

SAS Environment Upgrade: The PRA Experience

Gary Stevens, PRA Health Sciences, Reading, UK

ABSTRACT

During 2014, PRA Health Sciences was one of the first companies in the industry to evaluate the benefits of implementing the new SAS® Version 9.4. Billed as an upgrade with significant increases in functionality, we were tasked with a deep-dive gap analysis looking at the major enhancements SAS 9.4 would have over the existing SAS 9.1.3 environment. Safely migrating our current studies to the new environment with minimum impact whilst ensuring a seamless transition was of paramount importance. The final business decision was to implement a entirely, new SAS environment, including moving to a SAS Grid-enabled system accessed via a Citrix connection. SAS Enterprise Guide would be used as the front-end user interface, moving away from the typical SAS Windows-style Base SAS system. The encoding character dataset was to be upgraded to UTF-8 from wlatin1 along with an upgrade from a 32-bit system to a 64-bit system.

INTRODUCTION

With such a major change in the SAS infrastructure there were many challenges, decisions and unknowns to overcome. This paper details how PRA executed a successful transition to this new environment, the benefits we were able to realize from the increased functionality that SAS 9.4 offers, and a few points for consideration when upgrading your SAS system.

RATIONALE

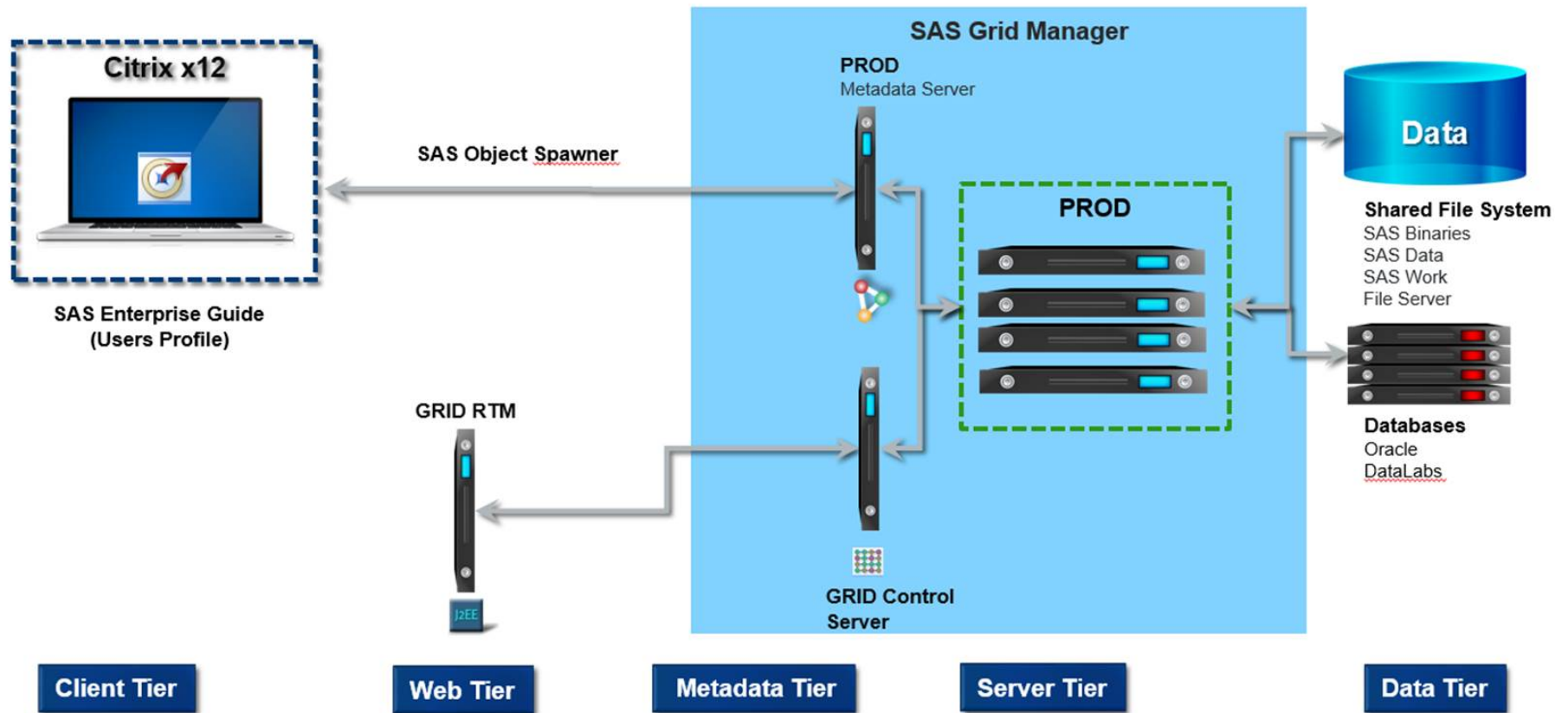
When upgrading our SAS environment we had two main drivers. Primarily we were looking for increased performance of our SAS application. With a growing user base the current server was becoming increasingly unable to meet our processing needs, and maintenance costs were becoming ever more expensive. Secondly we were looking for scalability. We needed an environment which would allow the flexibility to incrementally grow as our data and processing needs grew in the future. The decision to upgrade as indicated above would satisfy these requirements. Added incentives of this upgrade were the achievement of higher resource utilization and an overall lower cost of ownership.

