Paper 9920-2016

UCF SAS® Visual Analytics: Implementation, Usage, and Performance

Scott A. Milbuta, Carlos Piemonti, and Ulf Borjesson, University of Central Florida

ABSTRACT

At the University of Central Florida (UCF) we recently invested in SAS® Visual Analytics, along with the updated SAS® Business Intelligence platform (from 9.2 to 9.4), a project that took over a year to be completed. This project was undertaken to give our users the best and most updated tools available. This paper introduces the SAS Visual Analytics environment at UCF and includes projects created using this product. It answers why we selected SAS Visual Analytics for development over other SAS® applications. It explains the technical environment for our non-distributed SAS Visual Analytics: RAM, servers, benchmarking, sizing, and scaling. It discusses why we chose the non-distributed mode versus distributed mode. Challenges in the design, implementation, usage, and performance are also presented, including the reasons why Hadoop was not adopted.

INTRODUCTION

At UCF we recently upgraded our existing SAS BI environment from 9.2 to 9.4 and implemented a new SAS VA environment. This paper provides information on our existing SAS 9.4 BI environment and describes the implementation of our SAS VA environment and how UCF is getting up to speed using this new tool. We will be looking at the decisions regarding data structures for use in explorations and reports and some of the challenges we faced in our implementation of our SAS VA tool. We will also present snapshots of projects created using this tool and we will be answering why we are selected it over other SAS applications available.

Additional key points are some of the challenges in the design, implementation, usage, and performance. Detailed explanations of the technical environment for our non-distributed SAS VA setup including but not limited to RAM, servers, benchmarking, sizing and scaling, and why we chose this mode instead of a distributed SAS VA environment will be discussed. Finally, why Hadoop was not adopted for our solution.

UNIVERSITY OF CENTRAL FLORIDA – INSTITUTIONAL KNOWLEDGE MANAGEMENT

UCF's reporting needs are as varied as any other academic organization. At UCF Institutional Knowledge Management (IKM) is tasked with providing actionable information regarding our University to internal and external customers alike. The primary resource to facilitate all of UCF's reporting needs is the implementation of our SAS 9.4 BI environment and SAS VA suite.

To help in the endeavor two units within IKM (see Figure 1 below), Enterprise Decision Support (EDS) and Institutional Research (IR), provide the resources when it comes to report generation and support required to disseminate this information to our consumers.

UCF–IKM SAS ENVIRONMENT

Within IKM, EDS is mainly a SAS® EBI shop. We are currently running SAS® Enterprise BI Server 9.4 with a host of suite tools listed as follows. Also, screen capture, Display 1 below summarizes UCF–IKM's BI platform server.

- SAS® BI Dashboard 4.4
- SAS® Data Integration Studio 4.9
- SAS® Enterprise Guide® 7.1
- SAS® Information Map Studio 4.4
- SAS® Management Console 9.4
- SAS® OLAP Cube Studio 4.4
- SAS® Web Report Studio 4.4
- SAS Visual Analytics 7.1

