Data Warehouse Implementation with the SAS® System

Tony Brown, SAS Institute Inc., Dallas, TX

Abstract

The successful implementation of a data warehouse can be an extremely complex process, requiring significant planning, patience, and persistence. The Professional Services Division of SAS Institute Inc. has found that a good understanding of data warehousing concepts, coupled with a solid development methodology, and a technical infrastructure built around the SAS System can enable practitioners to achieve their goals of effective data warehousing. This paper addresses many of the challenges and issues that The Professional Services Division has confronted during the development of data warehouses. It explores the definition and purpose of data warehouses, common myths and mistakes in warehouse development, and presents the SAS Data Warehouse Model. It also provides information on staffing data warehouse projects, and finally introduces the SAS Rapid Data Warehouse Methodology.

This paper will be of interest to anyone who is planning or has recently initiated a data warehouse project. Several SAS products will be highlighted, but experience in their use is not necessary to derive benefit from this paper.

Introduction

The subject of data warehousing is beginning to dominate much of the focus and energy of the information systems world. It is a complex subject with many technical and methodological facets that determine the success or failure of decision support systems. In order to compete in today's marketplace, companies need easy access to value-added information. In this paper, the experience and knowledge gained by the Professional Services Division will be used to define what a data warehouse is, why it is necessary, and what it should do. More importantly, it will discuss how to successfully build a data warehouse using SAS System tools and SAS Institute Inc. methodologies. Some of the SAS System products discussed will include base SAS® software, SAS/MDDB Server™, SAS/EIS®, SAS/AF®, SAS/FRAME®, SAS/STAT®, SAS/GRAPH®, SAS/FSP®, SAS/ASSIST®, SAS/INSIGHT®, SAS/IntrNet®, SAS/GIS™, and the SAS/Warehouse Administrator™ product.

What is a Data Warehouse?

Data Warehouse vs. Transactional System

In defining the concept of a data warehouse, it is helpful to distinguish it from a transaction processing system. Online Transaction Processing Systems (OLTP) have the primary role of capturing business transactions, and the data relevant to them. The OLTP system is primarily concerned with adding, updating, and deleting data. This data represents "actual events", and is at a very atomic level. OLTP systems are typically implemented with hierarchical or relational database management systems, and are accessed using structured query language (SQL). They are typically single platform (usually mainframe) based, and managed by a central corporate information systems entity. In summary, their primary function can be thought of as getting transaction data into the information system quickly and accurately for clerical usage.

Since these systems must support huge volumes of adds, updates, and deletes (thousands per second in many systems - think about an airlines reservation system), they are very performance oriented. This performance orientation often involves utilizing a database schema that is entity-relationship based, and highly normalized. This type of schema provides a framework to support explicit data relationships, enforce referential integrity, and limit redundancy, to increase data access speed for single record operations.

In his book, "Building the Data Warehouse", W.H. Inmon presents the following definition of a data warehouse; "A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of

management's decisions."[1] An understanding of this statement begins to point out the major differences between the data warehouse and the transactional system. The purpose of the data warehouse is to provide information for managerial decision making that is distinct from, and does not interfere with the performance requirements of the OLTP systems. This is usually done by re-engineering OLTP system data around business subjects (for example Auto Sales), and making it easily available to the managerial community in a separate repository. This subject-based modeling allows a more logical view of the data from a business users perspective.

Before being loaded into the data warehouse, data must go through a transformation or "integration" process. This process has several distinct steps to make the data clean, uniform, and quality assured for use in the warehouse and will be covered later in this paper. In brief, integration ensures data fields are populated, data is correct, not out-of range, and not duplicated. Data warehouse data are not updated with the frequency of transaction data, therefore they are non-volatile (not updated in real time). Data will not change during the execution of an analysis, nor will two users get different answers requesting the same information. The data warehouse typically gets updated on a weekly, monthly, or bimonthly batch update cycle from the OLTP system. Since the data are non-volatile, the data warehouse does not need to employ the overhead aspects of the traditional RDBMS, such as locking, referential integrity checks, transaction logging, checkpoint/rollback, and so on.

