Paper SAS3801-2019

What's New in SAS® Data Management

Nancy Rausch, SAS Institute Inc.

ABSTRACT

The latest releases of SAS® Data Management provide a comprehensive and integrated set of capabilities for collecting, transforming, managing, and governing 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, and streaming. You also have the ability to perform extract, transform, load (ETL) and extract, load, transform (ELT) processes in diverse run-time environments such as SAS®, Hadoop, Apache Spark, SAS® Analytics, cloud, and data virtualization environments. There are also new artificial intelligence features that can provide insight into your data and help simplify the steps needed to prepare data for analytics. 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

One important new feature for data connectivity is the introduction of Cloud Data Exchange (CDE). CDE is a data connection capability offered in SAS® Data Preparation on SAS® Viya®. CDE can transfer high-volume data securely between on-premises data sources and the cloud. CDE stores on-premises data source credentials in a secured vault, so these never have to be stored or accessed outside the on-premises firewall. Figure 1 is a high-level diagram of CDE.

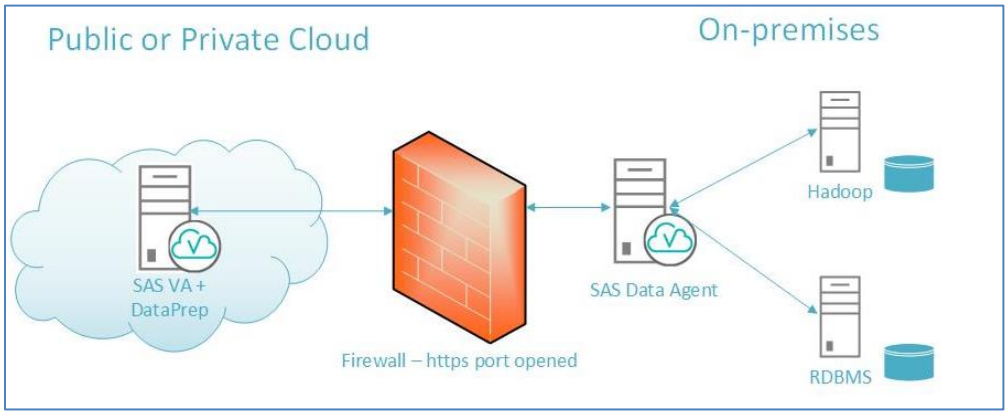


Figure 1: Cloud Data Agent Architecture

CDE offers several advantages for users that need to move data to and from cloud storage. One is that it enables you to open one secure port through your firewall rather than the many that would be required if you had multiple data connectors, which helps with security.

Another is that CDE's components can be placed in an organization's perimeter network. This enables you to monitor network traffic and provides isolation and protection of IT resources. Figure 2 shows an example of a system architecture before and after using CDE.

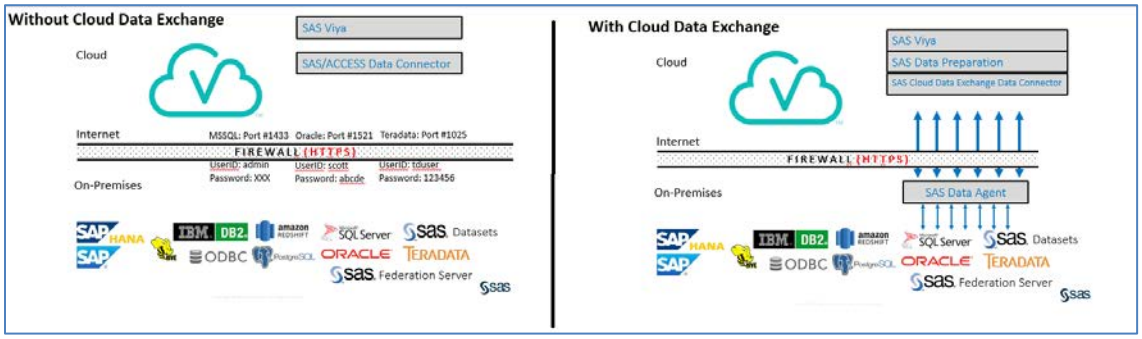


Figure 2: Firewall Port Requirements without and with Cloud Data Exchange

CDE also supports data connectors to many data sources including Apache Hive, Oracle, DB2, Teradata, PostgresSQL, SAS, Microsoft SQL Server, Open Database Connectivity (ODBC), Amazon Redshift, and others. Data can be transferred in serial or parallel, and there are several parallel transfer modes you can use. The following code shows how to set up a SAS library connection using CDE to SAS® Cloud Analytic Services (CAS).

