Life Cycle of a Data Warehousing Project in Healthcare

Ravi Verma, Jeannette Harper

ABSTRACT

Hill Physicians Medical Group (and its medical management firm, PriMed Management) early on recognized the need for a data warehouse. Management demanded that data from many sources be integrated, cleansed, and formatted. As a first step, an operational data store (ODS) was built and populated with data from the main transactional system; encounter data were added. The ODS has served its purpose well and has whetted management's appetite for more information and faster, more reliable access, all in one location. PriMed hired Annams Systems Consulting (Annams) for this effort. A team was formed, made up of consultants from Annams and members of PriMed's information services (IS) team. The "classical" approach is being taken: enhancing the ODS, which is largely normalized in structure, and integrating data from various sources, along with enforcing business rules. The team is designing and implementing data marts and a "star schema" style of data modeling—a useful tool for management to evaluate results before investing further.

KEYWORDS

- Operational data store (ODS)
- On-line transaction processing system (OLTP)
- Multidimensional data cube
- Metadata
- Data marts
- Start schema design

PriMed Management, Inc., is the management services company for Hill Physicians Medical Group, an independent physician association with over 380,000 members in Northern California. PriMed is headquartered in San Ramon and has regional offices in Sacramento and Redding. Annams System Consulting is a California-based corporation specializing in IT consulting, planning, and management. The company has extensive experience in implementing

decision support solutions for manufacturing, healthcare, distribution, customer relationship management (CRM), and e-commerce.

Problem Statement

The organization's problem is to most effectively and efficiently make use of data, that is, to turn them into information so that managers and physicians can base business decisions on facts. In 1996 PriMed replaced its transactional claim system and commenced an effort at data warehousing. The transactional system had been used for the processing of claims and the input of daily authorization requests for medical procedures. Reporting from the transactional system slowed the system down, interfering with claims and authorization processing; reporting was cumbersome and time consuming and mostly focused on daily factual reporting, not on analysis or decision support. With the new system, a series of scripts was written to download much of that information onto an Oracle platform with groups of tables that analysts could view and query against in Microsoft Access or, if more highly trained, in Oracle SQL. The original expectation of this newly created data warehouse was that monthly data would be available to create reports around claims processing, to track and trend data, and to respond to questions by senior management.

History

The original users of the data warehouse were staff in the Health Data Analysis Department—a department consisting of four staff members at the beginning. As the group learned to work with the data warehouse, they began to see where the errors were in the processing, where fields were repeated in tables—in effect, all the "warts" of a newly developed system. The output grew as did demand for more. More analysts were hired; more people were directed toward the maintenance and improvement of the data warehouse itself. At PriMed, we created a position for a database administrator and two staff fully dedicated to creating additional tables to be used for analytical purposes, streamlining tablespace, and training more analysts on the nuances of how the data would or should look.

The Health Data Analysis Department became very effective at responding to typical requests for information. A series of standardized monthly or quarterly reports was initiated: (1) physician compensation analysis, (2) physician profiling, (3) utilization (facility) reporting, (4) disease state management reporting, and (4) analyses of contract viability. As the department grew to its present number (ten), it became more and more evident that the same questions were being asked and answered. It also became apparent that many of the data necessary to answer questions were not strictly from the transactional system.

Islands of Information

Although the department became very adept at answering claims questions, more and more data were floating around that should have been captured and incorporated into the data warehouse. Examples were pharmacy information received from a number of our health plans, as well as laboratory and other information from a subcapitated lab provider. The data needed to be cleansed and incorporated into the data warehouse. That effort began in earnest in late 1998; we realized that the next phase of the data warehouse initiative needed to begin if we were to get the maximum usage of the data already captured.

We began a project to look at our current data warehouse, talk with users on the technical side, talk with the people who actually use the data that have been analyzed, and determine a strategy to get our current system upgraded, more powerful, and more user friendly. The first step was to become educated on the theoretical aspects of data warehousing.

