Paper # 125-30

Implementing Data Warehousing and Business Intelligence at McMaster University Using the SAS® Intelligence Value Chain

Debbie Weisensee, McMaster University, Hamilton, ON
Eric Matthews, McMaster University, Hamilton, ON
Anne McInnis, McMaster University, Hamilton, ON

ABSTRACT

This paper covers the incremental development of an enterprise-wide data warehouse at McMaster University, along with the deployment of end-user, web-based query and reporting tools using SAS® Value Chain Analytics. A goal of the initial data mart is to enable integrated reporting between the research funding and research financial databases. This integrated data is essential for managing, analyzing, reporting, assessing, and projecting performance at institutional, provincial, national and international levels.

This paper will review the process followed in designing the warehouse, based on enterprise-wide information needs. In developing the research data mart, we have built a foundation that is portable and uses standard industry protocol. This will enable us to be flexible and cope with changes in both hardware environments and business information requirements. Since the conceptual data model encompasses all of our current and anticipated information needs, and is flexible to meet future needs, it will help accelerate the pace of development for future data marts.

Using SAS® ETL Studio, we have mapped the logical data model into a research data mart. We have created a number of canned reports, using SAS® Enterprise Guide, which our stakeholders will be accessing as both static and interactive reports through SAS® Web Report Studio and the SAS® Add-In for Microsoft Office. Future stages include creating OLAP cubes for drill-down capability, and implementing SAS® Strategic Performance Management software (balanced scorecards to measure performance indicators).

INTRODUCTION

McMaster University's mission statement reflects the institution's commitment to communication, innovation, excellence and quality. To support these values, the university needs flexible and innovative information technology that will ensure the transfer and sharing of knowledge to inform and enhance all levels of decision making at the institution. Like many other organizations, McMaster University's transactional applications do not store data in easily accessible data models that can be easily transformed into comprehensive, meaningful information to support evidenced-based decision-making. Currently McMaster has several disparate systems on segregated platforms. They are transactional applications and do not store data in models that support on-demand and ad hoc aggregations. Even in other universities that use integrated systems, data warehouses are required to extract information for decision making. To rectify this situation, we're in the process of developing an enterprise-wide data warehouse to be deployed incrementally along with end-user, web-based query and reporting tools using the SAS® Intelligence Value Chain.

BACKGROUND

Successful data warehousing projects are designed to answer business questions posed by business users. Since data warehousing was new to the university, an external consultant was engaged to educate us about data warehousing best practices and guide us in developing a detailed requirements analysis for enterprise-wide information needs. As a result of that work, we have designed our conceptual data model to encompass all current and anticipated information needs across the enterprise.

Key stakeholders were interviewed in a series of focus group sessions and were asked to provide information on the following:

- Issues with current data/reports
- Data elements (candidate dimensions) required to support enterprise-wide information needs
- Report frequency and desired distribution method (electronic, web-based, paper)
- Amount of historical data required
- Data hierarchies
- Users of the information
- Security requirements
- Key performance measures
- Success criteria (to measure project's deliverables)

This information was summarized and used to develop enterprise-wide hierarchies and a decision support bus matrix. This detailed what information stakeholders wished to see, and how they wished to see it (for example, by time or geographic location). As part of the analysis, the model included an inventory of the underlying data sources required to support the information need, a data quality assessment of the source data and volumetric exercise (required for an initial sizing and expected growth of the data warehouse). What we found was that while the data was generally of a high quality, many process issues needed to be addressed, for example the establishment of enterprise-wide reporting standards. We expected to find that data volumes would be the major concern but instead found the complexity of the information needs to be a greater concern. Data from several disparate sources was generally required to fulfill the information requirement.

The following template was followed in evaluating the state of the data:

Information Needs Model – Color Codes

Informational Need	- Data exists and can be used as is, or with minor transformations - Identified as a priority, and data is in scope - Applicable for Data Warehouse - All view of the data is possible
Informational Need	- Data exists but requires more involved transformation - Some views related to data may not be possible due to integrity, consistency, data history problems or inaccuracy in the data - Identified as a priority - Applicable for Data Warehouse
Informational Need	- Data does not exist or cannot be used "as is" - Applicable for Data Warehouse - Some views related to data may not be possible due to integrity, consistency, data history problems or inaccuracy in the data
Informational Need	- Priority not applicable, out of scope

Based on this methodology, the following information needs model was derived for Research funding (grants/contracts applied for and awarded). Note that this is a graphical representation of the ease with which given information needs can be supported in the targeted environment (data mart).

