Paper SAS4079-2020

# What's New in SAS Data Management

Nancy Rausch, SAS Institute Inc., Cary, NC

## ABSTRACT

The latest releases of SAS® Data Management software provide a comprehensive and integrated set of capabilities for collecting, transforming, and managing your data. The latest features in the product suite include capabilities for working with data from a wide variety of environments and types including Apache Hadoop, cloud data sources, relational database management system (RDBMS), files, unstructured data, images, and streaming data, with the ability to perform extract, transform, load (ETL) and extract, load, transform (ELT) transformations in diverse run-time environments including SAS®, Hadoop, Spark, SAS® Analytics, cloud, and data virtualization environments. The SAS Data Management offering has been enhanced to include integration with SAS® Studio, including a new ETL flow building capability that can be used to build reusable data flow processes. Enhancements have also been added to leverage analytics and artificial intelligence (AI) to help automate data management tasks. This paper provides an overview of the latest features of the SAS Data Management product suite and includes use cases and examples for leveraging 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 preparation, and data management. This paper provides an overview of many of the new data management features.

## DATA CONNECTIVITY

Access to diverse data can provide deeper insights into business problems. In order to access diverse data, connectivity needs to support varied types and sources of data at different latencies, from historic to real time streams. Transformational capabilities also need to evolve to support new types and larger quantities of data.

Figure 1 provides a high-level overview of the many types and formats of data connectivity that SAS provides.

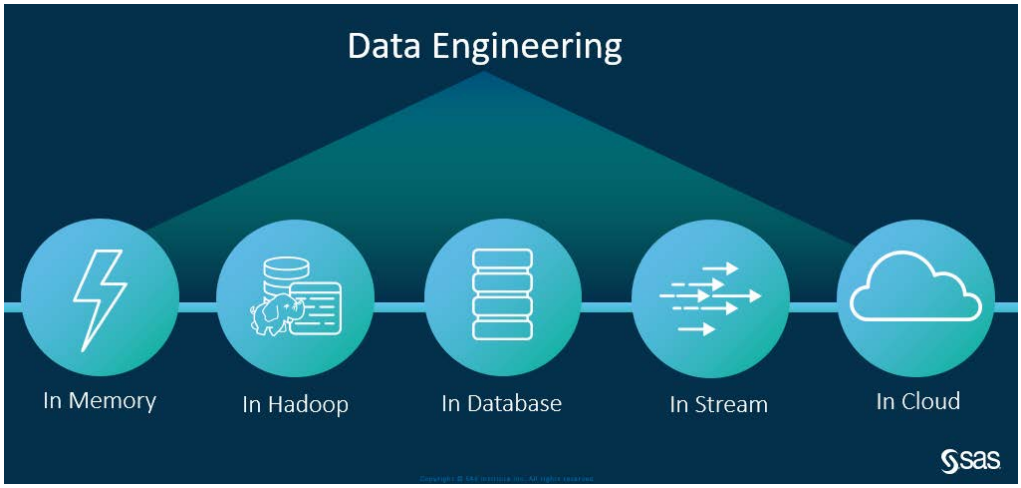


Figure 1: Data Types and Formats

Some important new data connectivity enhancements in the latest releases include:

- Read/Write of many new data types: JMP, CSV, TXT, XLS, XLSX, DTA, SAV on Local FS and DNF, ORC and CSV on Azure Data Lake Storage (Generation 2), and Parquet on S3, images, documents, multiple Excel worksheets, multiple delimited files in a directory, and unstructured data
- SAS® Data Connector to access data from SAS® Event Stream Processing (ESP) windows
- Support to securely move files, and directories of files, to and from cloud containers
- Support for file sharing through downloads and SAS® Drive folders
- Support for cloud sources including Google Big Query, Snowflake, MongoDB, and Salesforce

Additional details about some of these new features are further described below.

## COLUMNAR FILE SUPPORT: PARQUET AND ORC

One useful new feature is support for reading and writing columnar files. These files are particularly popular in Hadoop systems because they have faster read and write times compared to more traditional formats, they can be stored very efficiently in a distributed system, and they integrate well with third-party tools. They also are stored in a compressed format, which reduces storage requirements.

SAS has added support for two different columnar file types: PARQUET and ORC. Figure 2 below illustrates how these formats work:

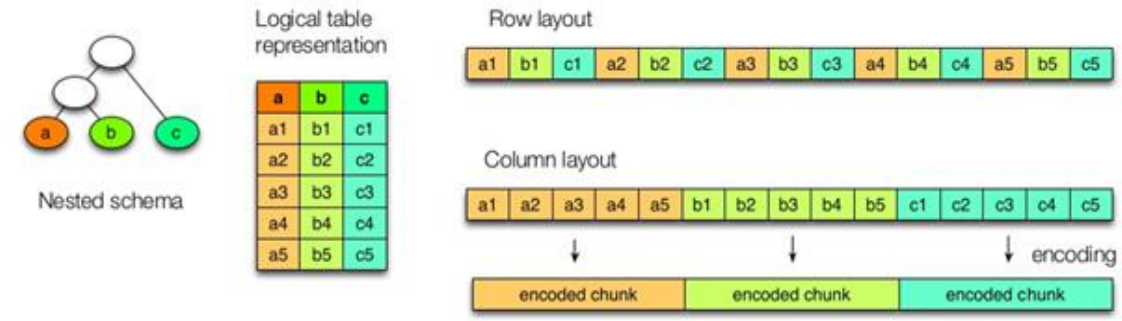


Figure 2: Columnar Storage Example

One reason for considering this storage format is to improve read and write performance. Figure 3 is an example of the potential performance difference between the SASHDAT format compared to Parquet.

| 4-nodes CAS, PATH CASLIB | Load time in sec. with default COPIES (1) | Load time in sec. with COPIES=0 |
|---|---|---|
| **SASHDAT – 11.6GB** | 97.12 | 54.38 |
| **Parquet – 483 MB** | 51.66 | 18.65 |
| **Times faster** | ~2 | ~3 |

Figure 3: Example Performance Using Columnar Storage

ORC and Parquet formats are specified when reading or writing data. Figure 4 is one example of how to specify reading a directory of ORC files into SAS in SAS® Data Explorer:

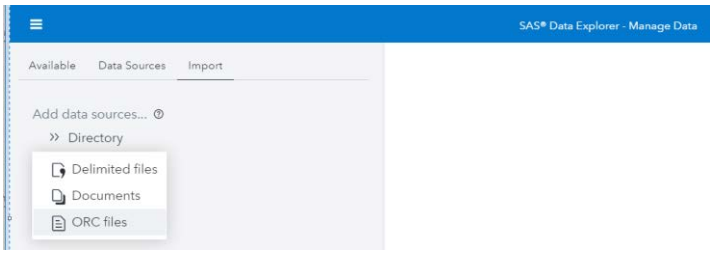


Figure 4: ORC Files Import Example

The following code snippet illustrates how to read or write this file format from SAS code:

```
libname orclib ORC '/dev/samspade'
      storage_account_name = sasuser
      storage_file_system = hdfflib
      storage_application_id="123456a6-19d8-4e97-94d3-91914f5c6917"
      storage_tenant_id="1234567c-3625-45b3-a430-95523796655";
data orclib.cars_orc;
   set sashelp.cars;
run;
```

Storage formats can also be specified when you save data, for example, on an import or when you save a table in SAS® Data Studio. This option is illustrated below.

