OLAP Techniques for Approximation and Mining Query Answering

Murtadaha M. Hamd **Waleed K. Hassan**
College of Computer
University of Anbar

Received on: 20/10/2010 **Accepted on: 10/11/2010**

ABSTRACT

Data warehouse (DW) systems have become a key component of the corporate information system architecture, in which they play a crucial role in building business decision support systems by collecting and consolidating data from a variety of internal and external sources. The content of a DW is analyzed by the On-Line Analytical Processing (OLAP) applications for the purpose of discovering trends, patterns of behavior, and anomalies as well as finding hidden dependencies between data . Through this research, theoretical concepts which are related with OLAP and Data Warehouse are studied in detail. In this work the SALES Data Warehouse (**SALESDW**) , procedures are implemented and algorithms are developed like ODBC (Open DataBase Connectivity) for different sources, and Data cleansing to carry out the consistency of Data Warehouse (DW). The aim of the research are design a prototype of a SALES Data Warehouse (SALESDW) by adding and implementing the essential concepts , and implementing the OLAP techniques (tools) on SALESDW.

Keywords: On-Line Analytical Processing (OLAP), SALES Data Warehouse (SALESDW), ODBC (Open DataBase Connectivity).

تقنيات OLAP للتقريب والتنقيب في الرد على الاستعلام

مرتضى حمد **وليد حسان**
كلية الحاسوب ، جامعة الانبار

تاريخ الاستلام: 2010/10/20 **تاريخ القبول: 2010/11/10**

الملخص

أنظمة مخازن البيانات أصبحت عنصرا رئيسيا في معمارية نظام المعلومات، والتي تلعب دورا حاسما في بناء نظم دعم القرار التجارية. من خلال جمع وتوحيد البيانات من مجموعة متنوعة من المصادر الداخلية والخارجية. محتوى مخازن البيانات يتم تحليلها عن طريق استخدام تطبيقات ما يسمى المعالجة التحليلية المباشرة (OLAP) لغرض اكتشاف اتجاهات وأنماط السلوك ، والشذوذ ، وكذلك لإيجاد التبعيات المخبأة بين البيانات. من خلال هذا البحث ، الكثير من المفاهيم النظرية التي ترتبط مع المعالجة التحليلية المباشرة (OLAP) ومخازن البيانات سيتم دراستها بالتفصيل. في النموذج الأولي مخزن البيانات (SALESDW)، سيتم تنفيذ العديد من الإجراءات والخوارزميات مثل ODBC (قاعدة البيانات المفتوح) لمصادر مختلفة ، تنظيف البيانات لإطلاق اتساق مستودع البيانات. الغرض من هذا البحث هو تصميم نموذج اولي لمستودع بيانات خاص بمؤسسة مبيعات معينة وتنفيذ المعالجة التحليلية عليها .

الكلمات المفتاحية: المعالجة التحليلية المباشرة ، مخزن البيانات (SALESDW)، ODBC (قاعدة البيانات المفتوح).

1. Introduction:

OLAP uses a Snapshot of a database taken at one point in time and then puts the data into a dimensional model. The purpose of this model is to run queries that deal with aggregations of data rather than individual transactions.[1] Data warehousing and On-

Line Analytical Processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Many commercial products and services are now available, and all of the principal database management system vendors now have offerings in these areas. Decision support places some rather different requirements on database technology compared to traditional On-Line Transaction Processing (OLTP) applications. [2] OLAP tools allow users to make ad hoc queries or generate canned queries against the warehouse database. The OLAP category has been divided further into the multidimensional OLAP (MOLAP) and relational OLAP (ROLAP) markets.[3]

2 .Literature Survey:

- **Dimitris Papadias, Panos Kalnis, Jun Zhang and Yufei Tao. Proposed in 2001, "Efficient OLAP Operations in Spatial Data Warehouses ".**

