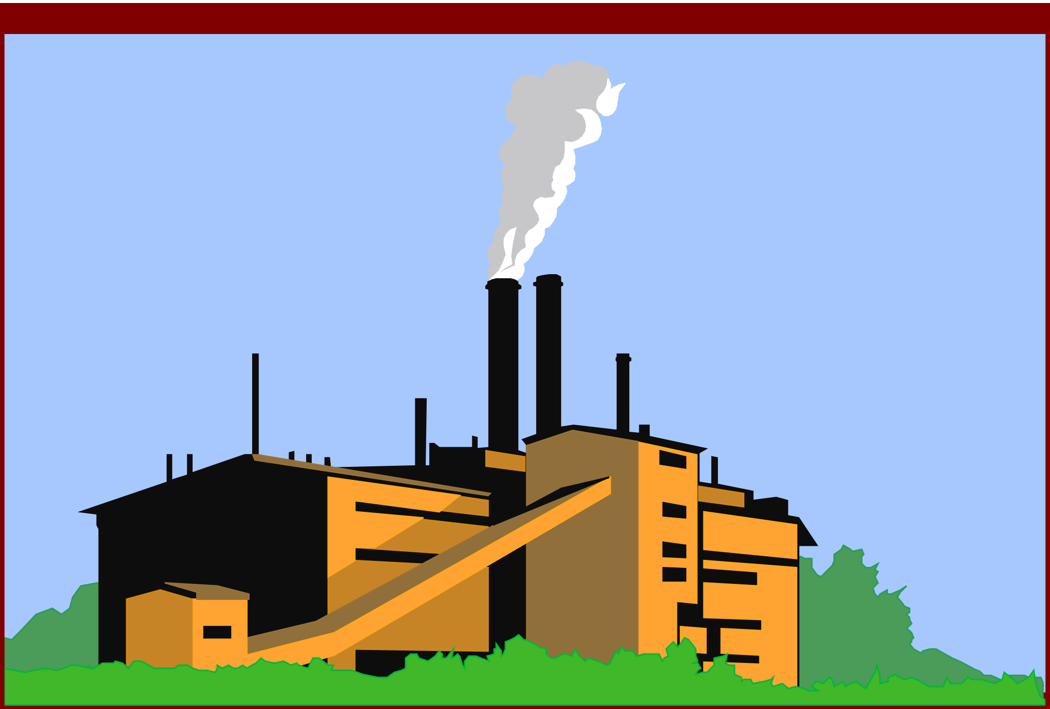
Data Wharehousing, OLAP and Data Mining

UNIT-5

Overview

- Part 1: Data Warehouses
- Part 2: OLAP
- Part 3: Data Mining
- Part 4: Query Processing and Optimization

Part 1: Data Warehouses

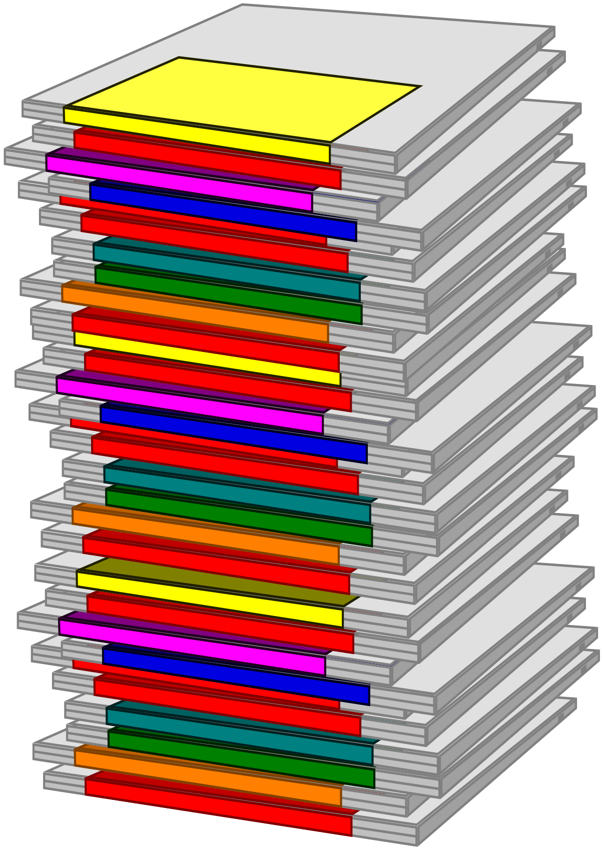


Data, Data everywhere yet ...

- 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]

Why Data Warehousing?

Which are our lowest/highest margin customers ?

What is the most effective distribution channel?

Who are my customers and what products are they buying?

What product prom- -otions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

What impact will new products/services have on revenue and margins?

Decision Support

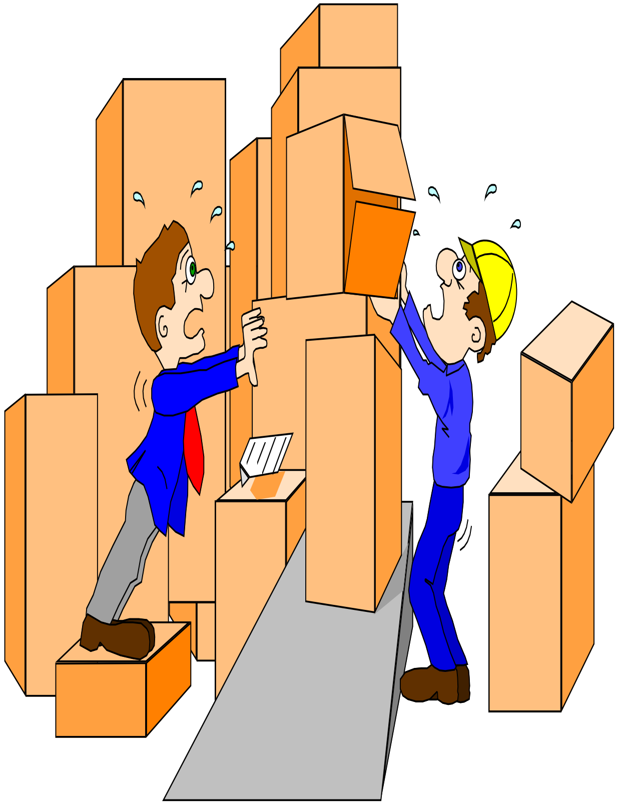
- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgements

Evolution of Decision Support

- 60's: Batch reports
  - hard to find and analyze information
  - inflexible and expensive, reprogram every request
- 70's: Terminal based DSS and EIS
- 80's: Desktop data access and analysis tools
  - query tools, spreadsheets, GUIs
  - easy to use, but access only operational db
- 90's: Data warehousing with integrated OLAP engines and tools

What are the users saying...

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



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

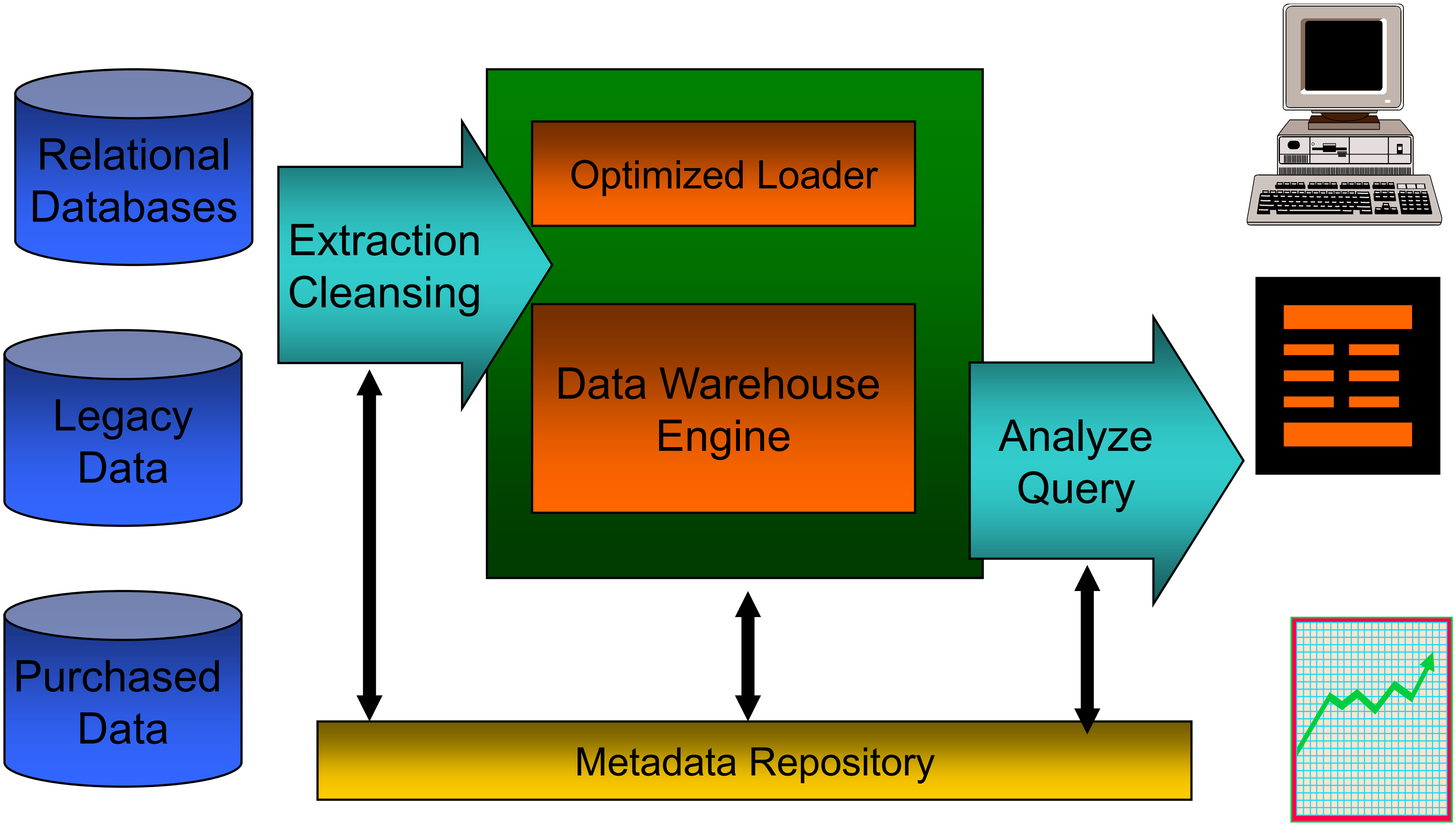
Traditional RDBMS used for OLTP

- 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
- Will call these operational systems

