EMPOWER YOUR USER COMMUNITY WITH A SAS INFORMATION WAREHOUSE
Stephen C. Scott, Scott Consulting Services, Inc.

ABSTRACT:
The idea of an Information Warehouse composed of SAS datasets and present on your platforms of choice offers many benefits to end-users, application developers and system administrators.
Performance benefits include CPU savings of 65% to 99% over conventional DBMS such as DB2 and elapsed time savings of 90% compared to DBMS products from Oracle and Sybase.
Whether your Information System is small (3 meg) or large (30+ gigabytes) you can achieve improved performance and improved business processes through the concepts of an Information Warehouse.
This paper explores the author's experiences with Information Warehouses, summarizes the most recent writings on the Data Warehouse and reviews some of the corporations using the SAS system as an integral part of their Information Warehouse strategy.

INTRODUCTION:
The ideas that are the foundation of the Information Warehouse certainly have existed for thousands of years. Long before there was writing, data from the world was arranged into information that could be presented, communicated, and passed on through oral tradition, the arts and teaching by example: legends and myths, arts from basket making to painting, dance, drama, music, practical trades such as building and agriculture, traditions of the seasons and of the society, all were ways of communicating information long before there was writing.
With writing and books came the concept of libraries where information could be stored, shared, and researched to develop new information. Long before there were computers there were great library systems. Within a University, for example, there could be a central library that contained most materials with a cataloging method that allowed research into subject areas. Departments could also have their own specialized collections as could an individual student or researcher. When an individual was researching or writing they would draw upon the central library, the specialized library and their own notes taken from their observations.
Yes we have a new tool... but the concepts are as old as humans themselves.

History of decision support systems
1960 - 1970 Storage on magnetic tape and punch cards. Master files and reports, proliferation of files, sequential access, long development time for reporting. Mostly Operational data. Reporting MIS department with COBOL.
1970 - 1980 DASD storage begins to allow random access to data. Still mostly operational data. On-line transaction processing and databases are developed. Reporting: MIS dept. with Cobol, begin development of user tools.
1990 - 1995 Information Warehouse as separate from the operational DBMS. Decision Support Systems as important as the Operational DBMS reporting. Multi-platform access to data: mainframe - midrange - workstations - PCs into one transparent system.

Why the Information Warehouse?
Operational data can be stored in application specific databases with OLTP. A corporation could have data stored in DB2, IMS, ADABAS, etc., on their mainframe. They may have other applications with servers where data is in Sybase, Oracle, or Informix. This data can continue to evolve - it is the original source of the warehouse.
A Data Warehouse would gather this data into a common model. The data would be stored separately from the operational data so that accessing it would not impact the other. The arrangement of the data and its indexing will be specifically for the needs of the business users.

The Information Warehouse may be distributed to several computer platforms - there may be a central warehouse on a mainframe and specialized departmental warehouses on RISC servers and local networks. There may even be individual warehouses on an analyst's PC or workstation.

**Designing the Information Warehouse**

The corporate data model is the basis for the design of the Data Warehouse model. Essentially, purely operational data is removed. The element of time is introduced so it becomes historical in nature. Stability factors are added to the design so that data that frequently changes is separate from data that does not change. Furthermore, data needs to be denormalized - the goal is to decrease I/O - so that tables that would have to be frequently joined to become one table. In the warehouse, redundant data is a positive thing and may need to be introduced in order to improve performance.

**When is 50% Correct ‘Excellent’?**

In designing a Data Warehouse, Bill Inmon, who is considered 'the father of the Data Warehouse', has written in his book "Building the Date Warehouse" that "if 50% of the first iteration of design is correct, the design effort has been a success". That is why RAD and JAD are both part of Data Warehouse building. RAD - in that instead of having a long traditional design effort you have a Rapid Application Development environment, where subsets of the warehouse are designed and evaluated within a JAD session, where the Joint Application Development concept of the experienced users and designers results in the production of Prototypes, where parts of the prototype are always saved since they have passed the acceptance of the group of designers and users. Only as the users can see and experience what can be created and have a chance to use the information can they decide if it is what they need and in the best form for their use.

**Think About It! Do It Now!**

If you have the organizational and political backing to form a corporate Information Warehouse, or are representing a smaller group of users, you can begin the process in small steps.

Small Scale: If you have programs that read and summarize extracts or databases routinely, consider doing a summarization into a SAS data set, for example, monthly. Once the data is in the SAS data set it is easy to use the full range of SAS Information Delivery and Presentation tools with it. By making the table historical, for example: "Sales Summary by Product and Region - April, 1995" and storing the table as a permanent SAS data set, then next month you will be able to do comparisons, and by next year it will be easy to do graphics and EIS applications with your tables.

