SAS® Workshop: Data Management

Course Notes

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.

SAS® Workshop: Data Management Course Notes

Copyright © 2015 SAS Institute Inc. Cary, NC, USA. All rights reserved. Printed in the United States of America. 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.

Book code E2835, course code SGF15DM, prepared date 4/2/2015. SGF15DM_001

Table of Contents

To learn more…

For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the web **at** http://support.sas.com/training/ as well as in the Training Course Catalog.

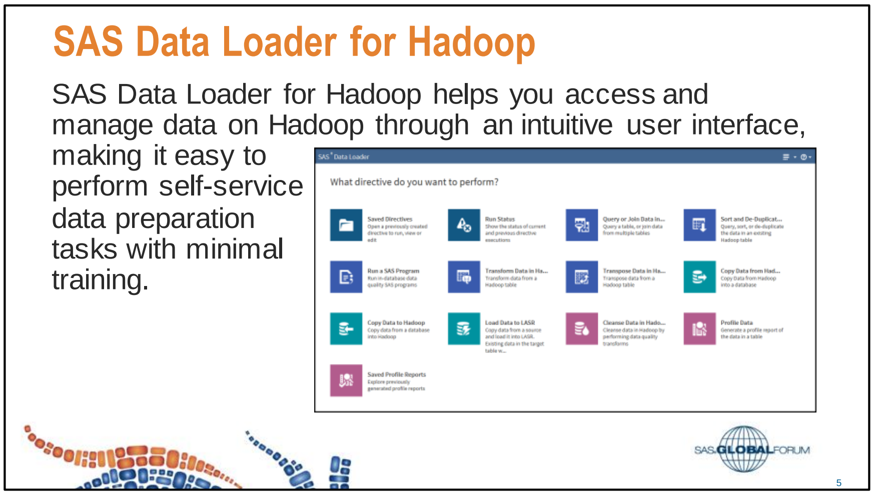
For a list of other SAS books that relate to the topics covered in this course notes, USA customers can contact the SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS o**ffice.**

Also, see the SAS Bookstore on the web at http://support.sas.com/publishing/ for a complete list of books and a convenient order form.

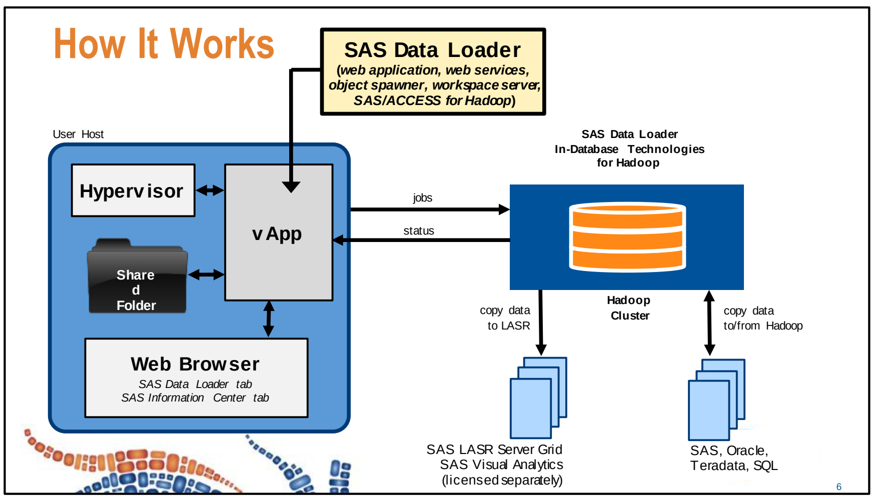
Chapter 1 SAS® Workshop: SAS® Data Loader for Hadoop

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

1.1 Introduction



SAS Data Loader for Hadoop uses a browser-based user interface. This solution runs "in cluster," taking advantage of the scalability of Hadoop by using a SAS Embedded Process (a lightweight SAS execution engine).



The SAS Data Loader for Hadoop web application runs inside the vApp. The vApp is started and managed by a hypervisor application called VMware Player Pro. 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 SAS Data Loader: Information Center does the following:

- starts the SAS Data Loader web application in a new browser tab.
- provides a single 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 your saved jobs, the JDBC drivers needed to connect to external databases, and the Hadoop JAR files that were copied to the client from the Hadoop cluster.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

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.

The SAS In-Database Technologies for Hadoop software is deployed to each node in the Hadoop cluster. The in-database technologies consist of a SAS Quality Knowledge Base for reference to data cleansing definitions, SAS Embedded Process software for code acceleration, and SAS Data Quality Accelerator software for SAS DS2 methods that pertain to data cleansing.

Workshop Setup

The machines for the workshop have

- VMware player installed
- Cloudera QuickStart installed
- SAS Data Loader for Hadoop (trial version) installed and configured to work with Cloudera QuickStart.

7

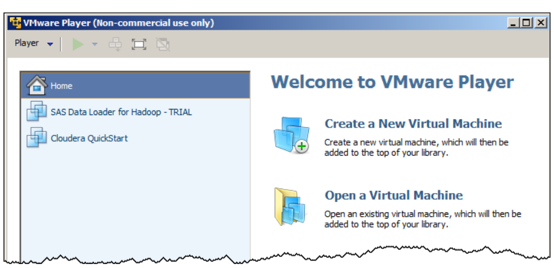
VMware player installed from https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/7_0

Cloudera QuickStart installed from https://downloads.cloudera.com/demo_vm/vmware/cloudera-quickstart-vm-5.3.0-0-vmware.7z

SAS Data Loader for Hadoop (Trial Edition) installed from https://support.sas.com/edownload/software/DPDLHFT01_VMware

Workshop Setup

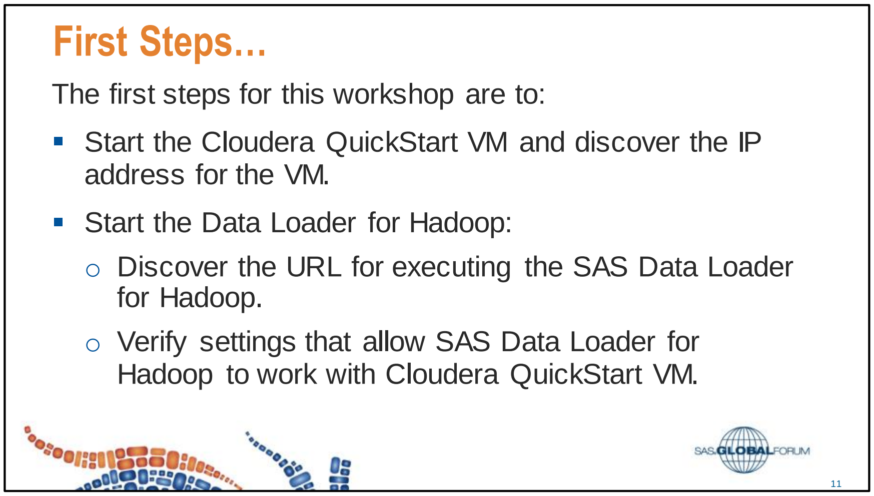
Both Cloudera QuickStart and SAS Data Loader for Hadoop (Trial Edition) have been added to VMware Player.



8

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

1.2 Working with SAS Data Loader for Hadoop

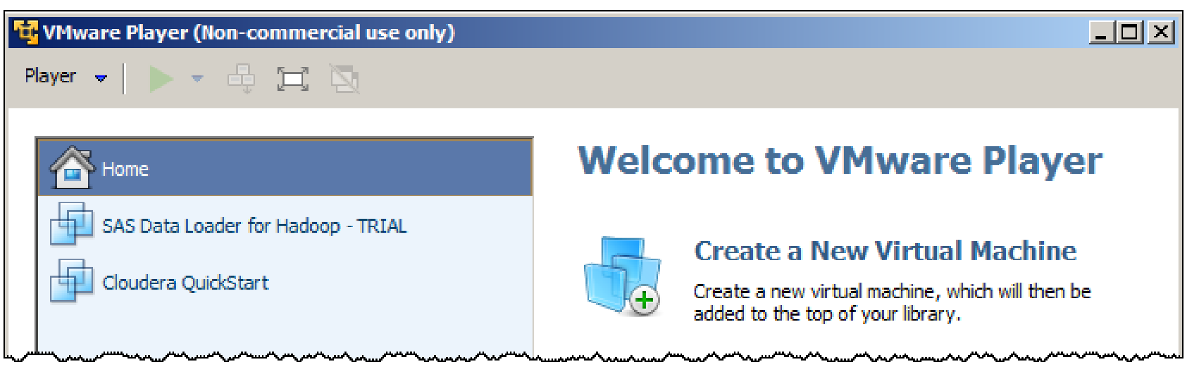


Verify Configuration of SAS Data Loader for Hadoop (Trial) with Cloudera QuickStart

1. Select **Start** ⇨ **All Programs** ⇨ **VMware** ⇨ **VMware Player**.

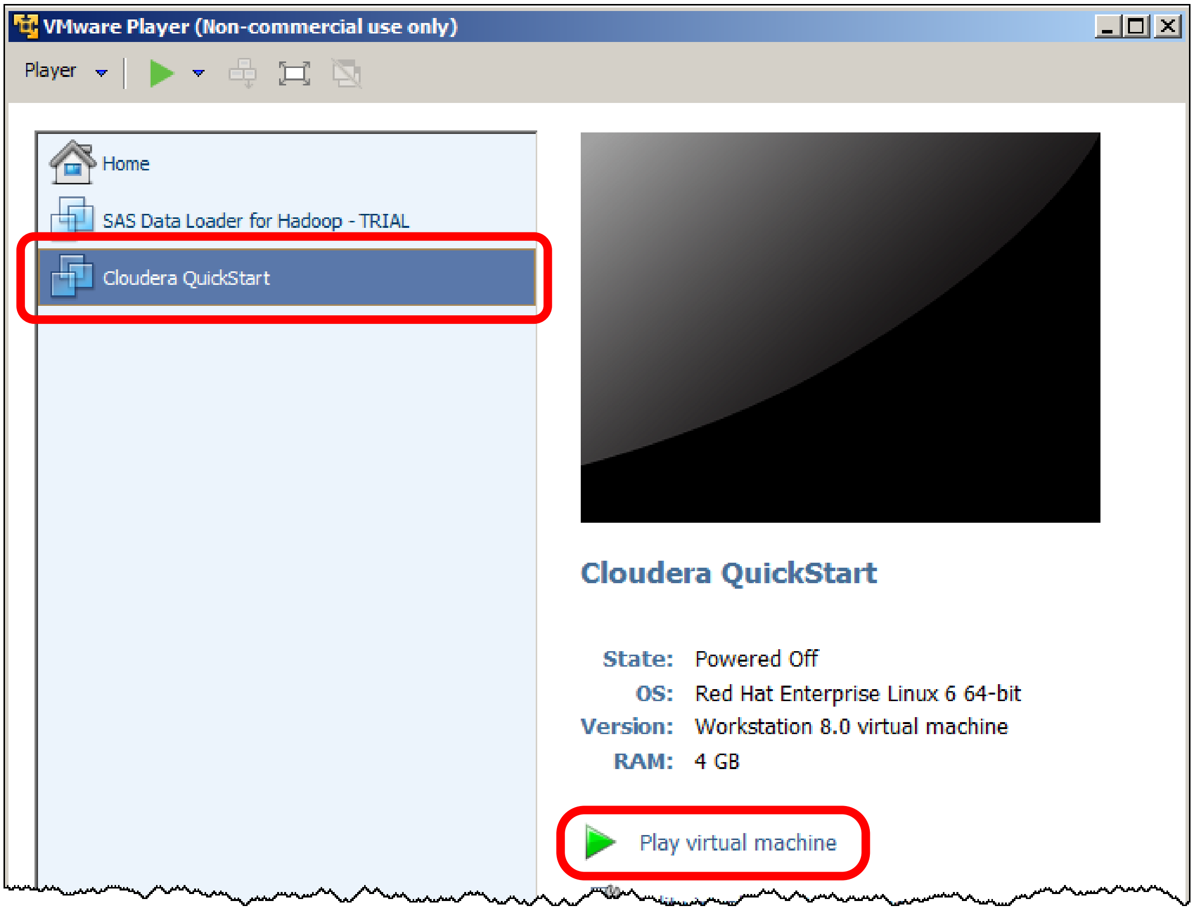
   There is also a shortcut ( ) on the desktop that you can double-click.

   The following should be displayed:



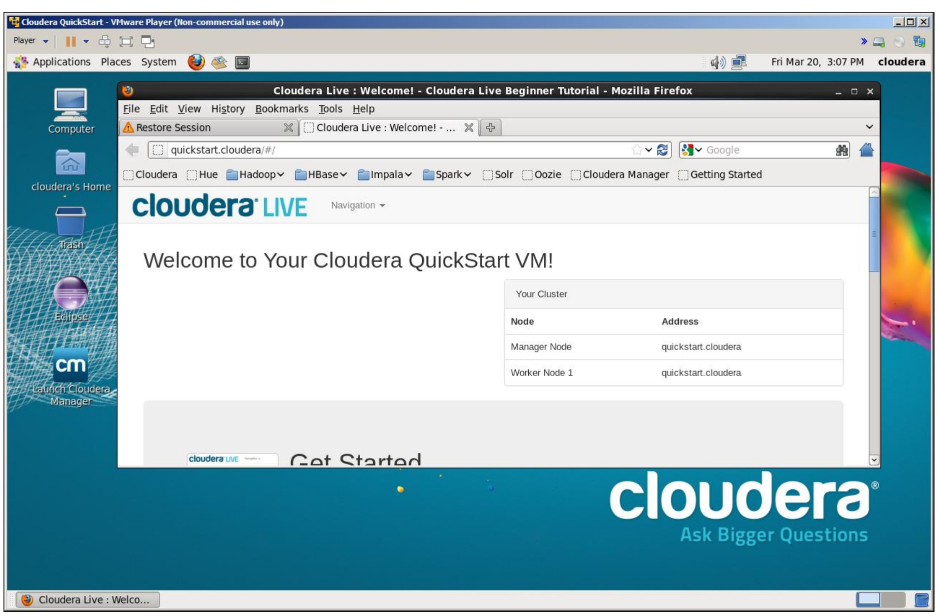
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

2. Single-click the virtual machine labeled **Cloudera QuickStart**, and then click **Play virtual machine**.



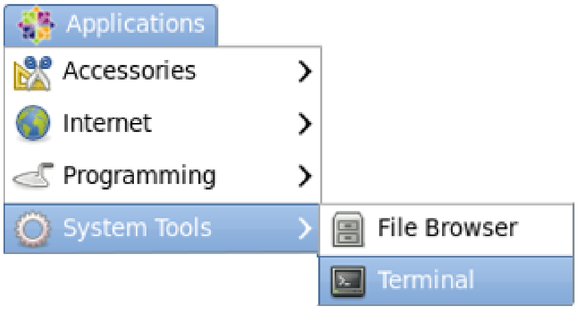
It takes a few minutes for the virtual machine to start.

After it is started, the VMware Player displays a desktop with a Firefox session open that has a "Welcome to your Cloudera QuickStart VM" message.

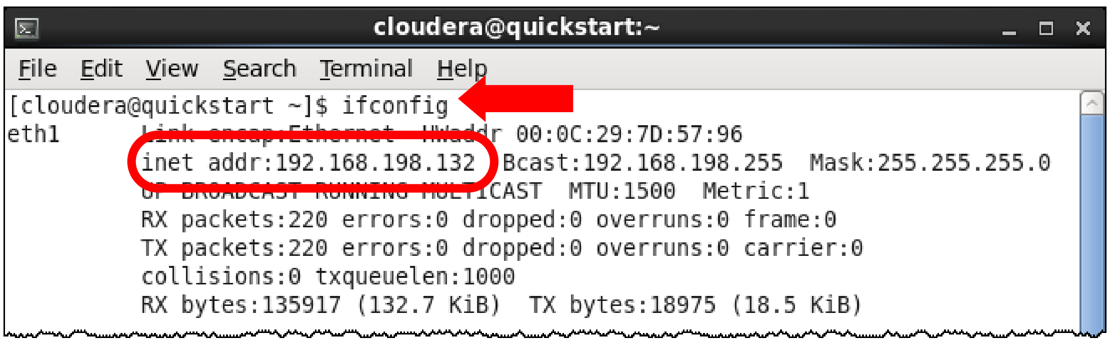


Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

3. Open a terminal window on the Cloudera QuickStart VM by selecting **Applications** ⇨ **System Tools** ⇨ **Terminal**.



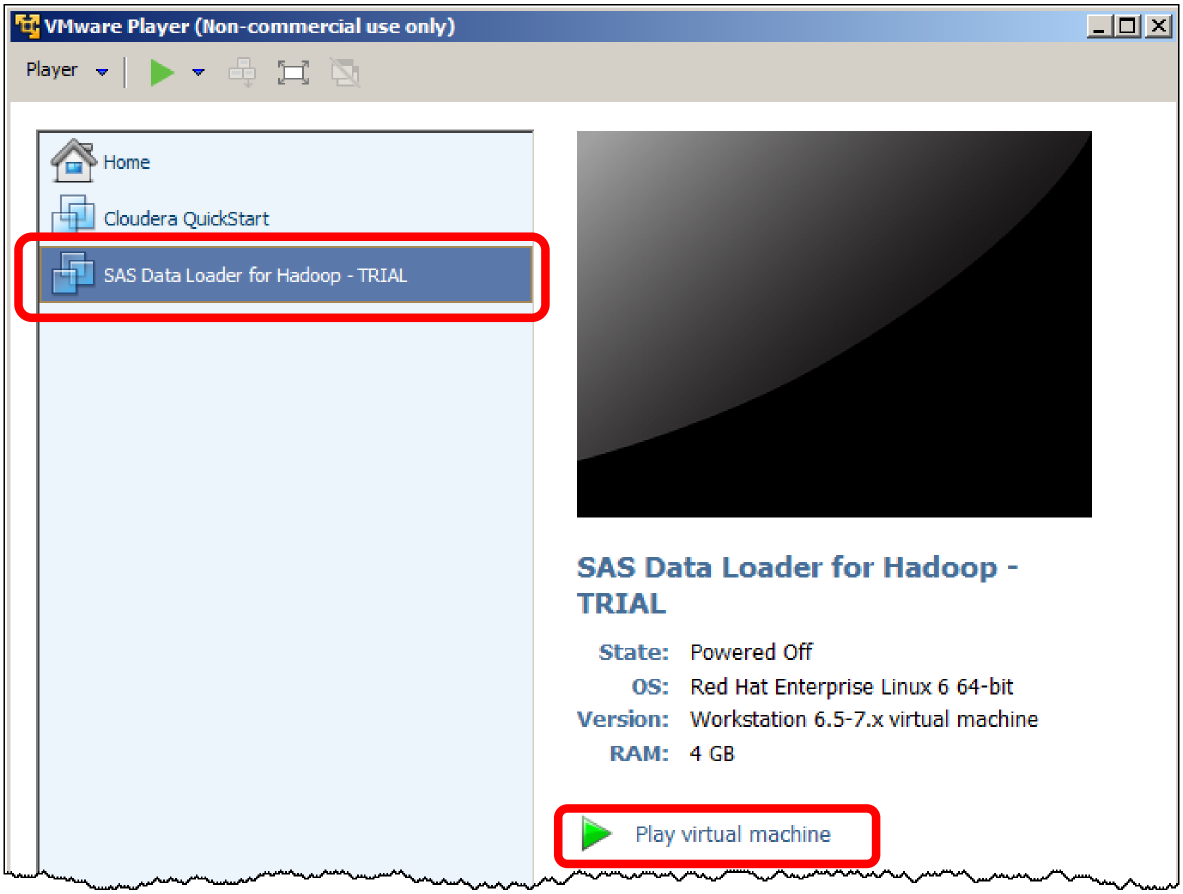
4. In the terminal window, type **ifconfig** and press Enter.
5. Locate the IP address for the virtual machine.



