5-2012

Designing OLAP Cubes: A Teaching Case

Gregg Greer
Lubbock Christian University, Gregg.Greer@lcu.edu

Designing OLAP Cubes: A Teaching Case

Gregg Greer

Assistant Professor
Lubbock Christian University
Gregg.Greer@lcu.edu

Doctoral Student
Dakota State University
jggreer@pluto.dsu.edu

ABSTRACT

This teaching case discusses the process used to create a "proof of concept" On-line Analytical Processing (OLAP) cube for admissions at small, private, university in the mid-western United States. The case assumes a rudimentary knowledge of Data Mining, Data Warehouses, Data Marts and OLAP and attempts to put those concepts in a context familiar to university students. Students will use database views to design star schemas for a prototype OLAP cube. The learning objective of this case is to build an understanding of the design and use OLAP cubes within a larger data mining course. To enhance student understanding, the case includes several student activities.

Keywords

Online Analytical Processing, Data Warehousing, Data Marts, Data Mining, Higher Education

INTRODUCTION

In this case, the student will play the role of a consultant for a small university. The consultant will examine the current university student database structure to determine if the university can use the existing data to create an Online Analytical Processing (OLAP) cube. Since many terms are similar, practitioners must have a clear understanding of the terminology.

Data mining analyzes detail datasets to highlight previously unknown relationships in the data. Data mining uses past behavior to predict future behavior (Power 2000).

A data warehouse stores information a separate data storage area designed to provide information to support decision-making in an organization (Power 2000). Data warehouses allow simpler processing of data and can consolidate data from different time frames and across organizational units.

Data marts differ from data warehouses in their narrower information focus. A data mart will often concentrate on the information for a department and not on the information needed by the organization as a whole. Organizations can implement data marts more easily than data warehouses because of their small scale and lack of needed input from the entire organization (Chaudhuri and Daval 1997).

Like pivot tables in Microsoft Excel, OLAP "cubes" summarize data across different dimensions, such as time, allowing for ad hoc complex analysis and fast query retrieval. OLAP cubes can process a large amount of data in a short amount of time. Practitioners refer OLAP structures as "cubes" and represent them using "Star Schemas" (Power 2000). Data mining, data warehouses, data marts and OLAP cubes deliver a rich view of an organization's data. Data mining is most effective with detail data but OLAP cubes summarize data. Tension can exist between these two methods (Berry and Linoff 2000).

THE SCENARIO

As of the twelfth day of classes each semester, the university reports on the enrolled student population. Throughout the academic year, the university also reports on the prospective student population for the next Fall. Entering Freshmen often apply to multiple colleges so they might have an easier time selecting from several options (Dixon, Gribbons and Meuschke 2002). When they know which students will most likely to attend, admissions personnel can more efficiently allocate scholarship funds and other recruiting resources.

The complexity of the data makes ad-hoc reporting from the university's existing database challenging. The university uses a third party software package to process its prospective and current student information. The vendor provides a detailed database which does an adequate job of online transactional processing (OLTP). However, several issues make it difficult to use the database for reporting. First, the database contains live data. Universities measure their enrollment according to the data on the twelfth class day. However, enrollment changes throughout the semester as students add and drop classes. The university makes a copy of the database for twelfth day reporting, saving each database separately. Separate databases make it difficult to compare twelfth day numbers over time. Second, the university's database contains thousands of normalized tables. Users must know the data and table structures intimately to create even simple reports. The vendor remedies this

with a series of reporting views. A view, or logical table of information, points back to selected columns and rows on multiple tables. Reporting users can query views without having to account for the nuances of all the tables on the database. However, users may not know which view to use and views degrade query performance.

Creating an OLAP cube for the university would enable users to pull information the various twelfth day datasets and compile it into one area for easier ad hoc reporting. OLAP processes would allow the university to quickly summarize students by program and by semester and application status and then allow them to "drill down" to relevant details to determine trends and relationships which drive enrollment (Hackney 1997).

The university could use data mining tools to look at prospective students to determine the ones most likely to attend. Data mining techniques might also reveal the students most likely to be retained by the university. In this case, the student plays the role of a consultant tasked with designing a proof of concept OLAP cube to determine if the university has all the pieces in place to for a data warehouse or data mart. The consultant wants to help the university find the format that will most effectively provide the needed information.