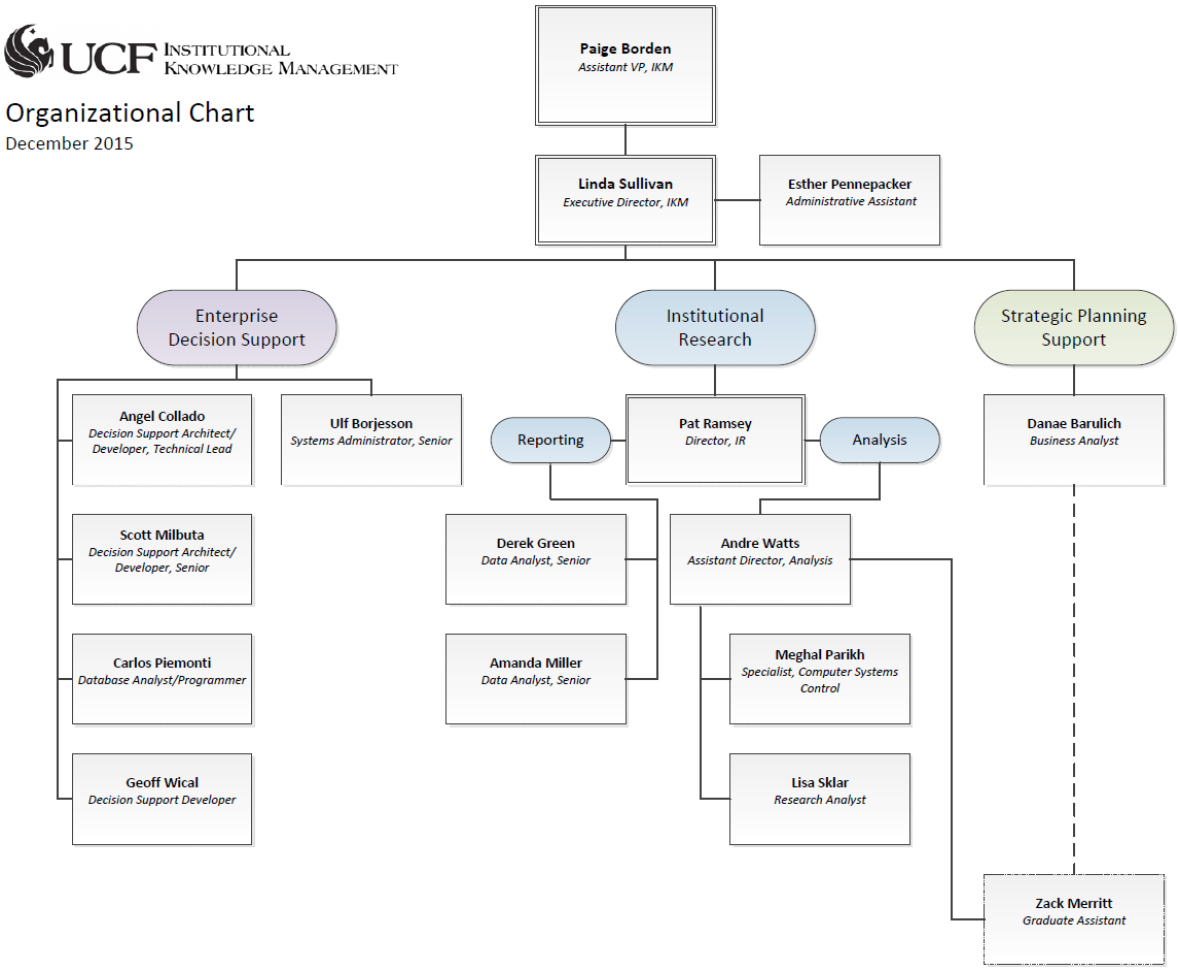
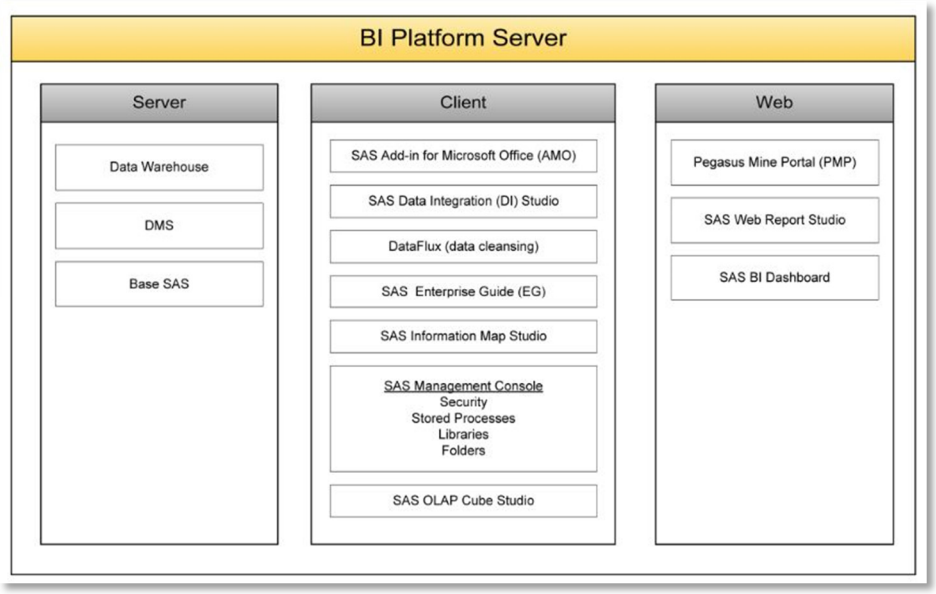