SAS GRID ARCHITECTURE

The steps involved to run a SAS program, from a technical point of view are:

1. The user connects to one of 12 Citrix servers where Enterprise Guide physically resides along with the user's profile. The assignment of the Citrix server is load balanced, which means the least busy server gets the connection.
2. The user selects and launches Enterprise Guide from the Citrix receiver desktop. This request is for a SAS Workspace Server, which goes to the SAS object spawner (listening daemon) and passes this request to the SAS Metadata server.
3. The user request goes through an authentication and authorization process to make sure the user is defined within the SAS Metadata Server and has the correct roles and capabilities assigned to them in order to launch a SAS Workspace Server.
4. Once authentication and authorization has been approved, the request is forwarded to the SAS Grid Control server, which takes the request and sends to one of the of four Production servers on the grid (SAS Nodes).
5. The Grid control server contains information with regards to load across the GRID and this request is then submitted to the least busy GRID Node.
6. Once submitted and the program has run, the log and any resulting output are sent directly back to the SAS Enterprise Guide session.
7. Once this initial connection has been made the EG session connects directly to the Grid bypassing the SAS object spawner and Metadata server for further submissions of code within the same session, on the same node assigned by the Grid control server originally, until the Citrix connection is terminated or a user launches another separate Enterprise Guide session.

PRA PROD SAS GRID ARCHITECTURE



PhUSE 2016

The servers installed are virtual machines, running Windows 2012 in a 64-bit configuration. The SAS work drive on each SAS Node Server is a high performance 200GB solid state drive. General storage is provided through a Windows based file server which is connected to PRA's storage area network.

64 BIT PROCESSING

One of the further enhancements was upgrading to a 64 Bit system along with SAS 9.4. Upgrading this component greatly increases the amount of memory the CPU can use. The number of bits in a processor refers to the size of the data types that it handles and the size of its registry. A 64-bit processor is capable of storing 2^{64} computational values, including memory addresses, which means it's able to access over four billion times as much physical memory than a 32-bit processor. As well as greater memory utilization and the speed that this offers there is also a much improved resolution in graphics and this matches our MS office products such as Word which also is running on a 64-bit system.

SAS 9.4 CONFIGURATION

One thing which was important in the upgrade to SAS 9.4 from SAS 9.1.3 was backward compatibility. Most of our clients were found to be using SAS 9.2 and it was imperative to be sure that any data sent to a client was readable by the client in these earlier versions of SAS. An internal analysis found that prior versions of SAS could not read datasets produced by SAS 9.4. Programs were no problem as long as they did not use the new functionality of SAS 9.4 since they are merely text files.

