Enterprise Data Management in an ‘In-Memory’ World

Tactics for Loading SAS® High-Performance Analytics Server and SAS® Visual Analytics
Table of Contents

Executive Summary ......................................................... 1
The Analytic Data Warehouse ................................................. 2
  Types of Analytics ......................................................... 2
  Finding the Right Data Structure for Advanced Analytics .......... 3
  Finding the Right Database Administration Skills ................. 3
Sourcing the Analytic Data Warehouse .................................... 4
  Homogeneous Point Data Sourcing ...................................... 4
    In-database processing uses existing assets ......................... 4
  Heterogeneous Point Data Sourcing .................................... 5
    Orchestrating multiple vendors’ databases .......................... 5
  Hybrid Ecosystem Data Sourcing ........................................ 6
    Analytical master data management ................................ 7
  Distributed Ecosystem Data Sourcing .................................. 8
    Master data management is key to distributed data ecosystems . 9
    Grid computing from SAS accelerates loading .................... 10
Conclusion ........................................................................ 10

Scott Chastain leads SAS Americas sales and consulting forces in positioning and integrating the SAS Business Analytics Framework infrastructure to address customer business initiatives. He often works with multifunctional customer teams comprising both IT and business to align technology to business initiatives ranging from data integration to big data analytics.

Gary Spakes focuses on both the technical aspects and business implications involving SAS High-Performance Analytics. An IT veteran, Spakes leverages his technology background to help organizations alleviate immediate pains without compromising long-term initiatives.

Mark Troester and Tapan Patel (SAS) also contributed to this paper.
Executive Summary

As organizations mature in their use of in-memory analytics, it is imperative to develop a more agile and efficient data management strategy. Organizations must understand several contributing factors so that data is managed as a valuable asset for making operational and strategic decisions:

- **Big data.** Big data is a relative term that describes a situation where the volume, velocity and variety of data exceed storage or computing capacity. To accommodate big data and to make accurate, timely decisions, organizations need to reconsider their whole approach to data management and analytics.

- **Data structure for analytics versus traditional enterprise data warehouse.** The primary goal of the enterprise data warehouse (EDW) is to store highly normalized data, which is suitable for transactional data. For advanced analytics, data needs to be stored in a denormalized fashion in an analytic data warehouse (ADW), wherein a column can appear in more than one table for one-to-many relationships. It is not unusual for an ADW to be terabytes in size – especially if it includes unstructured data. Processing these types of workloads requires a scalable analytic architecture, both from data management and analytics perspectives.

- **Information governance.** Enterprise data is often held in disparate applications across departments and geographies, which leads to poor service, redundancies, inaccuracies and, ultimately, a higher cost of doing business. An effective information governance strategy should include capabilities and services for data management (e.g., data stewardship, master data management) and analytics management (e.g., data preparation, model deployment). This approach allows organizations to maintain a single, consistent set of policies and processes for managing data, models and decisions.

Traditional data management strategies will not scale to effectively govern large data for high-performance analytics. Organizations must institute new paradigms for moving, cleansing and transforming data, without sacrificing metadata management. Long-term success with big data analytics projects is directly related to proper management of the initial data loading that’s required for SAS® In-Memory Analytics products such as SAS High-Performance Analytics Server and SAS Visual Analytics.

By investing in SAS Data Management solutions, organizations can effectively process and integrate large amounts of data required for analytics. In turn, they can pioneer new information management best practices and methodologies for loading a database appliance (e.g., Teradata or EMC Greenplum) or a Hadoop distributed file system.

This paper discusses tactics, best practices and architecture options associated with loading analytics-ready data required for SAS High-Performance Analytics Server or SAS Visual Analytics. The architectures described are:

- Homogeneous point data sourcing.
- Heterogeneous point data sourcing.
- Hybrid ecosystem data sourcing.
- Distributed ecosystem data sourcing.
The Analytic Data Warehouse

Data traditionally flows from an organization’s operational systems into an enterprise data warehouse (EDW). Data can then be used for business information reporting and querying – basic data summarizations that provide insight into what has happened. However, traditional EDW/data marts are not structured or set up for creating or updating analytical models. The result is unnecessary data movement – modelers extracting and repurposing data multiple times into their model development environments. This movement can lead to myriad information challenges, including data redundancy, data governance, auditability issues and more. Data management solutions have arisen to meet these needs. But as big data challenges mount, organizations need new approaches to augment existing methodologies.

Companies are searching for dynamic, agile systems that reduce data movement and data latency. They need analytic model development and deployment cycles to be more efficient and less prone to operational errors to avoid delayed and inadequate business decisions.