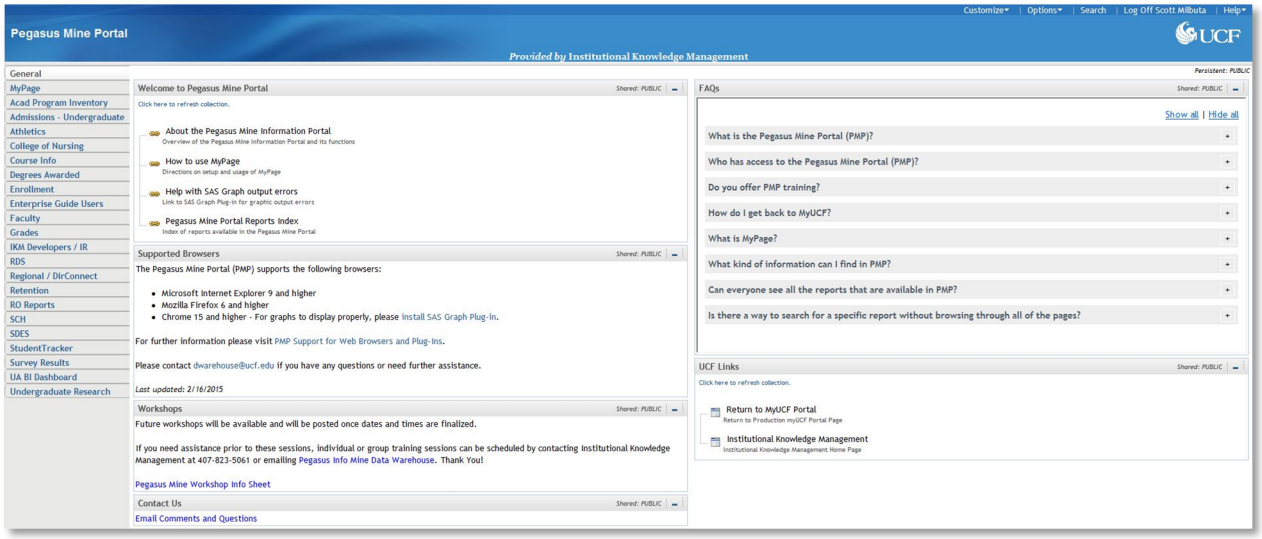
Figure 1. UCF–IKM's Organizational Chart as of December 2015



