

DATA WAREHOUSING AND DATA MINING

Lecture-2

Sajid Majeed

Why a Data Warehouse (DWH)?

2

- Data recording and storage is growing.
- History is excellent predictor of the future.
- Gives total view of the organization.
- Intelligent decision-support is required for decision-making.

Reason-1: Why a Data Warehouse?

- Size of Data Sets are going up .
- Cost of data storage is coming down .
 - The amount of data average business collects and stores is **doubling every year**
- Total hardware and software cost to store and manage 1 Mbyte of data
 - 1990: ~ \$15
 - 2002: ~ ¢15 (Down 100 times)
 - By 2007: < ¢1 (Down 150 times)

- Businesses demand Intelligence (BI).
 - Complex questions from integrated data.
 - “Intelligent Enterprise”

DBMS Approach

List of all items that were sold last month?

List of all items purchased by Sana?

The total sales of the last month grouped by branch?

How many sales transactions occurred during the month of January?

Intelligent Enterprise

Which items sell together? Which items to stock?

**Where and how to place the items?
What discounts to offer?**


How best to target customers to increase sales at a branch?

Which customers are most likely to respond to my next promotional campaign, and why?

Businesses want much more...

- What happened?
- Why it happened?
- What will happen?
- What is happening?
- What do you want to happen?

Stages of
Data
Warehouse



A complete repository of historical corporate data extracted from transaction systems that is available for ad-hoc access by knowledge workers.

Complete repository

- All the data is present from all the branches/outlets of the business.
- Even the archived data may be brought online.
- Data from arcane and old systems is also brought online.

Transaction System

- Management Information System (MIS)
- Could be typed sheets (NOT transaction system)

Ad-Hoc access

- Dose not have a certain access pattern.
- Queries not known in advance.
- Difficult to write SQL in advance.

Knowledge workers

- Typically NOT IT literate (Executives, Analysts, Managers).
- NOT clerical workers.
- Decision makers.

A data warehouse is a *subject-oriented, integrated, time-variant* and *non-volatile* collection of data in support of management's decision making process.

Subject oriented:

A data warehouse can be used to analyze a particular subject area. For example, sales, customer, products can be a particular subject.