In contrast to the OLTP system's "actual events", the data warehouse organizes data into dimensional slices of time (time variant - for example Software sales for June, 1996) to produce quantitative results of events at some time interval. In essence, the purpose of the data warehouse is to get data out of the system, and turn it into information. This analysis-driven, time variant data are often pre-summarized into levels usable for decision support, negating the need to summarize atomic level data in query operations. Industry has adopted a term encompassing the process of accessing and analyzing data warehouse data - OLAP.

Data Warehousing and OLAP

OLAP stands for Online Analytical Processing. Simply, it involves using a multidimensional database model, and appropriate access and analysis tools to quickly pull data from the warehouse, and turn it into meaningful information. OLAP was initially championed by E.F. Codd in 1994 as the answer to slow query reporting performance from traditional entity-relation database schemas. "OLAP's goal is to provide Fast Analysis of Multidimensional Information" or FASMI for short.[2]

There are currently two primary methods for implementing OLAP from a model/product perspective. The first method is composed of building subject-based models using a traditional relational database product. This approach is termed ROLAP, or Relational OLAP. In the SAS System, SAS data sets, and SAS/ACCESS views to database tables can be used to implement ROLAP. The second method uses a new type of multidimensional database or "cube" (MDDB - which will be described more fully in the coming paragraphs), and is termed MOLAP for multidimensional OLAP. SAS has developed the SAS/MDDB Server product that is offered in production in Release 6.12.

ROLAP and The Star Schema

The Star schema is often employed when the ROLAP approach is taken. Relational tables are used to model the subject of interest (for example sales, shipments, and so on). The schema utilizes implicit modeling of relationships instead of the explicit models of the entity-relation schema. It is currently one of the most popular subject-based models, involving a "fact" table which holds numeric information about a particular subject (for example Sales volume, profit, ROI, and so on). Because it holds large volumes of numeric factual data about the subject, the fact table is typically the largest table in a subject-based schema. This fact table is surrounded by accompanying "dimension" tables that describe attributes of the "facts" (for example Sales period, region, state, and so on).

Dimension tables are typically much smaller since they provide descriptive data. In order to process a query against a subject-based schema,

the keys of the surrounding dimension tables required to fulfill the query, are acquired and driven against the fact table. Figure 1.1 shows a Star schema, with the fact table, and the accompanying dimension tables surrounding it in a "star" design.

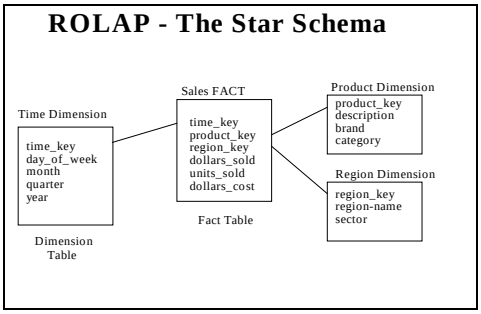

Figure 1.1

While introducing some increased data redundancy, the Star schema provides greater flexibility and simplicity in query generation. This is possible because the data redundancy allows the Star schema to utilize fewer tables to model a subject, requiring less joins to be made.

Another popular subject-based schema is the snowflake schema (not pictured). The snowflake schema, and the variations that can be made from it remove data redundancy in the Star schema by de-normalizing the tables. This is done to reduce storage requirements, but at a performance cost.

MOLAP and the MDDB

The multidimensional database (MDDB) is used to represent the subject-based organization of the data warehouse when the multidimensional OLAP or MOLAP approach is taken. The MDDB is a database structure that stores data into a multidimensional format like a "cube." This "cube" allows data to be stored in an array-like fashion. Reference figure 1.2. This example shows a 3-dimensional cube, the X-axis dimension is Color, the Y-axis depicts Model, and the Z-axis depicts Dealership. Each of the blocks, for example where the Jones dealership, the color Blue, and the model Mini-Van intersect, represents a "cell." The "Facts" or numerical information found in the dimensional database are located in these cells, and are addressed by the values of the "Dimensions" (Model, Dealership, Color). It becomes visually obvious that now one can "slice" through the cube to quickly find facts about blue mini-vans at the Jones dealership. The physical structure and indexing systems of cubes are designed to facilitate this "slicing and dicing" approach to data access. The SAS/MDDB Server multidimensional database works in this fashion. SAS/EIS offers facilities to access the SAS Multidimensional Database for analysis and viewing.

