

Unit 5 : Data Warehouse to Data Mining

Lecturer : Bijay Mishra

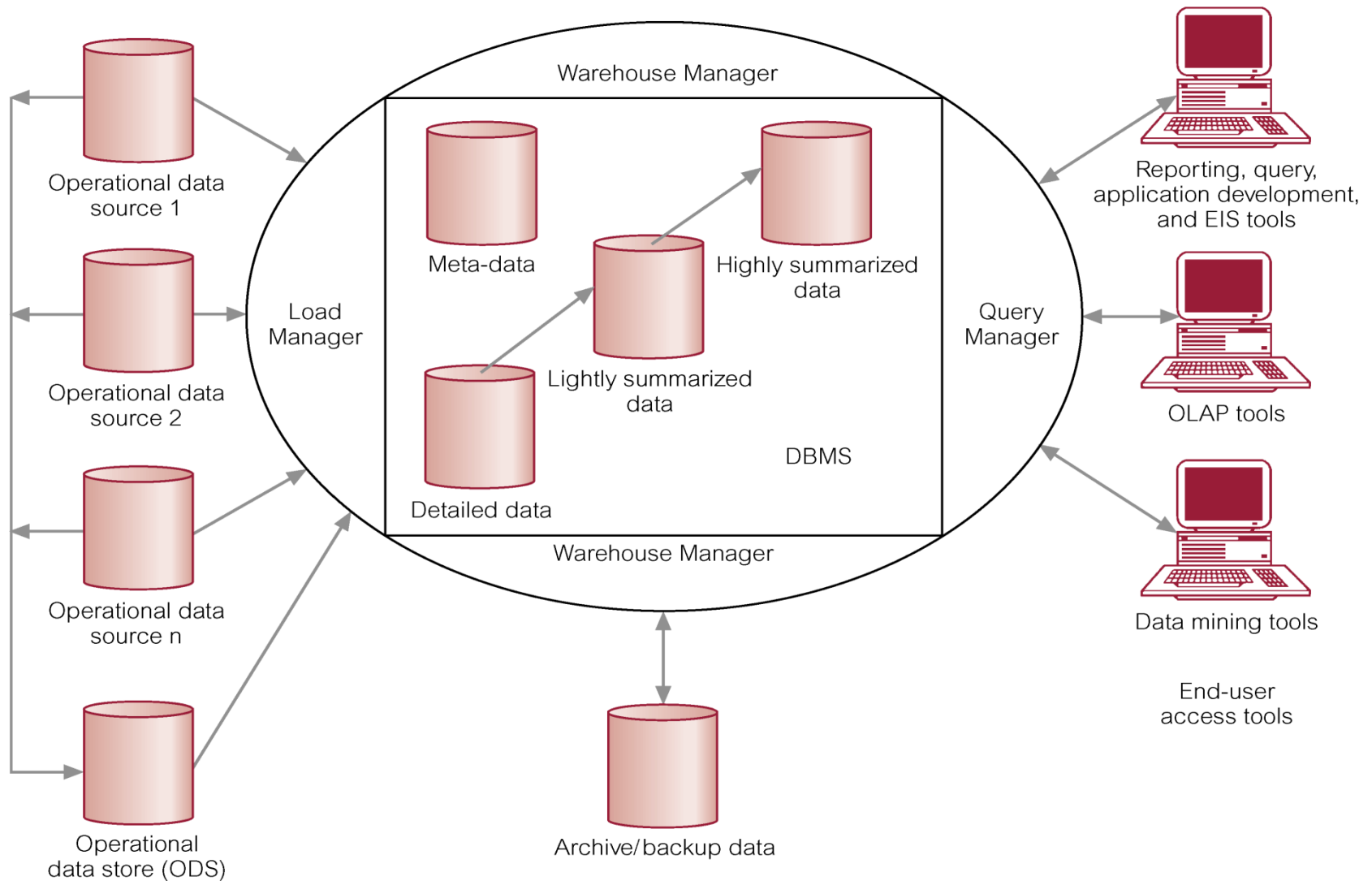
OLAP – Online Analytical Processing

A definition from **COMPUTERWORLD** An IDG company

NOVEMBER 30, 1998 - OLAP

Online analytical processing (OLAP) describes a class of tools that can extract and present multidimensional data from different points of view. Designed for managers looking to make sense of their information, OLAP structures data hierarchically -- the way managers think of their enterprises. OLAP functions include trend analysis, drilling down to more complex levels of detail, summarization of data and data rotation for comparative viewing.

Data Warehouse Architecture



- **OLAP** provides you with a very good view of *what is happening*, but can not 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 distinguishes OLAP from general-purpose query tools.
- 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.
- 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.

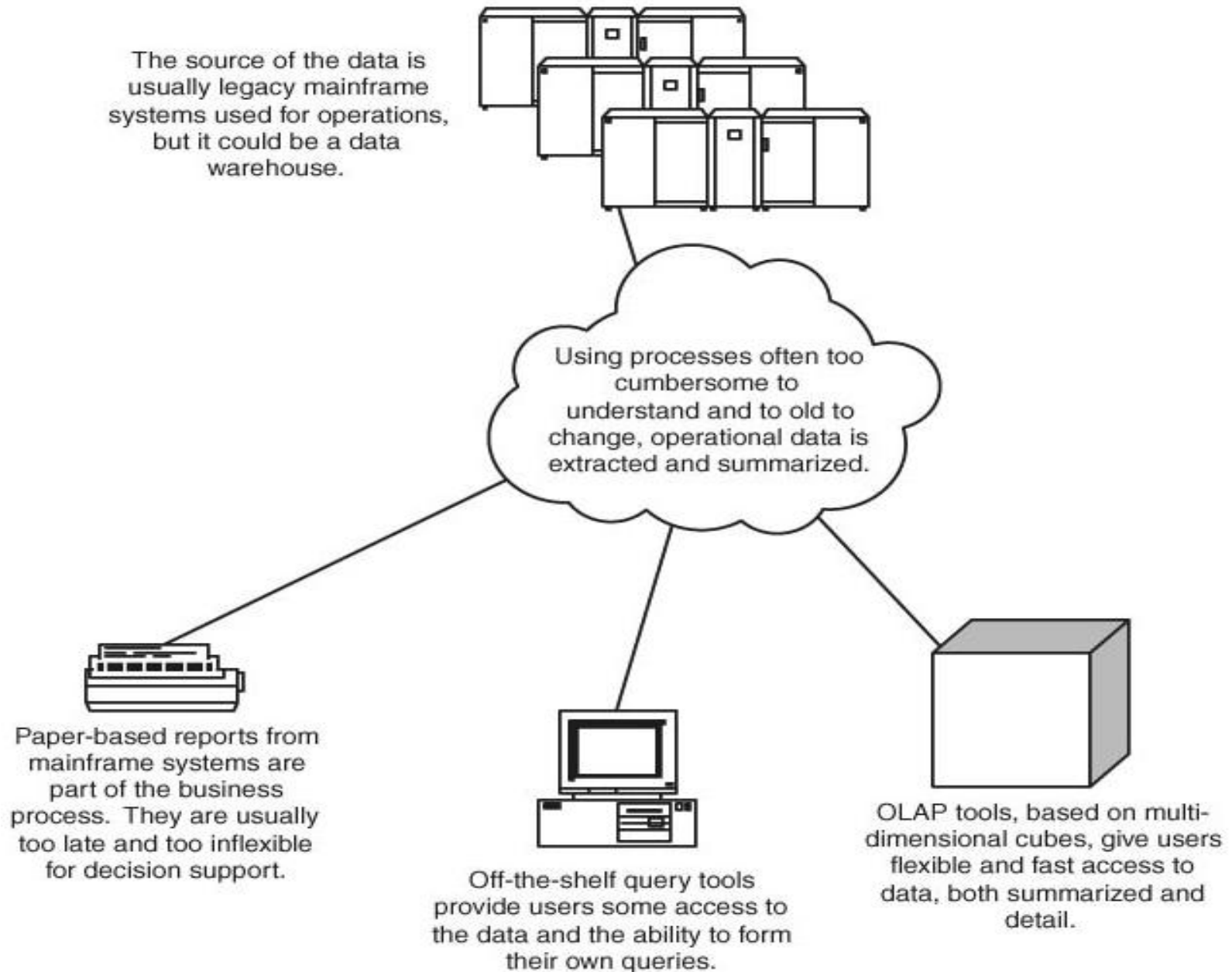
OLAP is **FASMI**

- Fast
- Analysis
- Shared
- Multidimensional
- Information

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

Where does OLAP fit in?



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

OLAP 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.

