Informatica®
10.2

Intelligent Data Lake User Guide

Informatica Intelligent Data Lake User Guide
10.2
October 2017

© Copyright Informatica LLC 2016, 2019

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

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

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

Portions of this software and/or documentation are subject to copyright held by third parties, including without limitation: Copyright DataDirect Technologies. All rights reserved. Copyright © Sun Microsystems. All rights reserved. Copyright © RSA Security Inc. All Rights Reserved. Copyright © Ordinal Technology Corp. All rights reserved. Copyright © Aandacht c.v. All rights reserved. Copyright Genivia, Inc. All rights reserved. Copyright Isomorphic Software. All rights reserved. Copyright © Meta Integration Technology, Inc. All rights reserved. Copyright © Intalio. All rights reserved. Copyright © Oracle. All rights reserved. Copyright © Adobe Systems Incorporated. All rights reserved. Copyright © DataArt, Inc. All rights reserved. Copyright © ComponentSource. All rights reserved. Copyright © Microsoft Corporation. All rights reserved. Copyright © Rogue Wave Software, Inc. All rights reserved. Copyright © Teradata Corporation. All rights reserved. Copyright © Yahoo! Inc. All rights reserved. Copyright © Glyph & Cog, LLC. All rights reserved. Copyright © Thinkmap, Inc. All rights reserved. Copyright © Clearpace Software Limited. All rights reserved. Copyright © Information Builders, Inc. All rights reserved. Copyright © OSS Nokalva, Inc. All rights reserved. Copyright Edifecs, Inc. All rights reserved. Copyright Cleo Communications, Inc. All rights reserved. Copyright © International Organization for Standardization 1986. All rights reserved. Copyright © ej-technologies GmbH. All rights reserved. Copyright © Jaspersoft Corporation. All rights reserved. Copyright © International Business Machines Corporation. All rights reserved. Copyright © yWorks GmbH. All rights reserved. Copyright © Lucent Technologies. All rights reserved. Copyright © University of Toronto. All rights reserved. Copyright © Daniel Veillard. All rights reserved. Copyright © Unicode, Inc. Copyright IBM Corp. All rights reserved. Copyright © MicroQuill Software Publishing, Inc. All rights reserved. Copyright © PassMark Software Pty Ltd. All rights reserved. Copyright © LogiXML, Inc. All rights reserved. Copyright © 2003-2010 Lorenzi Davide, All rights reserved. Copyright © Red Hat, Inc. All rights reserved. Copyright © The Board of Trustees of the Leland Stanford Junior University. All rights reserved. Copyright © EMC Corporation. All rights reserved. Copyright © Flexera Software. All rights reserved. Copyright © Jinfonet Software. All rights reserved. Copyright © Apple Inc. All rights reserved. Copyright © Telerik Inc. All rights reserved. Copyright © BEA Systems. All rights reserved. Copyright © PDFlib GmbH. All rights reserved. Copyright © Orientation in Objects GmbH. All rights reserved. Copyright © Tanuki Software, Ltd. All rights reserved. Copyright © Ricebridge. All rights reserved. Copyright © Sencha, Inc. All rights reserved. Copyright © Scalable Systems, Inc. All rights reserved. Copyright © jQWidgets. All rights reserved. Copyright © Tableau Software, Inc. All rights reserved. Copyright© MaxMind, Inc. All Rights Reserved. Copyright © TMate Software s.r.o. All rights reserved. Copyright © MapR Technologies Inc. All rights reserved. Copyright © Amazon Corporate LLC. All rights reserved. Copyright © Highsoft. All rights reserved. Copyright © Python Software Foundation. All rights reserved. Copyright © BeOpen.com. All rights reserved. Copyright © CNRI. All rights reserved.

This product includes software developed by the Apache Software Foundation (http://www.apache.org/), and/or other software which is licensed under various versions of the Apache License (the "License"). You may obtain a copy of these Licenses at http://www.apache.org/licenses/. Unless required by applicable law or agreed to in writing, software distributed under these Licenses is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licenses for the specific language governing permissions and limitations under the Licenses.

This product includes software which was developed by Mozilla (http://www.mozilla.org/), software copyright The JBoss Group, LLC, all rights reserved; software copyright © 1999-2006 by Bruno Lowagie and Paulo Soares and other software which is licensed under various versions of the GNU Lesser General Public License Agreement, which may be found at http:// www.gnu.org/licenses/lgpl.html. The materials are provided free of charge by Informatica, "as-is", without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability and fitness for a particular purpose.

The product includes ACE(TM) and TAO(TM) software copyrighted by Douglas C. Schmidt and his research group at Washington University, University of California, Irvine, and Vanderbilt University, Copyright (©) 1993-2006, all rights reserved.

This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (copyright The OpenSSL Project. All Rights Reserved) and redistribution of this software is subject to terms available at http://www.openssl.org and http://www.openssl.org/source/license.html.

This product includes Curl software which is Copyright 1996-2013, Daniel Stenberg, <daniel@haxx.se>. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http://curl.haxx.se/docs/copyright.html. Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

The product includes software copyright 2001-2005 (©) MetaStuff, Ltd. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http://www.dom4j.org/ license.html.

