Analyzing the Data for Electronic Shop by using OLAP tools

A Me Tun
University of Computer Studies (Pathein)
ametun1@gmail.com

Abstract

Data warehousing and online analytical processing (OLAP) are essential facilities for data analysis tasks supporting a user's decision in a business. Data warehousing approach is to integrate information and heterogeneous sources in advance, store the historical information in a warehouse and support complex multidimensional queries. This proposed system is developed by using multidimensional contingency tables. This model views data in the form of a data cube. To implement a framework that will use OLAP queries from multidimensional contingency tables. This system can provide the various sales analyzing reports and present data in various formats in order to accommodate the diverse needs of the manager for future sales plans.

Keyword: Data Warehousing, OLAP

1. Introduction

Data warehousing is architecture to help business executives to understand and organize data and make a business decision. On- Line Analytical Processing (OLAP) manages data warehouses for data analysis and provides calculations such as summarization and aggregation in advance, and manages information at different levels of granularity. OLAP has become very popular techniques to help users analyze data by providing multiple views of the data [1].

OLAP stands for Online Analytical Processing. It is an approach to quickly provide answers to analytical queries that are multidimensional in nature. OLAP is part of the broader category business intelligence, which also encompassed relational reporting and data mining. The typical applications of OLAP are in areas. The term OLAP was created as a slight modification of the traditional database term OLTP (Online Traction Processing) [2].

The item dataset involved many records. Each record contains a set of attribute. This attributes classifier and converts to the contingency tables. Database configured for OLAP employ multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time. The functionalities of OLAP are dynamic multidimensional analysis of consolidated data supporting end user analytical and navigational activities.

The rest of paper is organized as follow. In section (2) states the related work with my system. In section (3) describes background theory. Next, section (4) explains the overview of the system architecture and detail design. And then, states section (5) implementation of the system. In the last section, describes the conclusion and further extension.

2. Related Works

OLAP can be defined into three main types: ROLAP (Relational OLAP), MOLAP (Multidimensional OLAP) and HOLAP (Hybrid OLAP) [4]. The OLAP using in the proposed system can be put in ROLAP type. ROLAP is an alternative to the MOLAP (Multidimensional OLAP) technology. While both ROLAP and MOLAP analytic tools are designed to allow analysis of data through the use of a multidimensional data model, ROLAP differs significantly in that it does not require the pre-computation and storage of information. Instead, ROLAP tools access the data in a relational database and generate SQL queries to calculate information at the appropriate level when an end user requests it. With ROLAP, it is possible to create additional database tables (*summary tables* or *aggregations*) which summarize the data at any desired combination of dimensions [4]. The database using for the proposed system is carefully designed for ROLAP use.

There are several ways a data warehouse or data mart (known as database) can be structured: multidimensional, star and snowflake schema. Snowflake schema is used in this proposed system for the following benefits of snowflake schema [5].

- Some OLAP multidimensional database modeling tools that use dimensional data marts as a data source are optimized for snowflake schemas.
- If a dimension is very sparse (i.e. most of the possible values for the dimension have no data) and/or a dimension has a very long list of attributes which may be used in a query, the dimension table

may occupy a significant proportion of the database and snowflaking may be appropriate.

- A multidimensional view is sometimes added to an existing transactional database to aid reporting. In this case, the tables which describe the dimensions will already exist and will typically be normalized. A snowflake schema will hence be easier to implement.
- A snowflake schema can sometimes reflect the way in which users think about data. Users may prefer to generate queries using a star schema in some cases, although this may or may not be reflected in the underlying organization of the database.
- Some users may wish to submit queries to the database which, using conventional multidimensional reporting tools, cannot be expressed within a simple star schema. This is particularly common in data mining of customer databases, where a common requirement is to locate common factors between customers who bought products meeting complex criteria. Some snowflaking would typically be required to permit simple query tools to form such a query, especially if provision for these forms of query weren't anticipated when the data warehouse was first designed [4].

Mainly there are 6 types of OLAP functionalities. It doesn't mean that all those six functions needs to use to build OLAP application [2]. Suitable ones are only needed to use according to the application type. Those OLAP functionalities are described detail in section 5.1. Among those, slice, dice, drill up/down functions are mainly used in the proposed system.

3. Theory Background

Dataset is a collection of data, usually presented in tabular form. Each Column represents a particular variable. Each row corresponds to given member of the dataset. Its lists values for the each of the variables such as value of an objects. Each value is known as a data. This dataset must have instances for which you know the actual value of the target variable and the associated predictor variables. It might have to perform a controlled study to collect this data, or it might be able to obtain it form previously-collected historical records. The data it provides to DTRG for an analysis is called a "dataset". It consists of an entry for each case to be analyzed. Each case provides values for the target and predictor variables for a specific customer, patient, company, etc [3].

