UNIT-V(CHAPTER-I)

The Data Warehouse

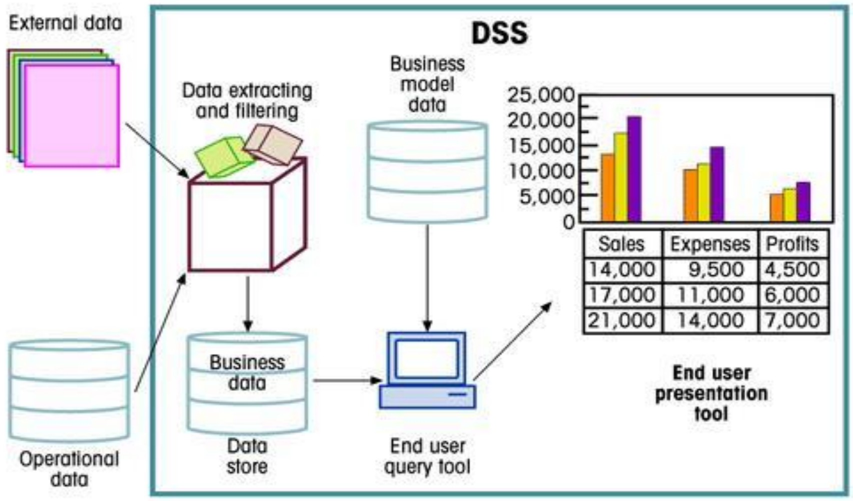
(I) The Need for Data Analysis:

1. Constant pressure from external and internal forces requires prompt tactical and strategic decisions.
2. The decision-making cycle time is reduced, while problems are increasingly complex with a growing number of internal and external variables.
3. Managers need support systems for facilitating quick decision making in a complex environment.
4. Decision support systems (DSS).

(II) Decision Support Systems:

1. Decision Support is a methodology (or a series of methodologies) designed to extract information from data and to use such information as a basis for decision making.
2. A decision support system (DSS) is an arrangement of computerized tools used to assist managerial decision making within a business.
    a. A DSS usually requires extensive data "massaging" to produce information.
    b. The DSS is used at all levels within an organization and is often tailored to focus on specific business areas or problems.
    c. The DSS is interactive and provides ad hoc query tools to retrieve data and to display data in different formats.
3. Four Components of a DSS
    a) The data store component is basically a DSS database.
    b) The data extraction and filtering component is used to extract and validate the data taken from the operational database and the external data sources.
    c) The end user query tool is used by the data analyst to create the queries that access the database.
    d) The end user presentation tool is used by the data analyst to organize and present the data.

Main Components Of A Decision Support System (DSS)



Sales	Expenses	Profits
14,000	9,500	4,500
17,000	11,000	6,000
21,000	14,000	7,000

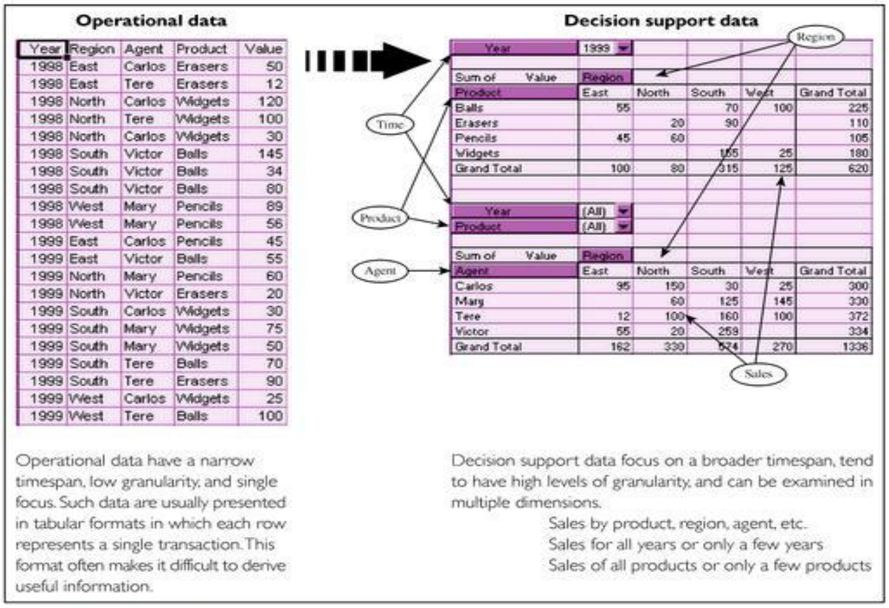
FIGURE 13.1 MAIN COMPONENTS OF A DECISION SUPPORT SYSTEM (DSS)

Operational Data vs. Decision Support Data:

a) Most operational data are stored in a relational database in which the structures tend to be highly normalized.
b) The operational data storage is optimized to support transactions that represent daily operations.
c) Whereas operational data capture daily business transactions, DSS data give tactical and strategic business meaning to the operational data.

