Paper 136-2011

Breaking Down the Silos - Data Management Across the Enterprise

Adrian Jones, SAS Institute Inc., Cary, NC

ABSTRACT

Most organizations have a legacy of data distributed around the organization, often held in disparate silos in what appears to be an unco-ordinated way. Traditionally, the business user would acquire the data through an ad-hoc process and often replicate more data to solve the business problem. This leads to many problems ranging from inability to access the right data through to lengthy delays obtaining the needed data. Much of the traditional approach focused on bringing the data to the user, and this is where the problems began.

There are many technologies and techniques available to the SAS® analyst to exploit the data around the organization. This paper looks at joining the dots between the different storage technologies within the organization and outlines the key areas to be considered when trying to address the problems encountered with data management in the organization.

INTRODUCTION

This paper looks at the traditional approaches to data storage and we look at how the data architecture supports this in a typical manner. Looking at the present, times are changing with some of the new techniques and technologies available to the analytic organization and we consider the impact that these will have on their options. What does it mean for data integration when we begin to reduce data movement to disparate analytical silos? This paper should provide some thoughts and strategic considerations when making changes to the data storage and data flows across the enterprise.

THE STORAGE LANDSCAPE

For many years, organizations have focused on building various layers of storage to support business and technology needs. In a simplified world following the traditional well recognized approaches, the key storage layers are as follows:

- The Operational Systems and Operational Data Store
- The Enterprise Data Warehouse
- Data Marts and the Analytical Data Store
- Desktop and other File Systems

Whilst these are common definitions for what is seen in practice, the reality is that there is no prescribed approach to defining the data architecture to support the business consumption of the data within the organization. We often see only some of these deployed, or we might see all of these deployed several times over within the organization. It is not uncommon to find some of these supporting the needs or covering the role of other layers. Ultimately, it is the combination of these layers that serves the business community and therefore these are major components of the data architecture and data strategy for the organization.

These layers are connected through some form of ETL process from ad-hoc user code through to managed Data Integration processes. It is this set of processes and transformations that allow us to build the lineage of the data and truly understand the data flow throughout the organization. Whilst it sounds easy to define such lineage, the practical reality is that many do not truly capture the required details due to incompatible

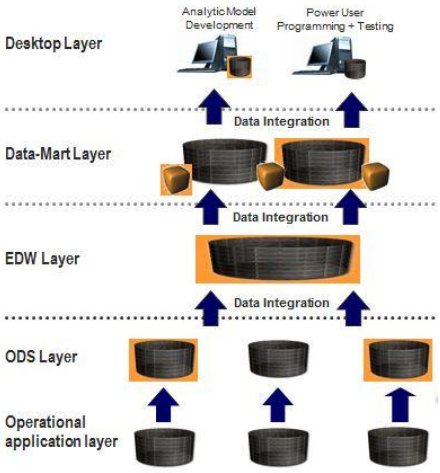

Figure 1. The Storage Landscape

technologies from previous departmental decisions. This has an impact on those maintaining the data as well as those wishing to trust and use the data.

To understand the different approaches seen in organizations, we need to first understand the characteristics of each of these types of storage. It is important to uncouple the technology from the storage layer – design decisions should be based on the intended function of the store, not the features of the technology.

OPERATIONAL SYSTEMS / OPERATIONAL DATA STORE

The Operational Layer exists to support and run operational applications (booking systems, sales systems, account systems, etc.). To do this they must be able to support millions of transactions, often with many small concurrent updates. There is no need for fast loading of large amounts of data, long running queries or supporting large numbers of users. The aim is to fully understand the workload and then to design and size the environment appropriately.

An associated concept is the Operational Data Store (ODS) which supports the operational systems. This is often made up of snapshots from operations with no historical data held. The ODS will often use the same RDBMS as the operational system and can take advantage of replication facilities to avoid load processes. The ODS will be used for backup purposes and will allow the data warehouse to load without putting strain on the operational RDBMS.

THE ENTERPISE DATA WAREHOUSE

The notion of the Enterprise Data Warehouse (EDW) is pursued to differing levels of success in many organizations globally. The concept is that this becomes the primary repository within the organization and supports the notion of a “single version of the truth”. The EDW collects data from operational RDBMS’ and systems over time, creating a large amount of historical data. By definition, there will be massive data growth over time. Load and query performance are key design criteria and focus leading to often application neutral designs. They will often include staging areas to further support the load performance. Some RDBMS vendors promote the notion of creating the ODS and EDW on the same platform to avoid the need for major data integration jobs between the layers.

