Type: String	16 %

3. Select columns to add from lookup sheet

contact_list
- ☐ last_name
- ☐ company_name
- ☐ address
- ☐ city
- ☑ state
- ☑ zip
- ☐ phone1

You can also select other columns to use as lookup keys. You must select columns of the same type.

The following image shows the selected lookup keys:



#	opt	opt_date	status	open_rate	last_email_sent	last_response	contact1	contact2	ID
1	Y	04/24/201..	A:Platinu...	0.5	04/24/2014	04/24/2014	msankara@informatica	6503003000	
2	N	02/25/201..	A: Gold	0.392	07/21/2014	04/19/2014	art@venere.org	856-636-8749	001i000000ILBvQAAW
3	N	02/27/201..	C: Bronze	0.583	04/03/2014	10/12/2014	lpaprocki@hotmail.com	907-385-4412	
4	TRUE	01/16/201..	C: Bronze	0.778	12/14/2014	10/13/2014	donette.foller@cox.net	513-570-1893	
5	TRUE	04/01/201..	C: Bronze	0.757	03/10/2014	08/24/2014	simona@morasca.com	419-503-2484	001i000000ILBvTAAW
6	TRUE	12/16/201..	B: Silver	0.414	03/06/2014	07/12/2014	mitsue_tollner@yahoo.com	773-573-6914	
7	TRUE	11/26/201..	C: Bronze	0.282	04/23/2014	03/07/2014	leota@hotmail.com	408-752-3500	
8	FALS...	07/14/201..	A: Gold	0.357	08/07/2014	11/01/2014	sage_wieser@cox.net	605-414-2147	
9	Y	11/25/201..	A: Gold	0.778	11/21/2014	06/17/2014	kris@gmail.com	410-655-8723	001i000000ILBvXAAW
10	TRUE	11/18/201..	A: Gold	0.137	02/08/2014	02/25/2014	minna_amigon@yahoo.com	215-874-1229	001i000000ILBvYAAW
11	TRUE	10/19/201..	B: Silver	0.049	11/13/2014	03/11/2014	amaclead@gmail.com	631-335-3414	
12	N	02/28/201..	A: Gold	0.638	07/21/2014	07/14/2014	kiley.caldarera@aol.com	310-498-5651	
13	Y	09/20/201..	B: Silver	0.022	02/19/2014	02/27/2014	gruta@cox.net	440-780-8425	
14	TRUE	01/07/201..	C: Bronze	0.476	02/11/2014	10/22/2014	calbares@gmail.com	956-537-6195	001i000000ILBvcAAG
15	TRUE	09/11/201..	C: Bronze	0.214	03/10/2014	06/23/2014	mattie@aol.com	602-277-4385	
16	TRUE	06/09/201..	D: Unknow...	0.332	09/05/2014	04/10/2014	meaghan@hotmail.com	931-313-9635	001i000000ILBveAAG
17	N	11/14/201..	B: Silver	0.358	05/06/2014	09/14/2014	gladys.rim@rim.org	414-661-9598	
18	Y	01/12/201..	D: Unknow...	0.497	07/26/2014	01/20/2014	yuki_whobrey@aol.com	313-288-7937	

Specify Lookup

1. Select a lookup sheet
   - contact_list
   - problemlist
2. Select a key column to lookup ⚠ Back to suggested keys...

email_list
- contact1
- contact2
- last_email_sent
- last_response
- open_rate
- opt
- opt_date

contact_list
- ID
- last_name
- phone1
- phone2
- state
- web
- zip

3. Select columns to add from lookup sheet

contact_list
- ☐ state
- ☐ zip
- ☐ phone1
- ☐ email
- ☐ web
- ☐ accountid
- ☑ ID

Performing a Lookup on Worksheets

A project must contain at least two worksheets to perform a lookup operation. You can modify the suggested columns to use as lookup keys. Both worksheets must exist in the same project.

You cannot perform a lookup on derived columns, worksheets derived from other, or sheets with filter options applied.

1. On the project page, click **Prepare**.
2. Click the **Blending** icon, and then select **Lookup**.
3. Select the worksheet containing the data you want to import from the list of worksheets.
4. Enterprise Data Lake displays the columns with the highest amount of overlap as suggested keys as indicated in the Approximate Overlap % column. Select the columns to use as lookup keys:
    - Select a suggested key pair.
    - Click **Select your own keys** to choose a column in each worksheet to use as a lookup key. You must select columns of the same type.
5. Select the columns to add to the current worksheet.
6. Click **Done**.

General Limitations for Lookup

The following limitations apply to the lookup operation.

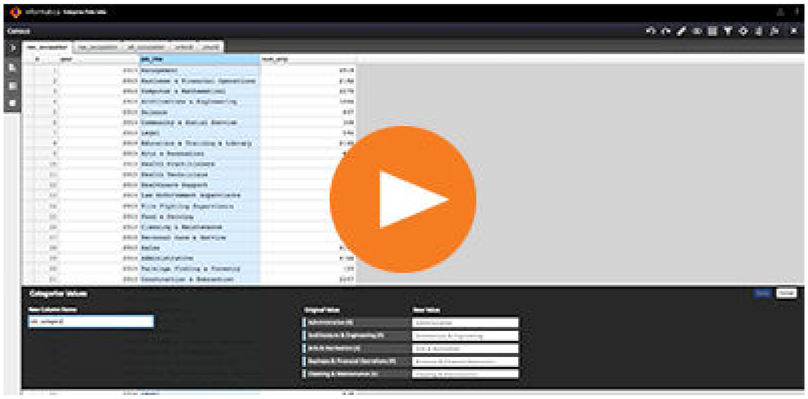
- Duplicate keys. You cannot perform a lookup on a worksheet that contains duplicate keys.
- Derived worksheets. You can perform a lookup only on existing worksheets within a project. You cannot perform lookup on worksheets derived from other worksheets using operations such as union, aggregate, or join. However, you can perform a lookup on copied worksheets.
- Filtered worksheets. Publication is disabled for lookup worksheets that use column filters. A lookup worksheet is the worksheet created by a lookup operation.
- Source filters. Publication is disabled for lookup worksheets that use source filters.
- New columns. You can perform a lookup only on existing columns. You cannot perform a lookup on new columns added to a worksheet, such as columns derived after performing operations on existing columns.
- Existing columns. You cannot perform a lookup on columns on which any operation has been performed.
- Data type mismatch. You cannot perform a lookup on two columns with different data types such as decimal and double data types.
- Duplicate rows removed. You cannot perform a lookup on a worksheet from which duplicate rows were removed.

Categorizing Column Data

You can categorize similar values in a column to make analysis easier.

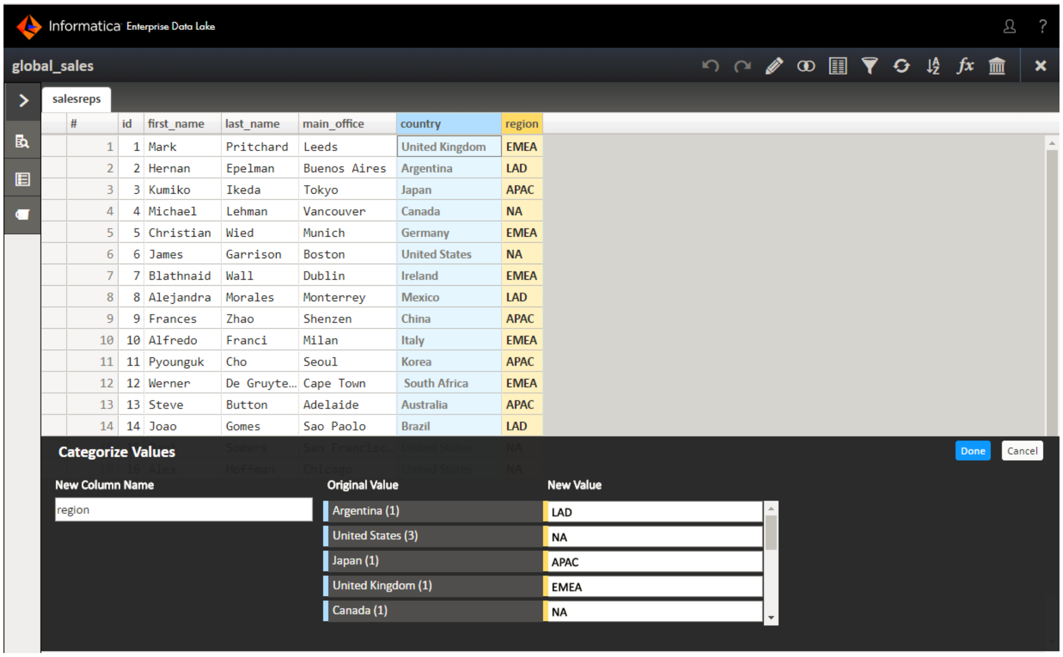
To categorize data, you map a category value to each value in a column. The category values are displayed in a new column you add to the worksheet.

Click the image below to view a video about categorizing column data:



The following example shows each country in a worksheet categorized by sales region. The corresponding category value for each country is shown in a new column named region that is added to the worksheet.

The following image shows the categorized data:



#	id	first_name	last_name	main_office	country	region
1	1	Mark	Pritchard	Leeds	United Kingdom	EMEA
2	2	Hernan	Epelman	Buenos Aires	Argentina	LAD
3	3	Kumiko	Ikeda	Tokyo	Japan	APAC
4	4	Michael	Lehman	Vancouver	Canada	NA
5	5	Christian	Wied	Munich	Germany	EMEA
6	6	James	Garrison	Boston	United States	NA
7	7	Blathnaid	Wall	Dublin	Ireland	EMEA
8	8	Alejandra	Morales	Monterrey	Mexico	LAD
9	9	Frances	Zhao	Shenzen	China	APAC
10	10	Alfredo	Franci	Milan	Italy	EMEA
11	11	Pyounguk	Cho	Seoul	Korea	APAC
12	12	Werner	De Gruyte...	Cape Town	South Africa	EMEA
13	13	Steve	Button	Adelaide	Australia	APAC
14	14	Joao	Gomes	Sao Paolo	Brazil	LAD

New Column Name: region

Original Value	New Value
Argentina (1)	LAD
United States (3)	NA
Japan (1)	APAC
United Kingdom (1)	EMEA
Canada (1)	NA

The number next to each value in the Original Value column indicates the occurrence of the value in the asset. An occurrence value of 0 indicates that the value is no longer available in the asset.