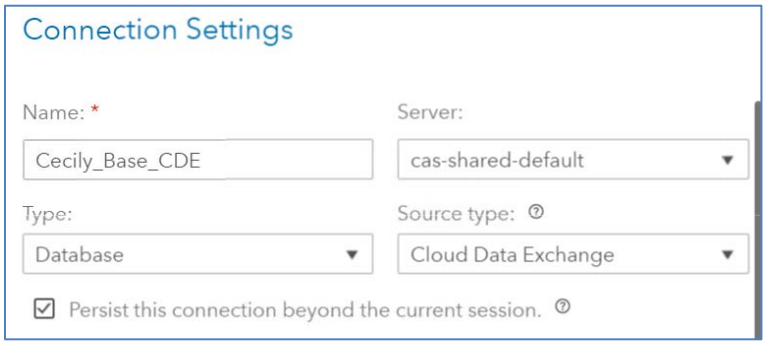
```
CAS mySession  SESSOPTS=( CASLIB=casuser TIMEOUT=99 LOCALE="en_US"
metrics=true);

/* Create a Session based Hive_CDE CASLIB */
proc cas;
  session mySession;
  action addCaslib / caslib="hive_cde"
             datasource={srcType="clouddex"
                         username="viyademo01", password="lnxsas",
                         dataAgentName="dagentsrv-shared01-default",
                         schema="CASDM",
                         conopts="dsn=hivesrv1"
                        };
run;
quit;
```

The following code shows an example of both serial and parallel load options for moving data.

```
/** Serial data Load to CAS */
proc cas;
    session mySession;
    action loadTable / casLib="hive_cde" path="cars"
               casout={ casLib="hive_cde" , name="cars_srl" } ;
run;
quit;

/*** Parallel data load to CAS */
   proc cas;
    session mySession;
    action loadTable / casLib="hive_cde" path="cars"
          datasourceOptions={numReadNodes=3, /* this triggers parallel
load*/
```

```
            useMetaTable=FALSE, /* you can control split options if you want
*/
            useMinMaxToSplit = TRUE,
            splitColumn="msrp",
           splitRange = 20000,
           traceFile="/tmp/cdetrace.log",
           traceFlags="SQL"
                        }
                casout={ casLib="hive_cde" , name="cars_splt" } ;
   run;
   quit;
```

There is also a command-line interface (CLI) for managing the CDE server, and CDE is available as a data connection type in the library connection windows in SAS Viya applications, as shown in Display 1.



Display 1. Cloud Data Exchange Connection Settings

There are also a number of other new data connectivity features. One new feature is the ability to load data to and from an Amazon Simple Storage Service (Amazon S3) location. Supported file formats are SASHDAT and CSV. The source type is S3, and you can specify the Amazon bucket in the connection string. Here is some example code that creates a connection to S3 sources.

```
/* Create a Session based CASLIB */
proc cas;
  session mySession;
  action addCaslib / caslib mys3 datasource=(
   srctype="s3",
   bucket="mybucket",
   awsConfigPath="/home/myuser/.aws/config",
   awsCredentialsPath="/home/myuser/.aws/credentials",
   region="US_East",
   objectPath="/mypath/"
) ;
run;
quit;
```

There are also new connectors to Java Database Connectivity (JDBC), MySQL, Spark, and Vertica for both CAS and the SAS®9 environments. JDBC is particularly handy if you have a JDBC driver for your database. Here is an example using the JDBC library to connect to a PostgresSQL database in CAS.

```
/* Create a Session based CASLIB */
proc cas;
  session mySession;
```

```
  action addCaslib / caslib jdbcpg desc='JDBC PostgreSQL'
   dataSource=(srctype='jdbc',
               url="jdbc:postgresql://mydb.example.com:5432/casdm",
               authenticationdomain="PGAuth",
               schema="public",
               class="org.postgresql.Driver",
               classpath="/opt/sas/jdbc") ;run;
quit;
```