The product includes software copyright © 2004-2007, The Dojo Foundation. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http://dojotoolkit.org/license.

This product includes ICU software which is copyright International Business Machines Corporation and others. All rights reserved. Permissions and limitations regarding this software are subject to terms available at http://source.icu-project.org/repos/icu/icu/trunk/license.html.

This product includes software copyright © 1996-2006 Per Bothner. All rights reserved. Your right to use such materials is set forth in the license which may be found at http:// www.gnu.org/software/ kawa/Software-License.html.

This product includes OSSP UUID software which is Copyright © 2002 Ralf S. Engelschall, Copyright © 2002 The OSSP Project Copyright © 2002 Cable & Wireless Deutschland. Permissions and limitations regarding this software are subject to terms available at http://www.opensource.org/licenses/mit-license.php.

This product includes software developed by Boost (http://www.boost.org/) or under the Boost software license. Permissions and limitations regarding this software are subject to terms available at http:/ /www.boost.org/LICENSE_1_0.txt.

This product includes software copyright © 1997-2007 University of Cambridge. Permissions and limitations regarding this software are subject to terms available at http:// www.pcre.org/license.txt.

This product includes software copyright © 2007 The Eclipse Foundation. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http:// www.eclipse.org/org/documents/epl-v10.php and at http://www.eclipse.org/org/documents/edl-v10.php.

This product includes software licensed under the terms at http://www.tcl.tk/software/tcltk/license.html, http://www.bosrup.com/web/overlib/?License, http:// www.stlport.org/doc/ license.html, http://asm.ow2.org/license.html, http://www.cryptix.org/LICENSE.TXT, http://hsqldb.org/web/hsqlLicense.html, http:// httpunit.sourceforge.net/doc/ license.html, http://jung.sourceforge.net/license.txt , http://www.gzip.org/zlib/zlib_license.html, http://www.openldap.org/software/ release/license.html, http://www.libssh2.org, http://slf4j.org/license.html, http://www.sente.ch/software/OpenSourceLicense.html, http://fusesource.com/downloads/ license-agreements/fuse-message-broker-v-5-3- license-agreement; http://antlr.org/license.html; http://aopalliance.sourceforge.net/; http://www.bouncycastle.org/ licence.html; http://www.jgraph.com/jgraphdownload.html; http://www.jcraft.com/jsch/LICENSE.txt; http://jotm.objectweb.org/bsd_license.html; . http://www.w3.org/

Consortium/Legal/2002/copyright-software-20021231; http://www.slf4j.org/license.html; http://nanoxml.sourceforge.net/orig/copyright.html; http://www.json.org/license.html; http://forge.ow2.org/projects/javaservice/, http://www.postgresql.org/about/licence.html, http://www.sqlite.org/copyright.html, http://www.tcl.tk/software/tcltk/license.html, http://www.jaxen.org/faq.html, http://www.jdom.org/docs/faq.html, http://www.slf4j.org/license.html; http://www.iodbc.org/dataspace/iodbc/wiki/iODBC/License; http://www.keplerproject.org/md5/license.html; http://www.toedter.com/en/jcalendar/license.html; http://www.edankert.com/bounce/index.html; http://www.net-snmp.org/about/license.html; http://www.openmdx.org/#FAQ; http://www.php.net/license/3_01.txt; http://srp.stanford.edu/license.txt; http://www.schneier.com/blowfish.html; http://www.jmock.org/license.html; http://xsom.java.net; http://benalman.com/about/license/; https://github.com/CreateJS/EaselJS/blob/master/src/easeljs/display/Bitmap.js; http://www.h2database.com/html/license.html#summary; http://jsoncpp.sourceforge.net/LICENSE; http://jdbc.postgresql.org/license.html; http://protobuf.googlecode.com/svn/trunk/src/google/protobuf/descriptor.proto; https://github.com/rantav/hector/blob/master/LICENSE; http://web.mit.edu/Kerberos/krb5-current/doc/mitK5license.html; http://jibx.sourceforge.net/jibx-license.html; https://github.com/lyokato/libgeohash/blob/master/LICENSE; https://github.com/hjiang/jsonxx/blob/master/LICENSE; https://code.google.com/p/lz4/; https://github.com/jedisct1/libsodium/blob/master/LICENSE; http://one-jar.sourceforge.net/index.php?page=documents&file=license; https://github.com/EsotericSoftware/kryo/blob/master/license.txt; http://www.scala-lang.org/license.html; https://github.com/tinkerpop/blueprints/blob/master/LICENSE.txt; http://gee.cs.oswego.edu/dl/classes/EDU/oswego/cs/dl/util/concurrent/intro.html; https://aws.amazon.com/asl/; https://github.com/twbs/bootstrap/blob/master/LICENSE; https://sourceforge.net/p/xmlunit/code/HEAD/tree/trunk/LICENSE.txt; https://github.com/documentcloud/underscore-contrib/blob/master/LICENSE, and https://github.com/apache/hbase/blob/master/LICENSE.txt.

