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

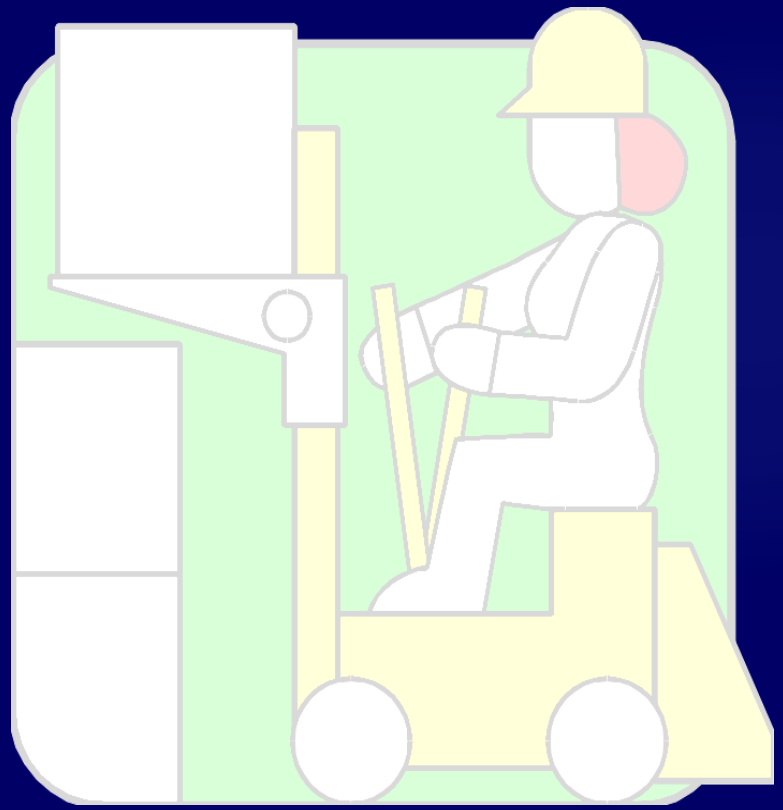
- 0. Introduction
- I. Data Warehousing
- II. Decision Support and OLAP
- III. Data Mining
- IV. Looking Ahead

- Demos and Labs

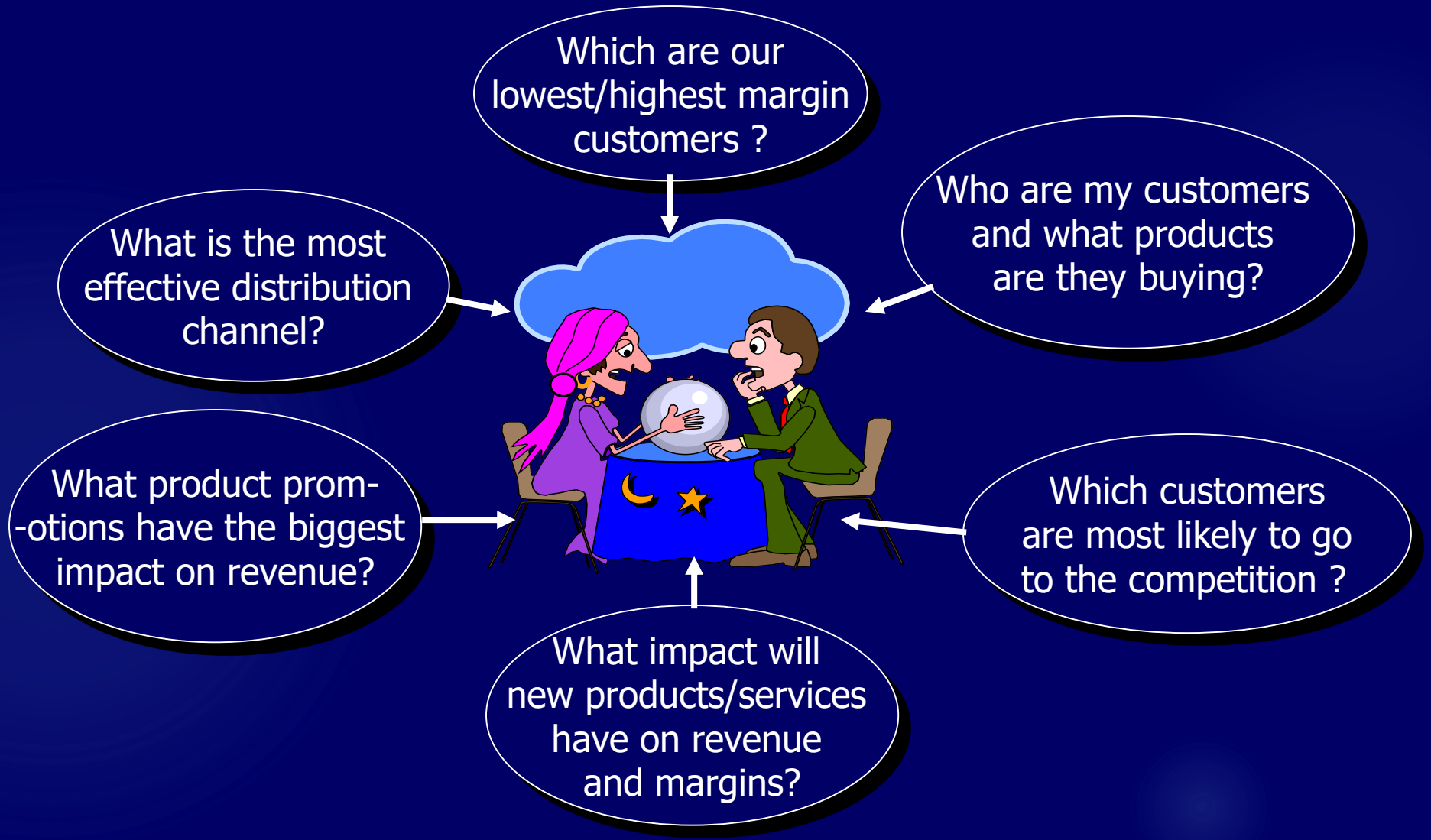
0. Introduction

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

- demos, labs

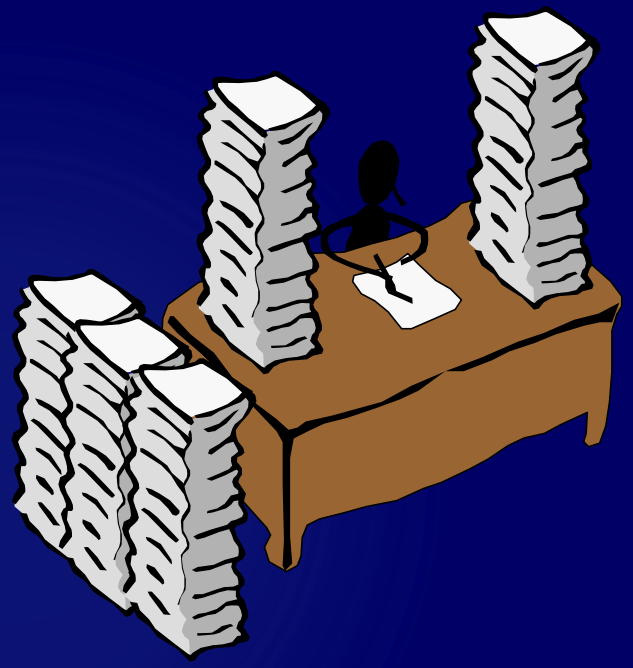


A producer wants to know….



Data, Data everywhere yet …

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

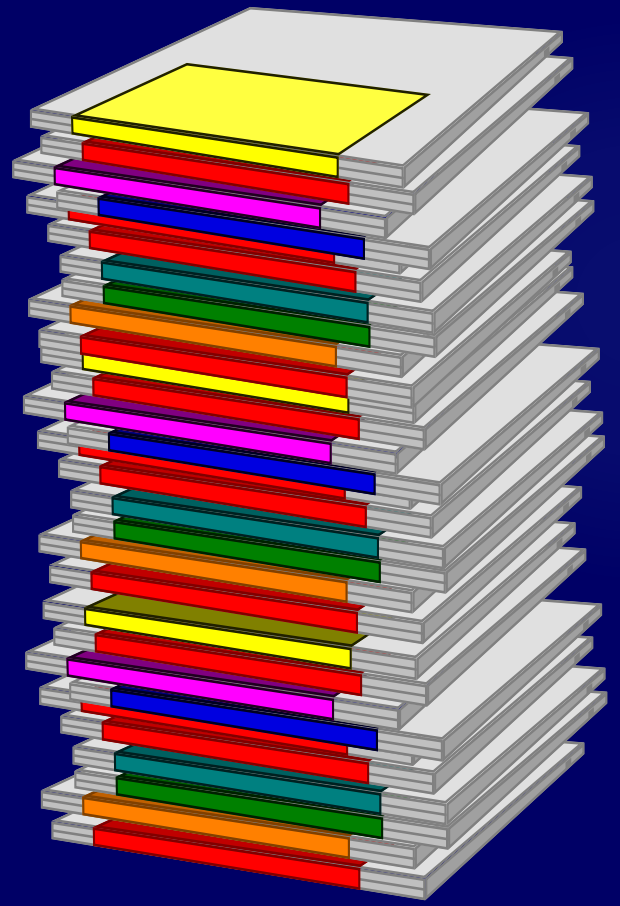


- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

What is a Data Warehouse?

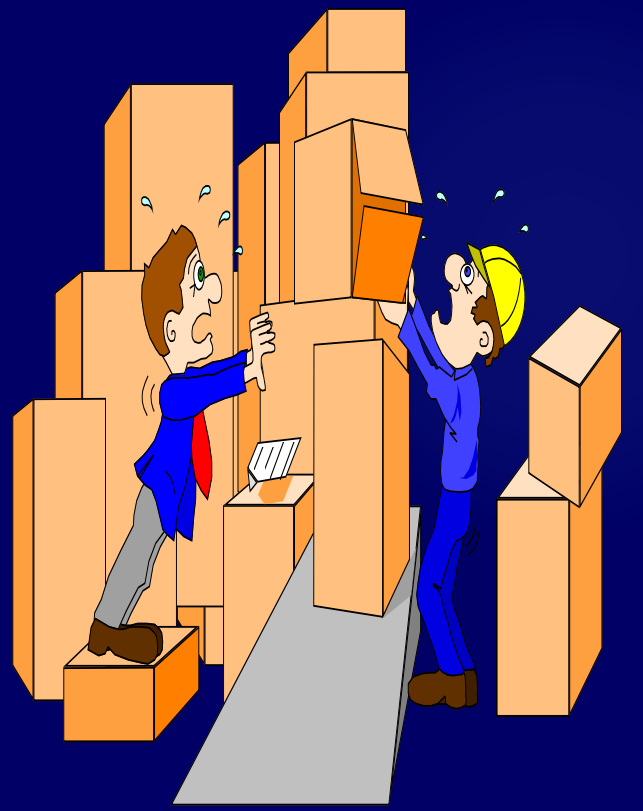
- A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

- [Barry Devlin]

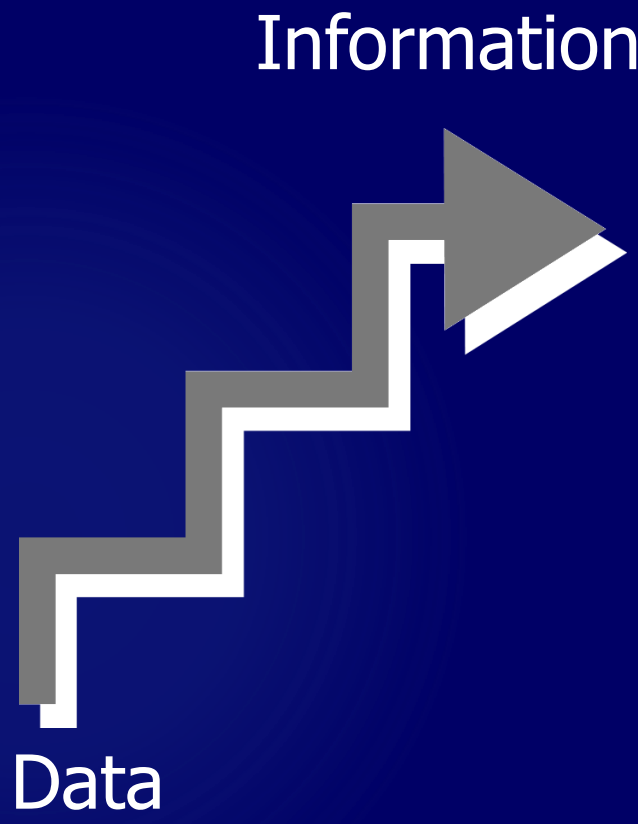


What are the users saying...

- 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?

