Paper 1733

A Need For Speed: Loading Data via the Cloud

Hadley Christoffels, Boemska Technology Solutions Ltd.

ABSTRACT

The value of the effective use of data is universally accepted, and analytical analysis methods such as machine learning make it possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results. However, before any such value can be realized, the data must be collected, moved, cleansed, transformed, and stored as efficiently and quickly as possible. SAS® Viya® not only addresses complex analytical challenges but can also be used to speed up data management processes. In this paper, we look at cloud infrastructure, code enhancements, and storage technologies that help you achieve this goal.

INTRODUCTION

“Effective use of data follows a kind of Maslow's hierarchy of needs”, says Jay Kreps. For those unfamiliar with the latter hierarchy, it is a concept introduced by psychologist Abraham Maslow in which he suggests that people are motivated by having their basic needs fulfilled before progressing to more complex needs. Kreps puts forward that similarly, the basic needs of effective data usage are to capture the data in a uniform way and have reasonable infrastructure on which to process it. My interpretation of the latter, which forms the basis of this paper, is that in order for powerful analytical methods like Artificial Intelligence and Machine Learning to produce value it needs to have quality and timely data made available to it.

There is currently a lot of hype, and deservedly so, around these methods but not enough is made of the advancements achieved in data management. Yet the adage, “80% of time is spent preparing data”, still rings commonly true today. So to do my bit for the cause (power to do the data !), this paper proposes ways to make the most of the agility and flexibility of cloud platforms, efficiencies gained through code enhancements and benefits gained through the use of cloud-native data storage solutions that support modern data and applications to speed up the loading of data.

CLOUD PLATFORMS

Cloud adoption has become a strategic imperative for enterprises. Lower upfront costs, greater agility and more rapid and cost-effective scaling are some of the benefits driving an increase in the migration of SAS 9 estates to the cloud and the rising embrace of cloud-based SAS Viya deployments. However, acquiring these platforms is just the beginning.

It is therefore important to have an understanding of the SAS applications being deployed in order to optimize performance and mitigate the risk of costly overprovisioning and unexpected expense being incurred. In addition, the understanding gained will enable application design that make the most of benefits associated with the cloud.

To support the goal of shortening the time it takes to prepare and load data, we look at both the elasticity and scalability of cloud infrastructure. The basic premise for both being to quickly and easily make additional processing power available, and only for as long as it is required. Either when the workload is unexpectedly high or specifically for processes with known high system resource demands.

ELASTICITY IN CLOUD COMPUTING

Simply put, infrastructure elasticity relates to the hypervisor being able to dynamically provision or remove virtual machines (VM) or containers and is one of the key factors that underpin the agility of cloud computing.

Cloud bursting is a technique that enable applications in a private cloud or on-premise data center to provision additional resources (or burst) into a public cloud when the demand for computing capacity spikes. Using this technique, it is possible to reliably and automatically augment capacity in an on-demand fashion to cater for unexpected workload and back off once things have stabilized. With this power at its disposal, businesses no longer need to provision infrastructure beforehand for the unexpected. A potential use case being SAS environments where heavy interactive usage could have an impact on batch processes, so when interactive usage spikes the workload is redirected to a burst resource.

As an alternative strategy to cloud-bursting, provisioning infrastructure for specific processes with expected high resource utilization is another way to go. For example, spinning up a virtual machine specifically for a heavily computational and data-intensive process or for entire, lengthy month-end batch schedules.

Again, having a good understanding of the underlying application is key before making decisions on the appropriate architecture design. It is also important to note that while all this sounds ideal in theory, it is not always possible as considerations around data availability, a lack of feature parity as well as underlying infrastructure and binary incompatibility could make achieving this unrealistic. The larger question being whether migrating all processing to the cloud is a better option.

SCALABILITY IN CLOUD COMPUTING

Scalability, while relying on elasticity, relates to increasing or decreasing capacity within existing infrastructure. There are typically two types of scaling:

- The first is to *scale vertically*, so to either move the workload to a bigger or smaller VM
- The second is to *scale horizontally*, by provisioning additional VM's and distributing the application load between them.