6. Start another instance of VMware Player by selecting **Start** ⇨ **All Programs** ⇨ **VMware** ⇨ **VMware Player**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

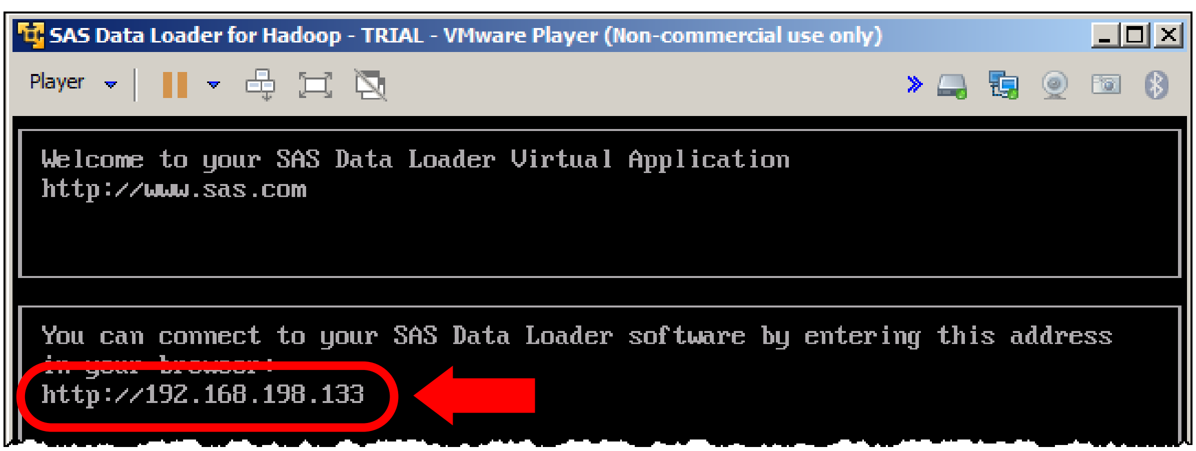
7. Single-click the virtual machine labeled **SAS Data Loader for Hadoop - TRIAL** and then click **Play virtual machine**.



It takes a few minutes for the virtual machine to start.

After it is started, the VMware Player displays a black background with a "Welcome to you SAS Data Loader Virtual Application" message.

8. Locate the address for accessing your instance of the SAS Data Loader.



IMPORTANT: To get a desktop cursor back, you might need to press Ctrl+Alt.

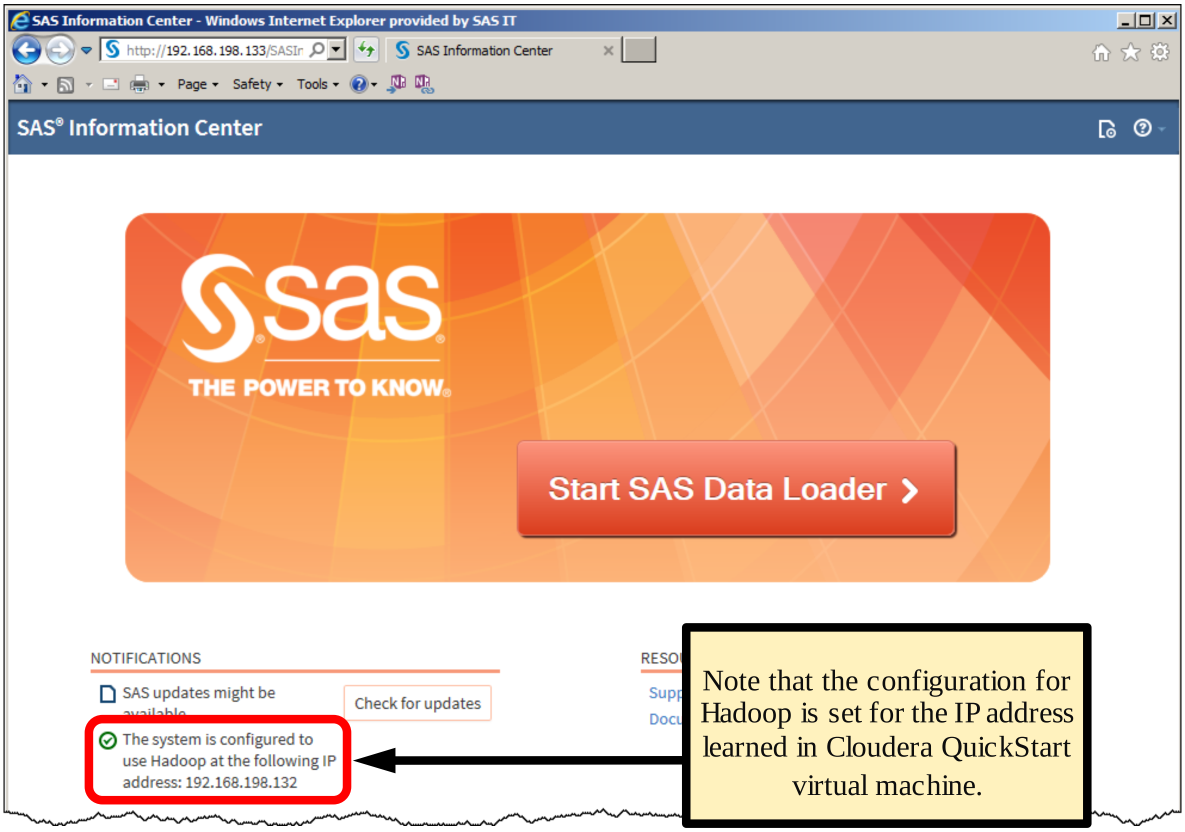
9. From the desktop machine, select **Start** ⇨ **Internet Explorer**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

10. Enter the address for the instance of the SAS Data Loader on ***your*** machine.

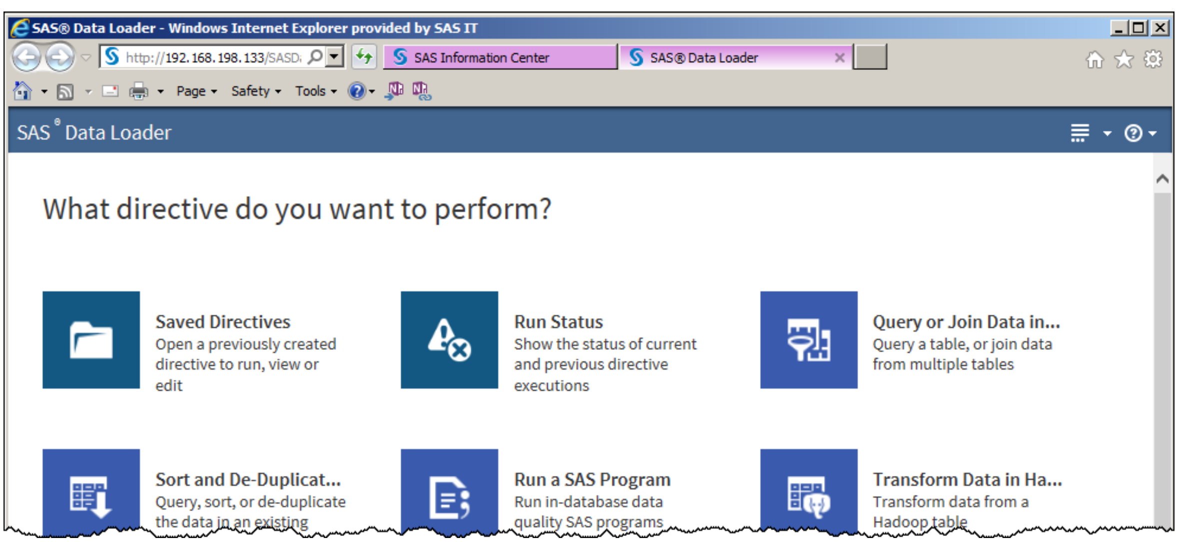
    It might take a few minutes for the SAS Data Loader for Hadoop to initialize with a connection to the Cloudera QuickStart sandbox.

    The SAS Information Center appears.



11. Click **Start SAS Data Loader**.

    The SAS Data Loader opens on a new tab in the web browser.



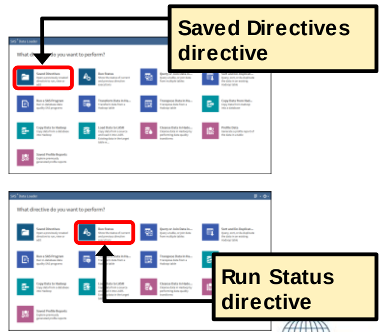
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

Directives

Data is managed through the development of ***directives*** that are executed as jobs.

Directives, after they are created, can be saved for reuse later or shared in multi-user environments.

When jobs are executing, you can monitor their status, and access the code and logs that are generated for each jobs.



13

SAS Data Loader for Hadoop builds directives as jobs. Each job generates and displays executable code, which can be edited and saved for reuse. SAS DS2 programs, DS2 expressions, and HiveQL expressions can be dropped into directives to repeat execution and simplify job management.

Directive Categories

The SAS Data Loader for Hadoop provides the following categories of directives:

- Copy data to and from Hadoop
- Manage data in Hadoop
- Profile data
- Manage jobs

14

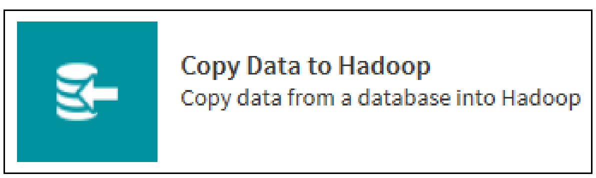
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

The SAS Data Loader for Hadoop provides the following categories of directives:

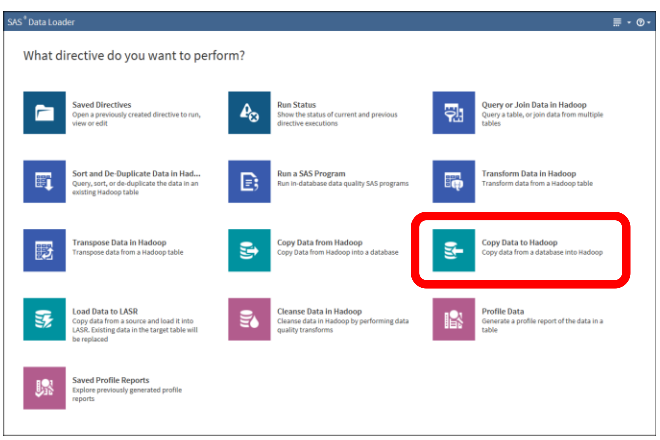
Copy data to and from Hadoop	Copy data as needed to and from SAS and databases outside of Hadoop. Also copy data out to SAS LASR Analytic Servers for analysis with SAS Visual Analytics and SAS Visual Statistics.
Manage data in Hadoop	Directives support combinations of queries, summarizations, joins, 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.
Profile data	Profile jobs examine the quality and content of tables and produce reports. The reports are stored and managed for future reference. When you select source and target tables for your jobs, you can open the profile reports of the tables that have been profiled.
Manage jobs	When jobs are created, they can be saved them for later execution and edit. When jobs are run, the Run Status directive shows the run status and enables the execution to terminate and return the job for edit in a directive.

Directives: Copy Data to Hadoop / Run Status

1. From the SAS Data Loader directives page, click the directive **Copy Data to Hadoop**.



In the list of directives, Copy Data to Hadoop is located in this position:

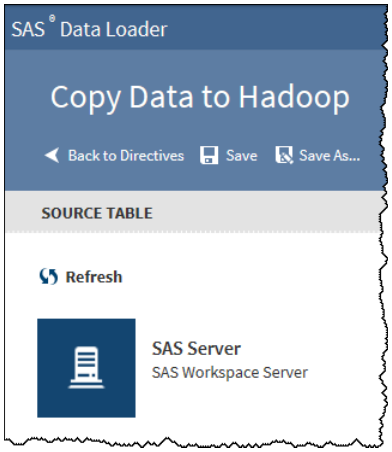


The Copy Data to Hadoop directive enables you to copy data to Hadoop using SQOOP.

We will work with a sample set of data provided as part of the download for the SAS Data Loader for Hadoop – Trial version.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

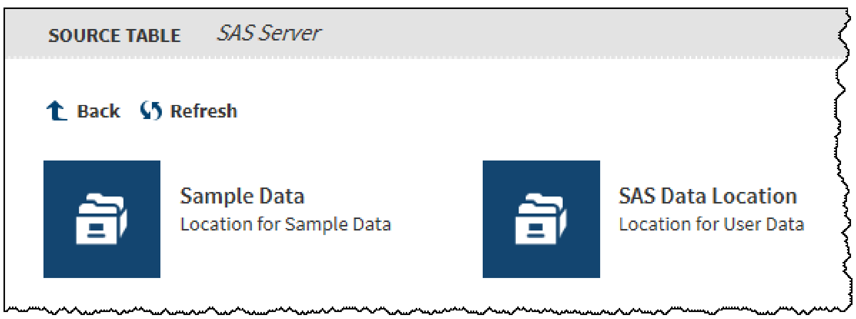
The view in SAS Data Loader should resemble the following:



We will copy a SAS data set from the defined SAS Server.

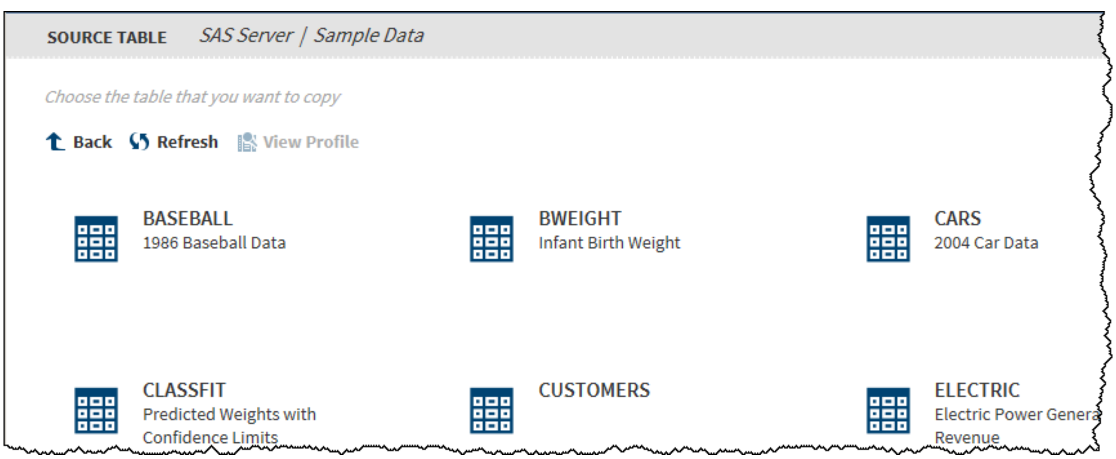
2. Click **SAS Server**.

The view in SAS Data Loader should resemble the following:



The data is located in the Sample Data schema.

3. Click **Sample Data**.

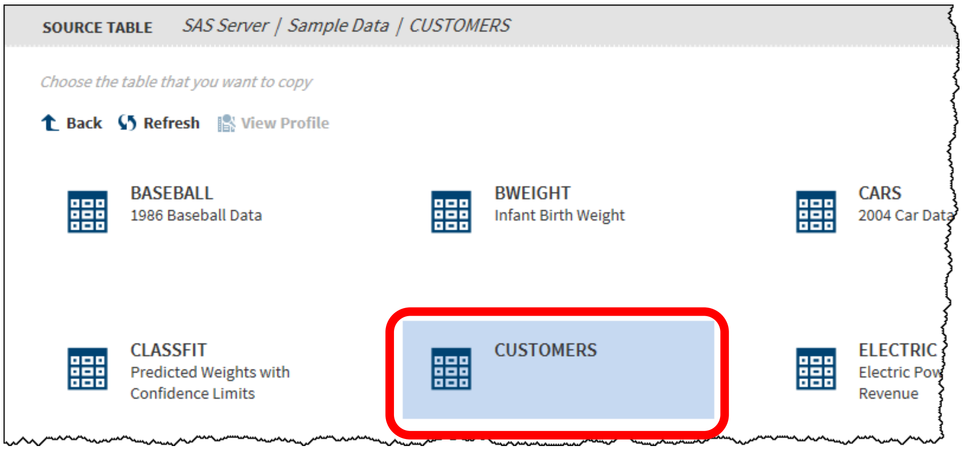


The first data we will copy is the **CUSTOMERS** table.

We have filtering and column options when selecting data.

