Paper SAS1669-2018

What's New in SAS® Data Management

Nancy Rausch, SAS Institute Inc., Cary, NC

ABSTRACT

The latest release of SAS® Data Management provides a comprehensive and integrated set of capabilities for collecting, transforming, and managing your data. The latest features include capabilities for working with data from a wide variety of environments including Apache Hadoop, the cloud, a relational database management system (RDBMS), unstructured data, streaming, Apache Spark, and the SAS Cloud Analytic Services server. New in the release is support and integration into SAS Viya. This paper provides an overview of the latest features of SAS Data Management and includes use cases and examples for leveraging the latest product capabilities.

INTRODUCTION

The latest release of SAS Data Management provides many new features that can help data warehouse developers, data stewards, and data scientists carry out data management tasks more efficiently and with greater control and flexibility. There are enhancements in the areas of data connectivity, data transformation, data quality and data governance. This release also introduces data management capabilities for SAS Viya, including some interesting new features for discovering data context.

DATA CONNECTIVITY

SAS Cloud Analytic Services (CAS) is the analytic server and associated cloud services in SAS Viya. A new CAS LIBNAME engine has been developed to connect a SAS9 session to an existing CAS session. The libref then becomes your link between SAS and the CAS server. When you assign a CAS engine libref, you are associating the libref with a CAS session and a caslib in order to work with CAS in-memory tables. The following example illustrates how to use the new engine to connect and work with tables in a CAS server.

```
/*Start a CAS session.*/
cas casauto host="myserver.example.com" port=5570;

/*Assign a CAS engine libref. If you don’t specify a caslib= option, then sas
uses the CASAUTO session from the SESSREF system option.*/
libname myCasLibrary cas;

/*Use the engine in SAS 9.4 to transfer data from SAS to the CAS server.
This reads the data from 9.4 SASHELP.CARS and send it to CAS.  The
promote=yes option makes the table global so that everyone can see it */
data mycas.cars (promote=yes);
   set sashelp.cars;
run;

/* you can go the other way too */
data sasuser.mycars;
  set mycas.cars;
run;
```

Data Integration Studio has developed a new Cloud Analytic Services Table Loader transformation that generates the code required use the CAS engine. This transformation replaces the Cloud Analytic Services Transfer transform. While existing jobs that use the Cloud Analytic Services Transfer transformation will continue to work, you can receive superior performance by leveraging this new transformation. Figure 1 is an example of the transformation in a Data Integration Studio job. The example illustrates using SAS to transfer data from a SAS table to a SAS CAS table.

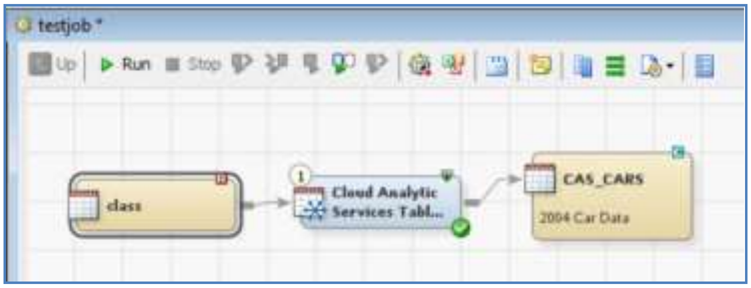


Figure 1: Cloud Analytic Table Server Transform

SAS Viya data connectors are enhanced to support reading and writing data, including new parallel read and write capabilities. When parallel load is configured, data is loaded directly into the Cloud Analytic Server worker nodes, as illustrated in Figure 2. If parallel load is not available, data is transferred into the controller node and then distributed. Parallel loads can speed up data loads, especially with large data sets.

