CONCEPT HIERARCHY_BASED CUBE AGGREGATION FOR ETL

PROCESS IN MATRICULATION WAREHOUSE

ZAR NI MG
M.C.Sc (Computer Science), MCTS(Microsoft Certified Enterprise Application)
UNIVERSITY OF COMPUTER STUDIES (MANALAY)

kongaung@gmail.com

Abstract: Data Warehousing and On-Line Analytical Processing are essential elements of knowledge worker (executive, manager, and analyst) to make better and faster decisions. Therefore, it becomes a focus of the database industry. There are many application areas of data warehousing technologies, such as manufacturing, financial services, telecommunication and healthcare. The main objective of our work is to use and understand the data warehousing and OLAP technologies. First this paper proposes a system for data warehouse. Then the detailed of ETL processes are followed. The data extracted from different sources are stored in data staging area, where the data is cleaned and transformed using cube aggregation. The cleaned data from staging area are loaded and organized as Matriculation Warehouse for the direct querying for the end users. Specifically, this paper emphasizes on how to conduct the ETL process with the use of concept hierarchy-based cube aggregation. The prototyped system is implemented based on Basic Education High School (B.E.H.S) student data for accessing the speed up OLAP operations because of cube aggregation.

1. Introduction

Database technology has evolved from primitive file processing to the development of database management systems with query and transaction processing. The progress has led to the increasing demand for efficient and effective data analysis and data understanding tools.

A data warehouse is very suitable to store the data like a subject-oriented, integrated, nonvolatile, time-variant collection of data for the purpose of analyzing of previous events, prediction and estimation for near future and later time [2]. The data in warehouse are modeled as Multidimensional with hyper-cubes [3] where each dimension represents attributes and cube cells represent measure. According to [Microsoft], there are four major process of the data warehouse. They are extract (bring operational data into data warehouse), transform (change into internal format and structure of data warehouse), cleans (change into quality data), and load (put into data warehouse).

Meanwhile the dataset are recommended to preprocess before mining, such as classification, clustering. Because some data may contain noise, other may contain missing value. Similarly some data are required to generalize to more higher conceptual level to save time and space in the analyzing and querying. There are many preprocessing techniques, such as data cleaning, data reduction, missing value detection, etc. Among them cube aggregation, one of the data reduction technique, can be used as a supporting in the querying of multidimensional data. At the same time, concept hierarchy can also be used as a tool for generalization and specialization of multidimensional data at different cube level.

This paper takes the advantages of concept hierarchy facility. Therefore, concept-hierarchy Based cube aggregation method is used for effectively query answering at different level of cuboids. In other words, this method is used for moving data from staging area onto data warehouse. First, data from Basic Education High School (B.E.H.S.) students of whole Myanmar are logically modeled into three dimensional, and then change into Matriculation warehouse using star and snow schema on Relational OLAP (ROLAP) server. Historical data of B.E.H.S students according to their registration to exam and pass/fail on these exams are stored in matriculation warehouse.

Before analyzing OLAP operations on these data, the data from warehouse are preprocessed with the use of Cube-Aggregation Method. Here pre-constructed concept hierarchy trees are used in Cube-Aggregation. The final result of the system is to present the useable results of the Implementation of Matriculation Warehouse System Based-on Online Analytical Processing.

The rest of the paper is organized as follows: Section 2 provides a brief overview of OLAP, ETL and aggregation. Section 3 proposes data warehouse system based on one of the existing warehouse architecture containing staging area. In Section 4 we describe the implementation details of our system and show how to speed up the OLAP operations. Finally we summarized our work and provide a brief discussion about future work in Section 5.

2. Related Work

According to [9], there are three data warehouse architecture: (i) basic architecture without containing staging area and data mart, (ii) architecture with staging area, and (iii) architecture with staging area and data mart. Architecture is proposed by [1] which stored not only summarized data but also operational data. [1] Considered the data warehouse architecture with some of the dimensions, such as data, information, technology and product dimensions.
Mostly, staging area is used in much warehouse architecture. According to [8], the staging area is everything between the operational source systems and the data presentation area. The staging area is also called operational data source (ODS). [8] also stated some of the key reasons why organizations incorporate a relational staging area into their data warehouse architecture. Staging areas are used to convert the data to be separate from transforming data, to facilitate moving data from different sources on different schedules.

Many proposed systems have been developed for more and more efficient OLAP operations. [6] proposed a method for efficient OLAP in distributed environment. [5] proposed a method for another efficient OLAP operations using Hidden Markor Model (HMM). They also proposed a data structure, aR-Tree for indexing and grouping spatial data.

