Saudi Aramco Exploration & Producing Data Warehouse: A Case Study

Muhammad S. Khakwani

Abstract— This paper focuses on the challenges faced when building a data warehouse dealing primarily with scientific or engineering data. Successfully applying data warehousing concepts to an engineering-specific database poses some challenges which vary from the usual textbook examples. This, coupled with the introduction of a new database technology (data warehouse), into a large organization with a mature set of users raised an interesting set of challenges. Some of these are highlighted and discussed using the E&P data warehouse at Saudi Aramco as a case study.

Index Terms— Data warehouse, Engineering data, Oil and Gas, Saudi Aramco.

Introduction

Saudi Aramco has been in the Oil and Gas business since the 1930s. It is committed to efficiently producing energy for the world's needs. Like all successful business entities, Saudi Aramco strives to meet the demands of a dynamic market in a cost-effective and timely manner.

Exploration and Producing (E&P) is a key business line within Saudi Aramco. The nature of E&P applications and data tends to be scientific, as its user community is comprised of professionals such as geologists, geophysicists, reservoir engineers, petroleum engineers, and drilling engineers. Most oil companies tend to have dedicated information technology experts who specialize in understanding and meeting E&P user needs. Saudi Aramco also has such a computer center dedicated for this purpose.

Saudi Aramco has a large, sophisticated, in-house developed E&P database which it has maintained for over 25 years. It is an integrated database which is rich in attributes, and contains information about the exploration and petroleum engineering business.

Manuscript received March 27, 2007. Saudi Aramco Exploration & Producing Data Warehouse, A Case Study;

M. Khakwani is with Saudi Aramco, E&P Database Services Division, Dhahran, Saudi Arabia (phone: +966 3 872 9382; e-mail: muhammad.khakwani@aramco.com)

Sophisticated as it may be, it is a relational database and contains normalized data structures to mainly support day to day operations and reporting needs. It does not lend itself easily to analytical processing.

Business is constantly changing. New technologies are being applied as to how exploration for oil and gas is conducted, how the wells are drilled, maintained, and produced and how the reservoirs are managed. These changes must also be reflected in the applications supporting these business functions. Over the decades, we have seen many changes in the IT technologies used to implement software systems. This is also true for the underlying database technologies. With business changing rapidly, asset managers were asking increasingly complex questions. The information was becoming harder to obtain from a relational database. Managers wanted their answers quicker, in order to make the right decisions. Support staff was getting busier and finding it impossible to keep pace with the requests.

It was time to think of a new approach. Could data warehousing technologies be leveraged to help the situation? Almost all the data warehousing cases that were studied were done in the financial, telecommunications or retail sectors. Theoretically, data warehousing seemed promising, but applying it to the E&P sector of an energy company dealing with engineering data raised some uncertainty. Applying any new technology adds an element of risk, especially when dealing with complex scientific (engineering) data structures containing years of historical data. However, in order to facilitate the E&P users' needs for high speed analytical information, we embarked on the design, creation, and population of a data warehouse instance to organize data in a new way that was conducive to analytical processing, hoping that it would result in empowering the end-user.

This paper highlights the process that was followed, the challenges faced, and how some of them were overcome. Things are noted where they could have been done differently. Where possible, it tries and highlight how implementing a data warehousing project for engineering

data is similar to, yet different from, the typical examples found in the literature.

Early Training

Formal training was very important in clarifying the concept of a data warehouse, and removing misconceptions that people had. For this initiative, the database designers, developers, data managers and some key users were trained in data warehouse design very early in the project. Even though the database designers would be responsible for dimensional modeling and implementing the star-schema structures, it was important that everyone understand how dimensional models were put together. Many of the senior developers admitted that attending training for data warehouse design cleared up many misconceptions they had about the capabilities of this technology. Formal training on data warehouse concepts and design techniques was a great way of generating interest and enhancing understanding for everyone involved in the project.

Having attended the training, key personnel were able to better understand data warehouse terminology such as dimensions, facts, star schemas, and the relationships between them. Developers benefited from the training in areas such as indexing schemes, query optimizing hints, concepts of drill-downs, rollups, and data cubes. It was now easier to have conversations about the data warehouse since everyone was on the same page, and had a better understanding of the overall concept, the terminology, and the various techniques used in the different stages of building a data warehouse.