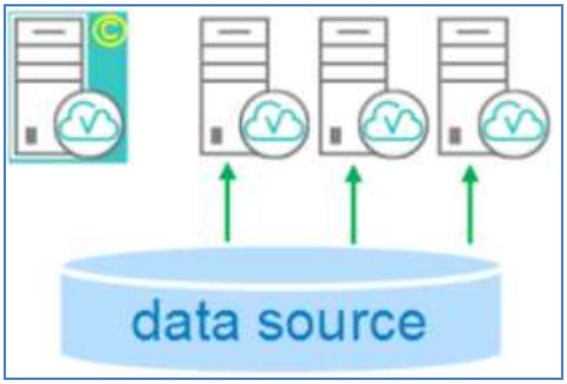
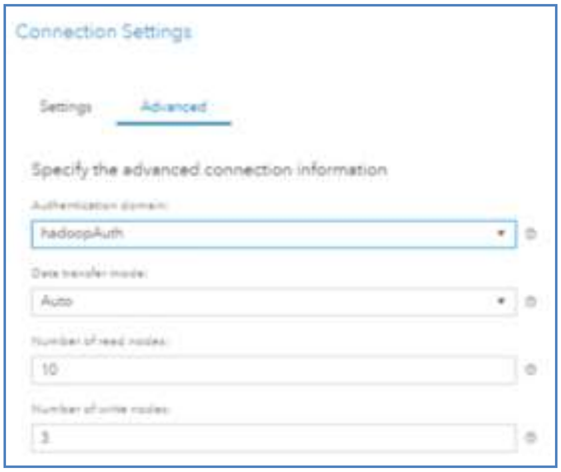


Figure 2: Example of Parallel Data Transfer

You can configure these read or write settings in a CASLIB statement via a new connection option, datatransfermode, as illustrated in the example below.

```
caslib myhiveclib
   datasource=(srctype='hadoop' datatransfermode='parallel'…
```

You can also use the Library Connection Settings dialog in the SAS Data Explorer dialog and application in SAS Viya to generate the correct code. This setting can be found on the Advanced settings of the Connections Settings dialog, as shown in Figure 3 below.



Connection Settings

Settings Advanced

Specify the advanced connection information

Authentication domain: hadoopAuth

Data transfer mode: Auto

Number of read nodes: 10

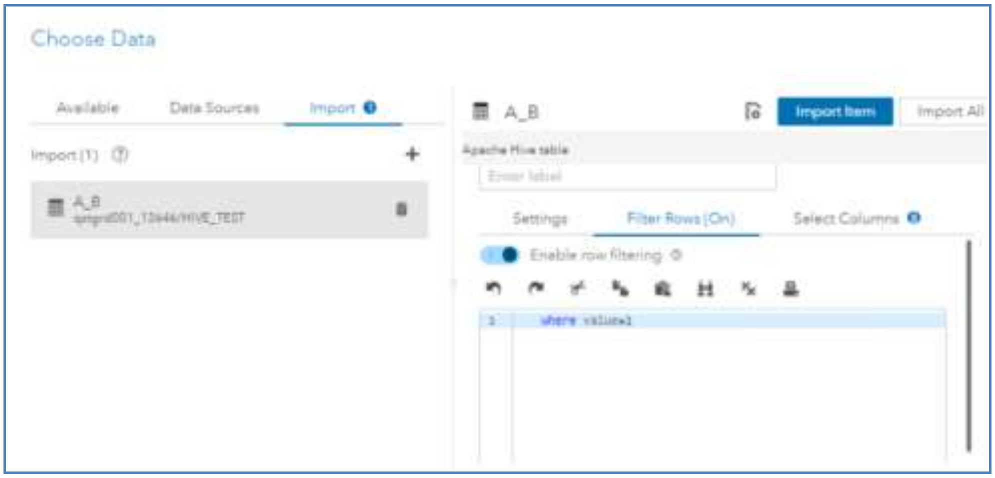
Number of write nodes: 3

Figure 3: Example of Connection Settings

There are four data transfer mode options available: default, serial, parallel, and auto.

