Paper SAS6424-2016

Getting Data into your SAS® Cloud Environment

Gary Mehler, SAS Institute Inc.

ABSTRACT

As more of your work gets performed in an off-premise cloud environment, understanding how to get the data you want to analyze and report on becomes important. In addition, working in a Cloud environment like Amazon Web Service may be something that is new to you or your organization when using abilities in a package like SAS® Visual Analytics. This presentation discusses how to get the best use out of Cloud resources, how to efficiently transport information between systems, as well as how to continue to leverage on-premise DBMS data in your future Cloud system.

INTRODUCTION

Just because your SAS environment is in a remote location doesn't mean you can't make your information easily available to it. Focusing on an application like SAS Visual Analytics, this paper discusses three paths to satisfy that requirement, from interactive methods any user can perform to more automated ways an administrator might help set up. The most general path is that individual users can upload files themselves, using an interactive feature in SAS Visual Analytics.

A second path is useful If larger files are being maintained by a data administrator, the SAS® vApp Data Manager can make it easy to transfer and check the status of data that's been uploaded. Once uploaded or otherwise available in our SAS Cloud instance, you can specify actions on that data as needed, interactively. Also, you can rest assured that once your data has been uploaded, it will be stored in a secure way so you can be confident that the right people will be granted access to it. This paper will use the SAS Cloud environment to highlight how this can be done.

INTERACTIVELY UPLOADING DATA

If you want to upload new data that is of reasonable size (less than two gigabytes or so) and have access to it from your local PC, you can use the self-service importer that is present in SAS Visual Analytics Designer, Explorer, and also in the SAS Visual Data Builder. This is the same process that would be used in a local, on-premise deployment, except that the data will be transmitted across the internet to your cloud-based SAS environment. Data from a range of file types, including SAS datasets, Excel workbooks, and text files that have delimited data are supported. Figure 1 shows the import options that are available from within the SAS Visual Analytics environment. We'll focus on the first section, Local data files, here.

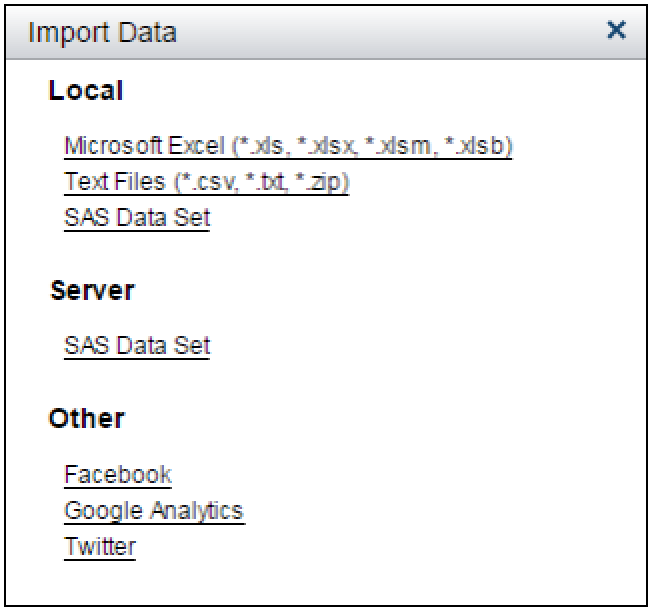


Figure 1. Selecting the type of data to import interactively

In this case, Local data refers to files that are accessible from your web browser. This means files that you have stored on your PC, or on network drives accessible from your PC. You can navigate to any accessible drive or server and select the file or files of interest to you.