They deal with the problem of providing OLAP operations in spatial data warehouses. Such warehouses should support spatial dimensions, i.e. allow the user to execute aggregation queries in groups, based on the position of objects in space. Although there exists well-known pre-aggregation techniques for non-spatial warehouses, which aim to speed up such queries. They presented an example of a traffic supervision system; other applications include decision support systems for cellular networks, weather forecasting, etc. [4]

- **Yufei Tao, Xiaokui Xiao proposed in 2008 "Efficient Temporal Counting with Bounded Error ".**

Temporal aggregation is an important operator for two reasons. First, aggregates are the direct target of analysis in a large number of applications of temporal databases. Second, the numbers of objects qualifying various range predicates are essential inputs to many sophisticated data mining tasks, such as association rule mining, decision tree learning, etc. motivated by the fact that precise aggregation demands expensive space or query overhead. They propose a novel technique for efficiently computing approximate results with good quality guarantees. [5]

3. Data Warehouse:

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. In addition to a relational database, a data warehouse environment can include an Extraction, Transportation, Transformation, and Loading (ETL) solution, Online Analytical Processing (OLAP) and Data Mining capabilities, Client Analysis Tools , and other applications that manage the process of gathering data and delivering it to business users . A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inmon: [6]

- Subject Oriented.**[6]**

Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a data warehouse that concentrates on sales. Using this data warehouse, you can answer questions such as "Who was our best customer for this item last year?" .

- Integrated.**[6]**

Integration is the most important. Data is fed from multiple disparate sources into the data warehouse. As the data is fed , it is converted, reformatted, resequenced, summarized, and so forth. The result is that data once it resides in the data warehouse

has a single physical corporate image . Figure -1 illustrates the integration that occurs when data passes from the application-oriented operational environment to the data warehouse.

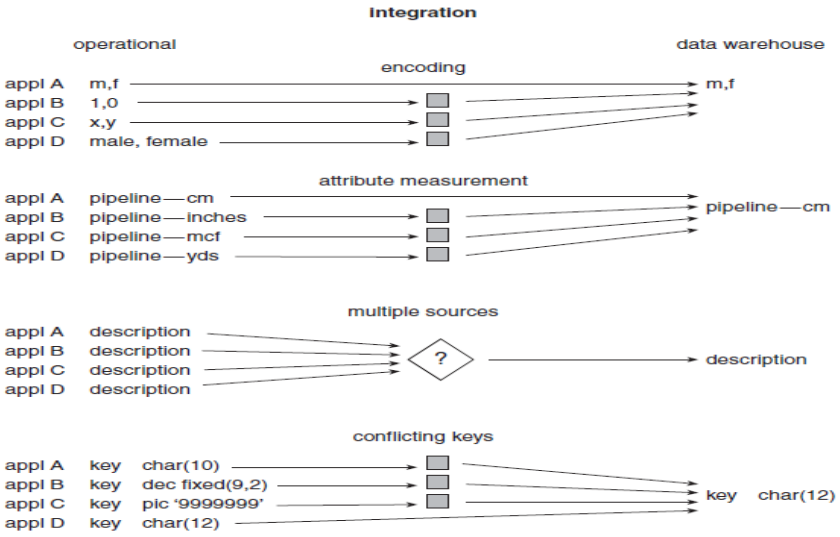

Figure -1The issue of integration

❋ Nonvolatile**.[6]**

Nonvolatile means that, once entered into the data warehouse, data should not change. This is logical because the purpose of a data warehouse is to enable you to analyze what has occurred.

❋ Time Variant **.[6]**

A data warehouse's focus on change over time is what is meant by the term time variant. In order to discover trends in business, analysts need large amounts of data.

4.OLAP :

Data warehouses and OLAP are necessary elements of Decision Support Systems (DSSs). They enable business decision makers to creatively approach, analyze and understand business problems. While data warehouses are built to store very large amounts of integrated data used to assist the decision-making process, the concept of OLAP, which is first formulated in 1993 by [7] to enable business decision makers to work with data warehouses, supports dynamic synthesis, analysis, and consolidation of large volumes of multidimensional data. Two of the most important ways to pursue high performance and usability of Data Cube computation are: [8]

a) Speeding Up , b) Reducing Storage Space.

