

Data Warehousing and Data Marts:

The Who, What, Where & Why of Accessing Critical Data

SAS Institute, Inc.

Abstract

Users are relying on computers to help them become more self-sufficient in getting their job done. They want to extract information from the company's enterprise systems with minimum intervention from the Information Systems (IS) staff. The solution to empowering these users and eliminating the impact on your enterprise databases is to provide them with access to a separate data repository of information to use for business intelligence purposes. This paper examines the role of the SAS® System, specifically SAS/ACCESS® software and SAS/CONNECT® software, in building a data warehouse of decision support information in a client/server environment.

Introduction

The balance of computer usage has shifted from data processing professionals to end users who want to be empowered to fulfill their own information needs. Companies are aligning IS goals with business goals in attempt to reduce the number of user requests to IS. This shift introduces challenges to IS to serve a different role in providing users with access to enterprise data. Their new role is to manage the enterprise resources and empower users with more effective decision support tools.

There are primarily two types of applications that contain critical information about an organization: operational and decision support. We distinguish between the two based on the primary use and audience addressed by the application. This paper explores the differences between these two application types and examines the problems associated with users needing to access them. These problems have caused organizations to move in the direction of establishing a data warehouse for use by the business users within the company. In fact, the Meta Group Inc., a consulting firm in Westport, Connecticut, conducted a survey on 250 I.S. professionals and found that in 1994,

"95% of companies surveyed plan to introduce or continue to use some kind of data warehouse in the next year."

The main goal of this paper is to describe the database features and client/server components of the SAS System that enable it to be a cost effective alternative to a commercial database management system as a data warehouse. We will achieve our objective by explaining how the SAS System can help you in each of the three phases of data warehousing:

- 1) Accessing operational data
- 2) Transforming operational data into decision support data
- 3) Populating the data repository for decision support

Information flows into a data warehouse from the operational environment where it is accessed by users of decision support applications. We depict this information flow in Figure 1 on the next page. As you can see in the diagram, the SAS System can help you access operational data, transform it into decision support data, store it in a warehouse, and develop reports and analyses.

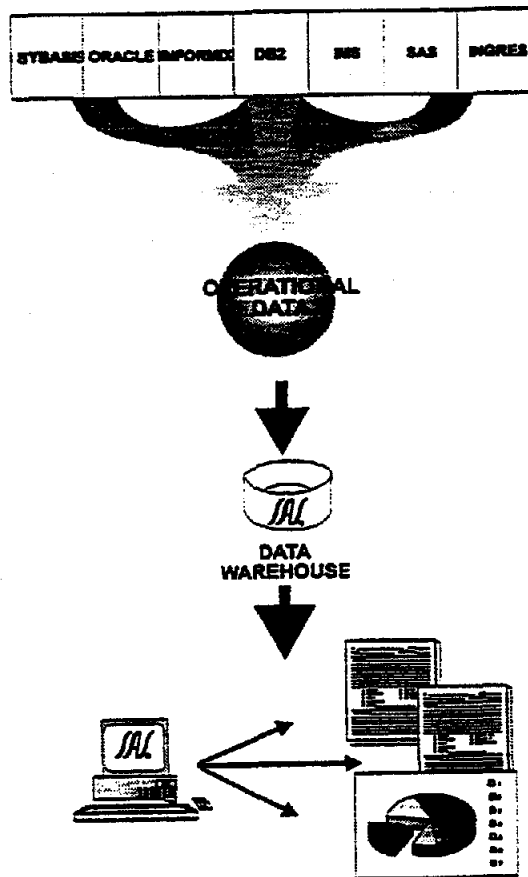


Figure 1 Data Warehouse Model

Statement of Problem

In implementing a successful data access strategy, it is important to recognize that there are appropriate and inappropriate ways to access data depending on the nature and distribution of the data and the types of applications requiring access to the data. Problems arise because of two facts:

- Data is both *operational* and *decision support* in nature
- Users need varying levels of access to data

Operational Data

Operational data often are stored in your company's production systems or other legacy databases or file structures usually written in COBOL. These systems are designed to handle the processing of large numbers of read/write and update transactions. For example, an insurance company might have operational applications for health, automobile, life and casualty claims. Many of these systems were developed over twenty years ago and the data are stored in non-relational form.

