Paper 113-2012

Best Practices for Managing and Monitoring SAS® Data Management Solutions

Gregory S. Nelson

President and CEO
ThotWave Technologies, Chapel Hill, North Carolina

Abstract

SAS® and DataFlux® technologies combine to create a powerful platform for data warehousing and master data management. Whether you are a professional SAS administrator who is responsible for the care and feeding of your SAS architecture, or you find yourself playing that role on nights and weekends, this paper is a primer on SAS Data Management solutions from the perspective of the administrator. Here, we will review some typical implementations in terms of logical architecture so that you can see where all of the moving parts are and provide some best practices around monitoring system and job performance, managing metadata including promotion and replication of content, setting up version control, managing the job scheduler, and discuss various security topics.

Introduction

SAS and its smaller sibling, DataFlux, have a come together to provide a full complement of technologies for data integration and data management. This paper will focus on the SAS side of this technology equation as found in two offerings: SAS Data Integration and SAS Enterprise Data Integration. The table below offers a comparison of the technologies that are available in each offering. For a better understanding of how SAS Data Integration can be used to solve real world problems, refer to the additional resources (Grasse and Nelson, 2006, Nelson, 2006, Nelson, 1999) referenced at the end of the paper.

SAS Data Integration	SAS Enterprise Data Integration
SAS Products • Base SAS • SAS/CONNECT • SAS Data Integration Studio • SAS Integration Technologies • SAS Management Console and SAS Metadata Bridges • SAS Metadata Server	SAS Products • Base SAS • SAS/CONNECT • SAS Data Integration Studio • SAS Integration Technologies • SAS Management Console and SAS Metadata Bridges • SAS Metadata Server • SAS Data Quality Server • SAS/ACCESS • SAS/SHARE DataFlux Products • Data Management Server for SAS • Data Management Studio

From a feature/ function perspective, SAS Data Integration Studio (DIS) is the primary user interface for ETL[1]/ Data Quality developers. DIS allows developers to perform a variety of tasks, including:

- Access data (from a variety of SAS and non-SAS sources such as databases and common file formats)

[1] ETL - Extract-Transform-Load

- Integrate data from multiple sources
- Manage metadata including information about data libraries, tables, jobs, servers, users, groups and roles
- Cleanse data through standard and user-written transformations
- Enrich data by augmenting fields with third-party or in-house enrichment techniques
- Extract, transform, and load (ETL)
- Integrate data with service-oriented architecture (SOA) and message queue technologies

From an administration perspective, there are a number of touch-points where SAS technologies integrate with the environment and require active administration. These include:

- Managing metadata – such as managing the folder structures and managing repositories
- Configuring change management - for multi-user repositories and promotion of metadata between environments
- Server Management – starting/ stopping of servers, log management, backup/ restore
- Managing Security - managing users/ groups/ roles
- Data library management – set up data definitions and access to external data libraries
- Scheduling in SAS – setup and maintenance of the scheduling server
- Integration with external message queues, FTP and HTTP external access and web services
- Event management – set up of job status handling and error handling
- Configuration of bulk loaders

Some of these activities are done once while others require active, ongoing maintenance. Throughout the remainder of this paper, we will focus on what we refer to as the “care and feeding” tasks for the environment. These include:

- Monitoring system and job performance
- Managing metadata including promotion and replication of content

