Paper 114-2012

How Does SAS® In-Database Analytics Impact Data Management?

Adrian Jones, SAS Institute Inc., Marlow, United Kingdom

ABSTRACT

Over the last few years, business analysts have been able to realize the benefits of SAS In-Database Analytics through both performance improvements and value generation. With SAS In-Database, users interact with the data changes along with the process to develop and deploy modeling outputs. This not only has an impact on the business processes to support analytical activity, but also on the underpinning storage and data management processes.

This paper takes a look back at the uptake of SAS In-Database Analytics and the impact it has had on the data architecture and data management processes required to support such activity.

INTRODUCTION

Figure 1 SAS In-Database Capabilities

SAS In-Database was introduced in 2007 and has now reached a level of maturity with good experiences from real world implementations. SAS In-Database appeals to many customers who have made investments in data warehousing technologies or are pursuing enterprise data warehousing strategies. Early development focused on the ability to move scoring and analytical processing to the data warehouse platform. This removed the need to either extract the data to a SAS environment or convert the analytical code to something that could be executed on the data platform.

As organizations began to adopt such techniques for analytical processing, it became more apparent that work was needed on the data processes to support analytics as well as the processes to support the deployment of the analytical models into a production environment. The value to the organization is in optimizing the end-to-end process in the analytics lifecycle and this needs to include the entire data lineage. This brings together data management with analytics management, which is a link into decision management. It is the combination of these disciplines that has become known as information management. SAS data integration is a key toolset that supports the breadth of needs of information management and is a key component to support and management in-database activities.

IN-DATABASE MISCONCEPTIONS

Before understanding the impact on data management, we need to clarify certain misconceptions that are often encountered in the field.

SPEED OF PROCESSING OR LENGTH OF PROCESS?

From initial implementations of SAS In-Database, we saw that most of the focus was on the particular processing speed when running a scoring function in the database. While this is of importance, it failed to overlook the savings that can be made in the end to process and, in particular, other efficiencies that are driven through appropriate optimization of associated data jobs.

We have seen sites where the in-database processing allowed the scoring process runtimes to be reduced from 30 minutes to 4 minutes. While this is impressive, when we start to look at the remainder of the process we see that the overall time taken including data movement and processing as part of this, then the overall process time has reduced from 3 hours to 12 minutes. This has been achieved by the removal of many interim data processing steps and transportation between the data platforms. With SAS In-Database, the data remains in the data warehouse and efficiency can be driven into the data shaping through more managed data processes.

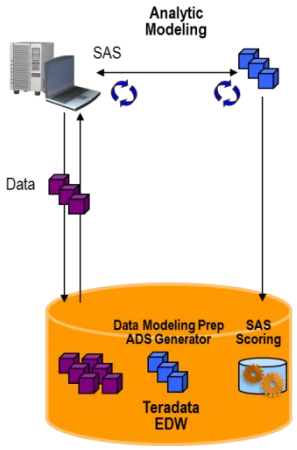

Figure 2 In-Database Processing

STORAGE VS PROCESSING

Many organizations have a focus in the current climate to reduce data layers in the organization. This is understandable from a cost and management aspect. However, it fails to acknowledge that certain processing types require very specific platforms.

When looking at in-database processing, it is important to separate the storage requirements from the analytical processing requirements. Certain analytical processes run more optimally on a SAS engine rather than on a database node. In these instances, attention needs to be placed on the data flow. It is important to use the engines appropriately and optimize across the entire process. For example, use the database for preparing the data and return only an appropriate subset of the data to the SAS server for analytical processing. If this is an ad hoc job, then this could be done at run time, however, if this is a job that is run many times per day, then it makes sense to physicalize the aggregated data on the SAS server to drive more efficiency.

ANALYSTS – THE UNDERGROUND DATA MANAGERS

As part of any analyst's job, there is a need to work with data and shape the data appropriately to allow for analytical activity.