Current Process

On-line transaction processing (OLTP) is a transaction-based, most often real-time system used for data gathering and reporting. In PriMed's business, OLTP is used for claims and authorization processing. An OLTP system is different from a data warehouse in that a data warehouse generates no data of its own and relies on the OLTP systems for data feed. A data warehouse lays out data from the OLTP systems in a way that facilitates reporting. Some features of a typical OLTP system are as follows:

- *It is function oriented.* It is always possible to assign a name to an OLTP system that clearly identifies what that system is trying to achieve. PriMed's is a Claims and Authorizations system; other industries have, for example, Airline Reservations systems, Credit Card Application systems, and so forth.
- *Business rules are embedded.* One useful and important feature of an OLTP system is that business rules are embedded in the application (database). At PriMed, certain claims may be paid only if an authorization is present, for example.
- *Transaction-level information is present.* At PriMed the transaction is a Claim line, indicating the procedure performed, the date, and the patient, as well as many other pieces of information.
- *It is instantaneously accurate.* The information is accurate at the moment the data are entered.
- *It is always available.* Most systems are available seven days a week, twenty-four hours a day, fifty-two weeks a year. In our model, the system is available about fifteen hours a day, seven days a week.
- *Performance is outstanding.* Most OLTP systems cater to operators who are on-line with users. A less-than-three-second response is the norm. Customer

service uses the OLTP system for querying; depending on the volume, the response time could be longer than three seconds.
- *The information is complete.* OLTP systems are usually complete with regard to the functional areas to which they cater.

This is not an exhaustive list of all the features of an OLTP system, but it will help us to understand data warehousing better in the pages to follow.

Getting to the Data Warehouse

A data warehouse is a database specially designed to facilitate business analysis and decision support. According to Bill Inmon, "A Data Warehouse is a repository where data is kept in a subject oriented, integrated, time variant and non volatile manner to facilitate decision support."[1] Data from operational systems are extracted, cleaned, and loaded into a data warehouse. Some important features of a data warehouse are as follows:

- *Equal importance of OLTP (the source) and the data warehouse (the destination).* People often ask, "Which is better, the OLTP system or the data warehouse?" They are equally important, but they have different audiences and different uses. Data are extracted from one or more OLTP systems and loaded into a data warehouse, where they are set up to facilitate business analysis.
- *Subject orientation.* Data warehouses are always oriented around subjects and are created to answer specific questions, as opposed to OLTP systems, which are oriented around functions. At PriMed the OLTP system deals with the functional areas of claims, referrals, and authorizations. Data marts are subsets of data that pertain to a specific business area. That specificity makes them smaller in size; a collection of data marts makes up the entire data warehouse. The data marts in our case will be built around the subject areas of Primary Care Physicians (PCPs), Members, and Claims. This shows that data warehouses may require only a subset of the information contained in the OLTP systems, and there may be a need to continuously enhance or augment the data warehouse to provide answers to more business questions.
- *Integrated data.* One of the most critical features of the data warehouse is that its information is always integrated. As stated earlier, the data warehouse can be the destination of data from more than one OLTP system. Sometimes, for example, a physician and a hospital may have different ideas of when an inpatient stay began. But when data are brought to OLTP, they must be cleaned, formatted, and reconciled at a single level of understanding. At PriMed, claim-related information is spread across systems like IDX, Encounter data, and Financial data, but all these data segments pertain to a member. For the purposes of analysis it is important that data from all these sources be integrated.
- *Time variance.* Data warehouses are always time variant; certain information about a member may be contained only once. For example, if a

member changes addresses, the OLTP system keeps only the latest address, but the data warehouse keeps the history of the member's moves from one address to another.

- *Nonvolatility.* Unlike an OLTP system that updates, inserts, and deletes entries, a data warehouse is not volatile. A typical data warehouse experiences only two types of operations: inserts and access to the information.
- *No instantaneous accuracy.* A data warehouse may contain data that are accurate as of the preceding day because of the way data are loaded. Unlike OLTP systems, in which the data are being inserted or updated in small transactions on a real-time basis, a data warehouse (depending on the frequency of load, which may range from a few minutes to a few days or even months) may not be accurate instantly. At PriMed, the data warehouse is currently being refreshed on a monthly basis.

The First Step

The beginning of the data warehouse for PriMed was an ODS. As mentioned earlier, a change in the transactional system (the OLTP) necessitated a change in the way reporting was accomplished. The series of scripts that was run monthly from that transactional system to create a set of Oracle tables was really the beginning stage of an ODS to the data warehouse. It began with those claims and authorizations only but soon added other islands of information when they became available. The ODS served to integrate those data as well as implement business rules. Data were coming in from laboratory companies that did not identify providers in the same way PriMed did. Research and programming were needed to get this provider-identification logic consistent. This type of programming needed to take place in a staging area so that data going into the ultimate data marts would be as clean as possible and, therefore, the resulting analyses would be as accurate as possible.