In order to address the many process issues identified through the data assessment exercise, university-wide standards for data collection, reporting and access to information were developed and adopted. Data warehousing procedures implemented include the capture, organization, storage, maintenance and cleansing of data, and will allow for data validation and checks for completeness and accuracy.

IMPLEMENTING A DATA WAREHOUSE ARCHITECTURE AND EXTRACTION, TRANSFORMATION AND LOADING (ETL) PROCESS (Using SAS® ETL Studio)

This section describes how we are using extraction, transformation and loading (ETL) processes and a data warehouse architecture to build our enterprise-wide data warehouse in incremental project steps. Before an enterprise-wide data warehouse could be delivered, an integrated architecture and a companion implementation methodology needed to be adopted. A productive and flexible tool set was also required to support ETL processes and the data warehouse architecture in a production service environment. The resulting data warehouse architecture has the following four principal components:

- Data Sources
- Data Warehouses
- Data Marts
- Publication Services

ETL processing occurs between data sources and the data warehouse, between the data warehouse and data marts and may also be used within the data warehouse and data marts.

Data Sources

The university has a multitude of data sources residing in different Data Base Management System (DBMS) tables and non-DBMS data sets. To ensure that all relevant data source candidates were identified, a physical inventory and logical inventory was conducted. The compilation of these inventories ensures that we have an enterprise-wide view of the university data resource.

The physical inventory was comprised of a review of DBMS cataloged tables as well as data sets used by business processes. These data sets had been identified through developing the enterprise-wide information needs model.

The logical inventory was constructed from "brain-storming" sessions which focused on common key business terms which must be referenced when articulating the institution's vision and mission (strategic direction, goals, strategies, objectives and activities). Once the primary terms were identified, they were organized into directories such as "Project", "Location", "Academic Entity", "University Person", "Budget Envelope" etc. Relationships were identified by recognizing "natural linkages" within and among directories, and the "drill-downs" and "roll-ups" that were required to support "report by" and "report on" information hierarchies. This exercise allowed the directories to be sub-divided into hierarchies of business terms which were useful for presentation and validation purposes.

We called this important deliverable the *"Conceptual Data Model"* (CDM) and it was used as the consolidated conceptual (paper) view of all of the University's diverse data sources. The CDM was then subjected to a university-wide consultative process to solicit feedback and communicate to the university community that this model would be adopted by the Business Intelligence (BI) project as a governance model in managing the incremental development of its enterprise-wide data warehousing project.

Data Warehouse

This component of our data warehouse architecture (DWA) is used to supply quality data to the many different data marts in a flexible, consistent and cohesive manner. It is a *'landing zone'* for inbound data sources and an organizational and re-structuring area for implementing data, information and statistical modeling. This is where business rules which measure and enforce data quality standards for data collection in the source systems are tested and evaluated against appropriate data quality business rules/standards which are required to perform the data, information and statistical modeling described previously.

Inbound data that does not meet data warehouse data quality business rules is not loaded into the data warehouse (for example, if a hierarchy is incomplete). While it is desirable for rejected and corrected records to occur in the operational system, if this is not possible then start dates for when the data can begin to be collected into the data warehouse may need to be adjusted in order to accommodate necessary source systems data entry "re-work". Existing systems and procedures may need modification in order to permanently accommodate required data warehouse data quality measures. Severe situations may occur in which new data entry collection transactions or entire systems will need to be either built or acquired.

We have found that a powerful and flexible extraction, transformation and loading (ETL) process is to use Structured Query Language (SQL) views on host database management systems (DBMS) in conjunction with a good ETL tool such as SAS® ETL Studio. This tool enables you to perform the following tasks:

- The extraction of data from operational data stores
- The transformation of this data
- The loading of the extracted data into your data warehouse or data mart

When the data source is a "non-DBMS" data set it may be advantageous to pre-convert this into a SAS® data set to standardize data warehouse metadata definitions. Then it may be captured by SAS® ETL Studio and included in the data warehouse along with any DBMS source tables using consistent metadata terms. SAS® data sets, non-SAS® data sets, and any DBMS table will provide the SAS® ETL tool with all of the necessary metadata required to facilitate productive extraction, transformation and loading (ETL) work.