- Default selects the best option based on the configuration of your system.
- Serial initiates parallel transfers if the number of nodes specified is greater than 1. This mode is called multimode, because it uses multiple CAS nodes to connect to a data source. A CAS controller node controls data transfer to and from worker nodes through concurrent connections with the data source. For reads, the controller node directs each CAS worker node to query the source data to obtain the needed data. Each worker node connects directly to the data source, which initiates multiple data streams to move data simultaneously. Note that in order to use this feature with databases, you must have connectivity configured on each of the worker nodes so that each of the worker nodes can connect to the database for the transfer. Multmode reads work by checking the target data source table for a numeric variable. SAS takes the first numeric variable that it finds and uses those values to divide the table into slices. Division is accomplished by using the MOD function and the number of nodes that you specify in the numreadnodes or numwritenodes option. The higher the cardinality of the numeric variable, the easier the data can be divided into slices
- Parallel leverages the SAS Embedded Process in-database technology to transfer data in parallel, if it is installed in your external system. In parallel mode, each CAS node connects directly to a data source node for parallel Read and Write. To use this mode, you must have access to a data connect accelerator. Data access is fastest with this method, as SAS Embedded Process to connect to individual slices of data on a data source node.
- Auto instructs SAS to first attempt data transfer using Parallel (SAS Embedded Process), and if that fails fall back to Serial (multimode) mode.

Another feature that can optimize performance when working with external data in a database is called pushdown. SAS Data Connectors now support pushdown of SQL queries to the underlying data source. You can specify WHERE clause queries in the loadTable action. You can also use the import feature in the Choose Data dialog or SAS Data Explorer in SAS Viya to generate the correct code to initiate pushdown for you. The Choose Data dialog and SAS Data Explorer support subsetting data by either rows and/or columns, as shown in Figure 4 below.



```
where value=1
```

Figure 4: Example of Choose Data Features

New data connectors are available to help you read and write data to and from a CAS server. In addition to the list below, you can define connections to any data source CAS can connect to by using the user-defined connection type. Below is a partial list of the new data connectors:

- SAS Data Connector to Amazon Redshift
- SAS Data Connector to DB2 for UNIX and PC Hosts
- SAS Data Connector to Microsoft SQL Server
- SAS Data Connector to SAP HANA
- SAS Data Connector to SPD Engine Files and SAS Data Connect Accelerator for SPD Engine Files
- SAS Data Connector to MySQL
- SAS Data Connector to SparkSQL
- SAS Data Connector to JDBC
- SAS Data Connector to Vertica

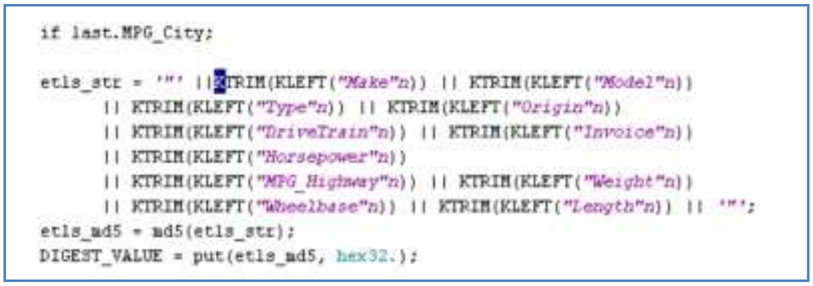
There are also several new SAS 9 access engine technologies available including SAS/ACCESS to PostgreSQL, SAP HANA, and Vertica.

DATA INTEGRATION AND TRANSFORMATION FEATURES

Data integration is the process of consolidating and transforming data from a variety of sources in order to produce a unified view of the data. New features enable you to transform and prepare your data for reporting or analysis.

SAS Data Integration Studio provides a powerful visual design tool for building, implementing and managing data integration and transformation processes for most types of data sources, applications, and platforms. An easy-to-manage, multiple-user environment enables collaboration on large enterprise projects with repeatable processes that are easily shared.