Where We Need to Go

Whereas a data warehouse can be at the enterprise level and serve the needs of the whole organization, a data mart can be at the department level and serve a small group of people. A data mart strives to analyze a particular area; a data warehouse can comprise a few or many data marts. Obviously, it is less expensive and less time consuming to build a data mart than a data warehouse. In addition, taking smaller chunks of data allows the building to be completed more quickly, showing parts of a total project in a more timely fashion. The time involved in putting together an entire enterprise data warehouse may make it obsolete before it even gets off the ground.

At PriMed the team decided to take the data mart approach for building the DSS. After the initial feasibility study, the team made a priority list of subject areas with which help was needed for their analytical needs. They chose

the following subject areas for the building of data marts: Membership, Revenue, Primary Care Physicians, and Claims.

Once the first data marts were decided on, the team needed to be more educated on the entire concept of moving from analysis within the ODS structure to something more flexible, consistent, and accessible (eventually) to more end users.

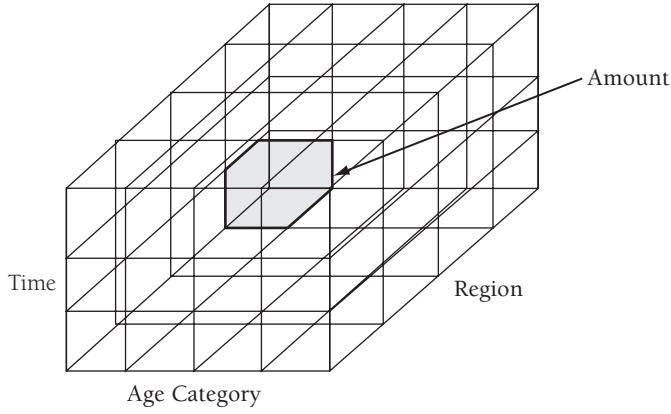
On-Line Analytical Processing (OLAP)

OLAP is all about a cube—a multidimensional data cube. E. F. Codd, the father of relational database systems,[2] originally came up with guidelines for OLAP systems. Unfortunately, there are no standard methods, such as SQL (structured query language) in the case of relational database management systems, for accessing data from an OLAP system. This is the layer where data are picked up from a data warehouse or a data mart and presented as a multidimensional view of information, that is, a data cube. In a relational database the information would be laid out as a series of columns and rows, as in Table 1. The same information in an OLAP system would be presented in the more cubelike multidimensional view shown in Figure 1.

Table 1. Relational Information

Time	*Region*	*Age*	*Claim Amount*

Figure 1. Multidimensional View



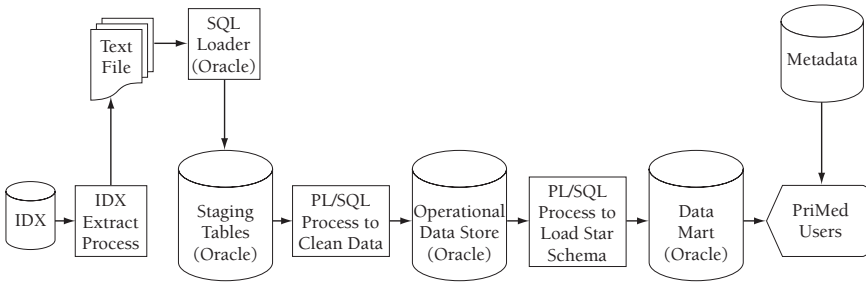
In the figure, the information has been arranged as a cube with the sides representing Time, Age Category, and Region; the subcells represent the amount incurred on behalf of the members. Sides of the cube are called Dimensions, and the information contained in the cells is called the Fact. There are always more than three dimensions in any business, so keeping data in this type of cube facilitates the analysis of information. Because it is critical to allow the analysts, managers, and executives to view the data, an end-user tool or OLAP tool helps interface with the multidimensional data cube explained in the previous section. The view has to be dynamic to allow the user to analyze the same data in more than one way. These tools are much more sophisticated than the usual OLTP reporting tools that are fixed in nature. A few subcategories within OLAP are as follows:

- *Relational OLAP.* People have managed to use the existing relational database management systems (RDBMS) technology for OLAP analysis by performing dimensional modeling. Although RDBMS technology has been around for a decade, it has matured to a great extent. Consequently, scalability and performance are good with most of the popular RDBMS products. It is relatively easy to find professionals who are proficient at understanding and implementing RDBMS products like Oracle, Sybase, Informix, or DB2. PriMed uses Oracle for housing the ODS and the data marts.
- *Multidimensional database.* These databases are designed for OLAP analysis and are different from relational databases in the way they keep the data. As we saw in a previous diagram, relational data are kept in a two-dimensional form as rows and columns, whereas multidimensional data are kept as a cube. PriMed is currently using Oracle Discoverer, and the goal is to move toward Web-based OLAP.
- *Web-based OLAP.* With the increasing popularity of the Internet and intranets, companies are finding it easier to deploy their information on the Web. Web-based OLAP tools allow users to perform their analysis using the ubiquitous Web browser.

