

Data Management with SAS[®] Special Collection



Foreword by
Ron Agresta

The correct bibliographic citation for this manual is as follows: Agresta, Ron. 2019. *Data Management with SAS®: Special Collection*. Cary, NC: SAS Institute Inc.

Data Management with SAS®: Special Collection

Copyright © 2019, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-64295-196-7 (PDF)

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.

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, NC 27513-2414

January 2019

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 software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

Table of Contents

Data Management in SAS® Viya®: A Deep Dive

By Wilbram Hazejager and Nancy Rausch

Doin' Data Quality in SAS® Viya®

By Brian Rineer

Is Your Data Viable? Preparing Your Data for SAS® Visual Analytics 8.2

By Gregor Herrmann

Ten Tips to Unlock the Power of Hadoop with SAS®

By Wilbram Hazejager and Nancy Rausch

Enable Personal Data Governance for Sustainable Compliance

By Vincent Rejany and Bogdan Teleuca

Data Management for Artificial Intelligence

By Todd Wright

Free SAS® e-Books: Special Collection

In this series, we have carefully curated a collection of papers that introduces and provides context to the various areas of analytics. Topics covered illustrate the power of SAS solutions that are available as tools for data analysis, highlighting a variety of commonly used techniques.



Discover more free SAS e-books!
support.sas.com/freesasebooks

 sas.com/books
for additional books and resources.

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. © 2017 SAS Institute Inc. All rights reserved. M1673525 US.0817


THE POWER TO KNOW.®

About This Book

What Does This Collection Cover?

Data may be the most valuable resource that your organization owns. *Data management* is the collection of practices used to acquire, validate, protect, store, govern, share, and process data. Data management involves more than just software; it encompasses the policies, procedures, and architecture that an organization develops to ensure that data is accessible, reliable, and secure. The volume and velocity of data is increasing, which means that managing data is a critical for the future. Successful data management in turn leads to successful analytics projects.

SAS offers many different solutions to manage your organization's data. The papers included in this special collection demonstrate how proper data management techniques can benefit every aspect of your organization's IT operations.

The following papers are excerpts from the SAS Global Users Group *Proceedings*. For more SAS Global Forum *Proceedings*, visit the [online versions of the Proceedings](#).

More helpful resources are available at support.sas.com and sas.com/books.

We Want to Hear from You

Do you have questions about a SAS Press book that you are reading? Contact us at saspress@sas.com.

SAS Press books are written *by* SAS Users *for* SAS Users. Please visit sas.com/books to sign up to request information on how to become a SAS Press author.

We welcome your participation in the development of new books and your feedback on SAS Press books that you are using. Please visit sas.com/books to sign up to review a book

Learn about new books and exclusive discounts. Sign up for our new books mailing list today at <https://support.sas.com/en/books/subscribe-books.html>.

Foreword

News about advanced analytics, machine learning, and artificial intelligence (AI) is seemingly everywhere. Rare is the organization that has not looked to advanced analytics and AI to automate processes, make better decisions, or reduce spending. SAS has an exceptional story to tell with our advanced analytics—in fact, we’re a market leader—but there’s a critical part of the analytics discussion that tends to be overlooked, without which no AI project will be successful.

That gap in the narrative is data. None of the promise of AI is possible without the ability to access, integrate, and transform data. SAS is intent on fundamentally changing the way our customers perform data management because changes in consumer expectations, and technology that drive them, continue to evolve at an incredible rate.

Just as conversations are evolving for the appropriate use of and analytics in everyday business processes, data management conversations are evolving as well. There have been several shifts in the data management market over the last few decades, and we’re in the middle of another one. There is new terminology that describes modern use cases, new personas involved in data management, and new platforms available that unite analytics, business intelligence, and data management.

Data access, data integration, data quality, and data governance—these are common and handy terms to describe the main subdomains of data management. In more modern parlance, however, these capabilities are described in terms of data lakes, data fabrics, and data hubs. Data integration and data quality become data preparation, and data governance is described in terms of data privacy and data protection. Regardless of the selected terminology, these are different names for data management activities that SAS has supported for years.

As for new personas, the growing recognition of data as a prized asset has given rise to different roles in organizations. Chief Data Officers and Chief Information Officers wield more influence in enterprise data strategy. Enterprise architects and Chief Information Security Officers are part of the conversation as well. Database administrators, coders, and ETL developers still exist but so do data engineers and data stewards. All these stakeholders are looking for a complete data management foundation on which to build diverse analytic-oriented projects.

Finally, the notion of independent data management platforms is waning. Increasingly, we see business intelligence vendors adding data management and analytics capabilities. We see data management vendors adding analytics. Other analytics vendors are extending their reporting and data management capabilities, too. These *insight platforms* that combine data management, visualization, and advanced analytics features are becoming much more common. SAS’ data management capabilities are best thought of as integral parts of the SAS platform, an insight platform that existed well before the new concept took hold.

So where are these changes taking us? We believe that most organizations will have to contend with the following in some way:

- Because data privacy has elevated awareness due to high-profile data breaches, we will see more attention given to data protection across enterprise applications, social media platforms, and cloud applications.
- We will see more organizations attempt to use AI and machine learning techniques to improve data quality and data management processes, but they will struggle to see meaningful results.
- We will see an increased desire for transparency about how data is being collected, aggregated, and shared. This will call for enhanced technology that can deliver detailed reports to organizations and their customers about data usage.

Those organizations that cope best with this changing data landscape dramatically increase their rate of success with analytics-driven projects.

It has been said that all data is big data now. It’s not necessarily data volumes that pose the biggest challenges—inexpensive technology to process billions of transactions is not uncommon—but what’s hidden in the data (good or bad) that can be difficult to resolve. Advanced analytics paired with good data management technology can help both detect threats and uncover untapped opportunities.

We will continue to see an increased use of even more advanced analytic capabilities to solve complex problems that in years past might have taken large teams and years of research to resolve. For this, it becomes even more critical to develop a comprehensive data management plan. We believe the content delivered here will help you do just that.

[Data Management in SAS® Viya®: A Deep Dive](#)

By Wilbram Hazejager and Nancy Rausch

This paper provides an in-depth look into the new SAS® data management capabilities in support of SAS® Viya® and SAS® Cloud Analytic Services. The paper includes an overview of SAS® Data Management in SAS Viya and contains details about how the feature set integrates with SAS Cloud Analytic Services. Examples and usage scenarios of how to best leverage the technology are also included.

[Doin' Data Quality in SAS® Viya®](#)

By Brian Rineer

SAS® Viya® introduces data quality capabilities for big data through data preparation and DATA step programming for SAS® Cloud Analytic Services (CAS). Learn how to configure SAS® Data Quality transformations in SAS® Data Studio and how to submit DATA step functions that are created in SAS Data Quality for execution in CAS. We also cover management of the vital SAS® Quality Knowledge Base in SAS® Environment Manager.

[Is Your Data Viable? Preparing Your Data for SAS® Visual Analytics 8.2](#)

By Gregor Herrmann

We all know that data preparation is crucial before you can derive any value from data through visualization and analytics. SAS® Visual Analytics on SAS® Viya® comes with a new rich HTML5 interface on top of a scalable compute engine that fosters new ways of preparing your data upfront. SAS® Data Preparation that comes with SAS Visual Analytics brings new capabilities like profiling, transposing or joining tables, creating new calculated columns, and scheduling and monitoring jobs. This paper guides you through the enhancements in data preparation with SAS Visual Analytics 8.2 and demonstrates valuable tips for reducing runtimes of your data preparation tasks. It covers integrating existing SAS® 9 tools and programs in your data preparation efforts for SAS Viya.

[Ten Tips to Unlock the Power of Hadoop with SAS®](#)

By Wilbram Hazejager and Nancy Rausch

This paper discusses a set of practical recommendations for optimizing the performance and scalability of your Hadoop system using SAS®. Topics include recommendations gleaned from actual deployments from a variety of implementations and distributions. Techniques cover tips for improving performance and working with complex Hadoop technologies such as YARN, techniques for improving efficiency when working with data, methods to better leverage the SAS in Hadoop components, and other recommendations. With this information, you can unlock the power of SAS in your Hadoop system.

[Enable Personal Data Governance for Sustainable Compliance](#)

By Vincent Rejany and Bogdan Teleuca

In the context of the European Union's General Data Protection Regulation (EU GDPR), one of the main challenges for data controllers and data processors is to demonstrate compliance by documenting all their data processing activities and, where appropriate, to assess the risk of these processes for the individuals. Such requirements cannot be achieved without being able to build an efficient data governance program. We use several processes developed in SAS® Data Management Studio to identify the personal data and update the governance view within SAS® Business Data Network and SAS® Lineage. We demonstrate several features in other products such as the Personal Data Discovery Dashboard in SAS® Visual Analytics and SAS® Personal Data Compliance Manager as it applies to Records of Processing Activities and the Data Protection Impact Assessment.

Data Management for Artificial Intelligence (Excerpt)

By Todd Wright

Machine learning systems don't just extract insights from the data they are fed, as traditional analytics do. They actually change the underlying algorithm based on what they learn from the data. So, the "garbage in, garbage out" truism that applies to all analytic pursuits is truer than ever. Few companies are already using AI, but 72 percent of business leaders responding to a PWC survey say it will be fundamental in the future. Now is the time for executives, particularly the chief data officer, to decide on data management strategy, technology, and best practices that will be essential for continued success.

We hope these selections give you a useful overview of the many tools and techniques that are available to incorporate data management best practices into your data analysis.

Ron Agresta

Director of Product Management – Data Management, SAS



As the Director of Product Management for all data management offerings at SAS, Ron Agresta works closely with customers, partners, and industry analysts to help research and development teams at SAS develop data access, data quality, data governance, data integration, and big data software and solutions. Ron holds a master's degree from North Carolina State University and a bachelor's degree from The Ohio State University.

Data Management in SAS® Viya®: A Deep Dive

Wilbram Hazejager and Nancy Rausch, SAS Institute Inc.

ABSTRACT

This paper provides an in-depth look into the new SAS® data management capabilities in support of SAS® Viya® and SAS® Cloud Analytic Services. The paper includes an overview of SAS® Data Management in SAS Viya, and contains details about how the feature set integrates with SAS Cloud Analytic Services. Examples and usage scenarios of how to best leverage the technology are also included.

INTRODUCTION

Self-service data preparation from SAS, on SAS Viya, helps you access, profile, cleanse, and transform data from an intuitive interface. SAS® Data Explorer copies data to SAS Cloud Analytic Services (CAS) and enables you to navigate and manage that data. SAS® Data Studio builds and executes collections of transformations on data that has been loaded to CAS.

This paper takes you through some common usage scenarios, and for each of those scenarios explains in more detail how the different architecture components are being used.

HIGH-LEVEL ARCHITECTURE

SAS® Data Preparation is built on SAS Viya, and encompasses a number of web applications, including the following (which is not an exhaustive list):

- SAS Data Explorer to manage data
- SAS Data Studio to build and execute collections of transformations on data
- SAS® Job Monitor to monitor the status of data management jobs

A number of additional web applications are available to the user as part of SAS Viya, including the following (which is not an exhaustive list):

- SAS® Environment Manager for managing a SAS Viya environment. It includes a dashboard view, which provides a quick overall look of your environment's health and status, as well as detailed views that enable you to examine and manage your environment in detail.
- SAS® Lineage Viewer to better understand the relationships between objects in your SAS Viya applications. These objects include data, transformation processes, reports, and visualizations.

SAS web applications talk to SAS Viya services, often referred to as microservices. SAS Viya includes services such as Audit, Identities, and Monitoring.

SAS Data Preparation uses SAS Cloud Analytics Services (CAS), as the run-time environment to interact with data sources and execute data transformations. CAS uses SAS® Data Connectors and SAS® Data Connector Accelerators to read from and write to data sources. Here is a high-level architecture diagram.

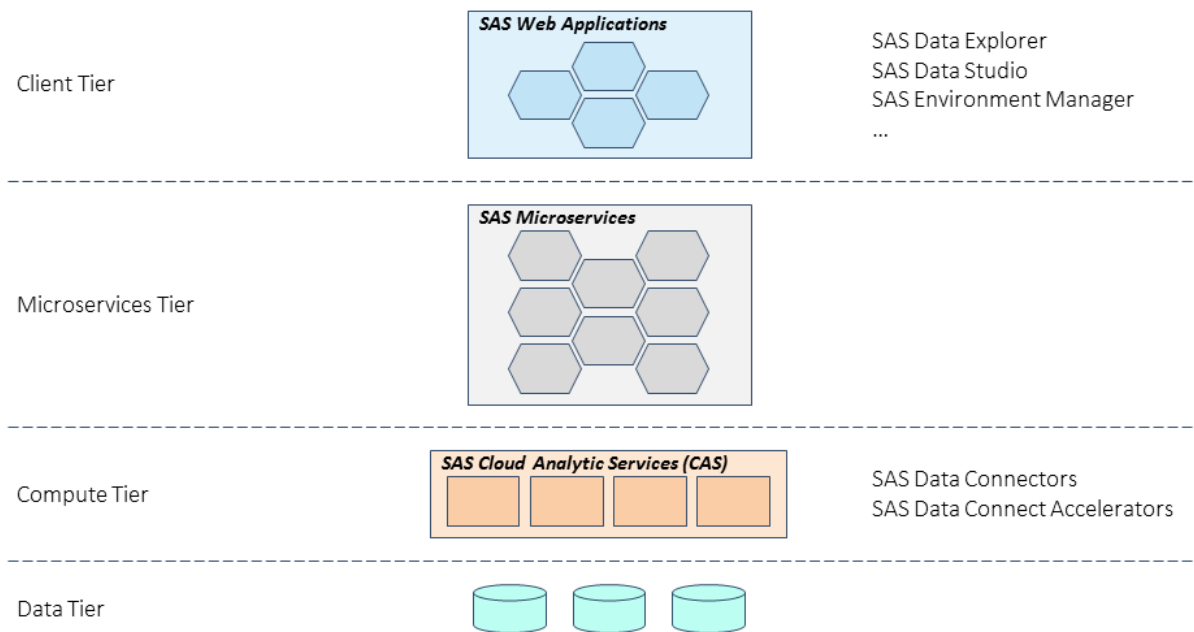


Figure 1. SAS Data Preparation: High-level Architecture

In this paper, we primarily focus on SAS Data Explorer and SAS Data Studio.

LOADING DATA IN CAS USING SAS DATA PREPARATION

SAS Data Preparation applications use CAS as the run-time engine. Using SAS Data Explorer, you can load data to CAS from a variety of sources, including files, databases, social media sources, and Esri. Profile metrics are available as well as can be seen in the following figure.

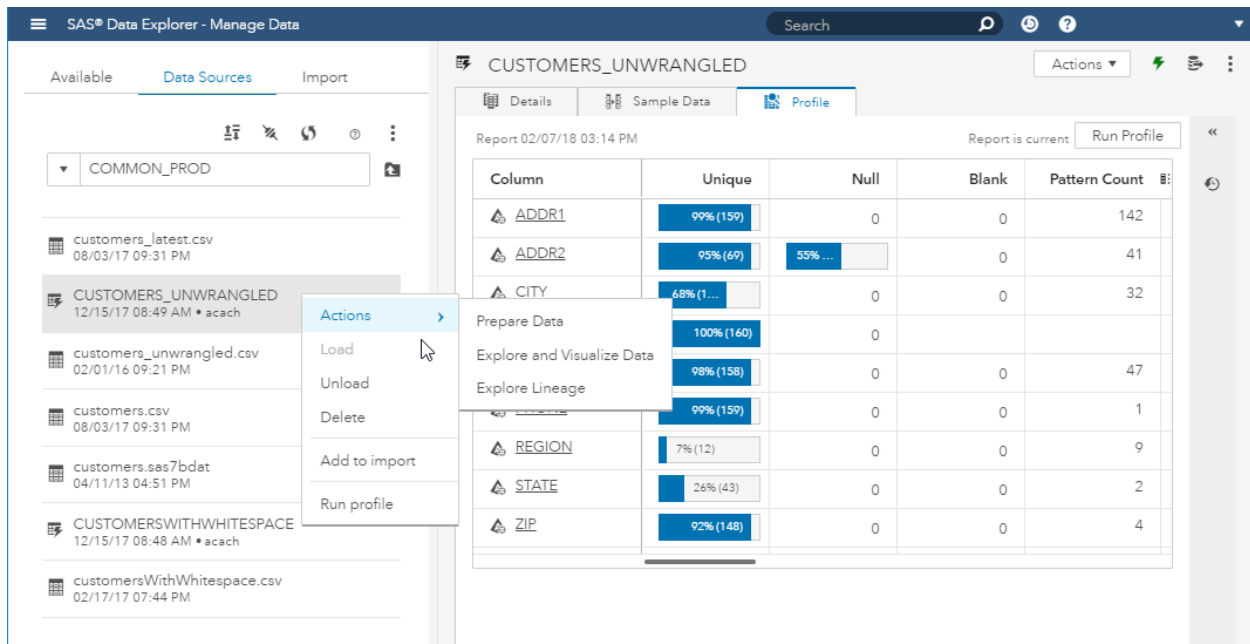


Figure 2. SAS Data Explorer: Profile Metrics

Loading Data from DBMS and Hadoop in CAS

Loading data from a DBMS or Hadoop in CAS is done using SAS Data Connectors and SAS Data Connect Accelerators.

SAS Data Connectors connect to the data source and can load data in a serial mode. In SAS Viya 3.3, these connectors have been enhanced to support a new data transfer mode called MultiNode. This mode allows multiple CAS worker nodes to connect to the data source at the same time. For more information about this data transfer mode, see *SAS® 9.4 and SAS® Viya® 3.3 Programming Documentation*.

SAS Data Connect Accelerators extend the functionality of SAS Data Connectors by enabling a parallel data load capability between the database clusters and CAS, while using SAS® Embedded Process framework to orchestrate such parallelization. Here is a diagram to illustrate this.

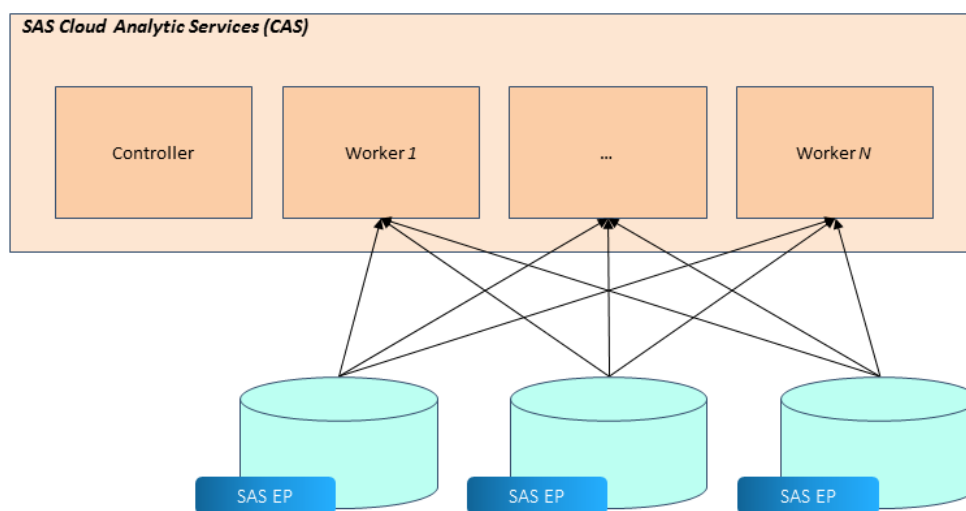


Figure 3. Parallel Loading from DBMS and Hadoop to CAS using SAS Embedded Process

SAS Embedded Process on each of the data source nodes has quick access to local data and can work directly with CAS worker nodes.

When using the SAS Data Connect Accelerator to Hadoop, the Hadoop ecosystem optimizes the request for data in such a way that each Hadoop node first tries to serve up its local data. This is to minimize data transfer between the Hadoop nodes. Therefore, the SAS Embedded Process transfers this local data to a CAS worker node. If CAS is co-located, then the SAS Embedded Process chooses the co-located worker, enabling optimal data transfer in terms of network usage.

While using the MultiNode approach, the data source is asked for a partition of the data, and that complete partition is then received by a single CAS node, the CAS node that requested the partition. Although MultiNode should be able to improve load times compared to using serial load, the SAS Data Connector Accelerator approach has more capabilities to optimize the parallel data load and if available is the recommend approach when working with large data sizes. Actually, when requesting data without the MultiNode option and SAS Data Connect Accelerator being available, CAS automatically uses parallel loading using SAS Embedded Process.

Starting with SAS Viya 3.3, these SAS data access components support both Read and Write access. The Write capabilities also apply to both of the parallel data transfer approaches discussed above.

Loading Files in CAS

SAS Data Preparation supports loading files into CAS. The file formats supported include Microsoft Excel, comma-separated value format (CSV format) files, and SAS data sets (sas7bdat and sashdat).

When the files reside on a file system that is directly accessible by CAS, we talk about server-side loading. This approach is used when you select a file from the Data Source panel in SAS Data Explorer. This data source requests a list of files from CAS and then lists server-side-available files.

When the files reside on the local file system of the machine where the browser is running, server-side processes, which includes CAS, typically cannot access that file system. However, using SAS Data Explorer, you can still load the data from the local file into a CAS table. We call this local file loading. This functionality is available using the Import tab of SAS Data Explorer. Under the covers, the data has to travel over the wire from web browser, to SAS Data Explorer mid-tier, and then (using microservices) to CAS as shown in the following diagram.

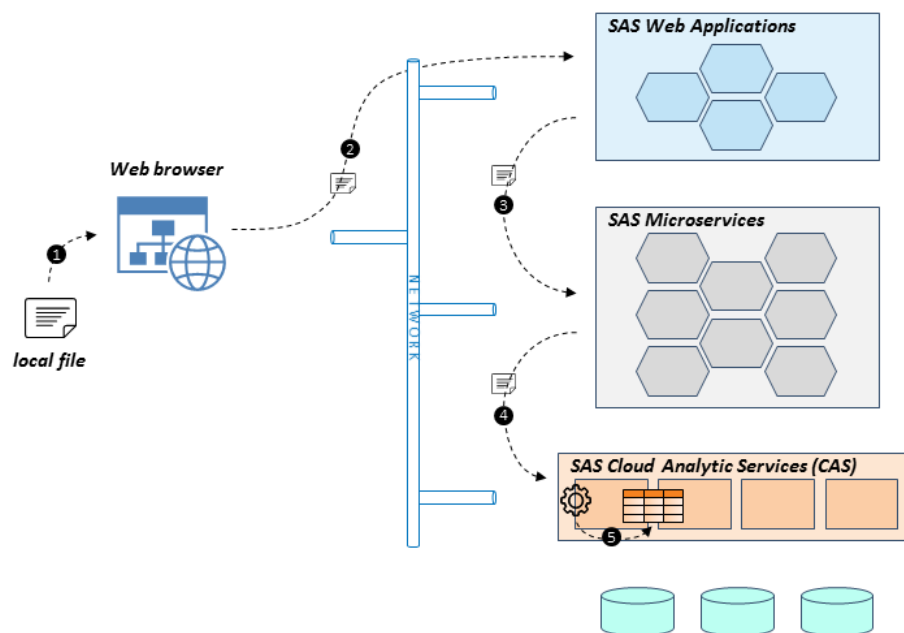


Figure 4. SAS Data Explorer: Local File Loading

Once the data arrives at the CAS server, functionality in the CAS server converts the file to a CAS table.

When loading large files, you should use server-side loading to minimize network usage.

Note that you can load data that is bigger than the total amount of memory in the CAS cluster.

TRANSFORMING DATA IN SAS DATA PREPARATION

SAS Data Studio is used to build and execute collections of transformations on data. It comes with a transformation library that includes common column transformations, table transformation to join and filter data, and data quality transformations. SAS Data Studio generates CASL code to execute the transformations and hands over the code directly to CAS. CASL is the CAS language and is used to invoke CAS actions. Actions are the equivalent of SAS procedures in the SAS language, and CAS supports a wide variety of actions. For more details about CASL, see *SAS® Cloud Analytic Services 3.3: CASL Reference*.

Because of the interactive nature of SAS Data Studio, it applies a transformation as soon as it has been added in the UI. Therefore, the user immediately sees the results of applying the transformation. These

intermediate results are stored in CAS session tables, and these tables have system-generated names. If you want to keep these results for future use, save the Data plan, which is the collection of all transformations that have been applied. When saving the Data plan, you are asked to provide a name for the output table. That name replaces the system-generated name. That output table is then also written to disk, so the data is still available after the CAS server restarts.

When a Data plan is executed, it is optimized such that all consecutive column transformations are executed in one DATA step action to minimize the number of times each record in the table is touched.

CUSTOM CODE TRANSFORMATIONS IN SAS DATA STUDIO

SAS Data Studio also supports Custom Code transformations, which allow you to write custom code that perform CAS actions or transformations on a table. There are two code languages available: CASL and DATA step. These Custom Code transformations support special references to use in your code to point to the output table of the previous transformation, and to indicate the table that is (going to be) created by your custom code, so the next transformation knows which table to use. Here is a diagram that shows this in pictures.

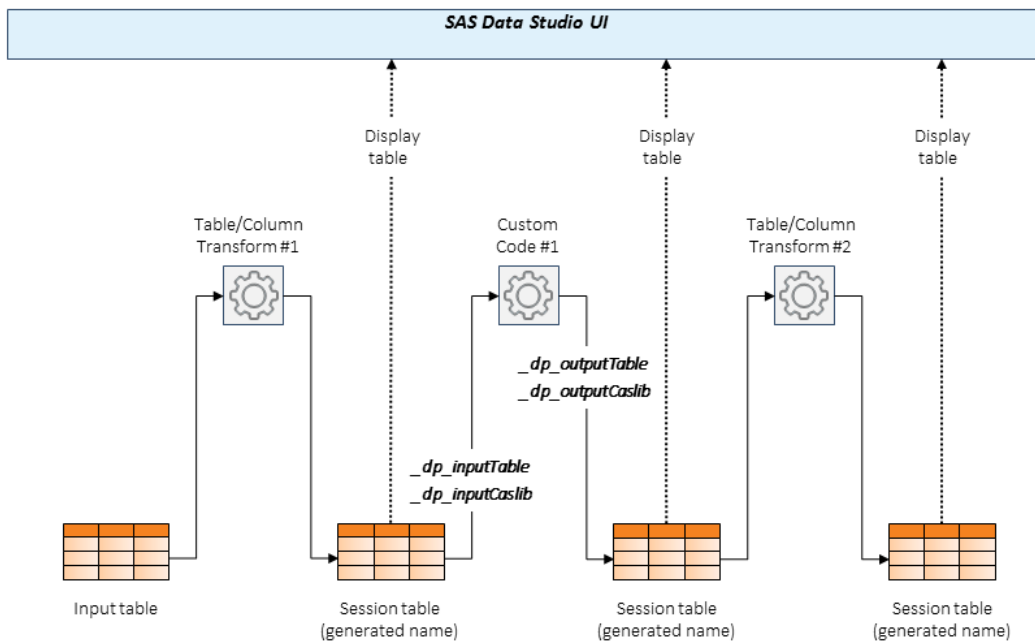


Figure 5. Referencing Input and Output Tables in Custom Code in SAS Data Studio

NOTE: These references are case-insensitive. We write them using special casing for readability purposes.

Later in this paper, we show code examples that use the `_dp_inputTable`, `dp_inputCaslib`, `_dp_outputTable`, and `_dp_outputCaslib` references.