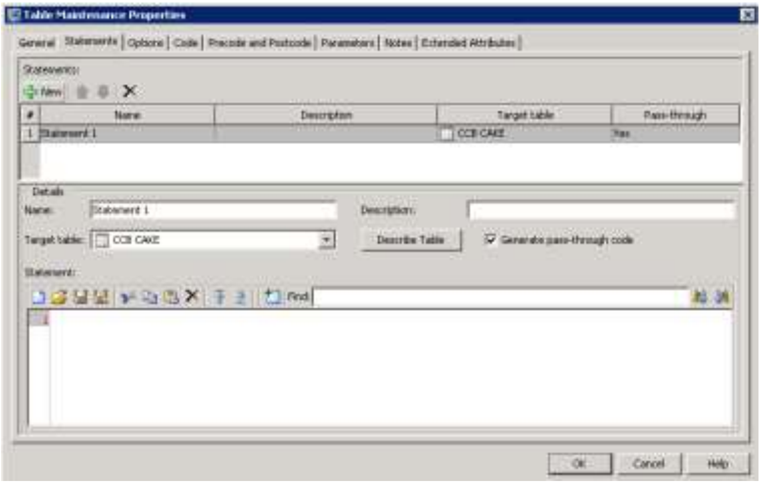
In addition to the new connectivity options documented above, SAS Data Integration Studio has additional new features. One key new feature is support for K-functions in code generation. K-functions enable you to manipulate DBCS string data. Data Integration Studio will now generate K functions where appropriate. The following example contains code that is generated during the Slowly Changing Dimensions Type 2 transformation.



```
if last.MPG_City;

etls_str = '"' ||KTRIM(KLEFT("Make"n)) || KTRIM(KLEFT("Model"n))
      || KTRIM(KLEFT("Type"n)) || KTRIM(KLEFT("Origin"n))
      || KTRIM(KLEFT("DriveTrain"n)) || KTRIM(KLEFT("Invoice"n))
      || KTRIM(KLEFT("Horsepower"n))
      || KTRIM(KLEFT("MPG_Highway"n)) || KTRIM(KLEFT("Weight"n))
      || KTRIM(KLEFT("Wheelbase"n)) || KTRIM(KLEFT("Length"n)) || '"';
etls_md5 = md5(etls_str);
DIGEST_VALUE = put(etls_md5, hex32.);
```

Figure 5: K-function Code Generation

Data Integration Studio has added a new Table Maintenance transform. This transform supports the ability to enter code to manage tables in a database, such as alter, update, drop, or truncate. The code can come from a file of commands, or can be entered into a statement window. The code can be either SAS PROC SQL code, or pass-through code to allow users to directly send statements to a database. Templates for various SQL statements such as CREATE, ALTER, and DROP for various database are also available to assist you in building code. Figure 6 is an example screenshot of the new transform.



Statements:

#	Name	Description	Target table	Pass-through
1	Statement 1		CCB CAKE	Yes

Details

Name: Statement 1

Description:

Target table: CCB CAKE

☑ Generate pass-through code

Statement:

Figure 6: Table Maintenance Transform

The list below describes additional features that are available in the latest version of SAS Data Integration Studio:

- SAS Data Quality jobs that are embedded into SAS Data Integration Studio jobs now support Integrated Windows Authentication.
- Hadoop configuration is now managed for all SAS features through a single environment variable, SAS_HADOOP_CONFIG_PATH. This simplifies deployment and integration of SAS and Hadoop clusters.
- Improvements in various code generation features for optimized performance. Some examples are:
    - The Hadoop Hive transform now enables you to delete or truncate target tables.
    - Insert OVERWRITE syntax is generated for Hive tables.
    - Support for cross schema joins in Hive.
    - Support in the Data Loader transform in SAS Data Integration Studio for the SAS JSON engine.
    - New macro variables to assist you when writing your own generated transforms. The macro variables are controlled via a check box on the generated transform definition.

