Global Research
Researcher Forum

Journal of Global Research Computer Science & Technology
(JGRCST)
Vol-I, Issue-III, August 2014
ISSN: 2349 - 5170

Data Mining and OLAP Operations

Pankaj Ramesh Vasnani
School of Computer and Systems Sciences
Jaipur National University
Jaipur

Antim Arora
School of Computer and Systems Sciences
Jaipur National University
Jaipur

Nidhi Mittal
School of Computer and Systems Sciences
Jaipur National University
Jaipur

Abstract: This research looks at the work, history and research work done in the field of Data mining and OLAP operations. Today, a large amount of data is present in the data warehouse, but this data is of no use until it is converted into useful information so that the managers can take fast and effective decisions. Data mining and its different techniques facilitates the managers by providing them with the relevant information. Along with Data mining, it also emphasises on KDD process. This research work also emphasises on the history of OLAP and its applications in the real business environment. The main objective of our research paper is to describe the real world practical applications of Data mining, OLAP and its operations.

Keywords: Data Mining, OLAP, Multidimensional Databases, OLAP cube

1 Introduction

Databases today can range in size into the terabytes - more than 1,000,000,000,000 bytes of data. But this data is of no use to organization unless it is converted into useful and purposeful information for the managers to take quick and smart decisions to achieve desired organisational goals and objectives. But then, how do we derive the meaningful information from a large amount of data stored in the data warehouse.

The newest answer is data mining, which is being used both to increase revenues and to reduce costs. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud.

Data Mining is defined as a collection of different tools and techniques which are applied to extract the useful and relevant information from a large amount of data stored in the data warehouse in the required structure and format such as reports,

charts, pie graphs, etc. that helps the managers to take fast and intelligent decisions to achieve desired organisational goals and objectives.

Today, managers aren't satisfied by getting direct answers to their direct questions. Instead due to the market growth and increase of clients, their questions became more complicated. Questions are like" How much profit from selling our products at our different centers per month". A complicated question like this isn't as simple to be answered directly. To answer these types of complex queries and problems, data mining techniques is used by the managers so that they can get a perfect solution to their queries and problems for taking fast and intelligent decisions.

The overall goal of the data mining process is to extract the useful information from a data set and transform it into an understandable structure for further use. Data Mining (DM) is the core of KDD process, mainly concerned with exploring the data, developing the model and discovering previously unknown patterns. The accessibility and abundance of data in the data warehouse makes knowledge discovery process and data mining technique a matter of considerable importance and necessity.

2 Work Related

The origin of OLAP technology was traced to **1962, Kenneth Iverso**n introduced the base foundation of OLAP through his book "A Programming Language" (APL), which defined a mathematical language with processing operators and multidimensional variables. The first product that performed OLAP queries was Express, which was released in 1970(acquired by Oracle in 1995 from Information Resources) and first OLAP product launched in 1975 by Information Resources**.** the first spreadsheet application —VisiCalc was introduced to the market in 1979 and was distributed by VisiCorp previously called Personal Software and developed by Dan Bricklin and Bob Frankson.

The term OLAP or Online Analytic Processing was coined in 1993 by Edgar F. Codd, who has been described as "the father of the relational database".

The products of OLAP are as follows:-

i. ***First OLAP Product-Express:*** In 1975 the first OLAP product—Express was launched by Information Resources. This was the first multidimensional tool to support marketing related demands or application needs. It later on evolved into a hybrid OLAP after its acquisition by Oracle and has thrived for more than three decades. It remains, even till date as one of the well-marketed multidimensional products.

ii. ***First Spreadsheet Program - VisiCalc:*** In 1979, the first spreadsheet application —VisiCalc was introduced to the market. This product was originally released for Apple II. VisiCalc was distributed by VisiCorp previously called Personal Software and developed by Dan Bricklin and Bob Frankson.

iii. ***OLAP for Financials - System W:*** By the year 1982, System W was the first OLAP tool to cater to financial applications and the first to apply hypercube approach in its multidimensional modeling. It was even less favored by technical people because it was more difficult to program in comparison with other software of its kind.

iv. ***Lotus 1-2-3:*** In 1983, Lotus 1-2-3 was launched. It was similar in structure to Visicalc but gained more sales and quickly replaced Visicalc. Lotus 1-2-3 became the mainstream spreadsheet application before the Windows. Lotus Software is now a part of IBM. Lotus 1-2-3 incorporated graphing and database functions, keyboard commands and menus much like spreadsheet applications today.

3 Knowledge Discovery in Database (KDD Process)

KDD process or Knowledge Discovery Process is defined as the process of discovering useful knowledge from a large collection of data.