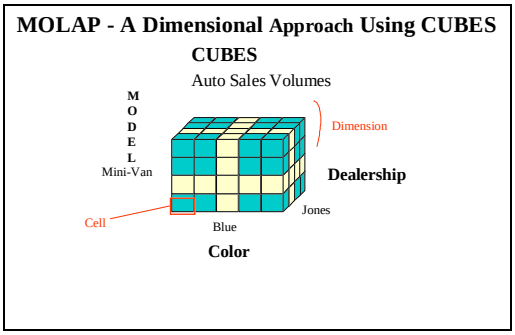

Figure 1.2

Data and Information Marts

Both subject-based tables (ROLAP), and the MDDB (MOLAP), are also used to create "Data Marts." A Data Mart is a body of decision support data (DSS) for a department that has an architectural foundation of a data warehouse. The primary difference is that the data mart is usually at a more refined level (aggregated, summarized, or subset). Data Marts make it unnecessary to access the entire data warehouse for a particular department's data. They reduce the processing load on a data warehouse, and they provide faster query response. They can be thought of as a departmental, "mini" data warehouse.

Some departmental entities also wish to store information in a finished output format (for example reports, graphs, and so on). Some storage structures are built to specifically store this output in a viewable format. These structures have been dubbed "Information Marts". They are useful for storing finished

information of standard reports, graphs, and multimedia that are widely used.

Why Build A Data Warehouse?

Now that we have defined to some extent what a data warehouse is, let's turn our attention to what benefits it can offer.

Quantitative Business Benefits

In an independent detailed study of 62 organizations, International Data Corporation (IDC) found that a staggering average three-year return on investment of 401% was realized by organizations building data warehouses. These figures varied widely by type of business, but the overwhelming evidence was carefully weighed by benefit categories. IDC divided cost benefits realized into three major categories: keepers, gatherers, and users. The keepers category represented savings attributed to the displacement of older, 'non-data warehouse' decision support systems and refocusing the staff involved. The gatherers category represented savings realized by reduced staff and analysts used in the information gathering process when a data warehouse was implemented. The users category identified specific instances when process savings or opportunity gains could be directly tied to using only the data warehouse for their implementation. The keepers category represented 20% of the total warehouse-related financial benefits, with the gatherers accounting for 30%, and the users accounting for 50%. This clearly spells out quantitative benefits to data warehousing!

Qualitative Business Benefits

In addition to the quantitative benefits, there are several qualitative benefits that are harder to measure, but no less dramatic. The first of these is a large one - providing standardized, clean, value-added data to create information from disparate sources. Combining data from these various sources, cleansing, standardizing, and enhancing it to create information is has benefits that cannot always be measured quantitatively.

Another benefit includes better managing the customer relationship or customer opportunity. Information about customers is being used in increasing ways to directly market to individuals. This information is being successfully used to acquire and retain customers, and gain or increase their usage of products and services. In addition, the data warehouse is making data available across corporate organizations for better understanding and usage between often non-integrated functions, such as marketing and sales. In companies where there are critical windows of time to make decisions, data warehouses provide the needed information quickly. Finally, most managers are relatively surprised with knowledge gained from the data warehouse, hidden trends discovered, and previous false assumptions exposed. This surprise is especially felt when control of processes and performance is gained through measurement. That measurement is only possible by turning raw data into useful information.

Turning Data Into Information

In order to compete in the business world, fast, accurate decision making is essential. The data warehouse provides access to corporate and organizational data to support cost-effective decision making. It provides the immediate, on-demand, high-performance response that quick decision-making requires, with clean, consistent data. The difficult process of cleansing and verifying data can also used to justify efforts to clean data in upstream transactional systems, to save doing it downstream where it is more costly, and provide cleaner transactional systems. This provides more, and higher quality data for competitive business benefit.