- An Inputs/Outputs tab to enable the generation of column mapping macros. When this feature is enabled, additional macro variables have been added so that you can access source table information. Below is an example of two of these new macro variables:

```
%let  OUTPUT col5 input0=customer since dt;
%let _OUTPUT_col5_input0_table=testlib.input_table0;
```

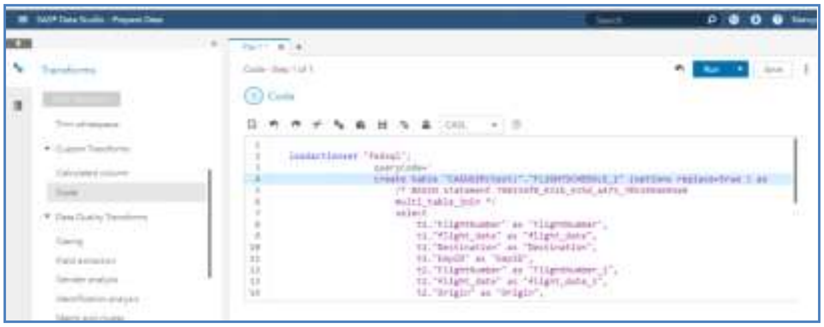
SAS Data Preparation is a web-based self-service application for SAS Viya that allows you to quickly prepare data for analytics– without coding, using specialized skills, or relying on IT. Prebuilt transformations and data cleansing functions run in memory in the SAS Cloud Analytics Server. The features are seamlessly integrated with SAS Data Mining and Machine Learning and SAS Visual Analytics. Data Preparation jobs can be scheduled to run, or can be orchestrated from other data management applications such as Data Integration Studio jobs via a documented REST interface. For example, you can create a SAS Data Preparation data plan and deploy it, and then run the data plan from Data Integration Studio using rest service calls. Below is an example.

```
proc http method="get"
          url="my.services.url/jobExecution/jobRequests?filter=eq(name,"MySav
edJob")"
          out=response
          headerout=headers headerout overwrite;
          headers "Authorization"="bearer
1iuasd9f9slssdks8fzI1NiIsImtpZCI6Imx..."
                  "Accept"="application/vnd.sas.collection+json";
run;

proc http method="post"
          url="my.services.url/jobExecution/jobRequests/12345674-7053-4712-
aae1-4d80d1234567/jobs"
          out=response
          headerout=headers headerout overwrite;
          headers "Authorization"="bearer
esadfsadfe9s8djsd8s0dksjfsfdI6Imx..."
                  "Accept"="application/vnd.sas.job.execution.job+json";


run;
```

SAS Data Preparation supports features to move data into and out of memory in serial or in parallel using SAS Data Connectors as explained above. There are also features to transform your data using data wrangling features like in-memory splits, join, append, transpose. There are data quality features such as standardize, parse, extract, and profile. Below is a list of the application components that make up SAS Data Preparation.

- SAS Data Explorer and the Choose Data Dialog – Used to move data into and out of the SAS Cloud Analytic Server and provide details about the data such as sampling and profiling.
- SAS Data Studio – Provides interactive data wrangling capabilities for transforming data in SAS CAS.
- SAS Lineage – Provides lineage and impact analysis features.
- SAS Job Monitor – Monitor jobs that are running in the SAS CAS server, stop and restart them interactively, and download logs and code.
- SAS Projects – Collaboration features for collaborating with other users and sharing content.

SAS Data Preparation includes a multi-lingual code generator that can generate various languages including SQL, SAS Cloud Analytics Server Language, and SAS DATA Step. You can write your own code and submit it through SAS Data Studio using the code node, DATA Step, and SAS CASL language. Below is an example of the code node in SAS Data Studio.



```
loadactionset "fedsql";
          queryCode=
          create table "[illegible]"."[illegible]" {options replace=true} as
            /* [illegible] statement [illegible]
            multi_table_join */
            select
              t1."flightNumber" as "flightNumber",
              t1."flight_date" as "flight_date",
              t1."Destination" as "Destination",
              t1."keyID" as "keyID",
              t2."flightNumber" as "flightNumber_1",
              t2."flight_date" as "flight_date_1",
              t2."Origin" as "Origin",
```