1. On the project page, click **Prepare**.

   The Sampling Preview panel appears along with the worksheet.

2. Select the column in the worksheet containing the data that you want to categorize.
3. Click the **Blending** icon and select **Categorize**.
4. Enter the name of the new column that contains the category values.
5. For each value in the Original Value column, enter the corresponding category value in the New Value column.

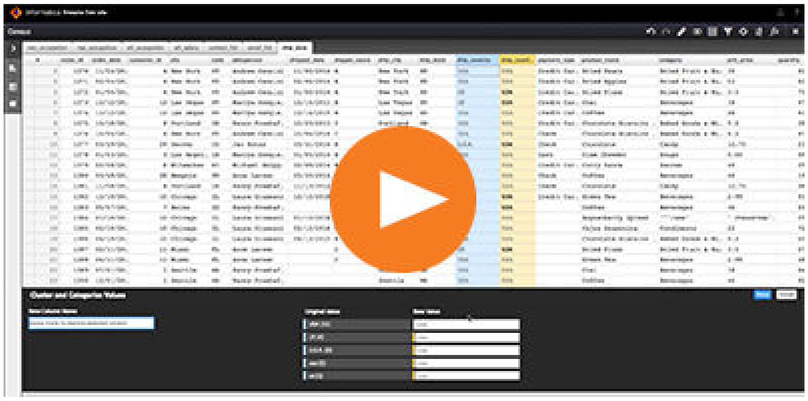
   If you do not enter a new column name, each value you enter in the New Value column overwrites the values in the Original Value column.

6. Click **Done**.

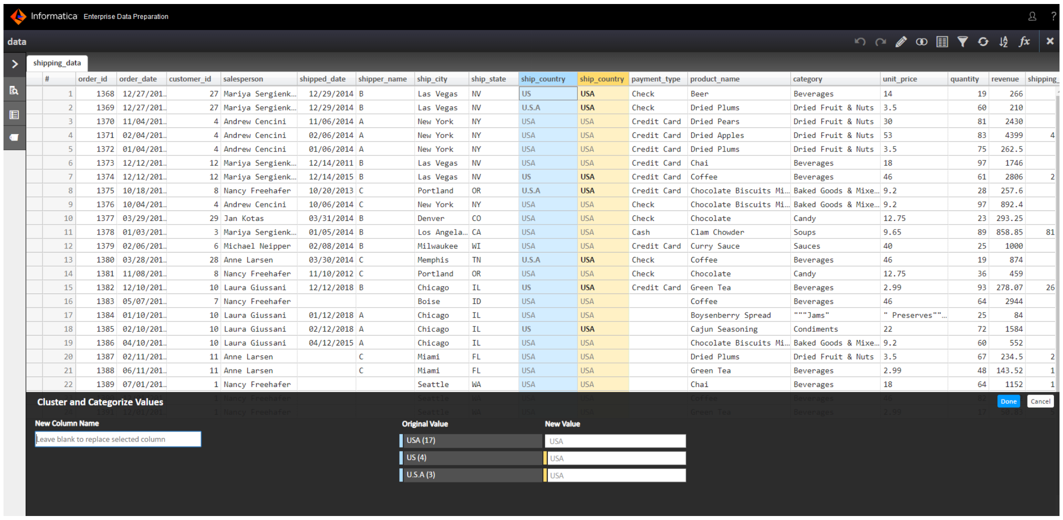
Clustering and Categorizing Column Data

You can cluster similar Text values in a column, and then categorize the values using values suggested by Enterprise Data Lake. The application uses a phonetic algorithm to identify similar values, and then suggests that you replace the less frequently occurring values with the most frequently occurring value.

Click the image below to view a video about clustering similar values in a column:



For example, if the string "USA" occurs 17 times in a column, and the strings "U.S.A" and "US" each occur fewer times, the application suggests that you replace the less frequently occurring values with "USA", as shown in the following example:



#	order_id	order_date	customer_id	salesperson	shipped_date	shipper_name	ship_city	ship_state	ship_country	ship_country	payment_type	product_name	category	unit_price	quantity	revenue	shipping_
1	1368	12/27/201..	27	Mariya Sergienk...	12/29/2014	B	Las Vegas	NV	US	USA	Check	Beer	Beverages	14	19	266	
2	1369	12/27/201..	27	Mariya Sergienk...	12/29/2014	B	Las Vegas	NV	U.S.A	USA	Check	Dried Plums	Dried Fruit & Nuts	3.5	60	210	
3	1370	11/04/201..	4	Andrew Cencini	11/06/2014	A	New York	NY	USA	USA	Credit Card	Dried Pears	Dried Fruit & Nuts	30	81	2430	
4	1371	02/04/201..	4	Andrew Cencini	02/06/2014	A	New York	NY	USA	USA	Credit Card	Dried Apples	Dried Fruit & Nuts	53	83	4399	4
5	1372	01/04/201..	4	Andrew Cencini	01/06/2014	A	New York	NY	USA	USA	Credit Card	Dried Plums	Dried Fruit & Nuts	3.5	75	262.5	
6	1373	12/12/201..	12	Mariya Sergienk...	12/14/2011	B	Las Vegas	NV	USA	USA	Credit Card	Chai	Beverages	18	97	1746	
7	1374	12/12/201..	12	Mariya Sergienk...	12/14/2015	B	Las Vegas	NV	US	USA	Credit Card	Coffee	Beverages	46	61	2806	2
8	1375	10/18/201..	8	Nancy Freehafer	10/20/2013	C	Portland	OR	U.S.A	USA	Credit Card	Chocolate Biscuits Mi...	Baked Goods & Mixe...	9.2	28	257.6	
9	1376	10/04/201..	4	Andrew Cencini	10/06/2014	C	New York	NY	USA	USA	Check	Chocolate Biscuits Mi...	Baked Goods & Mixe...	9.2	97	892.4	
10	1377	03/29/201..	29	Jan Kotas	03/31/2014	B	Denver	CO	USA	USA	Check	Chocolate	Candy	12.75	23	293.25	
11	1378	01/03/201..	3	Mariya Sergienk...	01/05/2014	B	Los Angela...	CA	USA	USA	Cash	Clam Chowder	Soups	9.65	89	858.85	81
12	1379	02/06/201..	6	Michael Neipper	02/08/2014	B	Milwaukee	WI	USA	USA	Credit Card	Curry Sauce	Sauces	40	25	1000	
13	1380	03/28/201..	28	Anne Larsen	03/30/2014	C	Memphis	TN	U.S.A	USA	Check	Coffee	Beverages	46	19	874	
14	1381	11/08/201..	8	Nancy Freehafer	11/10/2012	C	Portland	OR	USA	USA	Check	Chocolate	Candy	12.75	36	459	
15	1382	12/10/201..	10	Laura Giussani	12/12/2018	B	Chicago	IL	US	USA	Credit Card	Green Tea	Beverages	2.99	93	278.07	26
16	1383	05/07/201..	7	Nancy Freehafer			Boise	ID	USA	USA		Coffee	Beverages	46	64	2944	
17	1384	01/10/201..	10	Laura Giussani	01/12/2018	A	Chicago	IL	USA	USA		Boysenberry Spread	"""Jams"	" Preserves""...	25	84	
18	1385	02/10/201..	10	Laura Giussani	02/12/2018	A	Chicago	IL	US	USA		Cajun Seasoning	Condiments	22	72	1584	
19	1386	04/10/201..	10	Laura Giussani	04/12/2015	A	Chicago	IL	USA	USA		Chocolate Biscuits Mi...	Baked Goods & Mixe...	9.2	60	552	
20	1387	02/11/201..	11	Anne Larsen		C	Miami	FL	USA	USA		Dried Plums	Dried Fruit & Nuts	3.5	67	234.5	2
21	1388	06/11/201..	11	Anne Larsen		C	Miami	FL	USA	USA		Green Tea	Beverages	2.99	48	143.52	1
22	1389	07/01/201..	1	Nancy Freehafer			Seattle	WA	USA	USA		Chai	Beverages	18	64	1152	1

Cluster and Categorize Values

New Column Name: Leave blank to replace selected column

Original Value	New Value
USA (17)	USA
US (4)	USA
U.S.A (3)	USA

The features is also useful for replacing misspelled values. For example, the application might suggest that you replace "northdakota" or "North Dakkotta' with the more frequently occurring "North Dakota".

Note that updates are made to the sample data only. Values in rows not included in the sample data are published as is.

1. On the project page, click **Prepare**.

   The Sampling Preview panel appears along with the worksheet.

2. Select the column in the worksheet containing the data that you want to categorize.
3. Click the **Blending** icon and select **Cluster and Categorize**.

4. Enter the name of the new column that contains the category values.

   If you do not enter a column name, the application updates the values in the selected column.

5. For each value in the Original Value column, the application displays suggested values in the New Value column, with the most commonly occurring similar value shown first.

   You can modify a suggested value to suit your needs.

6. Click **Done**.

Summarizing Data

When you prepare your data, you can summarize the data to make analysis easier. Enterprise Data Lake offers you several options for summarizing data.

You can aggregate data in a worksheet by grouping selected columns together, and then performing calculations on columns containing number values to summarize the data.

You can reorganize and summarize data in selected columns in a worksheet using a pivot table. For example, you might want to summarize home sales data to analyze the average price of single family homes sold in each city by month.

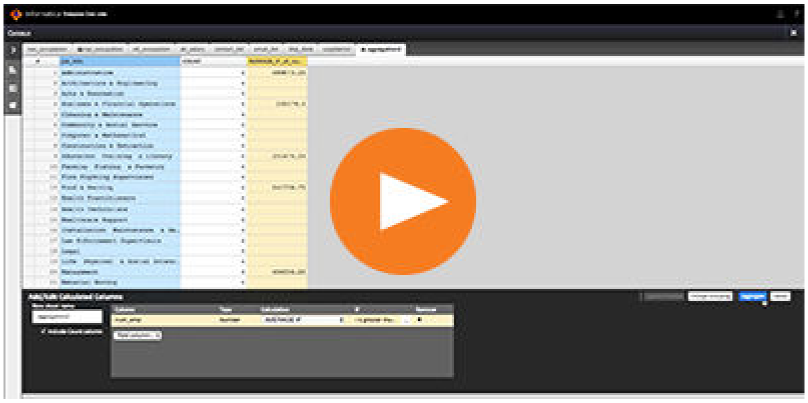
You can also transform columns in a worksheet into rows containing the column data in key value format. For example, if a worksheet includes columns containing price and supplier data for various products, you might want to summarize the data using products as the key, and price and supplier data as the values.

