Chapter 5

Data Warehouse to Data Mining

Data Warehouse Architecture

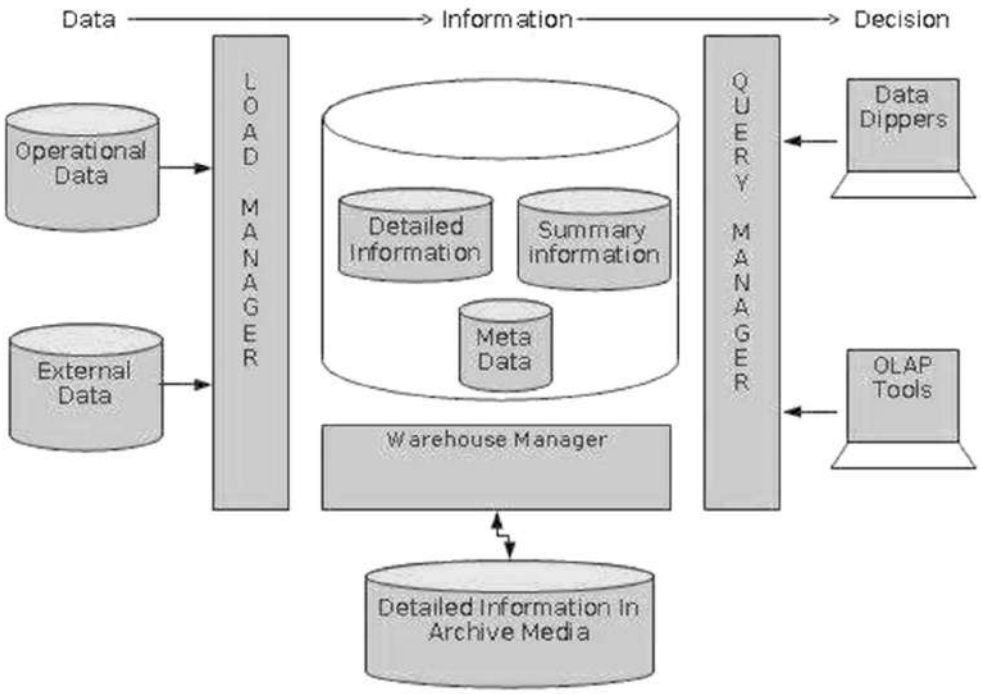


Figure: "Architecture of a Data Warehouse"

Load Manager: The system components that perform all the operations necessary to support the extract and load process. It fast loads the extracted data into a temporary data store and performs simple transformations into a structure similar to the one in the data warehouse.

Warehouse Manager: Performs all the necessary operations to support the warehouse management process. It analyzes the data to perform consistency and referential checks. It also transforms and merges the source data in the temporary data store into the published data warehouse with creating indexes and business views. Update all existing aggregations and back up data in the data warehouse.

Query Manager: Performs all the operations necessary to support the query management process by directing queries to the appropriate tables. In some cases, it also stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Detailed Information: Stores all the detailed information to determine the business requirements to analyze the level at which to retain detailed information in the data warehouse.

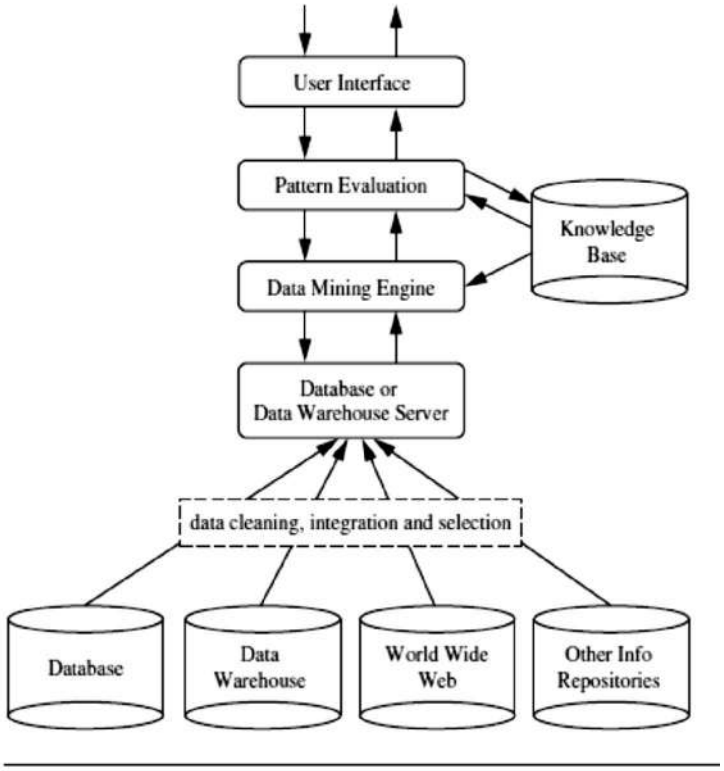
***Summary Informatio*n:** Stores all the predefined aggregations generated by the warehouse manager. It is a transient area which will change on an ongoing basis in order to respond to changing query profiles. It is essentially a replication to detailed information. The purpose of summary information is to speed up the performance of query.

