Optimizing Data Warehouse Design

THE RESPONSE SURFACE METHODOLOGY CAN REDUCE CYCLE TIME AND IMPROVE PERFORMANCE.

By Vladimir Boroditsky and Christine Molinare, Motorola

Modern production software development often involves reuse and customization of existing software components and solutions. In fact, many new projects are actually extensions of existing solutions to new application and business domains.

Production performance of new software components is directly affected by fundamental design decisions made in early development stages. Applying statistical analysis to the historical production performance data can help identify underlying dependencies used to develop working models. It can also provide software designers and developers with guidelines for optimal design characteristics for future projects.

In this study, we applied elements of the response surface methodology (RSM) to the database design of a corporate data warehouse. This approach helped estimate potential risks at the project initiation stage and provided a design optimization tool for the database architects at the project design stage.

Subject of the Study: Data Mart Database Design

The environment modeled in the study is a data warehouse, which is a family of data marts.[1] A data mart contains a snapshot of operational data that helps an organization strategize based on analyses of past trends and experiences. The data are grouped by certain logical denominators, are developed on their own schedule and integrated into the data warehouse.

There are three main types of data mart database tables:[2]

1. Dimension tables: These contain the textual descriptors of the business and usually a fairly small number of records and large number of data attributes that reflect relatively static characteristics of the business objects.

2. Facts tables: These are used when the numerical performance measurements of the business are stored and contain a large number of records. A fact record usually consists of references to corresponding dimension records and numeric characteristics.

3. Aggregations tables: These contain consolidated, denormalized, precalculated information ready to be presented through metrics and reports.

The overall information flow in a typical data mart implementation is shown in Figure 1 (p. 32). The extraction, transformation and loading (ETL) software component collects, validates, cleanses and consolidates data from external sources. The resulting data are stored in the dimension and fact tables. Then, various business, aggregation and filtering rules are applied to the data to calculate report-ready information and store it in the aggregation tables. The reports deliver the information to decision makers.

ETL performance is one of the most important constraints in data mart development. Many data marts are used to support both strategic and operational decision making. Therefore, it is imperative to have data marts updated in near real-time mode. Non-optimal data mart database design may make it impossible to refresh the data in the required operational windows.

There are many database design characteristics that influence ETL performance. In this study, we have assigned the characteristics to three distinct groups based on the degree of control a database designer has over them:

Group one: characteristics dictated by business requirements. The designer has almost no control over the number of external data sources to be polled by the data mart. The number of aggregation tables is dictated by the type or number of external reports to be supported by the data mart.

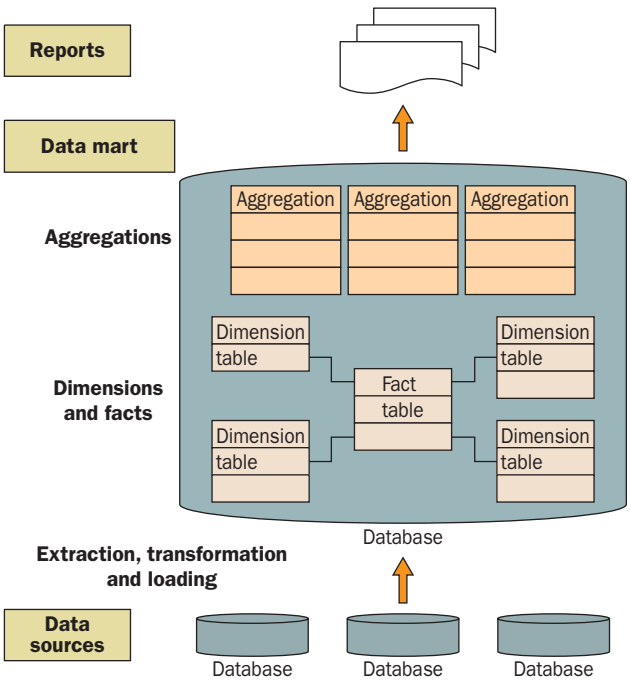
Group two: characteristics dictated by requirements. The designer has limited control over how often data loads run, both within a week and during a day.

Group three: internal characteristics. The designer has substantial control over the number of dimension tables and fact tables.

The goal of the study was to develop a model that would help designers identify optimal settings for the group two and three design characteristics while accommodating characteristics from group one. We also hoped to develop clear, simple reference materials for database design engineers that could be used at the early stages of a project. Therefore, we tried to find a technique that not only provided the needed statistical analysis foundation but also allowed the results to be represented graphically and dynamically.

After applying more traditional statistical studies methodologies, we selected RSM. Granted, RSM is customarily associated with design of experiments work, but we found it very useful in determining optimal settings of input parameters (design characteristics) and minimizing predefined response (ETL performance), while presenting the results graphically.