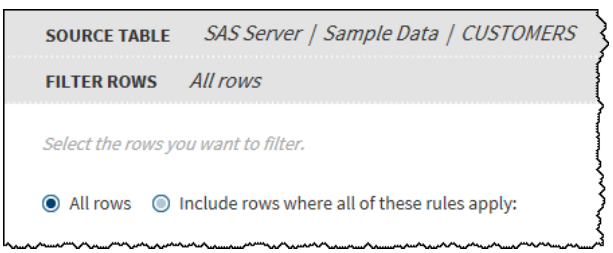
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

4. Click **CUSTOMERS**.
5. Click **Next**.



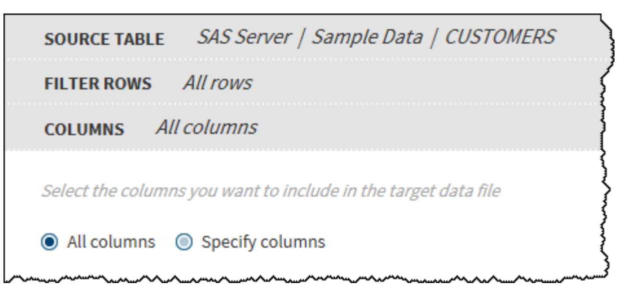
We will use all rows of data (the default selection).

6. Click **Next**.



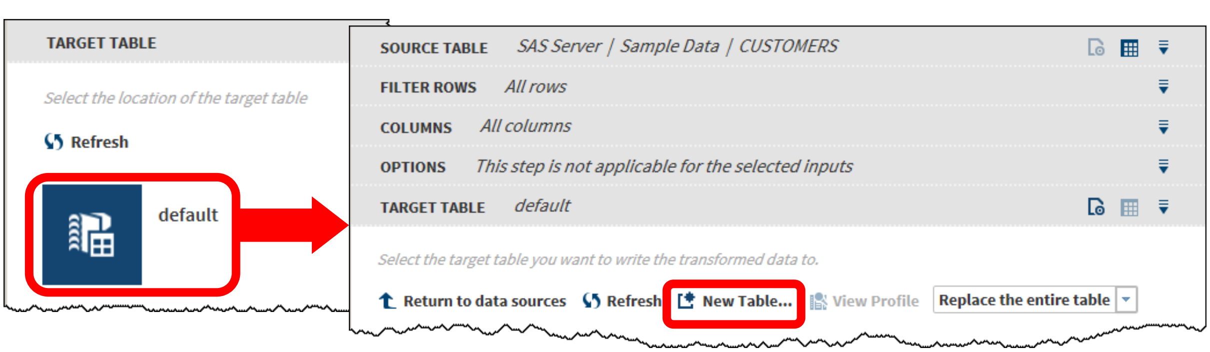
We will use all columns of data (the default selection).

7. Click **Next**.



We will build the table in the **default** schema.

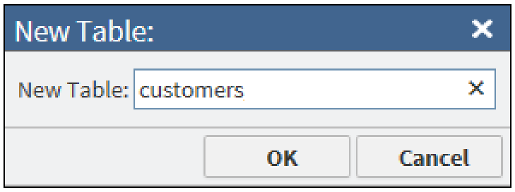
8. Click **default**.



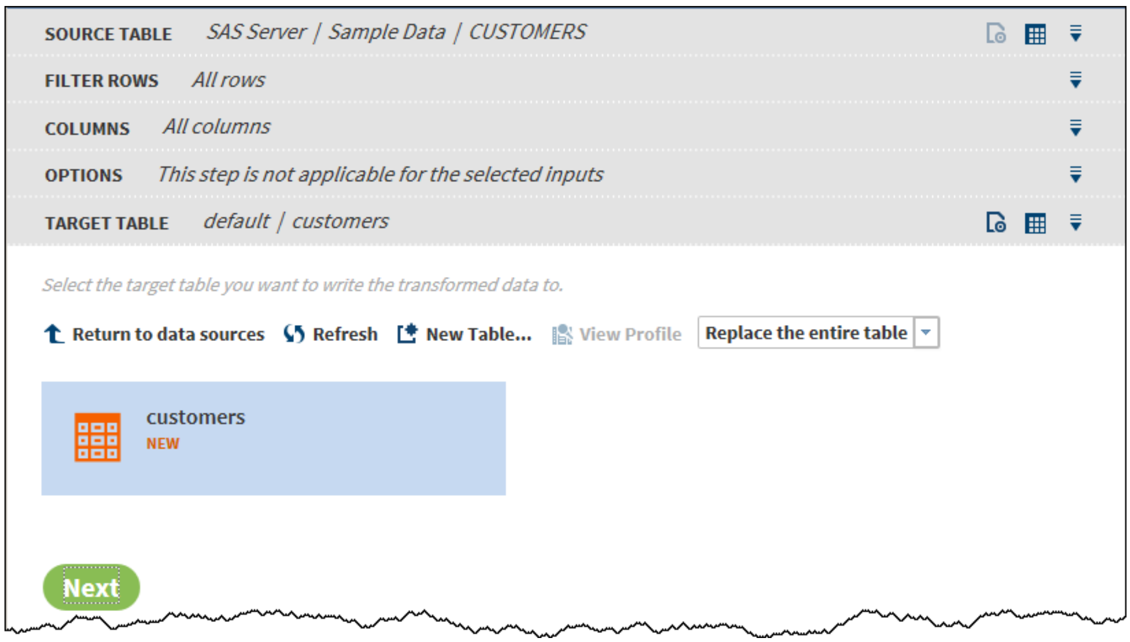
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

9. Click **New Table** to open the New Table window.

   a. Enter **customers**.

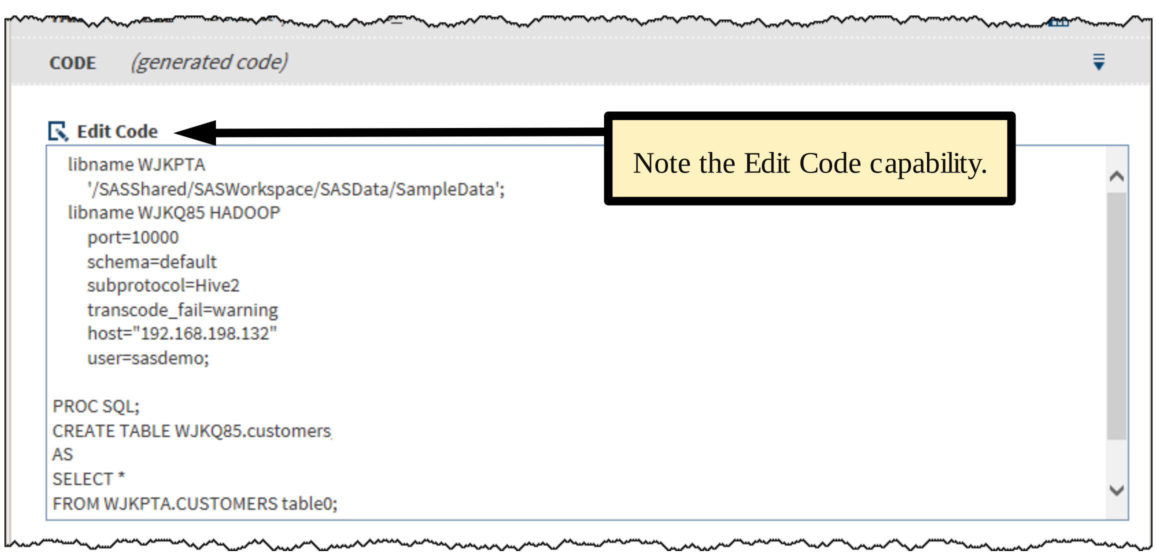


   b. Click **OK**.



10. Click **Next**.

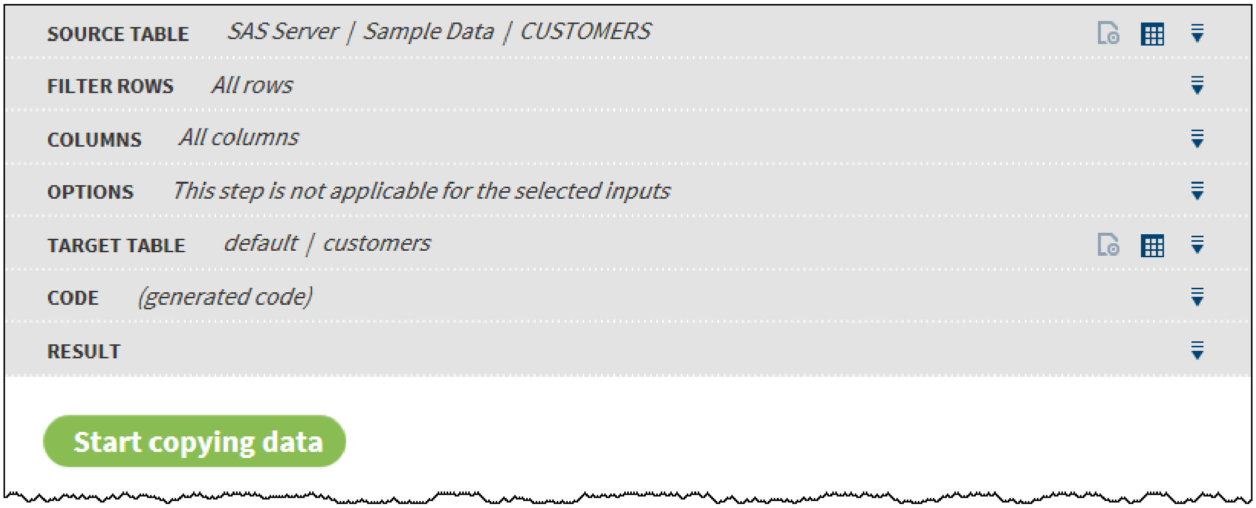
    We can now view the code that is generated to create this new table. This code is editable.



11. Click **Next**.

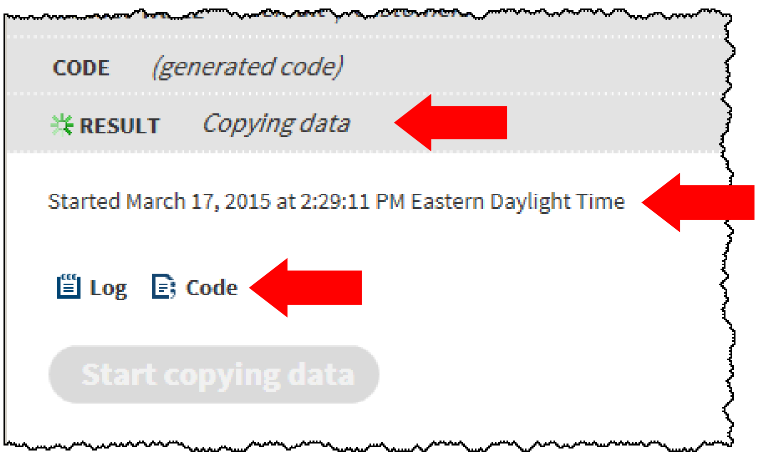
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

With all preparation steps specified, we are now able to start copying the data.

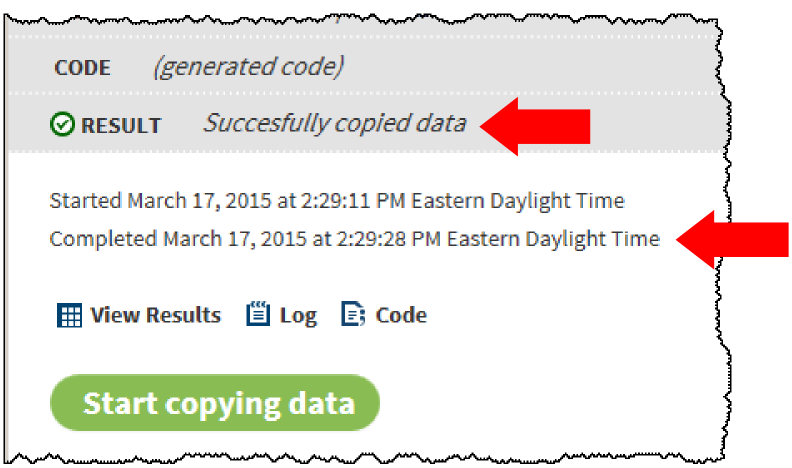


12. Click **Start copying data**.

We see a status of **Copying data** for the RESULT area, a start time, a code file, and a log file.



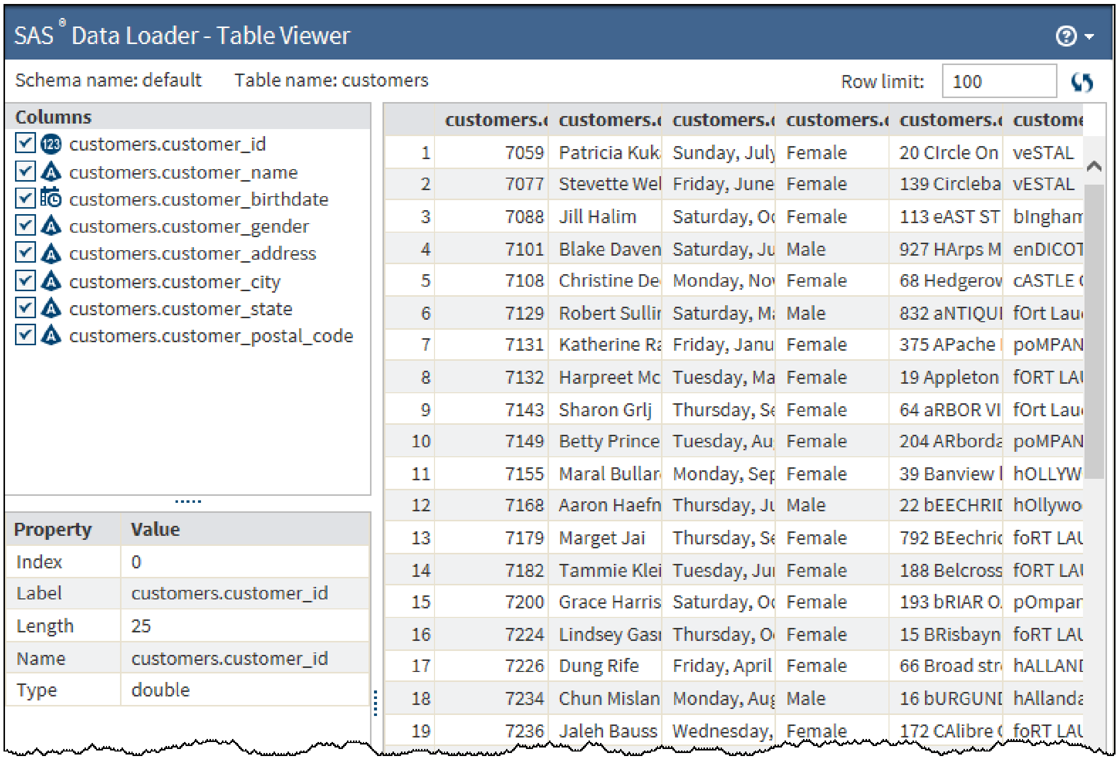
When the processing is complete, we see a **Successfully copied data** message, as well as a completion time.



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

13. Click **View Results**.

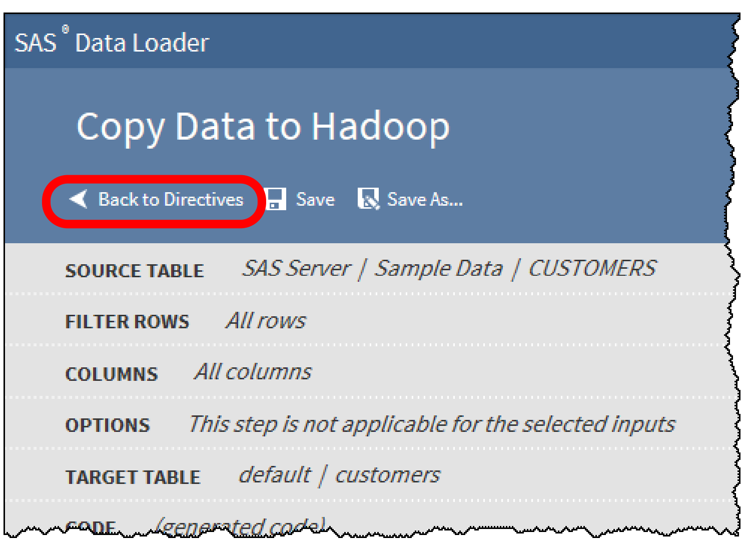
    We can see the result of the copied data in the Table Viewer.



This Table Viewer appears on a new tab in Internet Explorer.

Now we want to repeat these steps for the **ORDERS** table.

14. Click the Internet Explorer tab labeled **SAS Data Loader**.
15. Click **Back to Directives**.



16. From SAS Data Loader directives page, click the directive **Copy Data to Hadoop**.
17. Click **SAS Server**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

18. Click **Sample Data**.
19. Click **ORDERS**.
20. Click **Next**.

    We will use all rows of data (the default selection).
21. Click **Next**.

    We will use all columns of data (the default selection).
22. Click **Next**.

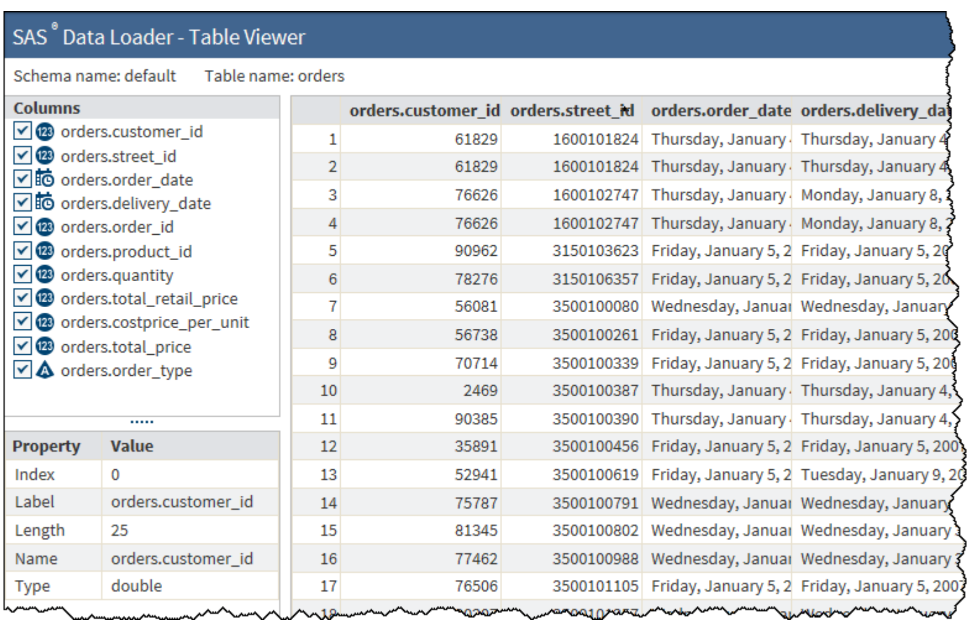
    We will build the table in the **default** schema.
23. Click **default**.