Project Definition and Scope

It is always easier to reach the destination, if one knows where to go and what to expect along the way. For this undertaking it was important to have a concise roadmap, in order not to jump in and start haphazardly. This task is best performed by someone who has actually done it before. Not having experience in-house, consulting advice was obtained from experts with experience in Petroleum Engineering data warehouse projects. A key objective was to validate the assumption that data warehousing concepts could be applied to engineering data.

Building dimensions looked similar to classic data warehouse designs. Objects such as wells, rigs, reservoirs, and plants could be easily defined. A time dimension is universal. But the engineering and scientific facts did not follow the classic definition of a data warehouse fact. They would not be additive in any but a few strict dimensions. Data like oil production volumes, pressure declines, water injection rates, well counts, etc. was required to perform the required analysis. With the help of the consultants and much research, it was decided that the pre-calculated, de-normalized data in the warehouse and the heavy use of indexing within star schemas for performance, would indeed give the E&P users the information they needed.

After completing this phase, it was easier to set expectations and get a “buy-in” from management.

Identify Objectives

From the onset, one must articulate the problems with the current system and clearly state what one hopes to achieve. In this case, E&P had a large, integrated database, so data consolidation from multiple sources (or creating a master data store) was not of primary concern. The objectives included:

- Performance gains for complex analytical queries
- Reduce the complexity for building analytical queries, thus empowering end-users to find information with minimal assistance
- Create a platform to support a new generation of Business Intelligence tools
- Record time-variant data for those performing analyses, ensuring they get better and more accurate information about the forces driving their business

Players and Roles

The user community was already very familiar with their interactions with the mature E&P relational database model, but the idea of a “data warehouse” was alien to most. Clearly identifying new roles and responsibilities was essential. Some of the notable players included Business systems analysts, data modelers, ETL programmers, DW educators, end user application developers, data stewards and DW quality assurance analysts. Key individuals from the IT and user communities were identified who would be needed to effectively carry out the responsibilities for each of these roles.

Warehouse vs. Data Marts

Key representatives from the major business areas such as Drilling Engineering, Reservoir Engineering and Petroleum Engineering were interviewed with the goal of identifying key business processes that would benefit from data warehousing. The conventional literature recommended large corporate data warehouses be planned and built. Then add on the specialized data marts. After reviewing the E&P user requirements and expectations, it was decided that a series of smaller, custom-built data marts per business area was a better solution. While they all used common dimensions, the data structures tended to be very business specific with only small areas of overlap. Dimensions would be shared (i.e., conformed) across data marts. The design team

started with designing the dimensions and two data-marts; one to measure drilling time performance, and one to analyze reservoir production.

Dealing with Engineering Data

Requirements for engineering data can often differ from financial or retail sales systems. Engineering deals with a lot of estimates. Two engineers could have different formulas for computing their estimates for two sections of the same reservoir. However, the data warehouse ETL routines should calculate information related to the reservoir in one standard way. This actually provides a good opportunity to standardize. At the same time it can lead to some passionate discussions.

Oil and Gas production analysis is very complex involving queries spanning various time periods, taking into consideration multiple parameters and generating statistics, all of which are used to manage key corporate assets. The focus was to provide information which would help reservoir managers do a better job of analyzing production trends. For example, there were two main fact tables. The first was the well-reservoir-month fact table. Each row would hold data on the monthly volumes of fluids produced, the percent of water vs. oil, the calculated daily rates based on operating days and calendar days, change in rates based on last month, last 12 months and last 18 months, and cumulative volumes produced from all time. The information was not in the actual volumes and rates, but on the trend analysis of how those rates of change varied over the past 2-3 years. A similar fact table was needed at the (oil) field-reservoir-monthly level. Several of the facts could be rolled-up from the wells table, like adding all the monthly well volumes, however most of the attributes on a row had to be recalculated in the ETL process. These attributes include active well counts, average producing rates, inactive well counts, field level water cuts, etc. All of this leading to answer important questions about where to drill to maximize production and monitor which wells need attention.