Having the ability to utilize standard structured query language (SQL) views on host DBMS systems and within SAS® is a great advantage for ETL processing. The views can serve as data quality filters without having to write any procedural code. The option exists to "*materialize"* these views on the host systems or leave them *"un-materialized"* on the hosts and *"materialize"* them on the target data structure defined in the SAS® ETL process. These choices may be applied differentially depending upon whether you are working with "current only" or "time series" data. Different deployment configurations may be chosen based upon performance issues or cost considerations. The flexibility of choosing different deployment options based upon these factors is a considerable advantage.

Data Marts

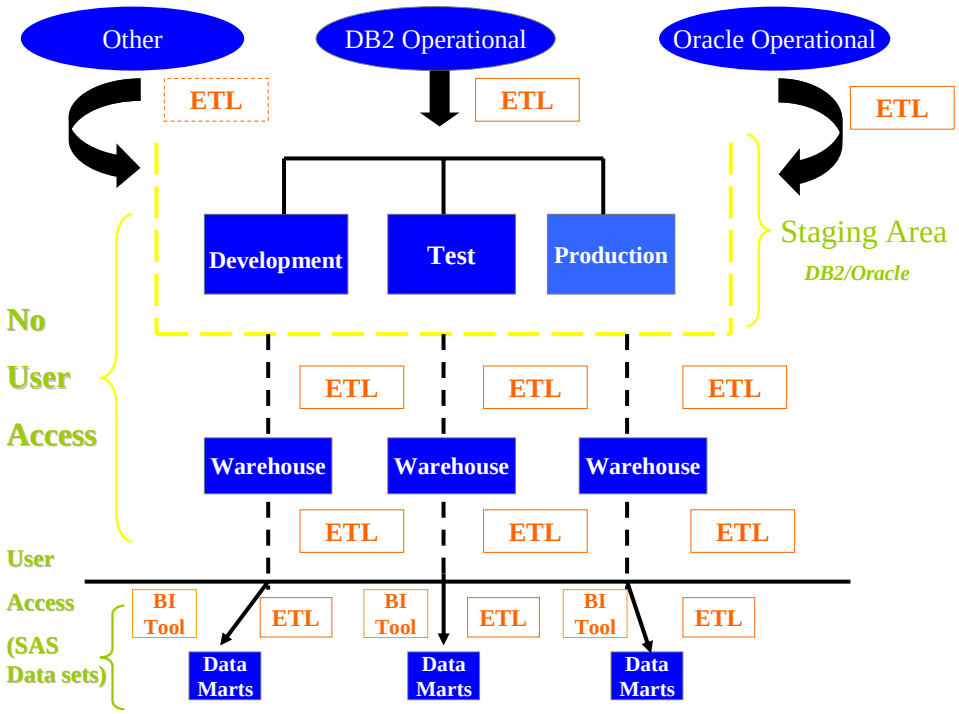
This component of the data warehouse architecture may manifest as the following:

- Customer “visible” relational tables
- OLAP cubes
- Pre-determined parameterized and non-parameterized reports
- Ad-hoc reports
- Spreadsheet applications with pre-populated work sheets and pivot tables
- Data visualization graphics
- Dashboard/scorecards for performance indicator applications

Typically a business intelligence (BI) project may be scoped to deliver an agreed upon set of data marts in a project. Once these have been well specified, the conceptual data model (CDM) is used to determine what parts need to be built or used as a reference to conform the inbound data from any new project. After the detailed data mart specifications (DDMS) have been verified and the conceptual data model (CDM) components determined, a source and target logical data model (LDM) can be designed to integrate the detailed data mart specification (DDMS) and conceptual data model (CMD). An extraction, transformation and loading (ETL) process can then be set up and scheduled to populate the logical data models (LDM) from the required data sources and assist with any time series and data audit change control requirements.