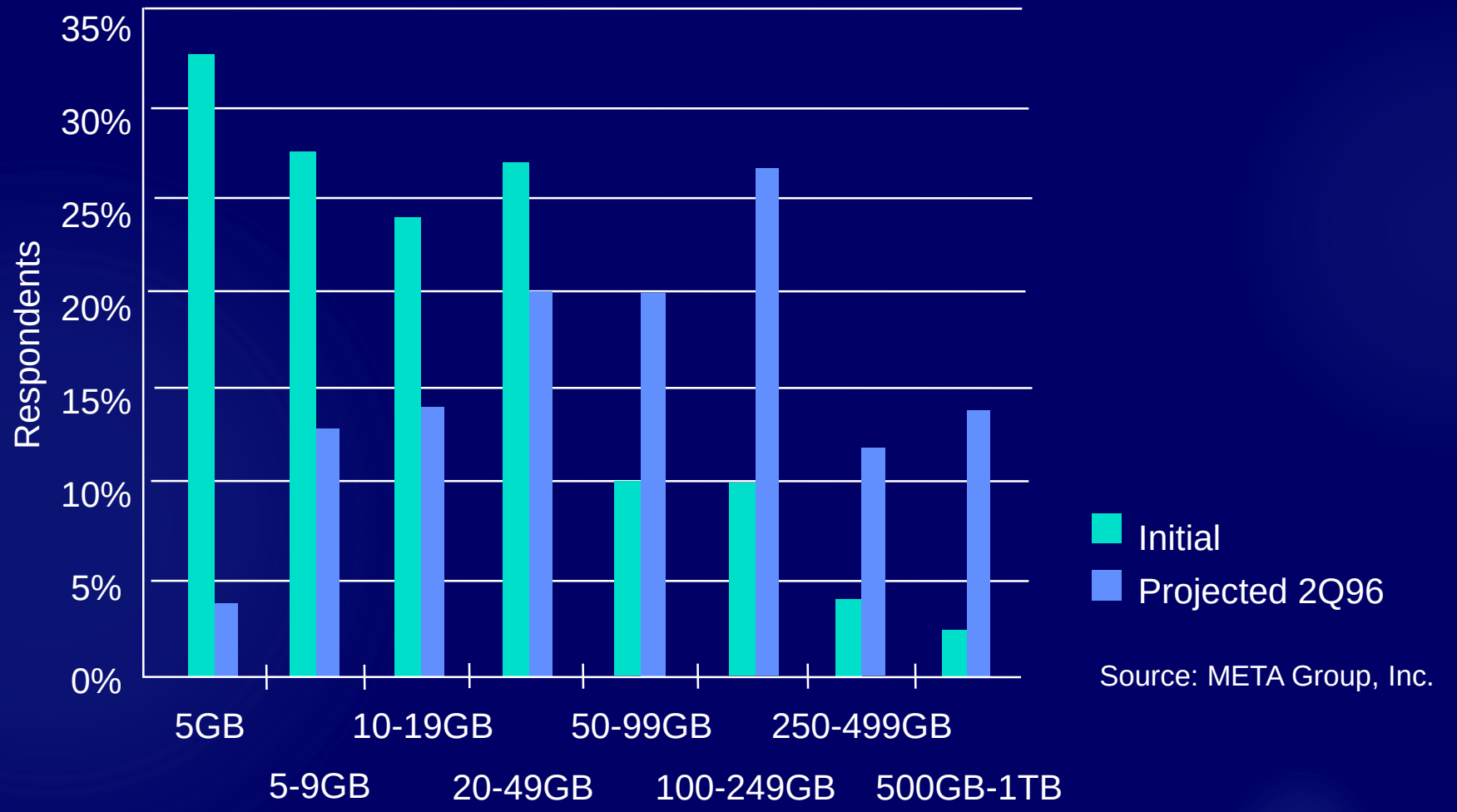


- A process of transforming data into information and making it available to users in a timely enough manner to make a difference
- [Forrester Research, April 1996]

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

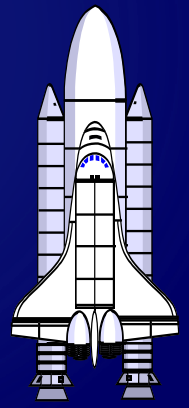
Data Warehouse

- 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

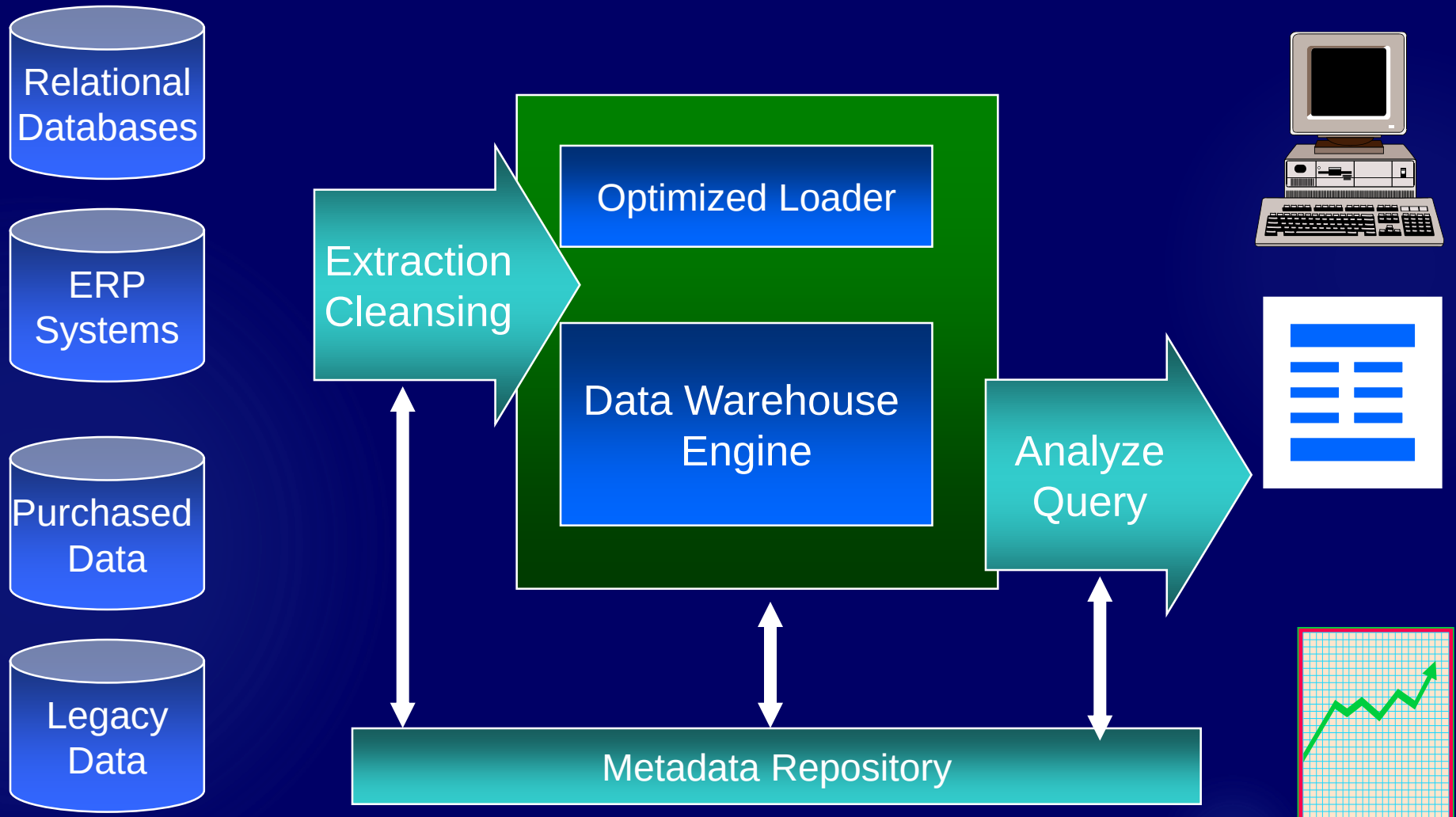
Tourists: Browse information harvested by farmers

Farmers: Harvest information from known access paths



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

Data Warehouse Architecture



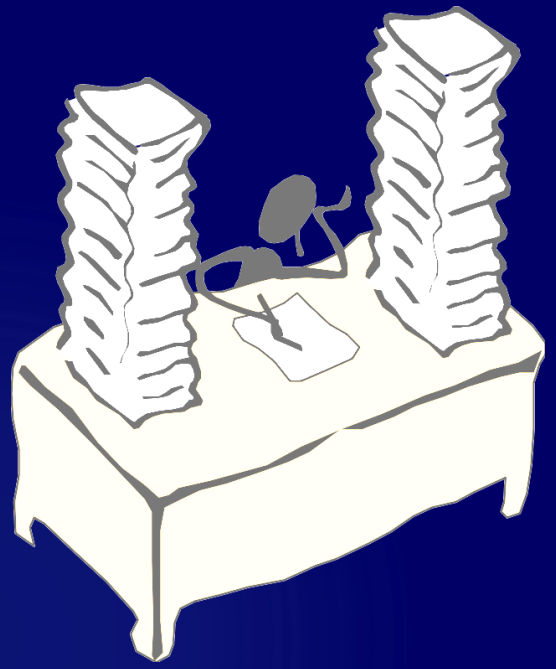
Data Warehouse for Decision Support & OLAP

- 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?

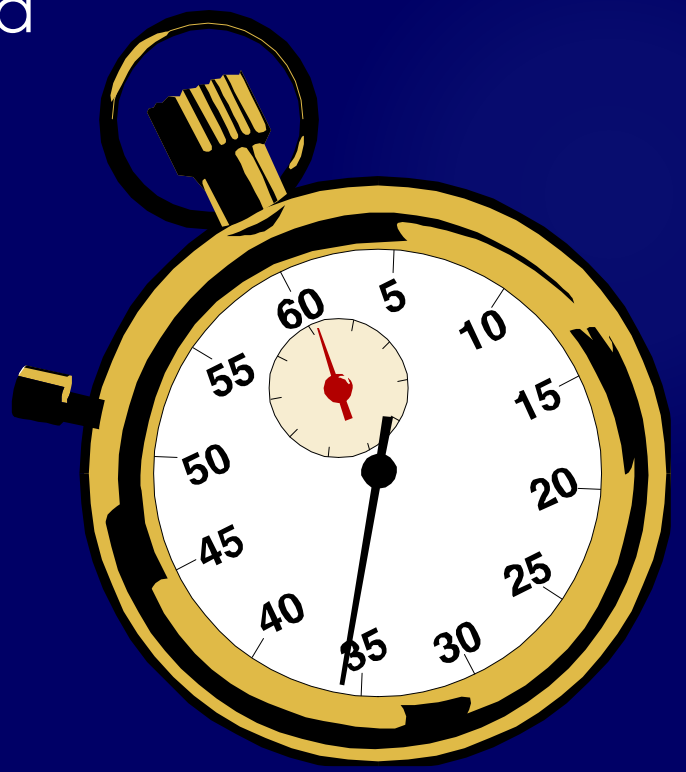
- They are OLTP systems
- Run mission critical applications
- Need to work with stringent performance requirements for routine tasks
- Used to run a business!

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

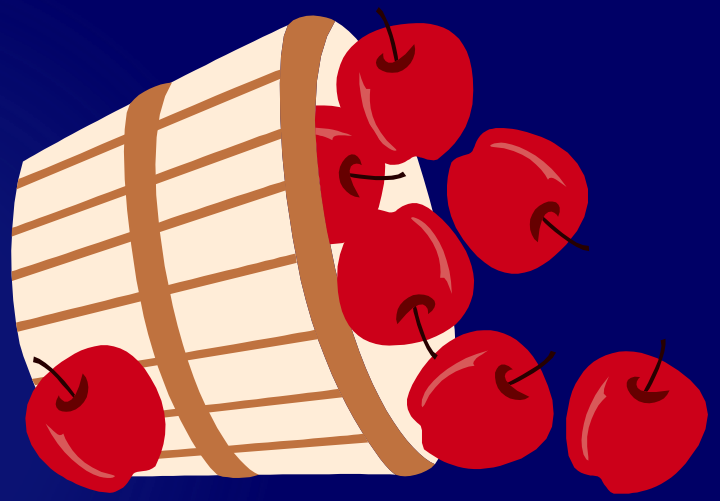
- Run the business in real time
- Based on up-to-the-second data
- Optimized to handle large numbers of simple read/write transactions
- Optimized for fast response to predefined transactions
- Used by people who deal with customers, products -- clerks, salespeople etc.
- They are increasingly used by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