Steps to create a "proof of concept" OLAP Cube

The consultant took several preliminary steps. First, the consultant requested permission to access the university's data. The university Provost and the Vice President for Technological Advancement expressed concern about off-site access of the university's data. All participants agreed that the consultant would have read-only access to the data and only access it through campus computers. Second, the university's Database Administrator granted access to one of the reporting databases. Third, university personnel installed Microsoft SQL Server Manger and Microsoft Business Intelligence Server on a desktop computer which they made available to the consultant in an otherwise unused office.

The consultant performed the following seven steps to create and evaluate the "proof of concept" OLAP cubes. First, the consultant examined the views on the reporting database to find available information. The consultant created preliminary designs of the cubes. Second, the consultant used this information to create a test version of the data. Third, the consultant designed the data warehousing views. Fourth, the consultant created a "project" in Microsoft's SQL Server Business Intelligence development Studio to access the test data. Fifth, the consultant created a new data source in the SQL Server database and designed a view of the data source. Sixth, the consultant defined an OLAP cube from the data source view.

Student Activity One: Address Security Concerns

This is an activity that should be completed by the student and turned in to the instructor for grading. The Family Educational Rights and Privacy Act (FERPA) governs student privacy rights. Universities often use outside consultants. Students should briefly respond to the following questions: What other security precautions might the university take when giving consultants access to student data? How can universities best balance the value of outside expertise and student privacy?

STEP ONE: ANALYZE VIEWS ON REPORTING DATABASE

The university's Database Administrator suggested using a view which had information on people in various states of admission and acceptance at the university. At the end of this paper, Table 1 contains an adapted version of the database view. A prospective student goes through three stages in the university's admission process. In the first stage, the student inquires at the university and receives information from the university. In the second step, the student applies to the university. In the third step, the university accepts the student.

STEP TWO: CREATE TEST DATA

The consultant created a randomized Microsoft Access database which simulated the information in the selected view on the university database. The database contained completely random records with fake identification numbers, fake names, and fake addresses. The consultant used no information from the university database to create the data in test database.

Student Activity Two: Complete Design of Database View

Students should add names and data types for the following fields. What naming conventions should the consultant use to describe the field names? What fields on the view seem unnecessary? What additional pieces of information should the consultant add?

People_Code_I	Prefi:	First Nam	Middle Nam	Last Nam	Suffi:	Preferred_Addre	Deceased_Da	Release_In	Address Line
P908567288	Ms.	Ginger	Hollie	Matthews		Parent/Guardian		Authorized	9019 Estes Street
P527385041	Ms.	Tiffany	Vivian	Glass		Local		Limited	9658 Bennett Street
P410458650	Mr.	Darwin	Marco	Randolph	Dr.	Permanent		Authorized	6195 Delgado Street
P442460827	Mr.	Chase	Mauro	Stephenson		Permanent		Department Authorized	1539 Spencer Street

Figure 1. Selected View from the University Database (Test Data)

STEP THREE: DESIGN THE DATA WAREHOUSING VIEWS

Analysts find OLAP cubes difficult to represent graphically because they contain more than three dimensions. Analysts often use a star schema to describe the structure of an OLAP database. A schema is a graphical representation of the structure of a database. Analysts call OLAP Schemas "Star Schemas" because of the design created when several tables point back to a central table (Power 2000). According to Hackney (1997), a simple star schema will typically contain the following elements. First, the 'heart' of the star schema is the facts table which contains the detail transactions from the transaction database as well as the 'foreign' key fields which will connect the fact table to the other tables in the star schema. Fact tables typically only contain foreign keys and metrics to summarize. Second, the dimensions of the star schema describe how the organization analyzes its data, often including such dimensions as time, place, constituent, or product. Dimension tables typically include descriptions, such as course names; hierarchies, to facilitate drill-down capabilities; and measurements, such as 'to date', or 'prior period'. Power (2000) suggests the following the steps for creating a Star Schema. First, define the process the OLAP cube will represent. Second, determine the granularity or degree of the process to measure. Third, determine the dimensions of the cube. Fourth, determine which facts to summarize in the cube (Power 2000).