Display 1. UCF–IKM's BI Platform Server

UCF–IKM'S PEGASUS MINE PORTAL

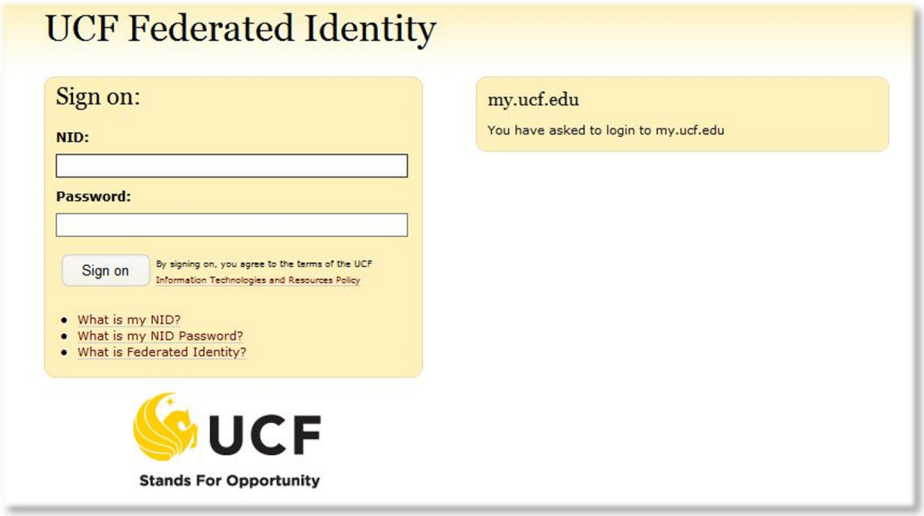
UCF–IKM uses the SAS Information Portal, called the Pegasus Mine Portal (PMP) (see screen capture, Display 2. UCF–IKM's Pegasus Mine Portal below), for delivery of static reports, information dashboards and stored processes with Excel and PDF output to our consumers on campus.



Display 2. UCF–IKM's Pegasus Mine Portal

Much of the data surfaced for our reporting needs comes from state census files generated from our operational data store. The primary transaction data store from the university student enterprise system is an Oracle Database 12c environment. Additional data resources for snapshot information come from a Microsoft SQL server environment running a COGNOS RDS overlay.

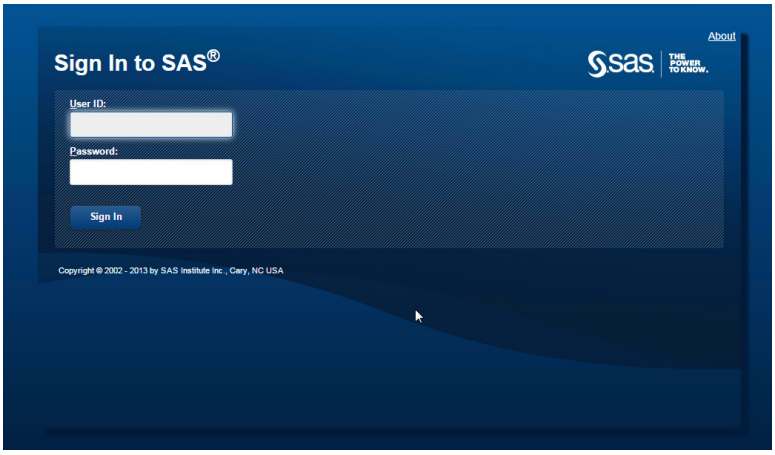
The PMP is the gateway to access university data and information in a secure web environment. It is sitting behind UCF's firewall with a single sign on security (see screen capture, Display 3. UCF's Single Sign below). All UCF full-time faculty and staff are automatically given access to PMP. OPS and part-time employees can request access via their department's security authorizer.



Display 3. UCF's Single Sign On Security

UCF–IKM VA ENVIRONMENT

In addition to all of the information and resources available to UCF with its current portal environment, a new reporting method was needed to meet some of the more analytical demands the executives required in order to make quick actionable decisions. In 2014 UCF–IKM initiated the process of installing our new SAS VA 6.4 server and necessary tools. Sign on is required to execute this environment (see screen capture, Display 4 below).



Display 4. SAS VA Sign On Security

UCF–IKM VA DATA GOALS

Once the VA server was in place there were some important initial goals UCF–IKM wanted to achieve with this new environment. The first step in setting up our environment was ensuring we had properly scaled it to meet our planned growth needs.