Over time as more and more data marts and logical data models (LDM’s) are built the conceptual data model (CDM) becomes more complete. One very important advantage to this implementation methodology is that the order of the data marts and logical data models can be entirely driven by project priority, project budget allocation and time-to-completion constraints/requirements. This data warehouse architecture implementation methodology does not need to dictate project priorities or project scope as long as the conceptual data model (CDM) exercise has been successfully completed before the first project request is initiated.

McMaster’s Data Warehouse design



Publication Services

This is the visible presentation environment that business intelligence (BI) customers will use to interact with the published data mart deliverables. The SAS® Information Delivery Portal will be utilized as a web delivery channel to deliver a *"one-stop information shopping"* solution. This software solution provides an interface to access enterprise data, applications and information. It is built on top of the SAS Business Intelligence Architecture, provides a single point of entry and provides a Portal API for application development. All of our canned reports generated through SAS® Enterprise Guide, along with a web-based query and reporting tool (SAS® Web Report Studio) will be accessed through this publication channel.

Using the portal's personalization features we have customized it for a McMaster *"look and feel".* Information is organized using pages and portlets and our stakeholders will have access to public pages along with private portlets based on role authorization rules. Stakeholders will also be able to access SAS® data sets from within Microsoft Word and Microsoft Excel using the SAS® Add-In for Microsoft Office. This tool will enable our stakeholders to execute stored processes (a SAS® program which is hosted on a server) and embed the results in their documents and spreadsheets. Within Excel, the SAS® Add-In can:

- Access and view SAS® data sources
- Access and view any other data source that is available from a SAS® server
- Analyze SAS® or Excel data using analytic tasks

The SAS® Add-In for Microsoft Office will not be accessed through the SAS® Information Delivery Portal as this is a client component which will be installed on individual personal computers by members of our Client Services group. Future stages of the project will include interactive reports (drill-down through OLAP cubes) as well as balanced scorecards to measure performance indicators (through SAS® Strategic Performance Management software). This, along with event notification messages, will all be delivered through the SAS® Information Delivery Portal.

Publication is also channeled according to audience with appropriate security and privacy rules.

SECURITY – AUTHENTICATION AND AUTHORIZATION

The business value derived from using the SAS® Value Chain Analytics includes an authoritative and secure environment for data management and reporting. A data warehouse may be categorized as a "collection of integrated databases designed to support managerial decision making and problem solving functions" and "contains both highly detailed and summarized historical data relating to various categories, subjects, or areas". Implementation of the research funding data mart at McMaster has meant that our stakeholders now have electronic access to data which previously was not widely disseminated. Stakeholders are now able to gain timely access to this data in the form that best matches their current information needs. Security requirements are being addressed taking into consideration the following:

- Data identification
- Data classification
- Value of the data
- Identifying any data security vulnerabilities
- Identifying data protection measures and associated costs
- Selection of cost-effective security measures
- Evaluation of effectiveness of security measures

At McMaster access to data involves both authentication and authorization. Authentication may be defined as the process of verifying the identity of a person or process within the guidelines of a specific

security policy (who you are). Authorization is the process of determining which permissions the user has for which resources (permissions). Authentication is also a prerequisite for authorization. At McMaster business intelligence (BI) services that are not public require a sign on with a single university-wide login identifier which is currently authenticated using the Microsoft Active Directory. After a successful authentication the SAS® university login identifier can be used by the SAS® Meta data server. No passwords are ever stored in SAS®. Future plans at the university call for this authentication to be done using Kerberos.

At McMaster aggregate information will be open to all. Granular security is being implemented as required through a combination of SAS® Information Maps and stored processes. SAS® Information Maps consist of metadata that describe a data warehouse in business terms. Through using SAS® Information Map Studio which is an application used to create, edit and manage SAS® Information Maps, we will determine what data our stakeholders will be accessing through either SAS® Web Report Studio (ability to create reports) or SAS® Information Delivery Portal (ability to view only). Previously access to data residing in DB-2 tables was granted by creating views using structured query language (SQL). Information maps are much more powerful as they capture metadata about allowable usage and query generation rules. They also describe what can be done, are database independent and can cross databases and they hide the physical structure of the data from the business user. Since query code is generated in the background, the business user does not need to know structured query language (SQL). As well as using Information Maps, we will also be using SAS® stored processes to implement role based granular security.