A caslib is an in-memory space to hold tables, access control lists, and data source information. All data is available to CAS through caslibs, and all operations in CAS that use data are performed with a caslib in place.

When you use the DATA step language, you can take advantage of your SAS DATA step skills to write transformations, and those transformations will run in multiple threads in parallel in CAS. Note that not all SAS language elements are supported in CAS. See the “DATA Step Programming for CAS” topic in the *SAS 9.4 and SAS Viya 3.3 Programming Documentation* for more details around which elements are supported.

Given the breadth of CAS actions available, the Custom Code transformation is typically used to take advantage of CAS actions that are not covered by the standard transformations. Another usage scenario is one where you need to perform similar transformations many times to the same data record. Often this is easier to accomplish writing a small piece of Custom Code of type DATA step, compared to adding the same transformation many times and filling in the transformation properties each time.

CUSTOM CODE TRANSFORM OF TYPE DATA STEP

Here is a simple example where the Custom Code transformation of type DATA step can be used. In this example, a table contains answers for each question in a multiple-choice questionnaire where those answers were “compressed” into a single field. The following screenshot shows the input data and the Custom Code transform before the transformation is run.

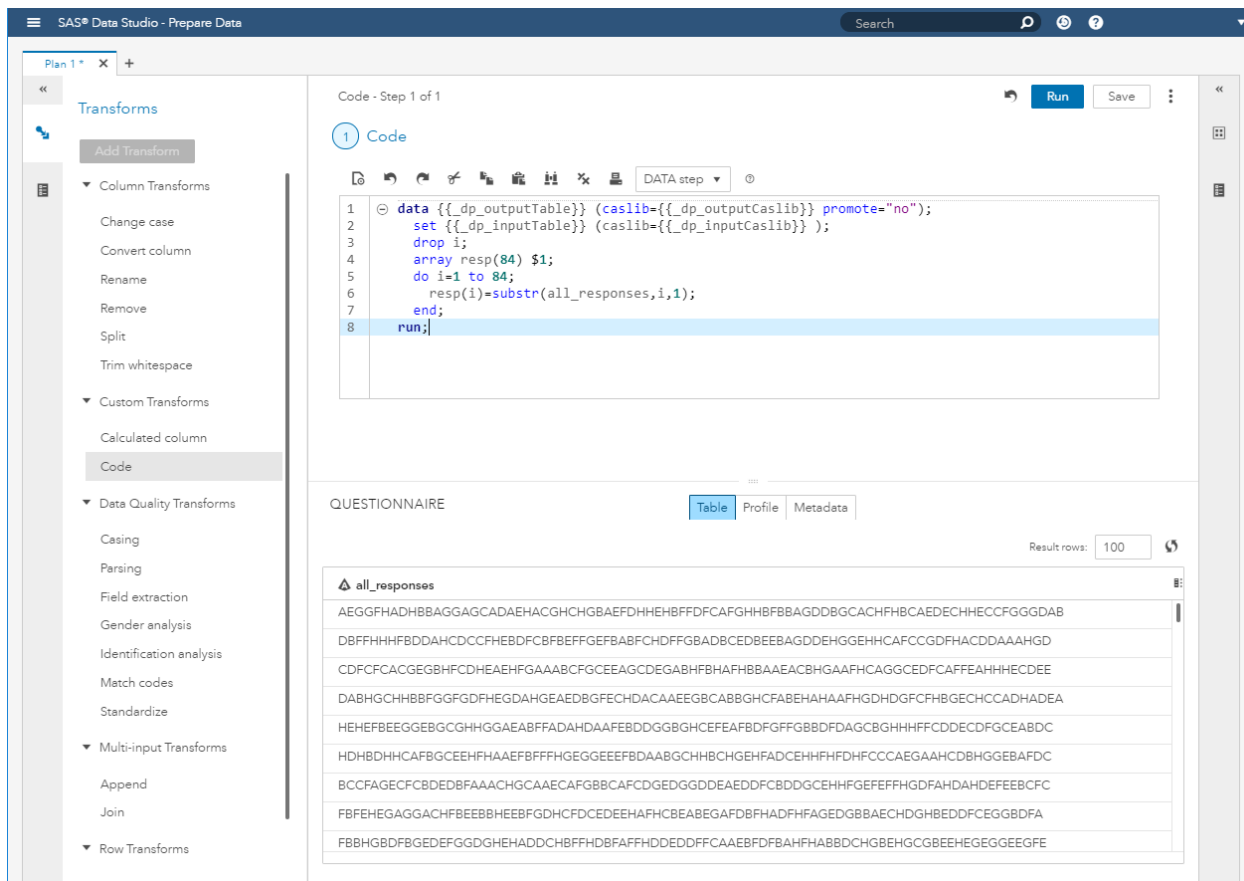


Figure 6. SAS Data Studio: Data before Custom Code Transform

Here is the DATA step code that can be used to create dedicated fields for each question.

```

data {{_dp_outputTable}} (caslib={{_dp_outputCaslib}} promote="no");
set {{_dp_inputTable}} (caslib={{_dp_inputCaslib}} );
drop i;
array resp(84) $1;
do i=1 to 84;
  resp(i)=substr(all_responses,i,1);
end;
run;

```

As mentioned earlier, Custom Code transformations support special references to point to input and output tables (and their corresponding caslib). Note that for the Custom Code transformation of type

DATA step, these references need to be enclosed in double curly braces as shown in the preceding code sample.

- `{{_dp_inputTable}}` points to the (system-generated) name of the table that was created by the previous transform.
- `{{_dp_outputTable}}` points the (system-generated) name of the table that the transformation is going to create. This table will be input for the next transform.

When the Data plan runs, the application replaces these references with actual table names. Here is screenshot showing the data after transformation.

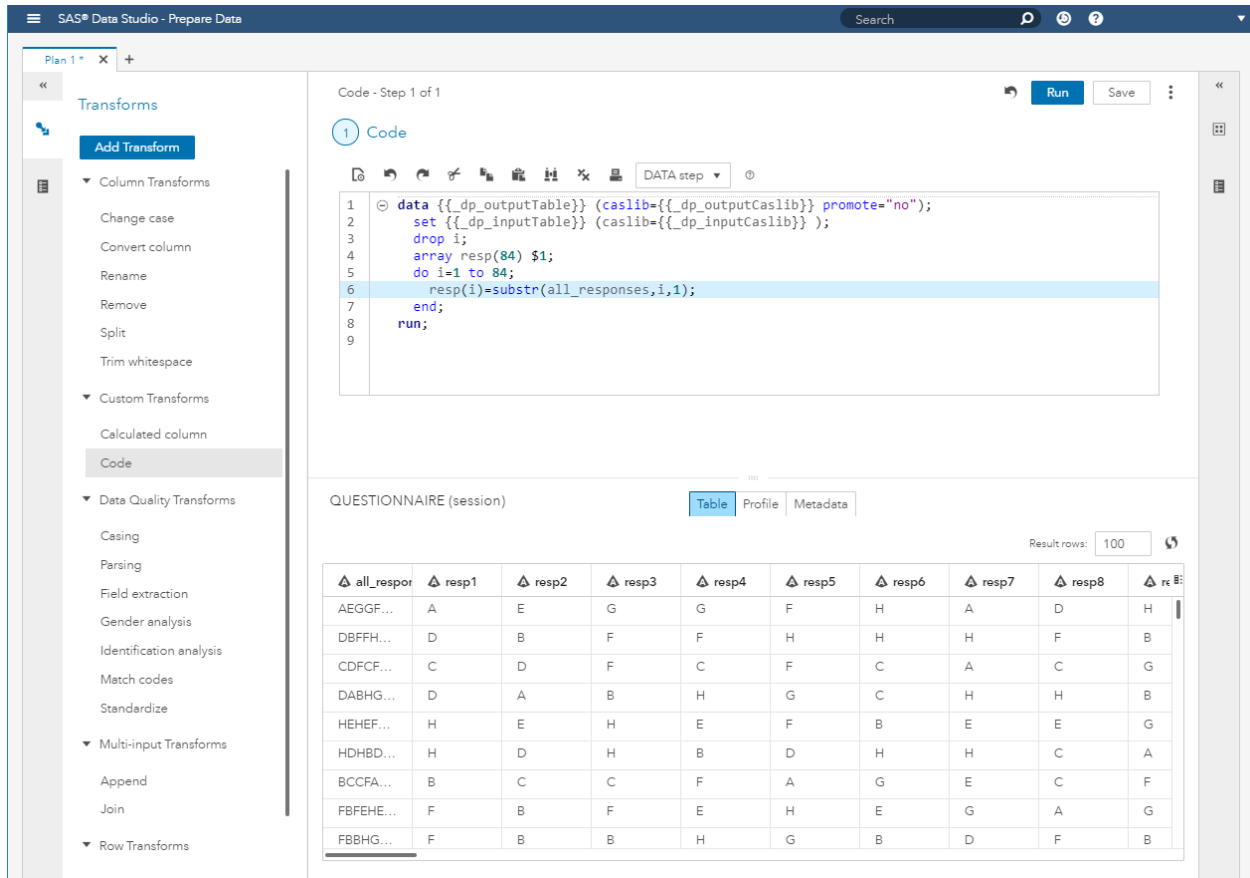


Figure 7. SAS Data Studio: Data after Custom Code Transform

When the transformation is the first one in a Data plan, the `{{_dp_inputTable}}` reference will resolve to the input table name for the Data plan. This table was specified during creation of the plan. Similarly, when a transformation is the last one in a Data plan and the user specified to save the plan, `{{_dp_outputTable}}` will resolve to the output table name that was specified by the user in the UI as part of the save plan interaction.

CUSTOM CODE TRANSFORM OF TYPE CASL

When using Custom Code of type CASL, you should not use double curly braces around the special references. When the Data plan runs, the application will prefix your code with some CASL code that creates CASL variables with the names `_dp_inputTable`, `_dp_inputCaslib`, `_dp_outputTable`, and `_dp_outputCaslib`, and fills in the appropriate values. These names can then be directly referenced in your CASL code.

Here is an example that uses the partition action from the table action set.

```

loadactionset('table');
partition status=rc /
  table={caslib=_dp_inputCaslib, name=_dp_inputTable}
  casout={caslib=_dp_outputCaslib, name=_dp_outputTable, replace=true}
;

```

Using the Custom Code transformation of type CASL, you can implement complex logic and invoke multiple CAS actions.

CONCLUSION

We hope that the information provided in this paper gives you a better understanding of how the SAS Data Preparation applications interact with SAS Cloud Analytic Services and some of the important options available for optimizing usage, especially when large amounts of data need to be handled.

RECOMMENDED READING

- Maher, Salman. 2018. "What's New in SAS® Viya Data Connectors." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/resources/papers/proceedings18/SAS1906-2018.pdf>.
- Rausch, Nancy. 2018. "What's New in SAS® Data Management." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/resources/papers/proceedings18/SAS1669-2018.pdf>.
- SAS Institute Inc. 2017. *SAS® Cloud Analytic Services 3.3: CASL Reference*. Cary, NC: SAS Institute Inc. Available at http://documentation.sas.com/?cdclid=pgmsascdc&cdcVersion=9.4_3.3&docsetId=proccas&docsetTarget=titlepage.htm&locale=en.
- SAS Institute Inc. 2017. *SAS® Cloud Analytic Services 3.3: DATA Step Programming for CAS*. Cary, NC: SAS Institute Inc. Available at http://documentation.sas.com/?cdclid=pgmsascdc&cdcVersion=9.4_3.3&docsetId=casdspgm&docsetTarget=p1eyivn5kal7qwn1drrdt71v21ml.htm&locale=en.
- SAS Institute Inc. 2017. "Data Connectors." *SAS® Cloud Analytic Services 3.3: User's Guide*. Cary, NC: SAS Institute Inc. Available at http://documentation.sas.com/?cdclid=pgmsascdc&cdcVersion=9.4_3.3&docsetId=casref&docsetTarget=n01iumvu56308zn1bud38udhg8w5.htm&locale=en.
- SAS Institute Inc. 2017. *SAS® 9.4 and SAS® Viya® 3.3 Programming Documentation: CAS User's Guide*. Cary, NC: SAS Institute Inc. Available at http://go.documentation.sas.com/?cdclid=pgmsascdc&cdcVersion=9.4_3.3&docsetId=casref&docsetTarget=titlepage.htm&locale=en.
- SAS Institute Inc. 2017. *SAS® Data Studio 2.1: User's Guide*. Cary, NC: SAS Institute Inc. Available at <http://documentation.sas.com/api/docsets/datastudioadv/2.1/content/datastudioadv.pdf?locale=en#na meddest=home>.
- SAS Institute Inc. 2017. *SAS® Viya® 3.3: Data Preparation*. Cary, NC: SAS Institute Inc. Available at <http://documentation.sas.com/?cdclid=dprepcdc&cdcVersion=2.1>.

CONTACT INFORMATION

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

Wilbram Hazejager
100 SAS Campus Dr

Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
Wilbram.Hazejager@sas.com
support.sas.com

Nancy Rausch
100 SAS Campus Dr
Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
Nancy.Rausch@sas.com
support.sas.com

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.

SAS2156-2018

Doin' Data Quality in SAS® Viya®

Brian Rineer, SAS Institute Inc

ABSTRACT

SAS® Viya® introduces data quality capabilities for big data through Data Preparation and DATA step programming for SAS® Cloud Analytic Services (CAS). In this session, a Senior Software Development Manager at SAS shows how to configure SAS® Data Quality transformations in SAS® Data Studio and how to submit DATA step functions that are created in SAS Data Quality for execution in CAS. We also cover management of the vital SAS® Quality Knowledge Base in SAS® Environment Manager.

INTRODUCTION

Data quality capabilities long available in SAS® Data Quality Server are now implemented in SAS Viya. SAS Data Quality in SAS Viya provides support for big data quality with the distributed processing power of CAS.

SAS Data Quality enables analysis and cleansing of structured text data. SAS Data Quality operations help you discover the semantic types of your data, break down data into constituent parts, standardize records into consistent formats, and identify potential duplicates.

If you're a SAS Data Quality Server user, you'll recognize a familiar set of data quality operations in the SAS Data Quality offering in SAS Viya. As with SAS Data Quality Server, you can analyze and categorize your data and cleanse it in preparation for analytics and investigation, or simply for better data hygiene and cleaner reporting. If you're new to data quality, see *SAS® Data Quality: Getting Started* for an overview and examples of SAS Data Quality operations.

In this paper, we briefly show how these data quality operations are accessed in SAS Viya.

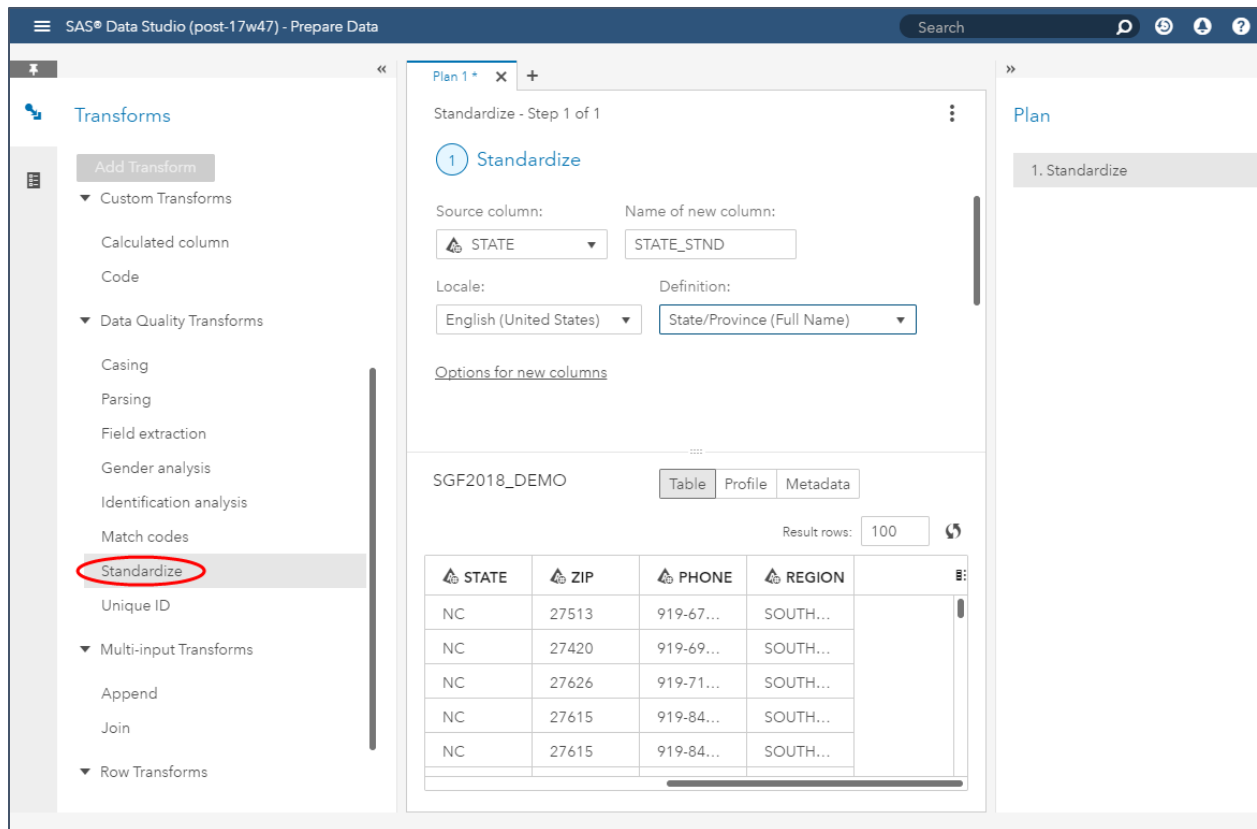
DATA QUALITY IN SAS DATA STUDIO

SAS Data Studio is a SAS® Data Preparation application in which you define transforms that are applied to your data in CAS. You can specify transforms to prepare your data for your analytics pipeline or to meet whatever data cleansing you might have.

In SAS Data Studio, you'll define transforms with an easy-to-use web browser interface. This interface enables you to create and modify transforms any time, in any environment. Once created, the set of transforms you've defined constitute a "plan" that is submitted for execution in CAS.

The data quality operations that are available in SAS Data Studio mirror the data quality operations available in traditional SAS products such as SAS Data Quality Server and DataFlux® Data Management Studio. These operations include identification analysis, standardization, parsing, extraction, casing, gender analysis, and matchcode generation.

An example of a data quality transform created in SAS Data Studio is shown below. In the example, we create a simple standardization transform. We do this by choosing to add a **Standardize** transform from the side menu, clicking to select a column to transform, and then specifying a context known as a "definition" and a locale for that definition.



Display 1: A Standardize Transform Applied to a STATE Column in SAS Data Studio

In our example, we specify that we will standardize the values in a STATE column. We select the **State/Province (Full Name)** definition and the **English, United States** locale to inform the software that the values to be transformed are names of US states.

To standardize another column, we add a second transform to the plan:

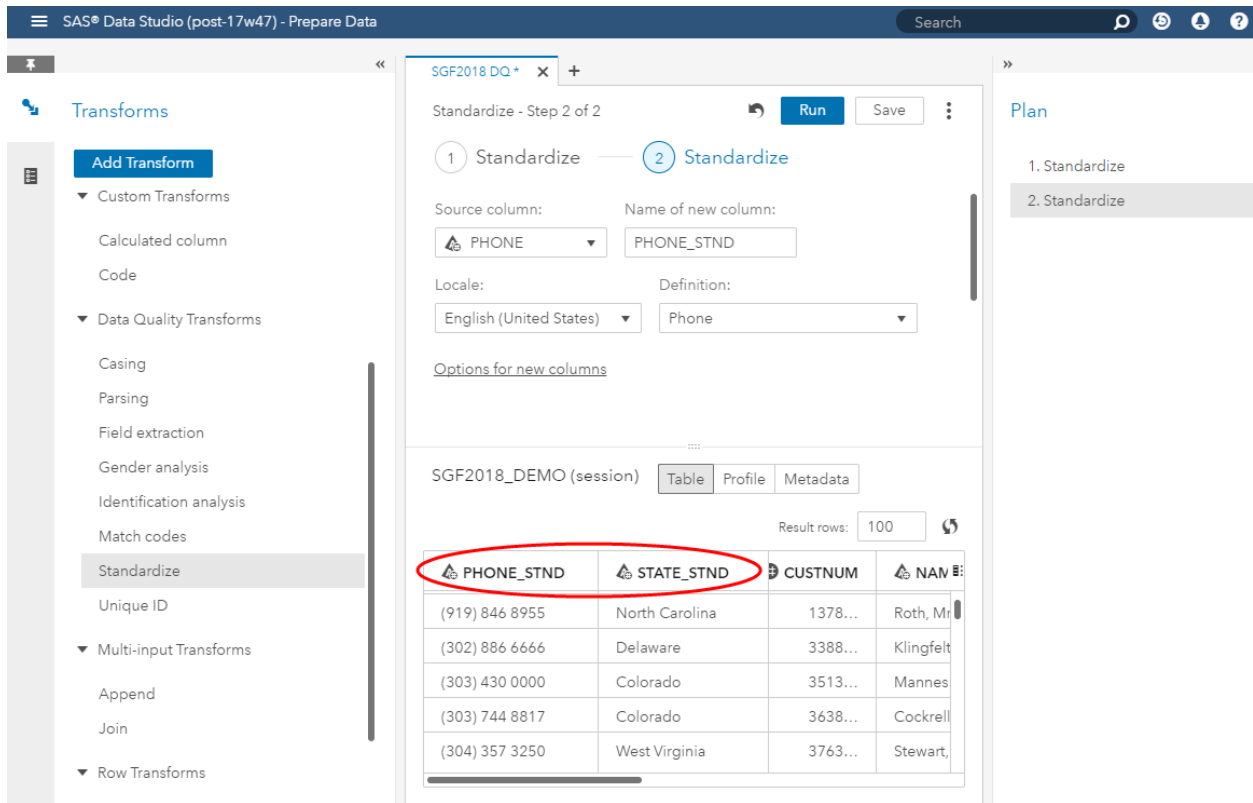
The screenshot shows the SAS Data Studio interface. On the left, the 'Transforms' sidebar is open, with 'Standardize' selected under 'Data Quality Transforms'. The main workspace shows the configuration for a second 'Standardize' transform. The 'Source column' is 'PHONE' and the 'Name of new column' is 'PHONE_STND'. The 'Locale' is 'English (United States)' and the 'Definition' is 'Phone'. Below the configuration, there is a 'Plan' sidebar on the right showing a sequence of two 'Standardize' steps. At the bottom, a data preview table for 'SGF2018_DEMO (session)' is shown with columns STATE, ZIP, PHONE, and REGION. The table contains five rows of data.

STATE	ZIP	PHONE	REGION
NC	27513	919-67...	SOUTH...
NC	27420	919-69...	SOUTH...
NC	27626	919-71...	SOUTH...
NC	27615	919-84...	SOUTH...
NC	27615	919-84...	SOUTH...

Display 2: A Second Standardize Transform Added to a Plan in SAS Data Studio

In this example, we add a transform to standardize the PHONE column. Notice that the plan now contains two Standardize transforms.

The plan can contain many transforms that will be executed in series. When the complete plan is ready, we click **Run** to submit it to CAS for execution on the CAS massive parallel processing grid. When the plan has been run, we see a preview of the two new columns that contain the output of our transforms:



Display 3: Columns Created by Standardize Transforms in SAS Data Studio

For more information about SAS Data Studio and about SAS Data Studio transforms and plans, refer to *SAS Data Studio 2.1: User's Guide*.

DATA STEP

If you're a SAS Data Quality Server user, you'll be glad to know that the same data quality DATA step functions exist in the SAS Data Quality offering in SAS Viya as in SAS Data Quality Server, with the same familiar syntax. But with SAS Data Quality in SAS Viya, you can submit your DATA step program for execution in CAS.

To submit your DATA step code for execution in CAS, you'll use a CAS action called *dataStep.runCode*. The syntax for your functions will be the same as in SAS Data Quality Server. Here is an example:

```
data CUSTOMERS;
  set SGF2018_DEMO;
  STATE_STND = dqStandardize(STATE, 'State/Province (Full Name)');
run;
```

This code applies the **State/Province (Full Name)** definition to standardize data in the STATE variable in the SGF2018_DEMO data set.

To submit this step for execution in CAS, you can wrap it in a call to the *dataStep.runCode* CAS action, and submit the call by invoking the *cas* procedure:


```

proc cas;

action dataStep.runCode /
code="
data CUSTOMERS;
  set SGF2018_DEMO;
  STATE_STND = dqStandardize(STATE, 'State/Province (Full Name)');
  ";
run;

```

You can submit this program in SAS Studio or anywhere you can execute SAS code. When you submit the program in your SAS environment, your DATA step code is routed to CAS and your data is processed in CAS.

For information about the *dataStep.runCode* CAS action, see the documentation for the DATA Step CAS Action Set.

For information about SAS Data Quality DATA step functions, with examples and syntax, see *SAS® Data Quality 3.3* and *SAS® 9.4 Data Quality Server: Language Reference*. Be aware that at the time of this writing, not all data quality functions are available for execution in CAS. In particular, functions that apply schemes are not yet supported in CAS.

PERFORMANCE

Regardless of whether you create data quality transforms in SAS Data Studio or invoke data quality DATA step functions using the *dataStep.runCode* CAS action, you will be harnessing the power of the CAS massively parallel processing architecture.

In CAS, your data is distributed across multiple worker nodes on the CAS grid. Each worker node processes the records that are located on that node. The processing occurs in parallel, meaning that the time required to process a large data set with SAS Data Quality in SAS Viya is a fraction of the time required to process the same data set with SAS Data Quality Server. Also, performance scales linearly in CAS according to the number of worker nodes available in the system.

While the time needed to execute data quality operations in SAS Viya depends on many factors — size of data set, complexity of data, grid load, and so on — in internal tests, we have seen data quality operations executed on millions of records in a few seconds on a large CAS grid.

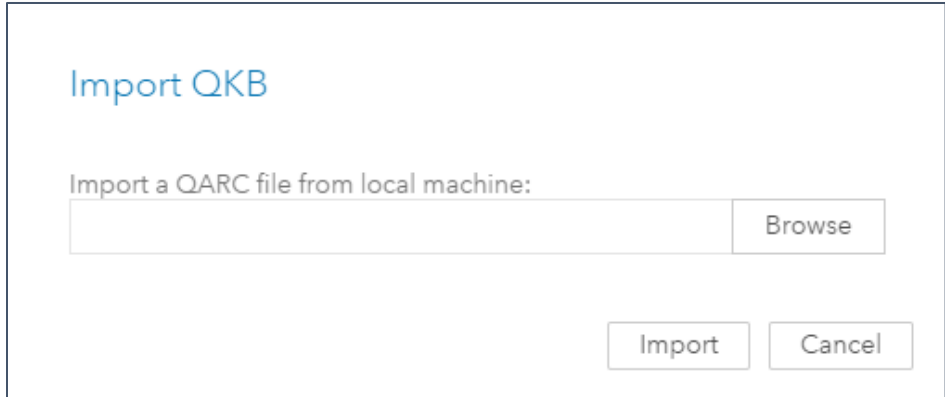
THE SAS QUALITY KNOWLEDGE BASE

As with SAS Data Quality Server, SAS Data Quality in SAS Viya requires a SAS Quality Knowledge Base (QKB). The QKB supplies rules and reference data used to analyze and transform your data.