OLAP Guidelines: [1, 10]

Multidimensionality is at the core of a number of OLAP systems (database and front-end tools) available today. Dr. E. F. Codd, the "father" of the relational model, has formulated a list of guidelines and requirements as the basis for selecting OLAP systems:

1- Basic Features (Multidimensional Conceptual View, Intuitive Data Manipulation , Accessibility , Batch Extraction versus Interpretive Extraction , OLAP Analysis Models , Client / server architecture , Client / server architecture , Transparency , Multi-user support) .

2- Special Features (Treatment of Nonnormalized Data , Store OLAP Results ,

Treatment of Missing Values).

3- Reporting Features (Flexible Reporting , Uniform Reporting Performance)

Categorization of OLAP [1]

ROLAP (Relational OLAP):

This type uses relational databases (RDMS) to store the data, sometimes by using a star schema or snowflake schema. ROLAP tools present sophisticated SQL and navigational methods on top of traditional relational databases. Relational OLAP tools present scalable, manageable technologies for very large data. ROLAP databases can easily handle dimensions with high cardinality. [9] Figure -2 shows the architecture of the ROLAP model. What you see is three-tier architecture.

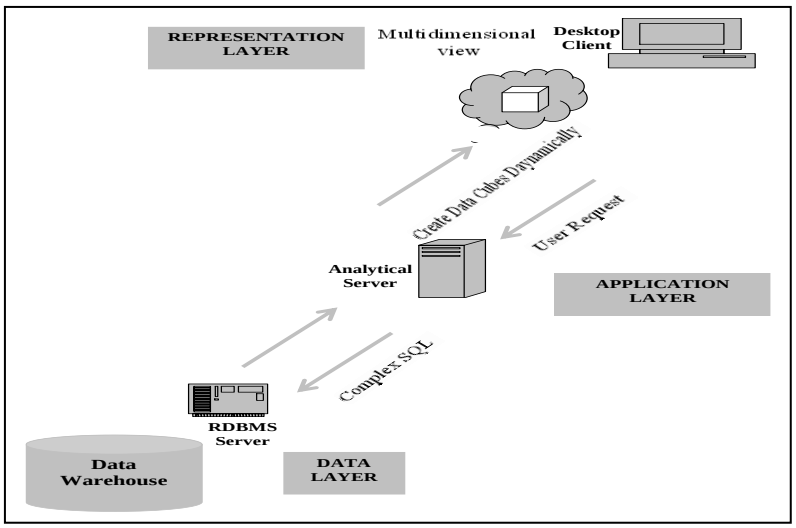

Figure-2 the architecture of the ROLAP

Advantages: [10]

- It can handle large amounts of data.
- It can leverage functionalities inherent in the relational database.

Disadvantage : [10]

- Performance can be slow because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database.
- It is limited by SQL functionalities because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs .

MOLAP (Multidimensional OLAP):

In the MOLAP, data is extracted from the data warehouse and aggregated into a data structure, commonly referred to as a cube, for analysis.[11]Uses a specialized data store with preaggregated summaries to store the data. The MOLAP data store is built specifically to handle multidimensional queries and offers fast, efficient, and manageable access to multidimensional data.[9] Figure -3 shows the architecture of the MOLAP model.