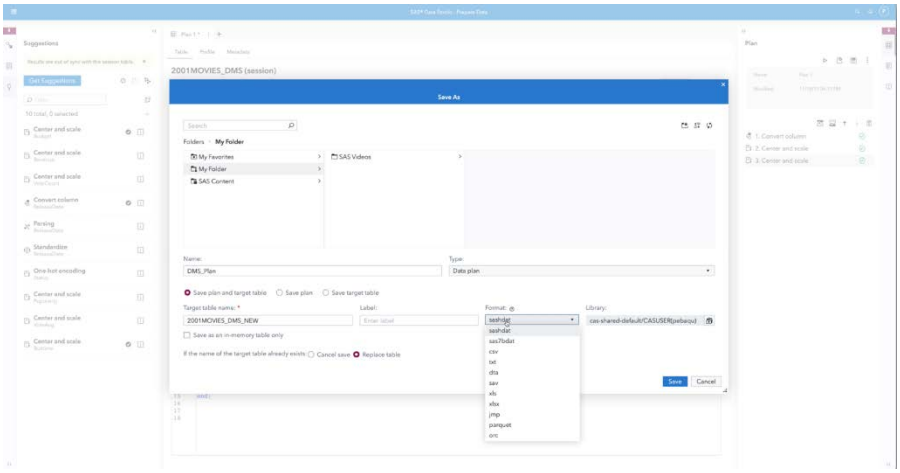


Figure 5: Example of Various Storage Formats Available When Saving a Dataset

## IMAGE FILES

SAS® Viya® image actions let you read and write images, process images, and manipulate image data. Support has been added to import directories of image files in the SAS import user interface (UI). You can also preview image files in the UI. Each imported image file is stored in a binary format in a single row in the target table. Additional fields in the row include metadata about the image: a unique identifier, size, name, and location of the file. Image file import is illustrated in Figure 6 and Figure 7 below.

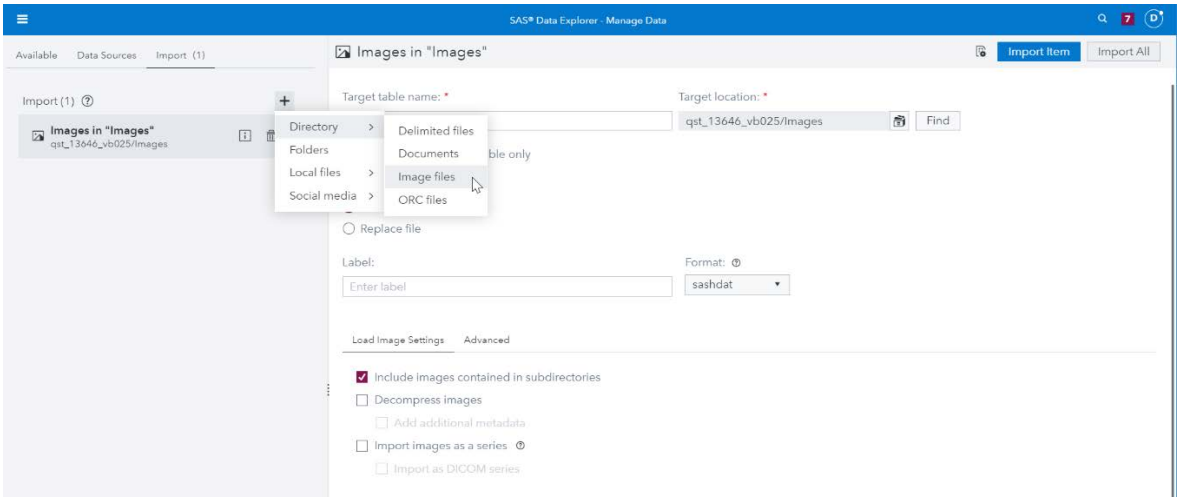


Figure 6: Image Files Import Example

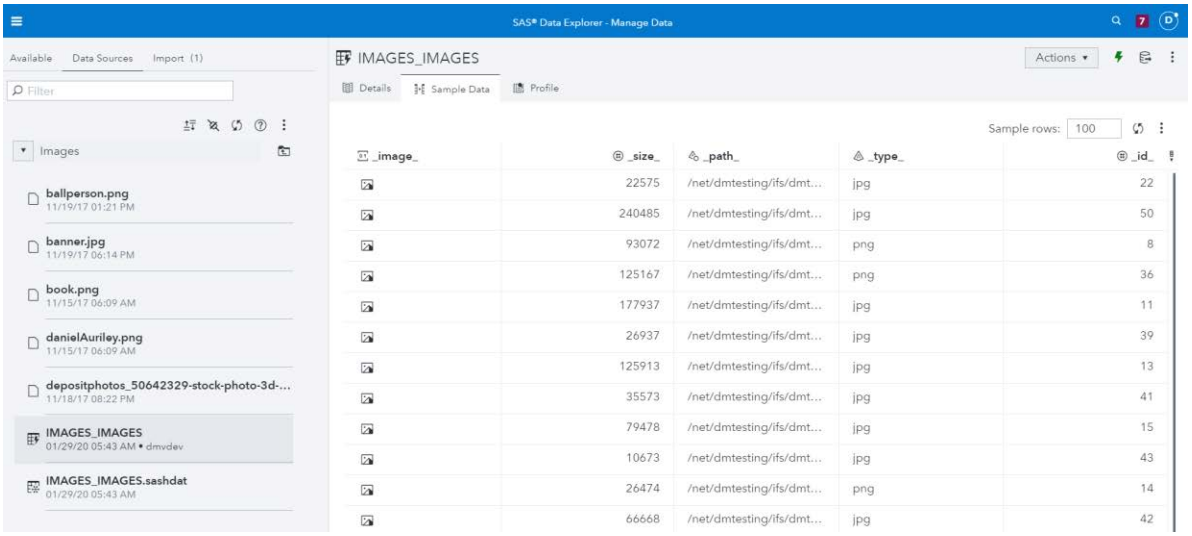


Figure 7: Image Table Showing Each Image as a Row in the Dataset

The image preview action in the sample data tab allows you to preview each image in the table. Figure 8 shows an example of the image preview.