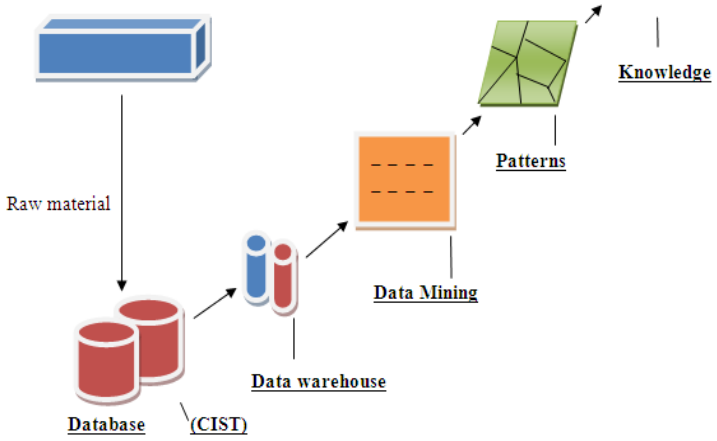

Fig 1: Knowledge Discovery In Database (KDD) process

Practical Example of KDD Process: Firstly, store manager will collect the information regarding the customers and the products purchased by different customers from different sources which is then converted in normalized form by Data Cleaning, Data Integration, Data Selection and Data Transformation. According to the example, in **Data Cleaning process**, store will remove the data of those customers who does not frequently purchase products from them and will also remove the data of those products which, according to store, are not going to be purchased by the customers. Then, in **Data integration process**, it will integrate the data of customers and products collected by it by using different techniques such as interview,

market survey, questionnaires, etc. Then, in **Data Selection process**, it will select only the data of those products which are frequently purchased by the customers for their further decision making process.

Then, in **Data Transformation process**, the information regarding the products and customers is being converted in a structured format such as any data structure like stack, cube, or any other format so that the store manager can use this information for further decision making process and also store this information in the refined format in the **Data Warehouse** to facilitate extraction of relevant and purposeful information for later reference.

The information regarding the different products and customers so obtained in the refined format or structured format is stored in **Data Warehouse** so that relevant and purposeful information can be extracted from it to facilitate the store manager to effective and efficient decisions. After this, the store manager will use different data mining tools and techniques to extract the relevant information about those products which are frequently purchased by the customers and also of those customers who are frequently visiting and purchasing the products from the store.

After applying the different data mining tools and techniques on the information stored in data warehouse to fetch the relevant information regarding the different products and customers, the useful and purposeful information will be generated in the required structure and format such as summarized report, charts, graphs or any other format as requested by store manager to take fast and intelligent decisions. After getting the required and relevant information in the desired structure, the store manager will take the effective decision regarding what new features should be added in the products frequently purchased by the frequent customers or what offers should be offered on the products being frequently purchased by the customers so that he can increase its profit margin to achieve competitive advantage over its competitors in the rapidly changing environment.

4 Online Analytical Processing(OLAP)

Before exploring the OLAP technology, it is very important to discuss about the important terms **Multidimensional Databases** and **OLAP cube** which plays a very important role in storing and organizing the aggregated data and easily provides it as and when required, for performing different OLAP operations on it.

Multidimensional Databases is defined as a variation of the relational model that uses the multidimensional structure to organize the data and express the relationship between the data. In a multidimensional database, the data is presented to its users through multidimensional arrays and each individual value of data is contained within a cell which can be accessed by multiple indexes. It mainly uses the concept of data cube (also known as a hypercube) to represent the dimensions of data

available to the end-users. These types of databases are mainly designed to assist with decision support systems. The detailed storage, classification and organization of data allows the end-users to generate complex and advanced queries to fetch the required information while ensuring effective and efficient performance.

This type of database is mainly used for optimizing OLAP and data warehouse applications. For example- the sales data can be stored and organized according to the different dimensions such as cities, year and product as depicted in the following Fig 2.

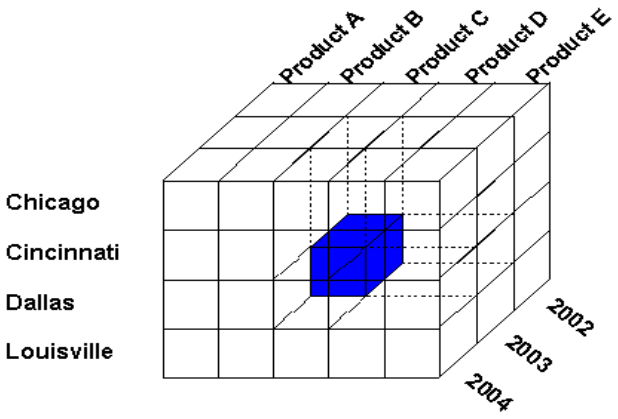

Fig 2: Sales Data Organized According to the different dimensions

In the above example as the sales data is organized according to cities, year and products, end-user can easily issue advanced and complex queries like "What was the sale of Product A in 2002 and 2003 in Chicago? Or what was the sale of Product A in Chicago, Cincinnati, Dallas and Louisville? "Corresponding to the requested query, the end-user will get the desired Information in less amount of time so that he/she can take fast and intelligent decisions like strategies to be made to increase the sales, offers to be provided to the customers to increase the sale, etc.