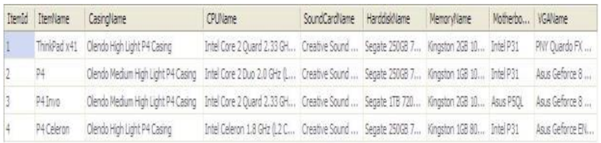
Classification is an automated process to group related records together. Related records are group together on the basic of having similar values for attributes. Classification is also segments customer records into distinct segments called classes. A classification analysis requires that end-user know ahead of time how classes are defined. It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because of each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use [3].

The foundation of all OLAP systems is a concept of and OLAP cube, also called a multidimensional cube. It consists of numeric facts called measures which are categorized by dimensions. The cube metadata is typically created from a star schema or snowflake of tables in a relational database. Measures are derived from the records in the fact table and dimensions are derived from the dimension tables. An OLAP system is used for data analysis by knowledge workers including managers, executives and analysis. An OLAP system provides facility for summarization and aggregation and stores and mange information at different level. This system makes the data easier to use in informed decision making [3].

3.1 Dataset

A sales data set is used to determine the accuracy of the model. Usually, the given data set is divided into Items Data and Sales History data sets, with Items Data sets used to build the model and the History Data set is used to validate it.

Item Dataset is a collection of records. Each records contains a set of attributes, one of the attributes is the class.



ItemId	ItemName	CasingName	CPUName	SoundCardName	HarddiskName	MemoryName	Motherbo...	VGAName
1	ThinkPad x41	Olendo High Light P4 Casing	Intel Core 2 Quard 2.33 GH...	Creative Sound ...	Segate 250GB 7...	Kingston 2GB 10...	Intel P31	PNY Quardo FX ...
2	P4	Olendo Medium High Light P4 Casing	Intel Core 2 Duo 2.0 GHz (L...	Creative Sound ...	Segate 250GB 7...	Kingston 1GB 10...	Intel P31	Asus Geforce 8 ...
3	P4 Invo	Olendo Medium High Light P4 Casing	Intel Core 2 Quard 2.33 GH...	Creative Sound ...	Segate 1TB 720...	Kingston 2GB 10...	Asus P5QL	Asus Geforce 8 ...
4	P4 Celeron	Olendo High Light P4 Casing	Intel Celeron 1.8 GHz (L2 C...	Creative Sound ...	Segate 250GB 7...	Kingston 1GB 80...	Intel P31	Asus Geforce EN...

Figure 2. Example of the Item Dataset

It is a Item Dataset

Used Attributes

CPU	Motherboard	Casing
Memory	Sound Card	
Hard Disk	Graphic Card	

This dataset involved many records and 7 attributes

Figure 3. This Dataset involved Many Records and 7 Attributes.

3.2. Classification

Classification is used for OLAP data analysis. The following figure (figure 4) shows the analysis view of computer model with same CPU.

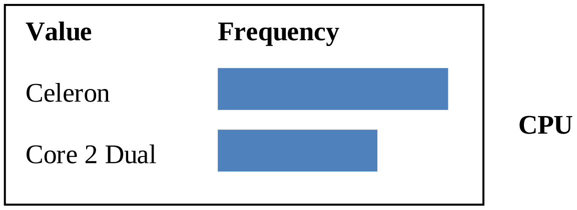


Figure 4. Example of the Classification

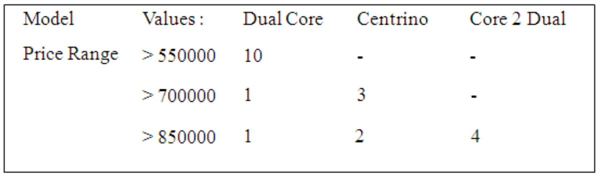
3.3. Contingency Tables

A contingency table is a table of counts from a database consisting of n rows, each comprising values for a fixed set of, say, binary attributes a1,...,a_k, the contingency table is the histogram of counts for each of the 2k possible settings of these attributes. Contingency tables are essentially equivalent to OLAP cubes, which cast traditional relational databases as a high-dimensional cube with dimensions corresponding to the attributes [4].

