Paper 3470-2015

SAS® Visual Analytics: The Value in Leveraging Your Preexisting Datasets

Shaun Kaufmann, Farm Credit Canada.

ABSTRACT

As a risk management unit, our team has invested countless hours in developing the processes and infrastructure necessary to produce large, multipurpose analytical base tables. These tables serve as the source for our key reporting and loss provisioning activities, the output of which (reports and GL bookings) is disseminated throughout the corporation. Invariably, questions arise and further insight is desired. Traditionally, any inquiries were returned to the original analyst for further investigation. But what if there was a way for the less technical user base to gain insights independently? Now there is with SAS® Visual Analytics. SAS Visual Analytics is often thought of as a big data tool, and while it is certainly capable in this space, its usefulness in regard to leveraging the value in your existing datasets should not be overlooked. By using these tried-and-true analytical base tables, you are guaranteed to achieve “one version of the truth” since traditional reports match perfectly to the data being explored. SAS Visual Analytics enables your organization to share these proven data assets with an entirely new population of data consumers--people with less "traditional data skills” but with questions that need to be answered. Finally, all this is achieved without any additional data preparation effort and testing. This paper explores our experience with SAS Visual Analytics and the benefits realized.

INTRODUCTION

SAS Visual Analytics is a product for analyzing, exploring and visualizing data. Consisting of a suite of products, SAS Visual Analytics is a web-based solution that leverages the SAS LASR Analytics Server to enable high-speed data access. Specifically, the LASR Analytics Server provides a multiuser environment for concurrent access to data that is loaded into memory and this is what makes Visual Analytics different from other SAS products. In this approach data resides in RAM instead of on disk, greatly reducing response time and allowing users to explorer huge volumes of data very quickly. Additionally, Visual Analytics enables interactive analysis of data by users lacking a background in traditional data skills, thus enabling an entirely new population of data consumers.

In order to ensure that your organization’s SAS Visual Analytics implementation actually enables its users, two requirements must be met: you must have data worth analyzing, and you must have a sufficient memory allocation to achieve the analysis. Visual Analytics is a powerful tool for exploring data. However, providing *data worth exploring* is the key to achieving business value. Put another way, the data you load into the system must be appropriate to provide the insights that you seek. Once you have determined the datasets necessary to effectively enable the user base, you must also ensure that the system has sufficient RAM to support the subsequent analytical tasks.

Striving to enable the most users possible per gigabyte of RAM should also be a primary goal when determining what data to load into the LASR Analytics Server. It is important to establish a formal strategy regarding dataset selection and memory management. Choices made regarding the mechanism of load, data volume, refresh frequency, data source, data granularity and data element selection will have an impact on a broad range of areas including server sizing, system stability, data quality and user satisfaction.

This paper explores the areas of consideration necessary to determine an effective strategy for data selection and memory management, with a goal of maximizing the business value returned by your investment in SAS Visual Analytics.

THE ROLE OF SAS VISUAL ANALYTICS IN YOUR DATA ENVIRONMENT

VISUAL ANALYTICS AND YOUR EXISTING DATA WAREHOUSE

Visual Analytics is a powerful tool for data exploration and knowledge discovery. However, it is not a replacement for a corporate data warehouse. As with every business intelligence product, Visual Analytics provides the greatest value when it leverages the investment that an organization has already made in its data infrastructure. Significant effort has been expended to ensure the data contained in a warehouse repository is cleansed, transformed and consolidated in a manner that supports your organization's specific business requirements. Data preparation of this nature is a time consuming task, so there is significant efficiencies gained any time you can leverage preexisting data manipulation processes (often referred to as extract, transform, and load processes or simply ETL).

ANALYTICAL BASE TABLES

Data warehouses implementation approaches vary. But regardless of whether your organization takes a dimensional approach or the normalized approach, one key fact is likely true; the data to answer any real business question is not contained in a single table. In this situation an analyst's query must join a variety of tables to retrieve the required data. This requires sufficient knowledge of both database concepts and the underlying warehouse structure. Unfortunately not every analyst has the skills necessary to work efficiently in such an environment. One way to address this issue is through the creation of analytical base tables (ABTs). These ABTs are completely denormalized tables, specifically created to enable analysts to answer business questions from a single source table. This approach not only enables analysts with less developed data skills, but also increases the efficiency of even the most proficient data user.

