Data Generalization and on-Line Analytical Processing

S . SAJIDA , M.C.A.,M.Tech
Asst.Professor, KMM ITS COLLEGE,
TIRUPATI

Prof. S . Ramakrishna, M.Sc.,M.Phil.,Ph.D.,M.Tech(IT)
Professor, Dept.ofComputerScience
Sri venkateswara University,
Tirupati.

ABSTRACT: **- A dataware house (DW or DWH) is a database used for reporting and data analysis. The data stored in the warehouse are uploaded from the operational system (such as marketing and sales etc.,).The data may pass through an operational data store for additional operations before they are used in the DW for reporting. according to W. H. Inmon the definition of Dataware house is" "A data warehouse is a subject-oriented, integrated, time variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon**

DATA WAREHOUSE and Online Analytical Processing (OLAP) are essential elements of decision support.

The paper introduces the datawarehouse ,data generalization and the online analysis process with an accent on their new requirements. For managing the dataware house I describe back end tools and data cube operations for extracting, cleaning and loading data into the data warehouse, tools for metadata management.

Keywords: Data Warehousing, OLAP, OLTP, operations.

INTRODUCTION:

Datawarehouse is a subject-Oriented

Organized around major subjects, such as customer, product, sales .Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

Data Warehouse is a Time Variant

The time horizon for the data warehouse is significantly longer than that of operational systems

Operational database: current value data

Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years) Every key structure in the data warehouse Contains an element of time, explicitly or implicitly But the key of operational data may or may not contain "time element".

Data Warehouse is a Integrated

Constructed by integrating multiple, heterogeneous data sources. Relational databases, flat files, on-line transaction records Data cleaning and data integration techniques are applied. Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources E.g., Hotel price: currency, tax, breakfast covered, etc.

When data is moved to the warehouse, then it is converted.

Data Warehouse is a Nonvolatile

A physically separate store of data transformed from the operational environment Operational update of data does not occur in the data warehouse environment Does not require transaction processing, recovery, and concurrency control mechanisms

Requires only two operations in data accessing:
initial loading of data and access of data

Traditional heterogeneous DB integration: Build wrappers/mediators on top of heterogeneous Databases Query driven approach When a query is posed to a client site, a metadictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set Complex information filtering, compete for sources. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied data bases, which facilitates effective data extraction. The functional and performance needs of OLAP are quite different from those of the on-line transaction processing(OLTP) applications traditionally supported by the operational databases. The architecture of dataware housing is shown below fig(i)

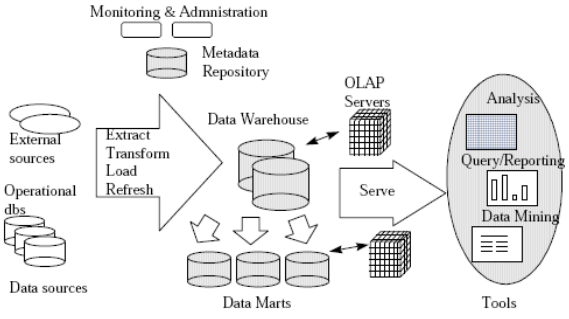


Figure i*Data warehouse systems use back-end tools and utilities to populate and refresh their data.*

DATA CLEANING:

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data Since a data warehouse is used for decision making, it is important that the data in the warehouse must be correct. Some examples where data cleaning becomes necessary are: inconsistent field length, inconsistent descriptions, inconsistent value assignments, missing entries and violation of integrity constraints.

Load

After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional preprocessing may still be required: checking integrity constraints; sorting; summarization; aggregation; and other computations to build the derived tables stored in the warehouse. In addition, load utility also allows the system administrator to monitor status, to cancel, to suspend and resume a load, and to restart after failure with no loss of data integrity.

Refresh

Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two sets of issues to consider: when to refresh and how to refresh. Usually, the warehouse is refreshed periodically. The refresh policy is set by the warehouse administrator, depending on user needs and traffic, and may be different for different sources. Refresh techniques also depends on the characteristics of the source and capabilities of the database servers. Replication servers can be used to refresh a warehouse when the sources change.