Horizontal scaling is additionally dependent on the application being able to distribute the load across the newly provisioned VM's. SAS Viya is such a platform and provides a robust, scalable, cloud ready distributed data management platform, i.e. it uses all cores on all compute nodes defined to the SAS Viya platform. To achieve this, it is necessary to ensure that all input and output datasets reside in the SAS Viya in-memory engine, Cloud Analytic Services (CAS).

In addition to the distributed nature of SAS Viya, it's multi-threaded capability is one of the most significant aspects of the platform. Particularly for improving the performance of SAS processes where the DATA step is heavily used (more on this later) and constraints exist due to older single -threaded environments.

To conclude, the inherent elasticity and scalability of cloud platforms coupled with SAS Viya that has been engineered to take advantage of these benefits is a major boon for SAS performance improvements. This is true not only for future application development but indeed also for existing applications.

PROGRAMMING INEFFICIENCIES

While migrating processing to the cloud, it is imperative to understand and address existing performance challenges, as attempting to migrate processes with known performance issues could be costly and only make matters worse. In addition to improving inefficient programs, we will also look at code amendments that make use of the new techniques and benefits now available in SAS Viya, such as the multi-threaded DATA step.

FINDING INEFFICIENCIES

SAS estates can be complex, and it is not always easy to gain the insight required to make meaningful efficiency improvements. We have found that the granular insight provided by a product called Enterprise Session Monitor (ESM) greatly simplifies this task and makes metric driven performance improvements for SAS environments possible with ease. It provides a unified platform with real-time and historical diagnostics that help identify metric driven actions for improving process efficiency and platform

performance. In addition, it provides a reporting mechanism that provide detailed usage analysis, a charge back facility, and performance and capacity trend analysis.

There are numerous features within the product, but we will focus only on those that related to our goal performance improvement. A subset of these are:

- *Real-time monitoring*

Figure 1 shows the detailed user session-level (top view) and node-level (bottom view) metrics collected and displayed at 2 second intervals. These are:

  - User session-level
    - Session CPU usage
    - Frequently updated SASWORK and UTIL measurements (*for that session*
    - Real-time Batch Warning and Error detection
    - Memory usage (physical and virtual)
    - User-defined code markers (for performance optimization)
    - Network device throughput
  - Node-level
    - Overall CPU usage
    - Lost CPU cycles (iowait)
    - Memory usage patterns (including cache and swap)
    - Disk throughput
    - Network device throughput

Using this facility, it is possible for end users to be aware of all sessions (including those that were orphaned) running under their username and understand the impact each of these sessions have on overall performance. Thereby enabling users to meaningful contribute to performance at a process level.

Figure 1. ESM real-time diagnostics interface

- *Batch Insight and Comparison*

The application also offers an option to see all the jobs that have run in a certain timeframe. This allows the user to get an interactive, graphical representation of the batch run and the option to compare this to a run in a different time (a day ago, a week ago), as shown in Figure 2. This is essential when trying to optimize and reduce the batch window

Figure 2. ESM Batch Comparison and Insight

- *Historical session-level view*

Because all the information that is offered in the live view is stored within ESM, the historical view can offer the same information as shown in Figure 3. This allows the user to look at the server at a specific point in the past to see for instance what exactly was going on at 02:00 in the morning. This is extremely useful for quickly finding root cause analysis of an issue that happened during the night, or to analyze the performance of individual jobs.

Figure 3. ESM historical diagnostics interface

- *Utilization reports*

Since all historical data is kept for a certain period, it is possible to see system resource utilization over a specified period and group these by, for example, teams or departments. This can offer various useful insights, but also offers the option to cross charge for actual usage and through this drive a culture change where the responsibility for overall performance is shared by the end-users. Figure 4 shows an example report

