Paper SAS5602-2016

Sparking Analytical Insight with SAS® Data Loader for Hadoop

Matthew Magne, SAS Institute Inc.

ABSTRACT

Sixty percent of organizations will have Hadoop in production by 2016, per a recent TDWI survey. It's become increasingly challenging to access, cleanse, and move these huge volumes of data. How can data scientists and business analysts clean up their data, manage it where it lives, and overcome the big data skills gap? It all comes down to accelerating the data preparation process. SAS® Data Loader leverages years of expertise in data quality and data integration, a simplified user interface, and the parallel processing power of Hadoop to overcome the skills gap and compress the time taken to prepare data for analytics or visualization. We cover some of the new capabilities of SAS® Data Loader for Hadoop, including in-memory Apache Spark processing of data quality and master data management functions, faster profiling and unstructured data field extraction, and chaining multiple transforms for improved productivity.

INTRODUCTION

Organizations today consist of increasingly heterogeneous architectures that blend traditional relational database systems with emerging big data analytics technologies like Apache Hadoop or Apache Spark. Different skills are required to manage data on Hadoop and take advantage of the parallel processing paradigm. Coding in Java using a batch-processing framework called MapReduce is one example. This slows the adoption and reuse of existing data management skills in this new environment. A recent TDWI survey estimates that 60% of organizations will have Hadoop in production by 2016, and Forrester is predicting total market saturation within two years.

At the same time, there is a growth in "self-service data preparation"—empowering business users to access, cleanse, and transform data from a designated area without burdening IT. This offloads some of the data provisioning and blending work previously performed by IT and improves the productivity of both parties in the process. Businesses can get faster access to the data they need for decisions and do more of the iterative transformations and aggregations that they would normally rely on IT to perform.

WHAT IS SAS DATA LOADER FOR HADOOP?

SAS Data Loader for Hadoop is a self-service, big data preparation solution that was created to mitigate the gap in big data skills and simplify profiling, cleansing, and transformation of data on Hadoop. It has a browser-based user interface that reduces the need for training by use of directives. Directives are discrete data processing activities like copying a SAS® data set to Hadoop, standardizing a state column, or lifting data into memory for analysis. These directives provide a guided path and require no coding to profile, blend, or move data on Hadoop. As a result, the time required to prepare data for analytics or visualization is compressed, and adoption of big data is accelerated.

While easy to use, SAS Data Loader also takes advantage of the parallel processing power of Hadoop to push down processing to the Hadoop cluster. Data scientists and SAS coders can use the included SAS® Data Quality Accelerator for Hadoop and SAS® In-Database Code Accelerator for Hadoop, which are installed on each node. These accelerators execute SAS DS2 code and data quality functions inside the nodes of the Hadoop cluster for improved performance, governance, and productivity.

In this paper, we'll cover some of the new capabilities of SAS Data Loader for Hadoop that speed data management processes with Spark, improve the productivity of data professionals, and enable organizations to manage data where it lives.