DATA GENERALIZATION:

is the process of creating successive layers of summary data in an evaluational database. It is a process of zooming out to get a broader view of a problem, trend or situation. It is also known as rolling-up data.

OLAP:

On-Line Analytical Processing (OLAP) as term was proposed by E. F. Codd, the "father of the relational database"

Relational databases put data into tables, while OLAP uses a multidimensional array representation.

The job of earlier on-line operational systems was to perform transaction and query processing. So, they are also termed as on-line transaction processing systems (OLTP). Data warehouse systems serve users or knowledge workers in the role of data analysis and decision-making.

Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are called on-line analytical processing (OLAP) systems.

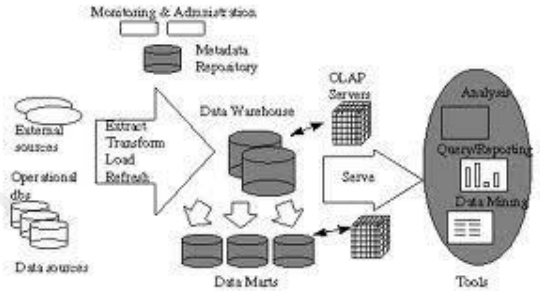
Need of data warehousing and OLAP

Data warehousing developed, despite the presence of operational databases due to following reasons:

• An operational database is designed and tuned from known tasks and workloads, such as indexing using primary keys, searching for particular records and optimizing . As data warehouse queries are often complex, they involve the computation of large groups of data at summarized levels and may require the use of special data organization, access and implementation methods based on multidimensional views. Processing OLAP queries in operational databases would substantially degrade the performance of operational tasks.

• An operational database supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanisms, such as locking and logging are required to ensure the consistency and robustness of transactions. While and OLAP query often needs read-only access of data records for summarization and aggregation. Concurrency control and recovery mechanisms, if applied for such OLAP operations, may jeopardize the execution of concurrent transactions.

• Decision support requires historical data, whereas operational databases do not typically maintain historical data. So, the data in operational databases, though abundant, is always far from complete for decision-making.



Typical OLAP architecture and process of design:

OLAP SERVER ARCHITECTURES:

Relational OLAP(ROALP)These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces. ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services. ROLAP technology tends to have greater scalability than MOLAP technology. The Microstrategy's DSS server and Informix's Metacube, for example, adopt the ROLAP approach.

Multidimensional OLAP (MOLAP) servers: These servers support multidimensional views of data through array-based multidimensional storage engines. They map

multidimensional views directly to data cube array structures. For example, Essbase from Hyperion is a MOLAP server. The advantage of using a data cube is that it allows fast indexing to pre computed summarized data. Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse. In such cases, sparse matrix compression techniques should be explored.Many MOLAP servers adopt a two-level storage representation to handle sparse and dense data sets: the dense subcubes are identified and stored as array structures, while the sparse subcubes employ compression technology for efficient storage utilization.

- Thereare other architectures Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer) and Specialized SQL servers (e.g., Redbricks)

OLAP Operations: Data Cube

The key operation of OLAP is the formation of a data cube.

A data cube is a multidimensional representation of data, together with all possible aggregates. By all possible aggregates, the meaning of aggregates that result by selecting a proper subset of the dimensions and summing over all remaining dimensions.

Example of Data Cube: Consider a data set that records the sales of products at a number of company stores (e.g. IBM, Real etc.) at various dates. This data can be represented as a 3 dimensional array

OLAP OPERATIONS:

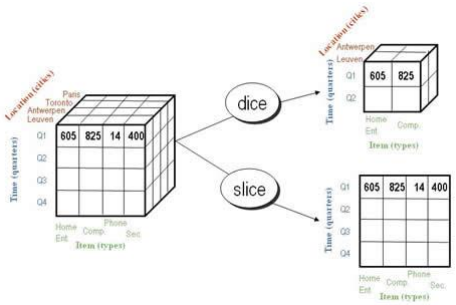
Slicing and Dicing

Slicing is selecting a group of cells from the entire Multidimensional array by specifying a specific Value for one or more dimensions.