A major strength of data warehouses is that they support ad-hoc queries, or unplanned explorations of data. A data warehouse gives business users the freedom and ability to analyze data in different ways, rather than being constrained by a limited set of predefined reports. With its architectural advantages and integration of capable exploitation tools, the data in the data warehouse can be separated and combined by means of every possible measure in the business. Additionally, the data act as the repository of the published corporate data.

Myths and Mistakes - Dooming Success from the Outset

Data warehousing is a complex subject that touches the core of business processes and utilizes the full scope of information technology required to make it possible. There are several mistaken beliefs about data warehousing, that when coupled with some common mistakes made when building data warehouses, doom efforts for success. The following sections address some of those myths and common mistakes that result in less than successful implementations, and some outright failures.

Common Myths About Data Warehouses

Mistaken notions are often behind some recurring myths that have become propagated concerning data warehousing. While many who are attempting to practice corporate data warehousing do not openly subscribe to these mistaken beliefs, they still seem to be prevalent. The following list discusses some of the data warehousing "myths" or "mistaken notions" and the reality that the Professional Services Division has encountered:

Myth: A data warehouse is a panacea for all Information Systems Management ills.
This is probably the biggest myth of all. A data warehouse fulfills a specific mission as stated earlier. It does not replace transactional systems or objectives, but fulfills the need for decision support information, and in the process illuminates a host of problems and challenges. It is extremely complex, with management issues of its own, not the least of which involves cleansing and integrating corporate transactional data for warehouse use.

Myth: A data warehouse is an application.
A data warehouse is not an application, but a complete functioning information system. It is much more complex and comprehensive than an application, and should be treated as such in conception, planning, development, and management. A data warehouse can involve the use of many applications in the capture, transformation, and loading of transactional data to the warehouse, as well as analytical access to the data by end-users.

Myth: As long as I use "bleeding edge" technology my data warehouse will be a success.
More data warehouses have failed because developers were "tool-centric" or "software-centric" instead of "solution-centric". Bleeding edge technology, like established technology, never solves problems unless it is appropriately implemented to solve the right business problem - at the right time. It can be the road to success for those with attainable, well specified plans and goals. Utilizing bleeding edge technology can have its drawbacks - being the pioneer with new technology can cost the practitioner with lessons learned by personal experience, instead of at the experience of others. These costs of these lessons learned come in both monetary and schedule currencies.

Myth: I should build my warehouse in one large effort, with an aggressive time schedule.
Professional Services Division experience shows warehouse development should be approached incrementally, a subject or two at a time. Incremental development emphasizing shorter term goals with high payback should be implemented first. Later development should learn from, and build on, these initial efforts, because plans can be justified by past successes and direction.

Myth: My data warehouse will decrease my storage requirements for information processing.
Some data warehouses can have significantly more data storage requirements than their upstream OLTP systems. Once users are accustomed to getting "slice and dice" answers, constant new demands are placed on the warehouse for access to new views and combinations of the data. This leads to an ever-growing repository of aggregated or summarized data, as well as a proliferation of specialized data and information marts.

Common Mistakes to Avoid

Many of the mistakes that render data warehousing projects ineffective are common to most development initiatives, and involve problem/opportunity definition, goals, project management, proper staffing, and so on Due to the multi-organizational involvement, and corporate visibility of data warehousing projects, these common mistakes become painfully amplified. There are some mistakes made that

are particular to the data warehousing initiative. This section addresses some of those common mistakes. Subsequent sections introduce the SAS Data Warehouse Model, Project Staffing, and the SAS Rapid Data Warehouse Methodology that helps prevent these common mistakes.

Mistake: Poor definition and approach. The data warehouse project itself may be poorly understood, improperly defined, and not well aligned with business goals. It is often approached with a "long term" mentality, and the short term goals and deliverables (which probably drove the effort in the first place) are overlooked. The end user is not involved enough to properly identify and participate in the solution of the appropriate business problem or opportunity. The SAS Rapid Data Warehouse Methodology, addressed later, will provide steps to help prevent these problems.

Mistake: Not having commitment from an executive sponsor. A data warehousing project is a large endeavor, consuming large amounts of capital and human resources. Without the top-level direction, sponsorship, support, and results monitoring from an executive sponsor, the project is in danger before it begins.