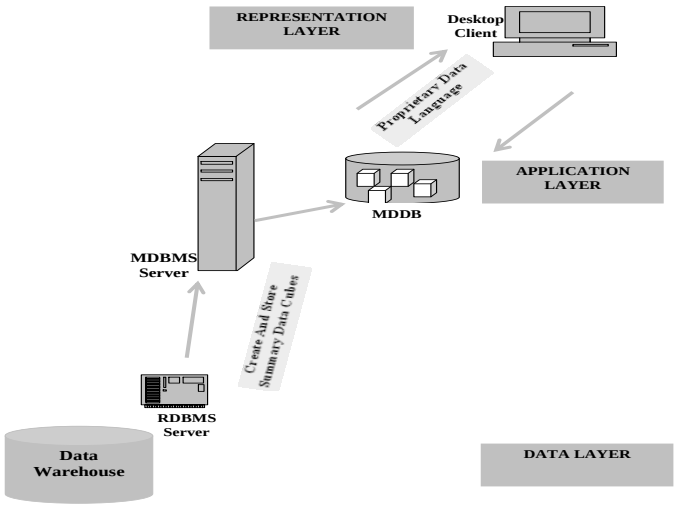

Figure-3 architecture of the MOLAP model

Advantages: [10]

- It has excellent performance: A MOLAP cube is built for fast data retrieval, and is optimal for slicing and dicing operations.
- It can perform complex calculations.

Disadvantage [10]

- Limited in the amount of data it can handle because all calculations are performed when the cube is built
- It requires additional investment: Cube technologies is often proprietary and do not already exists in the organization.

HOLAP (Hybrid OLAP):

It bridges the technology gap between ROLAP and MOLAP, enabling you to use both multidimensional data stores (MDDB) and RDBMS data stores. [9]

Proposed SALESDW Design

Essential files structures for the (SALESDW):

First of all, the tables must be created according to the proposed system (i.e. SALES Data Warehouse) (**SALESDW)** which is used for storing the system information. Table -1 explains the essential files structure for the SALESDW Data Warehouse.

Table -1 Essential files for the SALESDW Data Warehouse

File Names	Description
PRODUCTS	Dimension table
TIMES	Dimension table
COUNTRIES	Dimension table
CUSTOMERS	Dimension table
SALES	Fact table

The following steps illustrate the creation of tables and describe the operations for one table that are included in SALESDW Data Warehouse by using SQL Plus tool in Oracle Data Base 9i:

SQL> CREATE TABLE products(

1 prod_ID CHAR(6) PRIMARY KEY,
2 PROD_NAME VARCHAR2(50) NOT NULL,
3 PROD_DESC VARCHAR2(1000) NOT NULL,
4 PROD_SUBCATEGORY VARCHAR2(50) NOT NULL,
5 PROD_CATEGORY VARCHAR2(50)NOT NULL,
6 PROD_CAT_DESC VARCHAR2(1000) NOT NULL,
Table created.

System Design:

To start the system , go to start > all programs > oracle developer 6.0 > Form builder as in figure 4 :

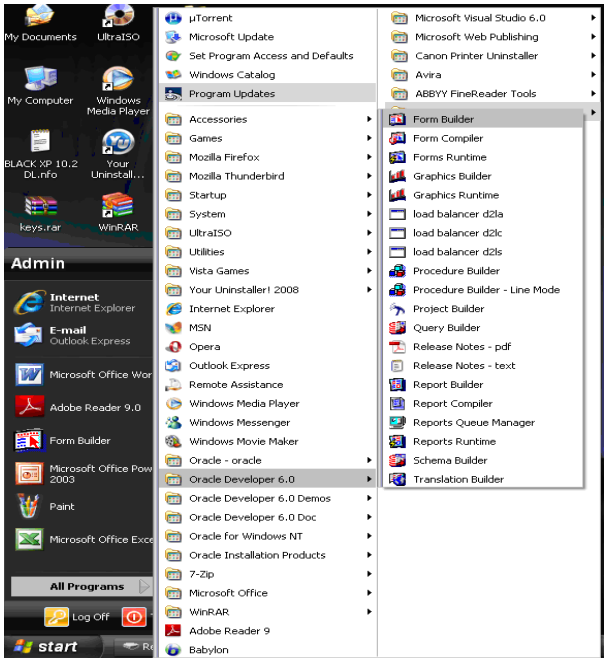

Figure 4

The project will be loaded to the form builder application, to run the project , we should connect to oracle Database as following (see figure 5):