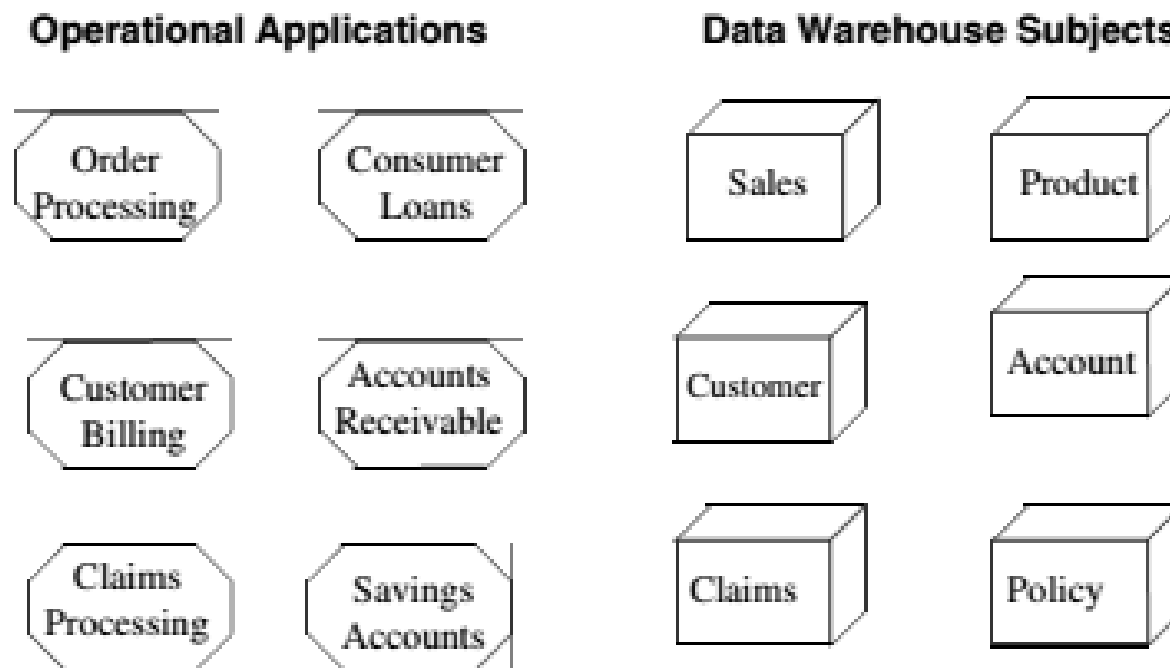


Figure 2-1 The data warehouse is subject oriented.

Integrated:

- loaded from different sources that store the data in different formats.
- checked, cleansed and transformed into a unified format to allow easy and fast access.

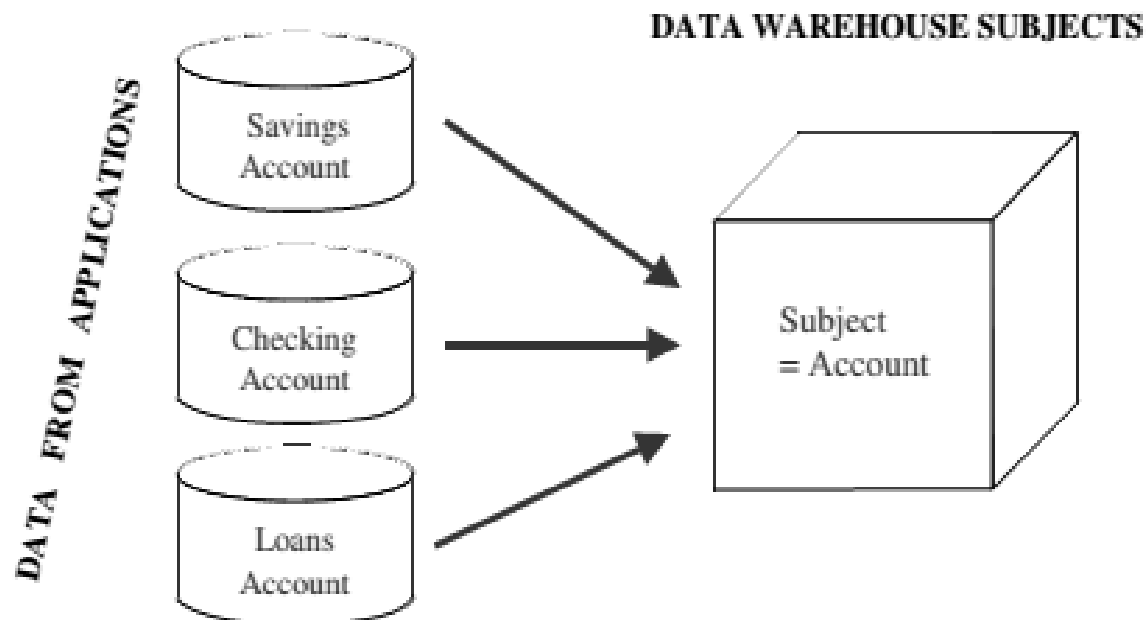


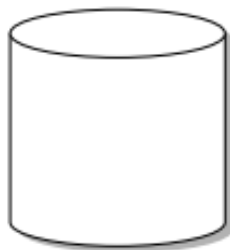
Figure 2-2 The data warehouse is integrated.

Time variant:

- Historical data is kept in a data warehouse.
- records that are created as of some moment in time.
- data warehouse as the data is loaded; the moment becomes its time stamp.

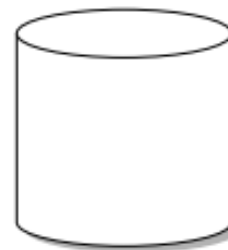
TIME VARIANCY

Operational



- Time horizon – current to 60–90 days
- Update of records
- Key structure may or may not contain an element of time

Data warehouse



- Time horizon – 5–10 years
- Sophisticated snapshots of data
- Key structure contains an element of time

Non-volatile:

after insertion data is neither changed nor removed.