Mistake: Tying the data warehouse to a specific development initiative. This is often done to make a data warehouse appear more politically or economically attractive. This often results in a mixing of transactional and informational objectives, since many of the development projects initiating a data warehouse originate in the transactional arena. This mixing of objectives can result in a difficult separation of tasks and objectives from the transactional to the informational arena - and can prove to be a great hindrance to the project.

Mistake: Using inexperienced personnel without appropriate skills or experience. Data warehousing requires an assorted skill set, disciplined project management, subject-based data modeling and data architecture for decision support, business matter expertise, business and data analysis expertise, OLTP system knowledge, transactional systems Data Base Administrators (DBA) talent, application development talent, query building skills, and so on. The absence or lack of solid experience in any of the above skill sets can introduce enormous problems. Data warehousing is a relatively new field, and combining the appropriate skills necessary can be one of the biggest challenges.

Mistake: Pre-occupation with technology instead of business goals . To add to the discussion in the Myths section above, business goals should drive the use of technology, not vice versa. In addition, when business goals require the use of bleeding edge, or unproved technology, the deliverable schedules should be carefully managed to allow time for technological "glitches" that most likely will occur.

Mistake: Poor design, and modeling. This is often due to ignoring the rules on subject-based modeling. An person experienced with subject-based modeling focuses on the users' needs, and performance of the warehouse to meet those needs, and model accordingly.

Mistake: Transactional data are not cleaned or integrated before loading into the warehouse. Because the subject-based models making up the data warehouse support the dimensional approach, the factual data in the warehouse, and the dimensions surrounding them, must be supported by accurate, and complete data. The data will be scrutinized, analyzed, and summarized from angles and perspectives never attempted before. Invalid or missing data will become extremely magnified in this process. If this is not addressed before the warehouse is built, the warehouse is virtually useless, as missing or bad data cannot be used to accurately generate information to make decisions.

Mistake: Underestimating refresh and update cycles. Many organizations often underestimate the time it takes to refresh and update the data warehouse from the transactional systems. Some companies have actually built data warehouses and then found it took so long to update the warehouse, it was unavailable to the end-users! This is a critical issue that must be estimated and planned in the warehouse design phase.

Mistake: Loading marginally useful data to the warehouse. This may degrade query

performance and needlessly increase storage space requirements.

Mistake: Ignoring external data and multimedia. Data warehouses often provide users with an architecture that permits a company to benchmark its performance against the performance of its peers.

Mistake: Expecting the data warehouse to remain static. Successful data warehouse projects never end. End-users continually evolve their requirements and expectations to ask for more data and analytical abilities to keep pace with constantly changing business needs.
Mistake: Focusing on only ad-hoc reporting and analysis. End-users, especially executives, may not have time to construct queries and run reports; providing those users with an email alert or a critical success factor should not be overlooked.

The SAS Data Warehouse Model

The SAS System provides a data warehouse model that addresses the entire scope of warehouse management, organization, and exploitation. It is a comprehensive model, supported by a broad array of products. Figure 1.3 shows the SAS Data Warehouse Model. Following, the major sections on the model explain and discuss their fit into the comprehensive approach needed for successful warehouse implementation.