Analytical base tables are populated from the underlying data warehouse tables and can be refreshed with a specific frequency. They are also fit-to-purpose tables, created to enable specific business processes. These ABTs can be developed to feed a variety of business processes including loss provisioning activities, economic capital calculations or monthly reporting activities. However, all these business processes share a common characteristic; additional analysis is often necessary to explain the results produced by these processes. It is the need for additional data analysis that presents an ideal opportunity to employ the power of Visual Analytics as part of these business processes.

OPTIONS FOR LOADING DATA IN THE LASR ANALYTIC SERVER

SAS Visual Analytics provides three ways to load data into the LASR Analytic Server. This section introduces each method and discusses the advantages and drawbacks to each approach.

SELF-SERVICE IMPORTS – TWO TYPES

Data imports that are performed by users are referred to as self-service imports. These self-service imports can be further subdivided into two types; Data imports into the Visual Analytics Designer or the Visual Analytics Explorer, and data imported using the Visual Data Builder. Both these approaches can be appropriate for proof of concept and prototyping work since they allow the power user to load data on their own without administrator involvement. However both approaches have drawbacks that make them less than idea for reoccurring processes. Specifically, using import functionality of the Visual Analytics Designer or the Visual Analytics Explorer tools requires the imported file to be transferred to the server and reloaded each time you want to work with it. Similarly, using Visual Data Builder to construct a query to assemble the dataset results in the query being executed each time you want to move the table into memory. This approach results in both unnecessary load on your data warehouse and unnecessary network traffic.

ADMINISTRATOR LOAD

The preferred approach for loading data into memory is by using the Visual Analytics Administrator product's *Interactive Load* function. This approach takes a single table that is registered in metadata and lifts it into memory. Using the Visual Analytics Administrator's capabilities to administer LASR tables is the most efficient approach to moving data in and out of the LASR Analytics Server's memory. However,

this approach is not without its drawbacks. It requires a single source table that is "VA ready", meaning that the table does not require any additional query processing (joining, filtering, etc.). This means that the table must be produced by some outside ETL process and the development of that process carries with it its own effort in terms of development and testing.

LASR SERVER - TYPES OF MEMORY USAGE

The RAM installed in the SAS LASR Analytical Server(s) meets two separate needs: there is a portion of memory required for system overhead, and there is a portion of memory available for dataset storage and analytical processing (often referred to as usable RAM).

The portion of RAM that is required for system overhead includes the memory utilized by both the operating system's processes and the SAS LASR Analytic Server's own processes. The portion of memory allocated for system overhead does not include the storage of datasets or any in-memory structures related to visualization or analysis of the datasets. As seen in Figure 1, in a non-distributed LASR Analytic Server environment (Single Server) the system overhead consumes approximately 30 percent of the physical memory on the server.

Usable RAM - SAS® Visual Analytics (Single Server Architecture)					
Number of Servers	Total Cores	RAM per Server	Total RAM	Usable RAM	Usable RAM (%)
1	4	64 GB	64 GB	~45	70%
1	6	96 GB	96 GB	~70	73%
1	8	128 GB	128 GB	~90	70%
1	12	192 GB	192 GB	~135	70%
1	16	256 GB	256 GB	~180	70%
1	32	512 GB	512 GB	~360	70%

Figure 1. SAS Visual Analytics Single Server RAM Configurations.

In a distributed LASR Analytic Server environment, system overhead is approximately 30% of the physical memory on each of the worker nodes and 100% for the root node. This is a function of the fact that the root node does not store any portion of the datasets. In a distributed environment the root node's role is to distribute data to the worker nodes and consolidate the results it receives from them. Possible SAS LASR Analytic Server distributed environment configurations (and the resulting usable RAM available for each) are presented in Figure 2 below.

