Paper 226-25

Application of SAS® Software in the Establishment of a Data Mart for Quality Analysis System in the Metallurgical Industry

Xiao Ji, Shanghai Baosteel Computer System Engineering Co. Ltd. Shanghai, P.R.C.
Ge Yu, Northeast University, Sengyang, P.R.C
Jay Cao, SAS Institute(Shanghai), P.R.C

ABSTRACT

Data warehousing or data mart technology is gaining acceptance in more and more industries in China. The technology provides business units a solution as to how to efficiently manage the vast amounts of data from OLTP. It also supports the informational needs for OLAP application, decision support and data mining in an enterprise. This work, using the actual application of a quality analysis data mart in the metallurgical industry as an example, explores using the star schema for a SAS data mart, and the related OLAP applications of the data mart. The establishment of the data mart meets the business needs of the quality control department. It offers business units a new alternative as to how to obtain information for decision support, and provides a model as to how to build a departmental or an enterprise-wide data warehouse. This data mart uses SAS/WA as an administrator, and SAS/IntrNet & SAS/MDDB for information exploration.

KEY WORDS Data mart, data warehouse, star schema, OLAP, SAS

INTRODUCTION

There exist an aggregation of traditional OLTP systems in a modern business enterprise. These OLTP systems meet the daily operation needs, and thus play a crucial role in the business activities within an enterprise. Due to the years' use of these systems, vast amounts of primitive operational data have accumulated in a business firm. This data reflects the business activity of the whole enterprise.

However, for various reasons, the data is always not efficiently used. Thus a better solution is needed in order to make good use of the data for statistical and analytical purposes, and to finally convert the data into useful business information for decision-makers. The problems faced are：different business units have their own operational systems, which are self-contained. The data generated do not share with each other, and thus form the so called different "information islands". The negative result is that it is difficult for different departments such production, technology, finance, sales and a lot others to efficiently access the data they need, which in turn makes it difficult for them to carry out further analytical work.

On the other hand, due to the lack of sophisticated analytical tools, IT engineers have to develop their own analytical applications, which may result into inefficiency and duplicate application developments. To overcome these problems, Shanghai Baosteel makes use of the data warehousing technology. Baosteel is one of the first metallurgical companies in China to use this technology. As early as 1997, Baosteel began to build a data mart for quality analysis to assist the Technology Department and other related business users in their quality control management.

A data mart is established based on the decision support needs of a specific department. Its audience is limited to a specific subject need. For example, the financial department may develop its own data mart for the purpose of financial reporting and analysis. The sales & marketing department may develop a data mart for the purpose of customer relationship management.

A data mart and a data warehouse are different in that the former usually uses the star schema in its database design, and mainly supports the departmental needs. The historical data in the data mart may be incomplete, and not detailed enough, and mainly for analysis at the departmental level. A big index may be built for speed purpose. A data warehouse is meant to support the needs of the entire enterprise. Usually the multidimensional data warehousing technique is used. The detailed data of an enterprise is stored in the data warehouse. The data model in a data warehouse is formalized, and mostly fit third normal form. A data warehouse contains more detailed historical information, and fit for a higher level analytical work. It is also suitable for the storage and management of huge amounts of data with a consistent structure and few indexes.

The establishment of an enterprise-wide data warehouse requires enormous manpower and budget with a long implementation cycle. Not all enterprises can afford the above. The establishment of a data mart requires shorter time can meet departmental needs well. We therefore start from a metallurgical quality analysis data mart, and after gaining enough experience will we further build other data marts, and eventually build an enterprise data warehouse.

This paper discusses techniques in establishing a data mart as well as OLAP applications based upon this data mart. The second part will mainly discuss data mart modeling. The third part introduces techniques in implementing a data mart. The fourth part explores an OLAP application, while the fifth summarizes the major points of this paper.

STAR SCHEMA MODELING & ITS FEATURES

Logical modeling represents a very important step in the implementation of a data mart, because it directly reflects the needs of business departments. It also plays an important instructive role in the physical implementation of the system. There exist a number of methods to build a data warehouse or a data mart such as entity relational model, summarized table, multi-dimensional data base, third normal form, star schema and snowflake schema. The star schema and third normal form are viewed to be the most common methods.