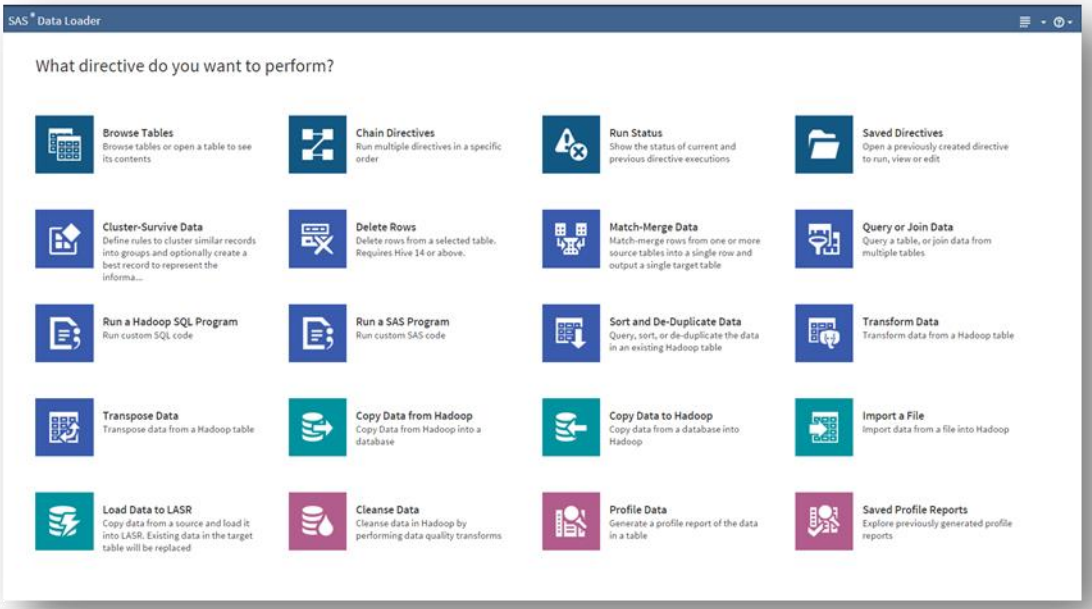

Figure 1. SAS Data Loader for Hadoop User Interface Provides Data Preparation on Hadoop via Directives

HADOOP BACKGROUND – WHAT IS HADOOP?

Apache Hadoop is a big data technology developed at Yahoo and then open sourced to the Apache Software Foundation (ASF) about eight years ago. It distributes processing to a group of commodity servers called nodes and has disrupted the cost of storing and processing huge volumes of data. Numerous SAS partners produce commercial distributions of Hadoop including Cloudera, Hortonworks, and MapR, to minimize organizational risk, accelerate development, and increase adoption. Many organizations have developed a coexistence strategy between their existing Enterprise Data Warehouses (EDW) and their Hadoop cluster that blends benefits of both traditional relational database systems and emerging big data technologies like Hadoop. SAS has embraced Hadoop and provides methods to access, transform, cleanse, and run high-performance analytics inside Hadoop. For more information about Hadoop, please visit the Hadoop tab of the SAS Big Data Insights page.

SPEED DATA MANAGEMENT PROCESSES WITH SPARK

One of the most exciting new developments in SAS Data Loader for Hadoop is the ability to take advantage of the Apache Spark, in-memory processing engine. The Cleanse Data, Transform Data, and Cluster-Survive Data directives all use the Apache Spark engine. The EEL (Expression Engine Language) functions found in the SAS Data Quality solutions are supported when using Spark.

WHAT IS SPARK?

Apache Spark is an open-source technology maintained by the Apache Software Foundation and designed to run data processes in-memory for improved performance. Spark takes advantage of memory on the Hadoop cluster to run processes up to 100x faster than MapReduce, the conventional distributed batch execution engine for jobs on Hadoop. Spark uses Resilient Distributed Data Sets, or RDDs, as the persistence mechanism for storing and processing data in-memory. Hadoop is not needed to run Spark, but the manner in which SAS Data Loader for Hadoop uses Spark requires related Hadoop technologies including Oozie, a workflow scheduler, and HDFS, a distributed big data file system, to store the output of the Spark job.

HOW DOES DATA LOADER TAKE ADVANTAGE OF SPARK?

The SAS® Data Management Accelerator for Spark is a component of SAS Data Loader for Hadoop that uses Spark to run data quality functions in the Cleanse Data, Transform Data, and Cluster-Survive directive in-memory.

SAS Data Management Accelerator for Spark is installed on every Spark client node in a process similar to that used to install the SAS® Embedded Process. (See References below.) It provides a mechanism to translate the Data Loader directive and associated transformations into RDDs. The data and transformations are then executed in-memory and saved back to a Hive table for viewing within Data Loader or chaining with other directives.

The preferred run-time target of the Cleanse, Transform, and Cluster-Survive directives can be changed from “MapReduce” to “Hadoop Spark” for a specific directive or globally using the configuration menu.

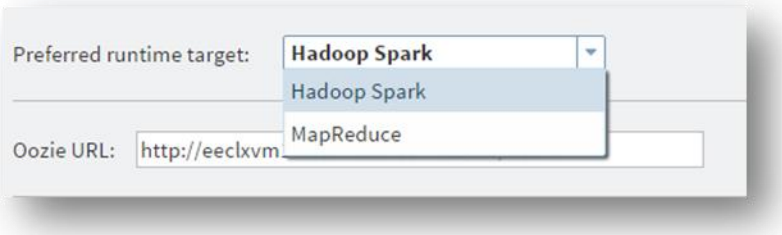

Figure 2. Preferred Run-time Target Can Now Use Hadoop Spark in Addition to MapReduce

Cleanse Data Directive

The Cleanse Data directive was created using SAS' years of expertise in bringing data quality and identity resolution solutions to market. An example of some of the data quality functions that Spark can run in-memory include guessing gender based on name (gender analysis), field extraction (extracting data is an entity of type Organization, Address, or Email), parsing one column into multiple columns, changing case, and identifying that data in a column is of a certain type based on data values (identification analysis).