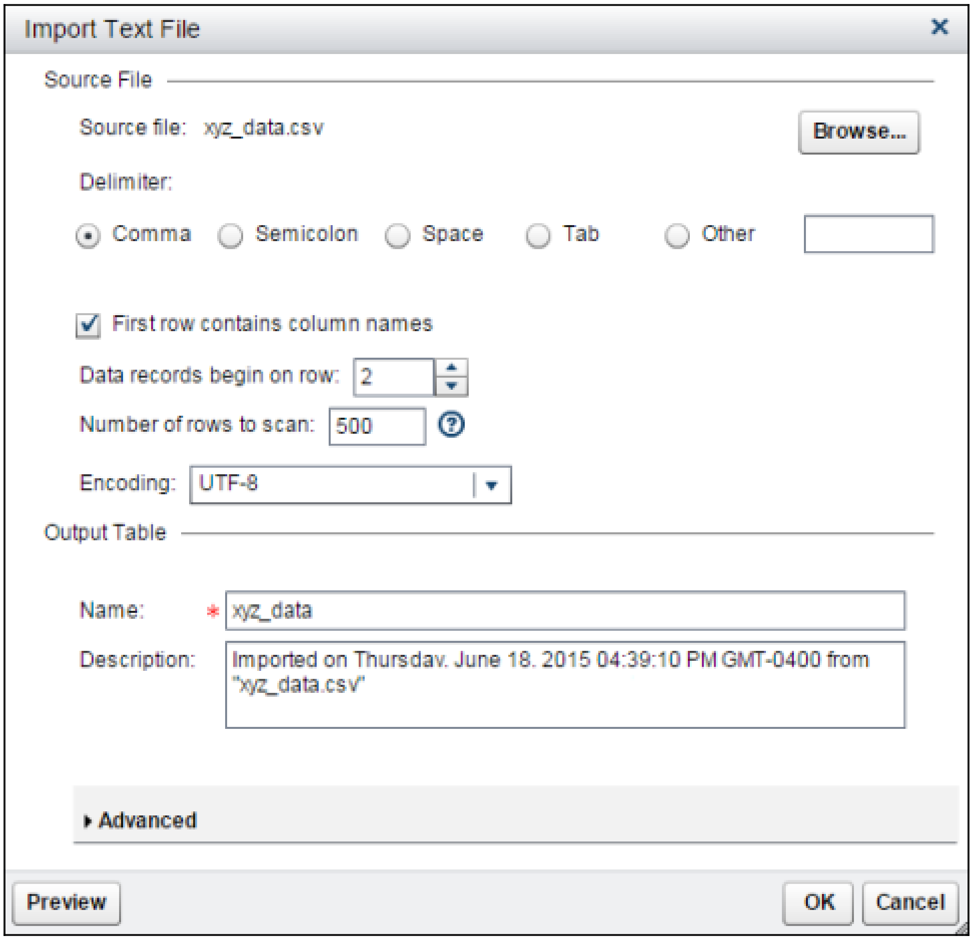
If your data is large and you wish to compress it before transport, or if you have a large number of small files you need to upload, placing them into a zip file archive can simplify that task.

Social media sources like Twitter can be accessed as well, by entering the needed information to complete the import. Figure 2 shows an example for a locally-accessible text file, allowing for selection of options to ensure the data is imported correctly.

WORKING WITH AN IN-MEMORY ANALYTIC SERVER

An important consideration when thinking about uploading your data is to understand that an application like SAS Visual Analytics requires that the data also be present in the in-memory analytic server to be useful. The data upload paths discussed in the paper take that into account. So, in addition to adding it to the in-memory analytic server, uploading the data as described also causes a copy of that data to be stored on your SAS Cloud system. This is useful when one thinks about the nature of an in-memory server. If your SAS Cloud system is powered down or restarts for any purpose, you'll want that data to remain available when the system has recovered.

Files you upload are copied to a known location (we'll look at that area in a later part of this paper) so that quick reloads can occur without the need to re-upload your data from your PC or local area network. You can certainly re-upload your data if you want to refresh its contents, but that's not a requirement when handling basic operation of keeping your data always in-memory for analytic use.



Import Text File

Source File

Source file: xyz_data.csv

Delimiter:

● Comma ○ Semicolon ○ Space ○ Tab ○ Other

☑ First row contains column names

Data records begin on row: 2

Number of rows to scan: 500

Encoding: UTF-8

Output Table

Name: * xyz_data

Description: Imported on Thursdav. June 18. 2015 04:39:10 PM GMT-0400 from "xyz_data.csv"

Figure 2. Specifying parameters to handle data file contents correctly.

UPLOADING SAS DATA SETS

If your data is large or you already have SAS data sets, the SAS vApp Data Manager can be used. If you are using a SAS Cloud environment, you may already be familiar with SAS App Central, which is a launching pad for various applications, including the vApp Data Manager. Figure 3 shows a view of SAS App Central, highlighting the SAS vApp Data Manager