Student Activity Three: Design a Star Schema

Design a star schema to represent the following process: inquiry, application and acceptance. Use the elements in Table 1.

Steps:	Student Description
Define the process represented by the OLAP cube.	
Determine the granularity or degree of the process measured.	
Determine the dimensions of the cube.	
Star Schema Diagram (See Figure 2.)	

STEP FOUR: CREATE A PROJECT TO ACCESS THE TEST DATA

The consultant created a project in Microsoft's SQL Server Business Intelligence Development Studio and connected it to an existing SQL Server Database. This allowed the consultant to create a data warehouse, with OLAP and data mining capabilities (Microsoft 2005).

Student Activity Four: Select Terminology

Review the definitions of Data Mining, Data Warehouse, Data Mart, and OLAP. The Administrators at the university requested a 'Data Warehouse'. Is this the right term, or does 'Data Mart' better describe the proposed system?

STEP FIVE: CREATE A NEW DATA SOURCE AND DATA SOURCE VIEW

The consultant added the test database as a data source to the project. The university can add multiple databases as data sources to a project. This would enable the university to tie to multiple databases containing twelfth day datasets from various periods. The consultant then created a *data source view* which allowed the project to access the information in the data source (Microsoft 2005).

Student Activity Five: Select Strategy

Students should answer the following question. How do data mining and OLAP cubes sometimes operate at cross purposes to each other? Can the university get the information they want without "data mining" per se? Would a different data structure suit them better than OLAP?

STEP SIX: CREATE AN OLAP CUBE

The consultant defined an OLAP cube using the *Solution Explorer*. The *Auto Build* function allowed the wizard to create the cubes automatically (Microsoft 2005). The consultant would need to build and deploy the cube before the querying the data

in the cube. In order to build and deploy the cube, the consultant needed to set up a new server, which was outside of the scope of the project. Other concerns included finding the information technology resources necessary to fully design and implement the cubes and train the appropriate staff in their use.

Organizations can chose to implement several types of OLAP tools. Multidimensional OLAP (MOLAP) tools use a dedicated database to store pre-defined dimensions and aggregations of the cube. Relational OLAP (ROLAP) tools use a relational database to generate cubes as necessary. This allows more flexibility but takes longer to process. Low-Level OLAP (Low-LAP) tools differ from MOLAP in that that they use strong client software to allow the end user to query the Low-LAP database. These limitations make Low-LAP queries more limited and not scalable (Hackney 1997).

Wang and Kuo (2010) suggest the following process for OLAP implementation: planning the project, defining requirements, modeling dimensions, designing physical architecture, designing processes to extract, transform and load data, designing technical architecture, selecting and installing a particular OLAP product, specifying and developing application for business intelligence, deploying and maintaining the system.

Student Activity Six: Select Software

Assume that the university cannot use Microsoft's SQL Server Business Intelligence Development Studio to develop a prototype OLAP cube. What free software would suit the purpose of creating an OLAP prototype?

CONCLUSION

This project investigated whether or not the university could create a data warehouse and OLAP cubes using their existing software and technology. The feasibility of setting up a data warehouse or data mart using OLAP cubes hung on the creation of an additional server to house the actual OLAP cube, dedication of information technology resources to design the production cubes and train the users in their use in the context of the university database.

ACKNOWLEDGMENTS

Although the table view was changed, the university's Information Systems vendor is to be commended for their initial design of their database and the corresponding views. The Information Technology Staff at the university gave willingly of their time and expertise in completion of the project despite their other significant priorities. The documentation which came with Microsoft's SQL Server Business Intelligence Development studio aided in preparing this information (Microsoft 2005). This case was prepared as a teaching case at the suggestion of a trusted mentor and teacher, whose suggestions were invaluable to the development of the paper.