The information contained in operational systems is critical to management's decision-making process. Operational data is often difficult to obtain because these systems are organized by application, like automobile claims, instead of by subject, such as customer name or policy number. There is a lack of consideration taken in developing the systems for analysis and reporting purposes. User requests to IS for reports against these data are time-sensitive and often divert IS from other systems development efforts because of their interrupt-driven nature.

Performance of operational systems is effected by poorly-constructed or ill-timed queries generated by users who have direct access authority. Data downloaded to the desktop for analysis frequently clog already overtaxed networks. Timely access to data in these systems is lacking because the business analysts compete for resources with the people entering transactions.

Decision Support Data

Decision support data is used for business intelligence, query and reporting systems, and analytical applications, all designed to augment your organization's decision-making process. These applications typically require read-only access to large volumes of data derived from the operational environment.

These data are usually scattered throughout your company in many forms; such as relational database management systems (DBMS), spreadsheets and word-processing documents. This type of data is hard to access because it originates in many locations.

Types of Users

Data is as much of an issue in providing users with access to information as the people who request it. The users of business information can be categorized into three groups:

1) Management

Management and decision-makers of today want computers to present to them a single vision of their organization's success. An executive information system designed with drill-down reporting on the critical success factors of the company will serve a manager's or executive's needs. They must receive information in a timely manner so they can make effective business decisions.

2) Business Analysts

Business analysts need ad-hoc query and reporting tools to access summary and detailed information stored in operational systems. Those interested in dynamically exploring relationships between these data want multi-dimensional analysis tools that give them the power to view and subset the data any number of ways.

3) Programmer Analysts

The programmer analyst needs access to more detailed information and possibly on a more frequent basis than management. Programmer analysts are interested in analyzing trends in data and performing statistical analytical functions on samples of detail data.

Introduction to Data Warehousing

The problems with users needing various levels of access to all types of data, and IS wanting to protect the company's data assets, present sound business reasons for creating a data warehouse. A data warehouse is an intermediate repository of summary and detail data organized by subject rather than by application to facilitate easier reporting against production systems. Data warehousing is an evolutionary process, and as a recent article in *Computerworld* states,

"You can't buy a data warehouse, you have to build it."

Computerworld, February 6, 1995.

Characteristics of a Data Warehouse

A data warehouse is a strategy and a process for staging corporate data defined by the physical and logical separation of decision support data and the operational data from which they are derived. Data in a warehouse is characterized as:

- Subject-oriented
- Time-variant
- Non-volatile
- Integrated

Only the data needed for decision support processing are integrated from the operational environment. The data are organized by *subject* to facilitate easier reporting. For example, information about customers can be summarized into one

table in the warehouse. Inquiries pertaining to customer activity become much easier because the data is stored in one place. In the operational systems where the data originates, the customer information can be stored in many files. It might be organized by applications like health, life or automobile policyholders, making reporting on information stored in these files more difficult.

