SAS® Data Loader 2.3 for Hadoop

User's Guide

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2015. *SAS® Data Loader 2.3 for Hadoop: User's Guide*. Cary, NC: SAS Institute Inc.

SAS® Data Loader 2.3 for Hadoop: User's Guide

Copyright © 2015, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication. The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

NOTICE: This documentation contains information that is proprietary and confidential to SAS Institute Inc. It is provided to you on the condition that you agree not to reveal its contents to any person or entity except employees of your organization or SAS employees. This obligation of confidentiality shall apply until such time as the company makes the documentation available to the general public, if ever.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202–1(a), DFAR 227.7202–3(a) and DFAR 227.7202–4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227–19 (DEC 2007). If FAR 52.227–19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513–2414.

Printing 1, July 2015

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Other brand and product names are trademarks of their respective companies.

With respect to CENTOS third party technology included with the vApp ("CENTOS"), CENTOS is open source software that is used with the Software and is not owned by SAS. Use, copying, distribution and modification of CENTOS is governed by the CENTOS EULA and the GNU General Public License (GPL) version 2.0. The CENTOS EULA can be found at http://mirror.centos.org/centos/6/os/x86_64/EULA. A copy of the GPL license can be found at http://www.opensource.org/licenses/gpl-2.0 or can be obtained by writing to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02110-1301 USA. The source code for CENTOS is available at **http://vault.centos.org/**.

With respect to open-vm-tools third party technology included in the vApp ("VMTOOLS"), VMTOOLS is open source software that is used with the Software and is not owned by SAS. Use, copying, distribution and modification of VMTOOLS is governed by the GNU General Public License (GPL) version 2.0. A copy of the GPL license can be found at **http://www.opensource.org/licenses/gpl-2.0** or can be obtained by writing to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02110-1301 USA. The source code for VMTOOLS is available at **http://sourceforge.net/projects/open-vm-tools/**.

Contents

What's New in SAS Data Loader 2.3 for Hadoop . . . . . *v*
Accessibility . . . . . *ix*

Chapter 1 • About SAS Data Loader for Hadoop **1**
What Is SAS Data Loader for Hadoop? 1
What Does It Help You Do? 2
How Does It Work? 3
How to Get Help for SAS Data Loader for Hadoop 5

Chapter 2 • Getting Started **7**
Prerequisites 7
First Tasks in SAS Data Loader 7
Create and Execute a Job Using SAS Sample Data 8
Naming Requirements for Schemas, Tables, and Columns 10

Chapter 3 • About the Directive Interface **13**
Using the Directives Page 13
Viewing Data Sources and Tables 14
Working with the Code Editor 18

Chapter 4 • Manage Data in Hadoop **19**
Overview of Data Management Directives 20
Browse Tables 20
Cleanse Data in Hadoop 23
Delete Rows 43
Query or Join Data in Hadoop 46
Sort and De-Duplicate Data in Hadoop 55
Transform Data in Hadoop 61
Transpose Data in Hadoop 70

Chapter 5 • Profile Data in Hadoop **73**
Overview of Profile Directives 73
Profile Data 75
Saved Profile Reports 80

Chapter 6 • Copy Data To and From Hadoop **89**
Overview of the Copy Data Directives 89
Copy Data to Hadoop 90
Import a File 102
Copy Data from Hadoop 108
Load Data to LASR 118

Chapter 7 • Run User-Written Programs in Hadoop **121**
Overview 121
Run a SAS Program 121
Run a Hive Program 123

Chapter 8 • Manage Jobs **125**
Overview of Job Management Directives 125
Run Status 126
Saved Directives 128

Chapter 9 • Maintaining SAS Data Loader **131**
Back Up Directives 131
Set Global Options 132
Troubleshooting 144

Chapter 10 • Using the vApp **151**
Overview of the vApp for SAS Data Loader 151
Play the vApp and Start SAS Data Loader 151
Tips for Running the vApp 152
Power Off the vApp 153

Recommended Reading **155**
Index **157**

What's New

What's New in SAS Data Loader 2.3 for Hadoop

Overview

The main enhancements for SAS Data Loader 2.3 include the following:

- Migration Support for SAS Data Loader 2.2
- New Support for Importing Delimited Files into Hadoop
- Enhanced Support for SAS LASR Analytic Server
- Profile Enhancements
- New Features for Data Quality Analysis
- Usability Enhancements
- Enhanced Support for Hadoop
- Enhanced Support for Apache Hive
- New Support for Active Directory (LDAP) Authentication
- Documentation Changes

Migration Support for SAS Data Loader 2.2

When you configure the vApp for SAS Data Loader 2.3, you can migrate your jobs, profiles, and configuration settings from SAS Data Loader 2.2. For more information, see the “Migrate From a Previous Version” chapter of the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

New Support for Importing Delimited Files into Hadoop

The new directive "Import a File" now imports and registers delimited source files in Hadoop. The contents of the imported file are saved to a Hive table. New tables receive columns that are derived from the source. Column definitions can be edited and stored for reuse in future imports.

Enhanced Support for SAS LASR Analytic Server

The directive "Load Data to LASR" now supports symmetric multiprocessing (SMP) with the SASIOLA engine when loading data into non-grid configurations of the SAS LASR Analytic Server software. In grid configurations, the directive continues to support massively parallel processing (MPP).

Profile Enhancements

The directives Profile Data and Saved Profile Reports have been enhanced:

- New configuration settings give you more control over how profile jobs are processed. For example, you can specify the number of parallel threads to use when processing a job. You can also choose to minimize the number of threads that are used. For more information about these settings, see "Profiles Panel" on page 142.
- When viewing data in a profile report, you can view trend graphs for the information in the report to quickly visualize how the profiled data has changed over time. For more information about using trend graphs in reports, see "Open Saved Profile Reports" on page 82.
- When you generate a profile report, SAS Data Loader now automatically performs data type analysis and includes the results in the report. The analysis detects and reports on the types of content in source columns. For example, the analysis can indicate that a column contains phone numbers or ZIP codes. For more information about profile reports, see "About Profile Reports" on page 80.

New Features for Data Quality Analysis

The directive "Cleanse Data in Hadoop" now provides the following new transformations: Change Case, Gender Analysis, Pattern Analysis, and Field Extraction.

New Support for Deleting Rows from a Selected Table

The new "Delete Rows" directive enables the deletion of specified rows within a source table, without writing the output to a target table. Rows can be deleted by one or more rules or by a user-written Hive expression. You can paste and edit existing expressions. An expression builder provides access to Hive function syntax and source column names. In order to use this feature, the Hadoop cluster must be configured with release 0.14 or later of the Apache Hive data warehouse software. This release supports transactional tables.

Usability Enhancements

You can do the following tasks, using features that are new to this release:

- display available tables and data sources in a list view or in the existing grid view.
- use a new Search field to locate tables and data sources in the list view.
- configure the list view and grid view to open the most recently used data source automatically. For more information, see " About the SAS Table Viewer" on page 16.
- select from the ten most recently accessed tables or data sources.
- delete individual jobs in the Run Status directive.
- filter rows using an expression builder or the existing rules mechanism in several directives.
- The SAS Table Viewer now displays short column names, not tablename.columnname. The short column names are easier to read.

Enhanced Support for Hadoop

New versions of Cloudera and Hortonworks are supported. Support for MapR has been added. Kerberos is not supported in combination with MapR.

If the version of Hadoop on your cluster has changed, you can now update the version of Hadoop that is specified in the vApp to match. For more information, see “Configuring a New Version of Hadoop” in the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

If the default temporary storage directory for SAS Data Loader is not appropriate for some reason, you can now change that directory. For example, if the default temporary directory is not writable, you can specify another directory. For more information, see the description of the **SAS HDFS temporary file location** field in the topic “Hadoop Configuration Panel” on page 132.

Enhanced Support for Apache Hive

The new directive Run a Hive Program enables you to paste and edit existing Hive programs, and then run those programs in Hadoop.

The Run a Hive Program directive provides a resource selector that enables you to browse and add Hive function syntax. The resource selector is also provided in the Hive expression builder in the following directives: Delete Rows, Query or Join Tables in Hadoop, and Sort and De-Duplicate Data in Hadoop.

You can now store data in the default Hive warehouse location, or you can specify an alternate location in the Hadoop file system. This location setting can be applied globally for all target tables that are generated. You can also override the global setting on a job-by-job basis for individual target tables. For more information, see “Change the File Format of Hadoop Target Tables” on page 145.

New Support for Active Directory (LDAP) Authentication

SAS Data Loader will now connect to a Hadoop cluster that is protected with Active Directory and LDAP (Lightweight Directory Access Protocol). For more information, see “Active Directory (LDAP) Authentication” on page 144.

Documentation Changes

The content that was formerly in the *SAS Data Loader for Hadoop: Administrator's Guide* is now in the “Administrator’s Guide for SAS Data Loader for Hadoop” section of the *SAS In-Database Products: Administrator's Guide*.

Accessibility

For information about the accessibility of this product, see Accessibility Features of SAS Data Loader 2.3 for Hadoop.

1

About SAS Data Loader for Hadoop

What Is SAS Data Loader for Hadoop? 1
What Does It Help You Do? 2
How Does It Work? 3
How to Get Help for SAS Data Loader for Hadoop 5
SAS Data Loader Support Community 5
Technical Support 5
Documentation and System Requirements 5

What Is SAS Data Loader for Hadoop?

SAS Data Loader for Hadoop is a software offering that makes it easier to move, cleanse, and analyze data in Hadoop. It enables business users and data scientists to do self-service data preparation on a Hadoop cluster.

Hadoop is highly efficient at storing and processing large amounts of data. However, moving, cleansing, and analyzing data in Hadoop can be labor-intensive, and these tasks usually require specialized coding skills. As a result, business users and data scientists usually depend on IT personnel to prepare large Hadoop data sets for analysis. This technical overhead makes it harder to turn Hadoop data into useful knowledge.

SAS Data Loader for Hadoop provides a set of “directives” or wizards that help business users and data scientists do the following tasks:

- copy data to and from Hadoop, using parallel, bulk data transfer
- perform data integration, data quality, and data preparation tasks within Hadoop without writing complex MapReduce code or asking for outside help
- minimize data movement for increased scalability, governance, and performance
- load data in memory to prepare it for high-performance reporting, visualization, or analytics

What Does It Help You Do?

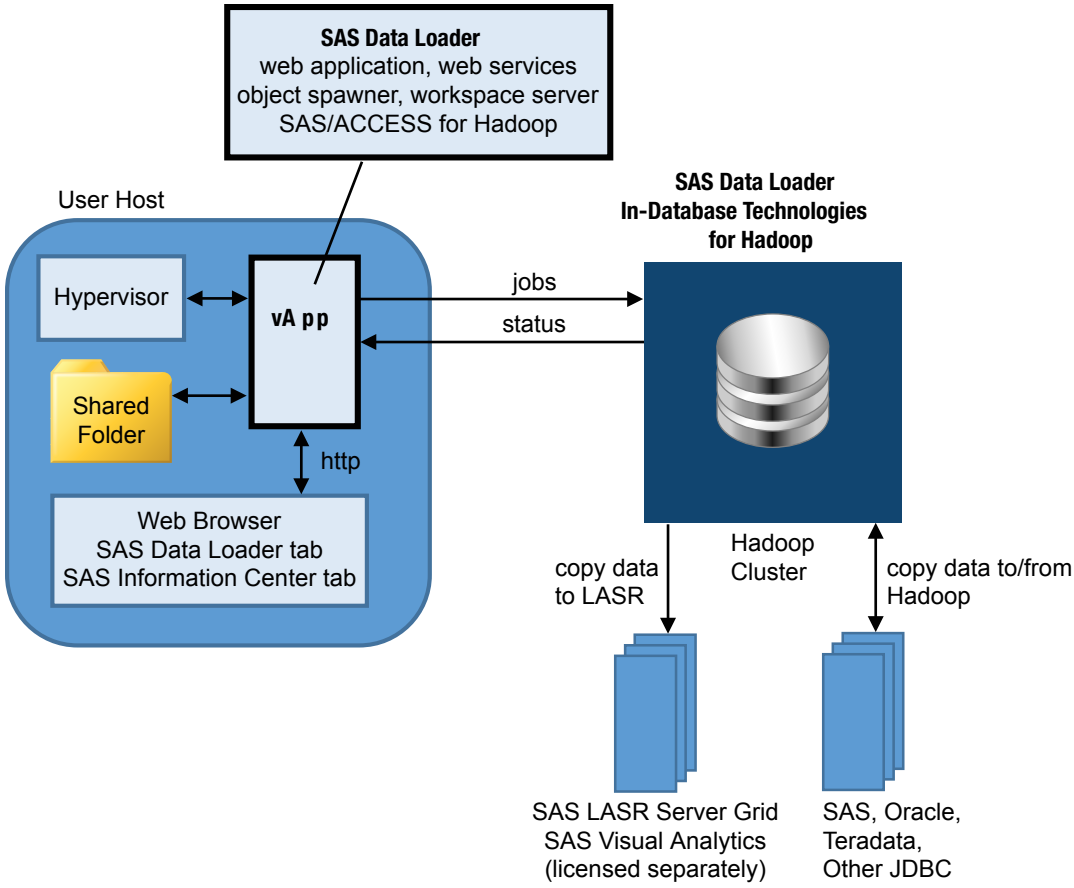
SAS Data Loader enables business users and data scientists to perform tasks such as the following:

Tasks	Description
Copy data to and from Hadoop	Copy relational databases and SAS data sets to and from Hadoop via parallel, bulk data transfer. For more information, see "Copy Data to Hadoop" on page 90, and "Copy Data from Hadoop" on page 108. Import data from delimited text files, such as comma-separated value (CSV) files. For more information, see "Import a File" on page 102.
Transform and transpose data in Hadoop	Transform data by filtering rows, managing columns, and summarizing rows. For more information, see "Transform Data in Hadoop" on page 61. Select columns and transpose or group them. For more information, see "Transpose Data in Hadoop" on page 70.
Cleanse data in Hadoop	Standardize, match, parse, and perform other data quality functions on data in Hadoop. For more information, see "Cleanse Data in Hadoop" on page 23. Use rules and expressions to filter data. For more information, see "About DS2 Expressions and the Advanced Editor" on page 42.
Sort or de-duplicate data in Hadoop	Sort data in an existing table and remove duplicate rows from the table. For more information, see "Sort and De-Duplicate Data in Hadoop" on page 55.
Query or join data in Hadoop	Query a table or join multiple tables without knowing SQL. For more information, see "Query or Join Data in Hadoop" on page 46. Run aggregations on selected columns. For more information, see "About the Aggregations in the Summarize Rows Transformation" on page 69. Power users can generate and edit a HiveQL query, or paste and run an existing HiveQL query. For more information, see "Run a Hive Program" on page 123.

Tasks	Description
Profile data and save profile reports	Analyze source columns from one or more tables to determine patterns, uniqueness, and completeness. For more information, see "Profile Data" on page 75. View data profile reports. Add notes to a data profile report to explain a result or ask a question.
Run user-written code	Use the Run a SAS Program directive to execute user-written Base SAS code or DS2 code. For more information, see "Run a SAS Program" on page 121. Use the Run a Hive Program directive to execute user-written Hive code. For more information, see "Run a Hive Program" on page 123.
Manage and reuse directives	Use directives to guide you through the process of creating and running jobs in Hadoop. View the status of current and previous job executions. For more information, see "Run Status" on page 126. Stop and start directives. Open their logs and generated code files. Run, view, or edit saved directives for reuse. For more information, see "Saved Directives" on page 128.
Load data to SAS LASR Analytic Server	Load specified Hadoop columns in memory onto the SAS LASR Analytic Server for analysis using SAS Visual Analytics or SAS Visual Statistics (licensed separately). For more information, see "Load Data to LASR" on page 118.
Specify global options	Specify server connections, data sources, global options, and other settings for SAS Data Loader. For more information, see "Set Global Options" on page 132.

How Does It Work?

SAS Data Loader for Hadoop is a software offering that includes SAS Data Loader, SAS/ACCESS Interface to Hadoop, SAS In-Database Code Accelerator for Hadoop and SAS Data Quality Accelerator for Hadoop. The following diagram illustrates an installed configuration.



The SAS Data Loader for Hadoop web application runs inside a virtual machine or vApp. The vApp is started and managed by a hypervisor application called VMware Player Pro. The web application uses SAS software in the vApp and on the Hadoop cluster to manage data within Hadoop.

The hypervisor provides a web (HTTP) address that you enter into a web browser. The web address opens the SAS Data Loader: Information Center. The Information Center does the following:

- starts the SAS Data Loader web application in a new browser tab.
- provides a Settings window to configure the vApp connection to Hadoop.
- checks for available vApp software updates and installs vApp software updates.

All of the files that are accessed by the vApp reside in the shared folder. The shared folder is the only location on the user host that is accessed by the vApp. The shared folder contains the JDBC drivers needed to connect to external databases, and the Hadoop JAR files that were copied to the client from the Hadoop cluster.

When you create a job using a directive, the web application generates code that is then sent to the Hadoop cluster for execution. When the job is complete, the Hadoop cluster writes data to the target file and delivers log and status information to the vApp. Saved directives are stored in a database within the vApp.

The SAS In-Database Technologies for Hadoop software is deployed to each node in the Hadoop cluster. The in-database technologies consist of the following components:

- SAS Quality Knowledge Base for reference to data cleansing definitions.
- SAS Embedded Process software for code acceleration.
- SAS Data Quality Accelerator software for SAS DS2 methods that pertain to data cleansing.

How to Get Help for SAS Data Loader for Hadoop

SAS Data Loader Support Community

If you need additional help with using SAS Data Loader for Hadoop, the SAS Data Loader for Hadoop Community is a great place to find answers. Join the community to ask questions and receive expert online support.

Technical Support

SAS provides customer support through self-help and assisted-help resources. See our Support page for more information about these resources.

Documentation and System Requirements

If you select **Help** from the Help icon at top right of most windows, the SAS Data Loader documentation page is displayed.

If you select **About SAS Data Loader** from the Help menu, version information, supported browsers, legal notices, and license information is displayed.

The SAS Data Loader for Hadoop product page includes links to additional information, such as technical papers, training resources, and videos that show you how to perform common tasks.

You can find system requirements for SAS products at the SAS Install Center on support.sas.com. On the SAS Install Center page, select the SAS release, such as SAS 9.4 for SAS Data Loader. A search window for the relevant documentation appears. Search for your product name (such as SAS Data Loader). A results page appears with links to the system requirements for your software.

2

Getting Started

Prerequisites 7
First Tasks in SAS Data Loader 7
Create and Execute a Job Using SAS Sample Data 8
Naming Requirements for Schemas, Tables, and Columns 10

Prerequisites

The following tasks must be completed before you can use SAS Data Loader:

- A Hadoop administrator installs SAS In-Database Technologies for Hadoop across the nodes of a Hadoop cluster. The administrator then provides the vApp installer with site-specific files and settings, including configuration files and JAR files required for the cluster. For MapR deployments, the administrator provides an additional file that contains user and password information for accessing the MapR cluster. For more information about these tasks, see the “Administrator’s Guide for SAS Data Loader for Hadoop” section of the *SAS In-Database Products: Administrator's Guide*.
- The vApp installer configures a hypervisor (VMware Player Pro) and the vApp for SAS Data Loader on each client host. The installer sets up the shared folder for the vApp and adds the Hadoop files that were provided by the Hadoop administrator to the shared folder. The installer also specifies a connection to the Hadoop cluster. SAS Data Loader will later use this connection to access the Hadoop cluster. For more information about these tasks, see the *SAS Data Loader for Hadoop: vApp Deployment Guide*.

  To verify that these prerequisites have been met, open SAS Data Loader. For more information, see “Play the vApp and Start SAS Data Loader” on page 151.

First Tasks in SAS Data Loader

Here are some of the first tasks you can do in SAS Data Loader:

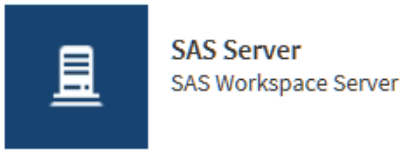
- Open SAS Data Loader. For more information, see “Play the vApp and Start SAS Data Loader” on page 151.

- Verify that you can connect to the Hadoop cluster. One way to do that is to browse tables on the Hadoop cluster. For more information, see "Browse Tables" on page 20.
- If you do not see the tables that you want to work with on the cluster, ask your Hadoop administrator if they should be there. You might want to copy tables into Hadoop. For more information, see Chapter 6, "Copy Data To and From Hadoop," on page 89.
- If you want to work with SAS LASR Analytic Server, see "Load Data to LASR" on page 118.
- If you want to review the global options for SAS Data Loader, see "Set Global Options" on page 132.
- To review what you can do with this software, see "What Does It Help You Do?" on page 2.

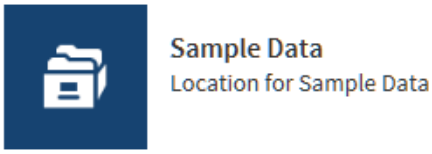
Create and Execute a Job Using SAS Sample Data

Follow these steps to copy a small SAS sample table into Hadoop and execute a transformation on that data.

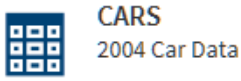
1. From the SAS Data Loader directives page, click the directive Copy Data to Hadoop.
2. In the **Source Table** task, click the **SAS Server** data source.



3. Click **Sample Data**.

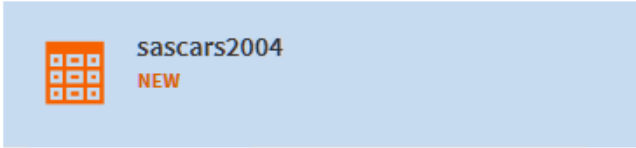


4. Click the CARS source table and click **Next**.



5. In the **Filter** task, click **Next** to include all SAS source rows in the Hadoop target table.
6. In the **Columns** task, click **Next** to accept the existing number and arrangement of columns.
7. In the **Target Table** task, click an appropriate location for the new table.

8 Click New Table... and enter a table name such as SASCars2004.



9 In the **Code** tab, browse and update the generated code as needed, and then click **Next**.

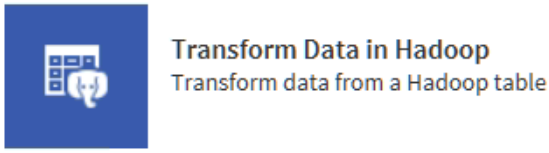
Note: Edited code is not retained when new code is generated.

10 In the **Result** task, click **Start copying data**.

11 Click **View Results** to see your new table in Hadoop.

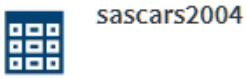
12 To transform your new table, click Back to Directives.

13 In the SAS Data Loader directives page, click **Transform Data in Hadoop**.

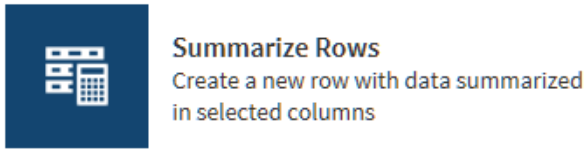


14 In the **Source Table** task, click the data source that you just used to store your new table.

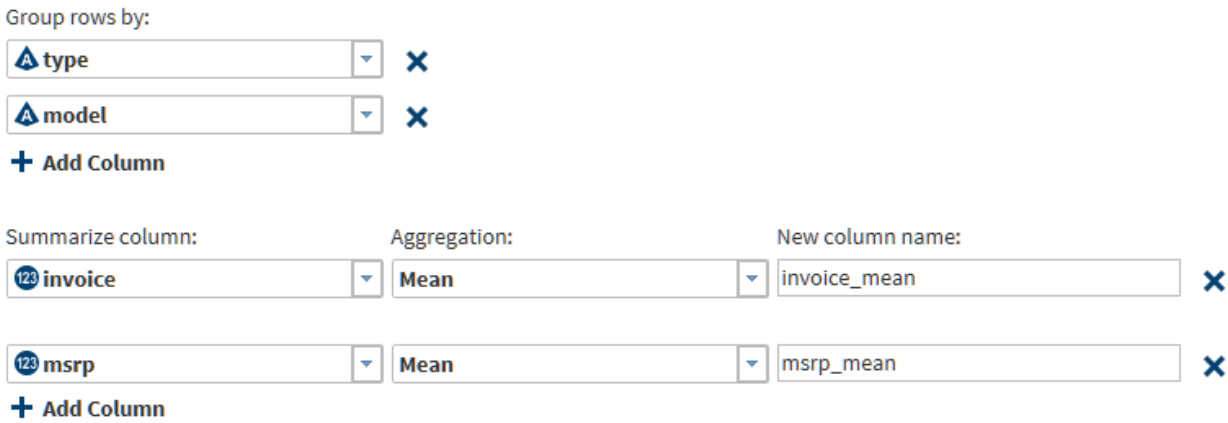
15 Click your new table, and then click **Next**.



16 In the **Transformations** task, select **Summarize Rows**.

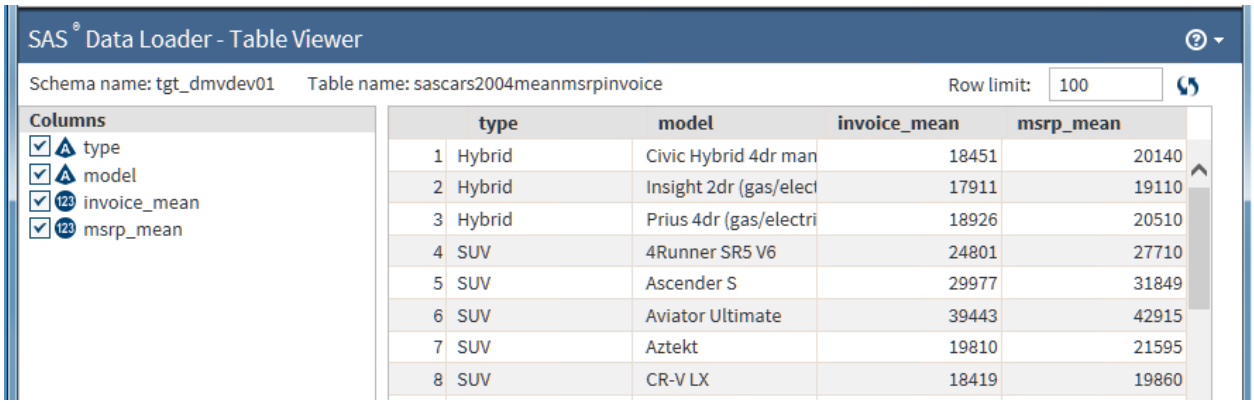


17 Select group-by rows, summaries and aggregations, and then click **Next**.



18 In the **Target Table** task, click the data source that you use for target data, click New Table... , enter a table name, and then click **Next**.

19 In the **Result** task, click **Start transforming data**. The job might run for a minute or so. Click **View Results** to see your first transformation in Hadoop using SAS Data Loader. Congratulations!



	type	model	invoice_mean	msrp_mean
1	Hybrid	Civic Hybrid 4dr man	18451	20140
2	Hybrid	Insight 2dr (gas/elect	17911	19110
3	Hybrid	Prius 4dr (gas/electri	18926	20510
4	SUV	4Runner SR5 V6	24801	27710
5	SUV	Ascender S	29977	31849
6	SUV	Aviator Ultimate	39443	42915
7	SUV	Aztekt	19810	21595
8	SUV	CR-V LX	18419	19860

Naming Requirements for Schemas, Tables, and Columns

Follow these requirements to name schemas, tables, and columns:

- Use alphanumeric characters in the names of schemas, tables, and columns.
  - Use underscore characters, but do not use an underscore character as the leading character in the name.
  - Do not use double-byte character sets in names.
  - Do not use Hive quoted identifiers (‘) in column names.
- Limit the length of the names of schemas, tables, and columns to 32 characters. This limit is required in the following directives and transformations:
  - Profile directive

- Transpose directive
- Summarize Rows transformation (a task in multiple directives)

3

About the Directive Interface

Using the Directives Page 13

Viewing Data Sources and Tables 14
Overview 14
About the SAS Table Viewer 16
About the Sample Table Viewer 17

Working with the Code Editor 18

Using the Directives Page

In the top-level web page for SAS Data Loader, you can browse and select directives. You can also select the following menus and icons:

Configuration
opens the Configuration window, with separate panels for configuring Hadoop connections, external database connections, SAS LASR Analytic Server connections, and several categories of user preferences. Some of these configurations are set during installation. You can also add a new database connection or add a connection to an instance of the SAS LASR Analytic Server software. For more information, see "Set Global Options" on page 132.

Back Up Directives
performs a backup of your saved directives. For more information about this option, see "Back Up Directives" on page 131.

Help
links to the SAS Data Loader documentation page on the SAS support website. If you select About SAS Data Loader from the Help menu, version information, supported browsers, legal notices, and license information is displayed.

Viewing Data Sources and Tables

Overview

For most directives in SAS Data Loader, data sources are Hive schemas that contain one or more tables. Data sources are defined in Hive by your Hadoop administrator. If you do not see the data source or table that you need, contact your Hadoop administrator. If needed, the administrator can add a new Hive schema and set appropriate user permissions for you to read and write data.

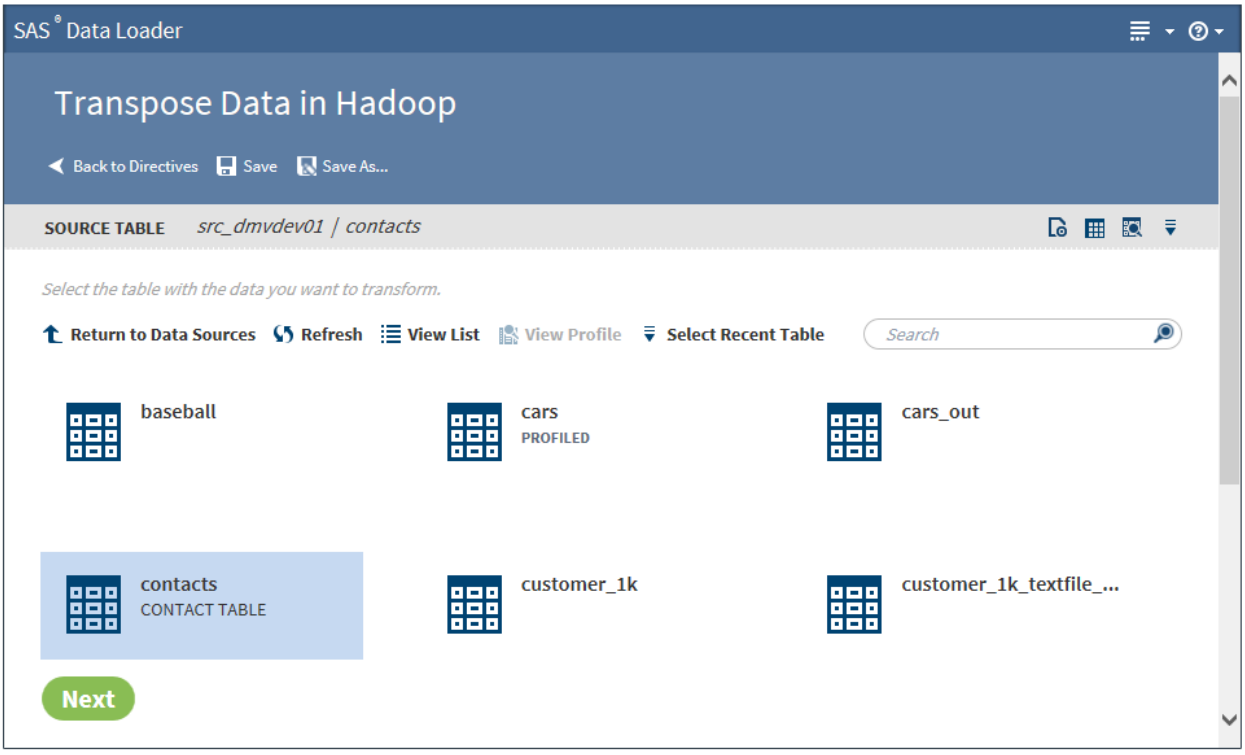
In some cases, data sources are not based on Hive schemas. For example, data sources for the Copy Data to Hadoop directive are RDBMS connections. Data sources for the Import a File directive are delimited files that are stored in the shared folder of the vApp.

When you open a directive to create a job that runs in Hadoop, you select a data source and a source table that is contained within that data source. If the directive produces output tables, you then select a data source and a target table at the end of the directive.

To protect your data, target tables do not overwrite source tables. Target tables are not required to be new tables each time you run your job. You can overwrite target tables that you created in previous job runs.

As the data is processed in each task in the job, you can view a sample on page 17 of the data that is produced in each task.

A typical Source Table task includes a graphical view of the tables in the selected data source.



SAS Table Viewer icon

Click to open the selected table in the SAS Table Viewer, which provides column information and sample data for the table.

View Data Sample icon

Click to display the first 100 rows of source data, as that data has been transformed up to that point in the job.

View List and **View Grid**

Click the View List icon to display data sources or tables as a list. When you view tables, the list format displays the table name and description, along with the dates on which the table was last profiled and last modified.

Note: The last modified date is displayed only when the **Identify each table as "new" when created or modified** setting is selected on the **General Preferences** panel of the Configuration window. For more information, see "General Preferences Panel" on page 143.

Click the View Grid icon to display data sources or tables in a grid.

View Profile

Click to view profile information for the selected table. If a profile exists for a table, PROFILED appears beneath the table name.

Select Recent Table

Click to choose from a list of recently used tables. If you select a table from a different data source, the source table information is adjusted accordingly. The table that you selected is automatically highlighted.

Search

Enter text in the search field to filter the list of data sources or tables. The search feature filters according to name when applied to data sources and according to name and description when applied to tables.

Click to return to the top of the page when viewing a long list of data sources or tables.

TIP If you frequently work with the same data source across multiple directives, you can have SAS Data Loader select the most recently used schema automatically. This can help you select source tables and target tables more quickly. To enable this feature, click , select **Configuration**, and complete the following steps:

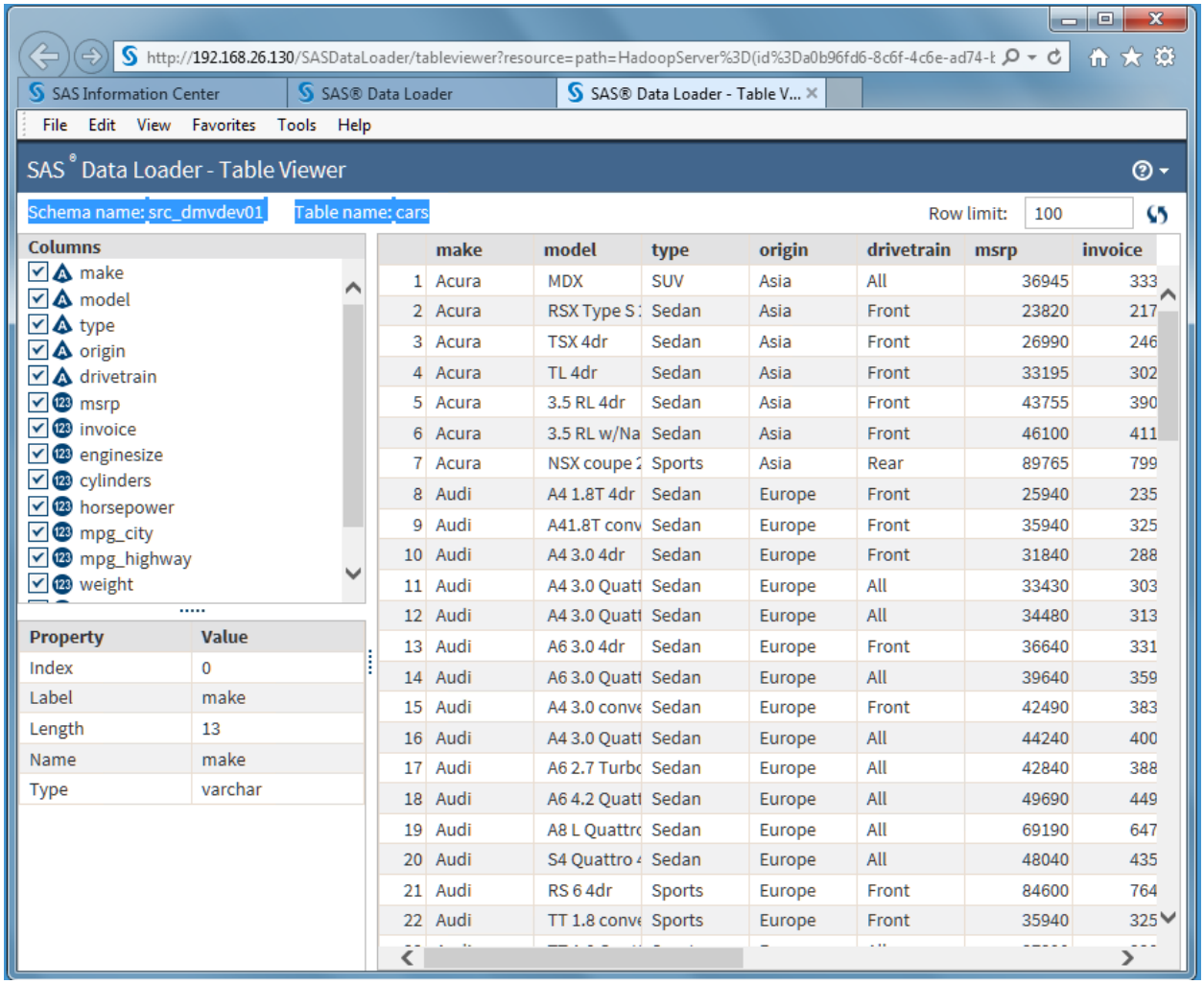
1. Click **General Preferences**.
2. Select **Automatically select the most recently selected hive schema**.

About the SAS Table Viewer

How It Works

The SAS Table Viewer displays sample data and column information for a selected table. The viewer is available when you select source or target tables or when you view results or status. The SAS Table Viewer opens in a separate tab in the browser, so you can continue to reference that information while working with directives.

To open the viewer, click the **Open the selected table in the table viewer** icon .



In the viewer, you can click a column name to display the properties of that column. You can also clear the check box next to the column name to temporarily remove that column from the sample data view.

To change the number of sample rows that are displayed, change the value of the **Row Limit** field.

To refresh the sample data after a directive has operated on that table, click the **Refresh** icon .

Column properties are defined as follows:

Index
: column number.

Label
: a shortened version of the column name that can be added to the data values for that column. If a label is not assigned, then the column name is used as the label.

Length
: the size of the table cell (or variable value) in bytes.

Name
: column name.

Type
: tThe type of the data in the column.

For information about data types and data conversions in SAS and Hadoop, see the chapter SAS/ACCESS Interface to Hadoop in the document *SAS/ACCESS Interface to Relational Databases: Reference*.

Usage Notes

- When viewing a SQL Server table, the following numeric data types are displayed in the **Columns** list with a character data type: datetime (datetime_col), money (money_col), smallmoney (smallmoney_col), numeric (numeric_col), and real (real_col).
- Viewing the source and target tables of transformations can show differences in decimal values. The source columns show no decimal values, and the target shows full double-precision values. This difference exists in the display only. In the Hadoop distributed file system (HDFS), the values are the same.

About the Sample Table Viewer

In directives that list tables for selection, you can click the **View a data sample** icon to display a subset of the source data, as that data has been transformed up to that point in the job. This gives you a preview of your data before you run your job against the full source table in Hadoop.

Data sample:

cust_number	cust_type	cust_entity_...	cust_status	cust_since_d...	cust_since
C0000000000...	Commercial	Organization	Active	2001-12-07	Dec 7, 2001
C0000000000...	Personal	Person	Active	1996-05-18	May 18, 199
C0000000000...	Personal	Person	Dormant	1992-06-27	Jun 27, 199
C0000000000...	Personal	Person	Active	2005-08-21	Aug 21, 200
C0000000000...	Personal	Person	Active	2008-04-03	Apr 3, 2008
C0000000000...	Personal	Person	Active	1991-11-12	Nov 12, 199
C0000000000...	Personal	Person	Dormant	2005-06-06	Jun 6, 2005
C0000000000...	Commercial	Organization	Active	1993-03-07	Mar 7, 1993
C0000000000...	Commercial	Organization	Active	2012-02-26	Feb 26, 201
C0000000000...	Personal	Person	Active	1994-06-17	Jun 17, 199
C0000000000...	Personal	Person	Active	2006-07-08	Jul 8, 2006
C0000000000...	Personal	Person	Active	2009-10-19	Oct 19, 200
C0000000000...	Commercial	Organization	Active	1990-01-13	Jan 13, 199

Next

In the data sample, you can click **Refresh** to display the latest data or click **X** to close the data sample.

Working with the Code Editor

You can edit and save changes to the code that is generated by directives. There are two ways to access code:

- use the code editor from the **Code** task within a directive

  Note: Some directives, such as Transform Data in Hadoop and Cleanse Data in Hadoop, do not include a **Code** task.

- download the code from a directive's **Result** task or from the Run Status directive. After downloading the code, you can work with it in a third-party text editor on your local machine.

The code editor is intended to be used only to implement advanced features. In normal use, there is no need to edit code. The code editor is a good way to see what will be running, but making changes can be problematic. If you make changes in the directive interface after you edit code, then your edits are lost when the code is regenerated. Also, your code edits are not reflected in the directive interface, which further complicates updates to edited code.

In addition to the code editor, SAS Data Loader provides two directives for user-written code. For more information, see Chapter 7, "Run User-Written Programs in Hadoop," on page 121.

4

Manage Data in Hadoop

Overview of Data Management Directives **20**

Browse Tables **20**
- Introduction 20
- Example 20

Cleanse Data in Hadoop **23**
- Introduction 23
- About Locales, Definitions, and the Quality Knowledge Base 23
- Select a Source Table 24
- Select a Data Cleansing Transformation 24
- Filter Data Transformation 25
- Change Case Transformation 28
- Field Extraction Transformation 29
- Parse Data Transformation 31
- Standardization Transformation 32
- Pattern Analysis Transformation 33
- Identification Analysis Transformation 35
- Gender Analysis Transformation 36
- Generate Match Codes Transformation 37
- Manage Columns Transformation 38
- Summarize Rows Transformation 40
- Select a Target Table and Run Your Job 41
- About DS2 Expressions and the Advanced Editor 42

Delete Rows **43**
- Introduction 43
- Prerequisites 43
- Example 44

Query or Join Data in Hadoop **46**
- Introduction 46
- Example 47

Sort and De-Duplicate Data in Hadoop **55**
- Introduction 55
- Example 55
- Using the Advanced Editor for Hive Expressions 60

Transform Data in Hadoop **61**
- Introduction 61
- Example 61
- About the Operators in the Filter Data Transformation 64
- About the Aggregations in the Summarize Rows Transformation 69
- Usage Notes 70

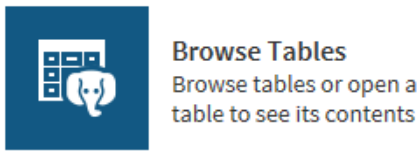
Transpose Data in Hadoop 70
Introduction 70
Example 70
Usage Notes 71

Overview of Data Management Directives

The data management directives support combinations of queries, summarizations, joins, transformations, sorts, filters, column management, and de-duplication. Data quality transformations include standardization, parsing, match code generation, and identification analysis, combined with available filtering and column management to reduce the size of target tables.

Browse Tables

Introduction

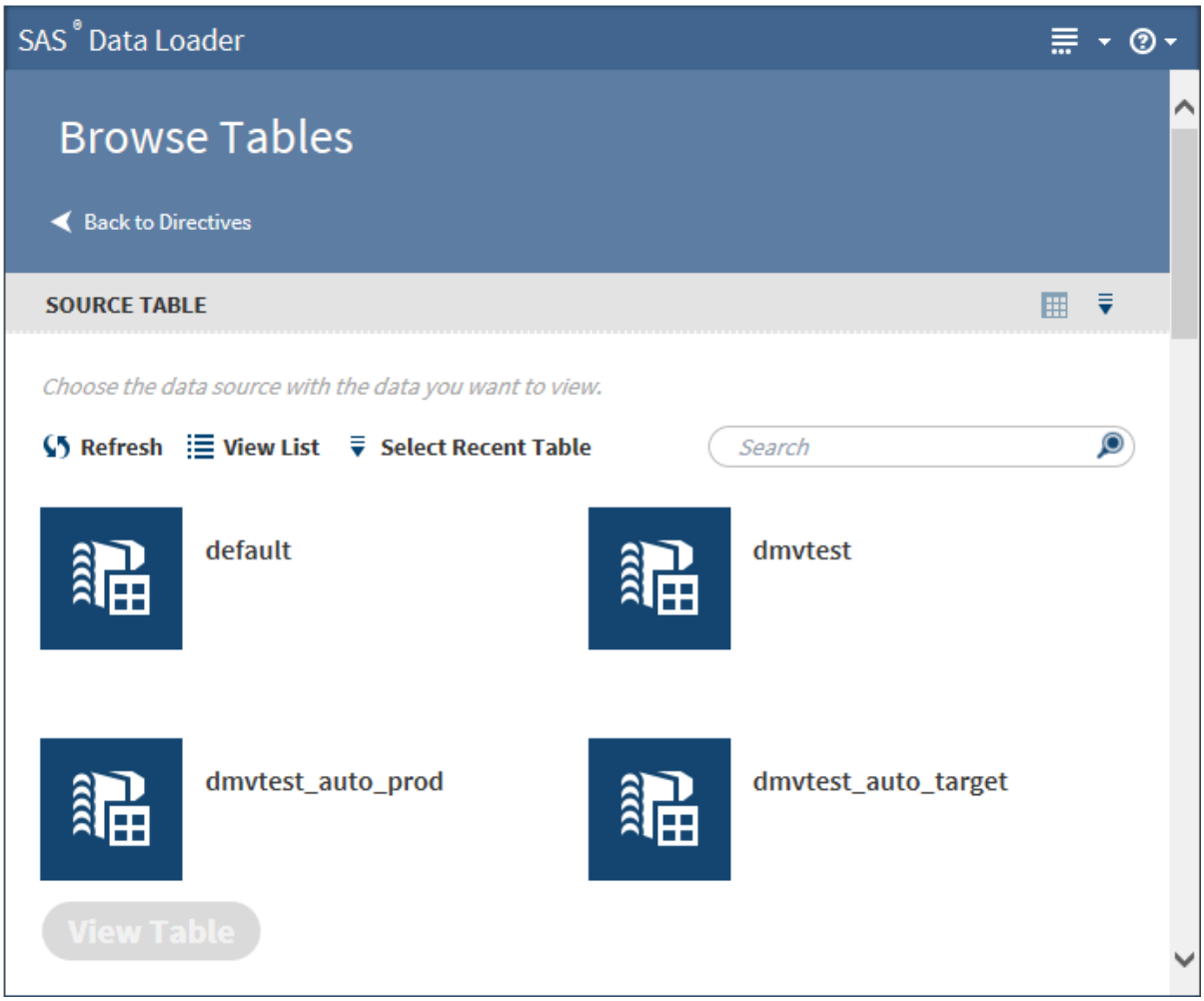


Use the Browse Tables directive to browse a list of the tables in a data source that is available on the Hadoop cluster. You can also view the contents of a table in the Table Viewer. With the Browse Tables directive, you can examine the data on quickly and conveniently before you begin working with the data in other directives.

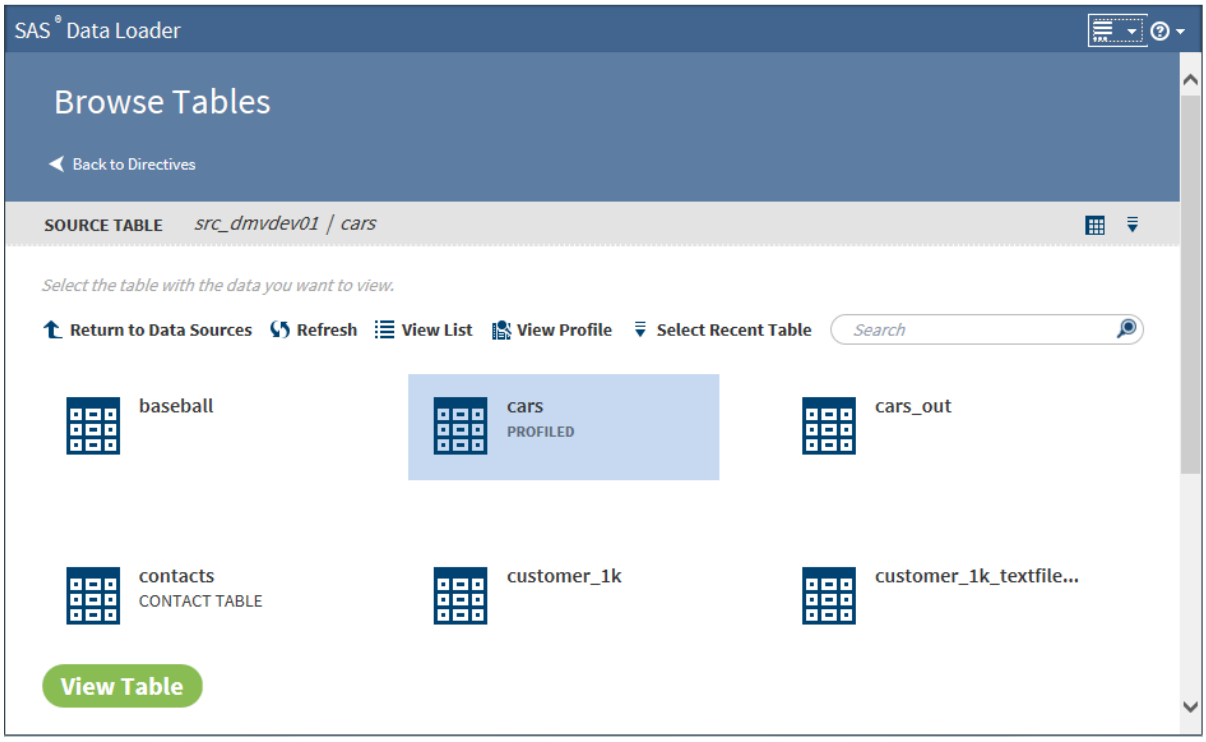
Example

Follow these steps to view the data in a table:

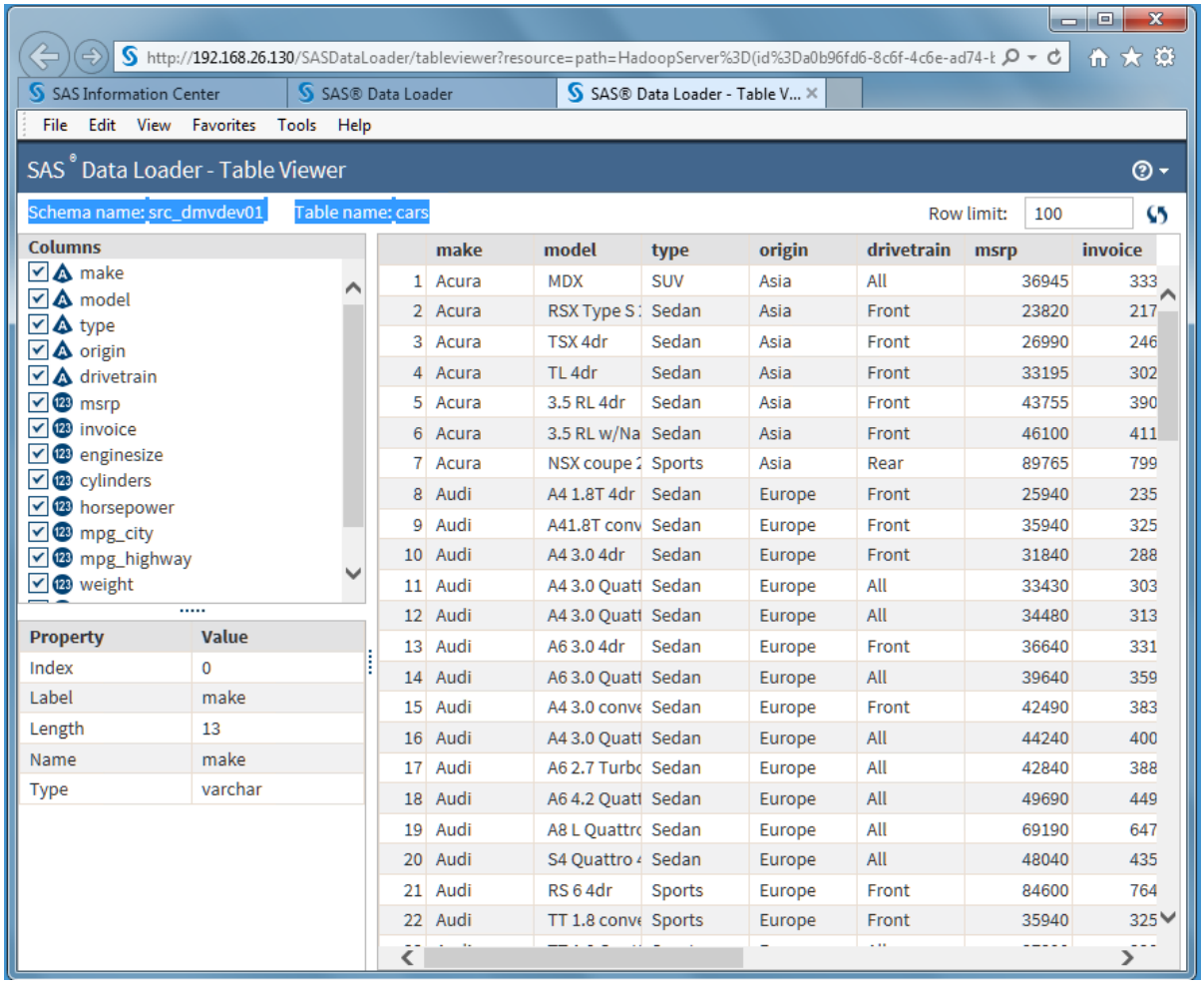
1 On the SAS Data Loader directives page, click the Browse Tables directive. The Source Table task that lists available data sources is displayed:



2 Click a data source to display its tables, and select the table that you want to view.



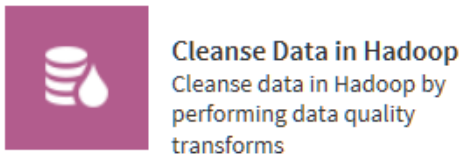
3 Click **View Table**.