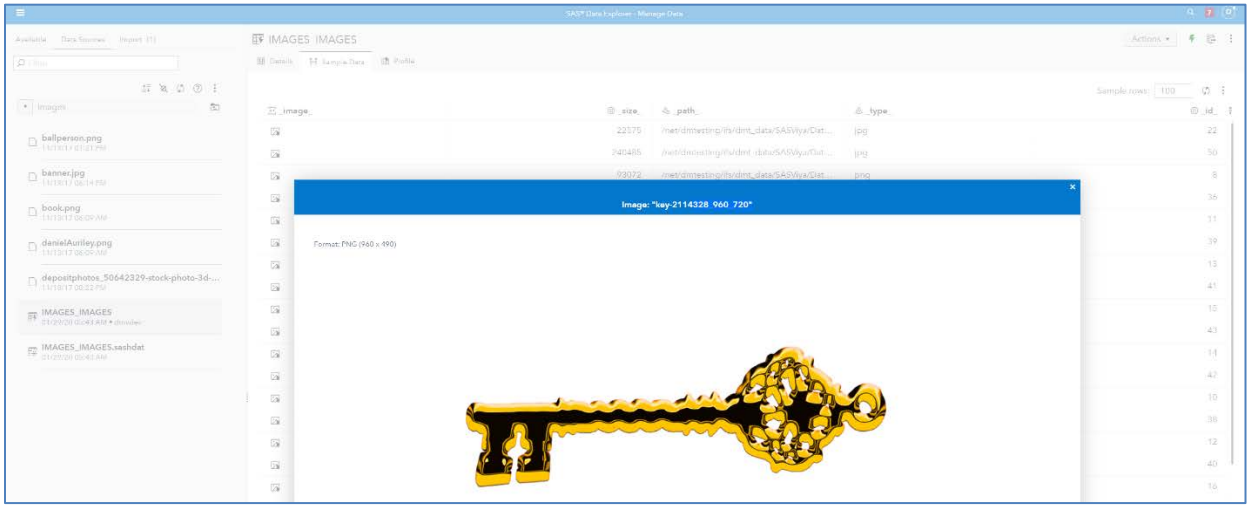


Figure 8: Example of the Image Preview Option

## SAS ESP DATA CONNECTOR

SAS ESP enables you to connect to and process real-time data. It has a variety of data connectors that support read and write capabilities, for both traditional and non-traditional data sources types. Some of the non-traditional sources include message queues, social media sources such as Twitter, NoSQL databases such as Cassandra, web sniffers, and change data capture engines. You can also manipulate and perform analytics on the data as it is read as part of the stream.

To connect to an ESP source, you first create an ESP project and name the node window that you want to pull data from. Then you create a CASLIB connection to that window. For the path setting, you use the following syntax <project_name>/<continuous_query_name>. For the server setting, use the hostname of your ESP server. For the port, use the pubsub port of your ESP server (typically 1111, if left unchanged).

An ESP connection is illustrated in the figure below. In this example, the node being used to retrieve data is the continuous query node, named cq1 in the example.

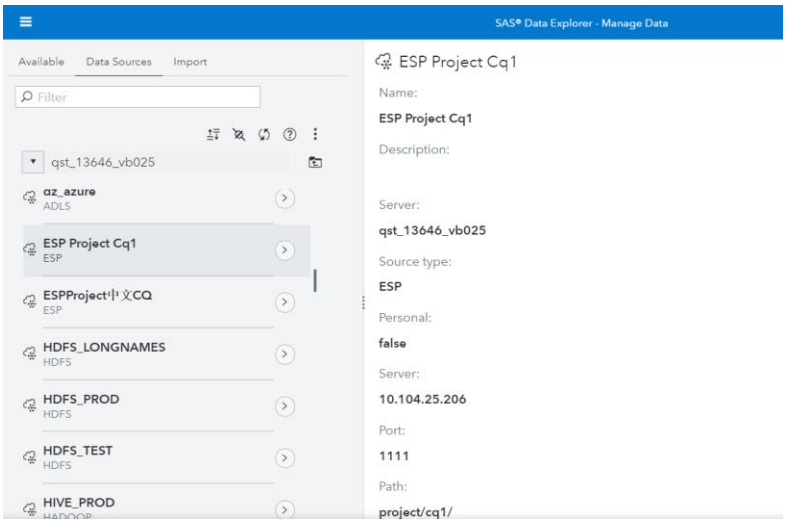


Figure 9: ESP Library Example

Once you have created an ESP connection, the many ESP connectors and adapters enable you to stream data through the ESP engine. They use a publish/subscribe API to interface directly with a variety of message buses, communication fabrics, drivers, and clients. When you read, data will be pulled through an ESP window into a CAS table. Each read will pull a new set of data, and you can append the data together for regularly scheduled data pulls.

You can also import data programmatically using SAS code. See the example code below.

```
caslib espsrv datasource=(srctype="esp" server="host.example.com"
port=55555)
  description="SAS ESP Server port 55555";
proc cas;

loadStreams.loadSnapshot /                                    /* 1 */
      casLib="espStatic"                                      /* 2 */
      espUri="trades_proj/trades_cq/Trades"                   /* 3 */
      casOut={caslib="mycas",name="TradesSnapshot"};          /* 4 */

table.fetch /                                                 /* 5 */
      table={caslib="mycas",name="TradesSnapshot",            /* 6 */
      vars={'tradeID*','security','quantity','price','traderID','time'}}
      index=false;

run;
```

## ADDITIONAL FEATURES

Other notable new features include the ability to import multiple worksheets from Microsoft Excel and append multiple delimited files into a single table. These features are illustrated below.