The latest production version of the SAS® Quality Knowledge Base for Contact Information is deployed with SAS Data Quality. The entire QKB is deployed with your software order, with support for all available locales. At the time of this writing, this product supports thirty-nine locales. For more information about the SAS Quality Knowledge Base for Contact Information, refer to *About SAS Quality Knowledge Base* in the SAS Quality Knowledge Base for Contact Information online help.

To make the QKB available for use in CAS, you'll need to import it into your CAS system after you start the system for the first time. You can import a QKB into CAS using the SAS Environment Manager.

To import a QKB using SAS Environment Manager, open the SAS Home page in a web browser. Then, on the side menu, select **Reference Data** and then **Quality Knowledge Bases**. Next, on the Quality Knowledge Bases page, select **Import QKB**:



Display 4: Import QKB Dialog Box in SAS Environment Manager

The import process takes a few minutes. Generally, it takes a little longer for CAS systems with many worker nodes than for CAS systems with fewer worker nodes. After import, the QKB is stored as a repeated table in CAS. This means that a full copy of the QKB is available to every worker node for optimized processing at run time. For further instructions, refer to the QKB Management topic in *SAS Viya Administration*.

If you already have a QKB that you would like to begin using with SAS Data Quality in SAS Viya — for example, a customized QKB that you used previously with SAS Data Quality Server or the DataFlux® Data Management Platform — you can import that QKB into CAS as well. Note, however, that you must first convert your QKB into a format that is ready for import into CAS. This format is called a QKB Archive, or QARC. To create a QARC from your QKB, you'll use a utility that is provided with CAS. For details, see the QKB Management topic in *SAS Viya Administration*.

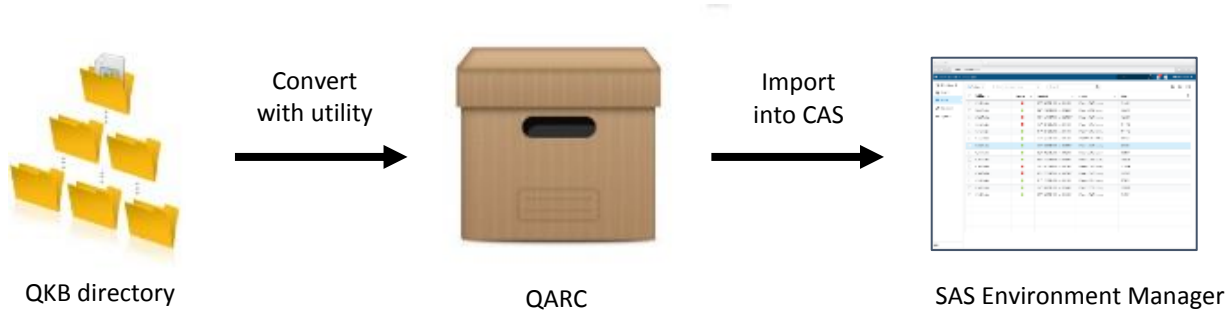


Figure 1: Process for Importing a Legacy QKB into CAS

After you have imported one or more QKBs into your CAS server, you can view the available QKBs using SAS Environment Manager:

Name	Server	Type	Product	Version	Default
CIQKB28	cas-shared-default	CAS	CI	v28	✓
PDQKB5	cas-shared-default	CAS	PD	v5	

Display 5: QKBs Viewed in SAS Environment Manager

Notice that there is a default QKB setting for your CAS server. If you open the properties screen for the default QKB, you can also find the default QKB locale setting. These settings tell SAS Data Studio which QKB you want to use when you execute data quality transforms, and which QKB and QKB locale you want to use when you submit data quality DATA step functions. (You can override these defaults in your DATA step program. See the *SAS Data Quality: Language Reference* for details.)

If you want to specify a default QKB and QKB locale, you can edit your CAS config file, and then restart your CAS server. Thereafter, you will see your default QKB and QKB locale settings in SAS Environment Manager. And the data quality transforms and DATA step functions will use these settings when they execute. For instructions on editing the default QKB and QKB locale settings in your CAS config file, see the QKB Management topic in the *SAS Viya Administration* documentation.

If you prefer to manage your QKBs programmatically rather than via a browser interface, you can call a CAS action to list QKBs, view QKB properties, and import or remove a QKB from CAS. An example is shown below. This example calls the *qkb.listQKBs* action to get a list of available QKBs for a given CAS server:

```
proc cas;
  action qkb.listQKBs;
  run;
quit;
```

For information about CAS actions for QKB management, refer to documentation for the QKB CAS action set.

GETTING STARTED

New to SAS Data Quality and eager to get started? Want a summary of SAS Data Quality functionality with links to references? If so, see *SAS® Data Quality: Getting Started*. This short book provides an overview of the various SAS Data Quality operations and what they do. It provides links to other books that contain product-specific details. It's a gateway to individual reference manuals that document the applications that surface SAS Data Quality operations and provide syntax and examples for programmatic interfaces.

CONCLUSION

With the introduction of SAS Data Quality in SAS Viya, you can harness the power of CAS to perform data quality operations on big data in a fraction of the time required by traditional data quality products. You can create data quality transforms with an easy-to-use web browser interface in SAS Data Studio, or write your own DATA step code to create custom programs. As always, a QKB is critical to SAS Data Quality. In SAS Viya, you can manage your QKBs via a web browser interface in SAS Environment Manager.

REFERENCES

DATA Step Action Set. Available at

<http://go.documentation.sas.com/?cdclid=vdmmlcdc&cdcVersion=8.11&docsetId=caspg&docsetTarget=cas-datastep-runcode.htm&locale=en>.

QKB Action Set. Available at

<http://go.documentation.sas.com/?cdclid=dqcdc&cdcVersion=3.3&docsetId=casactdq&docsetTarget=cas-qkb-TblOfActions.htm&locale=en>.

SAS[®] Data Quality 3.3: Getting Started. Available at

<http://go.documentation.sas.com/?cdclid=dqcdc&cdcVersion=3.3&docsetId=dqgs&docsetTarget=home.htm&locale=en>.

SAS[®] Data Quality 3.3 and SAS[®] 9.4 Data Quality Server: Language Reference. Available at

http://go.documentation.sas.com/?cdclid=pgmsascdc&cdcVersion=9.4_3.3&docsetId=dqclref&docsetTarget=titlepage.htm&locale=en.

SAS[®] Quality Knowledge Base for Contact Info online Help. Available at

<http://support.sas.com/documentation/onlinedoc/qkb/28/QKBCI28/Help/qkb-help.html>.

SAS[®] Data Studio 2.1: User's Guide. Available at

<http://go.documentation.sas.com/?cdclid=dprepcdc&cdcVersion=2.1&docsetId=datastudioadv&docsetTarget=titlepage.htm&locale=en>.

SAS[®] Viya[®] 3.3 Administration: QKB Management. Available at

<http://go.documentation.sas.com/?cdclid=dqcdc&cdcVersion=3.3&docsetId=calqkb&docsetTarget=titlepage.htm&locale=en>.

RECOMMENDED READING

- Rausch, Nancy. 2018. "What's New in SAS Data Management." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available at <https://support.sas.com/resources/papers/proceedings18/SAS1669-2018.pdf>.
- Rineer, Brian. 2015. "Garbage In, Gourmet Out: How to Leverage the Power of the SAS[®] Quality Knowledge Base." *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>.
- Rineer, Brian. 2016. "Get out of DATA Step Code and into Quality Knowledge Bases." *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. Available at <https://support.sas.com/resources/papers/proceedings16/SAS5644-2016.pdf>.
- *SAS Data Quality 3.3: Getting Started*. Available at <http://go.documentation.sas.com/?cdclid=dqcdc&cdcVersion=3.3&docsetId=dqgs&docsetTarget=home.htm&locale=en>.

CONTACT INFORMATION

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

Brian Rineer
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
+1-919-677-8000
Brian.Rineer@sas.com
<http://www.sas.com>

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.

Is Your Data Viable? Preparing Your Data for SAS® Visual Analytics 8.2

Gregor Herrmann, SAS Institute Inc.

ABSTRACT

We all know that data preparation is crucial before you can derive any value from data through visualization and analytics. SAS® Visual Analytics on SAS® Viya® comes with a new rich HTML5 interface on top of a scalable compute engine that fosters new ways of preparing your data upfront. SAS® Data Preparation that comes with SAS Visual Analytics brings new capabilities like profiling, transposing or joining tables, creating new calculated columns, and scheduling and monitoring jobs. This paper guides you through the enhancements in data preparation with SAS Visual Analytics 8.2 and demonstrates valuable tips for reducing runtimes of your data preparation tasks. It covers integrating existing SAS® 9 tools and programs in your data preparation efforts for SAS Viya.

INTRODUCTION

Sources of data have gone beyond the boundaries of IT-managed enterprise data warehouses. Organizations are facing the flux of ad hoc sources that business users need to make more informed decisions.

One of the key criteria for the successful use of a BI application is being able to import users' ad hoc data sources in a self-service manner for data analysis without depending on IT resources. In addition to access to these ad hoc data sources, there is an increasing need to enhance data suitable for the needs of analysis, without the need for the IT department to make changes to the centralized data source, which can often take a long time.

Business solutions that allow data access and data manipulation for business analysts are gaining more traction. SAS Visual Analytics comes with capabilities that can empower specifically enabled users to bring their own data into the environment and to further refine it by modifying existing data items or adding new data items. The goal of this self-service data management capability is to provide a managed, yet self-service way for users to provision and prepare their own data without always having to rely on IT. The subsequent sections of this document dive into more details about these data preparation tasks. In addition, integrating existing SAS 9 data preparation jobs into a workflow for making data available for analysis in SAS Visual Analytics 8.2 is covered.

SAS VISUAL ANALYTICS ON SAS VIYA TECHNOLOGY OVERVIEW

SAS Visual Analytics delivers analytical visualizations that use intelligent ways to help business analysts and nontechnical users to see patterns and trends and to identify opportunities for further analysis. SAS Visual Analytics is backed by the power of SAS Viya, which is available to users in a self-service and approachable manner. SAS Visual Analytics enables the creation and dissemination of dashboards, reports, and the results of investigative exploration, either to the web or to native mobile applications.



Figure 1. Analytical Visualizations

SAS Visual Analytics content can be augmented by advanced statistical methods and machine learning algorithms in one unified HTML5 interface by using additional modules of the SAS Viya product suite. SAS Visual Analytics includes the capability to prepare data before making it available to users, an interface for exploring your data (often known as data discovery), and an interface for building highly interactive and visual reports and dashboards.



Figure 2. SAS Visual Analytics Main Components

Because SAS Visual Analytics 8.2 is the second release on SAS Viya, it is important to understand the technical differences from the previous versions of SAS Visual Analytics running on SAS 9. Let's have a closer look at the underlying architecture:

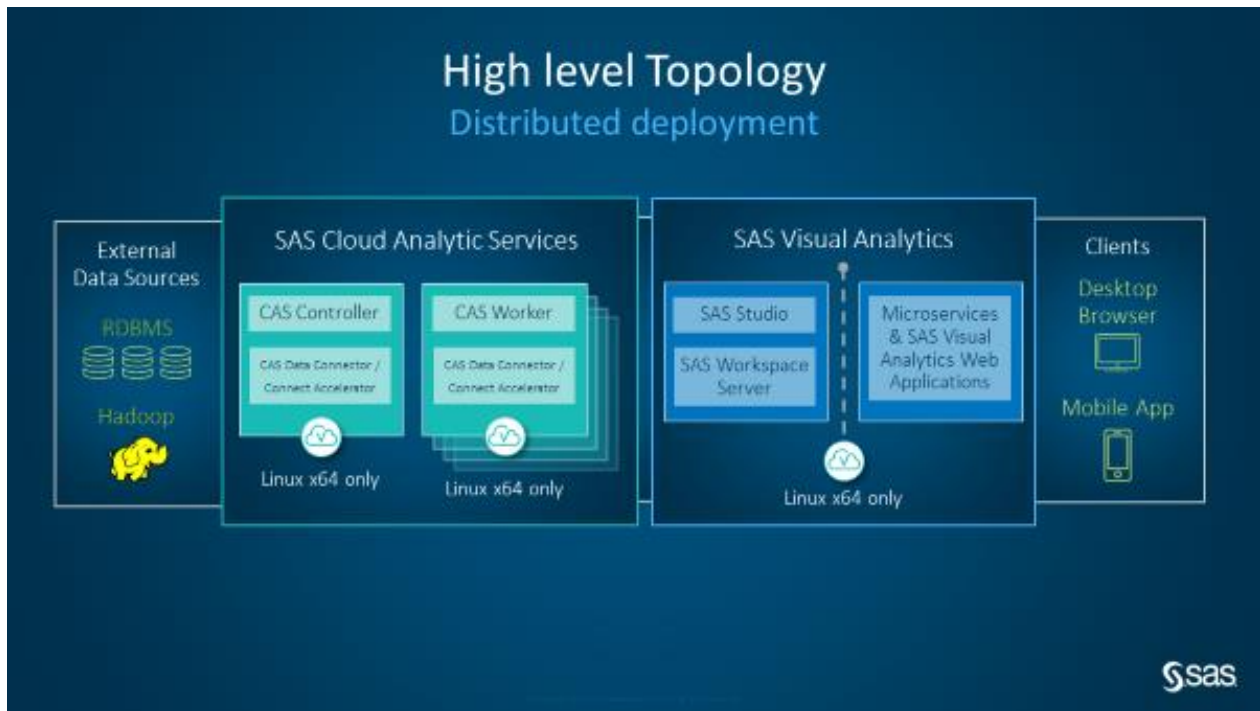


Figure 3. SAS Visual Analytics Topology

Compared to SAS Visual Analytics 7.x, the compute engine, SAS® LASR Analytic Server, has been replaced in SAS Visual Analytics in SAS Viya by the CAS server. You can call CAS the next generation in-memory compute server that brings new capabilities like failover and resiliency. The main difference from the SAS LASR Analytic Server is that data preparation is no longer executed by a SAS program on the head node. The CAS server brings its own data preparation capabilities, spreading its workload to all workers across a distributed environment. This allows data management transformations on larger data volumes with short execution times because the operations are done in parallel across the worker nodes.

DATA ACCESS

Data preparation always starts with accessing data. SAS Visual Analytics on SAS Viya supports a large variety of file formats in a standard configuration. Data can be imported from your local file system or from social media, or you can choose from data that is already available on the server.

SUPPORTED FILE FORMATS

Importing Local Files

Local files are files that can be accessed via the operating system of the machine on which you are running your browser to access SAS Visual Analytics. The following file formats can be imported from your local file system:

- Comma-separated values (CSV) files or TXT files.
- SAS data sets (SASHDAT or SAS7BDAT). SAS data set views (SAS7BVEW) cannot be loaded into CAS tables.

- Microsoft Excel workbook (XLSX) files and Excel 97-2003 workbook (XLS) files. You cannot import XLST, XLSB, XLSM, or other Excel file types. You cannot import pivot tables. To import native Microsoft Excel files, SAS Data Connector to PC File Formats is required.

Importing Social Media Data

With regard to social media, SAS Visual Analytics on SAS Viya supports the following data imports:

- Twitter
- Facebook
- Google Analytics
- YouTube
- Google Drive

To load data from the different social media channels, you must allow SAS Visual Analytics to access your account.

Accessing Server Files

If your data is already loaded onto the CAS server, it can be accessed via the **Available** data pane in the **Open Data Source** window. (See Figure 5.) Data that is already physically stored on the CAS server, but not yet loaded into memory, can be opened using the **Data Sources** pane of the same window. After clicking on **Data Sources**, a list of available CAS libraries is displayed. Drilling down into one of these libraries shows all available tables within that library. The icon beneath the table name indicates the table type.

- CAS table (a table already in the specific CAS format with extension .sashdat)
- Physical table (a text file usually in CSV format or a SAS 9 file with extension .sas7bdat)
- In-memory table (a table already loaded into memory on the CAS server; this file does not have any extension)

If you are loading SAS 9 tables that contain user-defined formats, they must be made available to your CAS environment before loading. The easiest way is to create a CAS table that contains all your user-defined formats, and then save it to the appropriate caslib. You can see the default settings for the caslib formats in SAS Environment Manager in Figure 4. SAS Environment Manager comes with a nice interface to check the availability of user-defined formats and make modifications if necessary.

Library	Server	Type	Path	Description	Personal
CPSAppData	cas-shared-default	PATH	/opt/sas/viya/config/data/cas/default/cpsAppData/	Stores data for th...	false
demodata	cas-shared-default	PATH	/opt/demodata/	General Caslib for...	false
Formats	cas-shared-default	PATH	/opt/sas/viya/config/data/cas/default/formats/	Stores user define...	false
mmLibs	cas-shared-default	PATH	/opt/sas/viya/config/data/cas/default/modelMonitorLi...	Library for Model ...	false
Models	cas-shared-default	PATH	/opt/sas/viya/config/data/cas/default/models/	Stores models cre...	false
ModelStore	cas-shared-default	PATH	/opt/sas/viya/config/data/cas/default/modelStore/	Stores model anal...	false
ProductData	cas-shared-default	PATH	/opt/sas/viya/home/share/productData/	Stores product da...	false

Figure 4. Default Caslibs in SAS Environment Manager

When you have selected the table that you want to open, you get a short summary of the table, including the number of rows and columns. You can switch to the **Sample Data** or **Profile Panel** to see more details of the content of your selected data source.

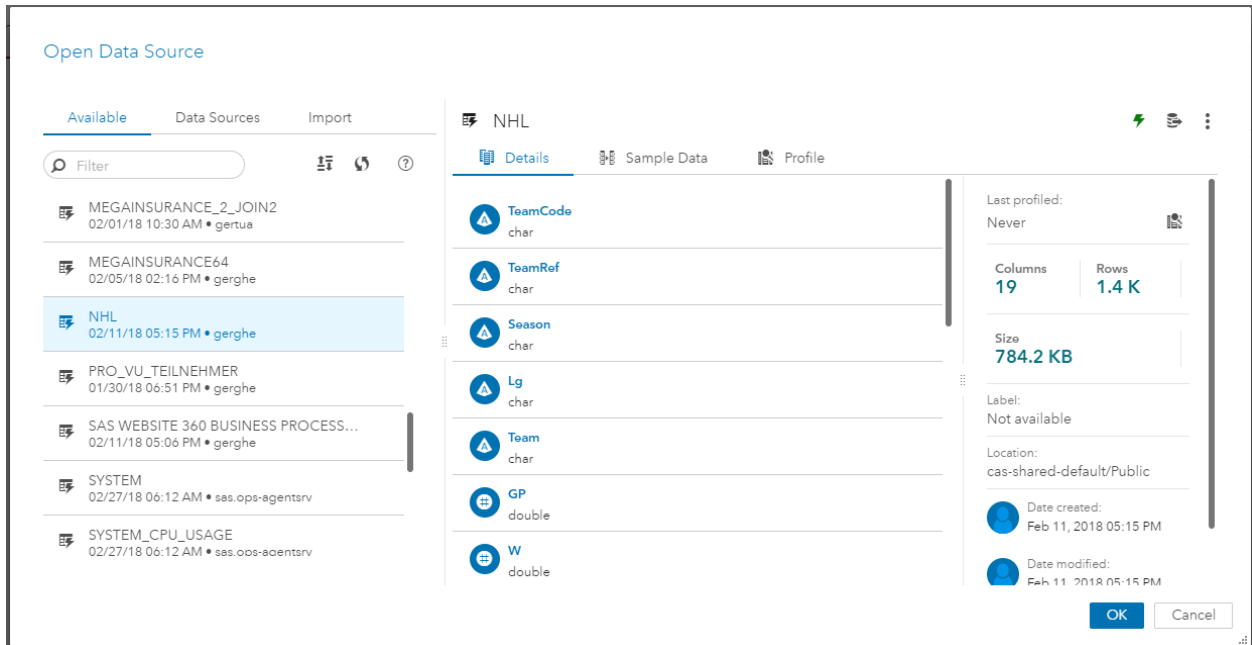


Figure 5. Open Data Source

DATA PREPARATION IN SAS VISUAL ANALYTICS 8.2

BASIC CONCEPTS

The SAS® Data Studio interface enables you to prepare and view data. Data preparation tasks in SAS Visual Analytics are stored as data plans. A plan is a collection of data transforms or actions performed on a table. SAS® Data Studio provides a convenient way for you to prepare data in tables, to keep track of the changes that you make to tables, and to modify or view the history of actions that you made to tables. If you start with a new plan, your first action is always adding a table to a plan. If you are not familiar with the content of your table, you should run a profile on your table. The profile gives you a good overview and might contain some hints if you are facing data quality issues (for example, variables that contain a lot of missing values).

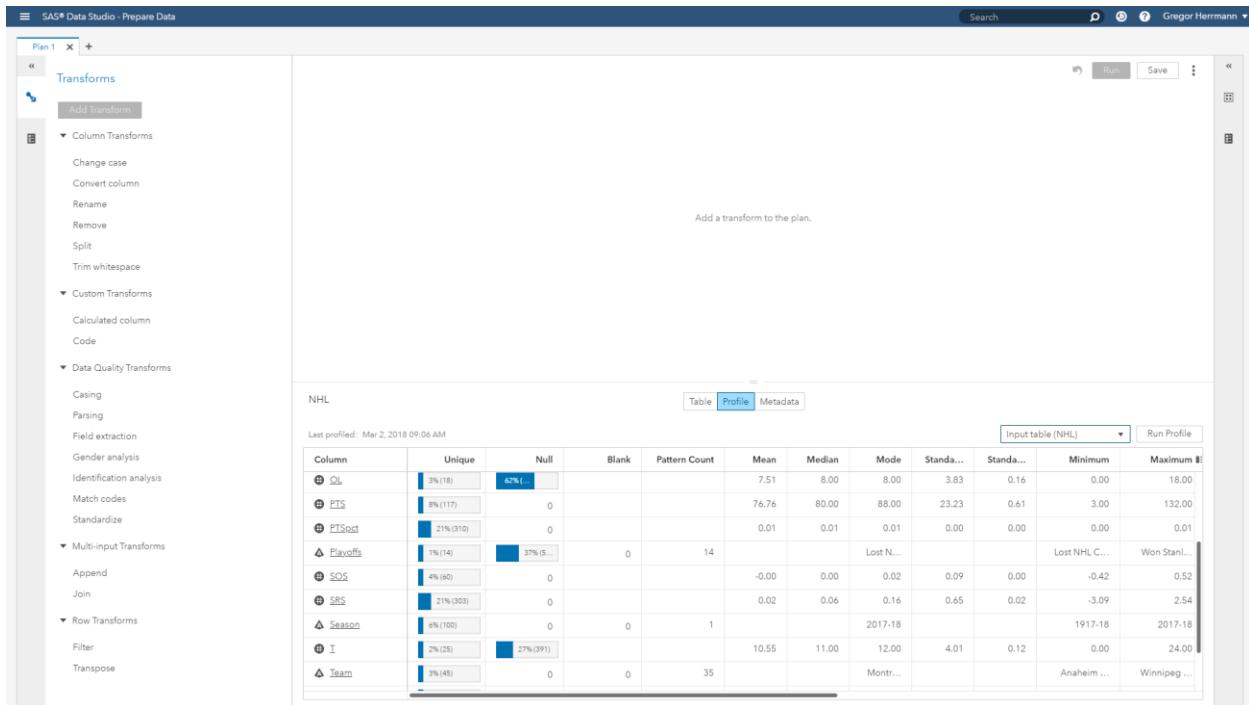


Figure 5. Viewing a profile in SAS Data Studio

DATA TRANSFORMS

You can choose from a large variety of data transforms to perform desired operations on your existing data. The transforms range from simple transforms like Rename, Change case, or Trim whitespace to more complex actions like Join or Transpose. They cover the most frequently used data transformations in terms of self-service data preparation. A sequence of data transforms can be saved as a data plan. Every transform requires initial modifications to be able to run the action. After all transforms have been executed successfully, you can save your plan.

If SAS® Data Preparation is licensed at your site, you can access the data quality transforms and integrate one or more of them into your data plan.

The data quality transforms use SAS® Quality Knowledge Base (QKB), which is a collection of locales and other information that is referenced during data analysis and data cleansing. The data quality transforms apply a QKB locale and a definition to a selected source column. Definitions define data formats for specific types of content and data cleansing. For example, a parse definition for a street address describes how a street address can be parsed into identifiable segments.

A locale reflects the language and linguistic conventions of a geographic region. These conventions can include word order or language selection for the country or region.

The screenshot shows the SAS Data Studio interface. On the left, a 'Transforms' sidebar lists various operations. The main workspace is titled 'Split - Step 3 of 4'. It shows a data plan with four steps: 1. Join, 2. Convert Column, 3. Split (highlighted), and 4. Gender Analysis. The 'Split' step configuration shows 'Source column' as 'Customer_Name', 'Split data' as 'On a delimiter', 'Delimiter' as 'Space', and 'Name of new column 1' as 'LEFT_Customer_Name'. Below this, a data table for 'ORDER_FACT (session)' is shown with columns: Order_Type, LEFT_Cus, RIGHT_Cus, Employee_ID, Street_ID, Order_Date, Delivery_Date, and Order_ID. The table contains 18 rows of data.

Figure 5. Data Plan

If you want to accomplish more complex data transformations, you can insert a code transform into your data plan. You can choose from two available code languages: CASL and DATA step. Each time you run a plan, table and library names might change. To avoid errors, you must use variables instead of table names and CAS library names in your code. Using variables instead of table names and CAS library names eliminates the possibility that the code will fail due to name changes. You can see the variables in the first line of the **CustomCode** transform in the following screenshot. After executing your CustomCode transform, you can download the log to check correctness.

The screenshot shows the SAS Data Studio interface with a 'CustomCode' transform selected. The main workspace is titled 'Code - Step 1 of 1'. It shows a code editor with the following DATA step code:


```

1 data {{_dp_outputTable}} (caslib={{_dp_outputCaslib}}); set {{_dp_inputTable}} (caslib={{_dp_inputCaslib}});
2 if _n_ = 1 then do;
3   _mult = 10 ** (int(log10(_NTHREADS_)) + 1);
4   retain _mult;
5   drop _mult;
6 end;
7 *UniqueID*n = _THREADID_ + (_N_ * _mult);
8 run;
    
```

 Below the code editor, a data table for 'CARS (session)' is shown with columns: EngineSize, Cylinders, Horsepower, MPG_City, PG_Highway, Weight, Wheelbase, Length, and UniqueID. The table contains 10 rows of data. On the right, a 'Result Table - CARS (session)' summary shows 16 columns, 428 rows, and a size of 103.7 KB.

Figure 6. CustomCode Transform

ADDITIONAL CAPABILITIES

From the toolbar, by clicking on the symbol with three dots, a drop-down menu is displayed, providing additional capabilities. You can download both the code and the log of the current data plan, you can change the source table of the data plan, or you can immediately start building a report or developing a model with either your source or your target table.

If you choose **Save As** from the drop-down menu, you can specify the location and the name of your target table and whether the table should be replaced if it already exists. To run a data plan on a regular basis in batch, you must create a job. After giving your job a name and saving it, the job can be scheduled to run in the background from SAS Environment Manager. In the current version, only time-based triggers are available.