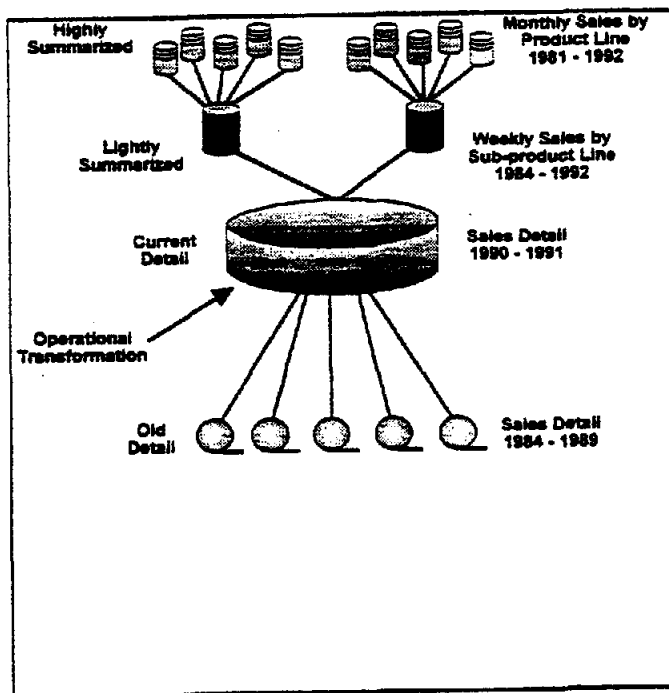
As data pass into the warehouse, they are *integrated* into a consistent structure for decision support. These data should have structure of varying levels of detail to support the types of reports needed by users. These levels of detail include:

- Highly summarized
- Lightly summarized
- Current detail
- Old or historical detail

In Figure 2, you see an example of how sales information might be stored in a data warehouse. By separating monthly, weekly and detailed sales data, users can quickly obtain *timely* and pertinent information needed for reports.

Figure 2 Levels of Detail in a Data Warehouse

The transformed data are placed directly into the warehouse at the current level of detail, where they are summarized, archived into the older detail data level, or purged. Data in the warehouse are *non-volatile* and usually not modified once they enter the warehouse.



Purpose Of A Data Warehouse

A data warehouse strategy benefits all three types of users identified in this paper: from management, to business users, to IS. The main reasons for creating a data warehouse are to:

- Prevent users from impacting production systems
- Alleviate users from having to learn data access methods
- Reduce interrupt-driven requests to IS
- Empower business users to become self-sufficient
- Summarize production data

A successful data warehouse strategy is dependent on the efficient use of existing infrastructure to effectively flow data in and out of the warehouse. As stated in *Datamation*,

"Concentrate on how information flows if you want to be successful with a data warehouse. Remember the point: to deliver information to make better business decisions amid rapidly changing business environments." *Datamation*, February 1, 1995.

If you are successful with the information flow in a warehouse, users will discover the efficiency of the warehouse as an integrated source of data that is easy to get to and that lends itself well to analysis of data over time. IS should be prepared for a gradual, but *dramatic* increase in the usage of a data warehouse for reporting purposes.

For the remainder of the paper, we will focus on the features of the SAS System that address each of the steps in building a data warehouse: accessing operational data, transforming it into decision support information, and populating the data repository.

Accessing Operational Data

The first step in creating a data warehouse strategy involves extracting data from your operational environment. With the SAS System's unique data access strategy, operational data is accessible, regardless of its form, through a single access method. The SAS System even has capabilities to transfer data across different hardware platforms without the user knowing that the data they have requested resides on a server or host machine.

Data Access Strategy

The enabling technology behind our data access strategy is *Multiple Engine Architecture*. The SAS/ACCESS family of products provides engines to over 50 relational, hierarchical, and network database management systems, data gateways and standard API's such as ODBC, and external file formats including VSAM. The engines translate SAS read and write requests for data into the appropriate calls for each database management system and file structure. Engines surface data to the SAS System in two forms:

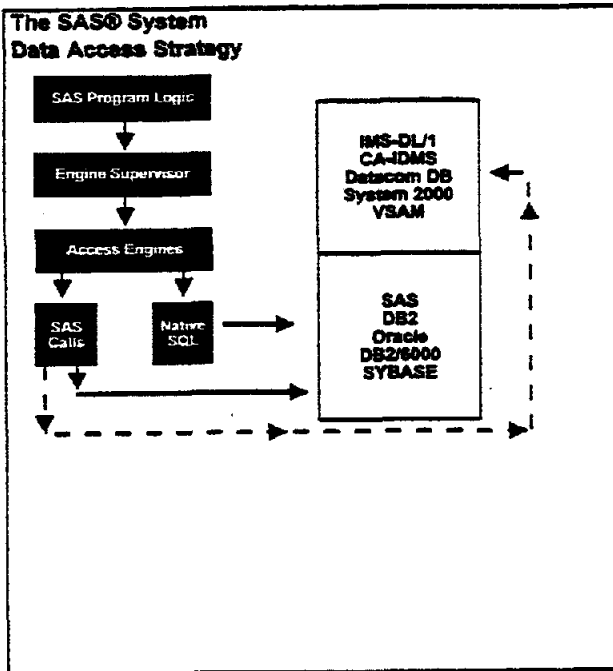
- 1) Logical views to the native data source
- 2) Extracts of native data into SAS data set form