Meta Data: Meta data is data about data which describes how information is structured within a data warehouse. It maps data stores to common view of information with the data warehouse.

Data Mart

- Data Mart is a subset of the information content of a data warehouse that is stored in its own database.
- Data mart may or may not be sourced from an enterprise data warehouse i.e. it could have been directly populated from source data.
- Data mart can improve query performance simply by reducing the volume of data that needs to be scanned to satisfy the query.
- Data marts are created along functional level to reduce the likelihood of queries requiring data outside the mart.
- Data marts may help in multiple queries or tools to access data by creating their own internal database structures.
- E.g.: Departmental Store, Banking System.

Data mining Architecture:



Architecture of a typical data mining system.

The architecture of a typical data mining system may have the following major components:

1. **Database, data warehouse, World Wide Web, or other information repository:** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
2. **Database or data warehouse server:** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
3. **Knowledge base:** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

4. **Data mining engine:** This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.
5. **Pattern evaluation module:** This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.
6. **User interface:** This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

OLAP (Online Analytical Processing) Architecture

The term OLAP or online analytical processing was introduced in a paper entitled "Providing Analytical processing to User Analysts," by Dr. E. R Codd, the acknowledged "father" of the relational database model. A particular definition for OLAP is as follows:
Online Analytical Processing (OLAP) is a category of software technology that enables analysts, managers 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.
This definition from the OLAP council contains all the key ingredients. Speed, interactive access, and multiple dimensional views-all of these are principal elements. An even simpler definition is that OLAP is a Fast Analysis of Shared Multidimensional Information. This definition (sometimes called FASMI) implies that most OLAP queries should be answered within seconds. Furthermore, it is expected that most OLAP queries can be answered without any programming.
OLAP stands for On Line Analytical Processing. It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. It is part of the broader category business intelligence, which also includes Extract transform load (ETL), relational reporting and data mining. The typical applications of OLAP are in business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas. The term OLAP was created as a slight modification of the traditional database term OLTP (On Line Transaction Processing).
Databases configured for OLAP employ a multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time. The output of an OLAP query is typically displayed in a matrix (or pivot) format. The dimensions form the row and column of the matrix; the measures, the values. OLAP tools are used to provide answer to analytical queries. OLAP has been growing in popularity due to the increase in data volumes and the recognition of the business value of analytics. Until the mid-nineties, performing OLAP analysis was an extremely costly process mainly restricted to larger organizations.
OLAP provides you with a very good view of *what is happening*, but cannot predict *what will happen in the future* or *why it is happening*. **OLAP** is a term used to describe the analysis of complex data from the data warehouse. **OLAP** is an advanced data analysis environment that supports decision making, business modeling, and operations research activities. Can easily answer 'who?' and 'what?' questions, however, ability to answer 'what if?' and 'why?' type questions distinguish OLAP from general-purpose query tools. OLAP Enables users to gain a deeper understanding and knowledge about various aspects of their corporate data through fast, consistent, interactive access to a wide variety of possible views of the data. OLAP allows users to view corporate data in such a way

that it is a better model of the true dimensionality of the enterprise. OLAP is a category of applications/technology for collecting, managing, processing, and presenting multidimensional data for analysis and management purposes.

OLAP is **FASMI**

- Fast
- Analysis
- Shared
- Multidimensional
- Information

The main characteristics of OLAP are as follows

1. Multidimensional conceptual view: OLAP systems let business Users have a multidimensional and logical view of the data in the data warehouse. It helps in carrying slice and dice operations.
2. Multi user support: Since the OLAP system is shared, the OLAP software should provide many normal database operations including retrieval, update, concurrency control, integrity and security.
3. Accessibility: OLAP acts as a mediator between data warehouse and front end. The OLAP software should be sitting between data sources (e. g., data warehouse) and an OLAP front-end.
4. Storing: OLAP results OLAP results are kept separate from data sources.
5. Uniform reporting performance: Increasing the number of dimensions or database size should not significantly degrade the reporting performance of the OLAP system. This is a good objective although it may be difficult to achieve in practice.
6. OLAP provides for distinguishing between zero values and missing values so that aggregates
7. are computed correctly.
8. OLAP system should ignore all missing values and compute correct aggregate values.
9. OLAP facilitate interactive query and complex analysis for the users,
10. OLAP allows users to drill down for greater details or roll up for aggregations of metrics
11. along a single business dimension or across multiple dimensions.
12. OLAP provides ability to perform intricate calculations and comparisons.
13. OLAP presents results in a number of meaningful ways, including charts and graphs.