David Vaskevitch writes in his book 'Client/Server Strategies' that there have developed three layers:

1. Transactionally oriented databases for operational data
2. Analytically oriented databases - transformed data stored separately - the central Information Warehouse concept.
3. Individual users tap into the analytically formed database and form other databases to meet their team and personal needs.

It's worth the time up-front that you spend. By analyzing a large amount of data and saving the results you have begun an Information Warehouse. You have made what you consider 'information' out of what was before operational data on sales.

---

Levels of Granularity

Granularity refers to the level of detail data - the more detail the lower the level of granularity. Ideally we would have access to all data - from the individual archived transactions to the daily, weekly, monthly and yearly summarizations. Practically we probably need to design our lowest level of access first. What is the finest level of detail we need? What is the finest level of detail that we can practically afford to access and store? Once we have decided upon this as our foundation we can then create higher levels of granularity that will allow drill down in an EIS system to these lower levels.

Storage of levels: The lowest level of granularity could be stored on a mainframe or large server. The mid-level of summarization could be on a departmental server. Or all levels of granularity for a specific business unit could be stored on a server - which is refreshed from a central warehouse.

Partitioning Data

This is the arrangement of data in smaller physical units. This allows easier management of the data. Examples of partitioning would be having tables separated by:

1. time periods
2. geographic locations
3. lines of business
4. organizational units

OLAP and EIS and the Data Warehouse

By having your data structured in a Data Warehouse it is much easier to develop Online Analytical Processing and Executive Information Systems. Specific applications accessing the Data Warehouse can be designed with SAS/SCL, SAS/AF, SAS/GIS, SAS/GRAPH. SAS/ASSIST software can give on-line ad hoc query and reporting capability to the warehouse. SAS/EIS can use the warehouse as its source for drill downs, problem monitoring, and trend analysis needs.

SAS Data Sets versus other DBMS for the Information Warehouse

While it is certainly acceptable and possible for you to build your Information Warehouse as relational tables stored in DB2, Oracles tables, Sybase tables, etc. - and access them with the full range of the SAS Information delivery system - I would recommend the consideration of building all or some of the Information Warehouse as SAS data sets.

Why? The Information Warehouse is a specialized environment - it is a read only environment where analysis is done. The SAS system with its summarization features can easily build the different levels of granularity and the partitioned data that you need. Once it is built why not leave it as SAS data sets? They have all the relational capabilities of the other products and are optimized to work best with the SAS system. They can be ported to any platform with ease and accessed. They can be accessed from any platform.

William Clifford, in his paper "Is the SAS System a Database Management System?" lists five benefits for storing data within the SAS system:

1. Faster Access to Data: The SAS system is optimized to deliver data to its procedures from SAS data sets.
2. Cost-effectiveness: No additional cost for a DBMS to store the data.
3. Reduction in number of vendors: By using the SAS system for both data analysis and data storage you provide a single source.
4. Product consistency across many platforms: The multi-vendor architecture of the SAS system provides a portable application environment independent of the host computer system.
5. Ease of transporting data to non-SAS applications: The SAS system's capabilities are extensive - with access to other DBMS systems and all types of file systems.

**SAS Information Warehouses - PART I - My Own Experiences**

In my papers published in SUGI 18 and SUGI 19 I present the stories of three corporations where SAS data sets were compared to DB2 tables for Information Warehouses at the global and local levels.

**Corp. A:** A major financial services company developed an 'Information Warehouse' for the purpose of portfolio analysis, marketing research, and risk management. The corporation's development methodology used was traditional:

A. They compare DBMS packages for the expected size of their system. From the two candidates of Oracle or DB2 they found best performance with DB2 for this very large system.

B. They select the SAS System for their corporate analytical tool.

C. They then begin the detailed research and building of the Information Warehouse.

1. Analysts from MIS researched all existing databases and their variables. From this list of all possible data elements and potential derived information a master list was formed.

2. Through interviewing the users of the new system throughout the corporation, the list was refined to contain the elements that they were most interested in.

3. By interviewing the potential users on how they would use the system and what type of queries they would do, different possible table structures for the elements could be evaluated.

4. Data was normalized and denormalized until table structures were formed that would allow most queries and reporting to come from 1 to 3 tables - thus minimizing the impact of joining.

5. The final database structure for the Warehouse is then submitted for approval to the users and technical staff.

6. Extracts are written to retrieve the data from the enterprise databases. Load programs are written to place the data into the relational tables.