REFERENCES

1. Berry, M.J.A., and Linoff, G.S. *Mastering Data Mining The Art and Science of Customer Relationship Management* Wiley Computer Publishing, John Wiley & Sons, Inc., New York, 2000, p. 494.
2. Chaudhuri, S., and Daval, U. "An Overview of Data Warehousing and OLAP Technology," *SIGMOD Record* (26:1) 1997, pp 65-74.
3. Dixon, P.S., Gribbons, B.C., and Meuschke, D.M. "Applicants Who Do Not Enroll--Fall 1999, Fall 2000, Fall 2001," Education Resources Information Center (ERIC), 2002.
4. Hackney, D. *Understanding and Implementing Successful Data Marts* Addison Wesley Developers Press, 1997, p. 430.
5. Microsoft "Microsoft SQL Server 2005 Analysis Services Tutorials," in: *Microsoft SQL Server 2005 Books Online*, Microsoft, 2005.
6. Power, D.J. "Decision Support Systems Hyperbook," Cedar Falls, IA, 2000.
7. Wang, Y.-H., and Kuo, T.-H. "A Financial Assets and Liabilities Management Support System," *Contemporary Management Research* (6) 2010, pp 315-339.

Table 1. Fields from the Selected Table View

Field Name	Field Type	Description	Example
		Person ID (Unique Person Identifier – Not Social Security Number)	123456789
		Name -- Salutation (Name Prefix)	Mr., Ms.
		Name – First (Person's first name)	Helga, Karl
		Name – Middle (Person's middle name)	Ramona, Henry
		Name – Last (Person's last name)	Dunn, Acosta
		Name – Suffix (Suffix to Person's name)	Jr., III
		Address -- Preference (Where does the person prefer to be contacted?)	Local, Permanent, PO box
		Date of Death (If any)	1/1/1900 is "unknown"
		Information Release (Has the person restricted the use of their address information?)	Authorized, Limited, Department Authorized
		Address – Row 1 (First line of preferred address)	4259 Macias Street
		Address – Row 2 (Second Line of preferred address)	Suite 1600
		Address – City (City in preferred address)	San Francisco
		Address – State (State in preferred address)	CA, FL, TX, NY
		Address -- ZIP Code (Nine digit ZIP code in preferred address)	123456789
		Address – Country (Country in preferred address)	USA
		Address – County (Person's county number)	195
		Address – Business Phone (Phone number during working hours)	1234567890
		Address – Home Phone (Phone number during non-working hours)	1234567890
		Address – Mail Ban (Has this person requested we send them no mail?)	Y, N
		Address – Business Phone Ban (Have they requested we not call their work phone?)	Y, N
		Address – Home Phone Ban (Have they requested we not call their evening phone?)	Y, N
		Address – E-mail (Person's e-mail address)	Helga_Dunn@infs830
		Semester – Year (The academic year the for which the person applied)	2013, 2012
		Semester – Term (The academic term for which the person applied)	Spring, Fall
		Semester – Session (The academic session for which the person applied)	01,02, Grad A, Grad B, Grad C
		Curriculum – Program (The person's overall academic program)	Graduate, Undergraduate, N/A
		Curriculum – Degree (Type of degree the person seeks)	Master of Arts, Master of Science
		Curriculum –Major (The type of degree the person seeks	Accounting, Biology, Chemistry
		Curriculum –College (College that houses the degree the person seeks)	Education, Liberal Arts
		Curriculum –Department (Department within the college that offers the degree.)	Behavioral Sciences, Business Administration, Communications
		Student – Classification (Determined by student's completed hours)	Freshman, Sophomore, Junior
		Student – Advisor (Staff or Faculty member that advises student.)	Person ID
		Student – Counselor (Staff member that Counsels student.)	Person ID
		Admission – Last Activity (Date of last change in status)	
		Admission – Interest Level (Level of interest as reported by the student.)	Definite, very sure, uncertain, unlikely, not coming
		Admission – Status of Inquiry (Has student Inquired to the university?)	Inquired, Applied, Deleted,
		Admission – Status of Inquiry – Date (Date when Status of Inquiry was updated)	Inquiry, Deleted, Applied.
		Admission – Date of Application (Date when student applied)	
		Admission – Status of Application (Where is the application in the process?)	Complete, Deleted, Rejected, Incomplete
		Admission – Date of Application Status (Date when application status was updated)	
		Admission – Decision (Admission decision based on student application.)	Accepted, Hold, Temporarily Accepted, Special Circumstances,
		Admission – Date of Decision (Date of Decision)	

ANSWERS TO STUDENT ACTIVITIES:

Answer to Student Activity One

The university could require consultants to sign non-disclosure agreements. Another option would involve giving the consultant limited access to the database. Then the consultant could only access the tables or views needed for the project. Some universities create test databases that have the ID numbers changed and the names removed or randomized. The university could have someone work closely with the consultant whenever they access student data.