We see many sites where the analysts are doing this, but also much more in the way of data management activity. On average, we find analysts spending between 60 and 80 percent of the time acquiring data and managing data stored on servers and desktops. This practice has often developed over time and has roots in previous issues with IT delivering the required data or inability to access data that they believed is required for their jobs. Needless to say, this is not an optimal use of the analysts' skills, and often leads to processes that are inefficient and not governed. Unfortunately, we often see the data providers in such organizations turning a blind eye to such activities.

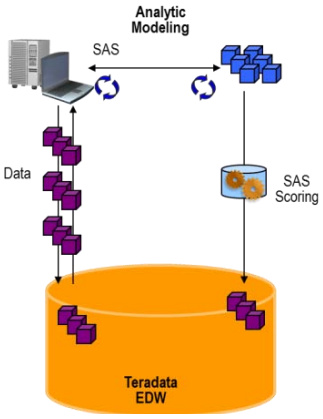

Figure 3 Traditional Approach

IN-DATABASE DRIVERS

Organizations adopt an in-database approach because they believe that it delivers a competitive advantage either through the ability to do a revenue-generating activity more quickly or by driving efficiency and cost savings in the IT delivery of such capability.

- Business Drivers focus on the ability to react quickly to market changes and competitive threats. To do this, there is a need to improve execution speed by driving more operational processes with analytics and thus reducing time to action. With such agility and speed, decision quality becomes important particularly from transparency and compliancy perspectives.
- IT Drivers focus on the ability to support the business requirements. Doing more and faster in a shorter timeframe are key to meet these expectations. Delivery and implementation risks need to be managed to drive quality throughout the processes, and ultimately all of these are driven by a desire for lower total cost of ownership (TCO) of data and analytical systems.

At many sites, this is where we can see the challenges in the organization, and we start to uncover the divide between Business and IT. Information management and the underpinning usage of tools like SAS Data Integration and SAS Model Manager help the teams supporting such processes to overcome the divisions by using a common platform that can support all of the different roles required to deliver such capabilities.

IN-DATABASE PRINCIPLES

The development effort to drive in-database processing by SAS and partners has been driven by three core principles:

1. Reduce Data Movement. This should be achieved by pushing data-intensive work to the database to make full use of database resources and technologies.
2. Preserve the SAS user experience. Users should be able to continue using their SAS Language skills and SAS procedures experience when working with in-database solutions.
3. Maintain SAS standards. SAS is known for numerical accuracy, precision, and statistical integrity and these should be maintained through software quality.

The above are commonly referred to as the in-database principles and it is important to take these into account when looking to best fit SAS In-Database into the organization. The important one here for data management teams is the first principle relating to data movement. From a data management perspective, this has been demonstrated to drive traditional, user managed data jobs into more centrally managed and optimized data jobs. Thus, by driving maturity of approach to the business analysts, this has the impact of creating work and processes for the data management teams.

IN-DATABASE ARCHITECTURE

We have already mentioned that one of the aims of SAS In-Database is to reduce data movement. As SAS Data Integration users, this might sound concerning, however, by understanding the bigger picture of SAS In-Database, you start to see that SAS Data Integration becomes more important to the organization in maturing the data processes and running analytical processes in a production fashion.