    We will name our new table **orders**.
24. Click **New Table** to open the New Table window.
25. Enter **orders**.
26. Click **OK**.
27. Click **Next**.

    The code is generated.
28. Click **Next**.

    We are ready to copy a second data set.
29. Click **Start copying data**.

    The data should copy successfully.
30. Click **View Results**.



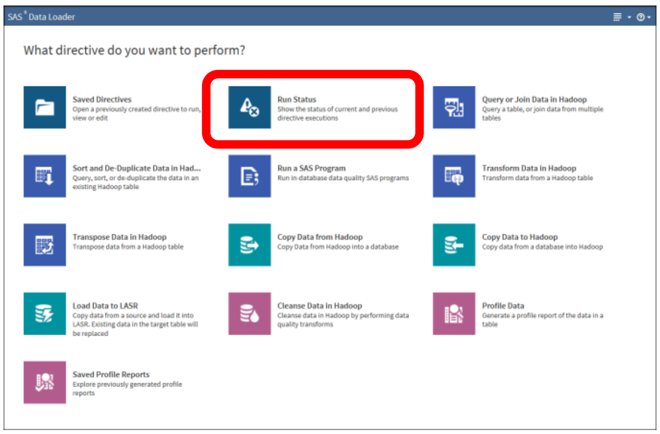
31. Click the Internet Explorer tab labeled **SAS Data Loader**.
32. Click **Back to Directives**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

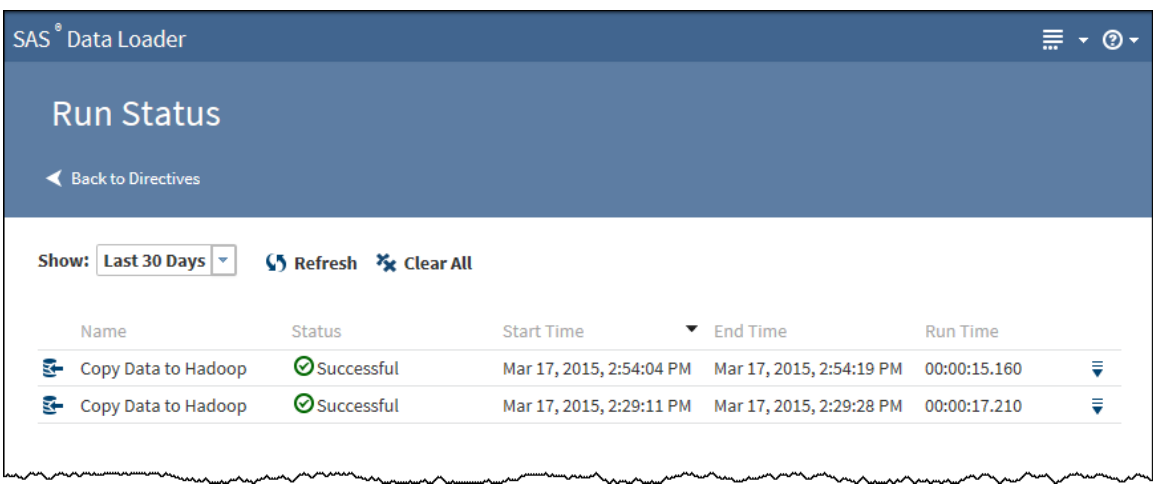
We want to investigate the status of the jobs that have run.

33. Click the directive **Run Status**.

In the list of directives, Run Status is located in this position:



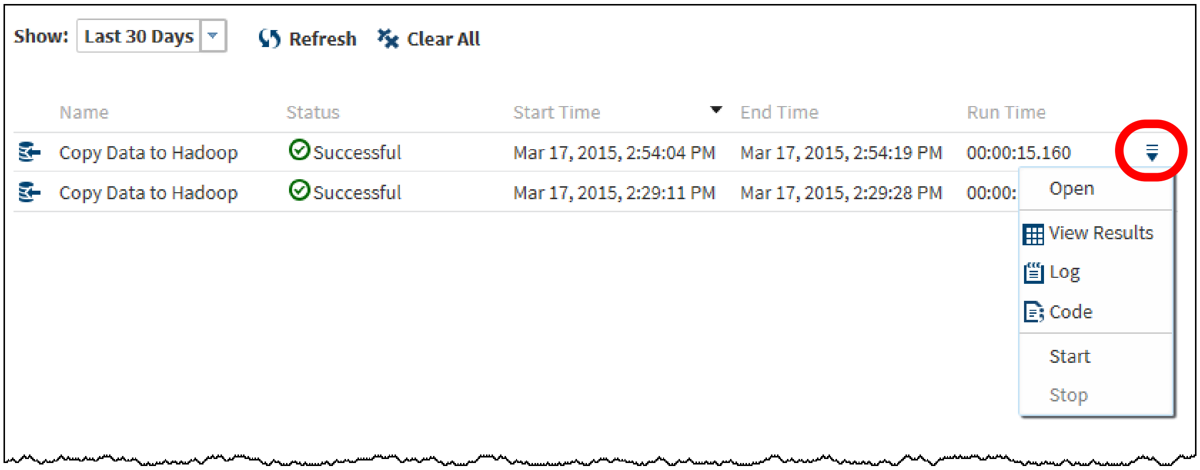
The Run Status directive shows us the status of each job that has been run.



For each job, if the job has run and if it was successful, we can see the starting and ending times, as well as the total execution time.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

34. Click to the right of the first job.

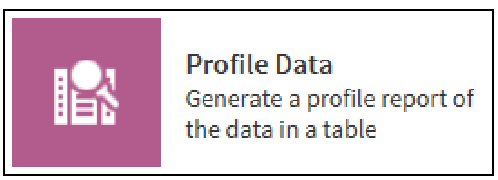


This menu enables you to view the results, the log, and the code. The log and the code are viewed in a text editor.

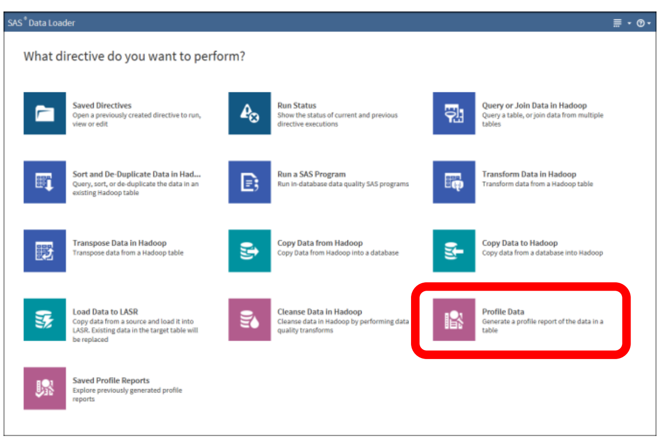
35. Click **Back to Directives**.

Directives: Profile Data / Saved Profile Reports

1. From the SAS Data Loader directives page, click the directive **Profile Data**.

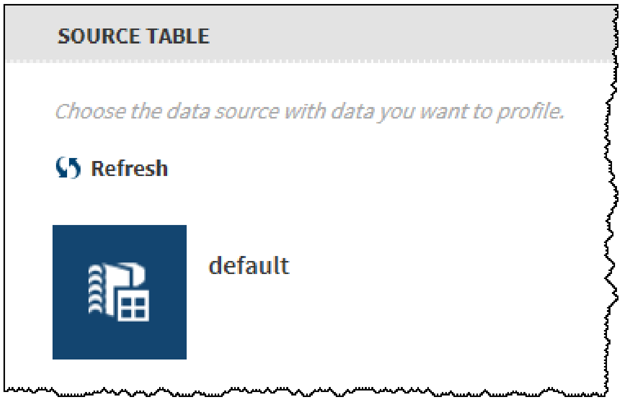


In the list of directives, Profile Data is located in this position:

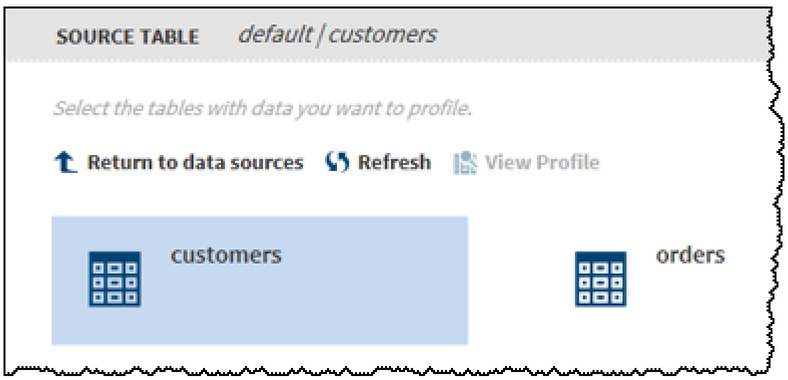


Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

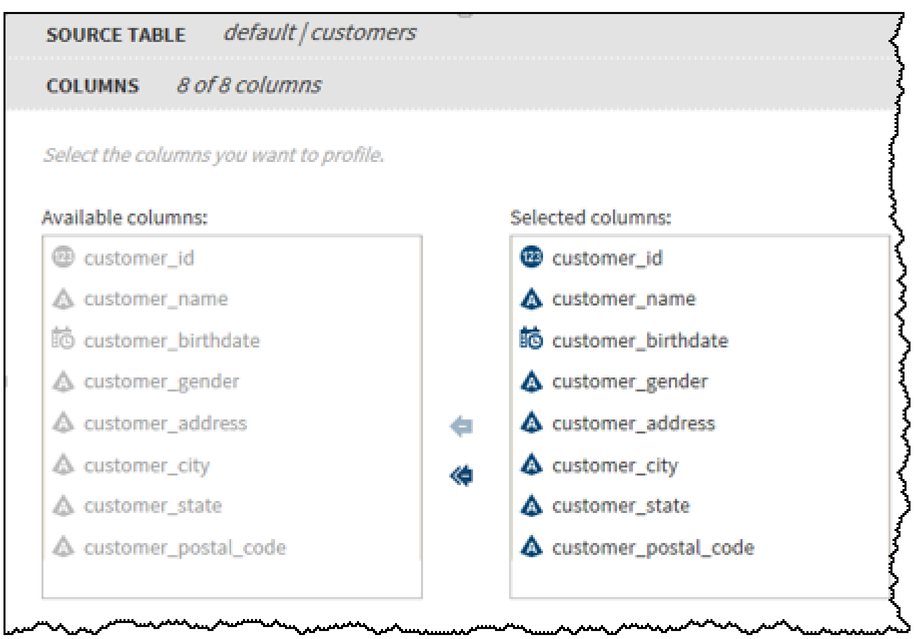
2. Click the **default** schema.



3. Click the **customers** table.



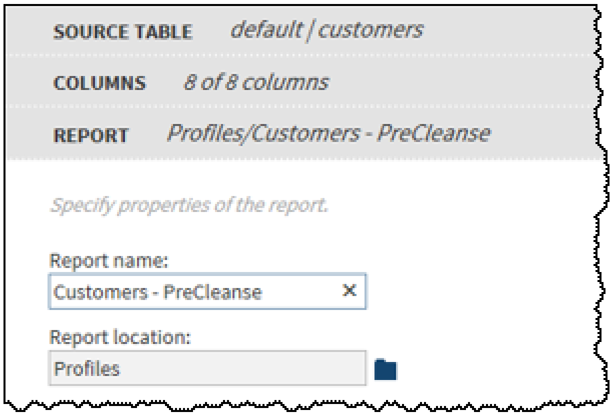
4. Click **Next**.
5. Verify that all columns are selected.



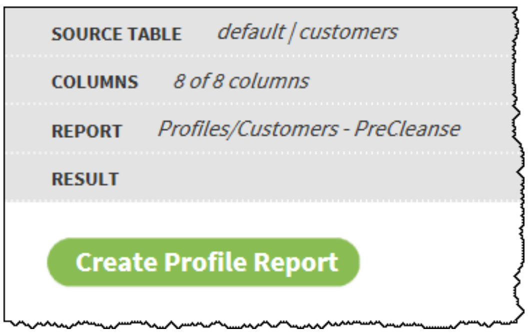
6. Click **Next**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

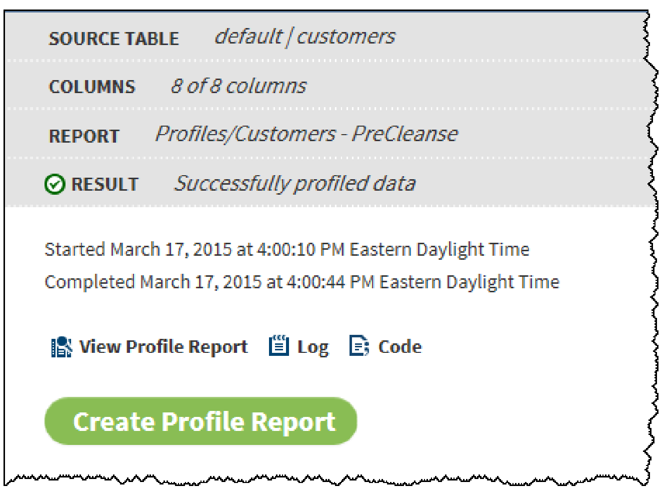
7. Enter **Customers - PreCleanse** for the report name.



8. Click **Next**.
9. Click **Create Profile Report** to start processing.

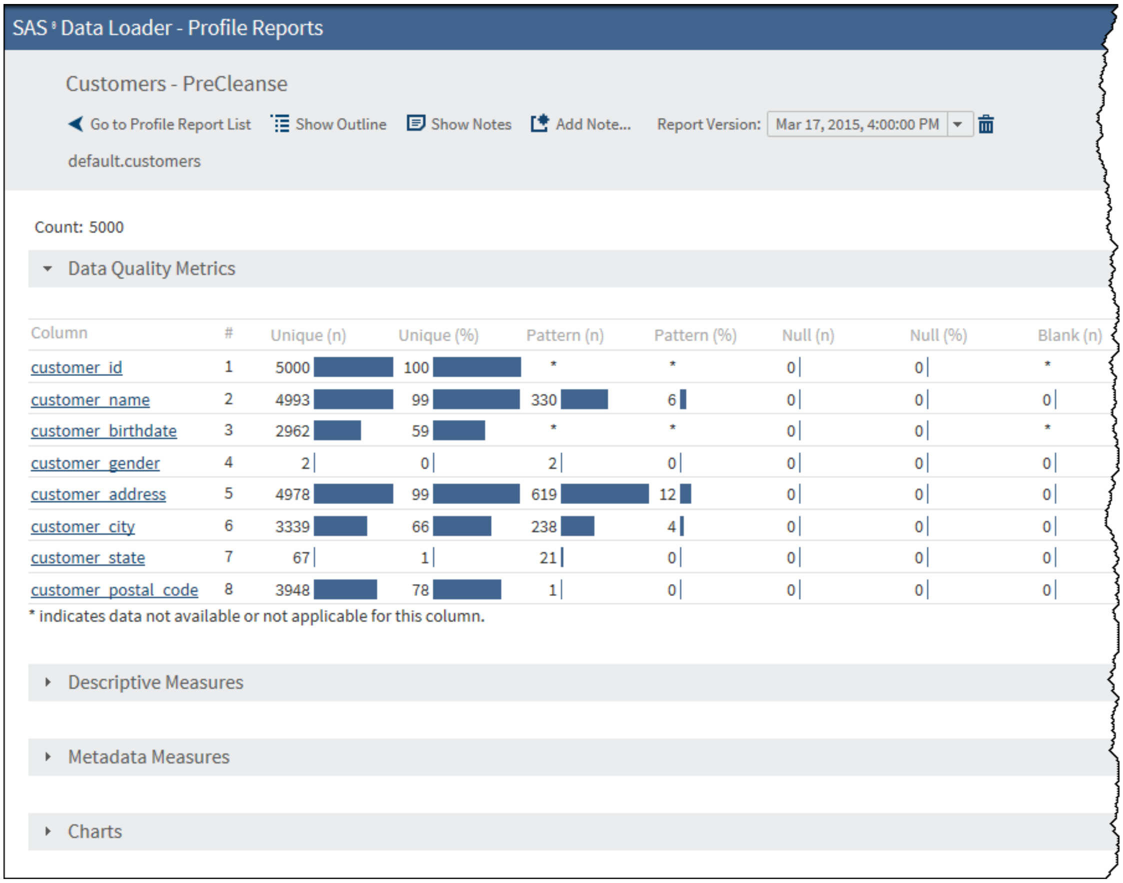


When the processing is complete, a completion time and a link to view the profile report appear.