Answer to Student Activity Two

Student field names will vary. Naming conventions may include standard abbreviations, and the inclusion of the database table within the name. All fields are text string fields, except for the following. Date fields include the following: Date of Death, Admission Last Activity date, Admission Status of Inquiry date, Admission Date of Application, Admission Date of Application Status, and Admission Date of Decision. Flag fields include the following: Address Mail Ban, Business Phone ban, Home phone ban,

Answer to Student Activity Three

Step One: The OLAP cube represents the process of inquiry, application and acceptance.

Step Two: Measure Time by a granularity of Year, Term and Session because the transactional system captures this information and it makes sense in the context of a university's twelfth day data.

Step Three: Cube should include the dimensions of Inquiry, Application, Decision, Academic Time, Program and Contact. The schema should use dimensions of inquiry, application and decision because they describe steps in the process and will allow aggregation along those dimensions. The schema should use time as a summary dimension. Adding a Contact dimension allows aggregation along geographical Lines. Program allows summarization among departments and programs at the university.

Step Four: Student Facts to summarize should include classification, advisor, counselor and any current activity. The fact table should include the keys to the dimension tables (Person_Code_ID) and the summary metrics: Person_Code_ID (number of students), class level, counselor and current_activity.

Answer to Student Activity Four

Since the university uses data from one computer system that primarily concerns one or two departments, the term 'data mart' fits better than 'data warehouse'. If the university pulled in multiple campuses or multiple computer systems, the term 'data warehouse' would apply more.

Answer to Student Activity Five

The most effective data mining techniques typically require detail-oriented data. By definition, OLAP cubes summarize data. (Berry et al. 2000). In this particular case, the university may get as much information simply by seeing trends across the different dimensions as by using traditional data mining techniques.

Answer to Student Activity Six

At this writing, students can find free OLAP software at download.cnet.com.

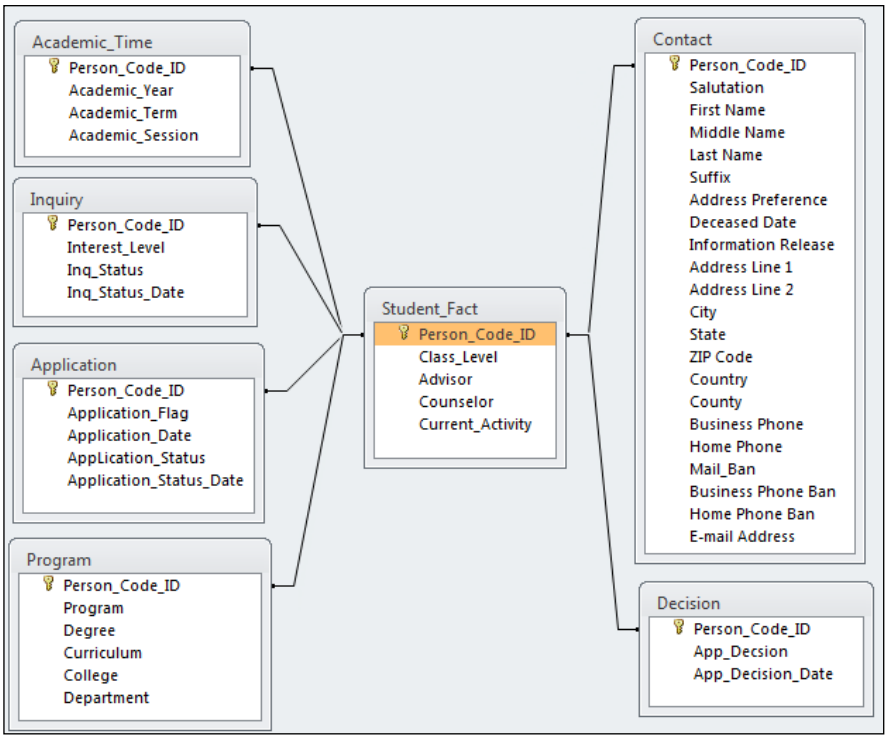

Figure 2. One Potential Solution for the Star Schema Diagram