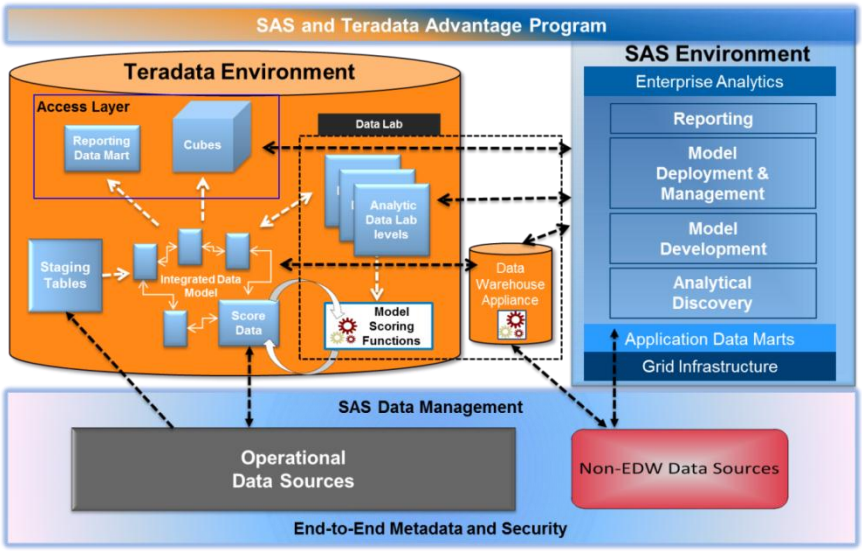

Figure 4. SAS and Teradata Joint Reference Architecture

This diagram represents the joint architecture in a SAS In-Database deployment. We see a representation of the data platform (in this case Teradata), as well as the SAS environment. The dotted lines with arrows represent the flows of the data into and around the environments.

DATA AND ANALYICAL PROCESSING

In this example there are a number of different types of processing required, and it comes down to which engines can support specific processing. Then it makes sense to design the data processes and stores in a manner which optimally supports this. It has been seen that this varies from site to site and often depends on the culture of the organization.

The Teradata environment is an SQL-based data platform that is optimal for this type of processing. Thus, it makes sense for this to be the primary data store running production-type analytical processes that can be published or pushed down.

The SAS environment runs the SAS engines and can be used for more analytical processing and ad hoc development type work. The SAS Data Integration Server environment exists here, although it would be expected that most of the data integration processing would be pushed to the database (ELT approach) in this scenario. The SAS Metadata Server environment would also exist here.

DATA ENVIRONMENTS

Data environments refer to the data platform that supports most of the traditional data activities. In this example, we are looking at a Teradata platform that supports a number of different activities that can be ring fenced by workload. The following components are then deployed on the platform:

- The Staging Tables are where the data is first landed. The data has not been transformed at this point and so they reflect the source systems. This is important to support ELT (Extract, Load, and Transform) type processing, which is supported by SAS Data Integration and is intended to land the data on a single platform and use the power of the SQL engines to perform the transformations, thus leading to more efficient loading of the warehouse.
- The Enterprise Data Warehouse (EDW) is represented by the Integrated Data Model. This is where the data is shaped into an appropriate data model, which could be a standard industry data model, or a bespoke model for the organization. This model is the source for all downstream consumption and supports other external systems as well as analytics and reporting activities. The normalized data model feeds other types of data structures downstream including cubes and de-normalized type structures.
- The Access Layer (or Presentation Layer) is there to support reporting and other applications. It is common to deliver cubes or aggregated structures for such a layer. The key to design for this is to understand the usage work patterns and how this impacts the workload on the warehouse. These can be delivered as physical tables and views or can be offloaded to other platforms as requirements specify.
- Analytical Data Labs are effectively a ring fenced area of processing and storage on Teradata that allow the analytical user to prepare and shape data ready for analytics. While on the same physical data platform, they are separated from the general EDW workload on the machine. This means that users can carry out sandbox-type activities without putting production workloads at risk. This is a core piece in supporting in-database development activities.

While these can all be deployed on a common data platform, consideration needs to be given to the data structures that are presented to the applications or users. The shape depends on the purpose of the data (for example, a third normal form type data model for the EDW, cubes and views for reporting and de-normalized data for analytic usage). Other transformations such as aggregation, needs to be considered to ensure that data processing is performed in the most appropriate place. SAS is working with partners to push appropriate processing to the data platform in the most appropriate manner, but it is important to understand the user requirements when designing such processes.

ANALYTICAL ENVIRONMENTS

Analytical Environments refer to the environments where SAS processing takes place specifically on SAS engines. In this example, there are two Analytical Environments shown:

- The In-Memory Analytics Appliance is represented as a Data Warehouse Appliance in this example. This is a separate appliance that is used specifically for In-Memory Analytics and is intended for solving complex problems with large amounts of data. This data is only for analytics and should be seen as separate to the EDW.
- The SAS Environment is to support processing that cannot be done inside the database or is more efficient on SAS Engines. In such a deployment, there is structured data delivered through managed processes as well as

some user-generated data. This should be data that is required for usage, and it should be remembered that in such an architecture, the EDW maintains the full data with history, while the SAS environment holds data that is fit for purpose.