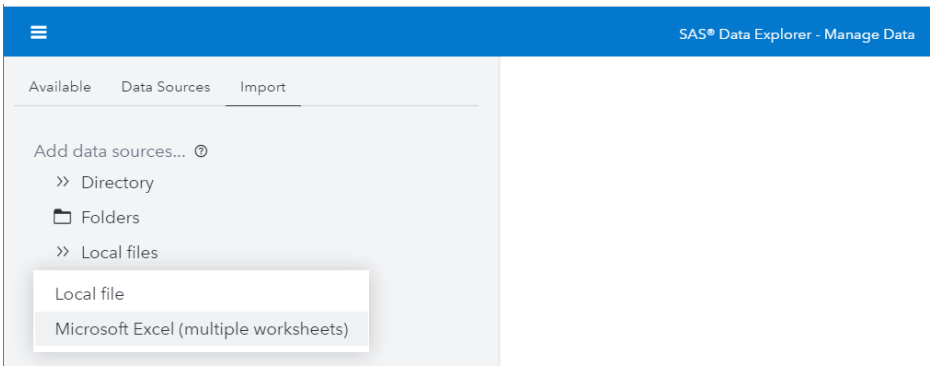


Figure 10: Other Import Options Available

SAS Drive is an interface that lets you manage your content in Viya®. When you create content in SAS Folders, it is available for you to view, open, and organize in SAS Drive. You can also upload small files to store and share in SAS Folders through the upload action.

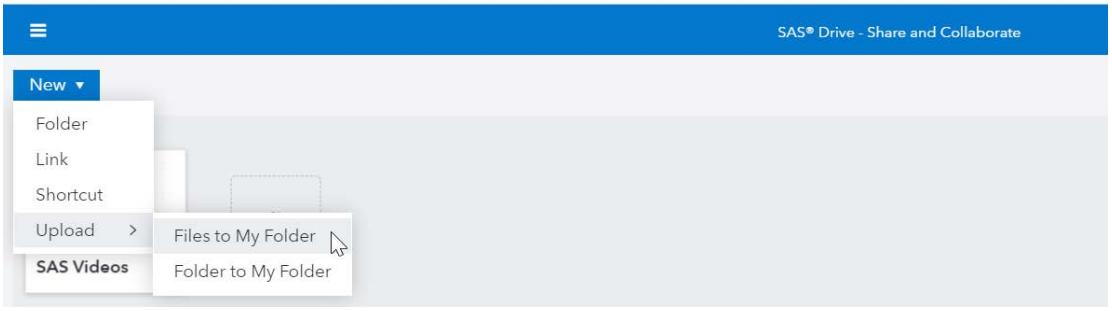


Figure 11: File Upload Example in SAS Drive

You can import files stored in SAS Drive from the import tab in the Choose Data dialog box or in SAS® Data Explorer. On the import tab, select Import Folders to bring up the Folder selection dialog box and select the File to import.

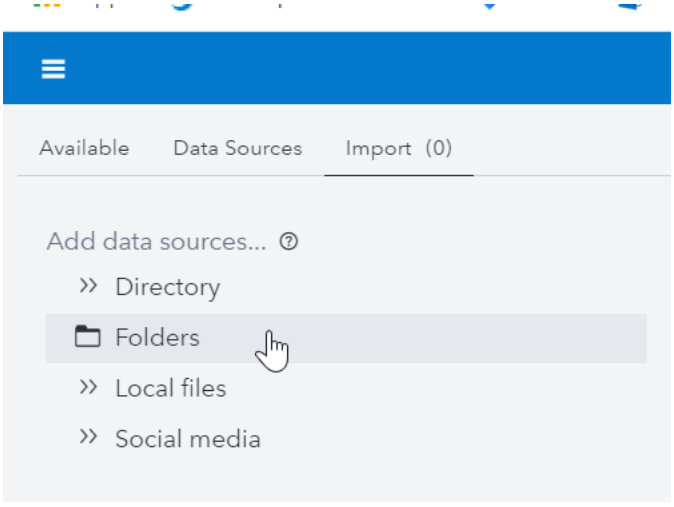


Figure 12: Importing Files from SAS Drive Storage

Another feature is the ability to download sample data to your local machine. You can choose the number of rows to download and the delimiter. This feature is accessed by using the Actions menu on the top right corner of the sample data tab. It is also available in SAS Data Studio, accessible from the prepare data action on the left navigation bar. Download can be disabled by using a permission in SAS® Environment Manager.

The following example illustrates this feature.

