Implementation of Multidimensional Data Models using Data Warehouse and OLAP Technology

Lwin Si Thu Nyunt
Computer University, Pathein
lwinstn@gmail.com

Abstract

Data warehousing and online analytical processing (OLAP) are essential elements of decision support system, which has increasingly become a focus of the database industry. A data warehouse is an integrated and time-varying collection of data derived from operational data and primarily used in strategic decision making by means of OLAP techniques. OLAP technology can use data from operational database to be executed data analysis queries. We describe back end tools for extracting, cleaning and loading data into a data warehouse multidimensional data model. In this system, we also implement multidimensional data models typical of OLAP. This proposed system intends to develop the data warehouse for the main crops of Pathein District in Ayeyarwaddy Division. And then develop relational online analytical processing (ROLAP) to retrieve information from the data warehouse via graphical user interface.

Keyword: Data Warehouse and OLAP

1. Introduction

A data warehouse is a repository of information collected from different sources. Querying of data warehouses for decision-making in areas such as sales and marketing planning is referred to as OLAP. OLAP is a method to support decision making in situations where raw data on measures such as sales or profit needs to be analyzed at different levels of statistical aggregation. In OLAP, queries are made against multidimensional cubes, called OLAP cubes.

Data warehouse systems have become a key component of the corporate information system architecture, the idea of a data warehouse is to extract data from operational databases and to store them separately. Data warehouses, in contrast, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Data warehouses contain historical, consolidated, and summarized data for supporting business decision systems at many levels. Data warehousing technology is currently a very active field of research.

Research problems associated with creating, maintaining, and using data warehouse technology are partially similar to those specific for database systems. In fact, a data warehouse can be considered as "large" database system with additional functionality. OLAP applications analyze data by means of complex queries ranging from a few to dozens operations of joining, filtering, grouping, and aggregating. Since these queries are very complex and they often read terabytes of data, their execution may take dozens of minutes, hours, or even days.

Data warehouse might be implemented on standard or extended relational DBMSs, called Relational OLAP (ROLAP) servers. These servers assume that data is stored in relational databases, and they support extensions to SQL and special access and implementation methods to efficiently implement the multidimensional data model and operations. In contrast, multidimensional data in special data structures e.g., arrays and implement the OLAP operations over these special data structures.

The rest of paper is organized as follows: This system is introduced in section 1. In section 2, related works is presented. We explain about data warehouse and OLAP in section 3. Proposed System Architecture Design is described in section 4. Section 5 shows experimental results. Finally this system is concluded in section 6.

2. Related Works

There has been an increasing amount of project similar and related with our process. Recently, a number of security models have been proposed for data warehouse and OLAP. For example, propose a security model based on mandatory access control for OLAP cubes [7]. Data Warehouses integrate data from multiple heterogeneous information sources and transform them into a multidimensional representation for decision support applications. All the data warehouse components, processes and data should be tricked and administered via a metadata repository.

Jarke presented a metadata modeling approach which enables the capturing of the static parts of the architecture of a data warehouse [6]. Inmon has been presented a data warehouse is a repository of integrated information, available for querying and analysis [4,5]. As relevant information becomes available or is modified, the information is extracted from its source, translated into a common model e.g., the relational model), and integrated with existing data at the warehouse.

3. Data warehouse and OLAP

A data warehouse is typically a dedicated database system. It draws most of it data from production systems (i.e., the "source" systems); yet it is separate in its operations. Users access the data warehouse rather than the data in the production systems.

OLAP can be performed in data warehouse/ marts using the multidimensional data model. OLAP constitutes a set of technologies and that are greating for summarizing and analyzing huge quantities of data OLAP allows users access to this data through such methods as ad hoc querying, calculations, time series analysis, and complex modeling.

3.1. Architecture and end to end Process

A typical data warehousing architecture is shown in figure 1. It includes tools for extracting data from multiple operational databases and external sources, for cleaning transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage. In addition to the main warehouse, there may be several departmental data marts.

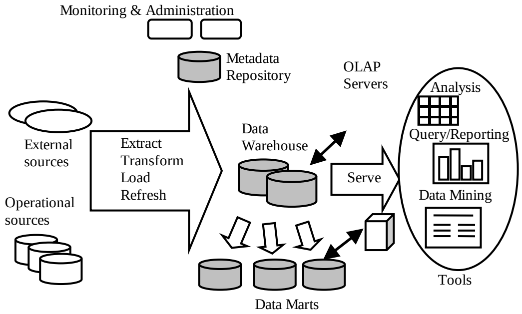


Figure1. Data Warehousing Architecture

Data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of front end tools: query tools, report writers, analysis tools, and data mining tools.