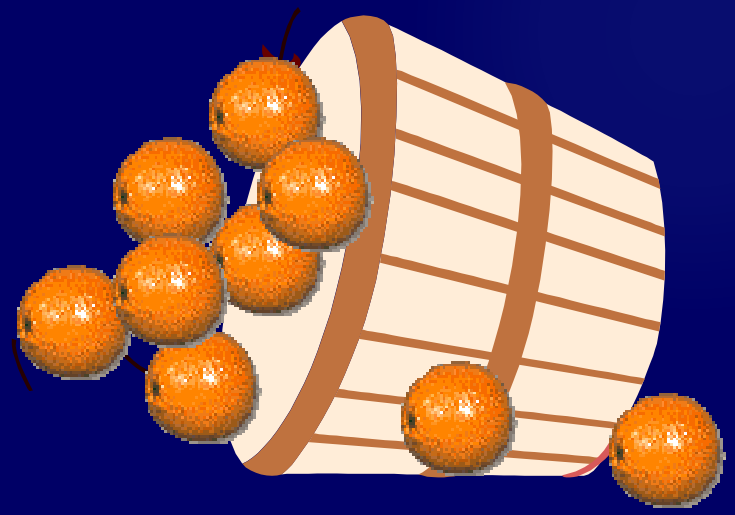


Examples of Operational Data

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

Vendor

Product

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*

Course Overview

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

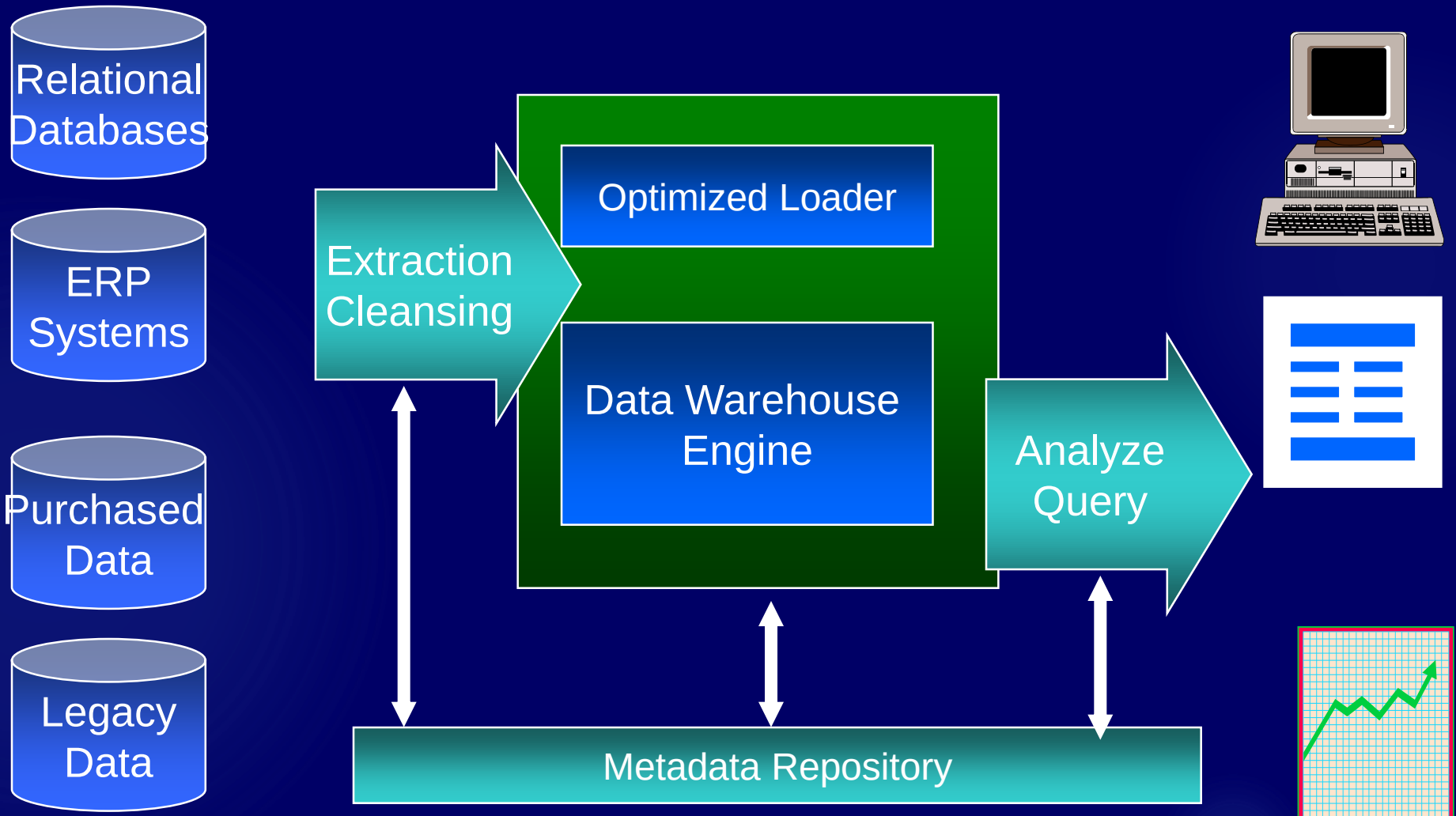
- Demos and Labs

I. Data Warehouses: Architecture, Design & Construction

- DW Architecture
- Loading, refreshing
- Structuring/Modeling
- DWs and Data Marts
- Query Processing

- demos, labs

Data Warehouse Architecture



Components of the Warehouse

- Data Extraction and Loading
- The Warehouse
- Analyze and Query -- OLAP Tools
- Metadata

- Data Mining tools

Loading the Warehouse

- Cleaning the data before it is loaded

Source Data

- Typically host based, legacy applications
  - Customized applications, COBOL, 3GL, 4GL
- Point of Contact Devices
  - POS, ATM, Call switches
- External Sources
  - Nielsen's, Acxiom, CMIE, Vendors, Partners

Data Quality - The Reality

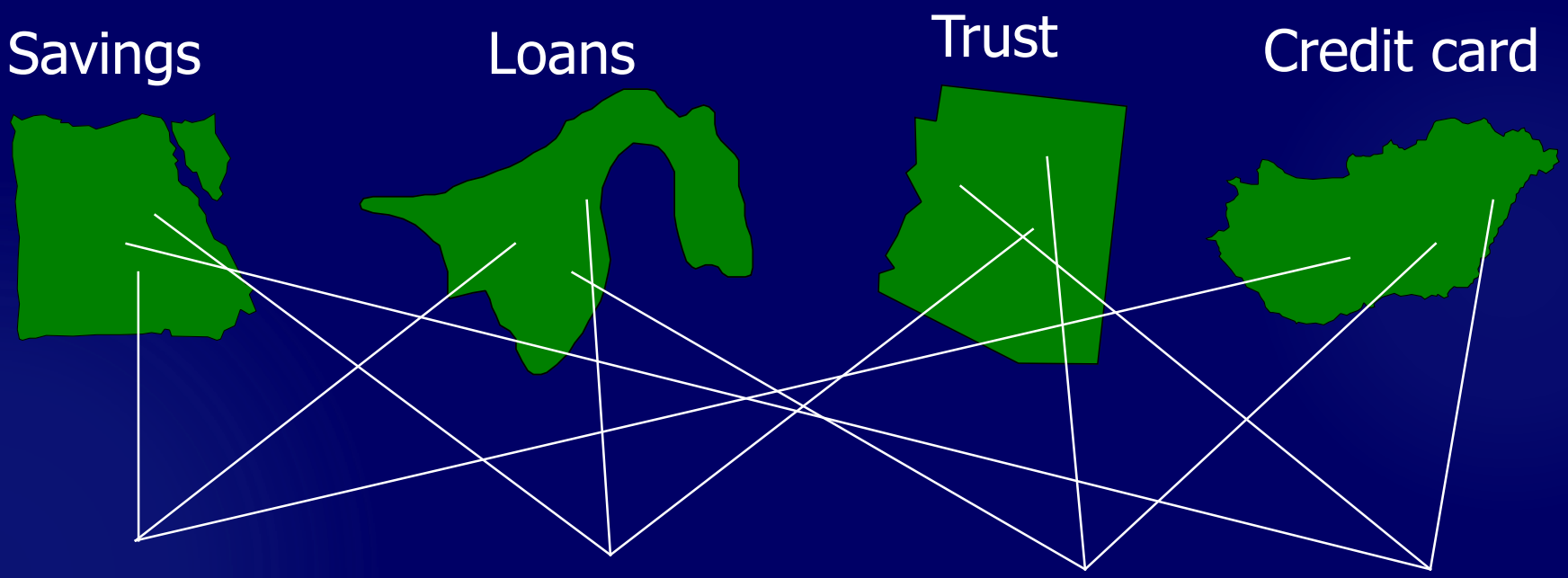
- Tempting to think creating a data warehouse is simply extracting operational data and entering into a data warehouse

- Nothing could be farther from the truth
- Warehouse data comes from disparate questionable sources

Data Quality - The Reality

- Legacy systems no longer documented
- Outside sources with questionable quality procedures
- Production systems with no built in integrity checks and no integration
  - Operational systems are usually designed to solve a specific business problem and are rarely developed to a a corporate plan
    - "And get it done quickly, we do not have time to worry about corporate standards..."

Data Integration Across Sources



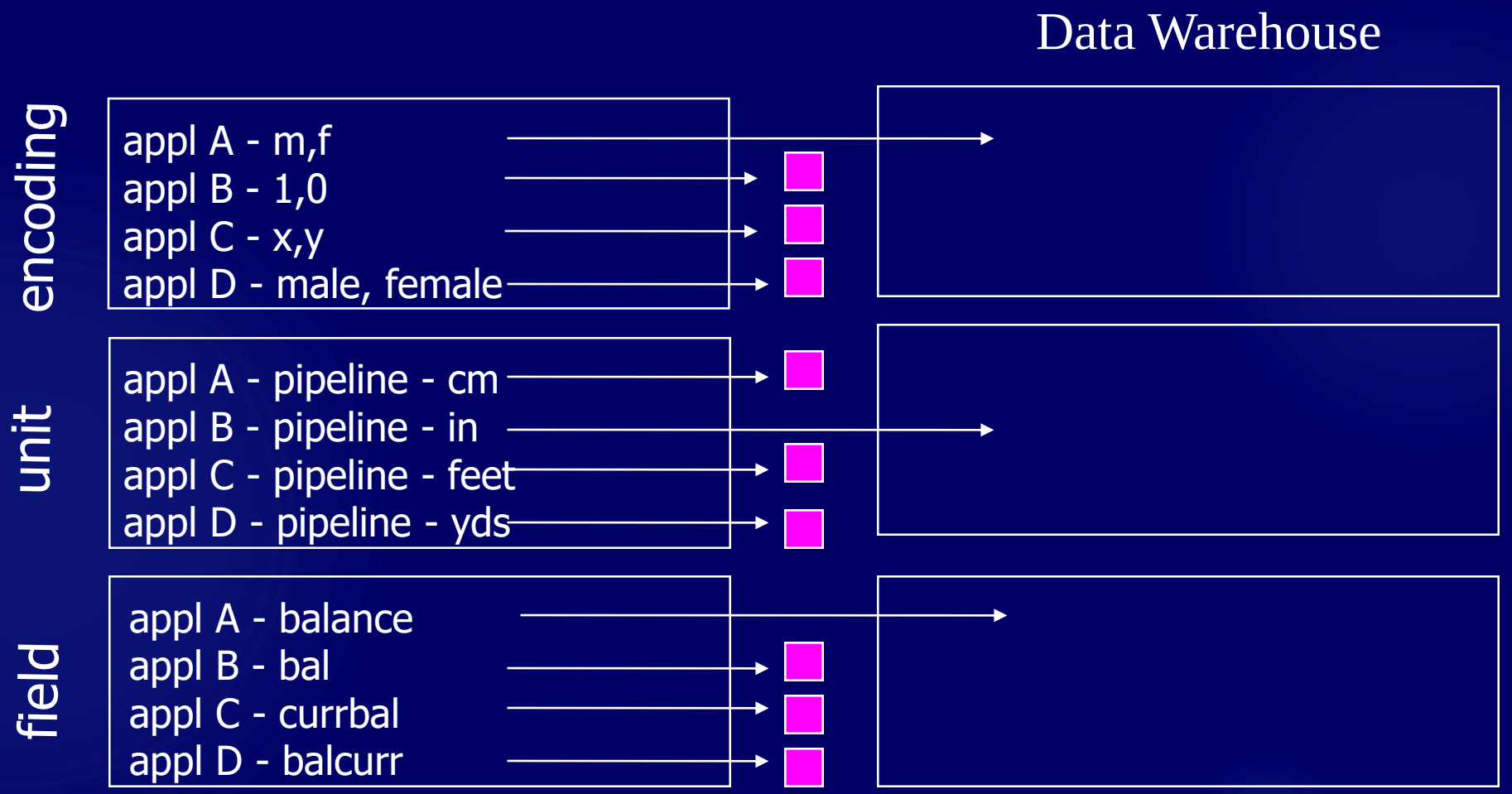
Same data different name

Different data Same name

Data found here nowhere else

Different keys same data

Data Transformation Example



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

- Extracting
  - Capture of data from operational source in "as is" status
  - Sources for data generally in legacy mainframes in VSAM, IMS, IDMS, DB2; more data today in relational databases on Unix
- Conditioning
  - The conversion of data types from the source to the target data store (warehouse) -- always a relational database

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 lakh catalogues at Rs. 50 each costs Rs. 50 lakhs. A 2% savings would save Rs. 1 lakh.

Data Transformation Terms

- Enrichment
  - Bring data from external sources to augment/enrich operational data. Data sources include Dunn and Bradstreet, A. C. Nielsen, CMIE, IMRA etc...
- Scoring
  - computation of a probability of an event. e.g..., chance that a customer will defect to AT&T from MCI, chance that a customer is likely to buy a new product

Loads

- After extracting, scrubbing, cleaning, validating etc. need to load the data into the warehouse
- Issues
  - huge volumes of data to be loaded
  - small time window available when warehouse can be taken off line (usually nights)
  - when to build index and summary tables
  - allow system administrators to monitor, cancel, resume, change load rates
  - Recover gracefully -- restart after failure from where you were and without loss of data integrity

Load Techniques

- Use SQL to append or insert new data
  - record at a time interface
  - will lead to random disk I/O's
- Use batch load utility

Load Taxonomy

- Incremental versus Full loads
- Online versus Offline loads

Refresh

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

- Full Extract from base tables
  - read entire source table: too expensive
  - maybe the only choice for legacy systems

How To Detect Changes

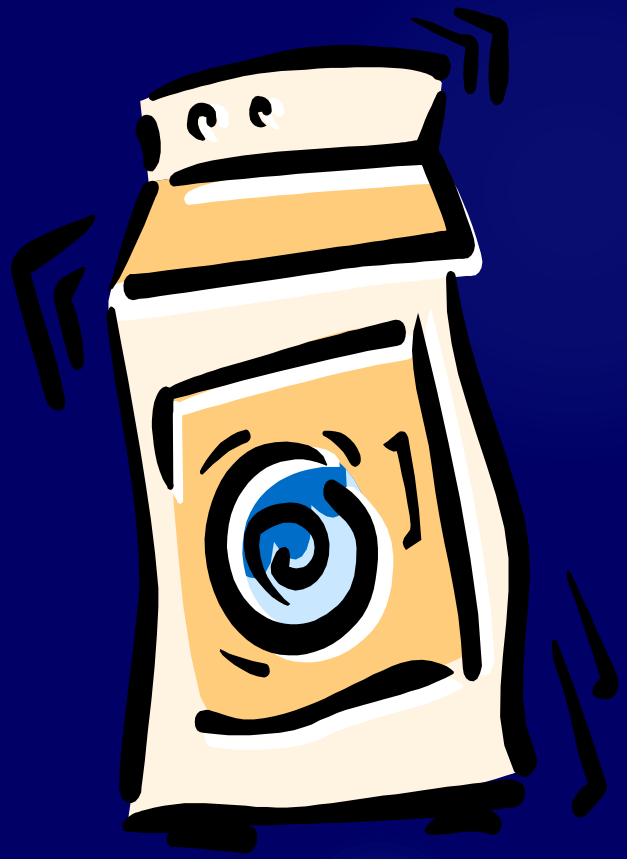
- Create a snapshot log table to record ids of updated rows of source data and timestamp
- Detect changes by:
  - Defining after row triggers to update snapshot log when source table changes
  - Using regular transaction log to detect changes to source data

