DATA WAREHOUSING AND DATA MINING

- S. Sudarshan
- Krithi Ramamritham

- *IIT Bombay*

- sudarsha@cse.iitb.ernet.in
- krithi@cse.iitb.ernet.in

Course Overview

- The course: what and how

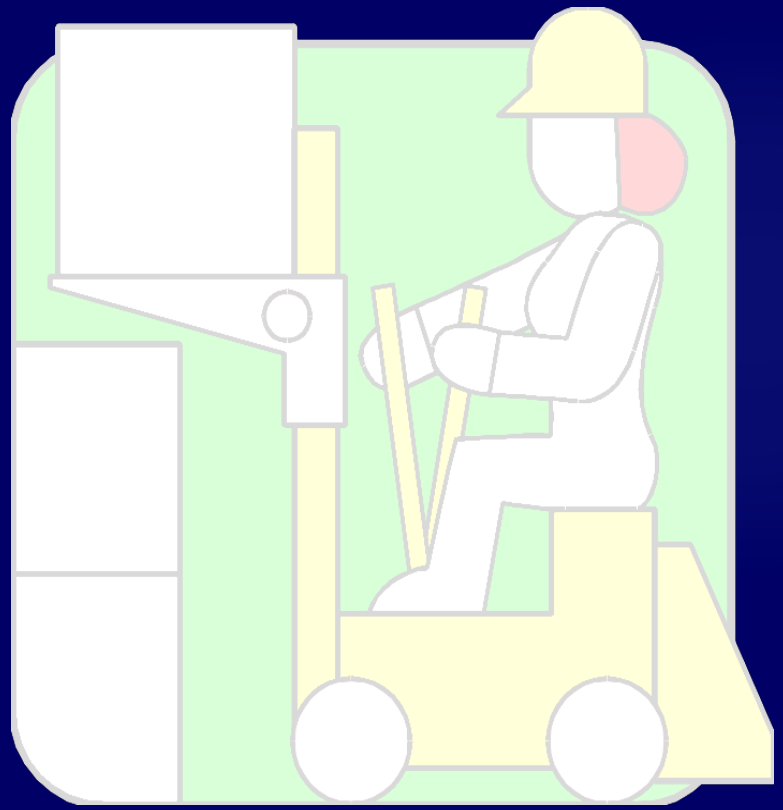
- 0. Introduction
- I. Data Warehousing
- II. Decision Support and OLAP
- III. Data Mining
- IV. Looking Ahead

- Demos and Labs

0. Introduction

- Data Warehousing, OLAP and data mining: what and why (now)?
- Relation to OLTP
- A case study

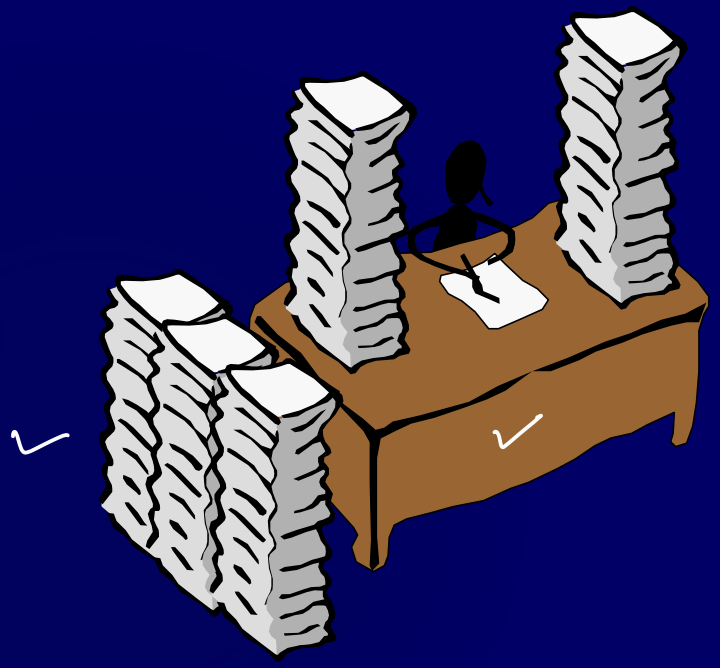
- demos, labs



A producer wants to know....