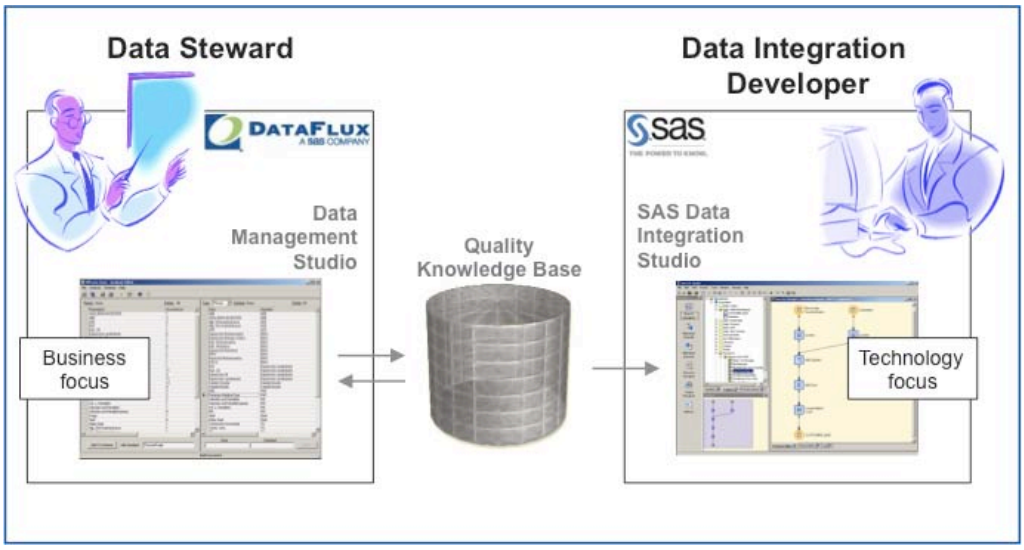
- Setting up version control
- Managing the job scheduler

In addition, we will discuss various security topics to ensure the environment is safe and secure.

Understanding SAS Data Integration

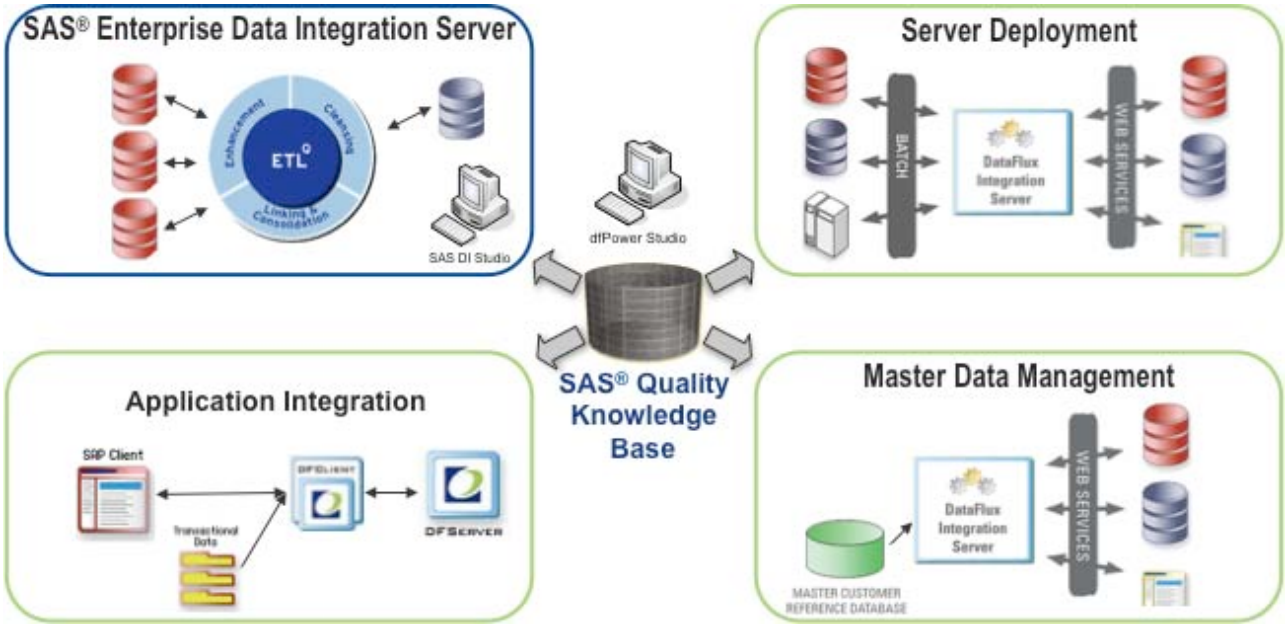
At a high level, SAS Data Integration is one of the many modern metadata-managed solutions from SAS. The information is stored in the metadata server, and users can utilize client tools like Data Integration Studio, OLAP Cube Studio and Information Map Studio to create data for use in other applications such as Web Report Studio or the OLAP viewer in the SAS Information Delivery Portal.

In the data management world, we often think of the data steward as a business analyst who is responsible for the cleansing of the data. These folks would use Data Management Studio, whereas traditional ETL developers would use the SAS Data Integration Studio. Both, however, would share a common quality knowledge base for the rules on how data is to be managed.