TIP Because the SAS Table Viewer appears in a separate browser tab, you can view the contents of multiple tables at the same time. For each additional table, just return to the Browse Tables directive in the **SAS Data Loader** tab and repeat the previous steps.

Cleanse Data in Hadoop

Introduction



Use the Cleanse Data in Hadoop directive to create jobs that improve the quality of your Hadoop data. Your jobs can combine any of the data quality transformations in any order. When you run your job, the transformations are executed in the order in which you defined them.

About Locales, Definitions, and the Quality Knowledge Base

Most of the data quality transformations ask you to select a source column, a locale, and a definition. A *locale* represents a distinct alphabetical language, combined with a specified regional usage of that language. For example, the English, United States locale applies only to that region. The locale English, England addresses different usage or data content for the same alphabetic language.

A locale consists of a collection of *definitions*. Definitions tell SAS how to cleanse data. For example, the Street Address definition for the English, United States locale describes the structure of the first part of an American mailing address. In the locale Spanish, Mexico, the Street Address definition accommodates differences in mailing address structure as well as the differences in language and alphabet.

Locales and definitions make up a *SAS Quality Knowledge Base*. A Quality Knowledge Base is deployed on your Hadoop cluster. When you run a data cleansing job in Hadoop, the SAS software on your cluster accesses the Quality Knowledge Base to transform your data.

In SAS Data Loader you specify a default locale, which should match the typical locale of your source data. The default locale is selected in the **QKB** panel of the Configuration window, as described in "QKB Panel" on page 142. You can override the default locale in any of the data quality transformations. The override applies only to the current transformation.

To learn more about the Quality Knowledge Base, refer to the related document titles in "Recommended Reading" on page 155.

To learn about the output that is generated by a given definition, refer to the online Help for the SAS Quality Knowledge Base, in the topic Global Definitions.

Select a Source Table

When you use a data cleansing directive to create and run a job, you begin by selecting a source table.

Follow these steps to select a source table:

1. Scroll through the list or grid of data sources or schemas, and then click the data source (also known as a schema) that contains your source table. You can also click **Select a Recent Table** and quickly choose from that list.
2. If you opened a data source, click the source table and then click **Next**.

   Note: To explore the contents of source tables, click a table and click **Data Sample**, Table Viewer, or (if available) View Profile
3. In the **Transformation** task, click a data cleansing transformation to begin building your job.

Select a Data Cleansing Transformation

In a new job, after you select a source table, click a data cleansing transformation. Click below to find usage information for your selected transformation:

- "Filter Data Transformation" on page 25.
- "Change Case Transformation" on page 28.
- "Field Extraction Transformation" on page 29.
- "Parse Data Transformation" on page 31.
- "Standardization Transformation" on page 32.
- "Pattern Analysis Transformation" on page 33.
- "Identification Analysis Transformation" on page 35,
- "Gender Analysis Transformation" on page 36.
- "Generate Match Codes Transformation" on page 37.
- "Manage Columns Transformation" on page 38.
- "Summarize Rows Transformation" on page 40.

Filter Data Transformation

Use the Filter Data transformation at the beginning of a job to decrease the number of rows that will be processed in subsequent transformations.

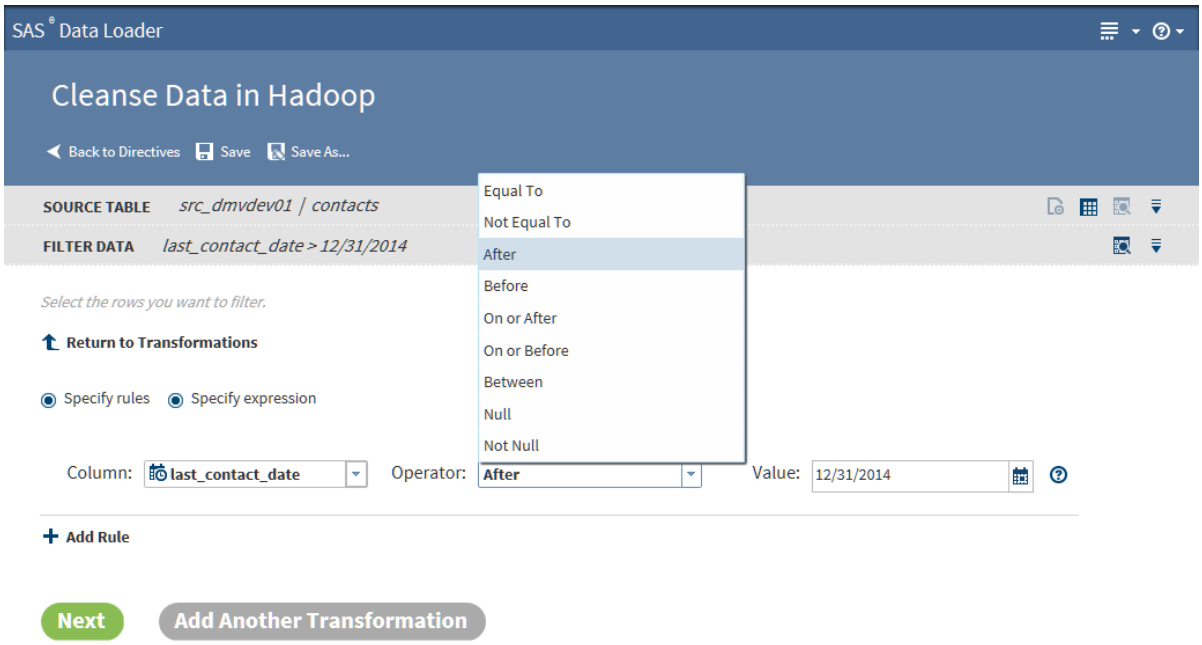
Follow these steps to use the Filter Data transformation:

1. If this is the first transformation in a new job, select a source table.
2. In the **Transformation** task, click **Filter Transformation**.

3. In the **Filter Data** task, choose one of the following:
   a. To filter rows using one or more rules, click **Specify rules** and proceed to the next step. You can specify multiple rules and apply them using logical AND and OR operators.
   b. To filter rows using a DS2 expression, click **Specify expression** and go to Step 5.

   TIP If the table that you selected has been profiled, an ellipsis button (...) appears next to the filter value selection. Click that button to view profile results while building your filters. For more information about generating profile reports for tables, see "Profile Data" on page 75.

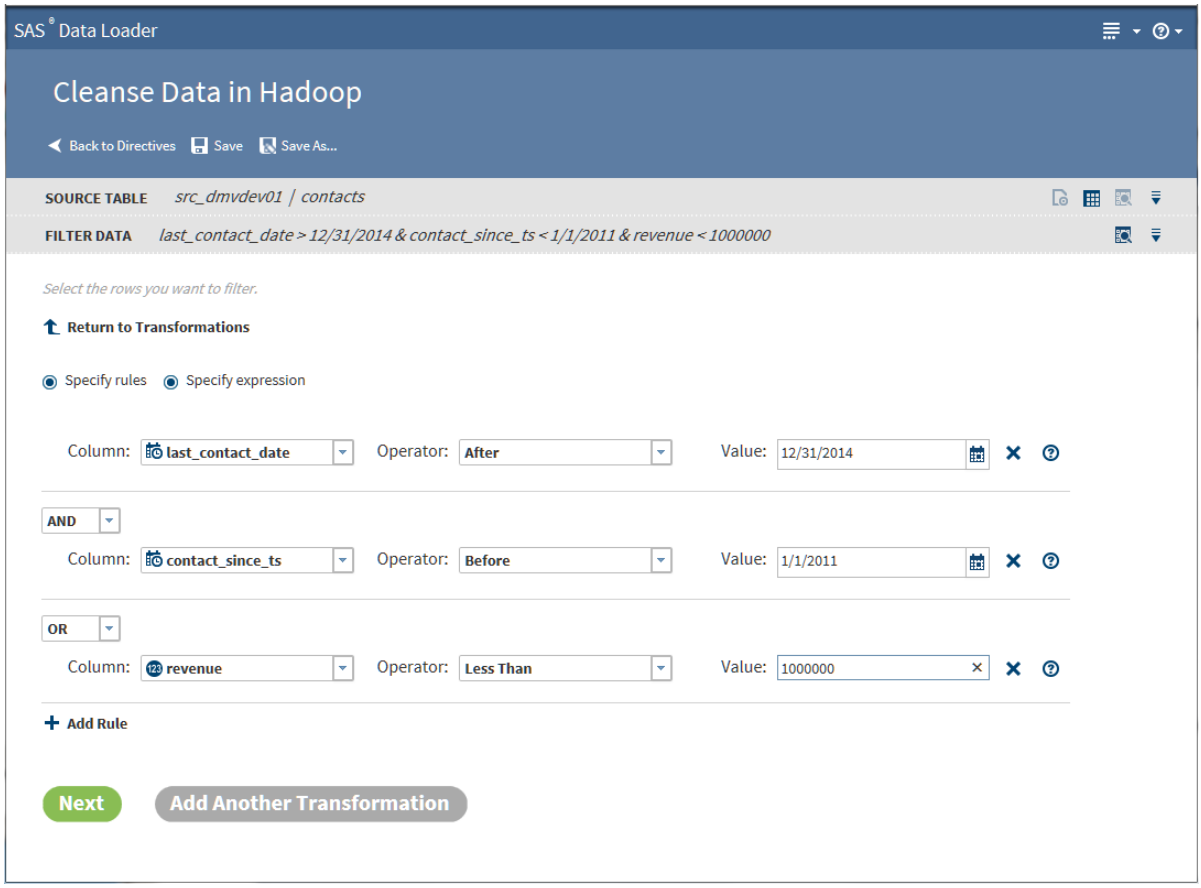
4. To filter rows by specifying one or more rules, follow these steps:
   a. Click **Select a column** and choose the source column that forms the basis of your rule.
   b. Click and select a logical **Operator**. The operators that are available depend on the type of the data in the source column. For example, the following image shows the operators that are available for the date/time data type:



c In the **Value** field, add the source column value that completes the expression. In the preceding example, the rule can be read as “Filter from the target all source rows with a last contact date after December 31, 2014.”

d Click **Add Rule** to add another rule. Select a different column, operator, and value.

e To filter rows when either the new rule or the preceding rule are true, change the **AND** condition to **OR**.

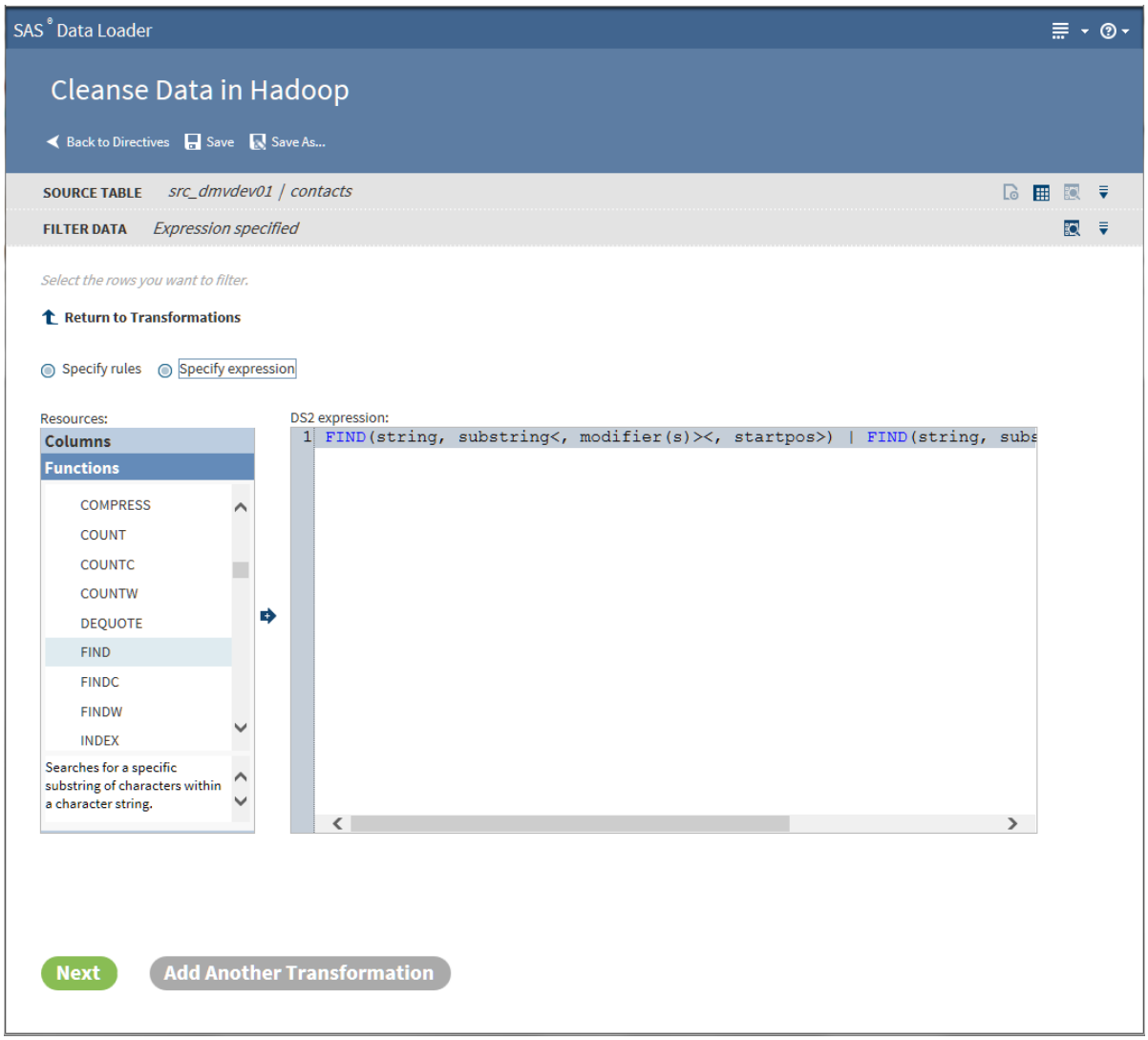


f When your rules are complete, go to Step 6.

5 To filter rows using a DS2 expression, follow these steps:

a In the **DS2 expression** text box, enter or paste a DS2 expression.

b To add DS2 functions to your expression, click **Functions** in the **Resources** box, expand a category, select a function, and click .



To add column names to your expression, position the cursor in the **DS2 expression** box, click **Columns** in the **Resources** box, click a source column, and then click .

6 When your rules or expression are complete, click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

Change Case Transformation

Use the Change Case transformation to standardize the casing of selected character columns. You can convert to ALL UPPERCASE, all lowercase, or Initial Capital Letters (or Proper Case).

Follow these steps to use the Change Case transformation:

1 If this is the first transformation in a new job, select a source table.

2 In the **Transformation** task, click **Change Case**.

3 In the **Change Case** task, accept or change the default **Locale**. The selected locale needs to reflect the language and region that applies to the content in the source table.

4 Click to **Select a Column**.

5 Select a **Type** of casing for the selected column.

6 Select the case **Definition** that best fits the content of your column. For the source column contact_full_name, and for **Proper** casing, you would select the case definition **Proper (Name)**.

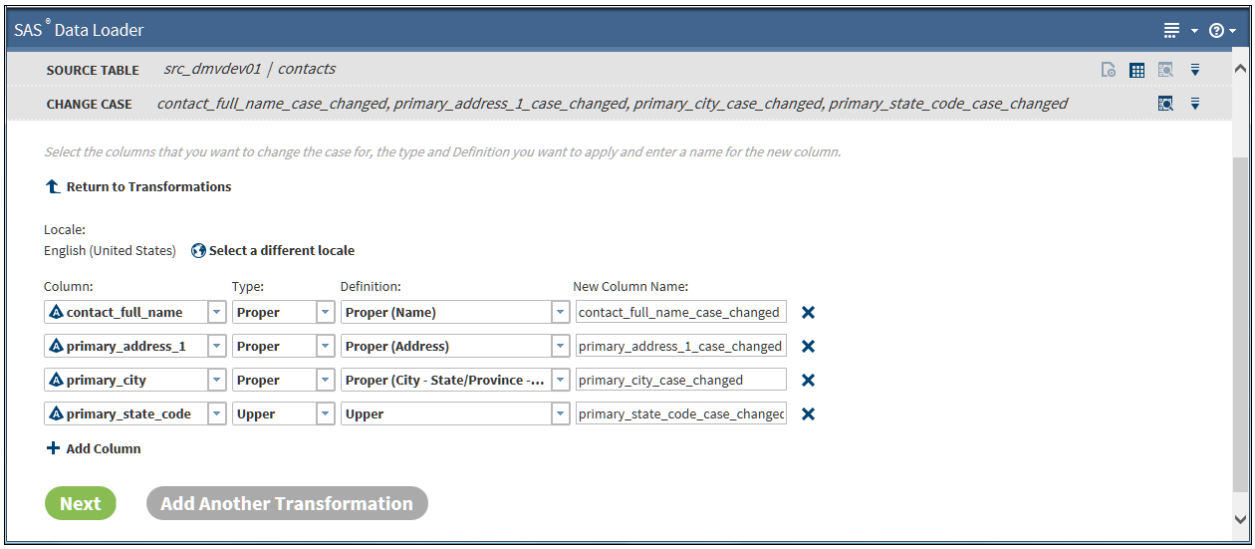
The case definition is part of the SAS Quality Knowledge Base that is installed on your Hadoop cluster. The case definition determines how case changes are applied to your data, based on your data content and selected locale.

7 Accept or change the default value in the field **New Column Name**.

8 Click **Add Column** + to define another case change.

9 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.



Field Extraction Transformation

Use the Field Extraction transformation to copy tokens from a source column to new columns in the target. Tokens represent types of content that can be extracted using an extraction definition. The available extraction definitions provide locale-specific information that enables extraction.

Follow these steps to use the Field Extraction transformation:

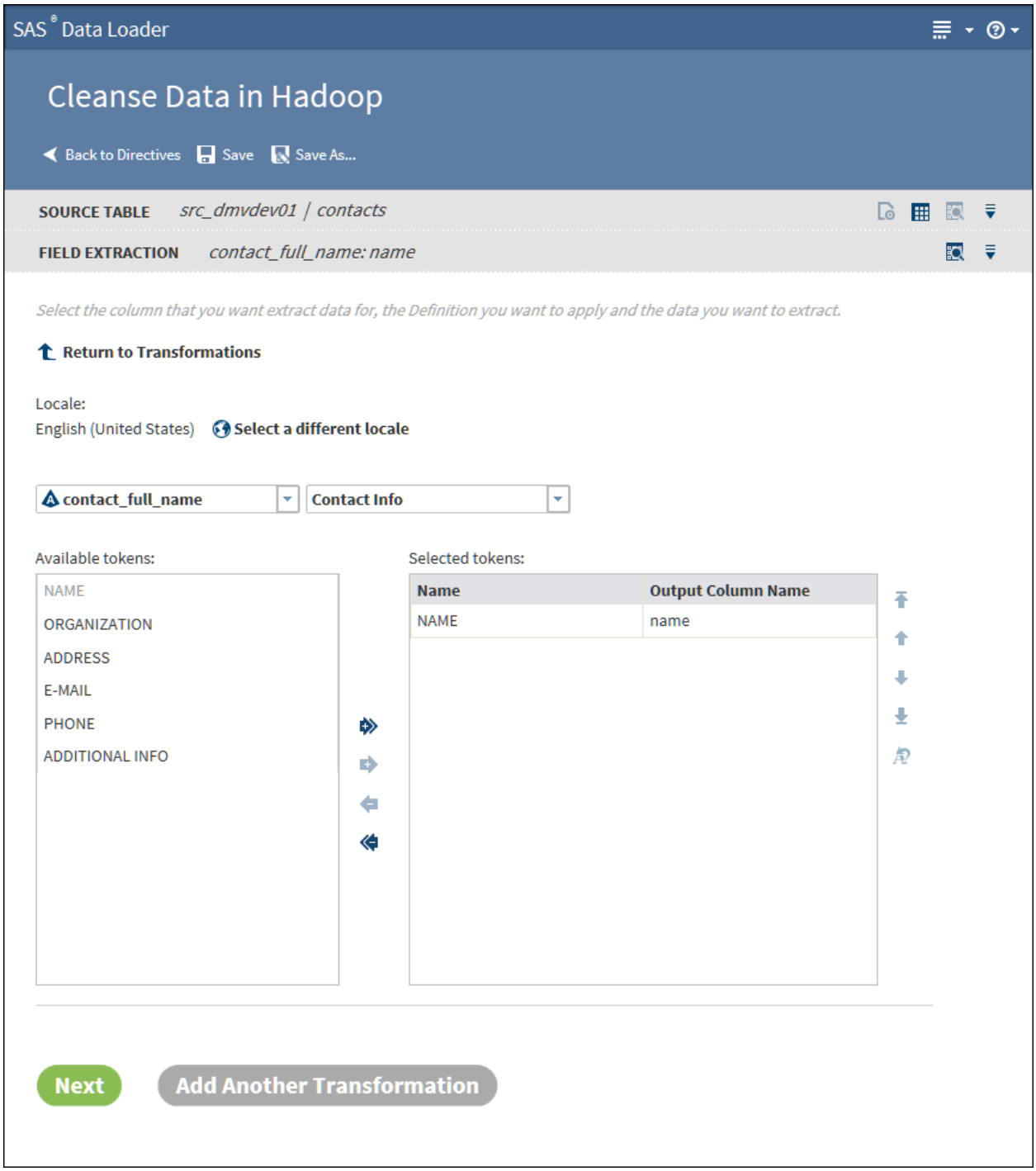
1. If this is the first transformation in a new job, select a source table.
2. In the **Transformation** task, click **Field Extraction**.

3. In the Field Extraction transformation, accept or change the default **Locale**.
4. Click **Column** and select a column from which you want to copy data to the target.
5. Click **Definition** and select the set of Field Extraction definitions that best fit your source data. Typical available selections include **Contact Info** and **Product Data**. The list of tokens that appear after you make your selection will show if you selected the appropriate definition.

   The tokens that you select are used to parse each source row and extract values of the specified type.
6. Click one or more tokens that you want to extract from the selected column and click . The tokens and default new column names appear in **Selected Tokens**.

   To select all tokens, click .
7. To change the default column name, click on the name in **Output Column Name**.
8. To reorder the columns in the target, click a row in **Selected tokens** and then click the up and down icons to the right of **Selected tokens**. The top row in **Selected tokens** specifies the first row in the target.
9. Click **Next** to select a target table and run your job.

   To add another data cleansing transformation, click **Add Another Transformation** and see “Select a Data Cleansing Transformation” on page 24.



Parse Data Transformation

Use the Parse Data transformation to extract tokens from a source column and add the token to a new column. A token is a meaningful subset of a data value that provides a basis for analysis. For example, for a column that contains phone numbers, you could extract the area code token and insert that value in a new column. You could then analyze the source table by grouping rows by area code.

Follow these steps to learn how to use the Parse Data transformation:

1 If this is the first transformation in a new job, select a source table.

2 In the **Transformation** task, click **Parse Data**.

3 In the **Parse Data** task, click **Select a column** and select a source column from the list.

4 Click the **Definition** field and click the definition that you will apply to the selected column.

5 In the **Available tokens** list, click the token that you will copy out to a new target column.

6 Click the right plus arrow to apply the token to a new column. You can change the suggested **Output Column Name**.

7 At this point you can choose other tokens to add to other new columns in the target table.

8 If you have multiple tokens, you can arrange the target columns using the up and down arrow icons.

9 To remove a token column, select it and click the minus arrow icon .

10 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

Standardization Transformation

Follow these steps with your own data to learn how to use the Standardization transformation. This example creates a job that standardizes a column of state names in a table of customer data.

1 If this is the first transformation in your job, select a source table.

2 In the **Transformation** task, click **Standardize Data**.

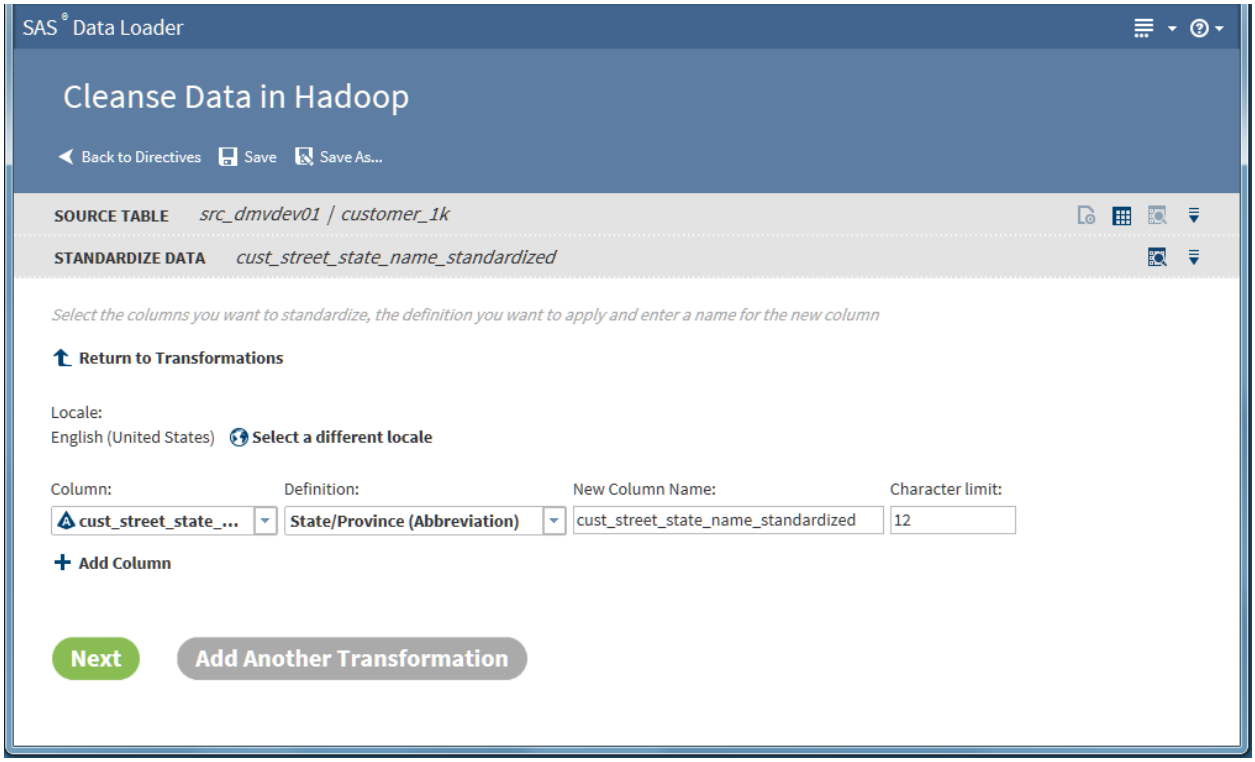
3 In the **Standardize Data** task, click **Select a Column** and select the column from the list.

4 Click **Select a Definition** and select the standardization definition to be applied to the selected column. Standardization definitions are available for certain character strings and numeric values. Also, standardization

definitions are available for generic actions that are independent of content, such as Space Removal and Multiple Space Collapse. To learn about the standardization definitions, seeStandardization Definitions in the online Help for the SAS Quality Knowledge Base.

5 Standardized values are applied to a new column in the target. You can change the default name of the new column by clicking **New column name**.

6 To save space or truncate long values, you can change the **Character limit** from its default value of 256.



7 The standardization transformation is now completely defined. By default, the target table contains both the original source column and the new standardized column. If you would prefer to remove the source column in the target or make other changes to target columns, add a Manage Columns transformation toward the end of your job.

Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

Pattern Analysis Transformation

The Pattern Analysis transformation reads a source row and generates a corresponding pattern value in the target. The content of the pattern value describes the content of the data. For example, character pattern analysis generate patterns that show if each character is uppercase, lowercase, or numeric.

The patterns form the basis for structural analysis. For example, you can apply a Filter transformation to the output of a pattern analysis. The filter can exclude the expected pattern and write to the target the rows that are structurally invalid.

Follow these steps to use the Pattern Analysis transformation:

1. If this is the first transformation in your job, select a source table.
2. In the **Transformation** task, click **Pattern Analysis**.

3. In the **Pattern Analysis** task, accept or change the default **Locale**. The selected locale needs to reflect the language and region that applies to the content in the source table.
4. Click **Select a column** and click the column that you want to analyze.
5. Click **Definition** and select a pattern analysis definition.

   Character
   generates patterns that represent the types of each character in the source. `A` indicates uppercase, `a` indicates lowercase, `9` indicates numbers, and `*` indicates other (punctuation, and so on). Blanks in the source are replicated as blanks in the pattern. Example: the source value `1 877-846-Flux` generates the pattern `9 999*999*Aaaa`.

   Character (Script Identification)
   generates patterns that identify the Unicode character set of each character in the source. Eleven or more character sets can be detected, including Latin, Arabic, Kanji/Han, katakana, Cyrillic, and Numeric. Uppercase and lowercase are detected for at least three character sets. Example: `(7F, SAS Institute)スズキイチロウ` generates `*9L* LLL Lllllllll*アアアアアア`.

   Note: The full mapping of pattern characters to Unicode character sets is provided in the Pattern Analysis Definitions in the online Help for the Contact Information Quality Knowledge Base.

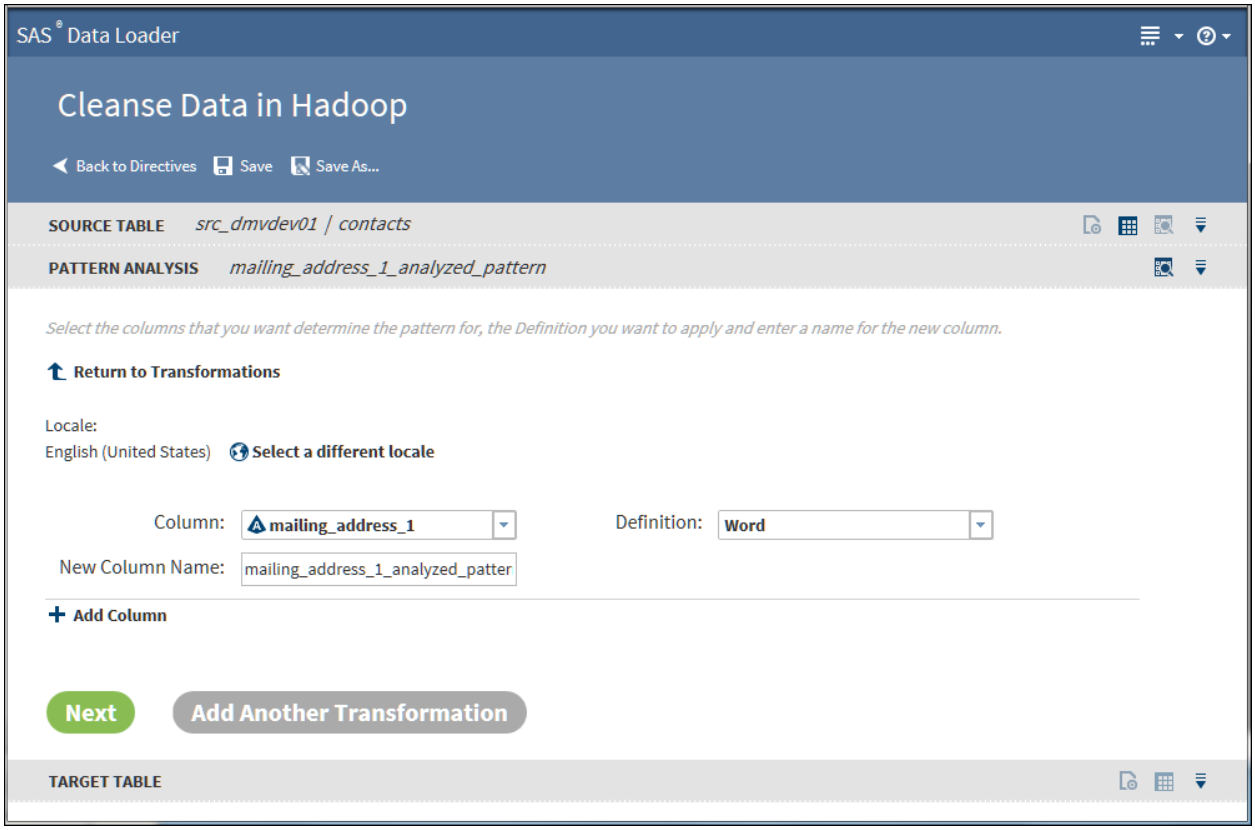
   Word
   generates patterns that represent the types of words in the source. `A` represents alphabetic words, `9` numeric, `M` mixed, and `*` other. Example: `216 E 116th St` generates `9 A M A`.

   Word (Script Identification)
   generates patterns that represent the Unicode character set of each word in the source. Eleven or more character sets can be detected, including Latin, Arabic, Kanji/Han, Katakana, Cyrillic, and Numeric. `W` indicates a potentially invalid word that contains multiple character sets. Example: `(7F, SAS Institute)スズキイチロウ` generates `*9L* L L*ア`.

6. Review and update the default **New Column Name**.
7. Review and update as needed the default **New Column Name**.
8. To generate patterns for other columns, click + Add Column.

9 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.



Identification Analysis Transformation

Use the Identification Analysis transformation to report on the type of the content in a given column. The content types that can be detected include contact information, dates, email, field names, offensive content, and phone numbers. The result of the analysis is added to a new column in the target table. You can analyze one column for multiple content types, and you can analyze multiple columns in the source table.

Follow these steps to use the Identification Analysis transformation:

1 If this is the first transformation in your job, select a source table.

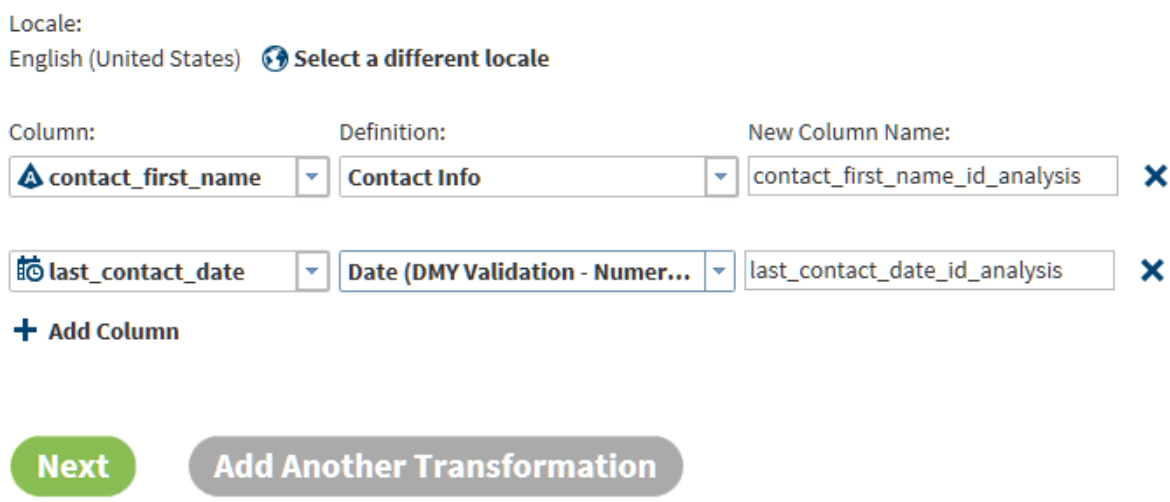
2 In the **Transformation** task, click **Identification Analysis**.

3 In the **Identification Analysis** task, click **Select a Column**, and then select a column for analysis.

4 Click **Select a Definition** and choose the content type that you want to apply to the source column.

5 In the **New Column Name** field, a name is suggested for the column that will be added to the target. The new column will contain the results of the identification analysis. Click in **New Column Name** field to change the suggested column name.

6 To analyze another column, or to analyze the same column with a different definition, click **Add Column**.



7 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

Gender Analysis Transformation

The Gender Analysis transformation analyzes columns of names and generates columns that indicate the probable gender of the names.

Follow these steps to use the Gender Analysis transformation:

1 If this is the first transformation in your job, select a source table.

2 In the **Transformation** task, click **Gender Analysis**.

3 In the **Gender Analysis** task, review and update the default **Locale** as needed to ensure that the locale matches the content of your source data.

4 Click **Select a Column** and click the column of name data in your source table.

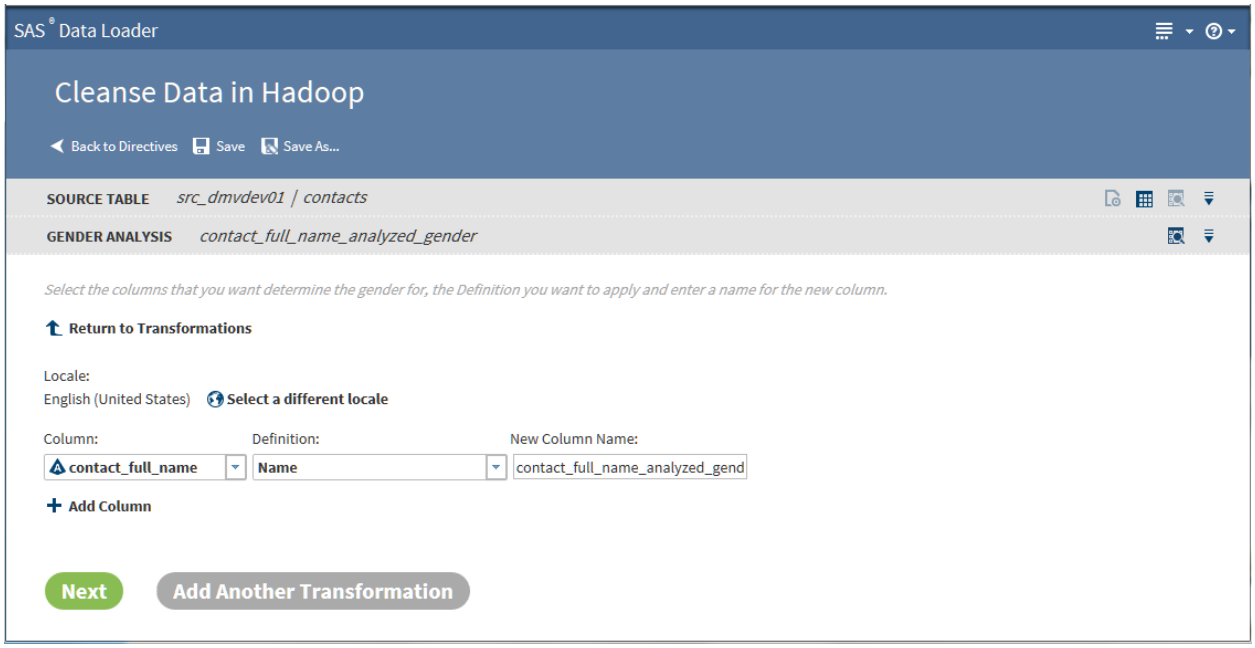
5 Click **Definition** and click **Name**.

6 To analyze a second column of name data, click + Add Column.

7 Review and update as needed the default **New Column Name**.

8 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.



Generate Match Codes Transformation

The Generate Match Codes transformation generates match codes for specified columns. The generated match codes are then added to new columns in the target table. The match codes are generated based on a definition and a sensitivity. The definition specifies the type of the content in the column. The sensitivity determines the degree of exactitude that is required in order for two data values to be declared a match. Higher sensitivity values specify that data values must be more similar to be declared a match. Lower sensitivity values enable matching with less similarity. The level of sensitivity is reflected in the length and complexity of the match codes.

Match codes can be used to find columns that contain similar data. For example, you can generate match codes for name and address columns, and then compare the match codes to detect duplicates.

Follow these steps to use the Generate Match Codes transformation:

1 If this is the first transformation in your job, select a source table.

2 In the **Transformation** task, click **Generate Match Codes**.

3 In the **Generate Match Codes** task, click **Select a Column** and then click the column for which you want to generate match codes.

4 Click **Select a Definition** and then click the definition that you want to use to generate match codes.

5 To change the default sensitivity value, click the **Sensitivity** field and select a new value. Lower sensitivity numbers give you more matches (less than identical match codes) and perhaps more matching errors. Higher sensitivity numbers produce the same match code only when data values are nearly identical.

6 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

Manage Columns Transformation

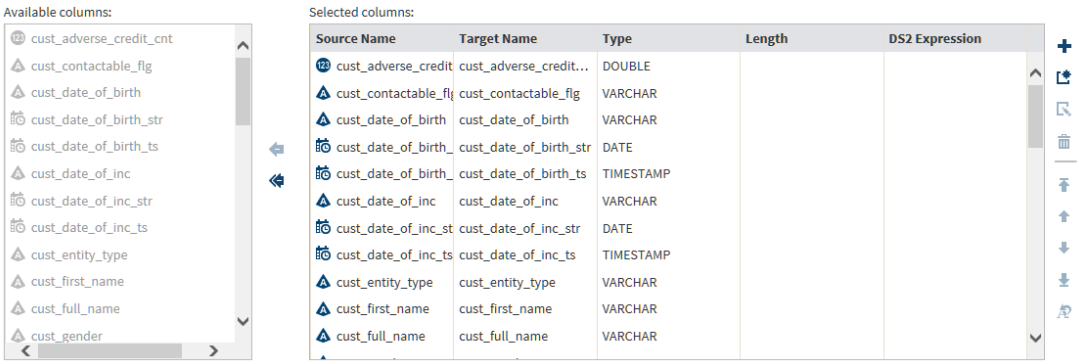
Use the Manage Columns transformation to remove, reorder, and rename source columns. You can also add new columns. The new columns contain generated values of a specified length and type. The values are generated by a DS2 expression that you supply, based on the values in each row. To learn more about DS2, see the *SAS 9.4 DS2 Language Reference*.

Follow these steps to learn how to use the Manage Columns transformation:

1 If this is the first transformation in your job, select a source table.

2 In the **Transformation** task, click **Manage Columns**.

3 In the **Manage Columns** task, columns are listed in order of appearance. The top column is the first or leftmost column.



Note the arrow icons between **Available columns** and **Selected columns**. To remove a column from the target, click the column name on the right and click the top arrow. To move all columns out of the target, click the double-arrow icon. After you remove a column, arrows will appear so that you can move columns from Available to Selected.

Initially, all columns are selected for the target table, including all of the new that you added in prior transformations.

4 Locate the icons on the right side of **Selected columns**. These icons provide the following functions:

- Add a new column and enter a DS2 expression for that column without using the Advanced Editor.
- Add a new column and specify a DS2 expression using the Advanced Editor.
- Edit the selected column using the Advanced Editor to modify its DS2 expression.
- Remove the selected column from the target table. Removed columns appear in **Available columns**.
- Move the selected column to the first column position in the target (leftmost).
- Move the selected column one position to the left in the target.
- Move the selected column one position to the right.
- Move the selected column to the last column position in the target (rightmost).
- Change the name of the selected target column.

5 If you want to add or paste a DS2 expression into an existing column, click the DS2 Expression field for that column and proceed. Any source data in that column will be replaced by the results of the DS2 expression.

6 If you want to use the Advanced Editor to define a DS2 expression, click and see “About DS2 Expressions and the Advanced Editor” on page 42.

7 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

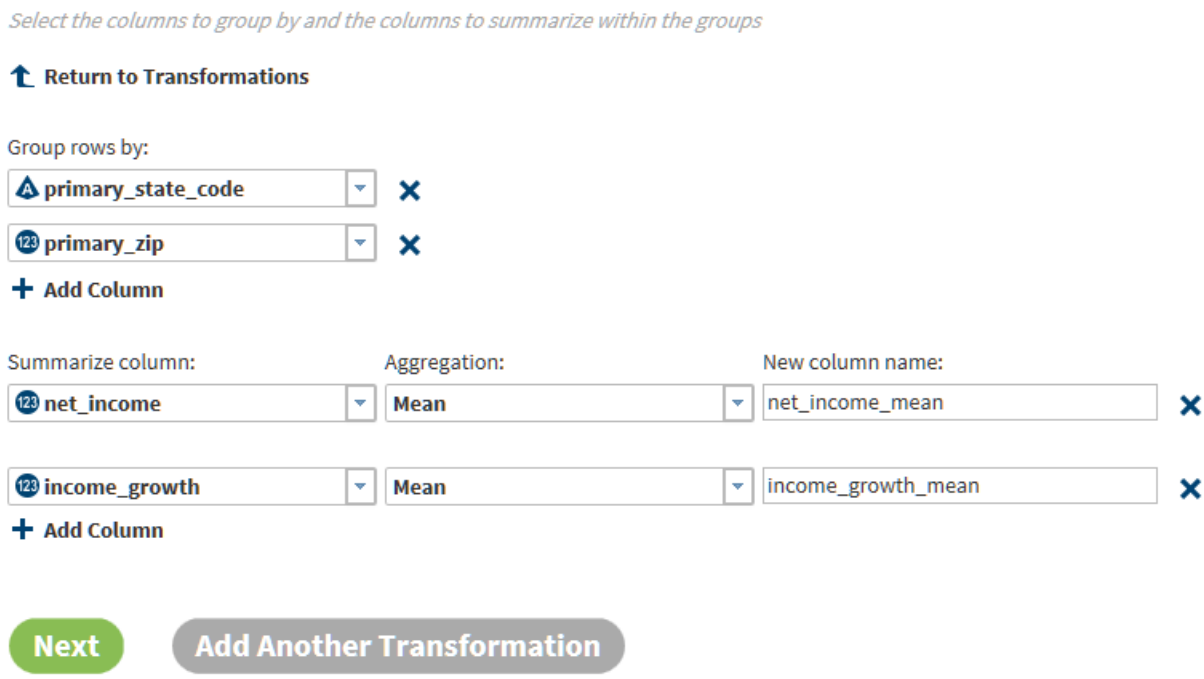
Summarize Rows Transformation

Use the Summarize Rows transformation to add summarized numeric values to your target table. To generate summaries, you first group rows by one or more columns. Then you select the columns that you want to summarize for each group and subgroup. The method of summarization is known as an aggregation. The number of aggregations depends on the column data type. Numeric columns have 13 available aggregations.

Follow these steps to learn how to use the Summarize Rows transformation:

1. If this is the first transformation in your job, select a source table.
2. In the **Transformation** task, click **Summarize Rows**.

3. In the **Summarize Rows** task, click the **Group rows by** field and choose the first column that you want to use to group rows for summarization. In the target table, rows with the same values in the selected column appear together, along with their summary values in new columns.
4. To further subset the initial set of groups, and to generate a second set of summary values, click **Add Column**. Select a second column. Add additional groups as needed.
5. Click **Summarize column** and select the first numeric column that you want to summarize.
6. Click **Aggregation** and select the aggregation that you would like to provide for the selected column.
7. To change the suggested name for the new column that will contain the aggregation values for each group, click **New Column Name**.
8. To add a second aggregation, click **Add Column**.



9 Click **Next** to select a target table and run your job.

To add another data cleansing transformation, click **Add Another Transformation** and see "Select a Data Cleansing Transformation" on page 24.

Select a Target Table and Run Your Job

After you click **Next**, follow these steps to select a target table and complete your data cleansing job:

1 In the **Target Table** task:

To select an existing target table (and completely overwrite any existing content), click the data source, click an existing target table, and then click Next. Or, you can click **Select Recent Table** and choose from a list of your recent targets.

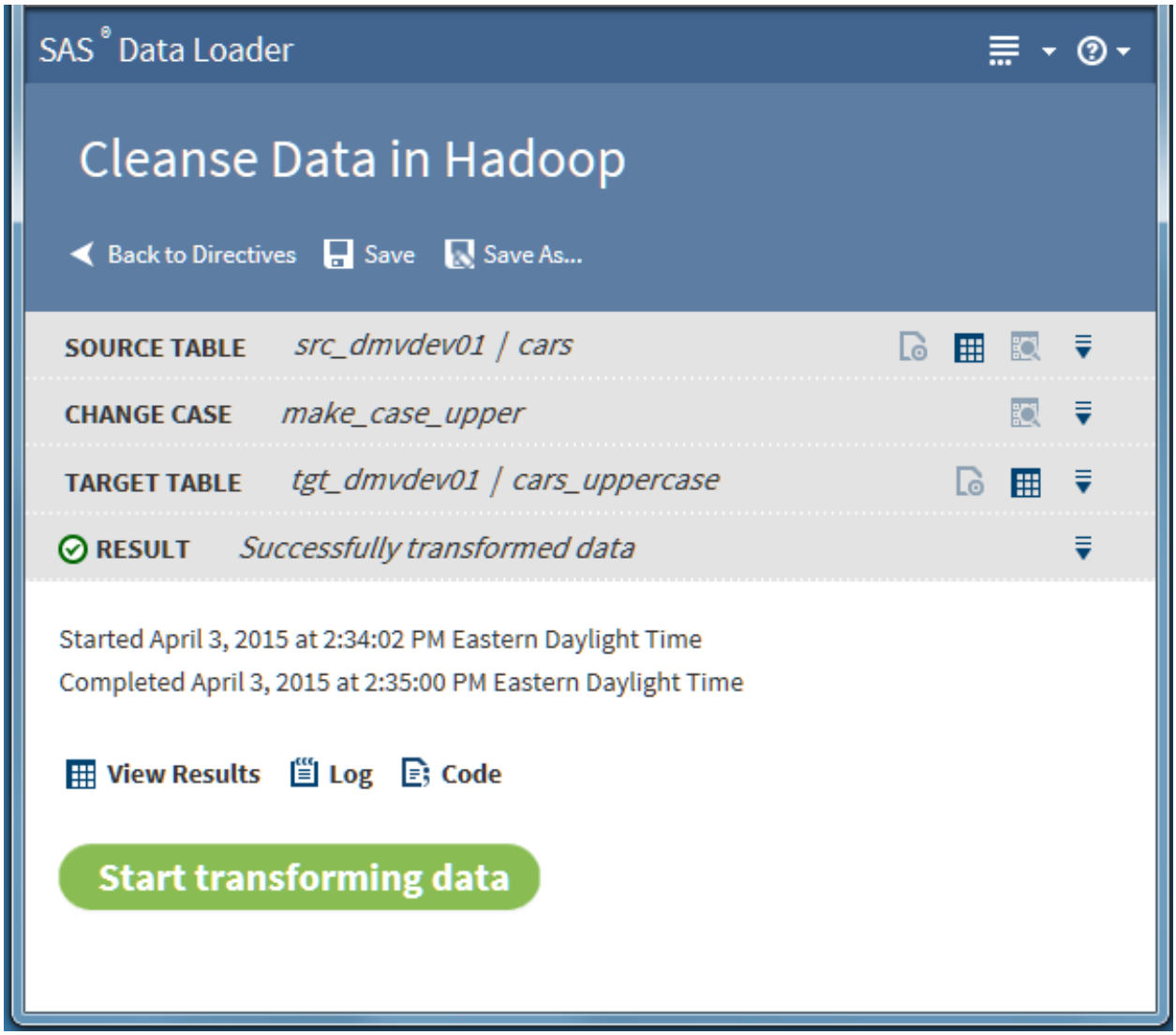
To create a new target table, click a data source, click **New Table** , and specify the table name in the New Table window. A new table of that name appears in the grid or list.