This product includes software licensed under the Academic Free License (http://www.opensource.org/licenses/afl-3.0.php), the Common Development and Distribution License (http://www.opensource.org/licenses/cddl1.php) the Common Public License (http://www.opensource.org/licenses/cpl1.0.php), the Sun Binary Code License Agreement Supplemental License Terms, the BSD License (http:// www.opensource.org/licenses/bsd-license.php), the new BSD License (http://opensource.org/licenses/BSD-3-Clause), the MIT License (http://www.opensource.org/licenses/mit-license.php), the Artistic License (http://www.opensource.org/licenses/artistic-license-1.0) and the Initial Developer's Public License Version 1.0 (http://www.firebirdsql.org/en/initial-developer-s-public-license-version-1-0/).

This product includes software copyright © 2003-2006 Joe Walnes, 2006-2007 XStream Committers. All rights reserved. Permissions and limitations regarding this software are subject to terms available at http://xstream.codehaus.org/license.html. This product includes software developed by the Indiana University Extreme! Lab. For further information please visit http://www.extreme.indiana.edu/.

This product includes software Copyright (c) 2013 Frank Balluffi and Markus Moeller. All rights reserved. Permissions and limitations regarding this software are subject to terms of the MIT license.

See patents at https://www.informatica.com/legal/patents.html.

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

NOTICES

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

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

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

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

Publication Date: 2019-09-03

Table of Contents

Preface . . . 7

Informatica Resources. . . . 7

Informatica Network. . . . 7

Informatica Knowledge Base. . . . 7

Informatica Documentation. . . . 7

Informatica Product Availability Matrixes. . . . 8

Informatica Velocity. . . . 8

Informatica Marketplace. . . . 8

Informatica Global Customer Support. . . . 8

Chapter 1: Introduction to Intelligent Data Lake. . . . 9

Intelligent Data Lake Overview . . . 9

Enterprise Data Lake Concepts. . . . 10

Data Lake. . . . 10

Data Asset. . . . 10

Project. . . . 10

Data Publication. . . . 11

Data Visualization. . . . 11

User Interface. . . . 11

Home View. . . . 12

Search Results View. . . . 12

My Activities View. . . . 13

Projects View. . . . 14

Data Discovery and Analysis Process. . . . 15

Chapter 2: Search. . . . 17

Overview. . . . 17

Search Strings. . . . 18

Search Results. . . . 18

Sort Search Results. . . . 19

Search Filters. . . . 19

Adding a Filter Option. . . . 20

Viewing All Filter Options. . . . 20

Editing Data Asset Properties to Assign Custom Attributes. . . . 20

Chapter 3: Discover Data. . . . 21

Data Discovery Overview. . . . 21

Data Asset Views. . . . 21

Asset Details View. . . . 22

Data Preview View. . . . 24

Lineage View. 25
Relationship View. 26
Copying a Data Asset. 27
Deleting a Data Asset. 28
Access to Data. 28

Chapter 4: Import and Export Data. 29

Overview. 29
Importing a Data Asset 29
Exporting a Data Asset. 30
Downloading a Data Asset. 31

Chapter 5: Manage Projects. 32

Overview. 32
Worksheets. 33
Creating a Project. 33
Adding a Data Asset to a Project. 34
Recommendations. 34
Editing a Project. 34
Sharing a Project. 35
Changing the Project Owner. 35
Locked Projects. 36
Deleting a Project. 36

Chapter 6: Prepare Data. 37

Overview. 37
Configuring Sampling Settings. 38
Configuring Sampling Settings For Delimited Files. 39
Editing Sampling Criteria. 40
Limitations and Constraints. 40
Suggestions and Previews. 41
Recipes and General Features. 42
Ingredients. 44
Steps. 44
Blending Data. 44
Joining Worksheets. 45
Performing a Lookup on Worksheets. 45
Data Aggregation. 46
Using Formulas. 47
Using Rules. 47
Deleting a Worksheet. 48

Chapter 7: Publish Data. 49
Data Publication Overview. 49
Publishing Prepared Data. 49
Operationalize Mappings. 50
Chapter 8: Visualize and Assess Published Data. 51
Overview. 51
Prerequisites. 52
Visualizing Published Data. 52
Sharing the Notebook. 53
Smart Suggestions. 53
Chapter 9: Upload Data. 54
Data Upload Overview. 54
Uploading Data. 54
Appendix A: Glossary. 57
Index. 59

Preface

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

Informatica Resources

Informatica Network

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

As a member, you can:

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

Informatica Knowledge Base

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

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

Informatica Documentation

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

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

Informatica Product Availability Matrixes

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

Informatica Velocity

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

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

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

Informatica Marketplace

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

Informatica Global Customer Support

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

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

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

CHAPTER 1

Introduction to Intelligent Data Lake

This chapter includes the following topics:

- Intelligent Data Lake Overview , 9
- Enterprise Data Lake Concepts, 10
- User Interface, 11
- Data Discovery and Analysis Process, 15

Intelligent Data Lake Overview

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

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

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

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

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

Intelligent Data Lake provides the following benefits:

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

Enterprise Data Lake Concepts

To successfully use Enterprise Data Lake, you must understand the key Enterprise Data Lake concepts.

Data Lake

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

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

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

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

Data Asset

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

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

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

Project

A project is a container that stores data assets and worksheets. When you add a data asset to a project, Intelligent Data Lake uses the data asset as an input source and creates a corresponding worksheet that

contains a sample of the data. When you publish the worksheet as a data asset to the data lake, the project displays the publication for the worksheet.