HADOOP INTEGRATION AND CONNECTIVITY

One of the important new features in Hadoop integration is enhanced support for Apache Spark. Spark is a Hadoop technology with an in-memory data processing engine. There is a new SAS access engine in SAS 9 and a SAS® Data Connector to Hadoop. In addition, SAS® In-Database Code Accelerator for Hadoop supports code running in Hadoop as either MapReduce or in Apache Spark. This means that you can run SAS programs inside of Spark just as you can when using MapReduce. You can also transfer data in serial or in parallel by using SAS® In-Database Technologies for Hadoop. A new system option has been added that specifies the run-time environment platform to use.

```
HADOOPPLATFORM=MAPRED | SPARK;
```

You can also use mapped SAS functions that run in-database when running in Spark, like you can do with other databases using SAS. This is a good performance tip because non-mapped SAS functions used during SQL implicit pass-through pulls rows out of Spark and into SAS for processing. Using mapped functions avoids this problem.

Another enhancement is in the TRANSPOSE procedure. This procedure can also be executed as either a MapReduce job or in Spark.

Following are several examples of connecting to Spark.

```
/* SAS9: This example uses the default Spark port and schema.*/
proc sql;
   connect to spark (user="myusr1" pw="mypwd1"
      server=sparksvr port=10016 schema=default);

/* SAS9:  This example specifies the Hive Kerberos principal to connect to
a
Kerberos secured HiveServer2 instance. It assumes that a Kerberos
'kinit' has been successfully performed.*/
proc sql;
   connect to spark (hive_principal='hive/_HOST@SPK.COMPANY.COM'
      server=sparksvr);

/* CAS:  create a connection between your Spark data source and SAS Cloud
Analytic Services using parallel mode.*/
caslib sparkcaslib desc='Spark Caslib'
   dataSource=(srctype='spark',
               dataTransferMode="parallel",
               hadoopjarpath="/hadoop/jars:/hadoop/spark/jars",
               hadoopconfigdir="/hadoop/conf",
               username="hive",
               server="thriftserver",
               schema="default");

/* Load a Spark data source using PROC CASUTIL.*/
proc casutil;
```

```
   incaslib="sparkcaslib" sessref=mysess;
   load incaslib="sparkcaslib" casdata="cars"
      casout="cars_CAS"
      options=(dataTransferMode="serial");
run;
quit;
```

The SAS In-Database Code Accelerator for Hadoop has a few other useful enhancements:

- support for the SCRATCH_DB option for a Hive database that is used when a temporary table is created
- SQL queries using a WHERE IN clause are now supported

SAS® Data Loader for Hadoop has also been enhanced. There is new support for Apache Hive High Availability, and you can load data to and from CAS using the Copy data to and from Hadoop directives. High availability (HA) support was added by generating the appropriate hive connection string based on the Hadoop configuration files and generating the URI= option to pass to SAS.

DATA INTEGRATION

Git is a popular open-source distributed version control system. SAS continues to enhance integration with open-source technologies such as Git. The latest release of SAS Data Integration Studio includes integration for storing content in Git. SAS Data Integration Studio previously supported Concurrent Versions System (CVS) and Apache Subversion (SVN). Since there are now three options for version control support, you need to choose which one you want to use because you can use only one version control system at a time. To configure Git integration, you need to remove the plug-in folders for the other two types. To do that, go to the SAS Data Integration Studio install location, typically C:\Program Files\SASHome\SASDataIntegrationStudio\4.7\plugins, and copy or rename the following two folders to some other place. You should copy or rename them in case you want to re-enable them again.

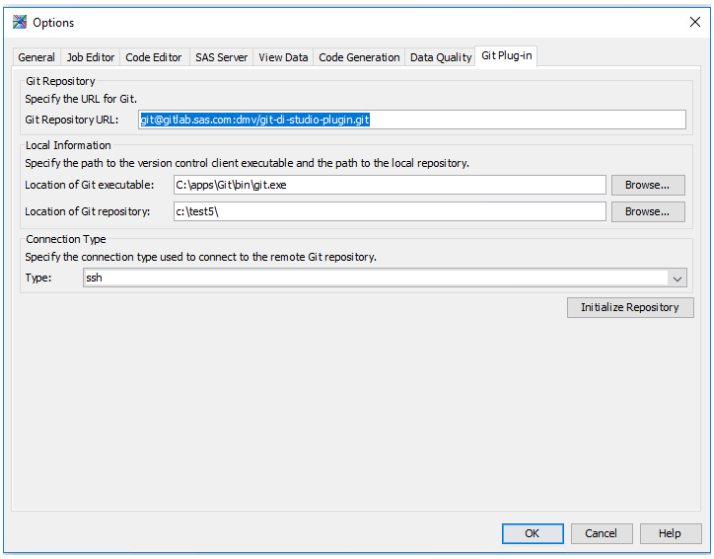
```
sas.dbuilder.versioncontrol.cvsplugin
sas.dbuilder.versioncontrol.svnplugin
```

You can now use the Git version control system. You might need to set up a secure connection to use Git. One way to do this is to connect using Secure Shell (SSH). For an SSH connection, you need to generate RSA keys and store them in C:\Users*userid*\.ssh.