OLTP vs Data Warehouse

OLTP	Warehouse (DSS)
Application Oriented	Subject Oriented
Used to run business	Used to analyze business
Clerical User	Manager/Analyst
Detailed data	Summarized and refined
Current up to date	Snapshot data
Isolated Data	Integrated Data
Repetitive access by small transactions	Ad-hoc access using large queries
Read/Update access	Mostly read access (batch update)

Data Warehouse Architecture



Relational Databases

Legacy Data

Purchased Data

Extraction Cleansing

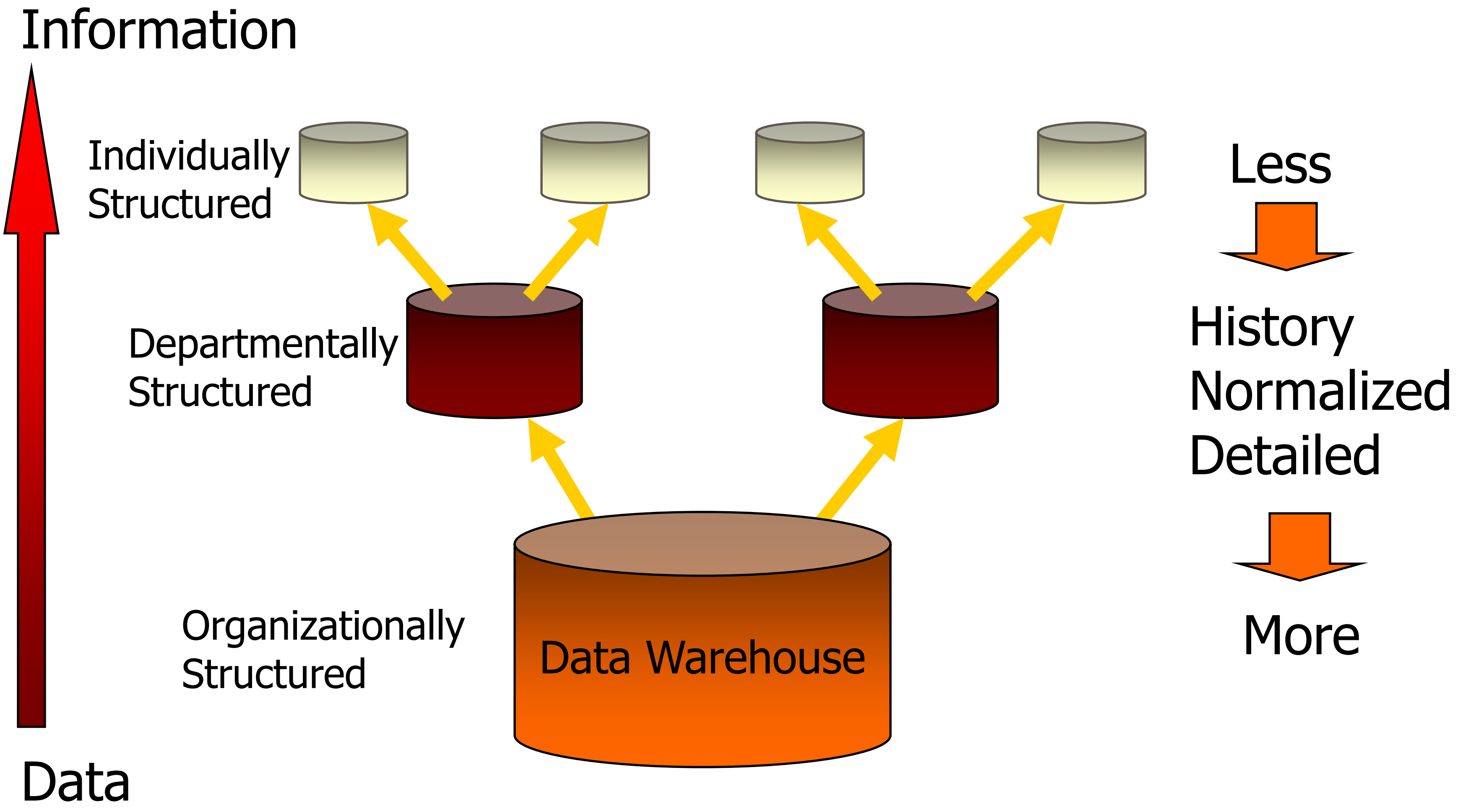
Optimized Loader

Data Warehouse Engine

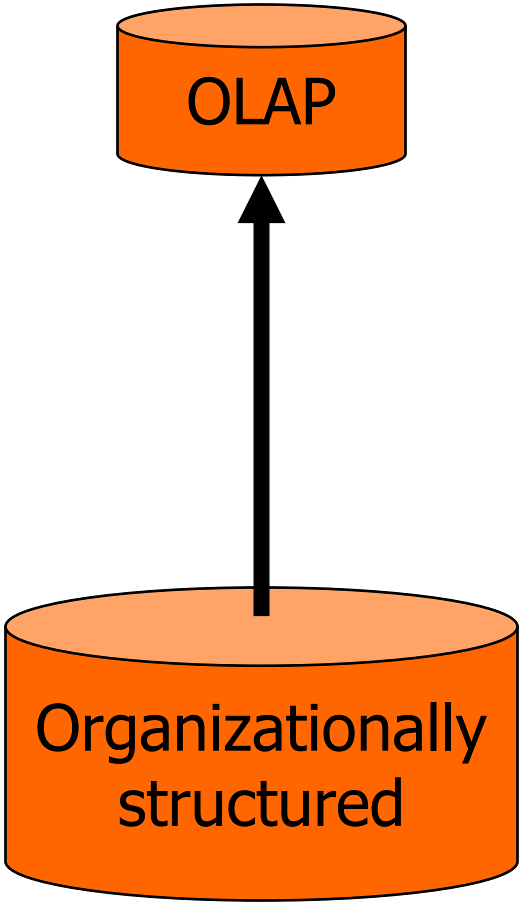
Analyze Query

Metadata Repository

From the Data Warehouse to Data Marts



Users have different views of Data

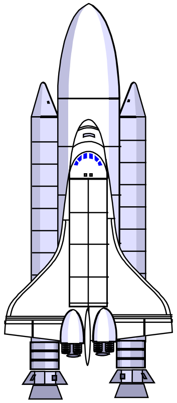


OLAP

Organizationally structured

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

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

Information as a Strategic Weapon

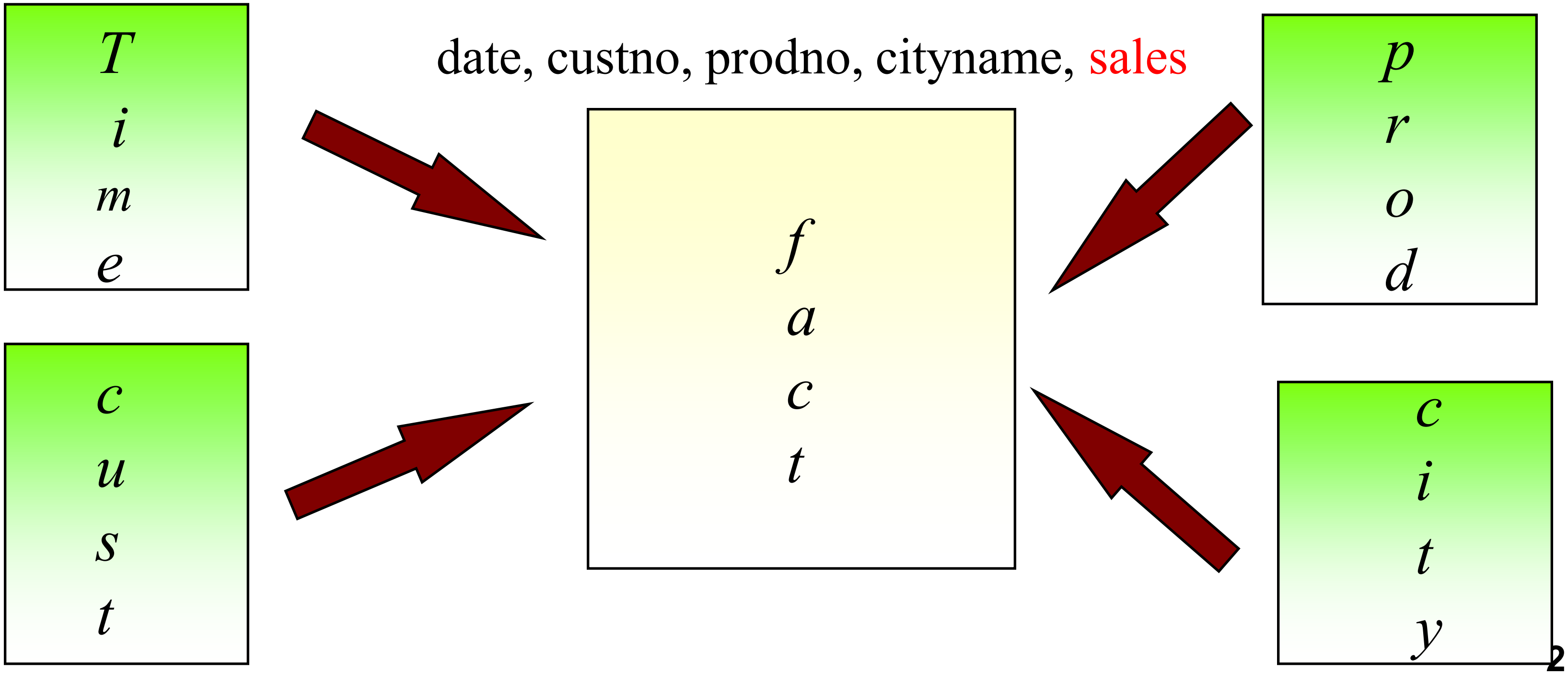
- Daily Summary of all Sales Information
- Regional Analysis of all Stores in a logical area
- Specific Product Sales
- Specific Supplies Sales
- Trend Analysis, etc.
- Wal*Mart uses information when negotiating with
  - Suppliers
  - Advertisers etc.