Note: To explore the contents of target tables, click a table and click **Data Sample** , Table Viewer , or View Profile (if available).

2 With a target table highlighted in the list or grid, click **Next**.

3 In the **Result** task, click **Save** or **Save As** to save your job, and then click **Start Transforming Data**.

4 When the job is complete, you can view the results, the log file, and the code that ran in Hadoop.



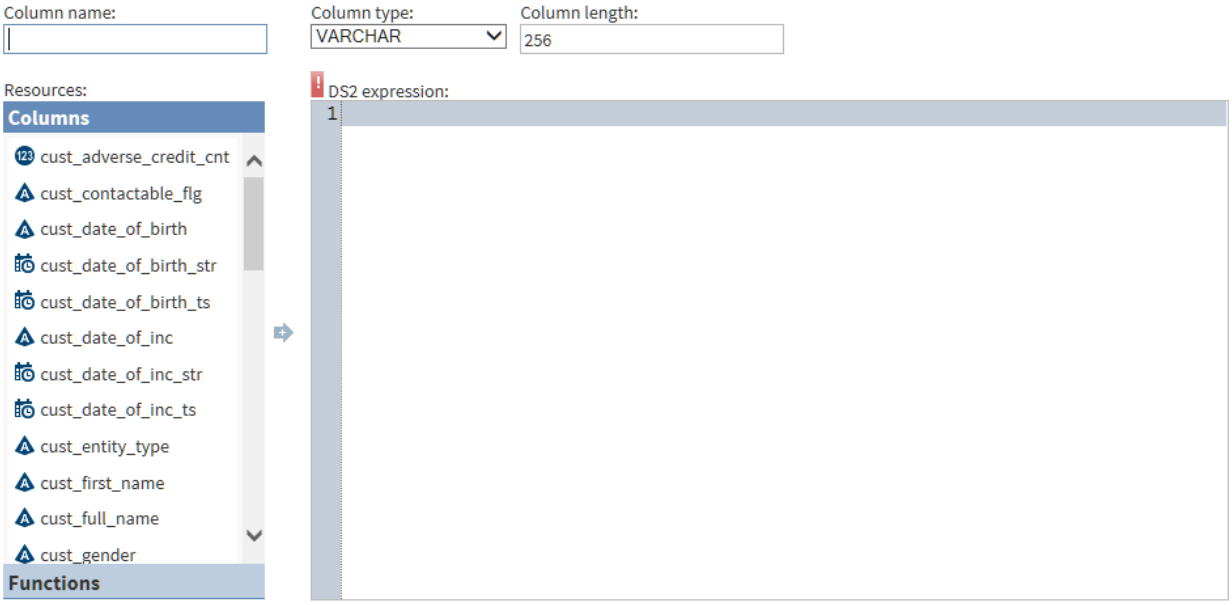
About DS2 Expressions and the Advanced Editor

In the Manage Columns transformation, you can add new columns and specify DS2 expressions for those columns. When you run your job, the DS2 expression is evaluated for each row and the result is added to the new column.

When you add a new column, you can enter or paste a DS2 expression directly into the **DS2 Expression** column (click +), or you can add your DS2 expression in the Advanced Editor (click). In either case, your expression uses DS2 expression syntax (and not SAS expression syntax.) For information about DS2 expressions, refer to the *SAS 9.4 DS2 Language Reference*.

Follow these steps to learn more about the Advanced Editor:

1 In the Manage Columns transformation, click to add a new column and open the Advanced Editor. Note that you can also select an existing column and click to replace the data in that column with the results of a DS2 expression.



2 Enter a name for the new column, a column data type, and the length of the column in bytes (if applicable.) The **Column type** is the data type of the result of your DS2 expression.

3 Define your DS2 expression using the columns and functions in the **Resources** list.

> **TIP** When you select a function, help is displayed for that function at the bottom of the **Resources** list.

4 When your DS2 expression is complete, click **Save** to return to the **Manage Columns** task. If you defined a new column for your DS2 expression, the new column appears at the bottom of the **Selected columns** list.

Delete Rows

Introduction

Delete Rows
Delete rows from a selected table. Requires Hive 14 or above.

Use the Delete Rows directive to delete data from a selected source table. Data is deleted in the source table itself rather than in a separate target table.

Prerequisites

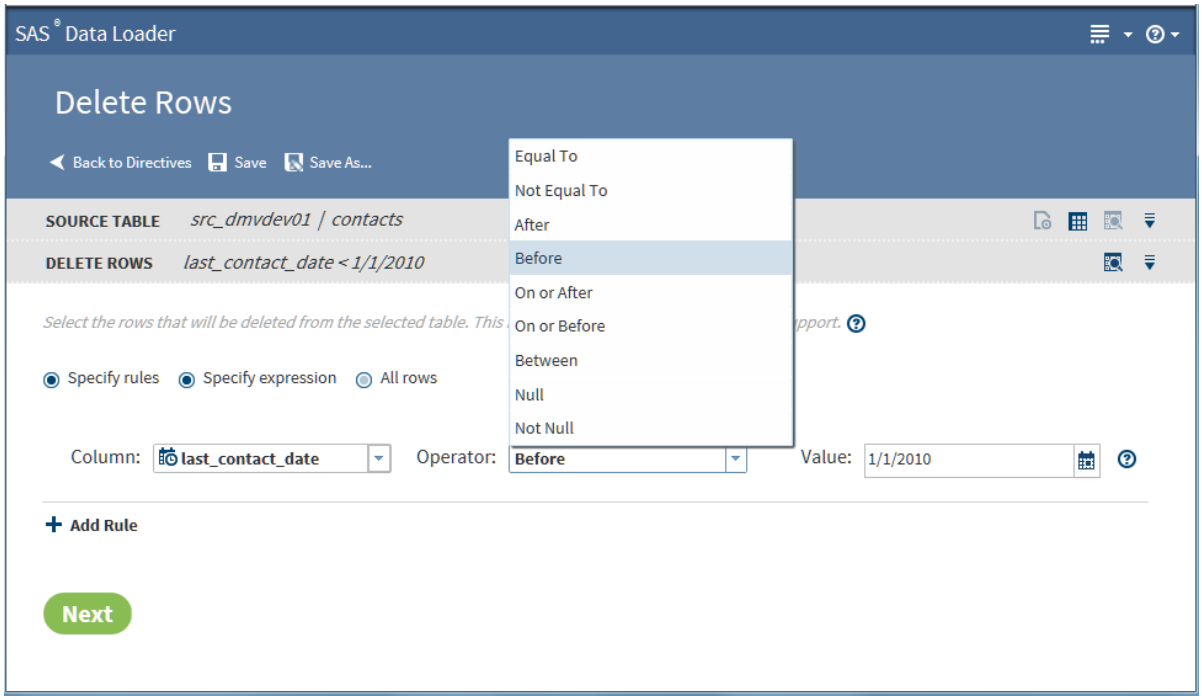
The prerequisites for the Delete Rows directive are defined as follows:

- The Hadoop cluster needs to be configured with release 0.14 or later of the Apache Hive data warehouse software. This release supports transactional tables.
- Source tables must use a Hive file format, preferably ORC (Optimized Row Columnar.)
- Source tables must be bucketed and partitioned. Bucketing clusters data based on the values in a specified (key) column. Partitioning creates individually accessible subsets of data based on the values in one or more source columns. To determine whether a source table has been bucketed and partitioned, contact your Hadoop administrator.

Example

Follow these steps to use the Delete Rows directive:

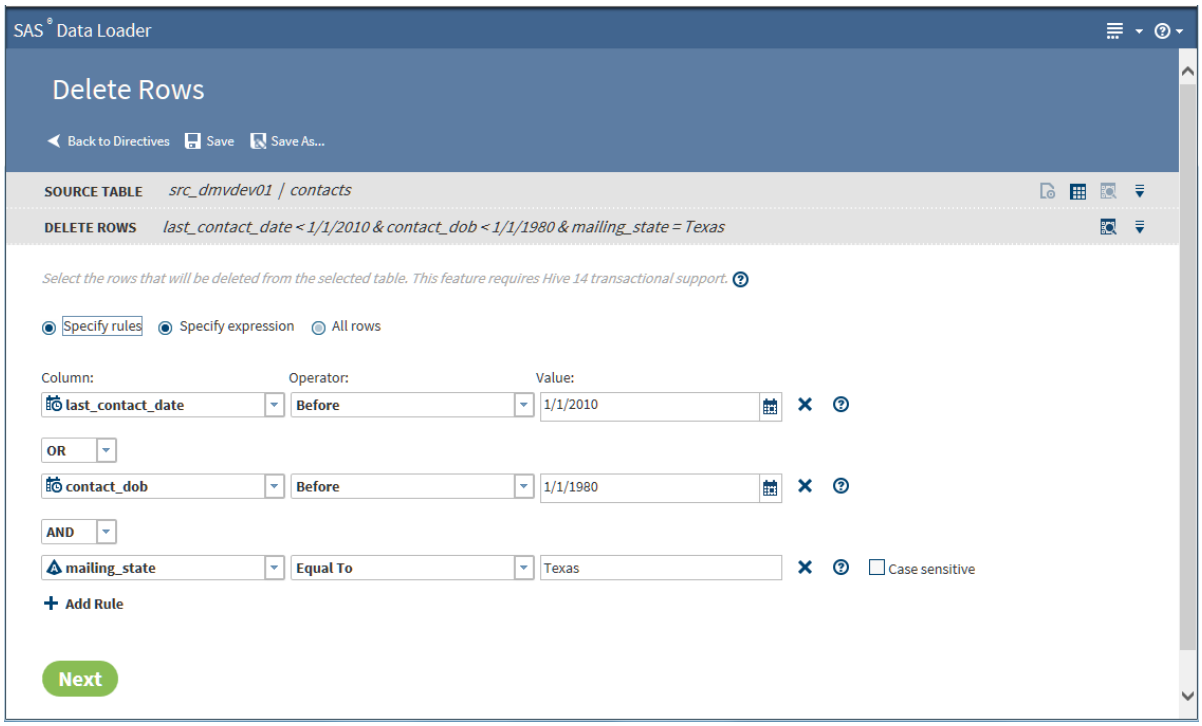
1. On the SAS Data Loader directives page, click **Delete Rows**.
2. In the **Source Table** task, select a data source and click **Next**, or click **Select Recent Table**. Refer to the prerequisites as needed.
3. In the **Delete Rows** task, choose one of the following:
   a. To delete all of the rows in the source table, click **All rows** and then click **Next**.
   b. To delete rows using one or more rules, click **Specify rules** and proceed to the next step. The Delete Rows job deletes rows when the specified rules are true. Multiple rules can be applied with logical AND and OR operators.
   c. To delete rows using a Hive expression, click **Specify expression** and go to Step 5 on page 46. Rows are deleted when the Hive expression returns true.
4. To delete rows by specifying one or more rules, follow these steps.
   a. Click **Select a column** and choose the source column that forms the basis of your rule.
   b. Click and select a logical **Operator**. The operators that are available depend on the type of the data in the source column. For example, the following image shows the operators that are available for the date/time data type:



c In the **Value** field, add the source column value that completes the expression. In the preceding example, the rule can be read as “Delete from the source table all rows with a last contact date prior to January 1, 2010.”

d Click **Add Rule** to add another rule. Select a different column, operator, and value.

e To delete rows when either the new rule or the preceding rule are true, change the **AND** condition to **OR**.

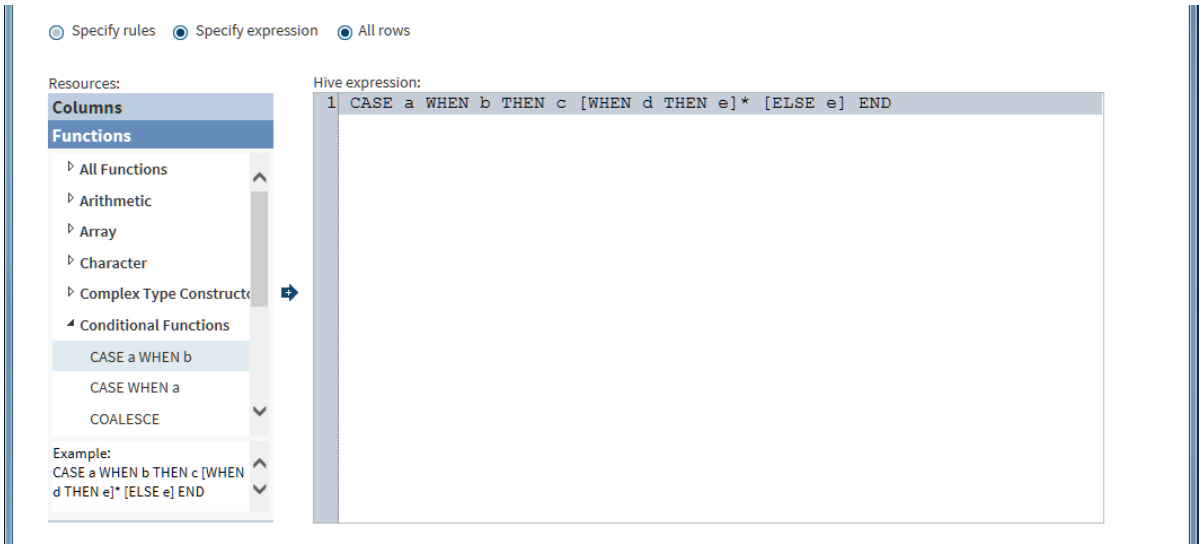


f When your rules are complete, click **Next** and go to Step 6 on page 46.

5 To delete rows using a Hive expression, follow these steps:

a In the **Hive expression** text box, either type or paste a Hive expression.

b To add Hive functions to your expression, click **Functions**, expand a category, select a function, and click .



To add column names to your expression, position the cursor in the **Hive expression** box, click **Columns** in the **Resources** box, click the source column, and then click .

6 When you have specified a rule or a Hive expression, click **Next**.

7 In the **Code** task, review the Hive code that will run in Hadoop. Click **Edit HiveQL Code** as needed.

Note: When you edit the Hive expression in the **Code** task, you will lose those edits if you then change the content of the **Delete Rows** task.

8 Click **Next** to open the **Result** task, and then click **Start deleting data**.

9 When the job is complete, click **Log** to confirm the deletion of rows.

Query or Join Data in Hadoop

Introduction

Query or Join Data in Hadoop
Query a table, or join data from multiple tables

Use queries to group rows based on the values in one or more columns and then summarize selected numeric columns. The summary data appears in new columns in the target table.

Use joins to combine source tables. The join is based on a comparison of values in “join-on” columns that are selected for each of the source tables. The result of the join depends on matching values in the join-on columns, and on the selected type of the join. Four types of joins are available: inner, left, right, and full.

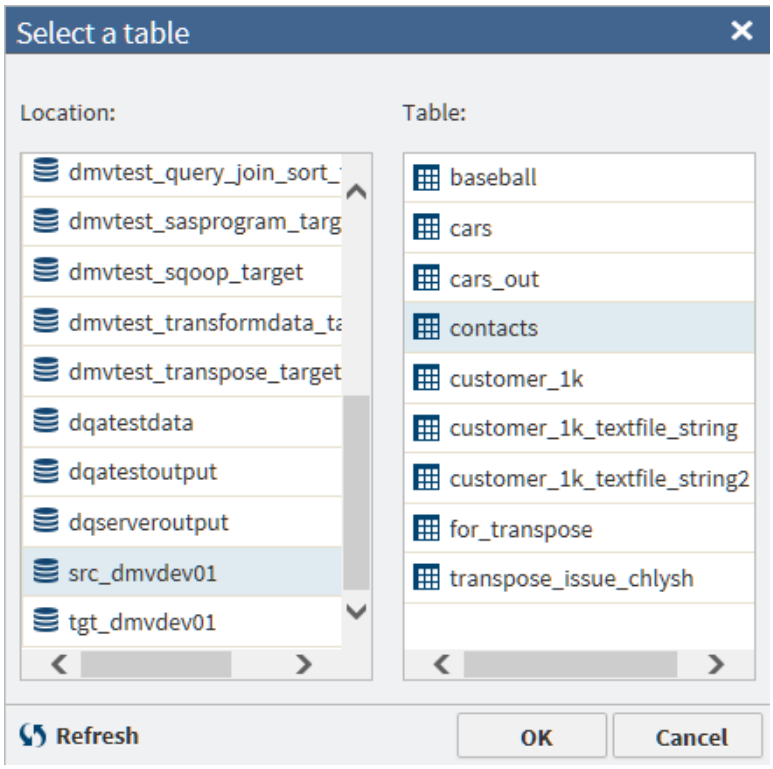
The Query or Join Data in Hadoop directive enables you to create jobs that combine multiple joins or queries. In the resulting table, you can remove unwanted rows and columns, remove duplicate rows, and rearrange columns. Before you execute the job, you can edit the generated Hive code and paste-in additional Hive code. The process of the directive is defined as follows:

- Select a source table.
- Join tables to the initial table as needed.
- Define queries that group columns and aggregate numeric values, again as needed.
- For jobs that do not include queries, use rules to filter unwanted rows from the target. (Queries require all rows.)
- For join-only jobs, select, arrange, and rename target columns.
- For join-only jobs, apply Hive SQL expressions in new or existing target columns.
- Sort target rows based on specified target columns.

Example

Follow these steps to use the Query or Join Data In Hadoop directive.

1. In the SAS Data Loader directives page, click **Query or Join Data in Hadoop**.
2. In the **Query** task, click the browse icon ...
3. In the Select a Table window, scroll through the **Location** list and click a schema. Then click a source table in the **Table** list, and then click **OK**.



4 If your job includes no joins, click **Next** to open the **Summarize Rows** task.

5 To join your source table with other tables, click **Add Join**, and then click **Next**.

6 In the **Join** row, click the browse icon ... and select the table for the join.

7 As needed, click the **Join** field and select a join type other than the default join type **Inner**.

Inner
: The inner join finds matching values in the join-on columns and writes one row to the target. The target row contains all columns from both source tables. A row from either source table is not written to the target if it contains a null value in the join-on column. A row is also not written to the target if the value in the join-on column does not match a value in the join-on column in the other source table.

Left
: The left or left-full join writes to the target all rows from the left table of the join statement. If a match does not exist between the join-on columns, null values are written to the target for the columns of the right table in the join.

Right
: The right or right-full join reverses the definition of the left join. All rows from the right table appear in the target. If no values match between the join-on columns, then null values are written to the target for the columns of the table on the left side of the join statement.

Full
: The full join combines the left and right joins. If a match exists between the join-on columns, then a single row is written to the target to represent those two source rows. If the left or right table has a value in the join-on

column that does not match, then the data for that row is written to the target. Null values are written into the columns from the other source table.

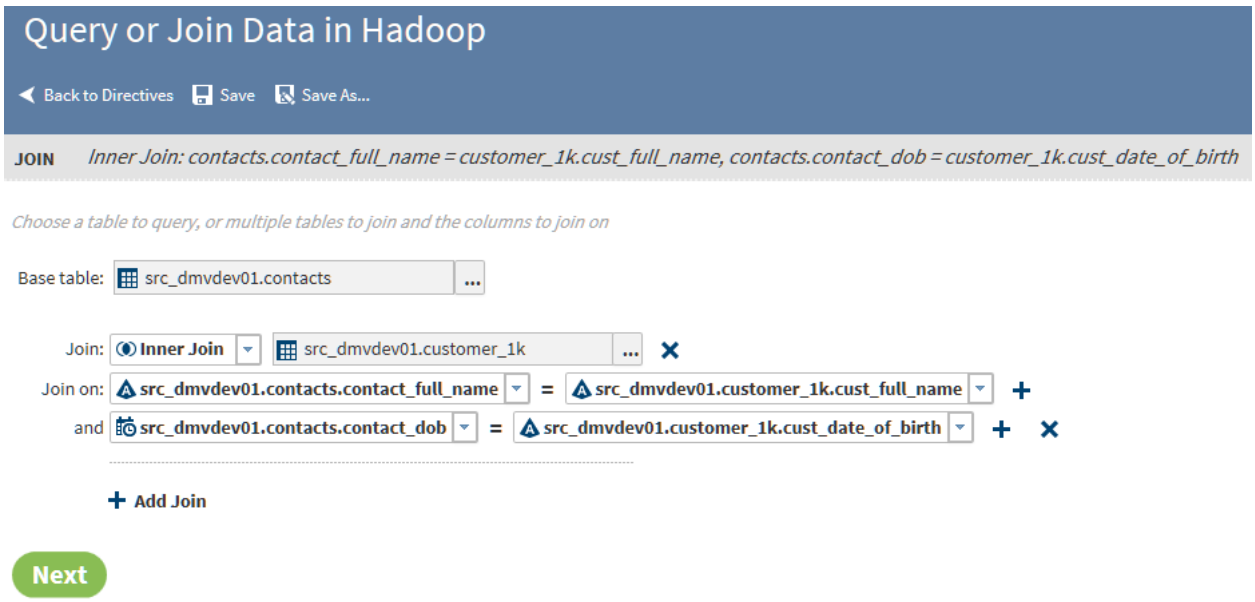
8 In the **Join-on** row, click the left join-on column and select a replacement for the default column, as needed.

Note: The left and right designations in the join-on statement define the output that is generated by the available left join and right join.

9 Click the right join-on column to select a replacement for the column, as needed.

10 To add more join columns, click the Add icon + at the end of the **Join-on** row. When you add a second pair of join-on columns, a match between the source tables consists of a match in the first pair of join-on values *and* a match between the second pair of join-on values.

11 To join a third table to the joined table that unites the two source tables, click **Add join**.



12 Click **Next** and wait a moment while the application assembles in memory the names of the joined columns.

13 In the **Summarize Rows** task, if you do not need to summarize, click **Next**.

Note: If your source data is in Hive 13 format or lower, the Summarize task will not handle special characters in column names. To resolve the issue, either rename the columns or move the source table into Hive 14 format.

14 To add summarizations, click the **Group rows by** field, and then click the column that you want to use as the primary grouping in your target table. For example, if you are querying a table of product sales data, then you could group rows by the product type column.

Note:

- If your job includes joins, note that the **Group rows by** list includes all columns from your source tables.
- If you intend to paste a Hive query into this directive, then you can click **Next** two times to display the **Code** task.

15 To subset the first group with a second group, and to generate a second set of aggregations, click **Add column**.

16 To generate multiple aggregations, you can add additional groups. The additional groups will appear in the target table as nested subgroups. Each group that you define will receive its own aggregations.

To add a group, click **Add Column**, and then repeat the previous step to select a different column than the first group. In a table of product sales data, you could choose a second group by selecting the column product_code.

17 In **Summarize columns**, select the first numeric column that you want to aggregate.

18 In **Aggregation**, select one of the following:

Count
: specifies the number of rows that contain values in each group.

Count Distinct
: specifies the number of rows that contain distinct (or unique) values in each group.

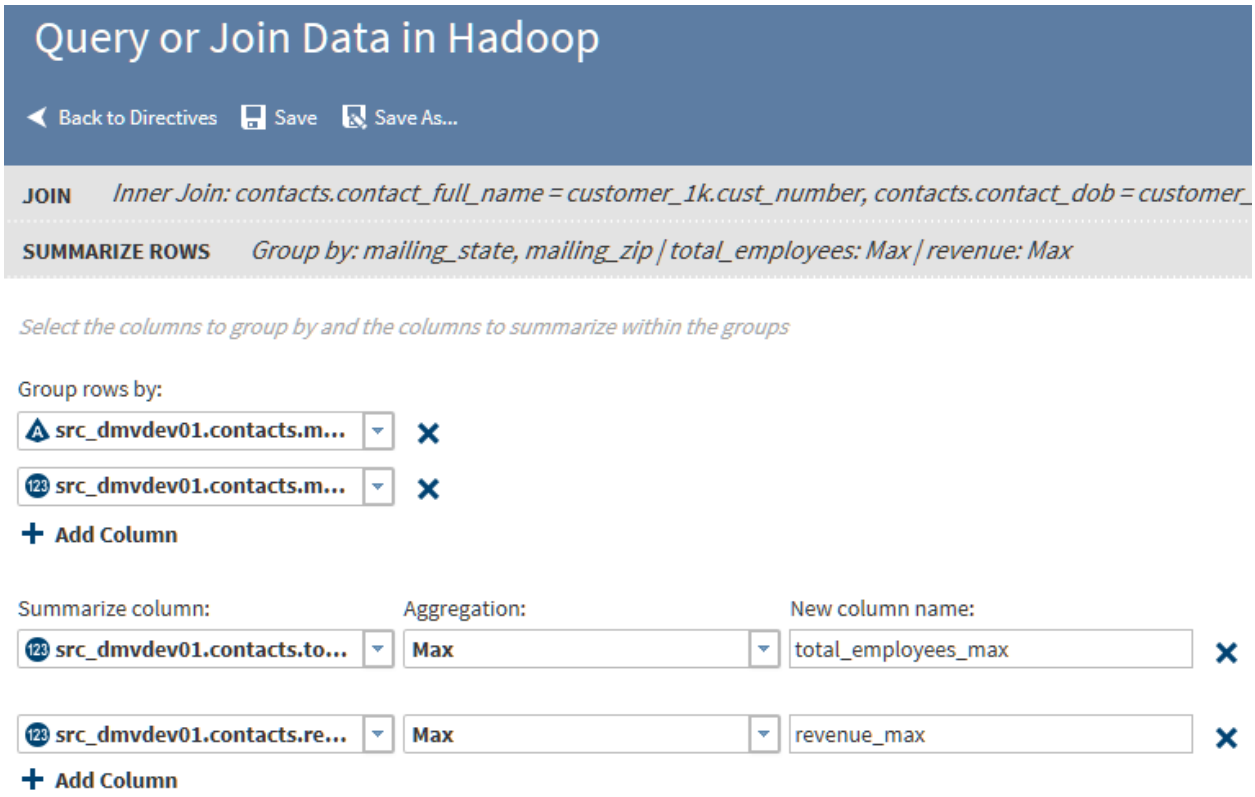
Max
: specifies the largest value in each group.

Min
: specifies the smallest value in each group.

Sum
: specifies the total of the values in each group.

19 In **New column name**, either accept the default name of the aggregation column, or click to specify a new name.

20 To add an aggregation, click **Add Column**.



21 When the aggregations are complete, click **Next**.

22 In the **Filter Data** task, all source rows are included in the target by default. To accept this default, click **Next**.

23 If your job includes joins but no summarizations, then you can select **No duplicate rows** to remove duplicate rows from the target. Older versions of Hive do not support the selection of both **No duplicate rows** and **All Rows**.

24 To filter rows from the target, choose one of the following:

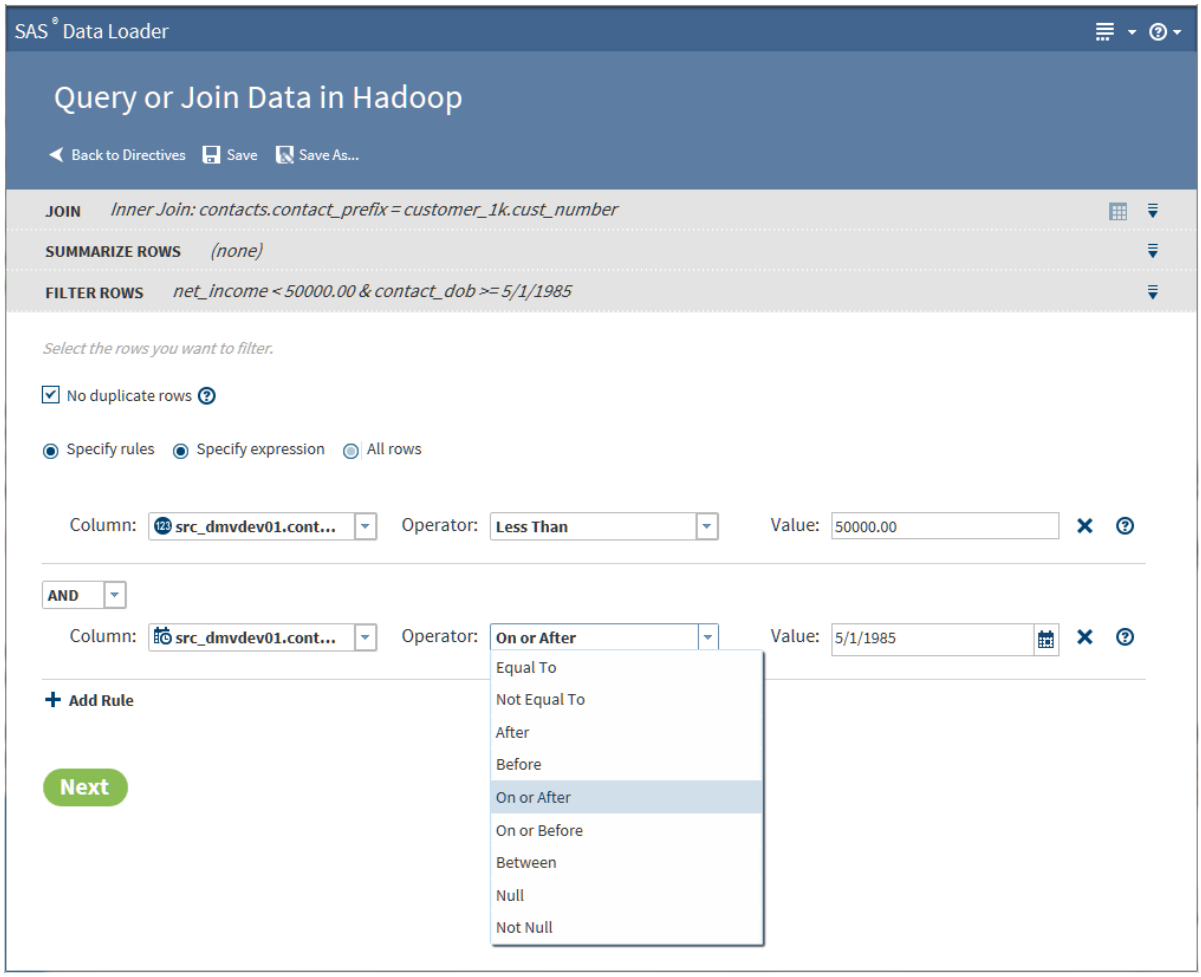
   a To filter rows using one or more rules, click **Specify rules** and proceed to the next step. You can specify multiple rules and apply them using logical AND and OR operators.

   b To filter rows using a Hive expression, click **Specify expression** and go to Step 26 on page 52.

25 To filter rows by specifying one or more rules, follow these steps:

   a Click **Select a column** and choose the source column that forms the basis of your rule.

   b Click and select a logical **Operator**. The operators that are available depend on the type of the data in the source column. For example, the following image shows the operators that are available for the date/time data type:



c In the **Value** field, add the source column value that completes the expression. In the preceding example, the two rules combine to read "Filter from the target all source rows with an income less than 50,000.00 and born on or after May 1, 1985."

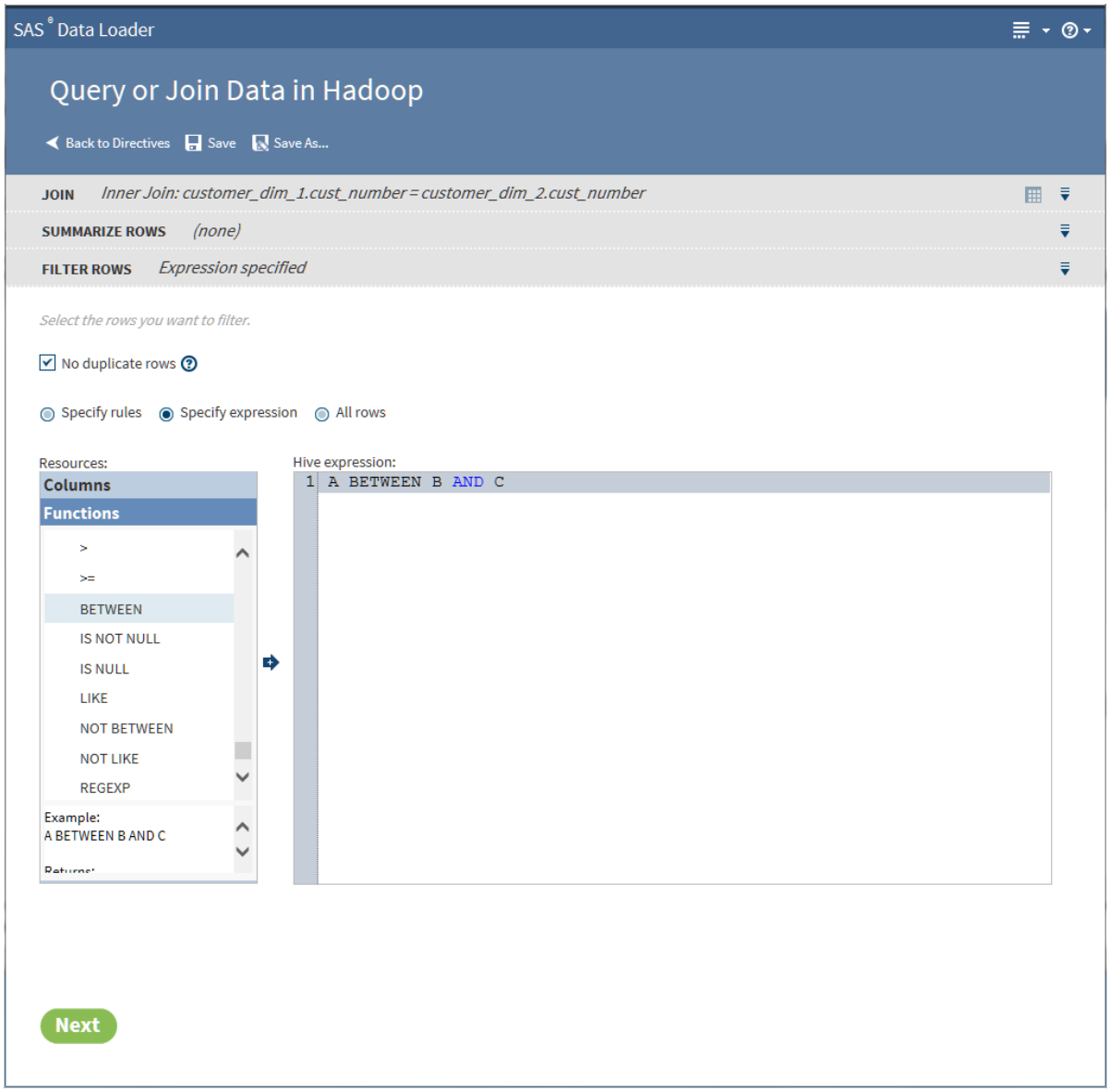
d Click **Add Rule** as needed to add another rule. Select a different column, operator, and value. To associate a new rule with the previous rules, either retain the default **AND** operator or click **AND** and select **OR**.

e When your rules are complete, go to Step 27 on page 53.

26 To filter rows using a Hive expression, follow these steps:

a In the **Hive expression** text box, either enter or paste a Hive expression.

b To add Hive functions to your expression, click **Functions** in the **Resources** box, expand a category, select a function, and click .



To add column names to your expression, position the cursor in the **Hive expression** box, click **Columns** in the **Resources** box, click a source column, and then click .

27 When your rules or expression are complete, click **Next** to open the **Columns** task.

28 Use the **Columns** task to select, order, and rename the columns that will be written into the target table. Also use the **Columns** task to select, order, and rename the columns that will be written into the target task to apply Hive expressions to new or existing columns.

Note: This **Columns** task is available only if your job *does not* contain summaries. If your job *does* contain summaries, then click **Next** to display the **Sort** task.

29 Use the Columns task to do the following:

- Select and order the columns in the target using these arrow icons (select all), (select one), or (remove one), and (remove all).

- Replace the suggested column names as needed by clicking in the **Target Name** column.
- Reorder columns by clicking a column and clicking (move to first column), (move column left one position), (move column right one position), and (move column to last position).
- Add new columns for Hive expressions. Click the **Add** icon , and then specify the column name.
- Add a new column and open the Advanced Editor to develop a Hive expression. Click .The expression uses the values in other columns to populate the new column.
- Edit an existing column by applying a Hive expression. Click .

30 Follow these steps to specify a Hive expression:

a In the **Column name** field, either enter a new column name or verify the name of edited column.

b Paste Hive code into the **Hive expression** box. Edit the text as needed. Add Hive functions and source column names using the **Resources** box.

31 Click **Next** to close the **Column name** task and open the **Target Table** task.

32 In the **Target Table** task, to learn about the contents of a table, click the table and click the **Table Viewer** icon .

33 To write your target data to an existing table, click that table and click **Next**. Any and all existing data is replaced.

34 To save data to a new target table, click New Table..., enter a table name in the New Table window, and click **OK**.

The names of tables must meet the naming conventions of SAS and Hadoop.

35 To display your target data as a temporary view, click Save as a View .

Saving as a view displays your target data in the Sample Data Viewer without saving the results to a table on disk.

When your target selection is complete, click **Next** to open the **Code** task

36 In the **Code** task, click **Edit HiveQL Code** to edit the generated code. Click **Reset Code** to restore the original generated code. Click **Next** to open the **Result** task.

Note: Edit your HiveQL code with care. The code in the editor is the exact code that will be executed by your job, regardless of previous selections.

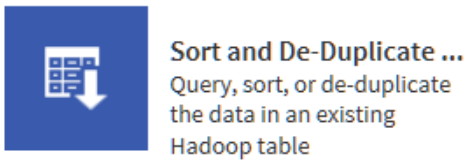
37 In the **Result** task, you can review the previous tasks by clicking on the gray taskbars at the top of the window.

38 Click **Save** or **Save As** to save your job.

39 Click **Start querying data** to execute your directive. To monitor the progress of your job, see the "Run Status"directive.

Sort and De-Duplicate Data in Hadoop

Introduction



Use the Sort and De-Duplicate Data in Hadoop directive to create jobs that include some or all of the following steps:

1 Group rows based on selected columns and then summarize numeric columns for each group and subgroup.

2 If not summarizing, specify the removal of duplicate rows and filter rows from the target.

3 Remove, reposition, and rename the columns in the target table. Add columns that receive the results of Hive expressions.

4 Sort target rows by selecting one or more columns for ascending or descending values.

Example

Follow these steps to use the Sort and De-Duplicate directive:

1 Open SAS Data Loader, as described in Chapter 2, “Getting Started,” on page 7.

2 In the **Source Table** task, click a table in your default data source and click **Next**. To select a source table from another data source, click

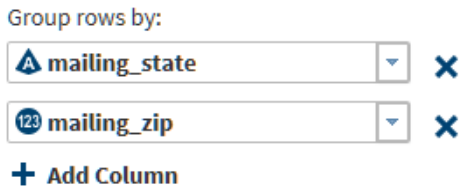
3 Use the **Summarize Rows** task to group rows in the target according to column values, and then summarize numeric values for each group or subgroup.

If you do not want to generate summary values for groups of rows, or if you want to remove duplicate rows, click **Next** to display the **Filter** task and go .

Note: If your source data is in Hive 13 format or lower, the Summarize task will not handle special characters in column names. To resolve the issue, either rename the columns or move the source table into Hive 14 format.

Follow these steps to use the **Summarize Rows** task:

a Click **Group rows by** and select a column. To generate nested groups with additional summary values, click **Add Column**.



b Click **Summarize column** and select a column that will be used to generate summary values for each group. The summarized values will appear in a new column in the target.

c Click **Aggregation** and select the summary type. The available summary types are defined as follows:

Count
: specifies the number of rows that contain values in each group.

Count Distinct
: specifies the number of rows that contain distinct (or unique) values in each group.

Max
: specifies the largest value in each group.

Min
: specifies the smallest value in each group.

Sum
: specifies the total of the values in each group.

d Click **New column name** to change the default column name for the new target column that will receive summarized data. The new target column will contain a summary value for each group and subgroup.

e Click **Add Column** to specify a second summary and target column.

f When your groups and summaries are complete, click **Next** to display the **Filter** task.

4 The **Filter** task enables you to remove duplicate rows and filter or remove unnecessary rows from the target.

If you specify summaries or if you do not need to filter rows from the target, then click **Next** to display the **Columns** task.

Follow these steps to use the **Filter** task:

a To remove from the target any rows that are identical to another row, click **No duplicate rows**.

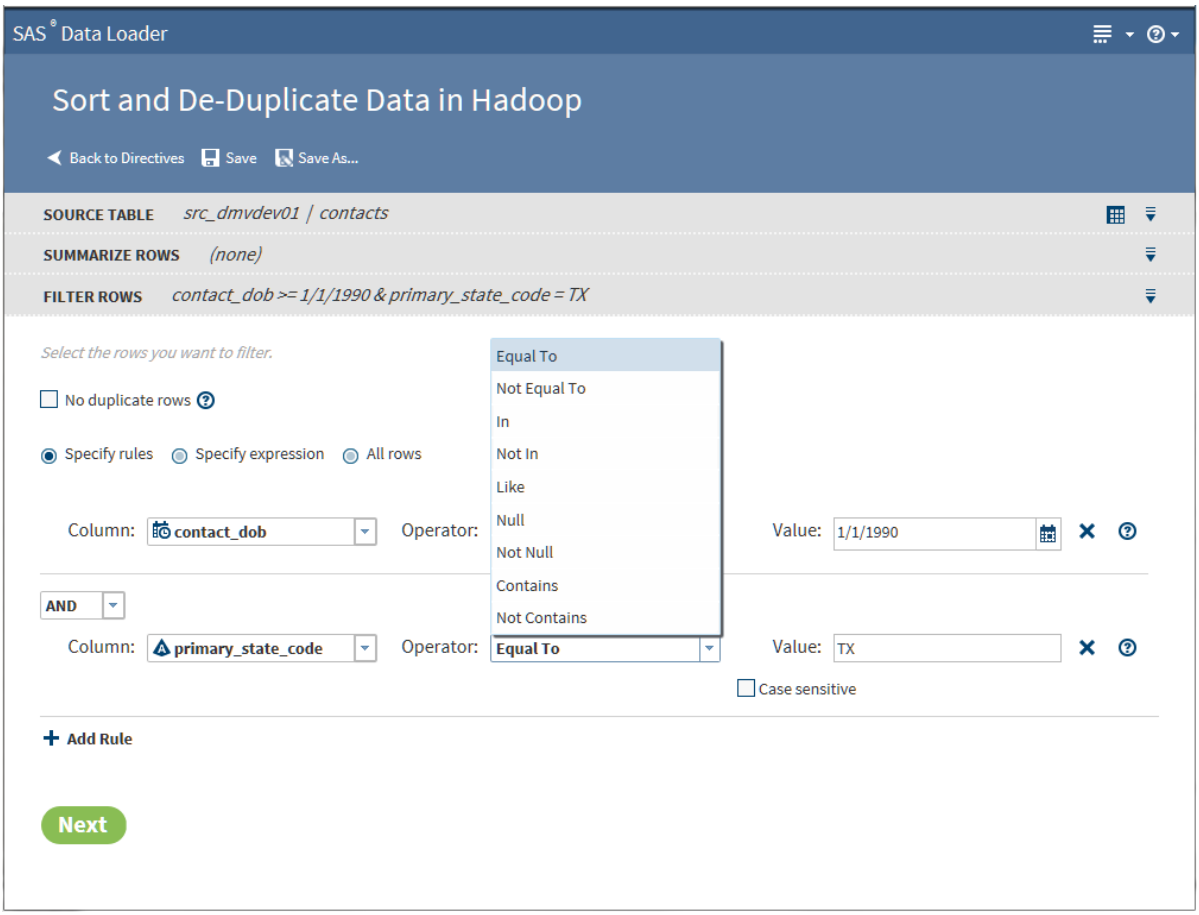
b To filter rows from the target using one or more rules, click **Specify rules**.

To filter rows using a Hive expression, click **Specify expression** and go to Step 4d .

c To filter rows by specifying one or more rules, follow these steps:

i Click **Column** and choose the source column that forms the basis of your rule.

ii Click and select a logical **Operator**. The operators that are available depend on the type of the data in the source column. For example, the following image shows the operators that are available for the character data type:



iii In the **Value** field, add the source column value that completes the expression. In the preceding example, the two rules combine to read “Filter from the target all contacts from Texas who were born on or after January 1, 1990.”

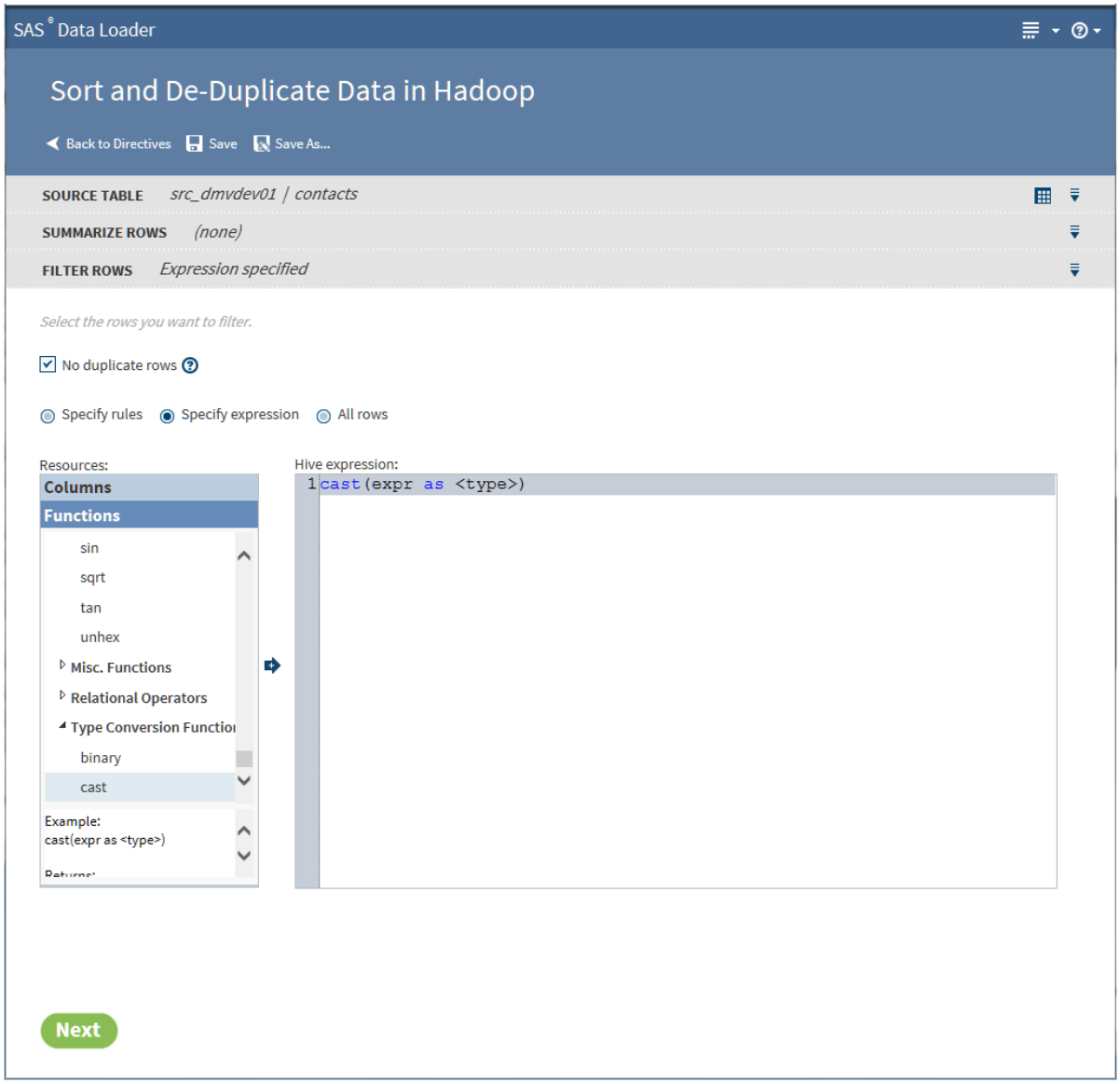
iv Click **Add Rule** as needed to add another rule. Select a different column, operator, and value. To associate a new rule with the previous rules, either retain the default **AND** operator or click **AND** and select **OR**.

v When your rules are complete, click **Next** to display the **Columns** task and go to Step 5.

d To filter rows using a user-written Hive expression, follow these steps:

i In the **Hive expression** text box, either enter or paste a Hive expression.

ii To add Hive functions to your expression, click **Functions** in the **Resources** box, expand a category, select a function, and click .



To add column names to your expression, position the cursor in the **Hive expression** box, click **Columns** in the **Resources** box, click a source column, and then click .

iii When your expression is complete, click **Next** to open the **Columns** task.

5 Use the **Columns** task to remove, reorder, or rename columns. You can also add columns that receive the results of user-written Hive expressions.

If you defined summaries in the **Summarize Rows** task, or if you do not need to modify columns, then click **Next** to display the **Sort** task and go to to Step 6.

Follow these steps to use the **Columns** task:

a Click **Specify Columns** to display the **Selected Columns** and **Available Columns**.

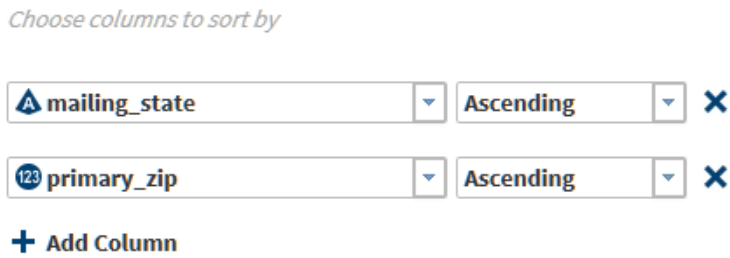
b In **Selected Columns**, click icons to perform the following tasks:

- To rename columns, click , or click and enter the new name in the **Target Name** column.
- To rearrange or reorder columns, click the up and down arrow icons. Note that the top row in **Selected Columns** is the leftmost column, or column 1, in the target table.
- To remove columns from the target, click the trash can icon or the left arrow icons.
- To add a new column and add data to that column using an existing Hive expression, click . In the **Hive expression** column, paste your Hive expression.
- To add new columns that contain a Hive expression, and to develop that expression using the Advanced Editor, click . The Advanced Editor helps you browse and select Hive functions and column names for your expression. To learn how to use the Advanced Editor, see "Using the Advanced Editor for Hive Expressions".
- To edit with the Advanced Editor columns that contain a Hive expression, click .

c When your columns are complete, click **Next** to display the **Sort** task.

6 If you have not defined any summaries in the **Summarize Rows** task, then the **Sort** task enables you to group rows based on ascending or descending values.

If you defined summaries, click **Next** to display the **Target** task and go to the next step.



7 In the **Target Table** task, select a location for the target table. When the table list appears, either select an existing target or click **New Table**. To generate a temporary table that is not saved to disk, select **Save as a View**. Click **Next**.

8 In the **Code** task, review and edit the generated Hive code. Note that if you edit the code, you will lose your edits if you change a task and regenerate code. Click **Next**.

9 In the **Result** task, click **Save** or **Save As** to save your job. You can then access that job in Saved Directives. Click **Start querying data** to run your job.

Using the Advanced Editor for Hive Expressions

In the directive Sort and De-Duplicate Data in Hadoop, in the **Columns** task, you use the Advanced Editor to add or edit user-written Hive expressions. The expressions are run by the Sort and De-Duplicate job to add data to new target columns. The Advanced Editor enables you to insert column names and Hive function syntax into your expressions.