7. For testing, certain data representing 10% of the corporation is selected to be run through the extracts and loaded into the database.

8. When this is completed the next phase begins: testing of the DBMS and querying the Data Warehouse.

**TROUBLE !!!!**

It is at this stage that the first sign of real problems appears. In phase 1 of stress testing of the system the front end that consultants built crashes - furthermore there is excessive CPU time seen by capacity management on the few queries that did execute.

---


SAS System Saves The Day
At this point I am called in to evaluate the problem and make recommendations.
1. The front end had to go - it was absurd - designed perhaps for a 1000 record system, not a major Data Warehouse with millions of records. A front end could be built in the future.
2. DB2 tables were compared with the same data in SAS data sets. To my surprise, there was a 90% drop in CPU time accessing SAS data sets.
3. We change the tables to SAS datasets. Stress testing is passed with ease. The system is approved for production. The schedule is met successfully. The entire database is loaded.
4. The production system goes to user acceptance. The users accept the system.

IN SUMMARY:
Corporation A: 30 gigabyte system with 5 years of data. Decision Support System accessed by a user community of 20 - 80 analysts. Originally designed with DB2 tables, the system was projected to take 100% of the CPU of a IBM/MVS type mainframe. By moving the tables to SAS data sets instead, the CPU usage was projected at 9% of the system - a savings of 91% of the CPU! Without SAS data sets this Data Warehouse could not have gone into production.
If you are interested in more information regarding this Data Warehouse please see my papers or write to me requesting a copy.

Review of Corporations Using SAS Data Sets in Information Systems and Warehouses.

Dun & Bradstreet Information Services
"We didn't need all the overhead of a database like SYBASE for our Data Warehouse. The SAS System had all we needed." Karen Grippo, DBIS
"Data warehousing, client/server, and the SAS System have all contributed to D & B's revenue...Communication is the real business problem we're addressing, and we couldn't have solved this problem without the SAS system." Dave Herman, DBIS.

Type: Unix server, PC
The world's largest commercial purveyor of information implemented a Data Warehouse in a client/server environment using the SAS system. Their first step was moving a 20 gigabyte SAS system-based Information Center database to a UNIX server from a mainframe.
In benchmarking they discovered that SYBASE and Informix would take 3 to 7 days, or 24 to 56 hours, to load the database while the SAS system took 8 hours to load and index the database. They decided on the SAS system for their Data Warehouse instead of SYBASE - which had overhead they did not need.
Later the Portfolio Applications Marketing Group integrated their data into the Warehouse. They use customized macros to request data from the server that is then brought to their workstations from the Data Warehouse. On their workstations they perform analysis and drill-downs since they have both summarized and detailed data available.

Auto/Oil Air Quality Improvement Research Program (AQIRP)

“We started from day one with the assumption that it would be a SAS database.” Jack Benson, General Motors

Type: Mainframe, Unix, OS/2, ... any configuration

A consortium of fourteen U.S. oil companies and three domestic automakers who funded a $40 million project to provide a public database on the interaction of fuels, vehicle emissions and air quality.

From day one of the project SAS was selected as the database - considering the amount of data and the need to make the data available to scientists around the country, the SAS system provided the best approach. The resulting research database in SAS data sets is kept at the Coordinating Research Council in Atlanta.

EDS receives data from emissions sites which is loaded onto a mainframe and run through a SAS application that has a SAS transport file as a final output. The transport files are sent to AS where they are loaded onto SAS on SPARK workstations.

At SAI, the results of data manipulation and analysis are included with the data set so that future workers do not have to repeat the work. The resulting SAS data sets are made into transport files and made available to all consortium members and as a public release database.

This ‘Data Warehouse’ comprises over 50% of the information being used by the EPA to develop its models for reformulated gas and is also being used by Europe.

Thos. Y. Pickett & Co.

“Our intentions are to build data sets where all information is stored and processed within the SAS system” Truett Phillips

Type: PC-LAN based

In 1993 TYP was contracted by the state of Colorado to perform a statewide property assessment study. The company licensed the SAS system for the analysis portion of the study. Their search for a product for data management, however, found that all systems tested failed to meet their strict criteria. Truett Phillips, operations coordinator said: “We had no idea that the SAS System could be used as a DBMS, and we certainly didn't think of it as an application design tool”. However, they discovered both of these attributes and standardized on the SAS system.

For the study over 5 million records were received from the counties of Colorado in a variety of formats. SAS is used to convert these and load them into SAS data sets. The SAS/AF software was used to create front ends for the data where users can drill down through graphical interfaces to detail levels and where they are given full reporting capability.