Data, Data everywhere yet …

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences

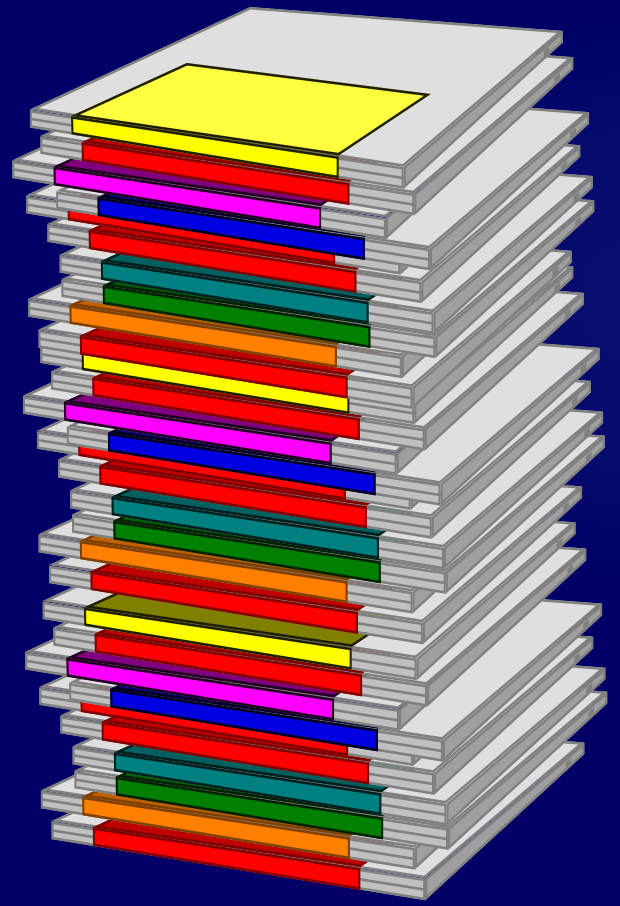


- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

What is a Data Warehouse?

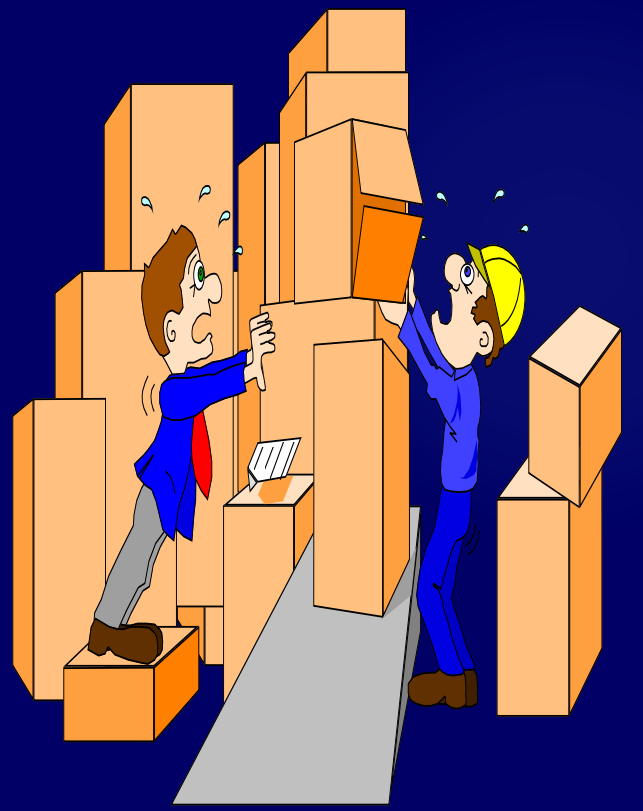
- A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

- [Barry Devlin]