Normal form is the basic theory in the logical model design in database, and the third normal form has a very strict mathematical definition. A relationship fit for the third normal form must meet the following three conditions:

- Only one value for each attribute, with only one single meaning;
- Each non-key attribute wholly depends on the entire primary key, and is not part of the primary key;
- Each non-key attribute cannot depend on that of other relationship; otherwise, this kind of attribute should be classified into other relationship.

In designing the logical model of data warehouse, most people tend to use the third norm form. In the physical implementation, in order to enhance the response speed, one has to de-normalize the logical model due to the restrictions in data base engine. This is achieved at the expense of an increased complication of the whole system, increased maintenance workload, and a deteriorated ability of dynamic inquiry of the system.

The star schema model is used to refer to a multi-dimensional data relationship. It consists of a fact table and a set of dimension tables.

These tables are de-normalized. Each dimension table has one dimension as its primary key, and all these dimensions make the primary key in the fact table. In other words, each attribute in the fact table is the foreign key of the dimension table. The non-primary key attribute in the fact table is called fact. They are normally nominal numbers or other calculable data, while dimensions are mostly data related to text or time.
In data warehousing implementation, the star schema method is frequently used. In a small-scale data warehouse such as a data mart, when the fact table is not big, the star schema model can be used for the actual physical modeling. Since the fact table is not very big, the data can be de-normalized entirely, and thus a simplest model can be established.

The star schema is a method, which can be used to deal with business needs and to convert them into being technically operational. It is easy for business users to understand, and can help business users to place data warehousing needs into star view, and can more easily see their needs, and to define them in a more simple way, and finally form so-called “fact” and “dimension.” In a word, using this method, it is easier to convert business needs into subjects and facts. These subjects are used for further analysis, and the users can easily understand them through the data view.

The inquiry on the data warehouse built on the star schema method is much more efficient. This is because it has already pre-processed all dimensions such as statistics, classification, and lining done in advance. Therefore, it is very quick to make reports based on the data mart of the star schema model.

When used for the large-scale data warehouse, the star schema has the following disadvantages: (1) in the physical implementation, due to the large volume of data merge, the response is mostly slow. That means that to the large-scale data warehouse, we only benefit from the formalized structure, but hardly realize it using the star schema model. (2) When subjects are established and further classifications are done, the actual business needs are conceptualized, and the direct relation with the data is lost. This means that there is no direct relationship between the business needs and the data warehouse. Therefore, there exist limitations in its assessment ability. (3) Since huge amounts of work has to be done in advance, the workload in building the model is huge. Besides, when business needs change, the original dimensions will not meet new requirements, under which case new dimensions will have to be added into. Since the keys in the fact table of the star schema model are composed of keys in the all dimension tables, this kind of dimensional change will be very complicated and time-consuming. (4) One more disadvantage of the star schema is that there are too many data redundancies.

However, a data mart is much smaller in scale than an enterprise data warehouse. Thus, a simple and easier star schema structure is frequently used.

IMPLEMENTATION OF DATA MART

After a comprehensive analysis of actual business needs and the actual IT infrastructure in Baosteel, we decide to use SAS/WA and other SAS products as the tools to integrate relevant source data, and to build a data mart of metallurgical quality analysis system in Baosteel.

In building the metallurgical quality analysis data mart in Baosteel, we collected huge amounts of production messages and transaction process data from all the existing OLTP systems, and stored them into the SAS/WA data warehouse management system, where SAS tools are used to aggregate, clean, store and manage data. Hundreds of subjects are established with a data mart established subject by subject to meet required OLAP needs. Based upon this data mart, further OLAP applications are developed. For example, we use the following SAS software such as SAS/ASSIST, SAS/GRAPH and SAS/STAT to do statistical analysis, and finally built a metallurgical quality analysis system for the purpose of quality control. Exhibit 1 illustrates the system structure of the metallurgical quality analysis system.