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

10. Click **View Profile Report**.



Column	#	Unique (n)	Unique (%)	Pattern (n)	Pattern (%)	Null (n)	Null (%)	Blank (n)
customer_id	1	5000	100	*	*	0	0	*
customer_name	2	4993	99	330	6	0	0	0
customer_birthdate	3	2962	59	*	*	0	0	*
customer_gender	4	2	0	2	0	0	0	0
customer_address	5	4978	99	619	12	0	0	0
customer_city	6	3339	66	238	4	0	0	0
customer_state	7	67	1	21	0	0	0	0
customer_postal_code	8	3948	78	1	0	0	0	0

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

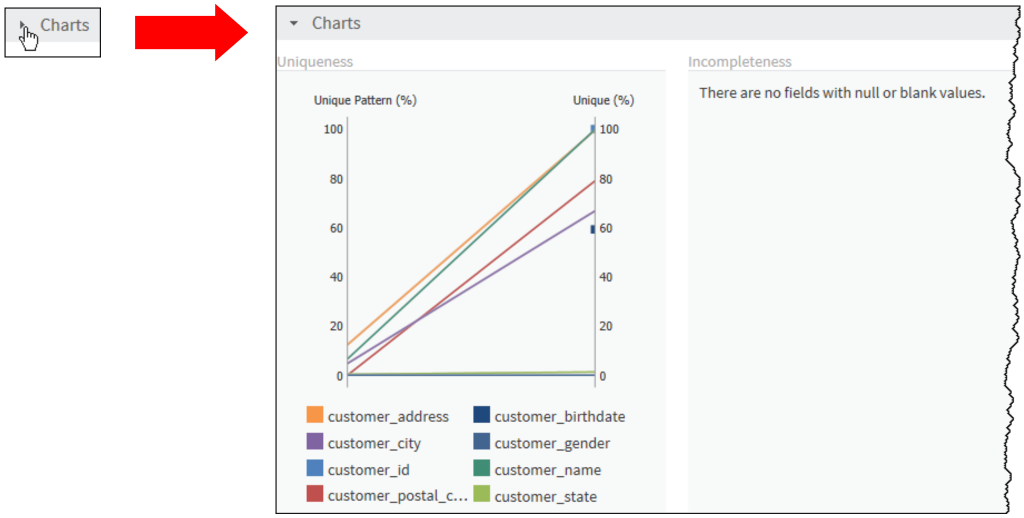
By default, the profile report displays overall table metrics in three groupings:

- Data Quality Metrics (expanded in above display capture)
- Descriptive Measures
- Metadata Measures

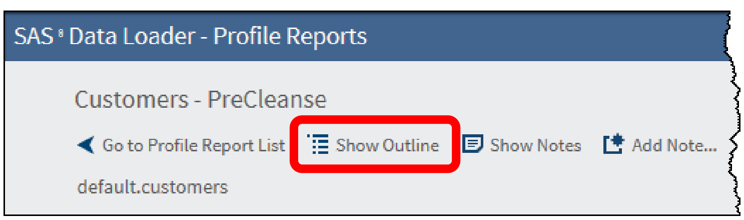
In addition, there is an area for two table graphics (Uniqueness and Incompleteness).

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

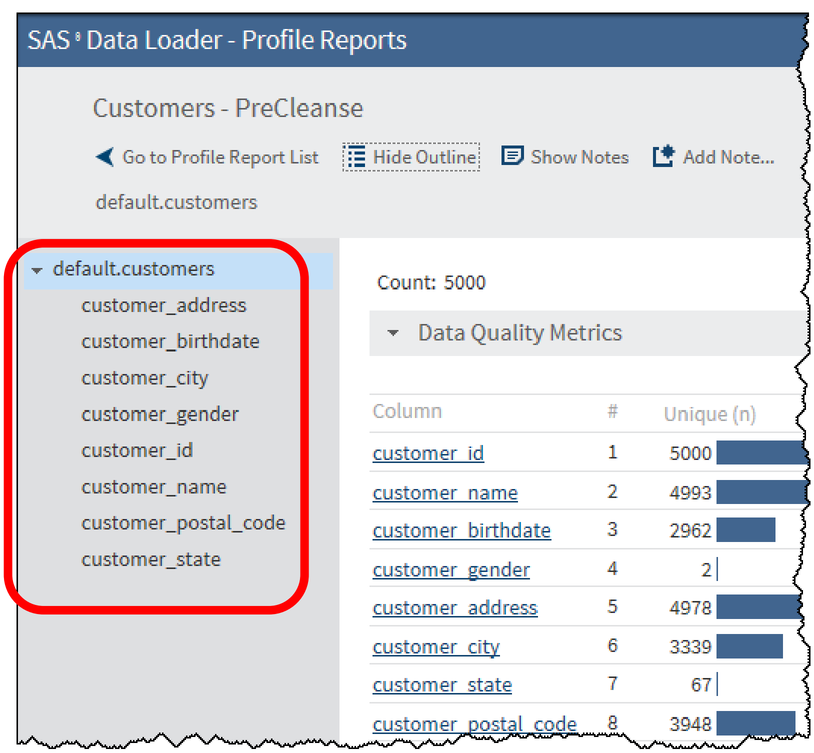
11. Click ▸ in Charts gray header bar to expand and display default graphics.



12. Click ☰ (**Show Outline**).



The left panel of the report viewer now is displayed, where we have easy access to a list of fields from the **customers** table. (You might need to click ▸ next to a table to display the fields.):



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

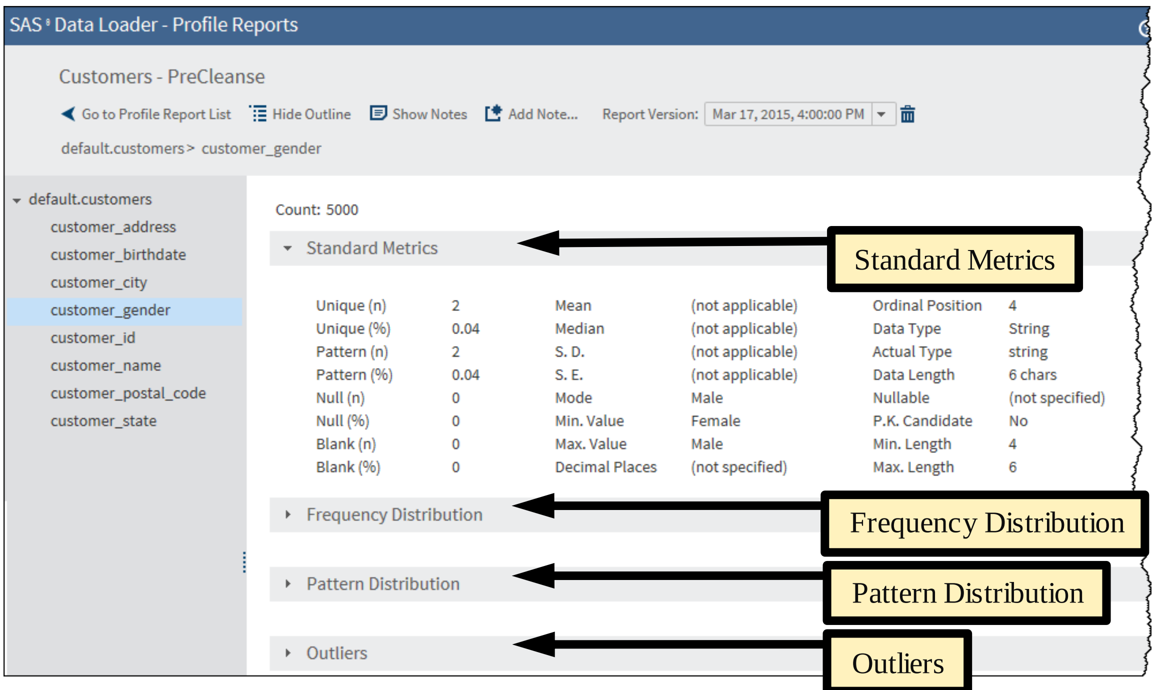
The Show Outline text next to the tool is now toggled to **Hide Outline**.

13. Click the **customer_gender** field in the left navigation pane (outline).

    This action drills the report view to look at metrics just for the **customer_gender** field. Field-level metrics are separated into four groupings:

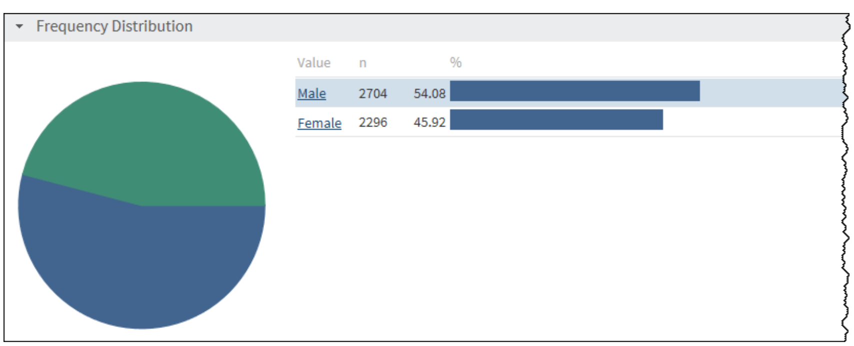
    - Standard Metrics
    - Frequency Distribution
    - Pattern Frequency Distribution
    - Outliers

    The Standard Metrics area is expanded by default.



14. Click the down-arrow button to collapse the **Standard Metrics** gray header bar.

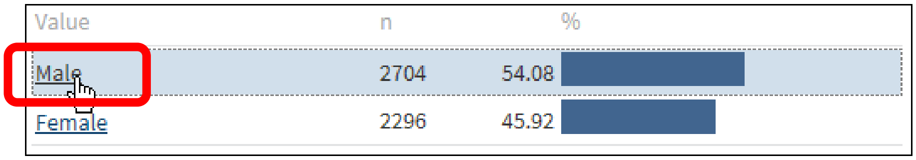
15. Click the right-arrow button in **Frequency Distribution** gray header bar to expand and display default graphic and table.



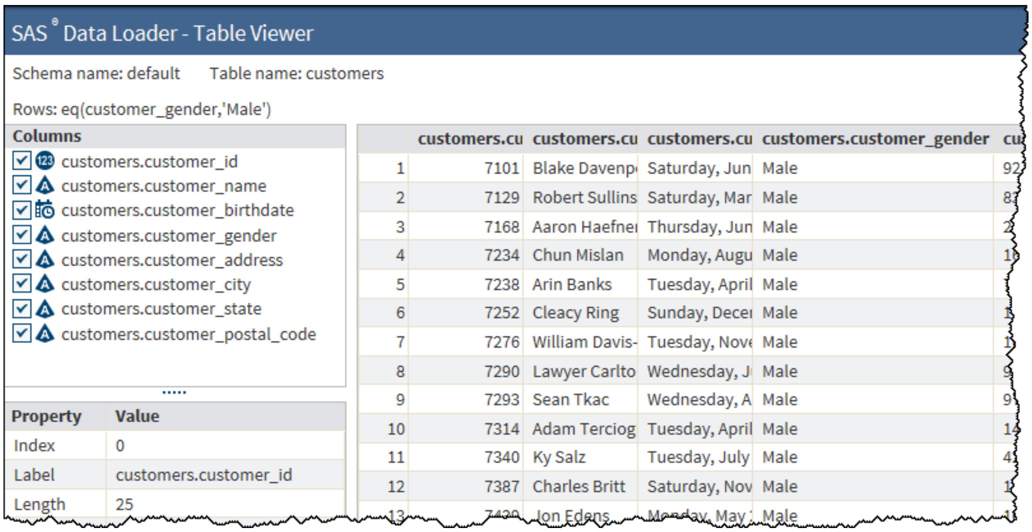
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

16. Place the mouse pointer on a slice of the pie chart and note that the same information is highlighted in the details table.

    a. Select the **Male** value in the details table.



This opens the table viewer (on a new tab) that has subset the data for **customer_gender** = '**Male**'.

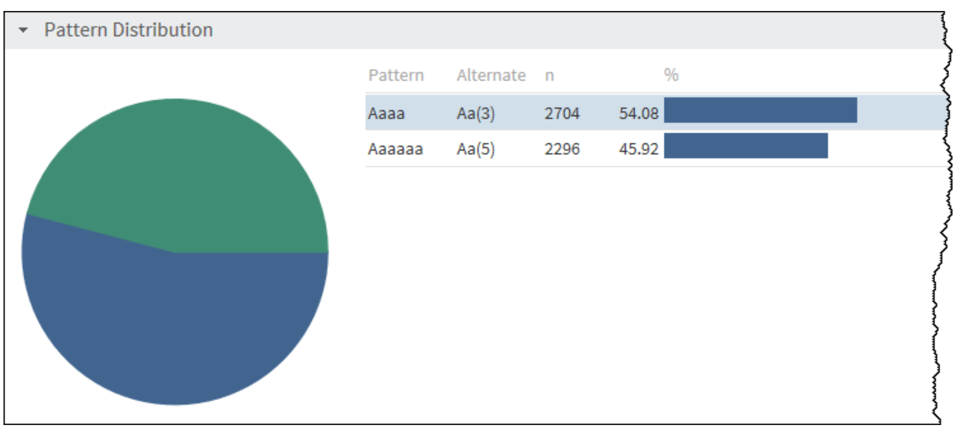


17. Close the tab in Internet Explorer that displays the table.

18. If necessary, click the tab in Internet Explorer that displays the profile report.

19. Click to collapse the **Frequency Distribution** gray header bar.

20. Click in the **Pattern Distribution** gray header bar to expand and display the default graphic and table.



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

21. Click the **customer_state** field in the left navigation pane (outline).

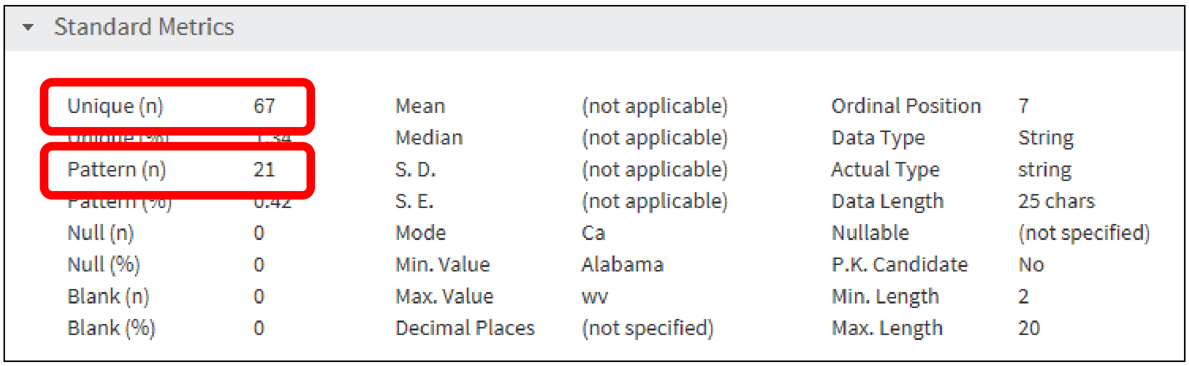
    QUESTION: Using the Standard Metrics list of values, what is the unique count [Unique (n)] for the **customer_state** field?

    ANSWER: 67. We know that there are not 67 unique states. Thus, further investigation of the state values is needed.

    QUESTION: Using the Standard Metrics list of values, what is the pattern count [Pattern (n)] for the **customer_state** field?

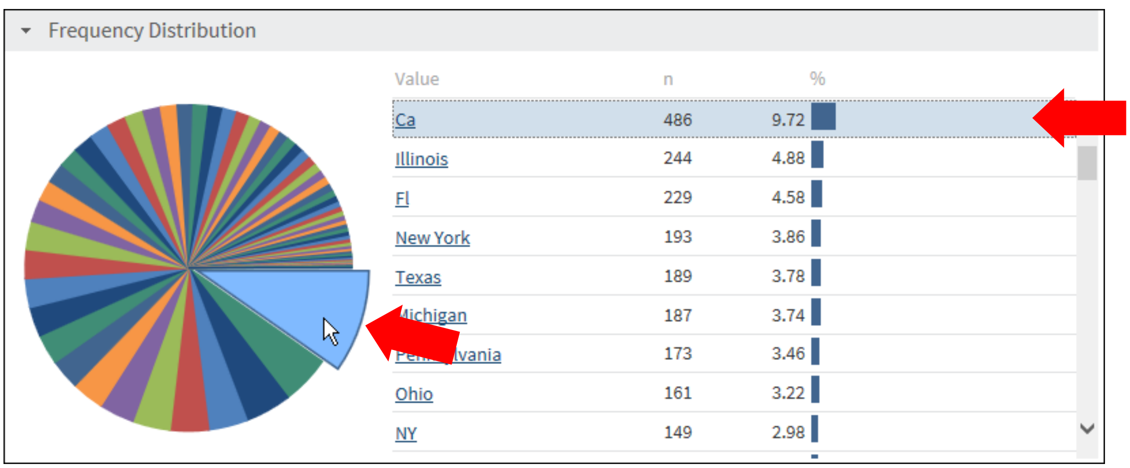
    ANSWER: 21. It is desired that all state values be two capital letters.

    The displayed standard metrics should resemble the following:



    QUESTION: Using the Frequency Distribution graph and report, which value of **state** has the highest frequency count?

    ANSWER: Place the mouse pointer on the largest slice of pie and note that the CA value is highlighted in the table.

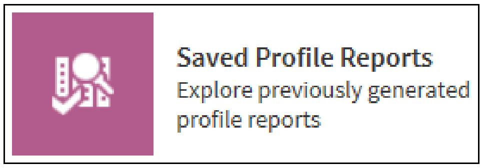


22. Close the tab in Internet Explorer that displays the profile.

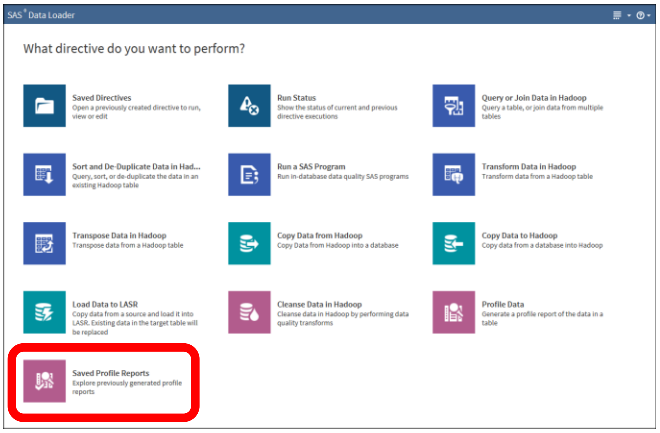
23. If necessary, click the tab in Internet Explorer that is the main SAS Data Loader tab (where the directives are shown).