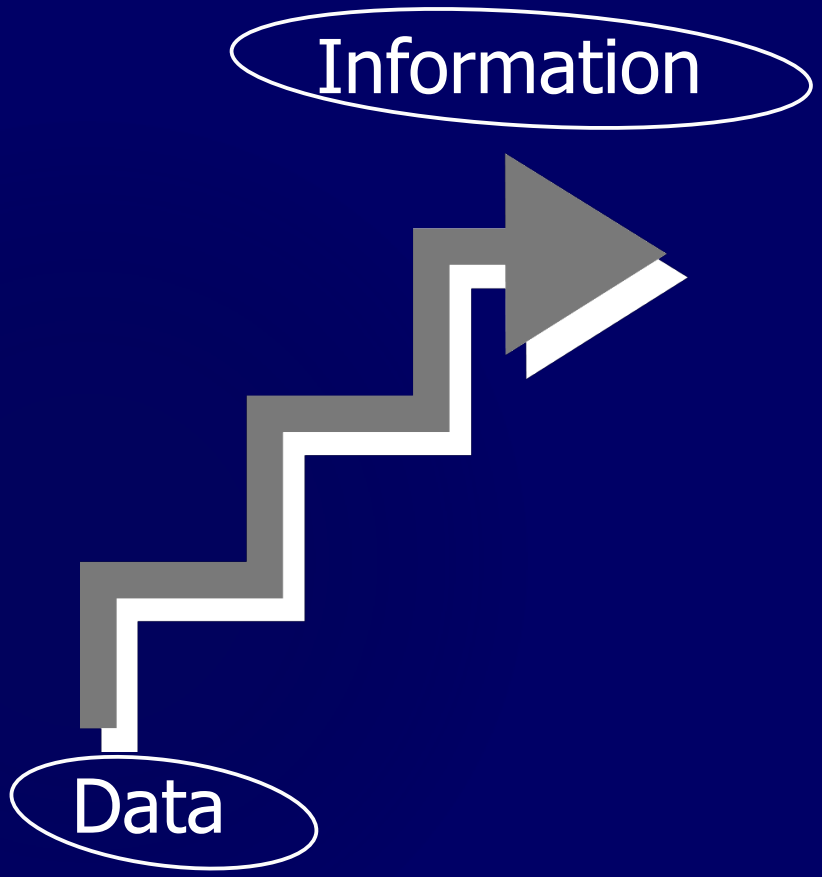


What are the users saying...

- Data should be integrated across the enterprise
- Summary data has a real value to the organization
- Historical data holds the key to understanding data over time
- What-if capabilities are required



What is Data Warehousing?



