Paper 214-31

Understanding the System Architecture of SAS/ETL and SAS/EBI Server for Federal Enterprise Architecture requirements.

By

John W. McCue, Congressional Research Service

Abstract: In order to document our Agency's Systems Architecture we need to understand how the design of the SAS System affects the other Systems in our infrastructure. This paper describes some of the key points learned through: many conversations with SAS Institute, and persistence to understand how the SAS Enterprise Business Intelligence Server and the Extraction, Transformation, and Loading Server work. These new Enterprise offerings of SAS Institute require planning to install and knowledge of your system infrastructure to configure. Because of this, SAS Institute personnel are often utilized to facilitate this process. But following two old axioms about teaching a man to fish, and learning more from your mistakes than your successes we installed the Server software on our own. This paper documents some of what we learned about the SAS Intelligence Platform from configuring the new software.

Keywords: ETL, EBI, Systems Architecture, FEA, Version 9, Intelligent Architecture

Introduction

As the Federal government struggles to understand the requirements of the Federal Enterprise Architecture (FEA), Chief Architects seek information that will describe the systems within their organization to others. Chief Architects will often request information, regarding the documentation of Agency systems under their domain, from the Data Architect. This paper was written to aide the Data Architect in understanding how the SAS Enterprise Business Intelligence Server and the Extraction, Transformation, and Loading Server work.

The task of documenting an Agency's data architecture is done in the hope that the agency will then be better positioned to share information where possible and not need to reinvent information that is time-consuming and expensive to generate. The various models that make up the FEA framework equip these Federal agencies with a common language, standardized procedures and a documented process for communicating with others. This architecture describes the analysis of IT investment to understand the amount of intra-agency collaboration. The goal of the task of documenting an Enterprise Architecture is to ultimately transform the Federal government into a citizen-centered, results-oriented, and market-based organization as set forth in the President's Management Agenda (PMA).

But SAS®, while an encompassing system is not as easily understood as communication protocols, databases, or other parts of the information systems infrastructure. Therefore when management asks a Data Architect to document the SAS® Architecture, it is best advised to take a holistic view to the task and not attempt to document SAS® apart from the entire infrastructure. Document the entire Enterprise Architecture and SAS® is simply a part of the larger whole and not something to be understood outside of the other systems in your organization. But even taken as a part of the entire enterprise, one must still address the issue of understanding the new foundation service layers of the SAS® Business Intelligence Platform and explaining these new services so that they fit within the common language of the FEA framework.

Federal Enterprise Architecture

The mission of the FEA working group is to use the knowledge that each agency gains from process of creating the Federal Enterprise Architecture to improve government efficiency and effectiveness. If this mission is to be successful then programs need a set of work products or models that provide a common language and that these data models are standards by which agencies agree to conform. The architecture should identify opportunities to streamline applications, processes, and data used both internally and externally to the agency.

Therefore, the Federal Enterprise Architecture can be thought of as the instrument used to reach the goal of making applications, processes, and data usage more efficient. But how is FEA supposed to identify opportunities for improvement? What are the steps that your Information Technology (IT) organization must take to document if you are reaching these goals? In order to identify opportunities for improvement an agency should first assess themselves against a

benchmark. In the case of FEA, the Office of Management and Budget (OMB) has four main assessment categories for the five reference Models. They are:

- **Change:** Assesses how well the FEA facilitates the management of change. Using the specific criteria provided in the Performance Reference Model (PRM), an agency identifies and describes the: A) *architectural approach* and B) *strategic direction*.
- **Integration:** Assesses how well the FEA ensures the standardization of interfaces, interoperation, information, and connectivity defined in the Data Reference Model (DRM). Using the specific criteria provided, an agency identifies the level best describing its standardization: A) *interoperability*, B) *data management practices*, C) *business logic*, and D) *interchange of information*.
- **Convergence:** Assesses how well the EA integrates the agency's IT as defined by the Technical Reference Model (TRM). Using the specific criteria provided, an agency identifies the level best describing its understanding of: A) *system components*, B) *technical platforms*, C) *performance monitoring*, and D) *security*.
- **Business Alignment:** Assesses how well the EA ensures alignment with the agency's strategic mission, direction, and plan. Using the specific criteria provided in the Business Reference Model (BRM), an agency identifies its alignment of: A) *strategic goals* and B) *business targets*.

