Data Warehouse and Web-Based OLAP for Hotspot Distribution in Indonesia

G. Hayardisi
PT Oto Multiartha
Jln Jend Sudirman
Kav 61-62 Jakarta 12190
Indonesia
e-mail: gananda_68@yahoo.com

I. S. Sitanggang
Computer Science Department,
Bogor Agricultural University,
Jl. Meranti, Wing 20 Level V,
Kampus IPB Darmaga,
Bogor 16680
Indonesia
e-mail: imas.sitanggang@gmail.com

L. Syaufina
Departement of Silviculture,
Bogor Agricultural University,
Jl. Lingkar Akademik
Kampus IPB Darmaga,
Bogor 16680
Indonesia
e-mail: syaufina@yahoo.com

Abstract—This work aims to develop a web-based OLAP (On-line Analytical Processing) integrated with a data warehouse for hotspot distribution data. The data warehouse development adopts the three-tier data warehouse architecture. The data are represented in multidimensional model using the star scheme which consists of one data cube with two dimension tables i.e. the dimension time and the dimension location, and number of hotspots as the measure. The application applied OLAP operations including roll-up and drill-down and the results are represented in form of crosstabs and graphs. The concept hierarchy defines the sequence mappings in each dimension. In the dimension time, the sequence ordered from the higher-level of hierarchy to the lowest from year, quarter, to month. Whereas, the sequence in the dimension location is from island, province, to regency. The web-based OLAP can be used to better organize and analyze the hotspot data as well as provide a decision-making supporting in forest fire management.

Keywords—data warehouse, hotspot distribution, three-tier architecture, web-based OLAP.

I. INTRODUCTION

In the last decade, forest and land fires occurred in large areas of Indonesia especially in Sumatera and Kalimantan. Those fires caused negative impacts into the environment including transboundary haze pollution, health and life problem, and material damages. Therefore, efforts need to be done to minimize fire occurrences. Hotspot distribution is one of indicators of fire occurrences. The satellite imagery indicates hotspots location in term of geographical coordinates. Hotspot distribution data have been stored in databases without further processing to extract information related to forest fires. Therefore, the work is needed to explore the data and to present the result in the more applicable format. A data warehouse integrated with an OLAP (On-line Analytical Processing) application is one of appropriate solutions to handle the accumulation of hotspot data. Therefore, the information concerning hotspot distribution will be available for the users for strategic decisions in forest fire control.

A data warehouse refers to a data repository system which is stored and maintained separately from daily operational databases. Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions [2]. The characteristics of data warehouses include subject-oriented, integrated, time-variant, and non-volatile [4]. Rather than focus on daily operations, data warehouses provide modeling and analysis of data in a particular subject for decision makers. In developing a data warehouse, usually organizations integrate the data from different sources such as relational databases, flat files, and spreadsheet. The term time-variant in data warehousing refers to an element of time contained in data warehouses either implicitly or explicitly [4]. Data warehouses are nonvolatile. The data in data warehouses are stored separately from operational and transactional databases. Therefore it usually requires only two operations: initial loading of data and access of data [4]. Figure 1 shows the data warehousing architecture [2].

![Data warehousing architecture](image)

Fig. 1. Data warehousing architecture [2].

Data from operational databases and external sources are extracted for data preprocessing: cleaning, transforming, and
integrating the data. The data are then loaded into the data warehouse. Periodically refreshing the data warehouse reflect updates at the sources and to purge data from the data warehouse [2]. In addition to the main data warehouse, the architecture may contain several data marts which represent the data in multidimensional model as in the data warehouse. Both the main warehouse and data marts will serve all tasks in front end tools: analysis, query/reporting, and data mining.

To facilitate data analyses, summarization, and visualization, data warehouses are usually integrated with OLAP applications (see Figure 1). For this purposes, data in data warehouses are represented in multidimensional model that organizes the data in form of cubes. A multidimensional data model consists of a set of numerical measures which are the objects of analysis, and dimensions which provide the context for the measure [2]. Figure 2 illustrates the multidimensional view of data cube sales.

![Example data cube](image)

In Figure 2, attributes product, date, and supplier are referred as dimensions, and attribute sales as a measure. Dimensions usually associate with the a concept hierarchy. For example, day → month → quarter → year is a hierarchy on dimension date that specifies various aggregation levels [1].

