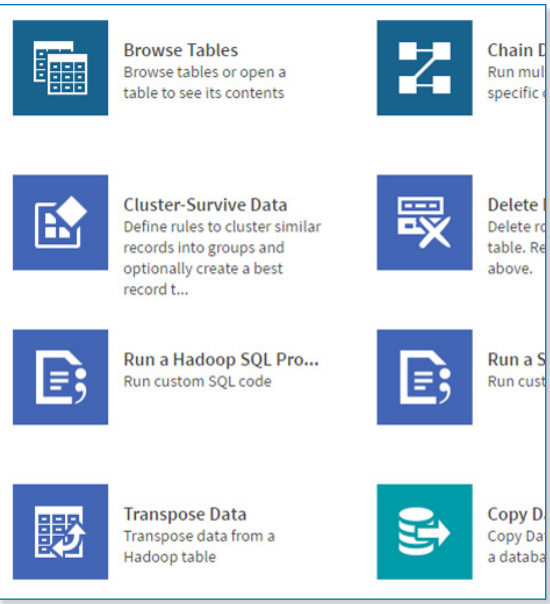
SAS® Data Loader for Hadoop

Take control of your data – and free up IT – with self-service data preparation



Browse Tables
Browse tables or open a table to see its contents

Chain D
Run mul
specific

Cluster-Survive Data
Define rules to cluster similar records into groups and optionally create a best record t...

Delete
Delete r
table. Re
above.

Run a Hadoop SQL Pro...
Run custom SQL code

Run a S
Run cust

Transpose Data
Transpose data from a Hadoop table

Copy D
Copy Da
a databa

What does SAS® Data Loader for Hadoop do?

SAS Data Loader for Hadoop helps you access and manage data on Hadoop through an intuitive user interface, so it's easy to perform self-service data preparation tasks with minimal training. Users who have technical skills can write and run SAS code on Hadoop for improved performance and governance.

Why is SAS® Data Loader for Hadoop important?

As more organizations turn to Hadoop for storing large amounts of data, they're finding it difficult to manage that data since Hadoop often requires specialized coding skills. SAS Data Loader for Hadoop bridges that skills gap, giving users easy access to their data regardless of technical ability.

For whom is SAS® Data Loader for Hadoop designed?

The solution is designed for business users, who can gain value from their big data without needing to write code, as well as SAS coders and data scientists, who use the solution for improved performance and productivity.

Organizations recognize the importance of big data. They know it can be used for analytics and other advanced technologies, so they've harnessed and stored it in systems like Hadoop.

Storing data is one thing – but the ability to manage it is quite another. Accessing data in Hadoop requires code that's difficult to create and maintain, which creates a gap in the skills needed to manage it. And if you can't get to the data you need, you lose the value you should have gained by collecting it. You're left with limited options: depend on someone from IT, learn to code yourself – or find a solution that bridges the gap.

SAS Data Loader for Hadoop has an intuitive interface for profiling, managing, cleansing and moving data in Hadoop – so you can manipulate data without knowing how to code. Likewise, IT will be freed up to focus on more technical benefits, such as boosting processing performance and improving data security.

Benefits

- **Manage data without knowing how to code**. No need to commit to advanced training or hire expensive talent. SAS Data Loader for Hadoop empowers you to perform data integration, data quality and data preparation tasks yourself, without leaning on IT.
- **Harness the power of big data.** Once skill set barriers are broken down, there's endless potential for what you can do with your data – and SAS Data Loader for Hadoop is the driving force. You'll be able to profile, cleanse, join and transform data to create high-quality information that powers advanced analytics.
- **Improve scalability and performance.** While business users focus on using SAS Data Loader for Hadoop to support analytics and decision making, data scientists and SAS coders use it to improve speed, efficiency and agility. The solution's code accelerator harnesses the power of Hadoop for faster performance. Plus, by minimizing data movement, you can increase the security of your data.
- **Strengthen big data security and governance.** You can share and secure saved directives, and use with SAS Lineage and SAS Data Integration Studio to visualize metadata relationships.
- **Operationalize data preparation tasks.** Once you've created a directive or profile, run it as part of a SAS Data Integration Studio job using common metadata that spans the worlds of ad hoc data preparation and discovery along with the more governed world of deployment.

Product Overview

SAS Data Loader for Hadoop is a bundle of SAS products that includes SAS Data Loader, SAS/ACCESS® Interface to Hadoop, SAS In-Database Code Accelerator for Hadoop and SAS Data Quality Accelerator for Hadoop – technologies encompassing data integration and data quality operations.