Figure 7: Example of the Code Node in SAS Data Studio

Transformation features are available in SAS Data Studio that transform your data in CAS, including the following:

- Change case by locale
- Field extraction
- Gender analysis
- Match and cluster to de-duplicate data
- Parse data into separate fields
- Standardize names, addresses, ZIP codes, country codes, phone numbers, email, and more
- Append and Join tables
- Analytical Partition data
- Generate unique identifiers
- Subset rows and columns
- Generate calculated columns

DATA QUALITY

SAS Data Management and Data Quality can help you cleanse your data to ensure that it is ready for integration and analysis. One new feature will return the name of the most likely locale by leveraging a confidence score for the locale that is most likely represented by a character value. For example, the following code uses three specified locales and returns the locale guess that has the highest score for the input data.

```
options dqlocale=(ENUSA ENGBR ENHKG);
data _null_;
locale = dqLocaleGuess(input, 'Wake County');
run;
```

Additional noteworthy new features include the following:

- List the locales supported by the SAS QKB with the DQLOCLST procedure.
- New features for customizing the SAS Quality Knowledge Base that contains user customizable data quality rules:

  - You can now copy and paste regular expressions in the Regex Library Edition.
  - You can import word, category, and likelihood values from external files in the Vocabulary Editor.
  - You can copy and paste rules and categories in the Grammar Editor.
- Netezza 7.2.0.5 has been added to the list of supported databases for data storage in Data Management Studio.
- DataFlux Expression Language 2.7 includes two new functions to calculate the distance between two geographical points: GEODISTANCE_COSINE, GEODISTANCE_HAVERSINE
- Increased security options, including enhanced support for FIPS compliance.
- Simplified software upgrades with upgrade in place migration support.

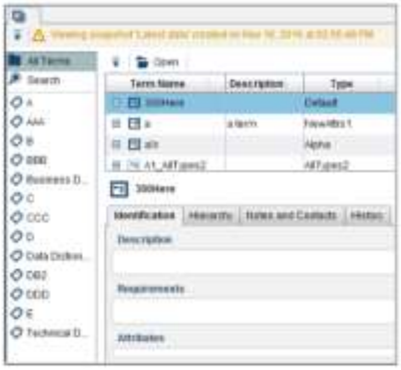
DATA GOVERNANCE

Data governance helps to ensure that important data assets are formally managed throughout an enterprise. SAS has a comprehensive suite of technologies that can be used to help you efficiently and effectively govern and manage your data assets. There are a number of new capabilities in this area.

Security is a key underpinning of data governance. One new feature in support of security is added support for Apache Sentry RecordService enabled clusters. This technique supports Hadoop cluster role-based row and column level security and dynamic data masking for data and metadata that is stored on a Hadoop cluster.

SAS has a complete solution in support of the European Union (EU) General Data Protection Regulation (GDPR). A key component of this offering is the SAS Federation Server, which supports role-based data access, row level security, and data masking features such as hashing, randomization, and encryption. The SAS GDPR solution enables you to implement a data protection methodology with minimal interruption to existing systems. SAS can also create customizable rules to identify customer-specific personal data categories using the SAS Data Quality Knowledge Base. This can help you both support and customize the handling of personally identifiable information (PII).

The SAS Business Data Network (BDN) provides capabilities for creating and managing an authoritative vocabulary that promotes a common understanding between stakeholders in an organization. There are a number of new features in SAS BDN. One key new feature is support for historical management of vocabularies through snapshots. An administrator can retain historical views of an entire dictionary and retrieve them when needed to see how information has changed. Figure 8 shows a view of a historical snapshot.