Some OLAP operations can be applied to a data cube including roll-up, drill-down, slice, dice and pivot [4]. A roll-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction [4]. A drill-down operation is the reserve of roll-up. A slice operation performs a selection on one dimension of the given cube, resulting in a sub cube. The dice operation defines a sub cube by performing a selection on two or more dimensions. Pivot is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

In modeling a multidimensional database, there are three popular schemes: a star scheme, a snowflake scheme, and a fact constellation scheme [4]. In the star scheme, a data warehouse contains a fact table that store the data, with no redundancy and a set of smaller tables i.e. dimension tables, one for each dimension [4]. Each dimension table is connected to the fact table which contain the key from each dimension table to form a star scheme. In addition to key of each dimension table, the fact table stores measures. The snowflake scheme is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables [4]. The purposes of normalization is to reduce redundancies in dimension tables. The fact constellation can be viewed as a collection of stars [4]. In this scheme, there are multiple fact tables that share dimension tables.

Data warehouses and OLAP technology have been implemented in many areas. The traditional Chinese Medicine (TCM) clinical data warehouse system was developed consisting of several key components: clinical data schema, extraction-transformation-loading tool, online analytical analysis (OLAP) and integrated data mining functionalities [7]. The system is based on the structured electronic medical record system and daily clinical data for TCM clinical research and medical knowledge discovery. A data warehouse system was also implemented utilizing Oracle 10g for air quality data with the monitor data of 86 main cities from the year of 2000 to 2007 in China [8]. The extension of star dimensional model was used in that work to design the system for managing the spatial data with ArcGIS 9.0. In addition, a web spatial online analytical processing (OLAP) has been implemented in order to improve the spatial visualization ability of analysis results [8]. Another application of data warehouse systems and OLAP technology is in drugs sales case study [3]. Microsoft SQL Server 2000 Analysis Services was selected as the data warehouse platform. This work developed the Web-based OLAP structure based on the drugs sales data cube in the data warehouse. The other work presented a data warehouse and an OLAP environment based on farm products transaction MIS in China in order to support the farm products safety testing data analysis [5].

II. DATA WAREHOUSE DESIGN

In this work we adopt the star scheme [4] in developing the data warehouse. The data warehouse contains one fact table, two dimensions: location and time, and the measure is number of hotspot. The hierarchy on the dimension location is district → province → island. The hierarchy on the dimension time is month → year. The data warehouse is then integrated to the OLAP application. The system is constructed in a personal computer with a processor Intel Centrino Duo 1.8 Ghz and 1.5 GB memory. Below are some softwares used in this work:

- Operating system: Windows XP SP 2
- DBMS: Microsoft SQL Server 2000
- OLAP server: Palo Server Win32 2.0
- Palo Excel Add-in 2.0 for managing the data cube
- Web server package WAMPP version 2.0
- Programming language: PHP 5
- JpGraph 1.20 for creating graphs to visualize results of OLAP operations.
- Web browser: Mozilla Firefox 2.0 and Internet Explorer 7
- Palo Excel Add-in 2.0 is a cell-related database that is multidimensional, hierarchical and memory-based [6].
Palo, a cube is a collection of cells, which are defined by two or more dimensions. We use Palo to create a data cube and to load the data into the data cube. In order to load the data stored in the relational database to OLAP server Palo, we write some php files:

- importLocation.php for importing name of islands, provinces, and districts to OLAP server Palo as elements of dimension location.
- importTime.php for importing the element of dimension time.
- monthlyImport.php containing queries to retrieve the data in DBMS SQL Server 2000, then the queries’ results are loaded into OLAP Server Palo using the library in Palo: palo_setdata.

III. WEB-BASED OLAP DESIGN

Our system adopts the warehousing architecture proposed by Chaudhuri and Dayal [2]. The system has three layers as shown in Figure 3.

In the bottom layer, relational databases and Ms. Excel files are integrated to construct the data warehouse. We perform the data preprocessing i.e. cleaning and transforming to improve the quality of the hotspot data before the integration step. The middle layer is the OLAP server implemented using Palo 2.0. The top layer is a front-end client layer which provides facilities for summarization and displaying results of OLAP operations. The results are visualized in forms of crosstabs and graphs.