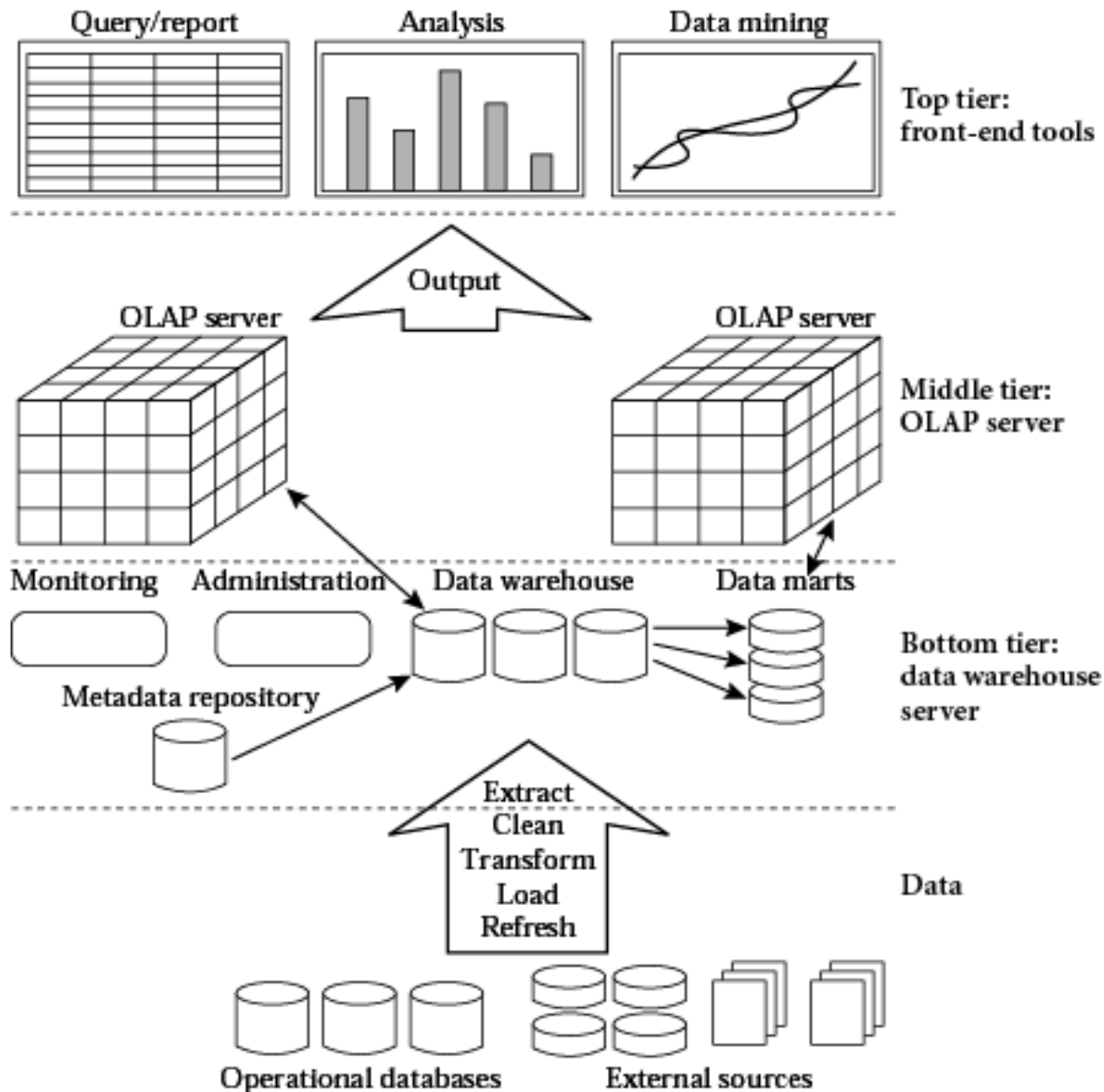
Strengths of OLAP

- 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

OLAP Architecture



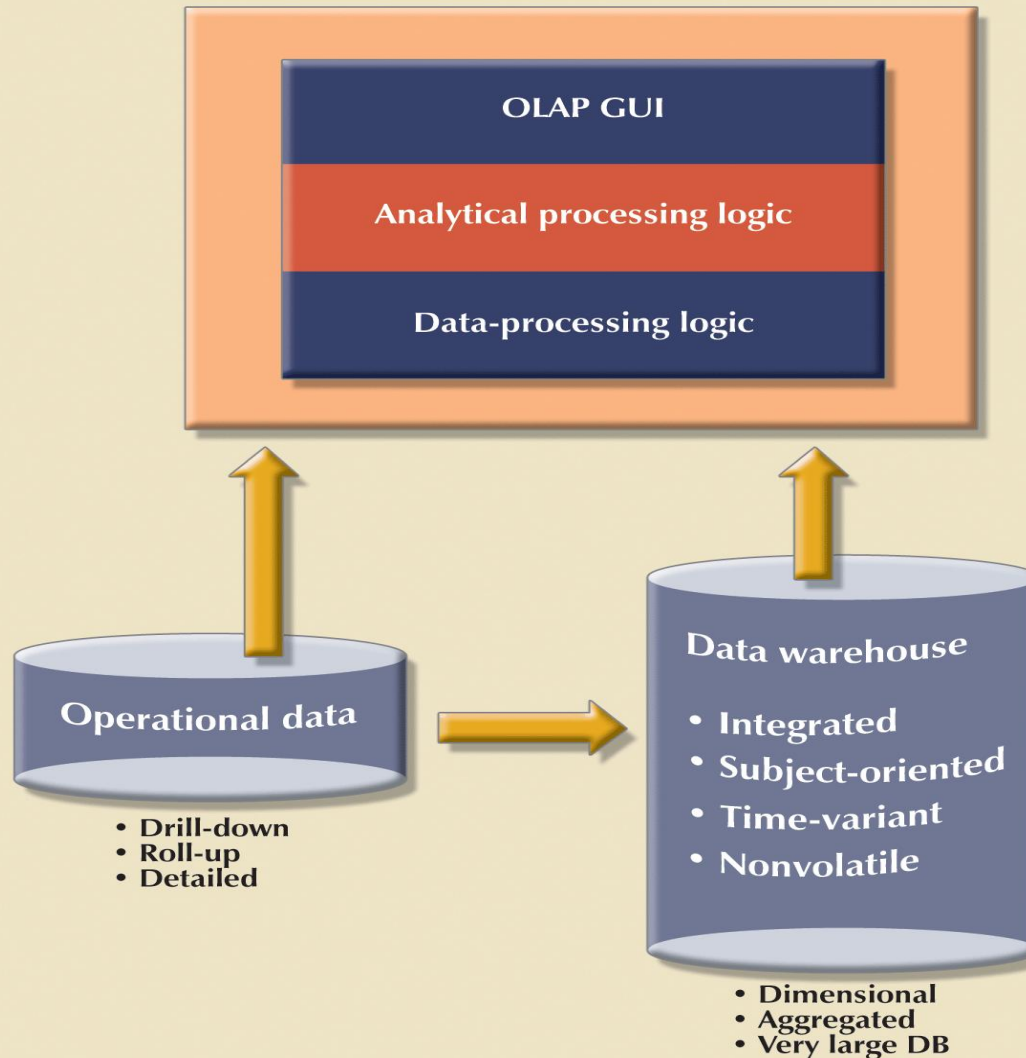
OLAP Architecture (continued)

- Designed to use both operational and data warehouse data
- Defined as an “advanced data analysis environment that supports decision making, business modeling, and an operation’s research activities”
- In most implementations, data warehouse and OLAP are interrelated and complementary environments

OLAP Client/Server Architecture

FIGURE
13.6

OLAP client/server architecture



OLAP System

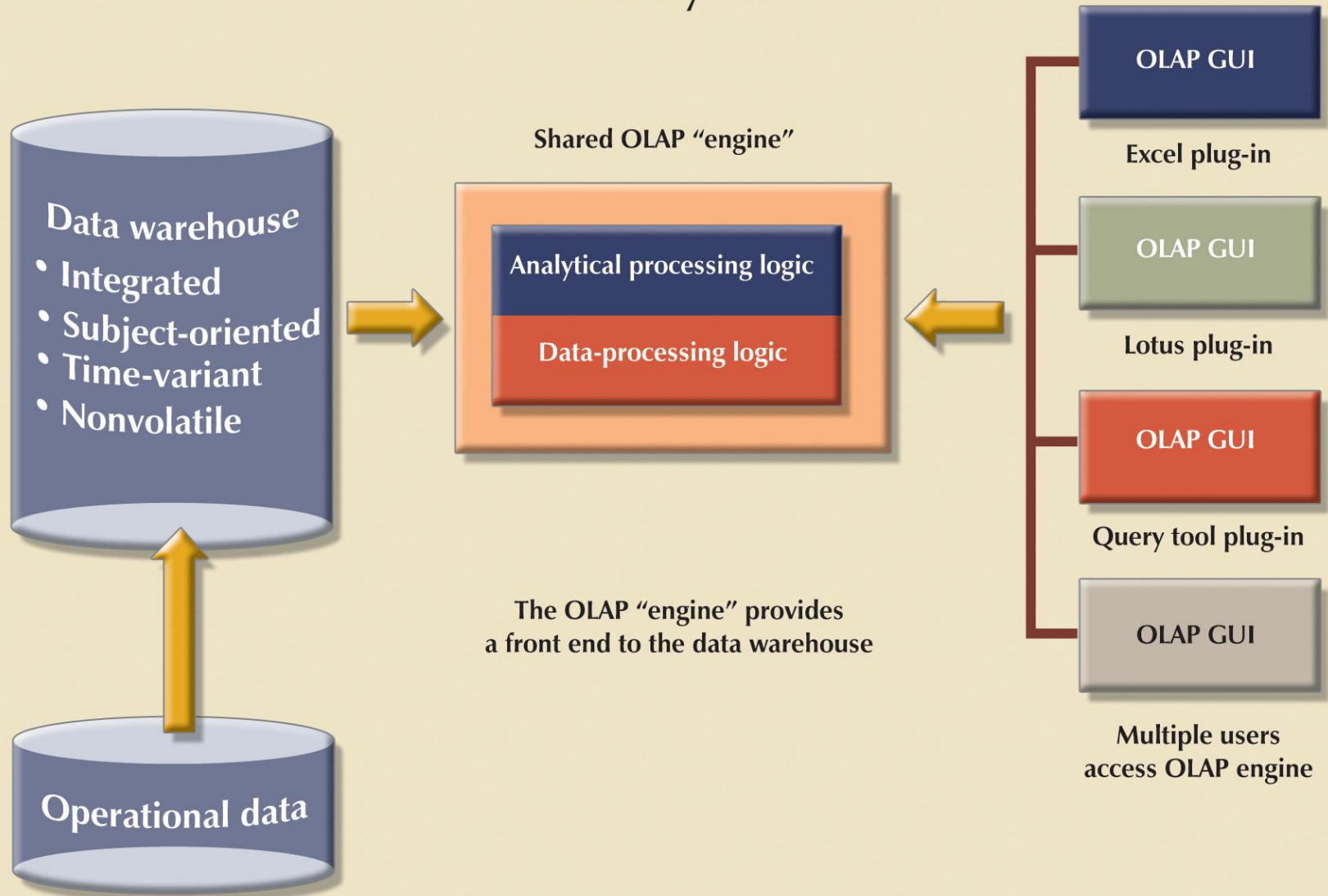
The OLAP system exhibits ...

- Client/Server architecture
- Easy-to-use GUI
 - Dimensional presentation
 - Dimensional modeling
 - Dimensional analysis
- Multidimensional data
 - Analysis
 - Manipulation
 - Structure
- Database support
 - Data warehouse
 - Operational DB
 - Relational
 - Multidimensional