Figure 1. **Data Mart Information Flow**



Data for Optimization Study

We applied RSM regression to the data set that combines historical ETL performance data with corresponding design characteristics. Our goal was to identify the main factors and interactions that affect performance and develop a model that would allow database designers to define an optimal ratio of dimension to fact tables based on the limited information they know from the requirements. All they know is the number of sources to pull data from and the number of aggregation tables to be presented in the design.

The data set includes 858 ETL performance data points for four data marts.[3] The numbers of dimensions and facts were replaced with ratio dimensions/facts. The final data set included the following data attributes:

- Data mart designator—change management data mart, test management data mart, inspection data mart, code count data mart.
- Run date—date when the ETL session runs.
- Run time—time when the ETL session runs.
- ETL duration—duration of the ETL session in minutes.
- Ratio dimensions/facts—number of dimension tables divided by the number of fact tables in the existing designs. This is the only factor that represents internal characteristics of the design (group three). The goal is to create a model to identify optimal ratio for the design with a given number or sources and aggregations for different days of week and run times.
- Number of aggregations—number of aggregation tables in the data mart.
- Number of sources—number of data sources that are polled by ETL procedures when the data are loaded into the data marts.

- Run time period—derivative of ETL run timestamp. A day is divided into three periods of eight hours each.
- Run day of week—number from one to seven (Sunday-Saturday). This is a derivative of ETL run date.

RSM Regression Application

We conducted the study using Minitab 1.4. We defined custom RSM design on the data set. Initial screening analysis identified that run day of week has no affect on the response and can be removed from the model. The resulting RSM analysis log is shown in Table 1.

The high R^2 value (89.1%) and R^2 adjusted (89%, penalizing for overfit) confirm a strong correlation among main effects, interactions and the response. The next step in evaluation of the model is to examine observed errors (residuals). The standardized residual plots are presented in Figure 2 (p. 34). The residual distributions against fitted values and order of the data are fairly random with no visible patterns. This indicates a reasonably good model fit and hardly detectable correlation among residuals.

On the other hand, the normality plot has clearly visible outliers. Equally important and alarming is that the distribution is not normal, as the Andersen-Darling P value is less than 0.05. Analysis of the collected data revealed the outliers can be linked to the observations made during abnormal conditions caused by network and database outages, such as occasional slow connectivity to remote data sources and exceeding database configuration limits while loading the data.

Table 1. **Response Surface Methodology Analysis Of ETL Performance Data**

Response surface regression: extraction transformation and loading (ETL) duration vs. run time

The following terms cannot be estimated, and were removed:

- Number of aggregations*number of aggregations.
- Ratio dimensions/facts*ratio dimensions/facts.
- Number of aggregations*ratio dimensions/facts.
- Number of sources*ratio dimensions/facts.

The analysis was done using uncoded units.

Estimated regression coefficients for ETL duration

Term	Coefficient	Standard error coef.	T	P
Constant	-824.851	121.467	-6.791	0.000
Run time period	46.568	11.882	3.919	0.000
Number of aggregations	-173.839	36.286	-4.791	0.000
Number of sources	380.719	85.181	4.470	0.000
Ratio dimensions/facts	343.074	31.672	10.832	0.000
Run time period*run time period	-3.492	1.844	-1.893	0.059
Number of sources*number of sources	-45.724	10.638	-4.298	0.000
Run time period* Number of aggregations	5.098	1.634	3.119	0.002
Run time period*number of sources	-4.884	3.006	-1.625	0.105
Run time period* Ratio dimensions/facts	-20.874	3.688	-5.659	0.000
Number of aggregations* Number of sources	28.603	5.345	5.352	0.000

S = 25.24 R^2 = 89.1% R^2(adj) = 89.0%

Analysis of variance for extraction, transformation and loading (ETL) duration

Source	DF	Seq SS	Adj. SS	Adj. MS	F	P
Regression	10	4413775	4413775	441377	692.97	0.000
Linear	4	4121760	742465	185616	291.42	0.000
Square	2	217657	13921	6961	10.93	0.000
Interaction	4	74358	74358	18589	29.19	0.000
Residual error	847	539484	539484	637		
Lack-of-fit	4	4697	4697	1174	1.85	0.117
Pure Error	843	534787	534787	634		
Total	857	4953259				

Seq SS = sequential sum of squares Adj MS = adjusted mean squares

While the effect of the special conditions can be lessened with proper network and database administration and monitoring, they cannot be eliminated completely. In this study, we decided to include the outliers in the considerations and preserve them in the model.

The analysis of variance table (Table 1, p. 33) shows there is reasonably low lack of fit (P-value is 0.117), while the regression F test statistic is quite high. This indicates a very low probability that the data are not showing true correlation. Also, the table indicates presence of the linear, quadratic and interaction components in the model. Contour and surface plots of the response against contributing factors exhibit a wide range of typical curves from straight parallel lines (linear without interactions) through rising ridge (quadratic with interactions).