You can generate these keys by signing in to Gitlab, selecting your name, and navigating to Settings>SSH Keys.

You can also connect via HTTPS if you prefer. You are generally prompted on first use of the Git plug-in for a user ID and password.

Display 2 shows an example of the configuration screen for the Git plug-n in SAS Data Integration Studio using SSH. The settings are the same for HTTPS, except for the connection type.



Display 2. Connection Setting for GIT Connector

The Git repository URL is the location in Gitlab where you want to store your source code. One easy way to get this URL is to go to the location in your browser, copy the URL to the clipboard, and paste it into the field, as shown in Display 3.

Display 3. URI Location of GIT Connector

The local repository is the folder on your system where you want to store and retrieve content. Once you have configured the panel, click the Initialize Repository button, which tests the connections to make sure you have connectivity. If you have connectivity problems, one thing to check is whether the SSH keys and known_hosts files exist in your user folder on your local machine. You can also look at the logs to determine whether the error message is coming back from Git. On Windows, the logs are in a place similar to this location: C:\Users*youruserid*\AppData\Roaming\SAS\SASDataIntegrationStudio\4.904\.

Another new feature in SAS Data Integration Studio is enhanced support for Oracle hints when SAS Data Integration Studio generates explicit SQL pass-through code. SAS Data Integration Studio already supported hints in implicit pass-through code and has added support for explicit pass-through code. There is a new PRESERVE_COMMENTS connection option for the CONNECT statement, the EXECUTE syntax has changed to EXECUTE BY <dbms>, and the Oracle hints /* hints */ are escaped by the %str() function to enable macro processor support. Following is an example snippet of the code that SAS Data Integration Studio generates with this enhancement. This example is from the SQL delete transform-generated code:

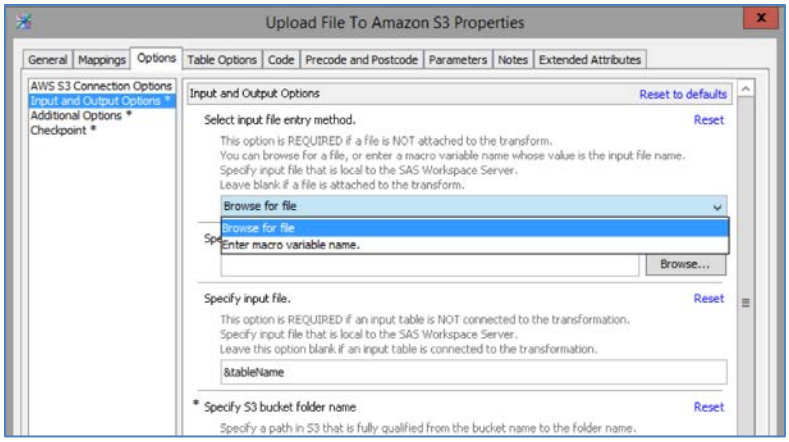
```
proc sql;
    connect to ORACLE
    (
        PATH=exadat12c AUTHDOMAIN="oracle_exadata_authdomain"
        PRESERVE_COMMENTS
    );
```

```
      execute by ORACLE
      (
   delete %str(/)%str(*)+ PARALLEL(CUSTOMER_DAY_1) %str(*)%str(/) from
"DMTEST"."CUSTOMER_DAY_1"
      );

      %rcSet(&sqlrc);

      disconnect from ORACLE;
   quit;
```

Support has also been added to support two new Oracle BULKLOAD options. One, the BL_DEFAULT_DIR= option, enables you to redirect where SAS writes all files that it creates during the bulk-loading process. Another enables you to add options in a text field to support any additional options you might want to specify during the bulk load process.