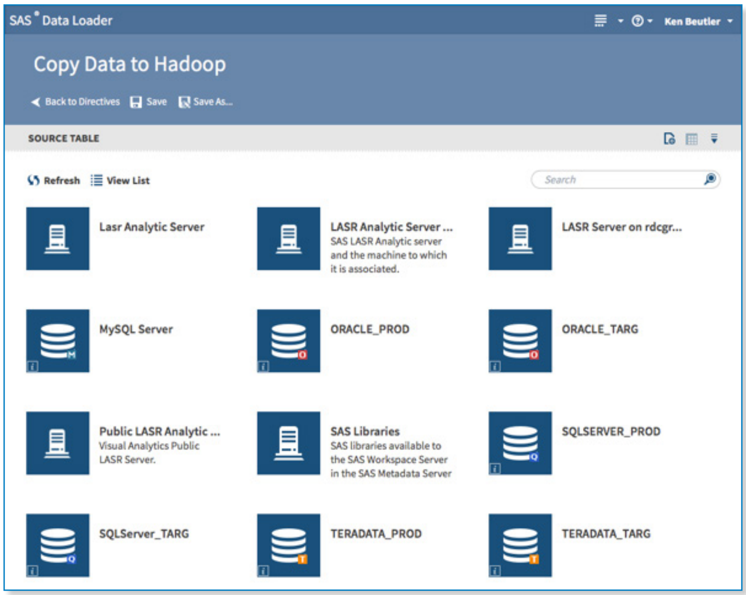
With its combination of user-friendly and highly technical features, SAS Data Loader for Hadoop is a solution that benefits both sides of the organization.

Intuitive user interface

SAS Data Loader for Hadoop is designed for the business user. Its intuitive, wizard-driven interface makes it easy to access and manage data that's stored in Hadoop, reducing the need to engage IT or hire Hadoop-specific talent to get the job done.

Purpose-built to load data to and from Hadoop

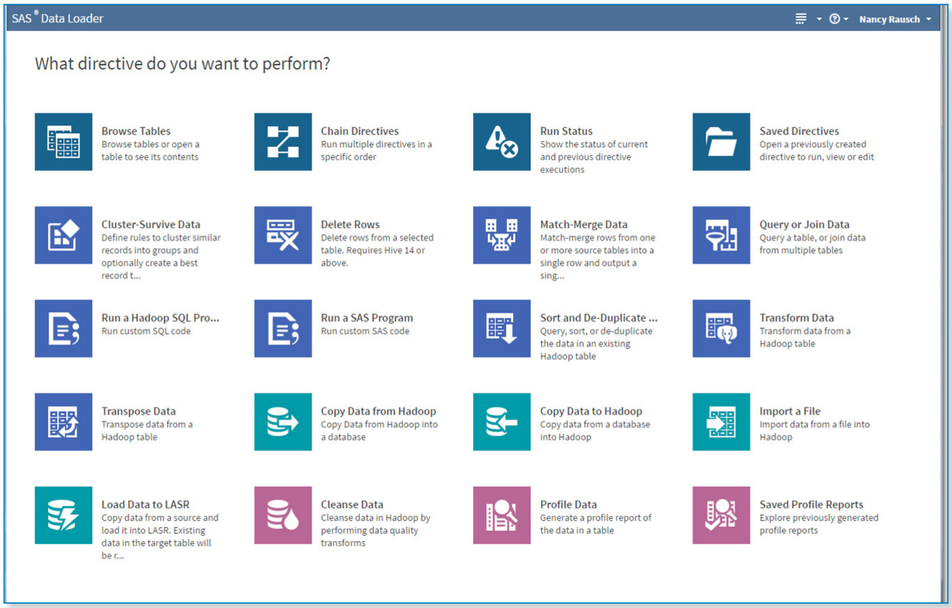
SAS Data Loader for Hadoop was built from the ground up to manage big data on Hadoop, not repurposed from existing IT-focused tools. Use SAS/ACCESS libraries to connect to dozens of cloud, big data and relational data sources including Amazon Redshift, Apache Hadoop, SAP HANA, Oracle, Teradata and IBM DB2.



Copy Data to Hadoop

Lasr Analytic Server	LASR Analytic Server ... SAS LASR Analytic server and the machine to which it is associated.	LASR Server on rdcgr...
MySQL Server	ORACLE_PROD	ORACLE_TARG
Public LASR Analytic ... Visual Analytics Public LASR Server.	SAS Libraries SAS libraries available to the SAS Workspace Server in the SAS Metadata Server	SQLSERVER_PROD
SQLServer_TARG	TERADATA_PROD	TERADATA_TARG

Figure 1: Copy cloud, in-memory, text files, SAS data sets, relational and big data sources to and from Hadoop.



What directive do you want to perform?