**FIGURE
13.7**

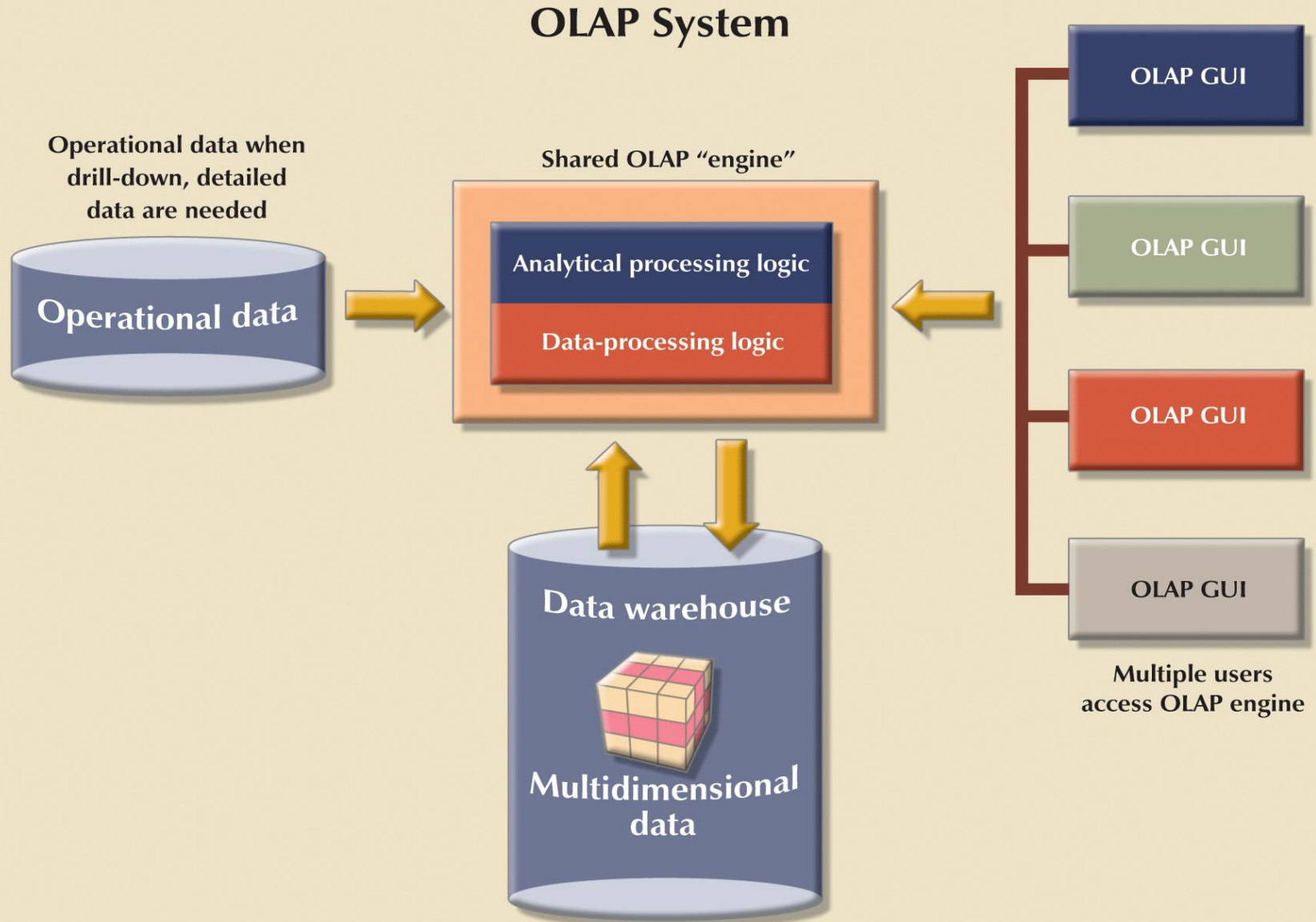
OLAP server arrangement

OLAP System



**FIGURE
13.8**

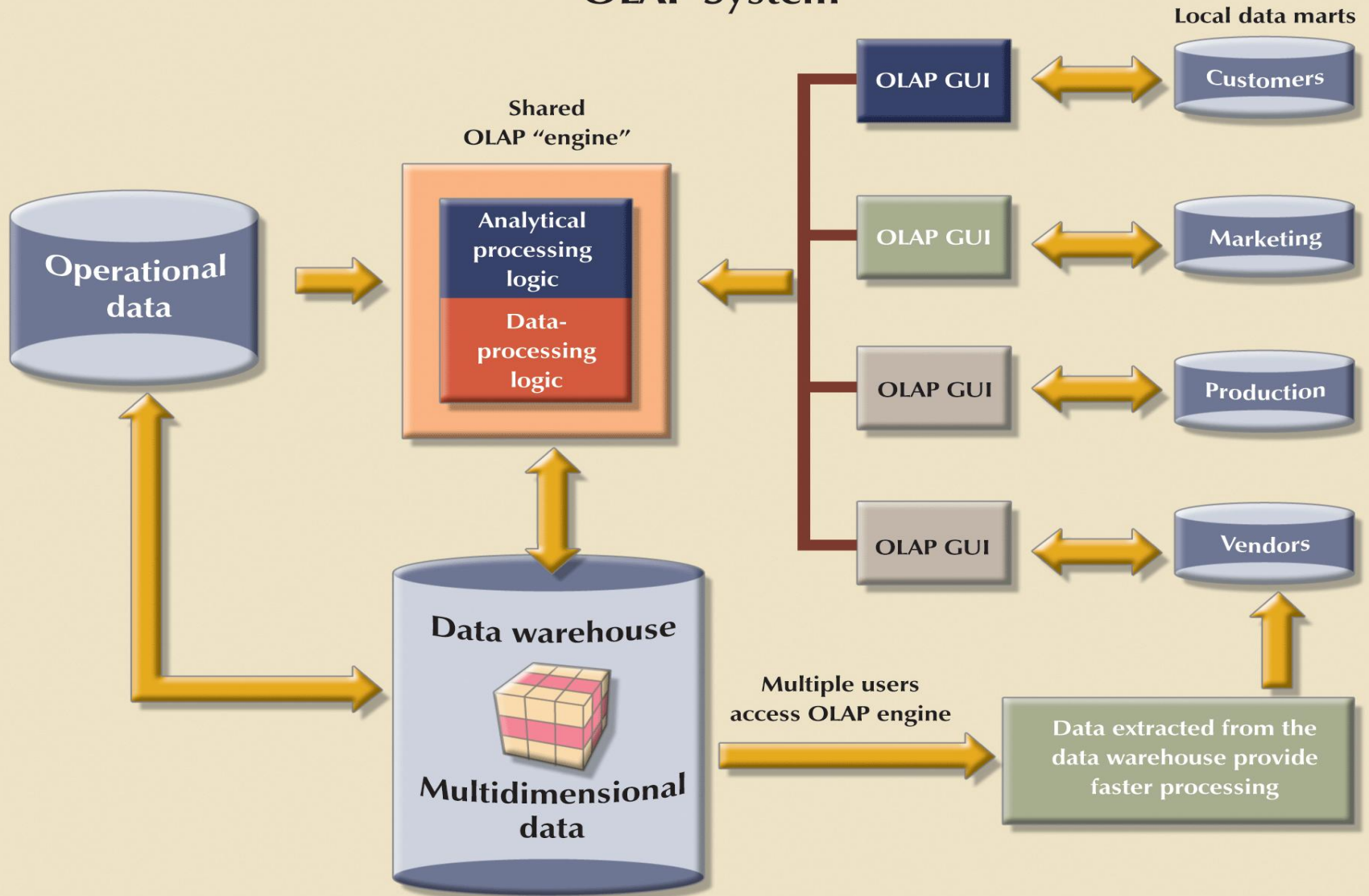
OLAP server with multidimensional data store arrangement



**FIGURE
13.9**

OLAP server with local mini data marts

OLAP System



On-Line Analytical Mining (OLAM)

- On-line analytical mining (OLAM) (also called OLAP mining) integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional databases.
- OLAM is particularly important for the following reasons:
 - High quality of data in data warehouses
 - Available information processing infrastructure surrounding data warehouses
 - OLAP-based exploratory data analysis
 - On-line selection of data mining functions

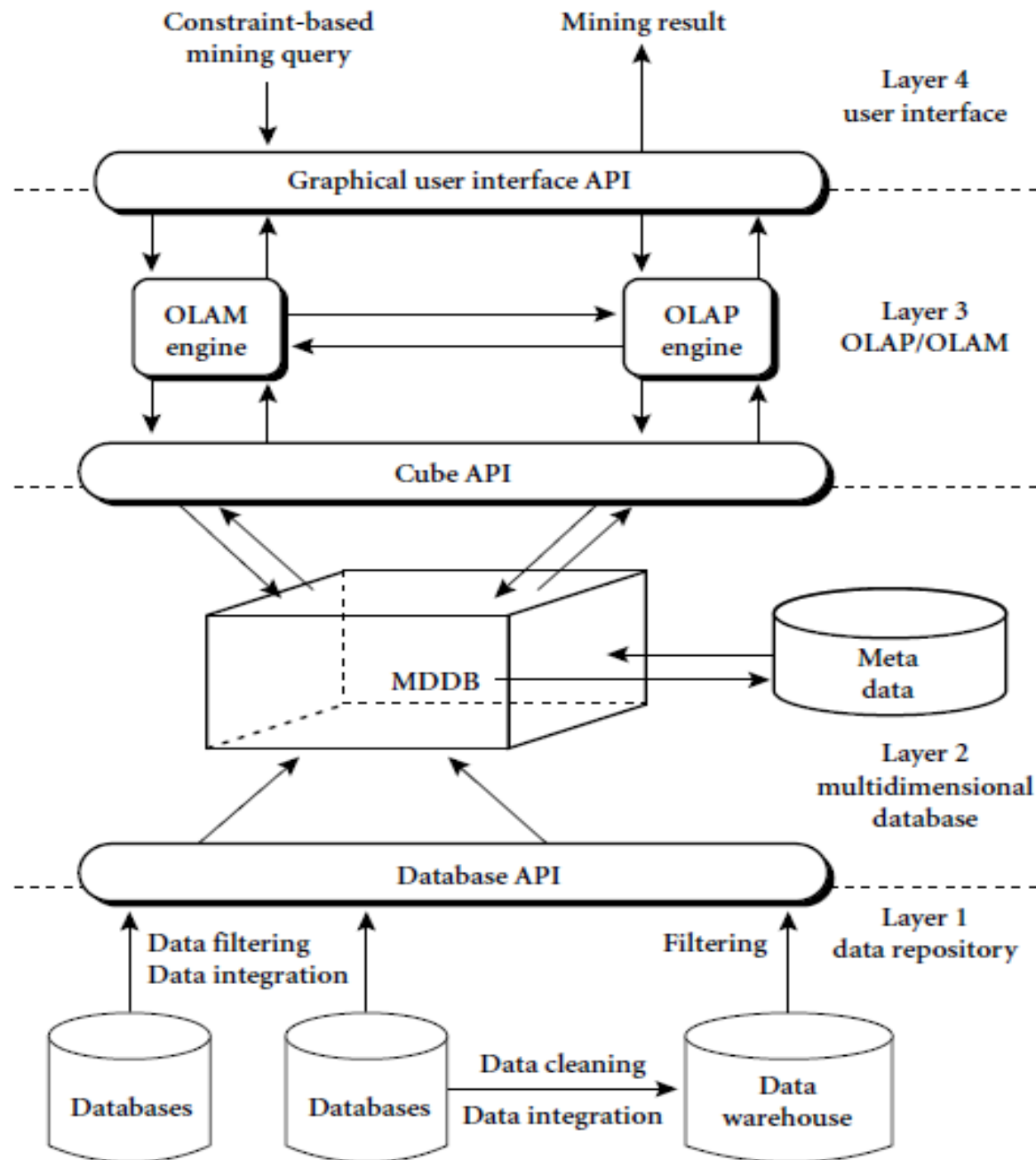


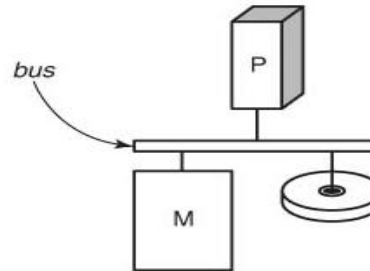
Figure An integrated OLAM and OLAP architecture.

- An OLAM server performs analytical mining in data cubes in a similar manner as an OLAP server performs on-line analytical processing.
- An integrated OLAM and OLAP architecture is shown in figure above, where the OLAM and OLAP servers both accept user on-line queries (or commands) via a graphical user interface API and work with the data cube in the data analysis via a cube API.

- A metadata directory is used to guide the access of the data cube.
- The data cube can be constructed by accessing and/or integrating multiple databases via an MDDDB API and/or by filtering a data warehouse via a database API that may support OLE DB or ODBC connections.

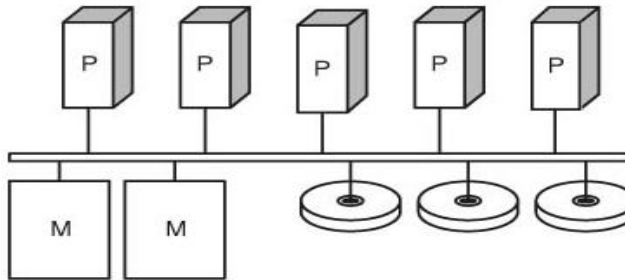
Server Options

- Single processor
(Uniprocessor)
- Symmetric
multiprocessor
(SMP)
- Massively parallel
processor (MPP)



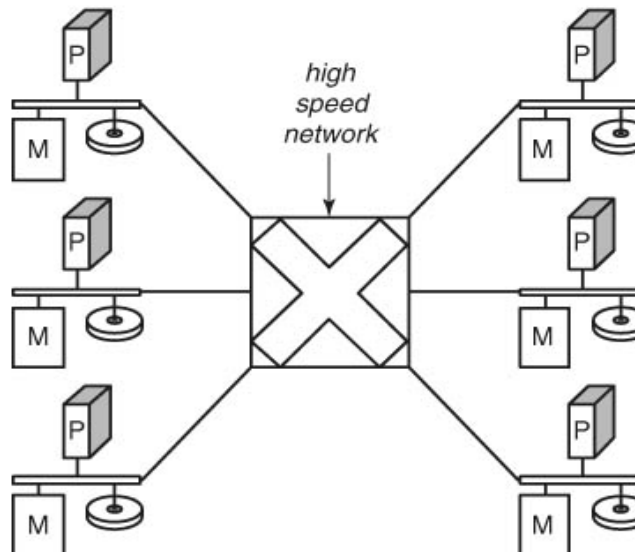
Uniprocessor

A simple computer follows the architecture laid out by Von Neumann. A Processing Unit communicates to Memory and Disk over a local bus. (Memory stores both data and the executable program.) The speed of the processor, bus, and memory limits performance and scalability.



SMP

The symmetric multiprocessor (SMP) has a shared-everything architecture. It expands the capabilities of the bus to support multiple processors, more memory, and more disk. The capacity of the bus limits performance and scalability. SMP architectures usually max out with fewer than 20 processing units.



MPP

The massively parallel processor (MPP) has a shared nothing architecture. It introduces a high speed network (also called a switch) that connects independent processor/memory/disk components. MPP architectures are very scalable but fewer software packages can take advantage of all the hardware.