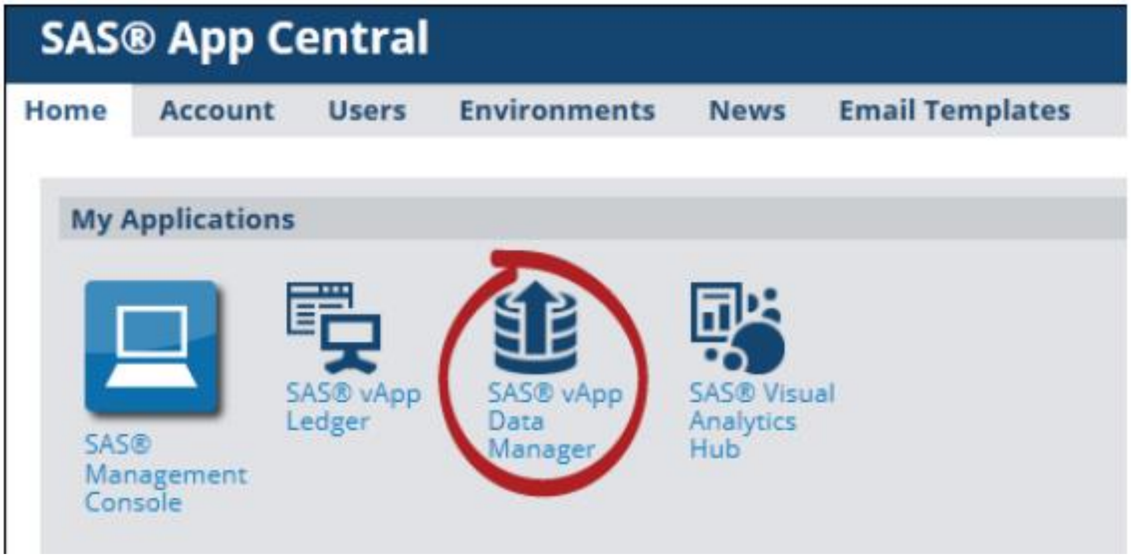
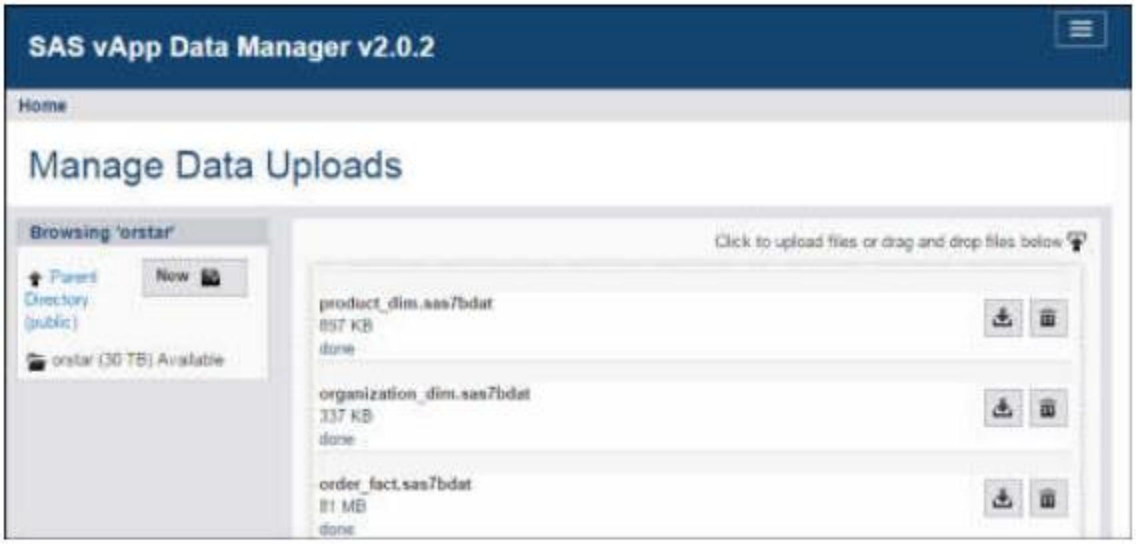


Figure 3. Selecting the SAS vApp Data Manager for uploading SAS data sets.

Within the Data Manager, you can perform various actions, like adding new data sets, downloading them again (back to your local PC), or deleting them. The area that the Data Manager shows is part of the SAS Visual Analytics Auto Loader area. The Auto Loader watches for changes in this area and when it detects one, causes that data set to be loaded to the in-memory analytics environment that is used by SAS Visual Analytics. This is another way SAS Visual Analytics keeps your data always available and accessible.

The Auto Loader uses a fixed interval (typically every five minutes) to check for new files, and then to load them. Using this area is a good way to keep things current if you already have SAS data sets, or are able to refresh the area automatically by other means like using SFTP to upload data files via scheduled scripting. Figure 4 shows an example view of files in the Auto Loader area and some of the actions that can be performed on them.