UCF–IKM's initial rollout of the VA server was set for approximately forty users. These users are comprised of developers, administrators, and eventually will include a selected executives to preview the type and scope of information delivered. As new development is generated, additional users will be included up to an estimate of four hundred non-concurrent users.

The reason for the limited scope of access is related to current information needs and future development explorations. Additionally, user access is being carefully managed as we gauge performance impact when new visuals are rolled out. A secondary reason for controlling user access is that in our current environment we only surface atomic student data in discreet reports within the RDS environment in the Portal. We are investigating the potential use of VA for generating explorations of atomic student level data incorporating row-level security.

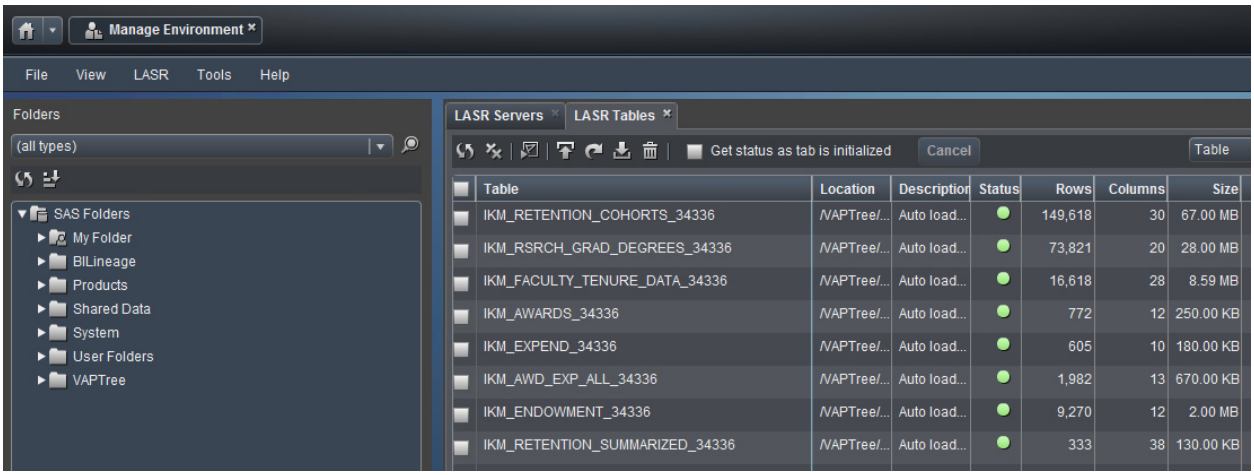
The second hurdle we faced was the 'what and how' regarding our current data structures: would they serve as intended for the reports as currently designed, or would brand new sets of data need to be developed for use exclusively on the VA server. Additional considerations needed to be examined regarding how users will use the environment and the overhead of their usage with regards to the reports they require and the impact on the server memory component.

UCF–IKM: OUR USE OF VA

Initially we looked at our existing BI data warehouse and many of the fact and dimension tables that we house there and decided to load the raw datasets into VA. Our initial understanding was that the same data structure used in the BI data warehouse would work just the same in VA. We quickly discovered that the application would require custom tables of pre-summarized data to be able to create reports in VA. This was accomplished by first creating the necessary datasets on the BI platform and then moving them to the VA autoload library.