Types of Analytics

Organizations rely on four main types of analytics capabilities. These include:

- **Descriptive.** Descriptive analytics allows you to visualize trends, patterns and relationships in existing data. There is no response (dependent) variable for which you must predict the value.
- **Diagnostic.** Diagnostic analytics helps you understand the causes of an outcome, often in the context of a process or related events. You can use various techniques and models to abstract and account for dependencies among causal factors.
- **Predictive.** Predictive analytics helps you find good rules (models) for predicting the values of one or more response (dependent) variables from the values of predictor (independent) variables in a data set.
- **Prescriptive.** Prescriptive analytics is used to develop a course of action in response to an event or series of events. A prescriptive model can help you define and articulate the ideal process to follow.

For organizations to move to the next level of analytic maturity, their data needs to be manipulated into an analytic data warehouse that supports more than descriptive analytics. Organizations often use predictive, diagnostic and prescriptive analytics to drive innovation throughout the enterprise. These business innovations affect both top and bottom lines, so organizations are evolving their infrastructures to support the growing analytical data demand.
Finding the Right Data Structure for Advanced Analytics

ADWs can differ from EDWs in several ways. The primary goal of the EDW is to be highly normalized. In this way, it can optimize data storage and minimize repeating values, with the basic assumption that data will be highly structured. Tables in the EDW may be long (meaning they have a large number of rows) – but they typically are not wide (they have a limited number of columns). This type of warehouse is designed to minimize duplicated data by using a series of keys and linkages. This approach provides some advantages by reducing storage costs and maintenance. But these structures are not optimal for advanced analytics.

In a typical analytic data warehouse, data is stored in a denormalized fashion. This means that the same column appears in more than one table and values will be repeated in the case of one-to-many relationships. For example, variables such as birthday and gender will appear multiple times per customer in one table. The table can then be used to apply customer analytics. Denormalization is required for analytics if data has to be merged together in one single table, which makes it long and wide.

ADW tables may be both long and wide, with an extreme number of rows and columns. It is not unusual for an analytic data set to be gigabytes, if not terabytes, in size – something that would cause most business intelligence reporting to take many hours. Analytic data tables may also consist of unstructured data from nontraditional data sources that are combined with structured data from operational systems. Data values may repeat with a row – perhaps several times – as data is coalesced from multiple data inputs. For example, an appropriate data structure for data mining will have all variables associated with a given observation in a single, materialized record. This structure is optimal for the iterative mathematical processing often used in advanced analytics.

Finding the Right Database Administration Skills

The database administrator (DBA) role is vital for supporting hub-and-spoke data ecosystems like those that are prevalent today. DBAs have various specializations ranging from transactional operational systems to decision support EDW and data stores.

The commoditization of data appliances, the explosion of enterprise data and the broad maturation of governed business analytics have led to a new breed of DBA – the analytical database administrator (A-DBA). Organizations that implement ADWs for specific purposes should tap an analytical database administrator to devise and manage the analytical data ecosystem required for high-performance analytics. In some organizations, this role is filled by a specialized data scientist with enhanced database expertise.
Sourcing the Analytic Data Warehouse

Homogeneous Point Data Sourcing

Homogeneous point data sourcing is when an enterprise data warehouse serves as the single data source for an analytic data warehouse. In addition to being a single-point source, the EDW and the ADW use the same storage and computing platforms. For example, a Teradata enterprise data warehouse acts as the sole data provider for a SAS High-Performance Analytics Server (based on a Teradata database appliance) solution.

See Figure 1 for an illustration of the simplest architecture. This architecture lends itself to significant use of EDW resources to transform and structure required data prior to direct porting to an ADW used by SAS High-Performance Analytics Server or SAS Visual Analytics.

Hadoop

Hadoop is a low-cost, open-source file system that has quickly gained importance because it allows organizations to use commodity-based hardware to store and process or compute large amounts of data. SAS can process, integrate and analyze Hadoop data (e.g., using SAS/ACCESS® and SAS® Enterprise Miner™). SAS also supports MapReduce, Pig and Hive so that users can employ Hadoop as part of an ELT/ETL strategy to prepare and load data for analytical processing. SAS uses Hadoop (storage and data persistence) as a way to deliver in-memory processing for data exploration, visualization and analytics.

Figure 1: Homogeneous point data sourcing is a simple architecture that takes full advantage of EDW resources and computing capacity.

In-database processing uses existing assets

Although numerous organizations have switched to enterprise toolsets that take advantage of data center assets, most companies are still challenged by ad hoc data extractions, extraneous data movement or hand-coded processes. A best practice is to use existing massively parallel processing (MPP) resources to transform and structure data prior to loading the ADW. But it is important not to lose or bury the metadata lineage within hand-coded programs. If that happens, organizations increase compliance risks, making it more difficult and costly to automate model deployment.
In a typical deployment, analytical base table structures may be represented by semantic views based on one or more other table structures within the EDW. Using views offers several advantages. First, it reduces data replication within the EDW. It also allows you to keep contributing tables’ locking strategies without affecting ADW loading. Finally, using views provides an opportunity not only to restructure the data for consumption, but also to apply any predictive models that were deployed prior to loading the ADW. Performing this restructuring and analytical enrichment in a single step reduces data movement and improves efficiency.