Schema Design

- Database organization
  - must look like business
  - must be recognizable by business user
  - approachable by business user
  - Must be *simple*
- Schema Types
  - Star Schema
  - Fact Constellation Schema
  - Snowflake schema

Star Schema

- A single fact table and for each dimension one dimension table
- Does not capture hierarchies directly



Time

date, custno, prodno, cityname, sales

prod

fact

cust

city

Dimension Tables

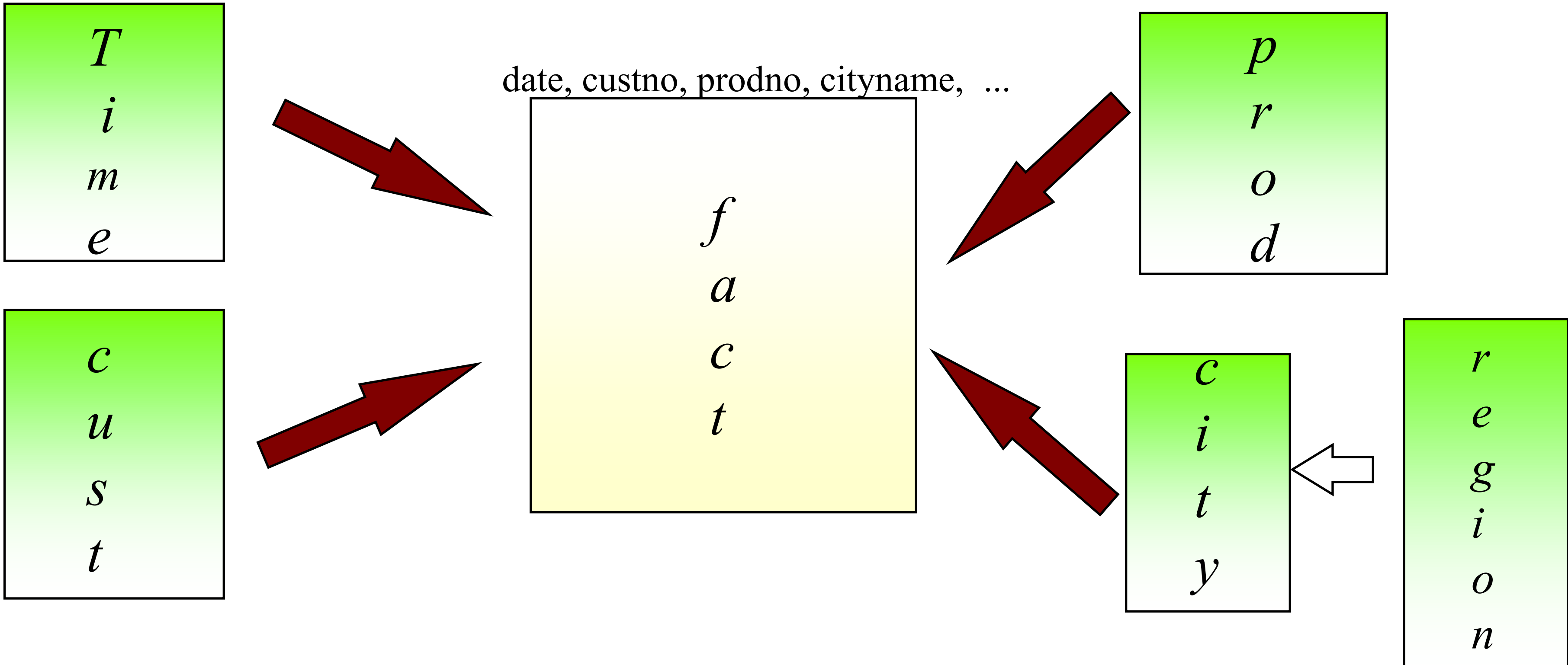
- Dimension tables
  - Define business in terms already familiar to users
  - Wide rows with lots of descriptive text
  - Small tables (about a million rows)
  - Joined to fact table by a foreign key
  - heavily indexed
  - typical dimensions
    - time periods, geographic region (markets, cities), products, customers, salesperson, etc.

Fact Table

- Central table
  - Typical example: individual sales records
  - mostly raw numeric items
  - narrow rows, a few columns at most
  - large number of rows (millions to a billion)
  - Access via dimensions

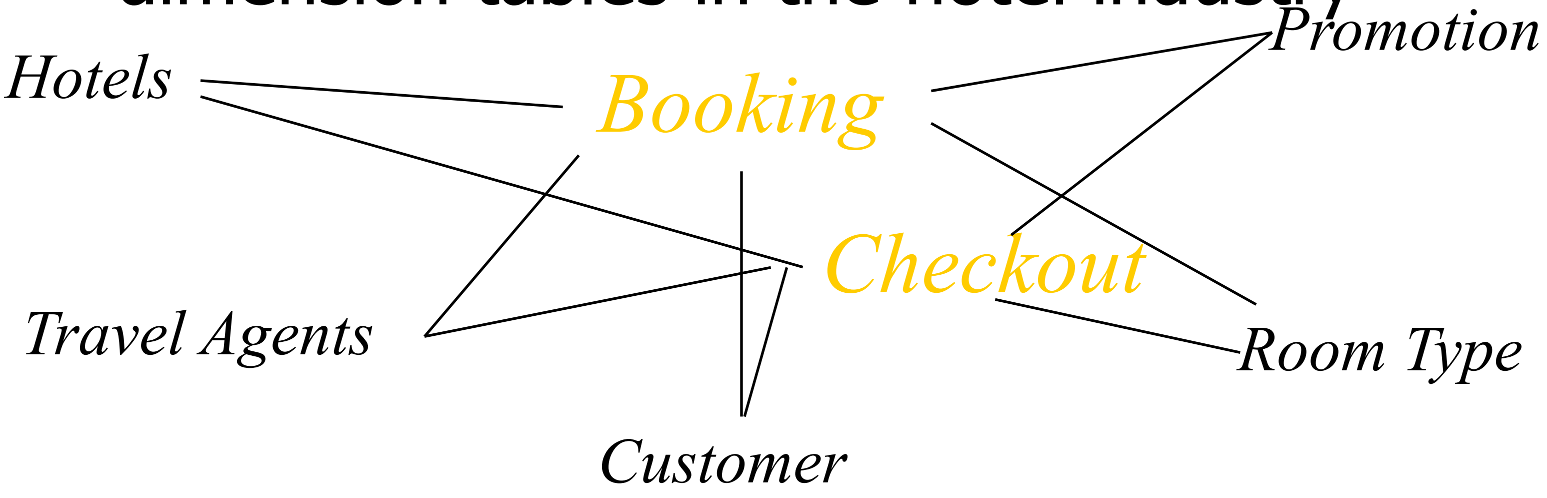
Snowflake schema

- Represent dimensional hierarchy directly by normalizing tables.
- Easy to maintain and saves storage



Fact Constellation

- Fact Constellation
  - Multiple fact tables that share many dimension tables
  - Booking and Checkout may share many dimension tables in the hotel industry