Data Extraction and Cleansing

- Extract data from existing operational and legacy data
- Issues:
  - Sources of data for the warehouse
  - Data quality at the sources
  - Merging different data sources
  - Data Transformation
  - How to propagate updates (on the sources) to the warehouse
  - Terabytes of data to be loaded

Scrubbing Data

- Sophisticated transformation tools.
- Used for cleaning the quality of data
- Clean data is vital for the success of the warehouse
- Example
  - Seshadri, Sheshadri, Sesadri, Seshadri S., Srinivasan Seshadri, etc. are the same person



Scrubbing Tools

- Apertus -- Enterprise/Integrator
- Vality -- IPE
- Postal Soft

Structuring/Modeling Issues

Data -- Heart of the Data Warehouse

- Heart of the data warehouse is the data itself!
- Single version of the truth
- Corporate memory
- Data is organized in a way that represents business -- subject orientation

Data Warehouse Structure

- Subject Orientation -- customer, product, policy, account etc... A subject may be implemented as a set of related tables. E.g., customer may be five tables

Data Warehouse Structure

Time is part of key of each table

- base customer (1985-87)
  - custid, from date, to date, name, phone, dob
- base customer (1988-90)
  - custid, from date, to date, name, credit rating, employer
- customer activity (1986-89) -- monthly summary
- customer activity detail (1987-89)
  - custid, activity date, amount, clerk id, order no
- customer activity detail (1990-91)
  - custid, activity date, amount, line item no, order no

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

- Can not answer some questions with summarized data
  - Did Anand call Seshadri last month? Not possible to answer if total duration of calls by Anand over a month is only maintained and individual call details are not.
- Detailed data too voluminous

Granularity in Warehouse

- Tradeoff 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

Vertical Partitioning

Acct. No	Name	Balance	Date Opened	Interest Rate	Address

Frequently accessed

Acct. No	Balance

Rarely accessed

Acct. No	Name	Date Opened	Interest Rate	Address

Smaller table
and so less I/O

Derived Data

- Introduction of derived (calculated data) may often help
- Have seen this in the context of dual levels of granularity
- Can keep auxiliary views and indexes to speed up query processing

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

Dimension Tables

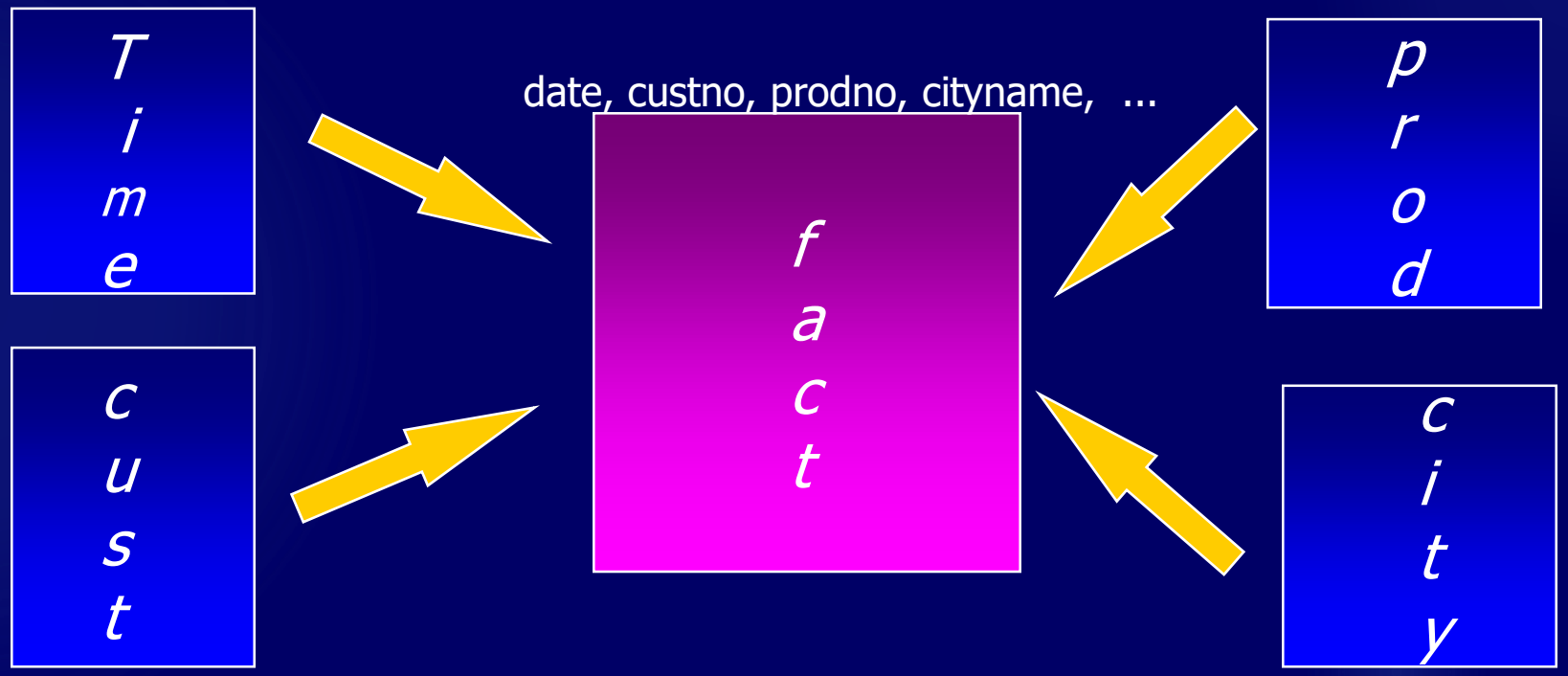
- Dimension tables
  - Define business in terms already familiar to users
  - Wide rows with lots of descriptive text
  - Small tables (about a million rows)
  - Joined to fact table by a foreign key
  - heavily indexed
  - typical dimensions
    - time periods, geographic region (markets, cities), products,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

Fact Table

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

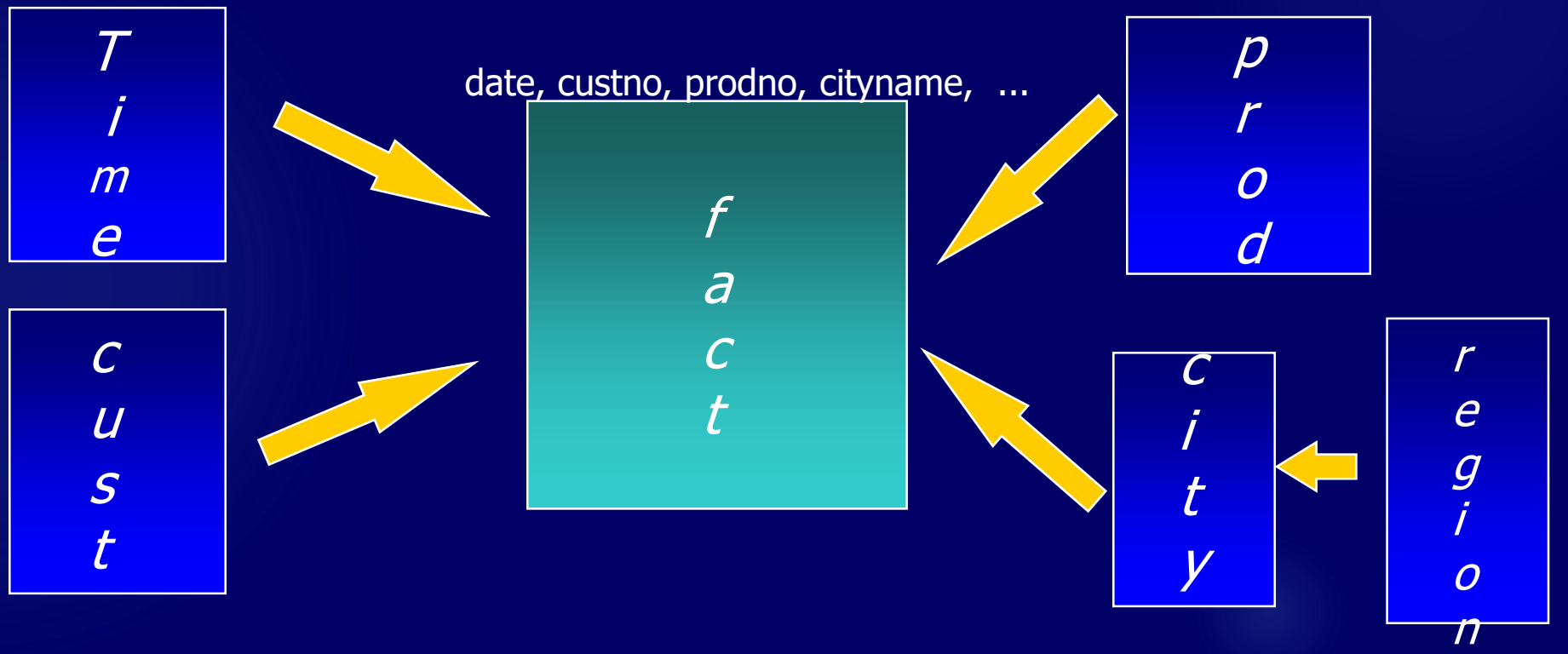
Star Schema

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



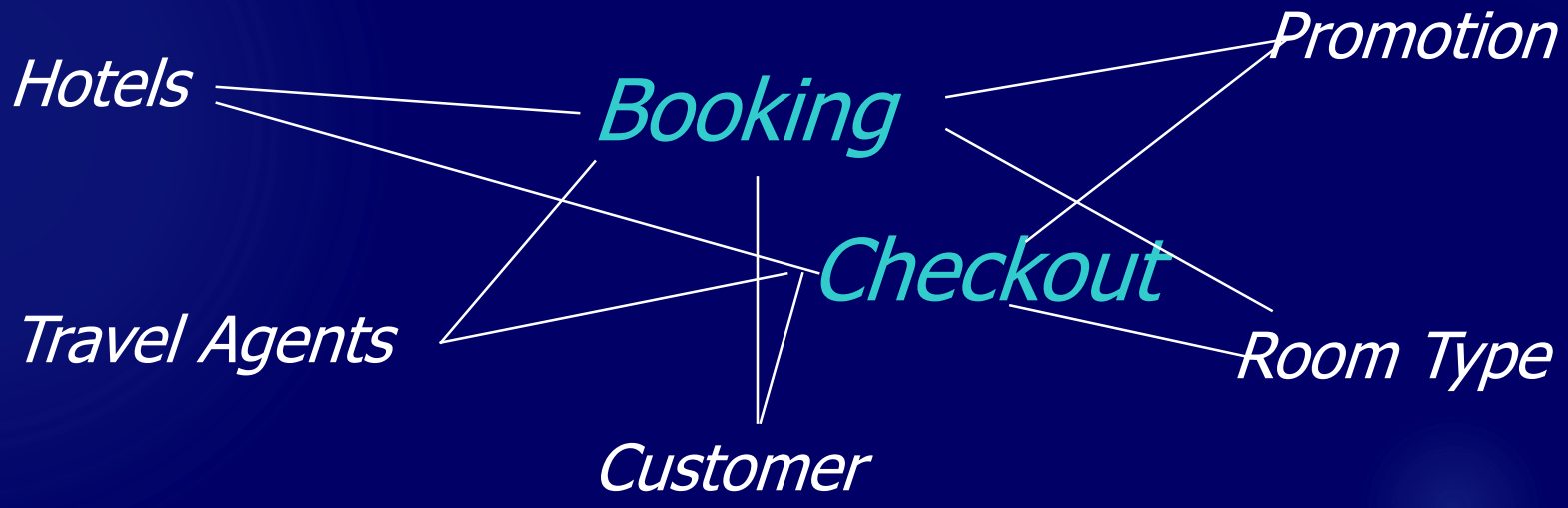
Snowflake schema

- 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



De-normalization

- Normalization in a data warehouse may lead to lots of small tables
- Can lead to excessive I/O's since many tables have to be accessed
- De-normalization is the answer especially since updates are rare

Creating Arrays

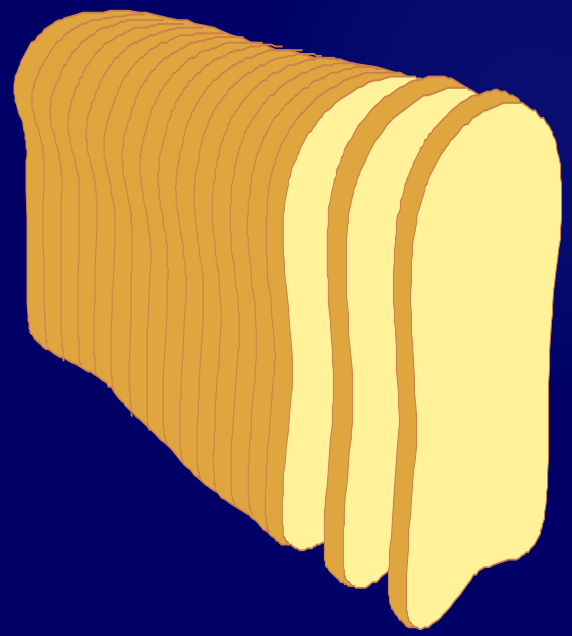
- Many times each occurrence of a sequence of data is in a different physical location
- Beneficial to collect all occurrences together and store as an array in a single row
- Makes sense only if there are a stable number of occurrences which are accessed together
- In a data warehouse, such situations arise naturally due to time based orientation
  - can create an array by month

Selective Redundancy

- Description of an item can be stored redundantly with order table -- most often item description is also accessed with order table
- Updates have to be careful

Partitioning

- Breaking data into several physical units that can be handled separately
- Not a question of *whether* to do it in data warehouses but *how* to do it
- Granularity and partitioning are key to effective implementation of a warehouse



Why Partition?