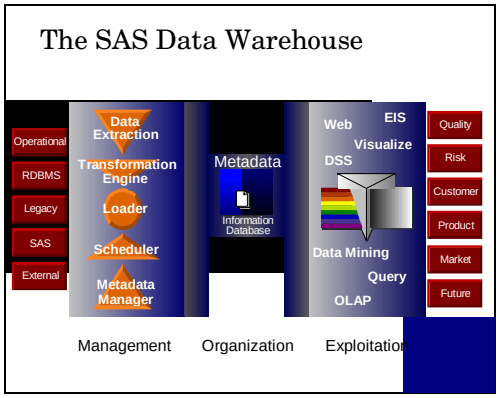

Figure 1.3

Warehouse Management

The warehouse Management process involves creation and administration of a data warehouse. The Management process consists of several sub-processes supported by and implemented with the tools and products available through the SAS System. These sub-processes include: data extraction, data transformation, data loading, job scheduling, and metadata management. The data extraction process is performed with the SAS/ACCESS Multiple Engine Architecture with SQL pass-through for relational legacy transactional systems, and Base SAS for others.
Data transformation is supported through the integration, validation, cleaning/scrubbing, reduction, and date/time variance processes on the data. The data loader permits full transfer or changed data-only transfer from legacy systems to the data warehouse. User-written processes are also supported. The job scheduler utilizes built-in tools or interfaces to third-party products. The Metadata Manager permits metadata to be built from SAS or external dictionaries.

Warehouse Organization

At the heart of the SAS Data Warehouse Model is the Warehouse Organization. It utilizes the SAS database engine, repository, and the accompanying metadata system. This system is ODBC compliant, and supports file creation, management, population, backup, data inventory services, query processing, and update processing. It also supports a relational data model to provide abstracting of data elements independent of application logic. The Management and Organization portions of the model are now manageable through a single interactive management interface in a new product called the SAS/Warehouse Administrator. The SAS/Warehouse Administrator provides data warehouse support staff with a consolidated front-end interface from which to manage, schedule, and maintain the extracting, cleansing, and loading programs for the data warehouse. It also provides metadata management facilities, for expanded metadata beyond what is found in the traditional CONTENTS PROCEDURE, or SAS dictionary.

Warehouse Exploitation

The Warehouse Exploitation portion of the SAS Data Warehouse Model includes an array of extremely powerful and flexible reporting, analysis, and statistical tools. These tools include point and click interfaces to mine (SAS/EIS), visualize (SAS/INSIGHT), and present data (SAS/EIS, SAS/GRAPH). The also include OLAP tools for interactive multidimensional analysis, interactive query tools (SAS/EIS, SAS/ASSIST), and client/server development tools for custom interfaces (SAS/AF, SAS/FRAME,SAS/FSP). Geographic information system capabilities are provided through SAS/GIS, and SAS/IntrNet provides Web access to data, reports, and graphics. In addition, the Warehouse Exploitation toolset includes excellent statistical modules for analysis and modeling (SAS/STAT).

Staffing the Data Warehouse Project

The Data Warehouse Project, requires a skilled staff to ensure a successful implementation. This section addresses the required members of a data warehousing project and their roles and responsibilities within the project. Single individuals can fulfill the demands of multiple roles if the project is smaller in size and scope. In large multi-departmental organizations, we recommended that different individuals fulfill the responsibilities as necessary. The types of team members include executive, core project, business unit, and Information Technology (I.T.) infrastructure personnel.

Team Membership

Executive Members

The executive member(s) includes the executive sponsor(s), whose task it is to provide overall corporate support for the project, financially, politically, and organizationally.

Core Project Members

The core project members consist of several key people. These people may reside within the I.T. department, the business unit, or be outside consultants. Their overall role is to manage, and implement the data warehouse project using input, information, and help from the business unit and I.T. transactional members. The core project members include a project manager, project leads, data architect, systems analyst, and programmers. A project manager, must holistically manage the project, including staffing, resources, tasks, and scheduling. Greater project success has been experienced by using a business representative as the project manager. This is due to the closer focus on solving the business problems they tend to have. Project leads may be necessary to assist the project manager depending on the size and scope of the project. Data modeling must be done by an experienced data warehouse modeling member. A systems analyst may be required to assist the project manager and leads in coordinating and conducting interviews with business and I.T. members to assemble requirements, and assist with design issues. Systems and application development programmers will be necessary for coding of the system.

Business Unit Members

It is the function of the business unit members to provide technical and business matter knowledge to the project team. They are the primary knowledge workers. They will supply insights from both a business and technical perspective concerning the business data environment, its usage, and end-user requirements. Members of the team may also play dual roles on the core project team. They include business unit managers and analysts, and the I.T. staff assigned to the business unit - OLTP workers, managers, programmers and the DBA. The business unit managers and analysts will provide the end-user requirements for the data warehouse and will be heavily involved in determining its evaluation criteria, and testing it for acceptance. The business unit OLTP worker will provide transactional arena information necessary for the data transformation and warehouse loading. The business unit I.T. manager, programmer, and DBA will provide support to implement the programming and other activities necessary to ensure data delivery to the warehouse. Working with the core project team members and data architect, they will ensure a “clean” handoff of transformed transactional data to the data warehouse for the initial and recurring loads.