Aggregating Data

You can aggregate data in a worksheet by grouping selected columns together, and then performing calculations on columns containing number values to summarize the data.

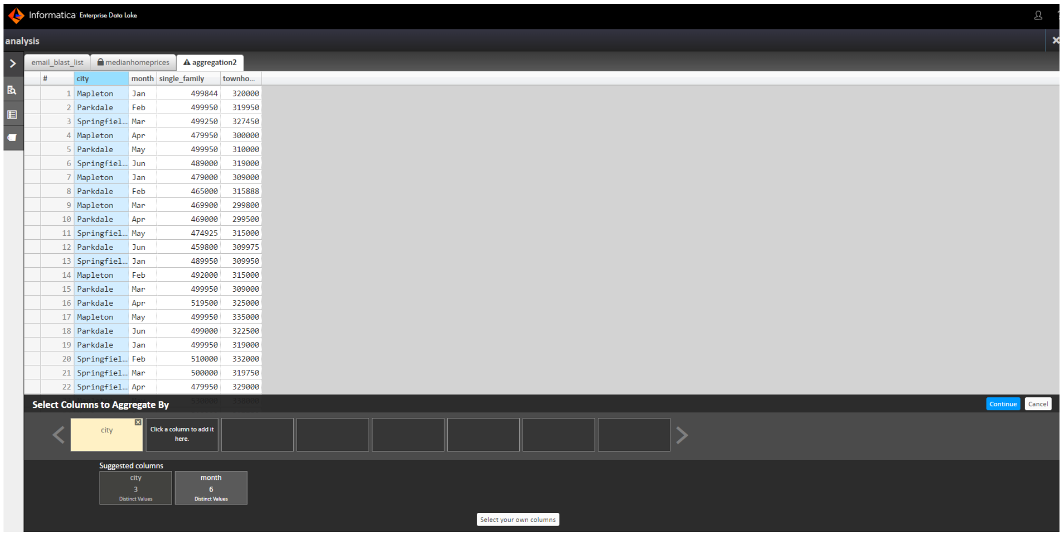
The aggregate operation adds a worksheet with a copy of the data in the original worksheet. The **Suggested columns** panel displays columns that Enterprise Data Lake suggests you group together. Enterprise Data Lake updates the worksheet with the columns you select.

Click the image below to view a video about the aggregate operation:



Example

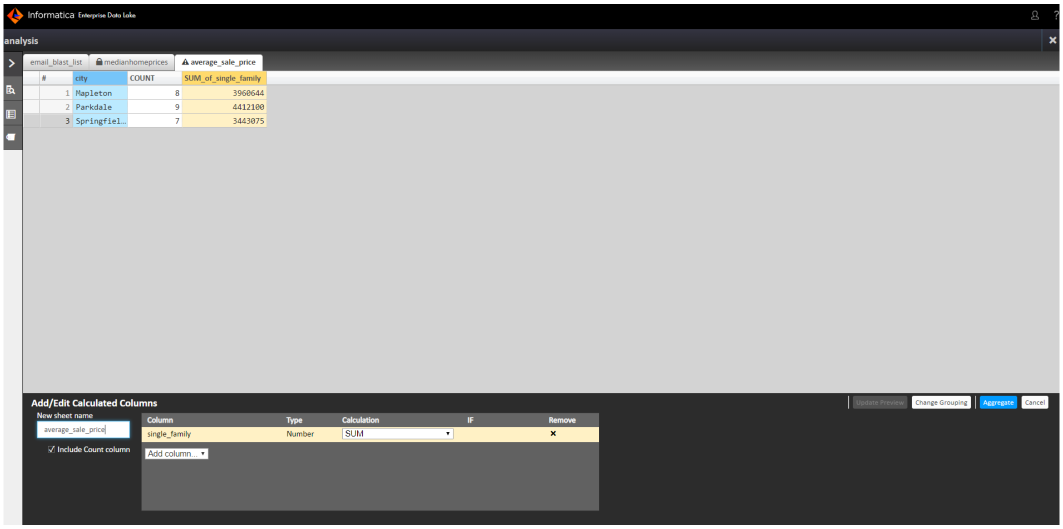
The following image shows an aggregation worksheet created in a project. Enterprise Data Lake suggests that you include the city and month columns in the grouping.



#	city	month	single_family	townho...
1	Mapleton	Jan	499844	320000
2	Parkdale	Feb	499950	319950
3	Springfiel...	Mar	499250	327450
4	Mapleton	Apr	479950	300000
5	Parkdale	May	499950	310000
6	Springfiel...	Jun	489000	319000
7	Mapleton	Jan	479000	309000
8	Parkdale	Feb	465000	315888
9	Mapleton	Mar	469900	299800
10	Parkdale	Apr	469000	299500
11	Springfiel...	May	474925	315000
12	Parkdale	Jun	459800	309975
13	Springfiel...	Jan	489950	309950
14	Mapleton	Feb	492000	315000
15	Parkdale	Mar	499950	309000
16	Parkdale	Apr	519500	325000
17	Mapleton	May	499950	335000
18	Parkdale	Jun	499000	322500
19	Parkdale	Jan	499950	319000
20	Springfiel...	Feb	510000	332000
21	Springfiel...	Mar	500000	319750
22	Springfiel...	Apr	479950	329000

city	month
3	6
Distinct Values	Distinct Values

Enterprise Data Lake displays the rows in the selected columns grouped by distinct values. The application adds a Count column indicating the occurrence of each value combination in the source data to the worksheet, as shown in the following image:.



#	city	COUNT	SUM_of_single_family
1	Mapleton	8	3960644
2	Parkdale	9	4412100
3	Springfiel...	7	3443075

New sheet name: average_sale_price

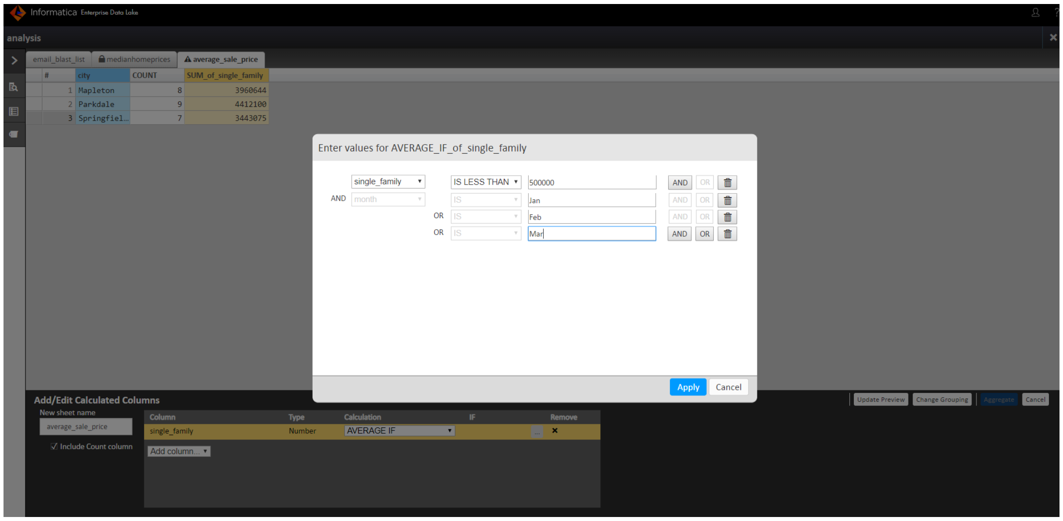
☑ Include Count column

Column	Type	Calculation	IF	Remove
single_family	Number	SUM		×

Select a column you want to include in the aggregation, and then select the calculation to apply to the column data. If you select an IF calculation, you can set conditions to restrict the calculation to a range of values. Use AND and OR logic to apply more than one condition.

- Use AND with all operators to include more than one column in a condition.
- Use OR with the IS, IS NOT and IS BETWEEN operators to include more than one value within a column in a condition.

The following example calculates the average price of single family homes sold for less than $500,000 in each city for first three months of the year. Use AND to include the single_family and month columns in the calculation, and then use OR to include specific values in the month column, as shown in the following image:



#	city	COUNT	SUM_of_single_family
1	Mapleton	8	3960644
2	Parkdale	9	4412100
3	Springfiel...	7	3443075

Enter values for AVERAGE_IF_of_single_family

single_family IS LESS THAN 500000
AND month IS Jan
OR IS Feb
OR IS Mar

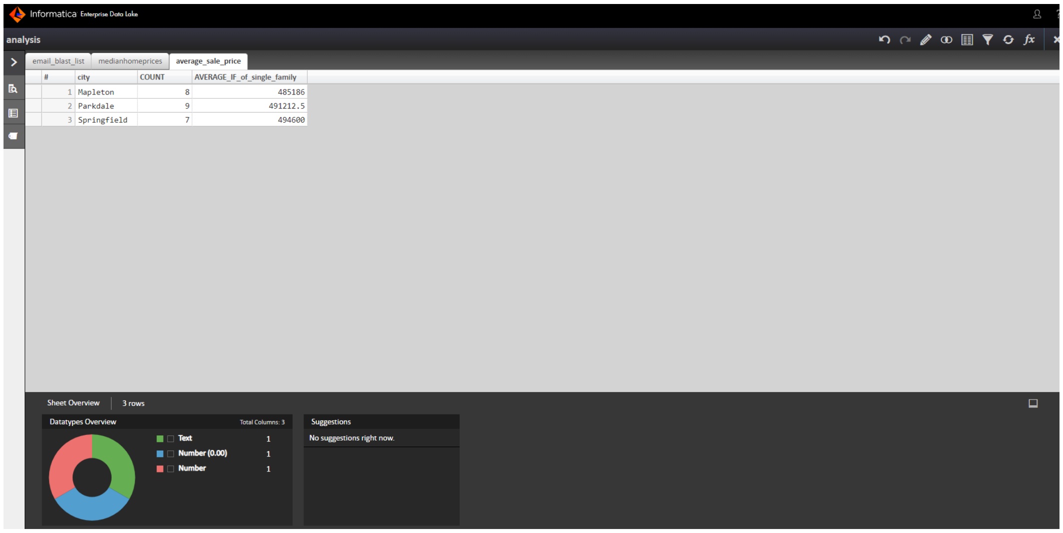
Add/Edit Calculated Columns

New sheet name: average_sale_price

☑ Include Count column

Column	Type	Calculation	IF	Remove
single_family	Number	AVERAGE IF	...	✕

Enterprise Data Lake aggregates the values in the calculated columns, as shown in the following image:



#	city	COUNT	AVERAGE_IF_of_single_family
1	Mapleton	8	485186
2	Parkdale	9	491212.5
3	Springfield	7	494600

🟩 ☐ Text	1
🟦 ☐ Number (0.00)	1
🟥 ☐ Number	1

No suggestions right now.