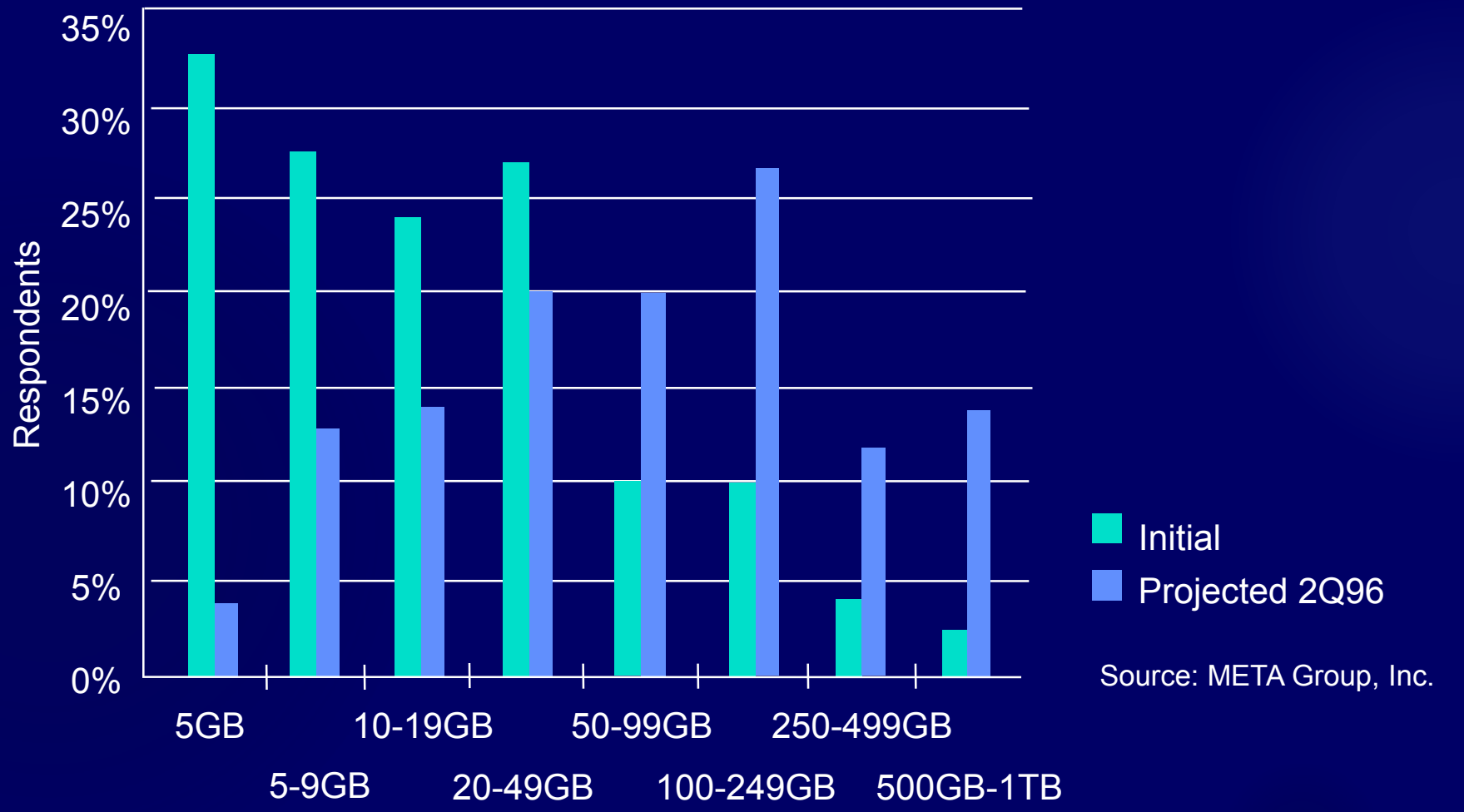
- A process of transforming data into information and making it available to users in a timely enough manner to make a difference

- [Forrester Research, April 1996]

Evolution

- 60's: Batch reports
  - hard to find and analyze information
  - inflexible and expensive, reprogram every new request
- 70's: Terminal-based DSS and EIS (executive information systems)
  - still inflexible, not integrated with desktop tools
- 80's: Desktop data access and analysis tools
  - query tools, spreadsheets, GUIs
  - easier to use, but only access operational databases
- 90's: Data warehousing with integrated OLAP engines and tools

Warehouses are Very Large Databases



Source: META Group, Inc.

Very Large Data Bases

- Terabytes -- 10^12 bytes:
- Petabytes -- 10^15 bytes:
- Exabytes -- 10^18 bytes:
- Zettabytes -- 10^21 bytes:
- Zottabytes -- 10^24 bytes:

- Walmart -- 24 Terabytes
- Geographic Information Systems
- National Medical Records
- Weather images
- Intelligence Agency Videos

Data Warehousing -- It is a process

- Technique for assembling and managing data from various sources for the purpose of answering business questions. Thus making decisions that were not previous possible
- A decision support database maintained separately from the organization's operational database

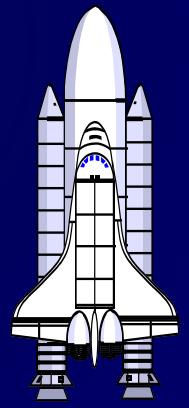
Data Warehouse

- A data warehouse is a
  - subject-oriented
  - integrated
  - time-varying
  - non-volatile
- collection of data that is used primarily in organizational decision making.
  - -- Bill Inmon, Building the Data Warehouse 1996