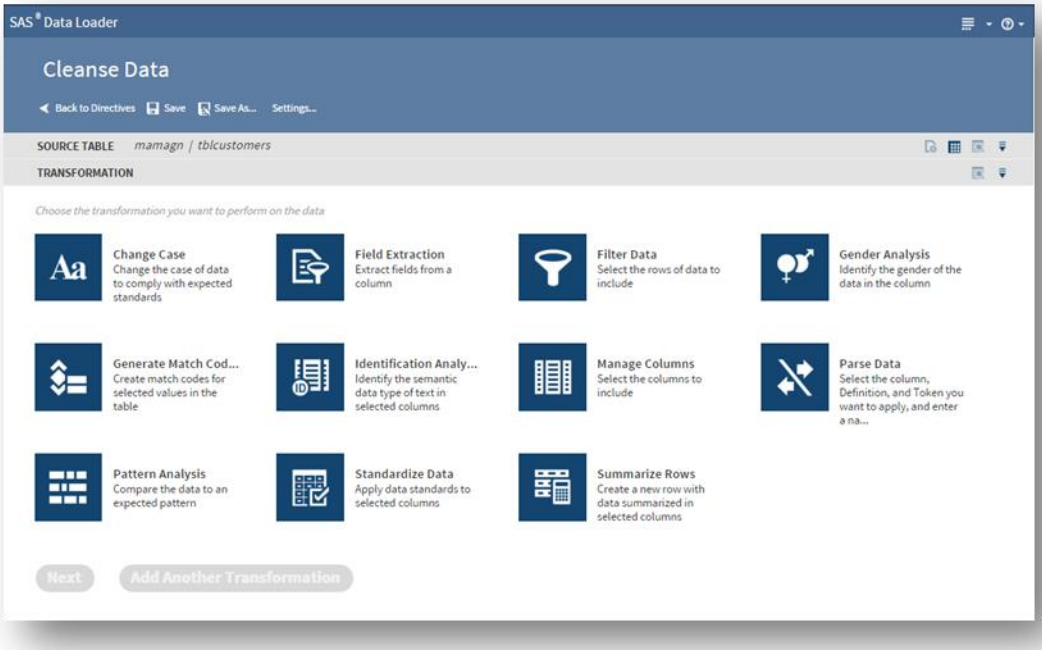

Figure 3. The Cleanse Data Directive Provides Ability to Run Numerous Data Quality Functions

CLUSTER-SURVIVE DIRECTIVE

The cluster-survive directive requires Spark and allows some of the clustering and survivorship capabilities conventionally found in SAS® Master Data Management solutions to be run on extremely large sets of data on Hadoop. It leverages years of experience in the identity resolution matching technology and pre-built rules found in the SAS® Data Quality solution. This “MDM for big data” capability is delivered as a simple wizard-driven directive that groups (clusters) similar records and then collapses or survives them into one "best record" based on a set of business rules.