Benefits:

- Increased productivity of end-users.
- Retention of organizational control over the integrity of corporate data.
- Reduced query drag and network traffic on OLTP systems or on the data warehouse.
- Improved potential revenue and profitability.

Strengths of OLAP

- It is a powerful visualization paradigm
- It provides fast, interactive response times
- It is good for analyzing time series
- It can be useful to find some clusters and outliers
- Many vendors offer OLAP tools

OLAP for Decision Support

- Goal of OLAP is to support ad-hoc querying for the business analyst
- Business analysts are familiar with spreadsheets
- Extend spreadsheet analysis model to work with warehouse data
  - Large data set
  - Semantically enriched to understand business terms (e.g., time, geography)
  - Combined with reporting features
- Multidimensional view of data is the foundation of OLAP

Examples of OLAP Applications in Various Functional Areas

Functional area	Examples of OLAP applications
Finance	Budgeting, activity-based costing, financial performance analysis, and financial modeling
Sales	Sales analysis and sales forecasting
Marketing	Market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation
Manufacturing	Production planning and defect analysis

Comparing OLAP and Data Mining

Characteristic	OLAP	Data Mining
Purpose	Supports data analysis and decision making	Supports data analysis and decision making
Type of analysis supported	Top-down, query-driven data analysis	Bottom-up, discovery-driven data analysis
Skills required of user	Must be very knowledgeable of the data and its business context	Must trust in data-mining tools to uncover valid and worthwhile hypotheses

OLTP (Online Transaction Processing)

OLTP is characterized by a large number of short online transactions (INSERT, UPDATE and DELETE). OLTP is used to carry out day to day business functions such as ERP (Enterprise Resource Planning), CRM (Customer Relationship Planning). OLTP system solved a critical business problem of automating daily business functions and running real time report and analysis. Systems is to put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second. The main emphasis for OLTP In OLTP database, there is detailed and current data, and schema used to store transactional databases in the entity model (usually 3NF).

OLAP Vs. OLTP

Facts	OLTP	OLAP
Source of Data	Operational Data	Data warehouse (From various database)
Purpose of data	Control and run fundamental business tasks	For planning, problem solving and decision support
Queries	Simple queries	Complex queries and algorithms
Processing Speed	Typically, very fast	Depends on the amount of data involved, data size, techniques and algorithms
Space requirements	Can be relatively small	Larger due to aggregated databases
Nature of users	Operations workers	Decision makers
Functions	Mission-critical	Management-critical
Nature of usage	Mostly repetitive	Mostly ad hoc
Nature of design	Application oriented	Subject oriented
Number of users	Thousands	Dozens

Nature of data	Current, detailed, relational	Historical, summarized, multidimensional
Updates	All the time	Usually not allowed
Database design	Highly normalized with many tables	De-normalized with fewer tables. Use of Star and/or snowflake schemes

OLAP OPERATIONS

In the multidimensional model, data are organized into multiple dimensions and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operation exist to show these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis.

- **Roll-up:** Roll-up is like zooming out on the data cube. It is required when the user needs further abstraction or less detail. This operation performs further aggregations on the data, for example, from single degree programs to all programs offered by a School or department, from single countries to a collection of countries, and from individual semesters to academic years. So, the roll-up operation (also called the ***"drill-up"*** operation by some vendors) performs aggregation on a data cube, either by climbing-up a concept hierarchy for a dimension or by dimension reduction.

