A Data Warehouse Implementation Using the Star Schema

Maria Lupetin, InfoMaker Inc., Glenview, Illinois

Abstract

This work explores using the star schema for a SAS data warehouse. An implementation of a data warehouse for an outpatient clinical information system will be presented as an example. Explanations of the many data warehouse concepts will be given.

The Goal of This Paper:

The purpose of this paper is to introduce the reader to data warehousing concepts and terms. It will briefly define concepts such as OLTP, OLAP, enterprise-wide data warehouse, data marts, dimensional models, fact tables, dimension tables, and the star join schema. The present study will also explore the implementation of a data mart for an outpatient clinical information system using the star schema After reviewing the concepts and approaches, one will conclude that the SAS family of products offers an end to end solution for data warehousing.

Why Implement a Data Warehouse or Data Mart?

In the past, a common system application has been of the OLTP kind: On-line Transaction Processing. Examples of OLTP systems are accounting, payroll, and billing systems where the lowest level of information is a transaction. These OLTP systems are at the heart of running an operation or department of a firm. In my belief, many of these basic operational needs have been satisfied by OLTP systems in place.

These systems contain vast amounts of data which can be difficult to access. However, this data is still critical for decision makers to better manage a firm. In many cases the data is spread across many disparate OLTP systems and is hard or impossible to access with one user friendly interface. This does not eliminate the need to have this data consistent, accessible and useable by the decision makers. Hence the formalization of data warehouse and data marts has occurred. The reader of is paper will be aware of this need because we have been providing information for decision makers using SAS for years.

A data warehouse is a data base that contains data that has been cleansed for quality and rganized for quick querying of information by many different views of the organization.

Data warehouses are subject oriented (i.e., customer, vendor, product, activity, patient) rather than functionally oriented, such as the production planning system or human resources system. Data warehouses are integrated; therefore, the meaning and results of the information is the same regardless of organizational source. The data is nonvolatile but can change based upon history. The data is always the same or history changes based on today's definitions. Contrast this to a database used for an OLTP system where the database records re continually updated, deleted, and inserted.

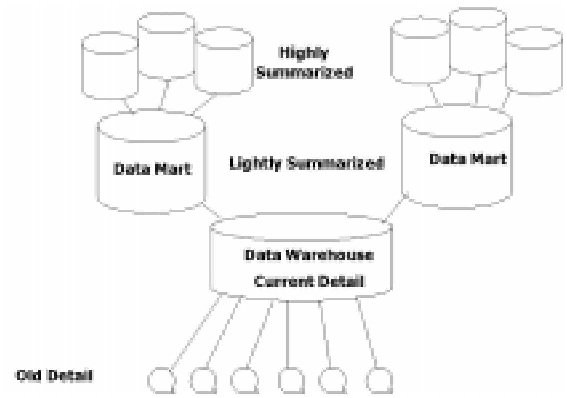
The data is consistence across the enterprise, regardless how the data is examined, "sliced and diced." For example, sales departments will say they have sold 10 million dollars of widgets across all sales regions last year. That volume will be confirmed by the finance department when they access the data warehouse and determine that indeed when the twelves months of last year's national sales are totaled, the result is 10 million dollars of widgets. The two perspectives and queries render the same result.

Data Warehouse vs Data Mart

An enterprise-wide data warehouse is meant to support the informational needs of the entire enterprise. These data warehouses will contain both a very detailed level of data and summarized data to serve decision makers. Several years of history will typically be stored in the data warehouse. It is not unusual to have at least ten years of granular data stored with its many perspectives such as geography, date, style, salesperson, etc.

Many times, development of such a reservoir of data is deemed too formidable, or not seen as a universal need throughout the organization. A data warehouse for a Fortune 200 firm can be several gigabytes to several terabytes in size. In those cases, many firms opt to develop data marts instead.

A data mart is in structure and purpose similar to a data warehouse, except its audience is limited to a departmental or specific subject need. For example, InfoMaker Inc. recently developed a data mart for the sales department of a Fortune 100 firm. The subject was limited to seven years of history of order and sales data at both an individual order and aggregate level.

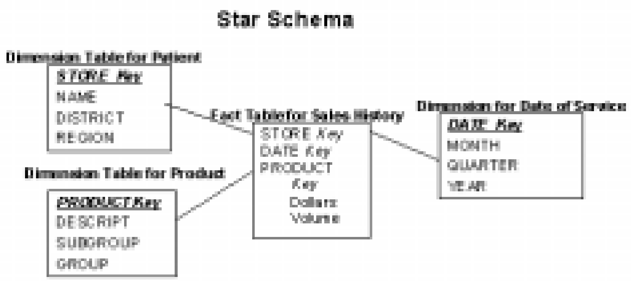


Data Warehousing Concepts

There are several ways a data warehouse or data mart can be structured: multidimensional, star, and snowflake. However, an underlying concept used by all the models method above is that of a dimension. A *dimension* is the different ways the information can be "cut" and summarized such as geography, time intervals, product groups, salesperson(s).