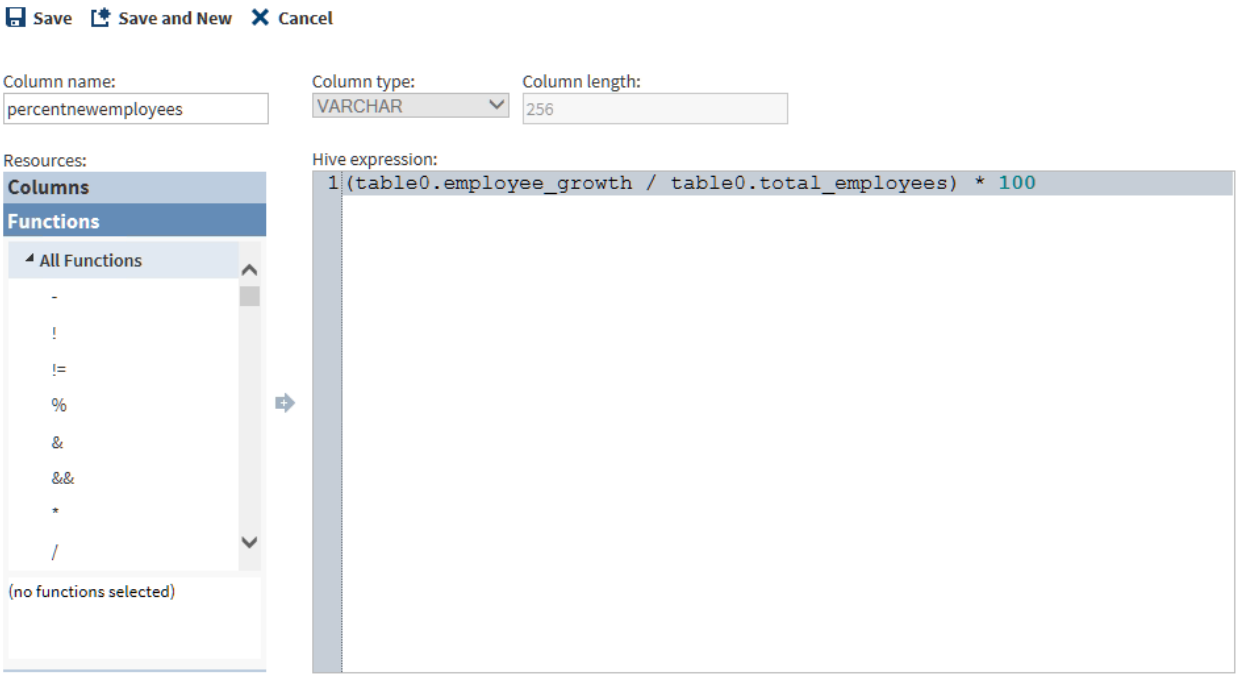
Follow these steps to use the Advanced Editor:

1 As needed in the **Columns** task, click or to open the Advanced Editor.

2 In the **Advanced Editor**, in the **Column Name** field, enter a name for a new column or rename an existing column. The fields **Column type** and **Column length** describe the selected column.

3 To build an expression, you can start by pasting Hive code from your clipboard. To edit or build your expression, click the column names and functions in the **Resources** box.

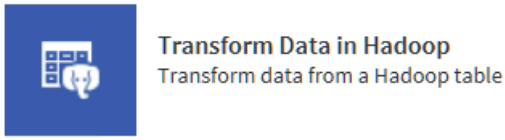
> **TIP** When you select a function, syntax help is displayed at the bottom of the **Resources** box.



4 To save your expression and return to the **Columns** task, click **Save**. To save and create another new column and expression, click **Save and New**. In the **Columns** task, new columns are displayed at the bottom of the **Selected Columns** box.

Transform Data in Hadoop

Introduction



Use the Transform Data in Hadoop directive to filter data, manage columns, and summarize data in one or more Hadoop source tables.

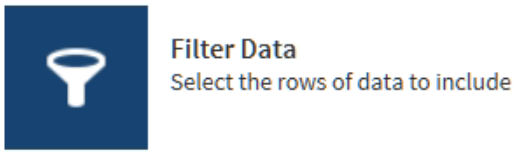
Example

The following example depicts the process of creating and running a directive that contains several transformations. The example opens a source table of customer information, selects columns for the target, and applies two filters.

1. In the SAS Data Loader directives page, click **Transform Data in Hadoop**.
2. In the **Source Table** task, click the schema that contains the source table that you will transform. When the tables appear, select the source table and click **Next**.
3. In the **Transformation** task, click a transformation:
   - Click **Filter Data** to define rules that include only desired data in the target.
   - Click **Manage Columns** to manage the columns in your target table. You can select source columns, reorder columns, and rename columns. You can also add or repurpose target columns to store the results of DS2 expressions. An advanced editor is provided to assist with the development of DS2 expressions.

     Note: To apply HiveQL expressions rather than DS2 expressions, see the Manage Columns transformation in the Query or Join Data in Hadoop directive.
   - Click **Summarize Columns** to group rows based on the values in one or more columns. For each group, you can generate summary aggregations from selected numeric columns.

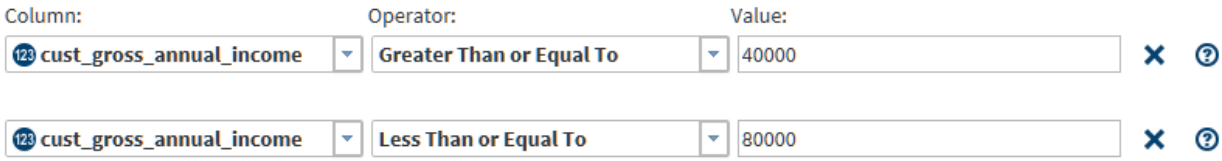
   Your job can consist of one or more transformations. Multiple transformations are executed in the order in which you define them. A logical order for all three transformations is filter data, manage columns, and summarize columns.
4. Click **Filter Data**.



5 Select the columns that you will use to filter the rows that will be written into the target table. For example, in a table of customer information, you could limit the data in your target to customers with incomes between $40,000 and $80,000. This filter requires two rules, and both rules must be true in order for the source row to be written to the target.

6 In the **Filter Data** task, accept the default value for the **Include** field: **Rows for which all of these rules apply**.

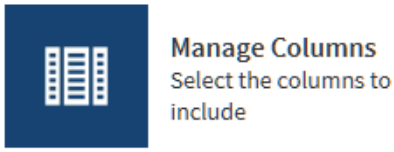
7 Select columns, operators, and values to define rules.



The operators that are available depend on the type of the column. To learn about available operators, see "About the Operators in the Filter Data Transformation" on page 64.

8 At this point, you could end a job that consists solely of a Filter Data transformation. You would click **Next** to select a target table and run your job. Instead, to see the other two available transformations, click **Add Another Transformation**.

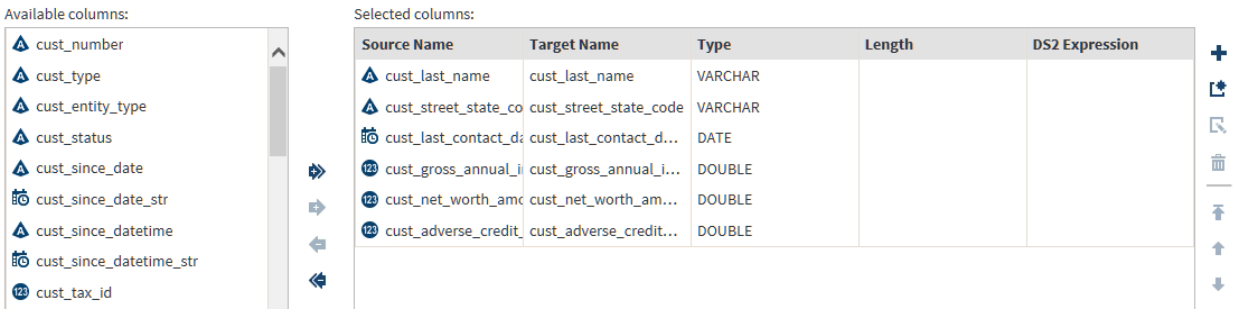
9 In the **Transformation** task, click **Manage Columns**.



10 Determine the columns that you want to see in your target table. In a table of customer data, you could choose columns for full name, gross annual income, net worth, number of adverse credit events, and state code. These columns include those that will be used in a Summarize transformation.

11 In the **Manage Columns** task, use the left and right arrow icons to click and move columns into and out of the **Selected Columns** list. Columns are listed vertically, with the first or leftmost column at the top, and the last or rightmost column at the bottom.

12 Use the vertical arrow icons to change the position of the columns.

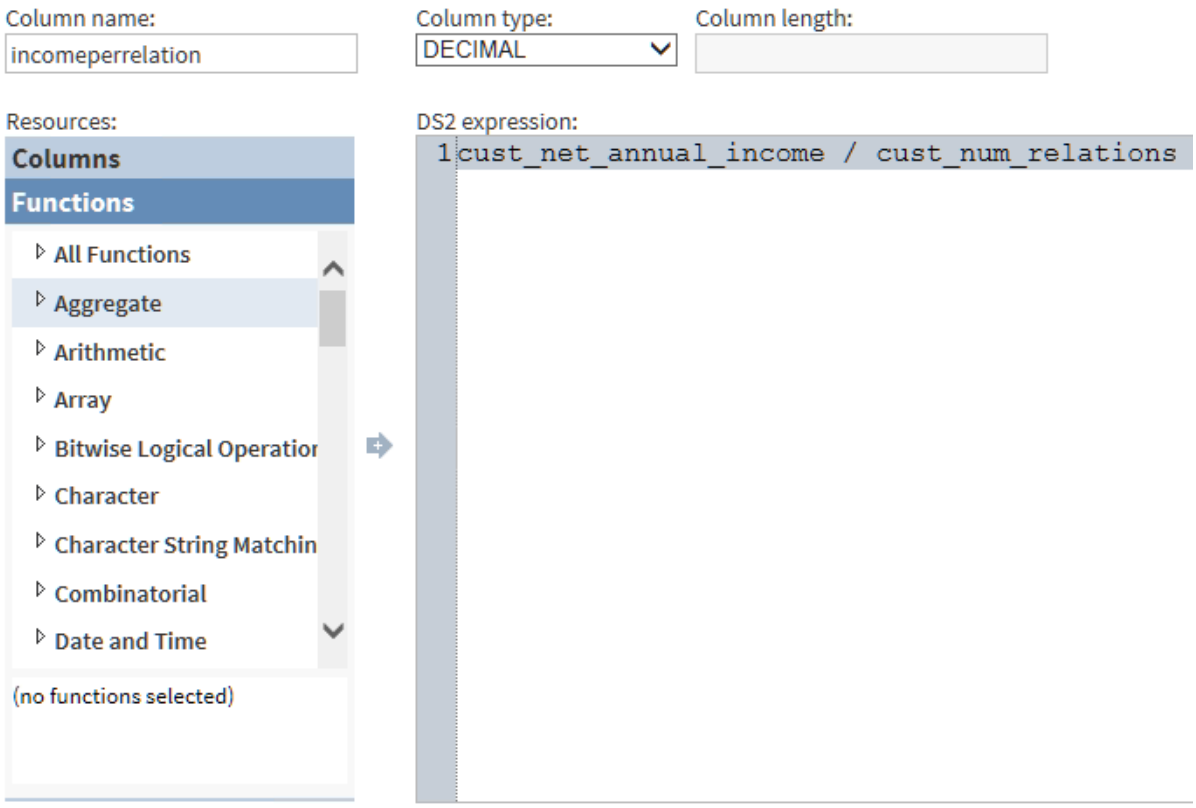


13 To rename columns, click and enter or paste the new name in **Table Name**.

14 To replace existing column data with data that is generated by a DS2 expression, click a selected column and click the **DS2 Expression** column. Enter or paste the DS2 expression.

15 To add a new column, and to use the Advanced Editor to generate a DS2 expression for that column, click .

16 To use the Advanced Editor, enter a **Column Name**, and then apply DS2 functions to specified target columns. When you select a column, syntax help appears at the bottom of **Resources**. When your DS2 expression is complete, select **Save** or **Save New** to return to the Manage Columns transformation. The new column appears at the bottom of **Selected Columns**.



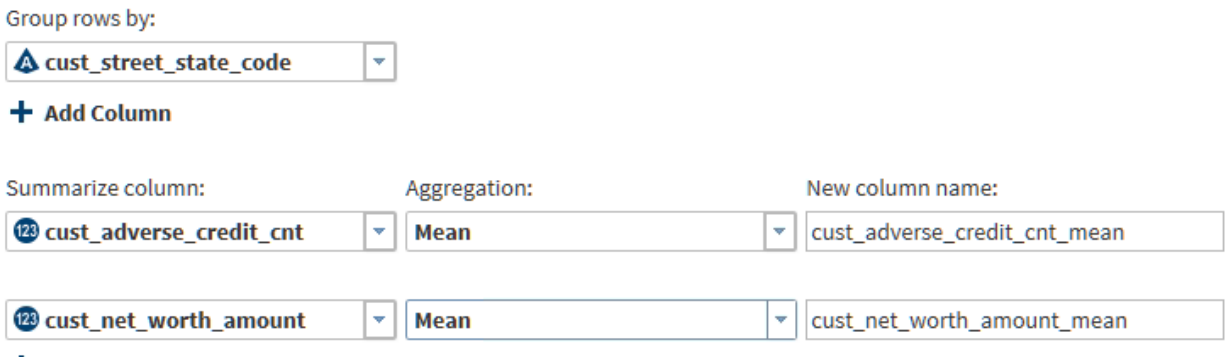
17 Click **Add a new transformation**, and then, in the **Transformation** task, click **Summarize**.

18 In the **Summarize Rows** task, click **Group rows by** to specify a column whose values will be used to group rows. You can specify additional columns that will form subgroups. Each group and subgroup will receive a value in each aggregation column.

Note: If your source data is in Hive 13 format or lower, the Summarize Rows task will not handle special characters in column names. To resolve the issue, either rename the columns or move the source table into Hive 14 format.

19 Click **Select a column** to specify a summarization, and then click and select an aggregation. to learn about the available aggregations, see "About the Aggregations in the Summarize Rows Transformation" on page 69.

20 Click **New column name** and enter or paste replacement names for the aggregation columns.



21 When your summaries are complete, click **Next** to conclude your job.

22 In the **Target Table** task, select the schema that contains or will contain your target table.

23 Click New Table... to create a new table, or click an existing table that will be overwritten by your job.

> **TIP** If you select a table and View Profile View Profile is enabled, you can click that icon to display a profile report for that table.

24 Click **Next** to display the **Result** task. In the Result page, click **Save** or **Save As** to store your job in your shared folder. If you want to run your job now, click **Start transforming data**. Otherwise, you can run your job later from "Saved Directives".

About the Operators in the Filter Data Transformation

The following table describes filter operators by the data type of the selected column.

Table 4.1 *Logical Operators in the Filter Transformation*

Operator	Source ColumnData Types	Description and Example

Equal To	The Equal To operator is available for use with all source data types, which include the following:
	Character
	Numeric
	Datetime
		The source value is accepted and its row is written to the target table only when the source value exactly matches the comparator.
		Character values can be case-sensitive. Blank spaces are included in the comparison.
		Datetime values in the comparator use the SAS format DATETIME(w.p).
		`Gender Equal To Male`
		`PrefCustomer Equal To 1`
		`SaleDate Equal To 5/1/2014`

Not Equal To		Accepts the source row when the column value is anything other than the comparator.
		`Region Not Equal To Europe`
		`NumChildren Not Equal To 0`
		`SaleDate Not Equal To 11/25/2013`

Null		Accepts the source row when the column value is NULL or if no source value is present.
		`CreditScore Null`
		`AnnualIncome Null`

Not Null		Accepts the source row when the column value is present and when the value is not NULL.
		`PostalCode Not Null`
		`PhoneNumber Not Null`

Operator	Source Column Data Types	Description and Example

In	A 123	Accepts the source row when the column value is included in its entirety within the comparator. The comparator consists of a list of constant values. The list consists of a vertical list of individual entries, without commas. Blank spaces are interpreted literally. Case sensitivity can be enabled.
		`CarManuf In`
		`BMW`
		`VW`
		`Benz`
		`WaistSize In`
		`32`
		`34`
		`36`
		`38`

Not In	A 123	Accepts the source row when the column value is not included anywhere within the comparator's list of constant values.
		`City Not In New York`
		`Chicago`
		`Los Angeles`
		`WaistSize Not In 32`
		`34`
		`36`
		`38`

Operator	Source Column Data Types	Description and Example

Like	A 123	Accepts the source row when the column value matches the result of an expression in the comparator. The source value and the comparator are compared on a character-by-character basis. For character columns, case sensitivity can be enabled.
		Use the pattern-matching character `%` to indicate any string of characters. Use the underscore character _ to indicate any single character in that position.
		Note that trailing blank characters are written to the target table when using `%` at the end of the comparator.
		Use the word `escape` to include literal instances of `%` and _ in the comparator.
		`SalesRegion Like NorthAmer%`
		`AnnualSales Like 199_`
		`CustSatisfaction Like 100 escape %`

Not Like	A 123	Accepts the source row when the column value does not match the result of an expression in the comparator. The source value and the comparator are compared on a character-by-character basis. For character columns, case sensitivity can be enabled. Pattern-matching characters `%` and _ and `escape` are valid as described for the Like operator.
		`Sports Not Like %ball`
		`FootballFieldLength Not Like 100%`

Contains	A 123	Accepts the source row when the column value is found within the character string of the comparator. Case sensitivity can be enabled.
		`Address Contains IL`
		`LicenseNumber Contains 7227`

Not Contains	A 123	Accepts the source row when the column value is not found within the character string of the comparator. Case sensitivity can be enabled.
		`Month Not Contains OctNovDec`
		`SalesMonthly Not Contains 0`

Operator	Source ColumnData Types	Description and Example

Between	123	Accepts the source row when the column value or date is between the two values or dates in the comparator, but is not equal to either.
		`GradeAverage Between 87.5 93`
		`DailySales Between December 20, 2014 December 27, 2014`

Greater Than	123	Accepts the source row when the column value is greater than the value of the comparator. `AnnualSales GreaterThan 100000`
Greater Than Or Equal To	123	Accepts the source row when the column value is equal to the comparator or greater than the comparator. `CarsInFamily Greater Than or Equal To 3`
Less Than	123	Accepts the source row when the column value is less than the value of the comparator. `GamerAge Less Than 30`
Less Than Or Equal To	123	Accepts the source row when the column value is equal to the value of the comparator, or less than the value of the comparator. `SalesYear Less Than Or Equal To 2010`
After		Accepts the source row when the column date is later than the date in the comparator. `HomePurchaseDate After January 1, 2013`
Before		Accepts the source row when the column date is earlier than the date in the comparator. `BirthDate Before March 17, 1980`
On Or After		Accepts the source row when the column date is later than, or the same date as, the date in the comparator. `DailySales On Or After January 1, 2014`

Operator	Source ColumnData Types	Description and Example
On Or Before		Accepts the source row when the column date is earlier than, or the same date as, the date in the comparator. `DailySales On Or Before December 31, 2013`

About the Aggregations in the Summarize Rows Transformation

The aggregations that are available in the Summarize Rows transformation are defined as follows:

Count
: the number of rows in the group that contain valid values.

Count Distinct
: the number of unique values in the column for each group.

Corrected Sum of Squares
: measures variability or dispersion around the mean. To learn more about this (and other) statistical summaries, see the *Introduction to Statistical Modeling with SAS/STAT Software*.

Covariance
: measures the strength of the correlation of the values in the group. A positive value indicates that values move in the same direction within the group. A negative value indicates that values move in opposite or random directions.

Max
: the maximum value in the column for each group.

Mean
: the calculated center value between the maximum and minimum values in the group.

Min
: the minimum value in the group.

Number of Missing Values
: the number of rows in the group that contain a blank or NULL value.

Range
: the difference between the lowest and highest values in the group.

Standard Deviation
: measures the degree of variance, or the degree in which the values in the group deviate from the mean. A small value indicates little deviation. The standard deviation is the square root of the Variance.

Standard Error
: measures the applicability or accuracy of the mean as it applies to the values in the group. A small value indicates that the mean is a more accurate reflection of the values in the group.

Sum
: adds the values in the group.

Variance
: the average of the squared differences from the mean, which measure diversity in the group

Usage Notes

If necessary, you can change the maximum length of character columns for input tables to this directive. For more information, see "Change the Maximum Length for SAS Character Columns" on page 147.

Transpose Data in Hadoop

Introduction

Transpose Data in Hadoop
Transpose data from a Hadoop table

Use the Transpose Data in Hadoop directive to transpose one or more columns in a source table into rows in a target table. The columns in the target are the values of a specified column in the source. For example, you could specify that the columns of the target be taken from the values of a source table column that contains customer ID numbers. Each unique customer ID value in the source becomes a separate column in the target.

You do not have to transpose all of the columns in the source. You can select source columns that will be copied directly to the target.

This directive contains embedded help that includes examples of transposed data.

CAUTION! Selecting columns with a high degree of cardinality (number of unique values) can decrease performance in Transpose jobs. To maximize performance, profile your source columns and filter your source rows. You can filter source rows in the Cleanse Data in Hadoop or Query or Join Tables in Hadoop directives.

Example

Follow these steps to use the Transpose Data in Hadoop directive.

1. In the **Source Data** task, click the data source that contains your source table, click the source table, and then click the Table Viewer .

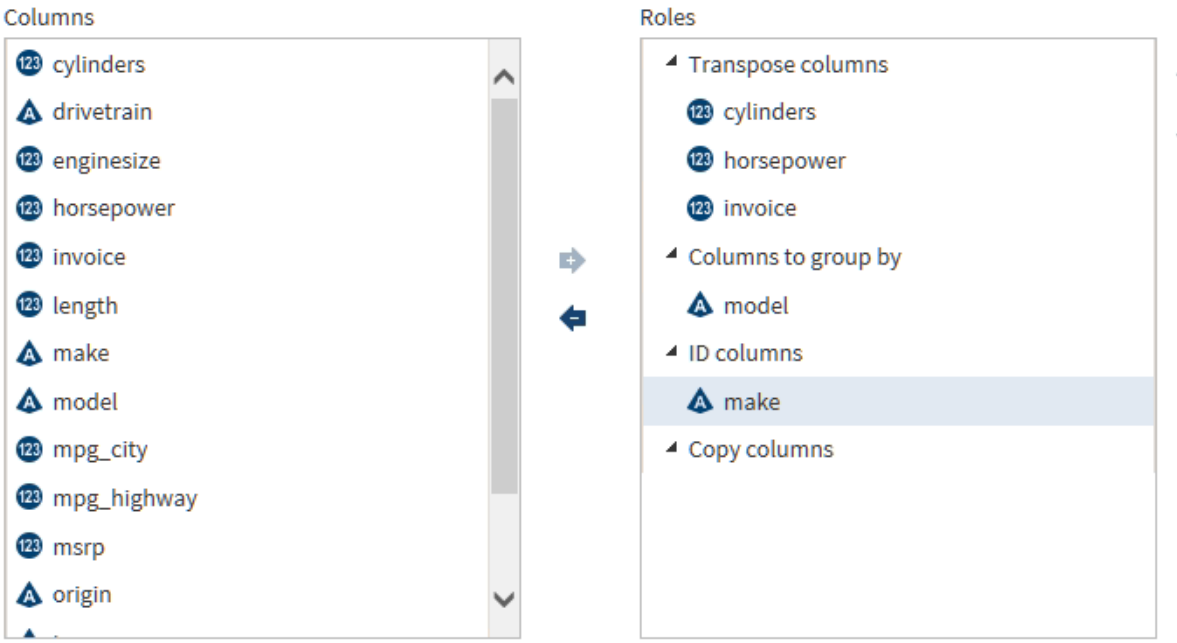
   Examine the source table to determine the roles for the columns.

   Note: Valid source table selections must have names that contain no more than 32 characters. Longer table names cause transpose jobs to fail.

2 In the **Transpose Data** task, click the required **Transpose data**, click the columns that you want to see as rows, and click the right arrow. If you transpose multiple columns, then you can arrange them in **Roles** using the up and down arrows.

3 Click the required **Columns to group by**, click an available column, and then click the right arrow. The group-by column becomes the leftmost column. Each row in that column receives a set of values from the transposed columns.

4 As needed, click **ID column**, click an available column, and then click the right arrow. The values of the ID column become column names in the target.



5 To copy a column from the source to the target, select **Copy column**, select an available column, and click the right arrow. The copied column will be positioned as the last, or rightmost, column.

Usage Notes

If necessary, you can change the maximum length of character columns for input tables to this directive. For more information, see "Change the Maximum Length for SAS Character Columns" on page 147.

5

Profile Data in Hadoop

Overview of Profile Directives 73

Profile Data 75
Introduction 75
Table Name Length Requirement 75
Create a Profile 75

Saved Profile Reports 80
Introduction 80
About Profile Reports 80
Open Saved Profile Reports 82

Overview of Profile Directives

Data profiling jobs help you assess the composition, organization, and quality of Hadoop tables. They help you recognize patterns, identify scarcity in the data, and calculate frequency and basic statistics. Data profiling can also aid in identifying both redundant data across tables and cross-column dependencies. All of these tasks are critical to optimal planning and monitoring.

The profile directives enable you to generate and view reports for one or more Hadoop tables. The reports display sample data, column information, and measurements of data quality. You create profile reports with the Profile Data directive and use the Saved Profile Reports directive to access and manage profile reports.

Here is an example of a profile report:

SAS 8 Data Loader - Profile Reports

contacts_report

Go to Profile Report List | Show Outline | Show Trends | Show Notes | Add Note... | Report Version: May 13, 2015, 2:09:00 PM

sgfdemo.contacts_cleanse

Count: 48

Data Quality Metrics

Column	#	Unique (n)	Unique (%)	Pattern (n)	Pattern (%)	Null (n)	Null (%)	Blank (n)	Blank (%)
contact_full_...	1	48	100	40	83	0	0	0	0
contact_first_...	2	38	79	9	19	0	0	0	0
contact_midd...	3	18	38	3	6	0	0	10	21
contact_last_...	4	48	100	10	21	0	0	0	0
contact_dob	5	45	100	*	*	3	6	*	*
company_name	6	32	67	8	17	0	0	16	33
primary_addr...	7	48	100	47	98	0	0	0	0
primary_city	8	45	94	42	88	0	0	0	0
primary_state...	9	37	77	15	31	0	0	0	0
primary_zip	10	5	100	*	*	43	90	*	*
primary_coun...	11	45	94	5	10	0	0	1	2
phone_number	12	42	95	5	10	4	8	2	4
fax_number	13	0	0	0	0	48	100	0	0

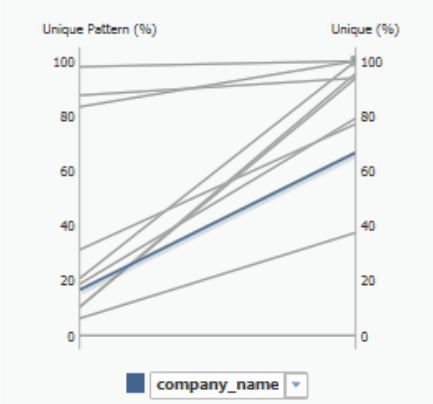
* indicates data not available or not applicable for this column.

Descriptive Measures

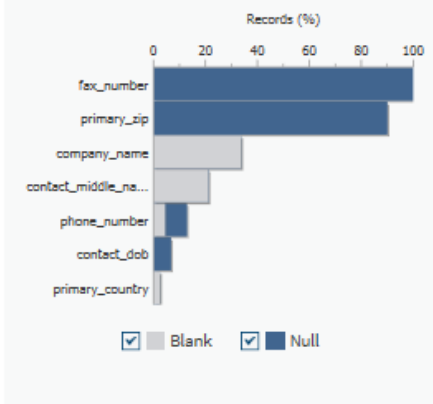
Metadata Measures

Charts

Uniqueness

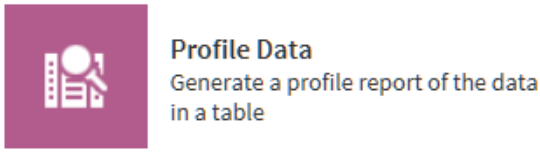


Incompleteness



Profile Data

Introduction



Use the Profile Data directive to generate profile reports for one or more tables. You can select a subset of the columns that you want to include in the profile report. The **Profiles** panel of the Configuration window enables you to change the default behavior of new profiles in order to improve performance. For example, you can limit the number of parallel processes that are used in new profile jobs. For more information, see "Profiles Panel" on page 142.

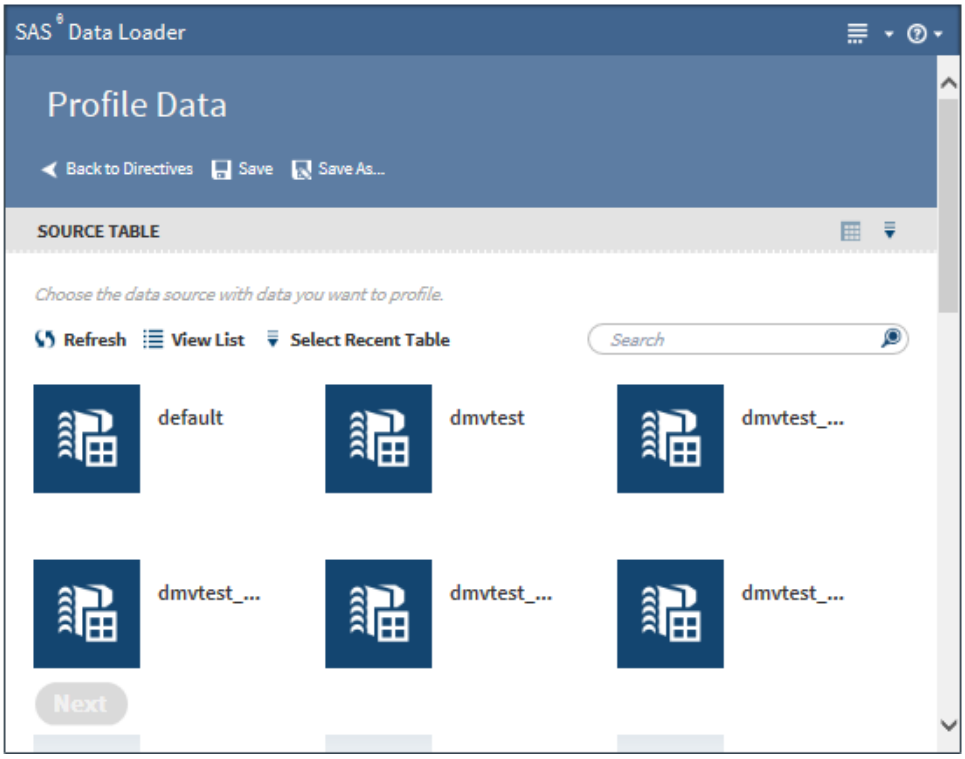
Table Name Length Requirement

Hive tables have a maximum table name length of 132 characters. Many of the SAS Data Loader directives can create tables with names that exceed the SAS table name length limit of 32 characters. The tables that you submit for profiling in the Profile Data directive must conform to the 32-character name length limit. Table names that exceed 32 characters generate error messages.

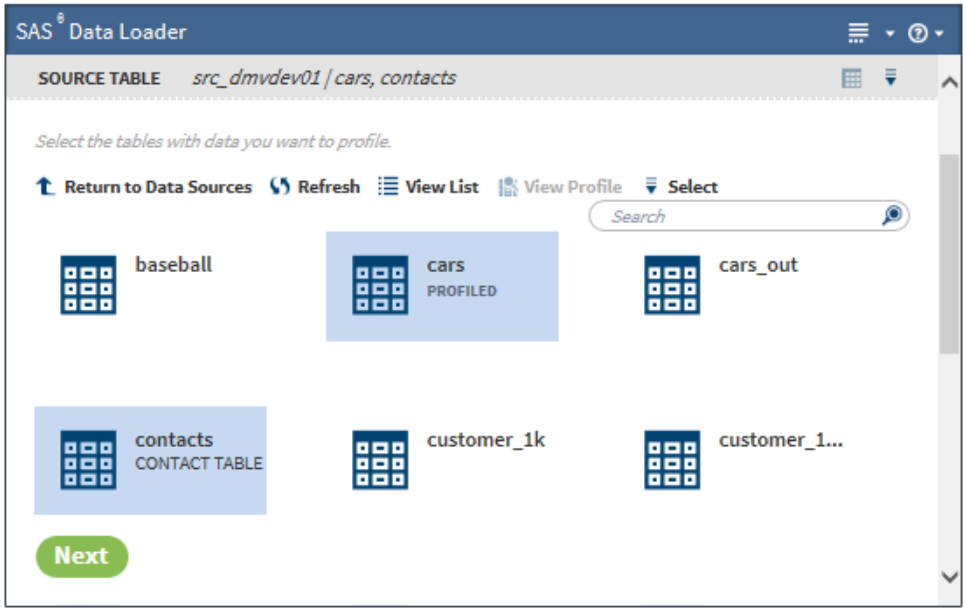
Create a Profile

To create a profile:

1 On the SAS Data Loader directives page, click the Profile Data directive. The **Source Table** task is displayed:



2 Click a data source to display its tables:



3 Select the table or tables for the profile report.

If a profile already exists for a table, PROFILED appears beneath the table name. You can view the existing profile by selecting the table and clicking **View Profile**.

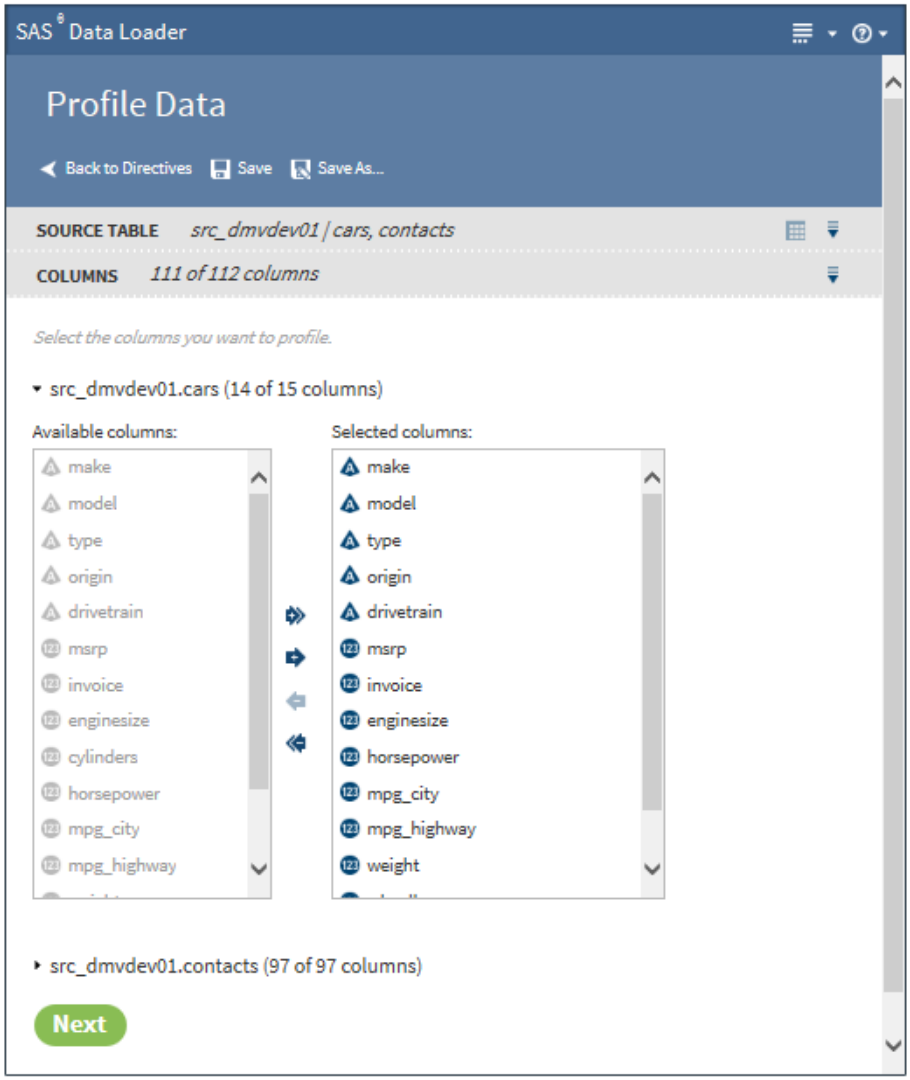
The **Select** menu (Select) provides several options to make selecting tables easier:

- **Select All New Tables** automatically selects all new tables in the current data source.

- **Select Recent Table** enables you to choose from a list of recently used tables. If you select a table from a different data source, the source table information is adjusted accordingly.
- **Deselect All Tables** deselects all tables that you have selected in the current data source.

TIP To view sample data from a table, select the table, and then click in the Source Table header to display the SAS Table Viewer.

Click **Next**. The **Columns** task is displayed:

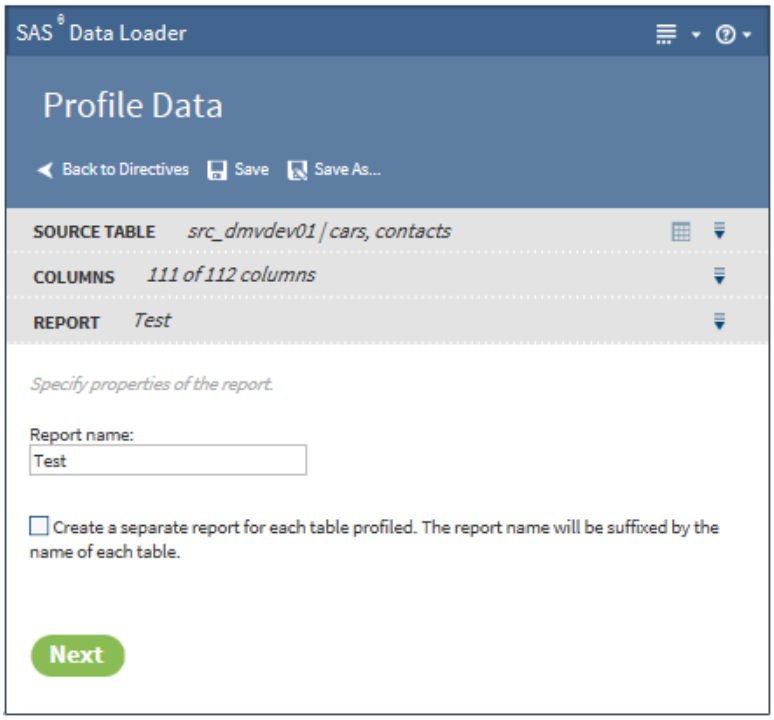


4 The **Columns** task displays the total number of columns that are to be processed in the profile report. If you selected more than one table for your report, the tables are listed by name. Click ▸ next to the tables to display the columns that are included in the profile report.

5 The column names in the **Selected columns** pane appear in the report. Select an individual column name and click ⬅ or ➡ to move the column name between the **Available columns** pane and the **Selected columns**

pane until the correct list of names appears in the **Selected columns** pane. Click or to move all column names at once.

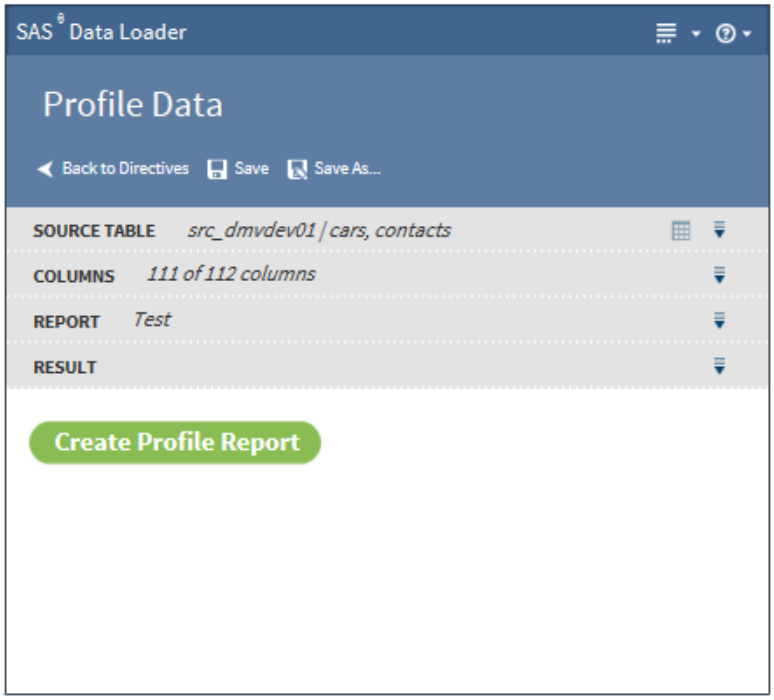
When the column selection is complete, click **Next**. The **Report** task is displayed:



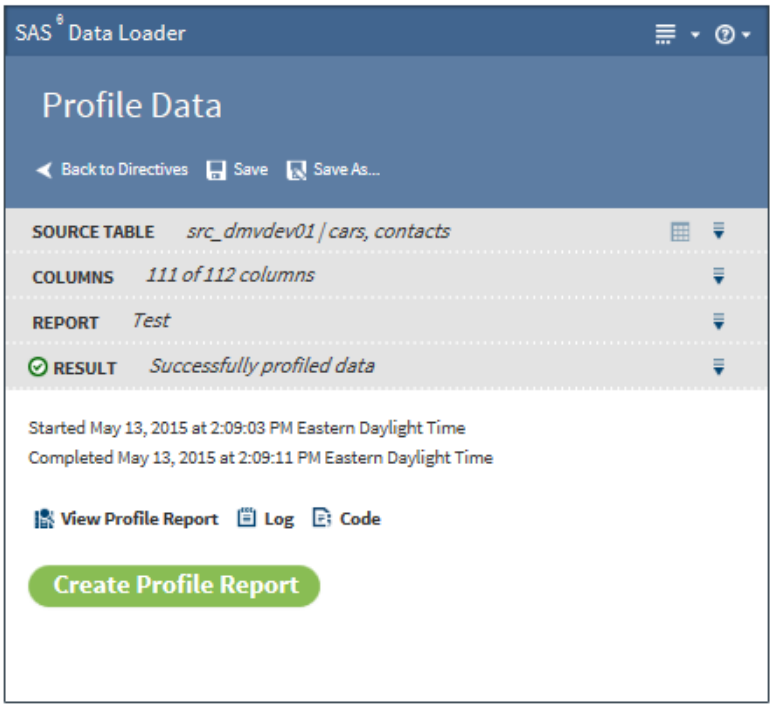
6 In the **Report** task, enter a name for the profile report in the **Report name** field.

If you selected multiple tables and want a separate report for each table, select the**Create a separate report for each table profiled** option.

Click **Next** to display the **Result** task:



7 Click **Create Profile Report**. After successfully creating any profile reports, a screen similar to the following is displayed:



The following actions are available:

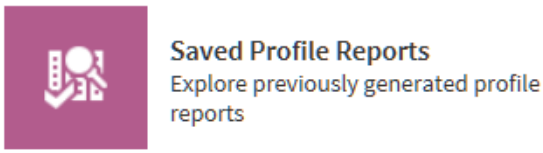
View Profile Report
enables you to view the Profile Report. See "Saved Profile Reports" on page 80 for more information about the profile report.

Log
displays the SAS log that is generated during the creation of the profile.

Code
displays the SAS code that generates the profile.

Saved Profile Reports

Introduction



Use the Saved Profile Reports directive to view the results of previously executed data profiles and to create notes about the results. The profiles are created with the Profile Data directive. The profile reports and notes are stored as XML documents on the file system. Saved Profile Reports displays these XML files in a readable format.

About Profile Reports

Profile reports can provide valuable information about a Hadoop table and help identify issues that might exist before you use the table for data management or analysis. A profile report includes a summary view with information about the table that was profiled and detail views with information about individual columns in the table.

Summary View

The summary view of a profile report includes the following information:

Count
the total number of rows in the table that was profiled.

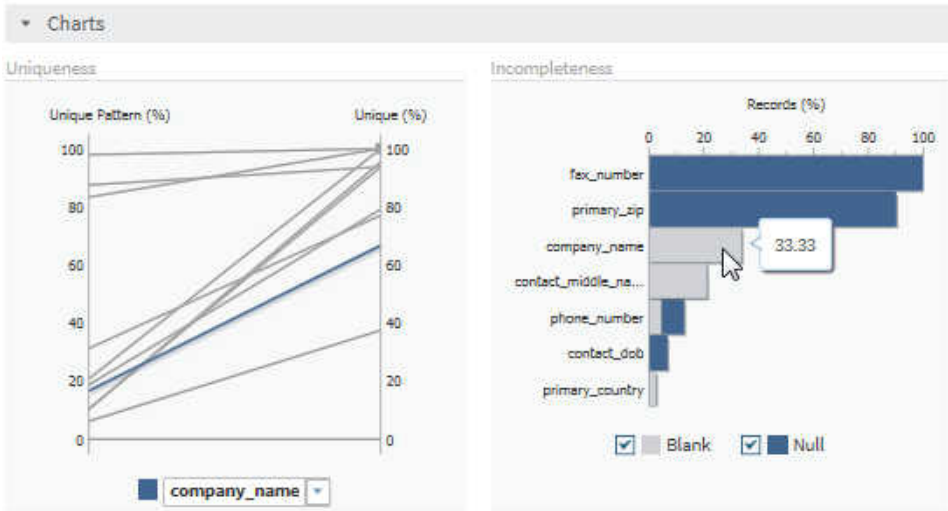
Data Quality Metrics
measurements of data quality for the columns in the table. Measurements include information about the uniqueness of column values, pattern analysis results, and completeness information, including null or blank values.

Note: The measurement of percent null (**Null (%)**) is rounded to the nearest tenth of a percent. Percentages of null values that are smaller than 0.01 are rounded to zero. Refer to the number of null values (**Null (n)**) as needed.

Descriptive Measures
descriptive statistics for columns in the table, including information about the central tendency of the data and how it is dispersed. Depending on the data type of the column, these measures might not be available.

Metadata Measures
metadata for the columns in the table, including the data type, the column length, and whether the column is a primary key candidate.

Charts
summary graphics that provide information about the uniqueness and incompleteness of column values.



Column Detail Views

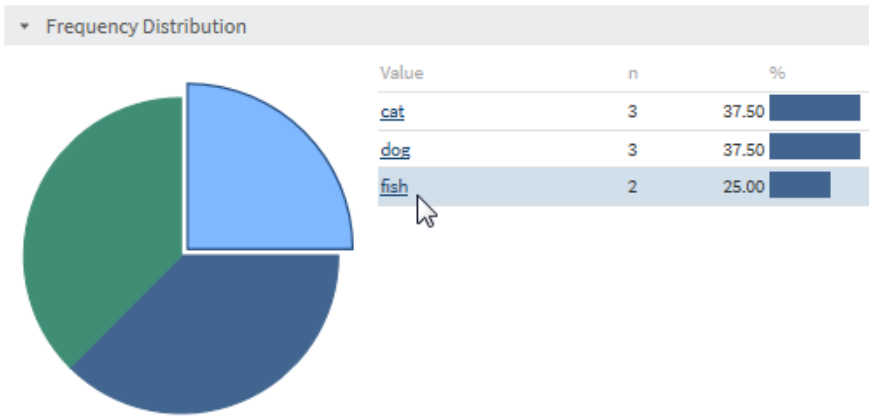
When you click on a column from the summary view in a profile report, another view is displayed that provides more detailed information about the selected column.

The detail view of a profile report includes the following information:

Count
the total number of rows in the table that was profiled.

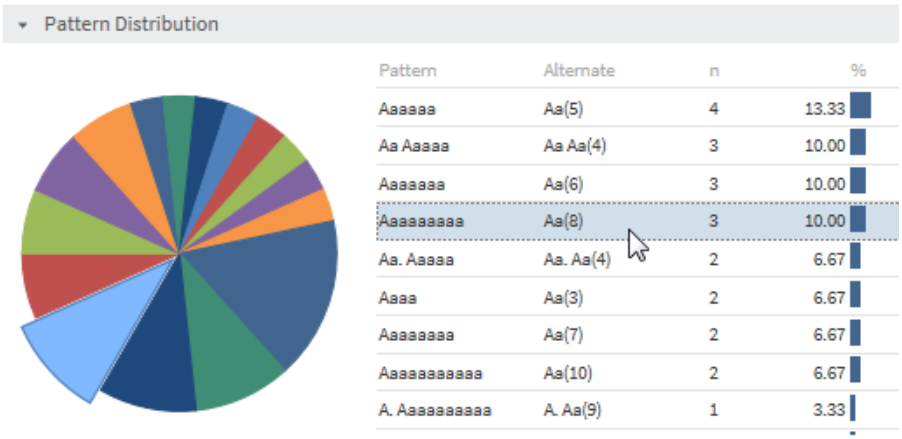
Standard Metrics
a combined listing of the data quality metrics, the descriptive measures, and the metadata measures for the column that were displayed on the summary view.

Frequency Distribution
a listing of the unique values for the column, including information about how frequently a value occurs in the table. When you select a value from the list, the associated section of the pie chart is highlighted.



Pattern Distribution

a listing of the distinct pattern values that were derived from performing pattern analysis on the values for the column. The content of the pattern value describes the content of the data and indicates whether each character is uppercase, lowercase, or numeric. When you select a value from the list, the associated section of the pie chart is highlighted.

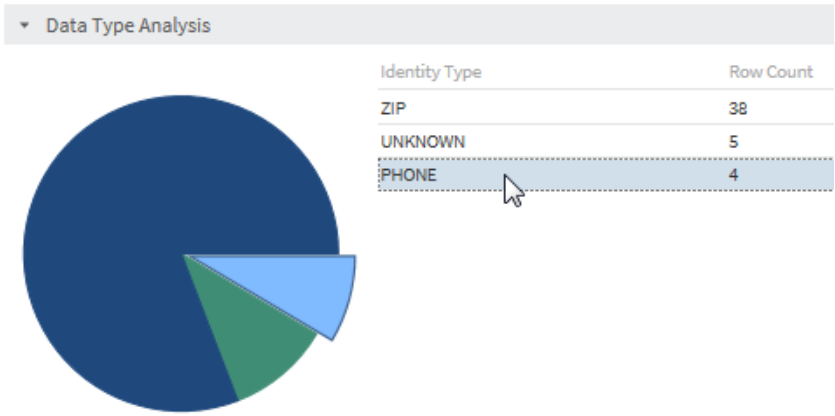


Outliers

a listing of extreme values for the column. By default, the 10 lowest values and the 10 highest values are saved, but you can change the number of outliers that are saved in the profile configuration settings. For more information, see "Profiles Panel" on page 142.

Data Type Analysis

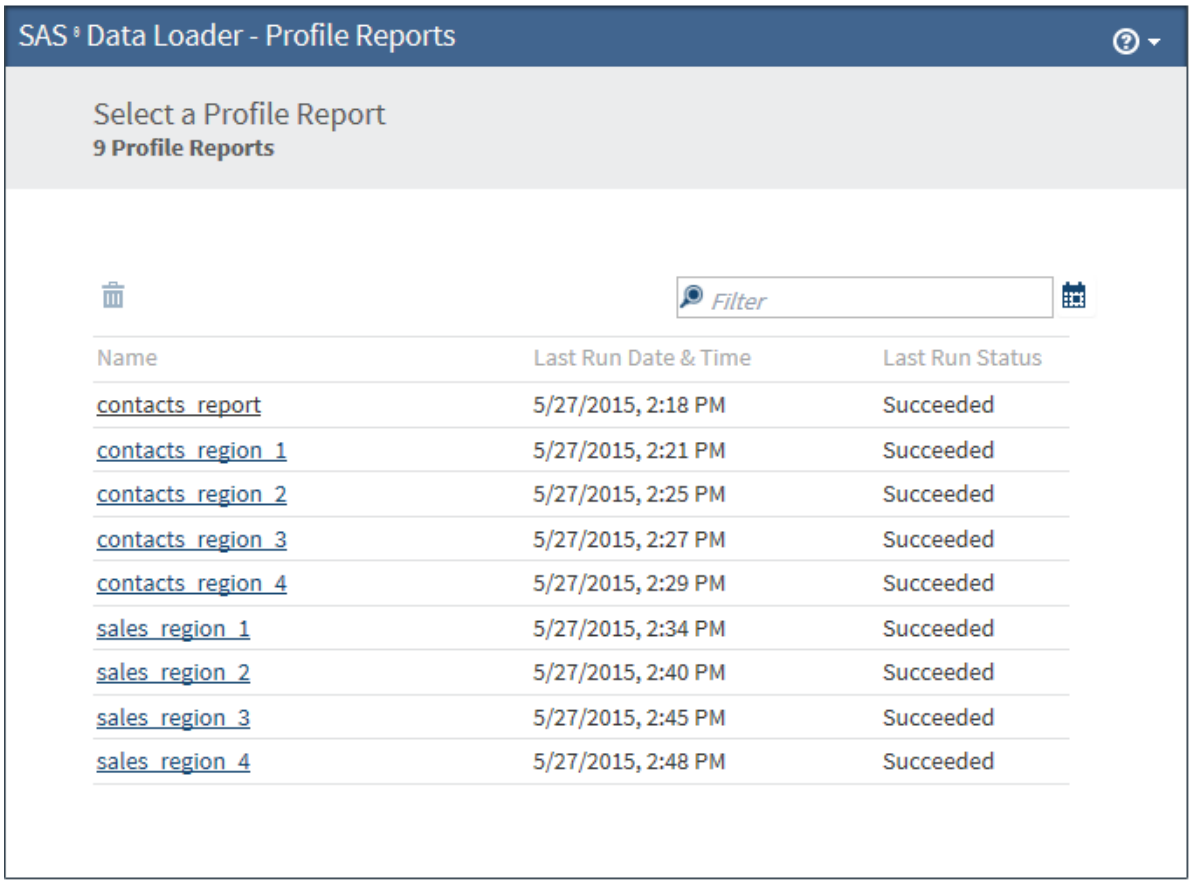
a listing of possible types of data for the information in the column, as determined by data type analysis that is automatically performed by SAS Data Loader. Results for data type analysis are available only for columns that contain string characters (for example, contact information such as name, address, state, ZIP code, and so on).