With a common platform for the EDW and ADW, you should use the database replication utilities to directly load the ADW. This method streamlines input and output, and does not require any in-flight transformation.

**Heterogeneous Point Data Sourcing**

Heterogeneous point data sourcing means there is a single data provisioning source for an ADW. But the ADW and data source provider are based on different storage or computing platforms. For example, the single point data source for SAS Visual Analytics (using a customized Hadoop distributed file system as an underlying disk persistence mechanism) could be an Oracle-based EDW running on Sun server hardware. In this situation, database vendor-supplied replication utilities are not optimal. You should employ some of the same techniques as in the homogeneous point sourcing architecture. Using views and transforming logic within the MPP architecture is still a best practice.

**Figure 2:** In heterogeneous point data sourcing, a majority of data comes from a primary EDW augmented by other sources. Organizations can use a master data management (MDM) hub to rationalize the data.
Orchestrating multiple vendors’ databases

When you are uniting multiple vendors’ databases that have incongruent platforms, you will need to use database utilities other than direct replication. Most ADWs support various high-performance utilities – such as bulk loading or upserting – specialized to the computing platform. You should take advantage of these optimized facilities when possible, to optimize system efficiency. Using rich metadata interfaces in tools such as SAS Data Integration Studio will allow you to orchestrate various database-optimized facilities while maintaining automated visual documentation and data lineage. In a heterogeneous environment, you can accomplish high-performance data movement by using a tool such as SAS Data Integration Studio to pull in multiple vendors’ tools simultaneously.

In addition, you should deploy change data capture (CDC) mechanics to minimize redundant data processing and to maximize I/O throughput. Organizations often use CDC mechanisms in transactional operational databases in the form of change logs or journals. But in most MPP database appliances optimized for EDW deployment, these logs or journals are not implemented. If the source database does not natively support CDC, then another CDC strategy should be implemented. For example, various data management products offer CDC capabilities that are database-independent. SAS Data Integration Studio supports a generic CDC capability that you can use to determine appropriate data loading requirements – and you can push that logic down to the source data system for optimal performance.

Combining these practices will reduce data redundancy and minimize the time required to load an analytic data warehouse.

Hybrid Ecosystem Data Sourcing

Hybrid ecosystem data sourcing for the ADW relies predominantly on the EDW for data, but it also pulls data from other sources. Many organizations have more than one data warehouse, and they rely on a variety of different data sources to support business analytics. A relatively mature organization may have a primary data repository – but business analytics often requires classic data management tools to access, explore, cleanse, transform and structure data from various internal and external sources. These other data sources are used to augment the primary data source. For example, to load a Teradata or EMC Greenplum database appliance used by SAS High-Performance Analytics Server, an analyst may be able to source approximately 80 percent of the data required for a business need from the EMC Greenplum EDW. But the analyst also has to access two Oracle databases, three text files and a spreadsheet to get the rest of the data.
Figure 3: An MDM hub can orchestrate ADW data strategy by helping to gather, cleanse, structure and link data from multiple sources. Organizations can use a primary source EDW to drive overall strategy.

Analytical master data management

To optimally support analytic techniques, data needs to be structured so that it can link various transactional, behavioral or otherwise relevant data to key entities such as customers, patients, accounts or products.

Numerous organizations have adopted master data initiatives as a way to transform their enterprise data paradigm. Enterprise master data management (MDM) solutions – such as SAS MDM – can be deployed in various architectures, ranging from operational systems synchronization to traditional EDW methodologies. MDM best practices are well-documented and are not covered in detail here. Instead, we focus on a growing application for master data management hubs that support decision-support systems.

Master data management holds the key to orchestrating the data strategy for high-performance analytics. Master data is the subset of data that defines key attributes about high-value data domain areas such as people, suppliers, products, accounts, etc. Many organizations have numerous sources that create or house critical data required to define and manage a consistent view of a master data entity, such as account. Many use an enterprise data warehouse as a major source. But advanced analytical problem solving often requires data from many other sources that must be gathered, cleansed, structured and integrated to load an ADW. MDM hub architecture, such as the SAS MDM solution, allows you to quickly link data from disparate systems. It works by maintaining the primary variables and the best record representation of key data entities, as well as the links to how individual best records are represented in contributing systems.
This hub methodology can manage multiple entity types in a single instance. For example, you can manage all the best attributes and linkage information for accounts, customers, and business partners in a single MDM hub. This hub becomes a key component that allows you to accurately load data into an ADW for high-performance analytics. The MDM hub creates agility by enabling you to link behavioral and transactional data to entities such as accounts or people, even though sources and systems may evolve over time. The hub maintains and persists the abstraction of identity resolution and linkage. The abstraction can be used for ADW applications and also across the enterprise to improve governance and flexibility of the ADW and future applications.