In a related paper entitled *"Serving SAS: A Visual Guide to SAS Servers"* (Nelson, 2012), we demonstrate how a number of SAS clients and servers work, including Data Integration Studio. You are encouraged to review this paper and the associated presentation to get a baseline understanding of how SAS Data Integration Studio works and what the server is doing in response to requests from the client application.

At a high level, the diagram below depicts the SAS Data Integration Server in a typical implementation depending on whether the focus is master data management, ETL processing or real-time data integration.



So now, let's turn our attention to some common administrative tasks.

Systems Administration Activities

System Monitoring

As with most SAS servers, there are a number of things that you'll likely want to monitor, such as checking the status of the servers. As outlined in Chapter 6 of the System Administration Guide, a number of tools are available to determine the status of your SAS server components. These include using SAS Management Console, server scripts or SAS programs that can be used to validate that instances of a SAS Metadata Server, SAS Workspace Server, SAS Pooled Workspace Server, SAS Stored Process Server, or SAS OLAP Server are set up correctly. We won't repeat those here, but in terms of best practice, we recommend the use of external tools like BMC Performance Manager, or tools from IBM, HP or third parties that actively monitor the environment.

Here we will focus on one approach that you can use out of the box with the tools that ship with SAS.

AUDIT AND PERFORMANCE MEASUREMENT

We often recommend utilizing the Audit and Performance Measurement capabilities with SAS (commonly referred to as ARM logging) to get more detailed information about what's going on in the server.

The SAS Audit and Performance suite of tools allows you to collect data from these various logs and provides advanced reporting capabilities to assess the load on the server; you also

have the capability of slicing and dicing the performance data.

The figure below shows an example as viewed in the SAS Information Delivery Portal, but note that SAS EBI is not required to generate reports for analysis of performance data.

It is important to note that jobs run interactively in Data Integration Studio utilize the Workspace Server (and object spawner) for logging information, whereas jobs run in batch (as part of a scheduled process) utilize the SAS Batch Server – where the logs will be written. If you want to collect ARM log information for SAS Data Integration Studio jobs that are run on a batch basis, you must enable logging for the batch server that executes the job, as these are not enabled by default. When jobs are run interactively in SAS Data Integration Studio, ARM logging is used to collect performance-related events, which are displayed in the SAS Data Integration Studio application. You do not need to change the initial server configuration to enable the capture and display of these events.

To run additional logging features, refer to Chapter 8 of *SAS® 9.2 Intelligence Platform System Administration Guide Second Edition*.

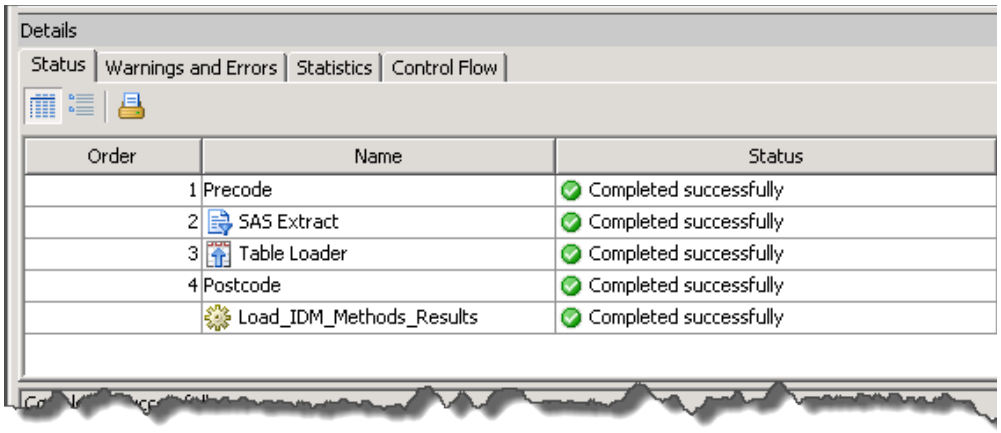
Monitoring Jobs

Just as monitoring the entire system is critically important, so too is monitoring specific jobs that are run. Users will often complain that their jobs are taking too long - interpreting this in order to really understand what might be going on can be a challenge. Several options are at your disposal that includes those options described above for monitoring your system.