Explorers, Farmers and Tourists

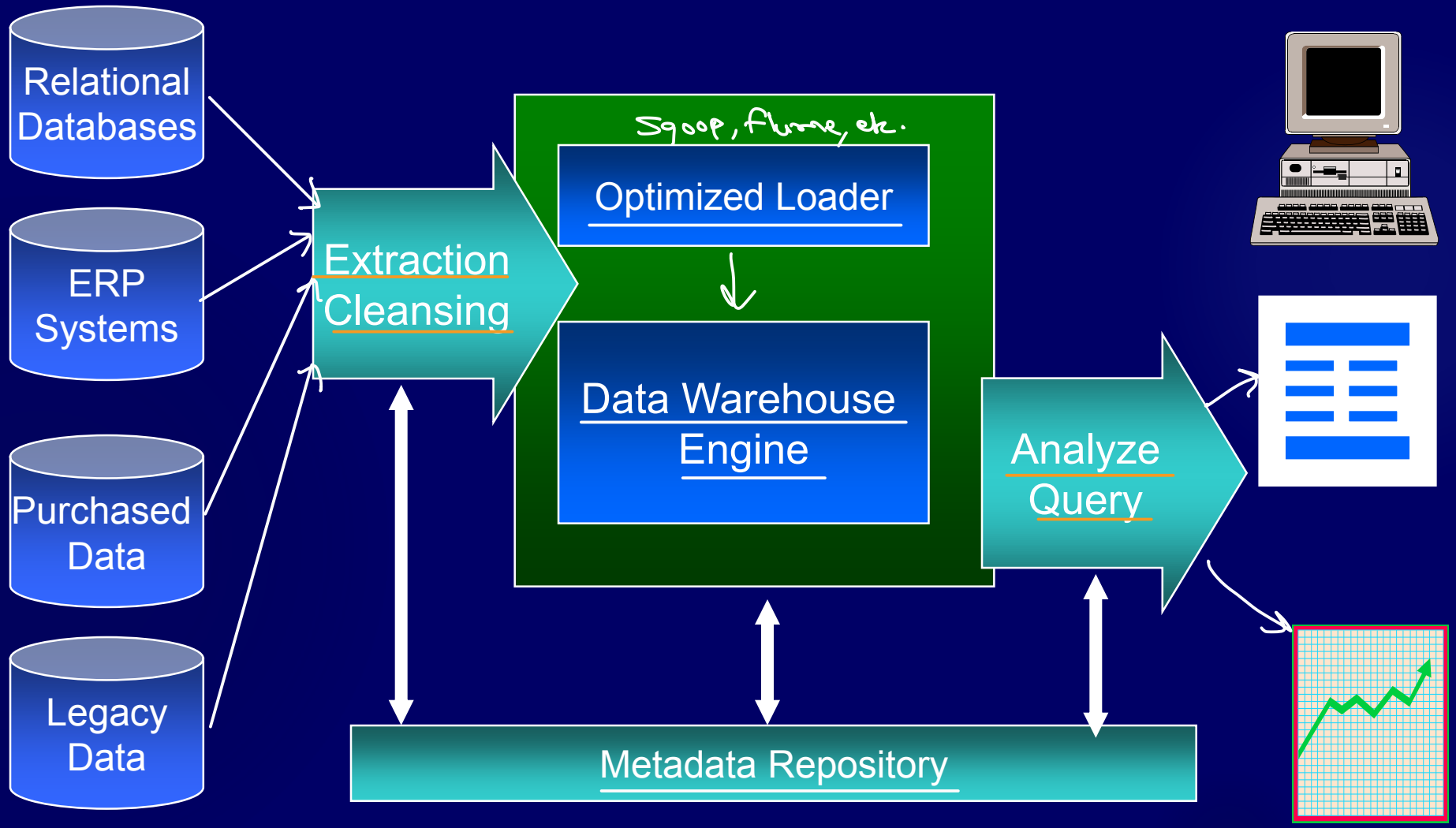
Tourists: Browse information harvested by farmers

Farmers: Harvest information from known access paths



Explorers: Seek out the unknown and previously unsuspected rewards hiding in the detailed data