I.T. Operations Members

The I.T. Operations members of the team will provide the necessary hardware, software, and networking information to accomplish the initial warehouse assessment, as well as ongoing support for developing the hardware, software, and networking environment related to the data warehouse. Figure 1.4 summarizes the staffing roles for the data warehousing project.

Data Warehouse Project Staff Roles

Staff Member	Staff Role
Executive Sponsor	Corporate, financial, political, and organizational support
Project Manager	Manage staffing, resources, tasks, progress and schedule
Project Lead	Assist Manager for large projects
Data Architect	Construct subject-based models to fulfill requirements
Systems Analyst	Coordinate and conduct interviews, assemble requirements, design inputs.
Systems/Applications Programmers	Design and write all needed code
Bus. Unit Managers & Analysts	Create evaluation and success criteria, convey requirements, acceptance testing.
Bus. Unit I.T. Manager	Provide I.T. resources and coordination
Bus. Unit I.T. Programmer	Design and write & schedule extract programs, all other needed transactional programs
Business Unit DBA	Perform transactional database tasks to support extracts
Operations Hardware Engineer	Provide hardware configuration & support
Operations Software/Systems Engineer	Provide systems software programming & support
Operations Network Engineer/Manager	Provide LAN/WAN support

Figure 1.4

The SAS Rapid Data Warehouse Methodology

Professional Services Division has developed a Rapid Data Warehouse Methodology to ensure a disciplined, iterative, approach in the management and implementation of data warehousing projects. This methodology consists of five distinct phases, designed to enable successful business and technical implementation of the data warehouse. These five phases are the Assessment, Requirements, Implementation, Training, and Review phases.

Assessment Phase

The first phase of the SAS Rapid Data Warehouse Methodology is the Assessment Phase. In this phase it is determined whether a realistic need or opportunity to develop and implement a successful data warehouse exists. It is composed of two primary stages. The first stage is a Project Definition Stage, which defines key subjects or business areas that could benefit from a data warehouse. In this phase several critical success factors are evaluated. These include:

- Choosing the highest corporate sponsor for the project
- Choosing a high-level interdepartmental team from business and IT to carry out the project (see Staffing above)
- Gaining resource commitment for the project
- Determining project ownership (who is ultimately responsible & will direct)
- Determining evaluation criteria for the project
- Ensuring unity between team members
- Determining the first project module to implement and grow from.

In the first stage the core members work with the business unit Members to define the data warehouse subject(s) to be implemented, provide the evaluation criteria for success, and the business unit resources needed for the project. The second stage of the Assessment Phase involves performing an initial assessment of the

I.T. infrastructure, to validate the ability of the computing environment to accommodate the demands and requirements of the data warehouse. This is conducted by the core members in concert with the I.T. personnel.

The outcome of the Assessment Phase is a formal document recommending whether or not to proceed with the project.

Requirements Phase

Once the Assessment Phase is complete and a decision is made to build a data warehouse, the Requirements Phase begins. In the Requirements phase, the sources of data for input to the warehouse are identified, a logical model of the data warehouse designed, the data transformation process and information delivery needs are documented, a data refresh strategy is built, gaps between needs and I.T. constraints are identified, and a project schedule is completed. This is done in two stages. The first stage is a Requirements Gathering stage where information is collected about a business unit's information needs through in-depth interviews with the business unit members. The components of the I.T. environment that currently services that business unit are then examined through interviews with the I.T. members. The second stage is a Reconciliation stage where a gap analysis is performed between business user requirements and I.T. infrastructure capabilities. The outcome of the Requirements Phase is a comprehensive Requirements Definition Document containing the logical and physical models, gap analysis, infrastructure diagram from OLTP extraction through warehouse data usage, transformation requirements, warehouse refresh strategy, scheduling tools used, and warehouse construction timeline.