SAS vApp Data Manager v2.0.2
Home
Manage Data Uploads
Browsing 'orstar'
Parent Directory (public)
New
Click to upload files or drag and drop files below

orstar (30 TB) Available

product_dim.sas7bdat
897 KB
done

organization_dim.sas7bdat
337 KB
done

order_fact.sas7bdat
81 MB
done

Figure 4. Looking at files in the Data Uploads area.

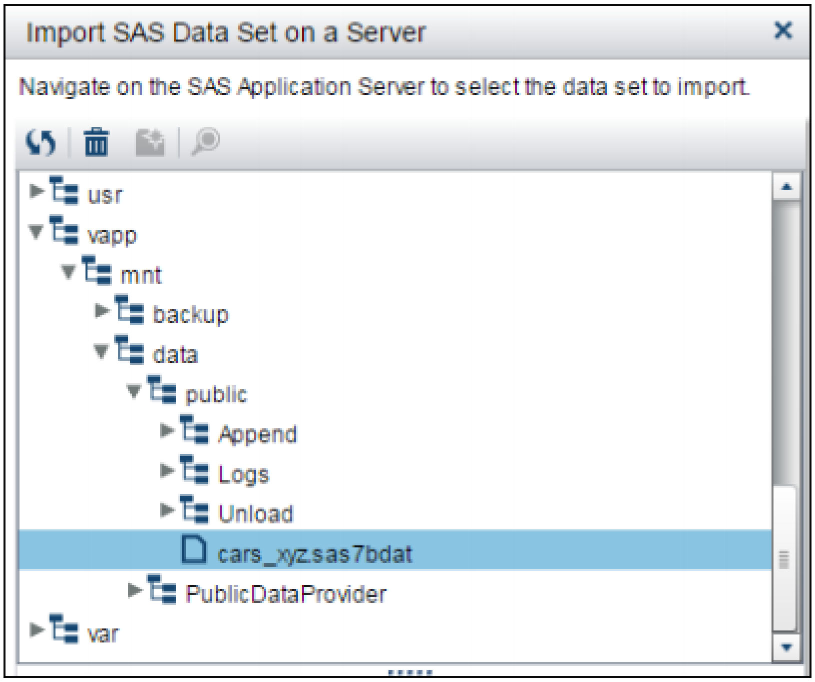
To understand more about the Auto Loader, it is helpful to understand that it uses its own directory structure to help perform a few key operations on data. This will be covered in more detail in the next section.

MANUALLY LOADING SAS DATA SETS

If your data has already made its way to your SAS Cloud environment but needs to be manually loaded, this can be performed interactively in SAS Visual Analytics, much like interactive file uploader discussed previously. Referring back to Figure 1, another option is to select Server data, meaning SAS data sets that are accessible from the SAS Cloud (server) environment. This brings up a way to look at the filesystem in the SAS Cloud environment and find data that was uploaded previously. Figure 5 shows how you can navigate the filesystem.

Why would you use this? As noted earlier, interactively uploading data causes that data to be loaded to the in-memory analytic server and be stored in a location that is automatically reloaded after any outage occurs. Usage of the SAS vApp Data Loader achieves the same always-available purpose by making use of an Auto Loader for data that keeps data fresh.

While these automatic methods are usually a good option, there may be other cases in which you don’t always want your data loaded into memory. A typical reason for this is that some data may be very large and infrequently used. If that’s the case, keeping it always resident in-memory might consume too much memory capacity and you might choose to only load on an as-needed basis, such as for running annual or other period reports. In this case, you’d want to specify when to load your data into memory in a more manual process.



Import SAS Data Set on a Server

Navigate on the SAS Application Server to select the data set to import.

Figure 5. Selecting the SAS vApp Data Manager for uploading SAS data sets.

Once the data is located, it can be loaded into the SAS in-memory analytics server for use with SAS Visual Analytics. You will note in Figure 3 that you are navigating the filesystem of the SAS Cloud environment, but there a few things to note about what is being seen there. First, you may notice that this only a small segment of the filesystem of the Linux environment on which SAS Cloud is running.

The filesystem view is filtered to only show “safe” areas and will never allow someone to see operating system or other files on the system. So, even if you need to browse around, you shouldn’t easily get lost when trying to find that that’s already been uploaded to your SAS Cloud environment.