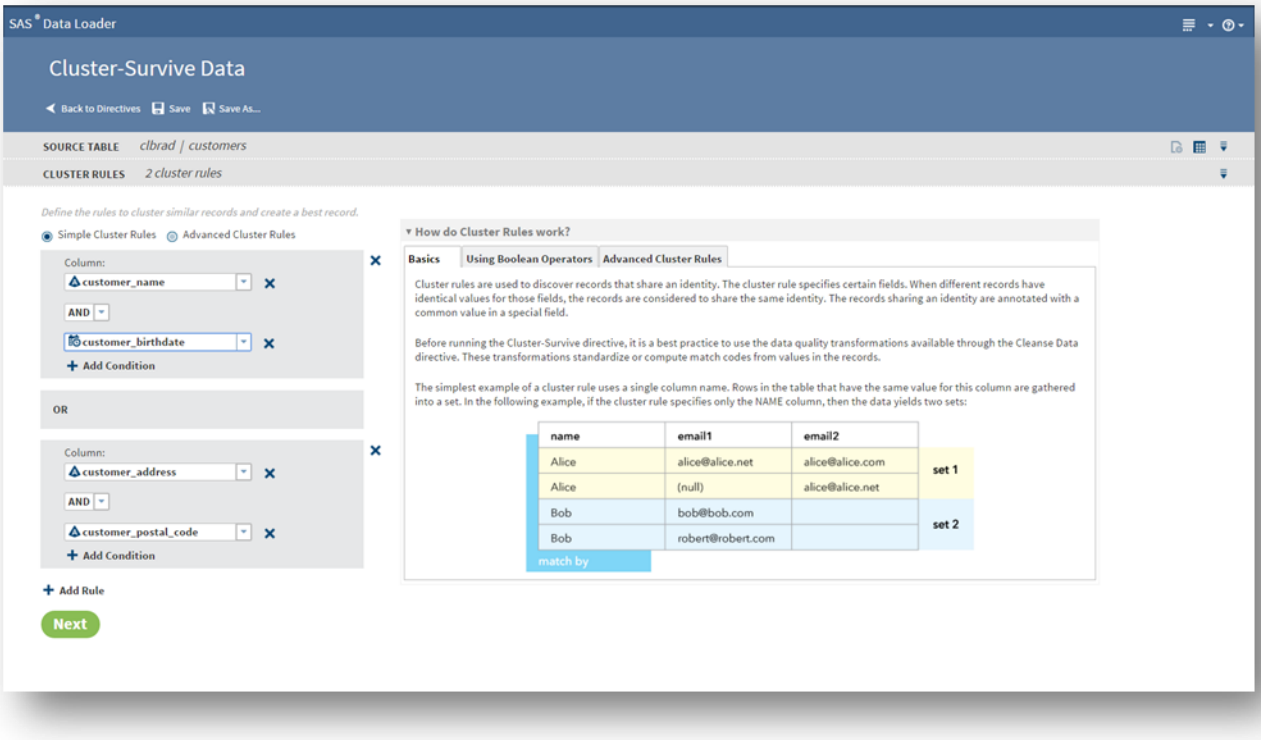

Figure 4. The Cluster-Survive Data Directive Uses Clustering and Survivorship Rules to Create a “Best Record”

IMPROVE PRODUCTIVITY OF DATA PROFESSIONALS

SAS Data Loader for Hadoop improves the productivity of data professionals by running processes faster using Spark (described above), allowing queries to be designed using Impala functions, chaining multiple directives together, and exposing directives for scheduling and execution via a RESTful API. Business users don’t have to learn multiple languages like MapReduce or HiveQL to take advantage of these capabilities. Power users can edit the code generated before running it on Hadoop.

RUN IMPALA SQL

Impala was open-sourced by Cloudera and enables data professionals to reuse their SQL skills and run them on Hadoop up to 60 times faster than MapReduce. The Query or Join Data, Sort and De-duplicate Data, and Run a Hadoop SQL Program directives all support Impala SQL.

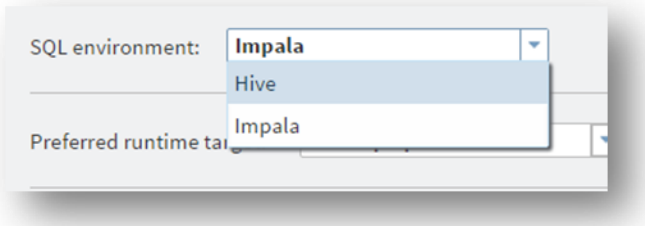

Figure 5. The SQL Environment Can Be Set to Use Impala or Hive

CHAINING DIRECTIVES AND USING THE REST API

We can also chain or group multiple directives together and run them in parallel or sequentially on Hadoop. For example, you might want to copy an Orders and Customers table from an Oracle database or a SAS data set into Hadoop; clean up state codes on the customer data; merge the data together; and then lift that data into SAS® LASR™ Analytic Server for visualization or analysis using SAS® Visual Analytics. This entire set of five or six directives can be executed together. With the new exposed REST API, these directives can be run by an external job scheduler overnight to further automate that process and save your analysts and data scientists time.