A final, important feature of OLAP systems is metadata, or data about data. Most of the OLAP tools map their own logical schema to heterogeneous physical data stores, access the data, and perform any conversions necessary to present a single, coherent, and consistent user view. In this way, the data warehouse team can begin to understand where particular data come from, how they were entered, who owns them, how often they are updated, and so on.

The technical architecture of the data warehouse for PriMed is shown in Figure 2, which should clarify the steps discussed so far, from the sources or OLTP systems through to the true data warehouse and OLAP systems.

Figure 2. Technical Architecture



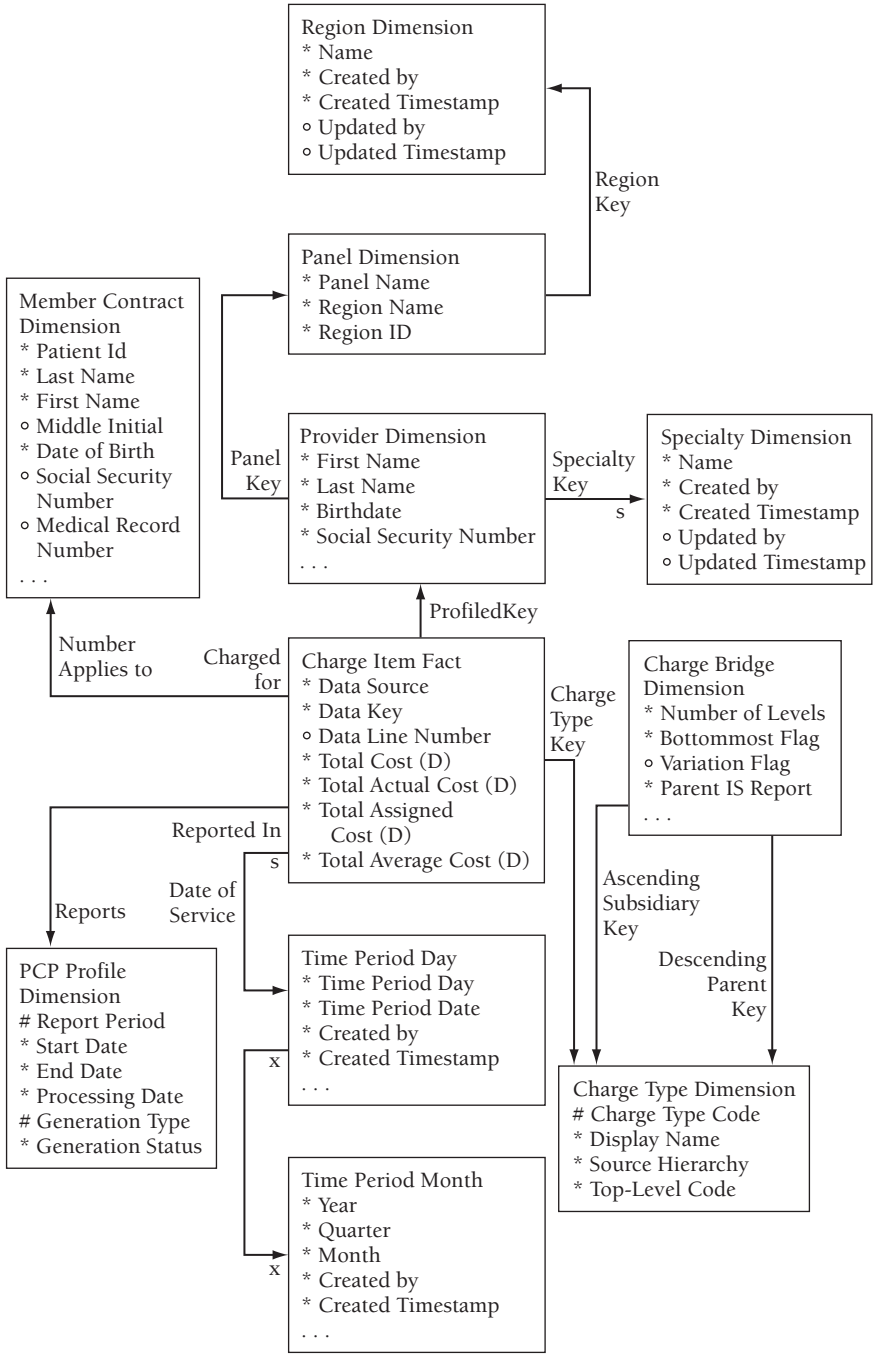
What a Data Warehouse Looks Like: A Multidimensional Conceptual View