The aggregation functions are classified into three [3]: distributive[COUNT,SUM,MIN,MAX], algebraic[AVG] and holistic[MEDIAN]. [4] proposed a set of aggregation algorithms on compressed data warehouses for multidimensional OLAP. Their algorithms operate directly on compressed data sets, which are compressed by the mapping-complete compression methods, without the need to first decompress them. [2] proposed a new compressed representation of the data cube which reduce storage requirements, does not require the discretization hierarchy and so on.

According to [7], aggregattion are precalculated summaries of data from leaf cells. Query response time can be improved by aggregation. The response may be immediate if the summarization data to answer queries has been precalculated. Precalculation of summary data is the foundation for the rapid response times of OLAP technology. Cubes are the way that OLAP technology organizes summary data into multidimensional structures. Dimensions and their hierarchies of attributes reflect the queries that can be asked of the cube.

[6] also stated that precalculation of all possible aggregations in a cube might provide the fastest possible response time for all queries. If no aggregations are precalculated (0%), the amount of required processing time and storage space for a cube is minimized, but the query response time may be slow because the data required to answer must be retrieved from the leaf cells and then aggregated at query time to answer each query.

3. System Architecture and Design

As illustrated in Section 2, they are some architecture for data warehouse. But they all favor a layer-based architecture. Our system is also constructed based on the following layer-architecture as in Figure 1.

Our system design is illustrated in Figure 2.

- ETL= Extract Transform Loading
- OLAP =On-Line Analytical Processing
- OP = Operational Data store
- DW=Data Warehouse

There are three main components in our system:(i) Data Collector, (ii) ETL Processor, (iii) OLAP Processor. The detail tasks of these components are as follow.

(i) Data Collector Collect the data from different locations of Myanmar. Data come with relational DBMS to our system.

(ii) ETL Processor ETL processes are taken place in staging area. There are two sub-managers in this component.

(a) Load Manager Performs all the tasks associated with extraction and loading data into the data warehouse. These tasks include simple transformations of source data into warehouse data.

(b) Warehouse Manager Performs all the tasks concerning with the management of data for the warehouse. The tasks of warehouse manager include: transformation and merging the source data from temporary storage into data warehouse table, generation of aggregation, and backing up and archiving of data.

(iii) OLAP Processor Accept the user’s queries and replies the answers using their basic functions of roll up, drill down, slice and dice.
4 System Set Up and Implementation

Student Registration data to sit the B.E.H.S Examination and Pass/Fail data of whole student in Myanmar are used for system implementation. There are two main data sources for this system, student registration data of the whole Myanmar to sit exam and B.E.H.S student’s pass/fail data of the whole Myanmar every year. The main assumption of the paper is students who have already registered.

4.1 Student Registration Data

Every year, there are total round about hundred to two hundreds thousands of BEHS students register to sit for the BEHS examination in Myanmar. The Department of Basic Education from MOE (Ministry Of Education) is responsible to record this large amount of data for the purpose of processing and analyzing. The attributes containing in students registration database is illustrated in Table 1 with the example records. The student registration entry form is illustrated in Appendix A1. Basically, personal information (such as NRC-No, Date of Birth, Parent Name, Address) and examination information (such as exam center, Scored Subjects group for exam, Roll No) are needed to store for individual.

<table>
<thead>
<tr>
<th>No</th>
<th>RollNo</th>
<th>Name</th>
<th>Fname</th>
<th>City</th>
<th>Division</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AMTS-1</td>
<td>Zar Ni</td>
<td>U Mg</td>
<td>AMTS</td>
<td>Mandalay</td>
<td>Science</td>
</tr>
<tr>
<td>2</td>
<td>AMTS-2</td>
<td>Phyo</td>
<td>MinZaw</td>
<td>U Hla</td>
<td>Mandalay</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>AMTS-4</td>
<td>Khine</td>
<td>Myo</td>
<td>U Myint</td>
<td>Mandalay</td>
<td>Science</td>
</tr>
<tr>
<td>5</td>
<td>AMTS-5</td>
<td>Thet</td>
<td>Naing</td>
<td>U Ning</td>
<td>Mandalay</td>
<td>Science</td>
</tr>
<tr>
<td>7</td>
<td>AMTS-7</td>
<td>Yar Zar</td>
<td>Tun</td>
<td>U zar tun</td>
<td>AMTS</td>
<td>Mandalay</td>
</tr>
</tbody>
</table>

4.2 Student Pass/Fail Data

After four months of BEHS examination has registered, the lists of Pass/Fail are announced via radio or Township Education Offices. These large amount of data are also recorded for man purpose such as

- To deliver the scored subjects’ marks for every examination passed students for their University Entrance Application
- To give exam passed certificate for student who apply the certificate to central office in Yangon
- To analyze who are top ten students in whole Myanmar, which state and division got the highest percentage on the number of passed students, etc...