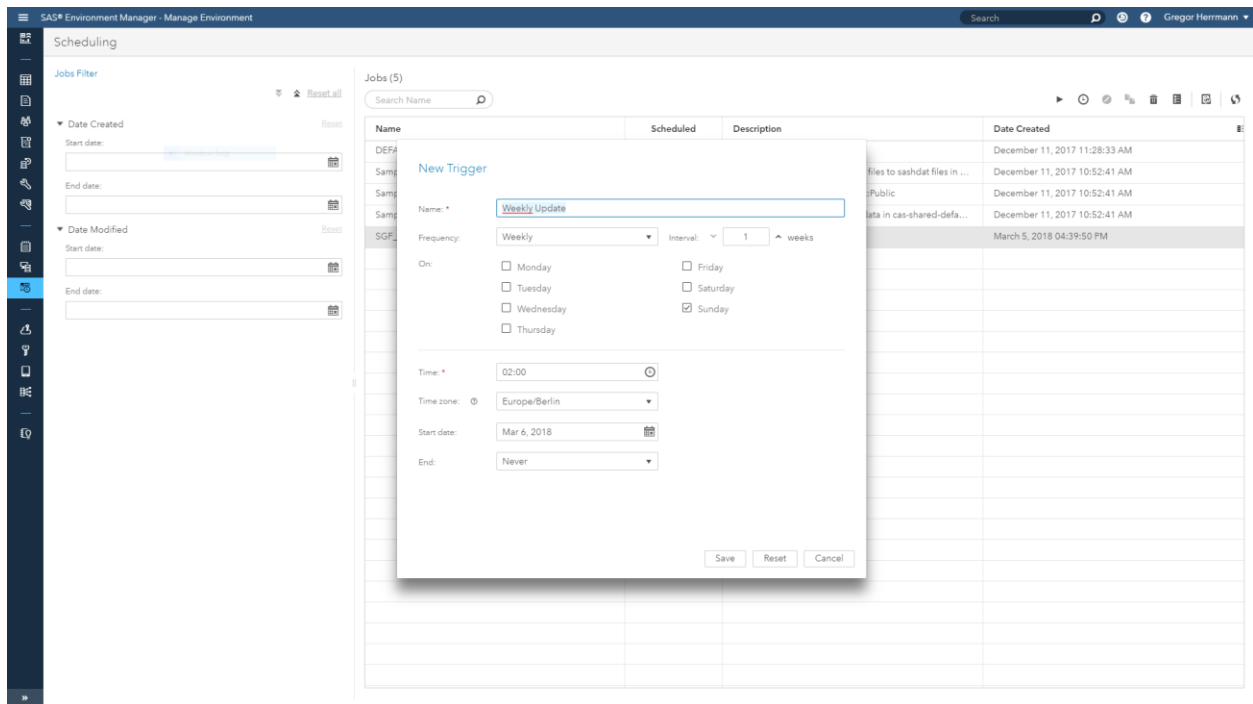


Figure 6. Scheduling a job

USING SAS 9 PROGRAMS AND TOOLS

You had a closer look at how data preparation works in SAS Visual Analytics on SAS Viya. Many of you, however, might have large amounts of existing SAS programs that do data preparation for you. What if you want to use these valuable assets?

Up to SAS 9.4M4, there was a way to execute SAS programs in a SAS Viya environment using SAS/CONNECT. It required SAS/CONNECT in both the SAS 9 and the SAS Viya environment, and it was not very easy to use. Beginning with SAS 9.4M5, executing programs in CAS from an existing SAS session got much easier. Let's dive deeper into it. The following screenshots show SAS® Enterprise Guide® as the interface to execute the SAS programs. The same code examples can be used from SAS® Studio or the Display Manager.

Connecting to CAS

To be able to execute any code on the CAS server, you have to make a connection first, which requires an identity that can authenticate against the operating system of your CAS server. In this example, connection is made from a Windows 10 laptop to a CAS server running on Linux using an .authinfo file that contains the credential information. It is basically a text file with a userid and encrypted password. A

similar mechanism is available for other operating systems as well. The three lines of code in the screenshot establish a connection to a CAS server, start a CAS session, and make all caslibs available in the SAS session in SAS Enterprise Guide.

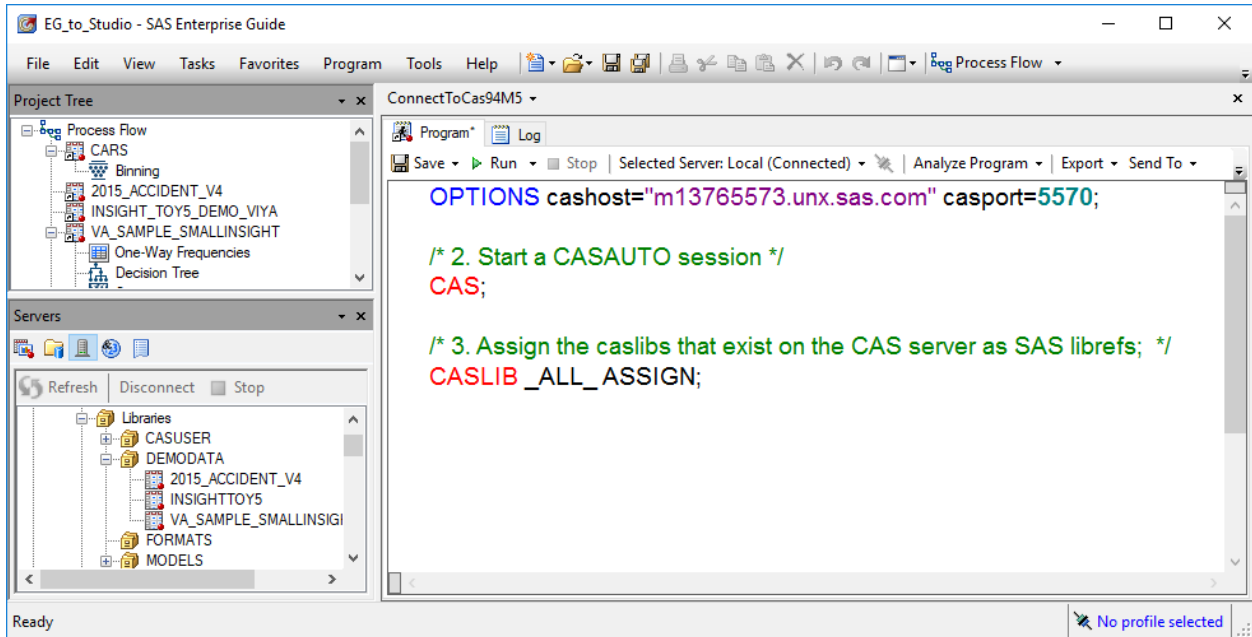


Figure 7. Connecting to CAS

Executing SAS Programs in CAS

A big advantage of being able to execute SAS programs in CAS is the fact that the DATA step and a large number of SAS procedures can be executed in CAS on a distributed system. This might lower execution times of your SAS programs significantly, especially if your data sizes are large. To make sure your programs execute in CAS, the procedure that you are using must be enabled in CAS and input tables and output tables must reside in a caslib. In the example below, you see two very simple DATA steps. In the first DATA step, both the input table and output table reside in a caslib. In the second DATA step, the input is coming from a SAS library.

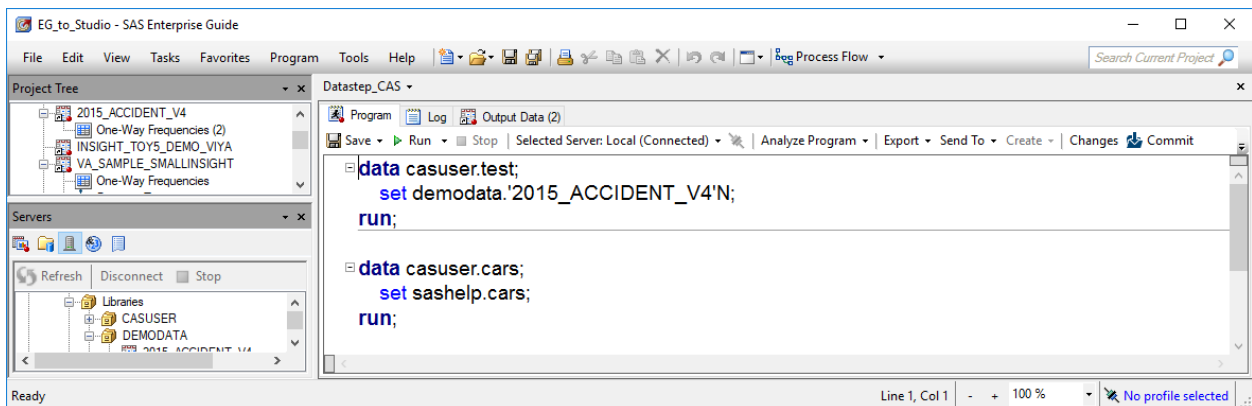


Figure 8. Executing a DATA step in CAS

If you look at the log, you can clearly see the difference: the first DATA step executes in CAS, whereas the second DATA step runs in a SAS session.

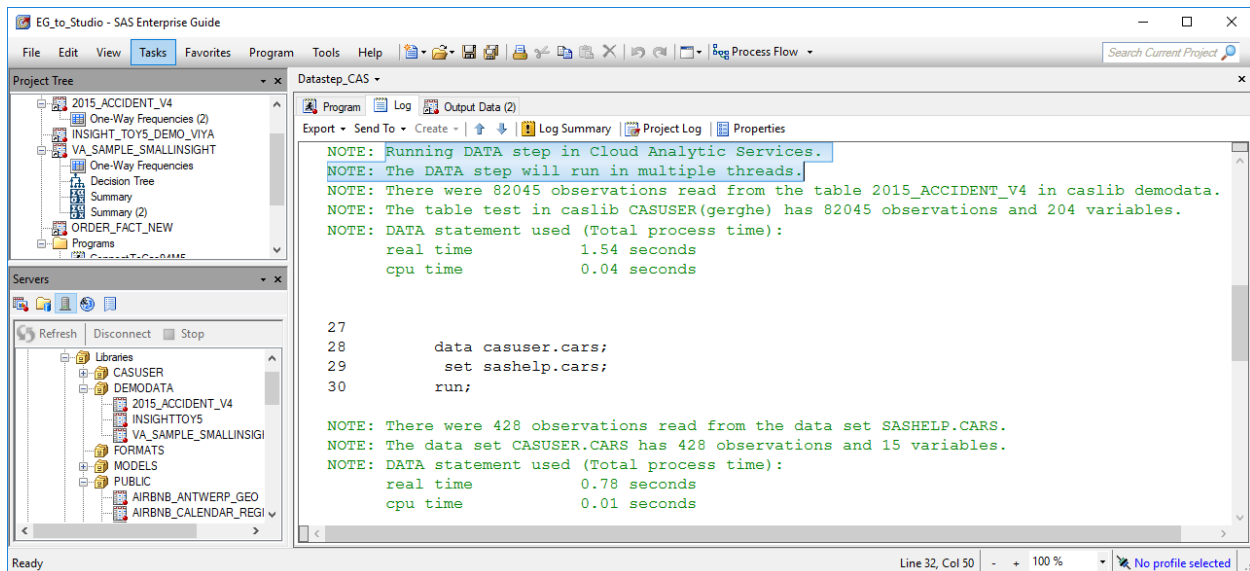


Figure 9. Examining the Log

As a best practice, upload your data to CAS first if you intend to execute DATA steps or procedures on large files. In terms of data preparation, be aware that PROC SQL is not enabled in CAS. If you intend to run SQL statements on the CAS server, you must use PROC FedSQL instead.

CONCLUSION

The need for tools that enable self-service data preparation capabilities will grow. SAS Visual Analytics on SAS Viya provides an easy-to-use interface for report authors and data scientists to access and prepare data without intervention from IT. The seamless integration of existing SAS programs enables existing customers to move to SAS Visual Analytics on SAS Viya and to benefit from new capabilities for analysis and visualization in a unified HTML5 interface.

RECOMMENDED READING

- Hazejager, Wilbram. 2018. "Data Management in SAS Viya." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC. SAS Institute Inc.

CONTACT INFORMATION

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

Gregor Herrmann
 SAS Institute Inc.
 SAS Campus Drive
 Cary, NC, 27513
gregor.herrmann@sas.com

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.

Ten Tips to Unlock the Power of Hadoop with SAS®

Wilbram Hazejager and Nancy Rausch, SAS Institute Inc.

ABSTRACT

This paper discusses a set of practical recommendations for optimizing the performance and scalability of your Hadoop system using SAS®. Topics include recommendations gleaned from actual deployments from a variety of implementations and distributions. Techniques cover tips for improving performance and working with complex Hadoop technologies such as YARN, techniques for improving efficiency when working with data, methods to better leverage the SAS in Hadoop components, and other recommendations. With this information, you can unlock the power of SAS in your Hadoop system.

INTRODUCTION

When traditional data storage or computational technologies struggle to provide either the storage or computation power required to work with large amounts of data, an organization is said to have a big data issue. Big data is frequently defined as the point at which the volume, velocity, or variety of data exceeds an organization's storage or computation capacity for accurate and timely decision-making.

The most significant new technology trend that has emerged for working with big data is Apache Hadoop. Hadoop is an open-source set of technologies that provide a simple, distributed storage system paired with a fault-tolerant parallel-processing approach that is well suited to commodity hardware. Many organizations have incorporated Hadoop into their enterprise, leveraging the ability for Hadoop to process and analyze large volumes of data at low cost.

SAS® has extensive integration options with Hadoop to bring the power of SAS to help address big data challenges. SAS, via SAS/ACCESS® technologies and SAS® In-Database Code Accelerator products, has been optimized to push down computation and augment native Hadoop capabilities to bring the power of SAS to the data stored in Hadoop. By reducing data movement, processing times decrease and users are able to more efficiently use compute resources and database systems.

The following recommendations describe some of the best practices to help you make the most of your SAS and Hadoop integration.

TIP #1: USING YARN QUEUES

Even if you have a large Hadoop cluster, resources are not unlimited, and all users have to share those resources. When you want regular SAS processes to have priority over long-running queries or certain other activities in your Hadoop environment, or perhaps you want to prevent your SAS processes from consuming too much of your Hadoop resources, then it's time to look at Apache Hadoop YARN (Yet Another Resource Negotiator).

YARN is the Hadoop cluster resource management system, whatever type of processing framework you are using (MapReduce, Spark, or Tez). The benefits promised by YARN are scalability, availability, resource-optimized utilization, and multi-tenancy.

A Hadoop administrator can define so-called YARN queues, and each queue has a set of associated resource settings. These settings specify minimum and maximum values for things like memory, virtual CPU cores, and so on. The Hadoop administrator can specify so-called placement policies that specify which default queue to use for users or groups of users.

The default queue for a user is used by SAS software unless your SAS application overrules this default. If you do override the default in your SAS application, the Hadoop administrator needs to have enabled queue overwrites; otherwise, the default queue for the user is used.

For more details about Apache Hadoop YARN, see <http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>.

SAS/ACCESS® Interface to Hadoop, when leveraging Hive, generates MapReduce code, and SAS Embedded Process runs purely as MapReduce jobs. Therefore, it can be completely managed by YARN.

When writing your own SAS code, you can modify the contents of the `mapred-site.xml` that is in the directory pointed to by the `SAS_HADOOP_CONFIG_PATH` environment variable and specify the MapReduce queue name to use. Here is an example:

```
<property>
  <name>mapreduce.job.queueName</name>
  <value>root.SASqueue</value>
</property>
```

Once you have done this, start your SAS session. From now on, whenever SAS code generates MapReduce jobs, this specific YARN queue is used.

You can also specify the YARN queue to use on your SAS LIBNAME statement as follows:

```
libname mydata HADOOP HOST='xxxx.yyyy.com' PORT=10000
  PROPERTIES="mapreduce.job.queueName=root.SASqueue";
```

See the SAS/ACCESS documentation for more details about the PROPERTIES option in the LIBNAME statement.

When using SAS® Data Loader 3.1 for Hadoop, the YARN queue to use is extracted from the `mapred-site.xml` Hadoop configuration file that is available on the SAS Data Loader middle tier. Whenever SAS Data Loader initiates processing in the Hadoop cluster, it specifies that queue. If no queue was defined in the Hadoop configuration file, then SAS Data Loader falls back to using queue default, which means that the (user-specific) default queue is used.

TIP #2: WORKING WITH HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

When working with Hive data in Hadoop, the actual content of the table is stored in the Hadoop Distributed File System (HDFS) as a file or set of files. The HDFS has a permissions model for files and directories that shares some things in common with a UNIX style POSIX model, where files and directories are owned by an owner and a group, and permissions are managed through those users.

In contrast to the POSIX model, there are no `setuid` or `setgid` bits for files. Instead, the sticky bit can be set on directories, which will prevent anyone except the superuser, directory owner, or file owner from deleting or moving files within the directory.

In SAS, this might manifest as a table that cannot be dropped or replaced by any user except the owner of the table. Multiple users can see, read, and open the table, but when they attempt to replace that table as a target, they will get a run-time error. Note that most of the vendor distributions are deployed with the sticky bit set.

SAS recommends that you turn off the sticky bit on either the `/tmp` directory or wherever the `HDFS_TMPDIR` is pointing to. The reason for this is because Work libraries will try to write, update, or delete temporary files in those locations. The permissions on the directory should be like this: `drwxrwxrwx`. To change the permissions using the command-line interface, as the HDFS superuser (usually `hdfs`), run a command similar to the following:

```
hadoop fs -chmod 0777 /tmp
```

HDFS HOME DIRECTORIES

Each user using the Hadoop cluster must have an HDFS home directory configured on each node in the cluster. Hadoop places files in that directory for some Hive operations. Also, because SAS Data Loader uses Oozie to run some types of jobs in Hadoop, including jobs that load data using Sqoop and Spark, it stores some temporary files in that directory.

Administrators should plan to create a user home directory and Hadoop staging directory in HDFS for each user. The user home directory in HDFS is typically `/user/myuser`. The Hadoop staging directory is controlled by the setting `yarn.app.mapreduce.am.staging-dir` in `mapred-site.xml` and defaults to `/user/myuser`. Change the permissions and owner of `/user/myuser` to match the UNIX user. The user ID must have at least the following permissions:

- Read, Write, and Delete permission for files in the HDFS directory (used for Oozie jobs)
- Read, Write, and Delete permission for tables in Hive

TIP #3: HIGH AVAILABILITY SUPPORT

In Hadoop, the HDFS is the primary storage system and is responsible for storing and serving all data. The HDFS has long been considered a highly reliable system. However, the HDFS has always had a well-known single point of failure because it relies on a single name node to coordinate access to the file system data. Sometimes, an HDFS outage can impact users. For this reason, you might want to consider adding high availability (HA) to the HDFS name node. You do this by adding a second name node in an active/passive configuration, so that if one goes down, the other can compensate. See your distribution vendor documentation for more details about how to configure an HA name node.

SAS supports HA for the HDFS for some products starting with the third maintenance release for SAS 9.4 and certain recent Hadoop vendor distributions. Some hotfixes might be required to support HA in some cases. HA support is available for a number of products, including the following:

- Base® SAS: FILENAME Statement for Hadoop Access Method (See **Limitation 1**)
- Base SAS: HADOOP Procedure (See **Limitation 1**)
- SAS/ACCESS Interface to Hadoop (See **Limitation 2**)
- SAS In-Database Code Accelerator for Hadoop

The expected behavior when enabled is the following:

- Functionality and communication with Hadoop are maintained when HDFS is configured for name node HA.
- In the case of a name node failover, SAS jobs in progress roll over to using the secondary name node, data processing continues, and there is no need to manually update client configuration.

Here are some known limitations:

Limitation 1: On Linux x64, with the connection to Hadoop orchestrated via the WebHDFS RESTFUL API (option `SAS_HADOOP_RESTFUL` set to 1), the FILENAME statement and HADOOP procedure might become unresponsive if the primary name node suffers a catastrophic failure. Restart of the SAS session might be required after the failover takes place.

Limitation 2: Performance of SAS/ACCESS Interface to Hadoop query might be affected if the primary name node suffers a catastrophic failure while the query is in progress. Queries submitted after the failover takes place are not affected.

A fix is also available through SAS Technical Support for SAS Data Loader to enable Hive HA. Similar to HDFS HA, Hive HA allows you to configure an active/passive configuration for the Hive server. Normally, SAS Data Loader generates JDBC connections using the form `jdbc:hive2://<hive-server>:10000`, which will fail in HA configurations if the Hive server is down.

With the fix in place, SAS Data Loader instead generates code similar to the following JDBC connect string if Hive HA information is found in the `hive-site.xml`:

```
jdbc:hive2://<hive.zookeeper.quorum>/serviceDiscoveryMode=zookeeper;zooKeeperNamespace=<hive.server2.zookeeper.namespace>
```

This instructs the JDBC driver to dynamically find a live Hive server. LIBNAME generation is also updated for this use case. Note that the value specified for SERVER= in the LIBNAME statement is ignored when explicitly pointing to Hive HA using the URI= option.

Support for generating Hive HA connections in SAS Data Loader can be disabled by setting the following advanced property for the “Data Loader Mid-Tier Shared Services” entry in SAS® Management Console:

```
sasdm.disable.hiveserver2.discovery
```

TIP #4: HIVE OPTIMIZATIONS – PART I – JOINS, FETCH TASK, FILE FORMATS

JOINS

When working with data, especially big data, it is best to avoid moving data around between different computing systems. Wherever possible, you want to move your data once, and then perform your joins and other transformations.

You can use PROC SQL to perform joins on your data in Hadoop. To ensure that your join keeps the data in Hadoop to perform the join, here are a few tips:

- When using SAS PROC SQL, SAS does not pass LIBNAME-referenced cross-schema joins to Hadoop. To pass a multiple-libref join to Hadoop, the schemas for each LIBNAME statement must be identical.

If you want to perform a cross-schema join with PROC SQL, you can use the SQL pass-through facility instead, for example:

```
proc sql;
  connect to hadoop (user="myusr1" pw="mypwd"
    server=hxpduped port=10000 schema=default);
```

- Another suggestion is to use SAS Data Loader for Hadoop. SAS Data Loader is a multi-lingual code generator for Hadoop, and it is designed to automatically select the best Hadoop language to generate based on the type of transformation you want to do. For joins, SAS Data Loader generates the Hive SQL syntax to perform joins in Hadoop. This can guarantee that your data does not move around because it is using the native features of Hadoop to work with data. Figure 1 shows an example of Hive SQL generated by SAS Data Loader.

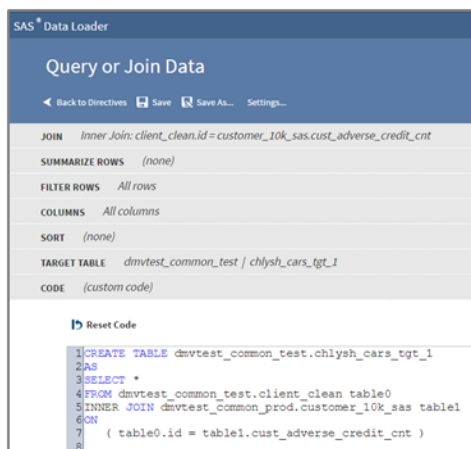


Figure 1. Example of Hive SQL Generated by SAS Data Loader

- Avoid the use of user-defined formats.
- Avoid the use of Hadoop invalid column names in your source or target.
- Some data set-related options such as DBNULL and DBTYPE will prevent pushdown.

- Function references where there is not a similar function available in Hadoop prevents pushdown; this is particularly common when using date and time type functions.

FETCH TASK

When working with smaller data sets in Hive, one option to speed up query performance is to consider using the fetch task. This option directly queries the data to give the result, rather than starting a MapReduce job for the incoming query. For simple queries like `select *`, it can be very fast, typically seconds instead of minutes because Hive returns the results by performing an HDFS get operation.

This feature is enabled via the following Hive configuration option:

```
hive.fetch.task.conversion = minimal;
```

In Hive 14, the fetch task is normally on by default in most distributions. When this Hive property is set, the Hive engine uses the action only for specific queries where it makes sense, like `select *`, and generates MapReduce jobs for other type of queries, when fetch would not be efficient. Statistics have to be enabled in the cluster for these settings to apply. For example, use the following code:

```
set hive.stats.autogather=true;
set hive.stats.dbclass=fs;
```

The fetch task works only on certain types of queries. The query has to be from a single data source without subqueries, aggregations, or views. It does not apply to joins, only queries. There is also a threshold value that determines the maximum size of the table that can be used with this technique. Above this threshold, Hive uses MapReduce instead. This is controlled via another setting:

```
hive.fetch.task.conversion.threshold = 1000000000;
```

In most distributions, the default is set to 1GB.

The Hive options can be set:

- at the Hadoop cluster level at the Hive server configuration level
- at the SAS level in the `hive-site.xml` connection file
- at the LIBNAME level with the `PROPERTIES` option; for example:

```
PROPERTIES="hive.fetch.task.conversion=minimal;
hive.fetch.task.conversion.threshold=-1";
```

For example, when a simple PROC SQL is submitted, without the property, two MapReduce jobs run, as shown in Figure 2.

application 1459332331704_0035	hpauser1	CREATE TABLE sasdata_05_27_26_138_00...TXT_1(Stage-1)	MAPREDUCE	default	Fri, 01 Apr 2016 09:27:27 GMT	Fri, 01 Apr 2016 09:30:45 GMT	FINISHED	SUCCEEDED
application 1459332331704_0034	hpauser1	SELECT * FROM `MEGACORP30M` (Stage-1)	MAPREDUCE	default	Fri, 01 Apr 2016 09:19:03 GMT	Fri, 01 Apr 2016 09:27:24 GMT	FINISHED	SUCCEEDED

Figure 2. Example of Job Status in Hadoop

When a simple PROC SQL is submitted with the fetch task enabled and the correct type of data query, only the MapReduce job corresponding to the actual SQL query runs, as shown in Figure 3.

application 1459332331704_0036	hpauser1	CREATE TABLE sasdata_05_34_13_512_00...TXT_1(Stage-1)	MAPREDUCE	default	Fri, 01 Apr 2016 09:34:14 GMT	Fri, 01 Apr 2016 09:36:41 GMT	FINISHED	SUCCEEDED
--------------------------------	----------	--	-----------	---------	-------------------------------	-------------------------------	----------	-----------

Figure 3. Example of a MapReduce Job Status

Because this is a Hadoop setting, this feature is transparent to SAS. This means that once it is set in your cluster, SAS just takes advantage of the feature. You don't have to write any special code to use it.

FILE FORMATS

Hadoop has various formats that can be used to store data in the HDFS. Each file type has pros and cons, and there are many factors that determine what file type is best for your specific usage scenario, such as your usage pattern, your Hadoop vendor and Hadoop version, and your data. Here is a list of the Hadoop native file formats supported by SAS:

- Delimited: This is the default type, which depends on the `hive.default.fileformat` configuration property
- SequenceFile
- RCFile: Available in Hive 0.6.0 and later
- ORC: Available in Hive 0.11.0 and later
- Parquet: Available in Hive 0.13.0 and later
- Avro: Available in Hive 0.14.0 and later

By default, SAS generates Delimited files when writing to Hadoop. Over time, SAS has introduced support for other file types in Hadoop including RCFile, ORC, and Parquet. When using SAS with Hive, Avro is supported as well.

You can use the `DBCREATE_TABLE_OPTS` option to specify the file type for your output table. This option is available in both a SAS `LIBNAME` statement and as a SAS data set option. For example, to have all your output tables in a specific SAS Hadoop library stored using the ORC file format, use the following statement:

```
libname mydata HADOOP ... DBCREATE_TABLE_OPTS="stored as ORC";
```

To have a specific table stored using the ORC file format:

```
data mydata.table_out (DBCREATE_TABLE_OPTS="stored as ORC");  
set mydata.table_in;  
run;
```

Some of the file types mentioned above take advantage of compression, which can reduce disk space. However, keep in mind that compression comes with a cost, especially when writing data. Both ORC and Parquet store data in columnar format, which can provide an advantage when reading only a subset of columns of a wide table or a subset of the rows. For usage patterns where you write the data once and then read it multiple times, like for reporting or analytics, this performance benefit for reading might outweigh the slower performance for writing.