Finally, there is a repository for storing and managing metadata, and tools for monitoring and administering the warehousing system. The warehouse may be distributed for load balancing, scalability, and higher availability. In such a distributed architecture, the metadata repository is usually replicated with each fragment of the warehouse, and the entire warehouse is administered centrally. An alternative architecture, implemented for expediency when it may be too expensive to construct a single logically integrated enterprise warehouse, is a federation of warehouses. Designing and rolling out a data warehouse is a complex process, consisting of the following activities [9].

- Define the architecture, do capacity planning, and select the storage servers, database and OLAP servers, and tools.
- Integrate the servers, storage, and client tools.
- Design the warehouse schema and views.
- Define the physical warehouse organization, data placement, partitioning, and access methods.
- Connect the sources using gateways, ODBC drivers, or other wrappers.
- Design and implement scripts for data extraction, cleaning, transformation, load, and refresh.
- Populate the repository with the schema and view definitions, scripts, and other metadata.
- Design and implement end-user applications.
- Roll out the warehouse and applications.

3.2. Data Warehouse Back-End Tools and Utilities

Data warehouse systems use back–end tools and utilities to populate and refresh their data. These tools and utilities include the following functions:

Data Extraction, which typically gathers data from multiple, heterogeneous, and external sources.

Data Cleaning, which detects errors in the data and rectifies them when possible.

Data transformation, which converts data from legacy or host format to warehouse format.

Load, which sorts, summarizes, consolidates, computes views, check integrity, and builds indices and partitions.

Refresh, which propagates the updates from the data sources to the warehouse.

Besides cleaning, loading, refreshing, and metadata definition tools, data warehouse systems, usually provide a good set of data warehouse management tools.

3.3. Data Warehousing Schema

A schema is a collection of the database objects, views, indexes and synonyms. There are a variety of ways of arranging schema objects in the schema model designed for data warehousing.

3.3.1. Star Schema

The star schema is the simplest data warehouse schema. It is called a star schema because the diagram of the schema resembles a star, with points radiating from a center. The center of the star consists of the one or more fact tables and the points of the star are dimensions tables in Figure 3.The schema is presented one fact table and dimension tables.

3.4. OLAP Operations in the Multidimensional data Model

In the multidimensional data model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies**.**

Roll-up: The roll-up operation performed on the central cube by climbing up a concept hierarchy for a dimension or by dimension reduction.

Drill-down: Drill-down is the reverse of roll-up. It navigates from less detail data to more detailed data.

Slice and dice: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. The dice operation defines a subcube by performing a selection on two or more dimensions.

Pivot: Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

Other OLAP operations: Some OLAP systems offer additional drilling operations. An OLAP engine is a powerful data analysis tool.

4. Proposed System Architecture

The proposed system, have main three components: (1) Excel file imported into transaction data. (2) The transaction data are loaded into the data warehouse using extract, transform, loading (ETL) process. (3) Retrieved the information from data warehouse as three dimensions using OLAP analysis.

The aim of this proposed system is the creation of data warehouse from data sources and OLAP.

The system can understand data warehouse and operational data source are different database structures. An overview of this system is shown in Figure 1.