There were situations where totals and cumulative values were not as simple as the formulas and functions provided by standard data warehousing tools. One such example was calculating total production from wells, and returning the number of wells which were summed. It seemed straightforward until the engineers looked at the numbers and decided that they did want to count the production from *all* the wells, but that the well count should *not* include wells from a certain region which produced less than a certain number of barrels of oil – they considered the wells to be non-contributors. So, even though we needed the production from *all* the wells, we did not count them *all*.

Users in drilling management were interested in measuring which service companies and rigs were the most (or least) efficient at drilling wells. They wanted to identify if drilling problems were limited to certain geographical areas, service providers, types of wells, or rigs. They needed optimization information such as which rigs were good at drilling horizontal sections, which rigs record the least lost time, which rigs have the longest (shortest) rig move times, etc.

Particularly important is handling the time dimension, since dealing with engineering data can be for events that span a considerable length of time. For example in the drilling data mart, there are drilling events which can span hours such as waiting on materials or lost circulation; or in the case of the production data mart, events can span months when evaluating the oil / water ratios from reservoirs, or even years when dealing with tracking fluid movement through rocks. The challenge then lies in *correlating* the information to present the overall picture.

Be prepared to handle many such exceptions when dealing with engineering data. A key point to always keep in mind; in the end information from these business processes will integrate because they are all parts of the overall E&P process. So you need to have dimensions which are *integrated*, and avoid building silos of statistics.

Risk Elements

Every project has risk associated with it. For this project, one of the major concerns was whether it would be better to spend time and energy working on enhancements to the existing database rather than embarking on a completely new initiative. Some designers and developers had a tendency to stay in “familiar waters” and felt they could achieve the same objectives with enhancements to the existing database such as creating various de-normalized structures and a “business layer” of views. Another challenge was to clearly communicate to the end users exactly what they could expect at the end of the activity.

It was important to identify the risk being undertaken and be upfront about the alternatives available. Uncertainties were identified and communicated to management. If management is to support an initiative, then one has the responsibility of making them aware of the choices.

Requirements Definition

A key area, and perhaps the trickiest, was being able to define what was of truly of interest to the user. It is important to coach the user to think in terms of providing requirements that enhance his business processes and not just solve short-term problems. From these requirements, derive the data attributes needed to build the warehouse.

Focus for Warehouse Requirements Gathering

Designers and analysts who have worked on projects gathering requirements for on-line systems, find it difficult (perhaps *different* is a better way of putting it) eliciting requirements for analysis. Some factors to pay attention to when collecting requirements for data warehouses are:

- What measures or facts are needed for the management of assets?
- What are the different ways users can query these facts?
- What is the source of the data elements?
- What is the time grain involved in the data elements?
- What transformations are required to produce the facts?
- What is the quality of the source data?

Approaches of Requirements Gathering

A combination of interviews and reviewing existing reports was used to gather the requirements. Conducting interviews with key-users gave an opportunity to explain the concept and objectives of the data warehouse project again, as well as prompting them for the information they used to manage their assets. Reviewing their "summary" types of reports (monthly, quarterly and yearly) helped identify the long-term trend analysis that they were using for asset management. There was no focus on the daily or operational reports. Instead the focus was on charts and reports being produced in support of trend analysis. This was quite often in spreadsheets.

It was intensive work to get through this phase. One needs to be open to statements like "Well, if you give me this, it would be nice if you could derive that also and put that next to it". The assumption was that being receptive to fulfilling such requirements would add value to the warehouse and get users more involved with the project. In retrospect, this could have been done a bit differently. The first releases of the data marts had a lot of work involved in building the infrastructure and bringing people up to speed on how to use them. Users were a bit disillusioned at the time it took for implementation. For future data mart designs, the focus will be on getting the key statistics in the first release, and addressing the wish list in subsequent releases.

Design Standards and Infrastructure

A data warehouse needs a new set of design standards before implementation can proceed. These can be enhancements to existing standards within an organization and should address:

Set up new database instances for warehouse testing and production.

Oracle RDBMS are used for databases at Aramco. The database setup and initialization parameters are different for a warehouse database instance as opposed to ones for an OLTP database instance. In order to leverage the technology built into the database engine to recognize and tune star-schema queries, new database instances are required.

Update standards to differentiate data warehousing objects.