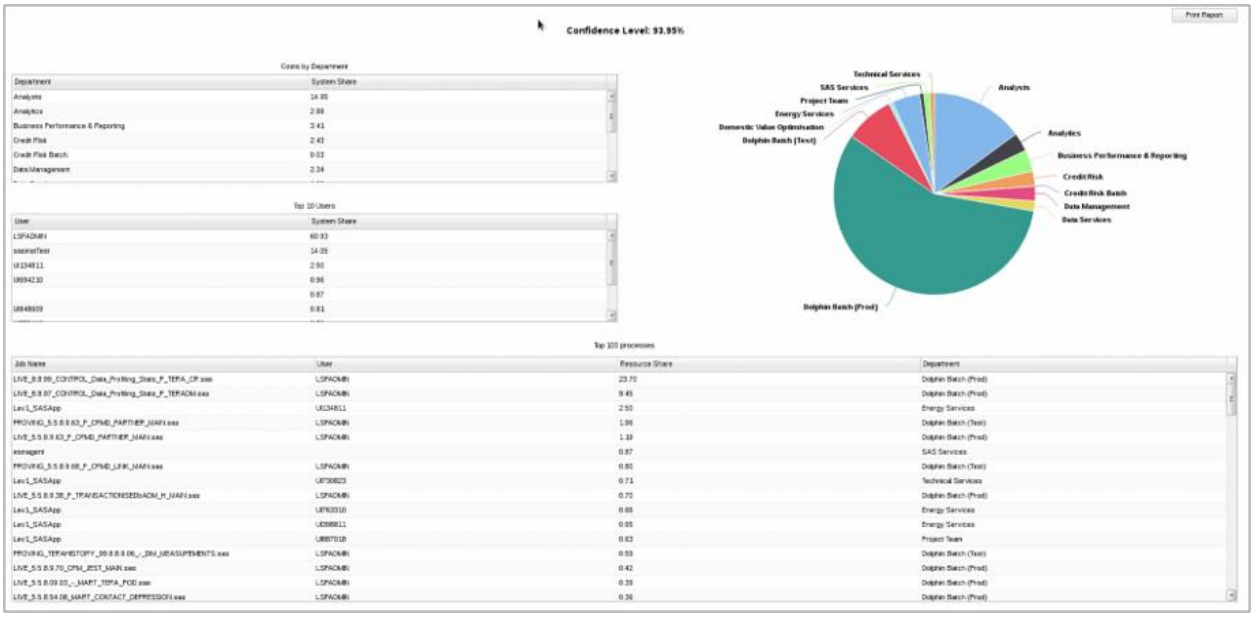

Figure 4. ESM resource utilization interface and report

OPTIMIZING INEFFICIENCIES

There are numerous ways to improve program performance in SAS, even more so with the right level of insight. For the purposes of this paper we will focus on the benefits to be gained from SAS Viya on a Cloud infrastructure. Specifically, around techniques for loading and manipulating data, but first an introduction to the platform.

CLOUD ANALYTIC SERVICES (CAS)

SAS Viya uses SAS Cloud Analytic Services (CAS) to perform tasks, it is a server that provides the run-time environment for data management and analytics with SAS. The smallest unit of work for the CAS server is a CAS action. CAS actions can load data, transform data, compute statistics, perform analytics, and create output.

CAS ACTION SETS

CAS actions are organized into action sets and each action set defines an application programming interface. Currently, the available action sets are:

- *Statistics* - provide summary statistics, clustering, regression, sampling, principal, component analysis, and more
- *Analytics* - provide text and numeric analytics
- *System* - run SAS code via DATA step or DS2, manage CAS libraries (caslibs) and tables, manage CAS servers and sessions, and more
- *Data Mining and Machine Learning* - support scoring, factorization, neural networks, support vector machines, graph/network analysis, text mining, and more

As we are interested in speeding up the load and manipulation of data, we will focus only on the *System* action sets. As a further subset, only a select list of action sets that will have the biggest impact within this area. These are:

- *Streaming Data Action Set*

The streaming data action set is used to capture data from event streams using SAS Event Stream Processing (ESP). We can use this along with the low latency, high-volume SAS ESP software to load data at high-speed. And the distributed, in-memory grid processing scales linearly as your data grows, optimizing hardware investments

- *DATA Step Action Set*

In SAS Viya, everything that can run multi-threaded does so by default, and the SAS DATA step has been enabled to do so. In fact, it takes very little modification of existing code to take advantage of the significant performance benefits this makes available.