OLAP Server Options/Categories of OLAP Tools

- OLAP tools are categorized according to the architecture of the underlying database.
- Three main categories of OLAP tools includes:
 - Relational OLAP (ROLAP)
 - Multi-dimensional OLAP (MOLAP or MD-OLAP)
 - DOLAP (Desktop OLAP)
 - Hybrid OLAP (HOLAP)

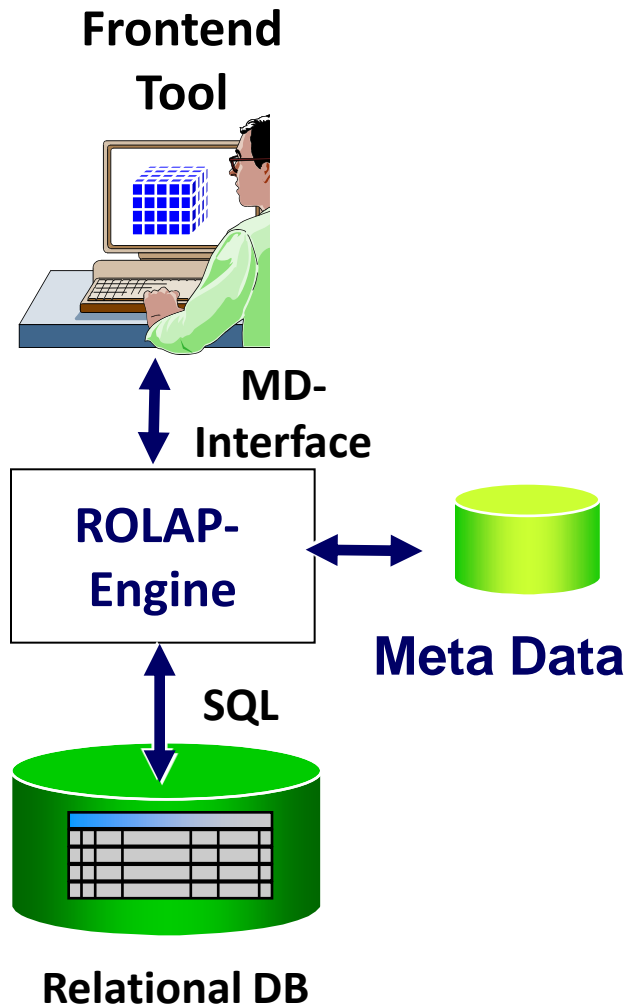
Relational OLAP (ROLAP)

Relational OLAP (ROLAP) implementations are similar in functionality to MOLAP. However, these systems use an underlying RDBMS, rather than a specialized MDDB. This gives them better scalability since they are able to handle larger volumes of data than the MOLAP architectures. Also, ROLAP implementations typically have better drill-through because the detail data resides on the same database as the multidimensional data .

The ROLAP environment is typically based on the use of a data structure known as a star or snowflake schema. Analogous to a virtual MDDB, a star or snowflake schema is a way of representing multidimensional data in a two-dimensional RDBMS.

The data modeler builds a fact table, which is linked to multiple dimension tables. The dimension tables consist almost entirely of keys, such as location, time, and product, which point back to the detail records stored in the fact table. This type of data structure requires a great deal of initial planning and set up, and suffers from some of the same operational and flexibility concerns of MDDBs. Additionally, since the data structures are relational, SQL must be used to access the detail records.

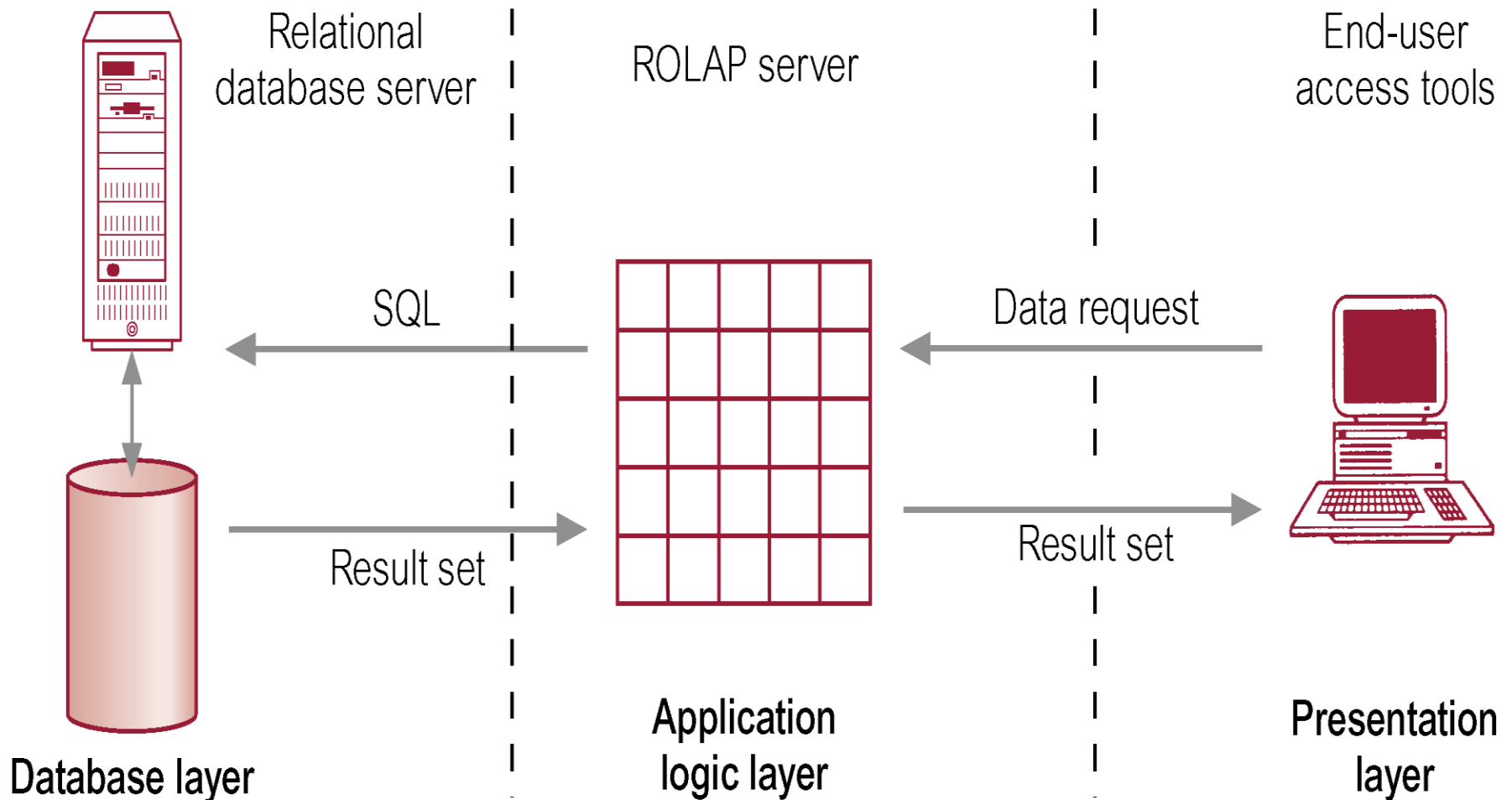
Therefore, the ROLAP engine must perform additional work to do comparisons, such as comparing the current quarter with this quarter last year. Again, IT must be heavily involved in defining, implementing, and maintaining the database. Furthermore, the ROLAP architecture often restricts the user from performing OLAP operations in a mobile environment



- Relational Online Analytical Processing (ROLAP)
 - OLAP functionality using relational database and familiar query tools to store and analyze multidimensional data
- Adds following extensions to traditional RDBMS:
 - Multidimensional data schema support within RDBMS
 - Data access language and query performance optimized for multidimensional data
- Support for Very Large Databases
- Tune a relational DBMS to support star schemas.

- ROLAP is a fastest growing style of OLAP technology.
- Supports RDBMS products using a metadata layer - avoids need to create a static multi-dimensional data structure - facilitates the creation of multiple multi-dimensional views of the two-dimensional relation.
- To improve performance, some products use SQL engines to support complexity of multi-dimensional analysis, while others recommend, or require, the use of highly denormalized database designs such as the star schema.

Typical Architecture for ROLAP Tools



- With ROLAP data remains in the original relational tables, a separate set of relational tables is used to store and reference aggregation data. ROLAP is ideal for large databases or legacy data that is infrequently queried.
- ROLAP Products:
 - IBM DB2, Oracle, Sybase IQ, RedBrick, Informix
- ROLAP Tools
 - ORACLE 8i
 - ORACLE Reports; ORACLE Discoverer
 - ORACLE Warehouse Builder
 - Arbors Software's Essbase

Advantages of ROLAP

- Define complex, multi-dimensional data with simple model
- Reduces the number of joins a query has to process
- Allows the data warehouse to evolve with rel. low maintenance
- HOWEVER! Star schema and relational DBMS are not the magic solution
 - Query optimization is still problematic

Features of ROLAP:

- Ask any question (not limited to the contents of the cube)
- Ability to drill down

Downsides of ROLAP:

- Slow Response
- Some limitations on scalability

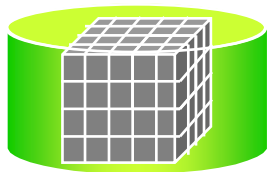
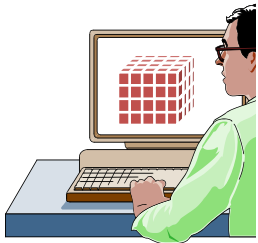
Multi-Dimensional OLAP (MOLAP)

The first generation of server-based multidimensional OLAP (MOLAP) solutions use multidimensional databases (MDDBs). The main advantage of an MDDB over an RDBMS is that an MDDB can provide information quickly since it is calculated and stored at the appropriate hierarchy level in advance. However, this limits the flexibility of the MDDB since the dimensions and aggregations are predefined. If a business analyst wants to examine a dimension that is not defined in the MDDB, a developer needs to define the dimension in the database and modify the routines used to locate and reformat the source data before an operator can load the dimension data.

Another important operational consideration is that the data in the MDDB must be periodically updated to remain current. This update process needs to be scheduled and managed. In addition, the updates need to go through a data cleansing and validation process to ensure data consistency. Finally, an administrator needs to allocate time for creating indexes and aggregations, a task that can consume considerable time once the raw data has been loaded. (These requirements also apply if the company is building a data warehouse that is acting as a source for the MDDB.)

Organizations typically need to invest significant resources in implementing MDDB systems and monitoring their daily operations. This complexity adds to implementation delays and costs, and requires significant IT involvement. This also results in the analyst, who is typically a business user, having a greater dependency on IT. Thus, one of the key benefits of this OLAP technology — the ability to analyze information without the use of IT professionals — may be significantly diminished.