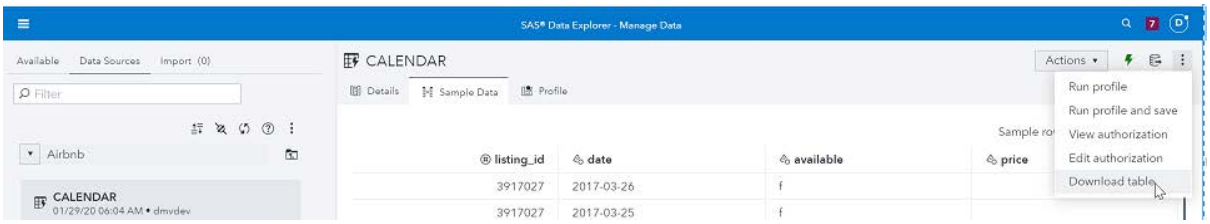


Figure 13: Download Content Option

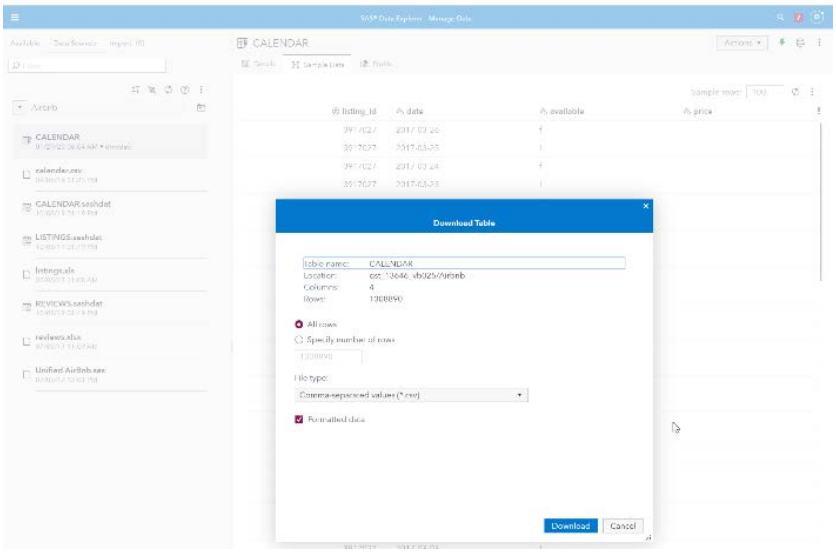


Figure 14: Download Options

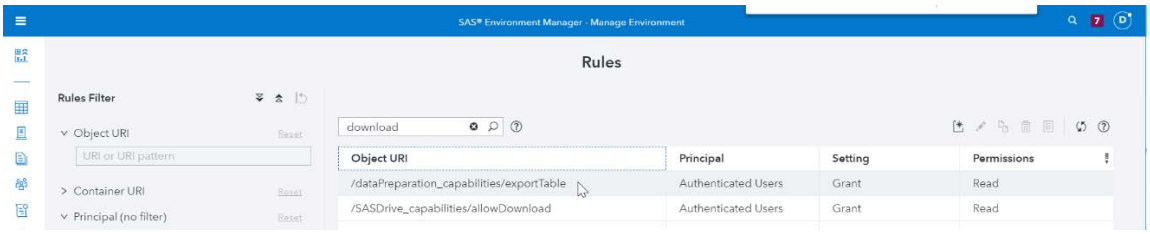


Figure 15: Permission Settings that Control Download

Additional new features include the ability to move files, including directories of files, to and from cloud storage using SAS Cloud Data Exchange.

Finally, note that there are a variety of connectivity and other types of examples, hints, and tips to be found on the SAS GitHub location at https://github.com/sassoftware.

## DATA PREPARATION

SAS® Data Preparation is an interactive, self-service product that allows you to access, blend, shape, and cleanse data to prepare it for reporting or analytics. You can use many of the data connectivity features, transform data in an interactive environment, view integration through lineage, and govern and manage the data life cycle, all through point and click interfaces that do not require programming knowledge.

There are many new features in the latest release. Here is a brief summary:

- Many usability enhancements including defer "run on open" action

- Insert, reorder, and remove steps in Data Plans in SAS Data Studio
- New Data Quality Transformations: Manage Columns and Remove Duplicates rows
- Support for simple random partitions of tables in addition to stratified partitioning
- Ability to add and view both column names and column labels

Figure 16 below illustrates some of these new features.

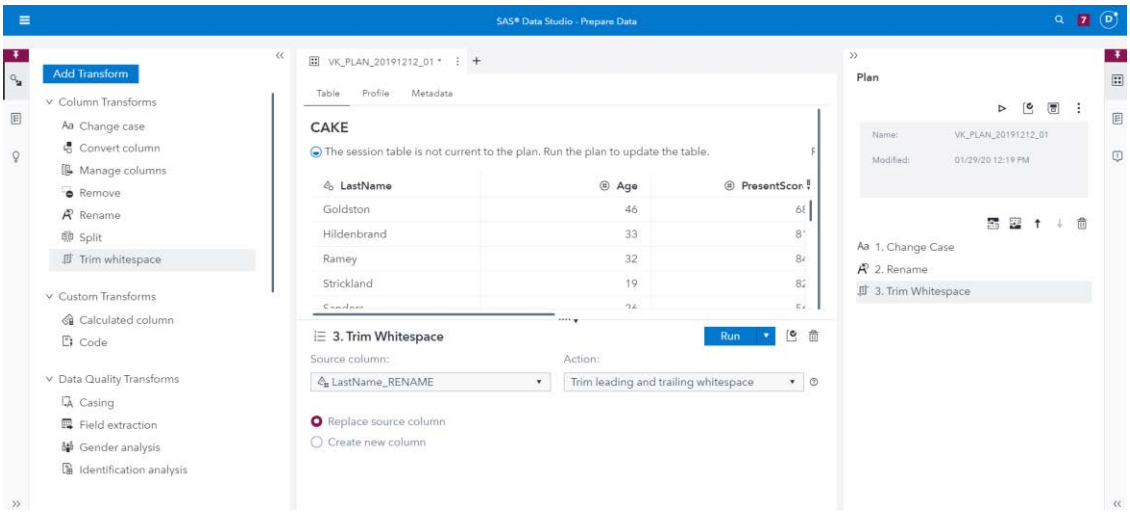


Figure 16: Updated Data Studio Features

## ARTIFICIAL INTELLIGENCE FOR DATA MANAGEMENT

Automation through AI can help to minimize the effort involved in basic or repetitive tasks, freeing you to focus on the actual problems you are trying to solve. SAS is investing in the use of AI to provide automation assistance in most of its product suite, including the automation of data management tasks. In SAS Viya 3.5, the new suggestions feature in SAS Data Studio, available with the SAS Data Preparation license, uses machine learning models to analyze your data and suggest transforms based on an analysis of problems found in your data set. You can choose to apply the suggested transforms to your data.

The suggestions button is in the top left corner of SAS Data Studio. You can also invoke the AI engine as an action on a specific column. The engine analyzes your data and provides transformation recommendations to clean and enrich your data. You can choose which transforms you want to apply. As you transform the data, you can continue to reanalyze, and the AI engine will perform the analysis on the updated results.