The main benefit of our data access strategy is that SAS application logic is abstracted from data access and management functions. As a result, users are shielded from having to learn Structured Query Language (SQL) or other database specific query languages to perform simple extracts of information. With the SAS

System, users have a single technology for accessing all their data sources. The SAS System also supports it's own version of SQL. When the SAS internal SQL processor is unable to optimize queries for the DBMS, the user can pass native SQL statements to the DBMS.

An illustration of the SAS System's data access strategy is presented in Figure 3. The diagram clearly shows the abstraction of data access logic from the application or program logic. This detachment of data access logic eliminates the need for users to perform data access and management functions.

Figure 3 The SAS System's Data Access Strategy



Once operational data has been accessed using logical views or by extracting native data into SAS table form, it is available to the SAS System for processing. To the user, a SAS/ACCESS view looks like a SAS table, except that it does not contain actual data.

As part of a data warehousing strategy, it is more efficient to allow applications developers to build and use views to extract relevant operational data and stage it in the data warehouse. Features provided by SAS/ACCESS views for extracting and staging operational data to the SAS System include:

- Instructions and a path to the target data source
- Mappings from target resource names into SAS names
- Mappings from target data types into SAS data types

SAS/ACCESS views provide the user with the most current version of the data for their application since they act as a map to the data source. This reduces redundant data and requires a small amount of storage space. Views allow the user to subset and filter the original data elements to reduce the amount of data being accessed. Using the SQL native to the database, users can combine dissimilar data sources across hardware platforms.

Four methods exist within the SAS/ACCESS family of products for accessing external data. Not all of the SAS/ACCESS engines have full functionality as described below.

- 1) Access and View Descriptors - Access descriptor holds database connection and data description information about the structure and attributes of the DBMS table. View descriptor defines some or all of the data described by one access descriptor
- 2) SQL Procedure Pass-Through facility - Passes SQL statements directly to the DBMS
- 3) DATA step interface - Embeds program logic into a view to pass to the DBMS
- 4) DBLOAD - Loads SAS data into a DBMS

The suite of SAS/ACCESS products provides transparent and dynamic access to a variety of data sources and operational database formats. When these data reside on different hardware platforms, SAS/CONNECT software provides tools for a seamless transfer of the data between machines.

Client/Server Services

The SAS System provides two sets of services to enable a client/server distribution of data and application logic across diverse hardware platforms:

- 1) Compute Services - Enable the utilization of remote computing resources, including hardware, software, and data, to execute an application
- 2) Remote data services - Enable access to data stored in a remote environment using data transfer services

The features of interest to us in accessing operational data across hardware platforms are the remote data services in SAS/CONNECT software. These services allow the SAS System to assume the entire role of accessing operational data for staging it in the warehouse.

Functionality in SAS/ACCESS software allows you to access data in the operational environment. What if these data reside on multiple hardware platforms, as they usually do? You can use the data transfer facilities in SAS/CONNECT software to move data from one machine to another. The client/server computing services of SAS/CONNECT software are highlighted in the next section on transforming operational data.

Transforming Operational Data

Only the data needed for decision support processing are carried over from the operational environment into the data warehouse. Some pre-processing must occur that summarizes, archives into older detail, or purges the data. Decision support applications are designed to present a broad view of what the data represent. They rarely need to display the detail level data, so it is not necessary to move all of the detail information from the operational setting. However, it is important to integrate the data into a consistent structure.