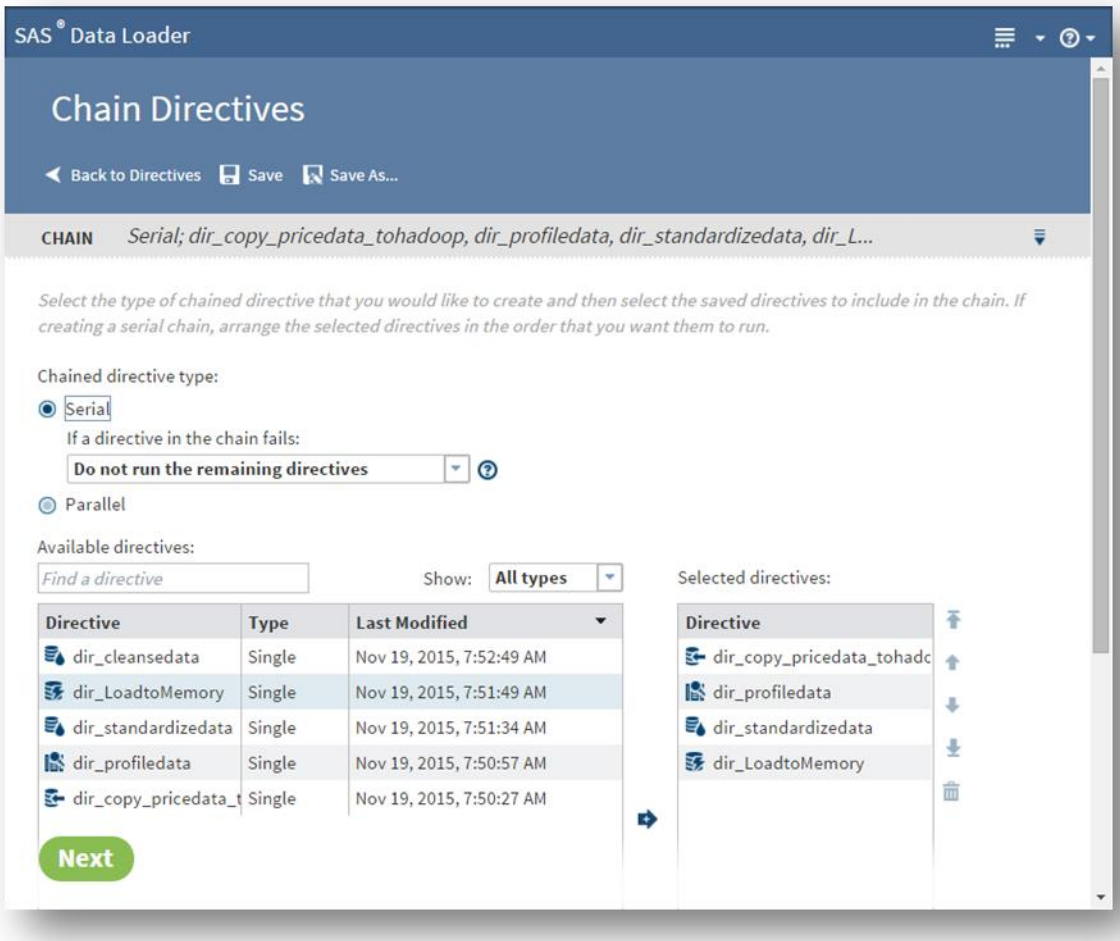

Figure 6. Directives Can Be Chained Together to Run in Both Serial and Parallel

MANAGE DATA WHERE IT LIVES

SAS Data Loader for Hadoop enables data professionals to manage data in and push processing to multiple execution environments and Hadoop distributions. This improves performance, governance, and security because less data is moved from the data source to the compute engine. SAS Data Loader for Hadoop runs the same functions and reuses the same skill sets and technologies across multiple environments via a portable execution engine called the SAS Embedded Process. Think of this as the secret sauce that enables us to run the same code and data quality functions in-memory, in-stream, in-database, or inside Apache Hadoop.

MATCH-MERGE DATA DIRECTIVE

The merge directive can execute additional logic beyond a standard SQL join and pushes down processing of almost 300 functions to Hadoop for improved performance. The Match-Merge Data directive uses SAS DS2 and the SAS Embedded Process to merge data inside Hadoop.