MORE ABOUT HOW THE AUTO LOADER WORKS

The second point to note in Figure 5 is the underlying directory structure of the “public” folder for data storage. This folder is also the Auto Loader storage area and the example shows that the SAS data set cars_xyz has been uploaded and should be available for usage. There are other directories available as well, and two of importance are “Append” and “Unload”. These directories have a special function with the Auto Loader in that they allow other operations to be performed than just loading a table.

Auto Loader logic will also detect when a new data set has been placed in that directory. So, if you were to re-upload cars_xyz, that data would get refreshed in the SAS in-memory analytic server to reflect the updated data set. In this way, periodic updates can be performed in these areas, either using the Data Manager or even by using a tool like SFTP to upload SAS data sets on a regular basis to make the latest data available.

Two other subdirectories are of interest as well. “Append”, as noted above has a special function in that if cars_xyz were already loaded and had data from, say, the last 3 years of data and you wished to add new rows to the end with data from the current year, that can be performed using the Append area. If a SAS data set with the same name as one that has already been loaded is uploaded to the Append directory, the SAS Auto Loader will perform an Append operation on the loaded data, making old and new data available for analysis.

“Unload” has a function that may not be difficult for the reader to guess: placing a SAS data set in this folder will cause an original SAS data set of the same name to become unloaded from the SAS in-memory analytic server. This can be helpful in cases in which data has become stale or needs to be unloaded for memory consideration. So, to complete our example, placing a SAS data set called cars_xyz in this folder will cause the original SAS data set cars_xyz to become uploaded from the SAS In-Memory Analytics Server.

MANUALLY LOADING SAS DATA SETS

The approaches described above work well for data that has the needed structure for reporting or analytics, and just needs to be uploaded to become available. What about cases in which data needs to be extracted from a DBMS or have actions performed on it like filtering to achieve the needed data structure? That’s a little more complicated, but there are few ways to think about that requirement. Environments like SAS Visual Analytics include an application for data manipulation called SAS Visual Data Builder. If your data needs to have some filtering performed, be joined with other data, or have data values modified or created, SAS Visual Data Builder can help with those activities. It can help create calculated columns to facilitate reporting or a range of other manipulations done with your data.

Visual Data Builder works well as long as the source data isn’t so large that uploading a large portion of it, just to have a large filtering operation subset it down, is impractical. That’s not always wise given the time it would take to upload extra data that’s not really needed. But if basic manipulations are all that’s required, it could be a good use of SAS Visual Data Builder.

If, on the other hand, the data is coming from a DBMS or is very large, then you should investigate on-premise methods to get the data extracted, or subset down to the right size to be used. If you have SAS Data Loader, it can help get data out of Hadoop environments. Or if you are using a SAS Data Management product, it should let you perform any extraction and on-premise data manipulations required. The output of your on-premise process just needs to be the starting point for the SAS Cloud environment to pick it up. At this point, you could then determine whether the interactive importer is best for you, or whether you should use the SAS vApp Data Manager to transport that information.

CONCLUSION

Getting data into a SAS Cloud environment is not a very complicated task, even though its physical location may be some distance away from the PC on which you see the user interface of a SAS application like SAS Visual Analytics. Small data can easily be uploaded using the interactive uploader in SAS Visual Analytics. If you have larger files of data (greater than four gigabytes in size), the SAS vApp Data Manager is a good path to pursue. This is also a good way to handle large numbers of data sets or other data files. SAS provides cloud-based support for these types of operations. If you need to do some on-premise manipulation before uploading data, keep other SAS applications in mind to help with those tasks.

There is a saying about people that absence or distance makes the heart grow fonder. Using a SAS Cloud environment will allow you to have your data further away from where it may have started, but it certainly won't make you *less* fond of your data. You should be able to use it just as much as you have been in an on-premise world, as long as you have some basic understanding of how it gets there and is maintained. In fact you might actually become more fond of it knowing that in such an environment, your data is well cared for. It is backed up, stored safely, and treated well. And it remains a reporting or analytic partner for your SAS activites.

REFERENCES

SAS Institute, Inc. "SAS Visual Analytics 7.3 for SAS Cloud Quick-Start Guide". Accessed March 1, 2016. Available at https://support.sas.com/documentation/onlinedoc/cloudgen/ .

SAS Institute, Inc. "SAS Visual Analytics Users Guide". Accessed March 1, 2016. Available at https://support.sas.com/documentation/onlinedoc/va/index.html#va73 .

CONTACT INFORMATION

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

Gary Mehler
SAS Institute, Inc.
Gary.Mehler@sas.com