Avro stores its schema as part of its metadata, which allows you to read the file differently from how you write the file. This is useful in cases where you are doing a lot of data transmissions. However, we have found during our testing that when copying a large data set into Avro format in Hive, a number of large temporary files get created in Hive. For example, an 8.6 GB table resulted in around 25 GB of persistent storage in Hive and 45 GB of temporary files. This is something to consider when moving data into and out of Avro format.

To quantify the performance advantages of the various file formats in Hadoop, we performed a series of simple performance tests. Note that this was a simple test, and your results might vary from our findings. The test environment we used was built using three servers and was performed as single user tests with no other workload. The data tables and environment were not tuned in any special way; tuning would more likely improve performance. We performed LOAD, READ, filtered READ, and JOIN tests using Hive/Parquet, Hive/SequenceFile, Hive/Avro, and Hive/ORC.

Figure 4 shows our findings using a 50 GB table:

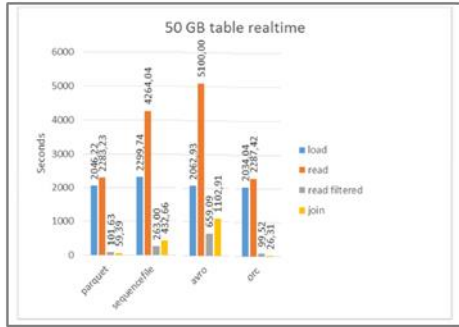


Figure 4. Performance Characteristics of Hadoop File Types

For larger tables, the ORC and Parquet formats provided the best overall performance for the data sizes and tests we performed. These two formats also have a compression rate of about a factor of two. Avro tables increase in size as table size increases, and SequenceFile format stays about the same (for example, no compression).

As always, your mileage and usage patterns might vary, so you will have to test your own usage patterns using your own data to see what performs best for you.

TIP #5: HIVE OPTIMIZATIONS – PART II – SAS® DATA INTEGRATION STUDIO RECOMMENDATIONS

SAS Data Integration Studio supports integration with Hadoop in a variety of ways. You can create your own Hive and PIG transforms, and you can run your own MapReduce jobs. In the fourth maintenance release for 9.4, you can now run SAS Data Loader jobs from SAS Data Integration Studio. Figure 5 shows the various Hadoop transformations available:

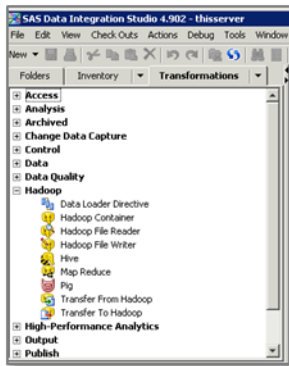


Figure 5. Hadoop Transformation available in SAS Data Integration Studio

A new transformation allows you to select which saved SAS Data Loader jobs you want to run. One advantage of this new feature is to support integrated impact analysis. You can now see impact analysis across both SAS and Hadoop environments. Figure 6 and Figure 7 illustrate some of these new features.

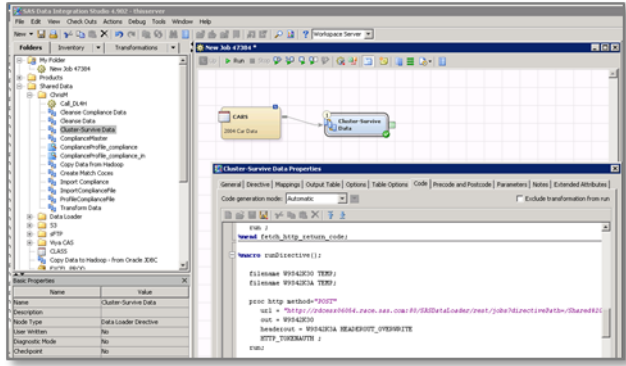


Figure 6. SAS Data Loader Transform Example

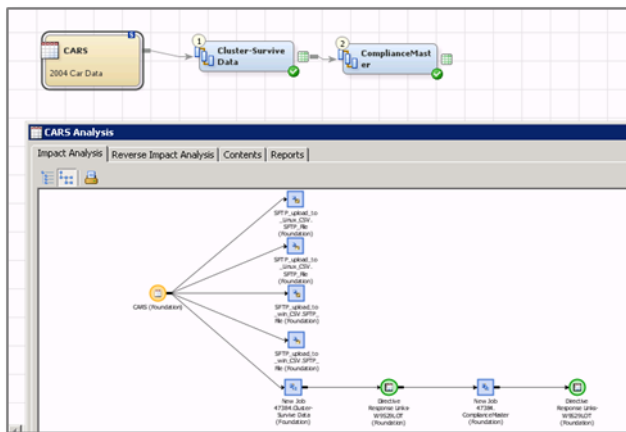


Figure 7. Impact Analysis Example

One tip when working with impact analysis: if you want to trace column level impact, go into the mapping tab on the SAS Data Loader transform and create the mappings to represent how data flows into and out of the system. This is illustrated in Figure 8:

#	Column	Column Description	Table	#	Column	Column Description
1	SALES	Retail sales in millio...	RETAIL (RETAIL)	1	href	
2	DATE		RETAIL (RETAIL)	2	method	
3	YEAR		RETAIL (RETAIL)	3	ordinal_links	
4	MONTH		RETAIL (RETAIL)	4	ordinal_root	
5	DAY		RETAIL (RETAIL)	5	rel	
				6	uri	

Figure 8. Mappings in the SAS Data Loader Transform

Several enhancements have been made to better ensure that the code generated in SAS Data Integration Studio pushes down to the Hadoop database. One useful setting allows you to disable the generation of column formats. This is important because some formats cannot be expressed in the Hadoop execution environment, and data would transfer unnecessarily between the two environments. By setting this option in SAS Data Integration Studio, you can avoid this extra data transfer. Figure 9 illustrates how to set this option on new and existing jobs.

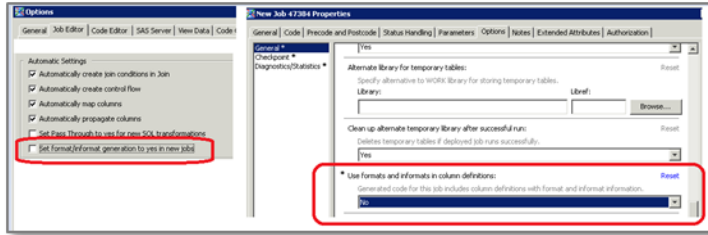


Figure 9. Option to Disable Column Formats

One debugging tip is to set MSGLEVEL=I in your code and look at the log. SAS documents where the code is run. Figure 10 is an example of the log output:

```

26      if _N_=1;
INFO: DATA Step contains subsetting-if.
27      if UnitReliability >0.95 then score=0; Else score=1;
28      run;
INFO: Could not run DATA Step in HADOOP.

```

Figure 10. MSGLEVEL Example

Another enhancement is available via SAS Technical Support when working with the Hive transformations in SAS Data Integration Studio. Figure 11 is an example of the Hive transformation:

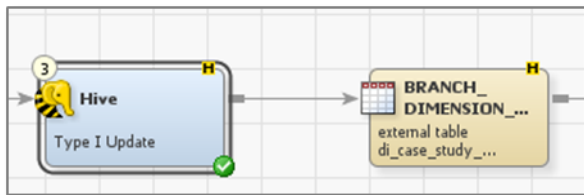


Figure 11. Hive Transform

An option has been added to the transform to allow you to turn on or off the DROP and CREATE statement, and the generated code has been modified to generate INSERT OVERWRITE instead of CREATE TABLE syntax. A partial list of the new code that gets generated is shown below:

```

INSERT OVERWRITE TABLE myschema.branch_dimension_new
/* Insert Unchanged Rows from Application Layer */
SELECT app.BRANCH_SK, app.BRANCH_TYPE_DESC, app.BRANCH_NUMBER,
       app.BRANCH_NAME, app.STREET_STATE_CODE, app.STREET_POSTAL_CODE,
       app.LOAD_DTTM, app.LAST_UPD_DTTM
FROM myschema.branch_dimension app
... other statements

```

Several changes have been made to the SAS Data Integration Studio SQL Join transform. One change now correctly handles generation of Hive SQL syntax when joining tables from multiple schemas to prefix the table name with the schema name. This is shown in the code below:

```

proc sql;
  connect to HADOOP
    (SERVER=pocserver PORT=10000);
  execute
  (
    INSERT INTO TABLE myschema.branch_dimension_updates
    SELECT
      ... other statements...
  )

```

```
) by HADOOP;
```

The join has also been updated to generate the pass through INSERT OVERWRITE for replace and INSERT INTO syntax for append when working with Hive data.

TIP #6: HIVE OPTIMIZATIONS – PART III – PARTITIONING

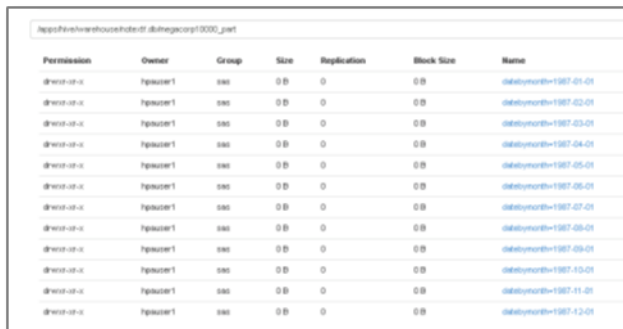
Data partitioning is useful to help improve the performances of queries. Hive supports partitioned data through the use of HCatalog (also referred to as HCat). This is a table management layer that exposes Hive metadata to other Hadoop applications.

Partitioning data in Hive can improve query performance, and it is recommended for low-cardinality variables. For example, you could partition a table of sales data by month, where each month is stored as a separate partition in Hadoop. Then, if you query with a WHERE clause based on the month, Hadoop will read only the data in the partition associated with that month.

The same option DBCREATE_TABLE_OPTS allows you to specify a PARTITION key, for example:

```
proc sql;
  create table myschema2.table_part
    (DBCREATE_TABLE_OPTS="PARTITIONED BY (datebymonth date)")
  as select * from myschema1.table_unpartitioned;
quit;
```

This code creates one file for each month in the HDFS, as shown in Figure 12:



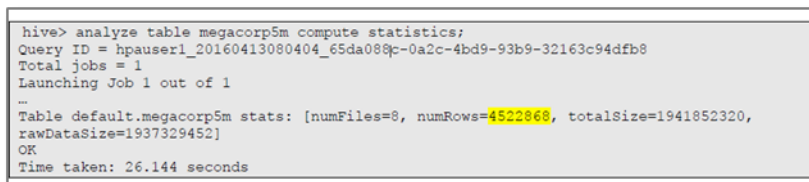
Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-01-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-02-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-03-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-04-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-05-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-06-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-07-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-08-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-09-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-10-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-11-01
drwxr-xr-x	hpauser1	sas	0B	0	0B	datebymonth=1987-12-01

Figure 12. Example of HDFS Partitions

TIP #7: HIVE OPTIMIZATIONS – PART IV – ENABLING HIVE STATISTICS

Statistics such as the number of rows of a table are important in Hive. One of the key uses is for query optimization. Statistics serve as the input to the cost functions of the optimizer so that it can compare different plans and choose among them.

Recent versions of Hive can store table statistics in the Hive metastore. You get this for free when a table is loaded, unless your administrator has turned table statistics off. A Hive analyze command can be run to compute table statistics at any time, as shown in Figure 13:



```
hive> analyze table megacorp5m compute statistics;
Query ID = hpauser1_20160413080404_65da089c-0a2c-4bd9-93b9-32163c94dfb8
Total jobs = 1
Launching Job 1 out of 1
...
Table default.megacorp5m stats: [numFiles=8, numRows=4522868, totalSize=1941852320,
rawDataSize=1937329452]
OK
Time taken: 26.144 seconds
```

Figure 13. Hive Analyze Command

You might want to consider running analyze commands in batch mode to keep your statistics updated and to benefit from this pre-computed information in your queries. Alternatively, you might consider using specific formats as ORC, Parquet, Avro, and so on, which are natively storing these types of aggregates.

When statistics are available, SAS uses them to determine the fastest way to transfer data to and from Hive. For small tables, SAS avoids the MapReduce job and uses direct fetch if SAS can determine information about table size from the statistics.

TIP #8: HIVE OPTIMIZATIONS – PART V – MAPREDUCE VERSUS TEZ

When running complex SQL queries using Hive, potentially many different MapReduce jobs get created that cause sub-optimal performance. Hadoop vendors are taking different approaches to trying to solve this. If your Hadoop vendor supports Hive running on Apache Tez, and your cluster has been configured to support it, then data transformations generated by SAS can take advantage of it.

Tez is an extensible framework for building high-performance batch and interactive data processing applications, coordinated by YARN in Apache Hadoop. Tez improves the MapReduce paradigm by dramatically improving its speed, while maintaining MapReduce's ability to scale to petabytes of data. When using Hortonworks Data Platform, Hive embeds Tez so that it can translate complex SQL statements into highly optimized, purpose-built data processing graphs that strike the right balance between performance, throughput, and scalability.

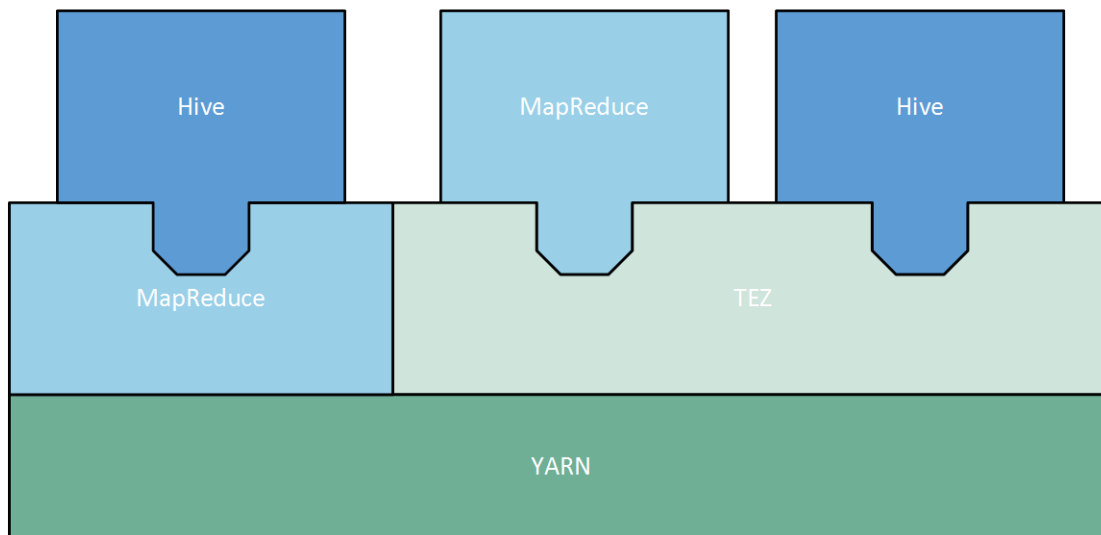


Figure 14. Hive on Tez

From a SAS perspective, if Tez is available in the Hadoop cluster, then it is possible to choose between a traditional MapReduce engine and Tez before submitting a SAS program that generates Hive SQL under the covers. See tip #2 in last year's SAS Global Forum paper *Ten Tips to Unlock the Power of Hadoop with SAS®* for more information about how to get details in the SAS log about what goes on under the covers.

By default, the engine set in the Hive server configuration (hive-site.xml) is used. You can overrule this default by using the PROPERTIES option in the LIBNAME statement. It allows you to explicitly choose the engine from the SAS client. Here is an example:

```
libname mydata HADOOP HOST='xxxx.yyyy.com' PORT=10000
        PROPERTIES="hive.execution.engine=tez";
```

Customers have seen considerable performance gains when using SAS procedures that push SQL into Hive, like SQL, SUMMARY, and FREQ. See the SAS In-Database Technologies for Hadoop documentation for more details about which Base SAS procedures support processing inside Hadoop. But, as always, your mileage might vary, so you will have to test your own usage patterns using your own data.

TIP #9: SUPPORT FOR MULTIPLE SASAPP IN SAS DATA LOADER

When SAS Data Loader 3.1 was introduced in the fourth maintenance release for 9.4, the architecture changed from a single user vApp (virtual application) deployment to a centralized deployment using SAS® Intelligence Platform. This means that infrastructure components used by SAS Data Loader are defined in SAS metadata and managed using SAS Management Console. As SAS Data Loader interacts with many different parts of the Hadoop ecosystem, either directly from the SAS Data Loader middle tier or indirectly via a SAS® Workspace Server, you now see definitions for various Hadoop components in SAS metadata. This could include Hive server, Impala server, or Oozie server.

Out of the box, SAS Data Loader 3.1 supports one SAS Workspace Server, and the workspace server to use is defined at deployment time. When working in a large organization where SAS projects are strictly separated, where each project has its own SAS Workspace Server and associated SAS LIBNAME definitions, and SAS metadata access rights are defined such that a user can only work with the SAS Workspace Servers and SAS LIBNAME definitions if they are member of the project, you might need to have multiple SAS Data Loader deployments.

SAS has enhanced the functionality of SAS Data Loader 3.1 to support multiple SAS Workspace Servers using a single deployment of SAS Data Loader. All of the SAS Workspace Servers to be used by SAS Data Loader need to be configured using the exact same set of Hadoop client JARS and the same set of Hadoop configuration files (*-site.xml). This feature is available by contacting your SAS administrator.

A typical setup would be as follows:

For each project:

- Create a dedicated schema in Hive to store project-specific data.
- Create a dedicated SAS Application Server environment with a SAS Workspace Server and make sure the name represents the project. See *SAS 9.4 Intelligence Platform – Application Server Administration Guide* for more details about defining multiple application servers.

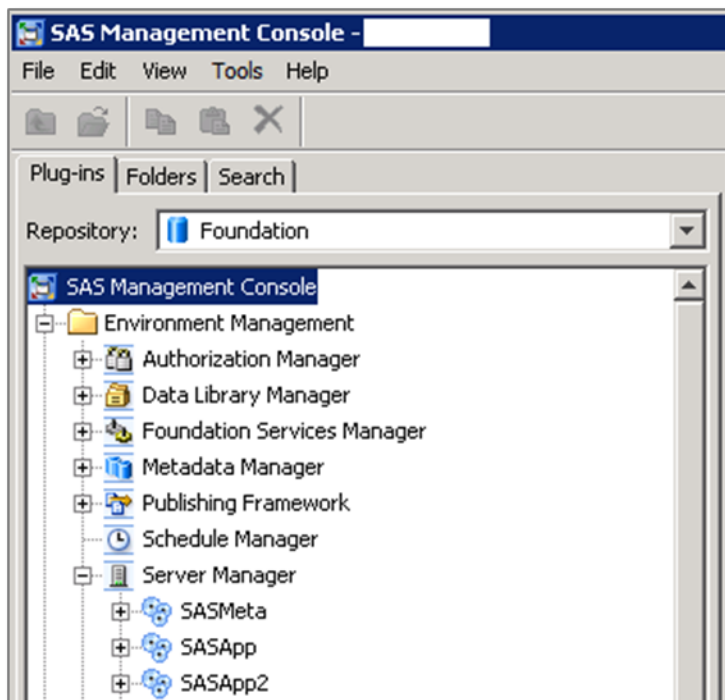


Figure 15. SAS Management Console-Server Manager–Showing Multiple SASApp Environments

Figure 15 shows that two SAS Application Server environments were deployed, SASApp and SASApp2, each representing a project.

- Using SAS Management Console
 - Create SAS LIBNAME definitions to represent source data locations and target data locations for the project.
 - Associate these SAS LIBNAME definitions with the corresponding SAS Workspace Server.
 - Note that a SAS LIBNAME definition can be associated with multiple SAS Workspace Server definitions (for example, to represent that multiple projects share a source data location).
 - Grant the users ReadMetadata access to the SAS Application Server, SAS Workspace Server, and SAS LIBNAME definitions for that project.
 - Note that users can be part of multiple projects.

In SAS Data Loader, the user can specify which of the SAS Workspace Servers to use using the Configuration menu. The list of available workspace servers contains only the servers that the user has been granted ReadMetadata access to.

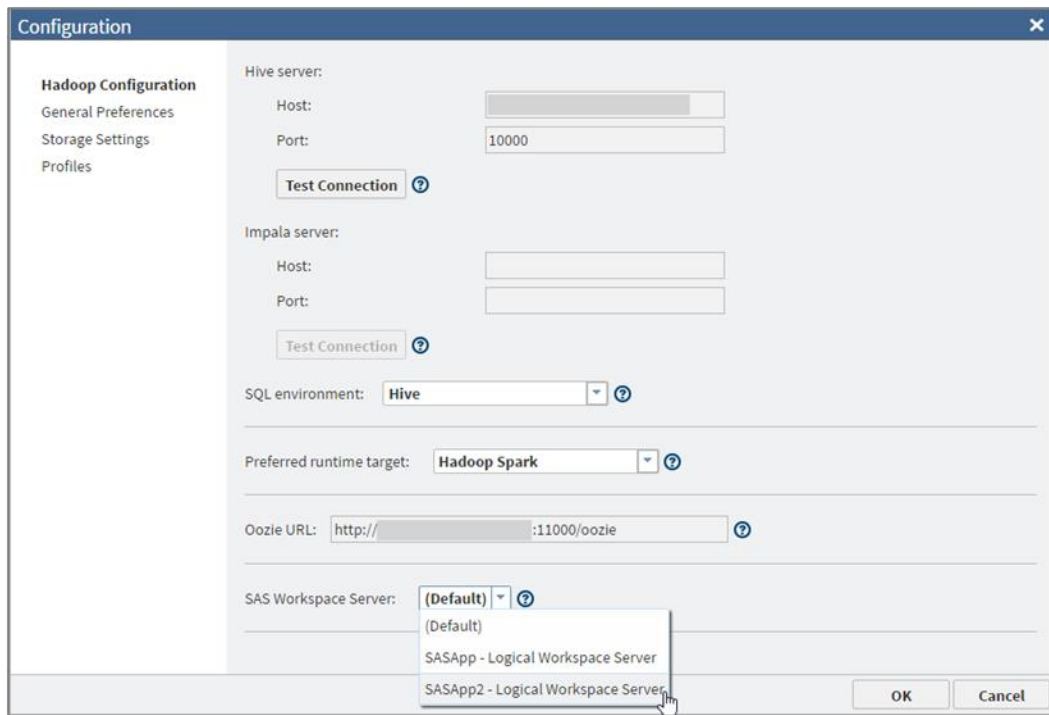


Figure 16. SAS Data Loader-Select SAS Workspace Server to Use

Note that (Default) means: revert to out-of-the-box behavior; therefore, use the SAS Workspace Server that was configured using SAS Management Console as the default to use for all SAS Data Loader sessions.

The SAS Workspace Server selected will be used whenever you use SAS Data Loader until you change your selection in the Configuration menu again.

Disclaimer: The described functionality in this section is not available in a standard SAS Data Loader 3.1 deployment. You need to contact your local SAS administrator or SAS Technical Support to download a fix that provides this functionality.

TIP #10: DEPLOYMENT TIPS AND TRICKS

The fourth maintenance release for SAS 9.4 significantly improved the SAS deployment experience when deploying SAS into Hadoop and collecting files for SAS/ACCESS and SAS Data Loader. One of the main improvements was to move SAS Data Loader into the SAS® Deployment Wizard, and to consolidate the various pieces to all deploy under the umbrella of the SAS® Deployment Manager. Many hours were spent validating the deployment instructions to ensure that they were correct for the many different configurations of Hadoop distributions and SAS configurations.

This first and most important tip in this section is to refer to the updated SAS deployment guides when deploying SAS and Hadoop, as they capture many recommendations and best practices. This section is short because most of our recommendations have been captured in the deployment guides. However, here are a few additional tips:

- Deployment requires steps from both the SAS and the Hadoop administrators. Review the pre-install checklist because there are steps outlined for both roles prior to starting any installation or upgrade.
- Remember that you need to collect new configuration files if you upgrade your Hadoop installation, you install a new Hadoop parcel, package, service, or component on an existing cluster, or you make any other type of major configuration change.
- You need to regenerate the Hadoop SAS Embedded Process configuration file (/sas/ep/config/ep-config.xml) when you collect new JAR files, you install or upgrade Hive or HCat, you upgrade the JDK or JRE that is used by the Hadoop processes, or any other significant change to the Hadoop cluster. In some cases, the SAS Deployment Manager can now help with this step.
- The JAR collection process requires Hadoop services running in order to identify the JAR files to collect. Occasionally, the JAR collection process under- or over-collects based on your cluster and what is running on it. If you run into issues, particularly in SAS Data Loader, when attempting to first connect or run jobs, this could be the reason. Review the deployment guide and then contact your SAS administrator.

TIP #11: BONUS

Hadoop is a complex system with many settings. Hadoop vendors provide recommendations for many configuration settings based on size, in terms of data nodes, of the cluster. The documentation contains tables for different ranges of number of nodes and for each range, there is a list of values for specific configuration settings.

It's important to keep in mind that these are intended as general guidelines and are not carved in stone. Especially if the size of your cluster is close to the upper-limit of a size range, we have seen in multiple client situations that increasing the values to the recommended settings for the next size up often resolves most (if not all) stability issues and causes certain performance issues to disappear as well.

Another thing to be aware of is that when running concurrent SAS jobs, each with multiple librefs to HiveServer2, each of these librefs keep connections open for the duration of the SAS job. We have seen jobs like this fail intermittently with connection errors being shown in the SAS log. This is often caused by low (default) values for options like `hive.server2.thrift.max.worker.threads`. Increasing the value of this parameter often resolves the intermittent connection issues. In this scenario, ZooKeeper errors might show up as well as each HiveServer2 client has a connection to ZooKeeper. The ZooKeeper `maxClientCnxns` is often set to a low default value as well, and when experiencing intermittent connection errors, increasing the value of this option might help.

TIPS FOR WORKING WITH SENTRY

Here are a few tips if you want to get Hadoop Sentry working with SAS:

1. Log on to Hive via Beeline, and authenticate using an ID member of a group defined as one of the admin groups for Sentry.

2. Create a role for Sentry admins (for example, CREATE ROLE ADMINS).
3. Grant membership to the role to a group your ID is member of (for example, GRANT ROLE ADMINS TO GROUP hive).
4. Grant the admin role the ability to execute grant or evoke commands at the server level (for example, GRANT ALL ON SERVER SERVER1 TO ROLE ADMINS). Note that SERVER1 is the default name used at the system level to recognize Sentry. Because the name is configurable, if the name is different from the default, use that name instead of SERVER1.
5. Create one or more additional roles for SAS users (for example, CREATE ROLE SASUSERS).
6. Grant membership to the SAS user roles to one or more OS groups (for example, GRANT ROLE SASUSERS TO GROUP 'r&d').
7. Grant at least SELECT privilege on a database to the SAS user roles to allow for the LIBNAME statement to connect successfully (for example, GRANT ALL ON DATABASE DEFAULT TO ROLE SASUSERS).
8. For the SAS/ACCESS engine to successfully create and load tables, grant Read and Write access to the temporary folder to the roles created for the SAS users:

```
/* without HA */
GRANT ALL ON URI 'hdfs://<name node>:8020/tmp' TO ROLE SASUSERS;
/* with HA */
GRANT ALL ON URI 'hdfs://nameservice1:8020/tmp' TO ROLE SASUSERS;
```

If a folder different from /tmp is being used, that folder needs to be specified in place of the default one. Same for the HDFS default port number. With HA, nameservice1 is the default name used by the system to identify the name node. Should a site use a name different from the default, that name must be used when running the grant.