Data Publication

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

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

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

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 of the project can access the notebook of the worksheet. You can share the notebook so that all users with adequate permissions can view the visualization of the worksheet.

User Interface

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

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

Home View

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

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

- View and create projects. The **Recent Projects** panel displays projects with the latest projects on top. To see a list of projects, click the **Switch to List View** icon . If you want to see all projects, click **All Projects**. To create a new project, click **Create New**.
- View recent 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.

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

Search Results View

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

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

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

My Activities View

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

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

- Click the name of a data asset to view its details. The data asset opens in the **Open Asset** view.
- Upload a data asset. Click **Upload Data** 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.
  - 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 the Chapter 7, “Publish Data” on page 49 chapter.

- 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.

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

Projects View

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

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

- Create a project. Click **Create New** to create a project.
- Switch the views on a project. By default, the Intelligent Data Lake application displays each project in a card view. Click the **Switch to List View** icon ( ) to switch to a list view. To return to card view, click the **Switch to Card View** icon ( ).
- Search for a project. Use the search box to search for a project.
- View all projects. To view all projects, click **All Pojects**.
- Click the name of a project to view details about the project. The project opens in a different view. You can perform the following tasks in this view:
  - Edit the name and description of the project. Click the edit icon ( ) in the **Overview** panel to edit these project properties.
  - Change the collaborators on the project. Click the add icon ( ) in the **Collaborators** panel to change the users who can collaborate on the project.
  - View or edit a worksheet in the project. Each project displays a worksheet for each data asset in the project. In the **Worksheets** panel, click **Prepare** and choose to **View** a worksheet or open it for editing.

    You can also select a worksheet in a project and click the Publish icon ( ) to publish the worksheet. 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.

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

Data Discovery and Analysis Process

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

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

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

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

1. Search for data assets.

   If you find data outside the data lake that you want to work with, ask your administrator to add the data to the data lake.

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

   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 GUID-36EB7523-E653-48CE-894E-96B20DF219F1.

CHAPTER 2

Search

This chapter includes the following topics:

- Overview, 17
- Search Strings, 18
- Search Results, 18
- Sort Search Results, 19
- Search Filters, 19

Overview

You can search for data assets in Intelligent Data Lake. Data assets can include items such as a flat file, table, or view.

You can search for and discover any data asset described in the catalog. You can search for, discover, and prepare data assets that are stored in the data lake.

You can search for a data asset by name, use a wildcard character in your search, and Intelligent 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. Intelligent 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, Intelligent 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. Intelligent 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`. Intelligent 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, Intelligent Data Lake suggests matching data asset names.

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

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

Search Results

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

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.

The following image shows the search results:

Sort Search Results

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

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

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

Search Filters

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

You can use the following search filters:

- **Asset Type**. The type of data asset, such as Table, Resource, or view.
- **Resource Type**. The resource type from where the data asset details were collected.
- **Last Updated**. The last time the details of the data asset were updated in the catalog. You can select from options such as Today, Yesterday, Last 7 days, and so forth.

- **Number of Rows**. The number of rows in the data asset.
- **Size**. The size of the data asset.
- **Asset Created By**. The user who created the data asset.
- **Asset Modified By**. The users who modified the data asset.
- **Asset Used By**. The users who used the data asset.
- The system attributes or custom attributes.

Adding a Filter Option

By default, Intelligent Data Lake does not list all the option under each filter. You can add an option that doesn't appear. You can then use this option to refine your search.

1. Click **Add**.

   A box appears where you can type the name of the option to be added.
2. Start typing the name of option in the box.

   A list of matching options appears.
3. Select the options that you want.

   Intelligent Data Lake refreshes the search results based on the new options selected.

Viewing All Filter Options

You can view all the options that are not listed under each filter. You can then add the required options to the list and refine your search.

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

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

Editing Data Asset Properties to Assign Custom Attributes

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

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

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

   The list of matching attributes appears.
4. Select the required attributes from the **Name** section.
5. Click **OK**.

CHAPTER 3

Discover Data

This chapter includes the following topics:

- Data Discovery Overview, 21
- Data Asset Views, 21
- Copying a Data Asset, 27
- Deleting a Data Asset, 28
- Access to Data, 28

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 discover all data assets in the catalog. You can prepare data assets that are stored in the data lake. To prepare data assets that are not stored in the lake, you must import them into the lake.

You can also download your data asset as a CSV file or Tableau TDE file for further analysis.

Data Asset Views

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

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

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

Asset Details View

The **Asset Details** view displays details of the selected data asset.

The following image shows the **Asset Details** view:

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

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

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

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

Data Asset	Description

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

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

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, Intelligent 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.

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.

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

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

Data Preview of HDFS Files

You can also preview the data in assets such as HDFS files.

You can view data from HDFS files in two ways:

- Tabular View: Similar to the Hive tables, the data is shown in a tabular format.

- Raw Preview: In addition to the data in the file, this preview also shows other additional content such as comments. This preview shows the file in its original format.

If the HDFS file is a delimited file, the tabular view is the default preview of the data asset. You can click **Raw Preview** to view the raw data in the data asset.