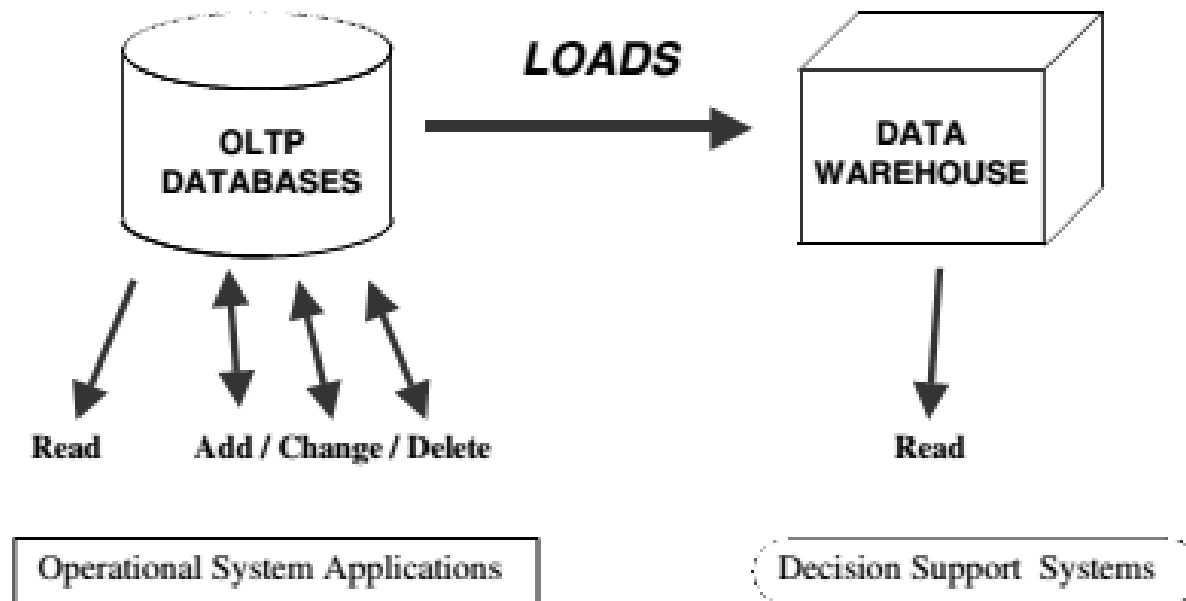
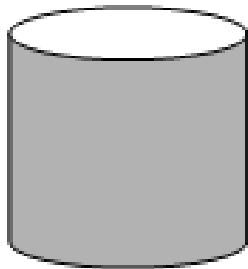


Figure 2-3 The data warehouse is nonvolatile.

Granularity

Data granularity in a data warehouse refers to the level of detail.

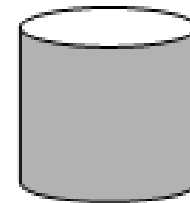


High level of detail—
low level of granularity

EXAMPLE:

The details of every
phone call made by a
customer for a month

GRANULARITY—
THE LEVEL OF DETAIL



Low level of detail—
high level of granularity

EXAMPLE:

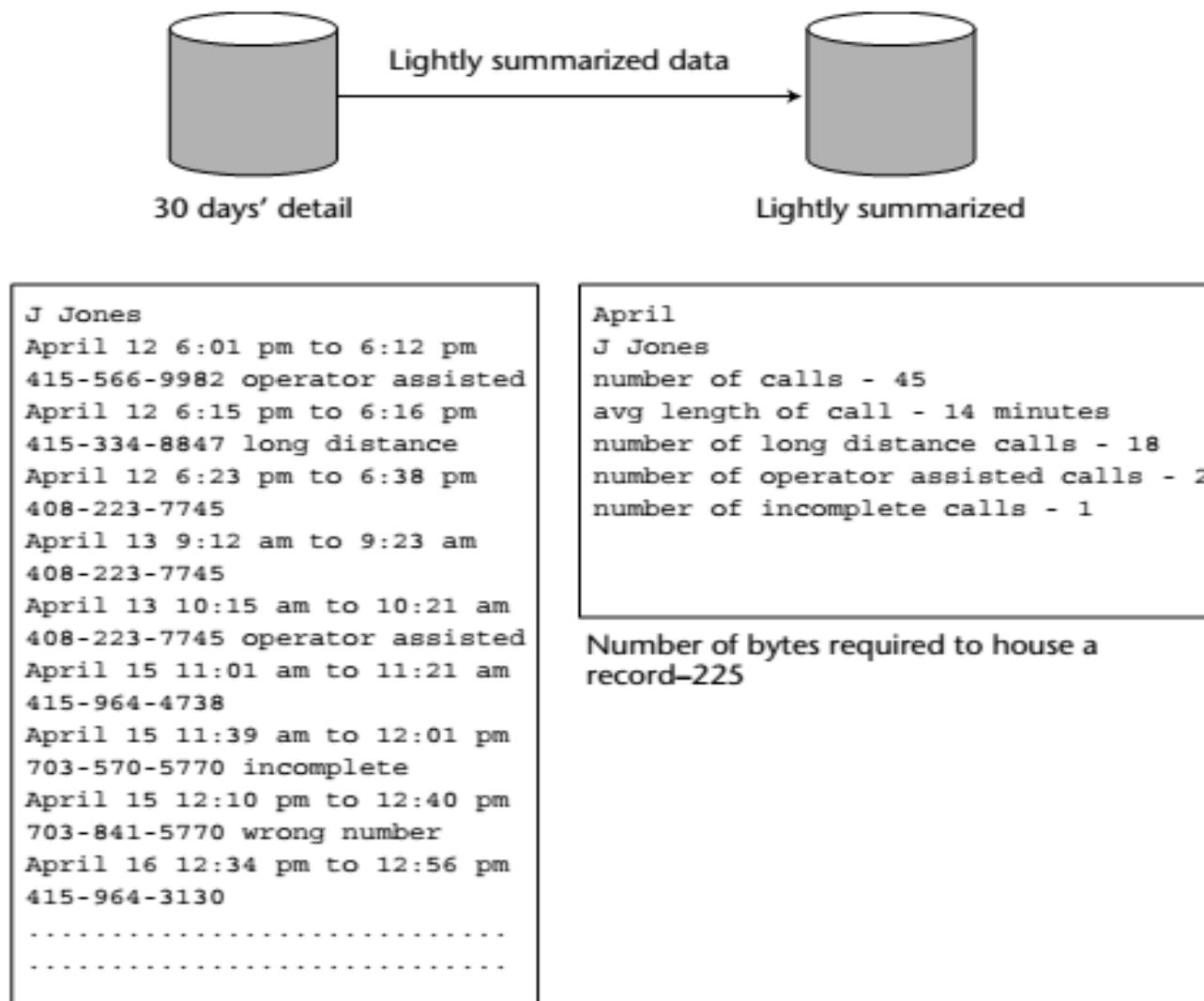
The summary of phone
calls made by a
customer for a month

THREE DATA LEVELS IN A BANKING DATA WAREHOUSE

<u>Daily Detail</u>	<u>Monthly Summary</u>	<u>Quarterly Summary</u>
Account	Account	Account
Activity Date	Month	Month
Amount	Number of transactions	Number of transactions
Deposit/Withdrawal	Withdrawals	Withdrawals
	Deposits	Deposits
	Beginning Balance	Beginning Balance
	Ending Balance	Ending Balance

Data granularity refers to the level of detail. Depending on the requirements, multiple levels of detail may be present. Many data warehouses have at least dual levels of granularity.

Figure 2-4 Data granularity.



For a single customer for a month, an average of 45,000 bytes are required to house 200 records.

Figure 2-16 With light summarization data, large quantities of data can be represented compactly.