- Flexibility in managing data
- Smaller physical units allow
  - easy restructuring
  - free indexing
  - sequential scans if needed
  - easy reorganization
  - easy recovery
  - easy monitoring

Criterion for Partitioning

- Typically partitioned by
  - date
  - line of business
  - geography
  - organizational unit
  - any combination of above

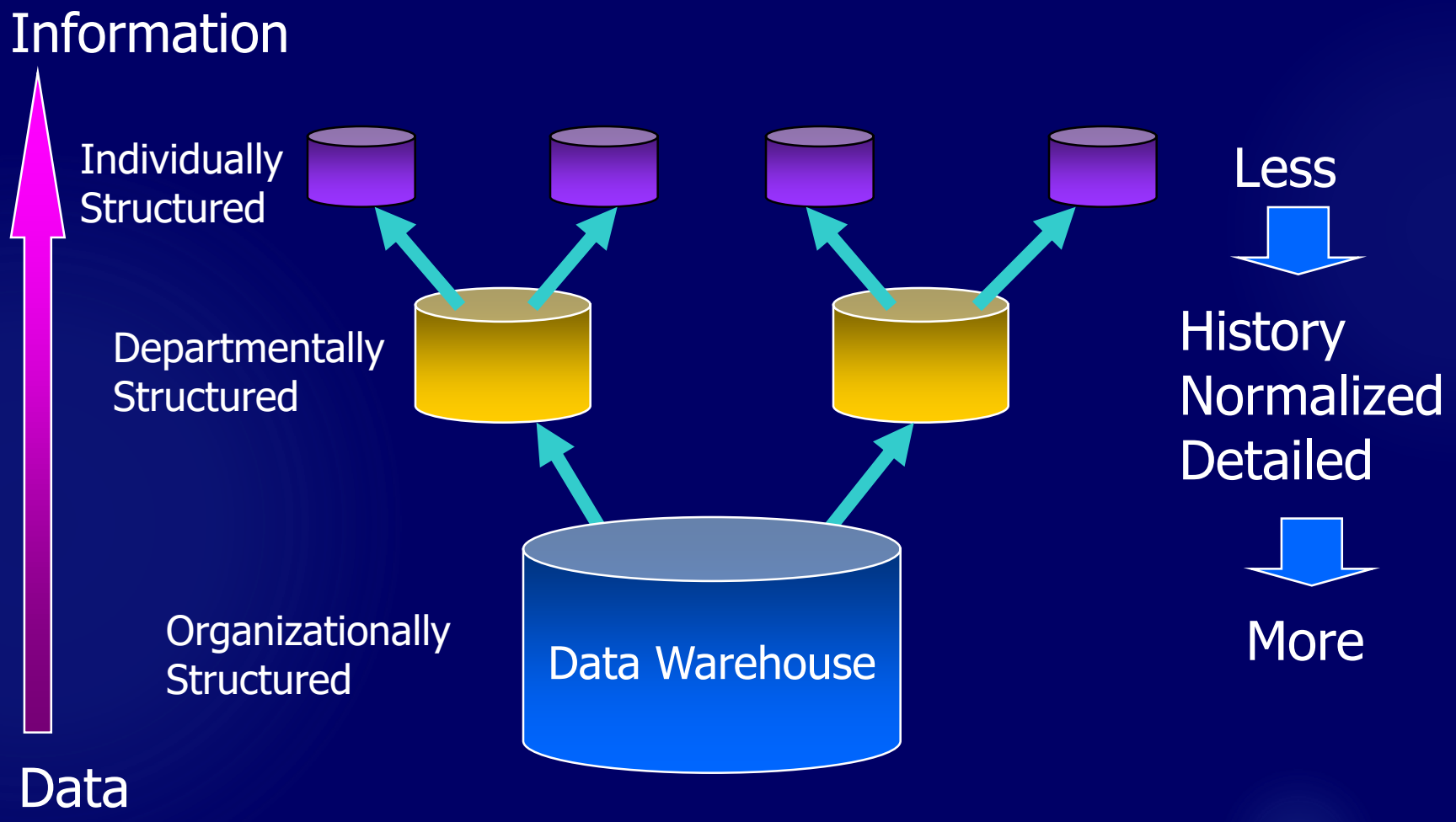
Where to Partition?

- Application level or DBMS level
- Makes sense to partition at application level
  - Allows different definition for each year
    - Important since warehouse spans many years and as business evolves definition changes
  - Allows data to be moved between processing complexes easily

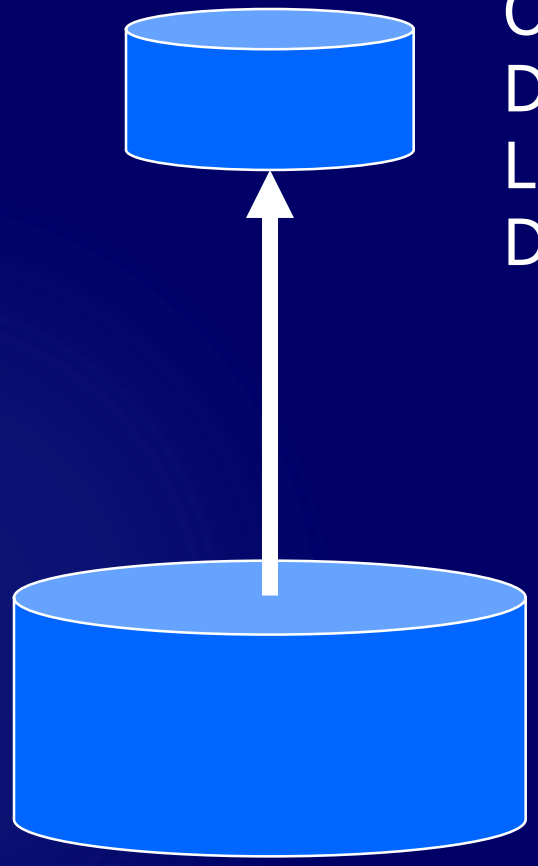
Data Warehouse vs. Data Marts

- What comes first

From the Data Warehouse to Data Marts



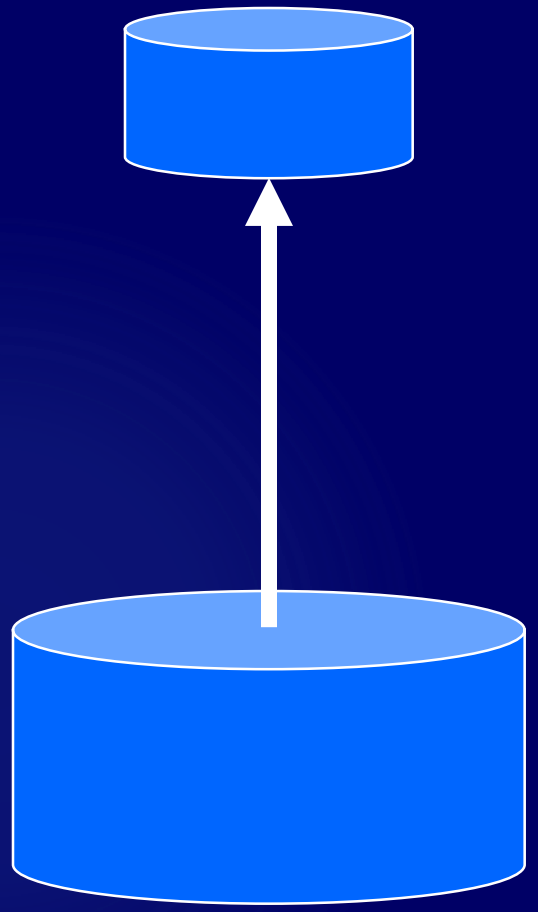
Data Warehouse and Data Marts



OLAP
Data Mart
Lightly summarized
Departmentally structured

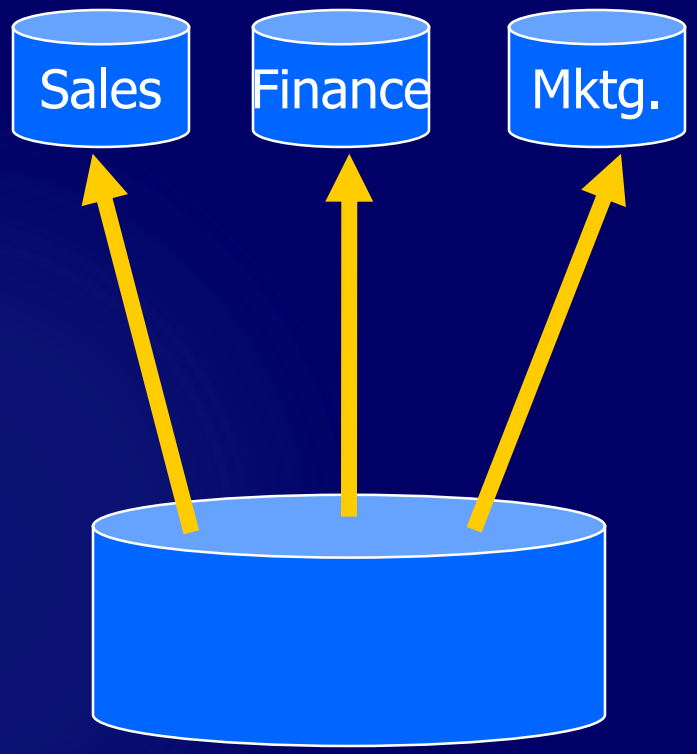
Organizationally structured
Atomic
Detailed Data Warehouse Data

Characteristics of the Departmental Data Mart



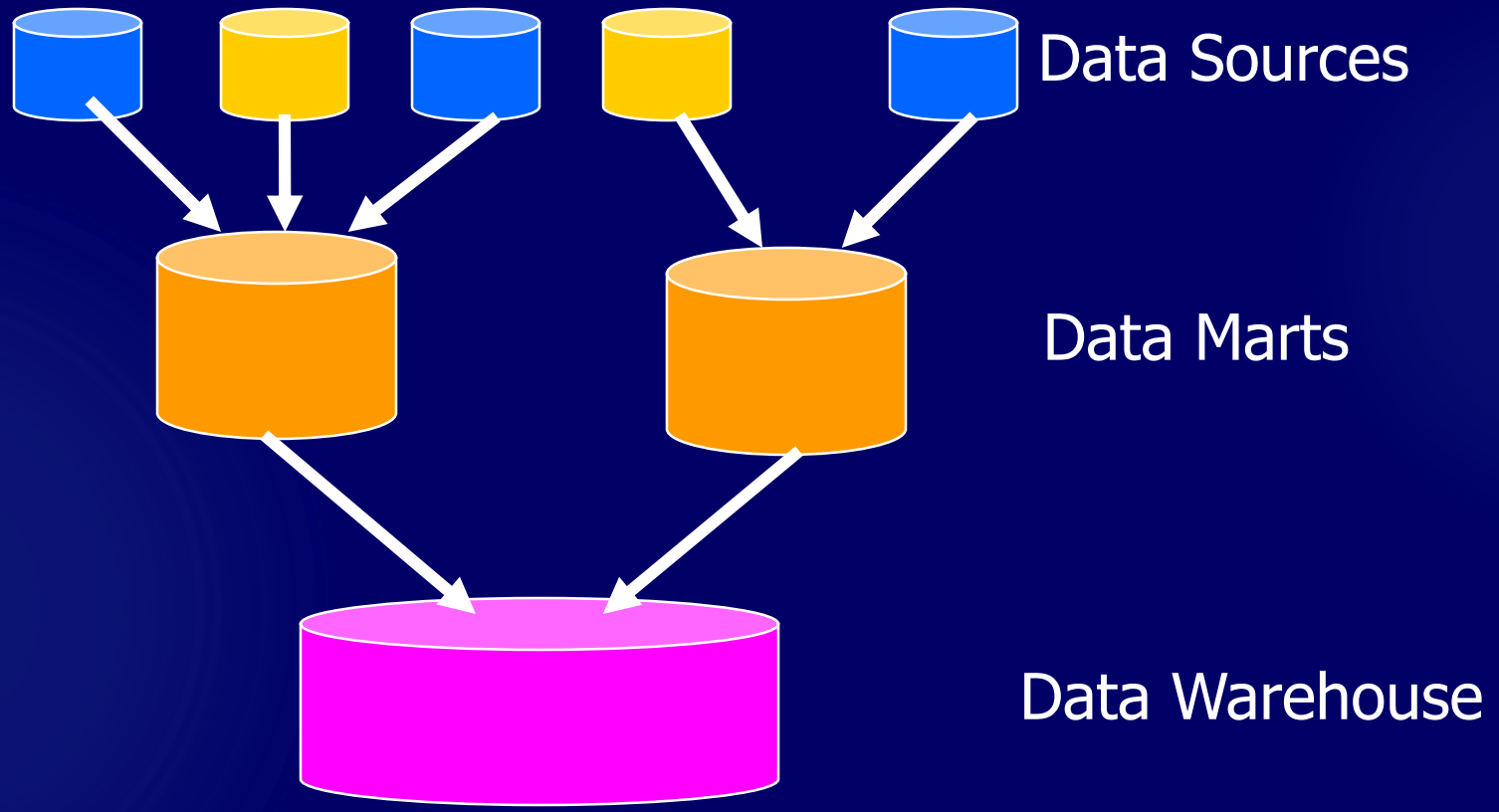
- OLAP
- Small
- Flexible
- Customized by Department
- Source is departmentally structured data warehouse