Dicing involves selecting a subset of cells by Specifying a range of attribute values.

– This is equivalent to defining a sub array from the Complete array.

In practice, both operations can also be accompanied by aggregation over some dimensions.



ROLL-UP AND DRILL-DOWN:

. – Attribute values often have a hierarchical structure.

– Each date is associated with a year, month, and week.

– A location is associated with a continent, country, state (province, etc.), and city.

– Products can be divided into various categories, such as clothing, electronics, and furniture.

This hierarchical structure gives rise to the roll-up

Roll up: for example, takes the current aggregation level of fact values and a further aggregation on one or more than dimensions. Which is equivalent to doing GROUP BY clause to this. Decreases a number of dimensions remove row headers

SELECT[attribute list], SUM[attribute names]

FROM[table list]

WHERE[condition list]

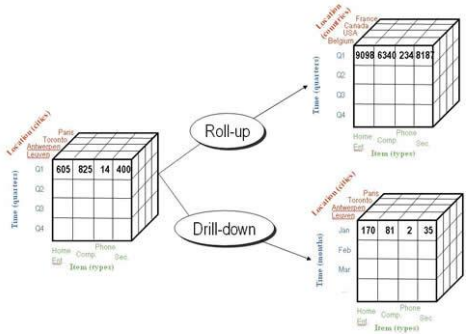
GROUP BY [grouping list];

DRILL DOWN:

#opposite of roll up

#summarizes data at a lower level of a dimension hierarchy, there by view in data in a more specialized level within a dimension.

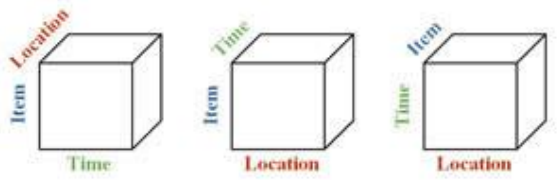
#increases a number of dimensions adds new headers



Pivot (or Rotate)

- This rotates the data axis to view the data from different persepectives.
- Groups data with different dimensions

- Below figure shows the example.



*Drill-across***:**

- Accesses more than one fact table that is linked by common dimensions
- Combines cubes that share one or more dimensions

Drill-through:

- Drill down to the bottom level of a data cube down to its back-end relational tables
- Cross –tab:
- Spreadsheet style row/column aggregates.

CONCLUSION:

Data generalization: Attribute-oriented induction

Data warehousing: A multi-dimensional model of a data warehouse. A data cube consists of dimensions & measures OLAP operations: drilling, rolling, slicing, dicing and pivoting, Data warehouse architecture OLAP servers: ROLAP, MOLAP, HOLAP Efficient computation of data cubes. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Data warehousing and online analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. OLTP is customer-oriented and is used for transaction and query processing by clerks, clients and information technology professionals. The job of earlier on-line operational systems was to perform transaction and query processing. Partial vs. full vs. no materialization Indexing OALP data: Bitmap index and join index

REFERENCES:

[1]C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003

[2]W. H. Inmon. Building the Data Warehouse. John Wiley, 1996

[3]R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002

[4]P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97

[5] Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998

[6] A. Shoshani. OLAP and statistical databases: Similarities and differences. PODS'00.

[7]S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94

[8]OLAP council. MDAPI specification version 2.0. In http://www.olapcouncil.org/research/apily.htm, 1998

[9]E. Thomsen. OLAP Solutions: Building Multidimensional Information Systems. John Wiley, 1997

[10]P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.

[11]J. Widom. Research problems in data warehousing. CIKM'95.

[12]Devlin, B. & Murphy, P. (1988) An Architecture for a Business and Information System, IBM Systems Journal, 27 (1), 60-80.