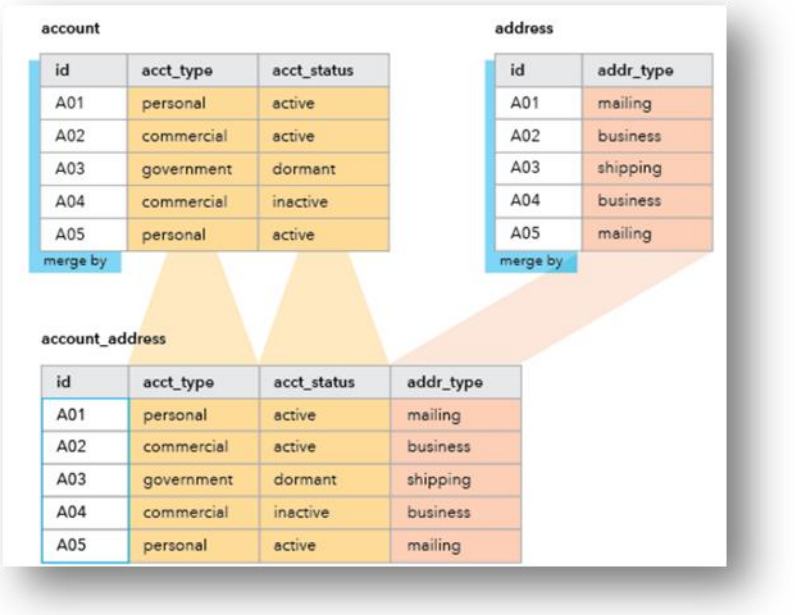

Figure 7. Match-Merge Directive Provides In-Cluster Data Matching and Merging Capabilities beyond Traditional SQL Joins

EXPANDED HADOOP DISTRIBUTION SUPPORT AND FREE TRIAL

There is expanded distribution support for Pivotal HD and IBM Big Insights in addition to existing support for Cloudera, Hortonworks, and MapR. Finally, a free trial is available at http://sas.com/dataloader that runs on production Hortonworks and Cloudera Hadoop clusters and converts to a production license without the need to reinstall.

CONCLUSION

SAS Data Loader for Hadoop provides self-service data preparation capabilities and in-cluster processing capabilities that speed performance, improve productivity, and improve governance. In its most recent version, users can now run data quality and master data management-type functions in-memory on Spark, design and execute queries using Impala SQL, merge data inside Hadoop, chain directives, call functionality via a RESTful API, and run on additional Hadoop distributions. SAS Data Loader for Hadoop can improve the performance of the data preparation process, the productivity of both business users and data scientists, and provide trusted data to drive better analytics.

REFERENCES

Gualtieri, Mike, and Noel Yuhanna. January 19, 2016. “The Forrester Wave™: Big Data Hadoop Distributions, Q1 2016.” Available at https://www.cloudera.com/content/dam/www/static/documents/analyst-reports/forrester-wave-big-data-hadoop-distributions.pdf.

ACKNOWLEDGMENTS

Thanks to Roger Barney and James Richardson for explaining the inner workings of SAS Data Loader for Hadoop.

RECOMMENDED READING

- Bailey, Jeff. 2016. “An Insider's Guide to SAS/ACCESS® Interface to Hadoop.” SAS3880, *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc.
- Ghazaleh, David. 2016. “Exploring SAS® Embedded Process Technologies on Hadoop.” SAS5060, *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc.
- Rausch, Nancy, et al. 2015. “What’s New in SAS Data Management?.” SAS1390, *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings15/SAS1390-2015.pdf.
- Rausch, Nancy, and Wilbram Hazejager. 2016. “Ten Tips to Unlock the Power of Hadoop with SAS®.” SAS2560, *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc.
- Renison, Keith. 2015. “Introduction to SAS® Data Loader: The Power of Data Transformation in Hadoop.” SAS1845, *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings15/SAS1845-2015.pdf.
- SAS Institute Inc. SAS Institute white paper. “Self-Service Data Preparation in the Age of Hadoop.” *Available at http://www.sas.com/en_us/whitepapers/tdwi-self-service-big-data-prep-hadoop-107635.html.*
- SAS Institute Inc. *SAS 9.4 In-Database Products: Administrator’s Guide*. See chapters about SAS Data Management Accelerator for Spark. Available at http://support.sas.com/documentation/cdl/en/indbag/69030/PDF/default/indbag.pdf.
- SAS Institute Inc. *SAS Data Loader for Hadoop: System Requirements.* Includes supported Spark versions*. Available at* http://support.sas.com/documentation/installcenter/en/ikdmddhadpvofrsr/68979/PDF/default/sreq.pdf.
- SAS Institute Inc. SAS Data Loader for Hadoop Documentation page. Available at https://support.sas.com/documentation/onlinedoc/dmdd/.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Matthew Magne
100 SAS Campus Drive Cary, NC 27513
Matthew.magne@sas.com
@BigDataMagnet
http://sas.com/data