Following investigation it was found the option EXTENDOBSCOUNTER affected the observation count attribute of datasets which impacted portability. In SAS 9.4 this option was defaulted to YES impacting readability with previous versions of SAS. As a result the PRA SAS 9.4 environment is configured with 2 autoexec files; the first called the global autoexec, which contains amongst other things, the EXTENDOBSCOUNTER set to NO which cannot be changed by the user, allowing SAS 9.4 datasets to be read by previous versions of SAS.

Functionality of the EXTENDOBSCOUNTER option:

- **NO** Specifies that the maximum observation count in a newly created SAS data file is determined by the long integer size for the operating environment. In operating environments with a 32-bit long integer, the maximum number is $2^{31}-1$ or approximately two billion observations (2,147,483,647). In operating environments with a 64-bit long integer, the number is $2^{63}-1$ or approximately 9.2 quintillion observations.
- **YES** requests an enhanced file format in a newly created SAS data file that counts observations beyond the 32-bit long limitation. For a SAS data file that is created for an operating environment that stores the number of observations with a 32-bit long integer, the data file behaves like a 64-bit file with respect to counters.

The global autoexec is used at every invocation of SAS. The purpose of the global autoexec file is to set up options and settings which are consistent across the programming environment. The second autoexec file is at the study level to set up study specific libname statements and other SAS options. This is created and maintained by the study lead programmer.

ENCODING

Another upgrade implemented by PRA was to change the encoding character dataset. The UTF-8 character dataset PRA switched to is a very broad character set containing over 100,000 printable characters. The previous character set used was WLATIN1 and this is commonly used by our clients and has only 256 characters. One of the reasons why UTF-8 was decided upon as the standard was to enable support for emerging markets such as China. There are, however, issues inherent in this move to this new character dataset. Clients not using UTF-8 encoding in their SAS environments would not be able to see the data as the encoding dataset has to match between PRA and the client.