DATA-MARTS / ANALYTICAL DATA STORE

Data marts are normally application or business specific and highly aggregated. They will contain a mixture of OLAP cubes and relational tables as needed. This will be the place where most queries are issued so there is a need to support fast and reliable query performance. Marts are often updated or rebuilt every time the data warehouse or EDW is updated, so loading performance is vital. Marts can exist on many different storage technologies within the organization.

The notion of the Analytical Data Store (ADS) is similar to data marts, however, rather than focusing on a specific business application, it should be a place to empower the user to explore the data in an analytic sense and to identify predictive patterns allowing the user to shape the data in a manner that is conducive with analytical activity. The ADS by definition should be a fluid set of data structures that evolves continuously to meet the analytical challenges of the organization.

DESKTOP / FILE SYSTEM

This is the least managed of the storage layers and is usually driven by the users themselves. Traditionally, this was the approach taken by analysts to support their work; however, this is often a legacy view that exists in organizations with users following such practices as the norm.

The desktop layer often consists of wide tables with many attributes for analysis and can be many gigabytes in size. They often hold historical information with no support for transactions. Modelers often use these for Rapid Application Development (RAD) activities when developing (not in production) and to avoid network latency issues. Being local, the user will often experience fast and reliable query performance with limited scalability.

DATA ARCHITECTURE – SUPPORTING THE STORAGE LANDSCAPE

The Data Architecture underpins the storage layers as defined previously and includes:

- The Storage Technology
- The Data Models and Structures
- Data Processes and Flows
- Data Strategy

There are many types of storage technology today that can support some or all of the layers as defined. These range from typical relational data base technologies, to columnar databases through to data appliances. All of these have their merits and have strengths for particular scenarios. It is because of this, that it is not uncommon to find a variety of different technologies deployed within the organization or featured in the technology strategy.

The Data Model is crucial to the adoption of the storage layer. It serves multiple functions and if not appropriate will impact the performance or uptake of the solution deployed. The data model needs to support the storage to ensure that is performant enough to meet the needs of the user or application. This should be a key link between the technology team and business users to ensure that the solution is fit for purpose. It is not uncommon to see data models used for a purpose that is different to what it was designed for with a resultant impact on performance. Whilst performance is important, the business analyst needs to be able to understand and manipulate the data to produce something of value. A model that focuses on either the operational systems or serving the needs of the entire business will need evolving into something that is more user friendly for the business analyst or they will not be productive. This is the point of divergence for many organizations, with the choice being to either “let the users do their own thing” or making an attempt to deliver data in suitable form for analytics through a more managed process. Needless to say both approaches have their shortfalls. With the first approach, the volumes of data is normally very large and it is difficult to manage, making this ideal for driving short term value, but making this difficult to sustain over time. With the second approach, the key is collaboration. This approach often struggles due to applying a rigid approach for structures that need to evolve to meet the continually changing needs of the business. All too often, it can be seen that good intentions often fail due to inflexible processes that are difficult to manage in traditional technology organizations.

The series of processes to move the data around the organization is effectively the glue that brings together the various data components into a common architecture. That said, this is an area of evolution within the organization where the actual end to end flow of the data from source system through to user consumption is rarely designed upfront in totality and this is because the end use is generally unknown at the time of creation of such data structures. Generally speaking, these data flows fall into three types:

- Loading the data warehouse
- Transforming the data for business use
- Data Preparation for analytics

Loading the data warehouse is primarily focused on speed of delivery of the data and therefore it makes sense to minimize data landing steps and this is why an ELT (Extract, Load, and Transform) approach can be faster than a more traditional ETL approach. These processes normally populate normalized data structures. At this point, the data will be modeled either in a neutral manner or being more a reflection of the source systems. The data will then normally be transformed into a more user friendly set of structures to support reporting or analytics. These will tend to be at function, application or departmental level and this is often maintained in star schema type structures. In many organizations these will be managed processes that will take into account the business use of the data. Finally, analysts performing exploratory or ad-hoc analysis will tend to create data structures to support the analytical modeling process and this tends to be stored as de-normalized or flattened structures, often with summary or aggregated data. Due to the nature of this work, these processes are often developed by the business user with little automation and with more focus on addressing the immediate data need.