Common to the star, and snowflake methods is the fact table. The *fact table* is a database table that contains the data (factual history) like sales volume, cost or quantity for usually for the smallest time period or a low level of all the dimensions.

This low level of data is call granular data or atomic data. It is usually not aggregated. An example of a record in a fact table for an outpatient clinical information system is information on a single event such as a medical procedure given to a patient on a given day at a specified charge.



In addition to the fact tables, there are also dimension tables in the database. These dimension tables describe the options to "cut" or view the data in the fact table. The star and snowflake schemata all use more than one dimension table in their database. The records in a single dimension table represent the levels or choices of aggregation for the given dimension. The classic data warehouse example used is the "sales" dimension. The records in the sales dimension table will indicate that the fact table data can be aggregated by salesperson, sales districts, sales regions, domestic, international, etc.

Another dimension would be date. Using the date dimension we would be able to analyze data by a single date or dates aggregated by month, quarter, fiscal year, calendar year, holidays, etc. For an outpatient clinical information system, a simple fact table would have the following columns (SAS variables).

PATIENT: unique patient identification number
LOCATION: where the procedure was performed
PROVIDER: doctor or facility that performed the procedure
PAYOR: the organization who pays the bill
CPTCODE: standard
CPT procedure code
DIAGNOS1: primary diagnosis as an ICD9 code
DIAGNOS2: secondary diagnosis as an ICD9 code
DATESERV: date of the procedure was performed
ADJUST: the adjustment to the charge for services
CHARGE: the final
actual charge for the procedure
AGE: age of patient at time of procedure
COUNT: the frequency of the event

The first eight variables from PATIENT to DATESERV are dimension variables. The final four variables, ADJUST, CHARGE, AGE and COUNT are numerical variables to be used for arithmetical or statistical operations.

Star Join Schema

The star join schema (also known as the star schema) is a database in which there is a single fact table and many dimension tables. These tables are not normalized. They are unlike traditional operational data bases where one attempts to normalize the tables. In the fact table there is one segment (a column, a SAS variable) for each dimension. The fact table uses a compound key made up of the group of the dimensions. In addition, the fact table usually contains additional variables which typically are additive numbers, i.e., numericfacts. In SAS terms, the dimension are like the "classes" in Proc Summary, and the facts in the fact table are the "variables" in Proc Summary.

In our outpatient clinical information example the individual dimension table would capture views by:

Patient (name and gender)
Location of service
Doctor or provider performing the procedure
(name and type)
Payor for the service (name and type)
Procedure performed (CPT code and groupings)
Diagnoses (ICD9 code and groupings)
Date procedure was performed (date, month, etc.)

For the full star schema of our outpatient clinical information system, see Exhibit 1 at the end of this paper.

Using the Star Schema for Building Datasets
Users of the Clinical Information System will want to look at the data summarized to various levels. Joining selected dimension tables to the fact table will provided the user with a dataset on which to aggregate the needed information. For example, to analyze the charge by patient, by quarter, by location ould require a the join of four tables: the fact table, the patient dimension table, the location dimension table, and the dataserv dimension table. The resultant data file will then be aggregated by using the Proc Summary step to produce a dataset for analysis. Below is a demonstration of this approach.

Our example Clinical Information System has 500,000 records in the fact table with 12 SAS variables, totaling to 49 megabytes of space.

The dimension tables are small in comparison.

Dimension Table	*records*	*kilobytes*
Patient	10,000	680
Payors	3,700	418
Providers	3,500	246
Procedures	4,000	213
Diagnoses	5,560	246
Locations	10	8
Dates	1,000	49

The SAS code used to generate the table necessary for analysis is as follows.

```
PROC SQL;
   CREATE TABLE PERM.JOIN4 AS
     SELECT FACT_TBL*.,
     PAT_TBL.PAT_NAME, PAT_TBL.PAT_SEX,
     DATE_TBL.MONTHYR,
     DATE_TBL.QUARTRYR, DATE_TBL.YEAR
     LOC_TBL.LOCNAME
FROM
   PERM.FACT_TBL, PERM.PAT_TBL,
   PERM_DATE_TBL, PERM.LOC_TBL,
WHERE
   FACT_TBL.PATIENT=PAT_TBL.PATIENT AND
   FACT_TBL.DATESERV =
   DATE_TBL.DATESERV AND
   FACT_TBL.LOCATION =LOC_TBL.LOCATION; RUN;

PROC SUMMARY DATA = Perm.JOIN4;
   Class Patient Quartryr Location;
   VAR Charges Adjust;
   OUTPUT OUT=Perm.Sum4;
RUN;
```

Since the datasets used are a realistic example, the reader will find the execution performance of interest too. The time to process the SQL code the above code took 2 minutes and 52 seconds on a Gateway 133 megahertz Pentium computer with 32 megabytes of RAM and one processor.