As a minimum, you would expect to see a SAS Environment alongside a Data Warehouse where organizations are carrying out in-database type work. Workloads and processes then need to be optimized across the platforms depending on the process or activity.

METADATA

Now that the data and analytics environments are closer together, it becomes possible to build a richer metadata lineage that combines the technical metadata with the business metadata. This metadata covers the end-to-end data processes from data acquisition from source right through to the final consumption via an analytical model or business report.

Working on a common metadata server, SAS Data Integration plays a key part in the creation and maintenance of such metadata to support business, technology, and governance processes.

DATA MANAGEMENT

With a SAS In-Database approach, we see that data management is really focusing on the capability to manage the data across the organization. In this context it is more than data integration and now includes all aspects including data governance, data quality to master data management (MDM).

DATA INTEGRATION

SAS In-Database has bought about advancements in the SAS Data Integration product as well as how this is used in the enterprise.

The ELT approach has been followed in the field and is promoted by the data platform vendors. SAS Data Integration has a number of features to specifically support this approach including specific table loaders, code pushdown, visual indicator processing, temporary tables in the database, and others. The main advantage of an ELT approach is to reduce data movement across the network between the data platform and the ETL server.

From a data flow perspective, the SAS Data Integration user has the ability to control and optimize the data flows between the platforms ensuring efficiency across the platforms. We often see a legacy of these processes created by Business or IT running as code-based jobs, which are difficult to manage and maintain.

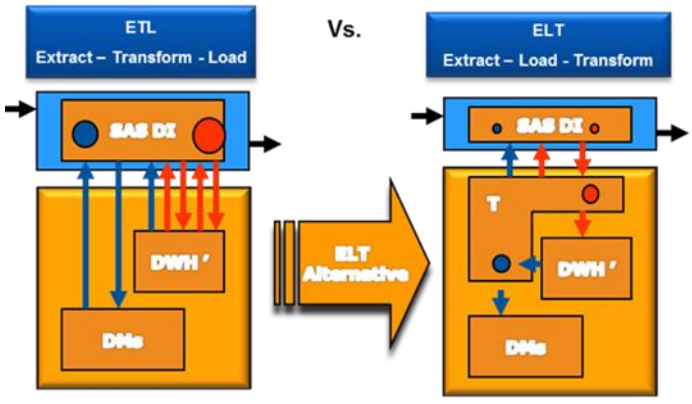

Figure 5. ETL and ELT Processing

DATA GOVERNANCE

In many organizations with a legacy of analytic driven processes, there are challenges with data governance and this is often one of the drivers behind moving toward an in-database approach. The data governance issues arise because users tend to acquire data through their own created processes and often move data to other environments that are convenient for them to work with. There are often many such “workarounds” in place with no formal approach or structure. Sometimes this is done for performance reasons, or just down to legacy and familiarity. Whatever the reasons, this leads to a situation where it is difficult to determine data lineage and who has done what to the data.

At such sites, SAS In-Database starts to bring some consistency across the data and analytic platforms and the SAS Data Integration user can start to manage the metadata through a common platform. This allows for central policies to be enforced and allows for monitoring to ensure that they are complied with.

DATA QUALITY

In-database projects tend to allow organizations to start looking at their data from a consistency and a reliability perspective. We often find that this is the first time that the business and technology teams truly engage to standardize the data across the platforms. These projects are often an opportunity to start building data quality activities into data integration processes that are being driven by business analytics requirements. This ensures a focus on and agreement of what to address to support the business needs.

MASTER DATA MANAGEMENT

The drive toward an in-database enterprise adds support to the desire for master data management. One objective of SAS In-Database is to cut down the number of data layers and isolated data silos by bringing these together on a common data platform where the analytical processes can be executed. The performance benefits of SAS In-Database gets the buy-in to the creation of an MDM vision, which has been proven to improve customer interactions through more appropriate data usage. An MDM approach drives data integrity and reduce management costs as data is centralized.

ANALYTICS MANAGEMENT

Analytics management is the ability to manage the analytic models and to manage the result of the model as an information asset. This is a new area for many working in data management, but this is an important area as the organization looks to deploy analytical models into production processes on larger platforms. This is an area where the IT organization can bring best practice to the production management of such activities.