In a hybrid ecosystem, the EDW contains the great majority of the data required to support SAS High-Performance Analytics Server or SAS Visual Analytics. But multiple systems outside of the EDW are still required. An analytical MDM hub will run alongside the EDW. A best practice is to synchronize the MDM hub information with the EDW. This practice serves two purposes. First, it gives you flexibility to explore and make quick use of systems outside the EDW. Also, the data residing in the EDW may be easily consumed via views or other in-database techniques for performance.

Synchronizing the hub and the EDW does create some data redundancy. But the relative amount of data is small, and the processing improvements gained by being able to link large quantities of detail data across master data entities far outweigh the increased storage costs.

**Distributed Ecosystem Data Sourcing**

Distributed ecosystem data sourcing uses dispersed data across the enterprise without any single source dominating the data sourcing for an ADW. For example, customer data is in a customer relationship management warehouse; product data resides in an enterprise resource planning application; and various other operational data is dispersed in other siloed applications. All this cross-domain data is required to support SAS High-Performance Analytics Server or SAS Visual Analytics.
Master data management is key to distributed data ecosystems

To consolidate master data and the linkages to detailed information stores, it’s crucial to use a master data management hub. This system provides the benefits of centralized business rules and streamlined data access across disparate systems without the need for extraneous data duplication.

It is important to design the MDM hub so that various business analytics processes can be easily supported, from data exploration to deployment. Another key is to design the hub so that it is scalable and relies on industry-standard load balancing and resiliency methodologies. The primary difference between this method and the hybrid paradigm deployment is that the hub data is not synchronized to any particular point source and will operate as a more independent system.

With a hybrid system or point system, an A-DBA can call on the power of the database to perform data work in an ELT fashion. But when data is equally distributed across several data sources, these methods are not as effective. This topology typically cannot make use of existing MPP-style resources, due to either enterprise or technical constraints. For example, there may not be enough resources to move sets from several systems to one system for ELT-type processing. The computing infrastructure must be defined and managed accordingly.

Figure 4: For distributed ecosystem data sourcing, an MDM hub plus a grid-enabled data strategy are necessary, because a primary EDW cannot be used to drive overall data strategy.
Grid computing from SAS accelerates loading

A master data management hub can provide all the required elements that link transactional or behavioral detailed data across various entity domains such as person, account, product, policy, etc. The distributed data ecosystem typically does not have a single system that contributes the bulk of the required data, nor does it have the bandwidth to drive data from multiple disparate systems through one of the contributors as the hybrid approach does.

This style of implementation requires extensive computing resources to extract, transform, cleanse, structure and load data into the ADW with as few touches as possible. The goal is to “one-touch” the data by performing all required work on it while it’s in flight, to minimize the loading time for SAS High-Performance Analytics Server or SAS Visual Analytics. A solution such as SAS Enterprise Data Integration Server offers some significant advantages for this loading strategy:

• First, the solution enables you to quickly use a drag-and-drop transformation to interface with a master data management hub, streamlining data access capabilities. The MDM hub can easily provide all the various source keys required to reach across any number of disparate sourcing systems within the ecosystem.

• Second, the solution includes embedded data quality capabilities that you can apply directly during data access or within the process to minimize any disk operations that would hamper performance.

• Third, any existing scoring models can and should be applied during the process. Because the data management interface has analytical scoring capabilities, it reduces data touches and minimizes latency.

This information management strategy is adaptable enough to be designed once and then scale across a dynamic SAS Grid Computing infrastructure. For example, if another data source needs to be added to the ecosystem, you could do that seamlessly.

Conclusion

Today, many organizations are evolving their enterprise architectures to address specific business analytics needs. It is important to devise and deploy appropriate data management strategies so that you can quickly maximize return on investment from SAS In-Memory Analytics products like SAS High-Performance Analytics Server and SAS Visual Analytics.

The four representative data management architecture approaches and best practices discussed in this paper will help IT organizations quickly realize the potential from big data – and make reliable information available where and when it is needed. As new data sources are added, it’s crucial to quickly assimilate, integrate and analyze relevant data in a consolidated manner so you will gain value from high-performance analytics. The architecture approaches described also provide a balance between requirements for data latency and frequency of decision cycles.
Using a metadata-rich data management solution, such as one that includes SAS Enterprise Data Integration Server and SAS MDM, will provide the core capabilities you need to get the most out of your analytic data warehouse, and to prepare your data for business analytics. By implementing a flexible data management approach, you will be well prepared to address a multitude of evolving business, operational and technical requirements.
About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.