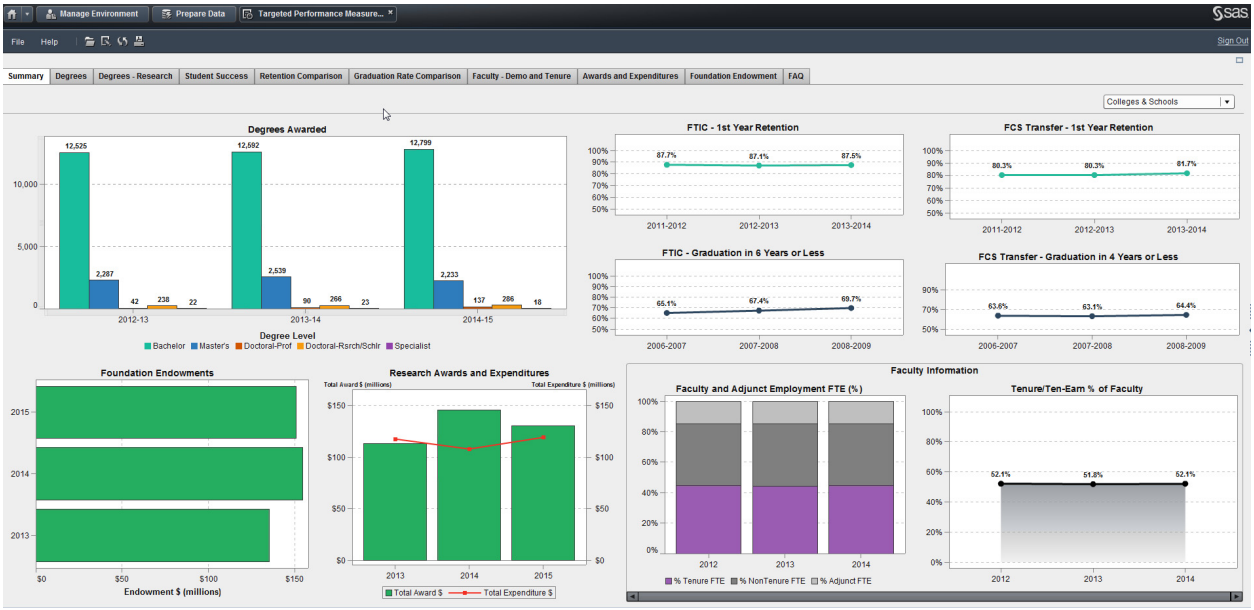
The next step was to create dynamic reports that would allow user interactivity and to present analytical visualizations of the data. One such report that is leading development is the “Targeted Performance Measure Dashboard”. This report is staged to serve as UCF’s executive dashboard displaying information on five selected metrics at the university level with the ability to drill into colleges and departments. This was a capability that we could not easily accomplish without using multiple reports with multiple maps in the traditional BI Portal environment.

The screen capture presented on Display 5 below shows the list of tables created for the “Targeted Performance Measure Dashboard”.



Display 5. “Targeted Performance Measure Dashboard” Tables

The following screen capture, Display 6, shows the summary page of “Targeted Performance Measure Dashboard” report.



Display 6. “Targeted Performance Measure Dashboard”: Summary Page

UCF–IKM VA environment is also being used for development on reports dealing with Peer Benchmarking analysis, University Strategic Planning, UCF Athletics information, and UCF Foundation fiscal information.

CHALLENGES

One of the challenges we faced within the new VA environment was learning best practices regarding our own data in order to know what would work.

Many of our reports use information maps (SAS BI environment) that have one-to-many joins depending on the data and prompt selections. We quickly realized that this was not viable because of the way VA establishes relations and query objects. We ran into difficulty regarding efforts to build reports using the star schema methodology due to the fact that tables within VA may only be joined on one key at a time. The output produced by the data query created large datasets, which utilized more memory than expected.

This created unexpected complications that we subsequently had to roll back from and develop large customized data sets specific to a report's needs. Multiple de-normalized data sets were developed in order to accommodate drilldown requirements for specific unit and organization reporting.

Another challenging area was embedded VA Data Query objects. Initially, several reports were built using the results of Data Query objects within the VA environment. However, upon a server reboot the query objects data set was no longer in memory and the query had to be re-run in order to regenerate the data set used for a report. This would happen every time the server was restarted. It was decided not to use embedded VA Data Query objects in our environment.

