# Data Warehousing and Data Mining

# Lecture # 3

**An environment, Not a product.**

- An ideal environment for data analysis and decision support.
- Flexible, and interactive
- 100 percent user-driven
- Very responsive and conducive to the ask–answer–ask–again pattern
- Provides the ability to discover answers to complex, unpredictable questions

A Blend of Many Technologies
The basic concept of data warehousing is:

- Take all data from different operational systems.
- If necessary, add relevant data from industry.
- Transform all data and bring into a uniform format.
- Integrate all data as a single entity.

**It is a blend of many technologies, the basic concept being:**

- Store data in a format supporting easy access for decision support.
- Create performance enhancing indices.
- Implement performance enhancement joins.
- Run ad-hoc queries with low selectivity.
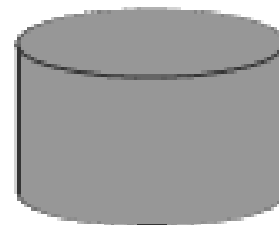
**OPERATIONAL SYSTEMS**

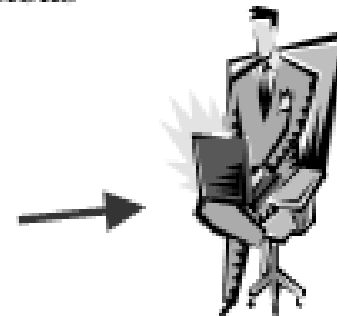Basic business processes

Extraction, cleansing, aggregation

**Data Transformation**

Key measurements, business dimensions

**DATA WAREHOUSE**

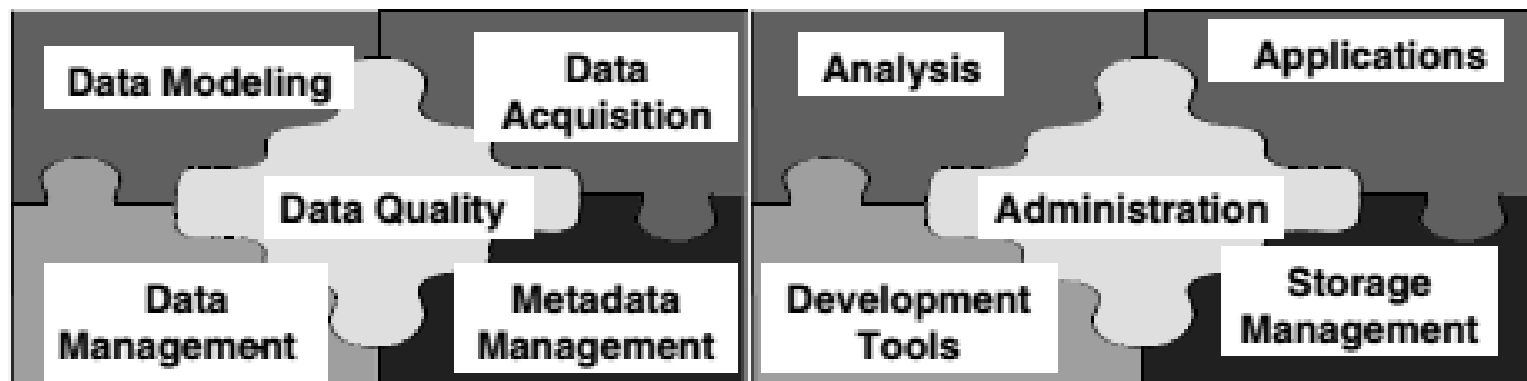Executives/Managers/ Analysts

## BLEND OF TECHNOLOGIES

Data Modeling

Data Acquisition

Data Quality

Data Management

Metadata Management

Analysis

Applications

Administration

Development Tools

Storage Management

**Figure 1-9** The data warehouse: a blend of technologies.

User needs
information

User hopes
to find the
right
answers

User requests
reports from IT

**THE FAMILIAR
MERRY-GO-ROUND
(4–6 weeks)**

IT sends
requested
reports
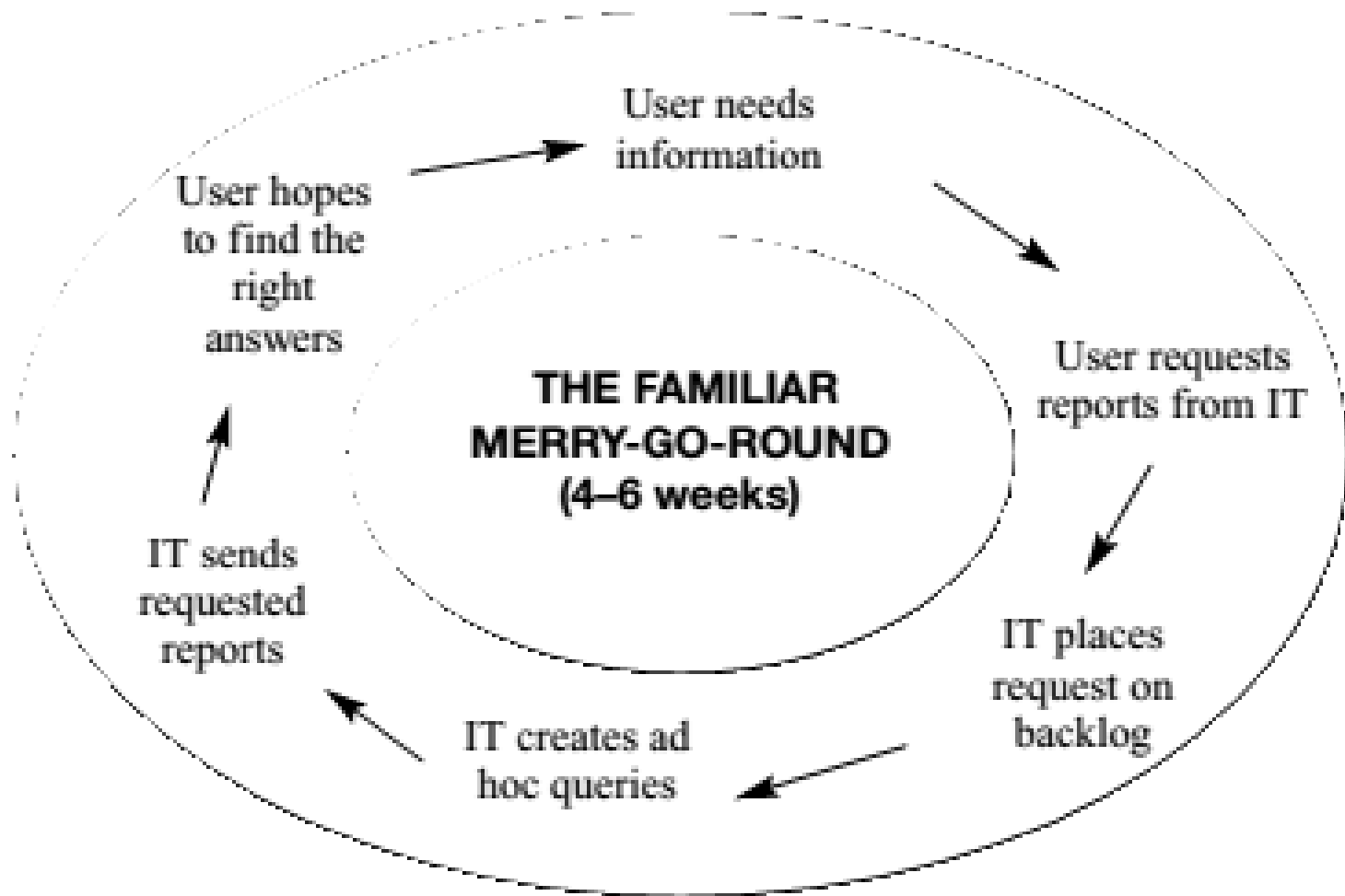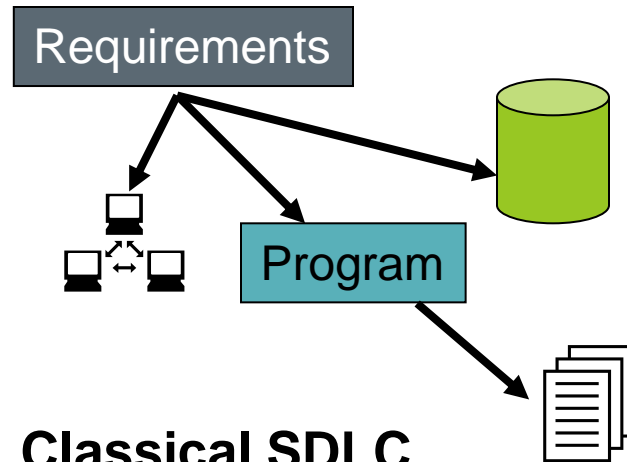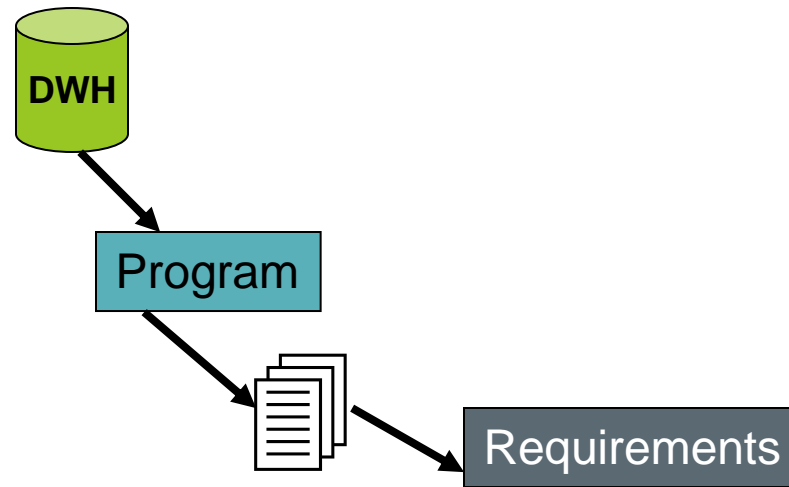
IT places
request on
backlog