Change and Business Alignment

Because these two assessment areas relate to the integration of management's direction and not the specifics of understanding how the SAS Intelligence Architecture integrates with the Enterprise Architecture we will not be expanding on these areas here. However, if readers want to learn more about these two areas examine https://www.feams.gov/ or http://www.core.gov .

Integration

Integration is the area with the most recently released information in the release of the second version of the DRM[1]. It is in this model that the context and integration of the data takes the form of sets of terms (in SAS this equates to a library reference) that are themselves organized in lists, hierarchies, or trees. The groupings of the hierarchies form a classification scheme. It is this scheme that is the organization of the data to be stored in the Extensible Markup Language (XML) of the open metadata repository.

The SAS System and other programs will "understand" the structure of the data by interoperating directly with the Open Metadata Repository (OMR). The interaction with the OMR can be done with PROC METASERVE or with the SAS tools that come with SAS ETL Server and SAS EBI Server.

A) Interoperability

In the DRM, interoperability is defined as the set of rules that are established to allow two software programs to communicate with each other. These rules establish how one software program requests information from another program, and how the queried software program will respond. These rules are referred to as the Application Programming Interface (API). A software program that writes to the API is one that uses the standard set of rules to ask for or provide information. If a software program writes to the API it should be able to communicate with other applications and the operating system.

The SAS system uses an API that communicates with the native operating system as well as remote operating systems and it also uses the Open Metadata API to communicate the standardization of the data for the FEA to other programs. The API provides to the SAS System the means to communicate with the operating system. The SAS System then utilizes an Input Buffer and the SAS Program Data Vector to process data on the computer system. The Open Metadata API uses Extensible Markup Language (XML) to communicate with other databases about the data. The XML libname engine can access externally produced XML files and the META libname engine can be used to create reports from the OMR.

```
libname library XML 'C:\Files\Data\QS\FEA_XML_Doc_Rev_1.2.xml'
PRESERVE_TAB_NAMES=YES;

libname omr libid=AC000001
   ipaddr='domain.agency.gov' port=8561
   userid=tuvwxyz pw=abcdefg;
```

In the case of some XML files you may have to first use the SAS tool XMLmapper in order for SAS to be able to process the file properly. In order for the SAS System to process data, it works at a much "lower" level in order read and interpret the data values so that it can process and output tables or reports/graphs.

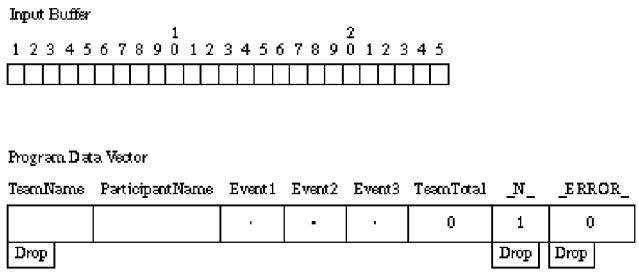
<u>SAS Program Data Vector</u>

When SAS DATA step statements are read and compiled, SAS determines whether it needs to create an input buffer. This decision is based on the review of the code and if the input file contains raw data. If the code does contain raw data in DATA step or as an external file, SAS creates an input buffer to hold the data before moving the data to the program data vector (PDV). If the input file is a SAS data set, however, SAS does not need to create an input buffer. Rather, the SAS System writes the input data directly to the PDV.

"The PDV contains all the variables in the input data set, the variables created in DATA step statements, and the two variables, _N_ and _ERROR_, that are automatically generated for every DATA step. The _N_ variable represents the number of times the DATA step has iterated. The _ERROR_ variable acts as a binary switch with a value of 0 when no errors exist in the DATA step and 1 when an error exists."[1] In the next figure a representation of the Input Buffer and the program data vector are shown after DATA step compilation of the SAS code:

```
DATA work.newset(drop=TeamName);
    input  @1 TeamName $8. @9 ParticipantName $11.
           @20 Event1 2. @22 Event2 2. @24 Event3 2.;
TeamTotal=sum(0,Event1,Event2,Event3);
datalines;

;
run;
```