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

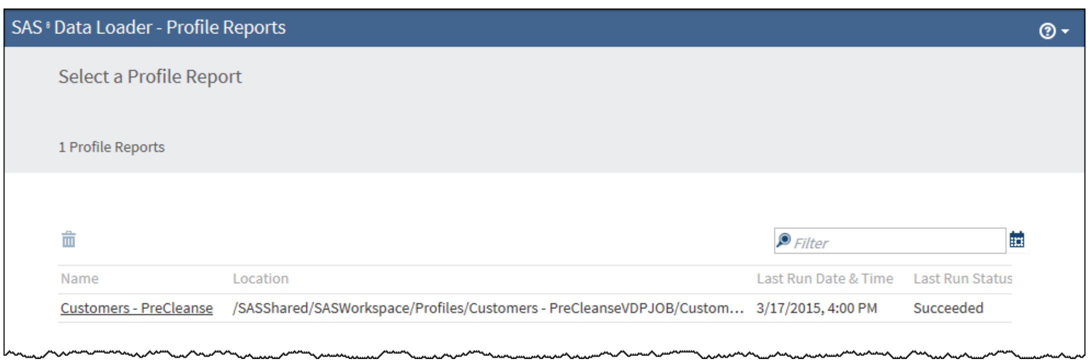
24. From the SAS Data Loader directives page, click the directive **Saved Profile Reports**.



In the list of directives, Saved Profile Reports is located in this position:



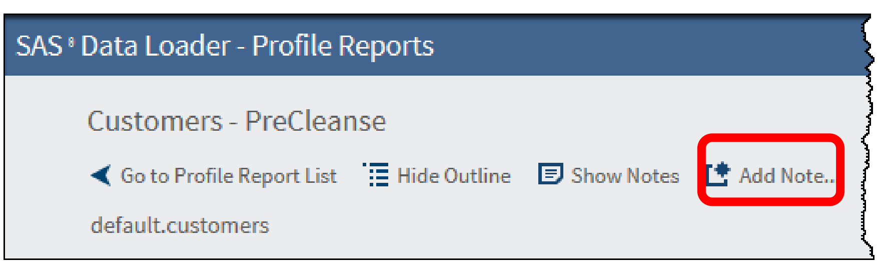
25. Verify that the one profile report created is listed.



26. Click the link (defined report name) **Customers - PreCleanse** to surface the profile report on a new Internet Explorer tab.

Any capabilities previously explored are all available when you open the profile report from this directive.

27. Click (**Add Note**).

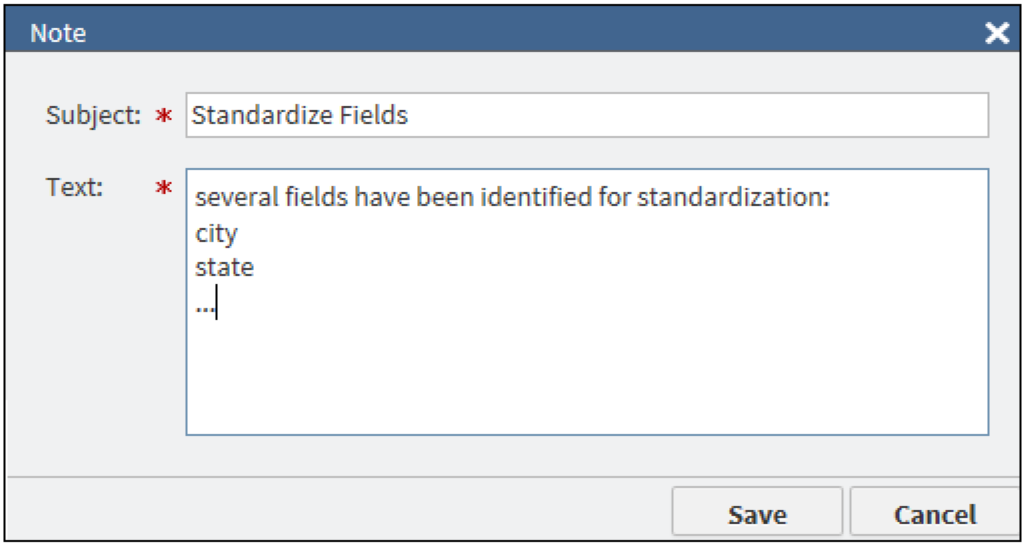


28. Specify note information.

a. Enter **Standardize Fields** in the **Subject** field.

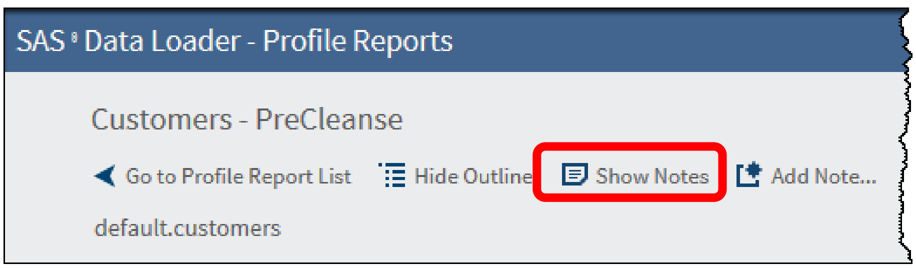
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

b. Enter **several fields have been identified for standardization: city state …** in the **Text** field.

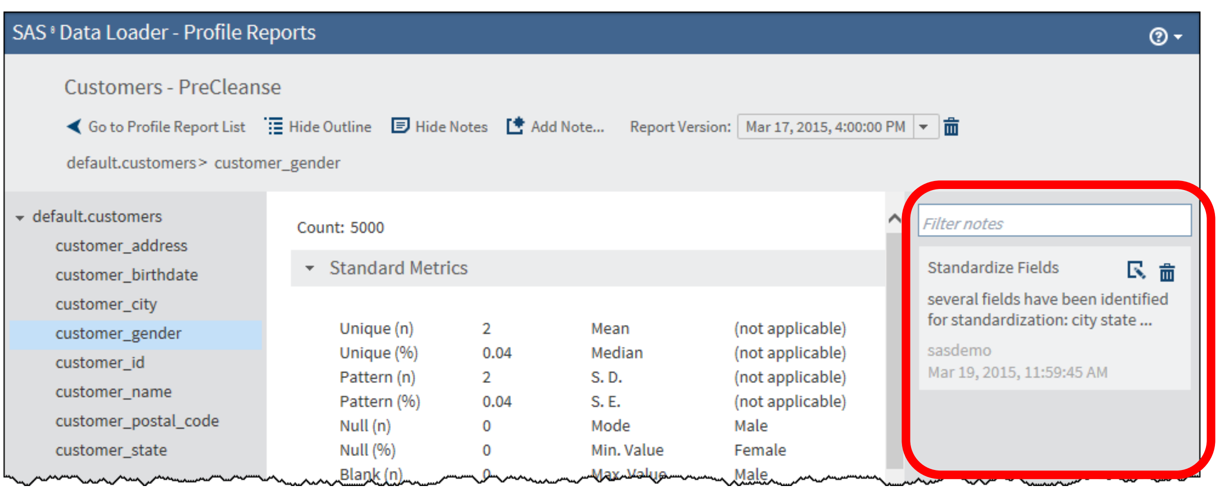


c. Click **Save**.

29. To surface any notes that have been added to the saved profile, click (**Show Notes**).



The ***right*** panel of the report viewer now displays the note entered:



The Show Notes text next to the tool is now toggled to **Hide Notes**.

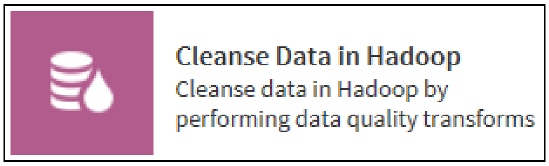
30. Close the tab in Internet Explorer that is viewing the profile.

31. If necessary, click the tab in Internet Explorer that is the main SAS Data Loader tab (where the directives are shown).

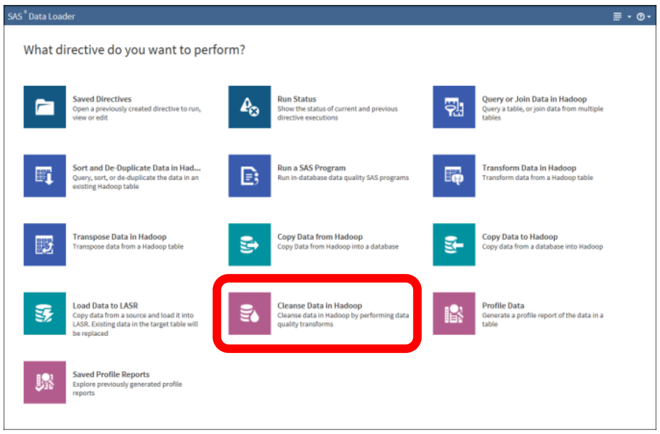
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

Directive: Cleanse Data in Hadoop

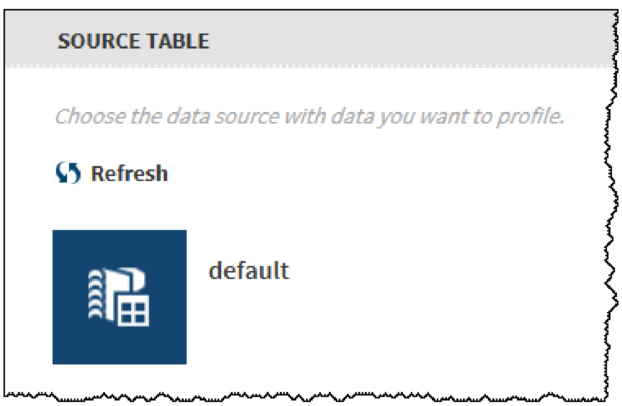
1. From SAS Data Loader directives page, click the directive **Cleanse Data in Hadoop**.



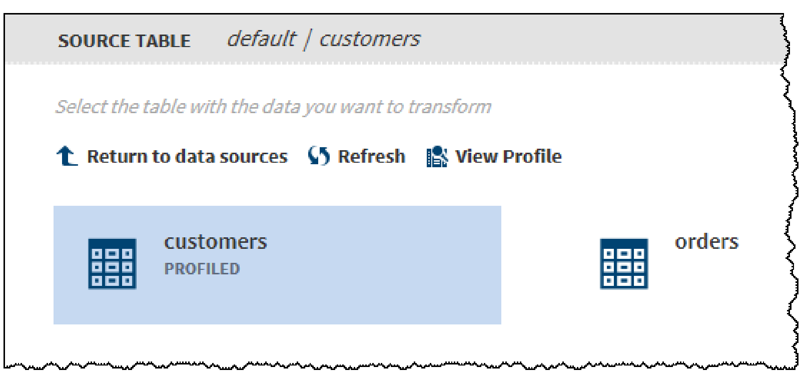
In the list of directives, Cleanse Data in Hadoop is located in this position:



2. Click the **default** schema.



3. Click the **customers** table.

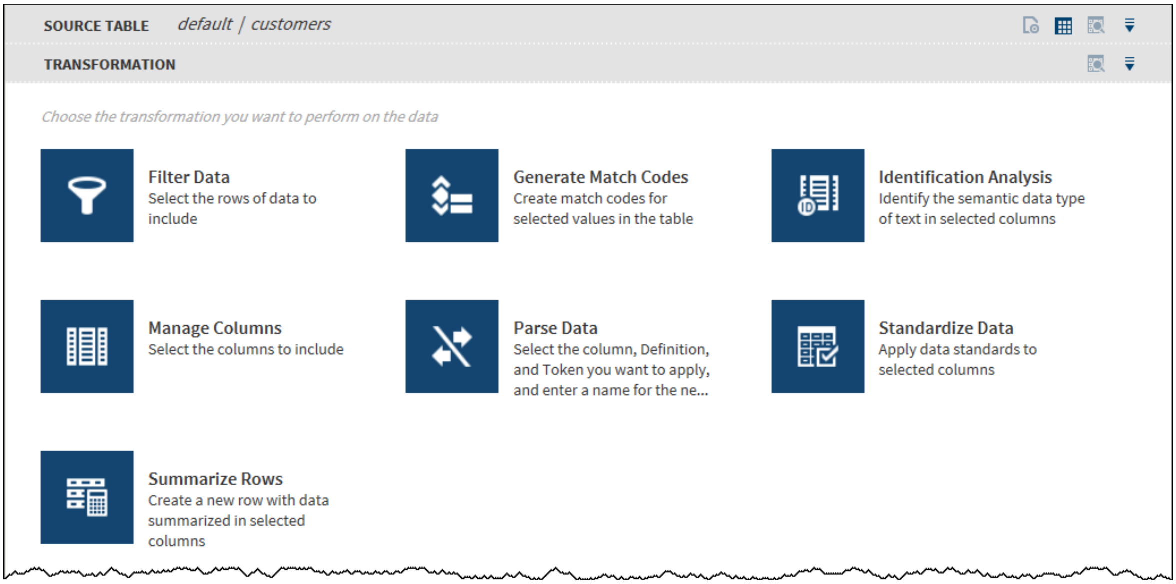


Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

The **customers** table was profiled, and this is indicated under the table name. This view has a link to View Profile for the selected table.

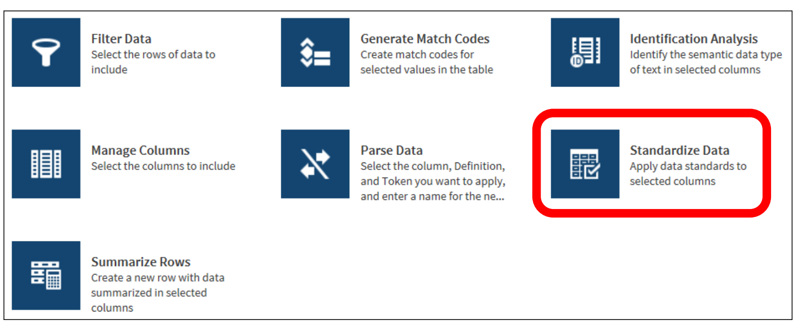
4. Click **Next**.

   With the source table selected, we need to select the type of cleansing to perform.



5. Click the **Standardize Data** transformation.

   In the list of transformations, Standardize Data is located in this position:

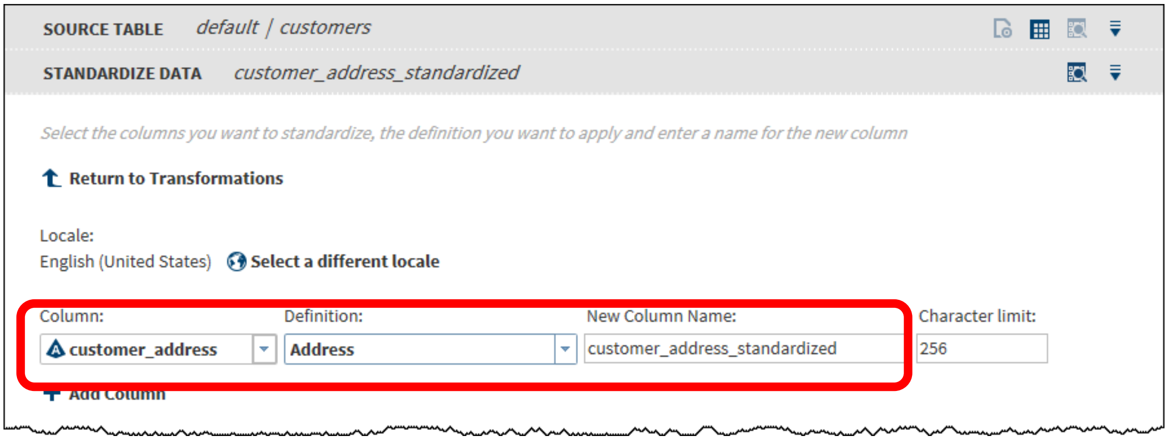


6. Click under **Column** and select **customer_address**.

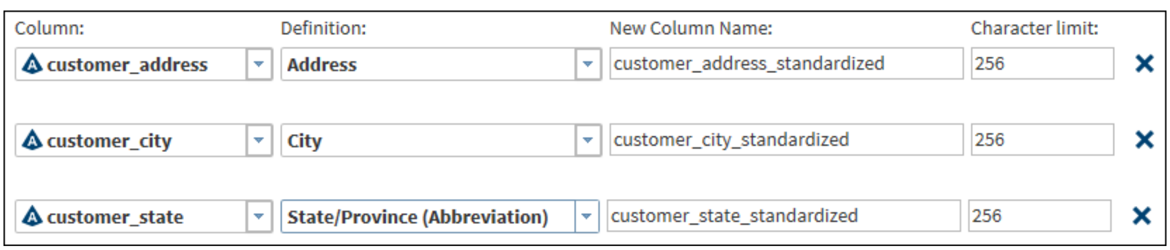
7. Click under **Definition** and select **Address**.

8. Verify that **New Column Name** is **customer_address_standardized**.

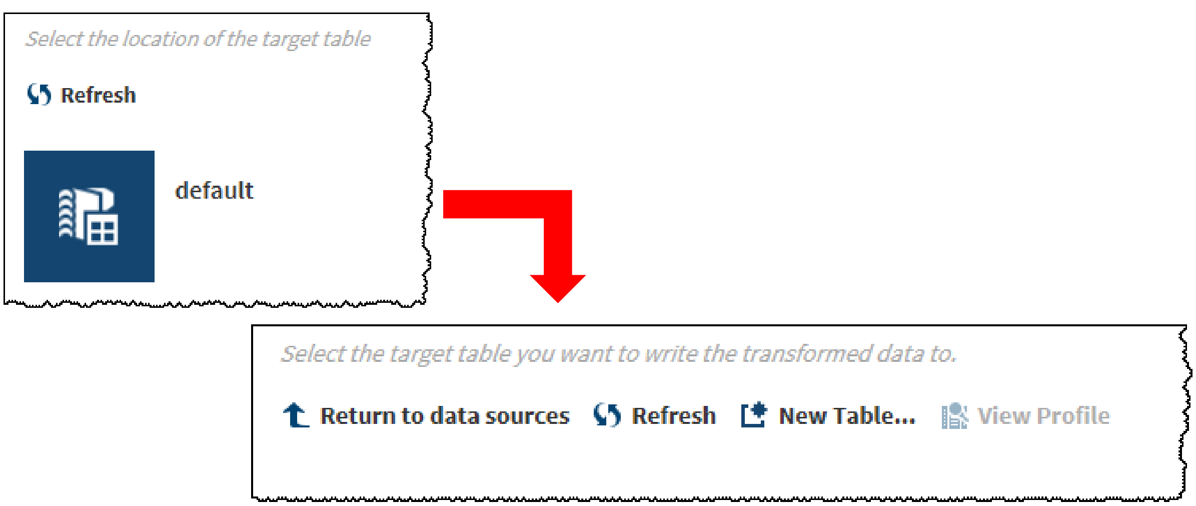
Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



9. Click **Add Column** to add another column to standardize.
10. Click under **Column** and select **customer_city**.
11. Click under **Definition** and select **City**.
12. Verify that **New Column Name** is **customer_city_standardized**.
13. Click **Add Column** to add another column to standardize.
14. Click under **Column** and select **customer_state**.
15. Click under **Definition** and select **State/Province (Abbreviation)**.
16. Verify that **New Column Name** is **customer_state_standardized**.