Aggregating Columns in a Worksheet

1. Open the worksheet containing the columns you want to aggregate.
2. Click the **Summarize** icon, and then select **Aggregate**.

   The application adds a worksheet with a copy of the data to the project.
3. The Suggested columns field displays columns that Enterprise Data Lake suggests you group together. Click each column you want to include in the grouping.

   Click **Select your own columns** to select other columns to include in the grouping.
4. Click **Continue**. Enterprise Data Lake displays the rows in the selected columns grouped by distinct value pairs.
5. Enter a name for the worksheet.

6. The application adds a Count column indicating the occurrence of each value combination in the source data to the worksheet. Clear the Count checkbox if you do not want to include the column in the aggregation.
7. Select the columns on which to perform an aggregate calculation from the **Add Column** menu, and then select the calculation to perform.
8. You can set an IF condition on certain calculations to limit the calculation to a range of values.
    a. Select a calculation containing IF in the calculation title.
    b. Enter the variables to use in the IF statement.
    c. Click **AND** or **OR** to add another condition.
    d. Click **Apply**.
9. Click **Update Preview** to update the worksheet with the calculation results.

    If you want to change the columns included in the operation, click **Change Grouping**.
10. Click **Aggregate**.

    Enterprise Data Lake adds the worksheet to the project.

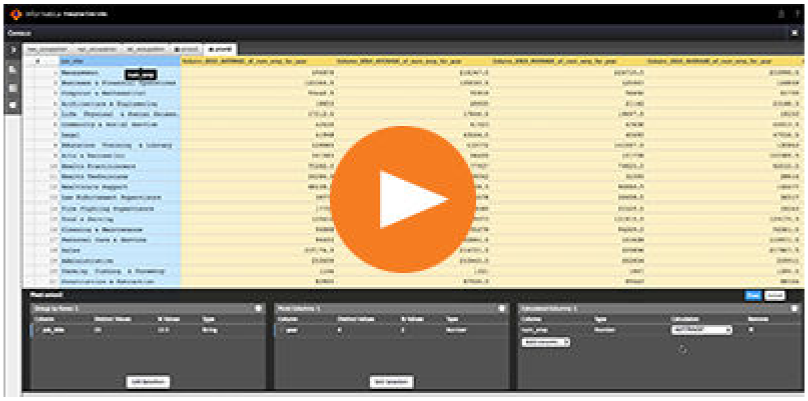
Pivoting Data

You can use the pivot operation to reshape the data in selected columns in a worksheet into a summarized format for analysis.

The pivot operation generates a pivot table in a new worksheet added to the project. To perform a pivot operation, you select one or more columns in the worksheet containing the values you want to group into rows in the pivot table. You also select one or more columns containing distinct values to pivot into new columns added to the pivot table.

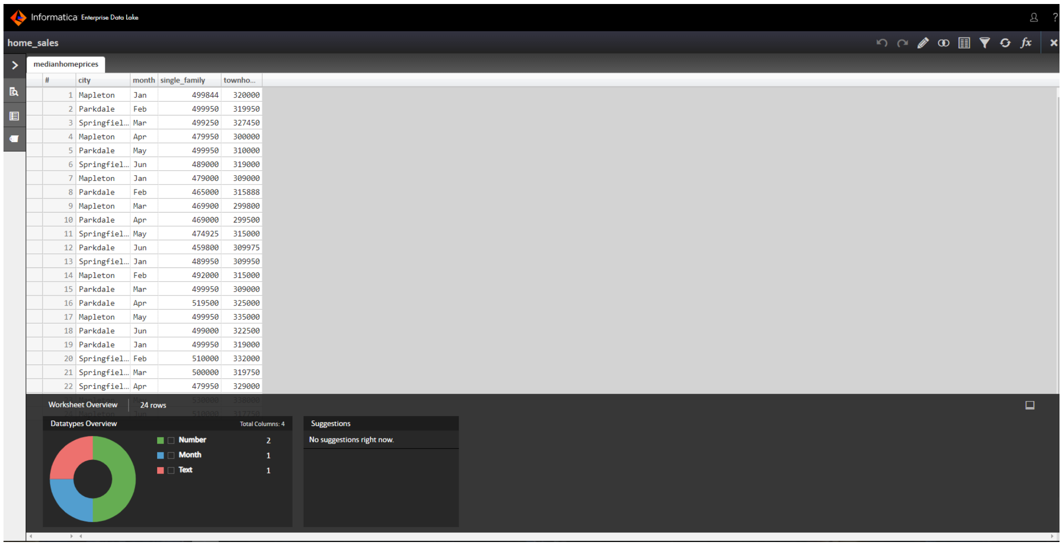
To summarize the data, you select the columns on which to perform a calculation, such as Sum or Average. Enterprise Data Lake displays the calculation results for each row in the corresponding column.

Click the image below to view a video about the union operation:



Example

The following example generates a pivot table summarizing the average price of single family homes sold in each city for the first six months of the year. The worksheet contains sale price data for single family homes and town homes sold in each city by month, as shown in the following image:

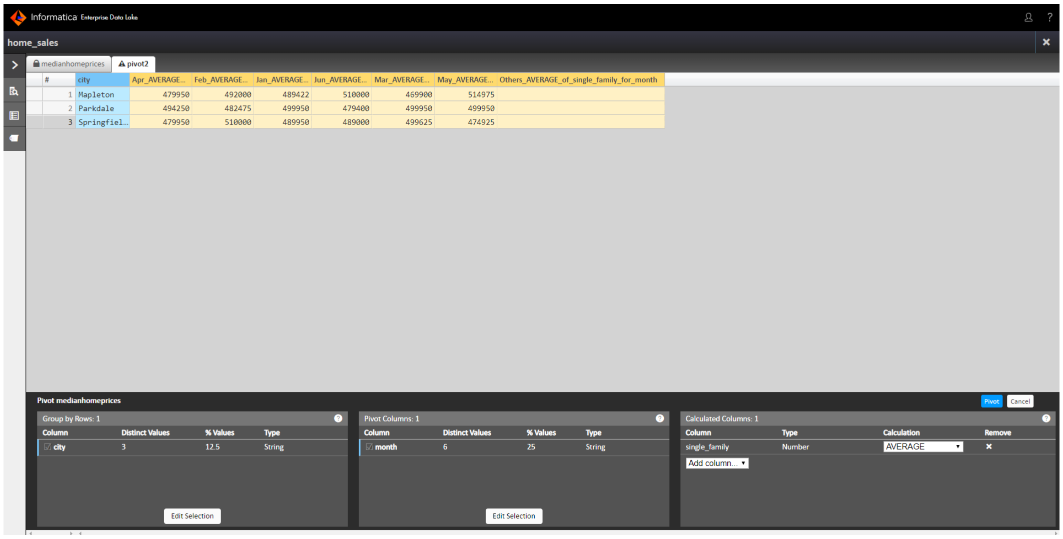


#	city	month	single_family	townho...
1	Mapleton	Jan	499844	320000
2	Parkdale	Feb	499950	319950
3	Springfiel...	Mar	499250	327450
4	Mapleton	Apr	479950	300000
5	Parkdale	May	499950	310000
6	Springfiel...	Jun	489000	319000
7	Mapleton	Jan	479000	309000
8	Parkdale	Feb	465000	315888
9	Mapleton	Mar	469900	299800
10	Parkdale	Apr	469000	299500
11	Springfiel...	May	474925	315000
12	Parkdale	Jun	459800	309975
13	Springfiel...	Jan	489950	309950
14	Mapleton	Feb	492000	315000
15	Parkdale	Mar	499950	309000
16	Parkdale	Apr	519500	325000
17	Mapleton	May	499950	335000
18	Parkdale	Jun	499000	322500
19	Parkdale	Jan	499950	319000
20	Springfiel...	Feb	510000	332000
21	Springfiel...	Mar	500000	319750
22	Springfiel...	Apr	479950	329000

To pivot the data:

- Select the city column as the group by row.
- Select the month column as the pivot column.
- Select single_family as the calculated column, and then select Average as the calculation to perform on the column values.

The following image shows the selected columns to pivot:



#	city	Apr_AVERAGE...	Feb_AVERAGE...	Jan_AVERAGE...	Jun_AVERAGE...	Mar_AVERAGE...	May_AVERAGE...	Others_AVERAGE_of_single_family_for_month
1	Mapleton	479950	492000	489422	510000	469900	514975	
2	Parkdale	494250	482475	499950	479400	499950	499950	
3	Springfiel...	479950	510000	489950	489000	499625	474925	

Group by Rows: 1

Column	Distinct Values	% Values	Type
☑ city	3	12.5	String

Pivot Columns: 1

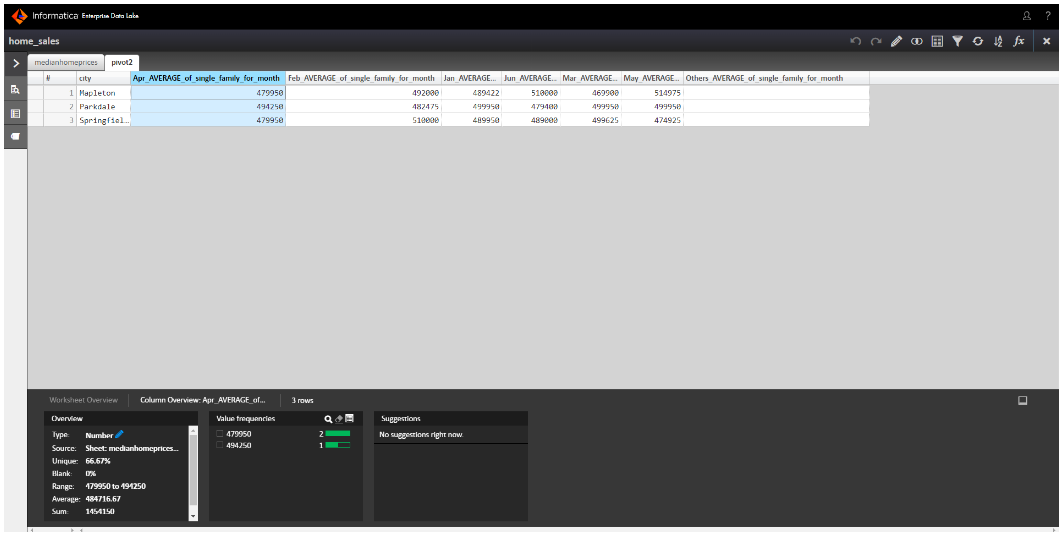
Column	Distinct Values	% Values	Type
☑ month	6	25	String

Calculated Columns: 1

Column	Type	Calculation	Remove
single_family	Number	AVERAGE	✕

Note that the Group by Rows panel and the Pivot Columns panel display the number of distinct values in each column, and the percentage of total values they comprise. You can use this information to determine the best columns to select.