Building the Decision Support Database
Similarly, other datasets could be generated for analysis. Using the building blocks of the fact table and the various dimension tables, one has thousands of ways to aggregate the data.

For expedient analysis purposes, frequently needed aggregated datasets should be created in advance for the users. Having data readily and easily available is a major tenet of data warehousing. For our Clinical Information System, some aggregated datasets could be:
*Charge and Adjustments by Procedure, Provider, and Date
*Patient count by Diagnosis, Gender, Age and Date
*Count of Procedures by Provider and Date
*Charge and Adjustments by Provider and Payor.

As one can see, the Star Schema lends itself well for custom analysis.

OLAP and Data Mining
On-line Analytical Processing (OLAP) is the analytical capabilities provided by the data warehouse or data mart. One can view granular data or various aggregations of data for business analyses using graphical-user-friendly tools.

Data warehouse and data marts exist to answer questions and find business opportunities. There are many ways to analyze data using SAS procedures such as Proc Freq, Proc Summary, Proc Univariate, Proc Means, Proc Tabulate for simple statistics, frequencies and summaries. For the more sophisticated user, SAS/STAT provides a wealth of statistical procedures to analyze the data from the data warehouse, including multiple regression, logistic regression and time series.

Finally, data mining is the name given to newer statistical techniques used to explore voluminous data stores. These techniques include decision trees and neural networks. These methods, like neural networks, can sometimes handle collinearity better than the older statistical techniques. The SAS Institute is offering a new product: Enterprise MinerTM that will contain these techniques. It will be available in production in early 1998.

Why SAS Is Such a Natural Choice for Data Marts

The most common implementation today across "Corporate America" of the data warehousing approach is the development and use of data marts. The SAS Institute provides software to create tables; clean, load, aggregate, maintain, query data; and data mine these data marts. Just think of the power we SAS developers, analyst, and users have had at our finger tips all these years!

SAS Procedures like Download, Copy, SQL, APPEND, FSEDIT, Format, and SUMMARY are only a few of the many tools available to build a data mart. Recently, the SAS Institute introduced new products such as the SAS/MDDB Server and SAS/Warehouse Administrator. The SAS/MDDB Server is used for the development of a multidimensional data warehouses and data marts. SAS/Warehouse Administrator is used to create and maintain table based data warehouses and data marts. The SAS Institute Provides an integrated end to end solution for data warehousing.

Why Data Warehouses Are Here to Stay

Simple, we need the answers now, accurate and consistent to run our businesses.

About the Author

Maria Lupetin is president of InfoMaker Inc., Glenview, Illinois. She has over twenty years of business experience in information technology and applied mathematics. Her background spans manufacturing, distribution, marketing, and healthcare applications assisting many Fortune 200 companies.

Ms. Lupetin is the founder of InfoMaker Inc., a consulting firm specializing in decision support systems and data warehousing. InfoMaker is a SAS Quality Partner, and partner with the Oracle and Informix companies. Before founding InfoMaker, Maria worked at Morton International and Union Pacific Railroad.

Maria has an MBA from the University of Chicago and a M.S. in Operations Research from the University of Minnesota.

Ms. Lupetin can be contacted at:

InfoMaker
Inc. 950 Milwaukee Avenue
Glenview, Illinois 60025
Phone: (847)390-6660
Fax: (847)390-6774
e-mail:lupetin@infomaker.com
http://www.infomaker.com

Note: Many thanks to Larry Greenfield, LGI Systems Inc., Northfield, IL, whose marvelous web page, "The Data Warehousing Information Center," was a starting point for this research. In addition, thanks to my staff: Lisa Surath, Sharon Terpstra and Oscar Hayes in the development and preparation of this paper.

References

Greenfield, Larry, LGI Systems Inc., (1997), "The Data Warehousing Information Center," http://pwp.starnetinc.com/larryg/index.html.

Inmon, W. H., (1996), Building the Data Warehouse, Second Edition, John Wiley & Sons, Inc.