OLAP cube is defined as a data structure that can be viewed as an array of data understood in terms of 0 or more dimensions. It provides an easy to use mechanism for querying data with quick and uniform response times. End users use client applications to connect to an Analysis server and query the cubes on the server. In most client applications, end users issue a query on a cube by manipulating the user interface controls, which determine the contents of the query. This spares end users from writing language-based queries. Precalculated summary data called aggregations provides the mechanism for rapid and uniform response times to queries. Aggregations are created for a cube before end users access it. The results of a query are

retrieved from the aggregations, the cube's source data in the data warehouse, a copy of this data on the Analysis server, the client cache, or a combination of these sources. Every cube has a schema which is the set of joins tables in the data warehouse from which the cube draws its source data. The central table in the schema is the fact table and the other tables are dimension tables

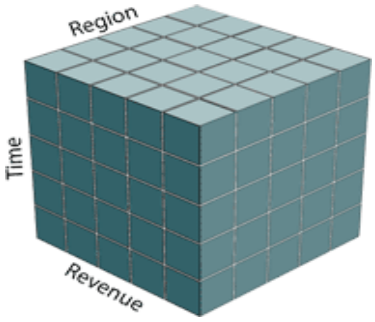

Fig 3: OLAP Cube

OLAP is a latest technology that presents the user with the useful information rather than just the raw data. They make it very easy for the users to search for the useful and interesting patterns of correct and relevant information, without the need for them to search through mountains of raw data. Typically, this analysis is driven by answering business questions such as "How our sales are doing this month in South America?" From these foundations, OLAP moves into different areas such as forecasting and data mining, allowing users to answer questions such as "What are our predicted profit next year?" Basically, On-Line analytical processing (OLAP) is a software technology that enables analysts, manager and executives to gain insight into data through fast, consistent, interactive access in 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.

4.1 Need of OLAP

There are many advantages that an organization can gain from using OLAP technology so as to change the way it retrieves the useful information from mountain of data stored in the data warehouse, which are as follows:-

1. Data can be viewed from different angles, which gives a broader perspective of a problem unlike other models.

2. It has been claimed that for complex queries OLAP cubes can produce an answer in around 0.1% of the time required for the same query on OLTP relational data. The most important mechanism in OLAP which allows it to achieve such performance is the use of *aggregations*.

3. Another main benefit of using OLAP technology is that it ensures consistency of information and calculations. No matter how fast data is processes through OLAP servers, the summarised report of relevant information is presented in a consistent presentation, so analysts and executives always gets accurate results for their queries and always knows what to look and where.

4. Another benefit of using OLAP technology is that it allows the manager to pull down the data from an OLAP database in broad or specific terms. In other words, reporting can be as simple as comparing a few lines of data in one column of a spreadsheet or as complex as viewing all aspects of mountain of data.

5. The information extracted from an OLAP database facilitates all the business needs i.e. planning, budgeting, forecasting, reporting and analysis.

4.2 OLAP Operations

1. Roll-Up –It is the first operation used in OLAP technology. This operation is an efficient way to summarize the data stored in the database. It performs aggregation on data cube by dimension reduction. Dimension reduction means one or more dimensions are removed from the cube. It means that through this operation, we can easily convert the detailed data into summarized data.

2. Drill Down – Drill Down operation is the opposite of Roll Up operation. This operation is performed by introducing new dimensions. It navigates from less detailed data to highly detailed data. For ex- taking the above example of sales, suppose the sales data is stored in the database year-wise and the end-user wants the data to be displayed month-wise. So Drill Down operation is performed by introducing more dimensions as sales data is displayed month-wise to the end-user.

3. Slice – Slice is defined as the operation of extracting a rectangular subset by choosing a single value for one of its dimensions. In other words, it means that the end-user can easily and selectively extract a subset of multi-dimensional array corresponding to a single value of one dimension. For ex –suppose the sales data is stored year wise and city wise in the data warehouse and the end-user wants to know the sales of all the years for the Jaipur city, then he or she can easily use slice operation to extract the sales information of all the years for the Jaipur city.

4. Dice –Dice is defined as the operation of extracting a rectangular subset by choosing specific values on more than one dimension. In other words, it means that the analyst can easily extract the subset of multidimensional array corresponding to specific values from more than one dimension. For ex- from the above example the analyst can easily extract the sales

information for the year 2003 for the Jaipur city by performing the dice operation on the sales information stored in the data warehouse.