Data Warehouse Architecture



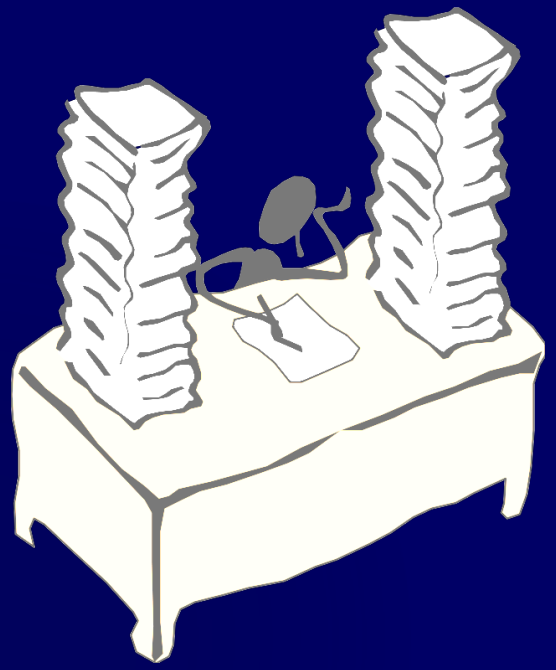
Data Warehouse for Decision Support & OLAP

- Putting Information technology to help the knowledge worker make faster and better decisions
  - Which of my customers are most likely to go to the competition?
  - What product promotions have the biggest impact on revenue?
  - How did the share price of software companies correlate with profits over last 10 years?

Decision Support

- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgements

Data Mining works with Warehouse Data



- Data Warehousing provides the Enterprise with a memory

- Data Mining provides the Enterprise with intelligence

We want to know ...

- Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
- Which types of transactions are likely to be fraudulent given the demographics and transactional history of a particular customer?
- If I raise the price of my product by Rs. 2, what is the effect on my ROI?
- If I offer only 2,500 airline miles as an incentive to purchase rather than 5,000, how many lost responses will result?
- If I emphasize ease-of-use of the product as opposed to its technical capabilities, what will be the net effect on my revenues?
- Which of my customers are likely to be the most loyal?

Data Mining helps extract such information

Application Areas

Industry	Application
Finance	Credit Card Analysis
Insurance	Claims, Fraud Analysis
Telecommunication	Call record analysis
Transport	Logistics management
Consumer goods	promotion analysis
Data Service providers	Value added data
Utilities	Power usage analysis

Data Mining in Use

- The US Government uses Data Mining to track fraud
- A Supermarket becomes an information broker
- Basketball teams use it to track game strategy
- Cross Selling
- Warranty Claims Routing
- Holding on to Good Customers
- Weeding out Bad Customers

What makes data mining possible?

- Advances in the following areas are making data mining deployable:
  - data warehousing
  - better and more data (i.e., operational, behavioral, and demographic)
  - the emergence of easily deployed data mining tools and
  - the advent of new data mining techniques.
    - -- Gartner Group

Why Separate Data Warehouse?

- **Performance**
  - Op dbs designed & tuned for known txs & workloads.
  - Complex OLAP queries would degrade perf. for op txs.
  - Special data organization, access & implementation methods needed for multidimensional views & queries.
- **Function**
  - Missing data: Decision support requires historical data, which op dbs do not typically maintain.
  - Data consolidation: Decision support requires consolidation (aggregation, summarization) of data from many heterogeneous sources: op dbs, external sources.
  - Data quality: Different sources typically use inconsistent data representations, codes, and formats which have to be reconciled.

What are Operational Systems?