This ranges from naming conventions for tables and views, to packages, public synonyms, and database links. It is alright to build on the existing standards for OLTP databases, but it is not a good idea to force them on a data warehouse where they don't fit. If something is different, it is an opportunity to show it distinctly. For example, a separate naming standard was created for views accessing the data warehouse. There are views in the OLTP databases that allow users to access data in the warehouse via database links. The naming standard helps to quickly identify these views and will help in identifying performance problems, and comparing execution plans, especially those that span multiple database instances.

Standardized procedures for performing ETL.

This tends to be one of the most complicated areas technically, involving extraction and transformation of data, and subsequently loading into data warehouse structures. The E&P ETL is very complex because of the engineering and scientific formulas involved. Given the size and complexity of the database, implementing the ETL layer was a daunting task. Several off-the-shelf tools were evaluated but it was very difficult to implement the complexity. They may have worked well for the retail or financial industries, but for upstream oil and gas engineering, it was not a good fit. In the end, customized scripts and infrastructure was built to handle the data as required.

Define procedures for performing Data Quality checks and notifications during the ETL phase.

This was another area that needed special attention since our data dates back a long time. Engineering data often depends on accuracy and needs to look at the entire history. For example, when reporting cumulative production for comparison to original reserves estimates, one needs to look at the entire history of a well, sometimes as much as 50 years. Data quality in any database varies over time and business rules also change. As with any data warehousing project, there were problems with data quality ranging from missing values to those outside acceptable ranges. There are decisions one needs to make on a case by case basis and resolve them. It is imperative to have a mechanism to:

- Identify the problems with data quality
- Report problems with data quality

- Resolve problems with data quality in the source data.

Implementation and Feedback

Going to production with smaller data marts helped to reduce the risk and address problems before proceeding with the next data marts. This approach has worked well. The drilling and production data marts were in areas where the users needed data urgently. This gave an opportunity to implement in an area where developers and users were eager to test and deploy applications. It is important to have a mechanism for feedback. Unlike an online transaction database, where one can count the records being inserted or updated, it is harder to monitor the data warehouse to see if the users are getting the required results.

The developers were quick to re-engineer some of their performance problem applications to use the data warehouse. This became the focal point as performance gains were realized and implemented, more people used the reports. As more people used the reports, data was quickly quality checked and fixed.

However, these changes were largely transparent to the user. To properly leverage the capabilities of a data warehouse, a front end Business Intelligence tool should also complement the data mart release. That is an ongoing project.

Documentation and User Training

Documentation is an important part of the implementation phase whether talking about technical or user documentation. User-level documentation should be provided along with the rollout. It needs to be detailed, business-oriented, have relevant examples, and also be quick and easy to reference. Several techniques can be used here such as creating training web-sites and distributing quick-cards.

This is a key activity which tends to get ignored as manpower and resources run thin and developers want to concentrate on the next release instead of documenting the rollout.

Outsourcing this work to experts who can develop customized course-materials, communicate with, and train the users, can be a big help. This phase was identified in the original plan. One avoidable mistake is to prepare for this step far in advance as instructors need time to understand and prepare materials for this activity.

Conclusion

The concept of a data warehouse can be successfully extended to engineering and scientific data. However, be prepared to handle situations which are not covered in the standard data warehouse courses. Certainly, when dealing with earth models, predicting trends, and analyzing oil and gas production, one has to be prepared for handling situations not described in textbooks. However, it *can* be done and in the end is worth it.

The first data mart released, based on drilling data, has been well received. We have successfully packaged easy to query, meaningful facts in the data mart while using conformed dimensions across a broad data warehouse concept. Users can quickly obtain key drilling performance measures. As more and more users and developers see the ease of use, requests for enhancements continue growing. We are in the process of designing new data marts and forging ahead with our strategy. Incidentally, this leaves increasing less time to do the needed documentation and conduct formal user-training.

Acknowledgment
Profound thanks to Wayne Wolfe, Petroleum Engineering Systems Consultant at Saudi Aramco, for his guidance and insight without which none of this would be possible.

References

[1] Bert Scalzo, “Oracle DBA Guide to Data Warehousing and Star Schemas,” *Prentice Hall, New Jersey*, 2003.

[2] Ralph Kimball, “The Data Warehouse Toolkit” *John Wiley and Sons, New York*, 2002.