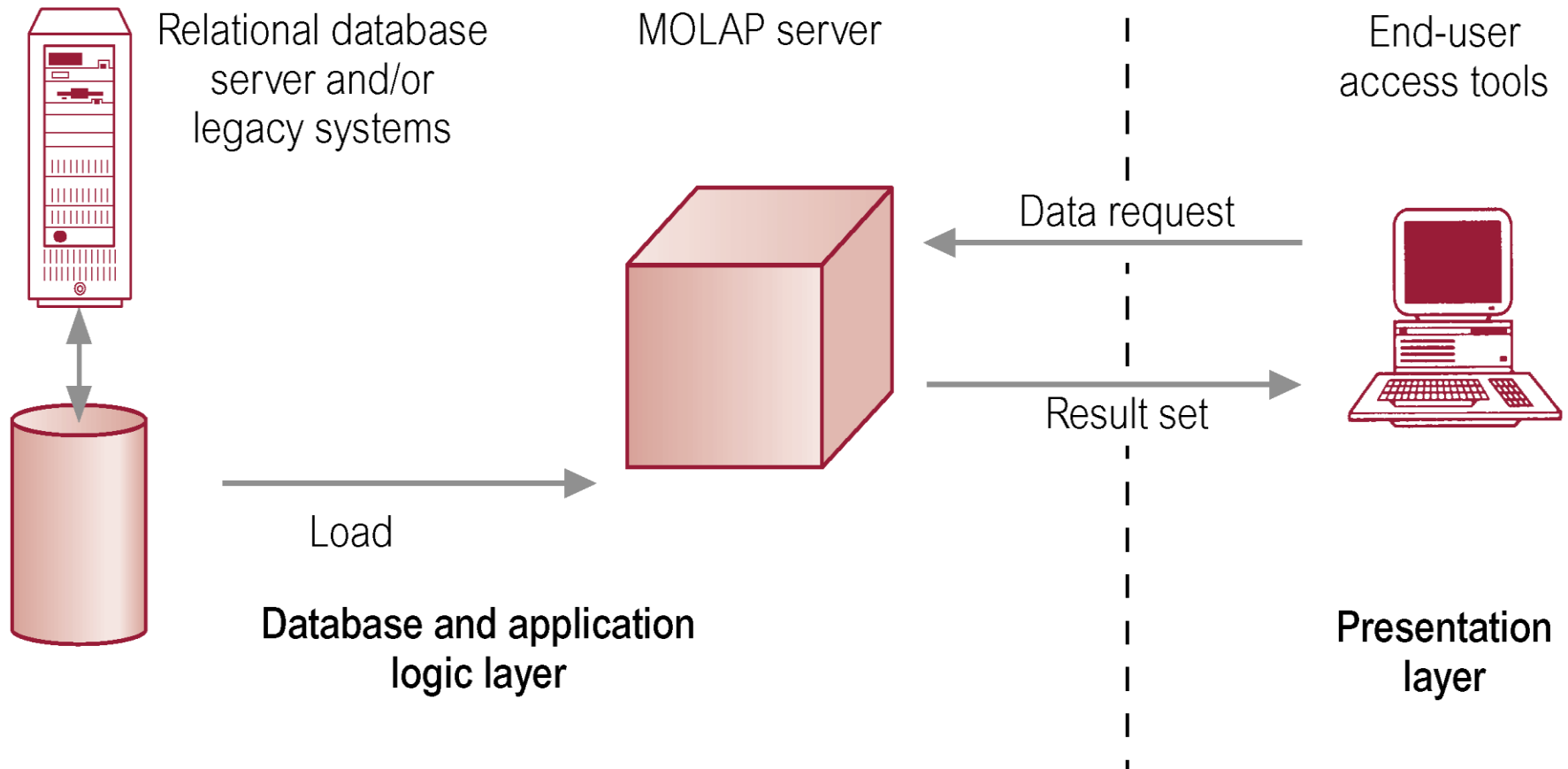
**Front-end
Tool**



**Multidimensional
Database**

- Uses specialized data structures and multi-dimensional Database Management Systems (MD-DBMSs) to organize, navigate, and analyze data.
- Use a specialized DBMS with a model such as the “data cube.”
- Data is typically aggregated and stored according to predicted usage to enhance query performance.

Typical Architecture for MOLAP Tools



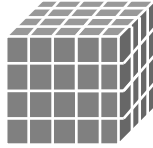
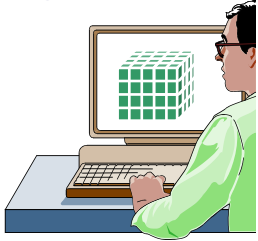
- Traditionally, require a tight coupling with the application layer and presentation layer.
- Recent trends segregate the OLAP from the data structures through the use of published application programming interfaces (APIs).
- MOLAP Products
 - Pilot, Arbor Essbase, Gentia
- MOLAP Tools
 - ORACLE Express Server
 - ORACLE Express Clients (C/S and Web)
 - MicroStrategy's DSS server
 - Platinum Technologies' Plantinum InfoBeacon

- Use array technology and efficient storage techniques that minimize the disk space requirements through sparse data management.
- Provides excellent performance when data is used as designed, and the focus is on data for a specific decision-support application.
- Features:
 - Very fast response
 - Ability to quickly write data into the cube
- Downsides:
 - Limited Scalability
 - Inability to contain detailed data
 - Load time

Desktop OLAP (or Client OLAP)

The desktop OLAP market resulted from the need for users to run business queries using relatively small data sets extracted from production systems. Most desktop OLAP systems were developed as extensions of production system report writers, while others were developed in the early days of client/server computing to take advantage of the power of the emerging (at that time) PC desktop. Desktop OLAP systems are popular and typically require relatively little IT investment to implement. They also provide highly mobile OLAP operations for users who may work remotely or travel extensively. However, most are limited to a single user and lack the ability to manage large data sets.

Client-OLAP



Stores in the form of cubes/micro-cubes on the desktop/client machine

Products:

Brio.Enterprise
BusinessObjects
Cognos PowerPlay

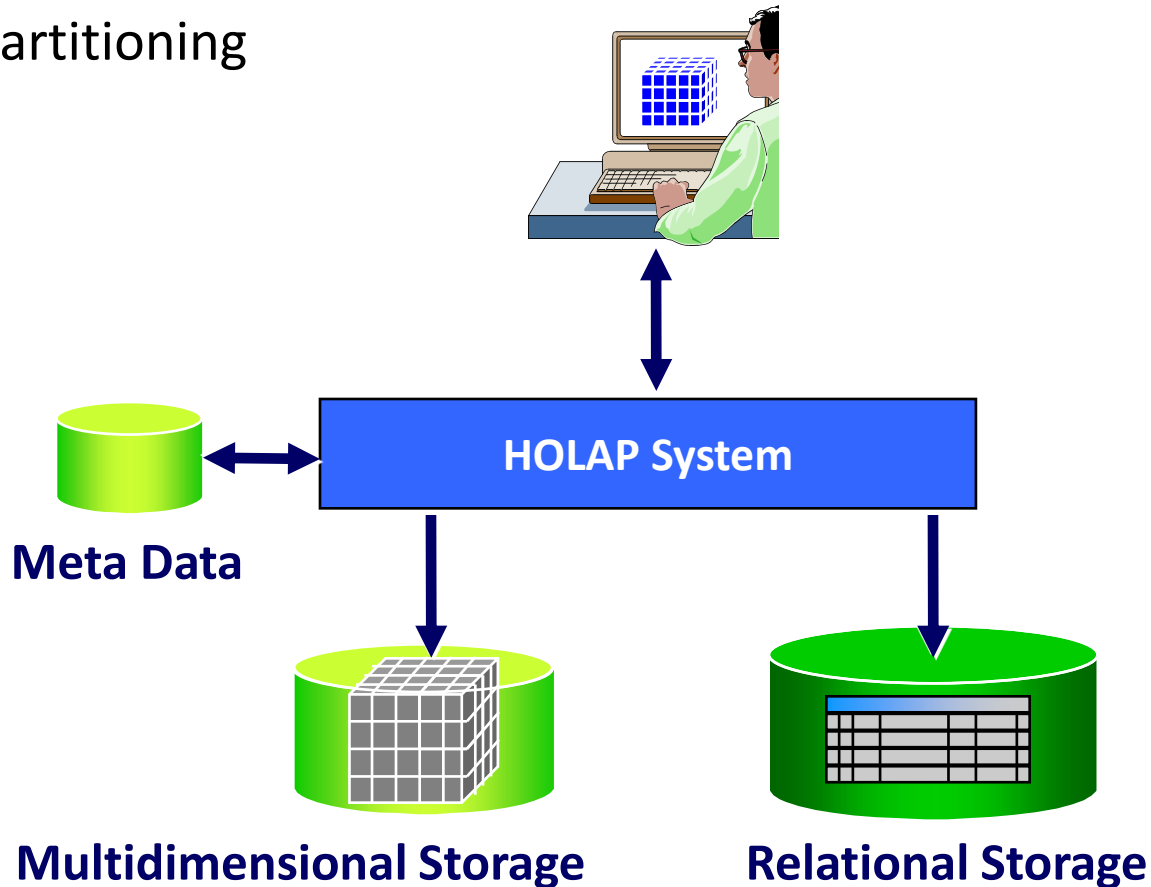
- proprietary data structure on the client
- data stored as file
- mostly RAM based architectures
- mobile user
- ease of installation and use
- data volume
- no multiuser capabilities

Hybrid OLAP (HOLAP)

Some vendors provide the ability to access relational databases directly from an MDDB, giving rise to the concept of hybrid OLAP environments. This implements the concept of "drill through," which automatically generates SQL to retrieve detail data records for further analysis. This gives end users the perception they are drilling past the multidimensional database into the source database.

The hybrid OLAP system combines the performance and functionality of the MDDB with the ability to access detail data, which provides greater value to some categories of users. However, these implementations are typically supported by a single vendor's databases and are fairly complex to deploy and maintain. Additionally, they are typically somewhat restrictive in terms of their mobility.

- Can use data from either a RDBMS directly or a multi-dimension server.
- Equal treatment of MD and Relational Data
- Storage type at the discretion of the administrator
- Cube Partitioning



- HOLAP combines elements from MOLAP and ROLAP. HOLAP keeps the original data in relational tables but stores aggregations in a multidimensional format.
- Combines MOLAP & ROLAP
- Utilizes both pre-calculated cubes & relational data sources
- HOLAP Tools
 - ORACLE 8i
 - ORACLE Express Serve
 - ORACLE Relational Access Manager
 - ORACLE Express Clients (C/S and Web)
- HOLAP Products:
 - Oracle Express
 - Seagate Holos
 - Speedware Media/M
 - Microsoft OLAP Services

HOLAP Features:

- For summary type info – cube, (Faster response)
- Ability to drill down – relational data sources (drill through detail to underlying data)
- Source of data transparent to end-user

OLAP Products

OLAP Category	Candidate Products	Vendor
ROLAP	Microstrategy	Microstrategy
	Business Objects	Business Objects
	Crystal Holos (ROLAP Mode)	Business Objects
	Essbase	Hyperion
	Microsoft Analysis Services	Microsoft
	Oracle Express (ROLAP Mode)	Oracle
	Oracle Discoverer	Oracle
MOLAP	Crystal Holos	Business Objects
	Essbase	Hyperion
	Microsoft Analysis Services	Microsoft
	Oracle Express	Oracle
	Cognos Powerplay	Cognos
HOLAP	Hyperion Essbase+Intelligence	Hyperion
	Cognos Powerplay+Impromptu	Cognos
	Business Objects+Crystal Holos	Business Objects

Typical OLAP Operations

Roll up (drill-up) or Aggregation: summarize data

- *by climbing up hierarchy or by dimension reduction*
- Data is summarized with increasing generalization
- dimension reduction: e.g., total sales by city
- summarization over aggregate hierarchy: e.g., total sales by city and year -> total sales by region and by year

Drill down (roll down): reverse of roll-up

- *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- going from summary to more detailed views
- Increasing levels of detail are revealed

Slice and Dice:

- *project and select*
- Performing projection operations on the dimensions.

Pivot (rotate):

- *reorient the cube, visualization, 3D to series of 2D planes.*
- Cross tabulation is performed

Other operations:

- **drill across:** *involving (across) more than one fact table*
- **drill through:** *through the bottom level of the cube to its back-end relational tables (using SQL)*

<i>location</i> = “Chicago”					<i>location</i> = “New York”				<i>location</i> = “Toronto”				<i>location</i> = “Vancouver”			
<i>item</i>					<i>item</i>				<i>item</i>				<i>item</i>			
<i>home</i>					<i>home</i>				<i>home</i>				<i>home</i>			
<i>time</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Table: A 3-D view of sales data according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars sold* (in thousands).