The processes associated with the data flow throughout the organization have impacts on the differing user communities. For the technology team their main aim is to support the timely delivery of data with a focus on throughput and speed, ever looking for efficiencies to help hit service targets. The business user is looking for data delivered in a timely manner that is suitable for reporting usage and will be looking for consistency in delivery times and accuracy of the data. For the analyst, they are less interested in how the data is structured and more focused on getting access to the data, with the ability to shape and combine the data as needed. As previously shown, these are often disconnected processes managed across several functional areas. A key element that can make this a more manageable set of processes is any notion of metadata (technical and business) to support these processes. For the technology team, they can perform impact analysis to truly understand how changes to data flows will affect the end

consumers. For the business user, they can understand the reporting structures and the associated business logic. For the analyst, it provides a means to understand the journey and transformations applied to a piece of data and this will allow them to understand how to use this in their analysis.

Thus, we can see that the ETL processes are multiple sets of flows that can be delivered by multiple technologies and approaches that will meet the end user needs, even with a disconnect. This is what is commonly seen in organizations whose data and analytics usage has grown organically over time.

STORAGE DRIVERS AND BUSINESS REQUIREMENTS

Many aspects drive decisions on data and storage in the organization when looking to deliver and store data for business applications. The two main considerations are:

- Where do I source the data?
- How will it be used?

The first focuses on an IT perspective which relates to how easy can I get the data and how fast can I deliver the data? Most technology projects run to tight budgets and the project team will be challenged to minimize the man-day effort and thus associated cost. This could pose a conflict with the ideal of considering how this integrates (rather than fits) with an enterprise data strategy. This is what often drives fragmented data architectures within the organization as the project focuses purely on the departmental or functional needs.

Usage of the data tends to fall into one of a number of categories – data warehouse, specific solution, analytics and query & reporting. Each has a very different access and usage pattern and this should be reflected in the data structures and technology employed. Most of these are business driven and it is often an area that is overlooked when designing the data flow and storage. If the approach is inappropriate, then the analytics or business user will find a more appropriate path to the data he requires, which is what leads to the many silos and sandpits that are seen in organizations. Once these workarounds become accepted, they quickly snowball and are often used for many purposes.

Looking at the conflicting interests of the business and technology aspects of any data project it is easy to see that without any external governance or influence, the project will follow an insular approach with no driver to align with strategic direction of the data architecture. Whilst this addresses the immediate needs of the project, in time, the data environments will become more challenging to manage and exploit due to unco-ordinated approaches.

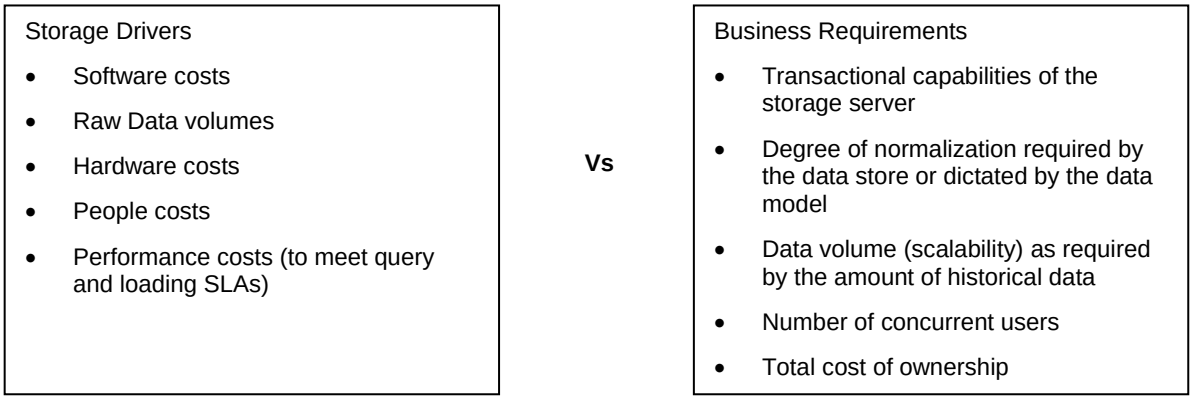

Figure 2. Storage Drivers v Business Requirements
The Conflicting Elements in Data Projects

The governance role in ensuring that the solution adheres to the data strategy and target architecture usually falls with the Enterprise Architect to provide a steer on direction. That said, all too often, the Enterprise Architects can only

provide guidance and do not have the power to enforce compliance. This leads to many projects self-exempting on the grounds of business needs and focusing purely on the immediate need rather than the bigger picture.