Data Mart Database Design Optimization Model

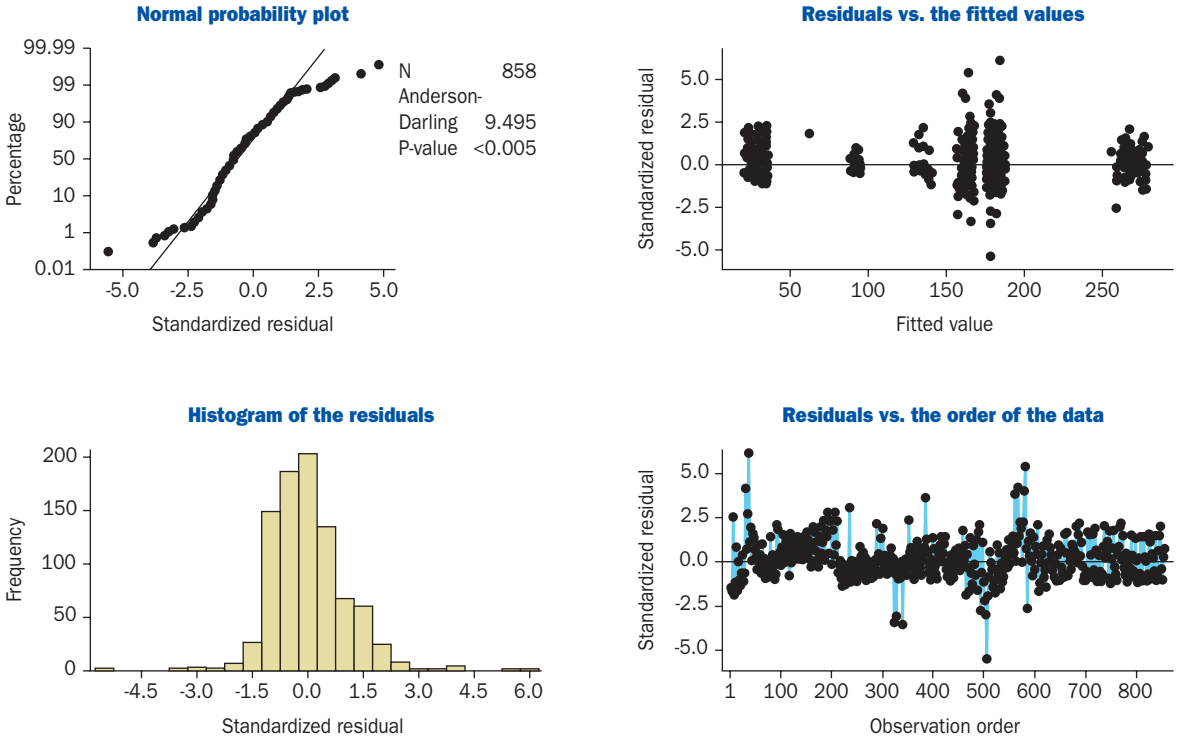
Among the practical deliverables of this study is a set of contour plots that show regions of acceptable ETL durations. An example of the plot is presented in Figure 3. The plot shows that to meet ETL duration requirements (ETL duration is no longer than 150 minutes) for the system extracting data from a predefined number of sources, a designer has to be very particular in selecting an appropriate ratio of dimensions to facts.

For example, for the number of sources equal to one, the ratio must be around two. In some cases (for the number of sources between three and six) the desired ETL performance cannot be achieved within the constraints of the model.

Minitab Response Optimizer can more accurately calculate the expected ETL duration (see Figure 4). It presents response-factor pair dependencies visually and allows dynamic change of one or more factor values with automatic recalculation of the response.

By sliding the red vertical lines through the response-factor dependency curves, an analyst can instantly see the effect on the response, so he or she can find factor settings to get as close as possible to the target response value. In one practical implementation, the data mart had to extract data from one source and support one aggregation table. Using the model, designers identified that the ratio of dimensions to facts had to be maintained at 2-1 to get a desir-

Figure 2. **Residual Plots for Extraction, Tranformation and Loading Duration**



able ETL duration of about 60 minutes.

The model is used for two distinct purposes in a typical data mart development project.

First, in the project initiation stage, it provides high level estimates of the ETL duration, which are considered in the initial risk assessment. At this stage, while trying to understand scope and customer requirements, an experienced data architect can identify the main data design characteristics, such as number of aggregation tables and number of data sources, to satisfy the data content requirements. The operational requirements usually contain requested performance characteristics, including ETL duration range.