- *DS2 Action Set*

DS2 is a SAS programming language that is appropriate for advanced data manipulation. A DS2 program can perform manipulations on multiple data rows concurrently, thus reducing the time required to process big data sets. While DS2 shares core features with the SAS DATA step, it exceeds it *by adding variable scoping, user-defined methods, ANSI SQL data types, and user-defined packages.*

- *TRANSPOSE Action Set*

Due to the nature of model input data structures, a large amount of time can be spent on transposing data. SAS Viya allows you transpose in a fraction of the time that it used to take as it now runs entirely multi-threaded and in-memory. Potentially saving significant amounts of processing time.

In addition to the above, *the concept of sorting goes away with in-memory processing. Meaning that, for the most part, nearly all sort routines can be removed from existing programs. Which of itself, can cut down processing time significantly.*

CLOUD-NATIVE DATA STORAGE

With the increased adoption of Cloud platforms, just about any software is now quickly and easily available on these platforms. Not only are software vendors making existing software compatible, but a lot of software is being written specifically for the platform. This trend includes storage technologies, and some relational databases have had a major revamp to ensure compatibility.

The use of NoSQL and Hadoop data stores are common place in the Cloud and are well integrated and documented so the focus here is on databases. In addition, databases are looked at from an analytical workload perspective and not operational/transactional. There is an insightful comparison of these technologies written by William McKnight, *Sector Roadmap: Cloud Analytic Databases 2017* for a more detailed view.

Performance challenges related to storage stem from a number of factors; for example, the inability to scale due to hard limitations of on-premise system resources such as memory and physical resource such as rack space, managing busy periods like month-end or heavy, unexpected analytical user workload, and aging storage technologies with outdated I/O throughput. Many of these challenges can of course be addressed by migrating these workloads to the Cloud. It is at this point, with an understanding of the SAS applications being migrated, that consideration should be given the most appropriate storage technology that 1) is most suitable to your workload and 2) maximizes the benefits of the Cloud.

Not all databases in the cloud are created equal, so the following are some of the factors that need to be evaluated:

- *In-database performance optimization*

A lot of work has gone into in-database processing and performance over the years and moving to the cloud should enhance and not impact this capability. It is important therefor to consider the extent of this capability when selecting a solution, particularly for newer entrants with databases built for the cloud

- *Database agility*

As elasticity and scalability forms the backbone of Cloud platform attraction, it is essential that the database mimics these benefits. It should therefore be able to contract or expand as required along with the infrastructure on which the SAS application is running. It should also not be necessary to create outage for any maintenance or upgrades that will minimize and disrupt the ability to execute processes.

There are a number of these solutions available and Snowflake is a great example of a cloud-native database (or data warehouse) suitable to analytical workloads done well. It is typically much faster than conventional data warehouses, separates compute from storage, capable of handling demanding analytical workloads, comes with adaptive optimization, supports impressive concurrency, scales easily and eliminate the need for manual tuning.

CONCLUSION

Cutting down processing time of data preparation and load processes will have a direct impact on how quickly downstream consumption applications can begin and therefore complete. The Cloud platform is increasingly being embraced to solve this problem. In order for the agility and power of the underlying infrastructure to take meaningful effect, applications must be able to take advantage of these benefits. SAS Viya has been engineered to do exactly that. Furthermore, the SAS Viya platform addresses SAS specific performance challenges by enabling performance improvements techniques that are commonly used and known to be resource intensive, such as the transposing of data and the transformation of data using the beloved DATA step. This takes care of the infrastructure and processing of data but not how and where data is stored. For that, we looked at cloud-native databases and the benefits it brings, particularly those that align to and make the most of the inherent benefits of the Cloud platform.

It is at the intersection of these technologies that maximum benefit will be realized. It is essential that the latter is architected with a deep understanding of the application being migrated to or developed for the Cloud.

RECOMMENDED READING

- *http://documentation.sas.com/?cdcId=pgmcdc&cdcVersion=8.11&docsetId=pgmdiff&docsetTarget=titlepage.htm&locale=en*

CONTACT INFORMATION

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

Hadley Christoffels
Boemska Technology Solutions Ltd.
+44789 424 8787
hadley@boemkats.com
www.boem.sk