Status and performance information is captured into job logs, and the performance reports and status dashboards have been developed to extract relevant information into tables that can be used for reporting. A set of pre-built reports for several deployment options is

shipped with SAS Data Integration (and Enterprise Data Integration). You can deploy these reports and get current and historical information on your jobs.

For interactive jobs, you can use the job status monitor. After you submit one or more jobs for execution, display the Job Status Manager window to view the name, status, starting time, ending time, and application server that will be used for all jobs that are submitted in the current session. From the SAS Data Integration Studio desktop, select the Status tab in the Details section of the Job Editor window (in older versions of SAS DI Studio this can be found under the Tools ►Job Status Manager) . This displays the status of each step in the job. The following figure shows a Status tab that confirms that all of the steps in a sample job were completed successfully.



In addition, the Statistics tab displays useful information about the job that is being processed. Statistics such as CPU, I/O and Memory can be displayed in both tabular and graphical format so you can quickly assess where jobs are consuming resources.

However, this ability is not turned on by default and additional logging options need to be configured which provide this level of introspection. See http://support.sas.com/documentation/cdl/en/bisag/64088/HTML/default/viewer.htm#a003314985.htm for information on collecting the required log information for interactive DI Studio jobs.

In monitoring jobs, you really want to know if jobs are performing as expected. Some SAS Administrators come from a SAS programming background, but if you are not SAS savvy, don't fret - here are a few suggestions that might help:

- Review the ARM logging (see above) to determine if the job time is really increasing
- Look at system monitoring tools (external to SAS which monitor the operating system, CPU, I/O and memory) and assess whether the load on the machine has changed and/or their might be bottlenecks somewhere in the system
- Have the user add options to their jobs so you can get the full picture; for example

`Options stimer fullstimer`[2]`;` - which produces detailed metrics on how the job is performing and what resources are being consumed. Compare the real-time values to the CPU time. If there is a difference of 20-25%, look at the hardware to see if there is a computer resource bottleneck.

A great paper is available from SAS with many more tips on improving performance; you are encouraged to read one or more of the papers specific to SAS performance tuning found at http://support.sas.com/resources/papers/tnote/tnote_performance.html

Metadata Management

A number of metadata management activities typically fall in the SAS Administrator's role. At a high level, these include:

- Setting up the SAS Folders structure
- Adding users/ groups and managing roles that are assigned
- Backup and restore of SAS metadata

Perhaps none is as important as backup and restore; you should establish a formal, regularly scheduled backup process that includes your metadata repositories as well as the associated physical files.

SAS provides backup and restore utilities that enable you to perform correct backup and restore of your metadata repositories, the repository manager, the metadata journal file, and the metadata server configuration files, while minimizing disruptions in service. It is also important to back up the physical data that is associated with the metadata, so that related information will be synchronized if a restore becomes necessary.

Before you back up your SAS Intelligence Platform, read Chapter 9 of *SAS® 9.2 Intelligence Platform System Administration Guide Second Edition*, "Best Practices for Backing Up and Restoring Your System," on page 123.

Change Management and Version Control

SAS Data Integration features some built-in capabilities for managing change. These generally fall into one of two buckets: change management within SAS Data Integration Studio, and using external version control systems for content management.

CHANGE MANAGEMENT

Change management in the context of SAS is about allowing a team of users to work simultaneously with a set of related metadata (such as Jobs, Flows, Data definitions, etc.),

[2] SAS Documentation for FullStimer results for UNIX http://support.sas.com/documentation/cdl/en/hostunx/61879/HTML/default/viewer.htm#o-f1.htm and Windows http://support.sas.com/documentation/cdl/en/hostwin/63285/HTML/default/viewer.htm#win-sysop-fullstimer.htm

and avoid overwriting each other's changes. With change management, metadata security controls who can add or update the metadata in a change-managed folder. Appropriately authorized users can add new content and check them in and out of the change-managed folder.