Three Main Areas in Which DSS Data Differ from Operational Data:

a. **Time span**
    i. Operational data represent current (atomic) transactions.
    ii. DSS data tend to cover a longer time frame.
b. **Granularity**
    i. Operational data represent specific transactions that occur at a given time.
    ii. DSS data must be presented at different levels of aggregation.
c. **Dimensionality**
    i. **Operational data focus on representing atomic transactions.**
    ii. **DSS data can be analyzed from multiple dimensions.**



Operational data

Year	Region	Agent	Product	Value
1998	East	Carlos	Erasers	50
1998	East	Tere	Erasers	12
1998	North	Carlos	Widgets	120
1998	North	Tere	Widgets	100
1998	North	Carlos	Widgets	30
1998	South	Victor	Balls	145
1998	South	Victor	Balls	34
1998	South	Victor	Balls	80
1998	West	Mary	Pencils	89
1998	West	Mary	Pencils	56
1999	East	Carlos	Pencils	45
1999	East	Victor	Balls	55
1999	North	Mary	Pencils	60
1999	North	Victor	Erasers	20
1999	South	Carlos	Widgets	30
1999	South	Mary	Widgets	75
1999	South	Mary	Widgets	50
1999	South	Tere	Balls	70
1999	South	Tere	Erasers	90
1999	West	Carlos	Widgets	25
1999	West	Tere	Balls	100

Operational data have a narrow timespan, low granularity, and single focus. Such data are usually presented in tabular formats in which each row represents a single transaction. This format often makes it difficult to derive useful information.

Decision support data

Year	1999				
Sum of Value	Region				
Product	East	North	South	West	Grand Total
Balls	55		70	100	225
Erasers		20	90		110
Pencils	45	60			105
Widgets			155	25	180
Grand Total	100	80	315	125	620

Year	(All)				
Product	(All)				
Sum of Value	Region				
Agent	East	North	South	West	Grand Total
Carlos	95	150	30	25	300
Mary		60	125	145	330
Tere	12	100	160	100	372
Victor	55	20	259		334
Grand Total	162	330	574	270	1336

Decision support data focus on a broader timespan, tend to have high levels of granularity, and can be examined in multiple dimensions.

Sales by product, region, agent, etc.
Sales for all years or only a few years
Sales of all products or only a few products

FIGURE 13.2 TRANSFORMING OPERATIONAL DATA INTO DECISION SUPPORT DATA

<u>The DSS Database Requirements:</u>

(1)Database Schema:

a) The DSS database schema must support complex (non-normalized) data representations.
b) The queries must be able to extract multidimensional time slices.

Ex: Ten Year Sales History For A Single Department,Millions Of Dollars

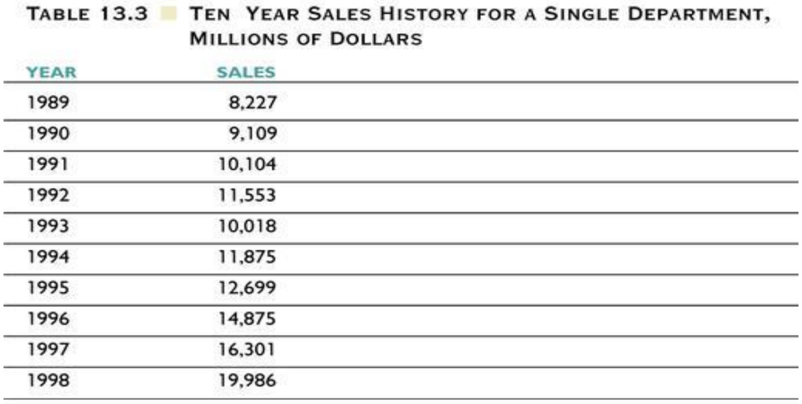


TABLE 13.3 TEN YEAR SALES HISTORY FOR A SINGLE DEPARTMENT, MILLIONS OF DOLLARS

YEAR	SALES
1989	8,227
1990	9,109
1991	10,104
1992	11,553
1993	10,018
1994	11,875
1995	12,699
1996	14,875
1997	16,301
1998	19,986

Yearly Sales Summaries, Two Stores and Two DepartmentsPer Store, Millions Of Dollars

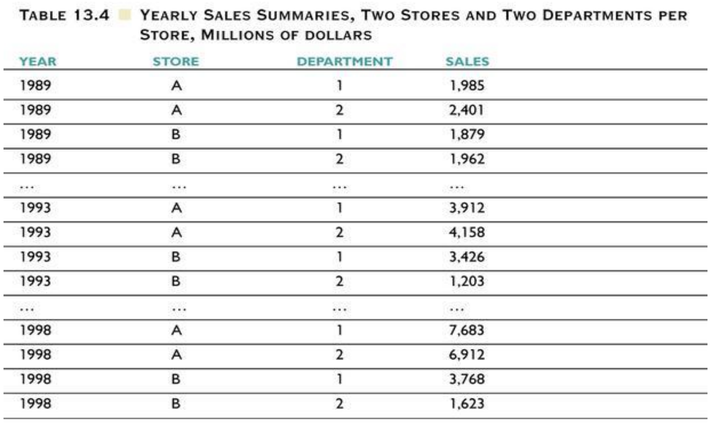


TABLE 13.4 YEARLY SALES SUMMARIES, TWO STORES AND TWO DEPARTMENTS PER STORE, MILLIONS OF DOLLARS

YEAR	STORE	DEPARTMENT	SALES
1989	A	1	1,985
1989	A	2	2,401
1989	B	1	1,879
1989	B	2	1,962
...	...	...	...
1993	A	1	3,912
1993	A	2	4,158
1993	B	1	3,426
1993	B	2	1,203
...	...	...	...
1998	A	1	7,683
1998	A	2	6,912
1998	B	1	3,768
1998	B	2	1,623