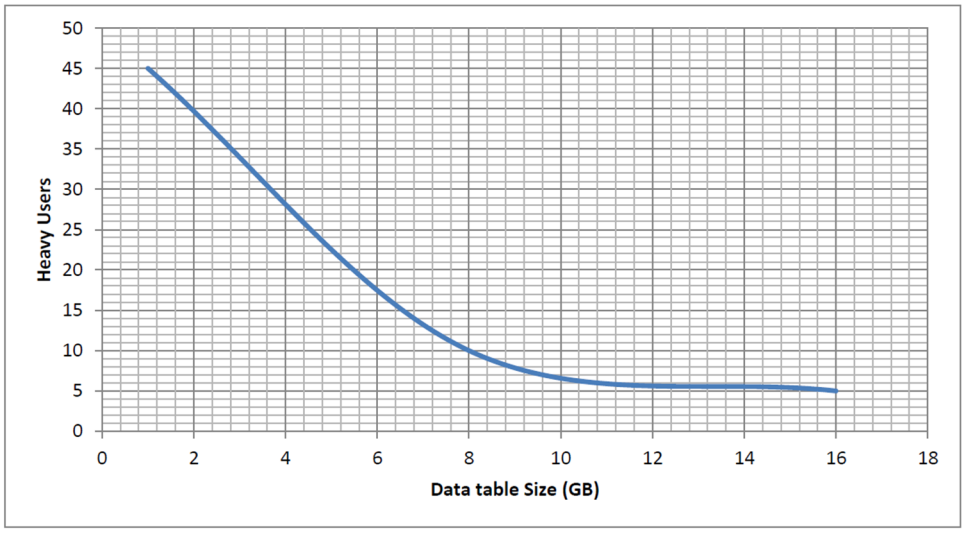
UCF–IKM VA TECHNICAL SPECIFICATIONS

SAS VA PERFORMANCE GUIDE

To support a large number of users on the SAS VA platform an organization needs to use a distributed environment of SAS VA. This setup runs on Linux and is highly expandable and nodes can be added to support an increase of users if when required.

Setup and support of a server cluster for a distributed implementation is more expensive than a non-distributed implementation of SAS VA. Additionally, support of a multi-server implementation can be more challenging than a one server setup and potentially more difficulties when patching and trouble shooting.

To help UCF–IKM make the environment decision we needed to look over our sizing information (see screen capture, Display 7 below). We also needed to examine our user base and decide how many users will be served in the VA portal and future use concepts. The administrators also considered the environment, either Linux or Windows, as well as funding.



Display 7. SAS User Estimated Matrix (source: SAS Server Sizing Analysis)

IMPLEMENTATION

UCF–IKM chose a hybrid of the two models.

A non-distributed environment was implemented based on funding allocations with regards to the scale and scope of our server. UCF–IKM also chose a Windows deployment for continuity with our existing Windows infrastructure which also reduced maintenance overhead. These two factors were the largest contributors in the decision to proceed with a non-distributed environment.

Additionally, UCF–IKM implemented a SAS Web server (reverse proxy server) in front of the SAS VA server to add an extra level of security with little cost associated with it. The UCF–IKM environment is in fact a two server implementation but still considered a non-distributed implementation because the servers do not share data processing resources. Only one VA server is in production and the other hosts the portal environment.

Hadoop was considered and reviewed but the decision was made not to implement it, primarily due to the size of the non-distributed VA environment. Hadoop was not a cost-effective solution for the size of our implementation.

HARDWARE

The hardware for the two servers is as follows.

The SAS Web server runs on VMware with a 2 vCPU allocation and 2GB of RAM with three virtual hard drives. The primary drive hosts the OS, MS 2008R2; it also hosts all other MS products including Net Backup and any other required software for maintenance of the system. The secondary drive hosts the SAS VA environment, and the third drive is used to exclusively host the Windows Page File. This provides an optimal configuration and allows the system to expand dynamically depending on real time usage and performance.