Lineage View

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

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

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

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

Expanded Lineage View

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

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

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

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.

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

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

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

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

Copying a Data Asset

You can copy a data asset stored in the 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, Intelligent 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 and Export Data

This chapter includes the following topics:

- Overview, 29
- Importing a Data Asset , 29
- Exporting a Data Asset, 30
- Downloading a Data Asset, 31

Overview

You can import data from tables in external sources (outside the Hadoop data lake) such as Oracle and Teradata into the data lake if these sources are already cataloged. You can also export a data asset or a publication to external targets (outside the Hadoop data lake) such as Oracle and Teradata.

The administrator needs to configure JDBC connections with Sqoop to the external sources and targets and provide access to the analyst. You can use these connection to preview the data, import the data from external sources, and export the data to external targets.

The Intelligent Data Lake application remembers your selected connections for future reuse for preview, import, or export purposes during a browser session.

During import or export, if the target already exists, you can append to it or overwrite on it. A new import activity or export activity is created. You can track these activities on My Activities page.

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.

Once the export activity is completed successfully, you can verify the exported data on the external target system. You cannot verify the exported data in the Intelligent Data Lake application if the table is not cataloged.

Importing a Data Asset

You can import a data asset from external sources into the data lake such as Oracle and Teradata databases.

1. On any data asset view, click **Import to Lake**.

   The **Import** dialog box appears.

2. If the asset has duplicates in the lake, a warning message appears. Click **Import Anyway** if you'd like to import the asset.
3. Select a connection for the asset from the specified resource. Click **Next**.

   Note: Contact your administrator if you do not see any connection in the list. You need to have access privileges to a connection and **Read** and **Execute** permissions for each connection.
4. Enter the target location details to import the asset using the connection.
   a. Select a schema from the drop down list.
   b. Enter the name of the table. 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.
   c. If the table already exists, select Append or Overwrite.
      - Append. Select this to append the data to an existing table. Make sure the column names are same, the data types of the columns are same, precision and scale of columns is same, and sequence of columns is same for the importing table and the existing table. The import will fail if any of these do not match.
      - Overwrite. Select this to overwrite the existing table. It drops the table and re-creates the table with the data from the importing table.
5. Click **Import** to import the table into the data lake.

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

   For more information about uploading data assets from local folder, see the *Upload Data* chapter.

Exporting a Data Asset

You can export a data asset or a publication to external targets such as Oracle and Teradata databases.

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

   The **Export** dialog box appears.
2. Select a connection for the asset. Click **Next**.

   Note: Contact your administrator if you do not see any connection in the list. You need to have access privileges to a connection and **Read** and **Execute** permissions for each connection.
3. Enter the target location details to export the asset using the selected connection.
   a. Enter the name of the schema.
   b. Enter the name of the table. The name can include letters, numbers, and the underscore character. The first character of the name cannot be an underscore or a number.
4. Click **Export**.
5. If a table does not exist with the same name, a new table is created. If a table exists with the name already, the table schema will be validated and the following options appear:
   - If the source table has the same schema structure as the target table, the following options appear:
     - **Append**. Select this option to append the data to an existing table.

- **Overwrite**. Select this option to overwrite the existing table. It drops the table and re-creates the table with the data from the exporting table.

- If the source table has a different schema structure, select to overwrite the table. This replaces the existing table with the source data.

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

Downloading a Data Asset

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

1. Click the name of the data asset under the **Input Sources** column.
2. On the Data Assets page, select the (⋮) icon and click **Download TDE** or **Download csv**.

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

CHAPTER 5

Manage Projects

This chapter includes the following topics:

- Overview, 32
- Worksheets, 33
- Creating a Project, 33
- Adding a Data Asset to a Project, 34
- Recommendations, 34
- Editing a Project, 34
- Sharing a Project, 35
- Changing the Project Owner, 35
- Locked Projects, 36
- Deleting a Project, 36

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.

Intelligent 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 My Projects view. The projects in the **Projects Overview** section appear as cards by default. The project card shows the following details:

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

- Worksheets. Displays the number of worksheets in this project.
- Publications. Displays the number of publications in this project.

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

Worksheets

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

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

- View a worksheet. Select the worksheet and click **Prepare** > **View**.
- Edit a worksheet. Select a worksheet and click **Prepare** > **Prepare**.

The worksheet opens in a spreadsheet interface.

The worksheets list displays the following details about each worksheet:

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

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. Intelligent 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

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

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

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

Editing a Project

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

To edit a project, you must have edit permissions.

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

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

Sharing a Project

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

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

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

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

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

Changing the Project Owner

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

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

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

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 Intelligent 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 Project

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

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

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

CHAPTER 6

Prepare Data

This chapter includes the following topics:

- Overview, 37
- Configuring Sampling Settings, 38
- Configuring Sampling Settings For Delimited Files, 39
- Editing Sampling Criteria, 40
- Suggestions and Previews, 41
- Recipes and General Features, 42
- Blending Data, 44
- Data Aggregation, 46
- Using Formulas, 47
- Using Rules, 47
- Deleting a Worksheet, 48

Overview

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

When you add a data asset to a project, Intelligent Data Lake creates a corresponding worksheet in the project. A worksheet has an interactive data-driven spreadsheet like interface. The worksheet contains the data and a recipe that tracks the changes you make to the data as you prepare it. Intelligent Data Lake loads sample data into the worksheet depending on 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, Intelligent Data Lake applies the recipe to the data in the input source and creates a new data asset.