(2)Data Extraction and Loading :

a) The DBMS must support advanced data extracting and filtering tools.
b) The data extraction capabilities should support different data sources and multiple vendors.
c) Data filtering capabilities must include the ability to check for inconsistent data or data validation rules.
d) The DBMS must support advanced data integration, aggregation, and classification capabilities.

Ex: Yearly Sales Summaries, 20 Stores, 10 Departments Per Store,Millions Of Dollars

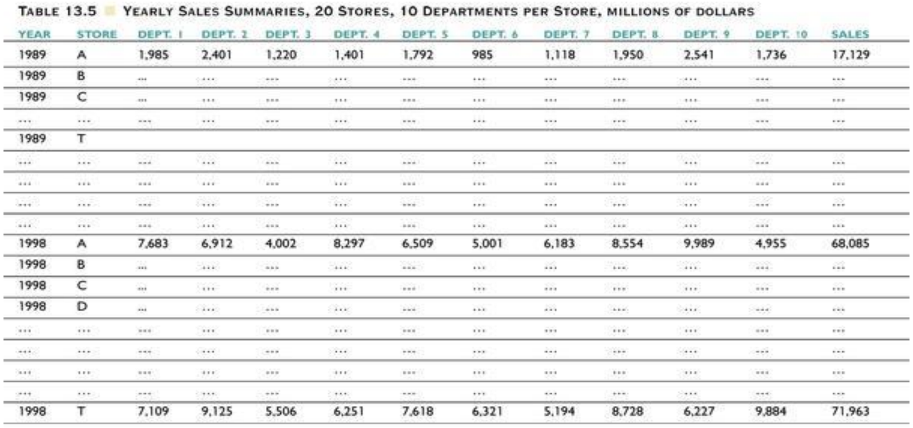


TABLE 13.5 YEARLY SALES SUMMARIES, 20 STORES, 10 DEPARTMENTS PER STORE, MILLIONS OF DOLLARS

YEAR	STORE	DEPT. 1	DEPT. 2	DEPT. 3	DEPT. 4	DEPT. 5	DEPT. 6	DEPT. 7	DEPT. 8	DEPT. 9	DEPT. 10	SALES
1989	A	1,985	2,401	1,220	1,401	1,792	985	1,118	1,950	2,541	1,736	17,129
1989	B	...	...	...	...	...	...	...	...	...	...	...
1989	C	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
1989	T											
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
1998	A	7,683	6,912	4,002	8,297	6,509	5,001	6,183	8,554	9,989	4,955	68,085
1998	B	...	...	...	...	...	...	...	...	...	...	...
1998	C	...	...	...	...	...	...	...	...	...	...	...
1998	D	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
1998	T	7,109	9,125	5,506	6,251	7,618	6,321	5,194	8,728	6,227	9,884	71,963

(3)End-User Analytical Interface:

a) The DSS DBMS must support advanced data modeling and data presentation tools, data analysis tools, and query generation and optimization components.
b) The end user analytical interface is one of the most critical components.

(4)Database Size Requirements

a) DSS databases tend to be very large.

b) The DBMS must be capable of supporting very large databases (VLDB).

c) The DBMS may be required to use advanced hardware, such as multiple disk arrays and multiple-processor technologies.

(III) <u>The Data Warehouse:</u>

The Data Warehouse is an integrated, subject-oriented, time-variant, non-volatile database that provides support for decision making.

a. Integrated

i. The Data Warehouse is a centralized, consolidated database that integrates data retrieved from the entire organization.

b. Subject-Oriented

i. The Data Warehouse data is arranged and optimized to provide answers to questions coming from diverse functional areas within a company.

c. Time Variant

i. **The Warehouse data represent the flow of data through time. It can even contain projected data.**

d. Non-Volatile

i. **Once data enter the Data Warehouse, they are never removed.**

ii. **The Data Warehouse is always growing.**

Creating A Data Warehouse

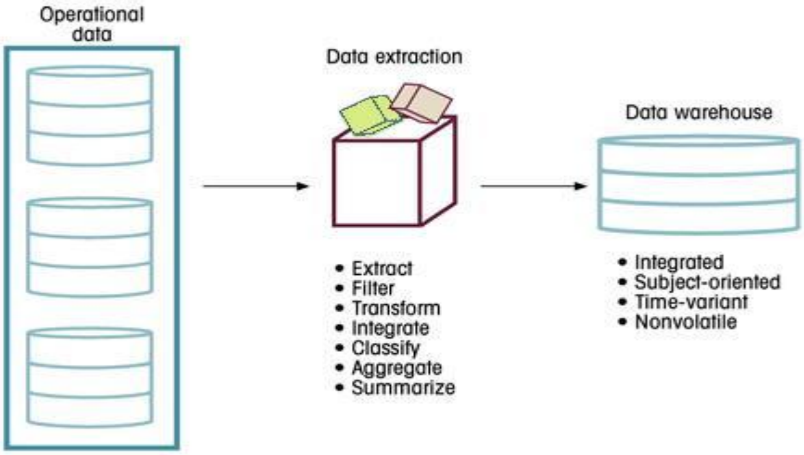


FIGURE 13.3 CREATING A DATA WAREHOUSE

Data Mart:

1. A data mart is a small, single-subject data warehouse subset that provides decision support to a small group of people.
2. Data Marts can serve as a test vehicle for companies exploring the potential benefits of Data Warehouses.
3. Data Marts address local or departmental problems, while a Data Warehouse involves a company-wide effort to support decision making at all levels in the organization.

Twelve Rules That Define a Data Warehouse:

1. The Data Warehouse and operational environments are separated.

2. The Data Warehouse data are integrated.

3. The Data Warehouse contains historical data over a long time horizon.

4. The Data Warehouse data are snapshot data captured at a given point in time.

5. The Data Warehouse data are subject-oriented.

6. The Data Warehouse data are mainly read-only with periodic batch updates from operational data. No online updates are allowed.

7. The Data Warehouse development life cycle differs from classical systems development. The Data Warehouse development is data driven; the classical approach is process driven.

8. The Data Warehouse contains data with several levels of detail; current detail data, old detail data, lightly summarized, and highly summarized data.

9. The Data Warehouse environment is characterized by read-only transactions to very large data sets. The operational environment is characterized by numerous update transactions to a few data entities at the time.

10. The Data Warehouse environment has a system that traces data resources, transformation, and storage.

11. The Data Warehouse's metadata are a critical component of this environment. The metadata identify and define all data elements. The metadata provide the source, transformation, integration, storage, usage, relationships, and history of each data element.

12. The Data Warehouse contains a charge-back mechanism for resource usage that enforces optimal use of the data by end users.

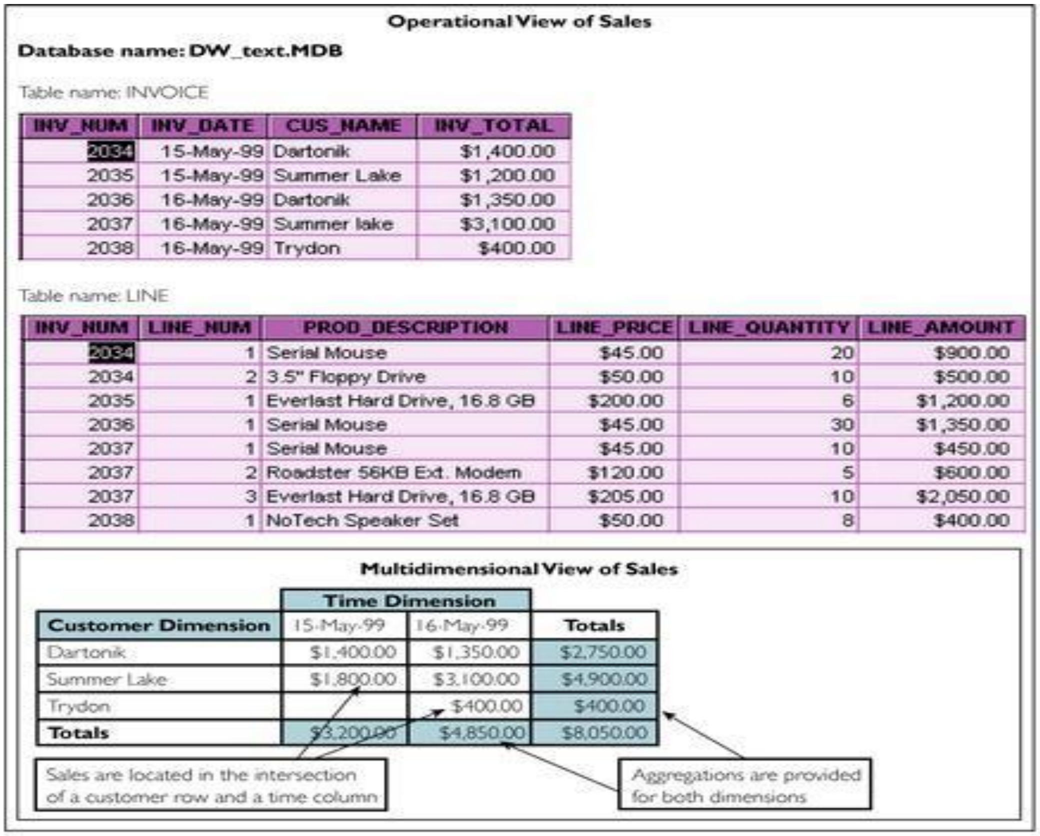
(IV) On-Line Analytical Processing:

1. On-Line Analytical Processing (OLAP) is an advanced data analysis environment that supports decision making, business modeling, and operations research activities.