Browse Tables Browse tables or open a table to see its contents	Chain Directives Run multiple directives in a specific order	Run Status Show the status of current and previous directive executions	Saved Directives Open a previously created directive to run, view or edit
Cluster-Survive Data Define rules to cluster similar records into groups and optionally create a best record t...	Delete Rows Delete rows from a selected table. Requires Hive 14 or above.	Match-Merge Data Match-merge rows from one or more source tables into a single row and output a sing...	Query or Join Data Query a table, or join data from multiple tables
Run a Hadoop SQL Pro... Run custom SQL code	Run a SAS Program Run custom SAS code	Sort and De-Duplicate ... Query, sort, or de-duplicate the data in an existing Hadoop table	Transform Data Transform data from a Hadoop table
Transpose Data Transpose data from a Hadoop table	Copy Data from Hadoop Copy Data from Hadoop into a database	Copy Data to Hadoop Copy data from a database into Hadoop	Import a File Import data from a file into Hadoop
Load Data to LASR Copy data from a source and load it into LASR. Existing data in the target table will be r...	Cleanse Data Cleanse data in Hadoop by performing data quality transforms	Profile Data Generate a profile report of the data in a table	Saved Profile Reports Explore previously generated profile reports

Figure 2: SAS Data Loader for Hadoop helps business users copy, view, prepare, cleanse and transform big data through a series of directives – without learning how to write code.

Big data quality

Take control of the data within your Hadoop environment. SAS Data Loader for Hadoop allows you to profile data to understand its overall quality. Then, you can standardize, parse, match and perform other core data quality functions inside Hadoop by using the SAS Embedded Process, a lightweight SAS execution engine.

Additional data quality directives include casing, gender analysis, pattern analysis and field extraction. This allows users to apply case changes to the data, guess the gender based on values to improve customer segmentation, and guess acceptable data patterns based on field values. Field extraction pulls useful tokens from unstructured or freeform text within a field – such as name, organization, address, email and phone number.

Profiling runs in parallel on the Hadoop cluster for improved performance, and adds trend graphs to track data changes over time. Identification analysis determines the type of data that's present in a column. For example, NC, North Carolina and N Carolina within a column's data would be categorized as "state" to aid in data exploration.

In-memory analytic server

There's no need to wait on IT to pull the data you need for reporting, visualization or analytics. SAS Data Loader for Hadoop allows business users to load data into the SAS® LASR™ Analytic Server in memory for visualization in SAS Visual Analytics.

Security and governance

SAS Data Loader for Hadoop provides secure access to Kerberos-enabled Hadoop clusters. You can also integrate with Active Directory or LDAP for user authorization. Users have role-based access to saved directives using SAS folders. You can also track metadata relationships by using with SAS Lineage and SAS Data Integration Studio.

In-cluster SAS® code execution

With SAS Data Loader for Hadoop, you can execute analytics processing within the Hadoop ecosystem, getting faster results at a lower cost than with more traditional solutions. You'll benefit from increased scalability and performance due to reduced data movement and parallel processing.

Better collaboration

Share directives using SAS folders and allow multiple user logins using SAS authentication.

Hadoop distributions

SAS Data Loader for Hadoop supports the following Hadoop distributions:

- Cloudera 5.8 or later.
- Hortonworks 2.5 or later.
- MapR 5.1 or later.
- IBM BigInsights 4.2 or later.
- Pivotal HD3.0.
- Kerberos will be supported for Cloudera, Hortonworks, IBM BigInsights and Pivotal HD.

Key Features

Transform and transpose data on Hadoop

- Copy relational databases and SAS data sets to and from Hadoop via parallel bulk data movement.
- Import data from CSV and other delimited files into Hadoop, and delete rows on Hadoop tables.
- Transform data by filtering rows, managing columns and summarizing rows.
- Transpose and group selected columns.
- Access dozens of cloud, big data and relational data sources through the use of SAS/ACCESS libraries including Amazon Redshift, Apache Hadoop, SAP HANA, Oracle, IBM DB2 and Teradata.

Secure and governed big data access

- Provides secure access to Kerberos-enabled Hadoop clusters.
- Supports Active Directory and LDAP-based user authentication.
- Lets you share and secure saved directives using SAS folders.
- Enables you to track metadata relationships when used in conjunction with SAS Lineage and SAS Data Integration Studio.

Cleanse data in Hadoop

- Standardize, de-duplicate, match and parse your data on Hadoop.
- Intelligent filtering allows import of values from Profile into Filter and Transform directives.
- Query, sort or de-duplicate the data in an existing Hadoop table.
- Speed data exploration by determining the type of data within a column based on its values.
- Using other data quality functions, you can apply casing, determine gender, conduct pattern analysis and extract tokens from unstructured text fields.

Query or join data in Hadoop

- Query a table or join multiple tables without knowing SQL.
- Run aggregations on selected columns and filter source data.
- Power users can generate and edit a HiveQL query, or paste an existing HiveQL query.