Implementation Phase

Once the Requirements have been set, a physical model is built to implement the logical model that has been designed. This is done in three stages that match the SAS Data Warehouse Model addressed above. The first stage is the Management stage, where transactional data are extracted, and transformed into the data that will load the warehouse. The transformation process is composed of several sub-processes: validation, scrubbing, and integration. In the validation process, invalid, missing, out-of-range, and duplicated data are identified. The scrubbing process corrects the data problems identified in the validation process by re-coding or removal. The integration process brings data variables to consistent meanings, values, and measurements. The scheduling interfaces are then written, and all relevant metadata captured.

The second stage is the Organization stage. In this stage the data are actually loaded into the warehouse, the data structures are indexed, and views to them are created. Metadata is captured and loaded.

The last Implementation phase is the Exploitation stage, in which the graphical user interfaces are developed for the analyses, reports, graphs, and so on, that will be used for the data warehouse. At the conclusion of the Implementation phase the warehouse and its access mechanisms are built and ready for personnel to be trained in its use.

Training Phase

The Training phase follows implementation, and consists of two major activities. The first is to create a high-level training document covering the data warehouse and the applications used to exploit it. The second activity is to provide training to warehouse users and the warehouse administrator (if not already part of the project team).

Review Phase

Once the Training Phase is complete and the system has been turned over to the end-users for production operation, an evaluation of the project's success or failure is necessary to quantify its impact on the organization. These findings should be documented for future reference in expanding or building other data warehouses.

Summary

Successful data warehouse implementation with the SAS System requires appropriate knowledge of data warehousing concepts, project management, modeling, methodologies, and SAS System usage. The Professional Services Division of SAS Institute Inc. has developed a comprehensive approach to successfully implement effective data

warehouses for its clients. This approach involves the use of the SAS Data Warehouse Model, which includes the abilities of the SAS System, and utilizes the SAS Rapid Data Warehouse Methodology. The SAS Rapid Data Warehouse Methodology provides a structured, iterative approach for successfully managing and implementing data warehouse projects. Usage of this disciplined methodology will help practitioners to achieve data warehousing success.

References

Inmon, W.H., *Building the Data Warehouse*, New York: Wiley Computer Publishing, 1996, Second Edition

Kimball, Ralph, *The Data Warehouse Toolkit*, New York: Wiley Computer Publishing, 1996

The OLAP Council

International Data Corporation, Graham, Stephen, "The Foundations of Wisdom: A Study of the Financial Impact on Data Warehousing", 1996

The Data Warehousing Institute
Hill, Neal, Cognos, Inc., "The Warehouse Is Not the Answer",
Zimmer, Harry, CTO, AT&T Global Information Systems, "Data Warehousing: Are You On a Path to Success or Failure?"

Performance Development Corporation, Shah, Arvind, D. & Milstein, Barney, M, "Data Warehousing: Practical Tips for Successful Implementation", 1996

SAS Institute Inc., "Data Warehousing, A Blueprint For Success", Cary, NC: SAS Institute Inc., Available through SAS Institute's Sales and Marketing Division.

SAS Institute Inc., "A SAS Institute White Paper: The SAS® Data Warehouse", Cary, NC: SAS Institute Inc., Available through SAS Institute's Sales and Marketing Division

SAS Institute Inc., "A SAS Institute White Paper: Rapid Warehousing Methodology", Cary, NC: SAS Institute Inc., Available through SAS Institute's Sales and Marketing Division

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Acknowledgments

The author wishes to acknowledge Kevin Simmons, Martha Johnson, Becky Webb, and John Kohl, from SAS Institute Inc., for their contributions to, and editorial support of this paper.

For Additional Information:

Tony Brown
Consulting Services Division
SAS Institute Inc., Dallas, TX
214/977-3916
SASABR@unx.sas.com

---

[1] Inmon, W.H., *Building the Data Warehouse*, New York: Wiley Computer Publishing, 1996, Second Edition, p33.

[2] The OLAP Council, Forsman, Sarah, Symmetry Corp., "OLAP Council White Paper", WWW.olapcouncil.org/whtpap.html