2. Four Main Characteristics of OLAP
    i. Use multidimensional data analysis techniques
    ii. Provide advanced database support
    iii. Provide easy-to-use end user interfaces
    iv. Support client/server architecture

(i) **Multidimensional Data Analysis Techniques**

a) The processing of data in which data are viewed as part of a multidimensional structure.
b) Multidimensional view allows end users to consolidate or aggregate data at different levels.
c) Multidimensional view allows a business analyst to easily switch business perspectives.



Operational View of Sales

Database name: DW_text.MDB

Table name: INVOICE

INV_NUM	INV_DATE	CUS_NAME	INV_TOTAL
2034	15-May-99	Dartonik	$1,400.00
2035	15-May-99	Summer Lake	$1,200.00
2036	16-May-99	Dartonik	$1,350.00
2037	16-May-99	Summer lake	$3,100.00
2038	16-May-99	Trydon	$400.00

Table name: LINE

INV_NUM	LINE_NUM	PROD_DESCRIPTION	LINE_PRICE	LINE_QUANTITY	LINE_AMOUNT
2034	1	Serial Mouse	$45.00	20	$900.00
2034	2	3.5" Floppy Drive	$50.00	10	$500.00
2035	1	Everlast Hard Drive, 16.8 GB	$200.00	6	$1,200.00
2036	1	Serial Mouse	$45.00	30	$1,350.00
2037	1	Serial Mouse	$45.00	10	$450.00
2037	2	Roadster 56KB Ext. Modem	$120.00	5	$600.00
2037	3	Everlast Hard Drive, 16.8 GB	$205.00	10	$2,050.00
2038	1	NoTech Speaker Set	$50.00	8	$400.00

Multidimensional View of Sales

	Time Dimension		
Customer Dimension	15-May-99	16-May-99	Totals
Dartonik	$1,400.00	$1,350.00	$2,750.00
Summer Lake	$1,800.00	$3,100.00	$4,900.00
Trydon		$400.00	$400.00
Totals	$3,200.00	$4,850.00	$8,050.00

Sales are located in the intersection of a customer row and a time column

Aggregations are provided for both dimensions

FIGURE 13.4 OPERATIONAL VS. MULTIDIMENSIONAL VIEW OF SALES

(ii) **Additional Functions of Multidimensional Data Analysis Techniques**

a) Advanced data presentation functions
b) Advanced data aggregation, consolidation, and classification functions
c) Advanced computational functions
d) Advanced data modeling functions

(iii) **Advanced Database Support**

a) **Access to many different kinds of DBMSs, flat files, and internal and external data sources.**
b) **Access to aggregated Data Warehouse data as well as to the detail data found in operational databases.**
c) **Advanced data navigation features such as drill-down and roll-up.**
d) **Rapid and consistent query response times.**
e) **The ability to map end user requests, expressed in either business or model terms, to the appropriate data source and then to the proper data access language.**
f) **Support for very large databases.**

(iv) **Easy-to-Use End User Interface**

a) Easy-to-use graphical user interfaces make sophisticated data extraction and analysis tools easily accepted and readily used.

(v) **Client/Server Architecture**

a) The client/server environment enables us to divide an OLAP system into several components that define its architecture.

(V)OLAP Architecture :

Three Main Modules

1. OLAP Graphical User Interface (GUI)
2. OLAP Analytical Processing Logic
3. OLAP Data Processing Logic

OLAP systems are designed to use both operational and Data Warehouse data.

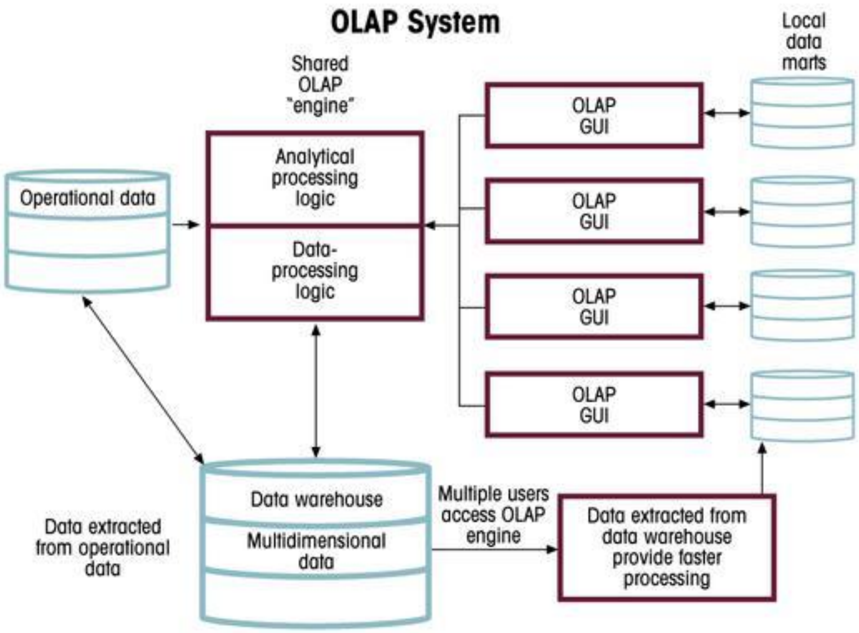

OLAP System
Shared OLAP "engine"
Local data marts
Analytical processing logic
Data-processing logic
Operational data
OLAP GUI
OLAP GUI
OLAP GUI
OLAP GUI
Data warehouse
Multidimensional data
Data extracted from operational data
Multiple users access OLAP engine
Data extracted from data warehouse provide faster processing

FIGURE 13.9 OLAP SERVER WITH LOCAL MINI-DATA-MARTS

(VI) Multidimensional OLAP (MOLAP):

MOLAP extends OLAP functionality to multidimensional databases (MDBMS).