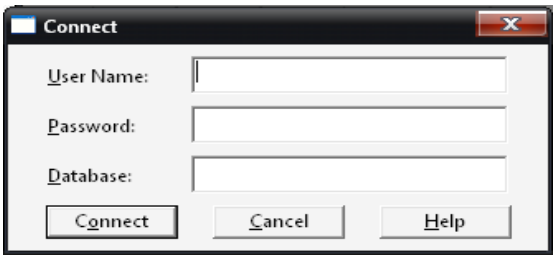

Figure 5

Then the project is connected to oracle Database and the main interface is comes up and is ready to execute queries, as showed in figure 6

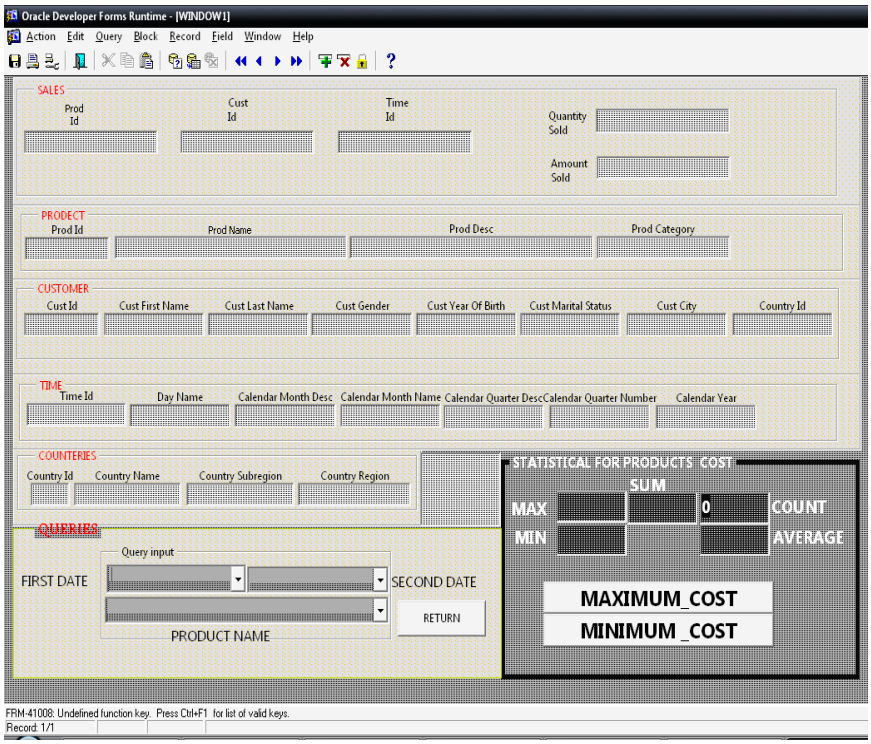

Figure 6 the main interface our proposed system

Implementation of Proposed System :

Executing the proposed system will pass through the following steps:

1- Chose the product name.

2- Chose the first time which is representing the start of the interval.

3- Chose the second time which is representing the start of the interval.

Figure -7 explains the resulting of the query .Also we can obtain the statistical resulting which is related to result of the query. Depending on statistical result we can build the DSS that clarify in the figures 8.