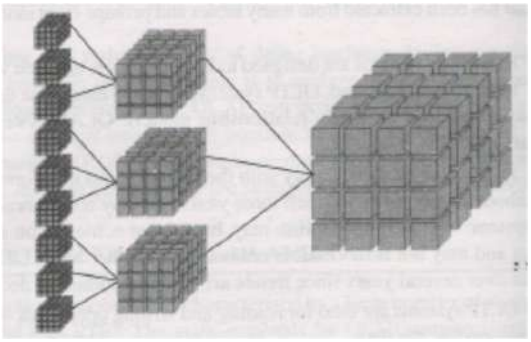


Fig. Roll-up

- **Drill-down:** Drill-down is like *zooming in* on the data and is therefore the reverse of roll-up. It is an appropriate operation when the user needs further details or when the user needs further details or when the user wants to partition more finely or wants to focus on some particular values of certain dimensions. Drill-down adds more details to the data. That is, Drill-down can be realized by either *stepping-down* a concept hierarchy for a dimension or introducing additional dimensions.
- **Slice and dice:** Slice and dice are operations for browsing the data in the cube. The terms refer to the ability to look at information from different viewpoints. A ***slice*** is a subset of the cube corresponding to a single value for one or more members of the dimensions. For example, a slice operation is performed when the user wants a selection on one dimension of a three-dimensional cube resulting in a two dimensional site. So, the slice operation performs a selection on one dimension of the given cube, resulting in a sub-cube.

The ***dice operation*** is similar to slice but dicing does not involve reducing the number of dimensions. A dice is obtained by performing a selection on two or more dimensions.

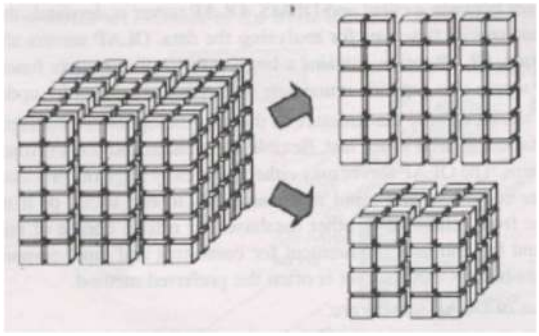


Fig Slicing (above) and dicing (below) a cube

- **Pivot or Rotate:** The pivot operation is used when the user wishes to re-orient the view of the data cube. It may involve swapping the rows and columns, or moving one of the row dimensions into the column dimension.

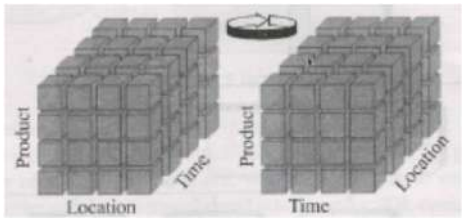


Fig Rotate

- **Other OLAP operations:** Some OLAP systems offer additional drilling operations. For example, drill-across executes queries involving (i.e., across) more than one fact table. The drill-through operation makes use of relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables.

Other OLAP operations may include ranking the top-N or bottom-N items in lists, as well as computing moving averages, growth rates, and interests, internal rates of return, depreciation, currency conversions, and statistical functions.

OLAP offers analytical modeling capabilities, including a calculation engine for deriving ratios, variance, etc., and for computing measures across multiple dimensions. It can generate summarizations, aggregations, and hierarchies at each granularity level and at every dimension intersection. OLAP also supports functional models for forecasting, trend analysis, and statistical analysis. In this context, an OLAP engine is a powerful data analysis tool.

OIAP SERVERS

OLAP server is a high-capacity, multi-user data manipulation engine specifically designed to support and operate on multi-dimensional data structures. A multi-dimensional structure is arranged so that every data item is located and accessed based on the intersection of the dimension members which define that item.

OLAP server sits between a client and DBMS. OLAP server understands the organization of data in the database and has special functions for analyzing the data. OLAP servers should support the data types like images, sound. OLAP server contains a large number of aggregate functions that are used for computations. OLAP server also supports immediate data refresh facilities for updating data load. The design of the server and the structure of the data are optimized for rapid ad-hoc information retrieval in any orientation, as well as for fast, flexible calculation and transformation of raw data based on standard relationships. The OLAP server may either physically stage the processed multi-dimensional information to deliver consistent and rapid response times to end users, or it may populate its data structures in real-time from relational or other databases, or offer a choice of both. Given the current state of technology and the end user requirement for consistent and rapid response times, staging the multi-dimensional data in the OLAP server is often the preferred method.

Main variations of OLAP servers are:

- ROLAP
- MOLAP

- HOLAP
- DOLAP

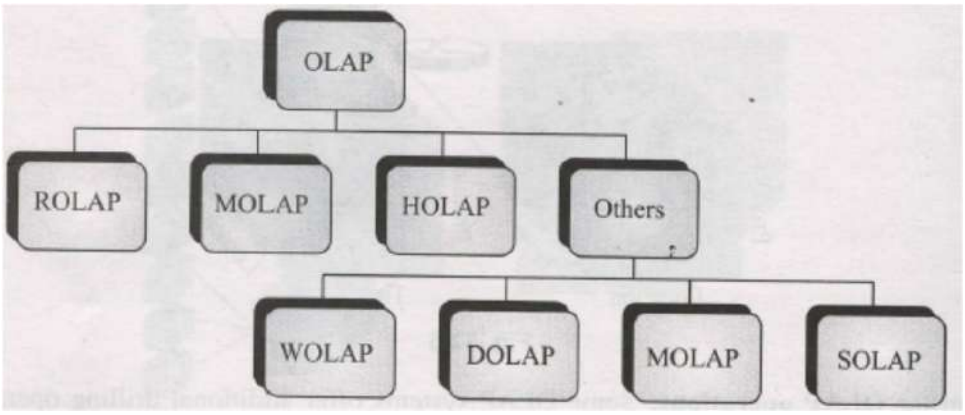


Fig.