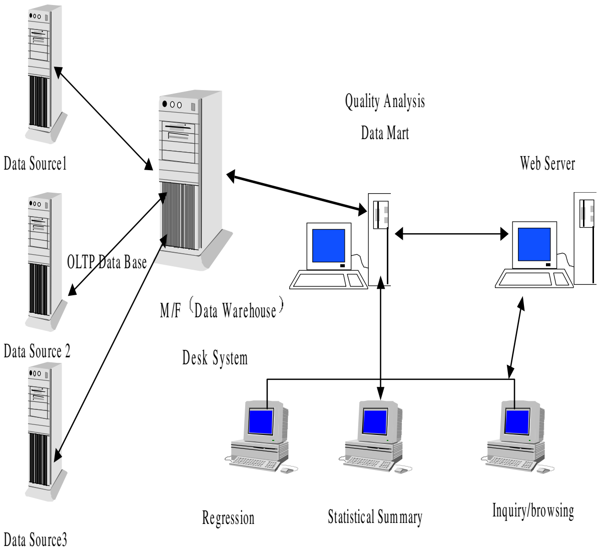
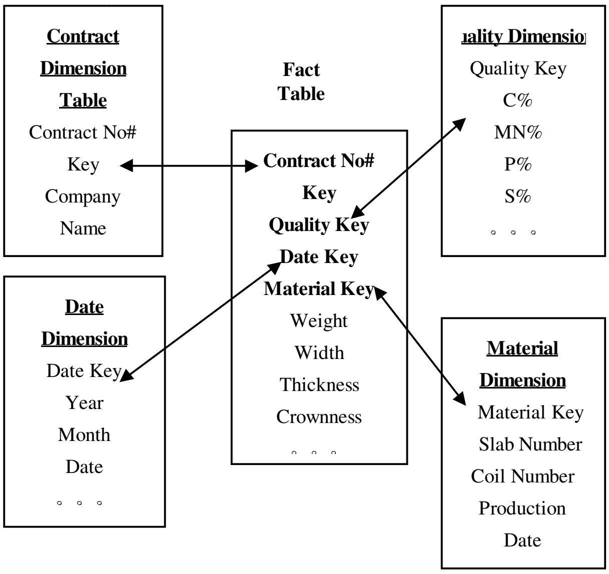


Exhibit 1 Data Mart System Structure

Exhibit 2 illustrates the star schema model in a data mart. Among all, the independent dimension tables include the following four kinds: (1) contract dimension table (order information), (2) material dimension table (material for production and processing information), (3) quality dimension table (quality information), (4) date dimension table (date of production information).

The fact table has the following variables:

- Contract No. key: PO identification, primary key
- Date key: to identify the material production date, primary key;
- Quality key: to identify the properties of the production material, index key;
- Material key: refer to material type, primary key;
- Weight: refer to the material weight, calculable nominal value;
- Width: refer to the material width, calculable nominal value;
- Thickness: refer to the material thickness, calculable nominal value;
- Crown: refer to the material crown, calculable nominal value.



Contract Dimension Table
Contract No#
Key
Company
Name

Fact Table

Contract No#
Key
Quality Key
Date Key
Material Key
Weight
Width
Thickness
Crownness
。 。 。

uality Dimensio
Quality Key
C%
MN%
P%
S%
。 。 。

Date Dimension
Date Key
Year
Month
Date
。 。 。

Material Dimension
Material Key
Slab Number
Coil Number
Production
Date

Exhibit 2 Star Schema in Quality Analysis Data Mart

The first four variables are dimensional, while the last four variables are nominal, and can be used to mathematical and statistical analysis.

End users in the business units always want to inquire multi-layer summary data. Thus, this data mart will enable the user to gain the information they need. For example, to analyze the production volume based on contract, steel grade number and time, four tables need to integrated: fact table, contract dimension table, steel grade number dimension table and product dimension table. The data mart thus entailed can be further used for analysis.