Hotels

Booking

Promotion

Checkout

Travel Agents

Room Type

Customer

Data Granularity in Warehouse

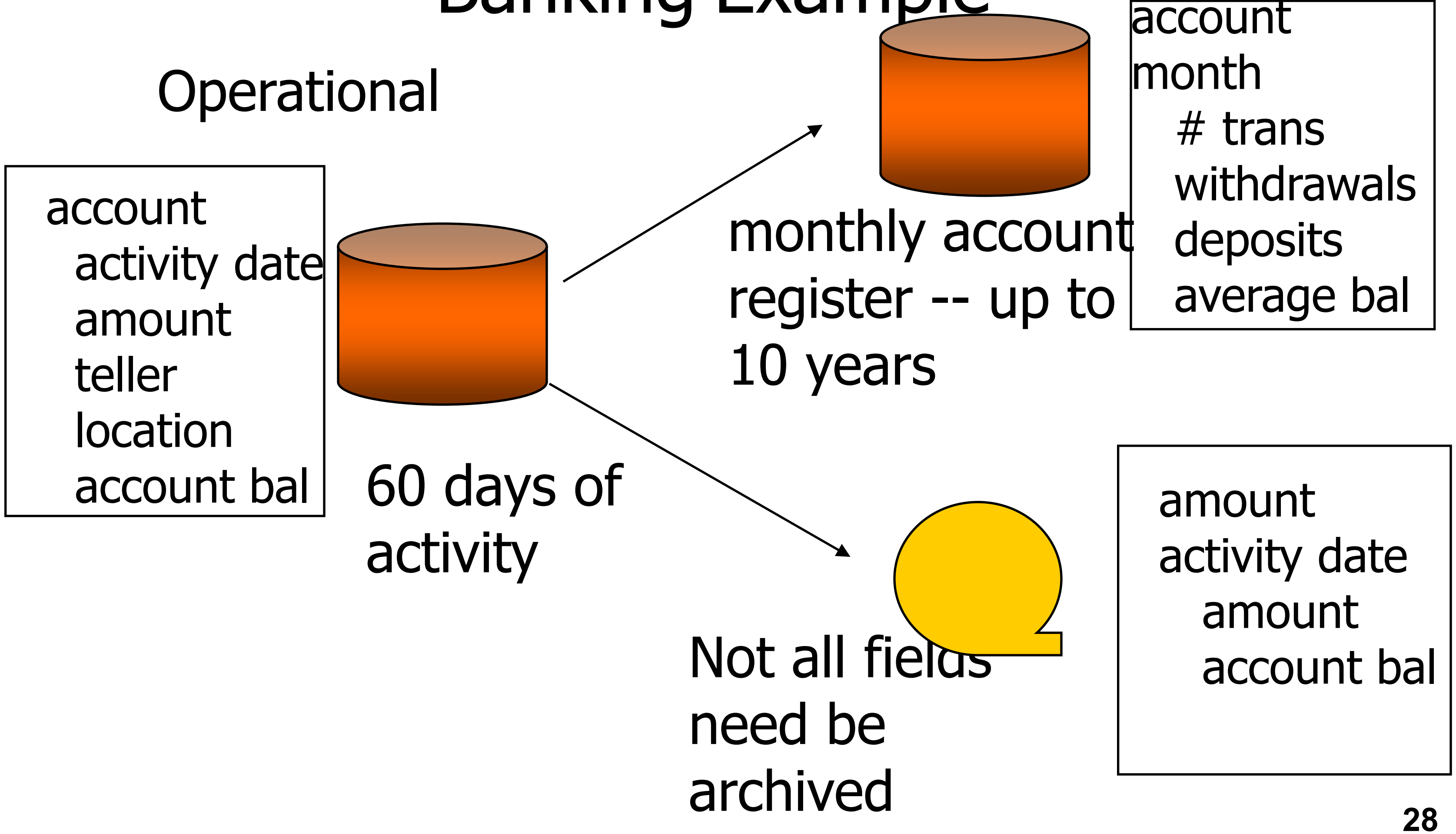
- Summarized data stored
  - reduce storage costs
  - reduce cpu usage
  - increases performance since smaller number of records to be processed
  - design around traditional high level reporting needs
  - tradeoff with volume of data to be stored and detailed usage of data

Granularity in Warehouse

- Solution is to have dual level of granularity
  - Store summary data on disks
    - 95% of DSS processing done against this data
  - Store detail on tapes
    - 5% of DSS processing against this data

Levels of Granularity

Banking Example



Operational

account
activity date
amount
teller
location
account bal

60 days of activity

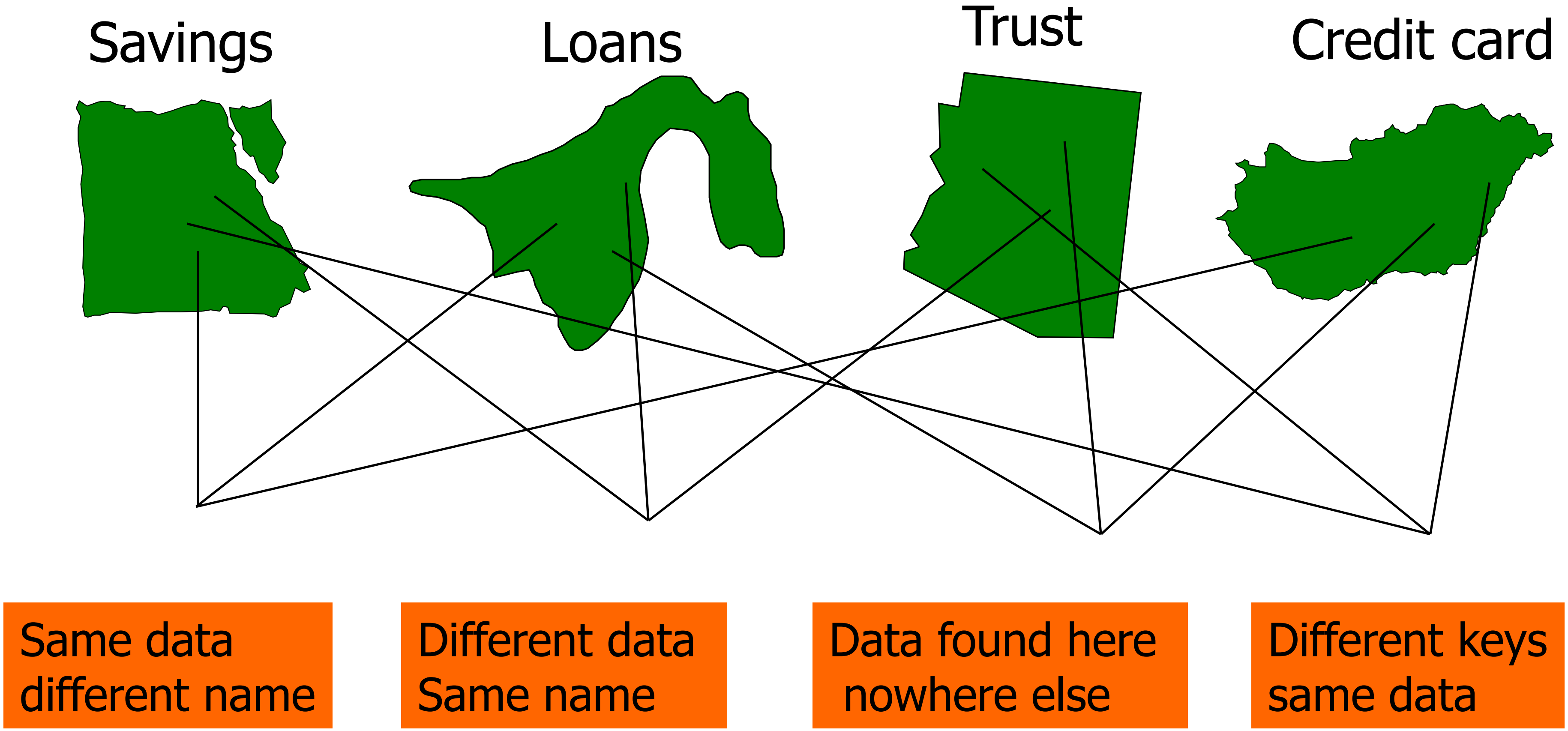
monthly account register -- up to 10 years

account
month
# trans
withdrawals
deposits
average bal

Not all fields need be archived

amount
activity date
amount
account bal

Data Integration Across Sources



Savings

Loans

Trust

Credit card

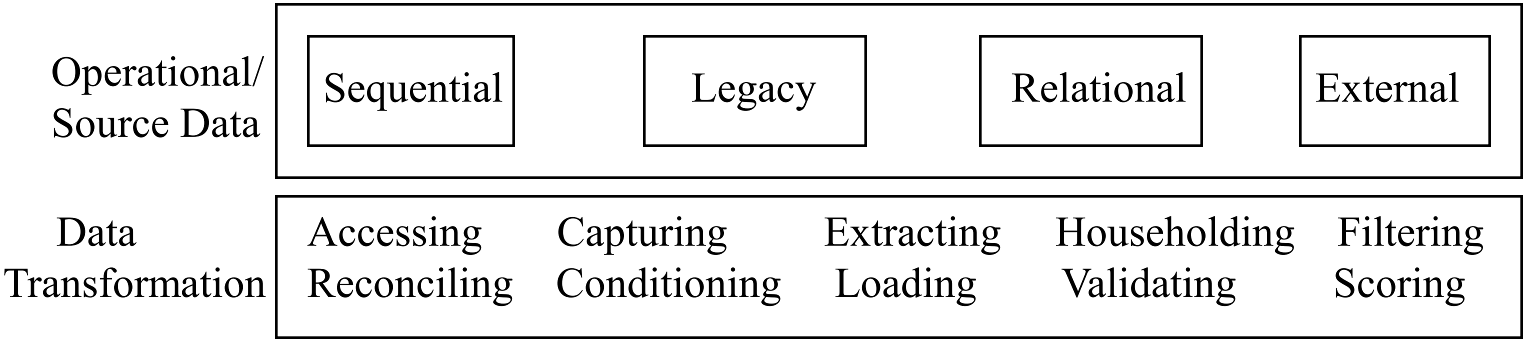
Same data
different name

Different data
Same name

Data found here
nowhere else

Different keys
same data

Data Transformation



Operational/ Source Data	Sequential	Legacy	Relational	External	
Data Transformation	Accessing Reconciling	Capturing Conditioning	Extracting Loading	Householding Validating	Filtering Scoring

- Data transformation is the foundation for achieving single version of the truth
- Major concern for IT
- Data warehouse can fail if appropriate data transformation strategy is not developed

Data Integrity Problems

- Same person, different spellings
  - Agarwal, Agrawal, Aggarwal etc...
- Multiple ways to denote company name
  - Persistent Systems, PSPL, Persistent Pvt. LTD.
- Use of different names
  - mumbai, bombay
- Different account numbers generated by different applications for the same customer
- Required fields left blank
- Invalid product codes collected at point of sale
  - manual entry leads to mistakes
  - "in case of a problem use 9999999"

Data Transformation Terms

- Extracting
- Conditioning
- Scrubbing
- Merging
- Householding
- Enrichment
- Scoring
- Loading
- Validating
- Delta Updating

Data Transformation Terms

- Householding
  - Identifying all members of a household (living at the same address)
  - Ensures only one mail is sent to a household
  - Can result in substantial savings: 1 million catalogues at $50 each costs $50 million . A 2% savings would save $1 million

Refresh

- Propagate updates on source data to the warehouse
- Issues:
  - when to refresh
  - how to refresh -- incremental refresh techniques

When to Refresh?