Note that the variables in WORK.NEWSET which are created by the INPUT and the Sum function statements (TeamName, ParticipantName, Event1, Event2, Event3) are set to missing initially, and TeamTotal would be set to missing if zero were not added to the missing values of Event1, Event2, and Event3. In addition, note that in this representation, numeric variables are initialized with a period and character variables are initialized with blanks. The automatic variable _N_ is set to 1; the automatic variable _ERROR_ is set to 0. The variable TeamName is marked Drop in the PDV because of the DROP= data set option in the DATA statement. Dropped variables are not written to the SAS data set. The _N_ and _ERROR_ variables are dropped because automatic variables created by the DATA step are not written to a SAS data set.

Services in the SAS Intelligence Architecture listen on predetermined ports for incoming requests either directly or through a Spawner. Verify standard SAS port assignments are not in use by other applications before starting the SAS Configuration Wizard. A standard practice is to document the port assignments used on the computer, in the file C:\WINDOWS\system32\drivers\etc\Services. The installation guide will detail the ports needed and the aforementioned file will document the assignments and serve as an artifact for the Architecture.

B) Data management practices

The way that the SAS System processes data is only one aspect of the data reference model. The second area is how the data is managed. In others words, how the tasks of organizing the data can be extracted from external sources, how data can be translated into different formats, and how data and metadata can be loaded. But the "how" of managing the data and metadata is only part of good data management practices. It is also necessary to understand the responsibilities of the people involved in the data management process. The role of the SAS Administrator is to support the Data Steward that manages the data assets of an

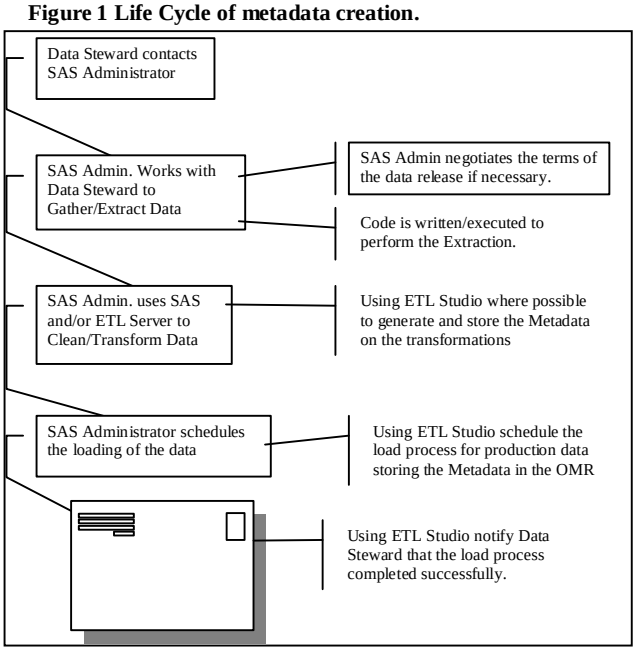
organization. The role of the data steward is to provide the standardized information that is to be stored in the metadata repository. These areas of standardization include the data description, the data context, and the data sharing.

A best practice that our agency learned regarding the accounts for the roles of the SAS Administrator and the Data Stewards is to create the accounts for these roles as Domain accounts as opposed to Local accounts. The accounts that are used as the server accounts (for programs run as services) should also be Domain accounts. These accounts should be created prior to installing the SAS software for ETL Server or EBI Server and the service accounts should not be used by any users to log into the server (these service account should also be denied local logon privileges). The account or accounts for the SAS Administrator role should be granted permission through a group with rights greater or equal to the Data Steward so that they can facilitate the storage of the metadata into the Open Metadata Repository (OMR).

SAS Administrators make up what the FEA describes as a Governance Board that validates the content of the data that has been checked into the data registry. In SAS this data registry is managed through the Management Console and is store as part of the OMR. It is the responsibility of the Data Steward however, to have the most current and correct information in the OMR.

As described (in Figure 1.), it is also important that you work with the Data Stewards, so that everyone understands the lifecycle of each data set prior to its addition to your data registry.

Figure 1 Life Cycle of metadata creation.



As much as possible, the information that is captured by the Data Stewards should be standardized across the agency.