There are a few other important new features in SAS Data Integration Studio:

- Four new database options were added to the SCD Type 2 transformation. These options enable you to specify one or more table options for the temporary work tables that are used to update the target database table. The specified table options are applied to the temporary work tables when explicit SQL pass-through update method is selected. These options can be used for Close Record, Match Record, New Record, and Type 1 Record work tables.
- A new option, Generate macro variables for external files, has been added to support transformations that allow external files as a source or target. These transformations include User Written Code, File Reader, File Writer, Hadoop File Reader, and Hadoop File Writer. If the option is set to Yes, then the code contains macro variables that provide information about the external file. You can use these macro variables in your own code to manipulate the files. When the option is set to No, the macro variables are not generated. The default is Yes.
- In the Cloud Analytic Services transformation, a new option enables you to add additional options for the LOAD DATA statement in PROC CASUTIL. For example, NCHARMULTIPLIER and TRANSCODE_FAIL options are available in the LOAD DATA statement. These options can be helpful when the local SAS session is running in an encoding other than UTF-8 and data needs to be transcoded before being sent to CAS. These options follow the same pattern as the CAS engine options of the same name.
- The Upload to Amazon S3 transform has been enhanced to add an option that enables you to browse for the input file name that is to be uploaded to S3. The option enables you to browse for a file, enter a file name, or paste a path and file name. You can also use a macro variable name whose value is the path and name of the chosen input file. This option is shown in Display 4.



Display 4. Amazon S3 Connection Settings Enhancements

- There are also several new source designers that have been added to support new access engines, including JDBC.

One useful new transformation capability in CAS is support in FedSQL for explicit pass-through queries. Explicit pass-through SQL lets you send SQL queries directly to the database for execution. In the CAS context, it means that you can run a specific SQL query inside the database, using the database syntax, and the results are loaded directly into CAS.

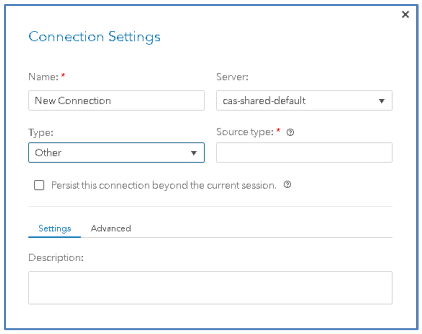
Following is a code example. The feature is triggered by the CONNECTION TO clause. The SQL enclosed in the CONNECTION TO clause is a database-specific SELECT-type query and must produce a result set. A data definition language (DDL) statement does not work.

```
/* Create a CASLIB */
caslib PG datasource=(srctype="postgres",authenticationdomain="PGAuth",
   server="mydb.example.com",database="dvdrental",schema="public")
libref=PG ;

proc fedsql sessref=mySession _method ;
   create table PG.myresults{options replace=true} as
      select film.title
   from connection to PG(select "title", "film_id" from "film");
quit ;
```

DATA PREPARATION

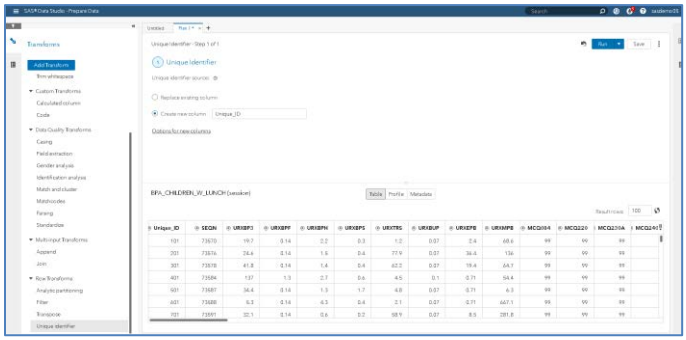
SAS Data Preparation provides an interactive, self-service set of products for users who need to access, blend, shape, and cleanse data to prepare it for reporting or analytics. The products integrate into SAS Viya and include features for connecting to and transferring data, transform data in an interactive environment, view integration through lineage, and govern and manage the data life cycle. The products also integrate with SAS Data Integration Studio. There are a number of new features in the latest release.

For data connectivity, support has been added for importing text content, which includes converting the text from a variety of formats such as PDF, txt, or png into tables. There are also enhancements in social media import connectors from sites such as Twitter and Facebook, and enhanced support for geocoding and geoenrichment using Esri. There is also a new feature (shown in Display 5) that enables you to define your own data connections if a preconfigured data connector does not exist, through a user-defined interface.