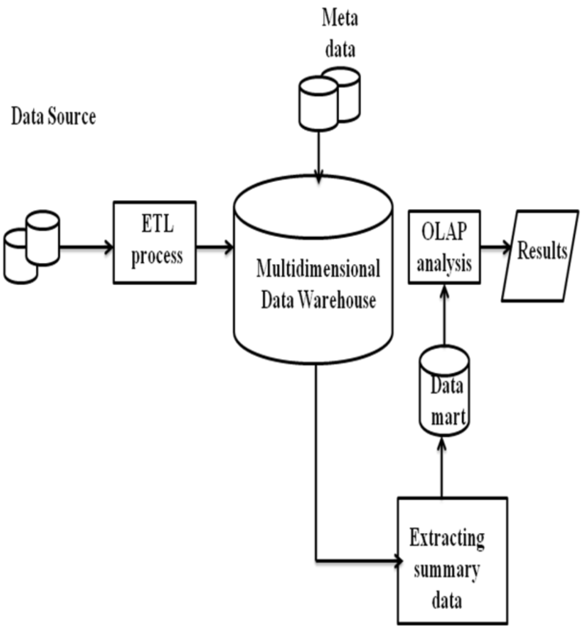
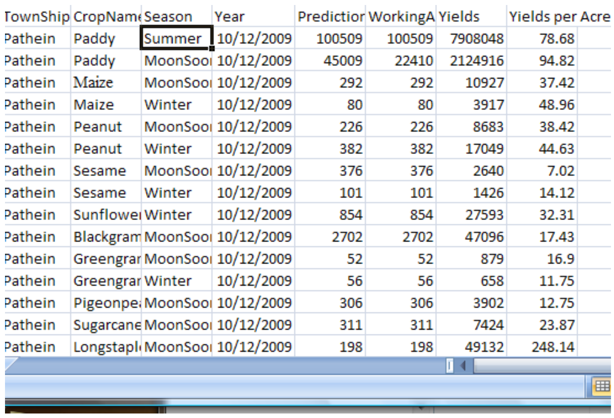


Figure 1. Proposed System Design

Step1: In this system, excel file is used as external source file is shown in Figure 2.



TownShip	CropNam	Season	Year	Predictior	WorkingA	Yields	Yields per Acre
Pathein	Paddy	Summer	10/12/2009	100509	100509	7908048	78.68
Pathein	Paddy	MoonSooi	10/12/2009	45009	22410	2124916	94.82
Pathein	Maize	MoonSooi	10/12/2009	292	292	10927	37.42
Pathein	Maize	Winter	10/12/2009	80	80	3917	48.96
Pathein	Peanut	MoonSooi	10/12/2009	226	226	8683	38.42
Pathein	Peanut	Winter	10/12/2009	382	382	17049	44.63
Pathein	Sesame	MoonSooi	10/12/2009	376	376	2640	7.02
Pathein	Sesame	Winter	10/12/2009	101	101	1426	14.12
Pathein	Sunflowei	Winter	10/12/2009	854	854	27593	32.31
Pathein	Blackgram	MoonSooi	10/12/2009	2702	2702	47096	17.43
Pathein	Greengrar	MoonSooi	10/12/2009	52	52	879	16.9
Pathein	Greengrar	Winter	10/12/2009	56	56	658	11.75
Pathein	Pigeonpe	MoonSooi	10/12/2009	306	306	3902	12.75
Pathein	Sugarcane	MoonSooi	10/12/2009	311	311	7424	23.87
Pathein	Longstapl	MoonSooi	10/12/2009	198	198	49132	248.14

Figure 2. External Data Source File

Step 2: After that operate the ETL process a data warehouse is created with these data. Data warehouse involves one fact table and dimension able. The fact table is Yields table and the dimension tables are Crops table, Township table, Season Table and Year table is shown in Figure 3.

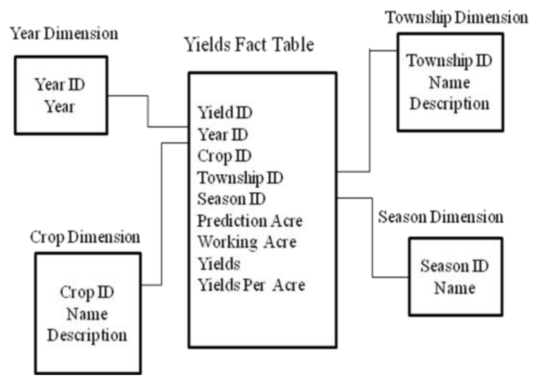


Figure3. Star Schema of Data Warehouse for Yields

Step 3: In this system, we show the information retrieval for three dimensions as shown in Figure 4.

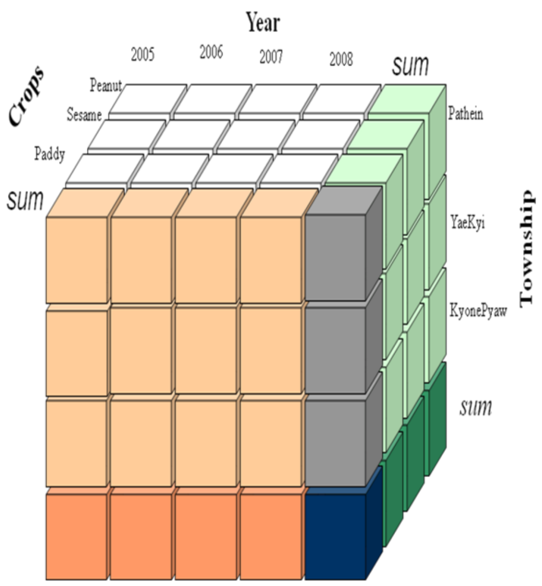


Figure 4. A Sample Data Cube

5. Experimental Results

The proposed system provides the constructing of data warehouse by operating the extract, transform, loading (ETL) process on operational external data source.