ANALYTIC APPROACH

The Analytic Model Lifecycle provides a good insight into the common practices that analysts use to solve business problems. While specific analytic skills are needed to build and validate the analytical model, it can be seen that many of the associated processes relate to data acquisition or deploying the model inside the database. In practice, we find these processes are often picked up and well managed by SAS Data Integration users who can support the analytical process. This adds benefits through the efficiency gains of repeatable processes, as well as allowing the analysts to focus on model development rather than data management.

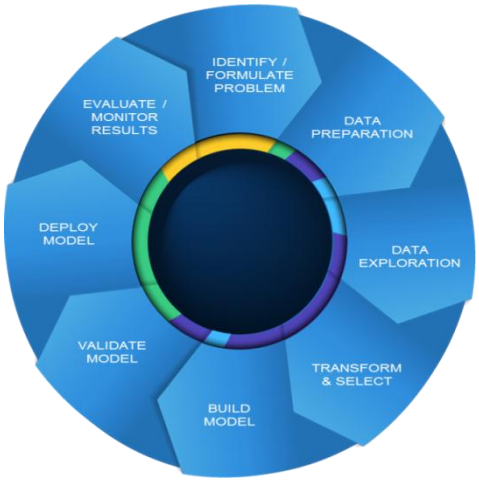

Figure 6. Analytic Model Lifecycle

MODEL MANAGEMENT AND MONITORING

SAS Model Manager is workflow tool that analysts use to manage the Model Development Lifecycle. This provides the appropriate levels of authority and sign off through-out the process. It is often seen in traditional analytics environments where there is little management or control around this. Because of this, there is little official monitoring of the models once they are in to production, and so little is understood of their true value or whether they should be retired or not. As part of an in-database approach, the data management team can begin to create and manage processes that support the model monitoring aspects and thus provide automated and consistent monitoring information.

MODEL DEPLOYMENT AND INTEGRATION

In traditional analytic environments, the analytical user would often run their own scoring processes on their own environments or hand the code over to IT to convert it to a suitable language for execution on the database. Both of these approaches carry business risk either from an environment failing or flawed logic conversion.

With SAS In-Database and SAS Model Manager, the scoring function is published to the database in a manner that can be processed in the database. This removes the conversion risks. However, there needs to be control and standards in place when deploying to a production data environment. This is where it makes sense for the data management team to be involved with the publishing process using an appropriate sign off process in SAS Model Manager. This means that different teams can carry out their individual roles to in a common user interface that shows the state of readiness of the model.

When running the scoring process inside the database, the SAS Data Integration user has a role to play in building and running the processes to generate the scores as needed. This is something that is normally outside of the scope of a typical analytical user.

THE SAS MODEL FACTORY

As part of the outcome of a number of SAS In-Database implementations, these activities have been brought together into a set of production like processes that have become known as the SAS Model Factory.

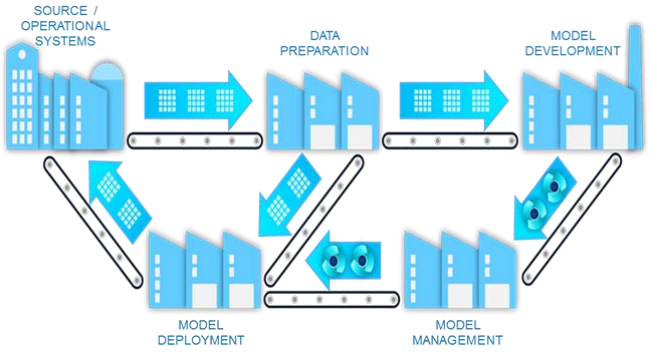

Figure 7. The SAS Model Factory

The SAS Model Factory moves organizations forwards to more managed, mature analytical modeling processes. It leads for consistency and efficiency in the development and deployment of models. SAS Data Integration plays a key part in this with regard to acquiring and developing the analytical data mart. SAS Data Integration is used to create and schedule the production scoring processes after model deployment.