MDBMS end users visualize the stored data as a multidimensional cube known as a data cube.

Data cubes are created by extracting data from the operational databases or from the data warehouse.

Data cubes are static and require front-end design work.

To speed data access, data cubes are normally held in memory, called cube cache.

MOLAP is generally faster than their ROLAP counterparts. It is also more resource-intensive.

MDBMS is best suited for small and medium data sets.

Multidimensional data analysis is also affected by how the database system handles sparsity.

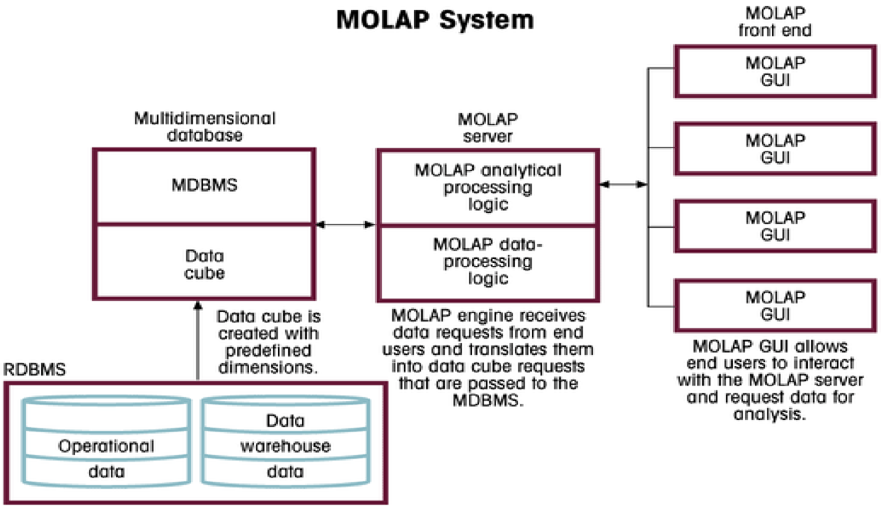

MOLAP System
MOLAP front end
Multidimensional database
MDBMS
Data cube
MOLAP server
MOLAP analytical processing logic
MOLAP data-processing logic
MOLAP GUI
MOLAP GUI
MOLAP GUI
MOLAP GUI
Data cube is created with predefined dimensions.
MOLAP engine receives data requests from end users and translates them into data cube requests that are passed to the MDBMS.
MOLAP GUI allows end users to interact with the MOLAP server and request data for analysis.
RDBMS
Operational data
Data warehouse data

FIGURE 13.11 MOLAP CLIENT/SERVER ARCHITECTURE

(VII) Star Schema:

1. **The star schema is a data-modeling technique used to map multidimensional decision support into a relational database.**
2. **Star schemas yield an easily implemented model for multidimensional data analysis while still preserving the relational structure of the operational database.**
3. **Four Components:**

a) **Facts**
b) **Dimensions**
c) **Attributes**
d) **Attribute hierarchies**

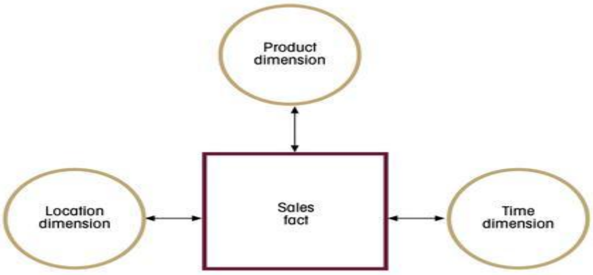


FIGURE 13.12 A SIMPLE STAR SCHEMA

(a) **Facts**

- **Facts are numeric measurements (values) that represent a specific business aspect or activity.**
- **The fact table contains facts that are linked through their dimensions.**
- **Facts can be computed or derived at run-time (metrics).**

(b) **Dimensions**

- **Dimensions are qualifying characteristics that provide additional perspectives to a given fact.**
- **Dimensions are stored in dimension tables.**

(c) **Attributes**

- **Each dimension table contains attributes. Attributes are often used to search, filter, or classify facts.**
- **Dimensions provide descriptive characteristics about the facts through their attributes.**