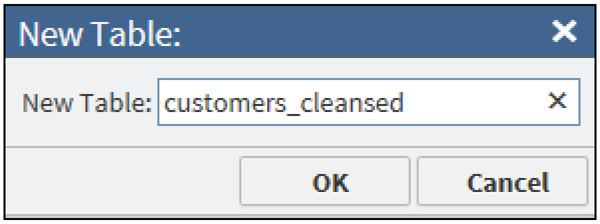
17. Click **Next**.
18. Click the **default** schema (for a target table destination).



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

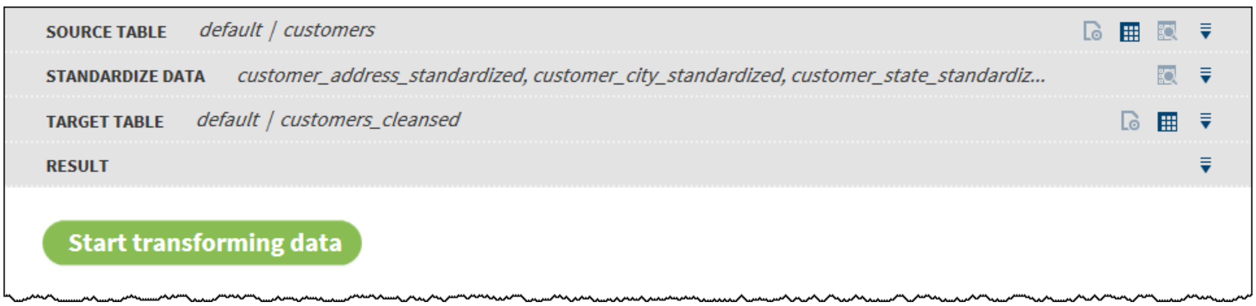
19. Click **New Table**.

    a. Enter **customers_cleansed** in the **New Table** field.



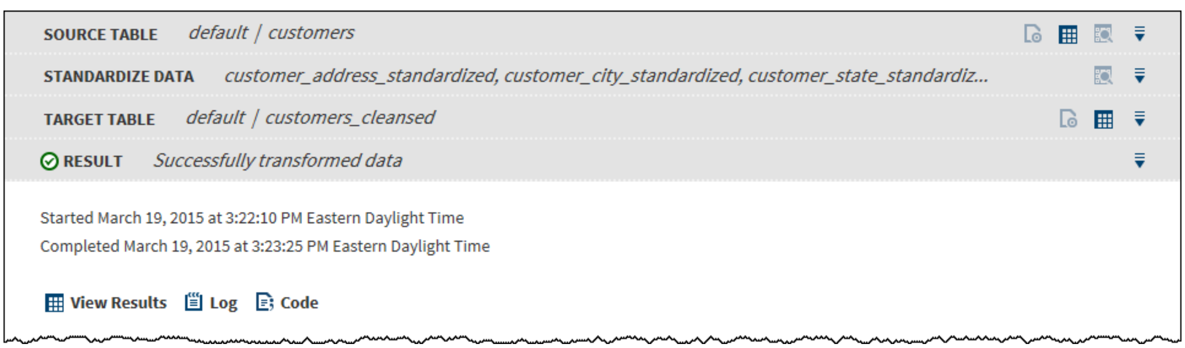
    b. Click **OK**.

20. Click **Next**.



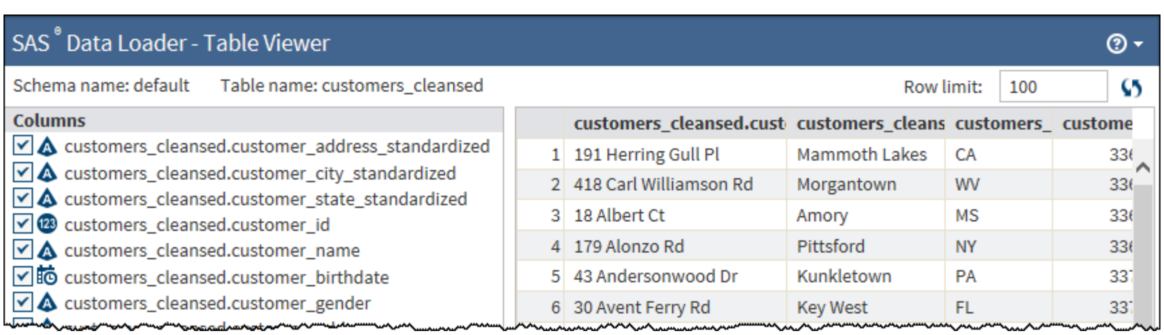
21. Click **Start transforming data**.

This process takes several minutes.



22. Click **View Results**.

The three new columns are listed to the left of the original columns.



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

The three columns that contain the result set of standardization are all now consistently cased, perhaps spelling errors corrected, and, for **customer_cleansed.customer_state_standardized**, we see that all state values seem to have the two-capital-letter pattern.

23. Close the tab in Internet Explorer that displays the profile.
24. If necessary, click the tab in Internet Explorer that is the main SAS Data Loader tab (where the directives are shown).

We will now profile the new cleansed data to see the improvements in the metrics.

25. From the SAS Data Loader directives page, click the directive **Profile Data**.
26. Click the **default** schema.
27. Click the **customers_cleansed** table.
28. Click **Next**.
29. Verify that all columns are selected.
30. Click **Next**.
31. Enter **Customers – PostCleanse** for the report name.
32. Click **Next**.
33. Click **Create Profile Report** to start processing.

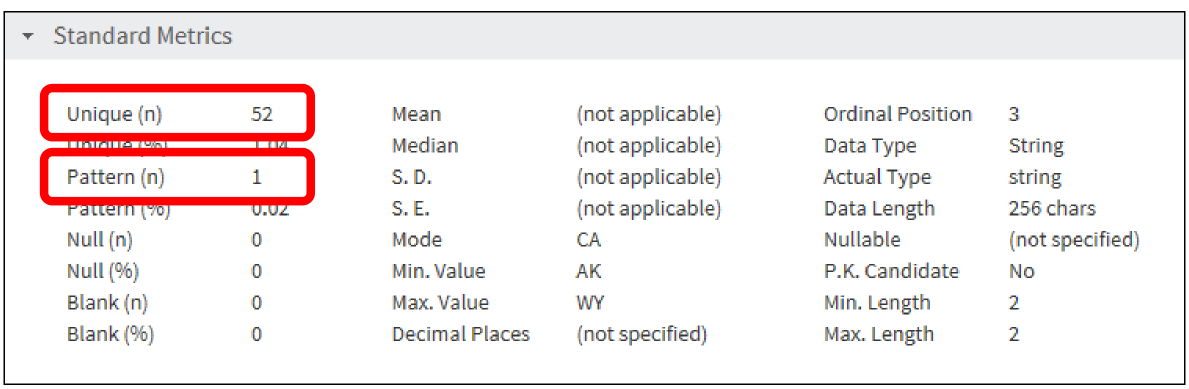
    When the processing is complete, a completion time and a link to view the profile report appears.
34. Click **View Profile Report**.
35. Click (**Show Outline**).
36. Click the **customer_state_standardized** field in the left panel (outline).

    QUESTION: Using the Standard Metrics list of values, what is the unique count [Unique (n)] for the **customer_state_standardized** field?

    ANSWER: 52. It is clear that the standardization of this field has reduced the unique count from 67 to 52. You should still verify that the 52 unique values are valid. Accomplish this by viewing the table in the Frequency Distribution area.

    QUESTION: Using the Standard Metrics list of values, what is the pattern count [Pattern (n)] for the **customer_state_standardized** field?

    ANSWER: 21. It is desired that all state values be two capital letters.



Standard Metrics

Unique (n)	52	Mean	(not applicable)	Ordinal Position	3
Unique (%)	1.04	Median	(not applicable)	Data Type	String
Pattern (n)	1	S. D.	(not applicable)	Actual Type	string
Pattern (%)	0.02	S. E.	(not applicable)	Data Length	256 chars
Null (n)	0	Mode	CA	Nullable	(not specified)
Null (%)	0	Min. Value	AK	P.K. Candidate	No
Blank (n)	0	Max. Value	WY	Min. Length	2
Blank (%)	0	Decimal Places	(not specified)	Max. Length	2

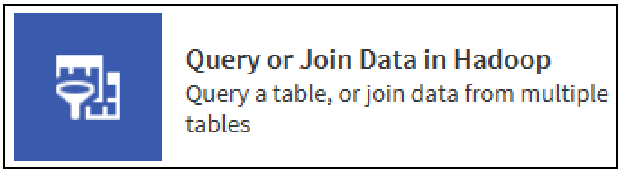
37. Close the tab in Internet Explorer that is viewing the profile.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

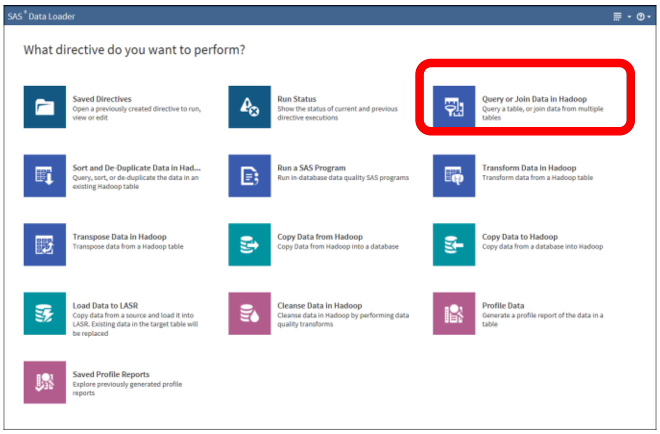
38. If necessary, click the tab in Internet Explorer that is the main SAS Data Loader tab (where the directives are shown).

Directive: Join Data in Hadoop

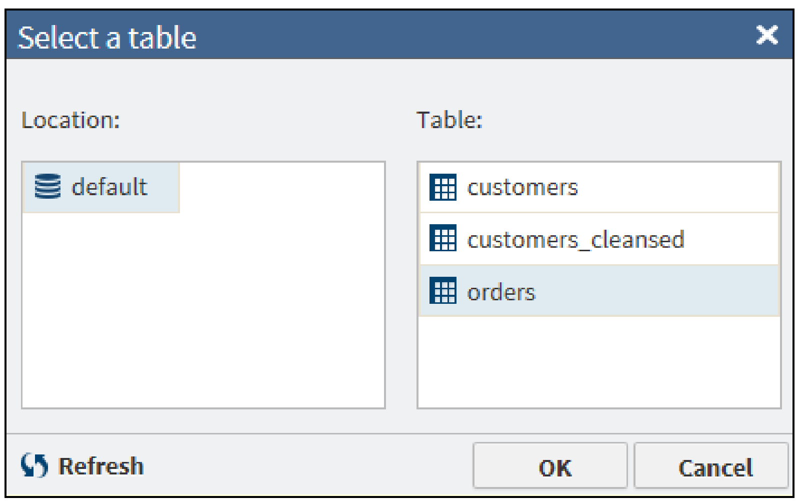
1. From SAS Data Loader directives page, click the directive **Query or Join Data in Hadoop**.



In the list of directives, Query or Join Data in Hadoop is located in this position:



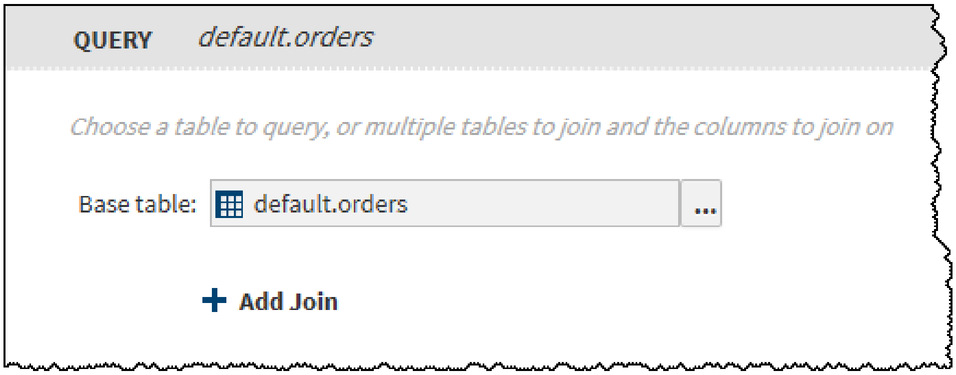
2. Click **...** next to **Base table**. This opens the Select a table window.
3. Click the **default** schema.
4. Click the **orders** table.



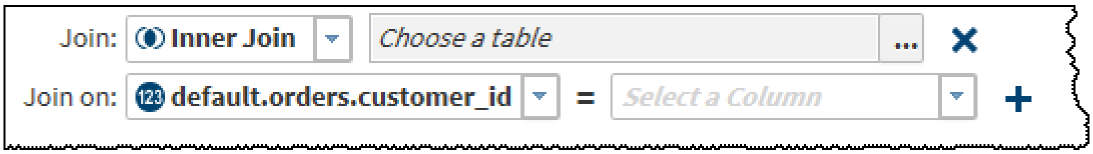
5. Click **OK**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

The result of the selection of the table should resemble the following:



6. Click **Add Join**.



7. Verify that the **Join** field is set to the value **Inner Join**.

8. Click ... next to ***Choose a table***. This open the **Select a table** window.

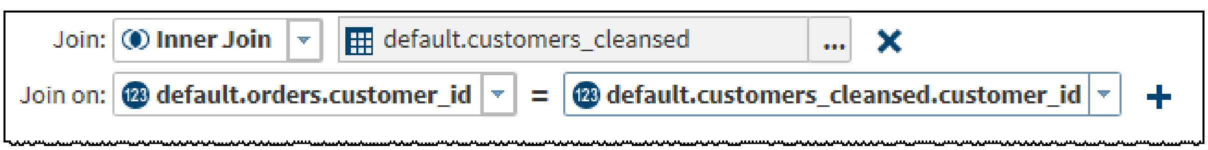
   a. Click the **default** schema.

   b. Click the **orders** table.

   c. Click **OK**.

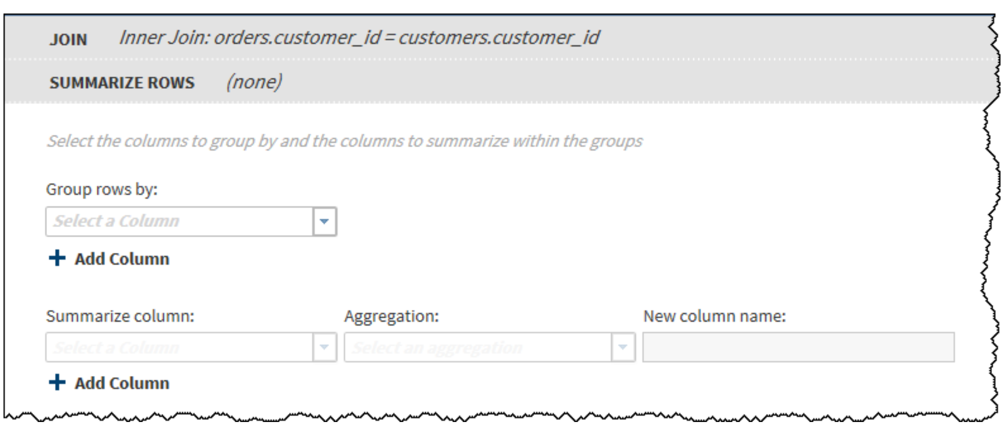
9. Click ▾ next to second field for the **Join on** criteria.

10. Click **default.customers_cleansed.customer_id**.



11. Click **Next**.

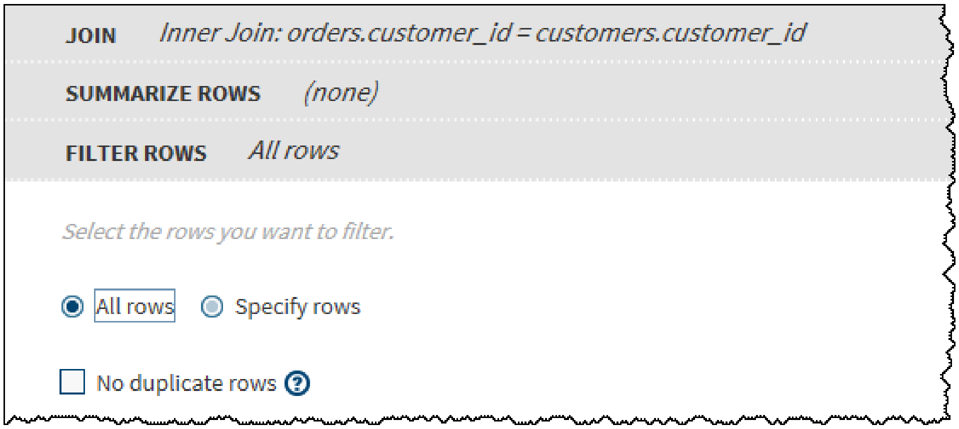
    The next item to consider is summarization. For this example, we will not be summarizing.



Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

12. Click **Next**.

    The next item to consider is filtering. For this example, we will be filtering for state values of *TX*.