Display 5. User-Defined Connections Window

When working with a table, sometimes you need to create a unique identifier for each row. This is difficult in a distributed, parallel execution environment like CAS because a simple sequential key might repeat in different segments of your table. One new feature in SAS Data Preparation is the ability to generate unique identifiers for a distributed table. You can generate a unique identifier when importing files, and there is also a transform available in SAS Data Studio. Display 6 is an example of the transform.



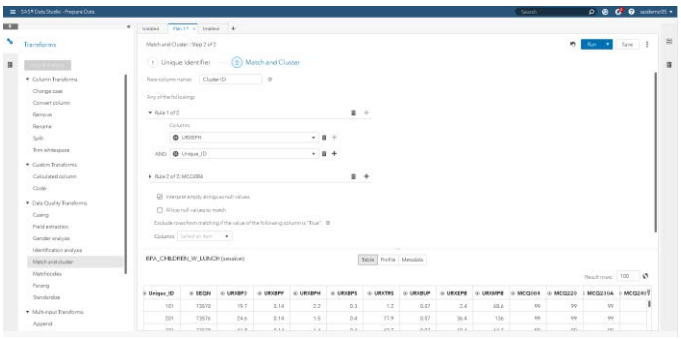
Display 6. Unique ID Transform

Another useful transform in SAS Data Studio is the new match and cluster transform. It supports creating clustering rules and conditions in CAS. This can be very helpful when your data contains multiple columns that you want to match together on. For example, if your data looks like that shown in Figure 3, you can build rules that enable you to match up the data on email and name.

name	email1	email2	
Alice	alice@alice.net	alice@alice.com	set 1
Alice	(null)	alice@alice.net	
Bob	bob@bob.com		set 2
Bob	robert@robert.com		

Figure 3. Example of Clustered Data

Display 7 shows an example of the transform.



Display 7. Example of the Cluster Transform

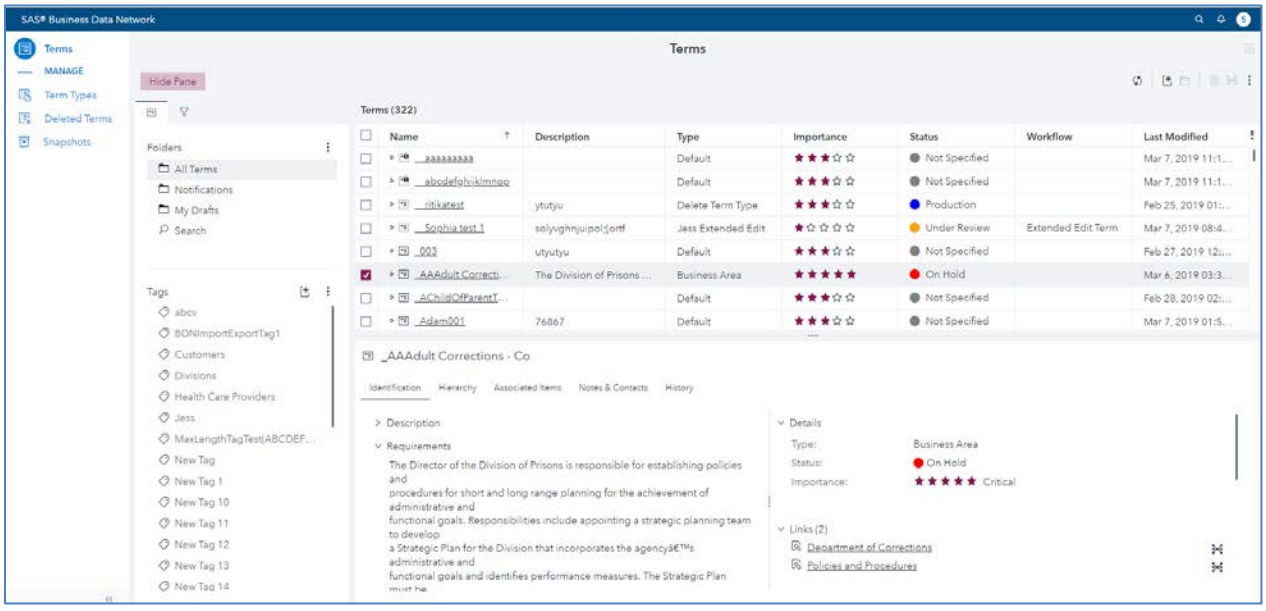
Other new features include the ability to create a segmented data partition of your data, enhanced support for right-to-left languages such as Arabic, and performance enhancements to support working with tables with thousands of columns.