9. For the SAS Embedded Process to successfully work with Sentry, grant Read and Write access to the temporary folder to the roles created for the SAS users:

```
/* without HA */
GRANT ALL ON URI 'file:///tmp' to ROLE SASUSERS;
/* with HA */
GRANT ALL ON URI 'hdfs:///tmp' to ROLE SASUSERS;
```

10. To make sure SAS Embedded Process can successfully load Hive tables, do the following:

If the default temporary folder is being used, create an OS group called supergroup on the name node, and connect the user hive to it. Define the user first in case the ID doesn't exist on the server. For example, for Linux:

```
groupadd supergroup
useradd/usermod -G supergroup hive
```

If a folder different from /tmp is being used to store temporary files, give ownership to the folder to the hive user, either at the user or at the group level.

Also, make sure that the permission bits are set to 777 and that the sticky bit is not set.

CONCLUSION

This paper discusses a set of practical recommendations for optimizing the performance and scalability of your Hadoop system using SAS. Topics include recommendations gleaned from actual deployments from a variety of implementations and distributions. Techniques cover tips for improving performance and working with complex Hadoop technologies such as YARN, techniques for improving efficiency when working with data, methods to better leverage the SAS in Hadoop components, and other recommendations. With this information, you can unlock the power of SAS in your Hadoop system.

ACKNOWLEDGMENTS

The authors wish to thank Alex Arey, Blake Russo, Chris Watson, Clark Bradley, Mauro Cazzari, and Raphaël Poumarede for their collective insights.

RECOMMENDED READING

- Hazejager, W., et al., 2016. “Ten Tips to Unlock the Power of Hadoop with SAS”. *Proceedings of the SAS Global Forum 2016 Conference*, Cary, NC. SAS Institute. Available at <http://support.sas.com/resources/papers/proceedings16/SAS2560-2016.pdf>
- Poumarede, R., 2016 “SAS with Hadoop: Performance considerations and monitoring strategies”. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/resources/thirdpartysupport/v94/hadoop/sas-hadoop-performance-strategies.pdf>
- Rausch, Nancy, et al. 2016. “What’s New in SAS Data Management.” *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/resources/papers/proceedings16/SAS2400-2016.pdf>
- Rausch, Nancy, et al. 2017. “What’s New in SAS Data Management.” *Proceedings of the SAS Global Forum 2017 Conference*. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/resources/papers/proceedings17/SAS195-2017.pdf>
- SAS® 9.4 Support for Hadoop website, available at <http://support.sas.com/resources/thirdpartysupport/v94/hadoop/>
- SAS® Data Management Community, Available at https://communities.sas.com/t5/Data-Management/ct-p/data_management
- SAS Institute Inc. 2016. *SAS® 9.4 In-Database Products User’s Guide, Seventh Edition*. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/documentation/cdl/en/indbug/69750/PDF/default/indbug.pdf>
- SAS Institute Inc. 2015. *SAS/ACCESS® 9.4 for Relational Databases: Reference, Ninth Edition*. Cary, NC: SAS Institute Inc. Available at <http://support.sas.com/documentation/cdl/en/acreldb/69580/PDF/default/acreldb.pdf>

CONTACT INFORMATION

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

Wilbram Hazejager
100 SAS Campus Dr
Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
Wilbram.Hazejager@sas.com
support.sas.com

Nancy Rausch
100 SAS Campus Dr
Cary, NC 27513
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444
Nancy.Rausch@sas.com
support.sas.com

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.

Enable Personal Data Governance for Sustainable Compliance

Vincent Rejany; Bogdan Teleuca, SAS Institute Inc. Cary, NC

ABSTRACT

In the context of European Union's General Data Protection Regulation (EU GDPR), one of the main challenges for data controllers and data processors is to demonstrate compliance by documenting all their data processing activities and where appropriate, to assess the risk of these processes for the individuals. Such requirements cannot be achieved without being able to build an efficient data governance program combining Legal driven top-down activities through personal data compliance and IT driven bottom-up operations through personal data mapping including personal data categories definition and discovery.

We use several processes developed in SAS® Data Management Studio to identify the personal data and update the governance view within SAS® Business Data Network and SAS® Lineage. We demonstrate several features in other products such as the Personal Data Discovery Dashboard in SAS® Visual Analytics, and SAS® Personal Data Compliance Manager as it applies to Records of Processing Activities and the Data Protection Impact Assessment.

INTRODUCTION

Data governance is not an old concept; at SAS we have been pitching data governance benefits for years. However, it is often seen as something that is nice to have, even though it is a recognized method for mitigating risk, increasing operational efficiency, and enabling innovation.

As of this writing, we are close to the deadline to implement the requirements brought by the European Union's new General Data Protection Regulation (GDPR). On 25 May 2018, not only will GDPR be enforced across the European Union, but it will also have a global impact on all organizations that deal with the information of EU citizens.

The objective of the regulation is to give citizens more control over their data and to create a uniform set of rules to enforce across the continent. The main priority for organizations will be to show accountability by regaining control of their data processes, especially the processes and reasons for collecting, processing, updating, archiving, and deleting personal data records. To achieve such a task, being able to size the effort and discover the type and location of personal data is essential. Such a perspective is a key element for addressing data protection impact assessment when one process represents a high risk to the rights and freedoms of individuals.

GDPR breathes data governance and calls for discipline, integrity, and trust. "In the middle of difficulty lies opportunities" said Albert Einstein, so GDPR should be embraced as an opportunity to create value for your business, to gain a competitive edge, to innovate, to reinvent the way you manage your customer relationship, and to start doing more with personal data. Knowing the information that you hold, its quality, the reason that you hold it, and the length of time you can retain it is key in terms of operational efficiency. Organizations that support this idea of transparency will gain one competitive advantage by differentiating themselves in the market.

WHAT IS PERSONAL DATA?

Personal data is any information that enables one person to be identified, directly or indirectly, by reference to an identifier such as a name, an identification number, location data, an online identifier, but also to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person. Different pieces of information, which collected together can lead to the identification of a person, also constitute personal data. These are examples of personal data:

- name and surname
- home address

- email address such as name.surname@company.com
- identification card number such as VISA, American Express, or a loyalty card
- location data
- network identifiers such as IP addresses, even if they are dynamic
- cookie ID
- advertising identifier of your phone

Personal data that has been rendered anonymous in such a way that the individual is not or is no longer identifiable is no longer considered personal data. For data to be truly anonymized, the anonymization must be irreversible.

The regulation also defines the concept of special categories of data, for which specific safeguards and requirements are specified, such as a higher level of consent. These special categories relate to personal data that are “particularly sensitive in relation to fundamental rights and freedoms” and, therefore, “merit specific protection.” These categories include data “revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade-union membership, and the processing of genetic data, biometric data to uniquely identify a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation.”

UNDERSTANDING THE NEED FOR PERSONAL DATA GOVERNANCE

How can you comply with GDPR? By applying **Good Data PR**actices! The question is not how to comply, but how to be compliant and remain compliant. The aim of data protection regulations such as GDPR is to change behaviors and mindsets. Taking that perspective, the accountability principle (in Article 5 of the GDPR) makes the data controller to be the one responsible for demonstrating compliance with these GDPR principles:

- Lawfulness, fairness, and transparency must exist in processes that manage personal data.
- Limitation of purpose. Personal data must be collected for specified, explicit, and legitimate purposes.
- Data minimization. There should be no reason to use more data than necessary for the defined purpose.
- Accuracy. Data quality must be ensured and personal data be kept up-to-date.
- Storage limitation. Personal data must be processed for no longer than is necessary.
- Integrity and confidentiality. Appropriate security measures must be taken.

So how do you prove and show accountability? Under GDPR, accountability can be proven by nominating a data protection officer, drafting your privacy notice, and responding to requests (such as access requests) from individuals. However, the main action that you must take is to document internally all your processing activities, and to make this documentation available to supervisory authorities upon request. This “record of processing activities” is required by GDPR Article 30, and will facilitate the compliance with the other principles.

Data controllers must also carry out data protection impact assessments (DPIAs) when data processes could represent a high risk to individuals’ rights and freedoms, particularly when new technologies are involved. The DPIA is required by Article 35 of the GDPR, and contains information about how a new or modified application might affect the privacy of personal information processed by or stored within the application.

Remember that any large organization has hundreds of systems, data assets, and processing activities, and thousands of personal data types to review daily, weekly, or monthly. Describing these items is a significant effort, but maintaining an up-to-date view of them is even more time-consuming and is prone to errors.

For data professionals (such as data owners, data stewards, and data controllers), the typical manual or semi-automatic steps no longer stand a chance when facing GDPR requirements.

SAS APPROACH FOR PERSONAL DATA GOVERNANCE

The challenges that companies face in complying with GDPR are both external (toward supervisory authorities) and internal. When supervisory authorities perform a review of a company's data protection status, they will require a company-wide overview of all data sources and processes. And data sources aren't just systems or databases – they include network drives, files on PCs, and everywhere else personal data can be stored!

To address this challenge, there is one fair and straightforward method: "Say what you do and do what you say," which matches the classic data governance top-down and bottom-up analysis. Personal data governance (Figure 1) embeds both personal data compliance and mapping efforts.

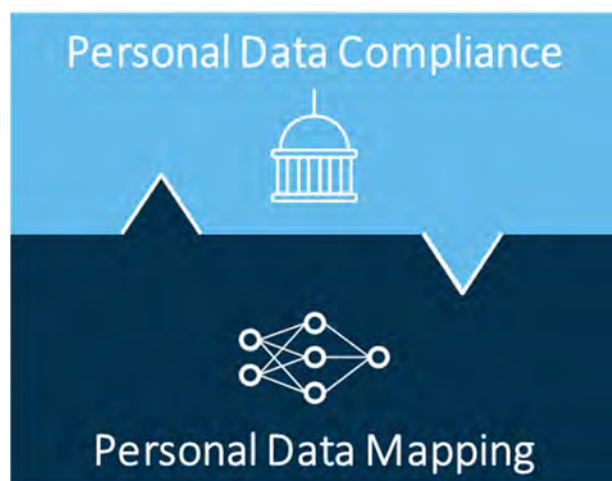


Figure 1. Personal Data Governance

Personal data compliance aims at addressing the legal requirements, mentioned previously, to document data sources, to record processing activities, and to list the potential risks of these activities.

Understanding data processes outside of IT is critical to capturing risks and potential control gaps. Such activities require advanced capabilities in the areas of structured methodologies shared across the organization, versioning, workflow for process management of approvals and reviews, and notifications. Because organizations can have thousands of data sources and hundreds of processes, relying on spreadsheets is not an option.

Personal data mapping is an approach that is proposed to facilitate the governance efforts and to significantly reduce the amount of time and effort needed to have the latest view of personal data. The location of personal data is essential information, to be as exhaustive as possible in your documentation and to show to the supervisory authority that you have established the processes needed to handle personal data. Moreover, recording the locations where personal data is stored will help your organization easily locate the information when an individual exercises his or her rights.

PERSONAL DATA COMPLIANCE

SAS® PERSONAL DATA COMPLIANCE MANAGER

SAS® Personal Data Compliance Manager is a recently released product based on the new SAS® Risk Governance Framework. SAS® Personal Data Compliance Manager is a workflow-driven and regulator-facing solution that automates the management of governance, risk, and compliance data. The product facilitates the entry, collection, transfer, storage, tracking, and reporting of operational losses, gains, and recoveries that are drawn from multiple locations across an organization.

The goal of SAS® Personal Data Compliance Manager is to provide organizations with customizable templates and workflows to document personal data processes and assess the risk of these processes.

The software has been developed based on GDPR requirements, as well as supervisory authorities' guidelines such as Working Party 29 (WP29), the CNIL (France), ICO (UK) and CPP (Belgium). SAS® Personal Data Compliance Manager is not specific to GDPR, and intends to address all data protection regulations, which are often strongly like the European law.

The product is intended to work in conjunction with the SAS® Data Management components and SAS® Visual Analytics. Together they provide a technology platform for participating firms to deal with the EU regulation on personal data protection

In this first release, SAS® Personal Data Compliance Manager can also be used to perform these tasks:

- document and maintain data processing activities
- define data controllers, processors, and data subject categories
- conduct data protection impact assessment
- describe and maintain data assets and systems
- define controls and security measures
- manage incidents, data breaches, and data subject correspondence

RECORDS OF PROCESSING ACTIVITIES

Legal Background

Data processing activity is defined in Article 4 (definitions 2 and 6) of the GDPR. Processing covers a wide range of operations performed on personal data, including both manual and automated processes. It includes the collection, recording, organization, structuring, storage, adaptation, alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, and erasure or destruction of personal data. The records shall be in writing, including in electronic form.

According to Article 30, the documentation of processing activities must include the following information:

- the name and contact details of the controller and, where applicable, the joint controller, the controller's representative, and the data protection officer;
- the purposes of the processing. In appendix, Table 1 lists some examples of activities, as recently provided by the Belgian supervisory authority.
- the basis of the processing that could be
 - necessary for the performance of a contract
 - A legal obligation
 - protection of the vital interests of the data subject
 - A task carried out in the public interest or in the exercise of official authority
 - legitimate interests pursued by the controller or by a third party
 - a result of data subject consent
- a description of the categories of data subjects and of the categories of personal data
- the categories of recipients to whom the personal data have been or will be disclosed including recipients in third countries or international organizations
- where applicable, transfers of personal data to a third country or an international organization, including the identification of that third country or international organization and, in the case of certain transfers the documentation of suitable safeguards
- where possible, the envisaged time limits for erasure of the different categories of data
- where possible, a general description of the technical and organizational security measures referred to in GDPR Article 32(1)

Recording data processing activities with SAS® Personal Data Compliance Manager Systems and Data Assets Definition

Within SAS® Personal Data Compliance Manager, the recording of processing activities starts with the definition of systems and data assets. One system is a general identifier for one application, such as ERP, Finance, CRM. One data asset is more precise and allows to differentiate software from file system or databases. One processing activity can cover more than one data asset, and one data asset can of course be linked to more than one processing. Display 1 illustrates the creation of one data asset “Finance”.

The screenshot displays the SAS Personal Data Compliance Manager interface for creating a data asset named "Finance". At the top, a workflow diagram shows the steps: Create (highlighted), Edit, Review, DPO Review, and Production. Below the workflow, the "Status" is set to "Create".

The main form is divided into several sections:

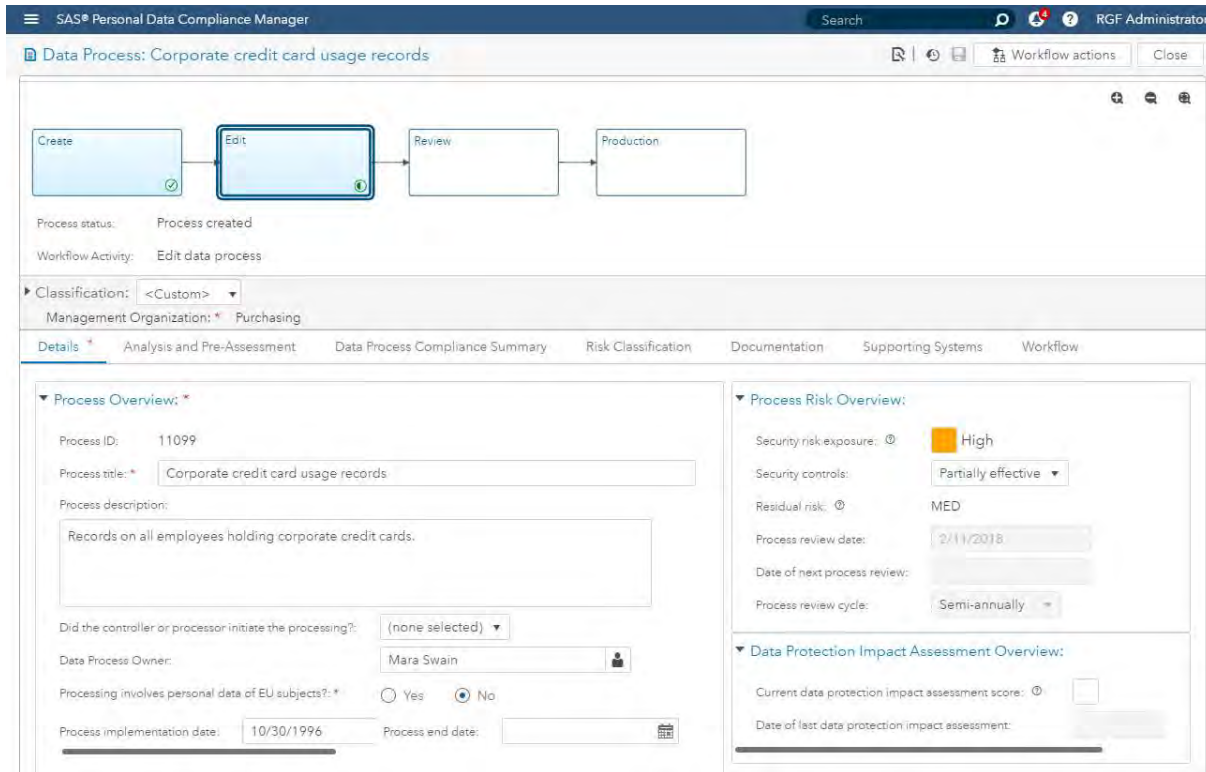
- Classification:** A dropdown menu set to "<None>".
- Management Organization:** A field with an asterisk indicating it is required.
- Personal Data Type:** A field with an asterisk indicating it is required.
- Details:** This section is expanded and contains:
 - Asset ID:** 10784
 - Asset Name:** Finance
 - Asset Description:** Finance System
 - Asset Type:** Software
 - Asset Status:** Active
 - Asset Usage:** A large empty text area.
 - Asset Implementation Date:** A date picker.
 - Asset Decommissioning Date:** A date picker.
- Asset Review Participants:** This section shows the "Asset Owner" as Andrew Dalton and a "Data Controller" table. The table has columns for "Actions", "User Name", "User ID", and "Title", but it is currently empty with "No results" displayed below it.
- Asset Review and Current Risk Ratings Information:** This section includes fields for:
 - Asset review date:** A date picker.
 - Review Cycle:** A dropdown menu set to "(none select)".
 - Next Review Date:** A date picker.
 - Last Review Date:** A date picker.
 - CIA Risk Rating:** A field with a help icon.
 - Residual Risk:** A field with a help icon.

Display 1. Personal Data Compliance – Data Asset Definition

Processing activities definition

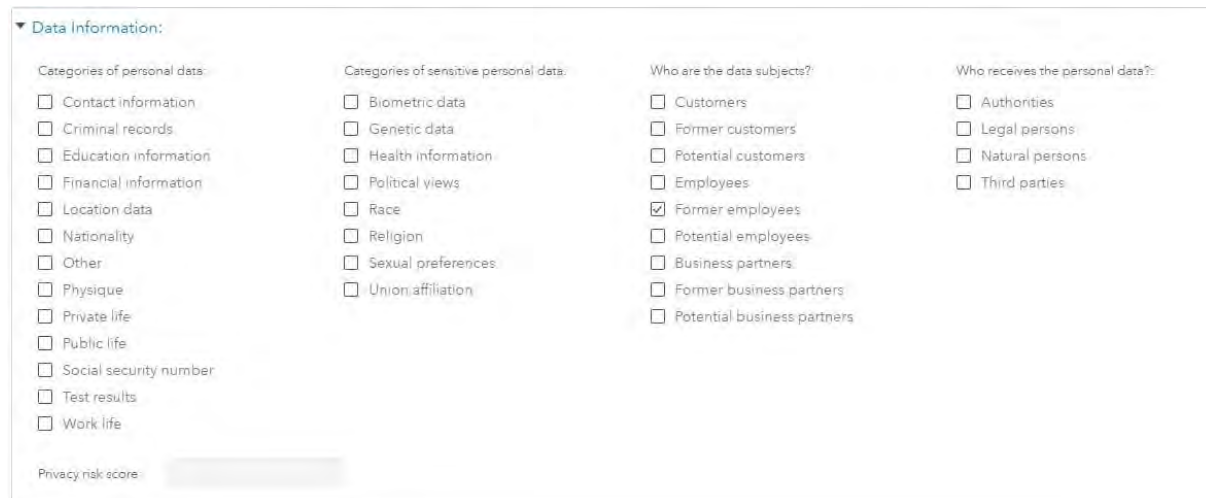
SAS® Personal Data Compliance Manager supports the recording of processing activities through one dedicated web form. By default, the definition of processing activities is workflow driven so each step can be validated by approved users. One processing activity can be linked to one or more data assets.

The different sections have been defined based on the Article 30 of EU GDPR and the recommendations of various European supervisory authorities such as the ICO (UK), CNIL (France) and CPP (Belgium CBPL/CPVP). Display 2 shows the first tab of the form with the details about the processing activity.



Display 2. Personal Data Compliance – Processing Activity Definition

One of the most important part of the processing activity creation is the specification of the personal data categories involved in the processing as well as the data subject categories concerned. Display 3 presents the related section within SAS® Personal Data Compliance Manager Data Process.



Display 3. Personal Data Compliance – Processing Activity - Data Information details

DATA PROTECTION IMPACT ASSESSMENT

Legal background

Data Protection Impact Assessment (DPIA), also known as Privacy Impact Assessment (PIA) is one the specific process required by the GDPR. Impact assessments are not new, as similar risk assessments are also required by ISO/IEC 27001. DPIAs aim at identifying the risks related to the use of personal within one or several processing activities by evaluating them versus the GDPR principles mentioned formerly. For each risk identified safeguards and security measures must be defined. Article 35 of EU GDPR defines three conditions for which one DPIA must be conducted:

- When evaluating a natural person using automated processing (including profiling) to make decisions or have legal impacts on the subject. The use of new technologies, i.e. Big Data or Artificial Intelligence over personal data, is typically that situation
- When processing large quantities of special categories of data, or personal data relating to criminal convictions and offences.
- When systematically monitoring a publicly accessible area on a large scale, i.e. CCTV

According to the UK Information Commissioner (ICO) one DPIA should contain:

- A description of the processing operations and the purposes, including, where applicable, the legitimate interests pursued by the controller.
- An assessment of the necessity and proportionality of the processing in relation to the purpose.
- An assessment of the risks to individuals.
- The measures in place to address risk, including security and to demonstrate that you comply.

Assessing data protection risk with SAS® Personal Data Compliance Manager

Within SAS® Personal Data Compliance Manager, one DPIA pre- assessment is required over each processing activity. This pre- assessment is key as the absence of DPIA would have to be justified to the supervisory authority in case of an audit.

Information collection has been structured through three tabs and is workflow driven. The first tab “Details” in Display 4 provides a view of the processing activities involved in the DPIA as well as an overview of the Privacy Risk Assessment. Depending on the risk severity of one assessment, regular reviews will be required. SAS® Personal Data Compliance Manager also supports the ability to define and schedule notifications for ensuring such review.

The screenshot displays the SAS Personal Data Compliance Manager interface for defining a Data Protection Impact Assessment (DPIA). The window title is "SAS® Personal Data Compliance Manager" and the user is "RGF Administrator". The main content area is titled "Data Protection Impact Assessment: Work/Life Balance Initiatives".

Classification: <Custom>
Management Organization: * HR

Details * | Data Protection Impact Assessment | Documentation

Basic Information: *

- Data Protection Impact Assessment ID: 10163
- Title: * Work/Life Balance Initiatives
- Description: Analysis of employee data to direct workplace initiatives designed to improve employee work/life balance
- Date of current assessment: * 2/11/2018
- Data Controller: Enter User...

Privacy Risk Assessment:

- Unauthorized erasure of data: Low
- Incorrect data: Low
- Unauthorized transfer or distribution of confidential data: Low
- Overall inherent risk: Low

Objects Under Assessment:

Related Data Processes

Actions	Process ID	Process title	Security risk exposure	Date of next process review
<input type="checkbox"/>				

Display 4. Personal Data Compliance – DPIA Definition

The Data Protection Assessment Tab, Display 5, proposes to conduct the DPIA through a series of questions covering topics such as: Collection, Use, Retention, Sharing and Transfer, Access and Security, Data Privacy Assessment. Relying on the answers provided, the risk (inherent and residual) will be calculated by SAS® Personal Data Compliance Manager. Documentation can be attached to the DPIA.

The screenshot shows the SAS Personal Data Compliance Manager interface. The title bar reads 'SAS® Personal Data Compliance Manager' and 'Data Protection Impact Assessment: DPIA for Data Process 3.2'. The interface includes a search bar and user information 'RGF Administrator'. The main content area is titled 'Data Protection Impact Assessment' and has tabs for 'Details', 'Data Protection Impact Assessment', and 'Documentation'. The 'Data Protection Impact Assessment' tab is active, showing a form with the following sections:

- Classification:** <Custom>
- Management Organization:** HR
- Collection:**
 - Why is the personal data being collected?: [Text area]
 - Will personal data be collected from external third-party sources?: (none selected) [Dropdown]
 - Comments: [Text area]
- Use:**
 - Is the personal data relevant and necessary to the data processing purpose?: (none selected) [Dropdown]
 - Does the processing consolidate data from multiple sources into one central location?: (none selected) [Dropdown]
 - What controls are in place to protect personal data and prevent unauthorized access?: [Text area]
- Retention:** (partially visible)

Display 5. SAS® Personal Data Compliance Manager – Data Protection Impact Assessment

PERSONAL DATA MAPPING

Personal data mapping is an essential task for complying efficiently with personal data protection regulations. Under GDPR, organizations must be fully accountable for all data flows, but visualizing and mapping data flows without the appropriate tools is very challenging. It is one key activity to demonstrate accountability by proving the reliability of your documentation effort to authorities, especially toward the recording of processing activities. Efficiency is important as organizations can have thousands of data assets, avoiding manual process through automated processes for discovering and mapping personal data attributes is required.

Data mapping is at the cornerstone of data governance, and data protection regulation require higher visibility regarding data processes. Data mapping enables you to get back control of your data management activities by supporting reverse engineering of your different data assets. You can then gain valuable insights for improvements in design and privacy compliance.

PERSONAL DATA MAPPING METHODOLOGY

Four steps have been defined to support Personal Data Mapping through the SAS® for Personal Data Protection end-to-end approach. SAS® for Personal Data Protection is an accelerator containing pre-built assets to identifying, governing and protecting personal data, which relies on SAS® Data Management, an industry-leading solution built on a data quality platform that helps you improve, integrate and govern your data.

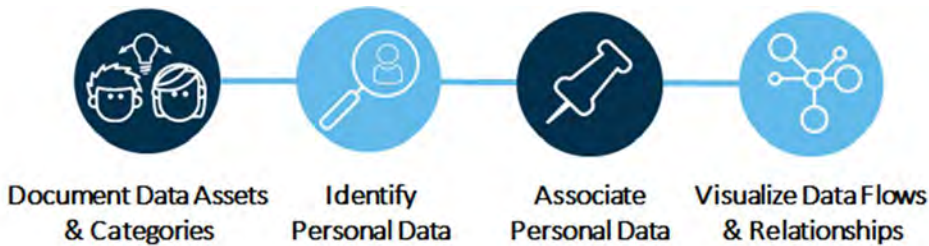


Figure 2. Personal Data Mapping Methodology

Document Data Assets & Personal Data Categories: From a risk and compliance point view this step has been fulfilled within SAS® Personal Data Compliance Manager. Therefore, these definitions are sourced from SAS® Personal Data Compliance Manager and presented into SAS® Business Data Glossary to be accessible for Business and IT people. Metadata related to these assets must be linked to applications or data owners.