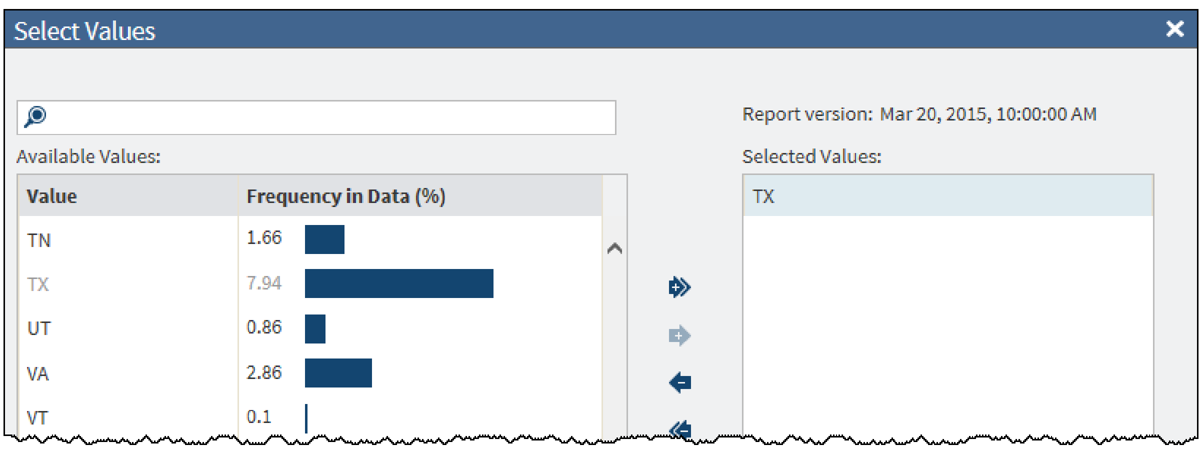
13. Click **Specify rows**.

14. Click under **Column** and click **default.customers_cleansed.customers_state_standardized**.

15. Verify that **Operator** is set to **Equal To**.

16. Click under **Value**.

    a. Locate the value of **TX** in the Available Values list and double-click it to move it to the Selected Values list.

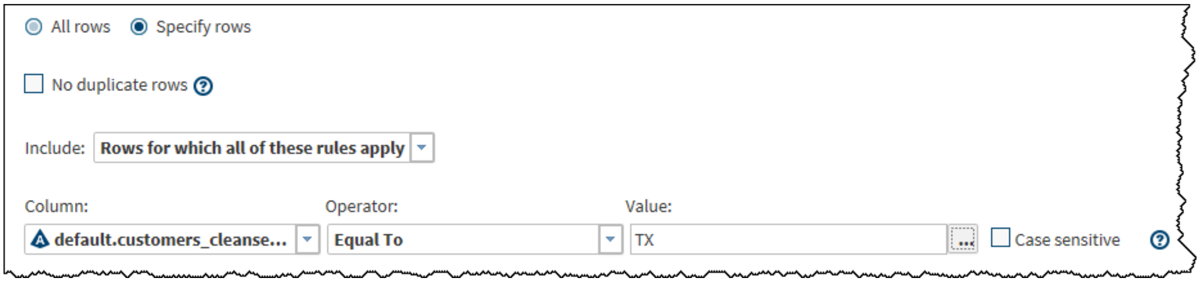


The list of distinct values is being retrieved from the profile report generated on this table.

    b. Click **OK**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

The Filter Rows area should now resemble the following:



17. Click **Next**.
18. Remove the following columns from the Selected columns list. (Click the column and then click).

    default.customers_cleansed.customer_address

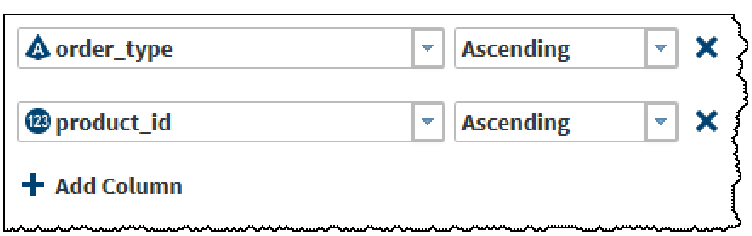
    default.customers_cleansed.customer_city

    default.customers_cleansed.customer_state
19. Rename the following columns in the Selected columns list. (Click the selected column in the **Target Name** field and remove the **_standardized** portion.)

Original Source Name	Renamed Target Name
default.customers_cleansed.customer_address_standardized	customer_address
default.customers_cleansed.customer_city_standardized	customer_city
default.customers_cleansed.customer_state_standardized	customer_state

    The Selected columns should now be a list of 18 total columns.
20. Click **Next**. We now have an opportunity to order the joined data.
21. Click next to **Select a Column** and select **order_type**.
22. Click **Add Column**.
23. Click next to **Select a Column** and select **product_id**.

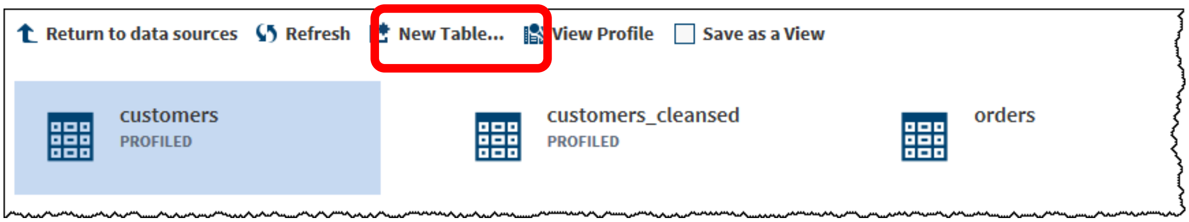
    The final sort information should resemble the following:



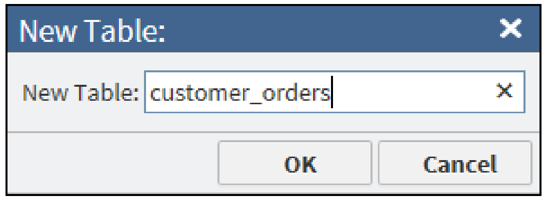
24. Click **Next**. We now need to specify the output table information.
25. Click the **default** schema.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

26. Click **New Table**.

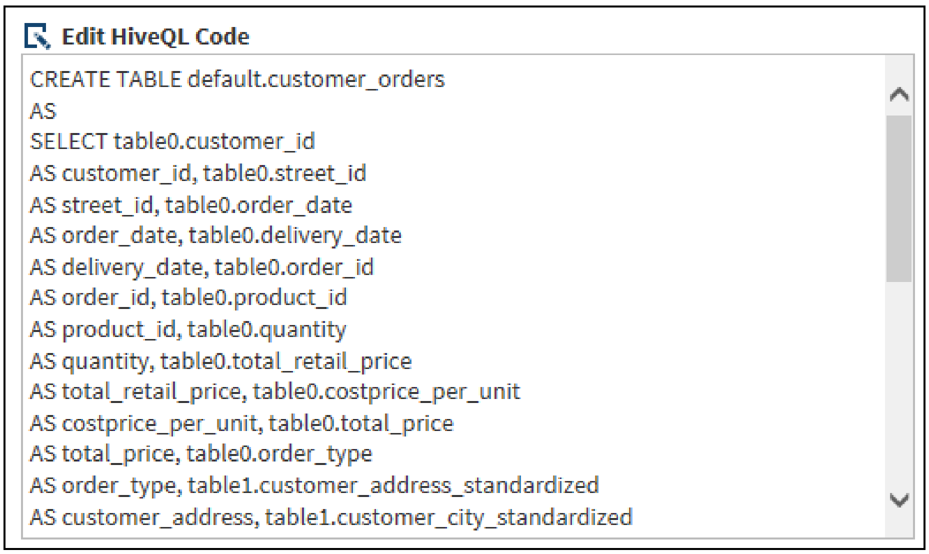


a. Enter **customers_orders**.

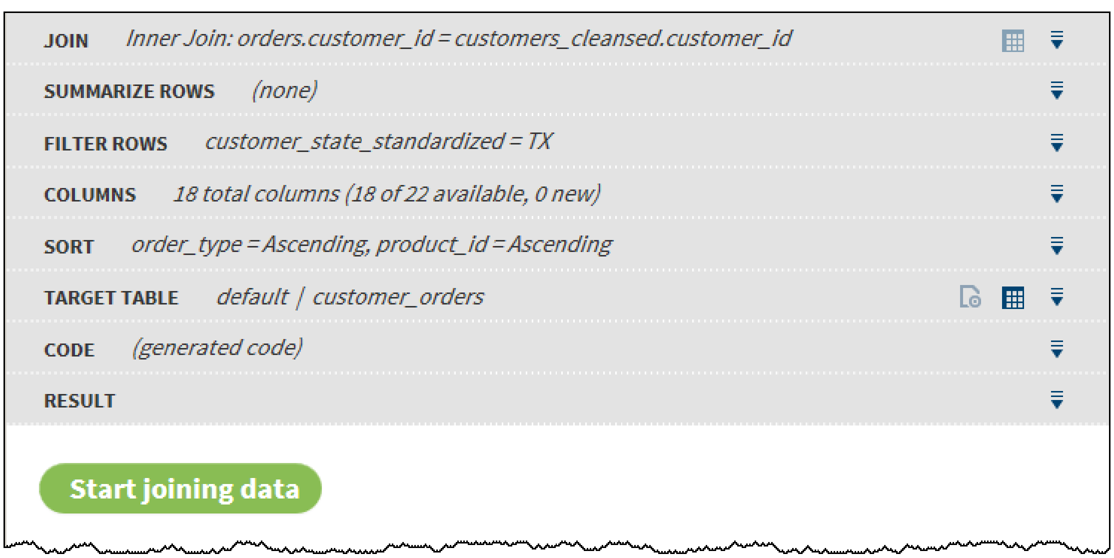


b. Click **OK**.

27. Click **Next**. We now see the generated HiveQL Code.



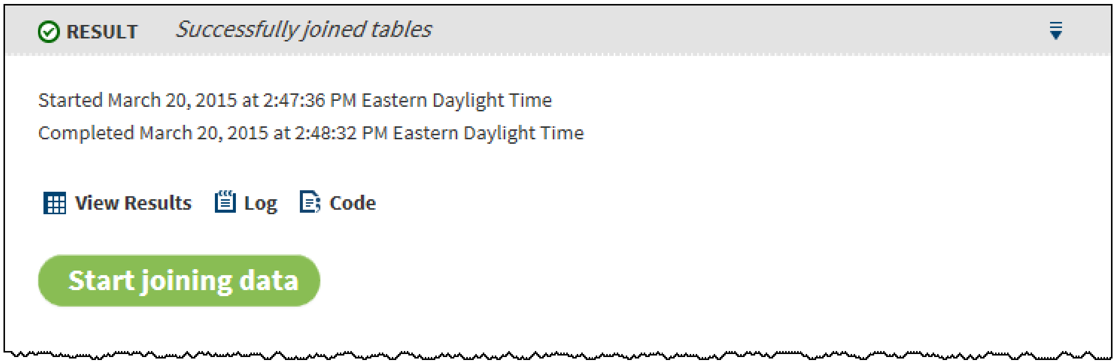
28. Click **Next**. We are now ready to join data.



29. Click **Start joining data**. This might take a few minutes.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

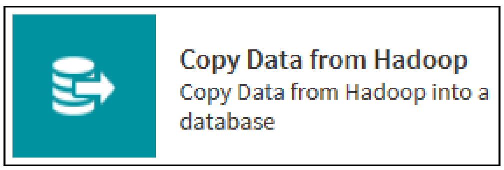
30. Verify that the result was generated successfully.



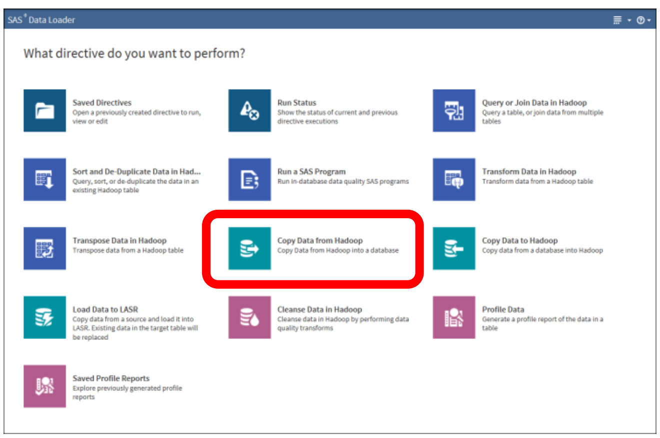
31. Click **Back to Directives** to return to the main directives page.

Directive: Copy Data from Hadoop

1. From SAS Data Loader directives page, click the directive **Copy Data from Hadoop**.



In the list of directives, Copy Data from Hadoop is located in this position:



The first portion of this directive is to define the Source Table.

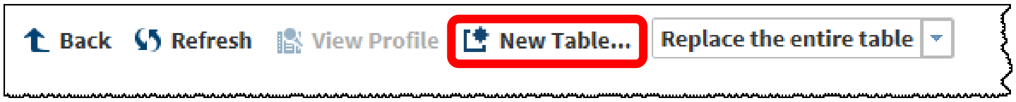
2. Click the **default** schema.
3. Click the **customer_orders** table.
4. Click **Next**.
5. Click **Next**. (We will accept the default number of processes of 1.)

   We now need to define specifics for the target table.

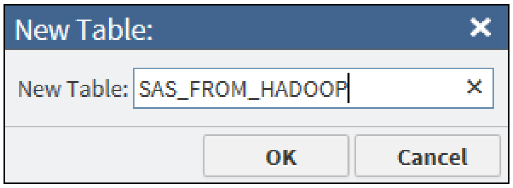
6. Click **SAS Server**.
7. Click **SAS Data Location**.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

8. Click **New Table**.

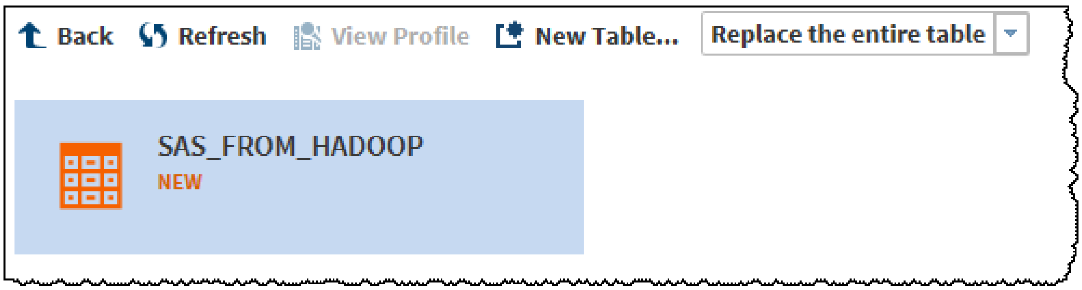


   a. Enter **SAS_FROM_HADOOP** in the **New Table** field.

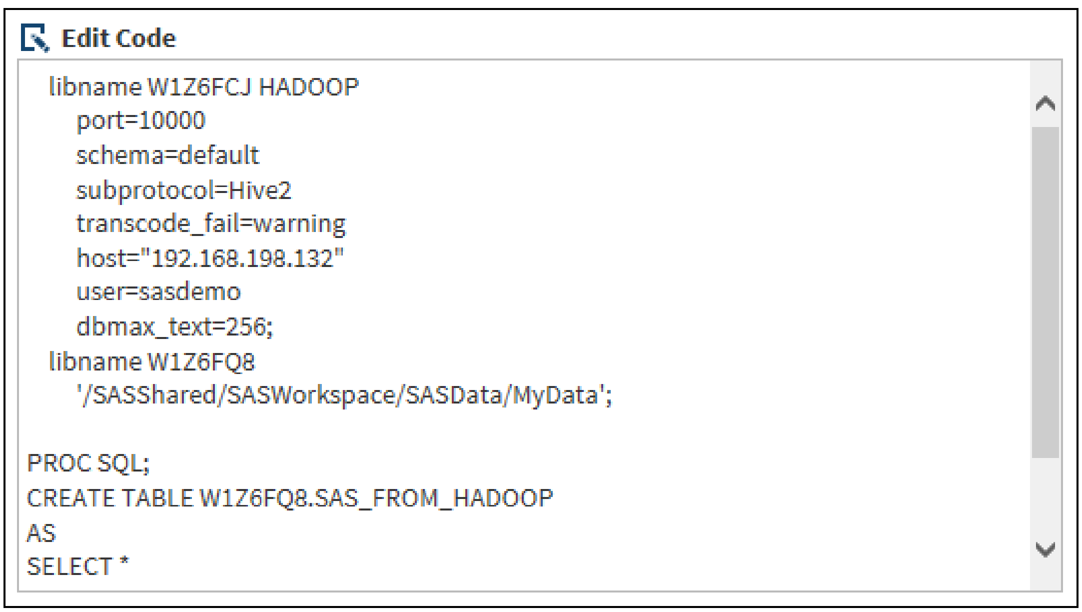


   b. Click **OK**.

   The new table name appears:



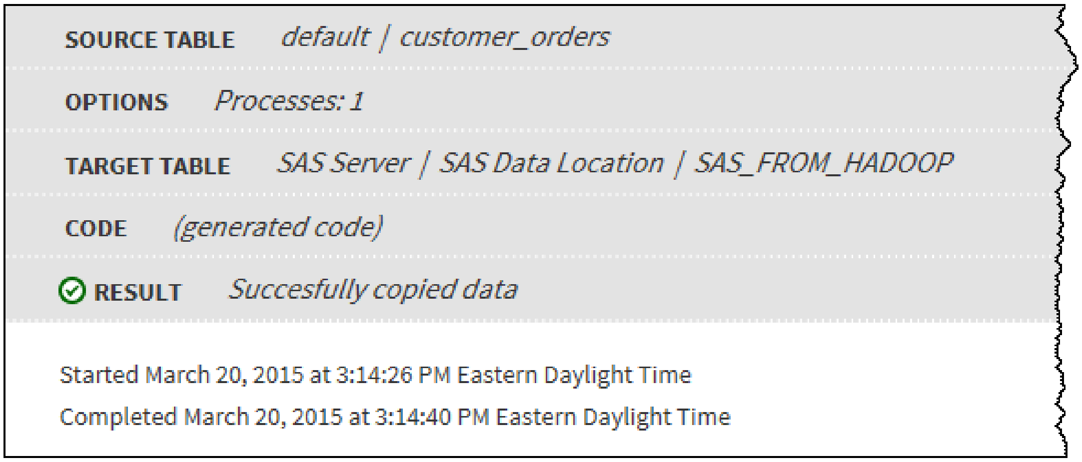
9. Click **Next**. We now see generated SAS code.



10. Click **Next**. We are now ready to move data from Hadoop.
11. Click **Start copying data**. This might take a few minutes.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

12. Verify that the data was successfully copied.



13. Click **Back to Directives** to return to the main directives page.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

Copyright © 2015, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.