- periodically (e.g., every night, every week) or after significant events
- on every update: not warranted unless warehouse data require current data (up to the minute stock quotes)
- refresh policy set by administrator based on user needs and traffic
- possibly different policies for different sources

Refresh techniques

- Incremental techniques
  - detect changes on base tables: replication servers (e.g., Sybase, Oracle, IBM Data Propagator)
    - snapshots (Oracle)
    - transaction shipping (Sybase)
  - compute changes to derived and summary tables
  - maintain transactional correctness for incremental load

How To Detect Changes

- Create a snapshot log table to record ids of updated rows of source data and timestamp
- Detect changes by:
  - Defining after row triggers to update snapshot log when source table changes
  - Using regular transaction log to detect changes to source data

Querying Data Warehouses

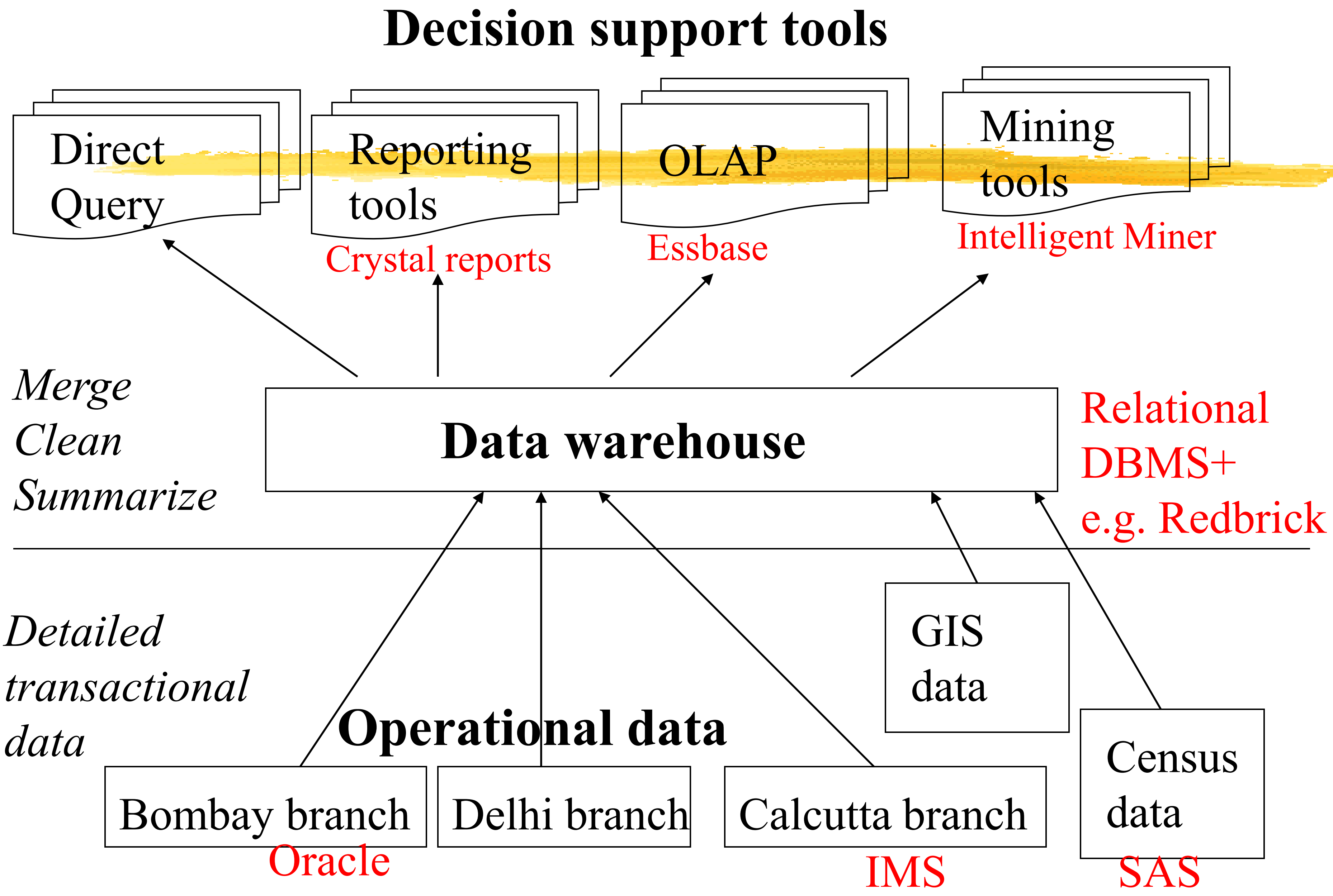
- SQL Extensions
- Multidimensional modeling of data
  - OLAP
  - More on OLAP later …

SQL Extensions

- Extended family of aggregate functions
  - rank (top 10 customers)
  - percentile (top 30% of customers)
  - median, mode
  - Object Relational Systems allow addition of new aggregate functions
- Reporting features
  - running total, cumulative totals

Reporting Tools

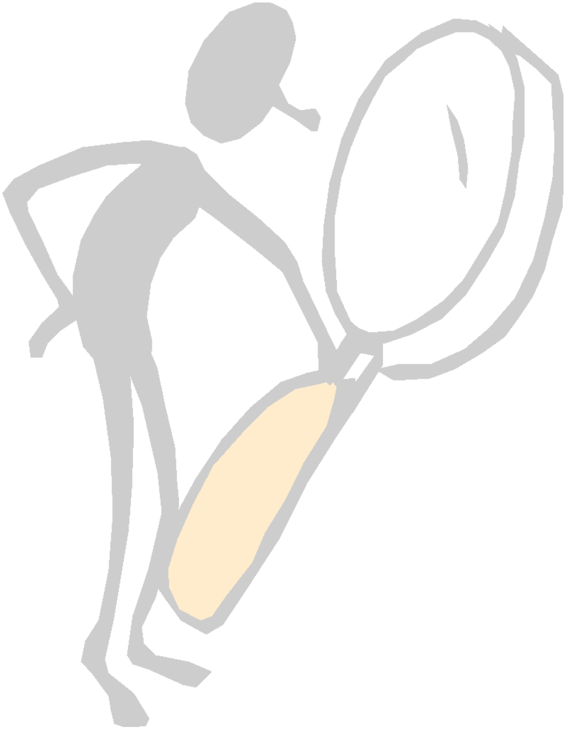
- Andyne Computing -- GQL
- Brio -- BrioQuery
- Business Objects -- Business Objects
- Cognos -- Impromptu
- Information Builders Inc. -- Focus for Windows
- Oracle -- Discoverer2000
- Platinum Technology -- SQL*Assist, ProReports
- PowerSoft -- InfoMaker
- SAS Institute -- SAS/Assist
- Software AG -- Esperant
- Sterling Software -- VISION:Data



Decision support tools
Direct Query
Reporting tools
Crystal reports
OLAP
Essbase
Mining tools
Intelligent Miner
Merge
Clean
Summarize
Data warehouse
Relational DBMS+
e.g. Redbrick
Detailed transactional data
Operational data
GIS data
Bombay branch
Oracle
Delhi branch
Calcutta branch
IMS
Census data
SAS

Deploying Data Warehouses

- What business information keeps you in business today? What business information can put you out of business tomorrow?
- What business information should be a mouse click away?
- What business conditions are the driving the need for business information?



Cultural Considerations

- Not just a technology project
- New way of using information to support daily activities and decision making
- Care must be taken to prepare organization for change
- Must have organizational backing and support

User Training

- Users must have a higher level of IT proficiency than for operational systems
- Training to help users analyze data in the warehouse effectively

Warehouse Products

- Computer Associates -- CA-Ingres
- Hewlett-Packard -- Allbase/SQL
- Informix -- Informix, Informix XPS
- Microsoft -- SQL Server
- Oracle – Oracle
- Red Brick -- Red Brick Warehouse
- SAS Institute -- SAS
- Software AG -- ADABAS
- Sybase -- SQL Server, IQ, MPP

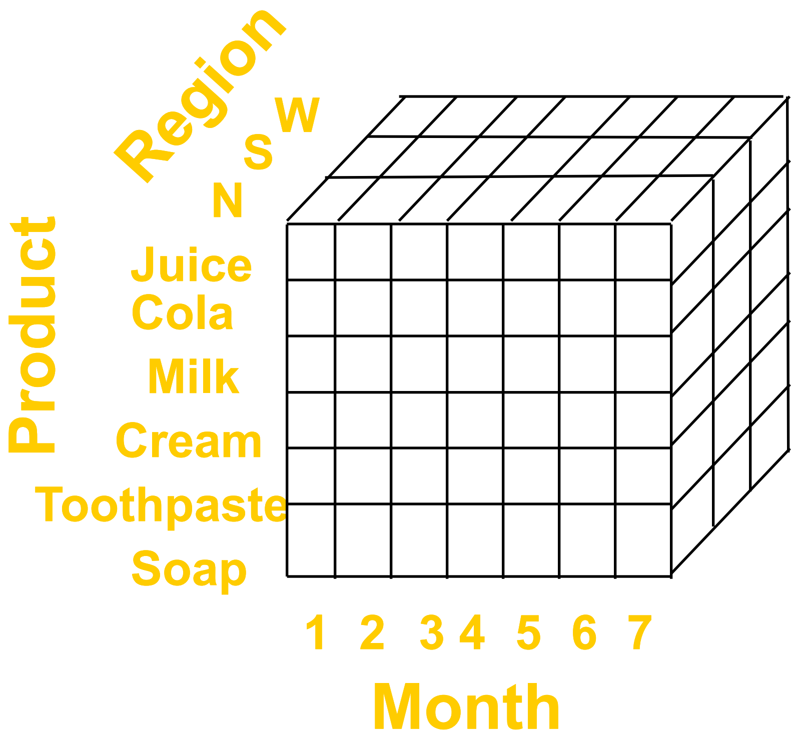
Part 2: OLAP

Nature of OLAP Analysis

- Aggregation -- (total sales, percent-to-total)
- Comparison -- Budget vs. Expenses
- Ranking -- Top 10, quartile analysis
- Access to detailed and aggregate data
- Complex criteria specification
- Visualization
- Need interactive response to aggregate queries