Petro Canada, Calgary

Type: OS/2 hosts, Windows clients

In 1994, the second largest petroleum company in Canada, employing 6000, implemented a distributed SAS information system with regional OS/2 hosts communicating with Windows based clients on LANS. The client/server Performance System contains data in SAS data sets that have been gathered from mainframe SAS applications providing financial indicators, DB2 databases providing sales information, and external data such as market research. The system contains about 20% of the existing data since it provides 80% of the usage - this data is placed on the regional servers.

The users of the system have a GUI front end to the SAS data sets built with SAS/AF and Frame entries. Through this they can use the SQL Query Window to do ad hoc queries. They can also do drill downs through standard reports that transparently access remote data.

---

8 Clearing the Air: A ground breaking air quality study uses the SAS System to overcome information barriers, SAS Communications, Volume XX, No. 4, Fourth Quarter, 1994, pp. 19-21.
LTV Steel, Cleveland, OH

“We have moved much of our data from COBOL systems into SAS data sets so all these people can use the SAS information database as a tool to get their information into a written format, without being captive to the Information Systems Group.” — Robert Scharl, senior statistician in Quality Assurance at LTV.

Type: Mainframe

SAS/ASSIST software and the SQL Query window are used for access, analysis and reporting of manufacturing, production, and process control information. Users are corporate types, engineers, quality technicians, mill foremen, etc. By using the SQL Query window to merge large datasets they have achieved a 50% decrease in turnaround time compared to using two sorts and a merge.

National Biological Survey, Ann Arbor, MI

“We believe that SQL processing in SAS is as powerful as the other application (Oracle) of the SQL language we use.”

“The prospect of fully performing both data management and data expression within SAS was attractive.” — Mary Fabrizio and Scott Nelson

Type: OS/2

A data set formed over 24 years containing information on lake trout in Lake Superior was converted to relational SAS data sets and accessed with Proc SQL on version 6.06 of SAS for OS/2. They found Proc SQL valuable for solving data expression problems and for data management of SAS data sets.

Abbey Life Assurance Company Limited, United Kingdom

“It was concluded that an MIS environment with data held in DB2 and accessed via SAS would be viable, but so would one with data held in SAS datasets. It was also confirmed that for read only purposes SAS datasets offered similar relational database advantages and accessing them generally consumed significantly less CPU time.” — Robin H. Wilson

Type: Mainframe

Data relating to the life assurance and pensions of a million clients forms the basis for a Management Information Database composed of SAS data sets. The database provides a consistent and validated source for IT professionals and business users to query via SAS/Assist and Base SAS. Both batch reports and point-and-click interfaces to the data, are provided for a range of users.

Benefits of the information database include:

• MIS provides a source of data accepted as the ‘corporate picture’.
• Productivity gains between 5 and 10 times that of a 3GL language.
• Anyone authorized can access the data.
• Data-driven design allows changes easily.
• Empowers management to make better decisions through information previously not available.

10 Charting New Courses for Information Delivery, SAS Communications, Volume XX, No. 3, 1994, pp. 3-4.
11 Founding New Courses for Information Delivery, SAS Communications, Volume XX, No. 3, 1994, pp. 4-5.
12 Fabrizio, Mary and Nelson, Scott, Find the Fish: Using Proc SQL to Build a Relational Database for a Mark-Recapture Study, SUGI19 Proceedings, pp. 217-222.
Hungarian Ministry of Finance
Hungarian Central Statistical Office

"In the Hungarian statistical system micro-economic databases including financial and business accounting data of enterprises have an important role. The database was developed in HP-UNIX environment using SAS software as a database management system. The database includes data of about 100 thousand enterprises which is a nearly complete coverage of the Hungarian business sector." Marianne Rudas & Edit Majtenyi, HCSO

Data is stored in SAS data sets which are viewed and stored as relational tables. There are 600 basic and derived variables in 12 table types. There are four different levels of aggregation for each table – different users have access to different levels of aggregation.

Nomenclatures contain code values of classifications with their naming and are stored in another set of SAS dataset tables that can be used when reports with good textual information are needed.

Meta data: Information on stored data and nomenclatures are also stored in relational tables (SAS data sets). They assist in orientation through the database. Meta data helps users through a SAS user-friendly interface that helps them make requests of the database.

Finmatic, Milano, Italy

Type: Mainframe, PC-LAN

Finmatic is a Technological and Administrative Company offering services to the financial companies operating in the Italian RAS Group. A SAS database is on the mainframe (formed by accessing other databases nightly) with a synthesized SAS database on the server. PCs running a SAS/AF application access both server and mainframe SAS databases transparently depending on the level of detail requested by the user.