All these variations relate to the way data for OLAP. The processing is still online analytical processing; only the storage methodology is different.

ROLAP

ROLAP stands for ***relational online analytical processing.*** ROLAP uses a relational DBMS to implement an OLAP environment. It may be considered a bottom-up approach which is typically based on using data warehouse that has been designed using a star schema. The data therefore is likely to be in a de-normalized structure. A normalized database avoids redundancy but is usually not appropriate for high performance. The summary data will be held in aggregate tables.

The data warehouse provides the multidimensional capabilities by representing data in fact tables(s) and dimension tables. The fact table contains one column for each dimension and one column for each measure and every row of the table provides one fact. An OLAP tool is then provided to manipulate the data in these data warehouse tables. This tool essentially groups the fact table to find aggregates and uses some of the aggregates already computed to find new aggregates.

Figure below shows the architecture of the ROLAP.

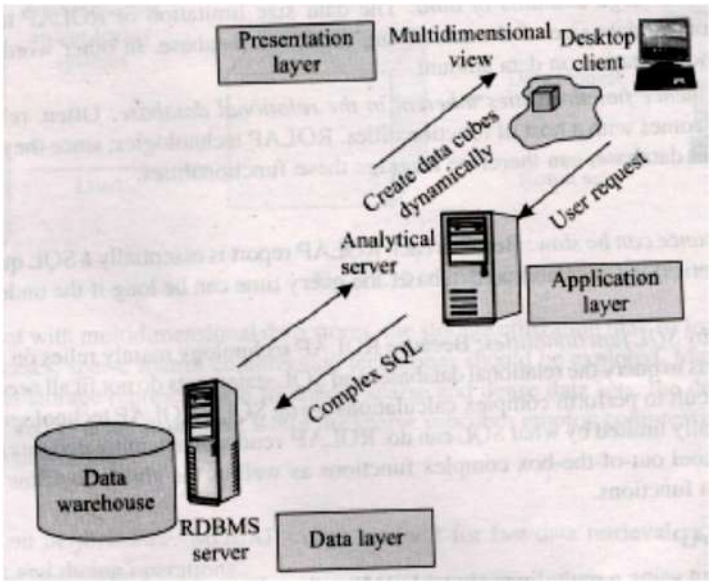


Fig. ROLAP model

This is three-tier architecture. The analytical server in the middle tier application layer creates multi-dimensional views on the fly. The multi-dimensional system at the presentation layer provides a multidimensional view of the data to the users. When the users issue complex queries based on this

multidimensional view, the queries are transformed into complex SQL directed to the relational database.

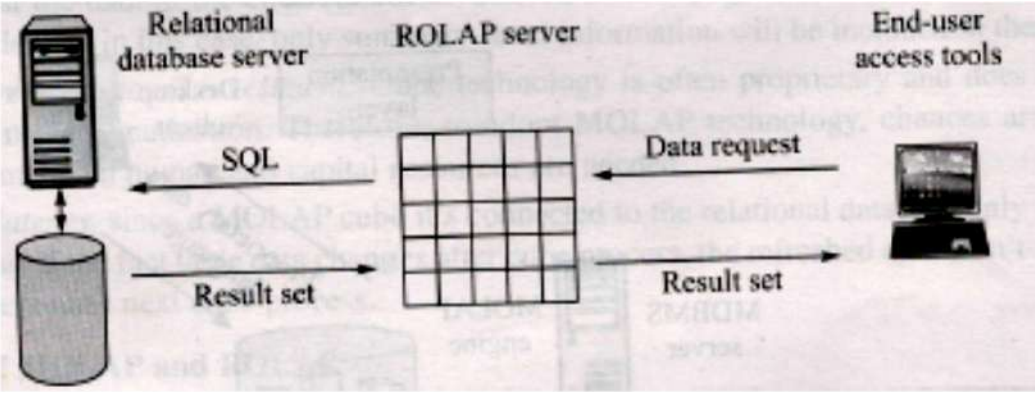


Fig.