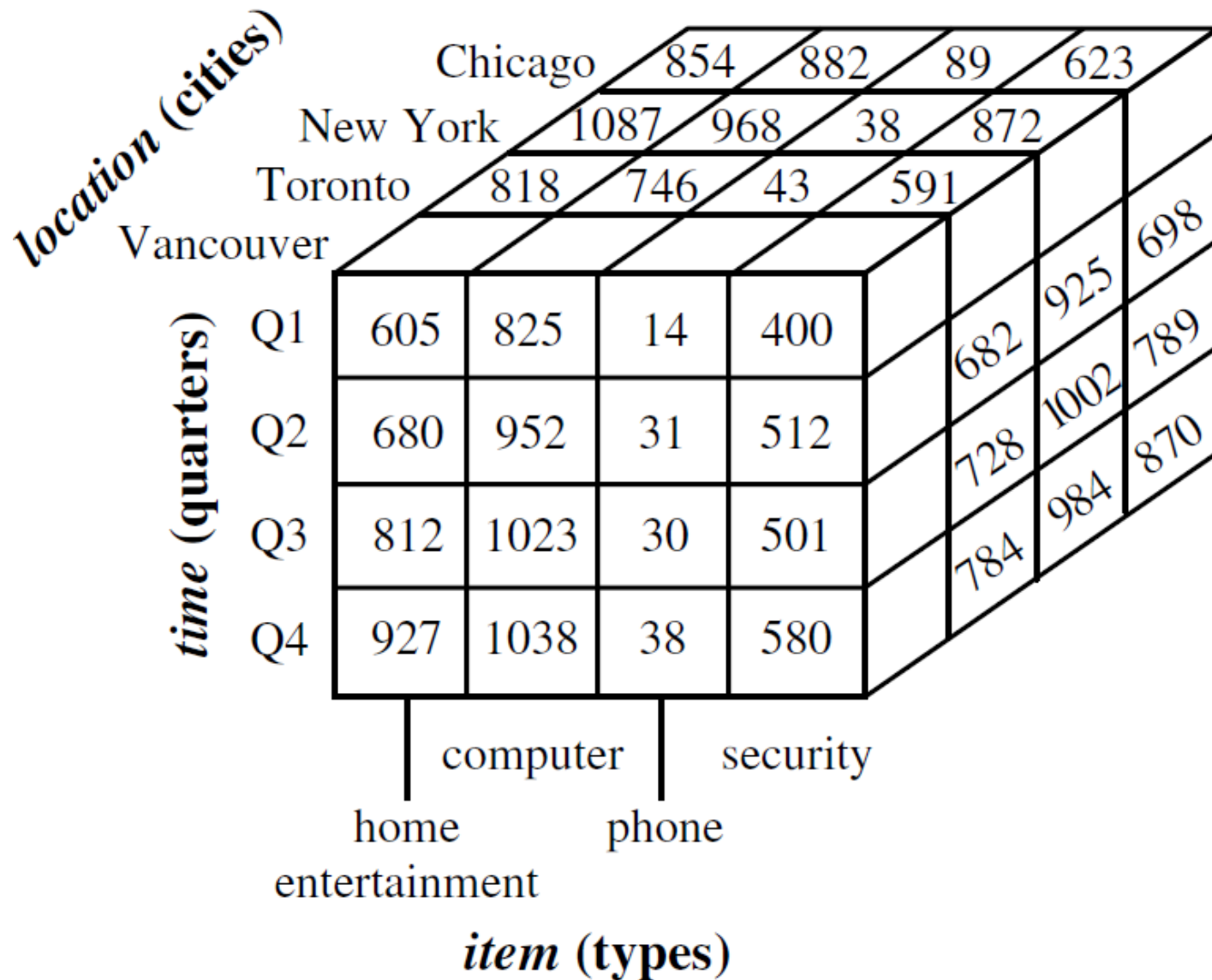
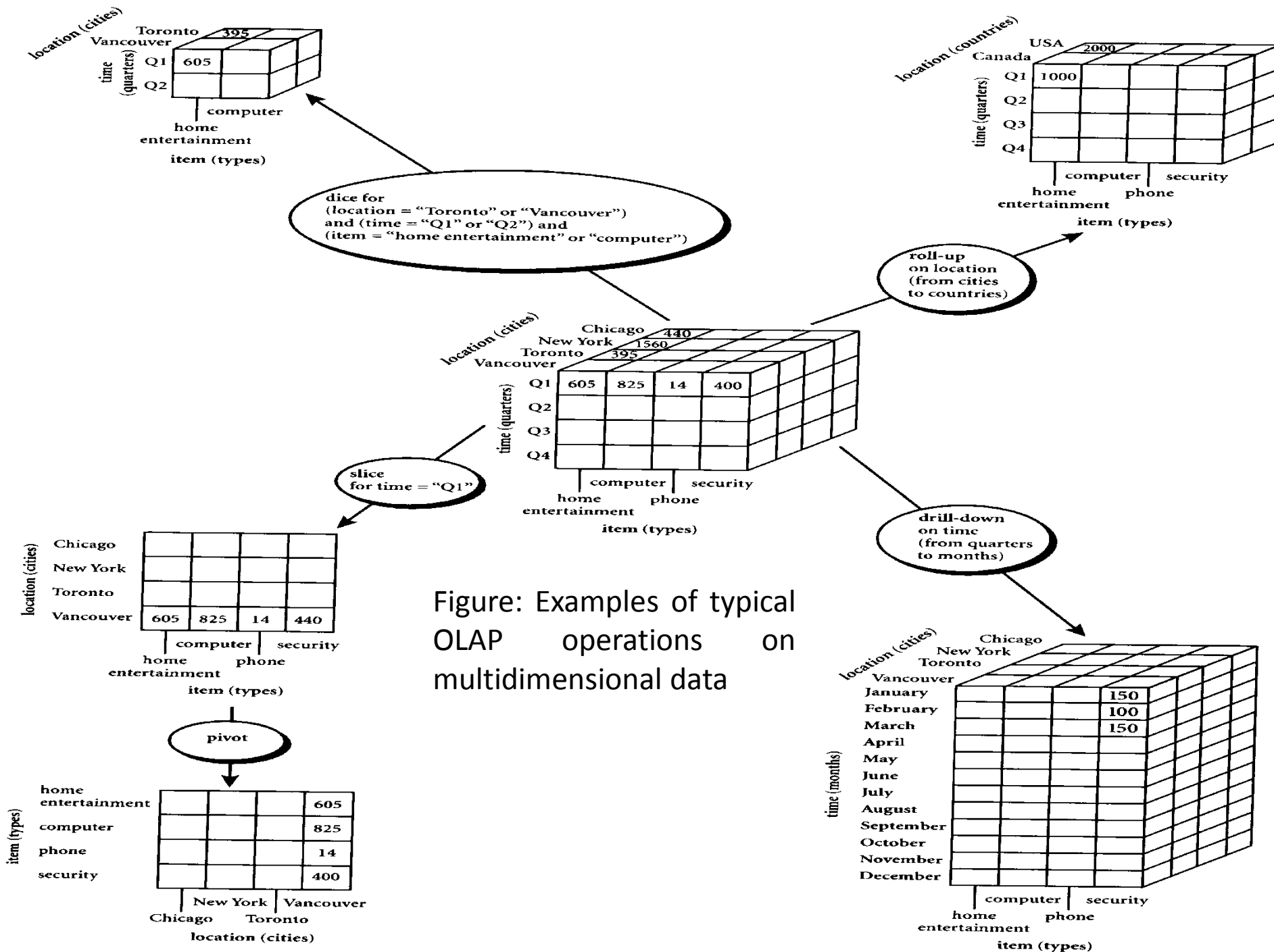


Figure: A 3-D data cube representation of the data in the table above, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars sold* (in thousands).

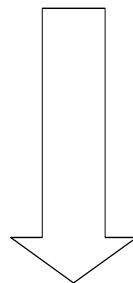


Roll-up and Drill-down

The **roll-up** operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction such that one or more dimensions are removed from the given cube.

Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions.

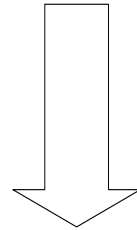
	Food Line	Outdoor Line	CATEGORY_total
Canada	29,116.5	69,310	98,426.5
Mexico	12,743.5	24,284	37,027.5
United States	102,561.5	232,679	335,240.5



Roll-Up

	Food Line	Outdoor Line	CATEGORY_total
North America	144,421.5	326,273	470,694.5

	Food Line	Outdoor Line	CATEGORY_total
Asia	59,728	151,174	210,902

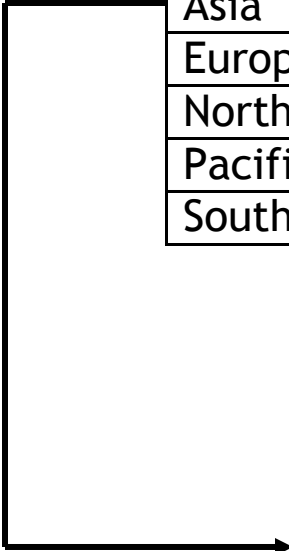


Drill-Down

	Food Line	Outdoor Line	CATEGORY_total
Malaysia	618	9,418	10,036
China	33,198.5	74,165	107,363.5
India	6,918	0	6,918
Japan	13,871.5	34,965	48,836.5
Singapore	5,122	32,626	37,748
Belgium	7797.5	21,125	28,922.5

Drill Down Example

Region	Sales variance
Africa	105%
Asia	57%
Europe	122%
North America	97%
Pacific	85%
South America	163%



Nation	Sales variance
China	123%
Japan	52%
India	87%
Singapore	95%

Sample OLAP Drill down online report

Product	All Product	Promotion	All Promotion Name
Store	All Store	Time	All Time
Yearly Income	All Yearly Income		

			MeasuresLevel	
- Country	- State Province	- City	Lname	Store Sales
All Customer	All Customer Total			1,079,147.47
+ Canada	Canada Total			98,045.46
+ Mexico	Mexico Total			430,293.59
	USA Total			550,808.42
	+ CA	CA Total		154,513.49
	+ OR	OR Total		128,598.50
		WA Total		267,696.43
		+ Anacortes	Anacortes Total	1,338.23
		+ Ballard	Ballard Total	5,301.58
		+ Bellingham	Bellingham Total	1,679.21
		+ Bremerton	Bremerton Total	25,927.72
		+ Burien	Burien Total	5,091.41
		+ Edmonds	Edmonds Total	4,583.23
		+ Everett	Everett Total	5,427.29
		+ Issaquah	Issaquah Total	4,583.63
		+ Kirkland	Kirkland Total	6,013.32
		+ Lynnwood	Lynnwood Total	5,199.78
		+ Marysville	Marysville Total	4,851.97
		+ Olympia	Olympia Total	27,800.70
		+ Port Orchard	Port Orchard Total	25,207.47
		+ Puyallup	Puyallup Total	23,123.39
			Redmond Total	5,158.29
			Abbey	30.33
			Alstorn	104.98
			Autobee	
			Bagwell	108.43
			Banks	8.04
			Bateman	35.12
			Bates	169.90
			Beerbaum	85.24
			Berner	31.57

- USA	- WA	- Redmond	
-------	------	-----------	--

Double-click a member to drill up or down.

- Drill Down
- Drill Up
- Member Properties...
- Customer info

Figure:
Example of drill-down

Quarterly Auto Sales Summary		
<u>Region</u>	<u>Units Sold</u>	<u>Revenue</u>
Northeast	_____	_____
Southeast	_____	_____
Central	_____	_____
Northwest	_____	_____
Southwest	_____	_____
	_____	_____

Quarterly Auto Sales Summary			
<u>Region</u>	<u>State</u>	<u>Units Sold</u>	<u>Revenue</u>
Northeast	Maine	_____	_____
	New York	_____	_____
	Massachusetts	_____	_____
Southeast	Florida	_____	_____
	Georgia	_____	_____
	Virginia	_____	_____

Figure:
Example of Roll up

Quarterly Auto Sales Summary

<u>Region</u>	<u>State</u>	<u>Units Sold</u>	<u>Revenue</u>
Northeast	Maine		
	New York		
	Massachusetts		
Southeast	Florida		
	Georgia		
	Virginia		

Quarterly Auto Sales Summary

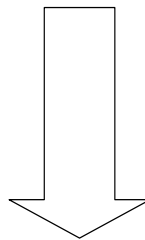
<u>Region</u>	<u>Units Sold</u>	<u>Revenue</u>
Northeast		
Southeast		
Central		
Northwest		
Southwest		

Slice and Dice

The **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.