HOW DOES THIS LOOK IN PRACTICE? YOU REAP WHAT YOU SOW?

So what does this look like in the real world? What does the data architecture of analytics hungry organizations look like? Where are they on their journey? Needless to say, to answer these questions one needs to understand the aspirations and directions of these organizations to truly gauge the level of success of the architectures. There are some common trends and issues that one should take into consideration when reviewing one's own situation.

1. Does your strategy focus on data storage or is it usage of the data? Is the RDBMS king? Many organizations struggle across these two different perspectives and as such this often becomes more of a political fight rather than an approach to deliver the best of both. It is not uncommon to find large user created data stores at sites where the organization focuses purely on the storage and data model. This often highlights a split between the IT and Business communities within the organization. The reality of this scenario is that often neither approach is fully realized in the delivered solution.
2. Departmental perspective v Enterprise strategy. We have already touched upon the pressures pulling a project in a certain direction, however, whilst departmental solutions meet the immediate needs, the more complex and distinct they become, the more problematic the challenge of trying to ultimately align and integrate solutions across the organization. Eventually, the effort required to pool such data will ultimately be the prohibitive factor. Even more so when you consider that the existing approach is already delivering value – how does this compare with the integration costs?
3. Data Redundancy is often cited as an issue in the organization. However, like all issues, there can also be positive aspects. This issue is often raised by those who maintain the data warehouse and from their perspective it is an issue. That said, additional copies of the data allow the data to be deployed on an appropriate platform for either improved performance ("horses for courses") or to facilitate creative usage of the data. Both of which can be seen as positives for the business user.
4. Complexity of data management may be or may not be recognized in organizations with data silos and flows that are not connected. The reason that this might not be recognized is often workaround approaches are hidden away or performed by areas that are not supposed to provide support functions and so this activity goes unnoticed or is just accepted by the organization. A classic example of this is the number of analysts who still spend more time managing data stores rather than spending time performing analytical tasks. Further to this, due to the lack of understanding of the consumption of the data, it is difficult for support staff to understand the business impact of any changes to the data.
5. Data Latency occurs because of the time taken to move and transform for the data between storage solutions in the organization. Appropriate data design will address some of this and will improve some of the load performance. Only moving required data for the user will also reduce latency. Most of this can be addressed at the design stage of the data structures and the data flows. As already said, the end usage is not always known at the outset. Some organizations get around this by closeness between IT and business to evolve this, but more often, hand coded workarounds address the changing the requirements. Leaving approaches that are not sustainable and leaving the IT team more distant from the user requirements, unaware of the need for change.
6. When there are multiple approaches to data storage and movement, we often see that complexity of the interfaces and the nuances of the various data models become prohibitory to the use of the data. This means that either the data goes untapped or unexplored or might be sourced from alternative locations. Either way, this is time lost for the analyst or the opportunity cost is the missed opportunity to derive value from the use of the data. This will also have an impact on new data projects and will possibly extend the design stage of the project.
7. With multiple strategies for the data flow, there is little chance of consistent end to end data lineage and also potentially no reconciliation of the data. Without such reconciliation it becomes difficult to have full confidence in the accuracy of the data used and makes any activity using the data difficult to audit. This lack of auditability could lead to regulatory issues in these times of ever greater regulation regarding the usage of data. The lack of reconciliation could lead to an inappropriate decision being made with further consequences.
8. By allowing the growth of multiple storage approaches, data volumes for storage will grow at a greater volume than source data acquired. Whilst storage is relatively deep, when you consider the total cost of ownership to include hardware, maintenance and support, it soon becomes apparent that money could be saved by understanding the bigger picture. On the other hand, one has to consider the additional resilience that the extra data brings as well as the potential of better performance by splitting the workload over different systems.