Open Saved Profile Reports

To open a saved profile report:

1 In the SAS Data Loader directives page, click the Saved Profile Reports directive to open a new browser tab. The Select a Profile Report page is displayed on the new tab:



Note: Any profile job that runs longer than five days is deleted from the Select a Profile Report page.

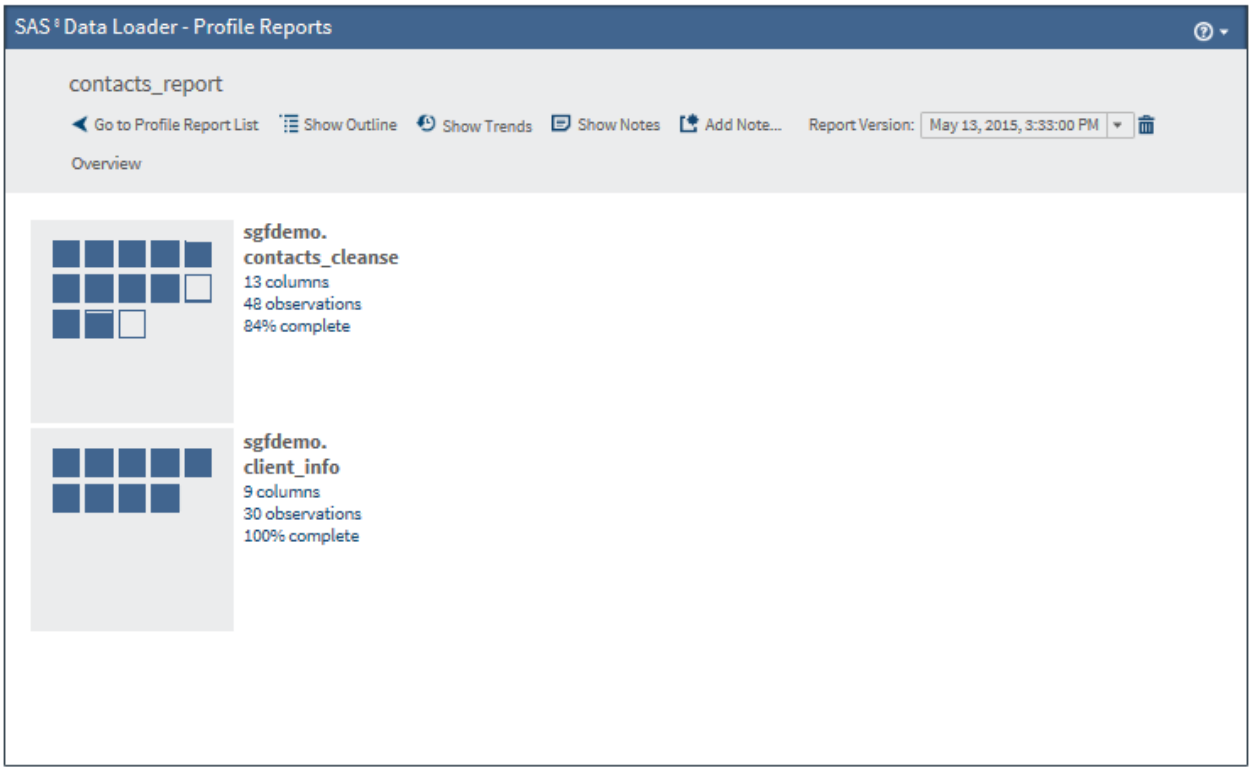
2 You can filter the list of reports using the following methods:

- Click and select a date. This filter displays profile reports that were generated on or after the selected date.
- Enter a text string into the search field.
- Click to remove the filter and restore the full list.

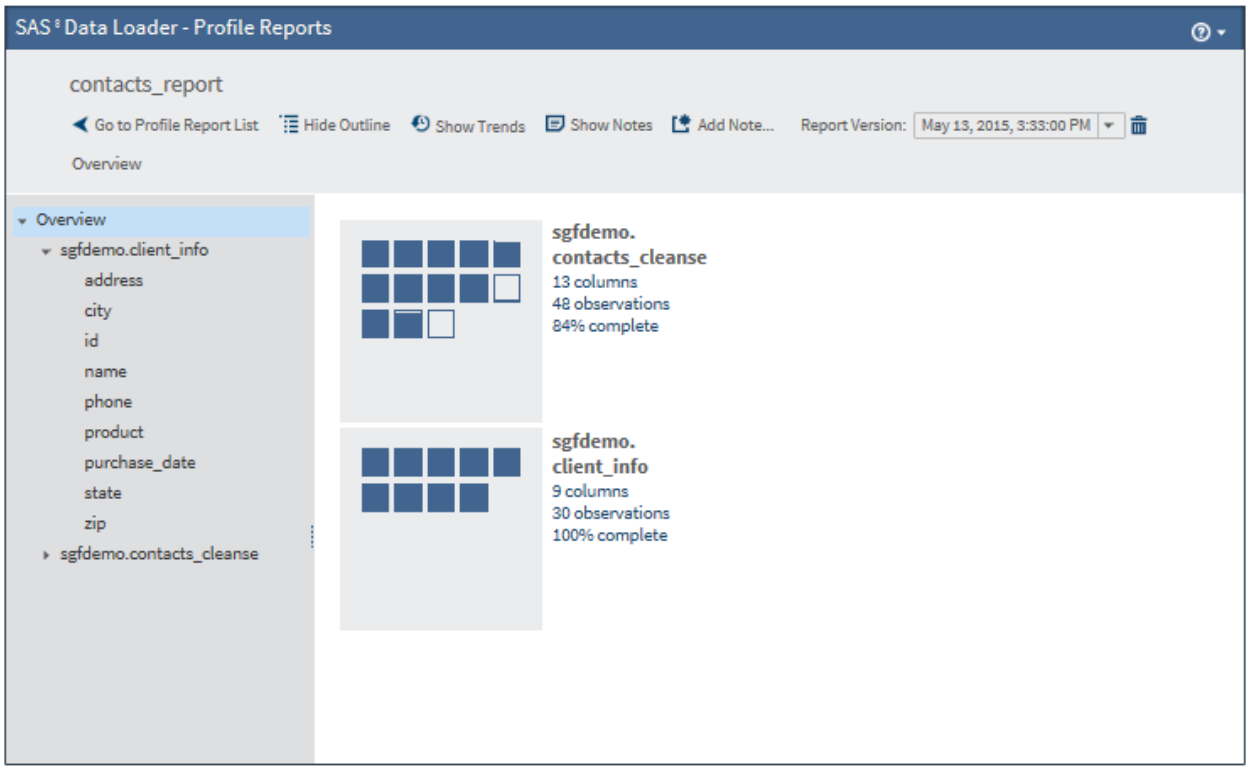
3 To delete profile reports, select one or more reports and click .

4 To open a profile report, click its name.

- If the report contains a single table, the table opens directly in the detail view shown in Step 6.
- If the report contains multiple tables, the table opens in an overview:



5 You can click a table to go directly to a more detailed view or you can click to open the outline view:

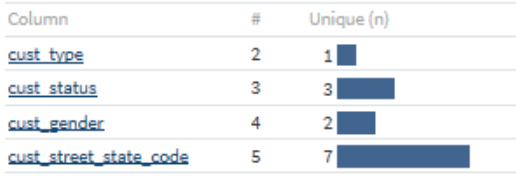


The following actions are available:

Go to Profile Report List
returns you to the Profile Report List.

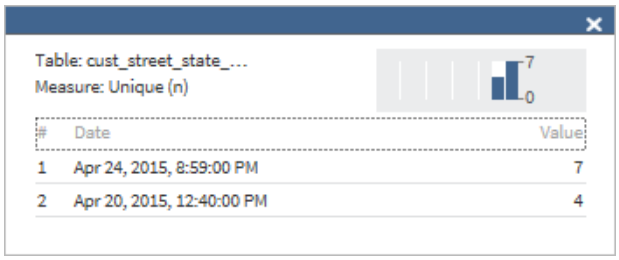
Show or Hide Outline
displays or hides the outline in the left pane.

Show or Hide Trends
displays or hides the trend graphs for data that is presented in the summary view. You can use trend graphs to quickly visualize changes in the data across multiple versions of the same report. When trend graphs are not displayed, the current value of the metric is shown. For example:



When trend graphs are on, each graph displays the 10 most recent values of a metric, as determined by the selected version of the report. For example:

To view the complete list of values for the metric, you can click the trend graph. A window is displayed:

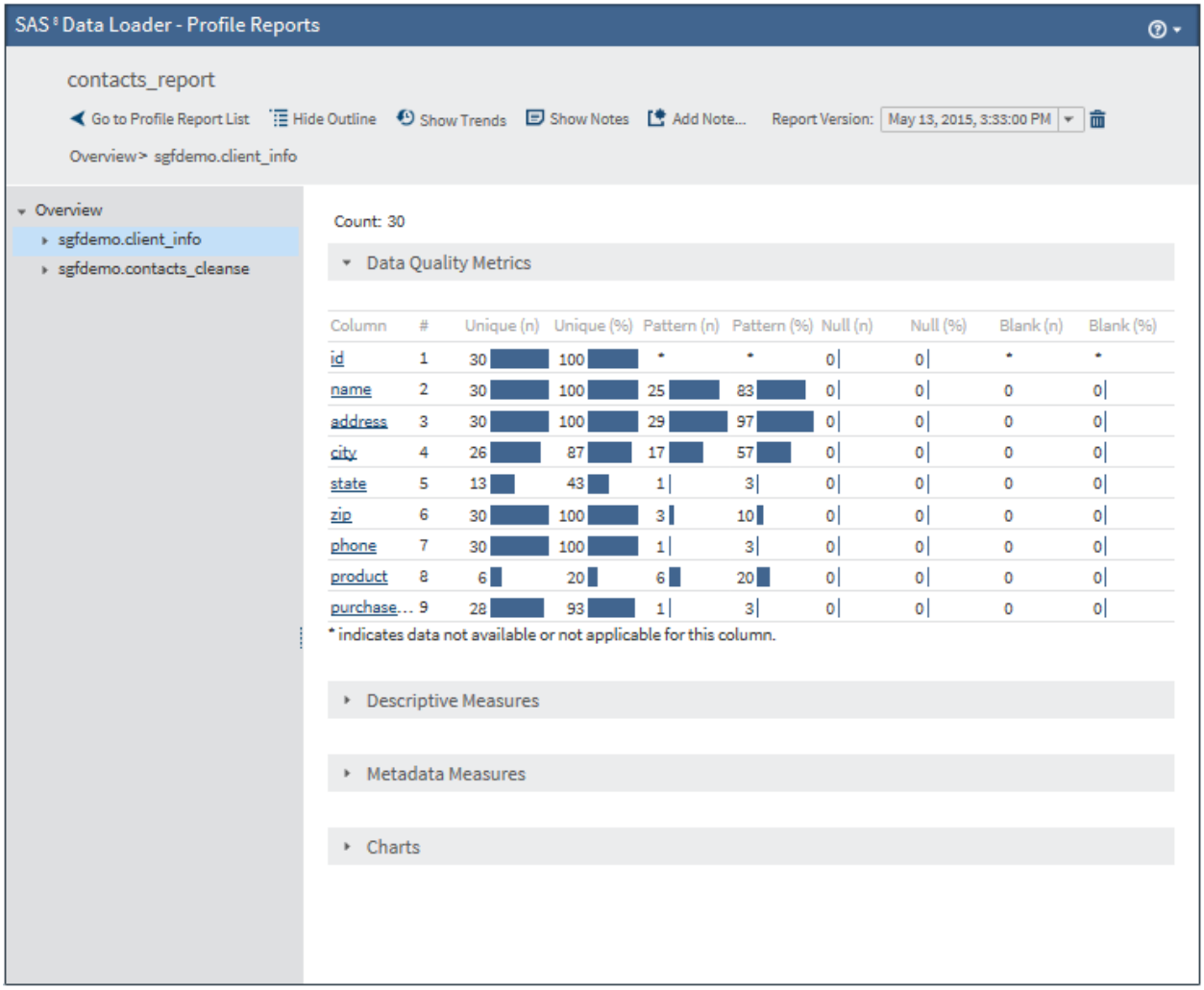


Show or Hide Notes
displays or hides notes in the right pane. You can filter the notes by entering a text string into the filter field.

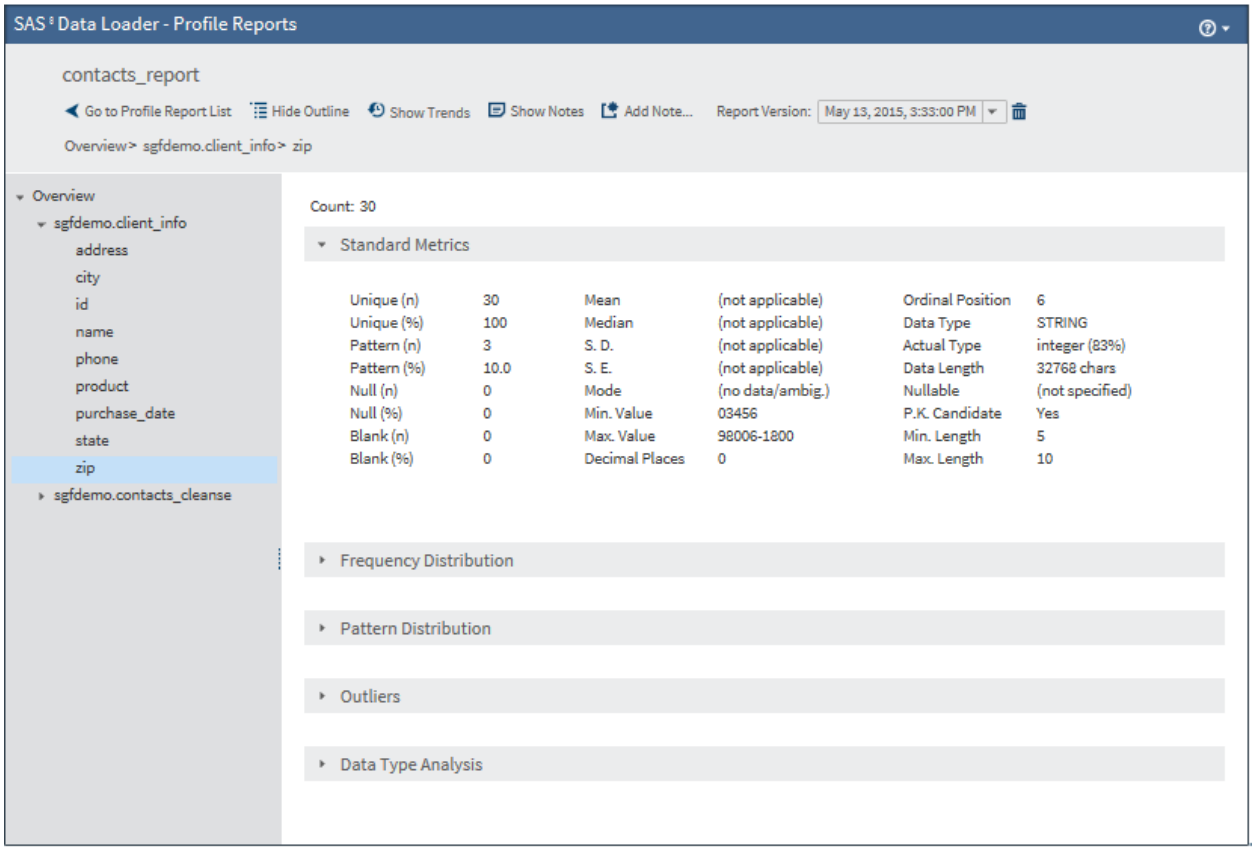
Add Note
opens a dialog box in which you can add a note.

Report Version
enables you to select the version of the report by date.

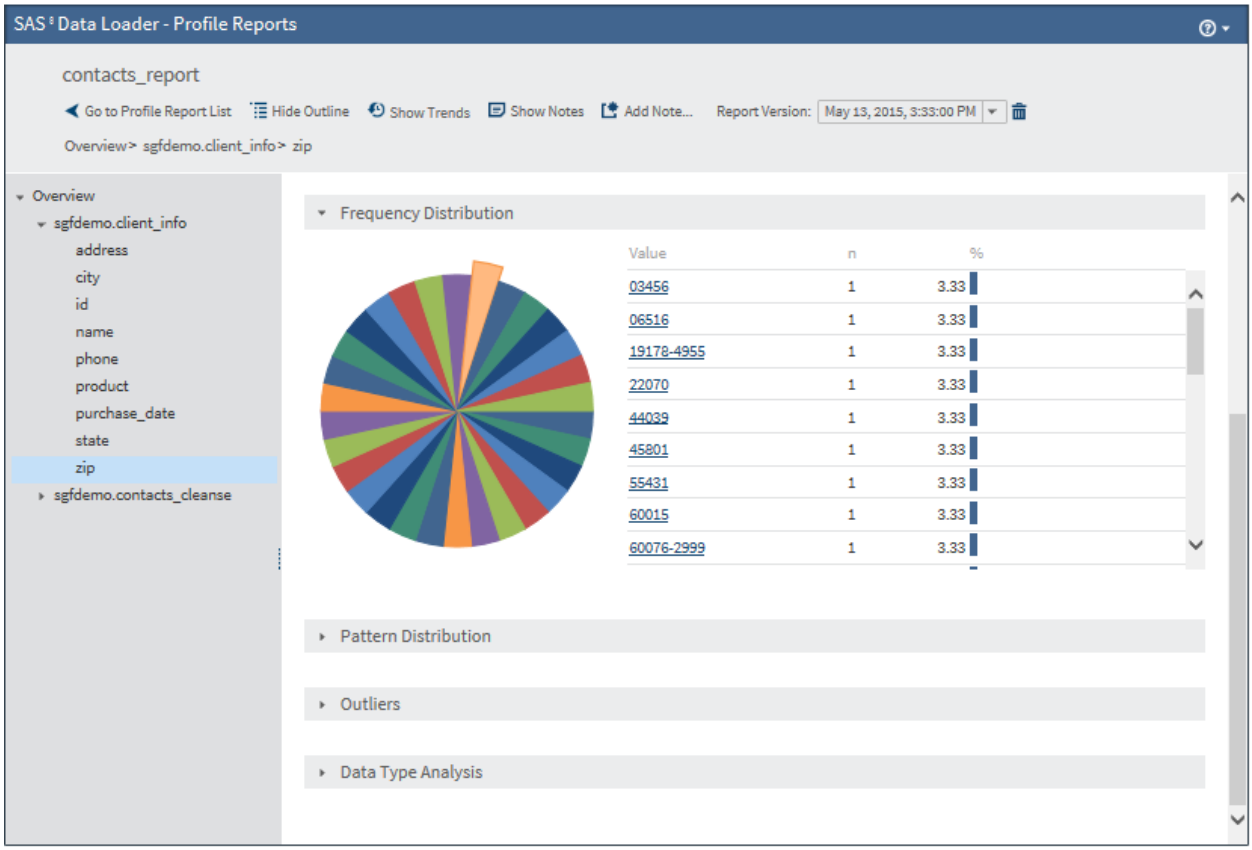
6 Select a table in the **Overview** pane or click directly on the table icon to display detailed information in the right pane. The Data Quality Metrics are displayed by default.



7 Click ▸ next to a table name to display columns. Select a column to display detailed column information in the right pane:



8 Click ▸ in the gray header bars to display the metrics in those sections. For example, clicking on Frequency Distribution icon displays the following metrics.



Clicking links in the detail view opens SAS Table Viewer.

6

Copy Data To and From Hadoop

Overview of the Copy Data Directives 89

Copy Data to Hadoop 90
Introduction 90
Prerequisites 90
Example 90
Usage Notes 101

Import a File 102
Introduction 102
Example 103

Copy Data from Hadoop 108
Introduction 108
Prerequisites 108
Example 108
Usage Notes 117

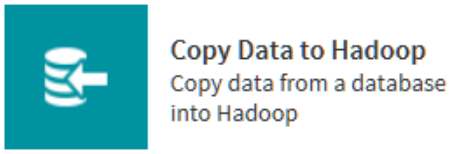
Load Data to LASR 118
Introduction 118
Prerequisites 119
Example 119
Usage Notes 119

Overview of the Copy Data Directives

The directives Copy Data to Hadoop, Import a File, and Copy Data from Hadoop enable you to move data from your files system or database management systems into and out of Hadoop. The copy directives require database connections to be defined on the **Databases** panel of the Configuration window. For more information, see “Set Global Options” on page 132. The Import a File directive helps you import miscellaneous files from your file system into Hadoop as columnar tables.

Copy Data to Hadoop

Introduction



The Copy Data to Hadoop directive enables you to copy data from your database management systems into Hadoop. You can also copy SAS data into Hadoop.

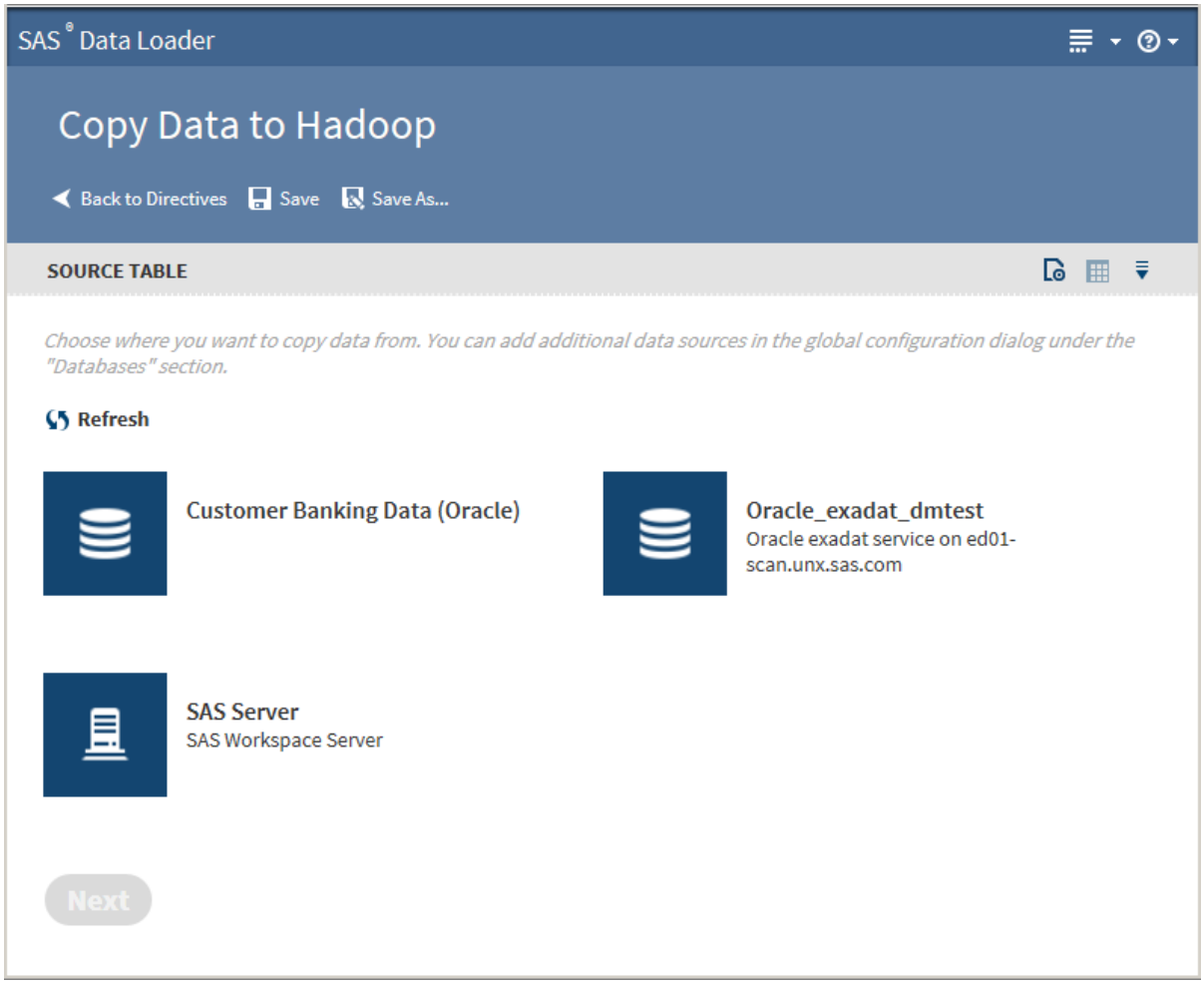
Prerequisites

When you open the Copy Data to Hadoop directive, the **Source Tables** task shows the data sources that are currently defined in SAS Data Loader. If you do not see the database from which you want to copy, you must add a connection to that database. See "Databases Panel" on page 139 for more information.

Example

Follow these steps to copy data into Hadoop from a database:

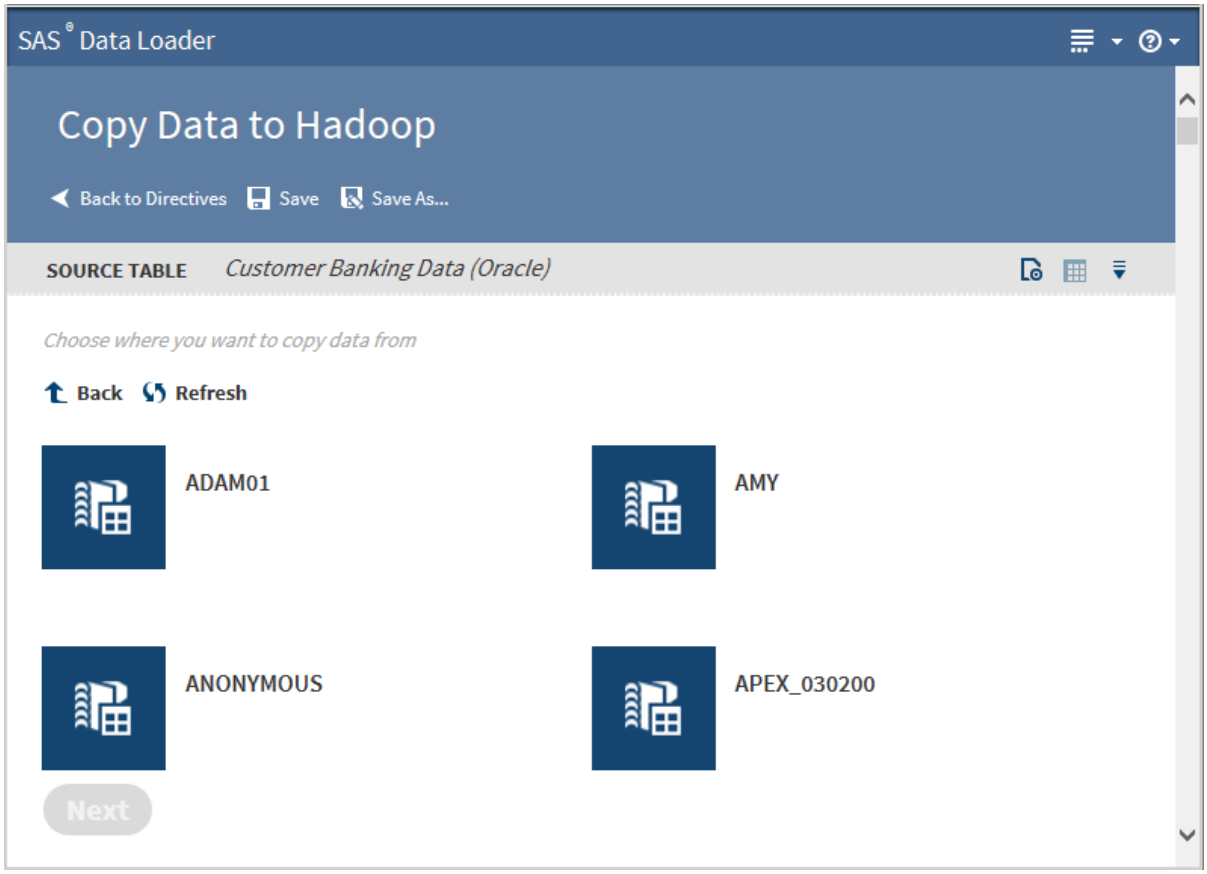
1 On the SAS Data Loader directives page, click the Copy Data to Hadoop directive. The **Source Table** task that lists available databases is displayed:



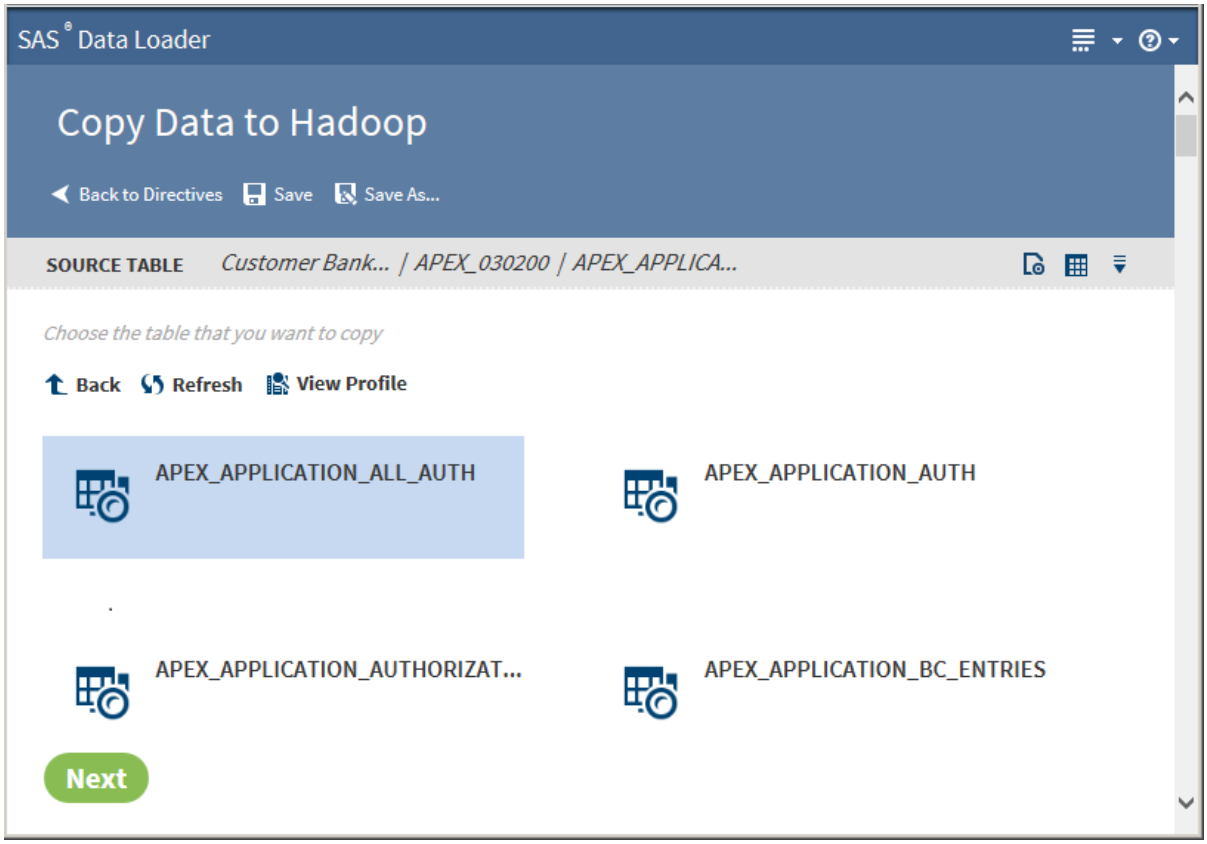
Note that the **SAS Server** data source points to the following location on the vApp host: `vApp-shared-folder/SASData/SAS Data Location`. To copy SAS data to Hadoop, all source tables must first be copied to this location.

Note: When you select the **SAS Server** folder, filenames are not translated. For locales other than English, this means that files that exist in the **SAS Server** folder are not displayed for selection. To work around this issue, you can import entire SAS files. In the **Source Table** task, select a file outside of the **SAS Server** folder and click through the directive. Select or create a target table. In the **Code** task, open the Code Editor to change the source file information, and then run the job.

2 Click a database to display its data sources:



3 Click a data source to display its tables:



4 Select the table from which to copy data.

> **TIP** If a profile already exists for a table, PROFILED appears beneath the table name. You can view the existing profile by selecting the table and clicking **View Profile**.

Clicking the **Action** menu enables the following actions:

Open
opens the current directive.

Table Viewer
enables you to view sample data from a table. Select the table, and then click to display SAS Table Viewer.

Advanced Options
opens a dialog box that enables you to modify the advanced options. The advanced options enable additional character variable length to accommodate converted non-UTF8 encoding.

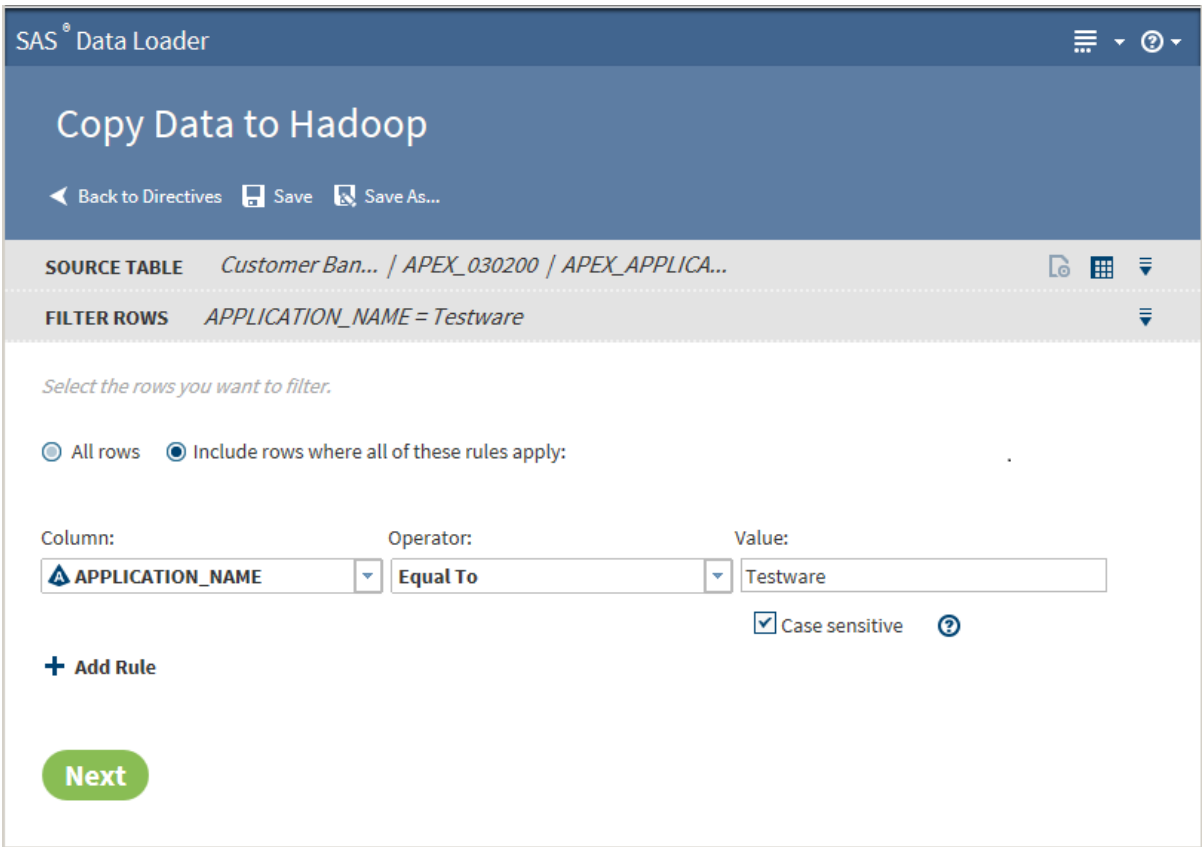
> **TIP** It is recommended that you use UTF8 encoding in SAS data when copying data from SAS to Hadoop. The vApp always uses UTF8 encoding. If you copy a non-UTF8 encoded data set from elsewhere, then the Hadoop target table is not able to accommodate all the characters. This limitation is due to the increased number of bytes when the data is converted to UTF8 encoding.

Note: Modify only one of the following two advanced options. If you fill in both fields, then the value in the multiplier field is ignored.

Number of bytes to add to length of character variables (0 to 32766)
Enter an integer value from 0 to 32766.

Multiplier to expand the length of character variables (1 to 5)
Enter an integer value from 1 to 5.

Click **Next**. The **Filter Rows** task is displayed:



5 The **Filter Rows** task enables you to filter the rows to be copied. You can select **All rows** or create filter rules. To create filter rules:

a Select **Include rows where all of these rules apply**.

b Select a column and an operator from the drop-down lists.

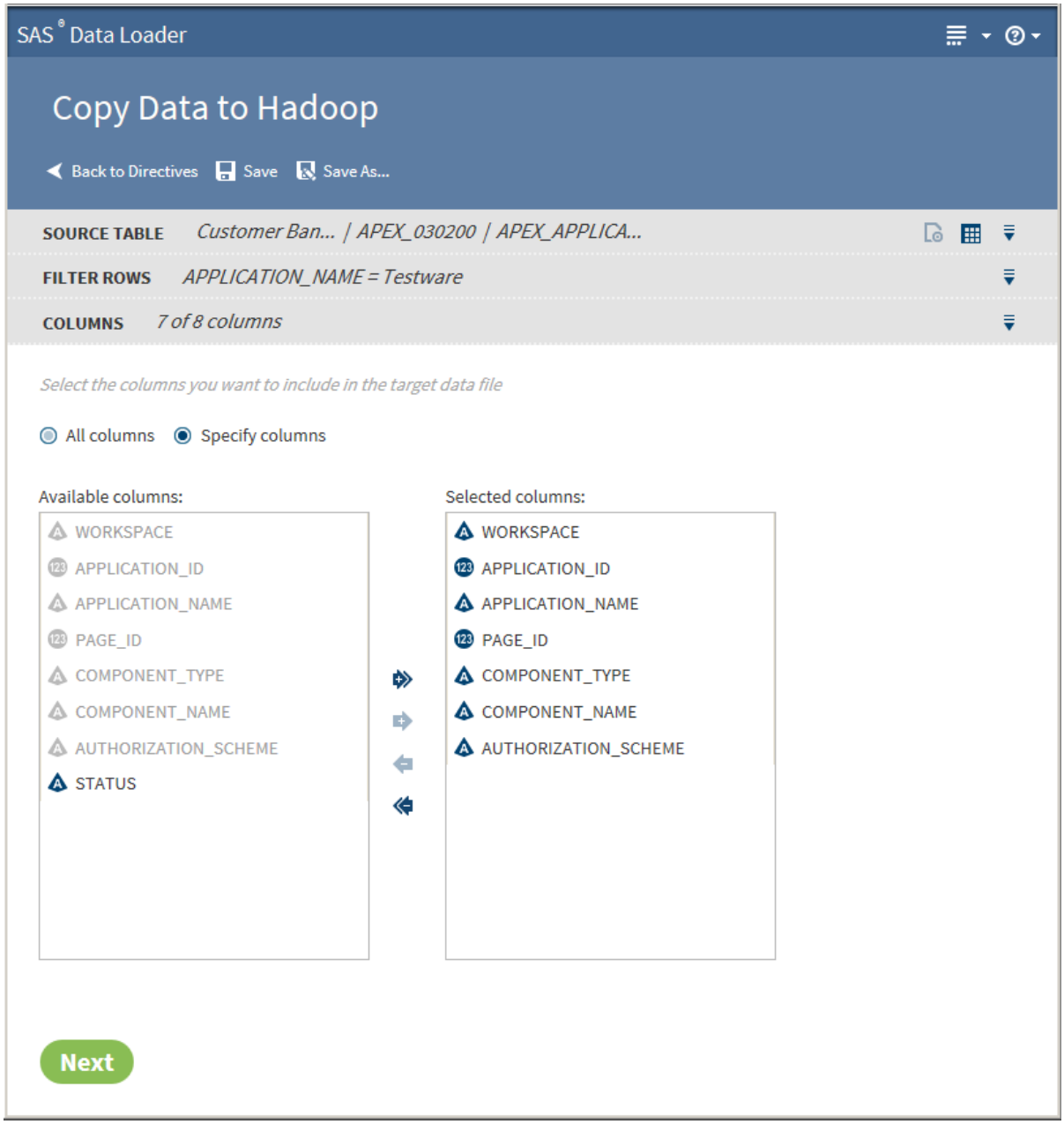
Note: If the table for which you are defining a filter is in the OTHER database format, the database might not support all operators. You should use only those operators that are supported by your database in the filter.

c Enter a value in the **Value** field.

d If appropriate, select **Case sensitive** for a string value.

e If you want to filter with additional rules, click **Add Rule**.

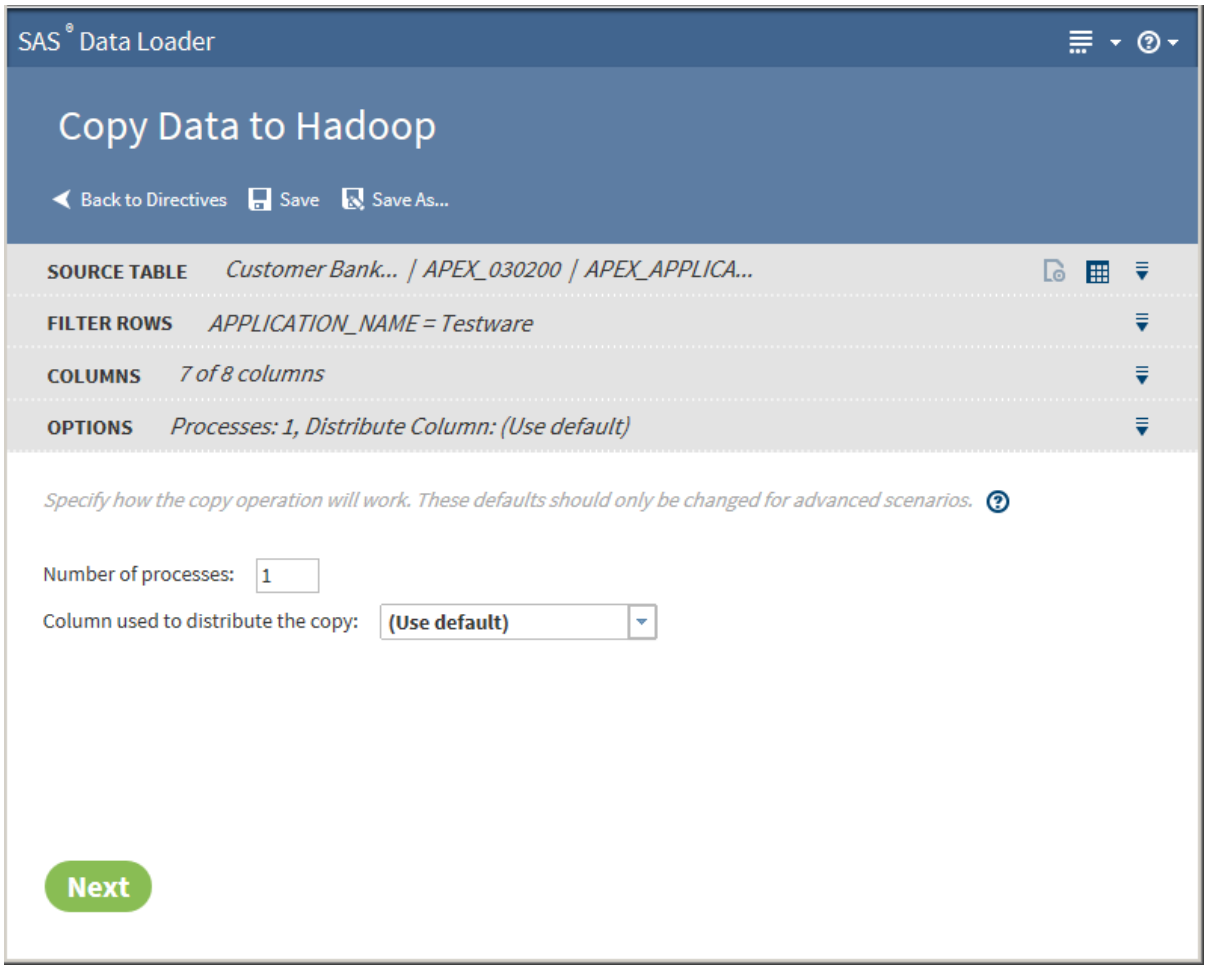
Click **Next**. The **Columns** task is displayed:



6 The **Columns** task enables you to choose the columns to be copied. You can select **All columns** or **Specify columns**.

The columns in the **Selected columns** pane are copied to Hadoop. Select an individual column name and click or to move the column name between the **Available columns** pane and the Selected columns pane until the correct list of names appears in the **Selected columns** pane. Click or to move all column names at once.

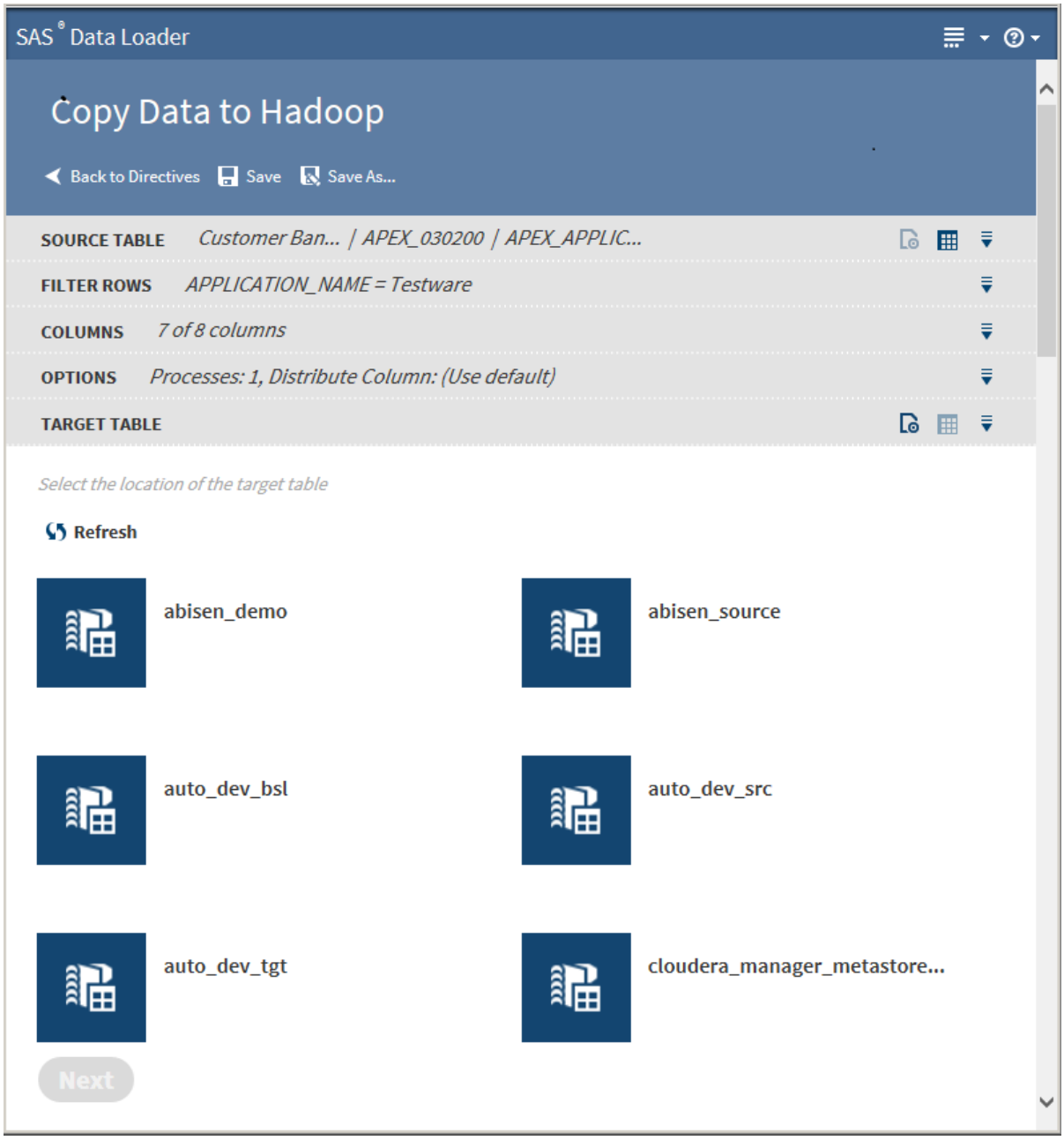
When the column selection is complete, click **Next**. The **Options** task is displayed:



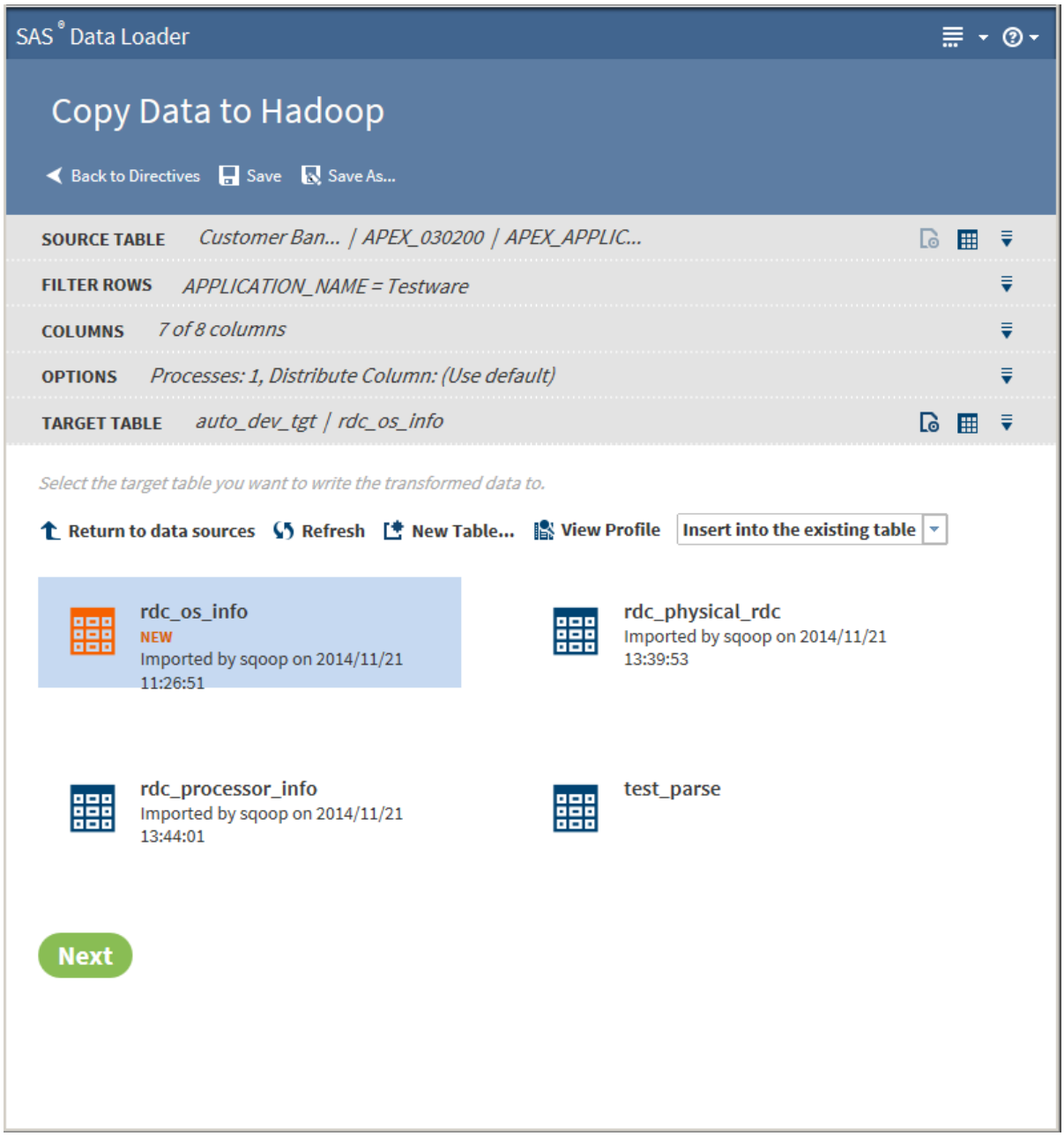
7 The values in the **Options** task should not be changed unless you have advanced knowledge of database operations.