True ROLAP has three distinct characteristics:

- Supports all the basic OLAP features and functions discussed earlier
- Stores data in a relational form
- Supports some form of aggregation

The DSS server of Microstrategy and Metacube of Informix, for example, adopt the ROLAP approach.

Advantages

- *Can handle large amounts of data:* The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on data amount.
- *Can influence functionalities inherent in the relational database:* Often, relational database already comes with a host of functionalities. ROLAP technologies, since they sit on top of the relational database, can therefore leverage these functionalities.

Disadvantages

- *Performance can be slow:* Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large.
- *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 (for example, it is difficult to perform complex calculations using SQL), ROLAP technologies are therefore traditionally limited by what SQL CAN DO. ROLAP vendors have mitigated this risk by building into the tool out-of-the-box complex functions as well as the ability to allow users to define their own functions.

MOLAP

MOLAP is based on using a multidimensional DBMS rather than a data warehouse to store and access data. It may be considered as a top-down approach to OLAP. The multidimensional database systems do not have a standard approach to storing and maintaining their data. They file often use special-purpose file systems or indexes that store pre-computation of all aggregations in the cube. .

Now consider the architecture for the MOLAP. Pre-calculated and prefabricated multidimensional data cubes are stored in multidimensional databases. The MOLAP engine in the application layer pushes a multidimensional view of the data from the MDDBs to the users.

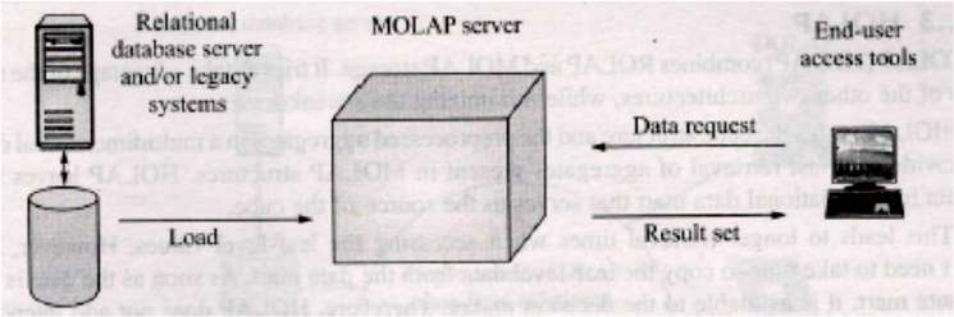


Fig.

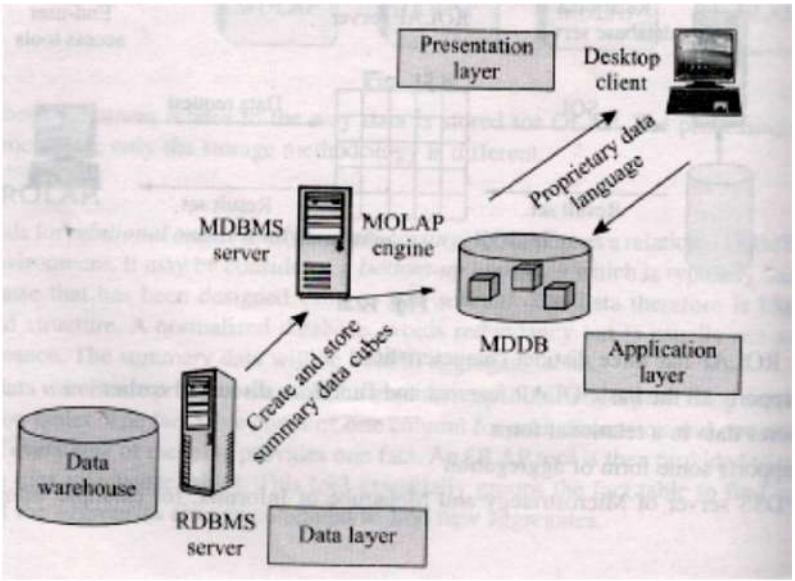


Fig. The MOLAP model

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 OLAP servers adopt a two-level storage representation to handle sparse and dense data sets: the dense sub-cubes are identified and stored as array structures, while the sparse sub-cubes employ compression technology for efficient storage utilization.

Advantages

- *Excellent performance:* MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
- *Can perform complex calculations:* All calculations have been pre-generated when the cube is created. Hence, complex calculations are not only doable, but they return quickly.