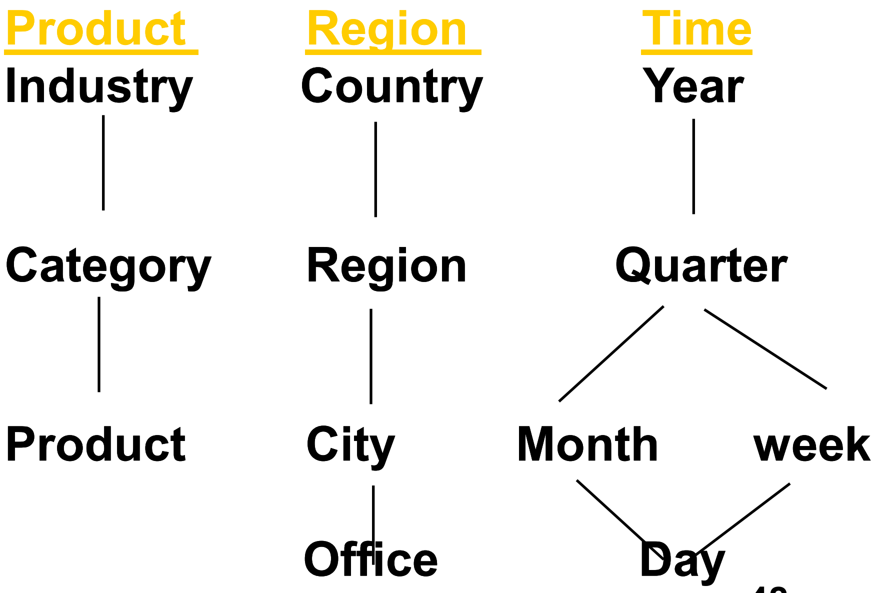
Multi-dimensional Data

- Measure - sales (actual, plan, variance)



Dimensions: Product, Region, Time

Hierarchical summarization paths



Product — Industry — Category — Product

Region — Country — Region — City — Office

Time — Year — Quarter — Month, week — Day

Conceptual Model for OLAP

- Numeric measures to be analyzed
  - e.g. Sales (Rs), sales (volume), budget, revenue, inventory
- Dimensions
  - other attributes of data, define the space
  - e.g., store, product, date-of-sale
  - hierarchies on dimensions
    - e.g. branch -> city -> state

Operations

- Rollup: summarize data
  - e.g., given sales data, summarize sales for last year by product category and region
- Drill down: get more details
  - e.g., given summarized sales as above, find breakup of sales by city within each region, or within the Andhra region

More Cube Operations

- Slice and dice: select and project
  - e.g.: Sales of soft-drinks in Andhra over the last quarter
- Pivot: change the view of data
  -

	Q1	Q2	Total
L	22	33	55
S	15	44	59
Total	37	77	114

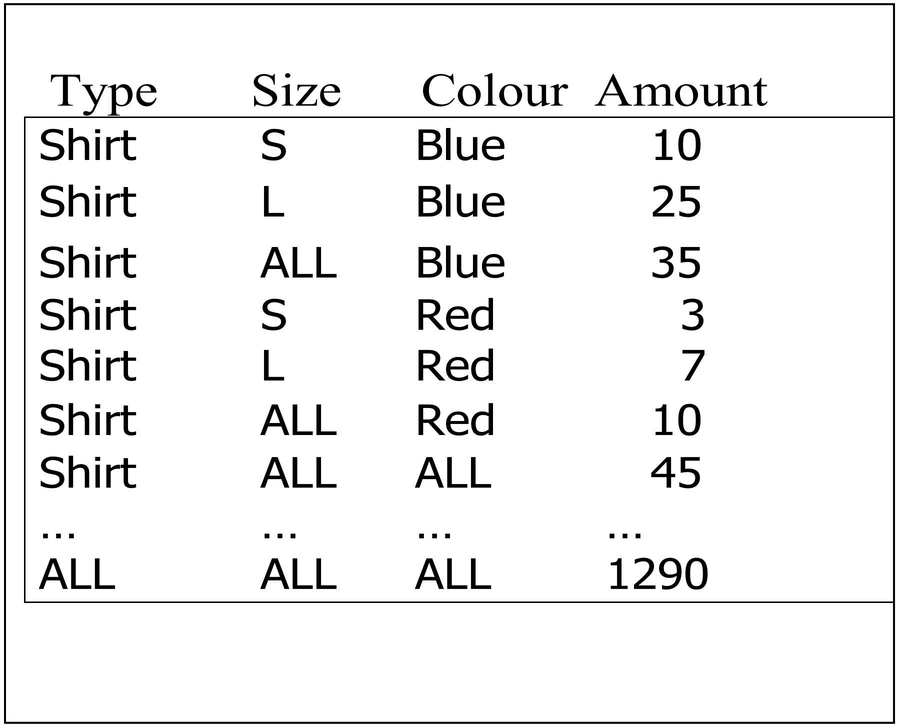
	L	S	Total
Red	14	07	21
Blue	41	52	93
Total	55	59	114

More OLAP Operations

- Hypothesis driven search: E.g. factors affecting defaulters
  - view defaulting rate on age aggregated over other dimensions
  - for particular age segment detail along profession
- Need interactive response to aggregate queries
  - => precompute various aggregates

MOLAP vs ROLAP

- MOLAP: Multidimensional array OLAP
- ROLAP: Relational OLAP



Type	Size	Colour	Amount
Shirt	S	Blue	10
Shirt	L	Blue	25
Shirt	ALL	Blue	35
Shirt	S	Red	3
Shirt	L	Red	7
Shirt	ALL	Red	10
Shirt	ALL	ALL	45
…	…	…	…
ALL	ALL	ALL	1290

SQL Extensions

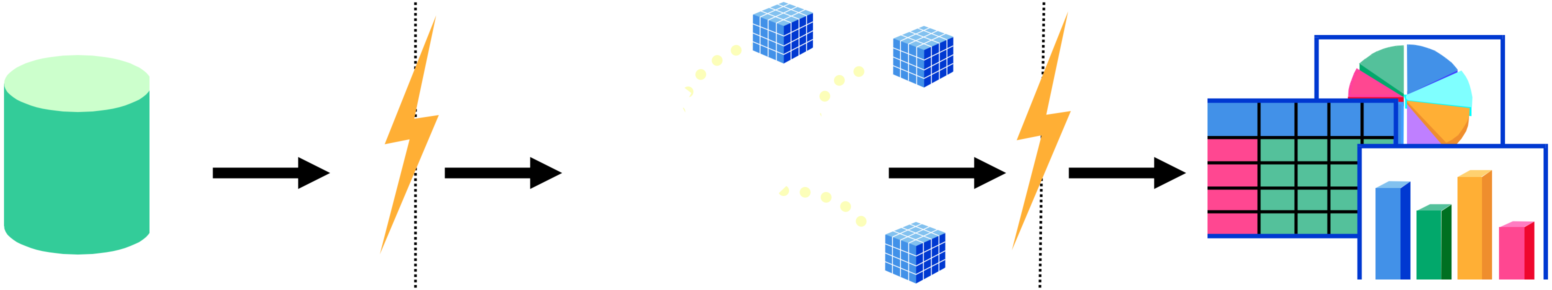
- Cube operator
  - group by on all subsets of a set of attributes (month,city)
  - redundant scan and sorting of data can be avoided
- Various other non-standard SQL extensions by vendors

OLAP: 3 Tier DSS

Data Warehouse

OLAP Engine

Decision Support Client



Database Layer

Application Logic Layer

Presentation Layer

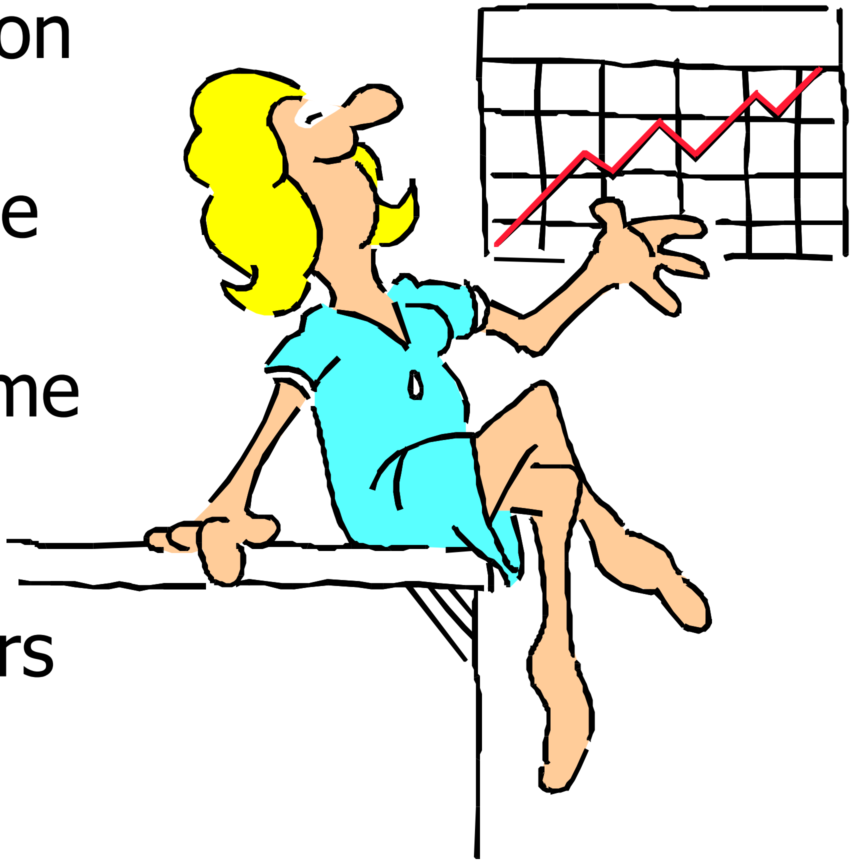
Store atomic data in industry standard Data Warehouse.