The throughput we get on these drives is about 600MB/s read and a 200MB/s write. The SAS Web server serves as a gateway to the SAS VA environment with minimal overhead impact and no performance degradation under load. A final item of note is the data Queue Length on the drive configuration. In our testing and production environment under load, a constant value below 0.05 can be maintained 95% of the time peaking near 0.25 at maximum.

Our SAS VA server is a Microsoft 2008R2 server with a 10 core x 2 socket setup with 256GB of RAM. The 24 Serial Attached SCSI (SAS) hard drives on a Dell 715xd are setup in a raid 5 configuration to maximize the data environment storage. The trade off in performance for storage on a raid 5 configuration was acceptable in the study of our environment.

The base line performance for our current environment is approximately 2.2GB/s read and 2.0GB/s write. Although this is much lower when compared to a SSD or Fusion card implementation, this has provided great performance at significantly reduced cost. The data throughput can load approximately 10GB of data from the drive bank to memory in about 12 minutes. Finally, the Queue Length on this array stays below 0.01 for 95% even when all applications are loading concurrently. Our server infrastructure also provides communication on a 10GB fiber network which reduces lag and data transfers times.

CONCLUSION

Performing SAS upgrades or installations may become a challenging task. There is much preparation, research, and frustrating attempts involved in the process. SAS platform versatility can be adapted to virtually any system, thus making its level of personalization and tweaking a very intricate process. As a consequence, many hours have to be invested communicating with SAS to solve ‘issues’ that are no more than mere specific setup steps or very specific conditions, not necessarily disclosed with standard installation processes due to the variety of ‘possible’ scenarios.

In this sense, we had the fortune at our institution to rely on a person that had the disposition and patience to deal with this entire process until its very end. The only person in the US that at the time of completion (according to a few SAS employees in a casual elevator conversation during SAS Global

Forum 2015) performed an upgrade to SAS EBI 9.4 and an installation of SAS VA successfully (outside the SAS Institute). All the merit goes to Ulf Borjesson and the team that supported him along the way.

We just adopted VA and we still have a long way to go into it. Our initial steps, carefully performed and documented, helped us to successfully deploy reports never created before in our institution. We are looking forward to developing and improving the quality of the information and services we are sharing with our consumers and with the public in general.

To choose between a non-distributed versus a distributed implementation of SAS VA one needs to consider key factors within the organization (current environment, funding, etc.) and what would serve consumers best. At UCF–IKM our decision to go with a non-distributed environment was the right choice that has allowed us to successfully grow our VA environment in many areas much more quickly.

REFERENCES

Vitron, Christine. Holman, James. “Considerations for Adding SAS® Visual Analytics to an Existing SAS® Business Intelligence Deployment.” SAS Global Forum 2014. March 23-26, 2014. Washington, DC. Available at: http://support.sas.com/resources/papers/proceedings14/SAS146-2014.pdf

ACKNOWLEDGMENTS

Sincere gratitude and appreciation is extended to Dr. Paige Borden, Dr. Linda Sullivan, Ms. Patricia Ramsey, Ms. Evangeline Collado, Mr. Andre Watts, Ms. Danae Barulich, Ms. Amanda Miller, Mr. Meghal Parikh, and Ms. Lisa Sklar; for all their input and persistent help during the implementation of the SAS BI and SAS VA environments.

RECOMMENDED READING

- *Hadoop Illuminated*

CONTACT INFORMATION

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

Scott A. Milbuta
Decision Support Architect/Developer – Senior
E-mail: Scott.Milbuta@ucf.edu

Carlos Piemonti, MBA
Database Analyst/Programmer
E-mail: Carlos.Piemonti@ucf.edu

Ulf Borjesson
Senior Systems Administrator
E-mail: Ulf.Borjesson@ucf.edu

University of Central Florida
Institutional Knowledge Management / Enterprise Decision Support
12424 Research Pkwy, Suite 215
Orlando, FL 32826
Web: https://ikm.ucf.edu/