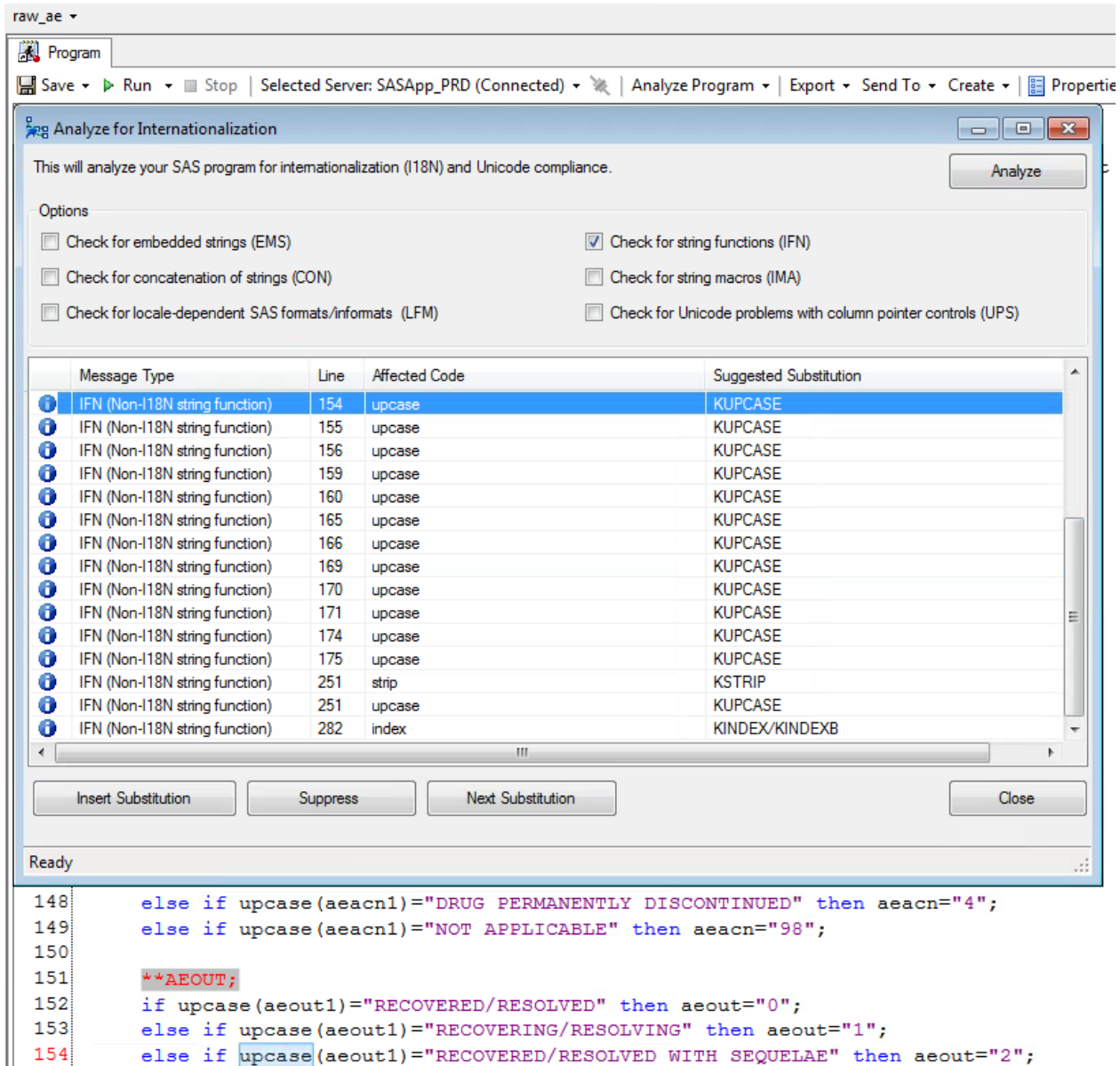
WLATIN1 is a single-byte encoding character dataset. This means one character is 1 byte. UTF-8 is a variable-width multi-byte encoding. One character in the UTF-8 encoding can be 1 byte, 2 bytes, 3 bytes, or even 4 bytes.

The SAS system provides many string functions and call routines that can be used to manipulate characters and strings. The original SAS string-handling functions assume this size of a character is always one byte, which is true with a single-byte encoding such as WLATIN1. However, using the original SAS string-handling functions with double or multi-byte data such as UTF-8 can lead to unexpected behaviour, such as data truncation.

To resolve issues that these string functions can cause for multi-byte data, SAS provides a set of string functions, called K functions, which do not make assumptions about the size of a character in a string. Any SAS function which computes a length counts the number of bytes. This suits WLATIN1 encoded data. A K Function however counts the number of characters. The K Functions are simply the normal SAS function preceded with a K. i.e. KLENGTH,

PhUSE 2016

KTRIM, KLEFT, KCOMPRESS etc. Most of the SAS functions which process a length exist as K functions and can be replaced to prevent such issues. Before replacing all of the original SAS string-handling functions with K functions, examine your SAS code; if the string function processes data that only contains single-byte characters, there is no need to use K functions. SAS Enterprise Guide (Version 6.1) has a tool which can look through code for such KFUNCTION issues