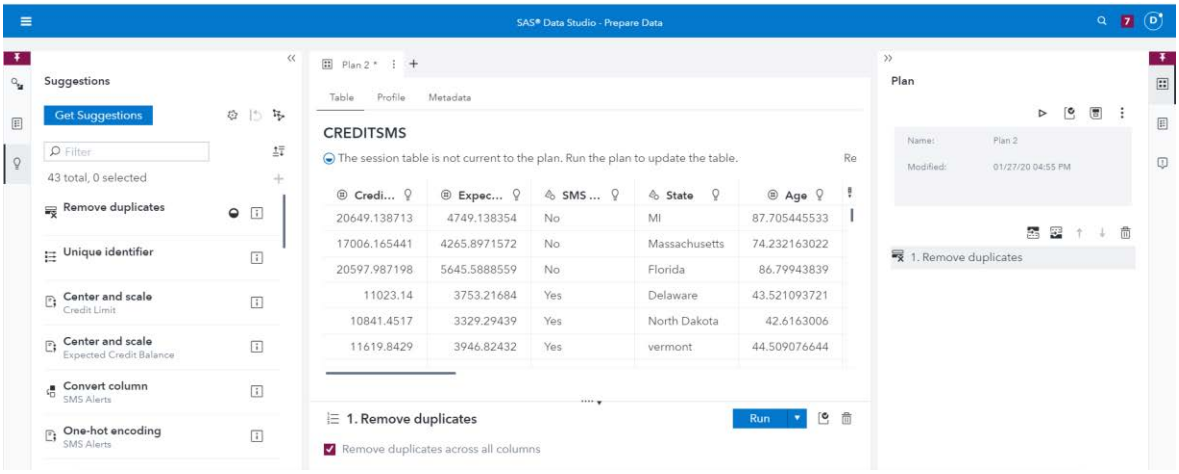


Figure 17: Suggestions AI Feature Example

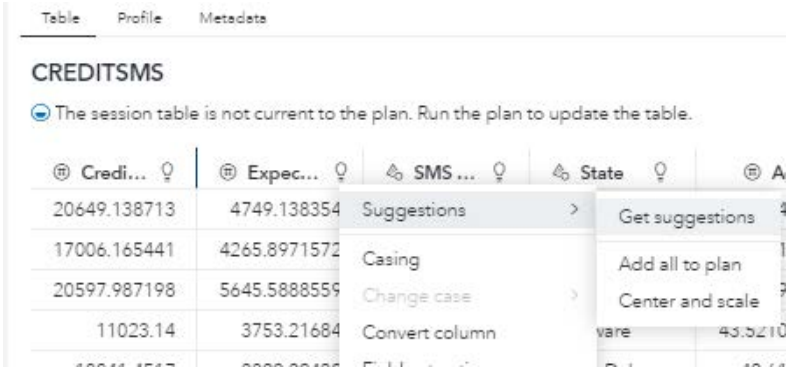


Figure 18: Suggestions are Also Available as a Per-column Action

Figure 19 shows some of the AI models and transformations available in the engine. Some models invoke existing transforms available in SAS Data Studio, and others generate code to perform actions. Models exist for data standardizing, enrichment, duplicate rows detection, one-hot encoding, center and scale, imputation, and other actions.

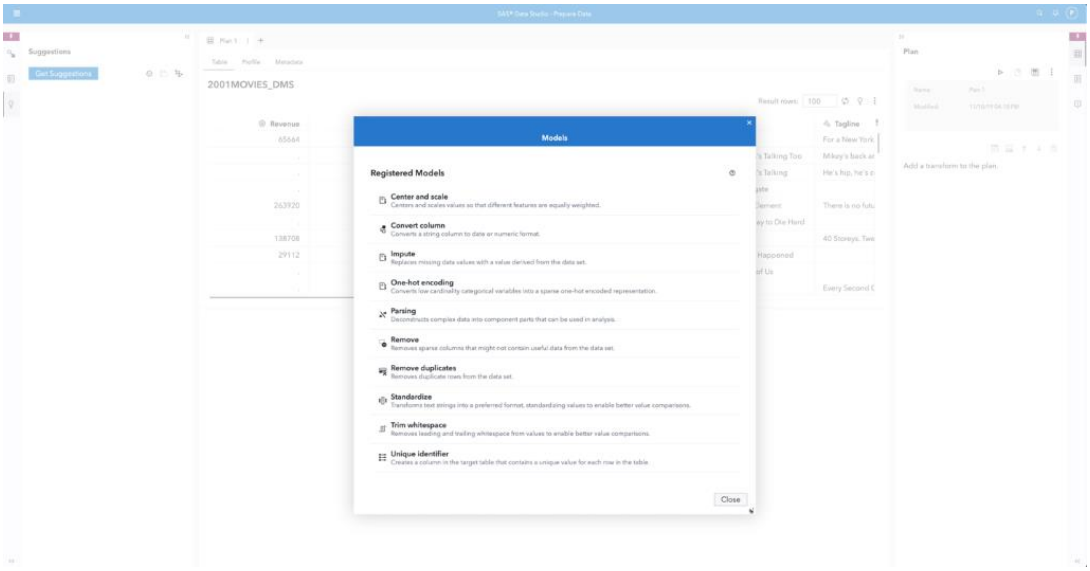


Figure 19: Some of the Available AI Suggestions

A feedback system exists in the AI engine that allows it to improve recommendations over use. As you select actions to perform on the data, the engine learns from those actions and increases the likelihood of making similar suggestions on similar data in the future. This is illustrated in the figure below.

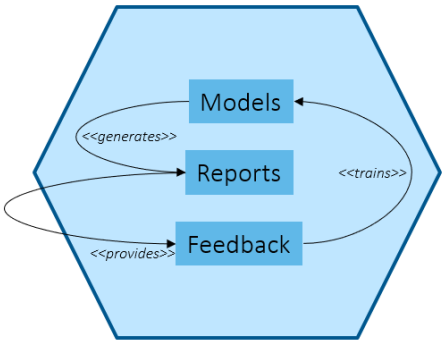


Figure 20: AI Feedback Loop

So how does the AI engine work? The engine is based on a Tree-Heuristic algorithm using Contextual Bandit Decision Trees (Elmachtoub, 2017). The algorithm can be summarized as follows. Suppose there are k data actions we want to suggest. We maintain k datasets that each correspond to a data action. The dataset corresponding to data action i contains observations where action i was suggested. An observation is the feature vector output by the characterization task for a column along with a value y=1 if the suggestion was used and y=0 if not. The datasets do not necessarily need to have the same schema. They only need to contain the relevant information for suggesting the data action. Then k decision trees are trained using the k datasets. We prune the decision trees to avoid overfitting. We then deliver out-of-the-box pre-trained models for a variety of common quality problems. In future releases, you will be able to extend the model set to add your own actions.

## DATA QUALITY

There are many new enhancements to SAS® Data Quality. One important addition is the ability to call data quality functions from more runtime environments and languages. For example, the DQ functions are now callable from the SAS DS2 language. The following code snippet illustrates how to call the functions from DS2:

```
%DQLOAD(DQLOCALE=(ENUSA),
DQSETUPLOC='/opt/sas/spre/home/share/refdata/qkb/ci/31
');

proc ds2 libs=work;
data _null_;

  /* init() - system method */
  method init();
    declare varchar(255) name; /* method (local) scope */
name = 'Doctor Sam Spade';
stdName=dqStandardize(name, 'Name');
    put stdName;
  end;
enddata;
run;
quit;
```

Data quality functions are currently available in SAS® Business Rules Manager and will soon be available in the SAS® Micro Analytic Service. Figure 21 is an example of some of the many DQ functions available. A complete list of the available functions can be found in the references at the end of this paper.