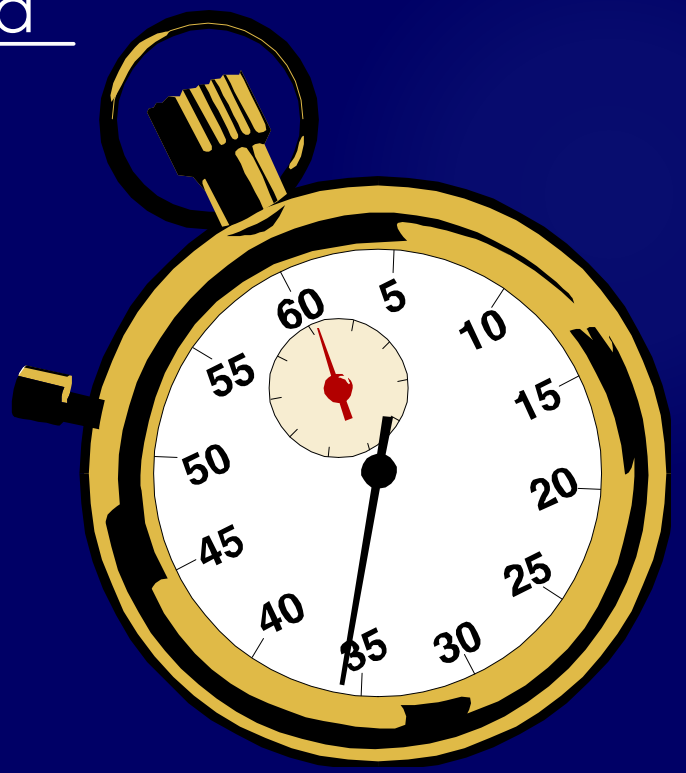
- They are OLTP systems
- Run mission critical applications
- Need to work with stringent performance requirements for routine tasks
- Used to run a business!

RDBMS used for OLTP

- Database Systems have been used traditionally for OLTP
  - clerical data processing tasks
  - detailed, up to date data
  - structured repetitive tasks
  - read/update a few records
  - isolation, recovery and integrity are critical

Operational Systems

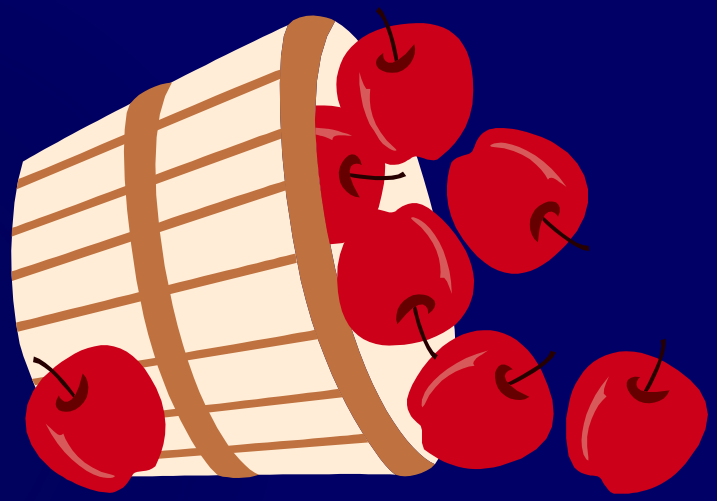
- Run the business in real time
- Based on up-to-the-second data
- Optimized to handle large numbers of simple read/write transactions
- Optimized for fast response to predefined transactions
- Used by people who deal with customers, products -- clerks, salespeople etc.
- They are increasingly used by customers