UPDATED USER INTERFACES

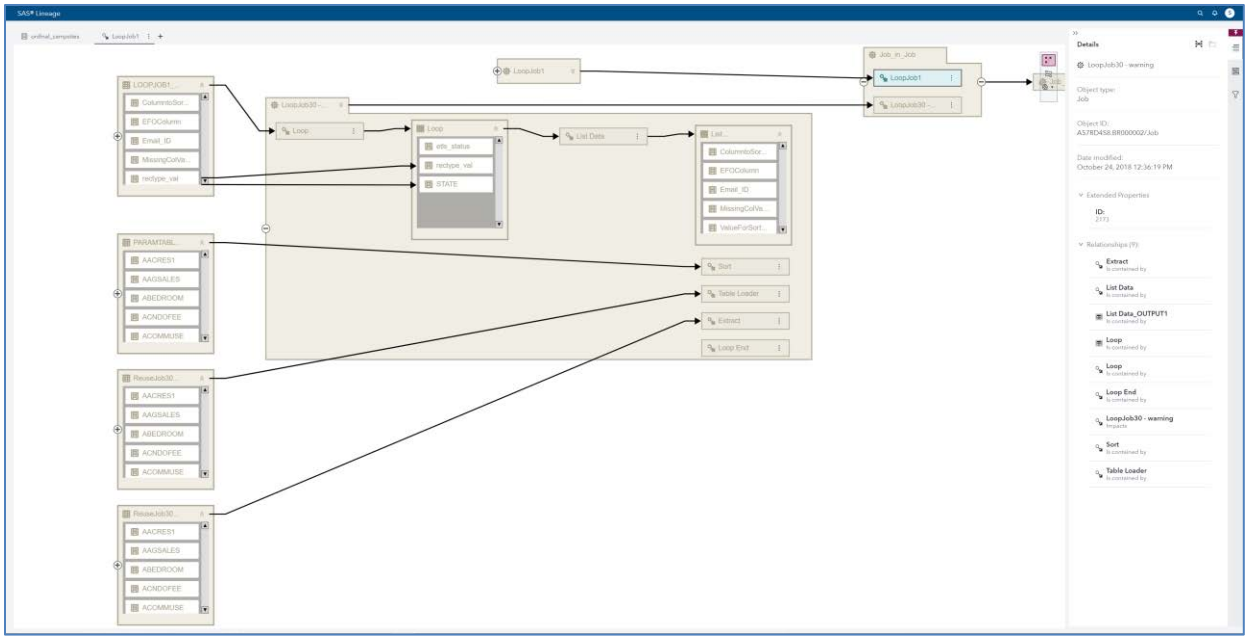
SAS is replacing existing user interfaces that rely on Adobe Flash with interfaces based on HTML5. SAS has upgraded the SAS® 9.4 server with maintenance release 6 (SAS® 9.4M6) to support these new clients. We are updating interfaces for many existing clients, including the following:

- SAS Business Data Network
- SAS Lineage
- SAS Data Remediation and Task manager
- SAS Reference Data Manager
- SAS Federation Server Manager

Display 8 and Display 9 show some examples of the new clients.