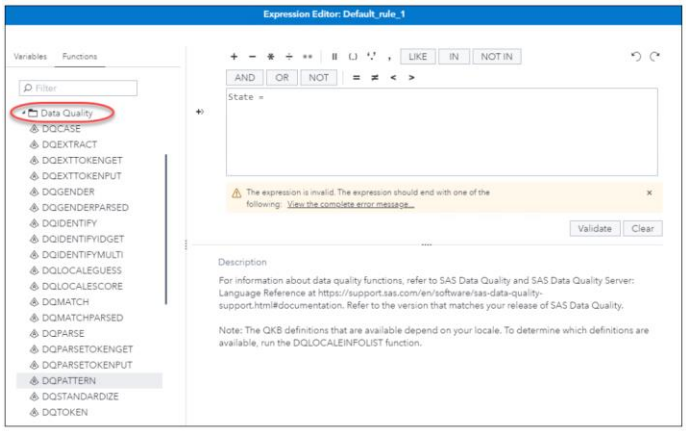


Figure 21: Some of the Many Available Data Quality Functions

The new PROC DATAMETRICS supports data profiling capabilities in SAS code. You can use the PROC to recognize patterns, identify sparsity in the data, generate frequency distributions, and calculate basic statistics.

You can also invoke identification analysis to guess the content of the data. For example, you can use identification analysis to identify personal information that may be contained in your data such as name, phone numbers, email, and various government ID values across many different languages. Additional columns are added to the data to store the classification result. The following code is an example of how to call the new proc.

```
proc datametrics data=sashelp.demographics out=my_results
frequencies=100 minmax=12 threads=8 median format;
identities qkb='/opt/sas/spre/home/share/refdata/qkb/ci/31' locale='ENUSA'
def='Field Content' multiidentity;
variables name GNI;
run;
```

The following table shows an example of the output.

| Input | Identity |
|---|---|
| William Smith | NAME |
| 500 SAS Campus Drive | ADDRESS |
| +1 (919) 677–8000 | PHONE |
| 123-456-7890 | Government ID (SSN) |

Another new feature is cross-field clustering support. This is best illustrated with an example. Suppose you want to cluster the following values in a table:

| Name | HomePhone | MobilePhone | DOB |
|---|---|---|---|
| Michael T Smith | 919-887-0099 | 919-766-5566 | |
| Mike Smith | 919-887-0099 | | 31OCT1960 |
| Michael Smith | | 919-766-5566 | 10-31-1960 |

Here is a code example illustrating this feature:

```
proc cas; session SASCAS1;
  entityres.match /
      algorithm="Auto"
      clusterIdType="Int"
      clusterId="clustId"
      MatchRules= {
    {Rule={{columns={'Name', 'HomePhone'}},
           {columns={'Name', 'MobilePhone'}},
          }
},
    {Rule={{columns={'Name', 'HomePhone', 'DOB'}},
           {columns={'Name', 'MobilePhone', 'DOB'}}
          }
    }
}
      inTable={name="customers"}
      outTable={name="recognized"}
      ;
  run;
```

The result is a table that identifies all three records as the same person. You can use this technique to match and identify duplicate or similar records in a dataset.

Other enhancements in data quality include updates to the SAS® Quality Knowledge Base (QKB) for Contact Information (CI):

- QKB Field Name and Field Content definitions updates for GIVEN NAME, FAMILY NAME, and MONTH
- QKB Field Content definitions have been added for the United Kingdom and Norway
- QKB locale updates for Argentina, Singapore, and Japan

## DATA INTEGRATION

### DATA FLOWS

Data flows in SAS Viya is a key feature that is under development. This feature will support complex data flow processing, such as data movement, transformations, and preparation. Figure 22 is an example of the data flow editor. The data flow editor is part of the upcoming release of SAS Studio and includes hundreds of transforms, tasks, and code snippets to support data integration activities.

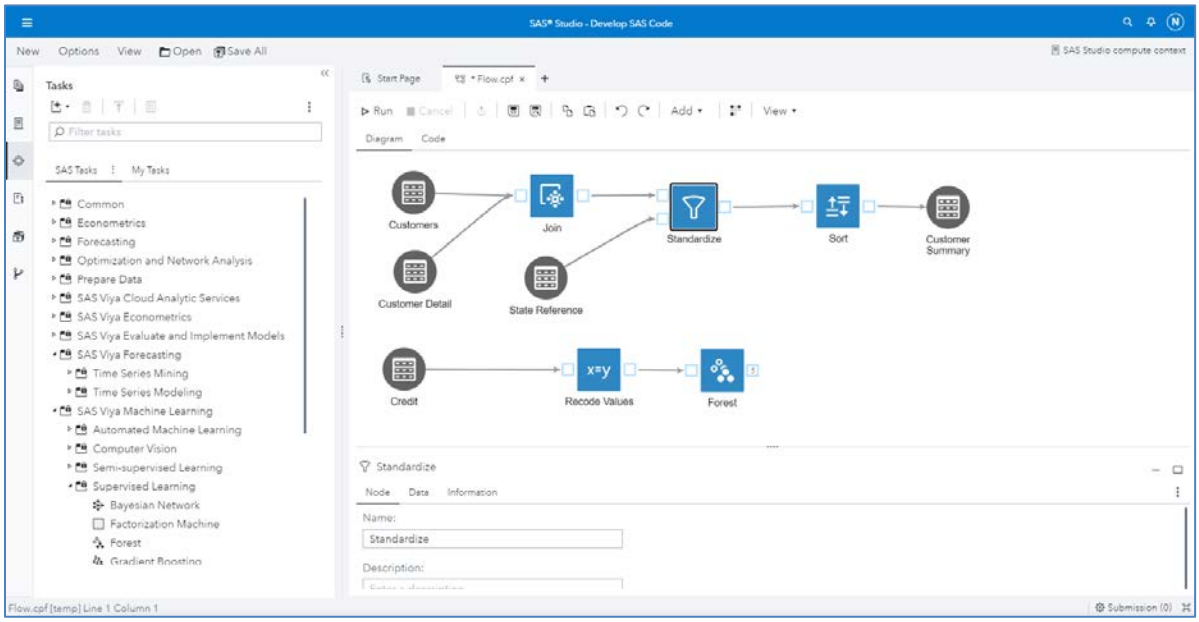


Figure 22: Example of Data Flow in SAS Viya

## JOB FLOWS

SAS Viya 3.5 now supports the ability to create job flows. A job flow is a group of jobs, dependencies, and conditions that are organized in a sequence and can be scheduled or executed as a group. Flows contain jobs, gates and events, and order is indicated with connections. You can add logic gates such as AND/OR gates, create time-based triggers, and schedule a flow. In the future, data flows will also be able to be added to job flows to create increasingly complex jobs.