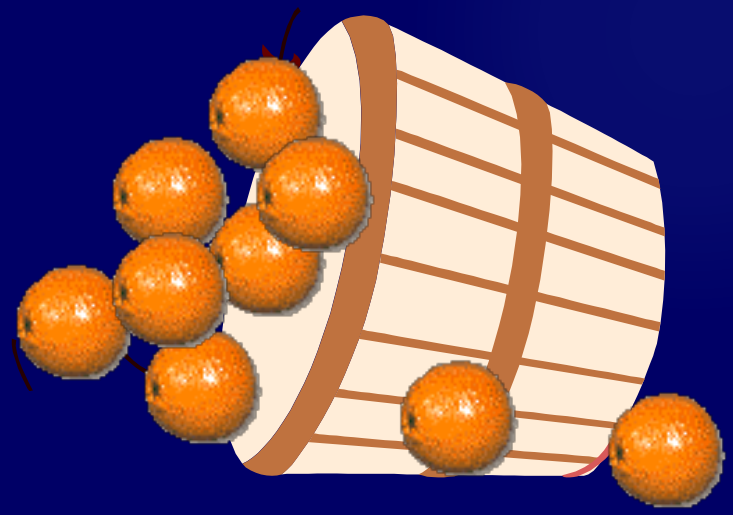


Examples of Operational Data

Data	Industry	Usage	Technology	Volumes
Customer File	All	Track Customer Details	Legacy application, flat files, main frames	Small-medium
Account Balance	Finance	Control account activities	Legacy applications, hierarchical databases, mainframe	Large
Point-of-Sale data	Retail	Generate bills, manage stock	ERP, Client/Server, relational databases	Very Large
Call Record	Telecomm-unications	Billing	Legacy application, hierarchical database, mainframe	Very Large
Production Record	Manufact-uring	Control Production	ERP, relational databases, AS/400	Medium

So, what's different?





Application-Orientation vs. Subject-Orientation

Application-Orientation

Operational Database

Loans

Credit Card

Trust

Savings

Subject-Orientation

Data Warehouse

Customer

Product

Vendor

Activity

OLTP vs. Data Warehouse

- OLTP systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
  - e.g., *average amount spent on phone calls between 9AM-5PM in Pune during the month of December*

OLTP vs Data Warehouse

- OLTP
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Repetitive access
  - Clerical User
- Warehouse (DSS)
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Ad-hoc access
  - Knowledge User (Manager)

OLTP vs Data Warehouse

- OLTP
  - Performance Sensitive
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - No data redundancy
  - Database Size 100MB - 100 GB
- Data Warehouse
  - Performance relaxed
  - Large volumes accessed at a time(millions)
  - Mostly Read (Batch Update)
  - Redundancy present
  - Database Size 100 GB - few terabytes

OLTP vs Data Warehouse

- OLTP
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety
- Data Warehouse
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

To summarize …

- OLTP Systems are used to *"run"* a business

- The Data Warehouse helps to *"optimize"* the business

Why Now?

- Data is being produced ✓
- ERP provides clean data ✓
- The computing power is available ✓
- The computing power is affordable ✓
- The competitive pressures are strong ✓
- Commercial products are available ✓

Myths surrounding OLAP Servers and Data Marts

- Data marts and OLAP servers are departmental solutions supporting a handful of users
- Million dollar massively parallel hardware is needed to deliver fast time for complex queries
- OLAP servers require massive and unwieldy indices
- Complex OLAP queries clog the network with data
- Data warehouses must be at least 100 GB to be effective
  - Source -- Arbor Software Home Page

Wal*Mart Case Study

- Founded by Sam Walton ✓
- One the largest Super Market Chains in the US ✓

- Wal*Mart: 2000+ Retail Stores
- SAM's Clubs 100+Wholesalers Stores

  - ☒This case study is from Felipe Carino's (NCR Teradata) presentation made at Stanford Database Seminar

Old Retail Paradigm

- Wal*Mart
  - Inventory Management
  - Merchandise Accounts Payable
  - Purchasing
  - Supplier Promotions: National, Region, Store Level
- Suppliers
  - Accept Orders
  - Promote Products
  - Provide special Incentives
  - Monitor and Track The Incentives
  - Bill and Collect Receivables
  - Estimate Retailer Demands

New (Just-In-Time) Retail Paradigm

- No more deals
- Shelf-Pass Through (POS Application)
  - One Unit Price
    - Suppliers paid once a week on ACTUAL items sold
  - Wal*Mart Manager
    - Daily Inventory Restock
    - Suppliers (sometimes SameDay) ship to Wal*Mart
- Warehouse-Pass Through
  - Stock some Large Items
    - Delivery may come from supplier
  - Distribution Center
    - Supplier's merchandise unloaded directly onto Wal*Mart Trucks

Wal*Mart System

- NCR 5100M 96 Nodes;
  - *24 TB Raw Disk; 700 - 1000 Pentium CPUs*
- Number of Rows:
  - *> 5 Billions*
- Historical Data:
  - *65 weeks (5 Quarters)*
- New Daily Volume:
  - *Current Apps: 75 Million*
  - *New Apps: 100 Million +*
- Number of Users:
  - *Thousands*
- Number of Queries:
  - *60,000 per week*