The following image displays the resulting pivot table with the summarized data:


Informatica Enterprise Data Lake
home_sales
medianhomeprices
pivot2
Worksheet Overview
Column Overview: Apr_AVERAGE_of...
3 rows

#	city	Apr_AVERAGE_of_single_family_for_month	Feb_AVERAGE_of_single_family_for_month	Jan_AVERAGE...	Jun_AVERAGE...	Mar_AVERAGE...	May_AVERAGE...	Others_AVERAGE_of_single_family_for_month
1	Mapleton	479950	492000	489422	510000	469900	514975	
2	Parkdale	494250	482475	499950	479400	499950	499950	
3	Springfiel...	479950	510000	489950	489000	499625	474925	

Overview

Type: Number
Source: Sheet: medianhomeprices...
Unique: 66.67%
Blank: 0%
Range: 479950 to 494250
Average: 484716.67
Sum: 1454150

Value frequencies

☐ 479950	2
☐ 494250	1

Suggestions

No suggestions right now.

The name of each output column is derived from the pivot column value and the calculated data. Special characters in pivot column values are replaced with an underscore in the output column name. Examples of special characters include @,#,*,-, and %. To avoid underscores in column names, remove any special characters from pivot columns before performing a pivot operation.

The pivot operation is performed on the sampled data. To accommodate distinct values beyond the number of rows to load specified in the sampling criteria for the source worksheet, the operation adds a placeholder column named other_<calculationName> to the pivot worksheet. The placeholder column enables the additional values to be added to the Hive table when the asset is published.

Pivoting Columns in a Worksheet

1. Open the worksheet containing the columns you want to pivot.
2. Click the **Summarize** icon, and then select **Pivot**.

   The application adds a new worksheet containing a pivot table to the project.
3. Select the columns containing the keys to group into rows in the **Group By Rows** panel.
4. Click **Apply Selection**.

   The application groups the selected columns into rows in the pivot table.
5. Select the columns to pivot into new columns in the **Pivot Columns** panel.

   The operation creates a column in the pivot table for each distinct value in the selected columns. By default, the operation adds a maximum of 500 columns to the table. To limit the number of columns created in the pivot table to a manageable level, select columns that have lower distinct value counts. You can view the distinct value count for each column in the DISTINCT_VALUES column in the panel.
6. Click **Apply Selection**.

   The application adds a pivot column for each value in the selected columns to the pivot table.
7. Select each column on which to perform an aggregate calculation, and then select the calculation to perform in the **Calculated Columns** panel.

   The calculated values are displayed in the pivot columns in the worksheet. Note that columns used as group by or pivot columns cannot be selected as calculation columns.
8. Click **Pivot**.

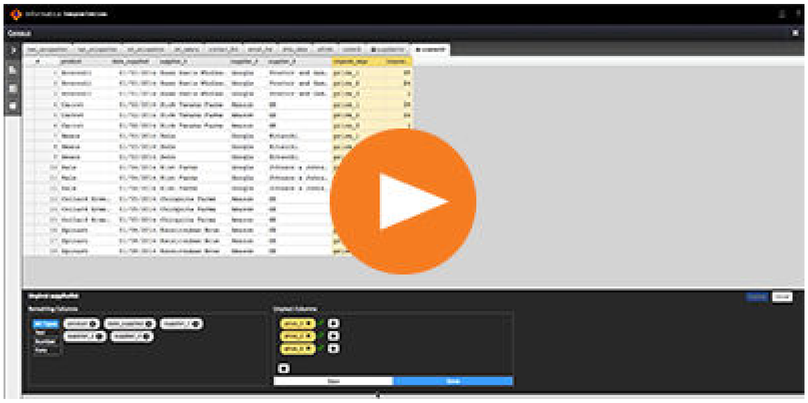
   Enterprise Data Lake generates the pivot table.

Unpivoting Data

You can use the unpivot operation to transform columns in a worksheet into rows containing the column data in key value format. The unpivot operation is useful when you want to visualize and organize data based on keys and corresponding values.

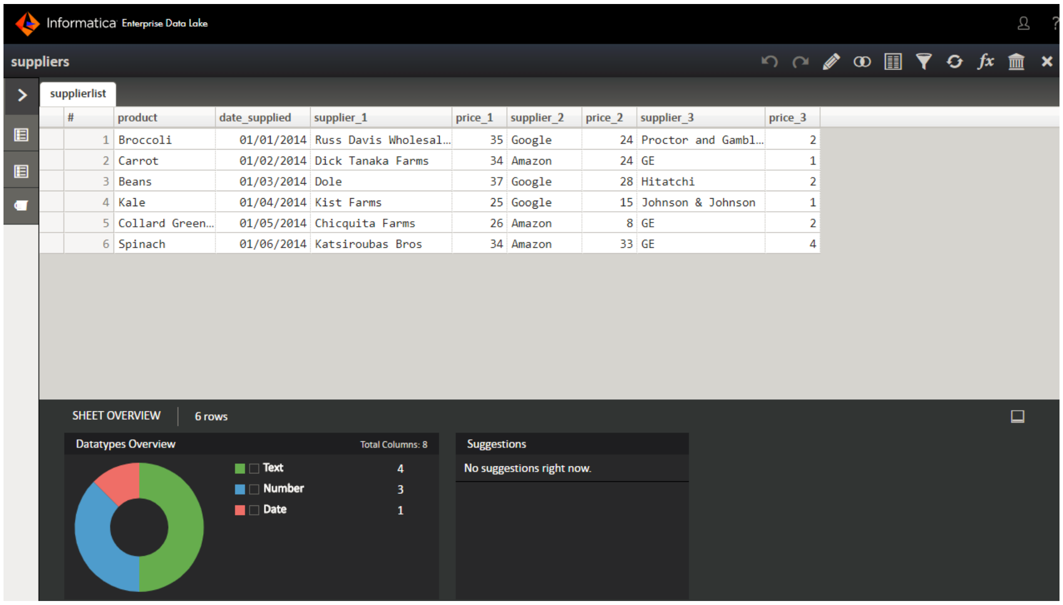
The unpivot operation adds a worksheet containing key and value columns to the project. The key column contains the name of each column included in the operation. The value column contains the corresponding column values.

Click the image below to view a video about the union operation:



Example

The following example shows how related columns can be transformed into key value format. The following image shows a worksheet containing supplier and price data for various products:



#	product	date_supplied	supplier_1	price_1	supplier_2	price_2	supplier_3	price_3
1	Broccoli	01/01/2014	Russ Davis Wholesal…	35	Google	24	Proctor and Gambl…	2
2	Carrot	01/02/2014	Dick Tanaka Farms	34	Amazon	24	GE	1
3	Beans	01/03/2014	Dole	37	Google	28	Hitatchi	2
4	Kale	01/04/2014	Kist Farms	25	Google	15	Johnson & Johnson	1
5	Collard Green…	01/05/2014	Chicquita Farms	26	Amazon	8	GE	2
6	Spinach	01/06/2014	Katsiroubas Bros	34	Amazon	33	GE	4

SHEET OVERVIEW | 6 rows

Datatypes Overview — Total Columns: 8

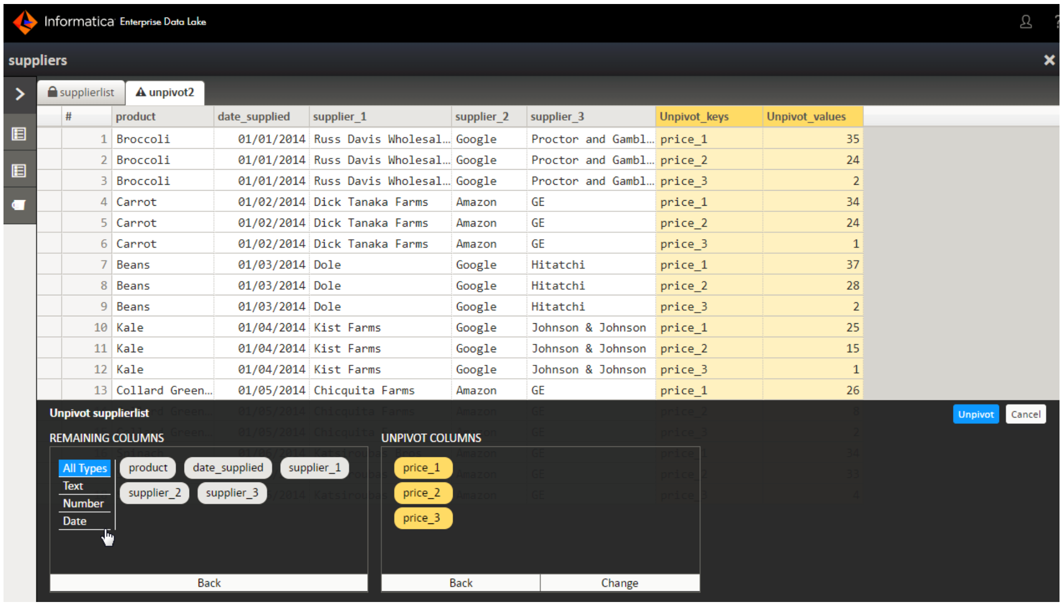
🟩	Text	4
🟦	Number	3
🟥	Date	1

Suggestions

No suggestions right now.

Enterprise Data Lake suggests that all number columns be included in the unpivot operation by default, and adds the columns in the **Unpivot Columns** panel. Columns not suggested for inclusion are displayed in the **Remaining Columns** panel. You can remove columns you do not want to include in the operation from the **Unpivot Columns** panel, or add columns in the **Remaining Columns** panel to the operation. All columns that you include in the operation must be of the same type.