During this stage of the data warehousing process, it is critical for those who are responsible for data migration strategies to examine the activities that users are performing on the data they use today. According to Bill Inmon, a reknown consultant for data warehousing issues,

"95% or more of decision support processing can be done against lightly summarized data and only 5% or less needs to go against the archived detail data". Bill Inmon, "Building the Data Warehouse." 1992.

It makes sense then to perform some data management logic on the detail data prior to placing it into the data warehouse repository. This may include activities such as:

- Cleaning
- Reconciliation
- De-normalization
- Summarization
- Loading

The SAS System provides the following tools for accomplishing these data management functions:

- PROC SUMMARY - Grouping by classification values
- PROC MEANS -Collapsing numeric data
- PROC FREQ - One-way, two-way and n-way classifications
- SQL - Groupby, orderby, and summary functions
- Data step programming - Arithmetic, random number and trig functions
- Open multiple input/output files simultaneously
- Look-ahead reads and table look-up logic

Client/Server Data Management

Data management functions can be performed in the operational environment or on the machine housing the data warehouse. The remote compute services of SAS/CONNECT software can help in distributing data management tasks in a client/server environment.

With SAS/CONNECT software, program logic can be submitted to the hardware environment most appropriate for performing data management tasks on your operational data. These remote compute services allow you to take advantage of the power of your hardware resources by distributing *program and application logic* across your organization's enterprise systems.

Metadata

After applying data management logic to the operational data to transform it into decision support data, you should consider developing a data dictionary; and more explicitly metadata. Metadata is information that describes how different structures and variables in a data warehouse relate to each other. A data dictionary is a subset of metadata and serves as an organized listing of data and associated descriptions.

Both metadata and the data dictionary will be developed by IS to make the views of data in the warehouse more meaningful to the user who does not understand database table name logic. These tools will become an invaluable reference tool for more advanced analysts who want to learn about the database and write queries that take advantage of existing tables, views and indexes.

The SAS System provides several methods for creating metadata and a data dictionary:

- 1) SAS® macro facility - To customize the SAS System for building metadata
- 2) SQL procedure - To generate dictionary tables for a data dictionary
- 3) CONTENTS procedure
- 4) DATASETS procedure

SAS Institute is working to improve the methods used for storing and retrieving metadata using a data dictionary in the SAS System.

Populating the Data Repository

Once data has been integrated into a consistent structure, it is then ready to be placed directly into a data repository. Organizations face an important decision in choosing their data repository for the warehouse. Many are conditioned to thinking that a DBMS is necessary for the data repository because of the essential features it provides for storing many types of data.

The intent of this paper is to suggest that it is not necessary to use a commercial DBMS as the data repository for a data warehouse. One SAS Institute customer, Diane Brown at Athena of North America, says they chose SAS over a DBMS as their data warehouse repository because,

"We didn't need rollback and recovery and other things that DBMS provide because we are getting our information from systems that have already done that."
Diane Brown, Director of Clinical Systems, Athena
Blue Cross Blue Shield

SAS Tables

Specific characteristics of SAS tables reinforce the SAS System's ability to serve as the data repository for the data warehouse. These features include:

- Relational SAS tables - Requests for information can be completed more quickly because fixed relationships between data are determined dynamically upon request

- Data set indexing - Processing time is reduced by aiding users in locating and selecting observations based on the value of a particular column
- WHERE-clause processing - Users receive only relevant data by specifying a set of criteria for select rows
- Compressed record format, or variable-length record - Maximizes query and storage resources by requiring less mass storage and fewer I/O operations
- Lean file format - Data can be stored on disk with low overhead

Multi-Vendor Architecture

Portability is an important issue when choosing the appropriate hardware platform for the data repository. With the Multi-Vendor Architecture (MVA) of the SAS System, companies can choose from a mix of platforms and know that the SAS System will perform. MVA assures that applications will run the same across all host, server and client machines supported by the SAS System.