Configuring Sampling Settings

When you are working with large amount of data, you must use a sample of the data to fit your data preparation needs. You can configure the sampling settings before preparing your data asset. You use this sample data to perform different data preparation operations.

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

   The Sampling Preview panel appears along with the worksheet.
2. In the **Columns** tab, click **Edit Selection**. From the list of columns, select the columns that you want to work with. If you want to search for a column, enter the column names in the Search text box. You can search for multiple columns by entering comma-separated lists.
3. Click the **Filters** tab.
   a. Click **New Filter**.
   b. In the **Create filter with** text box, enter the name of the column that you want to use for filtering.
   c. From the select drop down list, click one of the following options:
      - **Equals**. Select this 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. Only the first value is considered and the other entries are ignored.
      - **Contains**. Select this to filter and display data containing the entered value.
      - **Starts with**. Select this to filter and display rows with the column value that start with the entered value.
      - **Ends with**. Select this to filter and display rows with the column value that end with the entered value.
   d. Enter a value for filtering the column.
   e. Click **Save**.

   Note: Filters are not available when you try to configure sampling for worksheets that are HDFS files.
4. Click the **Sampling** tab.
5. Enter the number of rows to sample. The administrator can set the maximum number of rows in the Informatica Administrator tool.
6. Select the Sampling type:
   - **Random Rows**. Select this 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 this to view the rows from beginning based on the sampling criteria.

   Note: Random sampling is not available when you try to configure sampling for worksheets that are HDFS files.
7. Click **Apply Selection** if you made any changes in any of the tabs. Click **Edit Selection** to make any further changes.
8. Click **Done**.

   During publication, you can choose to apply these sampling filters (for column and row filters). When you publish, the recipe will be executed on the selected columns and rows according to the settings and not on the full data asset.

Example

For example, Sarah, a data analyst, wants to prepare a data asset that contains an email blast list of all customers residing in California using a particular email domain. Intelligent Data Lake displays a sample of the data in this data asset. Using the Sampling Settings, Sarah can choose to determine the sample that she

wants to work with. She can choose to include the name, telephone number, email address, state, and postal code details. She can choose to not select the IP address details. Since Sarah is only interested in customers residing in California, she can use the Filters tab to filter the state column to equal to California. Later, Sarah can enter the number of rows to sample and where to begin the sampling using the Sampling tab. Sarah can also choose to select random rows from the entire data asset.

Using the sample data, Sarah can later perform other operations such as finding the domain names of all customers in California. She can perform further operations to find out how many of them use a particular domain for email communication.

Configuring Sampling Settings For Delimited Files

You can also configure the sampling settings of the flat files you add to your project before preparing them. You use this sample data to perform different data preparation operations.

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

   The Sampling Preview panel appears along with the worksheets.
2. Enter the format details of the flat file.
   a. Start at row. Enter the row number from which you want to start the sampling.
   b. Column Names. Select **Get field names from first row** if the first row contains column names.
3. Enter the columns details of the flat file. In the **Columns** tab, click **Edit Selection**. From the list of columns, select the columns that you want to work with. If you want to search for a column, enter the column names in the Search text box.
4. Enter the sampling details.
   a. Click the **Sampling** tab.
   b. Enter the number of rows to sample. The administrator can set the maximum number of rows in the Informatica Administrator tool.
5. Click **Apply Selection** if you made any changes. Click **Edit Selection** to make any further changes.
6. Click **Done**.

   During publication, you can choose to apply these source filters (for column and row filters). When you publish, the recipe will be executed on the selected columns and rows according to the settings and not on the full data asset.

Example

For example, Sarah, a data analyst, wants to work on a flat file that contains an email blast list of all customers residing in California using a particular email domain. She can add this flat file to a project. Intelligent Data Lake displays a sample of the data in this data asset. Using the Sampling Settings for Flat Files, Sarah can choose to determine the sample that she wants to work with. She can choose to include the name, telephone number, email address, state, and postal code details. She can choose to not select the IP address details. Since Sarah is only interested in customers residing in California, she can use the Filters tab to filter the state column to equal to California. Later, Sarah can enter the number of rows to sample and where to begin the sampling using the Sampling tab. Sarah can also choose to select random rows from the entire data asset.

Using the sample data, Sarah can later perform other operations such as finding the domain names of all customers in California. She can perform further operations to find out how many of them use a particular domain for email communication.

Editing Sampling Criteria

When you are working with large amount of data, you might feel the need to resample the data. You can edit the sampling criteria while preparing your data asset.

You can edit the sampling settings in the following ways:

- You can change the columns selected.
- You can edit the filters selected by adding or removing them.
- You can change the sampling criteria.

After editing the sampling criteria, all the recipe steps will be automatically executed on the re-sampled data.

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

   The Recipe panel appears along with the worksheets.
2. In the Recipe panel, scroll to the Import step. This steps appears first in the recipe. When you hover on this step, a pencil icon appears. Click the pencil icon to edit the sampling criteria.