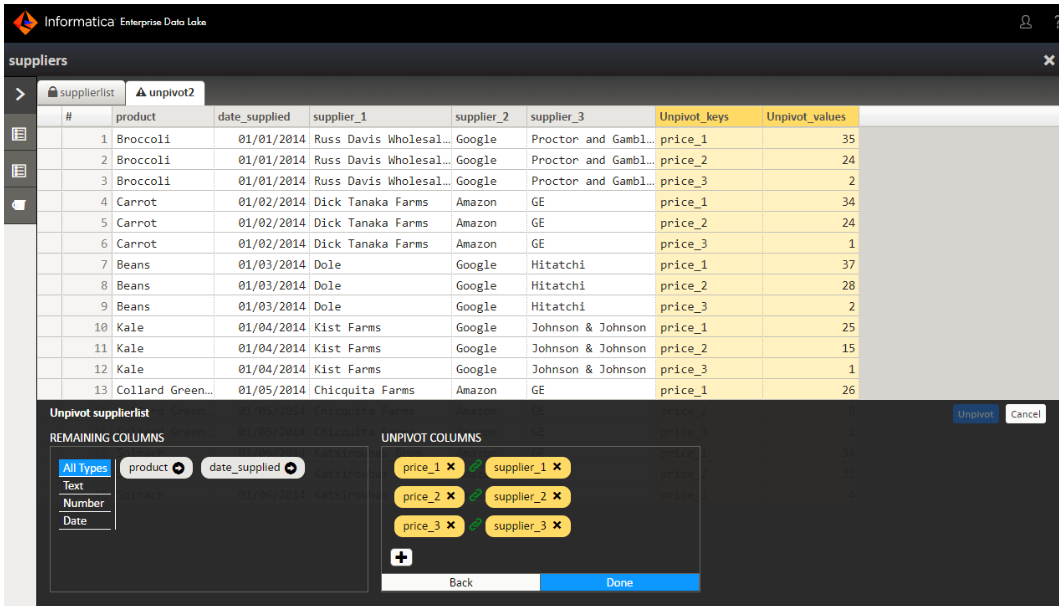
The following image shows the worksheet with the selected columns to unpivot:



#	product	date_supplied	supplier_1	supplier_2	supplier_3	Unpivot_keys	Unpivot_values
1	Broccoli	01/01/2014	Russ Davis Wholesal…	Google	Proctor and Gambl…	price_1	35
2	Broccoli	01/01/2014	Russ Davis Wholesal…	Google	Proctor and Gambl…	price_2	24
3	Broccoli	01/01/2014	Russ Davis Wholesal…	Google	Proctor and Gambl…	price_3	2
4	Carrot	01/02/2014	Dick Tanaka Farms	Amazon	GE	price_1	34
5	Carrot	01/02/2014	Dick Tanaka Farms	Amazon	GE	price_2	24
6	Carrot	01/02/2014	Dick Tanaka Farms	Amazon	GE	price_3	1
7	Beans	01/03/2014	Dole	Google	Hitatchi	price_1	37
8	Beans	01/03/2014	Dole	Google	Hitatchi	price_2	28
9	Beans	01/03/2014	Dole	Google	Hitatchi	price_3	2
10	Kale	01/04/2014	Kist Farms	Google	Johnson & Johnson	price_1	25
11	Kale	01/04/2014	Kist Farms	Google	Johnson & Johnson	price_2	15
12	Kale	01/04/2014	Kist Farms	Google	Johnson & Johnson	price_3	1
13	Collard Green…	01/05/2014	Chicquita Farms	Amazon	GE	price_1	26

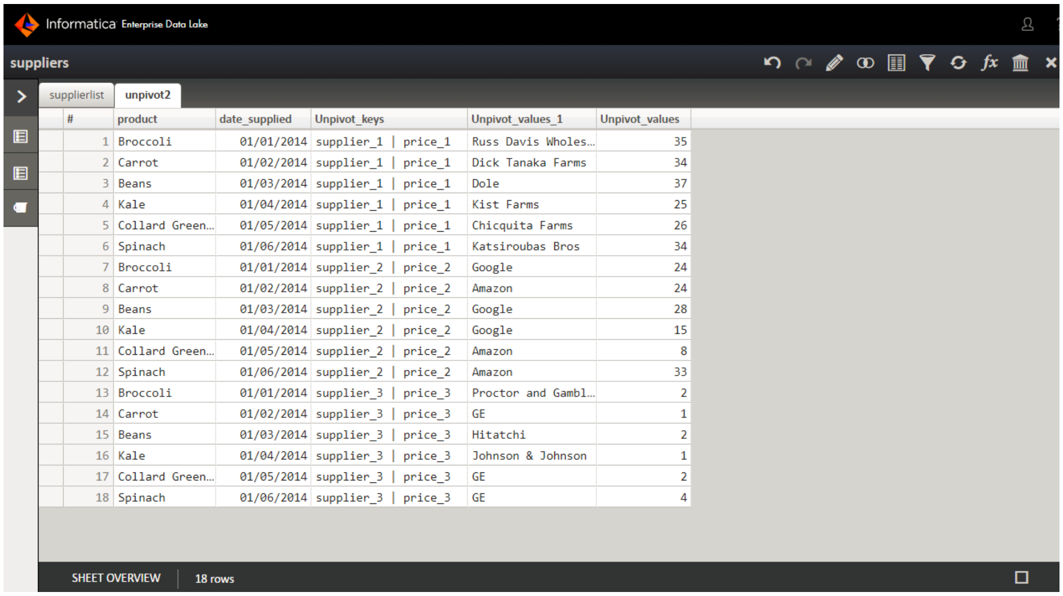
Because the price and supplier columns are related, you can link the columns in the **Unpivot Columns** panel. Note that if you link one column in the **Unpivot Columns** panel with a related column, you must link all of the columns in the panel with related columns.

The following image shows the linked unpivot columns:



#	product	date_supplied	supplier_1	supplier_2	supplier_3	Unpivot_keys	Unpivot_values
1	Broccoli	01/01/2014	Russ Davis Wholesal…	Google	Proctor and Gambl…	price_1	35
2	Broccoli	01/01/2014	Russ Davis Wholesal…	Google	Proctor and Gambl…	price_2	24
3	Broccoli	01/01/2014	Russ Davis Wholesal…	Google	Proctor and Gambl…	price_3	2
4	Carrot	01/02/2014	Dick Tanaka Farms	Amazon	GE	price_1	34
5	Carrot	01/02/2014	Dick Tanaka Farms	Amazon	GE	price_2	24
6	Carrot	01/02/2014	Dick Tanaka Farms	Amazon	GE	price_3	1
7	Beans	01/03/2014	Dole	Google	Hitatchi	price_1	37
8	Beans	01/03/2014	Dole	Google	Hitatchi	price_2	28
9	Beans	01/03/2014	Dole	Google	Hitatchi	price_3	2
10	Kale	01/04/2014	Kist Farms	Google	Johnson & Johnson	price_1	25
11	Kale	01/04/2014	Kist Farms	Google	Johnson & Johnson	price_2	15
12	Kale	01/04/2014	Kist Farms	Google	Johnson & Johnson	price_3	1
13	Collard Green…	01/05/2014	Chicquita Farms	Amazon	GE	price_1	26

When you apply the unpivot operation, the price and supplier column names are concatenated into keys. The corresponding column values appear in each row, as shown in the following image:



#	product	date_supplied	Unpivot_keys	Unpivot_values_1	Unpivot_values
1	Broccoli	01/01/2014	supplier_1 | price_1	Russ Davis Wholes...	35
2	Carrot	01/02/2014	supplier_1 | price_1	Dick Tanaka Farms	34
3	Beans	01/03/2014	supplier_1 | price_1	Dole	37
4	Kale	01/04/2014	supplier_1 | price_1	Kist Farms	25
5	Collard Green...	01/05/2014	supplier_1 | price_1	Chicquita Farms	26
6	Spinach	01/06/2014	supplier_1 | price_1	Katsiroubas Bros	34
7	Broccoli	01/01/2014	supplier_2 | price_2	Google	24
8	Carrot	01/02/2014	supplier_2 | price_2	Amazon	24
9	Beans	01/03/2014	supplier_2 | price_2	Google	28
10	Kale	01/04/2014	supplier_2 | price_2	Google	15
11	Collard Green...	01/05/2014	supplier_2 | price_2	Amazon	8
12	Spinach	01/06/2014	supplier_2 | price_2	Amazon	33
13	Broccoli	01/01/2014	supplier_3 | price_3	Proctor and Gambl...	2
14	Carrot	01/02/2014	supplier_3 | price_3	GE	1
15	Beans	01/03/2014	supplier_3 | price_3	Hitatchi	2
16	Kale	01/04/2014	supplier_3 | price_3	Johnson & Johnson	1
17	Collard Green...	01/05/2014	supplier_3 | price_3	GE	2
18	Spinach	01/06/2014	supplier_3 | price_3	GE	4

Unpivoting Columns in a Worksheet

1. Open the worksheet containing the columns to unpivot.
2. Click the **Summarize** icon, and then select **Unpivot**.

   The application adds a worksheet containing the suggested key and value columns to the project. The **Unpivot Columns** panel contains the suggested columns to unpivot. The **Remaining Columns** panel contains the columns not included in the operation.
3. Click **Change** in the **Unpivot Columns** panel to add or remove columns.

   Select each column to add in the **Remaining Columns** panel. Each column you add must be of the same type as the columns displayed in the **Unpivot Columns** panel.

   To link two related columns, click the **+** symbol next to a column in the **Unpivot Columns** panel, and then click the related column in the **Remaining Columns** panel.
4. Click **Done** when you finish selecting the columns to use in the operation.
5. Click **Unpivot**.

   Enterprise Data Lake transforms the unpivoted columns into key value format in the worksheet.

One Hot Encoding

You can use the one hot encoding operation to determine the existence of a string value in a selected column within each row in a worksheet. You might use the one hot encoding operation to convert categorical values in a worksheet to numeric values required by machine learning algorithms.

The operation creates output columns in the worksheet for the 500 most frequently occurring distinct values in the selected column. The output column for each value contains either a 1 indicating that the value exists in the row, or a 0 indicating that the value does not exist.

The existence of additional distinct values not included in the 500 most frequently occurring distinct values is indicated in an output column named Is_otherValue_<column_name>. A value of 1 in the Is_otherValue_<column_name> column indicates that additional distinct values exist in the selected column.