Identify Personal Data: Personal data discovery is the key part of personal data mapping. It aims at finding, cataloging, and analyzing personal data attributes, across data assets. Results of the personal data discovery process are surfaced into a SAS® Visual Analytics Dashboard.

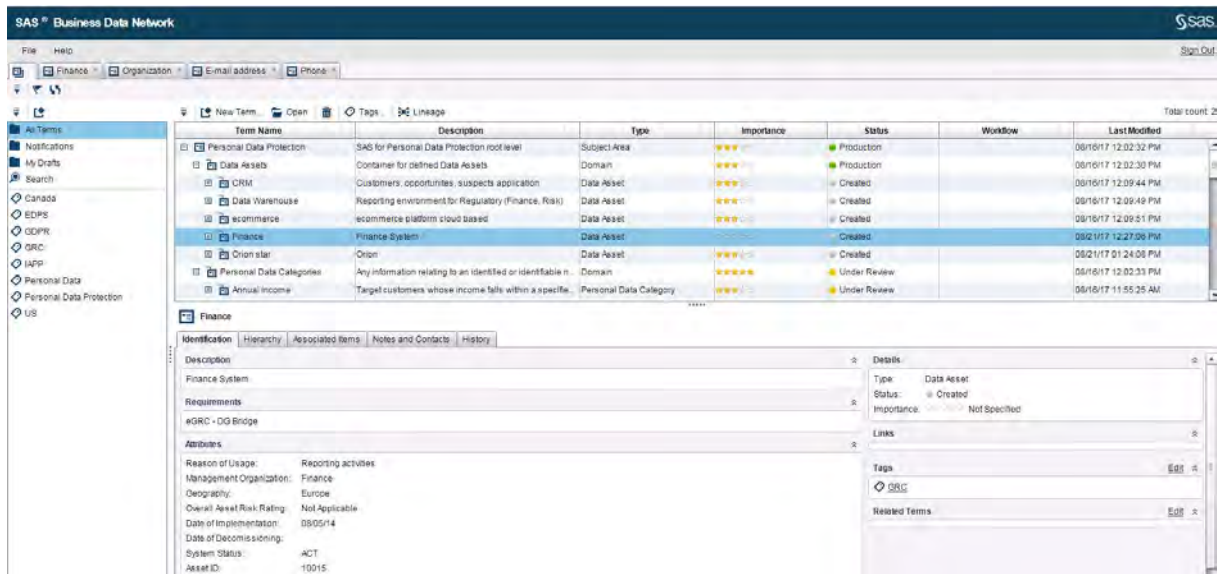
Associate & Review Personal Data: Once personal data have been identified within one data asset, related columns metadata will be automatically associated with their corresponding personal data category business term within SAS® Business Data Network. It will allow to document that such personal data attribute is present in this data asset and to get a clear mapping between metadata, personal data business terms, data assets and processes. Associations with a medium or low confidence are submitted for manual review

Visualize Data Flows & Relationships: Having centralized the definitions of the data assets, processes, data owners, personal data categories linked to databases and data jobs metadata, it now possible to surface a complete picture of the relationships between the objects. Such data flow is exposed into the processing activities definition with SAS® Personal Data Compliance Manager and the related personal data categories are provided by the personal data discovery step.

DOCUMENT DATA ASSETS & PERSONAL DATA CATEGORIES

After the initial work has been performed within SAS® Personal Data Compliance Manager, there is the need to share this perspective with a larger audience, beyond the Legal or Compliance departments. SAS® Business Data Network is a business data glossary, part of SAS® Data Management that enables collaboration between business, technical, and data steward users. SAS® Business Data Network can be used as a single-entry point for all data consumers to better understand and govern their data asset through the definition and the maintenance of business terms.

Business terms can be organized through hierarchies and relationships, and can be linked to different roles such as data or business owner or data steward. Different types of terms can be defined according to the information that needs to be documented.



Display 6: SAS® Business Data Network Main View

In the context of the personal data governance approach, seven term types have been defined and more than 700 terms related to data privacy have been already integrated into SAS® Business Data Network. Table 1, below, presents the term types implemented

Business Term Type	Description
Subject Area	Master entity identifying the subject or the project being described, such as personal data protection
Domain	Defines the second-level object types, such as data assets, processing activities, and personal data categories
Data Asset	Describes the databases or application potentially containing personal data, such as CRM, ERP, and data warehouse
Business Process	Defines business processes such as customer onboarding or marketing campaign, which is often the purpose of the processing activity
Processing Activity	Defines the processing activity including the information requested by the regulator
Personal Data Category	Contains the definition of the personal data to be identified, such as E-mail address or Network Address
Wiki	A generic term type used to enter glossary-like definitions, such as PII, consent, or data transfer

Table 1. Term Types Defined in SAS® Business Data Network

Personal Data Protection Bridge

The personal data protection bridge (PDP Bridge) is an asset developed by SAS® Data Management experts, that enables information exchange between SAS® Personal Data Compliance Manager and SAS® Business Data Network. It consists of a database and processes for extracting information to be synchronized between the two products. Objects defined within SAS® Personal Data Compliance Manager such as data assets and processing activities are automatically populated into SAS® Business Data Network and are available for use. Once the personal data categories are identified and associated with one data asset in SAS® Business Data Network, the categories are synchronized from SAS® Business Data Network to SAS® Personal Data Compliance Manager through the PDP Bridge.

The PDP Bridge uses the data export and import features available in SAS® Personal Data Compliance Manager and the SAS® Business Data Network REST API.



Figure 3: Personal Data Protection Bridge Overview

Data Asset Definition

Information about data assets can either be entered manually or imported from SAS® Personal Data Compliance Manager using the PDP Bridge. If you use either SAS® Federation Server or SAS® Data Quality as a personal data discovery engine, related connection information must be provided. This information is critical, because by default these connections are unknown in the SAS metadata.

SAS® Business Data Network

File Help

Finance *

View Edit Publish Lineage

Identification Hierarchy Associated Items Notes and Contacts History

Description

Finance System

Requirements

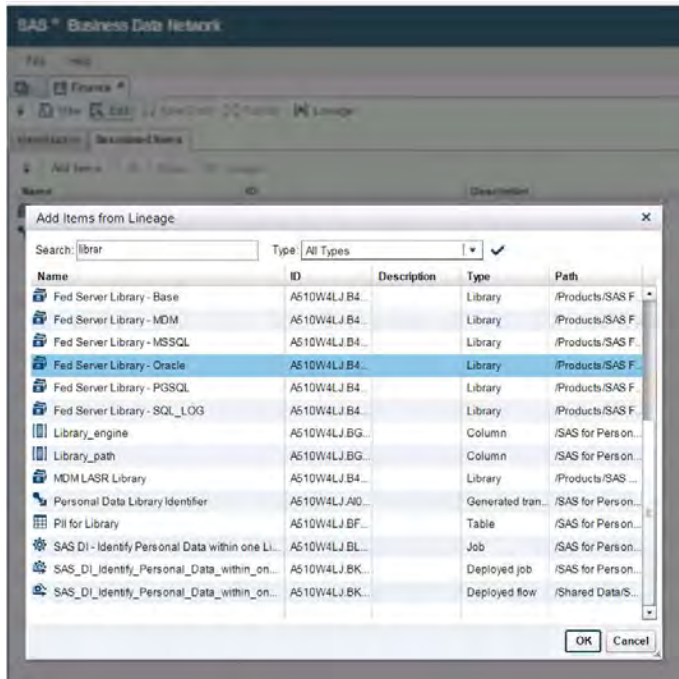
eGRC - DG Bridge

Attributes

Reason of Usage:	Reporting activities
Management Organization:	Finance
Geography:	Europe
Overall Asset Risk Rating:	Not Applicable
Date of Implementation:	08/05/14
Date of Decommissioning:	
System Status:	ACT
Asset ID:	10015
Source System:	MON
SAS Federation Server DSN:	ORACLE
SAS Data Quality DSN:	ORACLE

Display 7: Data Asset Example in SAS® Business Data Network

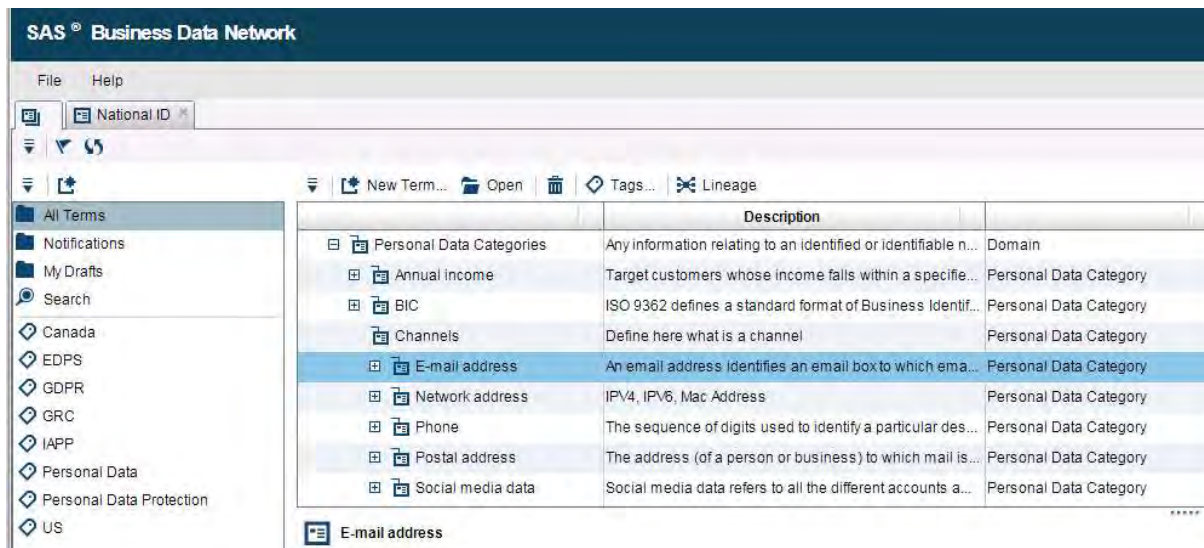
Next, you need to add the connection as a library type to the database, in the list of associated items of the term. Defined SAS libraries (Display 8) are available from SAS® Lineage, which is integrated with SAS® Business Data Network.



Display 8: SAS Library Association

Personal Data Categories Definition

A personal data category, such as **Phone**, **Payment Card Number**, **Delivery Address**, and so on, is the term that regroups the personal data that is identified in different sources. This also ensures that users understand the definitions behind the personal data categories.

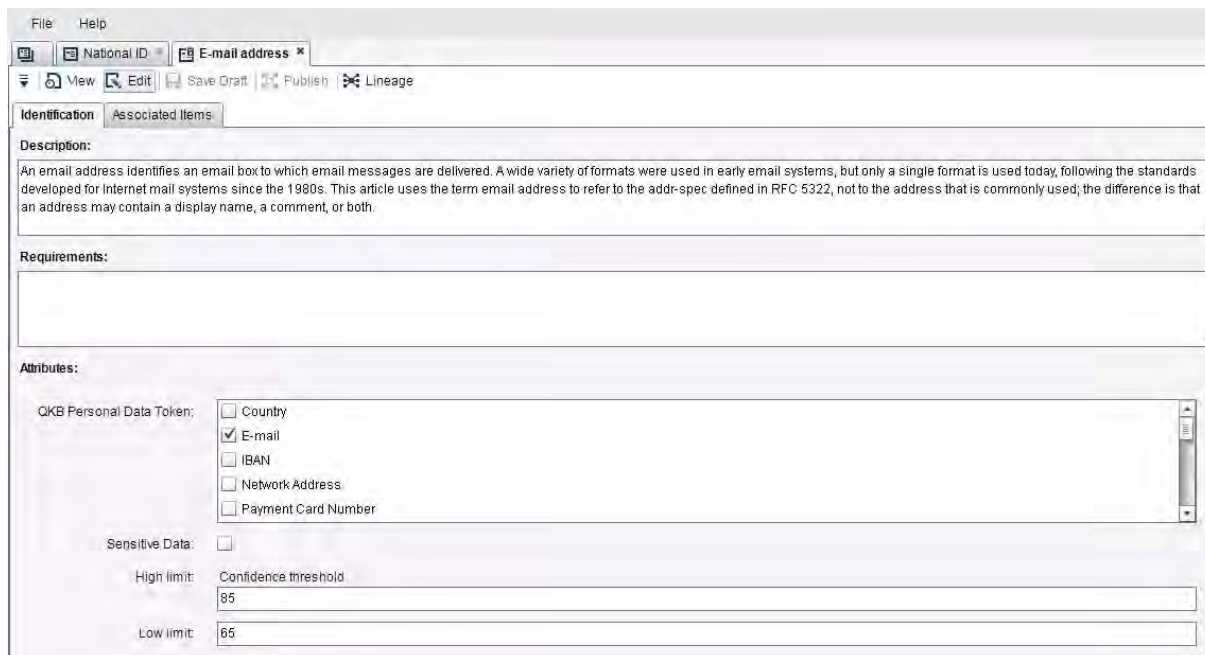


Display 9: Personal Data Category Example within SAS® Business Data Network

Firstly, you need to point all Personal Data Categories, such as an e-mail address, to a SAS® Quality Knowledge Base personal data token.

Secondly, you need to set a confidence level for the identification. The Personal Data identification uses fuzzy matching and the Quality Knowledge Base definitions. The usage of a match percentage is to link automatically all data with a match percentage over the high limit; send in a remediation queue (a manual confirmation process) all data with a match percentage between the low limit and the high limit and reject everything under the low limit. The SAS® Quality Knowledge Base is explained in the next paragraph.

You set, for example the QKB Personal Data Token to E-mail for the “E-mail address”, the High limit to: 85, the Low limit to: 65. In the section “Associate & Review Personal Data” we will see how these thresholds are used.



The screenshot shows a software window titled "E-mail address" with a menu bar (File, Help) and a toolbar (View, Edit, Save Draft, Publish, Lineage). The main content area is divided into sections: "Description" with a text box containing a definition of an email address; "Requirements" with an empty text box; and "Attributes" with a list of "QKB Personal Data Token" options: Country, E-mail (checked), IBAN, Network Address, and Payment Card Number. Below the list are checkboxes for "Sensitive Data" (unchecked), "High limit" (Confidence threshold, 85), and "Low limit" (65).

Display 10: Personal Data Category Definition

IDENTIFY PERSONAL DATA

Personal data discovery is the core part of the data mapping exercise. This process is about the categorization of personal data in structured data sources. The discovery report provides an overview of locations (databases, directories, columns etc.), types of personal data, indication of amount and other information. Unstructured data sources can also be analyzed but require another approach using text analytics. The approach is not covered in this paper.

Enable Personal Data Discovery with the SAS® Quality Knowledge Base

You can use the SAS® Quality Knowledge Base which consists of a repository that contains literally thousands of pre-built data quality rules applicable across the full range of data subject areas, including Customer and Product. The SAS® Quality Knowledge Base contains all the SAS® Data Management algorithms to standardize, identify, extract, fuzzy match data. Each locale offered by SAS has built-in data quality algorithms for typical data types (such as name and address conventions, phone standards and so forth). The algorithms vary from region to region. So far, 35 countries are currently covered.

SAS® Data Management Studio is used to identify the personal data and update the governance view in the business glossary.

For facing such a challenge in an efficient way, you rely on the capabilities present within the SAS Quality Knowledge Base for identifying personal data. One identification definition aims at categorizing one string based on vocabulary and structure analysis. Each possible token defined is score against the input string and the best score is returned. Figure 4 illustrates this principle.

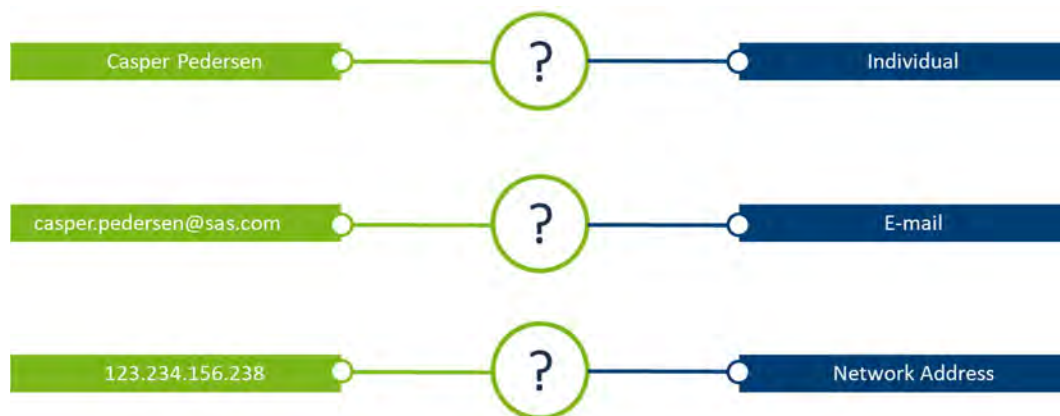


Figure 4: Personal Data Discovery

Thanks to the work already done by SAS, we are already able to identify more than 15 types of personal data categories in multiple languages. Depending on the country and language this list can contain additional personal data categories.

Table 2 presents the list of attributes identifiable in the context of the Personal Data Protection initiative.

Personal Data Tokens	
Country, Country (Iso2), Country (Iso3)	Individual
Date, Date/Time	Organization
E-Mail	Phone
Geographical Point	Postal Code
IBAN	National Id
Network Address	Vehicle Registration
Payment Card Number	Identity Card Number
URL	Passport Number
City	Tax ID
Delivery Address	

Table 2. PDP – Personal Data (Core) Identification Definition Tokens

The SAS Quality Knowledge Base features are available in multiple SAS products such Base SAS®, SAS® Data Quality, SAS® Data Integration Studio, and SAS® Federation Server. It is also available in Hadoop, Teradata, and SAS® Event Stream Processing.

Each locale can be extended to meet the needs of any organization by using the customize function of SAS® Data Management Studio to enhance or add algorithms as necessary. SAS works closely with its international offices to constantly refine existing locales and develop new versions

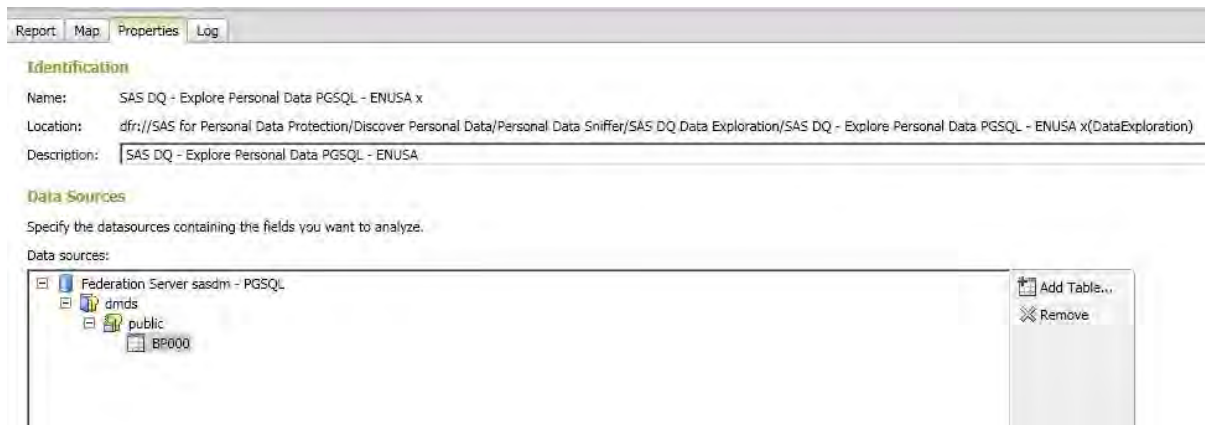
Automate Personal Data Discovery using SAS® DataFlux Data Management Studio

SAS® DataFlux Data Management Studio is a data management platform that combines data quality and data integration. It provides a process and a technology framework to deliver a single, accurate and consistent view of your enterprise data. With SAS® DataFlux Data Management Studio, you can establish an effective data governance through a wide range of activities, including personal data discovery.

SAS® DataFlux Data Management Studio offers multiple way for using the SAS® Quality Knowledge Base discovery capabilities presented formerly, including:

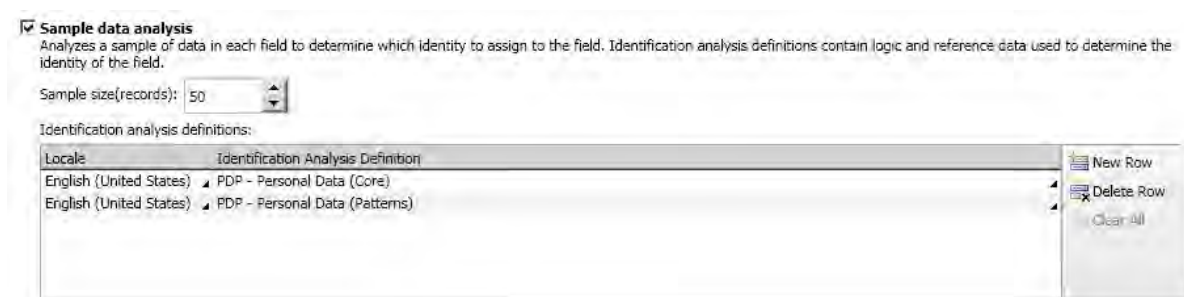
- Data Exploration
- Data Quality jobs

In the example below, we use Data Exploration to discover personal data in a table BP000.



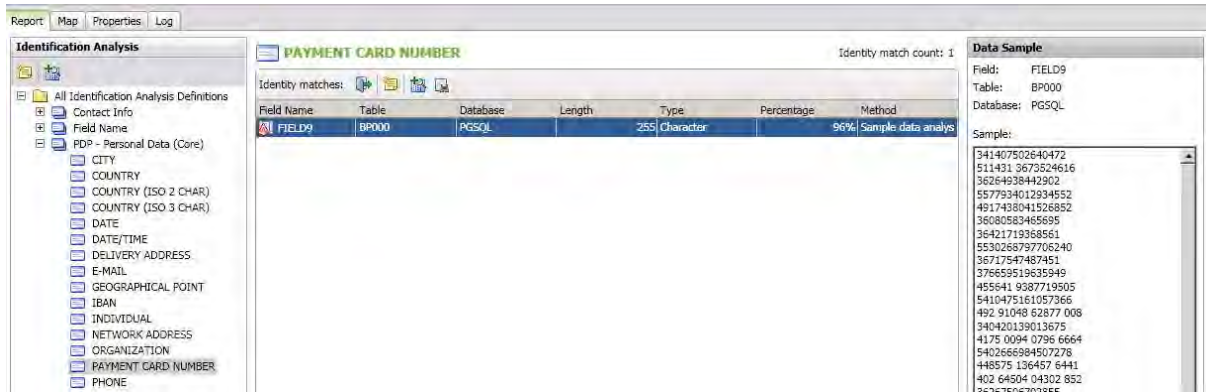
Display 11: Data Exploration – Data Source configuration

Display 12. Shows the configuration of the Data Exploration job to discover personal data based the field on two definitions, “PDP – Personal Data (Core)” and “PDP – Personal Data (Patterns)”. This last definition contains industry specific regular expressions to identify codes like VIN numbers, IMEI, or IMSI.



Display 12: Data Exploration – Identification Definition configuration

The identification analysis produces an overview of the results, by personal data category. You also get a sample view of the data that have been identified in the process. The field in the source, FIELD9, was identified a Payment Card Number with a 96% match percentage.



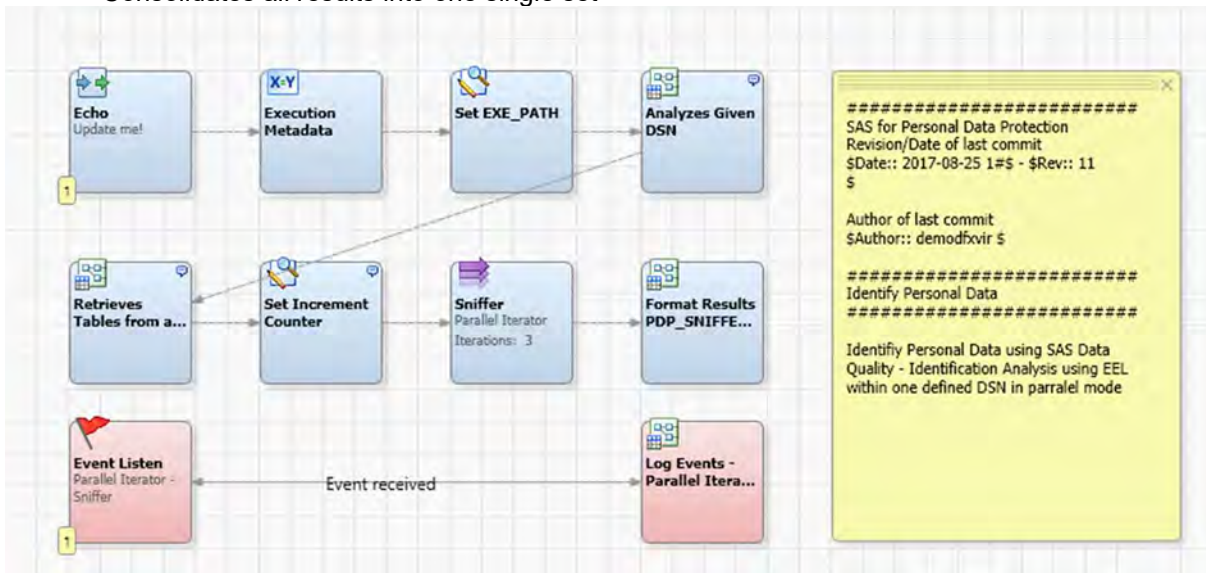
Display 13: Data Exploration – Identification Report

Data Quality Jobs for Personal Data Identification

Data quality jobs are the main way to process data in SAS® DataFlux Data Management Studio. Each data job specifies a set of data-processing operations that flow from source to target. These data processing operations can address multiple objectives such classic ETL operations, data quality operations like data standardization, matching and clustering, as well as personal data identification.