At the university some business intelligence (BI) services are targeted for particular roles such as researchers. The primary investigator role of a research project needs access to current and past research funding data at both the summary and detail levels for their research project. A SAS® stored process (a SAS® program which is hosted on a server) is used to determine the employee number of the login by checking a common university directory and then filtering the research data mart to selectively provide only the data that is relevant for the researcher who has signed onto the decision support portal.

Other business intelligence (BI) services are targeted for particular roles such as Vice-Presidents, Deans, Chairs, Directors, Managers and their Staff. SAS® stored processes are used as described above with the exception that they filter data on the basis of positions and organizational affiliations. When individuals change jobs or new appointments occur the authorized business intelligence (BI) data will always be correctly presented.

As the SAS® stored process can be executed from many environments (for example, SAS® Web Report Studio, SAS® Add-In for Microsoft Office, SAS® Enterprise Guide) authorization rules are consistently applied across all environments on a timely basis. There is also potential in the future to automatically customize web portals and event notifications based upon the particular role of the person who has signed onto the SAS® Information Delivery Portal.

ARCHITECTURE (PRODUCTION ENVIRONMENT)

We are currently in the planning stages for building a scalable, sustainable infrastructure which will support a scaled deployment of the SAS® Value Chain Analytics. We are considering implementing the following three-tier platform which will allow us to scale horizontally in the future:

Our development environment consists of a server with 2 x Intel Xeon 2.8GHz Processors, 2GB of RAM and is running Windows 2000 – Service Pack 4.

We are considering the following for the scaled roll-out of our production environment.

A. Hardware

1. Server 1 - SAS® Data Server

- 4 way 64 bit 1.5Ghz Itanium2 server

- 16 Gb RAM
- 2 73 Gb Drives (RAID 1) for the OS
- 1 10/100/1Gb Cu Ethernet card
- 1 Windows 2003 Enterprise Edition for Itanium

2 Mid-Tier (Web) Server

- 2 way 32 bit 3Ghz Xeon Server
- 4 Gb RAM
- 1 10/100/1Gb Cu Ethernet card
- 1 Windows 2003 Enterprise Edition for x86

3. SAN Drive Array (modular and can grow with the warehouse)

- 6 – 72GB Drives (RAID 5) total 360GB for SAS® and Data

B. Software

1. Server 1 - SAS® Data Server

- SAS® 9.1.3
- SAS® Metadata Server
- SAS® WorkSpace Server
- SAS® Stored Process Server
- Platform JobScheduler

2. Mid -Tier Server

- SAS® Web Report Studio
- SAS® Information Delivery Portal
- BEA Web Logic for future SAS® SPM Platform
- Xythos Web File System (WFS)

3. Client –Tier Server

- SAS® Enterprise Guide
- SAS® Add-In for Microsoft Office

REPORTING

We have created a number of parameterized stored processes using SAS® Enterprise Guide, which our stakeholders will access as both static (HTML as well as PDF documents) and interactive reports (drill-down) through SAS® Web Report Studio and the SAS® Add-In for Microsoft Office. All canned reports along with SAS® Web Report Studio will be accessed through the SAS® Information Delivery Portal.

NEXT STEPS

Next steps of the project include development of a financial data mart along with appropriate data quality standards, monthly frozen snapshots and implementation of university-wide financial reporting standards. This will facilitate electronic access to integrated financial information necessary for the development and maintenance of an integrated, multi-year financial planning framework. Canned reports to include monthly web-based financial statements, with drill-down capability along with budget templates automatically populated with data values and saved in different workbooks for different subgroups (for example by Department). The later will be accomplished using Microsoft Direct Data Exchange (DDE).

As well, we will begin the implementation of SAS® Strategic Performance Management Software to support the performance measurement and monitoring initiative that is a fundamental component of McMaster's strategic plan. This tool will assist in critically assessing and identifying meaningful and statistically relevant measures and indicators. This software can perform causal analyses among various measures within and across areas providing useful information on inter-relationships between factors and measures. As well as demonstrating how decisions in one area affect other areas, these cause-and-effect analyses can reveal both good performance drivers and also possible detractors and enable *'evidenced-based'* decision-making. Finally, the tool provides a balanced scorecard reporting format, designed to identify statistically significant trends and results that can be tailored to the specific goals, objectives and measures of the various operational areas of the University.