Open Architecture

The SAS System has provided an open architecture for accessing data since the introduction of engines in Version 6 of the SAS System in the 1980's. These engines can be used interactively or in batch mode to communicate with other applications and data sources across a wide range of hardware environments. SAS Institute is continuously adding support for new engines.

In 1995, we added an engine to support Microsoft Corporation's Open Database Connectivity (ODBC) standard. The SAS/ACCESS to ODBC facility enables users to access third-party ODBC-compliant data sources from within the SAS environment.

The strengths of the SAS System's open data access strategy for decision support reporting against a data warehouse include:

- 1) Access to only pertinent information
- 2) No impact on production systems
- 3) View to the most current data

Many companies expect that a repository of decision support information should be open to third-party applications. With our client support for ODBC (Introduced in January, 1995), a repository of SAS data can be accessed or updated by any ODBC-compliant third-party application using the SAS ODBC Driver.

For the first time in the history of SAS Institute, users are now able to access SAS System data from machines that are not licensed for SAS software. When the SAS data repository resides on a server platform, through the use of the SAS/SHARE*NET® facility, the SAS data can be accessed or updated by any ODBC-compliant application.

Decision Support Tools

A data warehouse is not complete without an effective set of decision support tools to report against the data. This step can be one of the most difficult parts of developing a data warehouse strategy. However, if you are considering the SAS system for your data repository, then you are probably familiar with the rich set of decision support tools available in the SAS System of software.

The SAS System provides tools for four major application areas:

- Business intelligence and On-line analytical processing (OLAP)

- Applications development
- Technical and scientific analysis
- Horizontal and vertical business solutions

More specifically, the SAS System provides an integrated suite of business intelligence and decision support tools for scheduling projects, controlling quality, analyzing data, building applications, presenting information, and querying databases. SAS Institute is in the business of delivering information across your enterprise.

Conclusion

The concept of the information database has existed for over 10 years; almost as long as our customers have been using the SAS System for data warehousing. However, the practice of implementing a data warehouse in a production environment is just beginning in most companies. Data warehousing may be a required step in incorporating client/server computing into your enterprise, because it makes effective use of your company's critical legacy systems as informational resources.

This paper is intended to make both users and decision-makers aware of the capabilities of the SAS System for both accessing data to populate a data warehouse and building applications for decision support purposes. The staff at SAS Institute is well-positioned to assist you as you begin the task of re-designing decision support systems to take advantage of data warehouses of your company's enterprise data. We understand that your ultimate goal is to deliver the right information to the right people in your organization in a timely manner. After all, the SAS System is the world's leading information delivery system.

References

Betancourt, Randy (1993), "The SAS System as an Information Database in a Client/Server Environment," Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990), *SAS® Language: Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1994), "The Data Warehouse: Storing Decision Support," *SAS Communications, Second Quarter*, Cary, NC: SAS Institute Inc.

Base SAS, SAS, SAS/ACCESS, SAS/AF, SAS/ASSIST, SAS/CONNECT, SAS/SHARE*NET are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

IBM, AIX, DB2 and DB2/6000 are registered trademarks of International Business Machines Corporation.

Other brand and product names are registered trademarks or trademarks of their respective companies.

SAS and The Data Warehouse - Panel Discussion

Susanne Daood - Mayo Foundation; Rochester, MN

Priscilla Van Grevenhof - Mayo Foundation; Rochester, MN

Larry Taub - Independent Consultant; Ely, MN

Randy Van Beek - First Data Card Services; Omaha, NE

Data warehousing is the current solution for grouping corporate data stored on disparate platforms in disparate databases to aid in making accurate and efficient business decisions. This panel has a wide range of experience and expertise using SAS tools to access, store, manipulate, analyze and display data. Aspects of the data warehousing process familiar to them will be discussed.

These topics include (but are not limited to) metadata, using a data warehouse for business decisions and marketing, tools for data grouping and reporting, issues of currency, changes and size in a data warehouse. Audience participation is encouraged.