- With 7 attributes, how many 1-d contingency tables are there?71 contingency tables
- How many 2-d contingency? 7-choose-2=7*10/2= 35
- How many 3-d tables?7*10/2 * 9/3 = 105

3.3.1. Example of the Contingency Tables

- **2-d contingency table**



Model	Values :	Dual Core	Centrino	Core 2 Dual
Price Range	> 550000	10	-	-
	> 700000	1	3	-
	> 850000	1	2	4

Figure 5. Model and Price Contingency Table

- **3-d contingency table**

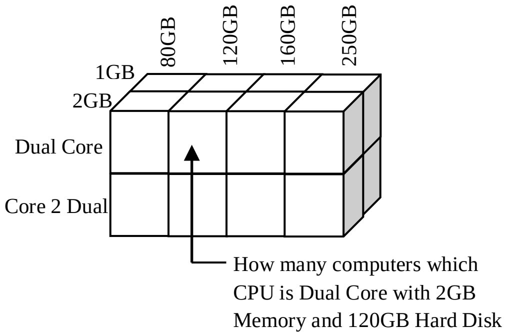


Figure 6. CPU and Memory and Hard Disk Table

4. Overview of the System Design

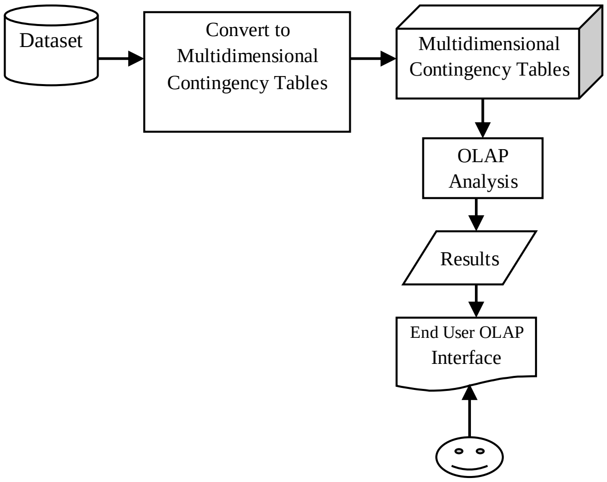


Figure.1. Process Flow of the System

In this system, two dimensions and three dimensions cube are used to sales analysis OLAP reports. The data warehouse using in the system is developed as business query view which is the perspective of data in the data warehouse from the viewpoint of the end user. The dimensional cubes used in the system are two dimensions and three dimensions; therefore, the system used dice function of OLAP functions.

5. Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) that enables analysts and executives to gain insight to data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user [2].

5.1. OLAP Functionality

Dynamic multi-dimensional analysis of consolidated data supporting end user analytical and navigational activities including:

- *Slice:* A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.
- *Dice:* The dice operation is a slice on more than two dimensions of a data cube (or more than two consecutive slices).
- *Drill Down/Up:* Drilling down or up is a specific analytical technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).
- *Roll-up:* A roll-up involves computing all of the data relationships for one or more dimensions. To do this, a computational relationship or formula might be defined.
- Reach-through to underlying detail data
- Rotation to new dimensional comparisons in the viewing area.

6. Experimental Results

Proposed system is written in Microsoft C#.NET 2008 under .NET Framework 3.5. The EXE file of the system can be run on every machine which installed window XP service pack 2, .NET Framework 3.5 and Microsoft SQL Server 2005 Developer Edition. To view the source code file or to make a modification, Microsoft Visual Studio 2.0 is needed to install. For the database, Microsoft SQL Server 2005 Developer Edition is used as the database server which provide for OLAP functionalities.

7. Conclusion

OLAP is a significant improvement over query systems. OLAP is an interactive system to show summaries of multidimensional data by interactive selecting the attributes in the multidimensional contingency tables. OLAP are useful facilities for a decision making of users. A data warehouse stores summarized and compressed information and data cubes are tools for access to a data warehouse efficiently. In this paper, we describe the usefulness of OLAP cube and reports with a practical approach.

8. References

[1] “What is OLAP?” by Nigel Pendse, Principal of OLAP Solutions and Co-author of the OLAP report.com. www.OLAPreport.com

[2] Decision Trees: Professor Andrew W.Moore. School of Computer Science, Carnegie Mellon University. www.cs.cmu.edu.

[3] Robert Wrembel and Christian Koncilia: Data Warehouses and OLAP. ISBN 1-59904-364-5

[4] Kimball, Ralph and Ross, Margy. *The Data Warehouse Toolkit* Second Edition (2002) John Wiley and Sons, Inc. ISBN 0-471-20024-7

[5] Hari Mailvaganam (2007). "Introduction to OLAP - Slice, Dice and Drill". DWreview. http://www. dwreview.com/OLAP/Introduction_OLAP.html. Retrieved on 2009-03-05.

[6] Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0471228524

[7] Bach Pedersen, Torben; S. Jensen (December 2001). "Multidimensional Database Technology" (PDF).ISSN 0018-9162. http://ieeexplore.ieee.org/ iel5/2/20936/00970558.pdf

[8] Codd E.F., Codd S.B., and Salley C.T. (1993). "Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate". Codd & Date, Inc.

[9] Alex Berson; Stephen J. Smith. Data Warehousing, Data Mining and OLAP.

[10] Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0471228524