In order to implement change management for SAS Data Integration Studio, as the SAS Administrator you will set up project repositories and a change-managed folder. This process is documented in "Administering Data Integration Studio" in the *SAS Intelligence Platform: Desktop Application Administration Guide.* *http://support.sas.com/documentation/cdl/en/bidaag/61231/HTML/default/viewer.htm#a002655688.htm*

A good overview of the capabilities of change management in metadata can be found in the SAS white paper entitled *"Best Practices for SAS®9 Metadata Server Change Control"* referenced at the end of this paper.
http://support.sas.com/resources/papers/MetadataServerchngmgmt.pdf

VERSION CONTROL

Version control, on the other hand, is the capability to manage files and objects through an external version control system such as CVS or Subversion. Versioning works by moving jobs and other content into a file (called a SAS package file), which uses a ZIP format for compression, and then archiving that file in a versioning system. You can version content independently (e.g. a single job) or with other objects to make up a package.

Users can see options for version control when they right click on an object in DI Studio when the SAS Administrator has added the correct plug-in to the plug-ins directory. The plug-in that is appropriately installed will have a tab for configuring the interaction with the 3rd party plug-in system in the available options panels.

To configure this, you need to have either Concurrent Versions System (CVS) server or Apache Subversion (SVN) server. Once installed and with a native repository created in one of these tools, you'll need to install a third party version control client, because SAS Data Integration Studio requires that a third party version control client be installed on the same client machine. Tortoise SVN is a common client used throughout the industry; see http://tortoisesvn.net/downloads.html

Once the version control client is installed, you will need to enable the appropriate SAS java plug-ins in DI Studio and configure DI Studio to communicate to your version control server. From here, your users will be able Create a Version, Review and Manage Versions, and Compare Versions - all within DI Studio.

Job Control and Scheduling

Earlier, we contrasted the interactive users of DI Studio from those that get scheduled. To

schedule jobs that are created by SAS Data Integration Studio, you can use either Platform Suite for SAS (PSS) or an operating system scheduler. There are some benefits of using PSS over that of the operating system scheduler, as PSS is tightly integrated with SAS.

In general, the process for scheduling a job from DI Studio involves creating a job (which is then stored in the SAS metadata repository); then the user "deploys" the job for scheduling, which creates a SAS program based on the job metadata and puts it in a directory on the SAS Batch Server (called a deployment directory.)

Once the job is deployed, the SAS administrator uses Schedule Manager in SAS Management Console to create and schedule a flow that includes the deployed job. Schedule Manager passes the schedule information and the associated commands to the designated scheduling server (either the Process Manager Server or an operating system scheduling server).

If you are using PSS, Platform Process Manager submits the scheduled job to Platform LSF for dispatching instructions and the SAS DATA Step Batch server executes the job. Job status information is stored on the scheduling server and the administrators can use Platform Flow Manager to view the status of the job.

If you are using an operating system scheduling server, the operating system submits the job at the designated time. Administrators can use operating system tools to view the status of all scheduled flows. Users who are not administrators can view and edit only the flows that they own.

In order to schedule a job, these essential components are required:

- the deployment directory – the location on the SAS Batch Server where the jobs will be stored
- SAS Batch Server definition in metadata

Because jobs are run in batch, it is often difficult to debug issues – especially when the user says that it worked fine in DI Studio. It may be a matter of following the trail by reviewing the SAS logs generated from the scheduled job. These are usually located along side the programs in the Deployment Directory. Common issues include users hard coding data library paths in their DI jobs rather than relying on the SAS metadata libname engine (MLE), or a user's credentials changing on the operating system that have not been updated in the deployed job.

Finally, it is important to note that if you are using SAS Grid Manager, jobs can be deployed so that they are aware of the grid machines for load balancing across the grid. You can leverage this capability in SAS Data Integration Studio and other SAS products by deploying the job to be scheduled and then using the scheduling server that manages the grid to schedule and run the jobs. The scheduling server manages the resources in the grid workload so that jobs are efficiently distributed across the available machines in the grid.

Security

When we talk about security in the context of SAS Data Integration, there are three general areas of concern:

- Managing groups and users - adding users and managing access; creating SAS administrators and regular SAS users; managing access to metadata, data, and application functionality (through roles). For details, see "Security Tasks" in the SAS Intelligence Platform: Security Administration Guide.
- Establishing connectivity to your data sources - enabling the client applications such as DI Studio to access your data sources (including SAS data sets and third-party relational databases). For details, see "Connecting to Common Data Sources" in the SAS Intelligence Platform: Data Administration Guide.
- Setting up your metadata folder structure - SAS clients use a hierarchy of SAS folders to store metadata for content such as libraries, tables, OLAP schemas, jobs, information maps, and reports. The initial structure provides private folders for individual users and a separate area for shared data. Within these folders, you should create a customized folder structure that meets your specific needs. For details, see Chapter 16, "Working with SAS Folders," on page 193 in *SAS 9.2 Intelligence Platform System Administration Guide Second Edition*.