Speed data management processes with Spark

- Data quality functions run in memory on Spark for improved performance.
- Spark matching and best record creation enables master data management for big data.
- Read and write to Spark data sets as needed.

Raise your data professionals' productivity to a new level

- Speed creation of Impala queries – a faster way to access data on Hadoop.
- Chain multiple directives together and run as a group.
- Permit external job scheduling with an exposed public API.
- Call SAS Data Loader directives and view profiles from SAS Data Integration Studio.

Manage your data where it lives

- Hadoop support is included for Pivotal HD and IBM BigInsights as well as Hortonworks, Cloudera and MapR.
- Match and merge directive allows in-database merging of multiple data sources.
- Improve performance by pushing processing down to the Hadoop cluster.

TO LEARN MORE »

To learn more about SAS Data Loader for Hadoop, download white papers, view screenshots and see other related material, please visit sas.com/dataloader.

Key Features (continued)

Profile data and save profile reports

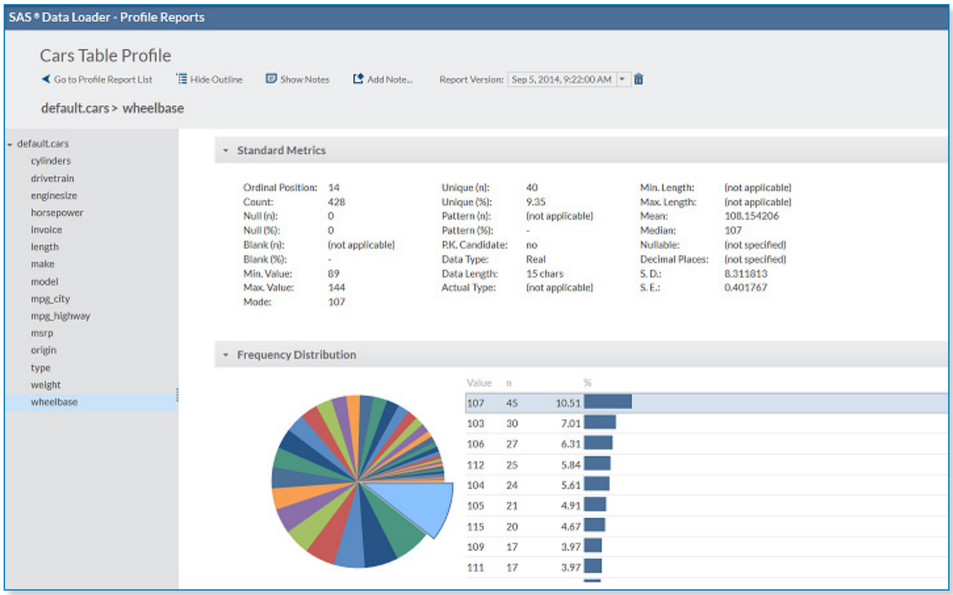
- Select source columns from one or more tables to determine uniqueness, incompleteness and patterns.
- List and open reports generated by the profile data directive.
- Create and save notes.
- Run profiling in parallel on the Hadoop cluster for improved performance.

Manage and reuse directives via wizard-driven user interface

- View list and status of directives and job logs.
- Stop and start directives, and open their log and generated code files.
- Run, view or edit saved directives for reuse.

Take advantage of your existing investment in SAS®

- Load specified Hadoop columns in memory onto the SAS LASR Analytic Server for analysis using SAS Visual Analytics or SAS Visual Statistics (licensed separately).
- Run SAS programs that use the DS2 language on Hadoop using the SAS Embedded Process, a lightweight SAS execution engine.



Report Version: Sep 5, 2014, 9:22:00 AM

default.cars > wheelbase

Standard Metrics

Ordinal Position:	14	Unique (n):	40	Min. Length:	(not applicable)
Count:	428	Unique (%):	9.35	Max. Length:	(not applicable)
Null (n):	0	Pattern (n):	(not applicable)	Mean:	108.154206
Null (%):	0	Pattern (%):	-	Median:	107
Blank (n):	(not applicable)	P.K. Candidate:	no	Nullable:	(not specified)
Blank (%):	-	Data Type:	Real	Decimal Places:	(not specified)
Min. Value:	89	Data Length:	15 chars	S. D.:	8.311813
Max. Value:	144	Actual Type:	(not applicable)	S. E.:	0.401767
Mode:	107				

Frequency Distribution

Value	n	%
107	45	10.51
103	30	7.01
106	27	6.31
112	25	5.84
104	24	5.61
105	21	4.91
115	20	4.67
109	17	3.97
111	17	3.97

Figure 3: Profiling processing is pushed down to the Hadoop cluster for improved performance.

To contact your local SAS office, please visit: sas.com/offices

THE POWER TO KNOW.