9. Collaboration and integration becomes challenging. Working across teams is difficult because they are working from different data in a different format and possibly with a different meaning. If analysts cannot work across teams then it will be challenging to make analytics pervasive in the organization meaning that the true value will not be found. Worse still, with different departments reporting or making decisions on different versions of the data, there is the potential for internal conflict and potentially mistakes that could have external impacts. In the current climates as organizations streamline and look to truly integrate recent acquisitions, this becomes challenging at best with additional time and cost required to align and integrate the businesses.
10. It has to be acknowledged that different users require different views of the data and as such "one size fits all" does not apply. Intentions to reduce duplication or replication of data are sometimes misplaced when applied to different functions within the user community. This often highlights that the data strategy is being driven by a perceived cost control rather than value creation. In practice, this is generally a false economy with savvy users creating their own silos to service their needs. This in turn creates more storage and data movement with associated (if not recognized) costs.
11. The debate of speed of load versus ease of use will always exist until there is true understanding and alignment between the IT and Business communities in the organization. Everyone has an opinion depending on where you sit. However, this really is a debate that needs rationalizing. The key driver is value and thus we should be looking more to right time movement of data. If meeting the speed challenge we deliver data that needs re-work or is unusable then the right speed has not been achieved. This often occurs when the ETL and Storage teams focus purely on the data warehouse and nothing beyond. This is more of a cultural aspect and often leads to division between the communities. With appropriate education and metrics for the delivery teams, this should be surmountable.
12. Business questions are unpredictable and constantly changing to meet market requirements. Is the approach flexible enough to accommodate this? Whilst there is a lot of benefit relating to manageability from production processes, this is often the Achilles of such systems. If you cannot drive business value from your data then the solution is purely a cost and therefore becomes redundant. Flexibility can be designed into processes with an appropriate cultural change to support this.

For each point above, for every negative perspective, a positive could be found in most situations. It is important to truly understand the objective within your organization and consider the positive or negative appropriately. All can and are mitigated against in practice, but it must be recognized that any workarounds only have a limited, sustainable lifespan, the length of which is determined by your data growth or analytical aspirations. Eventually, a point will be reached where the approach used will become a constraining factor. At this point, one has to consider more of the same or a change of approach or technology.

CHANGING APPROACHES AND NEW PROCESSING TECHNOLOGY

The approaches and associated experiences mentioned above reflect an approach that has remained largely unchanged for many years. Over time, more data was gathered, more data stores were created to support applications and fortunately, the technology grew proportionately to maintain the level of responsiveness that the organizations had come to expect. Bigger data meant bigger servers to support the bigger queries – but where does this all end and at what point do we recognize that this approach is constraining the exploitation of the data within the organization? At what point does the data management function become too costly or complicated that it becomes non-existent? At what point do organizations begin to fail due to lack of regulatory governance over the use of the data?

Recent technology developments began to look into ways to continue growth, but based on the concepts of optimizing across the various technologies within the landscape. Thus, by looking across the end to end data flow, one could exploit a number of different technologies by pushing the logic to the most appropriate platform, reducing data movement and optimizing processing. By breaking down the data and technology silos, the organization has the opportunity to look at a more holistic picture across the data landscape and make more appropriate choices and uses of technology to support the business requirements.

IN-DATABASE PROCESSING TECHNOLOGIES

SAS In-Database relates to the ability to move SAS processes to the database meaning that more work is done inside the database which results in less data movement occurring. There is also the potential of significant performance improvements when utilizing highly scalable data platforms.

In-Database is the ability to embed and use SAS functions, framework processes and applications inside the database including the SAS Format function, SAS Scoring Functions and Predictive Modeling Functions. This is separate to the more traditional Integration approach that SAS has followed for many years through the use of SAS/ACCESS technologies. In this approach, SAS applications are integrated to leverage standard database features including database specific SQL, SQL functions and Stored Procedures.

- SQL Pushdown – this is looking at the ability for SAS to natively produce more performant database SQL than traditionally. This helps to reduce data movement from both temporary and permanent perspectives. This allows the SAS user to further exploit the data platform whilst still working with interfaces that one is used to.
- Data Integration and ELT support – SAS DI Studio 4.2 introduced additional support for databases including supporting the notion of ELT which is complementary to the RDBMS vendors' perspectives. The DI developer can quickly see where the source and target data resides and where the transformations will take place. The DI jobs can be quickly optimized to take advantage of the approach being followed with the organization.
- SAS Analytics and Scoring Accelerators – provide the ability to take a SAS process and publish this to the database as a database object or function which can then be called by a database process, not just SAS. This is taking a statistical process and running it efficiently on a SQL engine without a manual and painful re-write. This will reduce data movement and move model scoring into a true production environment with the associated quality processes and performance benefits. The SAS DI developer will move towards creating specific jobs to support the production modeling process rather than ad-hoc movements of data to support the analytics development cycle.
- Data platforms and sandboxing – recently the database vendors have become more mature at managing mixed workloads (OLTP and ad-hoc analytics) as well as different data structures (normalized and de-normalized). This has enabled them to become more of a data platform and the benefits for the organization that employs a lot of heavy analytics is that the sandbox can now be hosted on the data platform with the analysts self-servicing their data needs directly from the warehouse. This leads to less data being moved between platforms and allows for greater visibility of the usage of the data. The methodologies to support this have become more sophisticated to allow flexibility within the sandbox whilst bringing in more management of the data. Needless to say, any experiments on samples in the sandbox can quickly be replicated to full volumes on the database.