5. Pivot –Pivot is an operation that allows the analysts to rotate the cube in different directions so as to view the data in different directions. For ex –the analysts can vertically arrange the sales data and horizontally arranges the time period(in years) data, while viewing the data for the sales in a particular city.

5. Practical Implementation of OLAP operations

The table which has been discussed below is created in SQL*Plus and used for the practical implementation of OLAP operation is as follows:-

a. Data of the Table:

CITY_NAME	PRODUCT_ID	YEAR	SALES
Jaipur	P101	2000	34000
Kanpur	P101	1990	45000
Udaipur	P102	2001	12000
Kanpur	P102	2002	3000
Delhi	P103	2002	30000
Pune	P103	2003	24000
Mumbai	P104	2004	34000
Bhopal	P104	2005	38000
Bengaluru	P105	2008	124000
Kolkata	P105	2010	31000

b. Roll-Up:

1. Query: Select sum (sales) “Total Sales” ,product_id from cities group by product_id;

2. Output:

TOTAL SALES	PRODUCT_ID
79000	P101
15000	P102
54000	P103

72000	P104
155000	P105

3. Explanation – In the above query, we have used **Roll Up** operation to show the summarized total sales of different products according to different cities grouped by their **PRODUCT_ID.**

c. Drill-Down:

1. Query: Select city_name, product_id , sales from cities;

2. Output:

CITY_NAME	PRODUCT_ID	SALES
Jaipur	P101	34000
Kanpur	P101	45000
Udaipur	P102	12000
Kanpur	P102	3000
Delhi	P103	30000
Pune	P103	24000
Mumbai	P104	34000
Bhopal	P104	38000
Bengaluru	P105	124000
Kolkata	P105	31000

3. Explanation – In the above query, we have used **Drill Down** operation to show the detailed information about the sales of different products in different cities.

d. Slice:

1. Query: Select * from cities where product_id='P101';

2. Output:

CITY_NAME	PRODUCT_ID	SALES
Jaipur	P101	34000
Kanpur	P101	45000

3. Explanation – In the above query, we have used **Slice** operation to show the detailed information about the sales of product id **P101** in Jaipur and Kanpur.

e. Dice:

1. Query: Select * from cities where product_id='P101' and city_name='Kanpur';

2. Output:

CITY_NAME	PRODUCT_ID	SALES
Kanpur	P101	45000

3. Explanation – In the above query, we have used **Dice** operation to show the detailed information about the sales of product id **P101** in only Kanpur.

f. Pivot:

1. Query: Select year, city_name, product_id, sales from cities;

2. Output:

YEAR	CITY_NAME	PRODUCT_ID	SALES
2000	Jaipur	P101	34000
1990	Kanpur	P101	45000
2001	Udaipur	P102	12000
2002	Kanpur	P102	3000
2002	Delhi	P103	30000
2003	Pune	P103	24000
2004	Mumbai	P104	34000
2005	Bhopal	P104	38000
2008	Bengaluru	P105	124000
2010	Bhopal	P105	31000

3. Explanation – In the above query, we have used **Pivot** operation to change the visual representation of the detailed information about the sales of different products in different cities.

6. Conclusion

Data mining and its techniques simplifies how to extract knowledge or information from data storage(data warehouse).Data mining is applied effectively not only in business environment but also in other fields such as weather forecast, medicine,

transportation, healthcare, insurance, government etc. The OLAP technology provides better performance for accessing multidimensional data. OLAP is a technology that can be distributed to many users on a variety of platforms. It provides the foundation for analytical processing through flexible information access. There are many on-going researches on Data mining and OLAP. Our purpose of this research is to give easy and understandable concepts about data mining and OLAP. Along with these we have also discussed the practical implementation of OLAP operations.

References

1. http://en.wikipedia.org/wiki/Data_mining
2. http://en.wikipedia.org/wiki/OLAP_cube
3. http://office.microsoft.com/en-in/excel-help/overview-of-online-analytical-processing-olap-HP010177437.aspx
4. http://resources.businessobjects.com/support/communitycs/TechnicalPapers/si_intro_olap.pdf
5. http://technet.microsoft.com/en-us/library/aa216365(v=sql.80).aspx
6. http://web.cs.wpi.edu/~cs561/s12/Lectures/IntegrationOLAP/OLAPandMining.pdf
7. http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf
8. http://web.mit.edu/profit/PDFS/SlaughterA.pdf
9. http://www.ifis.cs.tu-bs.de/webfm_send/922
10. http://www.olaphouse.com/documents/home/benefits-of-olap.xml?lang=en
11. http://www.slideshare.net/hasanshan/Multidimensional-Database-Design-Architecture
12. http://www.techopedia.com/definition/14710/multidimensional-database-mdb
13. http://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd
14. http://www.twocrows.com/intro-dm.pdf