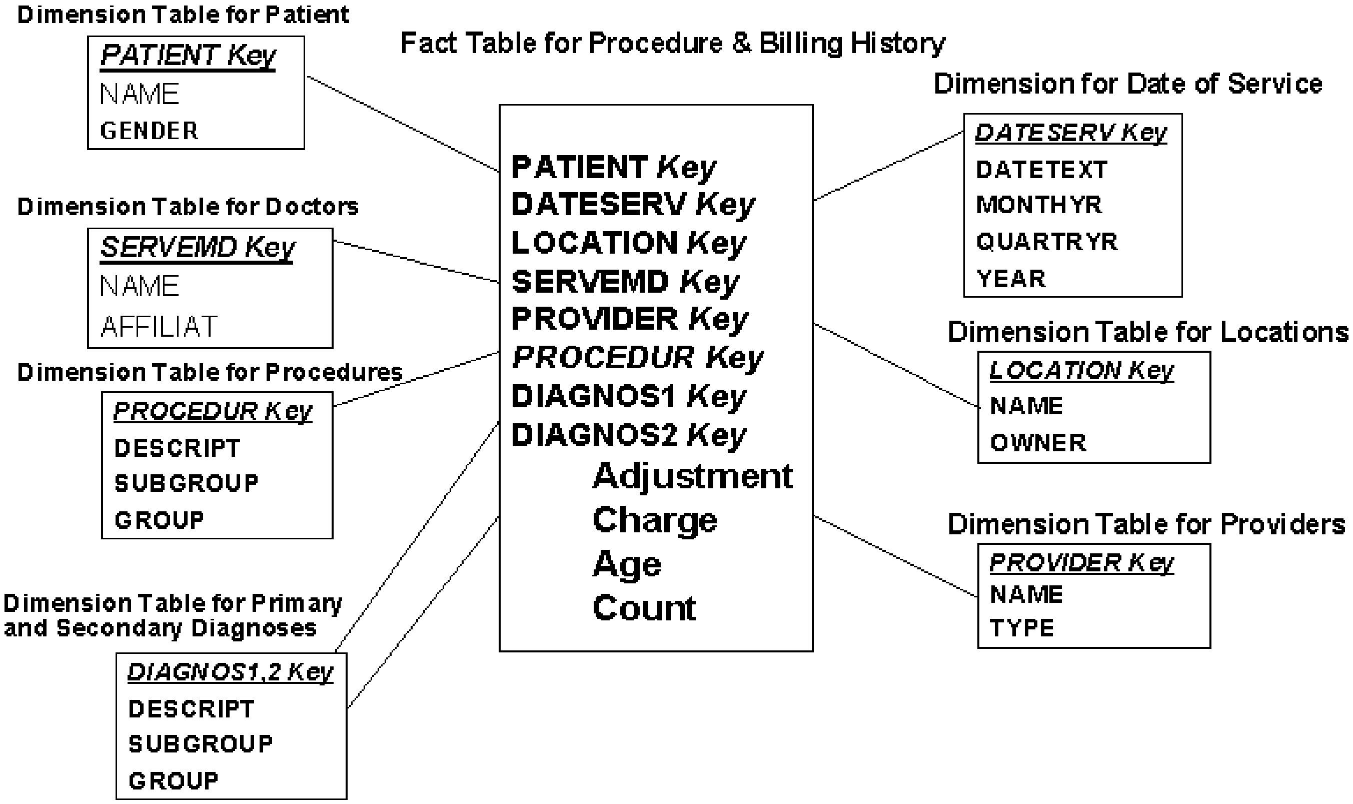
Kimball, Ralph, (1996), *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*, John Wiley & Sons, Inc.

SAS Institute Inc. (1996), "Implementing a Data Warehouse Using SAS Software Course Notes," Cary, NC., SAS Institute Inc.

Trademarks or Registered Trademarks:

Star Schema for Outpatient Clinical Information System

Exhibit 1



Dimension Table for Patient

PATIENT Key
NAME
GENDER

Dimension Table for Doctors

SERVEMD Key
NAME
AFFILIAT

Dimension Table for Procedures

PROCEDUR Key
DESCRIPT
SUBGROUP
GROUP

Dimension Table for Primary and Secondary Diagnoses

DIAGNOS1,2 Key
DESCRIPT
SUBGROUP
GROUP

Fact Table for Procedure & Billing History

PATIENT *Key*
DATESERV *Key*
LOCATION *Key*
SERVEMD *Key*
PROVIDER *Key*
PROCEDUR Key
DIAGNOS1 *Key*
DIAGNOS2 *Key*
Adjustment
Charge
Age
Count

Dimension for Date of Service

DATESERV Key
DATETEXT
MONTHYR
QUARTRYR
YEAR

Dimension Table for Locations

LOCATION Key
NAME
OWNER

Dimension Table for Providers

PROVIDER Key
NAME
TYPE