We can use SAS language and the star schema model to generate a data mart subject for the purpose of further analysis. The above program can be written as the following:

```
              PROC SQL;
              CREATE TABLE IH.HSMSU01 AS
              SELECT FACT_TBL.*,
                  ORD_TBL.ORD_NAME,
         ORD_TBL.ORD_LOC,
                  DATE_TBL.YEAR, DATE_TBL.MONTH,
                  QTY_TBL.STLGRD, QTY_TBL.C,
                  MAT_TBL.SLAB_NO, MAT_TBL.COIL_NO
              FROM
                  IH.ORD_TBL, IH.DATE_TBL,
                  IH.QTY_TBL,IH.MAT_TBL
              WHERE
                  FACT_TBL.ORDER_NO =
         ORD_TBL.ORDER_NO AND
                  FACT_TBL.QTY_KEY = QTY_TBL.QTY_KEY
         AND
                  FACT_TBL.DATE_KEY =
         DATE_TBL.DATE_KEY AND
                  FACT_TBL.MAT_NO = MAT_TBL.MAT_NO;
             RUN;
```

Similarly, we can use the same method to build subjects for other analyses. Using the fact table and different dimensional tables, people have thousands of methods to integrate data. For the purpose of convenience, those dimension tables, which are frequently needed to be integrated, should be built for the users in advance. To be easy to read and use is the major purpose of building a data warehouse or a data mart. As to our metallurgical quality analysis data mart, we can build some other summary data dimension tables, e.g.,

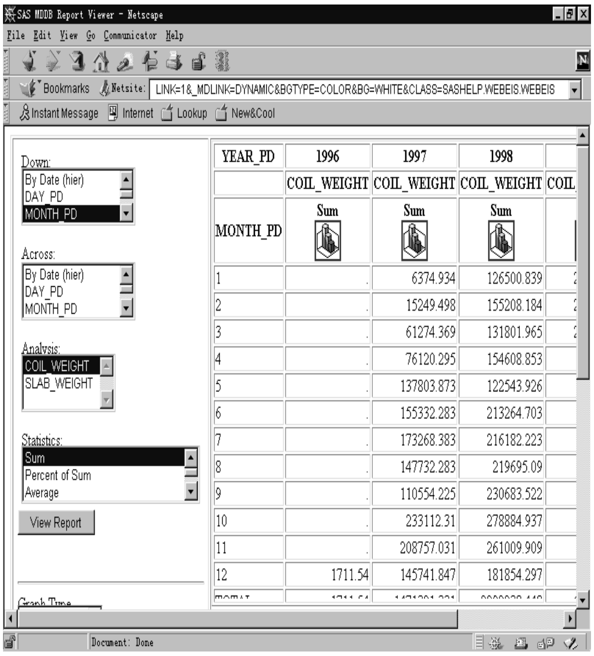
- Production volume by time and by contract.
- Production accuracy by steel mart and specification
- Property analysis by steel mark and steel property

Experience shows that the data mart of star schema model can well meet the needs of end users.

OLAP APPLICATION

OLAP（On-line Analytical Processing）can be seen as the analytical capabilities provided by the data warehouse or data mart. We can view granular data or various aggregations of data for business analyses.

Based upon the data mart of metallurgical quality analysis we built, a number of applications are developed, e.g., Quality Analysis System for Hot Rolling and Cold Rolling Products, and Cold Rolling Locking Management System. In addition, web applications are also initiated in the Intranet so that end-users can easily access the most up-to-date information in the data mart. The SAS/MDDB is also used to generate reports quickly. Furthermore, SAS calculation capability can be achieved dynamically through the web server. For the illustration, see Exhibit 3.



LINK=1&_MDLINK=DYNAMIC&BGTYPE=COLOR&BG=WHITE&CLASS=SASHELP.WEBEIS.WEBEIS

Down:
By Date (hier)
DAY_PD
MONTH_PD

Across:
By Date (hier)
DAY_PD
MONTH_PD

Analysis:
COIL_WEIGHT
SLAB_WEIGHT

Statistics:
Sum
Percent of Sum
Average

YEAR_PD	1996	1997	1998	
	COIL_WEIGHT	COIL_WEIGHT	COIL_WEIGHT	COIL
MONTH_PD	Sum	Sum	Sum	
1	.	6374.934	126500.839	2
2	.	15249.498	155208.184	2
3	.	61274.369	131801.965	2
4	.	76120.295	154608.853	
5	.	137803.873	122543.926	
6	.	155332.283	213264.703	
7	.	173268.383	216182.223	
8	.	147732.283	219695.09	
9	.	110554.225	230683.522	
10	.	233112.31	278884.937	
11	.	208757.031	261009.909	
12	1711.54	145741.847	181854.297	
[illegible]	[illegible]	[illegible]	[illegible]	

Exhibit 3 MDDB Multi-dimensional Analytical Report

The above illustrates the use of a multi-dimensional analytical tool for the quick generation of reports. Users can use browser through Intranet to produce single dimension or two dimension reports. Exhibit 3 shows a two-dimension production report and a corresponding graph can also be generated by simply clicking by mouse. Furthermore, selecting different analytical variables and analytical methods to satisfy the needs of the end user can produce statistical reports of various forms.