Usable RAM - SAS® Visual Analytics (Distributed Environment)					
Number of Servers	Total Cores	RAM per Server	Total RAM	Usable RAM	Usable RAM (%)
4	16	64 GB	256 GB	~135 GB	53%
4	24	96 GB	384 GB	~200 GB	52%
4	32	128 GB	512 GB	~270 GB	53%
4	48	192 GB	784 GB	~410 GB	52%
4	64	256 GB	1024 GB	~540 GB	53%
8	128	256 GB	2048 GB	~1.2 TB	59%
16	256	256 GB	4096 GB	~2.6 TB	63%
32	512	256 GB	8192 GB	~5.6 TB	68%
* Usable RAM (%) is approximately 70% per worker node.					

Figure 2. SAS Visual Analytics Distributed Environment RAM Configurations.

The remaining usable RAM must meet two separate needs: a portion is required to store the datasets that are loaded into memory, and a portion is required for dynamic usage.

When a table is loaded into memory with the SAS LASR Analytics Server engine, the server allocates physical memory to store the rows of data.

When a Visual Analytics user interacts with the system, physical memory is used as the server performs the requested analytics operations. One example of such an operation is the act of summarizing a table. The amount of memory required for this type of operation depends on the specific data in the table and the specific operation requested. Operations such as a table summarization that use group-by variables can require more memory than those that do not. If the number of unique values in the group-by variable is large (high cardinality) then the resulting summarized dataset will be proportionally large as well. In these cases the server allocates memory for data structures such as a decision trees or temporary tables to improve performance, thus consuming additional memory resources.

MEMORY MANAGEMENT STRATEGIES

FACTORS TO CONSIDER IN ESTIMATING MEMORY REQUIREMENTS

Of the three types of memory utilization requirements discussed above, only the last two are within the control of an organization. The RAM required to store datasets is the most straight forward to estimate and control because the memory required to store the dataset is directly related to the table's size and the data it contains. Careful selection of a dataset that contains the minimum number of attributes and rows to enable the maximum number of business users' needs is a equation that can be worked out with a fairly high level of certainty.

The amount of memory required for dynamic utilization however is much harder to predict. This equation has a larger and more fluctuating set of variables to consider. They include: the number of concurrent users, the type of activities each user is performing, the size of the dataset that each user is acting on, and even the specific values stored in the dataset that is being accessed (e.g. group-by variables with high cardinality).

A STRATEGY FOR CONTROLLING VARIABILITY

When determining server memory requirements, much of the uncertainty is due to the wide range of unknowns at play. If a way can be found to control this variability, then server memory requirements can be more accurately estimated. One approach that can be effective is to load tables with known characteristics. It is at this point that current business processes can lend significant insight. Often one of the main reasons for purchasing Visual Analytics is to enable existing business processes by making analysis faster and easier. In this situation we can look to the tasks and queries that the analyst currently performs for clues. For example, does the analyst routinely summarize the table by sales area? If so, then the cardinality of this group-by is no longer unknown since an organization surely knows how many sales areas it has. Does the analyst routinely group-by month? Then the cardinality of this summarization is known as well. The key word in this approach is *routinely*. Not every query will be routine, but it is likely that some large portion of queries that support reoccurring processes will be. By recording usage patterns a frequency distribution can be developed. Additionally, server memory can be monitored to determine the actual impact of each routine task. Similarly, there is likely a pattern to be discovered in regards to concurrent users. Reoccurring analysis is generally performed by a known set of analysts and with a known frequency. This provides insight into the number of concurrent users the system must be able to support. From this data a usage model can be developed that is specific to your organization.

Initial server size estimates are determined by requesting an official sizing exercise through your account executive. However, for organizations that acquire Visual Analytics to perform analysis in the support of routine and well understood business processes, there exists an opportunity to further refine your initial estimates. This opportunity is present because of the known characteristics of your users, processes and data.

PRODUCING "VA READY" TABLES

As mentioned earlier, the preferred method of loading data into memory is to utilize the Visual Analytic Administrator product's functions related to administering LASR tables. The main drawback of this approach is that the tables to be loaded must be in a "VA ready" state and in order to reach this state, significant data preparation in often necessary. This data preparation is generally referred to as an *extract*, *transform* and *load* process (or ETL for short).

Conceptually, an ETL process extracts data from data sources, transforms the data into the format and structure required for a specific purpose and loads the results into a target table. Unfortunately this brief description fails to communicate the complexity of the process, specifically in the transformation phase. In reality, the transformation process is usually non-trivial both in terms of complexity and the processing resources required.