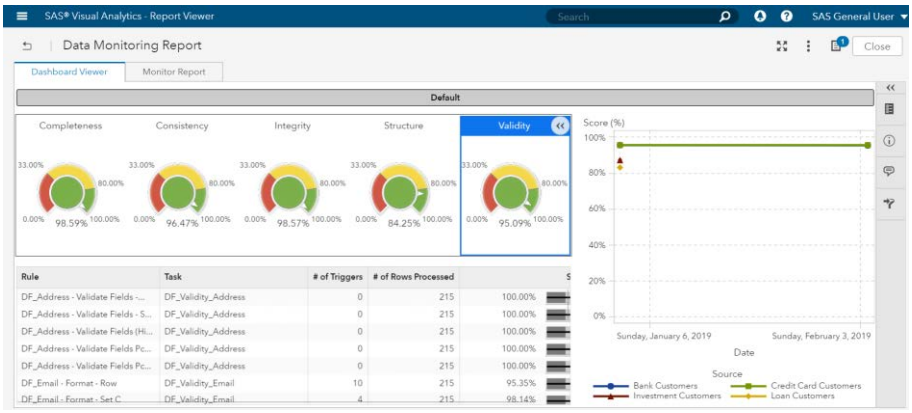


Display 8. Business Data Network in HTML5

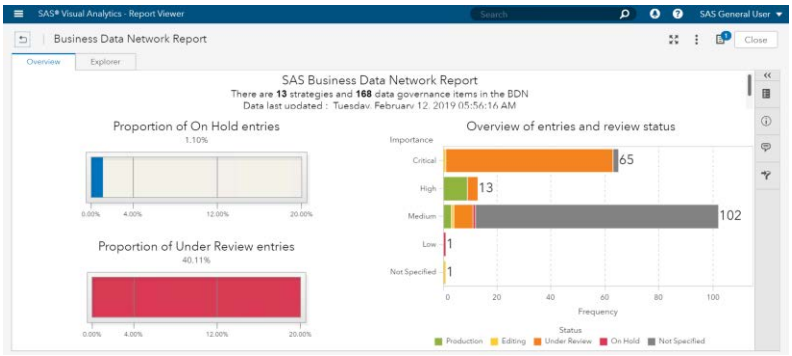


Display 9. Lineage in HTML5

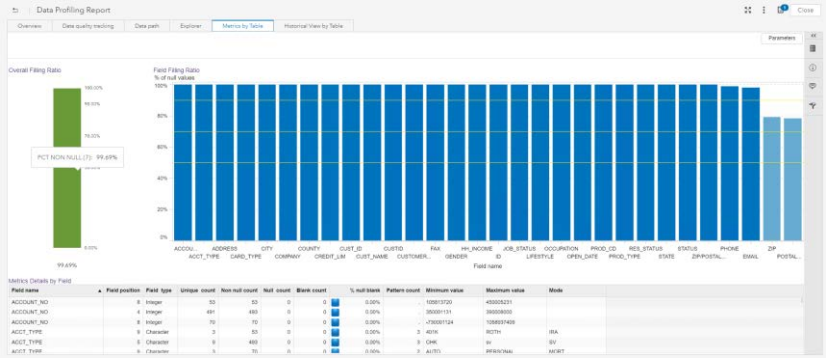
Additionally, new SAS® Visual Analytics reports are being developed to support data quality monitoring. The SAS® Visual Analytics reports are customizable to meet your specific needs. Display 10, Display 11, and Display 12 show some examples of these new reports.



Display 10. Dashboard UI Enhancements



Display 11. Business Data Network Report



Display 12. Profile Report Example

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 many areas. Customers can find many reasons to upgrade to the latest versions of SAS Data Management.

REFERENCES

Ghazaleh, David. 2019. "Execution of User-Written DS2 Programs inside Apache Spark Using SAS® In-Database Code Accelerator." *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/2019/3116-2019.pdf.

Hazejager, Wilbram., and N. Rausch. 2017. "Ten Tips to Unlock the Power of Hadoop with SAS®". *Proceedings of the SAS Global Forum 2017 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings17/SAS0190-2017.pdf

Hazejager, Wilbram., and N. Raush. 2018. "Data Management in SAS® Viya®: A Deep Dive." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1670-2018.pdf.

Hoffritz, C. 2019. "Boop-Oop-A-Doop! It's Showtime with SAS® on Apache Hadoop!" *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/2019/3019-2019.pdf

Maher, Salman, and C. Dehart. 2018. "What's New in SAS® Data Connectors for SAS® Viya®." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available 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 https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1669-2018.pdf.

Rineer, B. 2018. "Doin' Data Quality in SAS® Viya®". *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2156-2018.pdf.

Robert, Nicholas. "SAS® Viya® 3.4: What's New in CAS Data Access?." Available https://communities.sas.com/t5/SAS-Communities-Library/SAS-Viya-3-4-What-s-New-in-CAS-Data-Access/ta-p/490016. Last modified August 27, 2018. Accessed on February 26, 2019.

SAS Institute Inc. 2018. *Cloud Data Exchange 2.3 for SAS® Viya® 3.4: Administrator's Guide*. Cary, NC: SAS Institute Inc. Available https://go.documentation.sas.com/?docsetId=dataagentag&docsetTarget=p1bhvtbh9yip6zn1mf9ccgsaxy0d.htm&docsetVersion=2.3&locale=en.

SAS Institute Inc. SAS® Data Management Community. Available https://communities.sas.com/t5/DataManagement/ct-p/data_management.

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) 677-444