In addition, it is highly recommended that you review the *SAS Intelligence Platform: Security Administration Guide*. See "Checklists for a More Secure Deployment".

Summary

The SAS Administrator must be aware of a number of activities so that the environment performs well for users. From managing security to turning on features that the system has, this paper has featured some of the more common practices that may be useful to new or experienced administrators.

References and Recommended Reading

Administrative Documentation for SAS Data Integration Studio
http://support.sas.com/documentation/cdl/en/etlug/63360/HTML/default/viewer.htm#n0a1xzsuqksz2pn1pgu4au739qss.htm

Best Practices for SAS®9 Metadata Server Change Control
http://support.sas.com/resources/papers/MetadataServerchngmgmt.pdf

ETL Performance Tuning Tips, http://support.sas.com/resources/papers/tnote/tnote_performance.html

Grasse, D. and Nelson, G. *"Base SAS vs. Data Integration Studio: Understanding ETL and the SAS tools used to support it".* Invited Paper presented at the SAS Users Group International Conference. San

Francisco, CA. March, 2006.

Nelson, G. *"A Pragmatic Programmers Introduction to Data Integration Studio: Hands on Workshop".* Hands on Workshop presented at the SAS Users Group International Conference. San Francisco, CA. March, 2006.

Nelson, Gregory S. "*Implementing a Dimensional Data Warehouse with the SAS System*." Invited Paper presented at the SAS Users Group International Conference. San Diego, CA. March, 1999.

Nelson, Gregory S. "*Serving SAS: A Visual Guide to SAS Servers*." Invited Paper presented at the SAS Global Forum. Orlando, FL. April, 2012.

Scalability and Performance Papers from SAS authors http://support.sas.com/rnd/scalability/papers/

SAS Data Integration Studio 4.4: User's Guide
http://support.sas.com/documentation/cdl/en/etlug/65016/HTML/default/viewer.htm#titlepage.htm

SAS 9.2 Intelligence Platform: Desktop Application Administration Guide.
http://support.sas.com/documentation/cdl/en/bidaag/61231/HTML/default/viewer.htm#a002655688.htm

SAS® 9.2 Intelligence Platform System Administration Guide Second Edition.
http://support.sas.com/documentation/cdl/en/bisag/64088/PDF/default/bisag.pdf

Biography

Greg Nelson, President and CEO of ThotWave Technologies, LLC.

Greg is a certified practitioner with over two decades of broad Business Intelligence and Analytics experience gained across several life sciences and global organizations as well as government and academic settings. He has extensive software development life cycle experience and knowledge of informatics and regulatory requirements and has been responsible for the delivery of numerous projects in private and commercial environments. Greg's passion begins and ends with helping organizations create *thinking data*® – data which is more predictive, more accessible, more useable and more coherent.

His current area of interest is helping companies take advantage of the shifting world of convergence around data and systems and how modernization and interoperability will change the way that we discover new relationships, manage change and use data and analytics to improve organizational outcomes.

Mr. Nelson has published and presented over a 150 professional papers in the United States and Europe. He holds a B.A. in Psychology and PhD level work in Social Psychology and Quantitative Methods, along with certifications in project management, Six Sigma, balanced scorecard and healthcare IT.

Greg can be reached at greg@thotwave.com or www.linkedin.com/in/thotwave

About ThotWave

ThotWave Technologies, LLC is a Cary, NC-based consultancy and a market leader in real-time decision support, specializing in regulated industries, such as life sciences, energy and financial services. ThotWave works at the juncture of business and technology to help companies improve their operational and strategic performance and recognizes the difference between simply accessing data and making data work for business. Through products, partnerships and services, ThotWave enables businesses to leverage data for faster, more intelligent decision making.

Contact information:

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

Greg Nelson greg@thotwave.com

ThotWave Technologies, LLC

Chapel Hill, NC 27517 (800) 584 2819

http://www.thotwave.com