Standardized Vocabulary for Metadata Definitions

A well written metadata definition should include the following:

1. The name of the variable or item being defined.
2. The type of variable: Ordinal Character, Descriptive Character, Integer Numeric.
3. Broader term: classification to which the variable belongs.
4. Distinguishing characteristics: defining attributes with specific values.
5. Function Qualifier: how the item being defined is used.
6. Narrower Terms: refers to the classes below the variable/term being defined.
7. Related Terms: refers to a variable or term that has relevance to the variable or term being defined but is not a synonym.
8. Synonyms: variables or terms that mean nearly the same thing as the variable or term being defined.
9. Examples: an instance of the variable or term in everyday life.
10. Usage: a sample of using the variable or term in analysis or research.
11. Source: information including from where the definition came (date of the document, author, etc.)
12. Dates: Create, Modify dates as well as effective or expiration dates.
13. Replaced by: Indicates that you should not use this variable as it is continued for legacy purposes only, a value is entered only if the condition exists, otherwise you would not include this variable.
14. Approval Information (Data Stewardship): Information to track the governance trail.

C) *Business logic*

As you work with the Data Stewards, remember to require that the logic for the business rules be documented and captured into the Metadata Repository. The business rules and logic also include the context of the information to be stored. The SAS tool that can be used for this purpose is SAS Information Map Studio. When using this product, it is important to store the process flow of the data. This process flow should reflect the data's importance to your agency.

Deleted: s

The business logic can then be stored for access by users needing and having authority to the data using domain permissions. The metadata can be accessed by a wider group of users that need access to learn about the data, but may not need to actually process or report on the data.

D) *Interchange of Information*

One of the leading issues to the free exchange of information is the use of an Open Metadata API, for the SAS Open Metadata Repository and the Publish/Subscribe capabilities of Channels management. Data asset identification is the primary information that is created to populate the metadata repository as a requirement of the data exchange service definition in the FEA. Another requirement of the FEA is to communicate the information to users. As

part of the tools available in SAS EBI Server, the Portal (which is part of the Integration Technologies), gives Architects the ability to establish channels for communicating information on the source of the data as well as the data that is used in the actual analysis.

In order to create a syndication channel for RSS content related to external data sources that can be stored in the OMR you will first need a permission statement in the appropriate policy file. Then you will need to ensure that the appropriate permissions in the permission hierarchy are *stored in the OMR*. A best practice regarding this task is to avoid setting permissions at the person-level, instead opt for setting group-level permissions even if a group contains only one or two people. Managing user privileges via groups is likely to less labor intensive than at the person-level.

Convergence

Components

Recently and as a result of IT needing to be lean in its organization, more attention is concentrated on understanding the relationships of the various architectures and how they work together. Individuals that look at the FEA are just starting to see a thorough analysis of what it takes to document the logical data architecture and a contextual analysis of the metadata. Tools such as the Management Console let the SAS Administrator register the data into the Open Metadata Repository.

Technical Platform

In recent years, the problems of connecting machines and transferring data have been diminished. With the introduction of IPv6 connectivity this process is expected to be even easier. Still, as your Architect documents best methods for accessing data, the plans for improved connectivity should be reviewed and discussed.