Data Mining is a kind of decision support process. It is mainly based upon AI, automation learning, and statistical techniques. The data mining technique is used to analyze the primitive data in a firm, to draw synthesizing conclusion, to explore the underlying models, to predict customer behavior, to eventually help the decision-maker to make better business decisions.

Compared with data mining, OLAP may be seen to be at the shallower level. The difference in analytical model leads to the difference in the analytical ability between OLAP and the data mining technique. In Baosteel, the data mining technique has already been explored and used.

CONCLUSION

To build a data warehouse or a data mart, single models method is not the only technique. Under certain circumstance, it is easier for a firm to build dimension tables and in this case, the star schema method can be used. In another case, if the business needs in a firm are fixed and limited, the objective-oriented method may be more fit. If the star schema method is used for the detailed data in a data warehouse, and the objective-oriented method is used for the data mart, the selection of models to be used depends upon how to maximally satisfy the business needs of a firm. Overall, one point is vital, i.e., the data warehousing tool must be flexible enough to ensure the independence of these models, and the models themselves should also be sufficiently flexible to meet all those possible changes in the future.

The conclusion is that to build a departmental data mart for the purpose of answering repeated questions the star schema method is more fit for the predefined questions as in such a case when the data is not so big but large numbers of reports are to be generated based upon this data. The star schema method may not fit the case, which involves vast dynamic inquiries, high demand of a highly extendable system, and vast amounts of data.

Therefore, the star schema method entails more applications in a departmental data mart, which requires huge numbers of reports. The application of the metallurgical quality analysis data mart enables the business departments to access information conveniently, and it also offers a good preparation for data mining. Based upon our experience in implementing the metallurgical quality analysis data mart, we plan to build other data marts and eventually to build an enterprise-wide data warehouse.

REFERENCES

Cai,Yu & etc, (Jan.4, 1999), “Data Mining & OLAP,” *Network World*.

Inmon,W.H.,(1993),*Building the Data Warehouse,* John Wiley & Sons,Inc.

Lupetin,Maria,(1998), “A Data Warehouse Implementation Using Star Schema,” Proceedings of the Twenty Third Annual SUGI Conference.

SAS Institute Inc., *Data Warehousing Overview Theory and Business Concepts*, Cary, NC., SAS Institute Inc.

Wang,Cangzhou,(Sept.9, 1998), “Logical Modelling in Data Warehousing,” *Computer World*.

Wang, Cangzhou,(July 20, 1999), “Datamart≠Data Warehouse,” *Computer World*.

Welbrock,Peter R.,(1998), *Strategic Data Warehousing Principles Using SAS Software.*

Wang,Shan,(1998),*Data Warehousing Technique & OLAP*,

CONTACT INFORMATION

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

Author Name: Xiao Ji
Company: Shanghai Baosteel Computer System Engineering Co. Ltd. Shanghai, P.R.C.
Address: 688 Fujin road, Baoshan District
City state ZIP: 201900
Work Phone: 86-21-56191018
Fax: 86-21-56783985
Email: sam.ji@mailroom.com
Web: http://www.baosteel.com

Author Name: Ge Yu
Company: Northeast University, Sengyang, P.R.C
Address: Information Institute, NEU, Heping District
City state ZIP: 110006
Work Phone: 86-24-23895654
Fax: 86-21-23895654
Email: yege@mail.neu.edu.cn
Web: http://www.neu.edu.cn

Author Name: Jay Cao
Company: SAS Institute (Shanghai), P.R.C
Address: 803 Yun Hai Mansion,
1329 Huai Hai Zhong Road
City state ZIP: 200031
Work Phone: 86-21-64725536
Fax: 86-21-54560570
Email: jay@sas.com.cn
Web: http://www.sas.com