CAUTION! If you change the number of processes, you are required to select a distribution column. Changing the number of processes to greater than one expands the number of processes and source data connections that are used to import data. When running in this mode, a column must be identified in order to distribute the data across the parallel processes. This column is typically the primary key or index of the table in the data source. Only single columns are allowed. Numeric integer values that are evenly distributed in the data are recommended

Click **Next**. The **Target Table** task is displayed with data sources:



8 Click a target data source to display its tables:



9 Select the target table to which to copy data.

TIP

- You can create a new table by clicking **New Table**.
- If a profile already exists for a table, PROFILED appears next the table icon. You can view the existing profile by selecting the table and clicking **View Profile**.

Clicking the **Action** menu enables the following actions:

Open
opens the current directive.

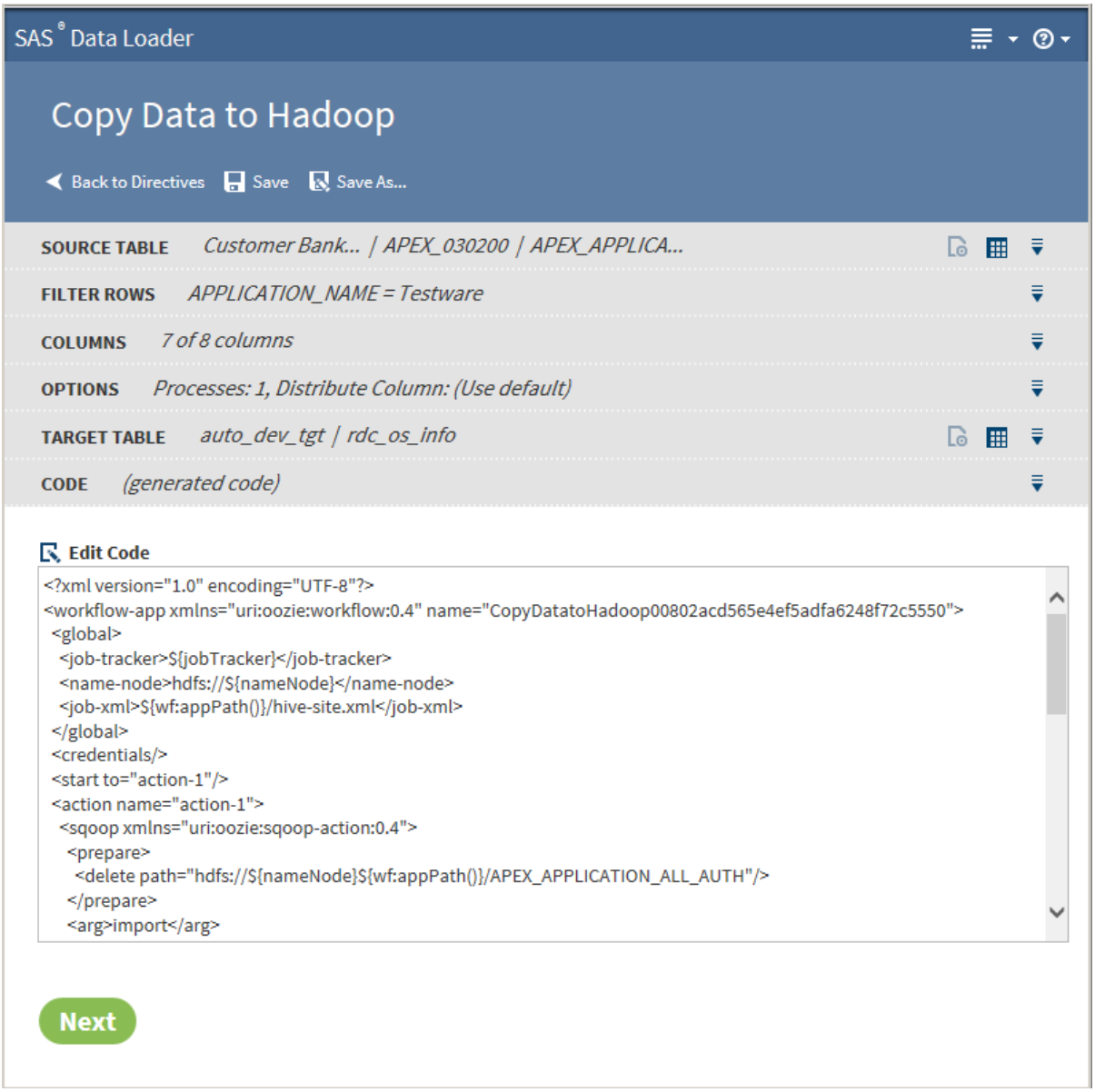
Table Viewer
enables you to view sample data from a table. Select the table, and then click to display SAS Table Viewer.

Advanced Options
opens a dialog box that enables you to modify the following advanced options:

Output table format
Use the drop-down list to select one of five output table formats: Hive default, Text, Parquet, ORC, or Sequence. The Parquet format is not support for MapR distributions of Hadoop.

Delimiter
Use the drop-down list to select one of five output table formats: Hive default, Comma, Tab, Space, or Other.

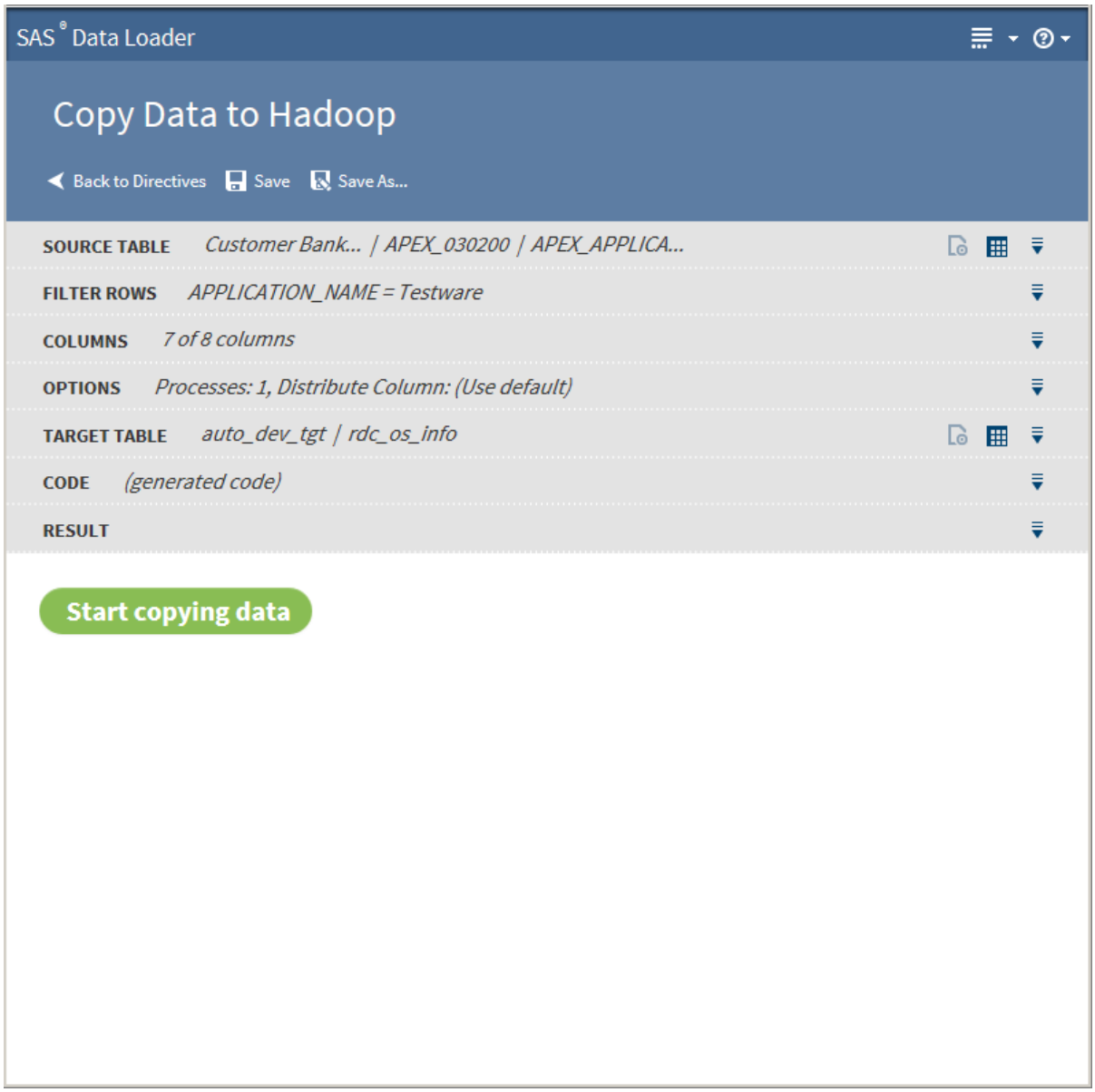
Click **Next**. The **Code** task is displayed:



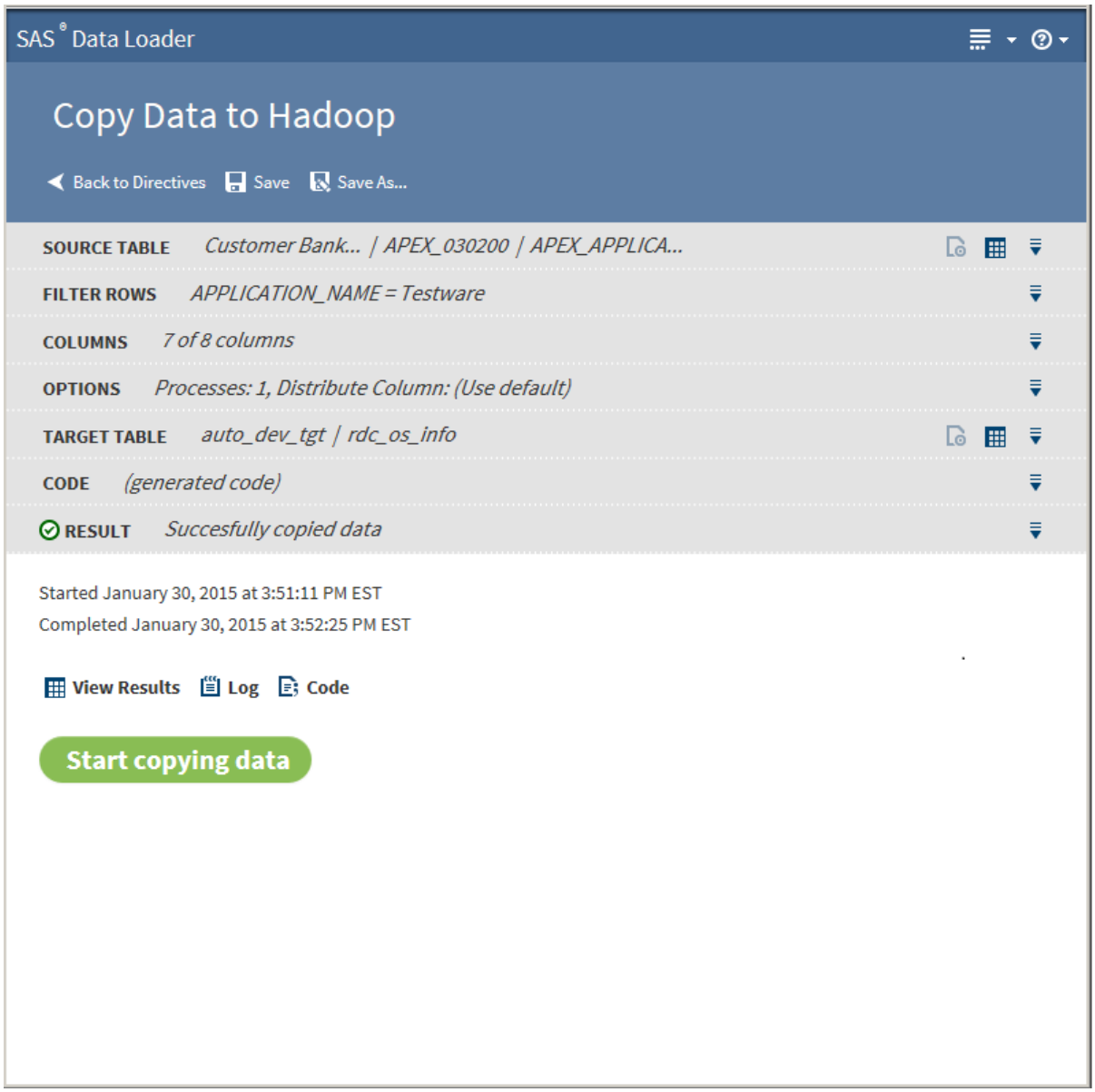
10 Click **Edit Code** to modify the generated code. To cancel your modifications, click **Reset Code**.

CAUTION! Code edits are intended to be used only to support advanced features. Code edits are not needed or required under normal circumstances.

11 Click **Next**. The **Result** task is displayed:



12 Click **Start copying data**. The **Result** task displays the results of the copy process:



The following actions are available:

View Results
enables you to view the results of the copy process in SAS Table Viewer.

Log
displays the SAS log that is generated during the copy process.

Code
displays the SAS code that copies the data.

Usage Notes

- If LDAP is used to protect your Hadoop cluster, you cannot use the Copy Data To Hadoop directive to copy data from a DBMS. For more information, see "Active Directory (LDAP) Authentication" on page 144.

- If necessary, you can change the maximum length of character columns for input tables to this directive. For more information, see "Change the Maximum Length for SAS Character Columns" on page 147.
- Error messages and log files that are produced by the Copy Data to Hadoop directive include the URL of the Oozie log file. Oozie is a job scheduling application that is used to execute Copy Data to Hadoop jobs. Refer to the Oozie log for additional troubleshooting information.
- When copying data from Teradata:
  - In Cloudera 5.2 or later, the Teradata source table must have a primary key that is defined, or you must specify a distribution column on the Options task.
  - In Hortonworks 2.1 or later, you are required to insert Teradata data into existing tables. The creation or replacement of tables is not supported. This is due to a limitation in the HortonWorks Sqoop connector. One workaround is to ask your Hadoop administrator to drop an existing table, and then create an empty table with the desired schema. At that point, you can use the Append option in the Copy Data to Hadoop directive to copy a Teradata table into the empty table. For more information, see Step 9 on page 98 in the Example section.
- When copying data from SQL Server, note that SQL Server does not support the SQL standard syntax for specifying a Date literal, which is as follows: `DATE 'date_literal'`. Edit the generated code and remove the word DATE that appears prior to the quoted date literal. For example, you would change `( table0.BEGDATE >= DATE '1990-01-01' )` to `( table0.BEGDATE >= '1990-01-01' )`. For more information about the Code task, see Step 10 in the Example section.
- When copying data from Oracle, note that Oracle table names must be uppercase.

Import a File

Introduction

Import a File
Import data from a file into Hadoop

Use the Import a File directive to copy a delimited source file into a target table in HDFS and register the target in Hive.

As you use the directive, it samples the source data and generates default column definitions for the target. You can then edit the column names, types, and lengths.

To simplify future imports, the Import a File directive enables you to save column definitions to a file and import column definitions from a file. After you import

column definitions, you can then edit those definitions and update the column definitions file.

The directive can be configured to create delimited Text-format targets in Hadoop using an efficient bulk-copy operation.

In the source file, the delimiter must consist of a single character or symbol. The delimiter must have an ASCII character code in the range of 0 to 255.

To learn more about delimiters and column definitions files, refer to the following example.

To copy database tables into Hadoop using a database-specific JDBC driver, use the "Copy Data to Hadoop" directive.

Example

Follow these steps to use the Import a File directive:

1. Copy the file to be imported, or copy a directory of files to be imported, into the directory `vApp-install-path\shared-folder\Files\MyData`. A common name for the vApp shared folder is `SASWorkspace`.
2. On the SAS Data Loader directives page, click **Import a File**.
3. In the **Source File** task, click to open folders as needed, click the file that you want to import, and click **Next**.

   TIP To open or save a copy of a selected file, click and select **Download**.

4. In the **File Specification** task, click View File. You will see the delimiter that separates the variable values. Check to see whether the delimiter is used as part of a variable value.

   Notes:

   - In the source file, all variable values that contain the delimiter character must be enclosed in quotation marks (`"`).
   - In Hadoop distributions that run Hive 13 or earlier, a backslash character (\ ) is introduced into the target when the delimiter appears in source values. For example, the source data `One, "Two, Three", Four` would be represented in the target as Column A: `One`, Column B: `Two\, Three`, and Column C: `Four`. In Hive 14 and later, the backslash character is not introduced into the target.

5. Click **Input format delimiter** to display a list of available delimiters. Click the delimiter that you see in your source file, or click **Other**. If you clicked **Other**, then enter into the text field the single-character delimiter or the octal delimiter that you see in the source file. Octal delimiters use the format *nnn*, where *n* is a digit from 0 to 9. The default delimiter in Hive is \001.

   Note: Input delimiters must have ASCII character codes that range from 0 to 255, or octal values that range from \000 to \177.

6. To efficiently register and store the source data in Hadoop using a bulk-copy, select **Use the input delimiter as the delimiter for the target table**. The bulk-copy operation is efficient, but the source data is not analyzed or

validated in any way. For example, the directive does not ensure that each row has the expected number of columns.

Note:

- The bulk-copy operation is used only if the next two options are not selected. If this condition is met, then the source file is bulk-copied to the target. The format of the target is Text. The Text format is used regardless of another format that might be specified in the **Target Table** task.
- If your source file uses `\N` to represent null values, you can preserve those null values in the target. A bulk-copy operation is required. In Hive, the default null value is `\N`.

CAUTION! Bulk-copies include newline characters in the target without generating error messages or log entries. Newline characters in the target can cause data selection errors. Remove newline characters from the source as needed to create a usable target in Hadoop.

7 If the source file is formatted accordingly, select **Check the input file for records wrapped in quotation marks (")**. Quotation marks are required when the delimiter appears within a variable value.

Quotation marks that appear inside a quoted value need to be represented as two quotation marks (**""**).

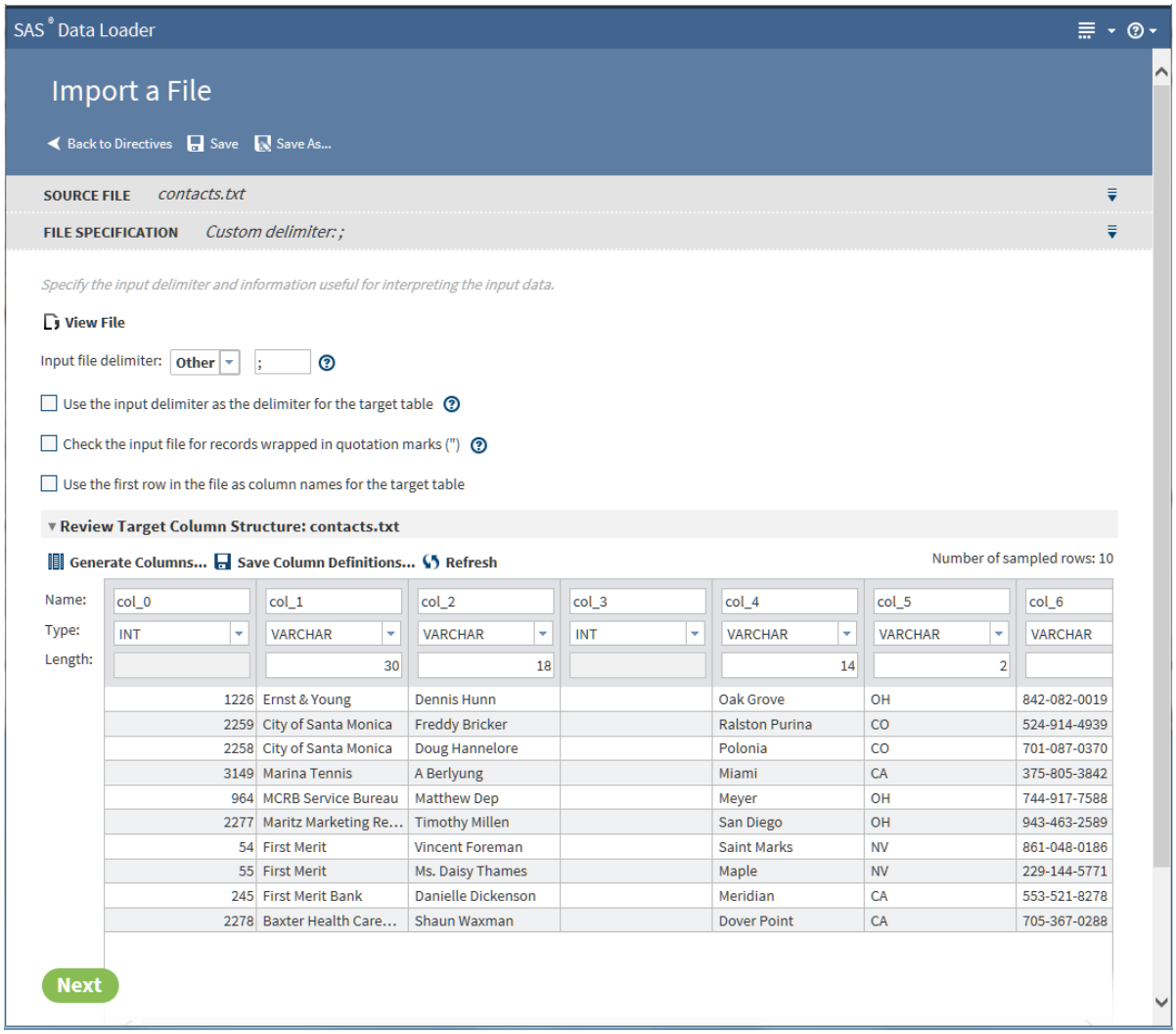
CAUTION! Except for bulk-copy operations, jobs fail if the source contains newline characters. For all jobs other than bulk-copies, ensure that the source file does not contain newline characters.

In all cases, newline characters within source values will cause the import job to fail.

8 If your source file includes column names in the first row, then select **Use the first row in the file as column names for the target table**.

9 Click **Review Target Column Structure** to display a sample of the target table. The target is displayed with default column names (unless they were specified in the source), types, and lengths (as available according to the type). Review and update the default column definitions as needed, or apply a column definitions file as described in subsequent steps.

TIP To display a larger data sample, click Generate Columns... . In the Generate Columns window, enter a new value for **Number of rows to sample**.



CAUTION! Time and datetime values in the source must be formatted in one of two ways in order for those columns to be assigned the correct type in the target. To accurately assign a column type, the directive requires that the source file use a DATE column format of `YYYY-MM-DD` and a DATETIME column format of `YYYY-MM-DD HH:MM:SS.fffffffff`. The DATETIME format requires either zero or nine decimal places after the seconds value `SS`. Source columns that do not meet these requirements are assigned the VARCHAR type. In the directive, the VARCHAR cannot be changed to a relevant Hadoop type such as DATE or TIMESTAMP.

10 When your columns are correctly formatted, you can save your column definitions to a file. You can then reuse that file to import column definitions the next time you import this or another similar source file. To generate and save a column definitions file, click Save Column Definitions... . In the Save Column Definitions window, enter a filename to generate a new file, or select an existing file to overwrite the previous contents of that file. Click **OK** to save your column definitions to the designated file.

11 If you previously saved a column definitions file, and if you want to import those column definitions to quickly update the defaults, then follow these steps:

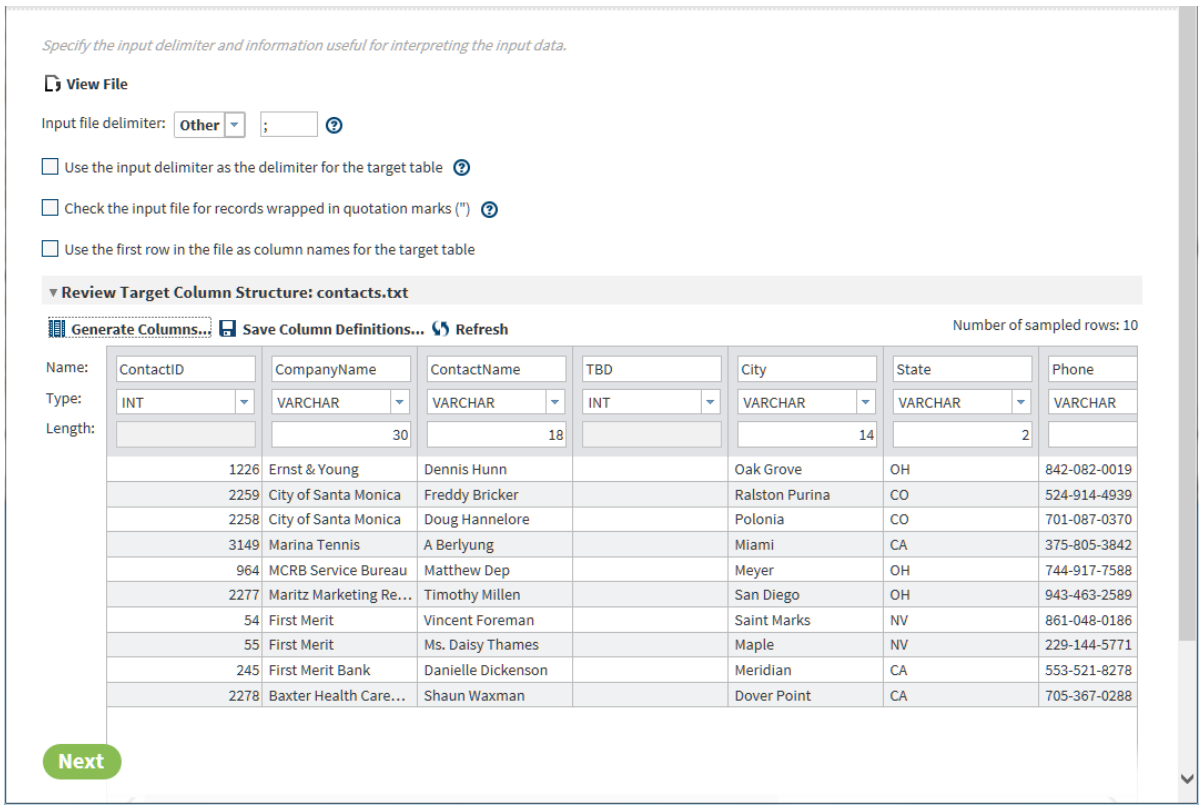
a Click Generate Columns... .

b In the Generate Columns window, click **Use column definitions from a format file**, and enter the filename or select the file using ... to display the Select a Format File window.

c As needed in the Select a Format File window, click to open folders, select a column definitions file, and click **OK**.

> **TIP** Use the Select a Format File to manage your column definitions files (refresh, rename, delete.) You can also download the files as needed.

d In the Generate Columns window, click **Generate** to close the window and format the target columns as specified in the column definitions file.



> **TIP** As is the case with the default column definitions, you can enter changes to imported column names, types, and lengths. You can then save your changes to the original column definitions file or to a new file.

> **TIP** During the definition of columns, you can replace your changes with the default column definitions at any time. Select Generate Columns..., click **Guess the columns based on a sample of data** , and click **Generate**.

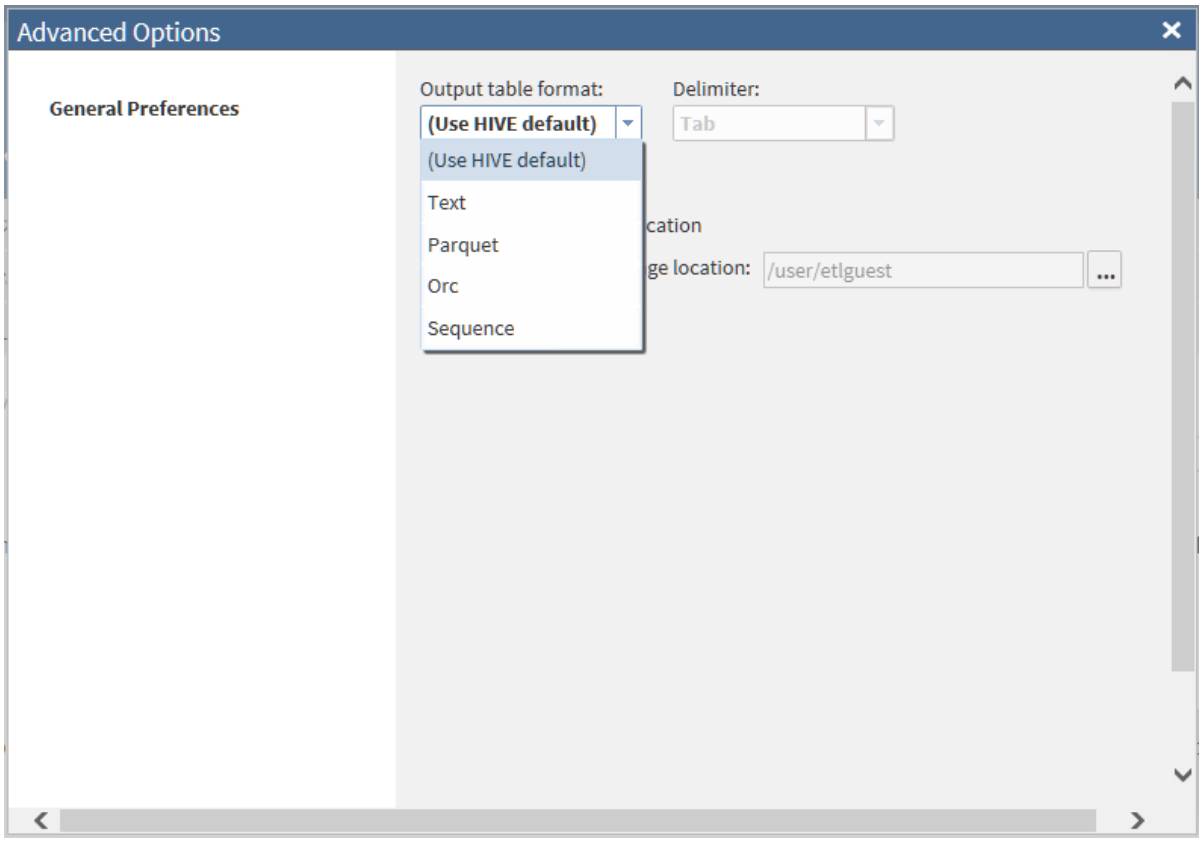
12 In the **Target Table** task, click to open a data source and select a target, or click Select Recent Table and choose a target. Existing targets are overwritten entirely when you run your job.

To name a new target table, select a data source and click New Table..., enter the new table name, and click **OK**.

13 The format of the target table is specified by default for all directives in the Configuration window. To see the default target format, click the **More** icon, and then select **Configuration**. In the Configuration window, click **General Preferences**.

To override the default target file format for this one target, click the target and select **Advanced Options**.

Note: If you are using a bulk-copy operation, as described in Step 6, then the target always receives the Text format, regardless of the selections in the Advanced Options and Configuration windows.



To browse a non-default Hive storage location, click **Specify alternate storage location**, and then click ... . You need appropriate permission to store your imported table or file to a non-default Hive storage location.

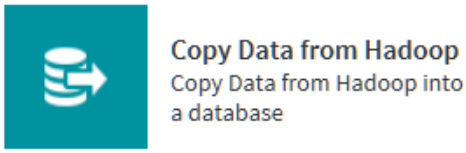
When your target selection is complete, click **Next**.

14 In the **Result** task, click **Start Importing Data** to generate code and execute your job. You can monitor long-running jobs in the Run Status directive. At the completion of execution, you can click the **Code**, **Log**, and possibly the **Error Details** icon to learn more about your job.

15 Click **Save** or **Save As** to retain your job in **Saved Directives**.

Copy Data from Hadoop

Introduction



The Copy Data from Hadoop directive enables you to copy data from Hadoop into database management systems such as Oracle and Teradata.

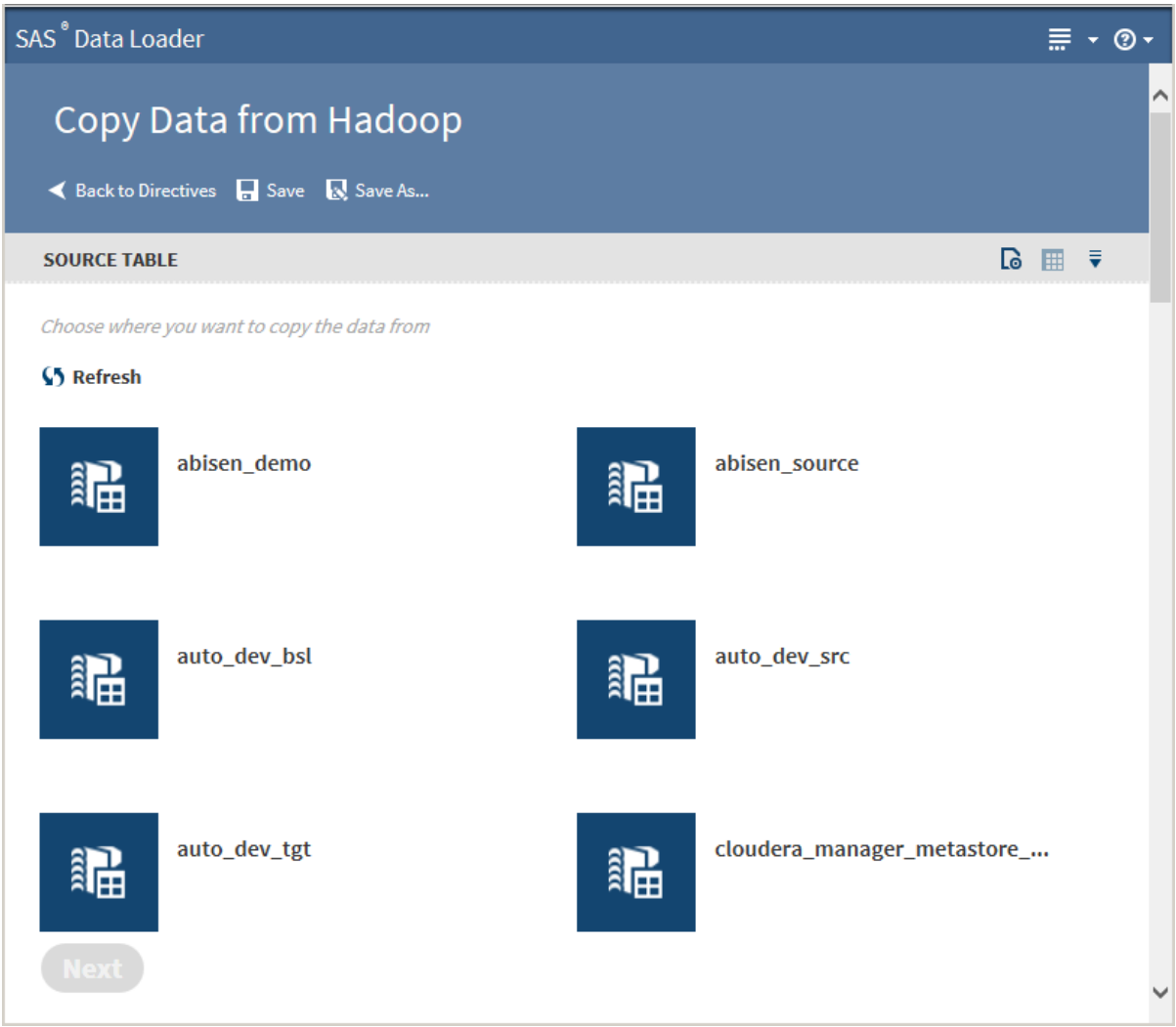
Prerequisites

When you open the Copy Data from Hadoop directive, the **Source Tables** task shows the data sources that are on the Hadoop cluster. When you come to the **Target Tables** task, you will see the databases that are defined in SAS Data Loader. If you do not see the database to which you want to copy, you must add a connection to that database. See "Databases Panel" on page 139 for more information.

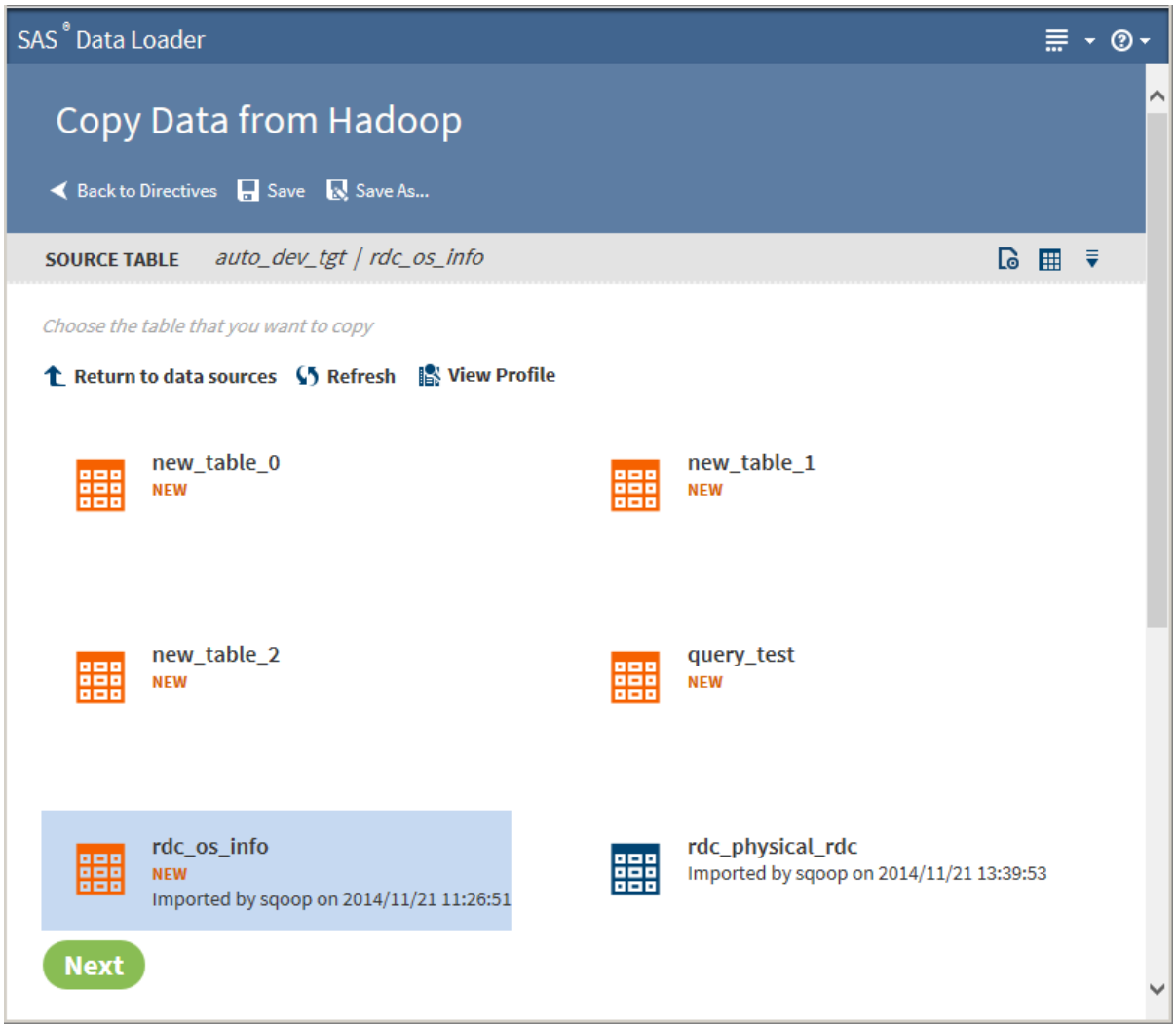
Example

Follow these steps to copy data from Hadoop into a database:

1 On the SAS Data Loader directives page, click the Copy Data from Hadoop directive. The **Source Table** task that lists available data sources is displayed:



2 Click a data source to display its tables:



3 Select the table from which to copy data.

Clicking the **Action** menu enables the following actions:

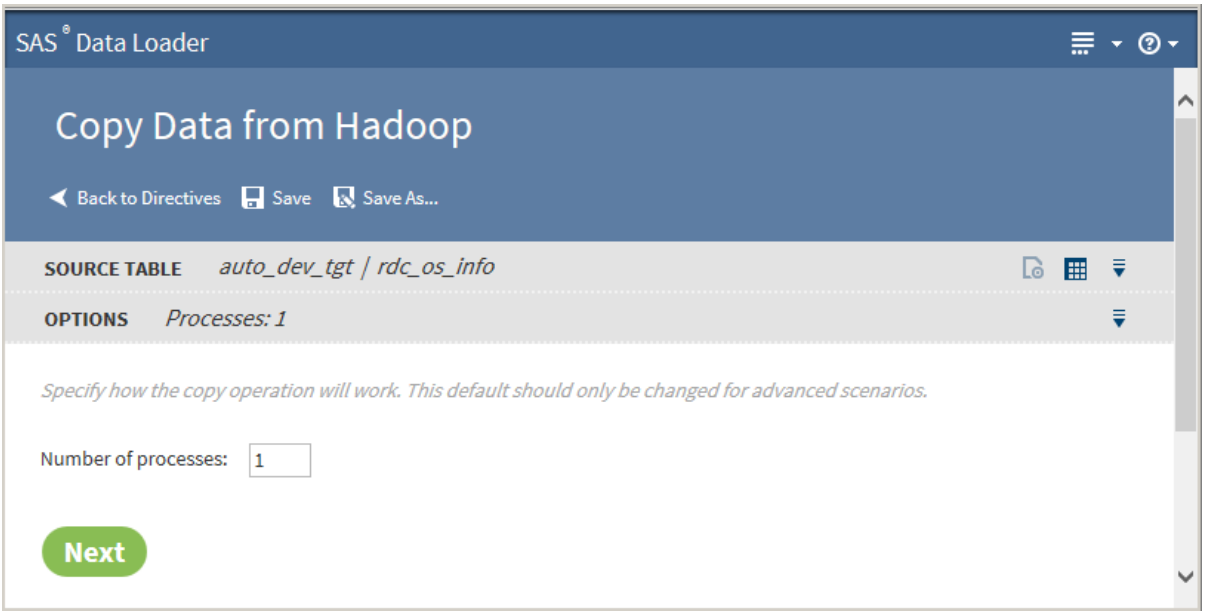
Open
opens the current task.

Table Viewer
enables you to view sample data from a table. Select the table, and then click to display SAS Table Viewer.

Advanced Options
opens a dialog box that enables you to specify the maximum length for SAS columns. Entering a value here overrides the value specified in the **Configuration** options.

Note: If the source table has String data types, the resulting SAS data set could be very large. The length of the target field in the SAS data set is determined by the value of this option.

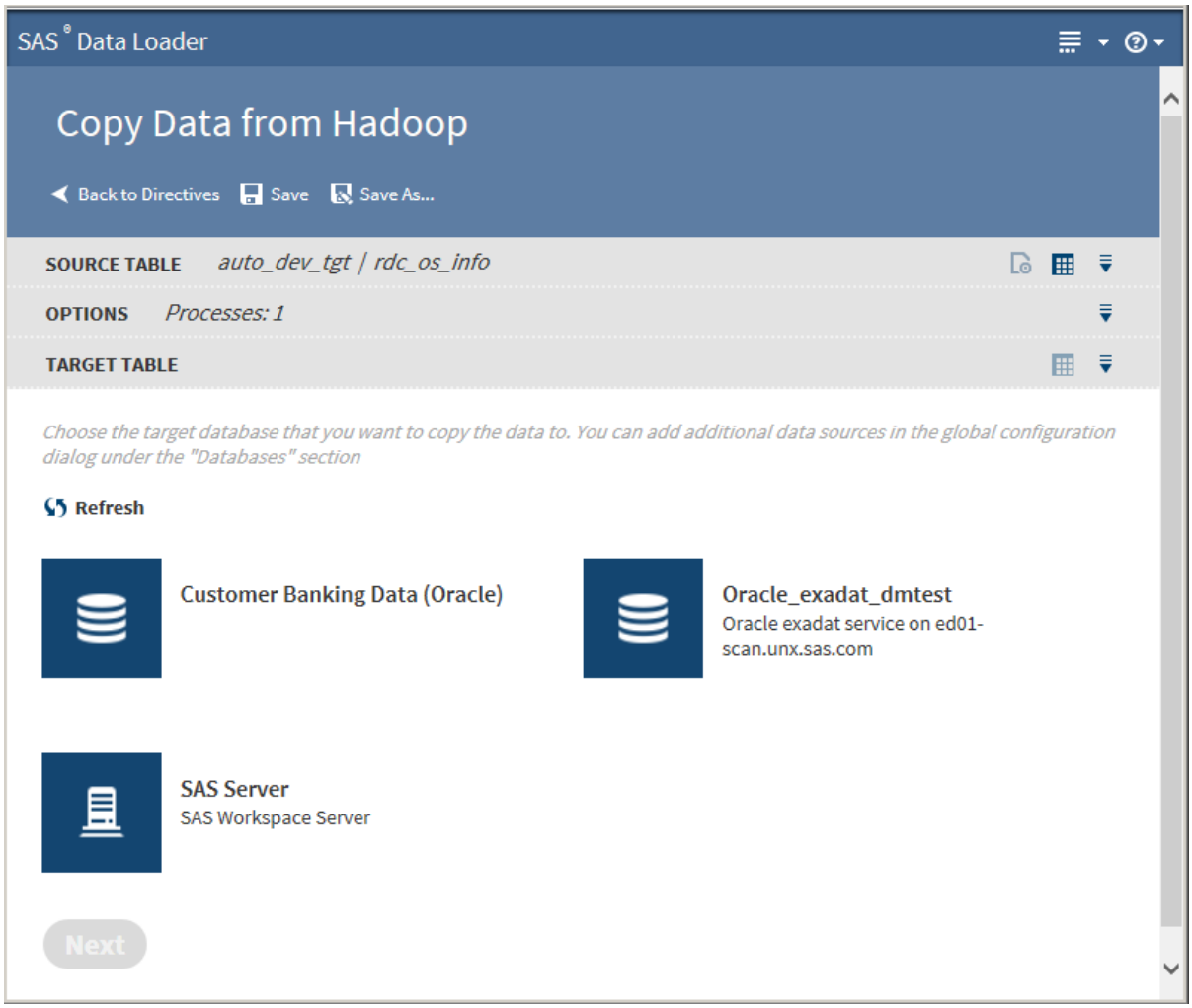
When table selection is complete, click **Next**. The **Options** task is displayed:



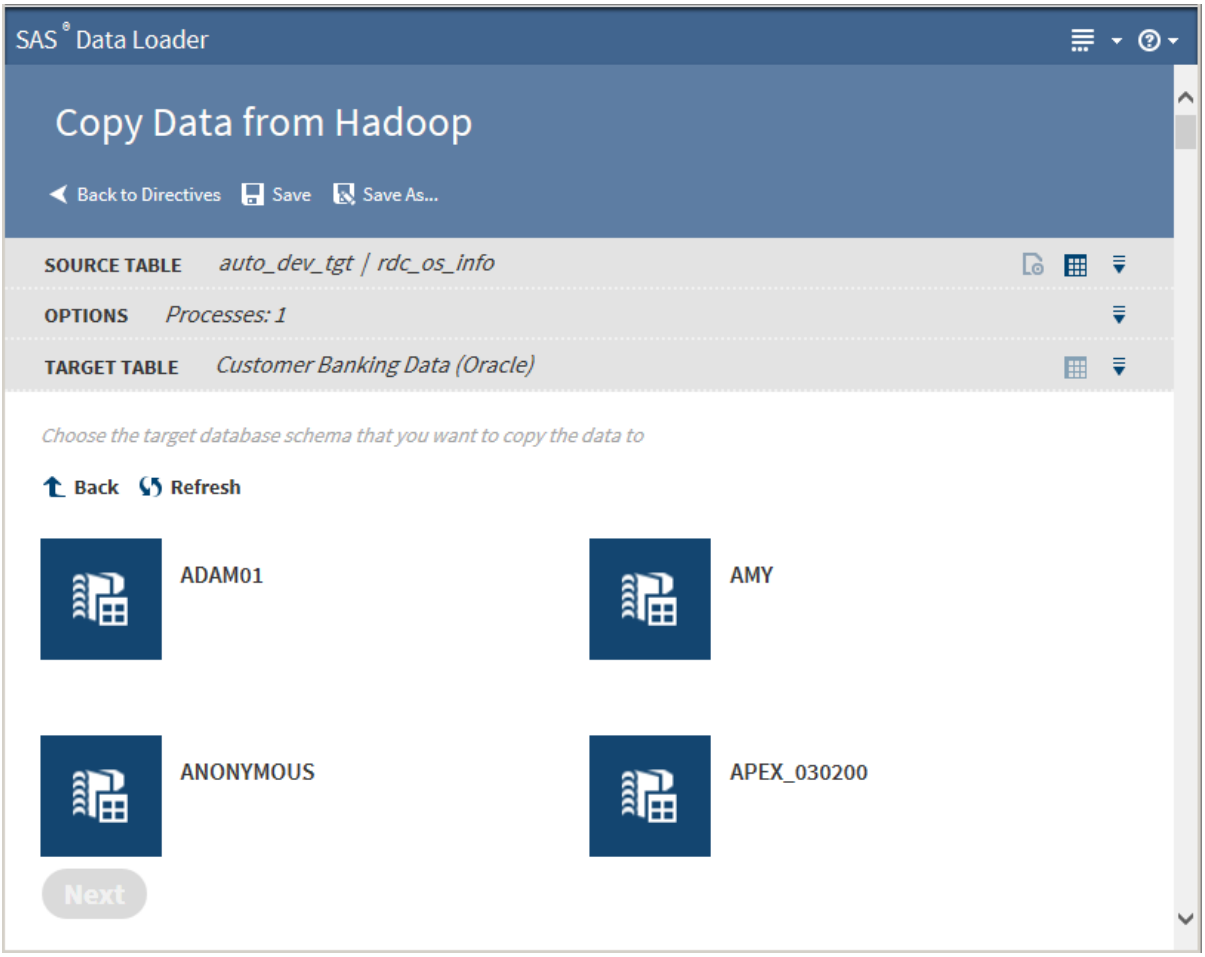
4 The value on the **Options** task should not be changed unless you have advanced knowledge of database operations.

Note: Changing the number of processes to greater than one expands the number of processes and source data connections that are used to import data.