Generate SQL execution plans in the OLAP engine to obtain OLAP functionality.

Obtain multi-dimensional reports from the DSS Client.

Strengths of OLAP

- It is a powerful visualization tool
- It provides fast, interactive response times
- It is good for analyzing time series
- It can be useful to find some clusters and outliners
- Many vendors offer OLAP tools



Brief History

- Express and System W DSS
- Online Analytical Processing - coined by EF Codd in 1994 - white paper by Arbor Software
- Generally synonymous with earlier terms such as Decisions Support, Business Intelligence, Executive Information System
- MOLAP: Multidimensional OLAP (Hyperion (Arbor Essbase), Oracle Express)
- ROLAP: Relational OLAP (Informix MetaCube, Microstrategy DSS Agent)

OLAP and Executive Information Systems

- Andyne Computing -- Pablo
- Arbor Software -- Essbase
- Cognos -- PowerPlay
- Comshare -- Commander OLAP
- Holistic Systems -- Holos
- Information Advantage -- AXSYS, WebOLAP
- Informix -- Metacube
- Microstrategies -- DSS/Agent
- Oracle -- Express
- Pilot -- LightShip
- Planning Sciences -- Gentium
- Platinum Technology -- ProdeaBeacon, Forest & Trees
- SAS Institute -- SAS/EIS, OLAP++
- Speedware -- Media

Microsoft OLAP strategy

- Plato: OLAP server: powerful, integrating various operational sources
- OLE-DB for OLAP: emerging industry standard based on MDX --> extension of SQL for OLAP
- Pivot-table services: integrate with Office 2000
  - Every desktop will have OLAP capability.
- Client side caching and calculations
- Partitioned and virtual cube
- Hybrid relational and multidimensional storage

Part 3: Data Mining

Why Data Mining

- Credit ratings/targeted marketing:
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
  - Identify likely responders to sales promotions
- Fraud detection
  - Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?
- Customer relationship management:
  - Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor? :

Data Mining helps extract such information

Data mining

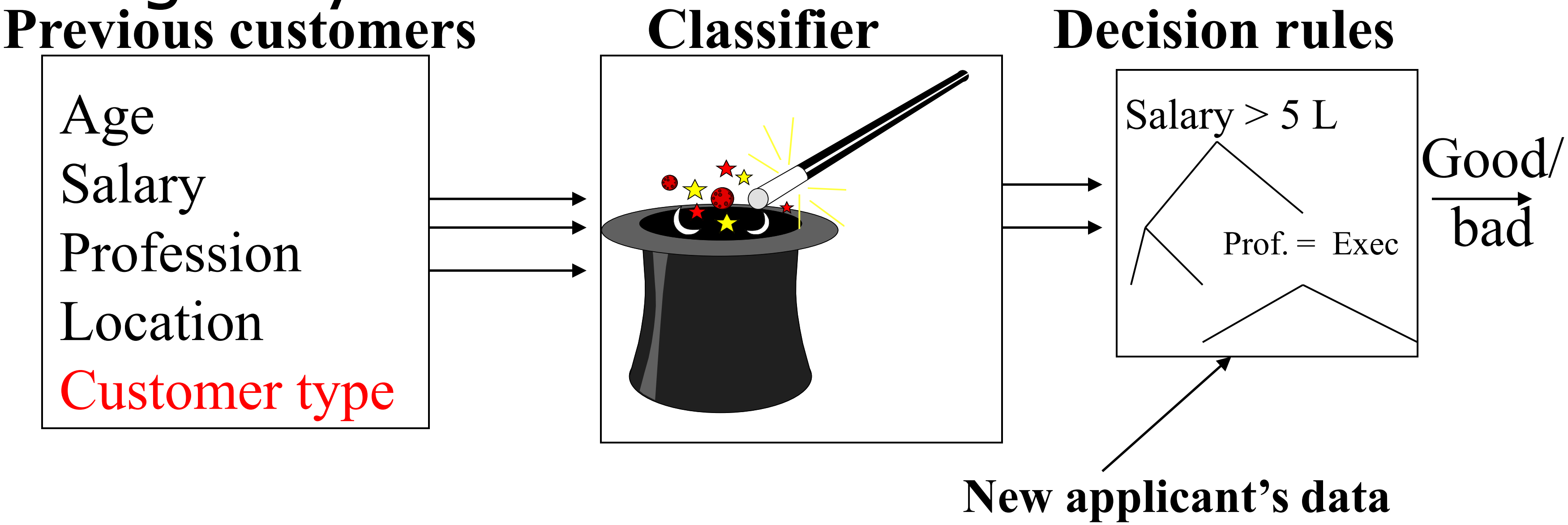
- Process of semi-automatically analyzing large databases to find interesting and useful patterns
- Overlaps with machine learning, statistics, artificial intelligence and databases but
  - more scalable in number of features and instances
  - more automated to handle heterogeneous data

Some basic operations

- Predictive:
  - Regression
  - Classification
- Descriptive:
  - Clustering / similarity matching
  - Association rules and variants
  - Deviation detection

Classification

- Given old data about customers and payments, predict new applicant's loan eligibility.



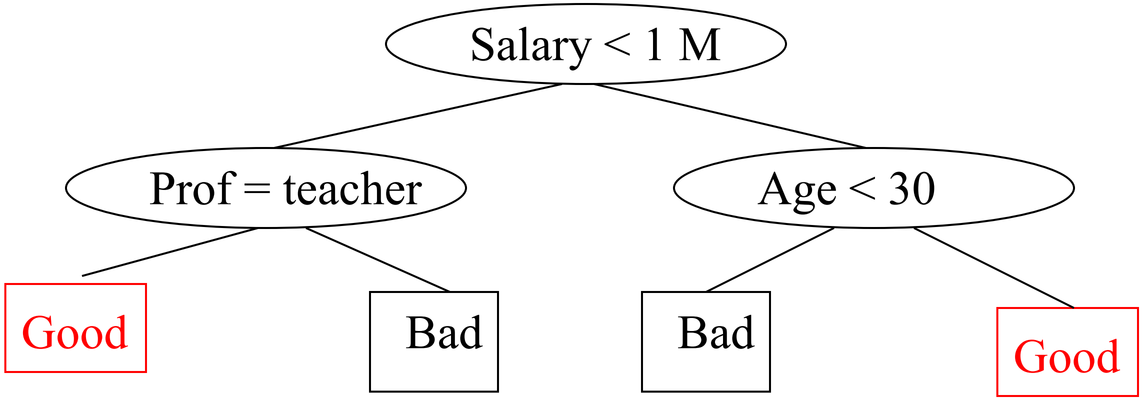
Classification methods

Goal: Predict class Ci = f(x1, x2, .. Xn)

- Regression: (linear or any other polynomial)
  - a*x1 + b*x2 + c = Ci.
- Nearest neighour
- Decision tree classifier: divide decision space into piecewise constant regions.
- Probabilistic/generative models
- Neural networks: partition by non-linear boundaries

Decision trees

⌘Tree where internal nodes are simple decision rules on one or more attributes and leaf nodes are predicted class labels.



Pros and Cons of decision trees

- Pros

  + Reasonable training time

  + Fast application

  + Easy to interpret

  + Easy to implement

  + Can handle large number of features

- Cons

  – Cannot handle complicated relationship between features

  – simple decision boundaries

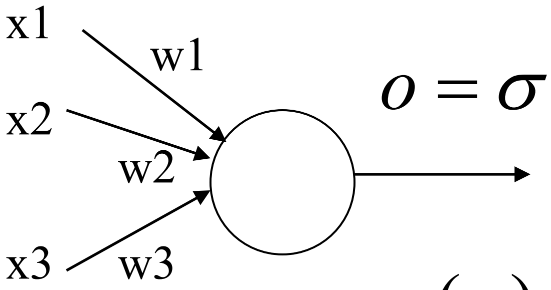
  – problems with lots of missing data

More information:
http://www.stat.wisc.edu/~limt/treeprogs.html

Neural network

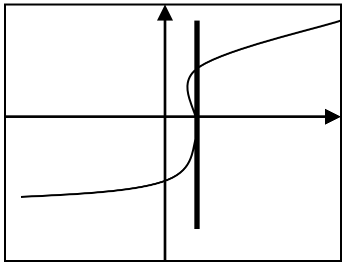
- Set of nodes connected by directed weighted edges

Basic NN unit

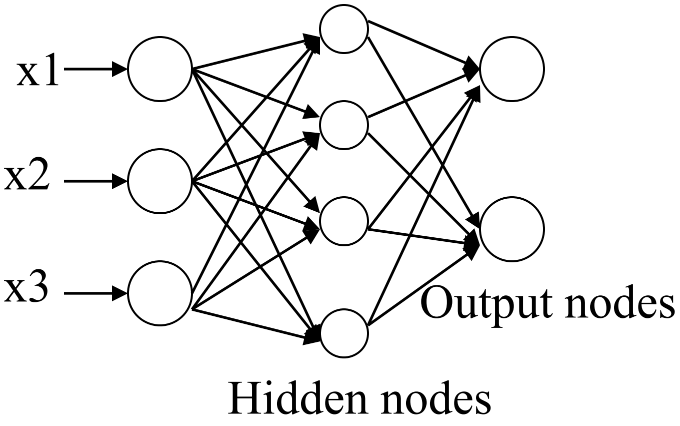


$$o = \sigma(\sum_{i=1}^{n} w_i x_i)$$

$$\sigma(y) = \frac{1}{1+e^{-y}}$$



A more typical NN



Pros and Cons of Neural Network

- Pros
  - + Can learn more complicated class boundaries
  - + Fast application
  - + Can handle large number of features

- Cons
  - – Slow training time
  - – Hard to interpret
  - – Hard to implement: trial and error for choosing number of nodes

Conclusion: Use neural nets only if decision trees/NN fail.