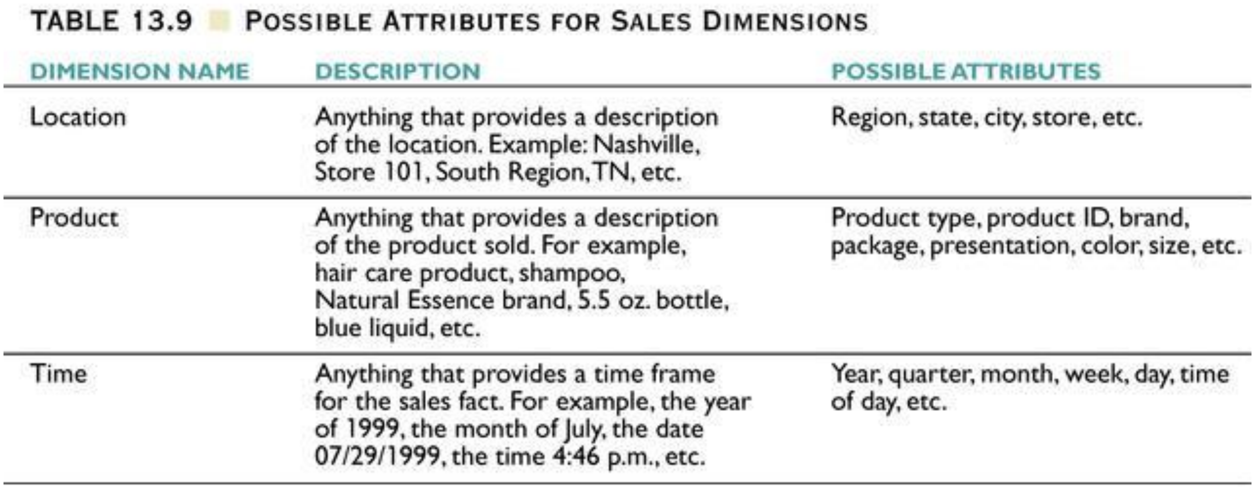
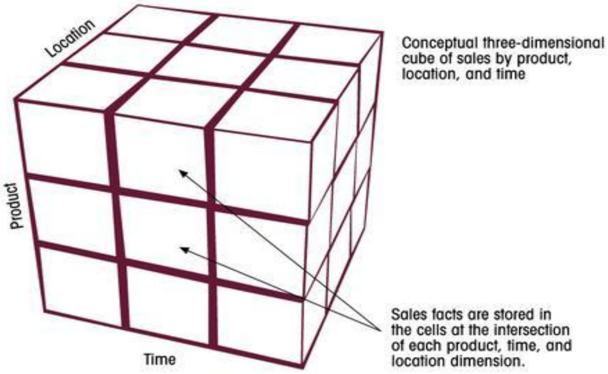


TABLE 13.9 POSSIBLE ATTRIBUTES FOR SALES DIMENSIONS

DIMENSION NAME	DESCRIPTION	POSSIBLE ATTRIBUTES
Location	Anything that provides a description of the location. Example: Nashville, Store 101, South Region, TN, etc.	Region, state, city, store, etc.
Product	Anything that provides a description of the product sold. For example, hair care product, shampoo, Natural Essence brand, 5.5 oz. bottle, blue liquid, etc.	Product type, product ID, brand, package, presentation, color, size, etc.
Time	Anything that provides a time frame for the sales fact. For example, the year of 1999, the month of July, the date 07/29/1999, the time 4:46 p.m., etc.	Year, quarter, month, week, day, time of day, etc.



Conceptual three-dimensional cube of sales by product, location, and time

Sales facts are stored in the cells at the intersection of each product, time, and location dimension.

FIGURE 13.13 THREE-DIMENSIONAL VIEW OF SALES

(d)Attribute Hierarchies

- **Attributes within dimensions can be ordered in a well-defined attribute hierarchy.**

- **The attribute hierarchy provides a top-down data organization that is used for two main purposes:**

  i. **Aggregation**

  ii. **Drill-down/roll-up data analysis**

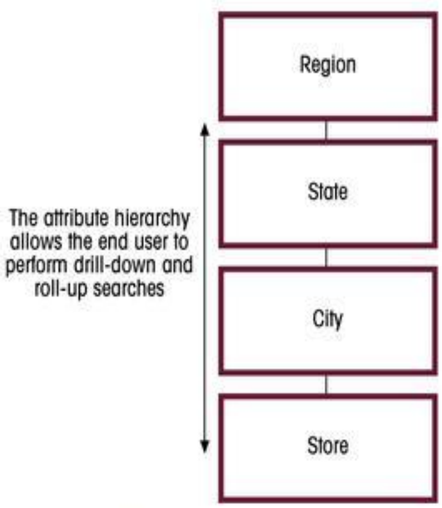

Region
State
City
Store
The attribute hierarchy allows the end user to perform drill-down and roll-up searches

FIGURE 13.15 A LOCATION ATTRIBUTE HIERARCHY

<u>Star Schema Representation:</u>

- **Facts and dimensions are normally represented by physical tables in the data warehouse database.**
- **The fact table is related to each dimension table in a many-to-one (M:1) relationship.**
- **Fact and dimension tables are related by foreign keys and are subject to the primary/foreign key constraints.**