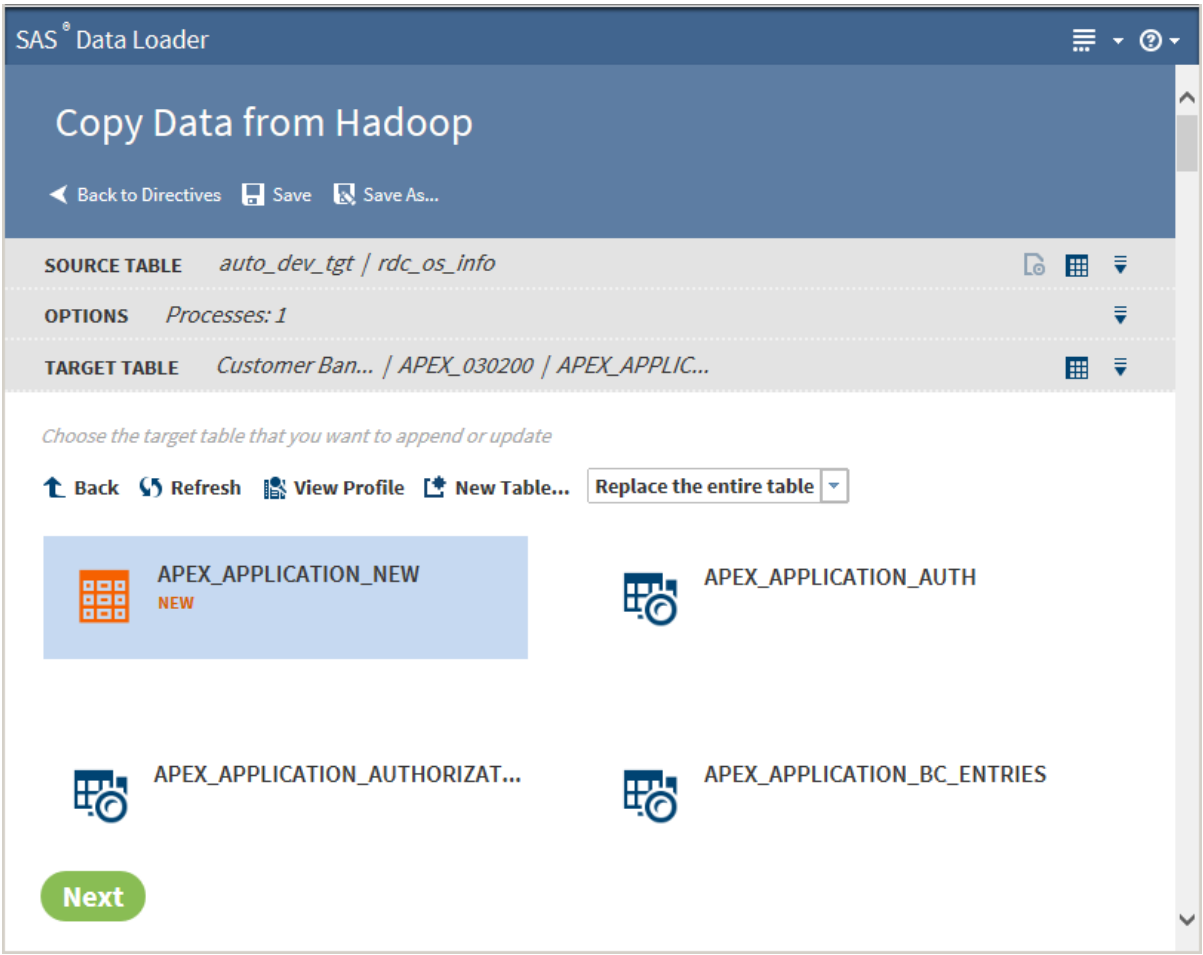
Click **Next**. The **Target Table** task is displayed with target databases:



5 Click a database to display its data sources:



6 Click a data source to display its tables:



7 Select the table from which to copy data.

TIP

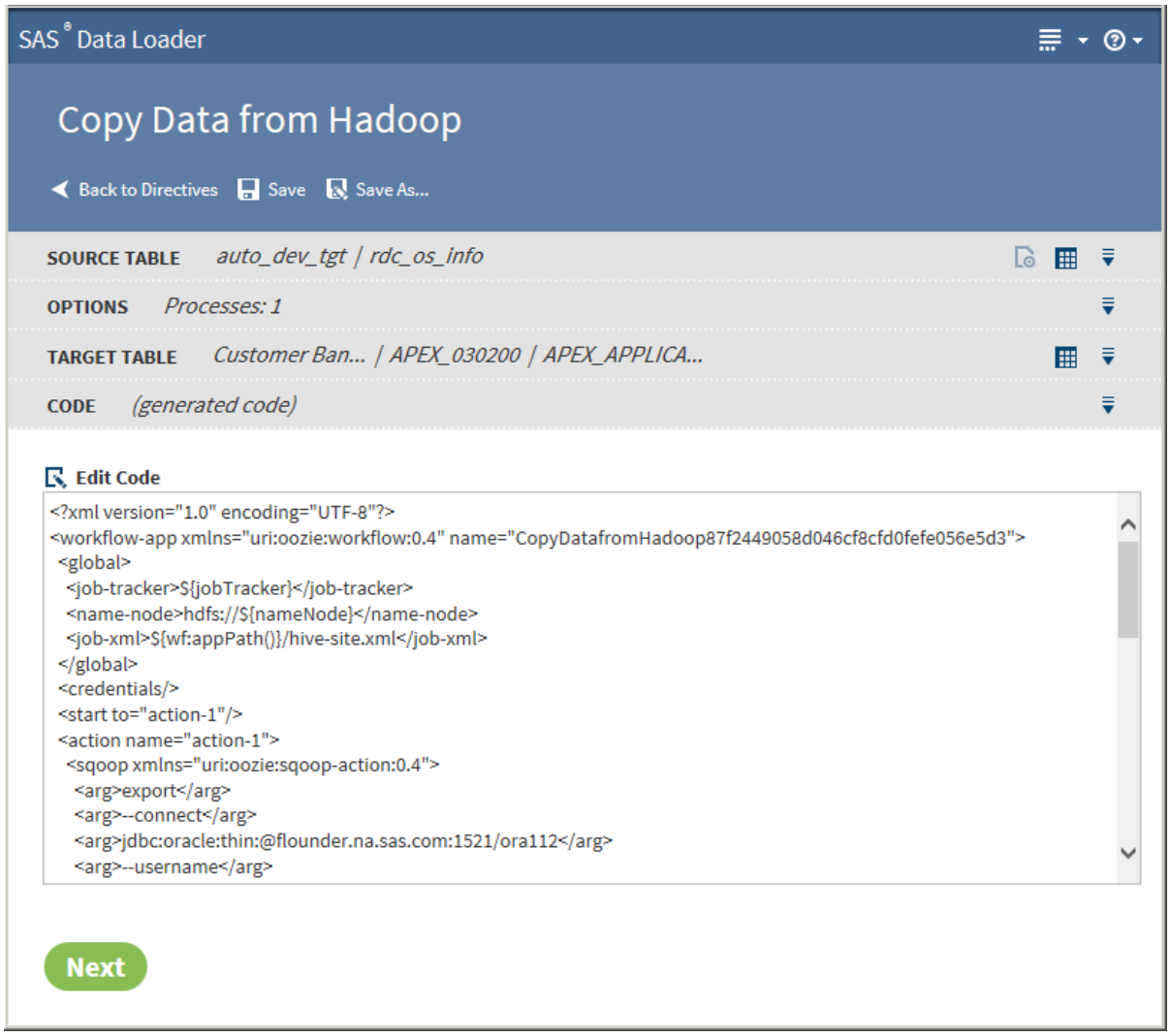
- You can create a new table by clicking **New Table**.
- If a profile already exists for a table, PROFILED appears next the table icon. You can view the existing profile by selecting the table and clicking **View Profile**.

Click to enable the following actions:

Open
opens the current task.

Table Viewer
enables you to view sample data from a table. Select the table, and then click to display SAS Table Viewer.

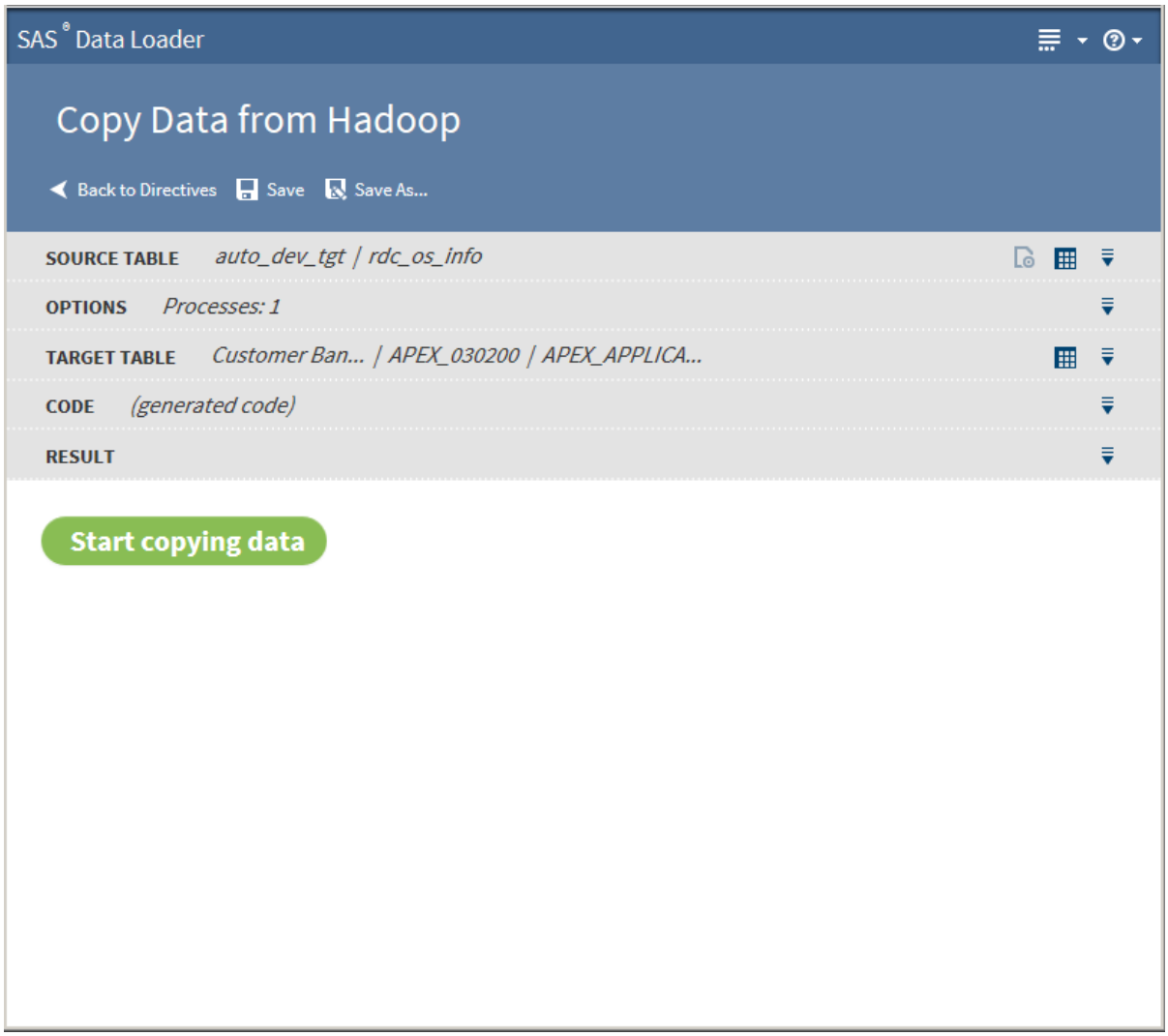
Click **Next**. The **Code** task is displayed:



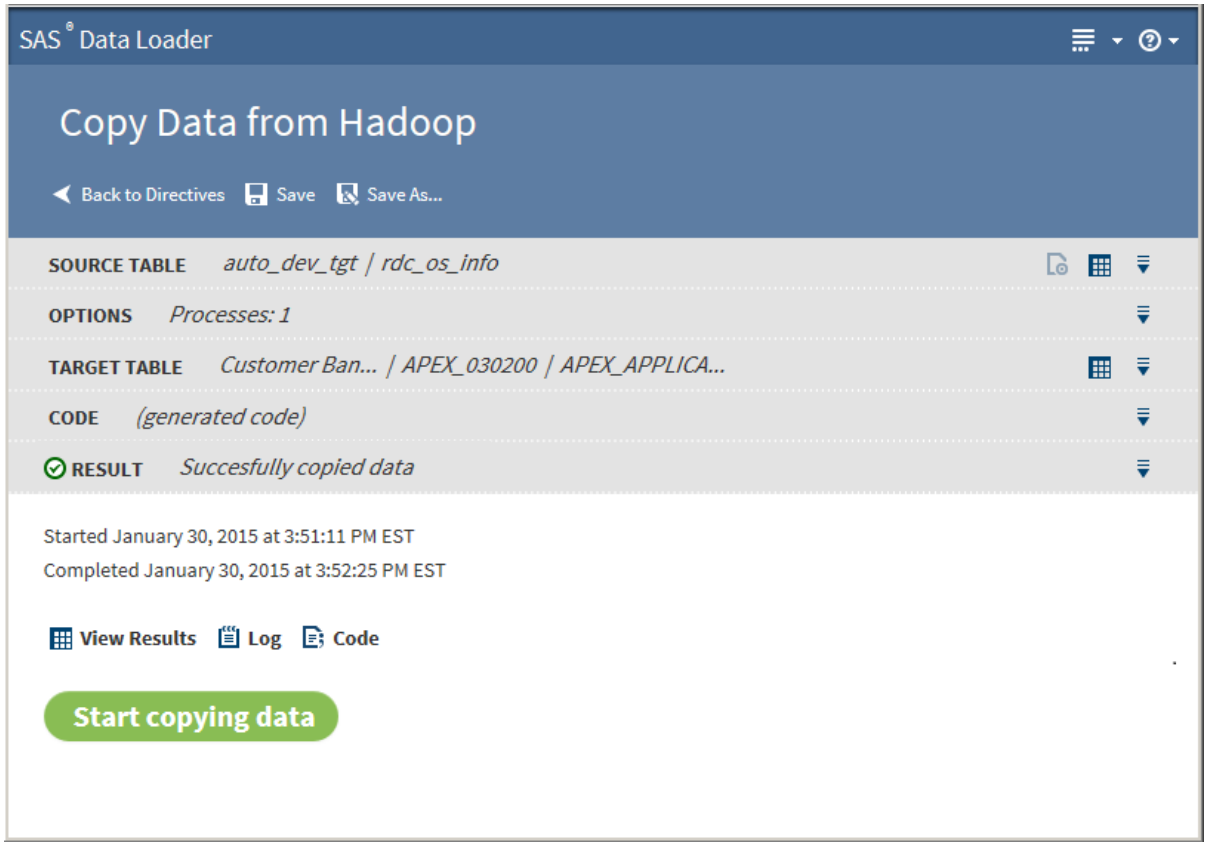
8 Click **Edit Code** to modify the generated code. To cancel your modifications, click **Reset Code**.

CAUTION! Edit code only to implement advanced features. Under normal circumstances, code edits are not needed or required.

9 Click **Next**. The **Result** task is displayed:



10 Click **Start copying data**. The **Result** task displays the results of the copy process:



The following actions are available:

View Results
enables you to view the results of the copy process in the SAS Table Viewer.

Log
displays the SAS log that is generated during the copy process.

Code
displays the SAS code that copies the data.

Usage Notes

- When copying a Hadoop table to a Teradata database, the name of the target table cannot exceed 20 characters. Longer target names cause the job to fail.
- If necessary, you can change the maximum length of character columns for input tables to this directive. For more information, see "Change the Maximum Length for SAS Character Columns" on page 147.
- Source tables with a large number of columns can cause Copy From Hadoop jobs to fail. The job runs until the target table reaches the maximum number of columns that are supported in the target database. To resolve the problem, reduce the number of columns that are selected for the target and run the job again.
- If one or more VARCHAR or STRING columns from a source Hadoop table contains more string data than the target database column, the Copy Data from Hadoop request times out. For example, a source Hadoop table might

contain a string column named myString and a target Oracle table might contain a varchar(4000) column also named myString. If data in the Hadoop myString column has a length greater than 4000, then the copy request fails.

- When copying a Hadoop table to a database, a column name specified in the array of STRUCT in the Hadoop table is not copied to the database table. This happens because of how STRUCT is mapped to VARCHAR in Sqoop.
- A copy from Hadoop is likely to fail if the name of a source column is also a reserved word in the target database.
- When copying a Hadoop table to Oracle, a mixed-case schema name generates an error.
- When copying a Hadoop table to Oracle, timestamp columns in Hadoop generate errors in Oracle. The Hive timestamp format differs from the Oracle timestamp format. To resolve this issue, change the column type in the Oracle target table from timestamp to varchar2.
- To copy Hadoop tables to Teradata, when the source contains a double-byte character set (DBCS) such as Chinese, follow these steps:

  1. Edit the default connection string to include the option `charset=utf8`, as shown in this example:

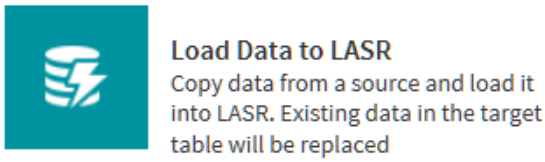
  ```
  jdbc:teradata://TeradataHost/Database=TeradataDB,charset=utf8
  ```

     To edit the configuration string, open the Configuration window , click **Databases**, and click and edit the Teradata connection.

  2. Ensure that the default character type for the Teradata user is UNICODE.
  3. In new Teradata tables, set VARCHAR CHAR columns to CHARACTER SET UNICODE to accommodate wide characters.

Load Data to LASR

Introduction



Use the Load Data to LASR directive to copy Hadoop tables to a single SAS LASR Analytic Server, or to a grid of SAS LASR Analytic Servers. On the SAS LASR Analytic Servers, you can analyze tables using software such as SAS Visual Analytics.

When you load data onto a single SAS LASR Analytic Server, you configure a connection that is optimized for symmetric multi-processing (SMP). When you load data onto a grid of SAS LASR Analytic Servers, you configure a connection that is optimized for massively parallel processing (MPP).

Note: The Load Data to LASR directive is distinct and separate from the Load to LASR capability that is provided by SAS LASR Analytic Server.

Prerequisites

In order to use the Load Data to LASR directive, you must specify a connection in the **LASR Analytic Server** panel of the Configuration window. For more information, see "LASR Analytic Servers Panel" on page 134.

Example

Follow these steps to create and run the Load Data to LASR directive:

1. In the Directives directives page, click **Load Data to LASR**.
2. In the Source Table task, click the schema that contains the source table that you want to load. Clicking the schema displays the tables in that schema. Click the table that you want to load into the SAS LASR Analytic Server software, and then click **Next**.
3. In the Target Table task, click the SAS LASR Analytic Server that you want to receive the target table. Clicking displays target table configuration fields and controls.
4. As needed, change the name in the **Target table name** field. The field defines the name of the table in the SAS LASR Analytic Server software.
5. Select options as needed to replace any existing table of the same name or to compress the target table in the SAS LASR Analytic Server software.
6. Click the **Locations** link to view or change the default storage options for the target table in the SAS LASR Analytic Server software.
7. In the Locations window, you can change the SAS folder, the library name, and the required tag that accompanies the table name.
8. In the Target Table task, click **Next**.
9. In the Result task, click **Start loading data**. SAS proceeds to generate code for the directive and displays the **Code** icon . Click the icon to open or save the text of the SAS code that comprises the directive.
10. During the execution of the directive, the **Result** task displays the **Log** icon . Click the icon to open or save the SAS log file that is generated during the execution of the directive.
11. At the conclusion of the directive, the Result banner receives a status icon that indicates the success or failure of the directive. To view the target table on the SAS LASR Analytic Server, click the **View Results** icon .

Usage Notes

In MapR distributions of Hadoop, massively parallel processing (MPP) is not supported in the LASR procedure. To load data from MapR Hadoop to a SAS LASR Analytic Server, the server definition must assert the SASIOLA option.

The SASIOLA option implements symmetric multiprocessing (SMP.) Server definitions are available in the SAS Data Loader Configuration window, in the **LASR Analytic Servers** panel. For more information about server definitions, see "Add or Update Connections to SAS LASR Analytic Servers" on page 137.

The Load Data to LASR directive moves entire tables. To improve performance, you can filter the rows and manage the columns before you load the table to the SAS LASR Analytic Server. To reduce table size, use the directives "Transform Data in Hadoop" or "Query or Join Data in Hadoop".

7

Run User-Written Programs in Hadoop

Overview 121

Run a SAS Program 121
Introduction 121
Example 122

Run a Hive Program 123
Introduction 123
Example 123

Overview

New file needs edit. The directives Run a SAS Program and Run a Hive Program enable you to execute existing code in SAS Data Loader. You can also create and execute new programs. SAS Data Loader enables you to receive and retain log files and save jobs for reuse.

Run a SAS Program

Introduction

The Run a SAS Program directive provides the primary means of submitting user-written SAS code in SAS Data Loader for Hadoop. The code runs as you submit it, without the code generation step that is used in other directives. The code that you submit generates the same log and error information as in other directives. Also, the running code is tracked in the Run Status directive, and you can save and reuse jobs in Saved Directives.

The code execution process begins and ends in the vApp. The Workspace Server inside the vApp runs the code and executes all Base SAS language

elements. If your code contains procedures that are enabled for DS2, or if your code contains native DS2 methods, then that code might be passed into the Hadoop cluster for execution. In your Hadoop cluster, DS2 code is executed by the SAS In-Database Code Accelerator for Hadoop.

Upon completion of DS2 execution on the cluster, the vApp receives notification and continues or concludes execution in the local Workspace Server.

For examples of DS2-enabled SAS code, refer to the code that is generated by directives such as Transform Data in Hadoop.

CAUTION! Data sets in Hadoop are of indeterminate size. Any data that is indiscriminately returned from Hadoop to the vApp can overload the client. To avoid overloading the vApp, your SAS programs need to minimize or eliminate the transfer of data from Hadoop to the vApp. It is generally preferable to define a result set or target table that remains in Hadoop. You can then analyze the data in Hadoop, or load data for further analysis onto a grid of SAS LASR Analytic Servers.

Note that you can generate code in any of the following software, and copy and paste that code into the Code task of the Run a SAS Program directive:

- SAS Data Management Studio
- SAS Enterprise Guide
- SAS Data Integration Studio

Conversely, you can copy the code that is generated in any SAS Data Loader directive and paste into any SAS text editor. One suggested location for pasting SAS Data Loader code is the SAS Code Node in DataFlux Data Management Studio.

To include DS2 syntax in your SAS programs, you can use a number of SAS procedures that support DS2 language elements, as described in the *SAS In-Database Products: User's Guide*. For DS2 syntax information, see the *SAS DS2 Language Reference*.

To run DS2 code directly in Hadoop using the SAS In-Database Code Accelerator, see the “SAS In-Database Code Accelerator for Hadoop” section of the *SAS In-Database Products: User's Guide*.

User-written SAS DS2 code can be submitted in an expression builder in the following directives:

- Delete Rows
- Cleanse Data in Hadoop (Filter Transformation)
- Transform Data in Hadoop (Filter Data task)

Example

Follow these steps to use the Run a SAS Program directive:

1. In the SAS Data Loader directives page, click **Run a SAS Program**.
2. In the **Code** task, enter SAS code, or right-click to cut and paste existing SAS code.

> **TIP** The pop-up menu enables you to display line numbers and to navigate to the beginning or the end of the program.

3 When your program is ready to run, click **Next**.

4 In the **Result** task, click **Start SAS program**. As your program runs, you receive start and end date/time information, along with **Log**, **Code**, and possibly **Error Details** icons. Click the icons as needed to resolve errors.

   The final status of the job is displayed in the **Result** taskbar.

5 Click **Save** to save your program for reuse. To edit or run your job in the future, go to the SAS Data Loader directives page and click **Saved Directives**.

Run a Hive Program

Introduction

Use the Run a Hive Program directive to create jobs that execute Hive programs in Hadoop. The directive enables you to browse available Hive functions and click to add function syntax into a text editor. You can also copy and paste existing Hive programs directly into the text editor. The user credentials that are specified in the Hadoop Configuration panel of the SAS Data Loader Configuration window are used to submit Hive code to the Hadoop cluster.

Note: User-written Hive code can also be submitted using the directives Delete Rows, Query or Join Data in Hadoop, or Sort and De-Duplicate Data in Hadoop. These directives provide a Hive expression builder that is used to filter or delete rows.

Example

Follow these steps to use the Run a Hive Program directive:

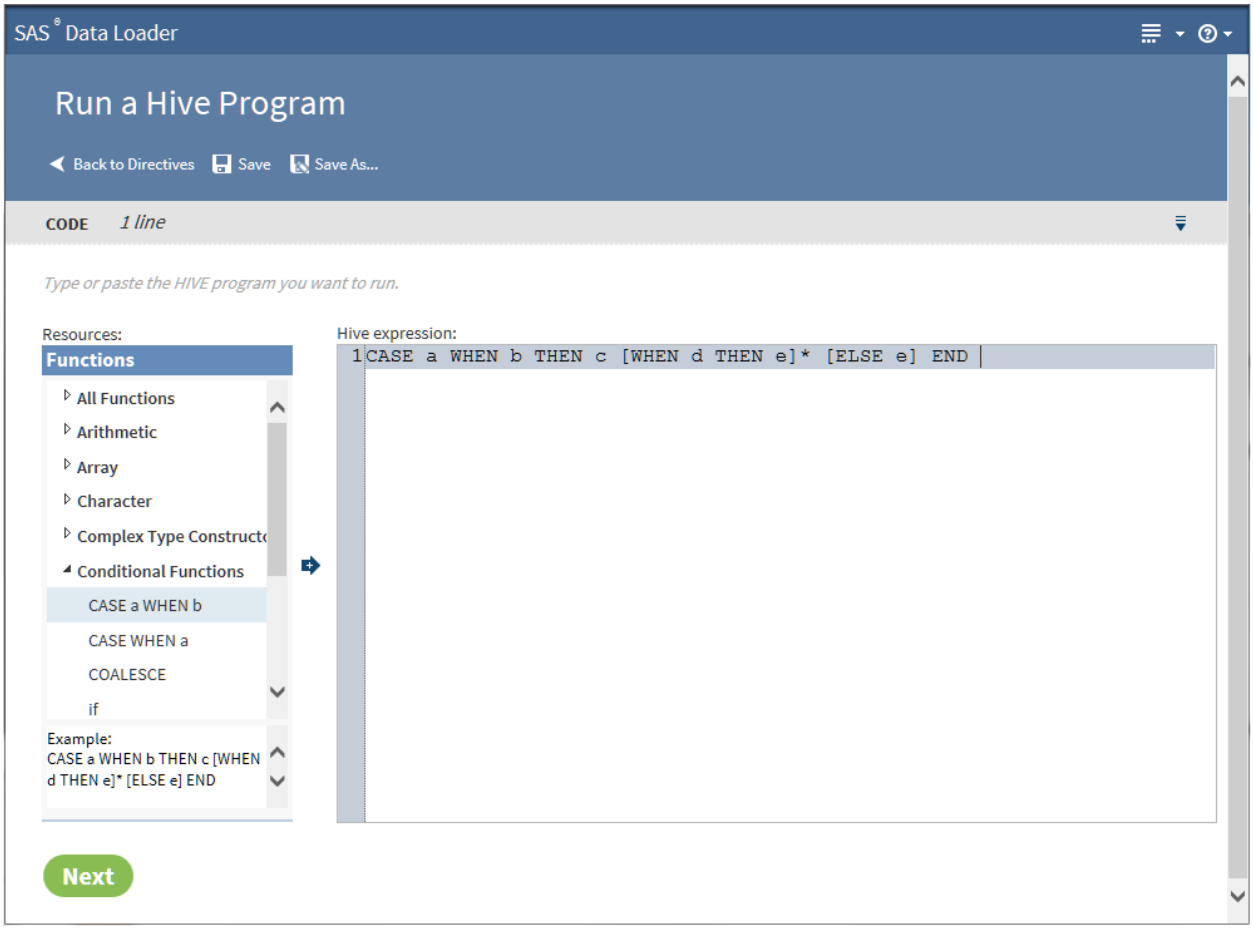
1 In the SAS Data Loader directives page, click **Run a Hive Program**.

2 In the **Code** task, click **Hive expression** and enter Hive code, or right-click to cut and paste existing Hive code from your file system.

   Note:

   - The pop-up menu also enables you to display line numbers and to navigate to the beginning or the end of the program.
   - The Hive program needs to explicitly define data sources and targets.

3 To add Hive functions to your program, click in **Resources** to expand categories and to display examples of function syntax. To move function syntax into your program, click the function and click .



4 When your program is ready to run, click **Next**.

5 In the **Result** task, click **Start HIVE program**. As your program runs, you receive start and end date and time information, along with **Log**, **Code**, and possibly **Error Details** icons. Click the icons as needed to resolve errors.

   The final status of the job is displayed in the **Result** taskbar.

6 Click **Save** to save your program for reuse. To edit or run your job in the future, go to the SAS Data Loader directives page and click **Saved Directives**.

8

Manage Jobs

Overview of Job Management Directives 125

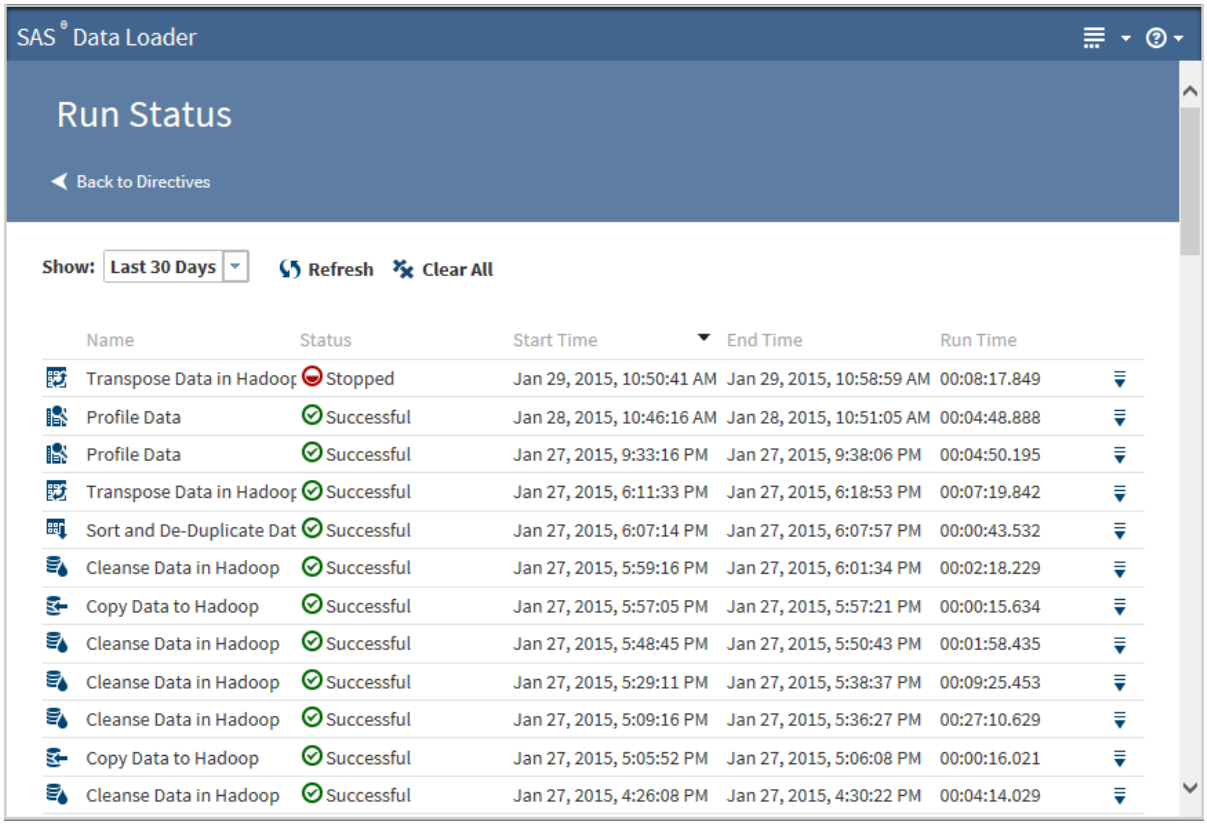
Run Status 126
Introduction 126
Using Run Status 126
About Unsaved Jobs 128
About Incomplete Jobs 128

Saved Directives 128
Introduction 128
Opening Saved Directives 128
Managing Saved Directories 129

Overview of Job Management Directives

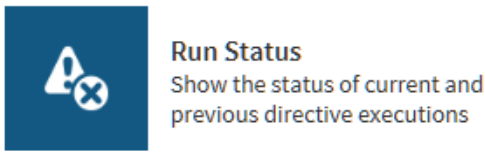
The job management directives enable you to view the status of current and previous jobs and to modify and execute saved directives. The Run Status directive displays information about the current execution state of jobs. The Saved Directives directive enables you to open, edit, and manage your existing directives.

Here is an example of the Run Status directive:



Run Status

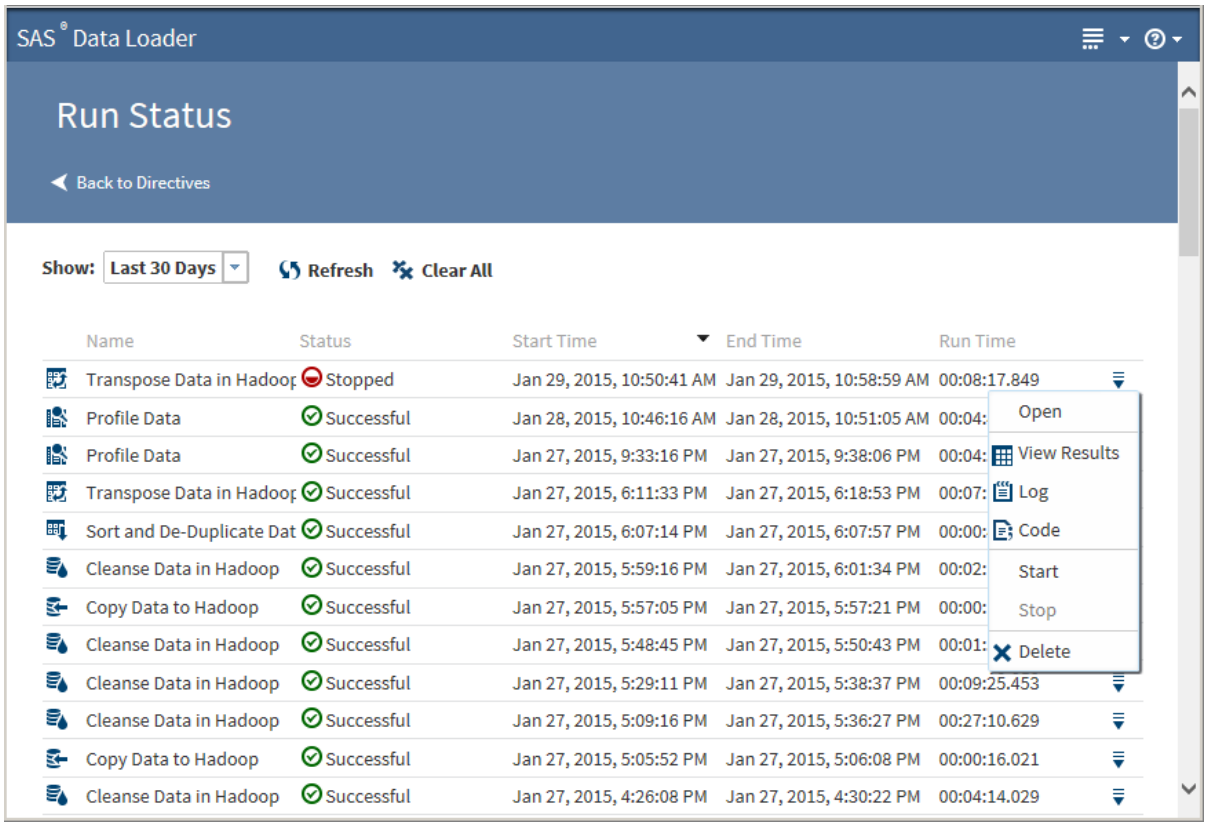
Introduction



Use the Run Status directive to view job runs. Each run is listed with its current execution status, start time, end time, and run time. The Status column value can be In Progress, Stopped, Failed, or Successful.

Using Run Status

In the SAS Data Loader directives page, click the Run Status directive. The Run Status page is displayed:



By default, the Run Status page displays all of the directives that have run in the past 30 days. The most recent runs appear at the top of the list. You can change the default of 30 days by selecting a new value from the **Show** drop-down list. Reports are identified by the given name or by the generic name of the directive (for example, Transform Data in Hadoop.) Given names are created when you save a directive.

When you click **Refresh**, you receive updates for all running jobs, including any that were started or completed after you opened the Run Status page.

Clicking **Clear All** clears all of the reports from the Run Status page. Clearing reports permanently removes the reports from the vApp database.

Clicking the **Action** menu for a job in the list enables the following actions:

Open
opens the directive associated with the job.

View Profile Report
for successful Profile Data jobs, enables you to view the Profile Report unless the report has been deleted from the Saved Profiles directive. See "Saved Profile Reports" on page 80 for more information about the profile report.

View Results
for completed transformations or queries, enables you to view a sample of the target table in the SAS Table Viewer.

Log
displays the SAS log that is generated during the execution of the profile job.

Code
displays the SAS code that is generated during the execution of the profile job.

Start
starts a failed or successful job.

Stop
stops an in-progress job.

Note: If you select **Stop**, your directive continues to display its In Progress status. In this situation, the directive is stopping, but it has not yet reached a suitable stopping point. Click **Refresh** periodically until the status changes to Stopped or reopen Run Status later to confirm the Stopped status.

Delete
clears a single report from the Run Status page. Clearing a report permanently removes the report from the vApp database.

About Unsaved Jobs

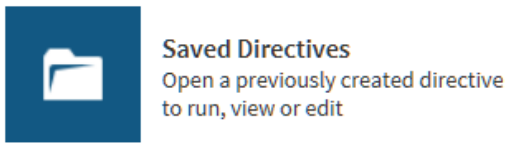
If you run a directive without saving it, the directive is displayed in Run Status like any other directive. When processing stops on the unsaved directive, you can select **Open** from its Action menu. You can then edit and save the unsaved directive.

About Incomplete Jobs

An incomplete job is one that you have stopped using the **Action** menu or one whose status is Failed. Depending on the type of the job and the point where execution ceased, log and code results might or might not be available.

Saved Directives

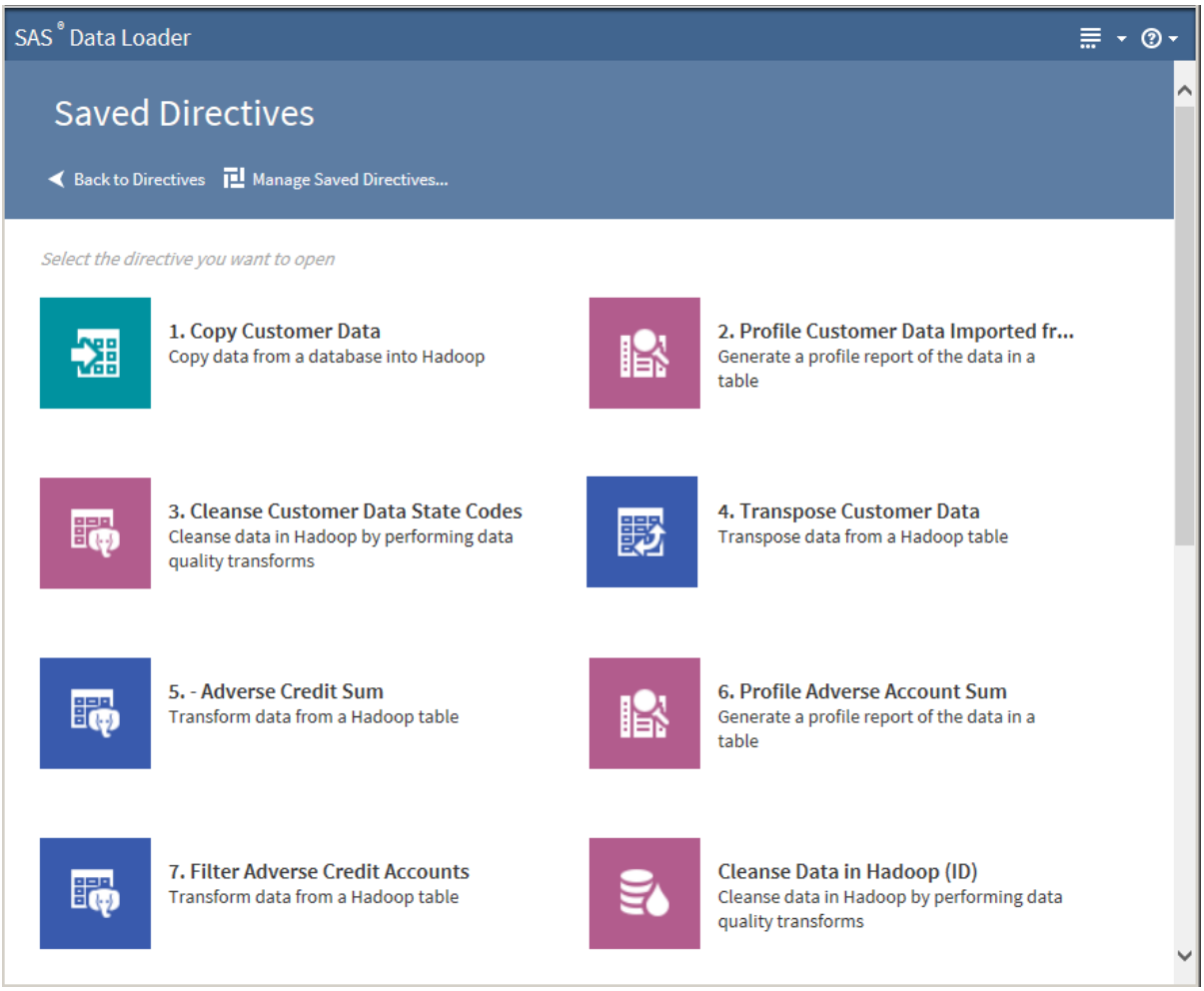
Introduction



Use Saved Directives to open, edit, and execute your saved directives. From the Saved Directives page, opening Manage Saved Directories enables you to open, duplicate, delete, refresh, or rename the selected directive.

Opening Saved Directives

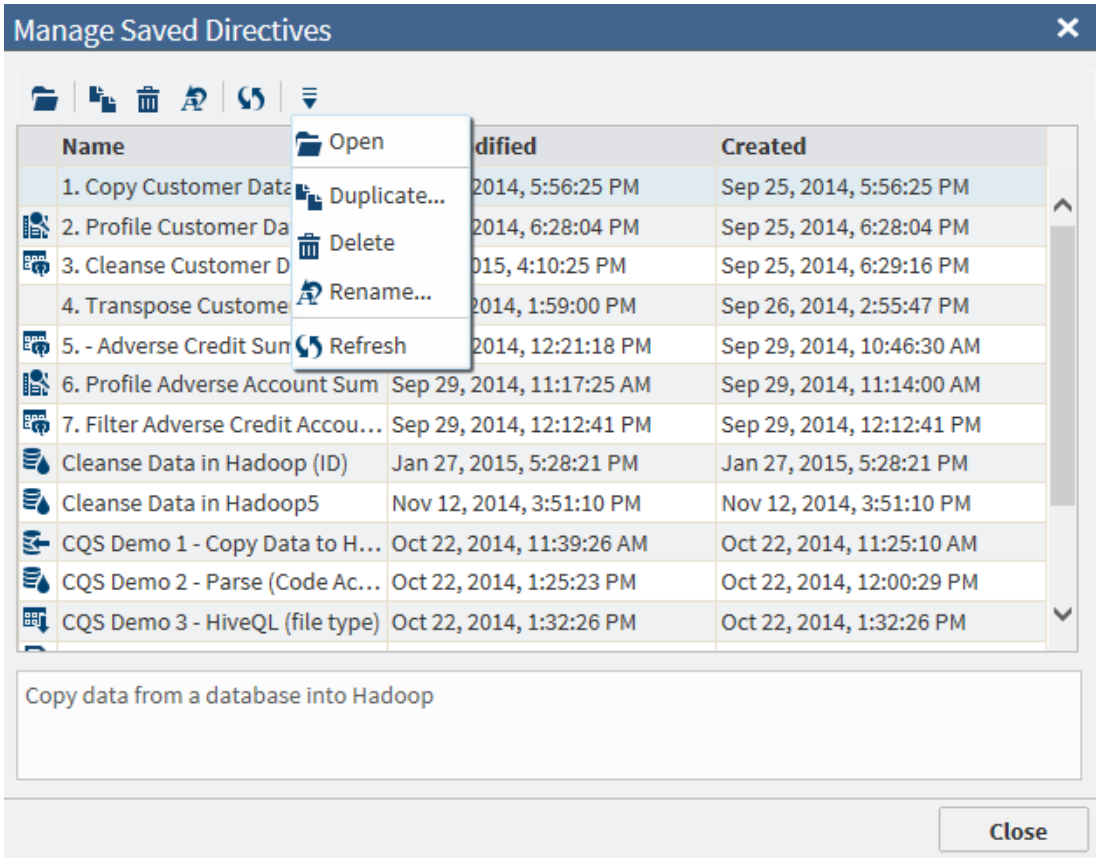
In the SAS Data Loader directives page, click the Saved Directives directive. A Saved Directives page similar to the following is displayed:



In the Saved Directives page, click a saved directive. The directive opens and can be edited or executed.

Managing Saved Directories

In the Saved Directives page, click **Managed Saved Directives** . The Managed Saved Directives dialog box appears:



Clicking the **Action** menu enables the following actions:

Open
opens the selected directive.

Duplicate
duplicates the selected directive by opening a dialog box that enables you to assign a new name to the duplicated directive.

Rename
renames the selected directive.

Delete
deletes the selected directive.

Refresh
refreshes the selected directive, or, if no directive is selected, refreshes all of the saved directives in the list. Any duplicate, rename, or delete actions that you have taken are then reflected in the saved directives list.

9

Maintaining SAS Data Loader

Back Up Directives 131

Set Global Options 132
Overview of the Configuration Window 132
Hadoop Configuration Panel 132
LASR Analytic Servers Panel 134
Databases Panel 139
QKB Panel 142
Profiles Panel 142
General Preferences Panel 143

Troubleshooting 144
Active Directory (LDAP) Authentication 144
Change the File Format of Hadoop Target Tables 145
Change the Maximum Length for SAS Character Columns 147
Change the Temporary Storage Location 147
Discover New Columns Added to a Source after Job Execution 148
Hive Limit of 127 Expressions per Table 148
Overriding the Hive Storage Location for Target Tables 148
Unsupported Hive Data Types and Values 149
Restarting a Session after Time-out 149

Back Up Directives

You can back up your saved directives for later use, in case you need to restore the vApp for SAS Data Loader.

To perform a backup, click the More icon on the SAS Data Loader page, and then select **Back Up Directives**.

Note:

- To perform a backup successfully, the backup location must be defined in the vApp settings before SAS Data Loader is started. For more information about setting up the backup location and restoring data, see *SAS Data Loader for Hadoop: vApp Deployment Guide*.
- SAS Data Loader supports only one backup. If you perform a backup, any previous backup data in the backup location is overwritten with the new data.

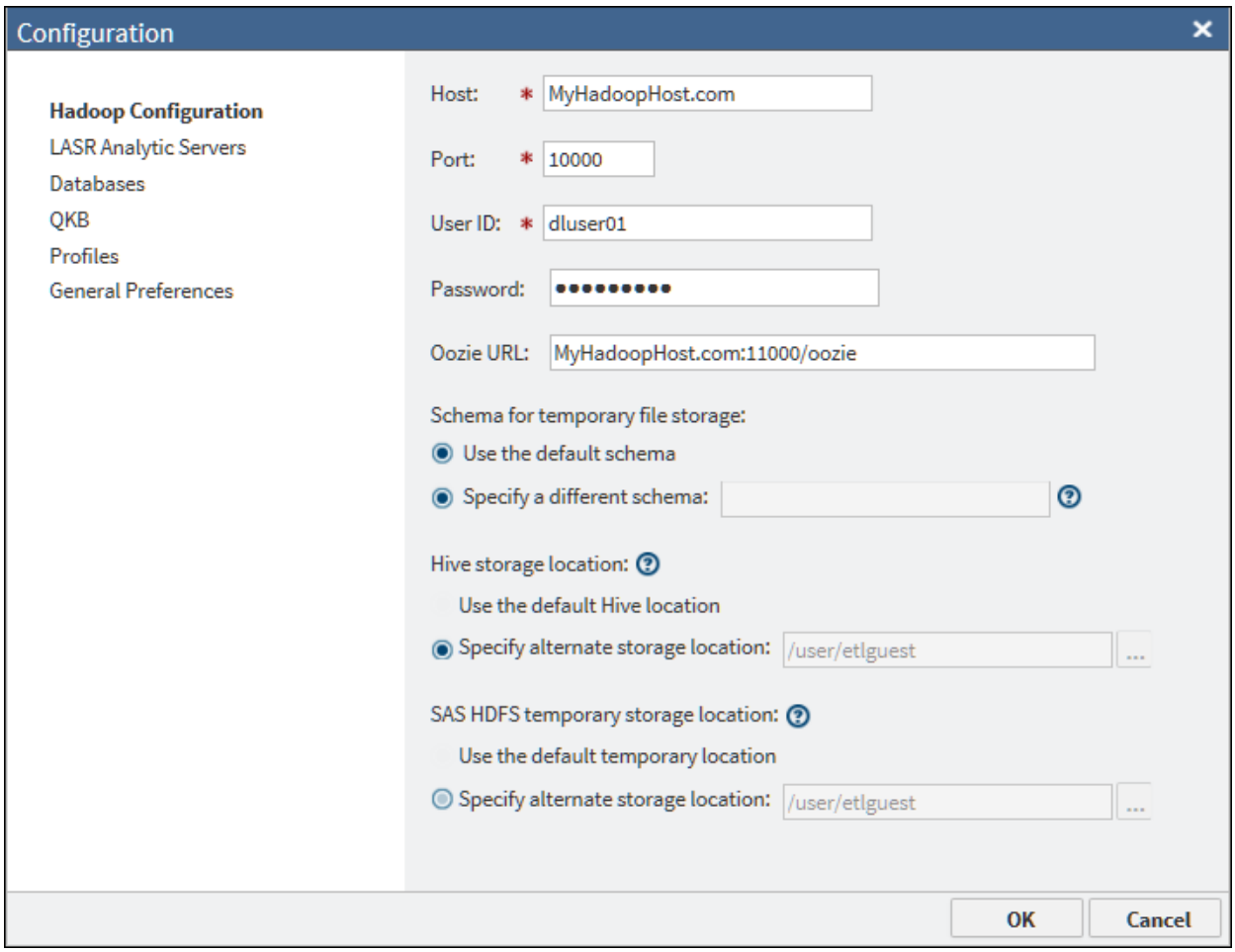
Set Global Options

Overview of the Configuration Window

You can use the Configuration window to specify server connections, data sources, global options, and other settings for SAS Data Loader. To display this window, click the More icon in the top right corner of SAS Data Loader. Then select **Configuration**. See the following topics for details about the options in each panel of the window.

Hadoop Configuration Panel

SAS Data Loader enables you to easily access, transform, and manage data stored in Hadoop. You can use the **Hadoop Configuration** panel of the Configuration window to specify a connection to a Hadoop cluster. You can also use it to change the default storage locations for various types of files.



The values for **Host**, **Port**, **User ID**, **Password**, and **Oozie URL** are entered in the vApp during the initial setup of SAS Data Loader, as described in the *SAS Data Loader for Hadoop: vApp Deployment Guide*. Typically, you will not change

these values except in consultation with your Hadoop administrator. The default values in the storage location fields will work in many cases, but you can change one or more of these locations for your site.

Note: To reconfigure SAS Data Loader for a different Hadoop cluster, you must copy a new set of configuration files and JAR files into the shared folder for the vApp. Then you can update these configuration settings for the new cluster. For more information about configuring a new version of Hadoop, see *SAS Data Loader for Hadoop: vApp Deployment Guide*.

The fields in the **Hadoop Configuration** panel are as follows:

Host
the fully qualified host name for HiveServer2 on your Hadoop cluster.

Port
the port number for HiveServer2 on your Hadoop cluster.