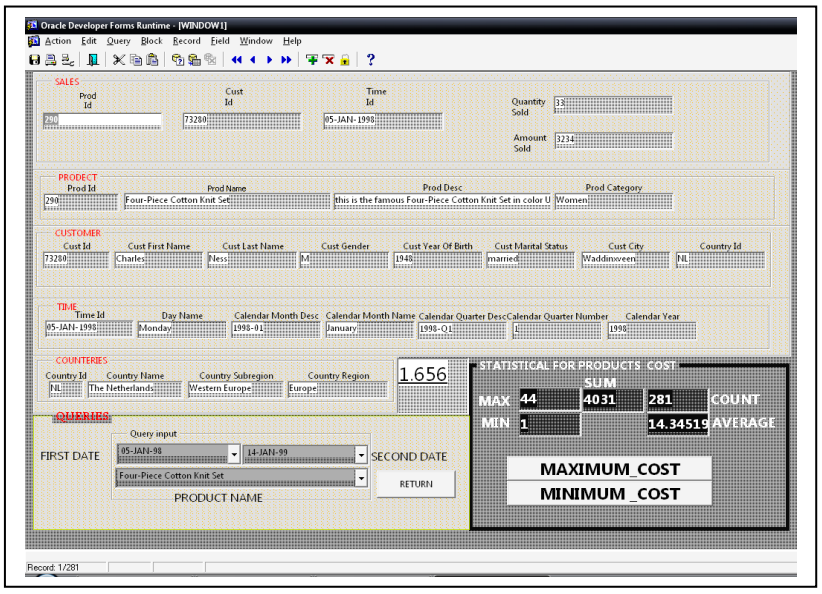

Figure 7 explains the resulting of the query

Figure 8 the response and suggested of DSS

The figure 9 shows the final results of the system which are related with statistical results

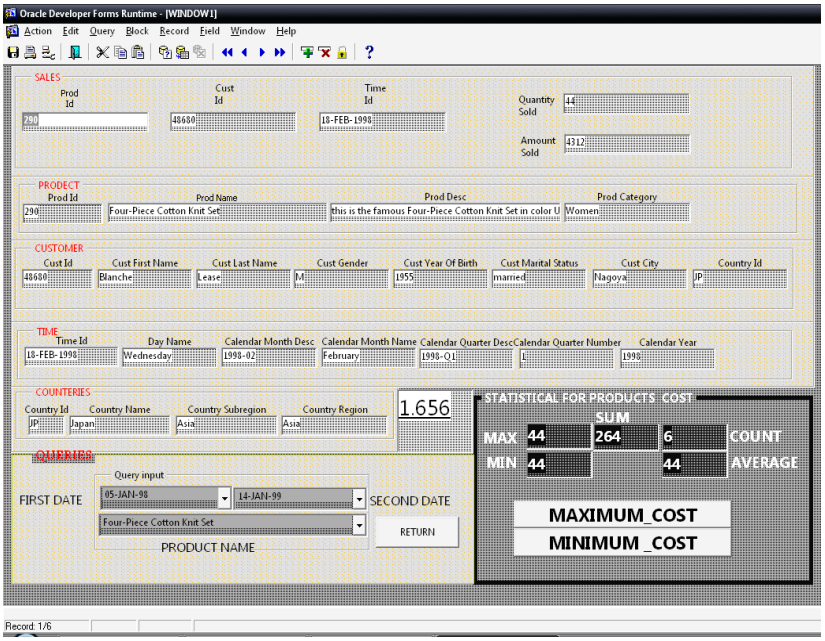

Figure 9 explains the Final Results of the System

Data Warehouse Schemas:

The most important two Schemas (Star Schema and Snowflake Schema in addition to Fact Constellation) as theoretical, these Schemas are implemented practically on SALESDW to determine the response time for the all Queries as shown in table- 2, in addition to the structure for all Queries using SQL in Visual Basic 6.0.
Figures 10 show the implementation of snowflake schema with some of cods the implemented using SQL in Visual Basic 6.0 and with number of records (250000, 500000, and 1000000).


```
SQL = "SELECT SUM(SALES_DOLLAR)AS SUM1,COUNT(*)AS DD,PNAME,CAT_NAME,DAY
FROM SALES, PRODUCT,CITY ,TIM,CUSTOMER,CATEGORY
WHERE SALES.PRODUCT_ID = PRODUCT.PRODUCT_ID
AND SALES.CITY_ID = CITY.CITY_ID
AND SALES.TIME_ID = TIM.TIME_ID
AND SALES.CUST_ID=CUSTOMER.CUST_ID
AND PRODUCT.CAT_ID=CATEGORY.CAT_ID
AND CUSTOMER.CUST_ID <=1000000
GROUP BY PNAME,CAT_NAME,TIM.DAY  "
```