Job flows are surfaced in SAS® Environment Manager. Figure 23 below shows the job flow interface.

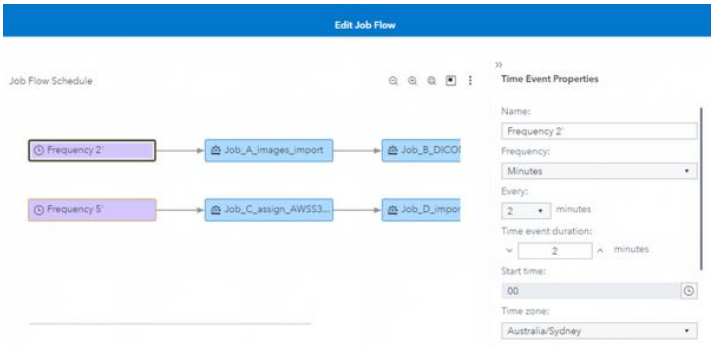


Figure 23: Example of Job Flows in SAS Viya

Jobs are linked through connections to form dependencies. You can modify the job dependencies to support different actions based on job run status, including checking for exit codes and expected maximum or minimum run time.

There is also support for scheduling flows. The scheduler supports time-based trigger events. Figure 24 is an example of the schedule manager view.

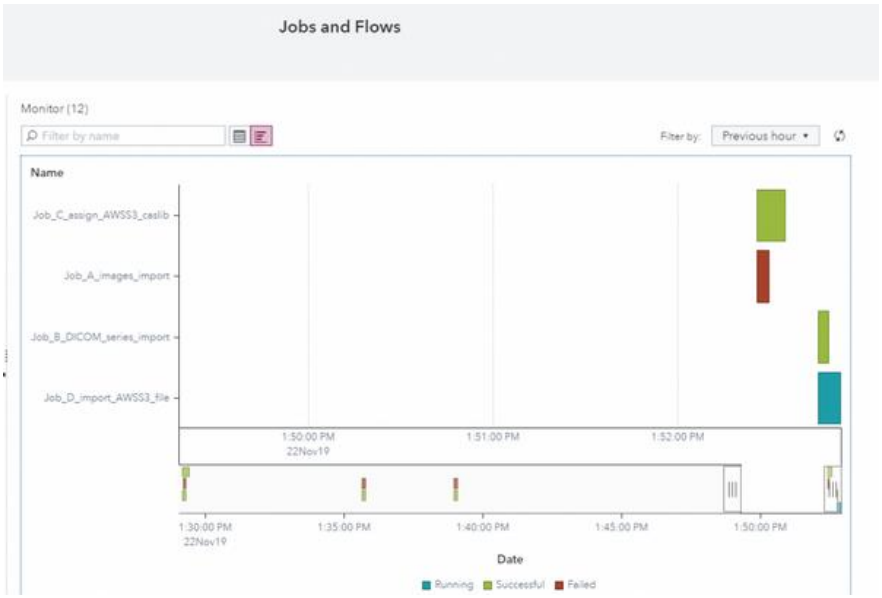


Figure 24: Integrated Scheduler

### STAR SCHEMA JOINS

New in the latest release of the SAS® Cloud Analytic Services (CAS) server is support for star schema joins using views. A star schema is a join between one fact table and one or more dimension tables. Each dimension table is joined to the fact table through a single column key.

You can use this type of view to improve query performance and reduce data movement across CAS workers. Performance improvements have been shown to be as much as 70% faster when using this type of join compared to a standard join with large tables.

The code snippet below shows an example of how to code a star schema join view.

```
proc cas;
table.view / promote=true name='mail_view_opt'
tables={
{name='mailorder_big_new', varlist={'qty'}, as='m'},
{keys={'m_pcode = p_pcode'},
name='products_big_new_rep',  varlist={'price', 'cost', 'descrip', 'type'},
{keys={'m_custnum = c_custnum'},
name='customers_big_new',
varList={'addr1','addr2','city','country',
'region','state','zip'}, as='c'},
{keys={'m_catcode = cat_catcode'},
name='catcode_rep', varlist={'catalog'}, as='cat'}
};
quit;
```

## CONCLUSION

The latest releases of SAS Data Management provide enhancements to help you carry out data-oriented processes more efficiently and with greater control and flexibility. Enhancements have been made in many areas. You can find many reasons to upgrade to the latest versions of the SAS Data Management software.

## REFERENCES

Ebersole, B. 2019Image Processing with SAS Viya. SAS Communities Library. Accessed June 21, 2019. Available https://communities.sas.com/t5/SAS-Communities-Library/Image-Processing-with-SAS-Viya/ta-p/568082.

Elmachtoub, A. N., R. McNellis, S. Oh, and M. Petrik. 2017. "A Practical Method for Solving Contextual Bandit Problems Using Decision Trees". *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Rausch, N. 2019. "What's New in SAS Data Management". *Proceedings of the SAS Global Forum 2019 Conference*. Cary, NC; SAS Institute Inc. Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3801-2019.pdf.

Teleuca, B. 2020. "Steer Your Hybrid SAS® Viya®/SAS® 9 Ship Toward the "Governed Data" Port". *Proceedings of the SAS Global Forum 2020 Conference*. Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/3801-2020.pdf

Teleuca, B. 2019. Go with the Job Flow in SAS Viya 3.5. SAS Communities Library. Available https://communities.sas.com/t5/SAS-Communities-Library/Go-with-the-Job-Flow-in-SAS-Viya-3-5/ta-p/612925

SAS Institute Inc. 2020. GitHub SAS Software. Accessed February 3, 2020. Available https://github.com/sassoftware/sas-access-samples.

SAS Institute Inc. 2019. What's New in SAS 9.4 and SAS Viya. Accessed December 20, 2019. Available https://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.5&docsetId=whatsnew&docsetTarget=n0x40wef7ky2b4n1bmn674w1om0d.htm&locale=en.

SAS Institute Inc. 2019. SAS Event Stream Processing Connectivity Sources. Accessed February 3, 2020. Available https://go.documentation.sas.com/?docsetId=espca&docsetTarget=p1swscq8yglnunn1p44y46w2rjtx.htm&docsetVersion=6.1&locale=en.

## ACKNOWLEDGMENTS

The author wishes to thank Mary Kathryn Queen, Bogdan Teleuca, JP Trawinski, Jeff Stander, and Vincent Rejany for their contributions.

## CONTACT INFORMATION

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

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