Extract is the process of reading data from a specified source file and extracting a desired subset of data. Transform is the process of converting the extracted/acquired data from its form into the form it needs to be placed into data warehouse. Loading is the transfer of processed data to the crops yields data warehouse as fact data and dimension data.

We implement the information retrieval for multidimensional crops, township and year for Pathein District in Ayeyarwaddy Division by using OLAP as shown in Figure 5 and 6.

Crops : -Select One- TownShip : KanGyiDaung Year : 2005 Search

Viewed by dimension (TownShip, Year)

Crop	Year	Season	TownShip	Prediction Acre	Working Acre	Yields	Yields Per Acre
Paddy	2005	Summer	KanGyiDaung	81136.00000	15570.00000	153675.00000	98.70
Sunflower	2005	Winter	KanGyiDaung	4379.00000	4379.00000	113635.00000	25.95
Peanut	2005	Winter	KanGyiDaung	3.00000	3.00000	139.00000	46.33
Greengram	2005	Winter	KanGyiDaung	1681.00000	1681.00000	23114.00000	13.75
Blackgram	2005	Monsoon	KanGyiDaung	330.00000	330.00000	3977.00000	12.05

Figure 5. Yields Data by two Dimension View

Crops : Peanut TownShip : Pathein Year : 2009 Search

Viewed by dimension (TownShip, Crops, Year)

Crop	Year	Season	TownShip	Prediction Acre	Working Acre	Yields	Yields Per Acre
Peanut	2009	Winter	Pathein	382.00000	382.00000	17049.00000	44.63
Peanut	2009	Monsoon	Pathein	226.00000	226.00000	8683.00000	38.42

Figure6. Yields Data by three Dimension View

6. Conclusion

This paper intended to study data warehouse structure and OLAP, to understand the different database structures of data warehouse and operational OLTP database, to know about the data warehouse usage and to improve the performance of query processing. This system provides multidimensional view with OLAP operations on the main crops data of Pathein District in Ayeyarwaddy Division. The system can be viewed according to the desired crops, township, and year for three dimensions. Moreover, this system can help the decision making for business firm to perform the right ad hoc decision.

7. References

[1] J.Han and K. Micheline" *Data Mining: Concepts and Techniques*" ISBN 1-55860-489-8, 2001 by Academic Press.

[2] J.Han, S.Chee, and J.Y.Chiang, " *Issuse for on_line analytical mining of data warehouse* ", In

Proc.1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD’98) pages2:1_2:5, Seattle, WA, June 1998.

[3] W.H Inmon, “*Building the Data warehouse”* ISBN-81-265-0271-1, 2001 by John Wiley&Sons.

[4] W.H. Inmon, *"EIS and the data warehouse: a simple approach to building an effective foundation for EIS. Database Programming & Design"*, 5(11): 70-73, November 1992.

[5] W.H. Inmon, "*Building the data bridge: the ten critical success factors of building a data warehouse. Database Programming & Design"*, 1992.

[6] M. Jarke, M.A. Jeusfeld, C. Quis, P. Vassiliadis,*"Architecture and quality in data warehouse. An extended repository Approach Information Systems"*, 24(3) 229-253 (1999). A previous version appeared in Proc. 10th Conf. of Advanced Information Systems Engineering (CASE 98), Pisa, Italy (1998).

[7] R. Kirkgöze, N. Katic, M. Stolba, andA M. Tjoa. *"A Security Concept for OLAP proceedings of the 8th. International Workshop on Database and Expert Systems Applications (DEXA’97)"*, IEEE Computer Society, 1997.

[8] R. Kimball, “ *The Data Warehouse Toolkit* “, NewYork : John Wiley & Sons, 1996

[9] C. Surajit and D. Umeshwar, “*An Overview of Data Warehousing and OLAP Technology*”, pp. 2-3, Microsoft Research, Redmond, 2000.