Disadvantages

- *Limited in the amount of data it can handle:* Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself. This is not to say that the data in the cube cannot be derived from a large amount of data. Indeed, this is possible. But in this case, only summary-level information will be included in the cube itself.
- *Requires additional investment:* Cube technology is often proprietary and does not already exist in the organization. Therefore, to adopt MOLAP technology, chances are additional investments in human and capital resources are needed.
- *High latency*, since a MOLAP cube it's connected to the relational database only during cube process. If the fact table data changes after cube process, the refreshed data won't be visible to the users until next cube Process.

Comparison of MOLAP and ROLAP

Property	MOLAP	ROIAP
Data structure	Multidimensional database using sparse arrays	Relational tables (each cell is a row)
Disk space	Separate database for data cube; large for large data cubes	May not require any space other than that available in the data warehouse
Retrieval	Fast (pre-computed)	Slow
Scalability	Limited	Excellent
Best suited for	Inexperienced users, limited set of queries	Experienced users, queries change frequently
DBMS facilities	Usually weak	Usually very strong

HOLAP

Hybrid OLAP (HOLAP) combines ROLAP and MOLAP storage. It tries to take advantage of the strengths of each of the other two architectures, while minimizing their weaknesses.

HOLAP stores the cube structure and the preprocessed aggregates in a multidimensional database. This provides the fast retrieval of aggregates present in MOLAP structures. HOLAP leaves the leaf-level data in the relational data mart that serves as the source of the cube.

This leads to longer retrieval times when accessing the leaf-level values. However, HOLAP does not need to take time to copy the leaf-level data from the data mart. As soon as the data updated in the data mart, it is available to the decision maker. Therefore, HOLAP does not add latency to the leaf-level data. In essence, HOLAP sacrifices retrieval speed on leaf-level data to prevent adding latency to leaf-level data and to speed the data load.

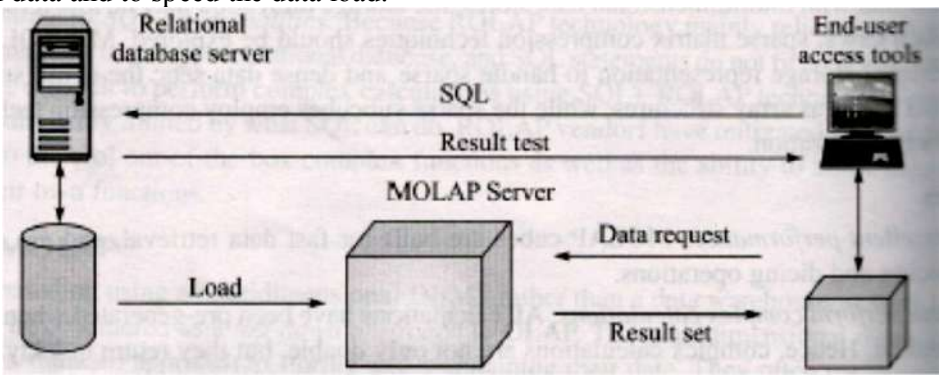


Fig. Typical Architecture for HOLAP Tools

Advantages

- HOLAP is best used when large amounts of aggregations are queried often with little detail data, offering high performance and lower storage requirements.
- Cubes are smaller than MOLAP since the detail data is kept in the relational database.
- Processing time is less than MOLAP since only aggregations are stored in multidimensional format.
- Low latency since processing takes place when changes occur and detail data is kept in the relational database.

Disadvantages

- As slow as ROLAP when you try to access leaf level data.
- Need to process when new records inserted.

DOLAP

Desktop OLAP is also called as **client-side OLAP**. Examples of this category are products that send data from a server to a client desktop computer. The data processing is mainly client-centric. It is used for end-user reporting and analysis. Its most distinguishing feature is its ability to download a relatively small hypercube from a central point (usually a data mart or data warehouse) and perform

multidimensional analyses while disconnected from the source. This functionality is particularly useful for mobile users who can't always connect to the data warehouse.