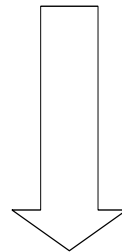
	Food Line	Outdoor Line	CATEGORY_total
Asia	59,728	151,174	210,902
Europe	97,580.5	213,304	310,884.5
North America	144,421.5	326,273	470,694.5
REGION_total	301,730	690,751	992,481



Slice

	Food Line	Outdoor Line	CATEGORY_total
North America	144,421.5	326,273	470,694.5

	Food Line	Outdoor Line	CATEGORY_total
Canada	29,116.5	69,310	98,426.5
Mexico	12,743.5	24,284	37,027.5
United States	102,561.5	232,679	335,240.5



Dice

	Food Line	Outdoor Line
Mexico	12,743.5	24,284
United States	102,561.5	232,679

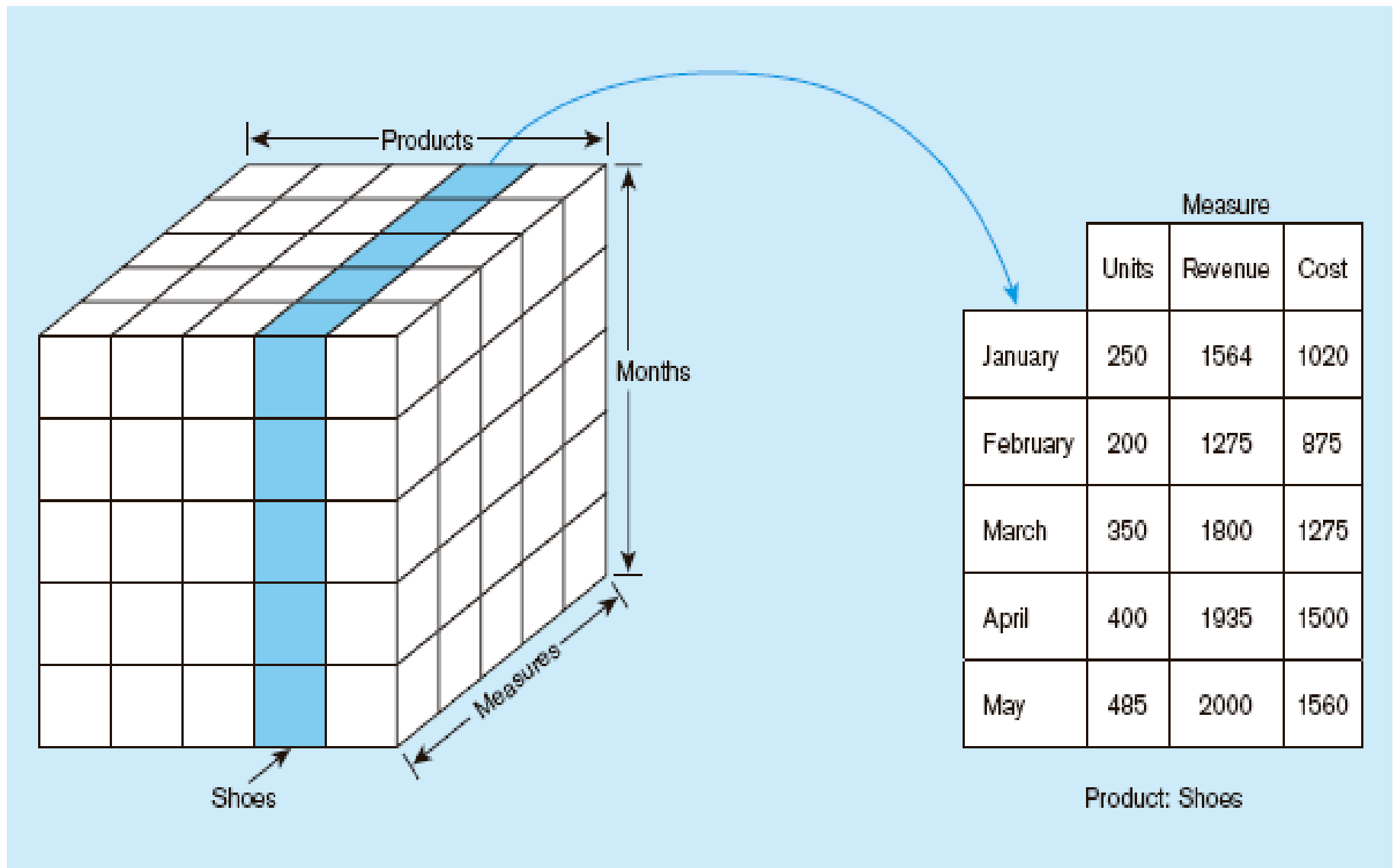
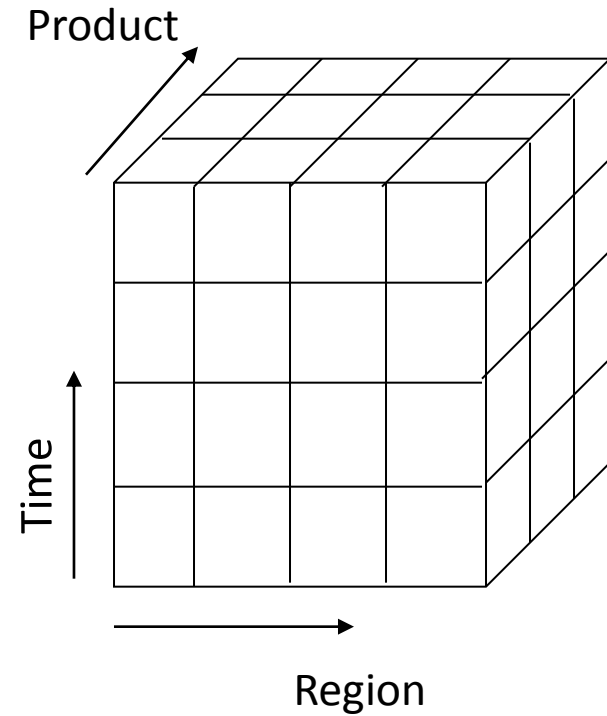
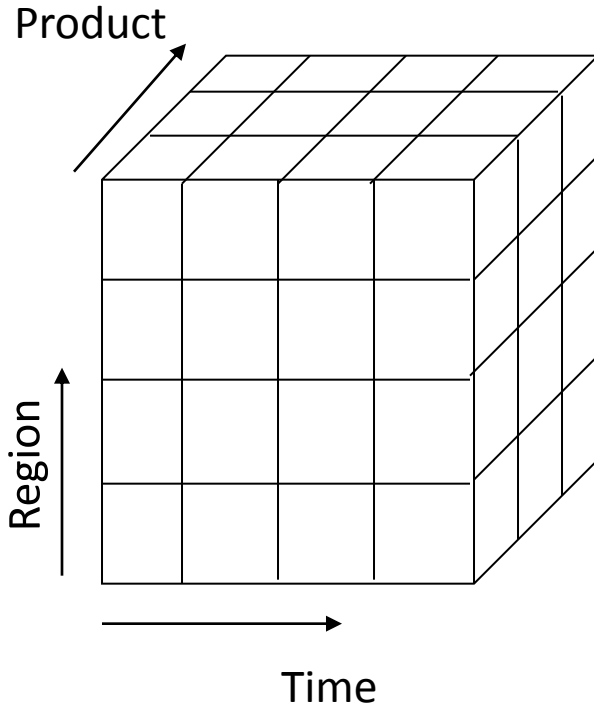


Figure: Slicing a data cube

Rotation (Pivot Table)

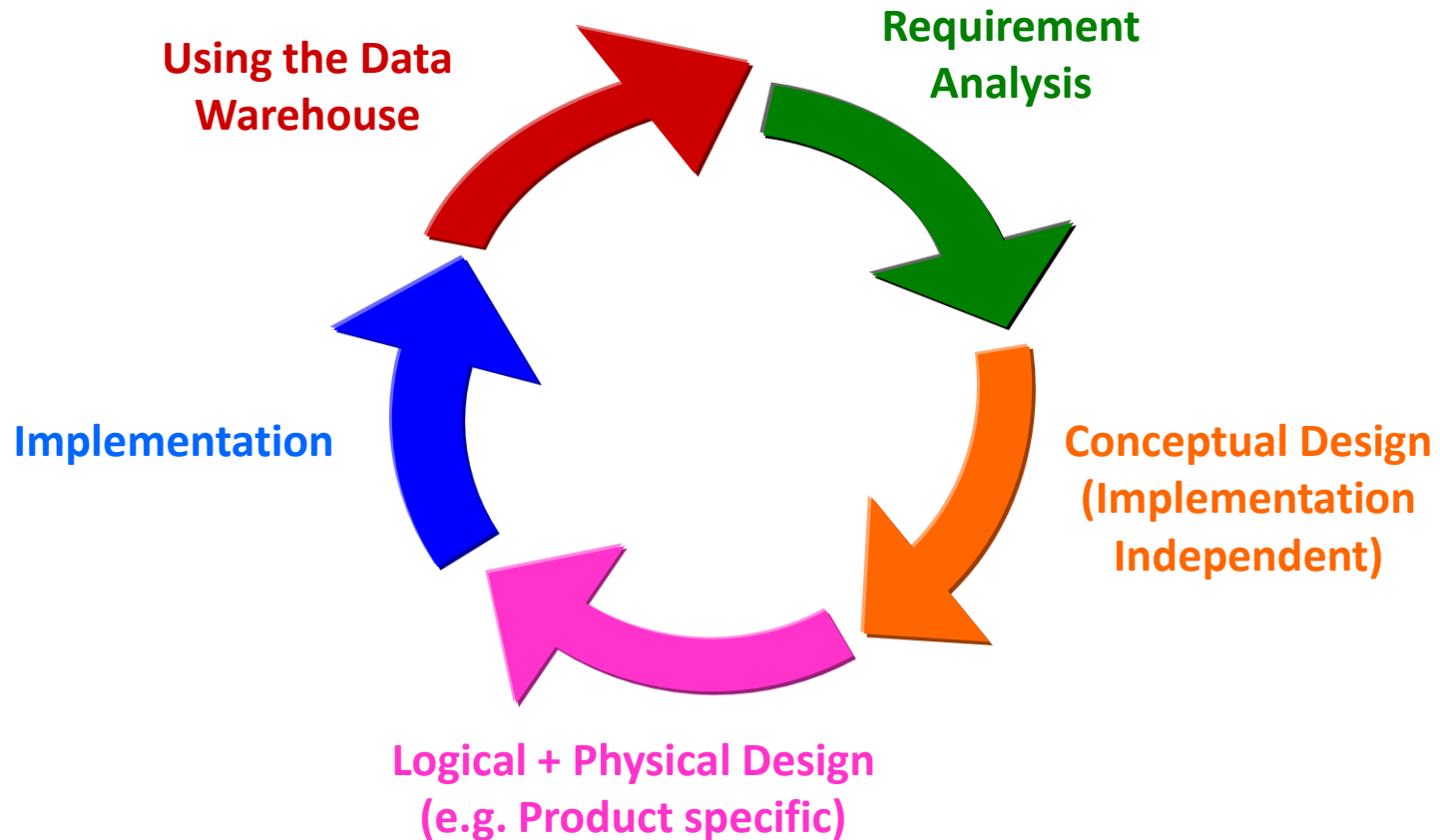


Example of Rotation (Pivot Table)

		Region			
Year	Data	Asia	Europe	North America	Grand Total
1995	Sum of Hardware	97	23	198	318
	Sum of Software	83	41	425	549
1996	Sum of Hardware	115	28	224	367
	Sum of Software	78	65	410	553
1997	Sum of Hardware	102	25	259	386
	Sum of Software	55	73	497	625
Total Sum of Hardware		314	76	681	1071
Total Sum of Software		216	179	1332	1727

		Year			
Region	Data	1995	1996	1997	Grand Total
Asia	Sum of Hardware	97	115	102	314
	Sum of Software	83	78	55	216
Europe	Sum of Hardware	23	28	25	76
	Sum of Software	41	65	73	179
North America	Sum of Hardware	198	224	259	681
	Sum of Software	425	410	497	1332
Total Sum of Hardware		318	367	386	1071
Total Sum of Software		549	553	625	1727

Design and Query Processing



Cube Operation

Cube definition and computation in DMQL

```
define    cube    sales[item,    city,    year]:  
    sum(sales_in_dollars)  
  
compute cube sales
```

Transform it into a SQL-like language (with a new operator **cube by**)

```
SELECT item, city, year, SUM (amount)  
FROM SALES  
  
CUBE BY item, city, year
```


- Aggregate a measure on one or more dimension
- Summarize at different levels of a dimension hierarchy (state - city)
 - Total sales per city aggregated to obtain Total sales per State - *roll-up*
 - Total sales per state probed further to obtain Total sales per city - *drill-down*
- *Slicing* - an equality selection on one or more dimensions, possibly also with some dimensions projected out
- *Dicing* - range selection

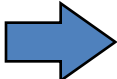
Note: k dimensions, lead to 2^k SQL queries

SQL extension for OLAP

Aggregates

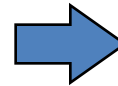
- Add up amounts for day 1
- In SQL: `SELECT sum(amt) FROM SALE
WHERE date = 1`

sale	prodlid	storeld	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

 **81**

- Add up amounts by day
- In SQL: `SELECT date, sum(amt) FROM SALE
GROUP BY date`

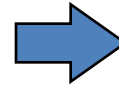
sale	prodlid	storeid	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4



ans	date	sum
	1	81
	2	48

- Add up amounts by day, product
- In SQL: `SELECT date, sum(amt) FROM SALE GROUP BY date, prodId`

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4



sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48



Operators for Aggregation: sum, count, max, min, median, avg

Data Mining Tools

The SPSS logo is displayed in white, bold, sans-serif capital letters on a solid red rectangular background.The ORACLE logo is rendered in a large, bold, red, sans-serif font. A registered trademark symbol (®) is located at the top right of the word.The Microsoft SQL Server 2012 logo features the word "Microsoft" in a small, black, sans-serif font above the words "SQL Server" in a large, bold, black, sans-serif font. The year "2012" is positioned to the right of "SQL Server" in the same large, bold, black, sans-serif font. A registered trademark symbol (®) is located at the top right of "Microsoft" and at the top right of "SQL Server".

WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

WEKA is free software available under the GNU General Public License.

Features:

- Written in JAVA
- Has graphical user interfaces
- Contains a collection of visualization tools and algorithms for data analysis and predictive modeling
- Supports standard data mining tasks like data preprocessing, clustering, classification, regression, visualization, and feature selection

Usage:

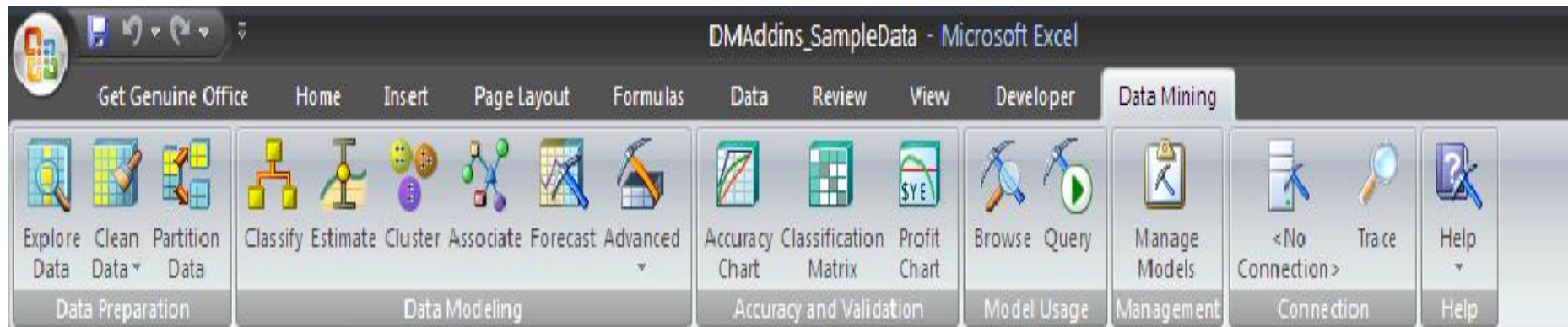
- Apply a learning method to a dataset & analyze the result
- Use a learned model to make predictions on new instances
- Apply different learners to a dataset & compare results

MS Excel

In order to bridge the gap between the common user and the complex data mining process, Microsoft has introduced a new and efficient data mining tool ,the Microsoft SQL Server 2005 Data Mining Add-Ins for Office 2007 putting data mining within the reach of every user or desktop.

The add-in can be downloaded from the following link: [DOWNLOAD LINK](#).

- The software pre-requisites for using the add-in are:
- Microsoft Office 2007 installed.
- Microsoft SQL Server 2005 or above installed.
 - Microsoft .NET 2.0 framework or higher (for SQL server 2008 only).
 - Microsoft PowerShell (for SQL server 2008 only)



Once the add-in is installed, you can see the DATA MINING tab in the EXCEL ribbon. The tab contains different options like:

Data Preparation

- Explore Data.
- Clean Data.
- Partition Data.

Data Modeling

- Classify.
- Estimate.
- Cluster.
- Associate.
- Forecast.
- Advanced.

Accuracy and Validation

- Accuracy Chart.
- Classification Matrix.
- Profit Chart.

Model Usage

- Browse.
- Query.

Management

Connection

- No Connection.
- Trace.

Help

Conclusion:

The Microsoft SQL Server Data mining add-in for Microsoft Excel provides users with an easy to use interface that is capable of performing complex data mining tasks with ease. The add-in can be extremely useful for both, people who just want to get more out of their data and also for those interested in serious data mining.

Microsoft SQL Server

Microsoft SQL Server is a relational database server, developed by Microsoft: it is a software product whose primary function is to store and retrieve data as requested by other software applications, be it those on the same computer or those running on another computer across a network (including the Internet).

- Microsoft has introduced a wealth of new data mining features in Microsoft SQL Server 2008 that allow businesses to answer their concerns with data and mining for information in them.
- The current version of SQL Server, SQL Server 2008, (code-named "Katmai") aims to make data management self-tuning, self organizing, and self maintaining.
- SQL Server 2008 data mining features are integrated across all the SQL Server products, including SQL Server, SQL Server Integration Services, and Analysis Services.
- Accessing the data mining results is as simple as using an SQL-like language called Data Mining Extensions to SQL, or DMX.

Oracle

The **Oracle Database** (commonly referred to as *Oracle RDBMS* or simply as *Oracle*) is an object-relational database management system (ORDBMS) produced and marketed by Oracle Corporation.

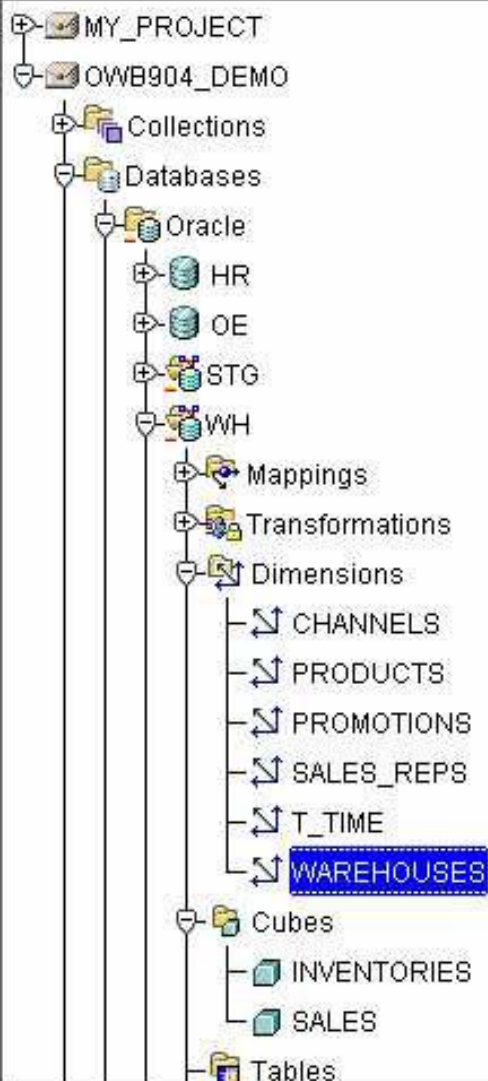
- **Oracle Data Mining (ODM)** is used to incorporate data mining with the Oracle database.
- ODM is used for both supervised (where a particular target value should be specified) and unsupervised (where patterns in data are observed) data mining.
- The results of Oracle Data Mining can be viewed by the Oracle Business Intelligence's reporting/publishing component.
- Oracle BI Standard Edition One is a product that is used to extract business information concealed in the data.

- **Oracle Warehouse Builder (OWB)** is used to create the logical and physical design of the data mart.
- The **Oracle BI Server** is used to build a repository of metadata from the data mart that was created using the Oracle Data warehouse builder.
- The users ultimately interact with Oracle BI Answers to extract useful information from the data mart created using OWB (Oracle Warehouse Builder).
- **Oracle BI Interactive Dashboards** are used to publish the data extracted from the data mart so that the users can have an easy access to it.
- **Oracle BI publisher** is used to create reports that are very essential to any kind of business.

Project Edit Object View Tools Window Help



ORACLE



SPSS

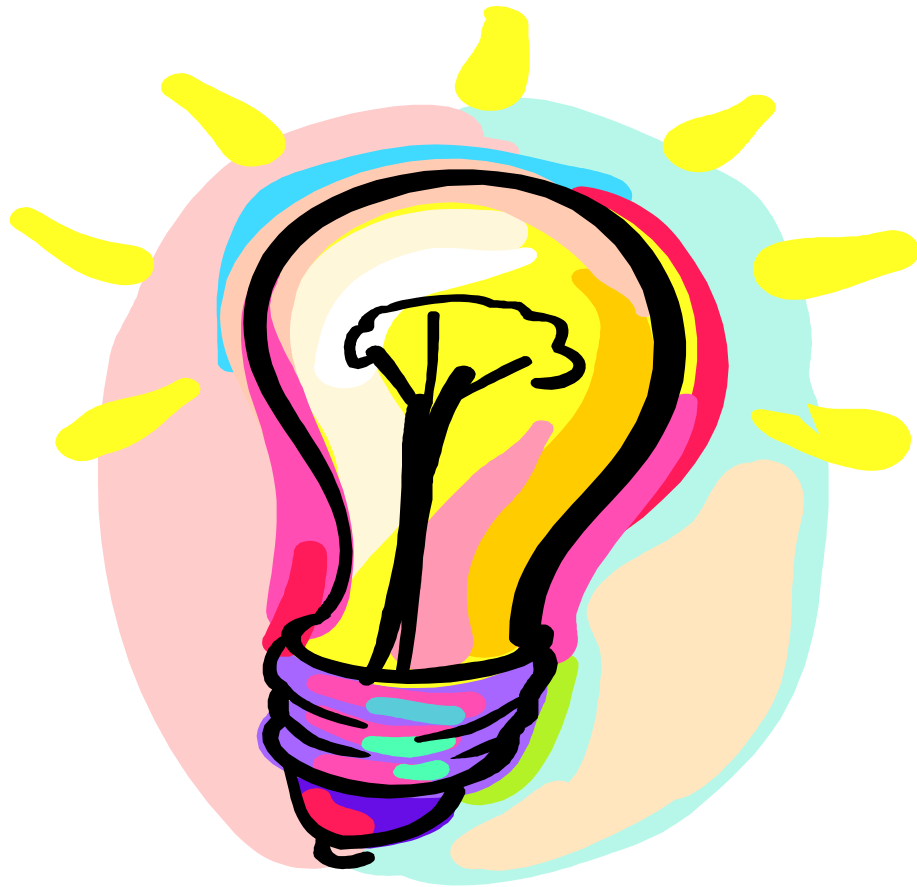
SPSS (originally, Statistical Package for the Social Sciences) is a computer program used for survey authoring and deployment (**IBM SPSS Data Collection**), data mining (**IBM SPSS Modeler**), text analytics, statistical analysis, and collaboration and deployment (batch and automated scoring services).

Assignments

1. Discuss the motivation behind *OLAP mining (OLAM)*.
2. In data warehouse technology, a multiple dimensional view can be implemented by a relational database technique (*ROLAP*), or by a multidimensional database technique (*MOLAP*), or by a hybrid database technique (*HOLAP*).
 - (a) Briefly describe each implementation technique.
 - (b) For each technique, explain how each of the following functions may be implemented:
 - i. The generation of a data warehouse (including aggregation)
 - ii. Roll-up
 - iii. Drill-down
 - iv. Incremental updating

Which implementation techniques do you prefer, and why?

Questions?



References

1. D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
2. A. Shoshani. OLAP and statistical databases: Similarities and differences. PODS'00.
3. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
4. C. Imhoff, N. Galemme, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
5. W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
6. R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
7. R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97

End of Unit 5





Thank you !!!