Figure **10 Explain Snowflake Schema with Daily Results and the Complex Query Used**

Table-2 explain in detail the comparing between the star and snowflake

RECORD NO.	GROUP BY	SNOWFLAKE SCHEMA / Time	STAR SCHEMA / Time
250000	YEARLY	7.0468	6.5937
500000		8.453	8.906
1000000		10.860	19.796
250000	MONTHLY	5.5146	5.9375
500000		6.375	7.4062
1000000		6.6718	17.421
250000	DAILY	7.06347	6.75
500000		7.9853	8.4843
1000000		8.0146	18.015

We obtained the following notices:

1. Snowflake schema will have opportunity more than star schema when we deal with a huge amount of records number.
2. Whenever the number of records decreases, this makes us convert to use the star schema.

Conclusions:

OLAP in the current systems allow analysts to improve the decision-making process by consulting and analyzing aggregated data. Data Cube computation, which has to handle massive data, is crucial for high performance in OLAP applications. Client / server architecture grant and allow the users to share data easily and to be able to use any front-end tool. The utility of representing data in the multidimensional space is that it is more natural to view certain features of the data in this way. The efficiency of OLAP in the current work is to conduct data analysis easily and rapidly has been recognized. OLAP allows for real-time access to pre-aggregated measures along important business dimensions. Applying the mining or approximating approach supports the OLAP tools with future predicable information .

Recommendations for Future Works:

Expanding the current idea of OLAP to include real and large companies of Sales. Dealing with the statistical analysis aspect more broadly to include the concept of prediction and expert or intelligent system, which increases the efficiency of the OLAP tools used in inducing information or knowledge that leads to making suitable decisions. Using the distributed DW concepts for this work to improve the OLAP efficiency.

REFERENCES

[1]. **Alex Berson,Stephen J** ,"**Data Warehousing,Data Mining , & OLAP** " , Wiley Publishing Inc ,pages:(205,206,247,248,250,251,252), (2008).

[2]. Surajit Chaudhuri, Umeshwar Dayal, "An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD Record (ACM Special Interest Group on Management of Data), Pages: 65 – 74, Volume 26 , Issue 1 (March 1997), USA.

[3]. Erik Thomsen ,"OLAP Solutions Building Multidimensional Information Systems " , Second Edition , John Wiley & Sons, Inc.page(5), (2002)

[4]. Dimitris Papadias, Panos Kalnis, Jun Zhang and Yufei Tao . "Efficient OLAP Operations in Spatial Data Warehouses " , Springer-Verlag London, UK , Pages: 443 – 459 (2001)

[5]. Yufei Tao , Xiaokui Xiao , " Efficient Temporal Counting with Bounded Error ",Springer-Verlag New York, Inc. Secaucus, NJ, USA , (2008).

[6]. Paul Lane," Oracle Database Data Warehousing Guide ", 10g Release 2 (10.2),(2005) .

[7]. Joe Celko's ,"Analytics and OLAP in SQL", Morgan Kaufmann Publishers is an imprint of Elsevier, pages:(58,62) , (2006).

[8]. Frank K. H. A. Dehne, Todd Eavis, Boyong Liang: Compressing Data Cube in Parallel OLAP Systems. Data Science Journal 6: pages:(184-197) ,(2007).

[9]. Ann Weinberger, Matthias Ender(SAS Institute Inc., Cary, NC), "The Power of Hybrid OLAP in a Multidimensional World ",(2000).

[10]. From internet : http://www.1keydata.com/datawarehousing/molap-rolap.html , accessed on 3 August 2009.

[11]. Michelle Wilkie and Arlene Zaima," cubes by design: ROLAP and HOLAP solutions using SAS and Teradata " http://www.teradata.com/tdmo/v08n03/Tech2Tech/AppliedSolutions/CubesByDesign.aspx accessed on 28 July 2009.