IT creates ad
hoc queries

**Figure 1-4**   Inadequate attempts by IT to provide strategic information.

- Does not follows the traditional development model

Requirements

Program

**Classical SDLC**

- Requirements gathering
- Analysis
- Design
- Programming
- Testing
- Integration

# How is it Different?

- Does not follows the traditional development model



**DWH SDLC (CLDS)**

- Implement warehouse
- Integrate data
- Test for biasness
- Program w.r.t data
- Design DSS system
- Analyze results
- Understand requirement

- Combines operational and historical data.

  ▪ Don't do data entry into a DWH, OLTP or ERP are the source systems.

  ▪ OLTP systems don't keep history, cant get balance statement more than a year old.

  ▪ DWH keep historical data, even of bygone customers. Why?

  ▪ In the context of bank, want to know why the customer left?

  ▪ What were the events that led to his/her leaving? Why?

  ▪ Customer retention.

# How much history?

- Depends on:
  - Industry.

  - Cost of storing historical data.

  - Economic value of historical data.

# How much history?

- Industries and history
  - **Telecomm** calls are much much more as compared to bank transactions- 18 months.

  - **Retailers** interested in analyzing yearly seasonal patterns- 65 weeks.

  - **Insurance** companies want to do actuary analysis, use the historical data in order to predict risk- 7 years.

- Usually (but not always) periodic or batch updates rather than real-time.

▪ For an ATM, if update not in real-time, then lot of real trouble.

▪ DWH is for strategic decision making based on historical data. Wont hurt if transactions of last one hour/day are absent.

# How is it Different?

- Rate of update depends on:
  - volume of data,
  - nature of business,
  - cost of keeping historical data,
  - benefit of keeping historical data.

- Summary

- What do we mean by strategic information. For a commercial bank, name some types of strategic objectives?

- Why are operational systems not suitable for providing strategic information? Give specific reasons and explain?

- A data warehouse is an environment, not a product. Discuss?

- For an airline company, how can strategic information increase the number of frequent flyers? Discuss giving specific details.