Bayesian learning

- Assume a probability model on generation of data.
- Apply bayes theorem to find most likely class as:

$$\text{predicted class}: c = \max_{c_j} p(c_j \mid d) = \max_{c_j} \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

$$c = \max_{c_j} \frac{p(c_j)}{p(d)} \prod_{i=1}^{n} p(a_i \mid c_j)$$

- Naïve bayes: Assume attributes conditionally independent given class value

Clustering

- Unsupervised learning when old data with class labels not available e.g. when introducing a new product.
- Group/cluster existing customers based on time series of payment history such that similar customers in same cluster.
- Key requirement: Need a good measure of similarity between instances.
- Identify micro-markets and develop policies for each

Association rules

- Given set T of groups of items
- Example: set of item sets purchased
- Goal: find all rules on itemsets of the form a-->b such that
  - support of a and b > user threshold s
  - conditional probability (confidence) of b given a > user threshold c
- Example: Milk --> bread
- Purchase of product A --> service B

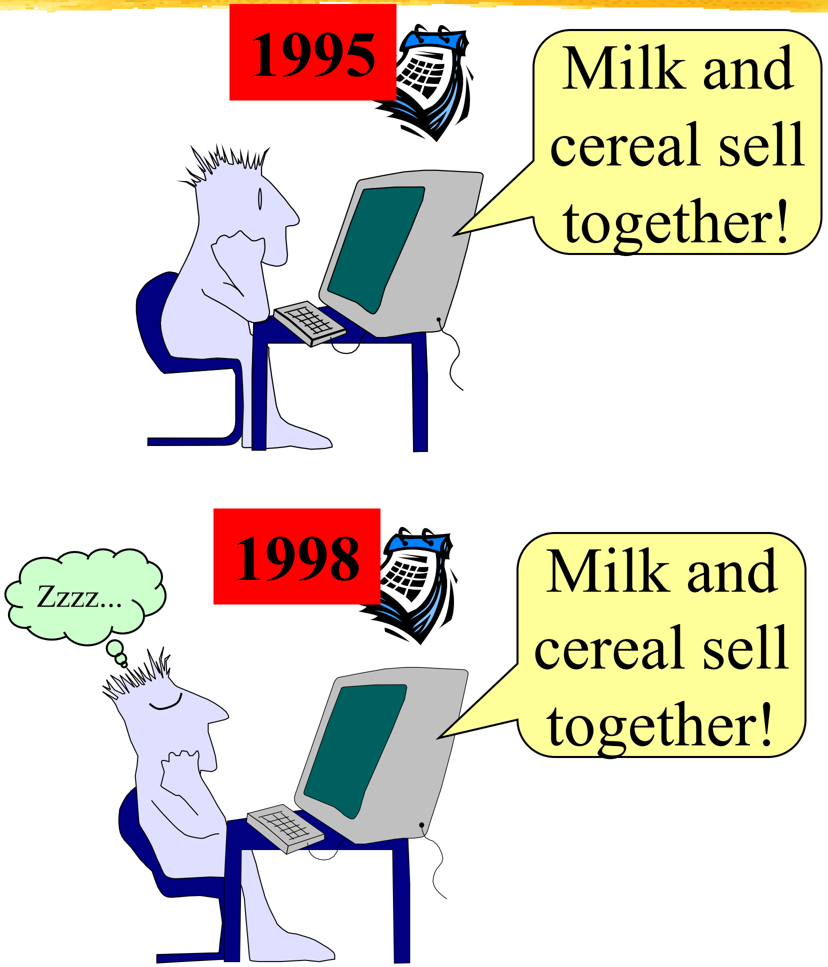
T
Milk, cereal
Tea, milk
Tea, rice, bread
cereal

Variants

- High confidence may not imply high correlation
- Use correlations. Find expected support and large departures from that interesting..
  - see statistical literature on contingency tables.
- Still too many rules, need to prune...

Prevalent ≠ Interesting

- Analysts already know about prevalent rules
- Interesting rules are those that *deviate* from prior expectation
- Mining's payoff is in finding *surprising* phenomena


1995
Milk and cereal sell together!
1998
Zzzz...
Milk and cereal sell together!

What makes a rule surprising?

- Does not match prior expectation
  - Correlation between milk and cereal remains roughly constant over time

- Cannot be trivially derived from simpler rules
  - Milk 10%, cereal 10%
  - Milk and cereal 10% … surprising
  - Eggs 10%
  - Milk, cereal and eggs 0.1% … surprising!
  - Expected 1%

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

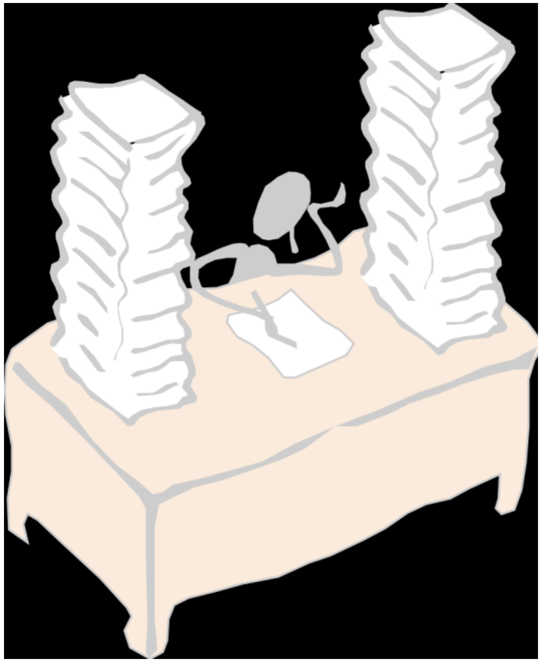
Data Mining in Use

- 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
- Target Marketing
- Holding on to Good Customers
- Weeding out Bad Customers

Why Now?

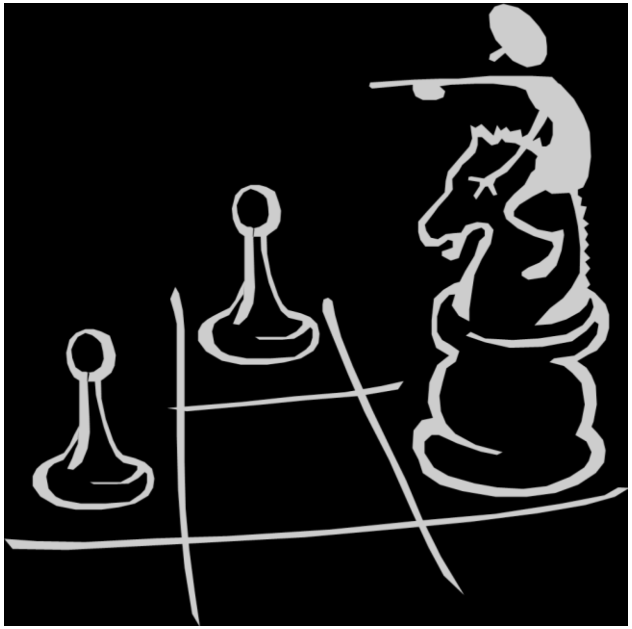
- Data is being produced
- Data is being warehoused
- The computing power is available
- The computing power is affordable
- The competitive pressures are strong
- Commercial products are available

Data Mining works with Warehouse Data



- Data Warehousing provides the Enterprise with a memory

- Data Mining provides the Enterprise with intelligence



Mining market

- Around 20 to 30 mining tool vendors
- Major players:
  - Clementine,
  - IBM's Intelligent Miner,
  - SGI's MineSet,
  - SAS's Enterprise Miner.
- All pretty much the same set of tools
- Many embedded products: fraud detection, electronic commerce applications

OLAP Mining integration

- OLAP (On Line Analytical Processing)
  - Fast interactive exploration of multidim. aggregates.
  - Heavy reliance on manual operations for analysis:
  - Tedious and error-prone on large multidimensional data
- Ideal platform for vertical integration of mining but needs to be interactive instead of batch.

State of art in mining OLAP integration

- Decision trees [Information discovery, Cognos]
  - find factors influencing high profits
- Clustering [Pilot software]
  - segment customers to define hierarchy on that dimension
- Time series analysis: [Seagate's Holos]
    - Query for various shapes along time: eg. spikes, outliers etc
- Multi-level Associations [Han et al.]
  - find association between members of dimensions

Vertical integration: Mining on the web

- Web log analysis for site design:
  - what are popular pages,
  - what links are hard to find.
- Electronic stores sales enhancements:
  - recommendations, advertisement:
  - Collaborative filtering: Net perception, Wisewire
  - Inventory control: what was a shopper looking for and could not find..