   The Sampling Preview panel replaces the Recipe and Ingredients panel.
3. In the **Columns** tab, from the list of columns, select the columns that you want to appear. If you want to search for a column, enter the column names in the Search text box. Select or deselect the columns to suit your sampling needs.

   Note: You cannot deselect columns that have dependencies. For example, if one of the recipe steps depending on this column, it will be disabled at this stage.
4. Click the **Filters** tab. Edit existing filters or create new filters.
5. Click the **Sampling** tab. Edit the existing sampling criteria.
6. Click **Apply Selection** if you made any changes. Click **Edit Selection** to make any further changes.
7. Click **Done**.

   During publication, you can choose to apply these source filters (for column and row filters). When you publish, the recipe will be executed on the selected columns and rows according to the settings and not on the full data asset.

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.

Suggestions and Previews

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

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

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

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

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

Recipes and General Features

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

The following image shows the worksheet interface:

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

Edit the worksheet and its columns.

Use the edit menu ( ) to edit the columns in the worksheet. You can choose to delete blank columns or columns with one value, or hide or unhide columns in the worksheet. You can also right-click the 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.

Combine worksheets.

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

Aggregate the data in the worksheet.

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

Filter the data in the worksheet.

Use the filter icon ( ) to filter the data on any column in the worksheet. You can choose to add a filter or modify a filter condition. You can also clear all filters. When you filter your data, you select a column

to filter and choose to search for a value in a column or create a custom filter and select your filter conditions. Different filters are available for String, Date, and Number columns depending on the column selected. You can filter using the top menu or you can use the filter panel available at the bottom or the window.

Apply or create formulas for data in the worksheet.

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

Refresh the data in the worksheet.

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

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

View your data preparation steps.

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

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

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

Sort the data.

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

Close the worksheet.

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

Ingredients

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

The following image shows an ingredient in a recipe:

Steps

Each action you take in a worksheet is added to the recipe as a step. You can add, edit, and delete recipe steps.

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

Select **Edit** from the menu within a step to make changes.

Select **Insert Step** from the menu in a step to insert a new step below. You can insert a step below any step except for the last step in the recipe. You cannot perform some operations when you insert a step, including aggregate, join, or union operations.

Click the **Delete** icon to remove a step.

You can also go back-in-time to a specific step in the recipe and see the state of data. You can refresh the recipe from the source.

Blending Data

You can create blended data sets by combining, importing, or merging data in worksheets within a project. Intelligent 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 perform operations such as inner join, left join, and right join on the worksheets. Intelligent Data Lake joins data from two different worksheets, based on a related column between them.

1. On the project page, click **Prepare**.
2. Click the blending icon and select **Join worksheets**.
3. In the Join window that appears, select the second worksheet that you'd like to join with the first worksheet.
4. Click **OK**. The Confirm key to join work panel appears.
5. If needed, edit to change the name of the newly joined worksheet.
6. Review the suggested selected key. You can view the approximate overlap percentage and results.
7. Click **Change Key** to select another join key if needed.
8. From the Results displayed, select the type of join:
   - Use inner join to join the rows matching in both the worksheets.
   - Use left join to join all of the rows from the first worksheet and the matching columns from the second worksheet.
   - Use right join to join all of the rows from the second worksheet and the matching columns from the first worksheet.
9. Click **Join**. To cancel the new join, click **Cancel**.

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 Intelligent 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, left join, or right join.
7. Click **Done**. To undo the editing of the join, click **Cancel**.

Performing a Lookup on Worksheets

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

1. Select a lookup worksheet from the list of worksheets. All the worksheets in the current project appear in the list.
2. Select a key column to perform lookup. Intelligent Data Lake displays the key columns with the highest amount of overlap and recommends them as key columns for lookup.
   - Select the recommended column.

- Or, click **Select your own keys** to view the list of all columns in both the data assets. Click a corresponding column name in both the assets. Click **Back to suggested keys...** to view the suggested key columns.

3. Select the columns to add from the lookup sheet to your current worksheet.
4. Click **Done**.

   Currently, you cannot perform a lookup on derived columns, derived sheets, or sheets with filter options applied.

For example, you might want to prepare an email blast list of all customers in California. The email blast list data asset does not contain contact information. You want to add the pin code details of each customer, which is available in the customer contacts list data asset. You can perform a lookup on the email blast list and select the contacts list as the second worksheet. You can then use the customer ID as the key column, and choose to add the postal code column to the email blast list data asset.

General Limitations for Lookup

Note the following limitations for performing lookup.

- Duplicate keys. When there are duplicate keys in the lookup sheet, you cannot perform lookup.
- Derived worksheets. You can perform lookup only on existing worksheets. You cannot perform lookup on new worksheets such as sheets derived from other worksheets using operations such as union, aggregate, or join. However, you can perform lookup on copied worksheets.
- Filtered worksheets. You cannot perform lookup on worksheets that have filtered columns.
- New columns. You can perform lookup only on existing columns. You cannot perform lookup on new columns such as columns derived after performing operations on existing columns.
- Existing columns. You cannot perform lookup on columns on which any operation has been performed.
- Source Filters. You cannot perform lookup on worksheets where source filters are enabled.
- Data type mismatch. You cannot perform lookup on two columns with different data types such as decimal and double data types.