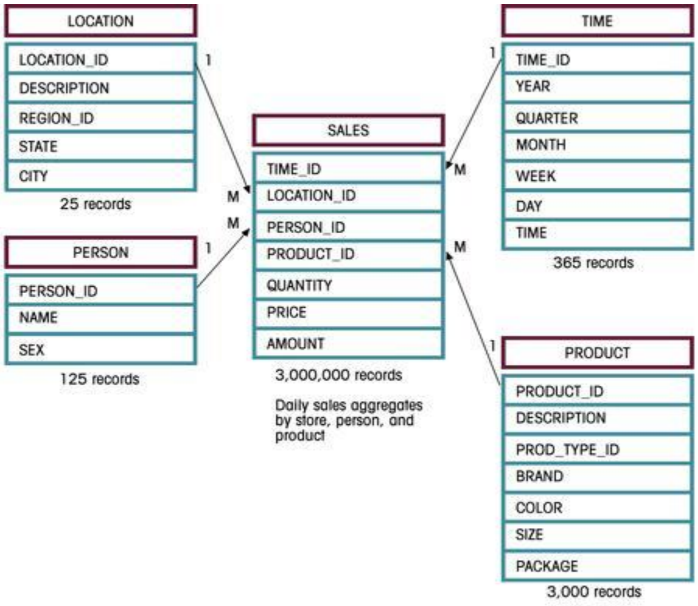


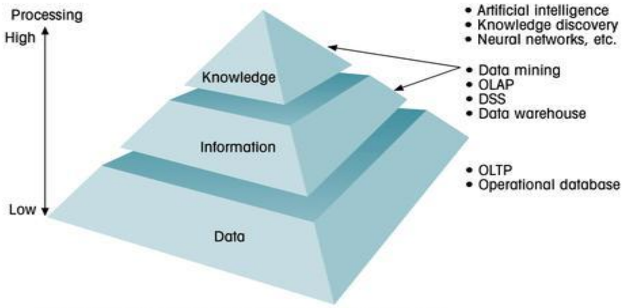
FIGURE 13.17 STAR SCHEMA FOR SALES

Performance-Improving Techniques:

- **Normalization of dimensional tables**
- **Multiple fact tables representing different aggregation levels**
- **Denormalization of fact tables**
- **Table partitioning and replication**

(VIII) Data Mining:

- In contrast to the traditional (reactive) DSS tools, the data mining premise is proactive.
- Data mining tools automatically search the data for anomalies and possible relationships, thereby identifying problems that have not yet been identified by the end user.
- Data mining tools -- based on algorithms that form the building blocks for artificial intelligence, neural networks, inductive rules, and predicate logic -- initiate analysis to create knowledge.



Data-mining tools use advanced techniques from knowledge discovery, artificial intelligence, and other fields to obtain "knowledge" and apply it to business needs. Knowledge is then used to make predictions of events or forecasts of values such as sales returns, etc. Several OLAP tools have integrated at least some of these data-mining features in their products.

FIGURE 13.22 EXTRACTION OF KNOWLEDGE FROM DATA

Four Phases of Data Mining:

1. Data Preparation

- **Identify and cleanse data sets.**
- **Data Warehouse is usually used for data mining operations.**

2. Data Analysis and Classification

- **Identify common data characteristics or patterns using**
  - **Data groupings, classifications, clusters, or sequences.**
  - **Data dependencies, links, or relationships.**
  - **Data patterns, trends, and deviations.**

3. Knowledge Acquisition

- **Select the appropriate modeling or knowledge acquisition algorithms.**
- **Examples: neural networks, decision trees, rules induction, genetic algorithms, classification and regression tree, memory-based reasoning, or nearest neighbor and data visualization.**

4. Prognosis

- **Predict future behavior and forecast business outcomes using the data mining findings.**

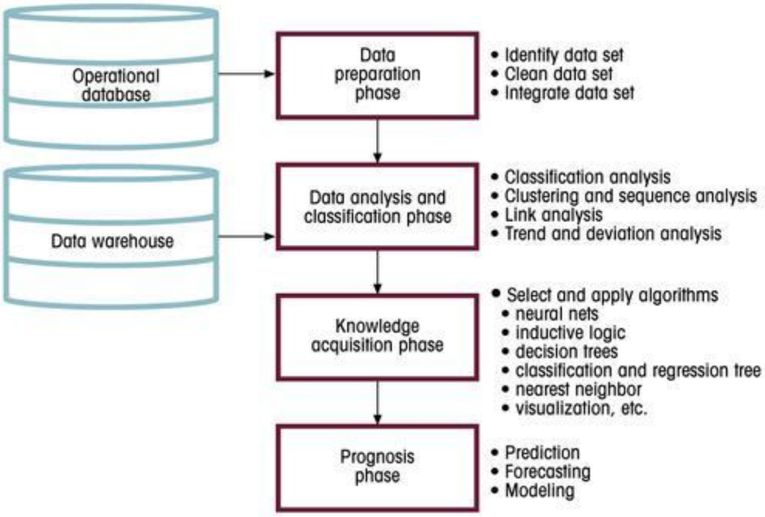


FIGURE 13.23 DATA-MINING PHASES