Figure 2 Component flow

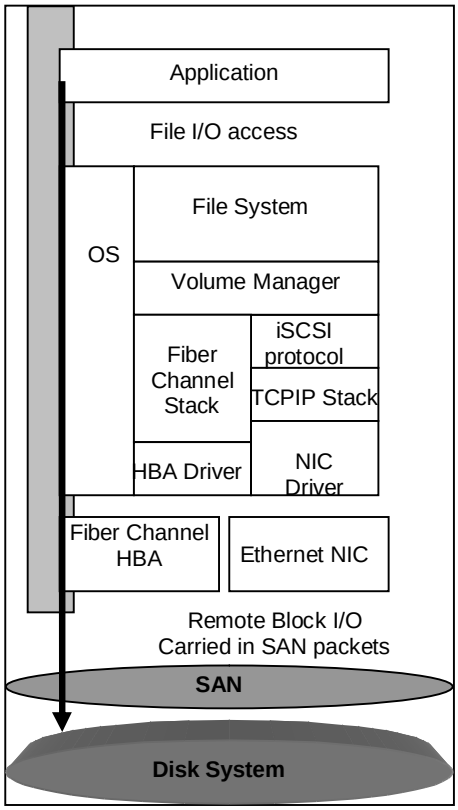


Figure 3 Services file

services.txt - Notepad

File Edit Format View Help

```
pptp            1723/tcp                          #Point-to-point tunnelling protocol
radius          1812/udp                          #RADIUS authentication protocol
radacct         1813/udp                          #RADIUS accounting protocol
js              1966/tcp    js_port               #Job Scheduler Port
nfsd            2049/udp    nfs                   #NFS server
kmetd           2053/tcp                          #Kerberos de-multiplexor
securepath      4987/tcp
sasmain_rs      5099/tcp    services_rmi_port     #SAS Remote Services Application Port
sasmain_ol      5451/tcp    olap_port sas_ol      #SAS OLAP Server
lsf_res_port    6878/tcp                          #LSF Scheduler Remote Execution Server
lsf_lim_port    6879/tcp                          #LSF Load Information Manager Service
lsb_mbd_port    6881/tcp                          #LSF Master Batch Daemon
lsb_sbd_port    6882/tcp                          #LSF Slave Batch Daemon
cnct            7551/tcp    sas_cs connect_port   #SAS Connect Server
tomcat          8080/tcp    websrv_port webapps   #Apache Jakarta Tomcat
default         8550/tcp    sasmain sasmain_as    #SAS Intrnet Default Application Dispatcher
sasmain_ss      8551/tcp    shareport sas_ss      #SAS Share Server
sasmain_ms      8561/tcp    omaport sasmain_os    #SAS Metadata Server port
spawner_lb      8571/tcp    spawner_loadbalancing_port #SAS Object Spawner load balancing port
spawner_op      8581/tcp    spawner_operator_port #SAS Object Spawner Operator Port
sasmain_ws      8591/tcp    iom_port sas_ws       #SAS Workspace Server
stp             8601/tcp    stp_port              #SAS Stored Process Server Port
stp1            8611/tcp    stp_port1             #SAS Stored Process Server Port1
stp2            8621/tcp    stp_port2             #SAS Stored Process Server Port2
stp3            8631/tcp    stp_port3             #SAS Stored Process Server Port3
man             9535/tcp                          #Remote Man Server
davsrc          9800/tcp    dav dav_port WebDAV   #SAS WebDAV Server Port
```

Performance

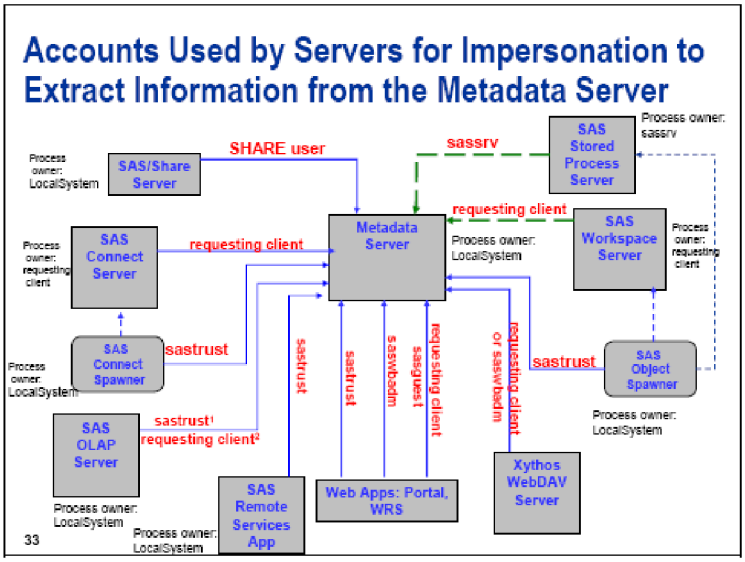
When OMB asked (in a RFI), "Does current search technology perform to a sufficiently high level to make an added investment in metadata technology unnecessary in terms of cost benefit?" The results were not overwhelming to one side or the other in the debate with 56% of the respondents believing that as of December, 2005 that computer "brute force" could solve the problem.[2] However, this relies on Moore's Law to stay true and for processing power to increase at a prodigious rate. Also, other areas of the infrastructure will have to improve performance or else they will become the bottlenecks.