All of these developments provide options for the organization and allow the SAS developers to work across platforms, exploiting what is available and defining the optimal approach to support the data processing needs of the organization.

IN-MEMORY PROCESSING TECHNOLOGIES AND GRID COMPUTING

SAS In-Memory Analytics and SAS Grid Manager provide highly efficient and highly performant approaches for analytical computations. Both accelerate the processing of SAS programs or increase the scale or scope (number of users, size of data sets, and frequency of analysis) of a particular SAS application. These innovations provide a more strategic approach to building and managing a lower-cost IT infrastructure based on commodity computing hardware that can flex to meet rapidly changing and growing computing requirements. This brings true modernization to traditional or legacy SAS environments.

WHERE DOES THIS GET US TO?

We are at a point where the landscape is changing due to the technology enhancements. What should the goal be for the data strategy within the enterprise? The analytical processing options now available allow us to truly focus on the business consumption of the data rather than the processing requirements. We can now truly look to utilize the assets capabilities rather than producing workarounds.

For the Enterprise Architect, this might challenge traditional perspectives of the data layer and will certainly require them to become closer to the analytical aspects of the organization. There will certainly need to be a change of thinking regarding standards within the organization relating to what processing occurs on what platform and which

storage supports which application. Either way, this is a positive note, because it provides more flexibility in approach and allows them to set an appropriate strategy. What it does mean is that the analytics world can become closer to the operational world again breaking down more of the barriers. Time will tell how successful we will be in adopting these changes.

For Data Integration professionals they can focus on ensuring that the right data is in the right place for the activity and providing flexibility to business consumer. Traditionally, the focus would have been on being told what storage to populate, whereas now the role should evolve to the DI team advising where to find the data and where is the right place for processing. Focusing on understanding the metadata and business logic will be the key to adding value to the process.

The Analytics Team should be able to focus more on the business challenge and developing the most value from analytics rather than on the data acquisition, movement and management. With the ability to sandbox on many data platforms, the analyst can often explore the data on its original platform before considering the movement of such data. This should unlock some data and platforms that were previously unreachable by the business user.

A CHANGING LANDSCAPE

With all that we have considered it is obvious that directions and focus are changing. Which raises the question of the validity of the traditional storage layers mentioned in the opening - do we still need all of the layers and do they still add value? Looking ahead, the answer can only be found in the objectives of the organization – will agility be the driver or is structure and foundation perceived to be the value creator? This should make us consider and possibly challenge the objectives from different perspectives. We can also see that there will be an ever more important role of data integration in bringing to together the data in the organization for consumption as opposed to traditional conveyor belt ETL. The ability to acquire data fit for purpose for business use is a primary linkage between the business and IT worlds.

Which brings us to the question of breaking down the silos? Many have tried to reduce the number of marts and silos within the organization and have been challenged by the user community. Whilst all agree that there is value to be had from bringing together and combining the data of the enterprise, we have previously approached this from a manner to support storage application needs rather than usage needs? The new technologies change this part of the game which means that the convergence of data could be logical and not necessarily physical. Thus, the enterprise view of the data is truly that – a view across all platforms within the organization rather than the attempt to build a monolithic data platform. This can only be achieved through a solid metadata base and a mature approach to Data Integration.

CONCLUSION

We now have many options to approach our storage and data flow needs. Look across your organization and look for data efficiencies. Consider only moving the data to locations that are best for performance and then optimize the process to meet the needs of the consumer. Flexibility will be the key to meet the constantly changing needs of the user and we now have more options to achieve this.

It is no longer a technology challenge, but more of a culture challenge. We all have a part to play in the data strategy, but we need to start educating, thinking and planning. The questions to ask yourself and your organization are:

- Where do you want to be? You need to understand your objectives and value drivers.
- Does your Data Strategy and Target Architecture Models Reflect this? Enterprise Architecture and Data Integration are vital in helping to chart an appropriate course.

CONTACT INFORMATION

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

Adrian Jones
SAS Institute Inc.
Cary, NC 27513
E-mail: adrian.jones@suk.sas.com
Web: www.sas.com