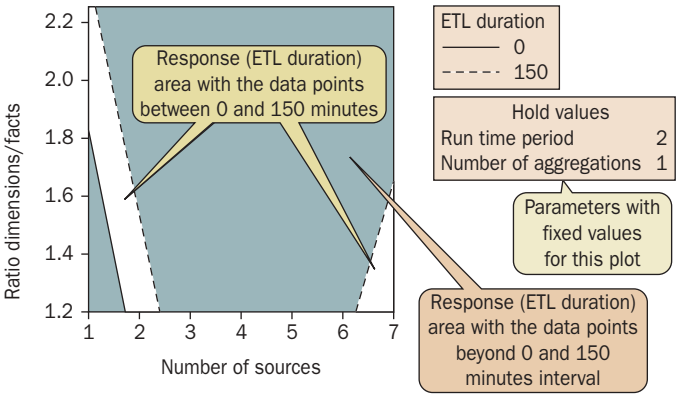
An architect applying the model can roughly predict whether the database defined by the data content requirements can be loaded within the required ETL operational window. When the answer is negative, either the requirements have to be relaxed (through negotiations with the customer) or additional effort and cost have to be built into the project plan to address the conflict.

The second purpose comes at the design phase of the project. The model is used to provide guidelines for the software designers working on the initial database schema variations. After the requirements are defined, the designers start developing the database schema. Typically, for large and complex projects, a few competing schemas satisfy data content requirements. The schemas can differ in several ways, including data primitives distribution among the staging and levels of normalization. Note that in data warehousing, the relational database schema should not be as strictly normalized as in typical transactional systems.

Designers applying guidelines derived from the RSM model can quickly identify the database schema that satisfies both data content and operational performance requirements. For example, a database schema for the data mart extracting data from two sources and having five dimension and four fact tables should not have more than one aggregation table to get ETL load time to less than 150 minutes.

Figure 3. **Contour Plot for ETL Duration Against The Ratio and Number of Sources**

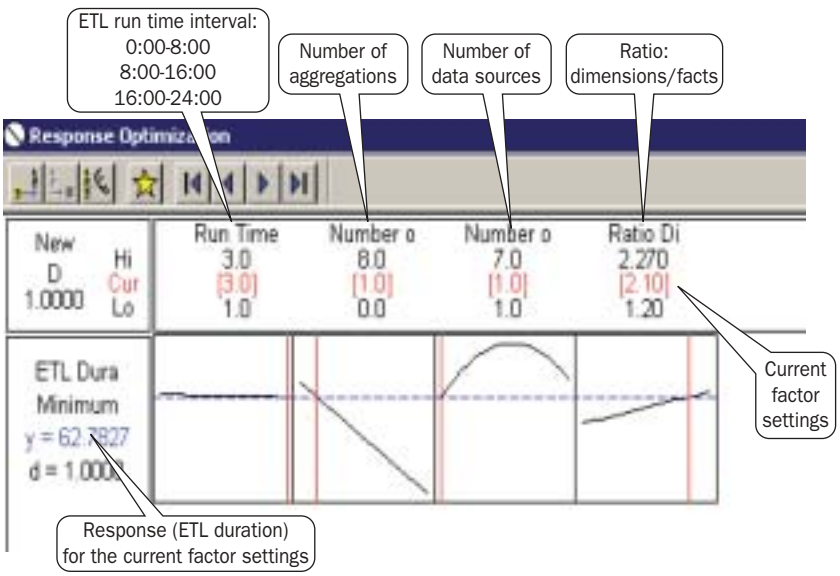
ETL = Extraction, transformation and loading



RSM Wins

Statistical analysis of historical software performance data at early development stages allows us to make predictions on the performance of the future

Figure 4. **Minitab Response Optimizer For Data Mart Database Design**



software components and help find the most optimal designs. While full and fractional factorial analyses provide a foundation for the research, we found RSM to be the most useful for investigating correlations and optimal settings for the data mart database design.

Of course, the limitation of this approach is that it is valid only as long as the hardware and operational software architecture remain fundamentally unchanged from the start of data accumulation. When the underlying data warehouse architecture is changed significantly, the study has to be conducted again on the historical set collected on the new system.

By applying identified optimal regions and a Minitab based response optimizer model to the practical data mart designs, we were able to estimate project risk. We could then mitigate the project risk by screening out nonoptimal design options at early project stages. As a result, we reduced the design cycle time and improved ETL performance characteristics of the final product.

REFERENCES

1. Vladimir Boroditsky and Christine Molinare, "Six Sigma Analysis in Data Warehouse Design, ETL Performance Prediction Model," *DM Direct*, June 3, 2005.
2. Ralph Kimball and Margy Ross, The Data Warehouse Toolkit, second edition, Wiley, 2002.
3. Boroditsky and Molinare, see reference 1.

WHAT DO YOU THINK OF THIS ARTICLE? Please share your comments and thoughts with the editor by e-mailing godfrey@asq.org.