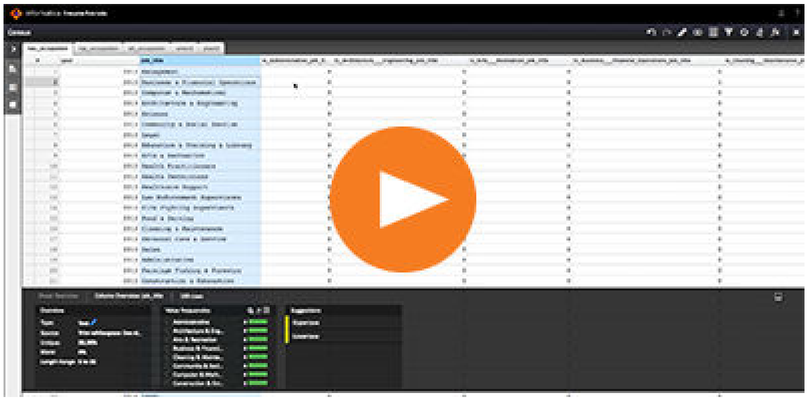
Output columns for NULL values are named Is_BLANK_<columnName>.

If you apply one hot encoding to a column with values that contain special characters, the special characters are replaced with an underscore in the output column names. Examples of special characters include @,#,*,-, and %. To avoid underscores in column names, remove any special characters from the columns.

You can copy a recipe step that includes a one hot encoding operation into a worksheet. Map the column in the copied step to a column in the worksheet that contains the same data as the column in the source worksheet. When you insert the copied step into a worksheet, the output columns created in the source worksheet are copied into the worksheet.

When you refresh the data in a worksheet that uses one hot encoding, the output columns for values removed from or added to the source worksheet are persisted in the worksheet.

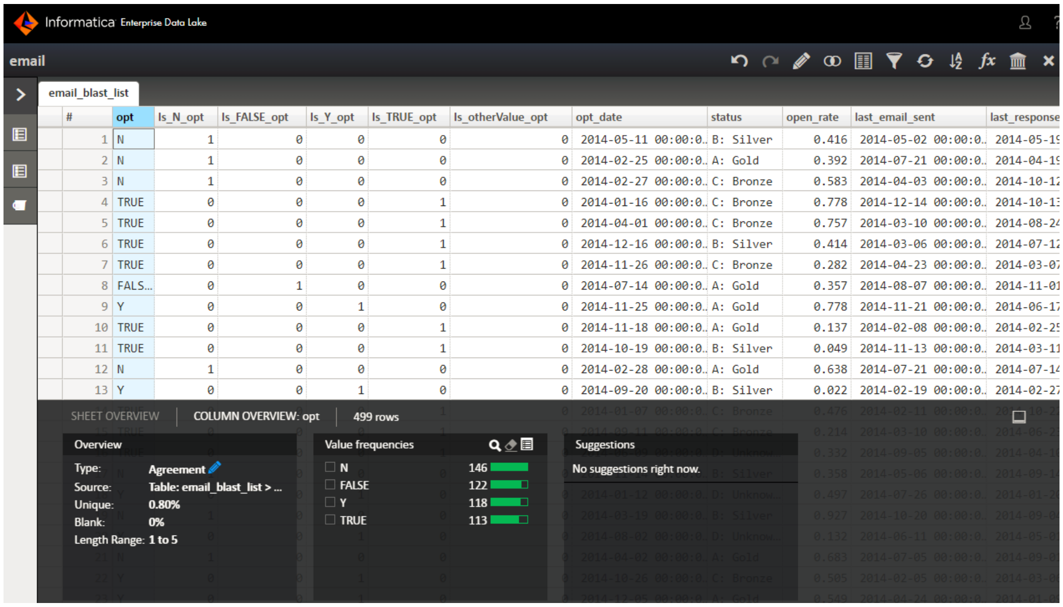
Click the image below to view a video about the one hot encoding operation:



Example

The following example shows the one hot encoding operation applied to the opt column in the worksheet. The existence of each value in the selected column is indicated in output columns added to the worksheet. Each output column is named Is_<valueName>_<columnName>.

The following image shows the one hot encoding operation applied to the worksheet:



#	opt	Is_N_opt	Is_FALSE_opt	Is_Y_opt	Is_TRUE_opt	Is_otherValue_opt	opt_date	status	open_rate	last_email_sent	last_response
1	N	1	0	0	0	0	2014-05-11 00:00:0..	B: Silver	0.416	2014-05-02 00:00:0..	2014-05-1
2	N	1	0	0	0	0	2014-02-25 00:00:0..	A: Gold	0.392	2014-07-21 00:00:0..	2014-04-1
3	N	1	0	0	0	0	2014-02-27 00:00:0..	C: Bronze	0.583	2014-04-03 00:00:0..	2014-10-1
4	TRUE	0	0	0	1	0	2014-01-16 00:00:0..	C: Bronze	0.778	2014-12-14 00:00:0..	2014-10-1
5	TRUE	0	0	0	1	0	2014-04-01 00:00:0..	C: Bronze	0.757	2014-03-10 00:00:0..	2014-08-2
6	TRUE	0	0	0	1	0	2014-12-16 00:00:0..	B: Silver	0.414	2014-03-06 00:00:0..	2014-07-1
7	TRUE	0	0	0	1	0	2014-11-26 00:00:0..	C: Bronze	0.282	2014-04-23 00:00:0..	2014-03-0
8	FALS...	0	1	0	0	0	2014-07-14 00:00:0..	A: Gold	0.357	2014-08-07 00:00:0..	2014-11-01
9	Y	0	0	1	0	0	2014-11-25 00:00:0..	A: Gold	0.778	2014-11-21 00:00:0..	2014-06-1
10	TRUE	0	0	0	1	0	2014-11-18 00:00:0..	A: Gold	0.137	2014-02-08 00:00:0..	2014-02-2
11	TRUE	0	0	0	1	0	2014-10-19 00:00:0..	B: Silver	0.049	2014-11-13 00:00:0..	2014-03-1
12	N	1	0	0	0	0	2014-02-28 00:00:0..	A: Gold	0.638	2014-07-21 00:00:0..	2014-07-1
13	Y	0	0	1	0	0	2014-09-20 00:00:0..	B: Silver	0.022	2014-02-19 00:00:0..	2014-02-2

Overview

Type: Agreement
Source: Table: email_blast_list > ...
Unique: 0.80%
Blank: 0%
Length Range: 1 to 5

Value frequencies

Value	Count
☐ N	146
☐ FALSE	122
☐ Y	118
☐ TRUE	113

Suggestions

No suggestions right now.

Using One Hot Encoding

1. Open the worksheet in which you want to apply one hot encoding.
2. Select the column on which you want perform the one hot encoding operation.
3. Click the **Edit** icon, and then select **Apply one hot encoding**.

   Enterprise Data Lake adds an output column for each distinct value in the selected column to the worksheet.

CHAPTER 7

Publish Data

This chapter includes the following topics:

Data Publication Overview

Data publication is the process of making prepared data available.

When you publish prepared data, Enterprise Data Lake applies the recipe to the data in the input source. Enterprise 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.

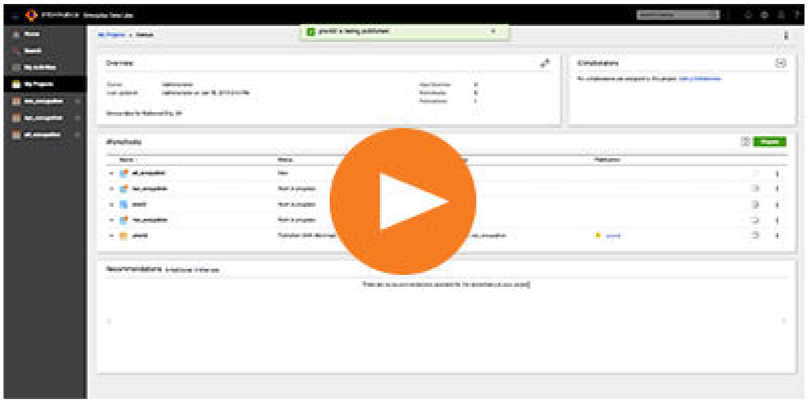
You can schedule publication of data assets. Scheduling publication activities enables you to publish updated data assets on a recurring basis. For more information about scheduling publication activities, see Chapter 9, “Schedule Export, Import and Publish Activities” on page 97.

You can also choose to save a recipe as a mapping that developers can modify and operationalize using Informatica Developer (the Developer tool). You can save the mapping to the Model repository associated with the Enterprise Data Lake Service, or you can save the mapping to an .xml file and import it into a different Model repository.

Publishing Prepared Data to the Data Lake

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

Click the image below to view a video about the publish operation:



1. In the **Projects** view, select the worksheet that you want to publish.
2. Click the **Publish** icon.

   The **Publish** dialog appears.
3. Select the Hive schema in the data lake in which to publish the data.
4. Enter a name for the Hive table.
5. If sampling filters are set on columns in the worksheet, the **Apply sampling filter on publication** option is selected by default. Clear this option if you do not want to apply the filters shown in the dialog when publishing the data asset.
6. Select **Enable Full Profiling** if you want to apply full profiling to the asset.

   Note that the checkbox is not enabled if profiling is disabled on the Hive resource. Contact your Enterprise Data Catalog administrator to enable profiling on the resource.
7. Click **Next**.
8. Specify when you want to publish the worksheet.
   - Select **Publish Now** to publish the asset now.
   - Select **Schedule Publication** to schedule the publish activity.
     If you decide to schedule the publish activity, specify what to do if the asset already exists in the data lake.
     - Select **Append** to append the data to an existing table. Verify that the column names, the data types of the columns, the precision and scale of columns, and sequence of columns is the same in the importing table and the existing table. The import will fail if any of these do not match.
     - Select **Overwrite** to overwrite the existing table. The application drops the table and re-creates the table with the data from the importing table.

     For more information about scheduling a publish activity by creating a schedule or using an existing schedule, see “Scheduling Overview” on page 97.
9. Click **OK**.

   Check the status of the publish activity on My Activities page.

Saving a Recipe as a Mapping

You can save a recipe as a mapping that developers can modify and operationalize using the Developer tool.

You can save the mapping to the Model repository associated with the Enterprise Data Lake Service. You must specify the Hive schema and the Hive table defined in the mapping output. The Data Integration Service writes the data to the specified schema and table when it runs the mapping.

You can also choose to save the mapping as an .xml file. Developers can import the file to another Model repository.