Techniques for Creating Departmental Data Mart

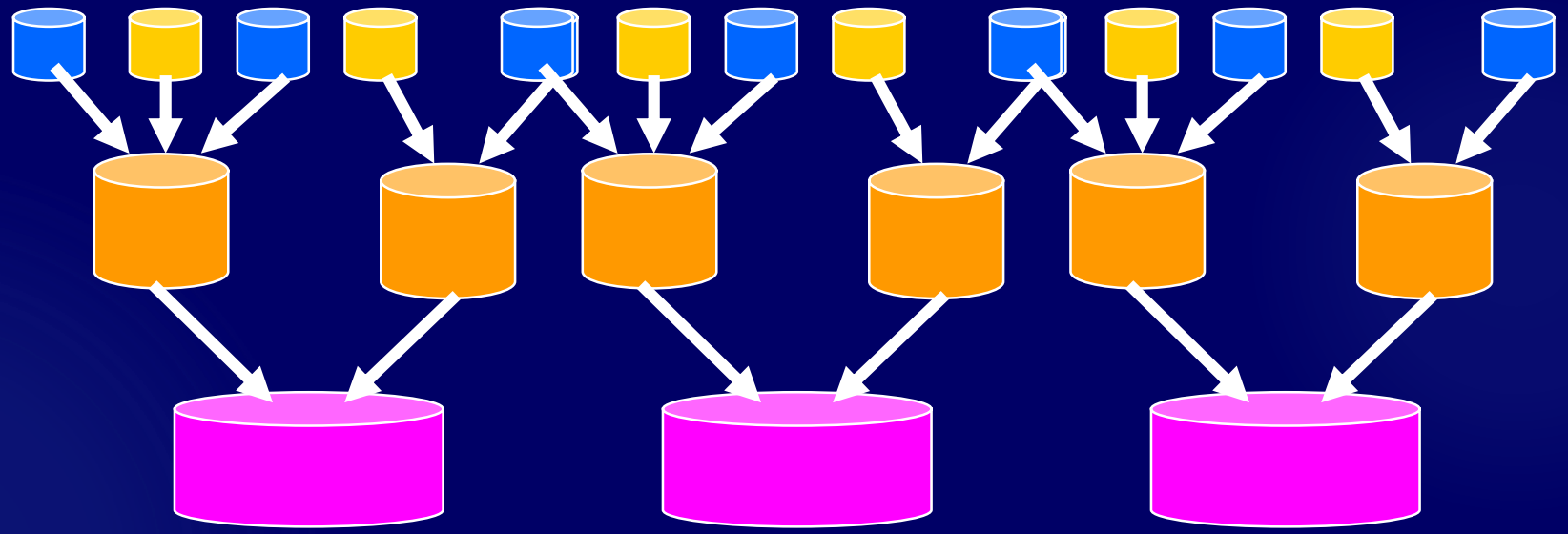


- OLAP
- Subset
- Summarized
- Superset
- Indexed
- Arrayed

Data Mart Centric



Problems with Data Mart Centric Solution



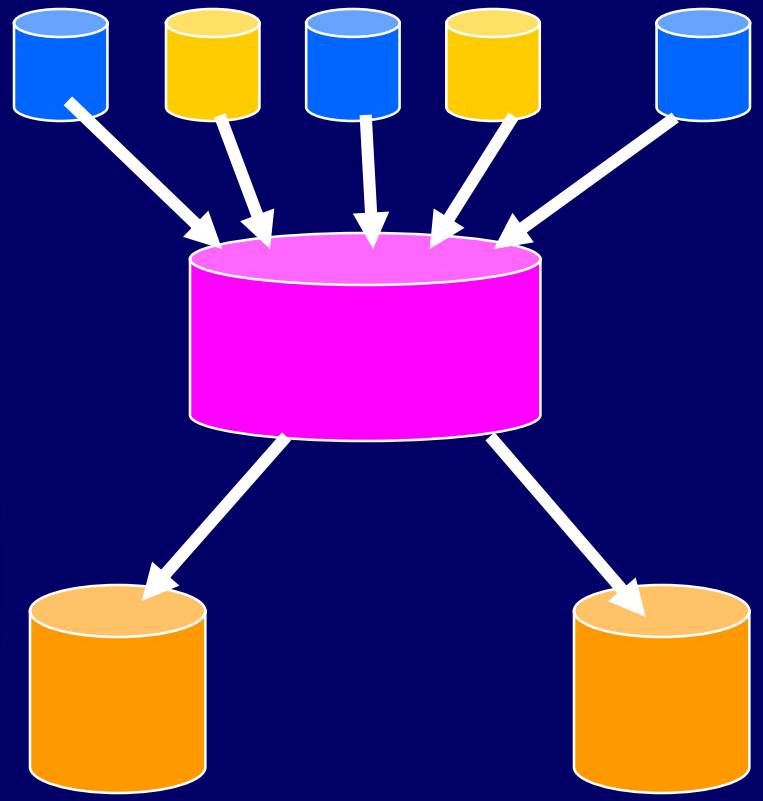
If you end up creating multiple warehouses, integrating them is a problem

True Warehouse

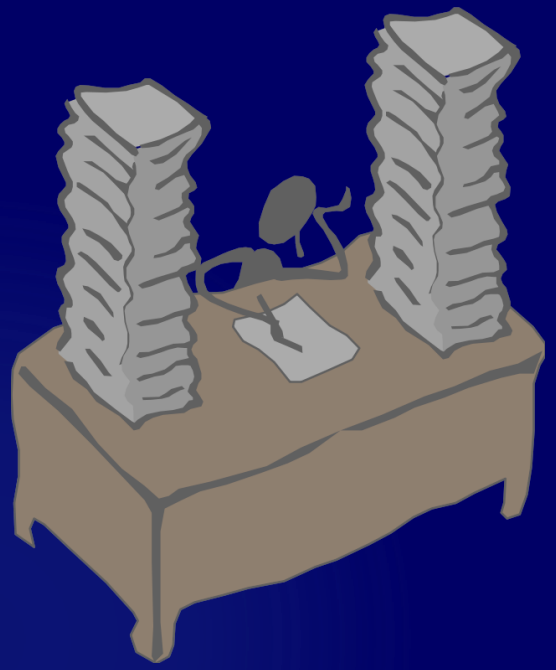
Data Sources

Data Warehouse

Data Marts



Query Processing



- Indexing

- Pre computed views/aggregates
- SQL extensions

Indexing Techniques

- Exploiting indexes to reduce scanning of data is of crucial importance
- Bitmap Indexes
- Join Indexes
- Other Issues
  - Text indexing
  - Parallelizing and sequencing of index builds and incremental updates

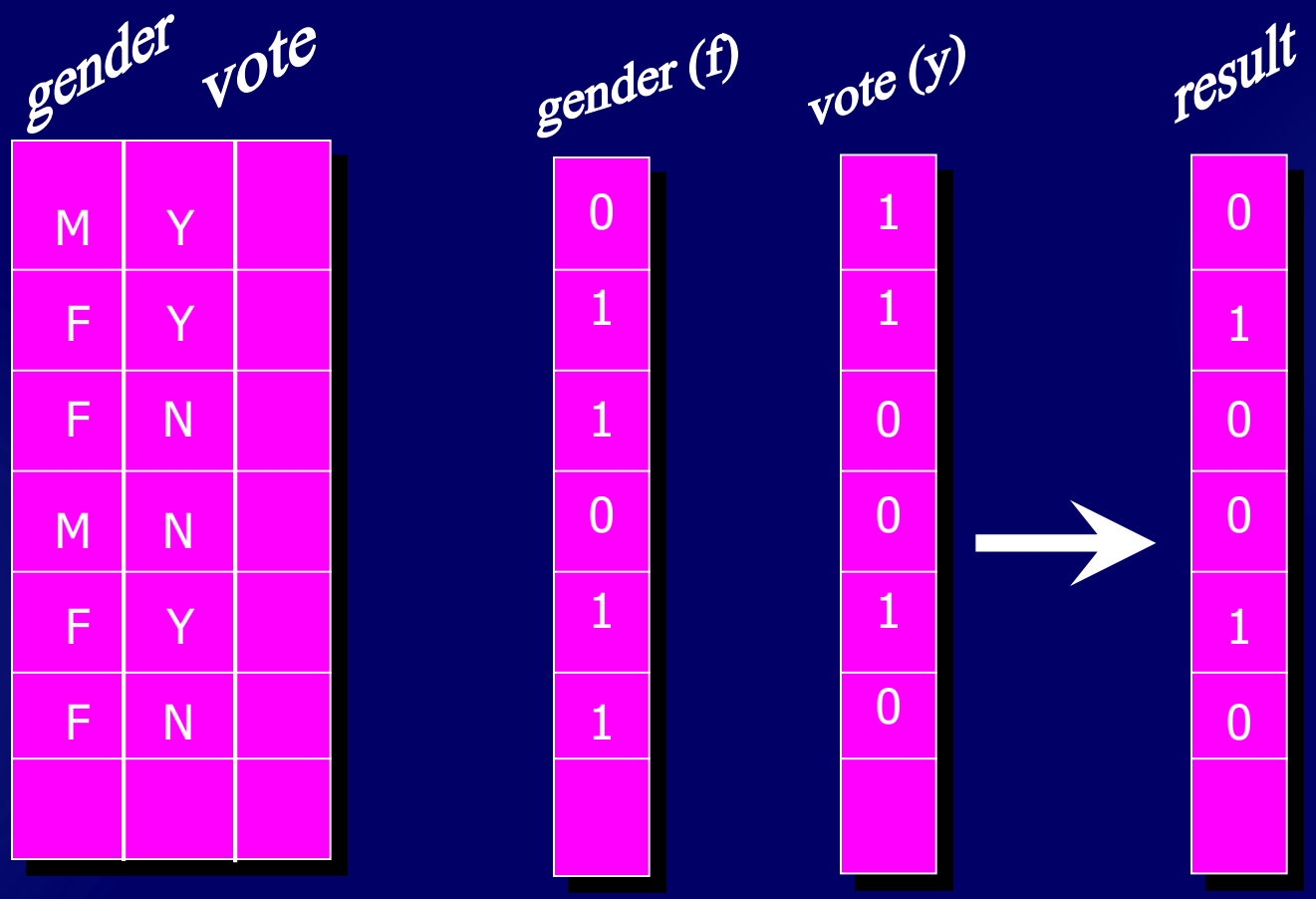
Indexing Techniques

- Bitmap index:
  - A collection of bitmaps -- one for each distinct value of the column
  - Each bitmap has N bits where N is the number of rows in the table
  - A bit corresponding to a value v for a row r is set if and only if r has the value for the indexed attribute

BitMap Indexes

- An alternative representation of RID-list
- Specially advantageous for low-cardinality domains
- Represent each row of a table by a bit and the table as a bit vector
- There is a distinct bit vector Bv for each value v for the domain
- Example: the attribute sex has values M and F. A table of 100 million people needs 2 lists of 100 million bits

Bitmap Index



Customer

Query : select * from customer where gender = 'F' and vote = 'Y'

Bit Map Index

Base Table

Cust	Region	Rating
C1	N	H
C2	S	M
C3	W	L
C4	W	H
C5	S	L
C6	W	L
C7	N	H

Region Index

Row ID	N	S	E	W
1	1	0	0	0
2	0	1	0	0
3	0	0	0	1
4	0	0	0	1
5	0	1	0	0
6	0	0	0	1
7	1	0	0	0

Rating Index

Row ID	H	M	L
1	1	0	0
2	0	1	0
3	0	0	0
4	0	0	0
5	0	1	0
6	0	0	0
7	1	0	0

Customers where **Region = W** ***And*** **Rating = M**

BitMap Indexes

- Comparison, join and aggregation operations are reduced to bit arithmetic with dramatic improvement in processing time
- Significant reduction in space and I/O (30:1)
- Adapted for higher cardinality domains as well.
- Compression (e.g., run-length encoding) exploited
- Products that support bitmaps: Model 204, TargetIndex (Redbrick), IQ (Sybase), Oracle 7.3