User ID
the name of the user account that is used to connect to the Hadoop cluster. If this field is editable, you can specify an ID that is provided by your Hadoop administrator.

The **User ID** is not editable if Kerberos security has been specified in the vApp, as described in *SAS Data Loader for Hadoop: vApp Deployment Guide*.

When your cluster uses a MapR distribution of Hadoop, the **User ID** field is populated from a configuration file when you start the vApp. To change the **User ID** field, first enter the new value in the file ***vApp-home\shared-folder*\hadoop\conf\mapr-users.json**. Next, restart the vApp to read the new value. Finally, open the **Hadoop Configuration** panel and enter the new user ID.

Password
the password for the user account that is used to connect to the Hadoop cluster. If your system administrator indicates that the cluster uses LDAP authentication, a password is required. Enter the password that is provided by the administrator. If the Hadoop cluster does not require a password for authentication, leave this field blank.

The **Password** is not editable if Kerberos security has been specified in the vApp.

Oozie URL
the URL to the Oozie Web Console, which is an interface to the Oozie server. Oozie is a workflow scheduler system that is used to manage Hadoop jobs. SAS Data Loader uses the SQOOP and Oozie components installed with the Hadoop cluster to move data to and from a DBMS.

- URL format: http://*host_name*:*port_number*/oozie/
- URL example (using default port number): http://my.example.com:11000/oozie/

Schema for temporary file storage
enables you to specify an alternative schema in Hive for the temporary files that are generated by some directives. The default schema for these files is **default**. Any alternative schema must exist in Hive.

Hive storage location
enables you to specify a location on the Hadoop file system to store your content that is not the default storage location. You must have appropriate

permissions to this location in order to use it. For more information, see "Overriding the Hive Storage Location for Target Tables" on page 148.

SAS HDFS temporary storage location
enables you to specify an alternative location on the Hadoop file system to read and write temporary files when using features specific to SAS. You must have appropriate permissions to this location in order to use it.

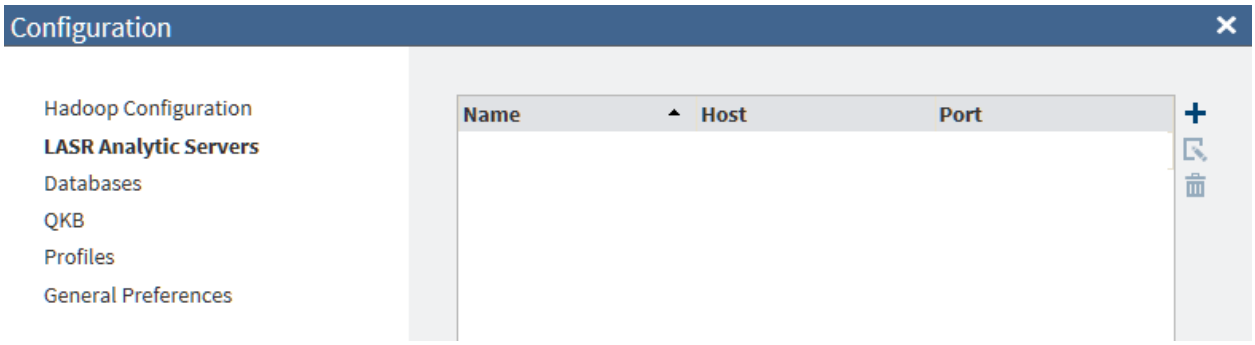
If the default temporary storage directory for SAS Data Loader is not appropriate for some reason, you can change that directory. For example, some SAS Data Loader directives might fail to run if they cannot write to the temporary directory. If that happens, ask your Hadoop administrator if the sticky bit has been set on the default temporary directory (typically `/tmp`). If that is the case, specify an alternate location in the **SAS HDFS temporary storage location** field. Your Hadoop administrator will tell you the location of this alternate directory. The administrator must grant you Read and Write access to this directory.

LASR Analytic Servers Panel

Overview

You can use the Load Data to LASR directive to copy Hadoop tables to a single SAS LASR Analytic Server or to a grid of SAS LASR Analytic Servers. On the SAS LASR Analytic Servers, you can analyze tables using software such as SAS Visual Analytics.

The Load Data to LASR directive requires a connection to SAS LASR Analytic Server software. To load data onto a grid of SAS LASR Analytic Servers, the directive requires a connection that is optimized for massively parallel processing (MPP). To load data onto a single SAS LASR Analytic Server, the directive requires a connection that is optimized for symmetric multi-processing (SMP). You can use the **LASR Analytic Server** panel of the Configuration window to create these connections.



To add connections for SAS LASR Analytic Servers:

- Verify that the appropriate SAS LASR Analytic Server software is available, as described in the general prerequisites below.
- Display the Hadoop Configuration panel of the Configuration window. If the **User ID** field is not editable, the Hadoop login for SAS Data Loader has been configured for Kerberos authentication. See the prerequisites below for Kerberos. Otherwise, see the prerequisites below for the case when Kerberos is not used.

- Use the information above to add a connection to SAS LASR Analytic Server software, as described below.

General Prerequisites for SAS LASR Analytic Server

Ask your SAS LASR Analytic Server administrator to verify that the following prerequisites have been met:

- SAS LASR Analytic Server must be release 2.5 or later. The server must be fully operational and configured to start automatically.
- SAS Visual Analytics 6.4 or later must be installed and configured on the SAS LASR Analytic Server.
- SAS LASR Analytic Server must be registered on a SAS Metadata Server.
- SAS LASR Analytic Server must have memory and disk allocations that are large enough to accept Hadoop tables. Jobs created with the Load Data to LASR directive cannot ensure that sufficient storage is available in SAS LASR Analytic Server.

Additional Prerequisites for Kerberos Authentication

Display the Hadoop Configuration panel of the Configuration window. If the **User ID** field is not editable, the Hadoop login for SAS Data Loader has been configured for Kerberos authentication. The following additional prerequisites apply.

- The user ID used to log on to the Hadoop cluster and the user ID used to log on to SAS LASR Analytic Server must be identical. Take note of the **User ID** that is specified in the **Hadoop Configuration** panel. Ask the SAS LASR Analytic Server administrator to create an account for that user ID on the SAS LASR Analytic Server.
- SAS Data Loader, the Hadoop cluster, and the SAS LASR Analytic Server must share a single Kerberos realm. The Kerberos realm for SAS Data Loader and the Hadoop cluster is specified in the SAS Data Loader: Information Center Settings window in the vApp. Ask the SAS LASR Analytic Server administrator to verify that the user ID on the SAS LASR Analytic Server is in the same Kerberos realm.
- When SAS Data Loader is configured, a Kerberos user ID and realm are entered into the SAS Data Loader: Information Center Settings window in the vApp. When this information is saved, a public key for that user is placed in the shared folder for SAS Data Loader. Ask the SAS LASR Analytic Server administrator to copy this public key to the SAS LASR Analytic Server or to the head node on the SAS LASR Analytic Server grid. The public key must be appended to the authorized keys file in the .ssh directory of that user.
- Review the fields in the LASR Server of the Configuration window. Ask the SAS LASR Analytic Server administrator to provide the information that is required to specify a connection in this window.

After these prerequisites have been met, you can add a connection to a SAS LASR Analytic Server. See “Add or Update Connections to SAS LASR Analytic Servers” on page 137.

Additional Prerequisites When Kerberos Authentication Is Not Used

Display the Hadoop Configuration panel of the Configuration window. If the **User ID** field is editable, the Hadoop login for SAS Data Loader has been configured for no authentication or for an authentication method other than Kerberos. The following additional prerequisites apply.

- The user ID used to log on to the Hadoop cluster and the user ID used to log on to SAS LASR Analytic Server must be identical. Take note of the **User ID** that is specified in the **Hadoop Configuration** panel. Ask the SAS LASR Analytic Server administrator to create an account for that user ID on the SAS LASR Analytic Server.
- The user account above must be configured with Secure Shell (SSH) keys on the SAS LASR Analytic Server.

Ask the SAS LASR Analytic Server administrator to perform these steps:

1. The administrator generates a public key and a private key for the SAS Data Loader user account and installs those keys in SAS LASR Analytic Server, as described in the *SAS LASR Analytic Server: Reference Guide*.
2. The administrator copies the public key file from SAS Data Loader at ***vApp-install-path\vApp-instance\shared-folder*** **\Configuration \sasdemo.pub**. A typical path is **C:\Program Files\SASDataLoader \dataloader-3p.22on94.1-devel-vmware.vmware (1)\dataloader-3p.22on94.1-devel-vmware\SASWorkspace \Configuration**.
3. The administrator appends the SAS Data Loader public key to the file ***~designated-user-account/*****.ssh/authorized_keys**.

   If SAS LASR Analytic Server is configured across a grid of hosts, then the public key is appended in the head node of the grid.

   CAUTION! To maintain access to SAS LASR Analytic Server, you must repeat step 3 each time you replace your installation of SAS Data Loader for Hadoop.

   Note: It is not necessary to repeat this step if you update your vApp by clicking the **Update** button in the SAS Data Loader: Information Center.

Review the fields in the LASR Server of the Configuration window. Ask the SAS LASR Analytic Server administrator to provide the information that is required to specify a connection in this window.

After these prerequisites have been met, you can add a connection to a SAS LASR Analytic Server. See "Add or Update Connections to SAS LASR Analytic Servers" on page 137.

Additional Prerequisites for SSL Connections

If you want SAS Data Loader to connect to a SAS LASR Analytic Server in a deployment where the SAS Web Server is secured with Secure Socket Layer (SSL), you must do the following tasks.

1. Contact the administrator who is responsible for SSL certificates for your site.
2. Obtain the SSL certificate file that is required to access the SAS LASR Analytic Server. The SSL file contains the public certificates for the trusted

certification authorities (CA) for your site. The CA file must be PEM-encoded (base64). The name of the file will be **cacert.pem**.

3 Locate the shared folder (**SASWorkspace** folder) on the SAS Data Loader host.

4 Create a subfolder named `cert` under the shared folder: **SASWorkspace \cert**.

5 Copy the SSL certificate file to the `cert` subfolder: **SASWorkspace\cert \cacert.pem**.

After these prerequisites have been met, you can add a connection to a SAS LASR Analytic Server, as described in the next section.

Add or Update Connections to SAS LASR Analytic Servers

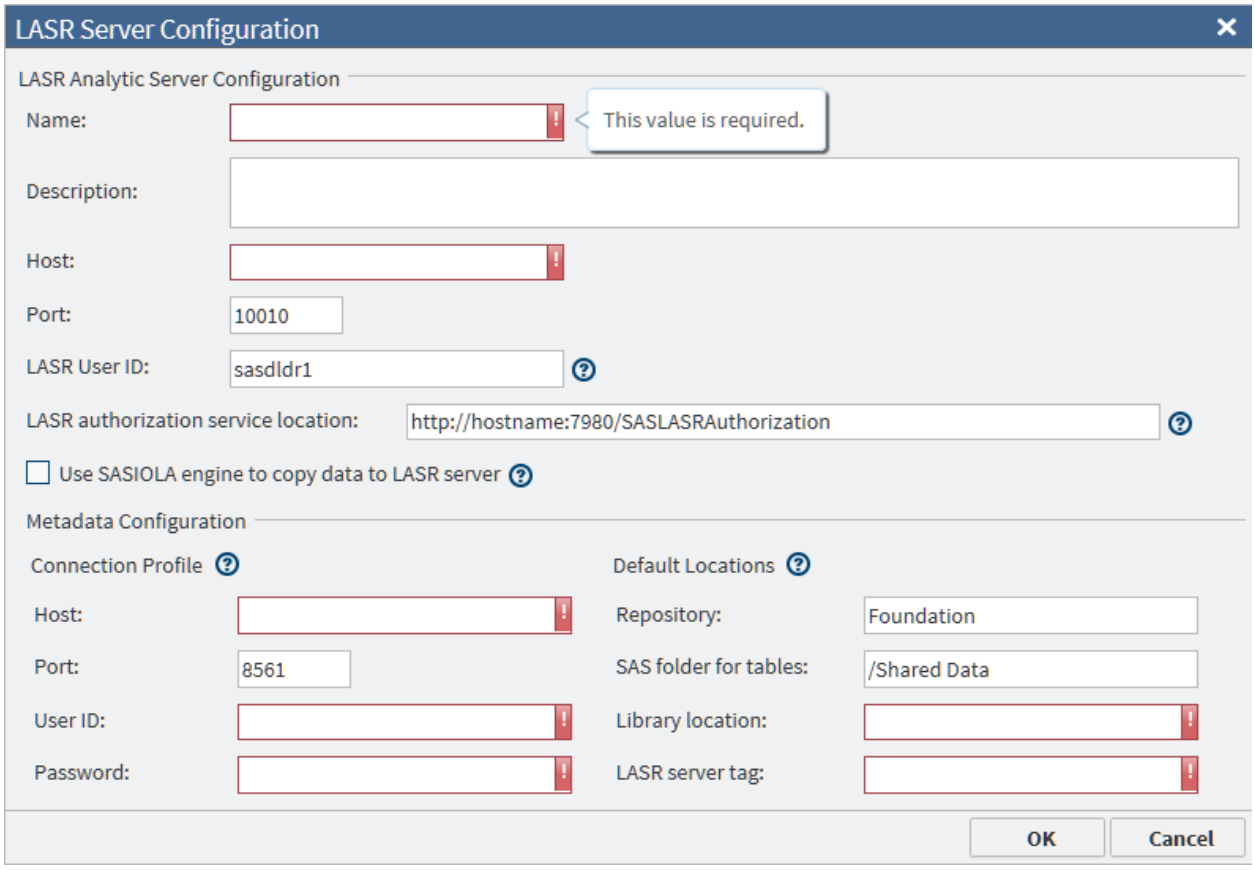
After the prerequisites above have been met, you can add a connection to a SAS LASR Analytic Server. Perform these steps:

1 In the SAS Data Loader directives page click the **More** icon and select **Configuration**.

2 Click **SAS LASR Analytic Servers**.

3 To configure a new connection to SAS LASR Analytic Server, click the **Add** icon . To change an existing connection to SAS LASR Analytic Server , click that connection in the list, and then click the **Edit** icon . To delete a connection to SASLASR AnalyticServer, select it and click the **Delete** icon .

4 In the LASR Server panel of the Configuration window, enter or change your choice of server name and description in the **Name** and **Description** fields.



5 In the **Host** field, add or change the full network name of the host of the SAS LASR Analytic Server. A typical name is similar to lasr03.us.ourco.com.

6 In the **Port** field, add or change the number of the port that the SAS LASR Analytic Server uses to listen for connection requests from SAS Data Loader. The default port number is 10010.

7 If your Hadoop cluster uses Kerberos for authentication, then the value of the **LASR User ID** field is not used. It is assumed to the same as the **User ID** that is specified in the **Hadoop Configuration** panel.

   If your Hadoop cluster does not use Kerberos for authentication, enter the name of the user account on the SAS LASR Analytic Server that received SSH keys, as described in "Additional Prerequisites When Kerberos Authentication Is Not Used" on page 136. Consult your administrator to confirm whether you should specify a user ID in this field and, if so, which user ID you should use. If no user ID is specified, the user sasdldr1 is used.

8 In the field **LASR authorization service location**, add or change the HTTP address of the authorization service. You can specify an HTTPS URL if you have done some additional set up. See "Additional Prerequisites for SSL Connections" on page 136.

9 If your SAS LASR Analytic Server is configured to run on a grid of multiple hosts, deselect **Use SASIOLA engine to copy data to LASR server**. Not selecting this field indicates that massively parallel processing (MPP) will be used in the SAS Data Loader jobs that use this connection.

If your SAS LASR Analytic Server supports symmetric multiprocessing (SMP) on a single host, click **Use SASIOLA engine to copy data to LASR server**.

10 Under **Connection Profile**, in the lower of the two **Host** fields, add or change the network name of the SAS Metadata Server that is accessed by the SAS LASR Analytic Server.

11 In the lower of the two **Port** fields, add or change the number of the port that the SAS Metadata Server uses to listen for client connections. The default value 8561 is frequently left unchanged.

12 In the **User ID** and **Password** fields, add or change the credentials that SAS Data Loader will use to connect to the SAS Metadata Server. These values are stored in encrypted form.

13 Under **Default Locations**, in the **Repository** field, specify the name of the repository on the SAS LASR Analytic Server that will receive data from Hadoop. The default value `Foundation` might suffice.

14 In the field **SAS folder for tables**, specify the path inside the repository on the SAS LASR Analytic Server that will contain the data that is loaded from Hadoop. The default value `/SharedData` might suffice.

15 In the **Library location** field, add or change the name of the SAS library on the SAS LASR Analytic Server that will be referenced by the Load Data to LASR directive.

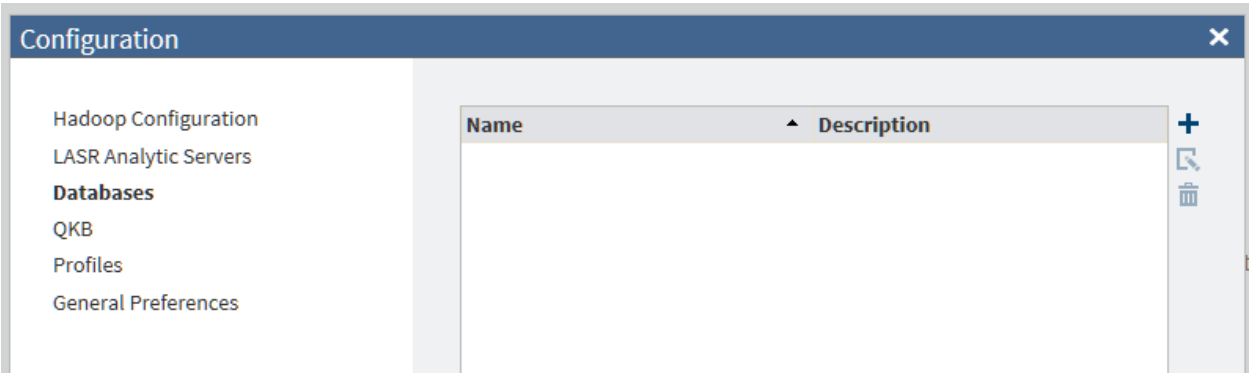
16 In the **LASR server tag** field, add or change the name of the tag that the SAS LASR Analytic server will associate with each table that is loaded from Hadoop. The tag is required to uniquely identify tables.

17 Review your entries and click **OK** to return to the Configuration window.

Databases Panel

Overview

Directives such as Copy Data to Hadoop and Copy Data from Hadoop require JDBC connections in order to access tables in databases. The **Databases** panel of the Configuration window enables you to maintain these connections.



To add connections for databases:

- Ask your Hadoop administrator to send you a copy of the JDBC drivers that are installed on the Hadoop cluster.
- Copy these JDBC drivers to the shared folder on the SAS Data Loader host.
- Contact the administrators of the databases to which you want to connect. Ask for the usual information that is required to connect to a database: host name, port, logon credentials, and so on.
- Add connections to the databases for which you have JDBC drivers.

Copy JDBC Drivers to the SAS Data Loader Host

SAS Data Loader uses the SQOOP and Oozie components installed with the Hadoop cluster to move data to and from a DBMS. SAS Data Loader also accesses the databases directly, using JDBC to select source or target tables and schemas. The Hadoop administrator installs appropriate JDBC drivers on the Hadoop cluster. The SAS Data Loader host must have the same version of the JDBC drivers in its shared folder.

Follow these steps to obtain JDBC drivers from the Hadoop administrator and copy them to the shared folder on the SAS Data Loader host:

1. Ask your Hadoop administrator for a copy of the JDBC drivers that are installed on your Hadoop cluster.
2. On the SAS Data Loader host, navigate to the shared folder and open the JDBCDrivers subfolder. Here is a typical path to the JDBCDrivers folder:

   ```
   C:\Program Files\SAS Data Loader\2.x\SASWorkspace\JDBCDrivers
   ```

   To verify the path to your shared folder, open the VMware Player Pro window and select **Player** ▸ **Manage** ▸ **Virtual Machine Settings**. In the Virtual Machine Settings window, click the **Options** tab, and then click **Shared Folders** (in the **Settings** list.) On the right side, the path to the shared folder is provided in the **Host Path** column.
3. Copy the files for the JDBC drivers into JDBCDrivers folder.
4. The vApp must be restarted so that it can pick up any new JDBC drivers. To restart the vApp:

   Check the “Run Status” directive to ensure that all jobs are stopped and saved.

   In VMware Player Pro, select **Player** ▸ **Power** ▸ **Restart Guest**. Wait for the vApp to restart.

   Note: Suspending the vApp is not sufficient to detect the new drivers. Restarting is the only way to restart services and ensure that the new drivers are detected.

SAS Data Loader now has access to these new JDBC drivers. The next task is to add connections to the databases for which you have new JDBC drivers.

Add Database Connections

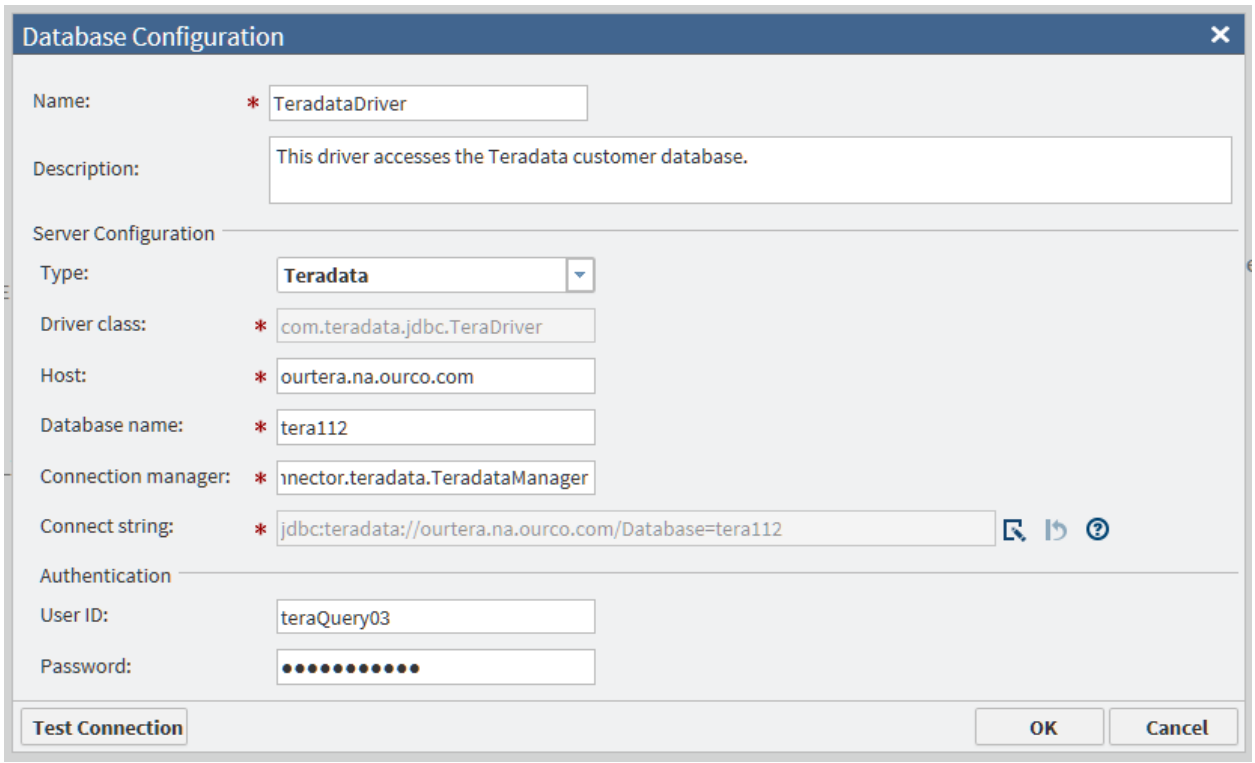
After you have copied the appropriate JDBC drivers into the shared folder on the SAS Data Loader host, you can add connections to the corresponding databases.

1 Contact the administrators of the databases to which you want to connect. Ask for the usual information that you would need to connect to a database: host name, port, logon credentials, and so on.

2 In SAS Data Loader, click the **More** icon and select **Configuration**.

3 In the Configuration window, click **Databases**. To add a new database connection, click **Add** +. To edit an existing database connection, click the name of the connection, and then click **Edit**.

4 The values of **Driver class** and **Connect string** are generated automatically when you select either Teradata or Oracle in the **Type** field. For an Oracle connection that requires a Service ID (SID), enter the SID in the **Database name** field. If you select **Other**, you must obtain these values from the JDBC driver provider.



5 When the configuration data is ready, click **Test Connection** to verify that the connection is operational.

6 If the test fails for a new Oracle connection, then examine the **Connect string** field. If the string has either of the following formats, then change the string to the other format and test the connection again.

```
jdbc:oracle:thin:@raintree.us.ourco.com:1521:oadev
```

```
jdbc:oracle:thin:@raintree.us.ourco.com:1521/oadev
```

One version uses a final colon character. The other version uses a final slash character.

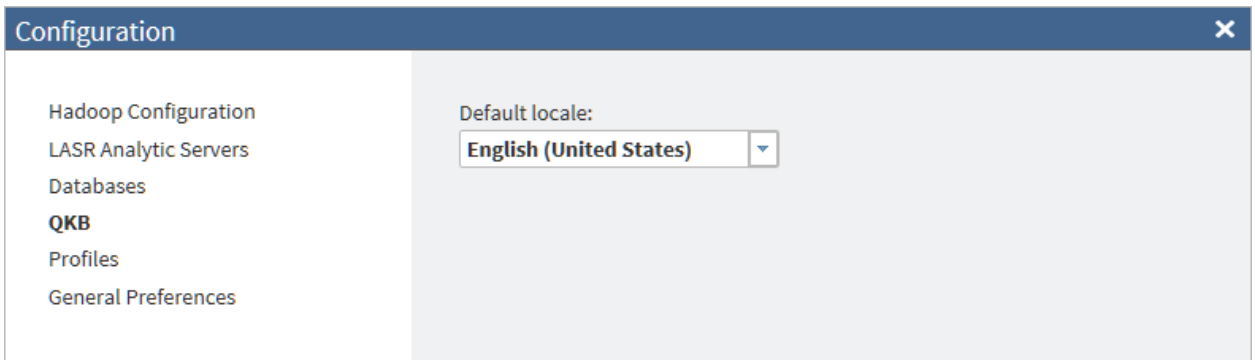
To edit the **Connect string** field, click **Edit**.

7 Click **OK** to close the window. SAS Data Loader directives can now use this database connection.

QKB Panel

A SAS Quality Knowledge Base (QKB) is a collection of files that store data and logic that define data management operations such as parsing, standardization, and matching. SAS software products refer to the QKB when performing data management operations, also referred to as data cleansing, on your data.

A QKB supports locales organized by language and country, for example, English, United States; English, Canada; and French, Canada. The **OKB** panel of the Configuration window enables you to select the default locale used by the transformations in the Cleanse Data in Hadoop directive. The default locale should match the typical locale of your source data.

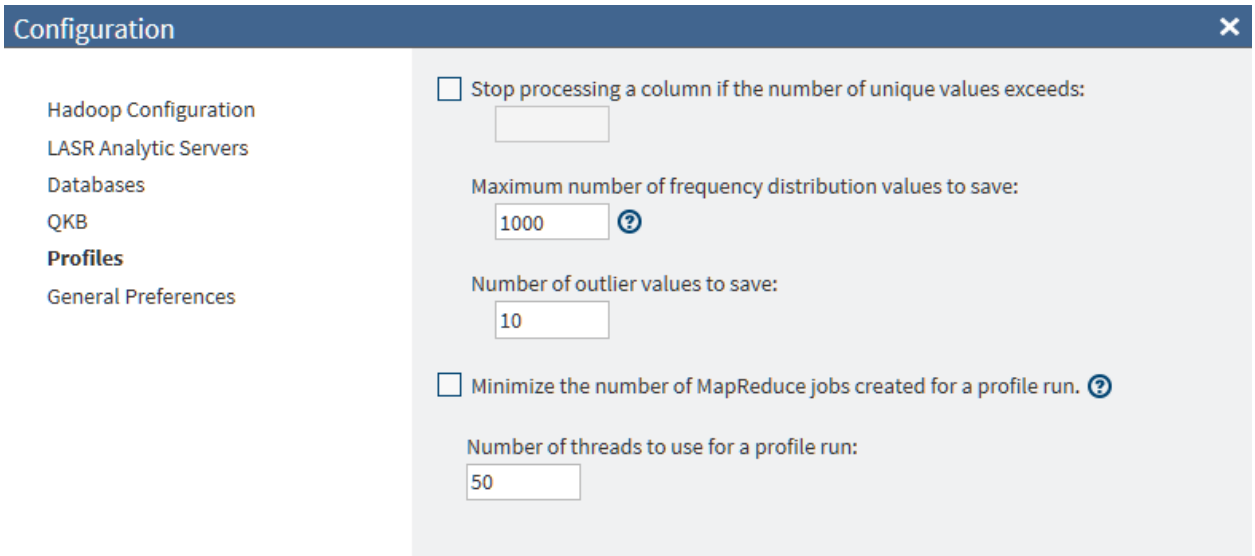


You can override the default locale in any of the data quality transformations in the Cleanse Data in Hadoop directive. For more information about this directive, see “Cleanse Data in Hadoop” on page 23.

Profiles Panel

SAS Data Loader profile directives enable you to assess the composition, organization, and quality of tables in Hadoop. For more information about these directives, see Chapter 5, “Profile Data in Hadoop,” on page 73.

Data profiling tasks can be resource intensive. Accordingly, the **Profiles** panel of the Configuration window enables you to change defaults that can improve the performance of new profile jobs.



You can change the following default options for profiles:

Stop processing a column if the number of unique values exceeds
stops processing a column if the number of unique values is greater than the number that you enter in the field. You can specify a value from 0 to 99999999. A value of 0 causes every observation in the table to be processed.

Maximum number of frequency distribution values to save
the maximum number of frequency distribution values (1–99999999) to save during the profile run. If there are more frequency distribution values than this number, the less-frequent values are combined into an Other frequency distribution.

Number of outlier values to save
the maximum number of outlier values (1–99999999) to save during the profile run.

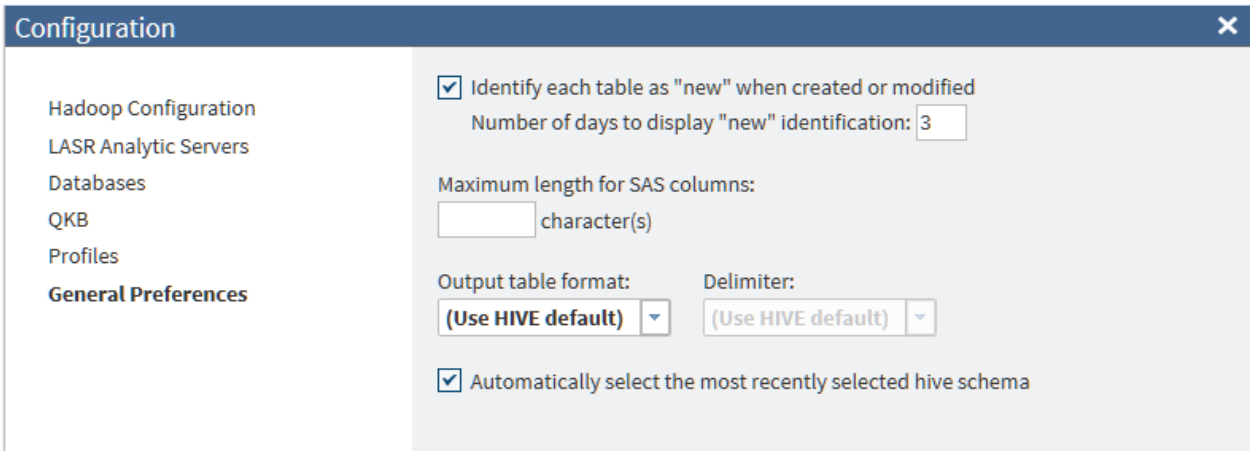
Minimize the number of MapReduce jobs created for a profile run.
limits the number of parallel processes that are used in a profile job.

Note: This option causes all processing to be performed by the vApp. For large data tables with many columns, performance can be affected, even to the point of causing the vApp to run out of memory. You should specify this setting only if you know that you have very small data tables.

Number of threads to use for a profile run
the maximum number of threads (1–99999999) to use for a profile job. If you select the option to minimize the number of MapReduce jobs for a profile run, this setting is disabled.

General Preferences Panel

The **General Preferences** panel of the Configuration window enables you to specify various global options for SAS Data Loader.



You can change the following default options:

Identify each table as "new" when created or modified. Number of days to display "new" identification:
specifies how long tables are identified as “new” in SAS Data Loader.

Maximum length for SAS columns
specifies a default length for character columns in input tables for some directives. The default length of 1024 characters should perform well in most cases. For more information, see “Change the Maximum Length for SAS Character Columns” on page 147.

Output table format and **Delimiter**
specifies the default file format and delimiter for directive target tables. Use the **Output table format** drop-down list to select one of five output table formats: Hive default, Text, Parquet, ORC, or Sequence. Use the **Delimiter** drop-down list to select one of five output table formats: Hive default, Comma, Tab, Space, or Other. For more information, see “Change the File Format of Hadoop Target Tables” on page 145.

Note: If your cluster runs a MapR distribution of Hadoop, then the Parquet output table format is not supported.

Automatically select the most recently selected hive schema
If you frequently work with the same data source across multiple directives, you can have SAS Data Loader select the most recently used schema automatically. This can help you select source tables and target tables more quickly.

Troubleshooting

Active Directory (LDAP) Authentication

If Active Directory and LDAP (Lightweight Directory Access Protocol) are used to protect your Hadoop cluster, an LDAP user and password must be specified in the Hadoop connection for SAS Data Loader. For more information about the Hadoop connection, see “Hadoop Configuration Panel” on page 132.

Oozie does not support LDAP authentication. SAS Data Loader uses the SQOOP and Oozie components installed with the Hadoop cluster to move data

to and from a DBMS. Accordingly, when LDAP authentication is used with your Hadoop cluster, directives that rely on Oozie, such as Copy Data To Hadoop, do not receive the authentication benefits provided by LDAP. However, the operation of these directives is otherwise unaffected.

Change the File Format of Hadoop Target Tables

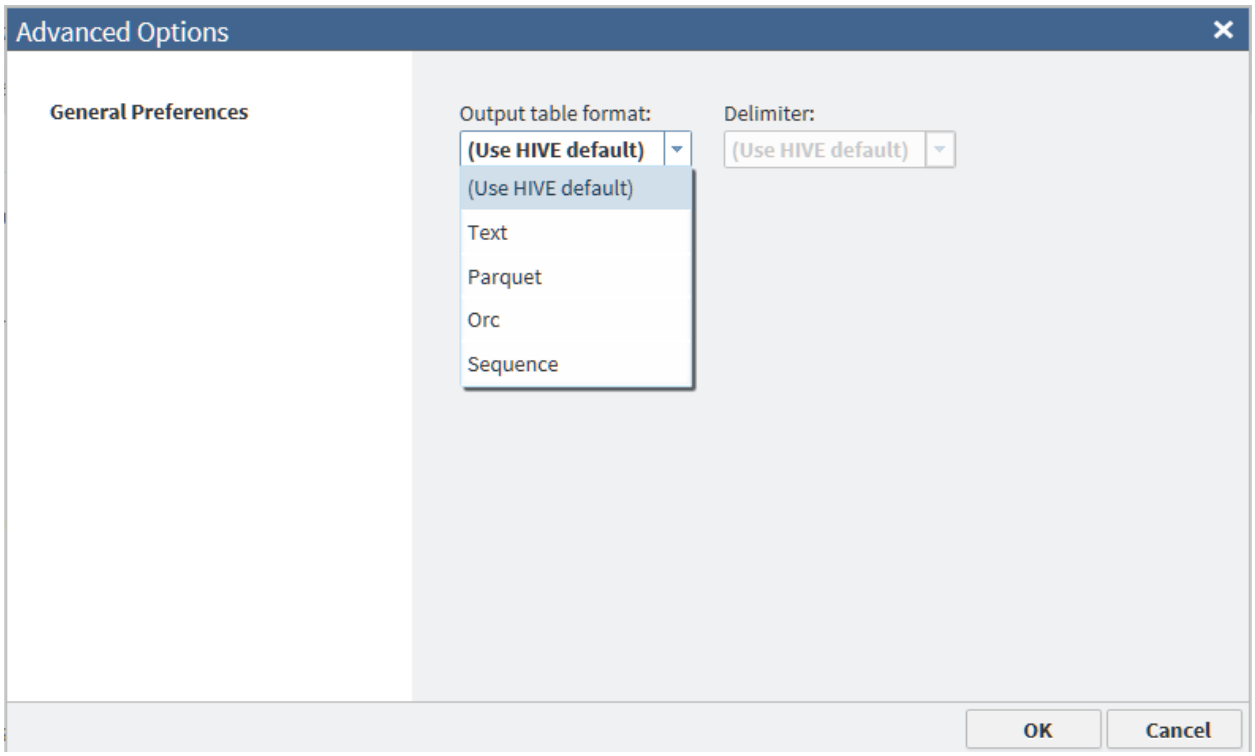
In Hadoop file system (HDFS), tables are stored as one or more files. Each file is formatted according to the Output Table Format option, which is specified in each file. In SAS Data Loader, when you create a new target table in Hadoop, the Output Table Format option is set by the value of the **Output table format** field.

You can change the default value of the **Output table format** field in the SAS Data Loader Configuration window. In any given directive, you can override the default value using the **Action** menu icon in the **Target Table** task.

The default format is applied to all new target tables that are created with SAS Data Loader. To override the default format in a new table or an existing table, you select a different format in the directive and run the job.

To change the default value of the **Output table format** field, go to the SAS Data Loader directives page, click the **More** icon , and select **Configuration**. In the Configuration window, click **General Preferences**. In the General Preferences panel, change the value of **Output table format**.

To override the default value of the **Output table format** field for a given target in a given directive, open the directive and proceed to the **Target Table** task. Select a target table, and then click the **Action** menu for that target table. Select **Advanced Options**, and then, in the Advanced Options window, set the value of **Output table format**. The format that you select applies only to the selected target table. The default table format is not changed. The default format continues to be applied to all new target tables.



The available values of the **Output table format** field are defined as follows:

Use HIVE default
: specifies that the new target table receives the Output Table Format option value that is specified in HDFS. This is the default value for the **Output table format** field in SAS Data Loader.

Text
: specifies that the new target table is formatted as a series of text fields that are separated by delimiters. For this option, you select a value for the **Delimiter** field. The default value of the **Delimiter** field is **(Use HIVE default)**. You can also select the value **Comma**, **Space**, **Tab**, or **Other**. If you select **Other**, then you enter a delimiter value. To see a list of valid delimiter values, click the question mark icon to the right of the **Delimiter** field.

Parquet
: specifies the Parquet format, which is optimized for nested data. The Parquet algorithm is considered to be more efficient than using flattened nested namespaces.

  Note: If your cluster uses a MapR distribution of Hadoop, then the Parquet output table format is not supported.

ORC
: specifies the Optimized Row Columnar format, which is a columnar format that efficiently manages large amounts of data in Hive and HDFS.

Sequence
: specifies the SequenceFile output format, which enables Hive to efficiently run MapReduce. This format enables Hive to efficiently split large tables into separate threads.

Consult your Hadoop administrator for advice about output file formats. Testing might be required to establish the format that has the highest efficiency on your Hadoop cluster.

Change the Maximum Length for SAS Character Columns

Some directives use a SAS Workspace Server to read or write tables. By default, the character columns for the input tables to such directives are expanded to 1024 characters in length. The valid range for the maximum length option is 1–32,767 characters. The default length should perform well in most cases, though there might be situations where a larger value is required. For example, if you have cells in your columns with data larger than 1024 characters, SAS directives that use DS2 truncates the data. In such situations, you should increase the maximum length.

Note: You should not change this value unless you need to because of your data requirements. As you increase the maximum length for SAS character columns, you also increase the likelihood that performance will be affected.

The affected directives are as follows:

- Transform Data in Hadoop
- Transpose Data in Hadoop
- Load Data to LASR
- Copy Data to Hadoop (when a data set from the SAS Server is selected as the input table)
- Copy Data from Hadoop (when the SAS Server is selected as the location for the target table)

If you want to change the default maximum length for SAS character columns for all directives, go to the SAS Data Loader directives page, click the **Action** menu , and select **Configuration**. In the Configuration window, select **General Preferences**, and specify the desired length for SAS character columns.

You can override the default maximum length for SAS character columns in a given directive without changing the default. In one of the directives listed above, open the **Source Table** task, click the **Action** menu, and select **Advanced Options** . In the Advanced Options window, specify the desired length for SAS character columns. The value that you specify applies only to the current instance of the current directive.

Change the Temporary Storage Location

If the default temporary storage directory for SAS Data Loader is not appropriate for some reason, you can change that directory. For example, some SAS Data Loader directives might fail to run if they cannot write to the temporary directory. If that happens, ask your Hadoop administrator if the sticky bit has been set on the default temporary directory (typically `/tmp`). If that is the case, specify an alternate location for temporary storage. For more information, see the description of the **SAS HDFS temporary storage location** field in the topic "Hadoop Configuration Panel" on page 132.

Discover New Columns Added to a Source after Job Execution

When you add columns to a source table, any directives that need to use the new columns need to discover them. To make the new columns visible in a directive, open the Source Table task, click the source table again, and click **Next**. The new columns are then available for use in the body of the directive, in a transformation or query, for example.

Hive Limit of 127 Expressions per Table

Due to a limitation in the Hive database, tables can contain a maximum of 127 expressions. When the 128th expression is read, the directive fails and the SAS log receives a message similar to the following:

```
ERROR: java.sql.SQLException: Error while processing statement: FAILED:
Execution Error, return
       code 2 from org.apache.hadoop.hive.ql.exec.mr.MapRedTask
ERROR: Unable to execute Hadoop query.
ERROR: Execute error.
SQL_IP_TRACE: None of the SQL was directly passed to the DBMS.
```

The Hive limitation applies anytime a table is read as part of a directive. For SAS Data Loader, the error can occur in aggregations, profiles, when viewing results, and when viewing sample data.

Overriding the Hive Storage Location for Target Tables

When you work with directives that create target tables, those tables are stored in a directory location in the Hadoop file system. The default location is defined in the configuration settings for SAS Data Loader, as described in "Hadoop Configuration Panel" on page 132.

If you prefer, you can override the storage location for a target table for an individual job.

To override the storage location for a target table:

1. Proceed through the initial tasks for the directive as usual. For example, if you are using the Transform Data in Hadoop directive, you would select a source table and specify any transformations for the data.
2. When you reach the Target Table task in the directive, click to open the Advanced Options window.
3. On the General Preferences page, select **Specify alternate storage location** for the **Hive storage location** setting, and then click ... to open the Select Directory window.
4. Navigate to a folder where you want to store the target table and click **OK**. You can also create a new folder, if needed.

   Note: To use the alternate location, you must have appropriate permissions to the selected directory.

5 Continue through the remaining tasks for the directive to submit the job.

Unsupported Hive Data Types and Values

The Hive database in Hadoop identifies table columns by name and data type. To access a column of data, SAS Data Loader first converts the Hadoop column name and data type into its SAS equivalent. When the transformation is complete, SAS Data Loader writes the data into the target table using the original Hadoop column name and data type.

If your target data is incorrectly formatted, then you might have encountered a data type conversion error.

The Hive database in Hadoop supports a Boolean data type. SAS does not support the Boolean data type in Hive at this time. Boolean columns in source tables are not available for selection in SAS Data Loader.

The BIGINT data type in Hive supports integer values larger than those that are currently supported in SAS. BIGINT values that exceed +/– 9,223,372,036,854,775,807 generate a stack overflow error in SAS.

Restarting a Session after Time-out

SAS Data Loader records periods of inactivity in the user interface. After a period of continuous inactivity, the current web page receives a session time-out warning message in a window. If you do not provide input within three minutes after you receive the warning, the current web page is replaced by the Session Time-out page. You can restart your session by clicking the text **Return to the SAS Data Loader application**.

When a session terminates, any directives that you did not save or run are lost.

To open an unsaved directive that you ran before your session terminated, follow these steps:

1 Open the Run Status directive.

2 Locate the entry for your unsaved directive.

3 If the unsaved directive is still running, click the Refresh button.

4 If the directive continues to run, either click **Stop** in the action menu , or wait for the completion of the run.

5 In the action menu, select **Open** to open the directive.

6 In the open directive, select **Save** from the title bar.

10

Using the vApp

Overview of the vApp for SAS Data Loader 151
Play the vApp and Start SAS Data Loader 151
Tips for Running the vApp 152
Power Off the vApp 153

Overview of the vApp for SAS Data Loader

The SAS Data Loader for Hadoop web application runs inside a virtual machine or vApp. The vApp is started and managed by a hypervisor application called VMware Player Pro. The web application uses SAS software in the vApp and on the Hadoop cluster to manage data within Hadoop. The topics in this section review basic tasks, such as starting and stopping the vApp.

The *SAS Data Loader for Hadoop: vApp Deployment Guide* is a complete reference to tasks that can be performed in the vApp, such as the following:

- configure the vApp and SAS Data Loader after installing your SAS software
- migrate from SAS Data Loader 2.2 to SAS Data Loader 2.3
- troubleshoot the vApp start process
- update your vApp software
- if the version of Hadoop on your cluster has changed, change the version of Hadoop that is specified in the vApp to match
- if the security settings for your cluster have changed, change the security settings in the vApp to match
- enable logging inside the vApp
- manage your SAS license

Play the vApp and Start SAS Data Loader

Perform these steps to play the vApp and start SAS Data Loader for Hadoop.

1 Open VMware Player Pro.

2 Click **SAS Data Loader for Hadoop** and then, when it appears, click **Play virtual machine**.

3 VMware Player Pro requires a minute or two to play the vApp. When the vApp is ready, the VMware Player displays the message `Welcome to your SAS Data Loader Virtual Application`.

   Note: If an informational Removable Devices window appears, review the information about removable devices and click **OK**.

4 In the window SAS Data Loader – VMware Player Pro, locate the HTTP address to connect to the SAS Data Loader.

5 Open a web browser, and enter the HTTP address in the browser's address bar. Press **Enter** to continue. The SAS Data Loader: Information Center opens in a new tab in your browser.

6 In the SAS Data Loader: Information Center, click **Start SAS Data Loader**. SAS Data Loader opens in a new tab in your browser.

Tips for Running the vApp

- You can close and open SAS Data Loader in a web browser without shutting down the vApp. The vApp continues to play until you shut it down in VMware Player Pro. If you close the browser tab for SAS Data Loader while the vApp is playing, any jobs on the Hadoop cluster continue to run. Also, their run status continues to be collected.
- After SAS Data Loader displays in your browser, you do not have to keep the browser tab for SAS Data Loader: Information Center open. While the vApp is still playing, you can close the tab for the Information Center at any time and still access SAS Data Loader.
- Do not close the VMware Player Pro while the vApp is playing.
- You can shut down SAS Data Loader for Hadoop as described in "Power Off the vApp" on page 153.
- If you shut down vApp processes with Windows Task Manager, the vApp might not restart properly. If this happens, shut down the vApp as described in "Power Off the vApp" on page 153. Then power on the vApp as described in "Play the vApp and Start SAS Data Loader" on page 151.
- If the vApp is playing, and you click in the window SAS Data Loader – VMware Player Pro, the cursor disappears. This behavior is expected; it ensures that you have to physically enter the web address in a web browser to open the SAS Data Loader: Information Center. To restore your cursor, press Ctrl+Alt.
- You must restart the vApp after you change the version of Hadoop or configure a connection to a new database.

CAUTION! If you suspend a guest, services might be interrupted. VMware Player Pro provides a capability to suspend guests. In VMware Player Pro, do not select **Suspend Guest** or **Player** ▸ **Power** ▸ **Suspend Guest**. Suspending the vApp can interrupt communications between the SAS Data Loader web client and

the Hadoop cluster. To resolve a suspended vApp, select **Player** ▶ **Power** ▶ **Restart Guest**.

Power Off the vApp

Perform the following steps to power off the vApp in VMware Player Pro:

1. In the browser, close the tab for SAS Data Loader if it is open.
2. In the SAS Data Loader – VMware Player window, click **Player** ▶ **Power** ▶ **Shut Down Guest**.

   Note: The term guest refers to the guest operating system that runs the vApp.
3. In the VMware Player dialog box, click **Yes** to confirm that you want to power off the vApp.

Recommended Reading

- *SAS Data Loader for Hadoop: vApp Deployment Guide*
- *SAS In-Database Products: Administrator's Guide*
- *SAS DS2 Language Reference*
- *SAS/ACCESS for Relational Databases: Reference*
- SAS Quality Knowledge Base for Contact Information: Online Help

For a complete list of SAS publications, go to sas.com/store/books. If you have questions about which titles you need, please contact a SAS Representative:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-0025
Fax: 1-919-677-4444
Email: sasbook@sas.com
Web address: sas.com/store/books

Index

C

clear Run Status entries 127
columns, discover new 148
Configuration window 137

D

directives
  incomplete 128
  troubleshoot 149
  unsaved 128
Directives page 13

H

Hive
  data types 149
  limit on expressions 148
  maximum integer value 149

I

incomplete directives 128

L

LASR Server Configuration window 137
Load Data to LASR directive 118

P

prerequisites
  SAS LASR Analytic Server 135
Profile Data directive 75
Profile, Saved Reports 80

Q

Query or Join Data in Hadoop directive 47, 61

R

Run Status directive 126

S

SAS Data Loader
  architecture 3
SAS Data Quality Accelerator for Hadoop 3
SAS LASR Analytic Server 135
  with Kerberos 135
  without Kerberos 136
SAS LASR Server
  LASR Server Configuration window 137
SAS Table Viewer 16
SAS Visual Analytics 135
SAS/ACCESS for Hadoop 3
Saved Directives 128
Saved Profile Reports directive 80
session time-out 149
shared folder 4
Summarize Rows transformation 69

T

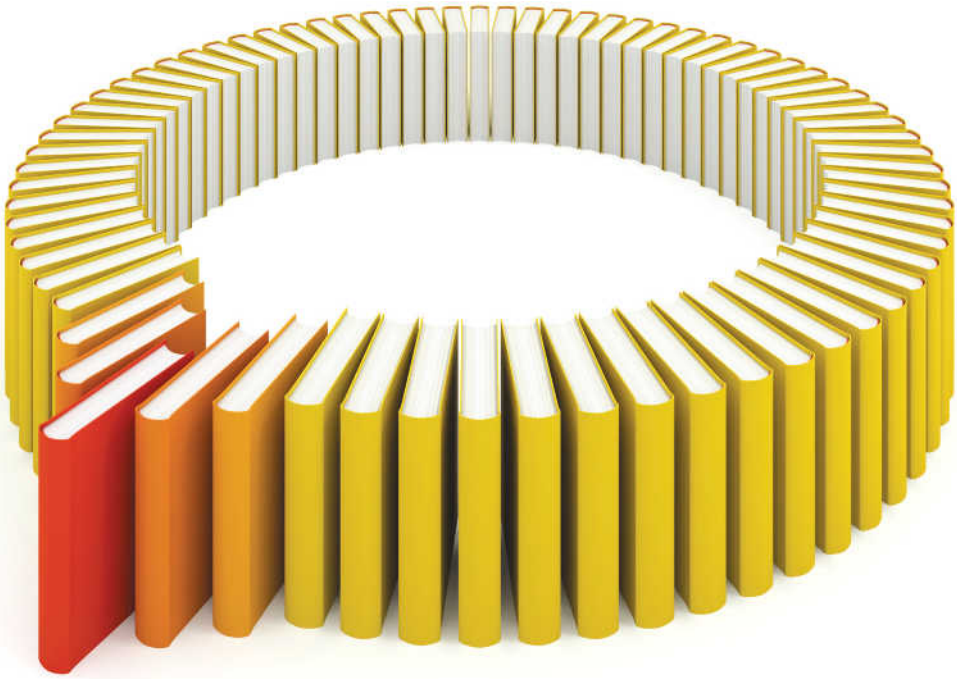
Transpose Data in Hadoop directive 70
troubleshoot directives 149

U

unsaved directives 128

V

vApp
how it works 3



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

sas
THE POWER TO KNOW.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613