The SAS Model Factory has been proven to bring the controls and processes to ensure that modeling activity is suitable managed for a production data environment. The SAS Model Factory starts to reflect the mission critical nature of such analytical systems.

INFORMATION MANAGEMENT - THE NEXT STEP

The move toward SAS In-Database adoption and the alignment of the data and analytical platforms can be seen to emphasis that organizations need to move from traditional data management approaches toward a more strategic and comprehensive information management approach. Information management brings together the disciplines of data management and analytics management with a focus toward driving decision management. Decision management in this context is the ability to operationalize the information and analytical results directly in the business applications or business process. This is where the real value is.

An information management approach addresses the strategy and governance needs and delivers the appropriate capabilities required to support this. SAS uses an integrated analytics platform approach that can support and drive the processing to the most appropriate environment. Data management teams are equipped to support and deliver such processes.

When delivering SAS In-Database, a series of processes have evolved to support the data and analytical activities. These have matured to reflect the environments on which they are now being deployed. It makes sense to bring the production processes and capabilities together on a common platform.The most effective way to support this is through an Analytics Competency Center that pulls skill sets from both the analytical and data management communities into a common team that can focus on delivery and management of information assets.

A unified information and analytics platform built on SAS In-Database and other SAS High-Performance technologies enables the organization to exploit the data through analytics. Information management provides the capability to take advantage of this in a sustainable manner.

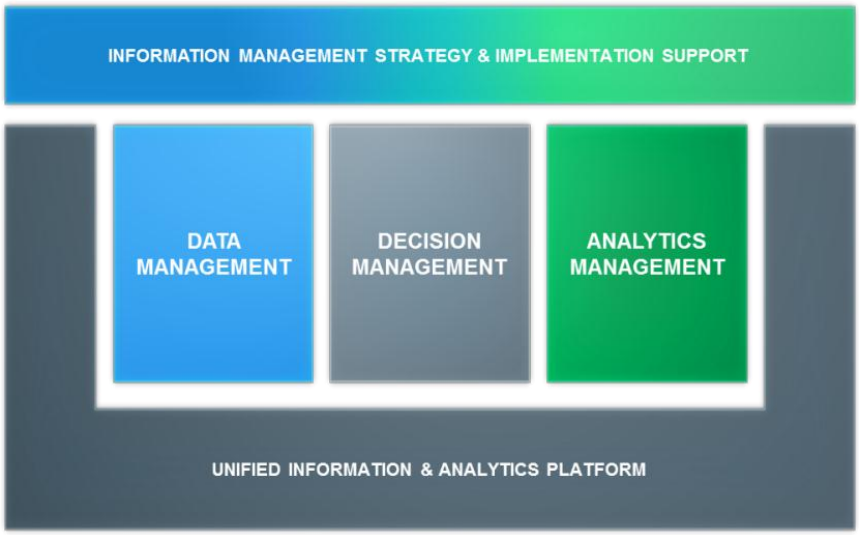

Figure 8. SAS Information Management

CONCLUSION

Technology is no longer a blocker to bringing data and analytics together and this means that the approach to managing data and information is evolving. The technologies to support data and analytics have gone through evolutionary changes over the last several years.

From a capability perspective, organizations are moving from stand-alone data integration related disciplines to a mindset where data integration, data quality, MDM, and data governance are leveraged and designed together. They have augmented the traditional ETL approach with an ELT approach that leverages the processing power of the data more effectively and minimizes data movement.

Also organizations are beginning to think more strategically and are moving from a project-based approach to a holistic, enterprise view, where information can be leveraged as a strategic asset. This is the move from a reactive approach to addressing data needs to a managed, and ultimately, a proactive approach to managing information. A key part of this move is to expand the notion of data management to include governance and to expand the scope of the strategy to include analytics and decision making.

This can be seen as an information continuum where we go from data to information to decision and insight (where the result of the decision is more information that helps optimize the information continuum). Then the organization can focus on truly creating value through data and information.

CONTACT INFORMATION

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

Adrian Jones
SAS Software
Wittington House, Henley Road
Marlow, SL7 3HA, United Kingdom
adrian.jones@suk.sas.com