Join Indexes

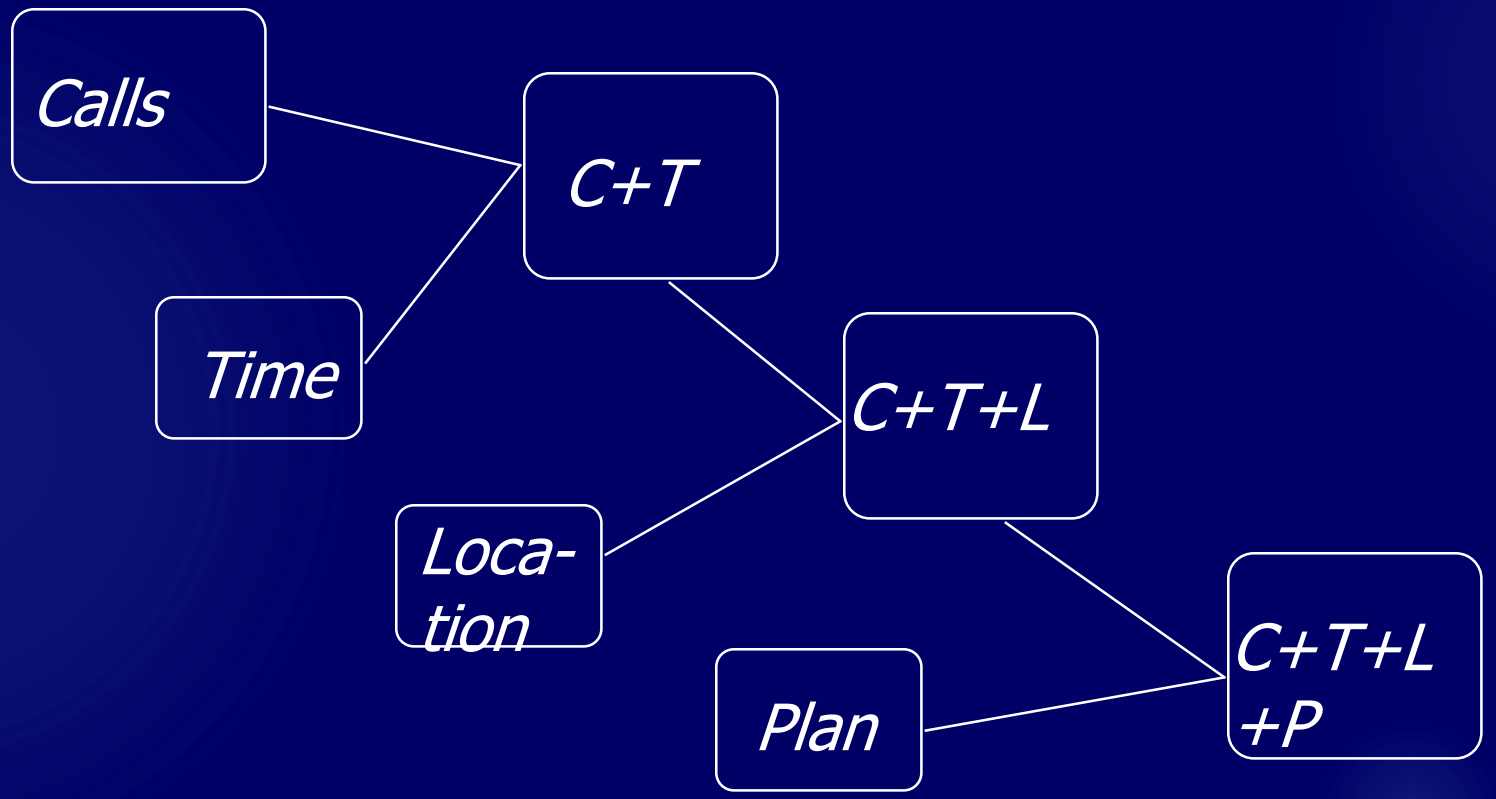
- Pre-computed joins
- A join index between a fact table and a dimension table correlates a dimension tuple with the fact tuples that have the same value on the common dimensional attribute
  - e.g., a join index on *city* dimension of *calls* fact table
  - correlates for each city the calls (in the *calls* table) from that city

Join Indexes

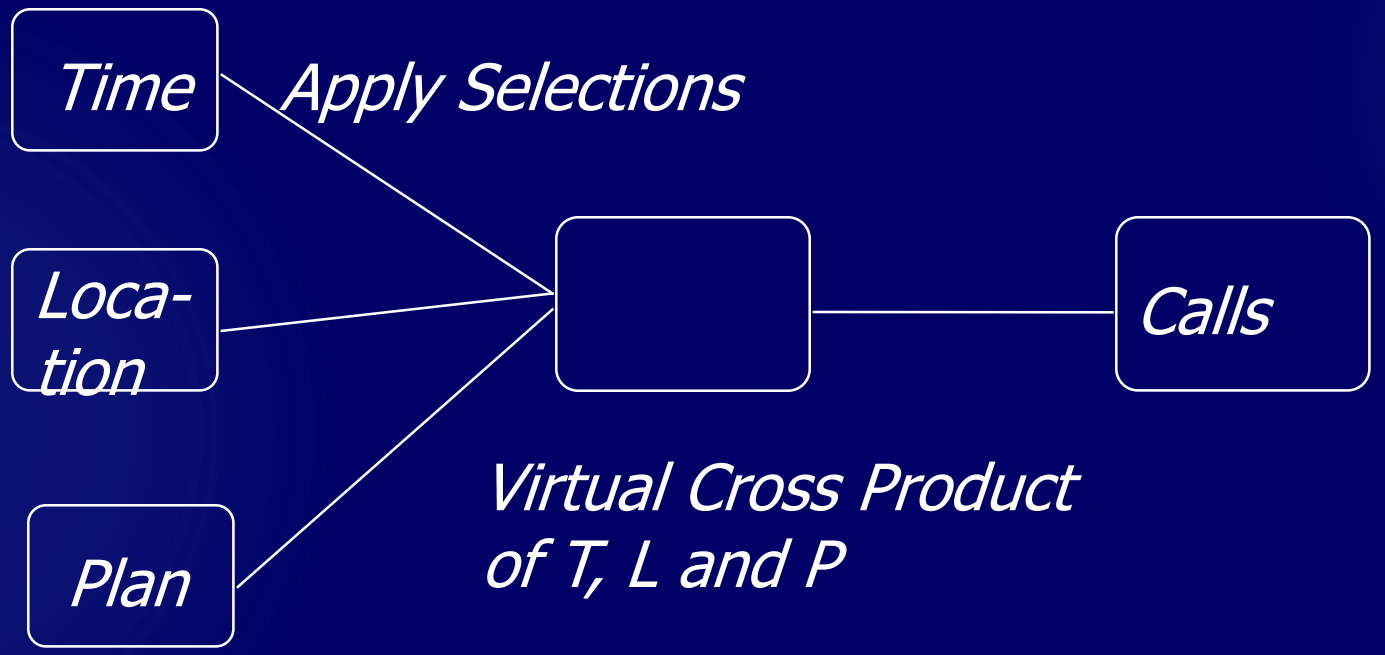
- Join indexes can also span multiple dimension tables
  - e.g., a join index on *city* and *time* dimension of *calls* fact table

Star Join Processing

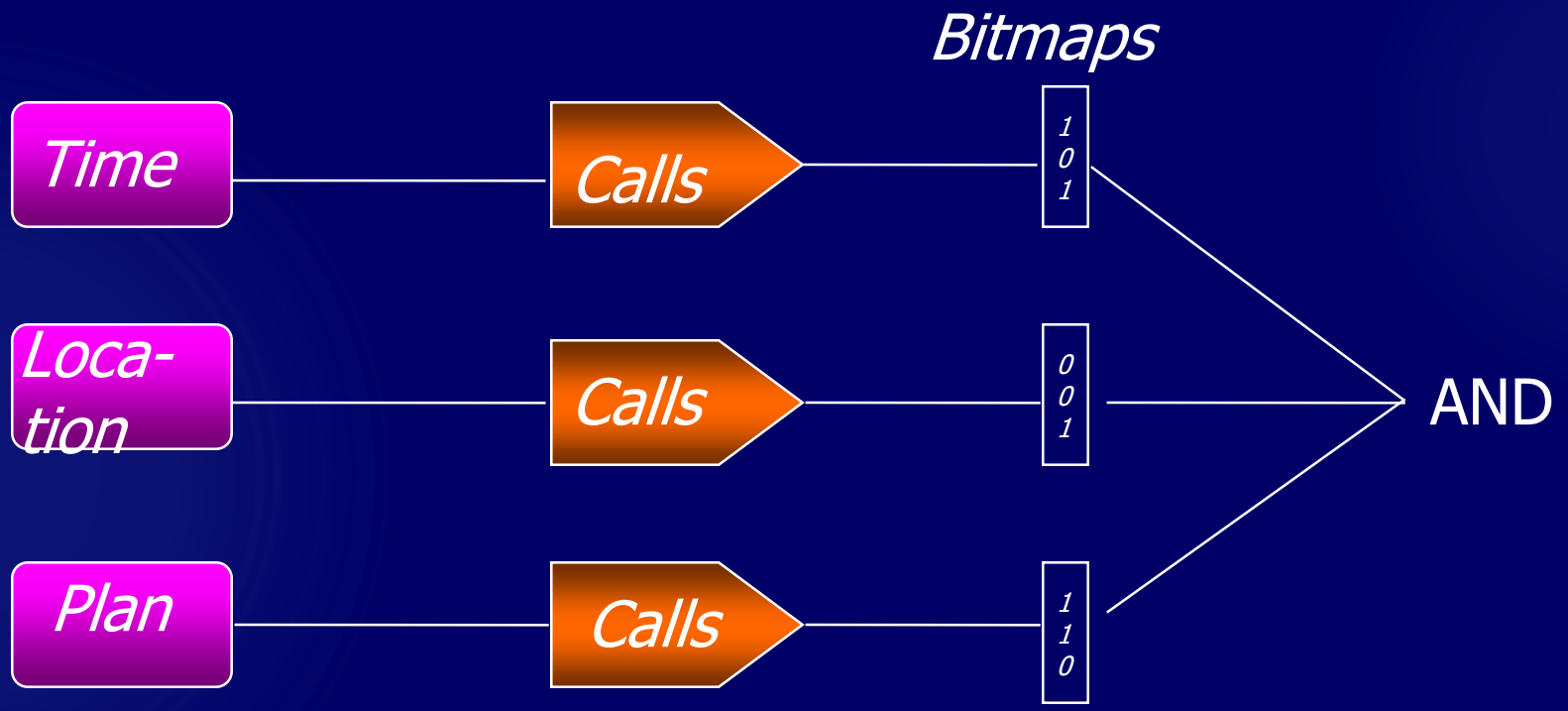
- Use join indexes to join dimension and fact table



Optimized Star Join Processing



Bitmapped Join Processing



Intelligent Scan

- Piggyback multiple scans of a relation (Redbrick)
  - piggybacking also done if second scan starts a little while after the first scan

Parallel Query Processing

- Three forms of parallelism
  - Independent
  - Pipelined
  - Partitioned and "partition and replicate"
- Deterrents to parallelism
  - startup
  - communication

Parallel Query Processing

- Partitioned Data
  - Parallel scans
  - Yields I/O parallelism
- Parallel algorithms for relational operators
  - Joins, Aggregates, Sort
- Parallel Utilities
  - Load, Archive, Update, Parse, Checkpoint, Recovery
- Parallel Query Optimization

Pre-computed Aggregates

- Keep aggregated data for efficiency (pre-computed queries)
- Questions
  - Which aggregates to compute?
  - How to update aggregates?
  - How to use pre-computed aggregates in queries?

Pre-computed Aggregates

- Aggregated table can be maintained by the
  - warehouse server
  - middle tier
  - client applications
- Pre-computed aggregates -- special case of materialized views -- same questions and issues remain

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

SQL Extensions

- Reporting features
  - running total, cumulative totals
- Cube operator
  - group by on all subsets of a set of attributes (month,city)
  - redundant scan and sorting of data can be avoided

Red Brick has Extended set of Aggregates


```
Select month, dollars, cume(dollars) as
run_dollars, weight, cume(weight) as
run_weights
from sales, market, product, period t
where year = 1993
and product like 'Columbian%'
and city like 'San Fr%'
order by t.perkey
```

RISQL (Red Brick Systems) Extensions

- Aggregates
  - CUME
  - MOVINGAVG
  - MOVINGSUM
  - RANK
  - TERTILE
  - RATIOTOREPORT
- Calculating Row Subtotals
  - BREAK BY
- Sophisticated Date Time Support
  - DATEDIFF
- Using SubQueries in calculations

Using SubQueries in Calculations


```
select product, dollars as jun97_sales,
(select sum(s1.dollars)
from market mi, product pi, period, ti, sales si
where pi.product = product.product
and   ti.year    = period.year
and   mi.city    = market.city) as total97_sales,
100 * dollars/
(select sum(s1.dollars)
from market mi, product pi, period, ti, sales si
where pi.product = product.product
and   ti.year    = period.year
and   mi.city    = market.city) as percent_of_yr

from market, product, period, sales

where year = 1997

and   month = 'June' and city like 'Ahmed%'

order by product;
```

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

II. On-Line Analytical Processing (OLAP)

- Making Decision Support Possible

Limitations of SQL

- "A Freshman in Business needs a Ph.D. in SQL"

-- Ralph Kimball

Typical OLAP Queries

- Write a multi-table join to compare sales for each product line YTD this year vs. last year.
- Repeat the above process to find the top 5 product contributors to margin.
- Repeat the above process to find the sales of a product line to new vs. existing customers.
- Repeat the above process to find the customers that have had negative sales growth.

What Is OLAP?