ETL processes require significant effort to develop, implement and test. This effort is related to a broad range of activities including defining and capturing the business requirements, determining appropriate data mappings, evaluating data quality and performing data cleaning, as well as the act of implementing the process. As with any complex development initiative, a significant testing effort is required to ensure the quality of the final product. All this requires time, effort and money.

A NOTE ABOUT DATA QUALITY

The quality of an analysis, visualization or report is directly related to the quality of the dataset on which it was based. Having erroneous values present in the dataset leads to artifacts that are visible in the visualizations and reports produced. For example, hierarchies are dynamically generated in the Visual Analytics Designer and Explorer tools. These hierarchies are only accurate when the data from which they were derived is accurate. Every effort should be made to ensure the quality of the data loaded into Visual Analytics.

ONE VERSION OF THE TRUTH

Additionally, it is worth noting that implementing logic in multiple places invariably leads to inconsistencies in the data. When inconsistencies exist in the data an opportunity as exists to produce conflicting results when an analysis is performed. This situation should be avoided at all cost because it can call into question the validity of the results produced by your analytics environment. Once your organization loses faith in the accuracy of the data it is extremely hard to regain. Therefore it is imperative that the data you load into Visual Analytics is consistent with the data that is produced for consumption by other business processes. Loading Visual Analytics with data that was produced by leveraging existing ETL processes ensures consistency and guarantees *one version of the truth*.

THE VALUE IN YOUR PREEXISTING DATASETS

The previous sections introduced several challenges that must be addressed in order to ensure that your organization can derive value from its investment in Visual Analytics. In this section we examine how your preexisting datasets allow you to address each of the issues identified above. Specifically, utilizing your organization's preexisting datasets:

- Allows for the use of the *Administrator Load* method of lifting data into the LASR Server's memory, since the dataset is already in a "VA ready" state.
- Allows you to avoid any additional effort related to developing a separate ETL process to create the "VA ready" table by leveraging your existing ETL processes.
- Ensures the same level of data quality that your other business processes enjoy.
- Ensures *one version of the truth* is achieved by having a single implementation of ETL processes and business logic.

Additionally, these preexisting datasets exist to enable preexisting business processes. These preexisting processes are generally well understood and this creates an opportunity to derive a memory utilization model. With an accurate memory utilization model you can find the right mix of datasets and

users in order to extract the maximum business value from your SAS LASR Analytics Server's memory allocation.

CONCLUSION

This paper explored the areas of consideration necessary to determine an effective strategy for data selection and memory management.

We demonstrated that by leveraging your preexisting datasets you can ensure that you have data worth analyzing in Visual Analytics. This is achieved without the creation of any additional ETL processes while simultaneously ensuring high data quality and *one version of the truth*.

Additionally, we discussed how examining your preexisting dataset (and the business processes they support) can lead to valuable insights that can help you develop an effective SAS LASR Analytics Server memory management strategy.

REFERENCES

SAS Institute Inc. 2015. "SAS® LASR Analytic Server 2.4 Reference Guide." Accessed March 24, 2015. http://support.sas.com/documentation/cdl/en/inmsref/67597/PDF/default/inmsref.pdf

SAS Institute Inc. 2015. "SAS® Visual Analytics 7.1 Administration Guide." Accessed March 24, 2015. http://support.sas.com/documentation/solutions/va/index.html

SAS Institute Inc. 2015. "SAS® Visual Analytics Sizing Guidelines." Accessed March 24, 2015. http://support.sas.com/documentation/onlinedoc/va/

SAS Institute Inc. 2015. "SAS® Visual Analytics 7.1 User's Guide." Accessed March 24, 2015. http://support.sas.com/documentation/solutions/va/index.html

ACKNOWLEDGMENTS

Thanks to Christine Gamble for her help in editing this paper. Her dedication to excellence greatly added to the quality of this paper.

RECOMMENDED READING

- SAS® LASR Analytic Server 2.4 Reference Guide.
- SAS® Visual Analytics 7.1 Administration Guide.
- SAS® Visual Analytics 7.1 User's Guide.
- SAS® Visual Analytics Sizing Guidelines.