Two other areas to watch are the operating system, and the network. Starting in 2006, there is the hope Windows Vista will improve the use of resources, and the increased use of fiber optics in networks and IPv6 will improve the performance of the network infrastructure. The increasing growth of data, however, will mean that ever increasing performance will be necessary.

Security

One lesson learned the hard way is that SAS Administrators need to understand the security rules that exist at their site prior to starting the installation of SAS Enterprise software. An early understanding of the process for authentication and/or encryption at your agency will save the SAS Administrator from having to reinstall the SAS software multiple times. During the SAS installation process when requested for the accounts to be used, enter the account as *domain\account* rather than simply the account as documented in the installation guide. This extra step will prevent having to reinstall the software when multiple servers are involved in the configuration. The SAS EBI Server services installation documentation shows the following diagram:

Figure 4

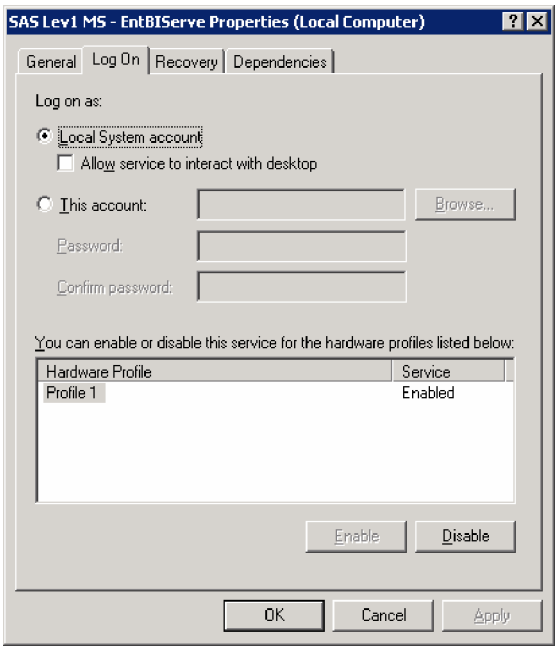


The diagram in figure 4 shows the Server Applications that are needed by EBI and ETL server.

In the diagram in Figure 4, the accounts listed in red represent the accounts that would act as the Process Owner to the service. The Process Owner is what should be used in the Log On tab of the Service Properties. In Figure 5, a service is set to use LocalSystem for its credentials. A Domain Local Account could have been set for the services where the process owner is not required to be LocalSystem as indicated above and in the installation manual. These are some of the needed Domain Local Accounts:

SASadm - Service Level SAS Administrator account can be used as the requesting client.
SASsrv – Service level account for remote service startup of SAS.
SASguest – Used for requests with minimal access to other systems.
SAStrust – Used for requests with a high level of access to other systems.
SASwbadm – Used for service requests controlling the web and portal administration.

Figure 5 – Windows Service properties, logon tab



Conclusion

The Federal Enterprise Architecture can provide a standardized way to look at the overall picture of how the SAS System integrates into an organization's Information Technology infrastructure. The SAS Intelligent Architecture tools that are provided as part of ETL Server and EBI server provide the ability to draw the map

Acknowledgments

I thank our System Administrators for helping gather information for this paper. I also thank our other SAS Administrators (Chris Zogby and Mikki Waid), our Chief Architect, my manager, and my wife (Robin) for reviewing the paper.

Contact Information

Your comments and questions are valued and encouraged. Contact the author:
John McCue
Congressional Research Service
101 Independence Avenue, LM413
Washington, DC 20540
Phone: (202) 707-2818
Email: jmccue@crs.loc.gov

References

[1] Data Reference Model, Version 2.0, November 17, 2005, Federal Enterprise Architecture Program
[2] "To Tag or not to tag", Joab Johnson, 01/09/06, GCN Government Computer News, pg. 25
3 "SAS ® 9. 1. 3 Integration Technologies Server Administrator's Second Edition", *http://support.sas.com/rnd/itech/updates/913/admin_oma_sp2.pdf*