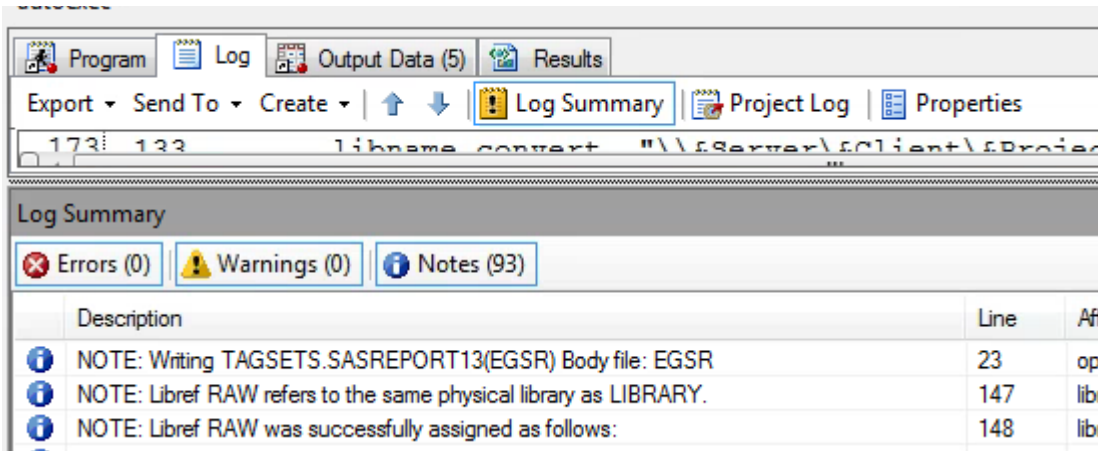


The above screen shot shows this tool in action. To access this tool in Enterprise Guide select the 'Analyze Program' tab at the top of the screen. Check the 'Check for string functions (IFN)' checkbox and then click on 'Analyze' and the tool will search the code for possible K function issues. You can see the "upcase" in the last line (line 154) of the program highlighted and you can decide to 'Insert Substitution', 'Suppress' or go to 'Next substitution'.

In any data sent to a client the encoding is an important aspect which needs to be considered. As most of our clients use WLATIN1 as their encoding character dataset we need to conform to this in deliverables. Along with this most of our clients prefer the data to be sent as SAS transport files (xpt files). These xpt files are only able to use ASCII or WLATIN1 encoding (single-byte character encoding) and there are issues when converting these to WLATIN1 xpt files if they contain characters which are not directly transcodable into WLATIN1. PRA has developed a suite of macros to tackle this issue converting any detected UFT-8 characters which are not directly transcodable, into WLATIN1 equivalents.

ENTERPRISE GUIDE

PRA uses SAS Enterprise Guide version 6.1 and is billed as a point-and-click, menu- and wizard-driven tool that empowers users to analyze data and publish results. It provides fast-track learning for quick data analysis, generates code for productivity and speeds your ability to deploy analyses and forecasts in real time. PRA however, simply uses SAS EG as a coding environment. There are however some features of SAS EG which PRA does use, such as the log summary which will analyze the log of a run program for Errors or Warnings.



The user opens the log in SAS EG and then run the Log Summary tool. The button's for Errors, Warnings and Note's toggle on and off. Turning these on, the ERROR/WARNING/NOTE will be shown and clicking on the ERROR/WARNING/NOTE will enable the user to navigate directly to the appropriate place in log. This tool unfortunately does not give undesirable WARNINGS such as numeric to character conversions, multiple by variables in merges etc. and so PRA has additional utilities to deal with these. It's more useful for program development. Another useful functionality of SAS EG is that the work datasets produced within a program are readily available under the 'Output Data' tab, removing the need to search for the dataset in order to open.

CONCLUSIONS

The process of migrating and adopting the new SAS environment took a little longer than we originally expected due to unforeseen technical issues. Once these were overcome, all projects were transitioned in 4 separate waves over a period of 4 months. Users are now transitioned to the new set-up, appreciating greater system stability and performance. As our clients upgrade to a similar set up, with regard to SAS version and encoding, we can begin to implement some of the new functionality we found when initially looking through SAS 9.4 documentation.

CONTACT INFORMATION

Contact the author at:

Gary Stevens
PRA Health Sciences
500 South Oak Way, Green Park,
Reading, RG2 6AD, United Kingdom
Work Phone: +44(0) 1189 181211
Email: stevensgary@prahs.com
Web: www.prahs.com

Brand and product names are trademarks of their respective companies.