The web-based OLAP development is implemented using PHP, Palo PHP API, and javascript. We use an additional library in SDK Palo 2.0 package (version 20080118_1000) to connect PHP with the OLAP Server Palo. The application has some php files:

- config.php. This module manages the route of web pages.
- olap.func.inc.php. This module is a part of the file palo_demo.php stored in Palo SDK (Software Development Kit) 1.0c. Palo_demo.php consists of the sub modules: Olap.func.inc.php, fetchData.inc.php, and olapCrosstab.php. Palo_demo.php plays a significant role in creating crosstabs and OLAP operations including roll-up and drill-down. Olap.func.inc.php stores the set of functions required by the system.
- fetchData.inc.php. This module is used in data fetching and creating crosstabs.
- olapCrosstab.php. This module represents the application in the HTML syntax.
- graph folder. It contains two modules: graphBar.php and graphPie.php. These modules utilize the jsGraph library to produce graphs for the results of OLAP operations.

The web-based OLAP provides summary hotspot data and information about the trend of hotspot distribution in Indonesia. The users can perform the OLAP operations such as drill-down and roll-up via a web browser. The application has some features:

- The system can be applied to other databases, data cubes, and dimensions, not limited to hotspot data.
- Crosstabs and graphs in form of bar plots and pie plots allow the users to analyze the hotspot distribution data in regions in Indonesia for a period of time such as yearly, quarterly, and monthly.
- The users can explore the data in different hierarchies on dimension time and location by executing OLAP operations roll-up and drill-down.
- The system has the dimension filter to select elements of dimension that will be displayed in the x-axis and the y-axis.

IV. OLAP OPERATIONS ANALYSIS

The web-based application provides easy access for users who want to get hotspots distribution in various locations and time as they can use this application either in Intranet or in Internet. The main page of the system is shown in Figure 4. In this page, the users can select a cube containing analyzed objects that will be processed by applying OLAP operations. In addition to a cube, the users can choose the dimension both for the x-axis and the y-axis. The filter menu will facilitate the users to display the summary or graphs for a particular element of selected dimension.

Figure 5 illustrates the graphs resulted from the drill-down operation on dimension time and location. The element of dimension time is year, whereas the element of dimension
**location** is island. When the hierarchy of element **island** is decreased to **province**, then the results will be plotted as in Figure 6.

![Fig. 5. Bar graph for the result of drill-down operation, where the element of dimension **time** is year and the element of dimension **location** is island.](image)

![Fig. 6. Bar graph for the result of drill-down operation, where the element of dimension **time** is year and the element of dimension **location** is province.](image)

Figure 5 indicates that Kalimantan island has the highest number of hotspot in 2002 with 17,938 hotspots, and Sumatra island at the second place with 8,291 hotspots. Refer to dimension **time**, Indonesia has the highest number of hotspot in 2002 of about 26,684 hotspots. Figure 6 shows that Central Kalimantan (Kalimantan tengah) has the highest number of hotspot in Kalimantan in 2002 with 10,946 hotspots.

![Fig. 7. Bar graph for the result of drill-down operation, where the element of dimension **time** is year and the element of dimension **location** is province.](image)

In addition to bar plots, pie plots represent hotspot distribution in percentage. Figure 7 illustrates number of hotspot (%) for the element of dimension **time** is year and the element of dimension **location** is province.

Our future work will focus on several improvements including 1) expanding the level in the concept hierarchy such that the lowest level becomes day and sub district for the dimension **time** and the dimension **location** respectively; 2) implementing the pivot operation in order to transpose the axis in crosstabs; 3) developing an updating module to handle new hotspot data.

**V. CONCLUSION**

The data warehouse for hotspot distribution has one data cube contains number of hotspot in regions in Indonesia with hierarchy of location from island to district for the period of 2000 to 2004. The data warehouse is integrated to the OLAP application in order to provide the users summarization and visualization for the results of OLAP operations for decision making related to the forest fires. The OLAP operations including roll-up and drill-down allow the users to explore the data cube with different hierarchies on both dimension **time** and **location**. The results are represented in crosstabs and graphs (bar plots and pie plots). The results indicates that Kalimantan island has the highest number of hotspot in 2002 with 17,938 hotspots, and Sumatra island at the second place with 8,291 hotspots. Refer to dimension **time**, Indonesia has the highest hotspot number in 2002 of about 26,684 hotspots. At the element **province**, Central Kalimantan has the highest number of hotspot in Kalimantan in 2002 with 10,946 hotspots.

**REFERENCES**