The Infocenter is an Executive Information System type application developed in SAS with a client-server architecture where PCs communicate through a LAN with a server, and the server communicates with a mainframe. The system provides graphical representations of data via the screen or through printing high quality graphics. Drill down analysis and user defined levels of aggregation are available by point and click - the user can store their profiles and use them again. They are guided by lists of analysis variables such as product lines.

Canadian Imperial Bank of Commerce, Toronto, Ontario

Type: Mainframe

The second largest bank in Canada (and one of the top 10 banks in North America) uses the SAS system and SAS datasets for their Data Warehouse. The Data Warehouse has been evolving for several years.

"The secret is to hold the data at the event level and summarize them to the level of granularity appropriate for specific queries. We realized that most queries were needed at a few specific levels of granularity. So we began to routinely summarize and hold data at that level. This substantially improved our response time and our operating costs." Boyd Carter, Superintendent of Productivity Management.

---

16 Data Warehousing at the Canadian Imperial Bank of Commerce, SAS Communications, Volume XX, No. 2, Second Quarter, 1994, p. 28.
IBM Microelectronics Division, Essex Junction, VT

Increasing requirements for complex analysis in monitoring reliability-related defects in a semiconductor process line led to the development of a structured SAS database for short and long term analytical usage. The result has been an impressive improvement in the outgoing quality of IBM's semiconductor products.

"The key to these improvements is the use of SAS databases for storing data in the appropriate format. Proper planning of the SASDB's enabled us to run all aspects of our analysis with outstanding results."17 Tracy L. Lord, Ken Wiggin

State Council of Higher Education for Virginia

A SAS Data Warehouse stored on an IBM RS-6000 provides ten years of valuable Virginia and National Educational data. Access to information is provided on the institutional, state and national levels. Data can be accessed via PCs over established state and national networks including the INTERNET.18

Bear Creek Corporation, Medford, OR

BCC is a mail-order specialty catalog company marketing fine fruits, confectioneries, roses and orchids. Operational databases include DB2, IMS, DATACOM, flat files, and SAS data sets. The Marketing Workbench project involves the creation of a Marketing Data Warehouse.

"We will integrate the marketing data we need from all the database products in the operational database into SAS data sets on UNIX workstations. From there, our clients will be able to easily perform what-if analysis or customer profiling against smaller sets of data that are summarized monthly."19 Bryan Jones

The users on Windows PCs access data through a SAS/AF front end and can bring data from the UNIX servers to the LAN. The goal is to develop front ends to empower users to do more sophisticated data access and analysis on their own.

Principle Financial Group, Des Moines, IA

 Ranked in the upper one percent of all life insurance companies in assets and premium income, the Principal Financial Group developed a SAS/EIS for regional officers where they can access over 2100 graphical displays to view sales trends and track the performance of their sales representatives. The basis of the system is an Information Warehouse composed of SAS datasets that is updated monthly with sales data downloaded from the IBM mainframe to the OS/2 SAS system. A SAS/AF front end system is automatically updated through the extended tables usage of SAS datasets. Through the PC network the regional officers access their own subsets of data.

"Preprocessing the data and storing them on the network reduces the amount of time it takes for the officers to generate a graph on demand and it takes up less storage space."20 Nancy Carley

---

17 Lord, Tracy and Wiggin, Ken, SAS Database Management System Used for Semiconductor Defect Learning, SUGI 19 Proceedings, pp. 231-233.
Group Health Cooperative, Seattle, WA

Group Health, the largest HMO in the northwest with $800 million in annual revenue providing services to 500,000 enrollees, created a Information Warehouse that "allow for data driven decision making to manage cost, utilization, and quality". The evolving system is composed of SAS datasets. The DSS is very successful. They have built SAS/AF and SAS/EIS front ends to the system, have a pool of business analysts trained in SAS to work with specific departments, and for less frequent users, have SAS/Assist software.

"It really changed how management works at Group Health."

ACKNOWLEDGEMENTS:
Base SAS, SAS/AF, SAS/SCL, SAS/GRAPH, SAS/EIS, SAS/ASSIST, are registered trademarks of SAS Institute Inc., Cary, NC, U.S.A. Oracle is a registered trademark of Oracle Corporation. IBM, DB2, OS/2 are registered trademarks of International Business Machines Corporation, Armonk, NY, U.S.A. Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT:
The author will be glad to answer questions and accept suggestions at the following address:

Stephen Scott
Phone: 608/276-7843
Scott Consulting Services, Inc.
346 Kent Lane #302
Madison, WI 53713