Most members of the IT community are comfortable with the techniques of entity relationship (ER) modeling. Most have also noticed that users have a difficult time understanding the most accurate ER diagrams that represent the business that is truly theirs. The reason for this is that users have a different perspective of the business; they see it through a multidimensional conceptual view. Star schemas or dimensional modeling are the answer to a multidimensional data cube in an OLAP system. For example, at PriMed there is often a need to view the cost incurred by PCPs by region and by age categories. This is a multidimensional view with these dimensions: Age Category (10–15, 25–35, and so forth), Region (Shasta, Sacramento, East Bay, and so on), PCP Name (a particular PCP), and the Fact (Cost). The OLAP system would show the cost incurred for some period of time (a day, week, month, or year): one might drill up to see the monthly amount or drill down to see the daily amount. In this way, an analyst could slice and dice the information to look at the business issues from many different perspectives. OLAP systems present the data in a multidimensional view that has arms (or dimensions) for Region, Age Category, PCP, and Time, and the cell of such a view would contain the Dollar Amount.

In addition to the slicing and dicing that a multidimensional model can produce, the ability to determine aggregate data is also an important feature. OLAP systems are easy to use for determining aggregates at different levels in the hierarchy of a dimension. If Profit is stored at the days level, it is easy to find Profit at the week, month, quarter, or year level by performing aggregation "on the fly," that is, without creating a standard report.

A star schema model (called such because it looks like a star) preserves the flavor of a multidimensional data cube, and it is possible to implement it in an RDBMS like Oracle or DB2. Figure 3 shows an example PriMed used in profiling PCPs. There is a large table in the middle (Charge_Item_Fact) that contains information about the total cost, total actual cost, total assigned cost,

Figure 3. Star Schema



and total average cost. The fact table is surrounded by dimension tables, namely, member contract, PCP profile, time period, charge type, and so on. Dimensions are connected to the fact table through foreign keys, which enforce the integrity rule (or constraint) between a fact table and a dimension table. For example, it would enforce that each claim belong to one and only one region, or it may enforce that every member have one and only one gender. This schema is not as efficient as a true multidimensional data cube would be for OLAP analysis and query, but it comes close because most of the queries involve the joining of two tables—a dimensional table and a fact table.

What Is Left to Do?

Currently, Annams is assisting PriMed with three separate data warehousing projects. The first is a data mart for providers that has had the challenge of integrating three different Access databases with information about participating (or network) providers. The three systems have been integrated, and the effort currently is to finish with the cleansing of "dirty data" and move into a maintenance period with this data mart.

The second project is a data mart for Revenue and Membership data. This project was the first to undergo a rigorous planning cycle, including the detailed understanding of source systems. The source systems include the OLTP system for membership and many Access databases and Excel spreadsheets for revenue. It has also included the OLAP tool, Oracle Discoverer, that currently the Revenue Analysis Department is using for standard queries. Once the data mart is stabilized, this end-user query tool will be delivered to the managers involved with further analysis of the information.

The third current project is a data mart specifically designed to look at PCP profiling. It is, by far, the most complex yet and has matched our strategy to move from something fairly simple and easy to understand to the more complicated and complex information. A full project plan is about one-half complete and includes the most rigorous of user-requirements documents yet. Once defined, the team has begun the design modeling process and, again, analyzing the source information systems and documenting their findings. This data mart is scheduled to be in a pilot phase by the end of this year; the team is working hard and the users-to-be are excited about seeing the product in a live environment.

References

1. Inmon, W. H. *What Is a Data Warehouse?* Sunnyvale, Calif.: Prism Solutions, Inc., 1995.
2. E. F. Codd Associates. *Providing OLAP to User Analysts: An IT Mandate.* Sunnyvale, Calif.: Hyperion, 1995.

About the Authors

Ravi Verma is a practice manager at Annams System Consulting, president of the Sacramento Oracle Users' Group, and an instructor at the University of California–Davis extension.

Jeannette Harper is director of health data analyses at PriMed Management. She holds an MBA in health service administration from Golden State University.