Firstly, these source data are needed to change into conceptual model. Then the actual data are needed to store on physical storage as warehouse. We can consider this conceptual model with multi-dimension as a cube. Therefore the following section explains how to model the source data into multi-dimensional tube.

4.3 Multidimensional Model for Matriculation Data Warehouse

The collected data can be categorized into three main groups: personal data (name, father name, DOB, National races, etc...), location data (address, exam center), time data (year of exam). Therefore three-dimension model is constructed based on the (i) student registration data and (ii) student pass/fail data as shown in Figure 3 and Figure 4.

![Figure 3. Registration Data Cube](image-url)

Measure = Number of Students

Table 2. Example of Student Mark Information

<table>
<thead>
<tr>
<th>No</th>
<th>Roll No</th>
<th>Name</th>
<th>Subject</th>
<th>Total Mark</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AMTS-1</td>
<td>Zar Ni</td>
<td>Science</td>
<td>320</td>
<td>Pass</td>
</tr>
<tr>
<td>2</td>
<td>AMTS-2</td>
<td>Phyo</td>
<td>Science</td>
<td>247</td>
<td>Fail</td>
</tr>
</tbody>
</table>
After that we need to choose database scheme for storing the data physically on ROLAP server. According to nature of our data, snowflake scheme are used for our system. The next section is the explanation of how to map the data with snowflake schema.

4.4 Mapping into Snowflake Schema

The most common modeling paradigm is the snowflake schema, in which data warehouse contains a large central tables (fact tables) containing the bulk of the data, with no redundancy and a set of smaller attendant tables (dimensional tables), one for each dimension.

Any table that references or is referenced by another table must have a primary key, which is a column or group of columns whose contents uniquely identify each row. In a snowflake schema, the primary key for the fact table consists of one or more foreign keys. A foreign key is a column or group of columns in one table whose values are defined by the primary key in another table.

When a database is created, the SQL statements used to create the tables must designate the columns that are to form the primary and foreign keys. The following Figure illustrates the relationship of the fact and dimension tables within a simple snowflake schema with a single fact table and three dimension tables. Here Main Report is fact table, Location, Year and Score Subject tables are dimension tables. Address table is the sub-dimension table of location.

Before the data are loaded onto ROLAP using snowflake schema, the data are summarized with the use of cube aggregation and concept hierarchy as in the following section. Figure 5. Explains of the Snowflake Schema

4.5 Cube Aggregation-Based Data Transformation

Some kinds of data reduction techniques are described in the following section
• Data cube aggregation, where aggregation operations are applied to the data in the construction of data cube.
• Dimension reduction, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
• Discretization and concept hierarchy generation, where ranges or the conceptual levels replace raw data values for attributes. Concept hierarchies allow mining the data at multiple levels of abstraction and are a powerful toll for data mining.

Among them, data cube aggregation method is chosen for data summarizing and transformation. Cube aggregation, one of the data reduction techniques, can be applied to obtain a reduced representation of the data set. The data set is much smaller in volume, closely maintains the integrity of the original data. So the mining on the reduced data set should be more efficient produce the same (or almost the same) analytical results.

4.5.1 Concept Hierarchy
The concept hierarchy for location dimension is constructed as follow.

{Lathar, Kyauktada, Innsein … } ⊂ Yangon
{Taungyi, Aung Pan, Kalaw… } ⊂ Shan
{Sagaing, MonYwa, Kalae … } ⊂ Sagaing
{Sittwe, Maungtaw, Yanbae… } ⊂ Rakhine
{Pyin Ma Na, Aung Myae Thar San} ⊂ Mandalay
{Pkokku, Magwe, } ⊂ Magwe
{LoiKaw, Ywarthit, Dimawso } ⊂ Kayar
{MyitKyina, Bamaw, Kathar } ⊂ KaChin
{Bago, Taungngu,Minla} ⊂ Bago
{Pathein, MaungMya, Hintada } ⊂ Ayeyarwaddy
{Htarwai, Yae, Myit,… } ⊂ Tahinntayi
{Harkhar, Tandalan, } ⊂ Chin
{Mawlamyine, Bae Lin, Chaungsone} ⊂ Mon
{Phahan, Mawine, Yardo } ⊂ Kayin

The detail and abstract level of concept hierarchy for location dimension is also illustrated in Figure 5.

```
Whole Myanmar
  State/Division
    City
      Exam Center
```