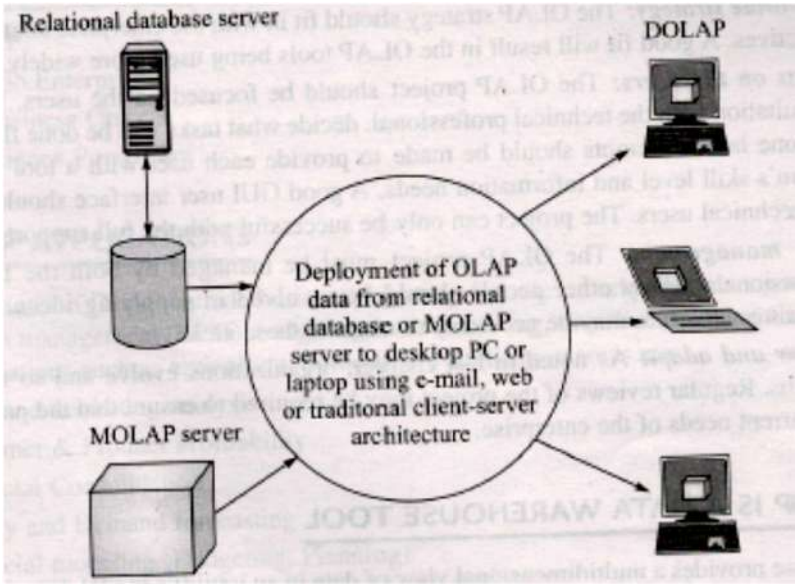


Fig. Typical Arc hitecture for DOLAP Tools

- Store the OLAP data in client-based files and support multi-dimensional processing using a client multi-dimensional engine.
- Requires that relatively small extracts of data are held on client machines. They may be distributed in advance, or created on demand (possibly through the Web).
- As with multi-dimensional databases on the server, OLAP data may be held on disk or in RAM, however, some DOLAP products allow only read access.
- Most vendors of DOLAP exploit the power of desktop PC to perform some, if not most, multidimensional calculations.
- The administration of a DOLAP database is typically performed by a central server or processing routine that prepares data cubes or sets of data for each user.
- Once the basic processing is done, each user can then access their portion of the data.

GUIDELINES FOR OLAP IMPLEMENTATIO N

Following are a number of guidelines for successful implementation of OLAP. The guidelines are somewhat similar to those presented for data warehouse implementation.

- ***Vision:*** The OLAP team must, in consultation with the users, develop a clear vision for the OLAP system. This vision including the business objectives should be clearly defined, understood, and shared by the stakeholders.
- ***Senior management support:*** The OLAP project should be fully supported by the senior managers. Since a data warehouse may have been developed already, this should not be difficult.
- ***Selecting an OLAP tool:*** The OLAP team should familiarize themselves with the ROLAP and MOLAP tools available in the market. Since tools are quite different, careful planning may be required in selecting a tool that is appropriate for the enterprise. In some situations, a combination of ROLAP and MOLAP may be most effective.
- ***Corporate strategy:*** The OLAP strategy should fit in with the enterprise strategy and business objectives. A good fit will result in the OLAP tools being used more widely.
- ***Focus on the users:*** The OLAP project should be focused on the users. Users should, in consultation with the technical professional, decide what tasks will be done first and what will be done later. Attempts should be made to provide each user with a tool suitable for that

person's skill level and information needs. A good GUI user interface should be provided to non-technical users. The project can only be successful with the full support of the users.

- ***Joint management:*** The OLAP project must be managed by both the IT and business professionals. Many other people should be involved in supplying ideas. An appropriate committee structure may be necessary to channel these ideas.
- ***Review and adapt:*** As noted in last chapter, organizations evolve and so must the OLAP systems. Regular reviews of the project may be required to ensure that the project is the current needs of the enterprise.

OLAP IS A DATA WAREHOUSE TOOL

A data warehouse provides a multidimensional view of data in an intuitive model designed to match the types of queries posed by analysts and decision makers. OLAP organizes data warehouse data into multidimensional cubes based on this dimensional model, and then preprocesses these cubes to provide maximum performance for queries that summarize data in various ways. For example, a query that requests the total sales income and quantity sold for a range of products in a specific geographical region for a specific time period can typically be answered in a few seconds or less regardless of how many hundreds of millions of rows of data are stored in the data warehouse database.

OLAP is not designed to store large volumes of text or binary data, nor is it designed to support high volume update transactions. The inherent stability and consistency of historical data in a data warehouse enables OLAP to provide its remarkable performance in rapidly summarizing information for analytical queries.

OLAP PRODUCTS

The list of some main OLAP products is as follows:

- **MOLAP**
  - SAS CFO Vision
  - Comshare Decision
  - Hyperion Essbase
  - Power Play Enterprise Server
- **ROLAP**
  - Cartesis Carat
  - Micro Strategy
- **HOLAP**
  - Oracle Express
  - Seagate Holos
  - Speedware Media/M
  - Microsoft OLAP Services
- **DOLAP**
  - Brio.Enterprise
  - Business Objects
  - Cognos Powerplay

OLAP APPLICATIONS

Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas, with new applications such as agriculture. Some of the application areas are:

- Sales and marketing analysis
- Customer & product profitability
- Financial Consolidation
- Supply and Demand forecasting
- Financial modeling (Budgeting, planning)

- Loss ratio analysis
- Claims experience analysis
- Human resources analysis
- Market share analysis
- Resource allocation and capacity planning
- Variance analvsis