Data Aggregation

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

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

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

she notices that the aggregation has twice as many rows because each geographic region now has a male row and a female row.

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

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

Using Formulas

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

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

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

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

Using Rules

You can apply rules to data in a worksheet to help you cleanse, transform, or validate the data.

Use rules to maintain a consistent standard for preparing a given data type or data set within or across organizations. Rules are available in a library that you can access during data preparation. Use the pre-built rules created in Informatica Developer or Informatica Analyst or import existing rules using Informatica Developer and run them to address your project objectives.

You can create rules using Informatica Developer or Informatica Analyst, or import existing rules using Informatica Developer.

To apply a rule, click the rule icon and select **Apply Rule**. Browse the rules library or search for the rule and select the rule that you want to apply. Enter the details of the input columns based on the rule requirements. Choose to select a few or all of the output columns. Click **Done** to apply the rule.

If you make a mistake, you can always go back to rectify it. Click the recipe icon on the left side, find your rule, edit the rule to modify it, or remove it from the recipe.

For example, you might want to validate postal addresses of all customers within a geographical location. You can use the address validation rules for validation. Using these rules, you can validate the accuracy of

the postal address data of each customer. Address validation corrects errors in addresses and completes partial addresses by comparing address records against reference data from national postal carriers. Address validation can also add postal information that speeds mail delivery and reduces mail costs.

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**.

CHAPTER 7

Publish Data

This chapter includes the following topics:

- Data Publication Overview, 49
- Publishing Prepared Data, 49
- Operationalize Mappings, 50

Data Publication Overview

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

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

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

Publishing Prepared Data

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

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

   The **Publish** dialog appears.
3. Use one of the following options to publish the data asset:
   - New: Selected by default. Use this option if the table is completely new for the selected schema.
   - Append: Use this option if the entered table name is same as the existing table name in the schema and if the schema of this table is same as the existing table schema.
   - Overwrite: Use this option if the entered table name is same as the existing table name in the schema. It drops the table and re-creates the table with the data from the published table.
4. Select the Hive schema in the data lake in which to publish the data.

5. Enter a name for the Hive table.
6. Select the **Apply source filter on publish** if you want to apply the source filters defined while configuring sampling settings while publishing the data asset.
7. Click **OK**.

Operationalize Mappings

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

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

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

CHAPTER 8

Visualize and Assess Published Data

This chapter includes the following topics:

- Overview, 51
- Prerequisites, 52
- Visualizing Published Data, 52
- Sharing the Notebook, 53
- Smart Suggestions, 53

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.

Intelligent 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, Intelligent Data Lake creates a Zeppelin notebook for the data asset. You can continue to work on the notebook and share it with other Intelligent 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:

Prerequisites

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

- Make sure that Zeppelin is configured by the Intelligent 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 Intelligent 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. You can write ad-hoc queries that run on the entire data asset.

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

Upload Data

This chapter includes the following topics:

- Data Upload Overview, 54
- Uploading Data, 54

Data Upload Overview

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

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

Uploading Data

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

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

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

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

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

The following table describes the file properties that you specify:

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

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

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

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

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

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

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

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

9. Click **Upload**.

APPENDIX A

Glossary

asset

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

catalog

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

data asset

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

data lake

A centralized repository of large volumes of structured and unstructured data. A data lake can contain different types of data, including raw data, refined data, master data, transactional data, log file data, and machine data.

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

data preparation

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

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

data publication

The process of making prepared data available in the data lake.

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

data upload

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

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

input source

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

project

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

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

recipe

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

When you publish prepared data, Intelligent Data Lake applies the recipe to the data in the input source. Intelligent Data Lake converts the recipe into an Informatica mapping and stores the mapping in the Model repository.

worksheet

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

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

INDEX

A

asset views
- data preview view 24
- lineage view 25
- overview view 22
- relationship view 26

assets
- associating a business term 23

D

data asset
- copying 27
- deleting 28

data discovery
- access to data 28
- asset views 21
- copying a data asset 27
- deleting a data asset 28

data preparation
- data aggregation 46
- data blending 44
- formulas 47
- ingredients 44
- recipes 42
- rules 47
- steps 44
- suggestions, previews 41

data publication
- operationalize mappings 50
- publishing 31, 49

data upload
- uploading data 54

data visualization
- visualizing 52

I

Intelligent Data Lake application
- data discovery 21
- data preparation 37
- data publication 49
- data upload 54
- projects 32
- search 17
- views 12–14

P

Project view
- worksheet 33

projects
- changing project owner 35

projects *(continued)*
- creating 33, 34
- deleting 36, 48
- locking 36
- recommendations 34
- sharing 35
- worksheet 33

R

recipes
- ingredients 44
- steps 44

S

search
- search filters 19
- search results 18, 19
- search strings 18

search by custom attributes
- assigning custom attributes 20

search filters
- editing data asset properties 20

search strings
- search suggestions 18
- wildcard search 18

V

views
- Home view 12
- My Activities view 13
- Projects view 14
- Search view 12

visualizing data 51

W

worksheet
- data aggregation 46
- data blending 44
- formulas 47
- recipes 42
- rules 47
- suggestions, previews 41