To industrialize and accelerate the personal data discovery process over one database, we have designed one dedicated data quality process job, named the “Personal Data Sniffer”. For one connection and schema, the “Sniffer”:

- Creates one folder for storing all the process output results and logs
- Analyzes the connection and parameters provided (sampling, randomization, definitions to apply)
- Lists all the tables contained in the connection and schema configured
- Initializes one parallel iterator to process multiple tables at the same time
- Executes the personal data identification over each table
- Consolidates all results into one single set

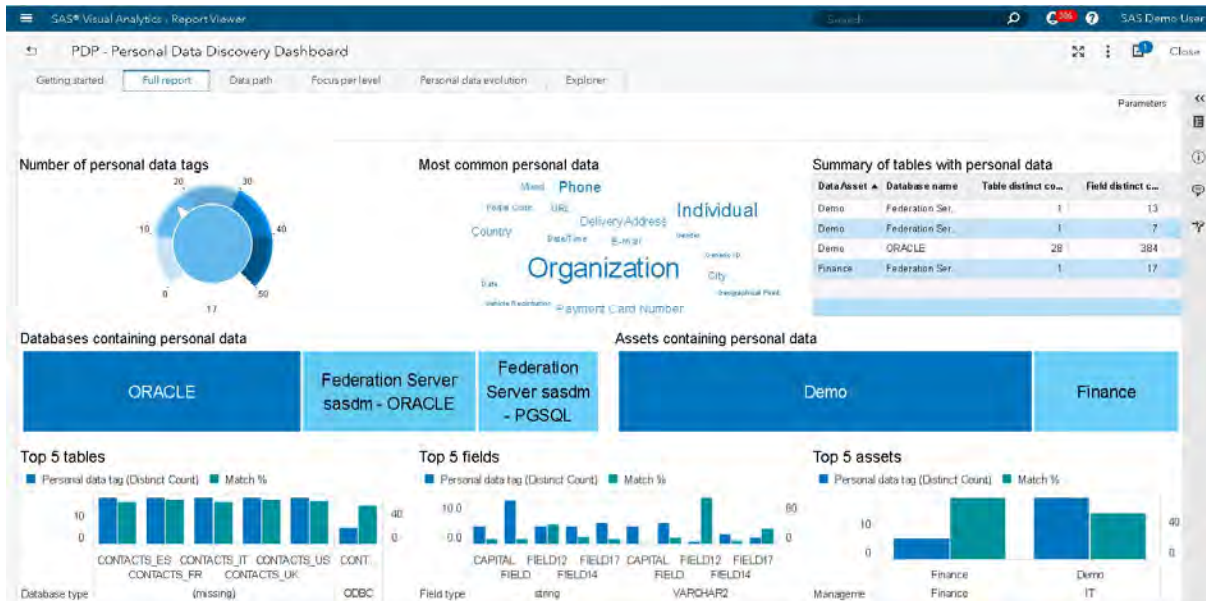


Display 14: Data Quality Job – Personal Data Sniffer

Report on personal data discovery in SAS® Visual Analytics

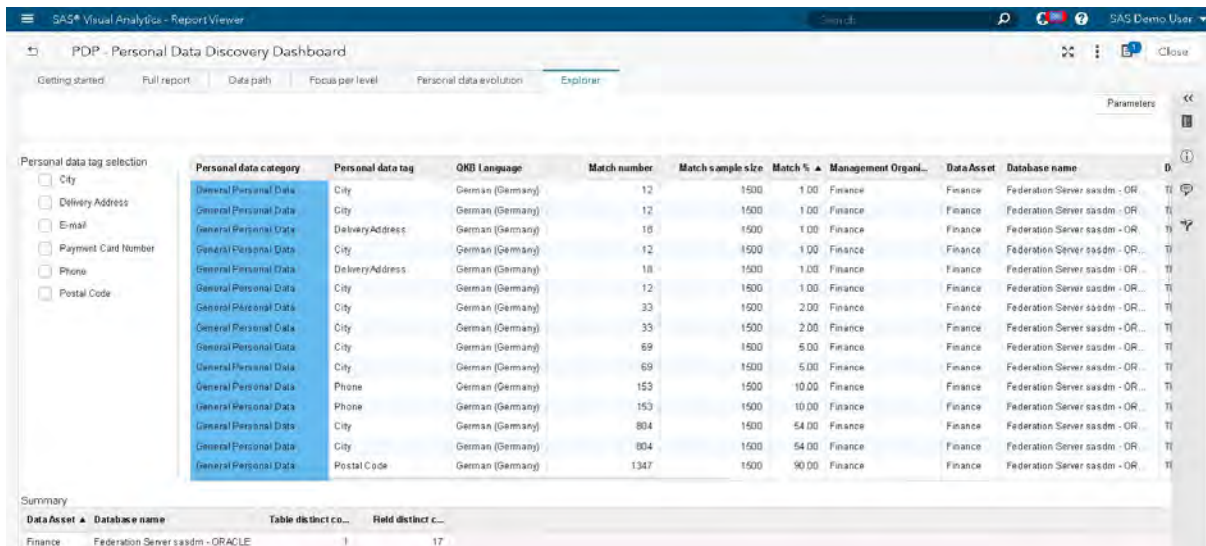
Results delivered by Data Exploration or Data Quality Job executions are challenging to analyze as the volume of information can be quite significant. In that perspective we have designed in SAS® Visual Analytics the Personal Data Discovery Dashboard to review the results of the different personal data discovery processes.

The dashboard provides a clear inventory of the personal data identified in any of the sources analyzed. The personal data categories are displayed by source, data asset, in which tables and columns they are present. Data Stewards and other users have a clear overview of the personal data contained in their applications, reducing their crypticity and making them transparent.



Display 15: Personal Data Discovery Dashboard – Full Report Tab

The “Explorer” tab allows you to investigate why a certain field has been identified as personal data category by the SAS® Quality Knowledge Base, in which table and with what match percentage.



Display 16: Personal Data Discovery Dashboard – Full Report Tab

ASSOCIATE & REVIEW PERSONAL DATA

Link Personal Data Discovery results with Personal Data Category Terms

In the first step of the methodology we have defined the personal data categories and data assets terms into SAS® Business Data Network. For each data asset we know the ODBC connection as well as the corresponding SAS® Library. The personal data discovery process provides one personal data attribute for each table and column analyzed. In the second step, the field metadata is associated to their corresponding personal data category term in SAS® Business Data Network.

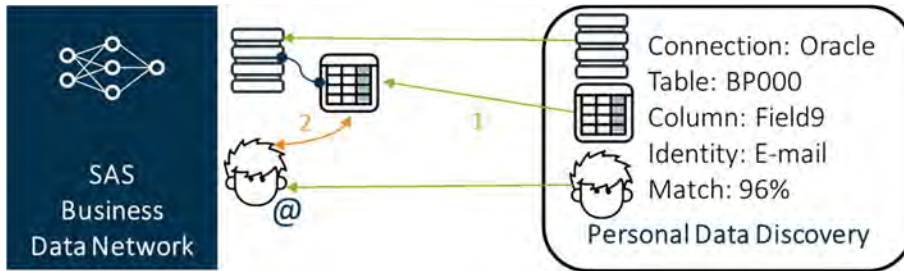
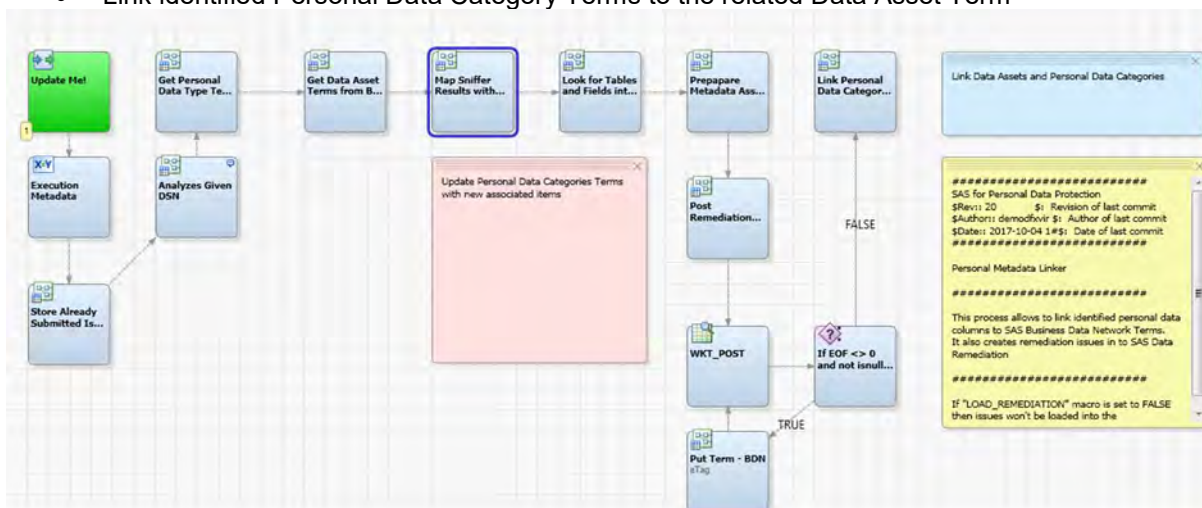


Figure 5: Metadata Linker Process – Overview

This process named the “Metadata Linker” connects data assets with the personal data categories automatically if personal data is identified in a table or file within the connection defined in the data asset. The level of confidence and the mapping with the SAS® Quality Knowledge Base token helps you validating automatically this association.

Display 17 illustrates the different steps of the process.

- Retrieve all defined personal data categories and data assets from SAS® Business Data Network
- Map Personal Data Discovery results with Personal Data Category Terms considering only the scores with the highest confidence (Figure 5 – Step 1)
- Search for the tables and columns into SAS® Lineage for each data asset according to the SAS Library mapped
- Associate column metadata with Personal Data Category Terms (Figure 5 – Step 2)
- Link identified Personal Data Category Terms to the related Data Asset Term



Display 17: Metadata Linker Process – Detailed view

Associations are automatically posted to SAS® Business Data Network using its REST API. For the Data Asset Terms, the results of the Metadata Linker process are visible in the Related Terms section. The “Finance” data asset presented in Display 7 now contains personal data categories (Display 18)

The screenshot shows the SAS Business Data Network interface. The main content area is divided into several sections:

- Description:** Finance System
- Requirements:** eGRC - DG Bridge
- Attributes:**
 - Reason of Usage: Reporting activities
 - Management Organization: Finance
 - Geography: Europe
 - Overall Asset Risk Rating: Not Applicable
 - Date of Implementation: 08/05/14
 - Date of Decommissioning:
 - System Status: ACT
 - Asset ID: 10015
 - Source System: MON
 - SAS Federation Server DSN: ORACLE
 - SAS Data Quality DSN: ORACLE
- Details:**
 - Type: Data Asset
 - Status: Created
 - Importance: Not Specified
- Links:**
- Tags:** GRC
- Related Terms:**
 - City (Personal Data Category)
 - Delivery address (Personal Data Category)
 - E-mail address (Personal Data Category)
 - Organization (Personal Data Category)
 - Payment card number (Personal Data Category)
 - Phone (Personal Data Category)
 - Postal code (Personal Data Category)

Display 18: Data Asset Term associated with Personal Data Category Terms

The full list of columns metadata identified as “E-mail” has been added automatically by the Metadata Linker to the Personal Data Category “E-mail”. You now have a complete view (identification, definition, governance, lineage) with just one process.

The screenshot shows the SAS Business Data Network interface with the 'Associated Items' tab selected. The 'Lineage' view displays a table of column metadata:

Name	ID	Description	Note	Type	Path
EMAILADDRESS	A510W4LJ.BG0004X3/Col...			Column	/Products/SAS Federation Serv...
EMAILADDRESS	A510W4LJ.BG0004XY-Col...	ORACLECONTACTS_ES	Personal Data Discovery	Column	
EMAILADDRESS	A510W4LJ.BG00051E-Col...	ORACLECONTACTS_US	Personal Data Discovery	Column	
EMAILADDRESS	A510W4LJ.BG0004X3-Col...	ORACLECONTACTS_DE	Personal Data Discovery	Column	
EMAILADDRESS	A510W4LJ.BG0004YT-Col...	ORACLECONTACTS_FR	Personal Data Discovery	Column	
EMAILADDRESS	A510W4LJ.BG0004ZO-Col...	ORACLECONTACTS_IT	Personal Data Discovery	Column	
EMAILADDRESS	A510W4LJ.BG00050J-Col...	ORACLECONTACTS_UK	Personal Data Discovery	Column	
FIELD	A510W4LJ.BG00062O/Col...			Column	/Products/SAS Federation Serve...
FIELD	A510W4LJ.BG00062U/Col...			Column	/Products/SAS Federation Serve...
FIELD9	A510W4LJ.BG000569-Col...	ORACLEITABLE_UK	Personal Data Discovery	Column	
FIELD9	A510W4LJ.BG00053O-Col...	ORACLEITABLE_ES	Personal Data Discovery	Column	

Display 19: Column Metadata associated automatically to a Personal Data Category Term

Remediate Low Confidence Association in SAS® Data Remediation

In SAS® Business Data Network, for each personal data category you had to specify a QKB personal data token and set a high limit and a low limit for the match percentage. After you specified these items, the personal metadata linker performs these tasks:

- links data with a match percentage over the high limit
- sends in a remediation queue (a manual confirmation process) all data with a match percentage between the low limit and the high limit
- rejects everything under the low limit

You can use SAS® Data Remediation to review propositions made by the personal metadata linker.

The screenshot shows the SAS Data Remediation interface. At the top, there's a header "SAS® Data Remediation" and a menu bar with "File" and "Help". Below that, there are navigation icons and a search bar containing "17 issues | Contains: phone". A dropdown menu shows "Grouped by (ungrouped)". The main area contains a table with the following data:

Issue	Package	Item	Subject Area	Application
Map Field Metadata to ...	GDPR Metadata Association	Associate NATIONALID with term Phone?	Metadata Linker	SAS PDP
Map Field Metadata to ...	GDPR Metadata Association	Associate FIELD4 with term Phone?	Metadata Linker	SAS PDP

Display 20: SAS® Data Remediation – Association Remediation main screen

When you open a proposition, you are being asked: “Associate Column with Term”, then, “Do you want to associate the column FIELD4 from the table TABLE_US in library ORACLE with term Phone?”. The term Phone is the term from the Personal Data Category in SAS® Business Data Network. You might have no clue what the FIELD4 contains, but you can review the content behind. A stored process behind opens a window with 20 sample records from the table TABLE_US, for the FIELD4 column.

Listing of FIELD4 from TABLE_US in Fed Server Library - Oracle

FIELD4
649-22-8483
637-92-5935
376-74-7719
171-80-2166
285-01-5025
243-48-9893
079-14-1865

Display 21: SAS® Data Remediation – Stored process showing column values

At this point you might recognize a telephone number pattern and decide that the proposed association is legitimate. You decide to associate the Field4 with the term Phone in SAS® Business Data Network. From SAS® Data Remediation, one REST API call is executed over SAS® Business Data Network to post the changes. The new associated field is visible in the “Phone” Personal Data Category Term (Display 22).

The screenshot shows the "Phone" term details in SAS Data Remediation. It includes tabs for "Identification", "Hierarchy", "Associated Items", "Notes and Contacts", and "History". The "Associated Items" tab is active, showing a table with the following data:

Name	ID	Description	Note	Type
FIELD4	A510W4LJ.BG00054E-Column	ORACLE\TABLE_FR	Personal Data Discovery	Column
FIELD4	A510W4LJ.BG00056Z-Column	ORACLE\TABLE_US	Personal Data Discovery	Column

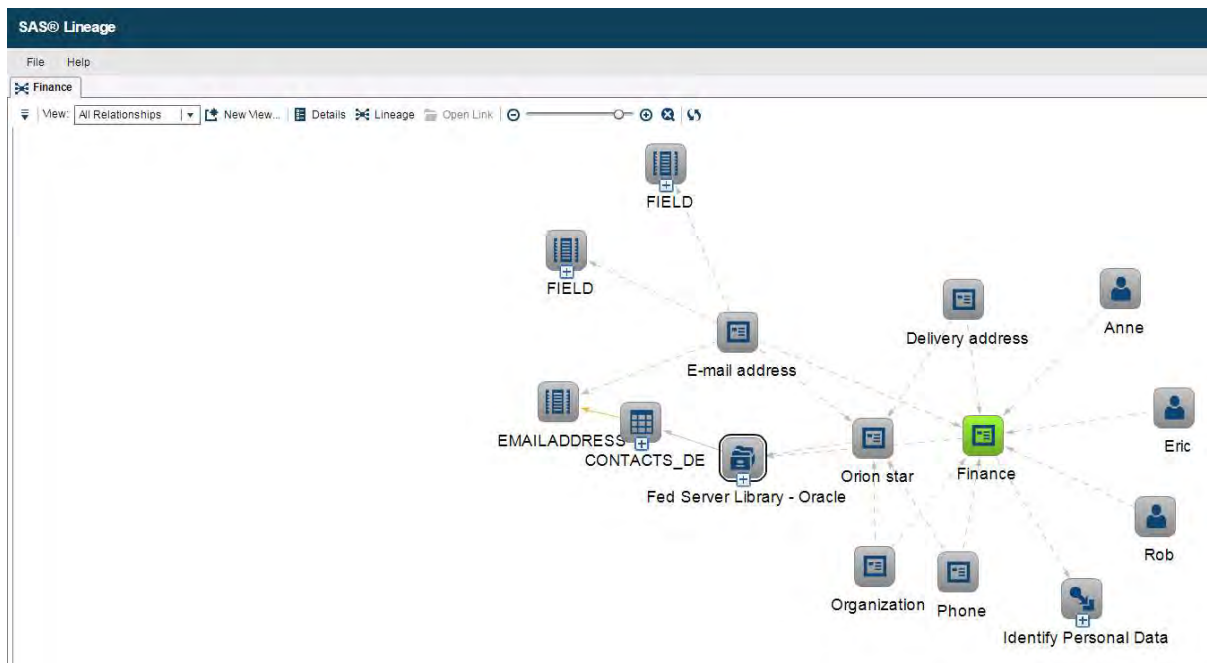
Display 22: SAS® Data Remediation – Manual Review

VISUALIZE DATA FLOWS & RELATIONSHIPS

SAS® Lineage

Once the former steps completed you can have a complete view of all the data assets and the personal data categories through SAS® Lineage. Lineage capabilities tie business and technical information together in a single and cohesive information store. It is a strategic component of a data governance plan, in the quest to understand the data assets in your organization.

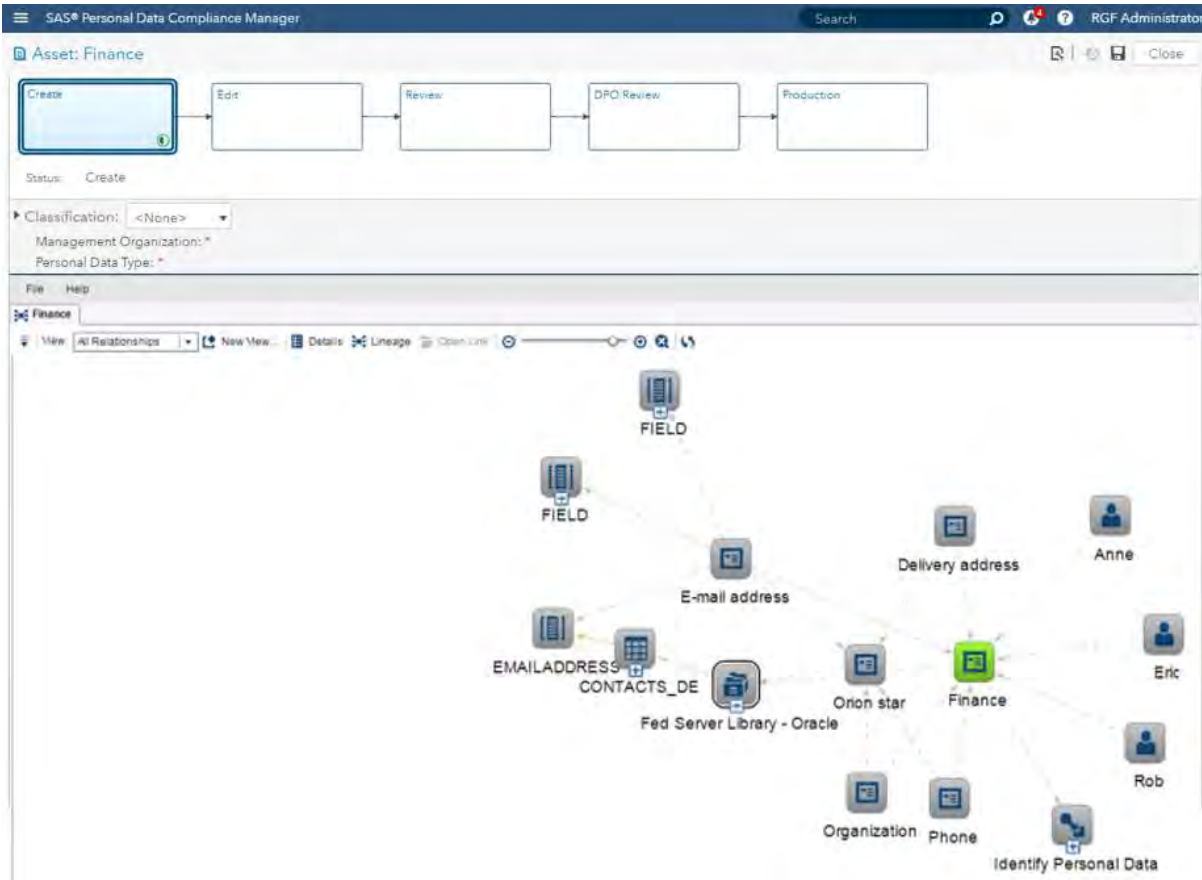
SAS® Lineage supports the management and analysis of object and metadata relationships, including dependencies and lifecycle. This management and analysis process reveals where data comes from, how it is transformed, and where it is going. It surfaces all the metadata present in the SAS® Platform including libraries, tables columns, but also SAS® Data Integration jobs, SAS® Data Quality jobs and SAS® Visual Analytics reports. Impact analysis can be activated to reveal which objects are affected when another object is changed or deleted.



Display 23: SAS® Lineage – Finance Asset Overview

SAS® Personal Data Compliance Manager integrated with SAS® Lineage

The Personal Data Protection Bridge (PDP Bridge) makes the Lineage View accessible in the Asset screen from SAS® Personal Data Compliance Manager. This opens the door to the synchronization of personal data categories discovered with the view available to a Data Controller.



Display 24: SAS® Personal Data Compliance Manager – SAS® Lineage Integration

CONCLUSION

We explained the legal background set by the General Data Protection Regulation, the provisions of the Article 30 for the records of processing activities and personal data definition. We then looked how data processing activities, data assets and data protection impact assessments are defined in the SAS® Personal Data Compliance Manager.

We introduced the need for a Personal Data Mapping Methodology. We demonstrated the need for SAS® Business Data Network and SAS® Lineage as a central point to maintain the personal data categories definitions, the sources of data where these can be found and the link with the data assets via related terms.

We proposed a personal data identification process with SAS® Data Management Studio to identify personal data using the SAS® Quality Knowledge Base definitions for personal data protection and highlighted the need for a Personal Data Discovery Dashboard in SAS® Visual Analytics to explore the results. Lastly, we introduced the usage of SAS® Data Remediation to solve low confidence proposals and stressed the importance of SAS® Lineage to visualize the links and associations.

The benefit for having the SAS platform at the core of the personal data protection initiatives is to reduce considerably the effort and the time to answer the regulator or the customers: what personal data is available in the sources, what are the processing activities and the risk posed by those to the liberties and rights of data subjects.

REFERENCES

- European Parliament and the Council of the European Union. 2016. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data*. Brussels, Belgium: European Union. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG
- *Information Commissioner's Office. 2017. Preparing for the General Data Protection Regulation (GDPR): 12 Steps to Take Now*. Cheshire, England: Information Commissioner's Office. Available: <https://ico.org.uk/media/1624219/preparing-for-the-gdpr-12-steps.pdf>
- EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide - *ITGP Privacy Team*

ACKNOWLEDGMENTS

We would like to express our special thanks to the SAS Personal Data Protection community and especially to Olivier Penel, Casper Pedersen, Arturo Salazar and Reece Clifford from the SAS EMEA DM Business Unit for driving this initiative. Many thanks also to Sorin Anghel from the SAS RQS team who has been a key player in the creation of SAS® Personal Data Compliance Manager, and to Camille Bouly for showing such expertise for designing wonderful SAS® Visual Analytics Dashboards.

RECOMMENDED READING

- Casper Pedersen, 2016. "SAS Solution for Personal Data Protection, Enabling compliance with new regulation". Available https://www.sas.com/content/dam/SAS/en_us/doc/other1/solution-for-personal-data-protection-108517.pdf
- Rineer, Brian, 2015. "Garbage In, Gourmet Out: How to Leverage the Power of the SAS® Quality Knowledge Base." Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc. Available <http://support.sas.com/resources/papers/proceedings15/SAS1852-2015.pdf>
- Hoffritz, Cecily, 2017. "I Spy PII: Detect, Protect, and Monitor Personal Data with SAS® Data Management." Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc. Available: <http://support.sas.com/resources/papers/proceedings17/SAS0639-2017.pdf>

CONTACT INFORMATION

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

Vincent Rejany
Domaine de Grégy
Grégy-sur-Yerres
77257 Brie Comte Robert Cedex
SAS Institute, Inc.
+33 (0)6 40 54 17 99
vincent.rejany@sas.com
<http://www.sas.com>

Bogdan Teleuca
Hertenbergstraat 6,
3080 Tervuren, Belgium
SAS Institute, Inc.
+32 475 36 33 81
bogdan.teleuca@sas.com
<http://www.sas.com>

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

Data Management for Artificial Intelligence



Contents

Introduction.....	1
Why AI is a hot topic.....	2
Why data management should be an equally hot topic.....	3
Five data management best practices for machine learning.....	4
1. Simplify access to traditional and emerging data	4
2. Drive smarter data integration with statistical AI.....	5
3. Scrub data to build quality into existing processes	6
4. Shape data using flexible manipulation techniques	6
5. Share metadata across data management and analytics domains.....	7
If you only remember four things.....	8
About SAS	9
Learn more	9

Introduction

Think back to the grade-school game where one person whispers something in the ear of the next person, and the phrase is then whispered from person to person around a circle.

The last person reveals what he or she heard, and it is always something wildly and hilariously different from the statement that started the cycle.

Working with bad data can be like that.

And if your process is some form of artificial intelligence (AI) - where the machine actually adapts the underlying algorithms based on what it learns from the data - bad data can really get you into trouble.

The results will be wildly skewed from the input that went into the cycle, but it's not hilarious at all.

Artificial intelligence is the science of training systems to emulate human tasks through learning and automation. With AI, machines can learn from experience, adjust to new inputs and accomplish tasks without manual intervention.

The explosion in market hype around the term is closely tied to advances in deep learning and cognitive science, but AI spans a variety of algorithms and methods. It doesn't require the flashiest new technologies to still be considered an AI application.

As a topic of interest for years - from science fiction plots to futurists' prophecies - the promise of AI has always been at the forefront of our minds. But what was once a distant vision is becoming reality as organizations embrace the value of AI now:

- By 2025, the artificial intelligence market will surpass \$100 billion. (Source: [Constellation Research](#))
- Seventy-two percent of business leaders believe AI will be fundamental in the future. (Source: [PwC](#))
- In the immediate future, execs are looking for AI to alleviate repetitive, menial tasks such as paperwork (82 percent), scheduling (79 percent) and timesheets (78 percent). (Source: [PwC](#))

Machine learning (a subset of artificial intelligence) automatically creates analytic models that adapt to what they find in the data. Over time, the algorithm "learns" how to deliver more accurate results, whether the goal is to make smarter credit decisions, retail offers, medical diagnoses or fraud detection.

We hope you have enjoyed this excerpt of “Data Management for Artificial Intelligence.”

To access the full whitepaper, please visit https://www.sas.com/en_us/whitepapers/data-management-artificial-intelligence-109860.html

Ready to take your SAS[®] and JMP[®] skills up a notch?



Be among the first to know about new books,
special events, and exclusive discounts.

support.sas.com/newbooks

Share your expertise. Write a book with SAS.

support.sas.com/publish

 sas.com/books
for additional books and resources.


THE POWER TO KNOW.®