1. In the **Projects** view, select the worksheet containing the recipe that you want to save as a mapping.
2. Click the **Manage Worksheets** icon, and then select **Save As Mapping**.

   The **Save As Mapping** dialog box appears.
3. Choose how you want to save the mapping.
   - Click **Model Repository** to save the mapping to the Model repository associated with the Enterprise Data Lake Service.
     Specify the Hive schema and the Hive table to which the mapping writes data. The application adds the table and schema names to the target definition in the mapping.
   - Click **Local File** to save the mapping as an .xml file.
4. If sampling filters are set on columns in the worksheet, the **Apply sampling filter on publication** option is selected by default. Clear this option if you do not want to apply the filters shown in the dialog when saving the data asset.
5. Click **Save**.

Downloading a Data Asset

Download a data asset or a publication as a comma-separated value (.csv) file or a Tableau Data Extract (.tde) file. You can use the file with a third-party application to run reports on the data.

1. Open the project containing the asset or publication you want to download.
2. In the Worksheets panel, click the name of a data asset in the **Input Sources** column, or of a publication in the **Publications** column.
3. On the Overview page, select the **Manage Data Assets** icon, and then click **Download CSV** or **Download TDE**.

   Enterprise Data Lake downloads the data asset or publication as a CSV file or TDE file.

Operationalize Mappings

You can work with your Informatica administrator to operationalize your data preparation steps so that your prepared data is uploaded to the data lake on a regular basis.

When you publish your prepared data to the data lake. Enterprise Data Lake generates a mapping from the worksheet recipe. The administrator can deploy the mapping and run it in the Hadoop environment to load the prepared data back into the data lake on a regular basis.

CHAPTER 8

Visualize and Assess Published Data

This chapter includes the following topics:

Overview

You can assess a published asset visually to make sure that the data is appropriate for analysis from content and quality perspectives. You can then choose to fix the recipe and republish the worksheet.

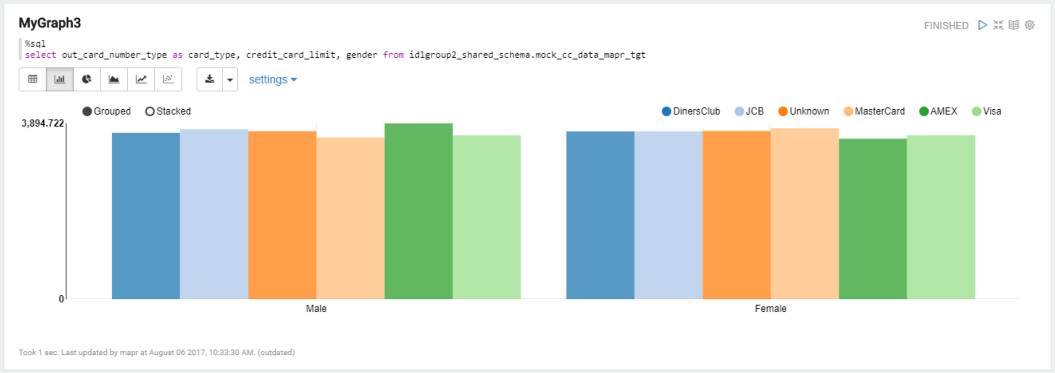
Enterprise Data Lake uses Apache Zeppelin to view the worksheets in the form of a notebook that contains graphs and charts. You can assess and validate data by performing the following activities:

- View summary statistics. You can view summary statistics for the newly created data assets and assess the completeness, conformity, consistency, integrity, timeliness, and uniqueness of the data. You can view summary statistics at table level and at column level. You can also view statistics for multiple columns.
- View relationships between different columns. You can view the relationships between different columns in the form of pie charts, graphs, and histograms.
- Validate the recipes. You can view the data in the form of graphics to identify any anomalies and correct them.
- Create multiple charts and graphs. You can run new queries on the existing columns to create new charts and graphs to validate the data.

When you click the visualize icon for a published worksheet in a project, Enterprise Data Lake creates a Zeppelin notebook for the data asset. You can continue to work on the notebook and share it with other Enterprise Data Lake users.

You can view the Zeppelin notebook from the project page or from the data assets page. You cannot make changes to the notebook when you view it from the Visualization tab of the data assets page.

The following image shows a visualized graph in a visualization notebook:



```
%sql
select out_card_number_type as card_type, credit_card_limit, gender from idlgroup2_shared_schema.mock_cc_data_mapr_tgt
```

Took 1 sec. Last updated by mapr at August 06 2017, 10:33:30 AM. (outdated)

Prerequisites

You must be able to access Zeppelin to prepare and view notebooks.

- Make sure that Zeppelin is configured by the Enterprise Data Lake administrator on Spark. For more information about Zeppelin configuration, see Apache Zeppelin documentation.
- Make sure the administrator has configured users to the Zeppelin configuration.
- Make sure you collect the Zeppelin login credentials from the Administrator and use them to log in to the notebook. You can log in to Zeppelin from Enterprise Data Lake.

Visualizing Published Data

After preparing data and publishing the worksheet, you can visualize the data before you operationalize the recipes created during publishing.

1. In the **Projects** view, select the worksheet that you want to visualize.
2. Click the visualize icon corresponding to the worksheet.

   Note: The Visualize icon will be enabled only if Zeppelin is configured in Informatica Administrator and the successfully published assets are allowed to be visualized.
3. Enter the username and password to access the Visualization page. You can collect these details from the Administrator.
4. Run queries on the notebook page for the worksheet to validate the data. You can run a query on an entire table, single column, or multiple columns in a table. You can also create queries establishing relationship between different columns in the worksheet.

You can delete the Zeppelin notebook if you don't need it for analysis and validation. On the Visualization page, click the menu and select **Delete**. Your assessment notebook associated with the data asset will also be deleted along with this Zeppelin notebook.

Sharing the Notebook

You can share the notebooks with other users of the lake. Other users can access the assessment notebook from the Visualization tab of the published data asset. Unless you share the notebook, only the collaborators of the project can view the notebook.

To share the notebook with all users who have permissions to access the data asset, click **Share**. Make sure you click **Share** every time you update the notebook by adding a graph or running a new query.

Click **Unshare** to avoid letting other users access the notebook. Only the collaborators of the project will be able to view and edit the worksheet, depending on their roles. If the visualization notebook is not shared from the project, a warning message appears indicating that you need to create and share the notebook for others to access it.

Smart Suggestions

When you open the notebook of a publication, you can see some smart recommendations.

When you open the notebook for the first time after it is published, you can see histogram visualizations of derived numeric columns. You can see a maximum of four derived numeric columns. If the publication does not contain any derived numerical columns, you can see a select * from table query in the first paragraph of the notebook.

CHAPTER 9

Schedule Export, Import and Publish Activities

This chapter includes the following topics:

Scheduling Overview

When you import, export, or publish a data asset, you can choose to complete the activity at that time, or you can choose to schedule the activity. Scheduling an activity enables you to import, export or publish updated data assets on a recurring basis.

If you decide to schedule an activity, you can create a new schedule, or you can select an existing schedule. Other users can use schedules that you create, and you can use schedules created by other users. If you use a schedule owned by another user, changes the owner makes to the schedule might impact your scheduled activity.

When you schedule an activity, you can specify what to do if an asset with the same name as the asset you are importing, exporting, or publishing already exists. You can choose to overwrite the asset, or you can append the new data to the asset.

The Scheduler Service must be running to create or use a schedule. Contact your Informatica administrator if you encounter an error indicating that the Scheduler Service is not available.

Creating a Schedule

You can create a new schedule when you import, export or publish a data asset. You can also create a new schedule from the Manage My Schedules page.

Contact your Informatica administrator if you do not have the privileges required to create a schedule.

1. To create a new schedule:

  - Select **Create New Schedule** in the import, export or publish wizard.
  - On the Enterprise Data Lake home page, select **My Activities**. Click the **Manage My Schedules** tab, and then click **New Schedule**.

2. Provide the schedule details.

   The following table describes the schedule detail properties:

Property	Description
Schedule Name	The required schedule name. The name can contain alphanumeric characters and any of the following special characters: _, ., -. or $.
Description	An optional description that might help other users decide whether to use your schedule.
Starts	The date and time the schedule takes effect.
Time Zone	The time zone the schedule is based on.
Frequency	The frequency at which to run the activity. The options displayed are based on the frequency value you select. If you select Weekly, you can select multiple days on which to run the activity.
Repeat Options	The time period for which to repeat the activity.

3. Click **Save**.

   The schedule is added to the repository used by the Scheduler Service and can be selected by other users.

Managing Scheduled Activities

Use the My Scheduled Activities page to delete or suspend a scheduled activity, or to select a new schedule to use for an activity.

1. On the Enterprise Data Lake home page, click **My Activities**, and then click the **My Scheduled Activities** tab.
2. Highlight the scheduled activity you want to modify.
   - Click the **Edit** icon to select a new schedule to use.
   - Click the **Delete** icon to delete all occurrences of the scheduled activity.
   - Click the **Suspend** icon to suspend all occurrences of the scheduled activity. You can resume the scheduled activity at a later time.

Managing Schedules

Use the Manage My Schedules page to modify, delete or suspend schedules that you own.

Modifications that you make to a schedule impact all activities that use the schedule. When you delete or suspend a schedule, the schedule is unavailable to all activities owned by other users, and the activities fail to run. Data Integration Service jobs and other service jobs that use the schedule also fail to run.

You can only manage schedules that you own. Users with administrator privileges can manage any schedule.

1. On the Enterprise Data Lake home page, click **My Activities**, and then click the **Manage My Schedules** tab.
2. Select the schedule you want to modify in the **All Schedules** panel.

   The details for the selected schedule load in the page.
3. Before you modify the schedule, compare the values shown for the Total Activities and Only My Activities properties to determine if activities owned by other users use the schedule.
4. Select the action you want to take from the **Actions** menu.
   - Select **Edit Schedule** to modify the schedule settings.
   - Select **Delete Schedule** to delete the schedule and all activities that use the schedule from the domain.
   - Select **Suspend Schedule** to suspend the schedule and all activities that use the schedule.

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 Enterprise 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, Enterprise Data Lake applies the recipe to the data in the input source. Enterprise 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. Enterprise 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, Enterprise Data Lake applies the recipe to the data in the input source. Enterprise 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, Enterprise 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, Enterprise 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, Enterprise Data Lake applies the recipe to the data in the input source and creates a new data asset.

INDEX