LESSONS LEARNED

Lessons learned include the importance of taking a consultative approach not only in assessing information needs, but also in building data hierarchies, understanding subject matter, and in prioritizing tasks to best support decision making and inform senior management. We found that a combination of training and mentoring (knowledge transfer) helped us accelerate learning the new tools. It was very important to ensure that time and resources were committed to complete the necessary planning and data quality initiatives prior to initiating the first project. When developing a project plan, it is important to build in appropriate time for research and development and a learning curve. Our core project team comprised a small cross-functional, high performance team. We learned the importance of branding for Business Intelligence (BI) products and services. The business intelligence website was an important communication mechanism for the project.

Notable challenges included the lack of dedicated project resources resulting in deliverables being too far apart. The need to balance enthusiasm of stakeholders with project planning is an on-going challenge. Ensuring a proper sequencing from the stages of proof of concept, to pilot, to production is important but difficult to manage.

CONCLUSION

The creation of the research funding datamart and deployment of business intelligence tools at McMaster University has been a critical first step in creating a *"University without boundaries".* The continued development of the data warehouse and increased deployment of user-friendly, web-based query and reporting tools through the decision support portal will provide the campus community with a greater capacity for data analysis, interpretation, flexible reporting to support decision making and facilitate greater accountability in an environment of devolved authority.

REFERENCES

Adamson C., Venerable M., Data Warehouse Design Solutions, John Wiley and Sons Inc., 1998

Breslin M., Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Innmon Models, Business Intelligence Journal Winter 2004

Data Warehouse Control and Security", Association of College and University Auditors LEDGER, Volume 41, No. 2, April 1997
http://www.tdwi.org/research/index.aspx, The Data Warehouse Institute

Introduction to the SAS® Business Intelligence Client Tools Course Notes, SAS Institute Inc., Carey NC, 2004

Kimball R., A Dimensional Modeling Manifesto: Drawing the Line Between Dimensional Modeling and ER Modeling Techniques, DBMS and Internet Systems, August 1977

Kimball R.,Reeves L.,Ross M., Thornthwaite W., The Data Warehouse Lifecycle Toolkit, John Wily and Sons Inc,1998.

Kimball R., Ross R., The Data Warehouse Toolkit, John Wiley and Sons Inc., 2002

Mimno P., How to Avoid Failures When Using ETL Tools, TDWI December 3, 2003 FlashPoint

Moody, D.L., Kortink, M.A.R., From ER Models to Dimension Models: Bridging the Gap between OLTP and OLAP Design, Part I, Business Intelligence Journal Summer 2003

Moody, D.L., Kortink, M.A.R., From ER Models to Dimension Models, Part II, Business Intelligence Journal Fall 2003

SAS 9.1.3 Intelligence Platform – Planning and Administration Guide, Second Ed. http://support.sas.com/documentation/configuration/iaplanning913.pdf

SAS Add-In for Microsoft Office: An Introduction and Overview (White Paper) http://support.sas.com/documentation/whitepaper/downloads/101814 0904.pdf

SAS Enterprise Guide – A Roadmap (White Paper) http://support.sas.com/documentation/whitepaper/downloads/101016 0603.pdf

SAS Information Delivery Portal (White Paper) http://support.sas.com/documentation/whitepaper/downloads/44182 0701.pdf

SAS Web Report Studio – An introduction and Overview (White Paper) http://support.sas.com/documentation/whitepaper/downloads/101717 1104.pdf

Sprague R.H, Carlson E.D., Building Effective Decision Support Systems, Prentice-Hall Inc.,1982

Wells D., Data Warehouse Architecture for the Rest of Us: Somewhere Between Dimensional and Normalized, TDWI October 20,2004 FlashPoint

ACKNOWLEDGMENTS

The authors would like to sincerely thank Arnold Cordeiro, David Ghan, Stephen Keelan, Fulton Lee and John Parsons from SAS Canada. Your support, encouragement, expertise and superb mentoring skills are greatly appreciated.

Contact Information

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

Debbie Weisensee
McMaster University
1280 Main Street West, T-13 room 101H
Hamilton, ON Canada L8S 4L8
(905) 525-9140 x26408
Fax: (905) 524-5288
Email: weisens@mcmaster.ca