Term Name	Description	Type
300Here		Default
a	a term	[illegible]
alt		Alpha
A1_AllTypes2		AllTypes2

Figure 8: Example of Snapshots in SAS Business Data Network

The following enhancements are also included:

- Ability to import content from CSV files
- Enhanced lineage and Business Data Network interoperability
- Publish support for import results into the SAS Relationship Service

- New fields for URL, DATE, and RTF for term types
- A documented Application Programming Interface (API) for programmatic access to dictionary content and workflow

DATA DISCOVERY

SAS has added new features that help automate content identification for data discovery. One new feature leverages SAS data quality capabilities to automatically identify and label data. For example, SAS can determine that data contains name values, and is able to distinguish between names of an organization versus the name of an individual. Figure 9 shows an example of this categorization in SAS.

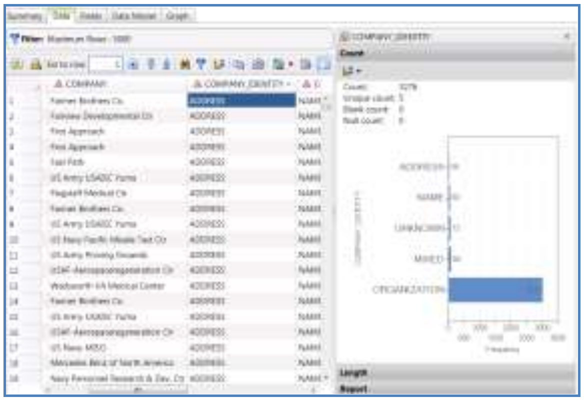


Figure 9: Example of Identification Analysis

You can identify data across different locales within a data set, as shown in the following example.

```
dqIdentify('CompanyA','Individual/Organization','ENUSA');
```

```
dqIdentify('CompanyB','Individual/Organization', DEDEU);
```

Identification analysis has been enhanced with additional definitions and locales. These include definitions for identifying personal information, such as government IDs and bank account numbers.

SAS Data Preparation leverages this feature to auto-categorize content for you. This can be very useful when working with large data sets where you want to understand the data context across a large amount of unknown data. Figure 10 illustrates how to enable this feature in SAS Data Preparation, and Figure 11 shows an example of the results.

Figure 10: Example of How to Enable Identification Analysis

Figure 11: Example Identified Columns

CONCLUSION

The latest releases of SAS Data Management products provide enhancements to help data specialists carry out data-oriented processes more efficiently and with greater control and flexibility. Enhancements have been made in all areas of the data management lifecycle. Customers will find many reasons to upgrade to the latest versions of SAS Data Management.

RECOMMENDED READING

- SAS Data Management Community, Available at https://communities.sas.com/t5/Data-Management/ct-p/data_management?nobounce
- Hazejager, W.,2018. "Data Management in SAS Viya: A Deep Dive". *Proceedings of the SAS Global Forum 2018 Conference*, Cary, NC. SAS Institute. Available at https://support.sas.com/resources/papers/proceedings18/SAS1670-2018.pdf
- Rineer, B.,2018. "Doin' DQ in SAS Viya". *Proceedings of the SAS Global Forum 2018 Conference*, Cary, NC. SAS Institute. Available at https://support.sas.com/resources/papers/proceedings18/SAS2156-2018.pdf
- Rausch, N., 2017. "What's new in SAS Data Management". *Proceedings of the SAS Global Forum 2017 Conference*, Cary, NC. SAS Institute. Available at http://support.sas.com/resources/papers/proceedings17/SAS0195-2017.pdf
- Hazejager, W. & Rausch, N., 2017. "Ten Tips to Unlock the Power of Hadoop with SAS". *Proceedings of the SAS Global Forum 2017 Conference*, Cary, NC. SAS Institute. Available at http://support.sas.com/resources/papers/proceedings17/SAS0190-2017.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 http://support.sas.com/resources/papers/proceedings16/SAS5644-2016.pdf.

CONTACT INFORMATION

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

Nancy Rausch
SAS Campus Drive
Cary, NC 27511
SAS Institute Inc.
Work Phone: (919) 677-8000
Fax: (919) 677-4444

Email: Nancy.Rausch@sas.com
Web: support.sas.com