Making the wheels of business turn

- ◆ Take an order
- ◆ Process a claim
- ◆ Make a shipment
- ◆ Generate an invoice
- ◆ Receive cash
- ◆ Reserve an airline seat

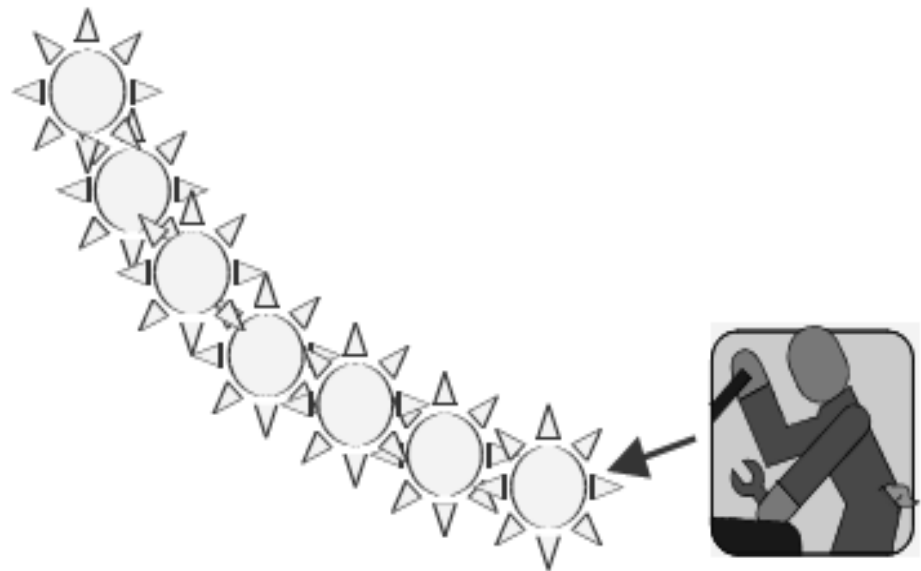


Figure 1-5 Operational systems.

Watching the wheels of business turn

- ◆ Show me the top-selling products
- ◆ Show me the problem regions
- ◆ Tell me why (drill down)
- ◆ Let me see other data (drill across)
- ◆ Show the highest margins
- ◆ Alert me when a district sells below target

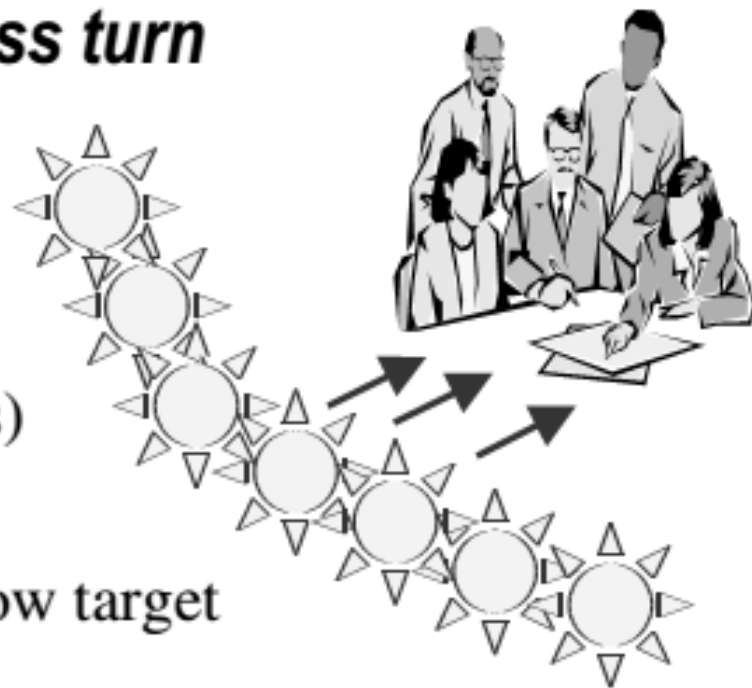


Figure 1-6 Decision-support systems.