Figure 6. Concept Hierarchy with Abstract Level

4.5.2 Cube Aggregation using Concept Hierarchy

The data from matriculation warehouse are aggregated at different cuboids levels: 3D-cuboids level and 2D-cuboids level. One of the example for cube aggregation at 3D-cuboid level can be shown in Figure 6.

```
<p>| Year = 2008, Score Subject = Science(S) |</p>
<table>
<thead>
<tr>
<th>Location</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>SanChaung</td>
<td>2000</td>
</tr>
<tr>
<td>Mandalay</td>
<td>1000</td>
</tr>
<tr>
<td>Matayar</td>
<td>50</td>
</tr>
<tr>
<td>Meikhtila</td>
<td>100</td>
</tr>
<tr>
<td>Hlaing</td>
<td>1500</td>
</tr>
</tbody>
</table>
```

Figure 7(a) 3 Dimensional Table (Before Aggregation)

Figure 7(a) becomes the following table as in Figure 7(b) when we conduct the cube aggregation at location dimension. Concept hierarchy is used to replace the detail description of city into their correspondent State and Division name (i.e., lower level description is summed up and replaced with higher level description).

Similarly, another cube aggregation at different cuboids level can be easily done with the help of concept hierarchy as illustrated above.

```
<p>| Year = 2008, Score Subject = Science(S) |</p>
<table>
<thead>
<tr>
<th>Location</th>
<th>No of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bago</td>
<td>3500</td>
</tr>
<tr>
<td>Mandalay</td>
<td>1150</td>
</tr>
</tbody>
</table>
```

Figure 7(b). Three Dimensional Table (After Aggregation)

4.6. OLAP Operations on Matriculation Data Warehouse

Normalized E-R models are often not well suited to addressing the needs of a business. Business information can be requested at the detailed or summary level. Summary-level information often provides input to the strategic direction of the business. In many cases, determining the business questions that are asked by managers as well as finding out how those questions are answered can identify important business information. These business queries may be satisfied by a simple set of criteria, which is often visualized as a three-dimensional cube. A cube is not limited to three dimensions; hypercube of greater than three sides are also possible.

Multidimensional OLAP systems provide a business-oriented solution for answering complex questions. This approach is highly successful when the answers are most often composed of metrics or quantitative data. A different design approach is needed to address the analytical information needs of the business. This approach is characterized by four qualities: a logical design technique for analytical querying; a predictable, standard framework; ability to withstand unexpected changes in user behavior; and existing queries and applications are unaffected by system changes.

4.6.1 OLAP Operations in the Matriculation Multidimensional Data Model

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. The following OLAP operations are taken place on different levels of cuboids.

4.6.1.1 Roll-Up The roll-up operation performs aggregation on a data cube, by climbing up a concept hierarchy for a dimension. The hierarchy was defined as the order of exam center< City< Division. The roll-up operation showed aggregate the data by ascending the location hierarchy from the level city to the level of country. Grouping the Data City, the resulting cube groups the data by Division.
Two aggregation functions SUM and GROUP BY CUBE are used in this Roll Up operations. Here SUM function calls the concept hierarchy.

4.6.1.2 Drill-down Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Either steeping down a concept hierarchy for a dimension or introducing additional dimensions can realize drill down. Drill-down occurs by descending the time hierarchy from the level of location to more detailed level of month. A drill-down adds more details to the given data; adding new dimension to a cube can also perform it.

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

5. Conclusion

In this paper we tries to use and understand the data warehouse architecture and OLAP technology. Therefore we proposed a system for warehousing which is based on the layer architecture of warehouse. The Matriculation Student Information System are used to built the system prototype. Source of data are obtained from different school of Myanmar. This department need to record historical data concerning with student registration and student pass/fail for analysis purpose.

In our system, cube aggregation is introduced to support in the ETL process. The outputs of cube aggregation are loaded onto the ROLAP as Matriculation Warehouse.

According to results, cube aggregation support to one of the aggregation function SUM. Another advantages of our system is to provide not only normal query (because the operational data are also recorded in our system), but also analyzed information to make decision in management level. As an extra, system provide automatically data backed up, archive and input data validation.

The system cannot support to another aggregation functions, such as AVG, MIN, and MAX. Our system only supports to one of the aggregation functions, SUM, with the use of concept hierarchy. This factor is the limitation of our system.

6. About the author.
Zar Ni Mg(M.C.SC, MCTS) is a programmer who has been developing Windows application software, Web developer in Hicom Company Limited, Myanmar since 2006.

REFERENCE