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

* Reference: http://www.arborsoft.com/essbase/wht_ppr/coddTOC.html

The OLAP Market

- Rapid growth in the enterprise market
  - 1995: $700 Million
  - 1997: $2.1 Billion
- Significant consolidation activity among major DBMS vendors
  - 10/94: Sybase acquires ExpressWay
  - 7/95: Oracle acquires Express
  - 11/95: Informix acquires Metacube
  - 1/97: Arbor partners up with IBM
  - 10/96: Microsoft acquires Panorama
- Result: OLAP shifted from small vertical niche to mainstream DBMS category

Strengths of OLAP

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

OLAP Is FASMI

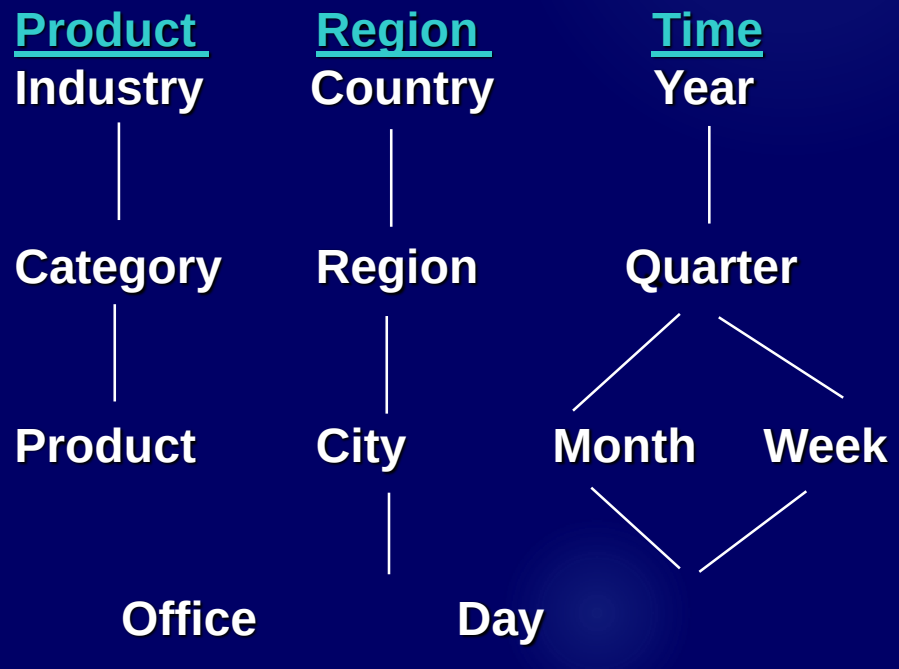
- Fast
- Analysis
- Shared
- Multidimensional
- Information

Nigel Pendse, Richard Creath - The OLAP Report

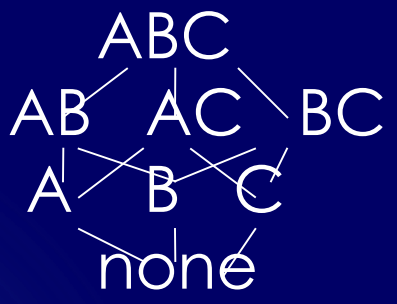
Multi-dimensional Data

- "Hey…I sold $100M worth of goods"

Dimensions: Product, Region, Time
Hierarchical summarization paths

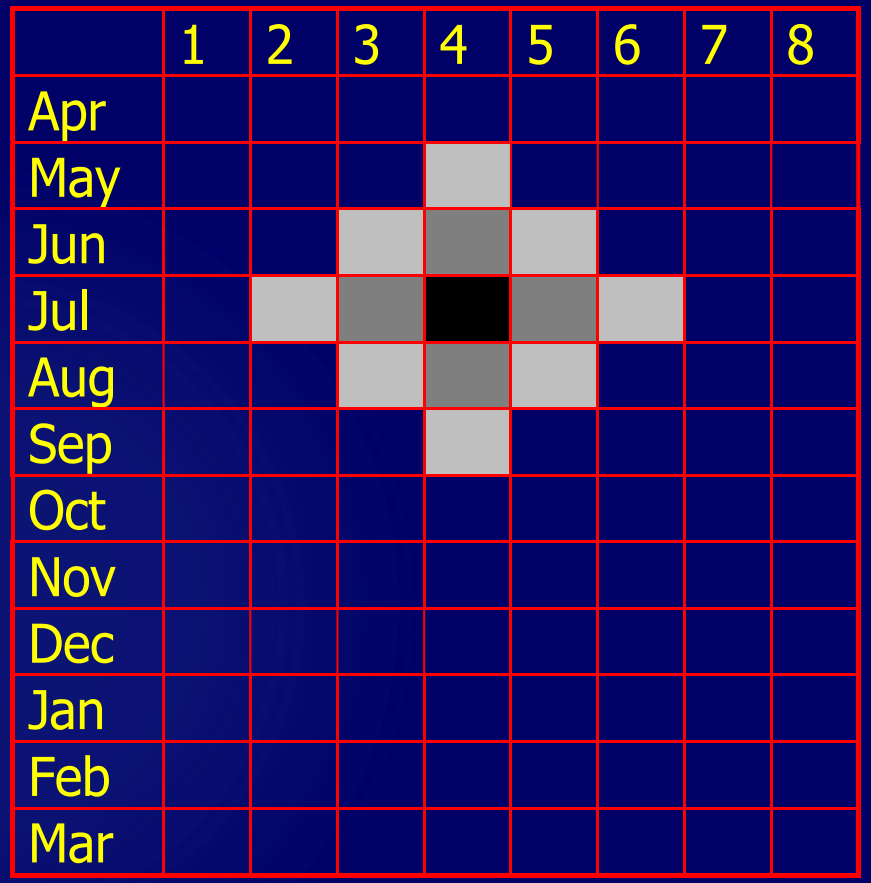


Data Cube Lattice

- Cube lattice
  - 

- Can materialize some groupbys, compute others on demand
- Question: which groupbys to materialze?
- Question: what indices to create
- Question: how to organize data (chunks, etc)

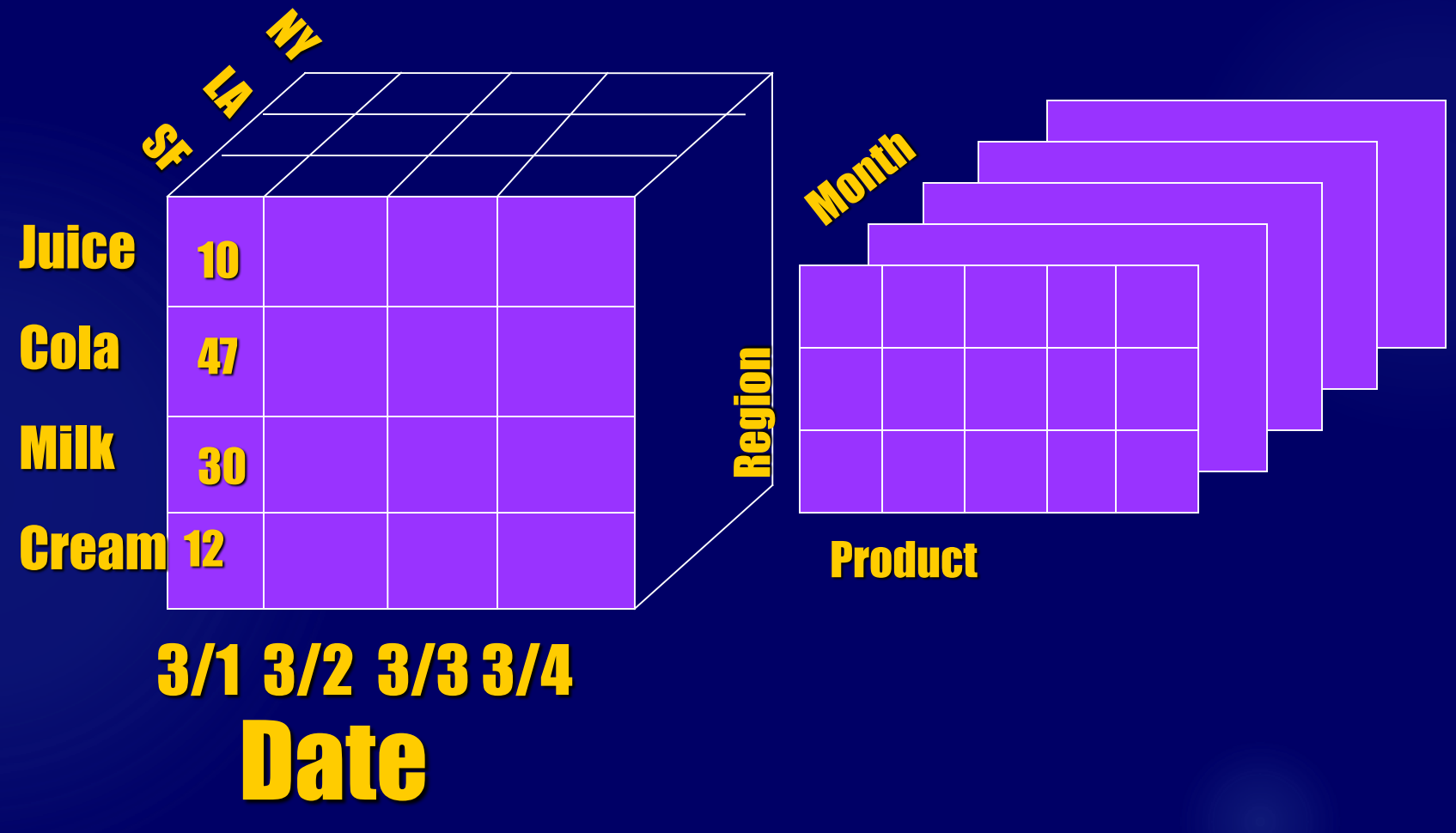
Visualizing Neighbors is simpler



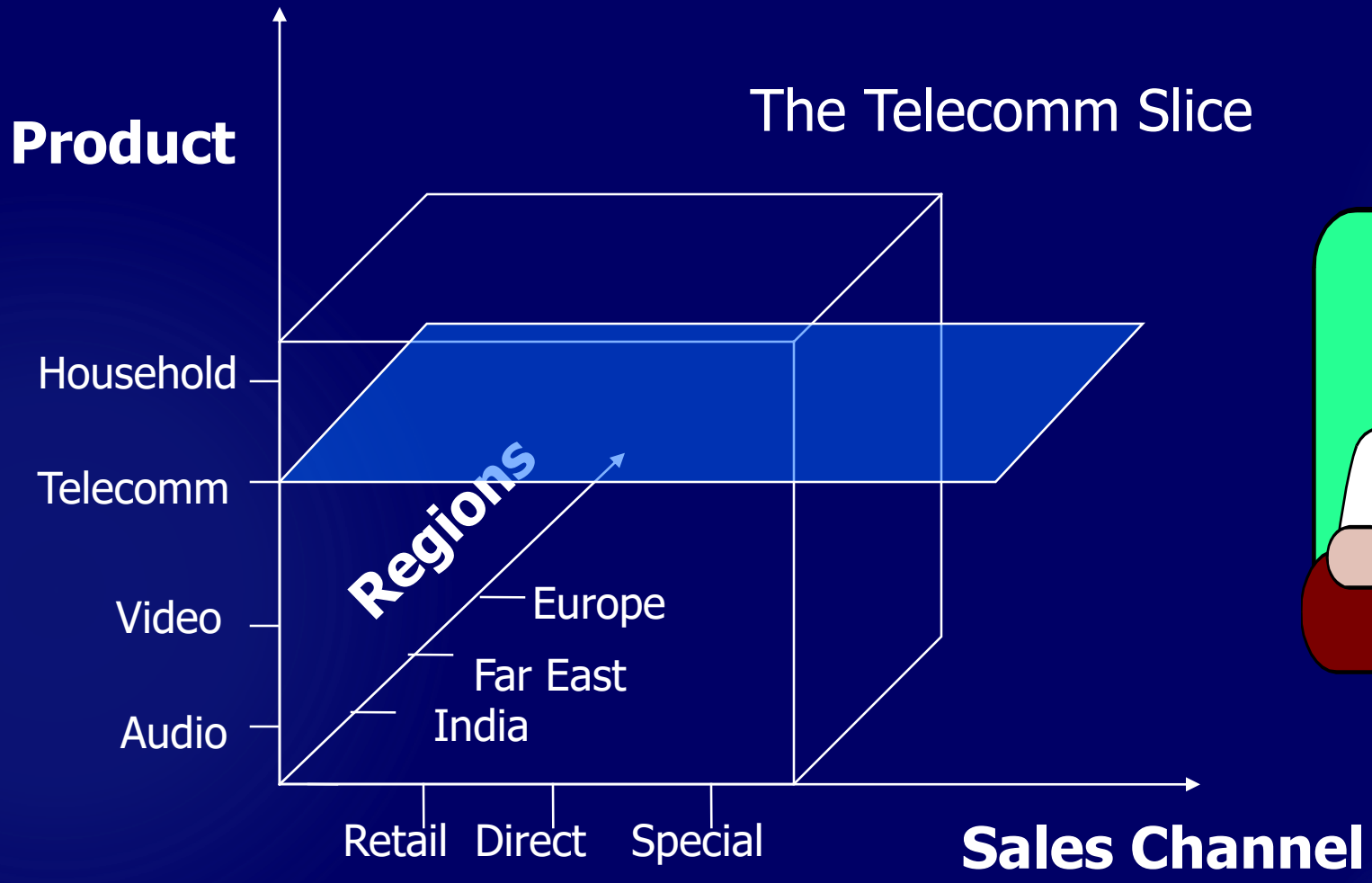
	1	2	3	4	5	6	7	8
Apr								
May								
Jun								
Jul								
Aug								
Sep								
Oct								
Nov								
Dec								
Jan								
Feb								
Mar								

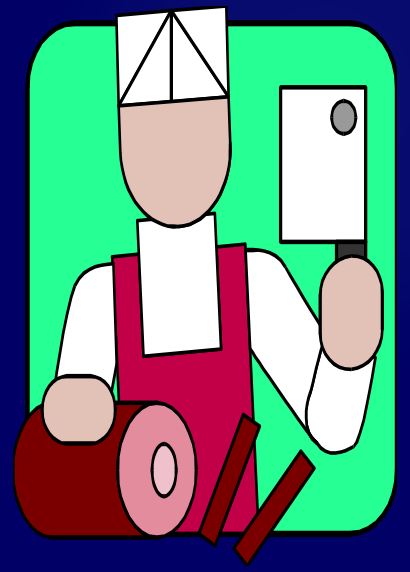
Month	Store	Sales
Apr	1	
Apr	2	
Apr	3	
Apr	4	
Apr	5	
Apr	6	
Apr	7	
Apr	8	
May	1	
May	2	
May	3	
May	4	
May	5	
May	6	
May	7	
May	8	
Jun	1	
Jun	2	

A Visual Operation: Pivot (Rotate)

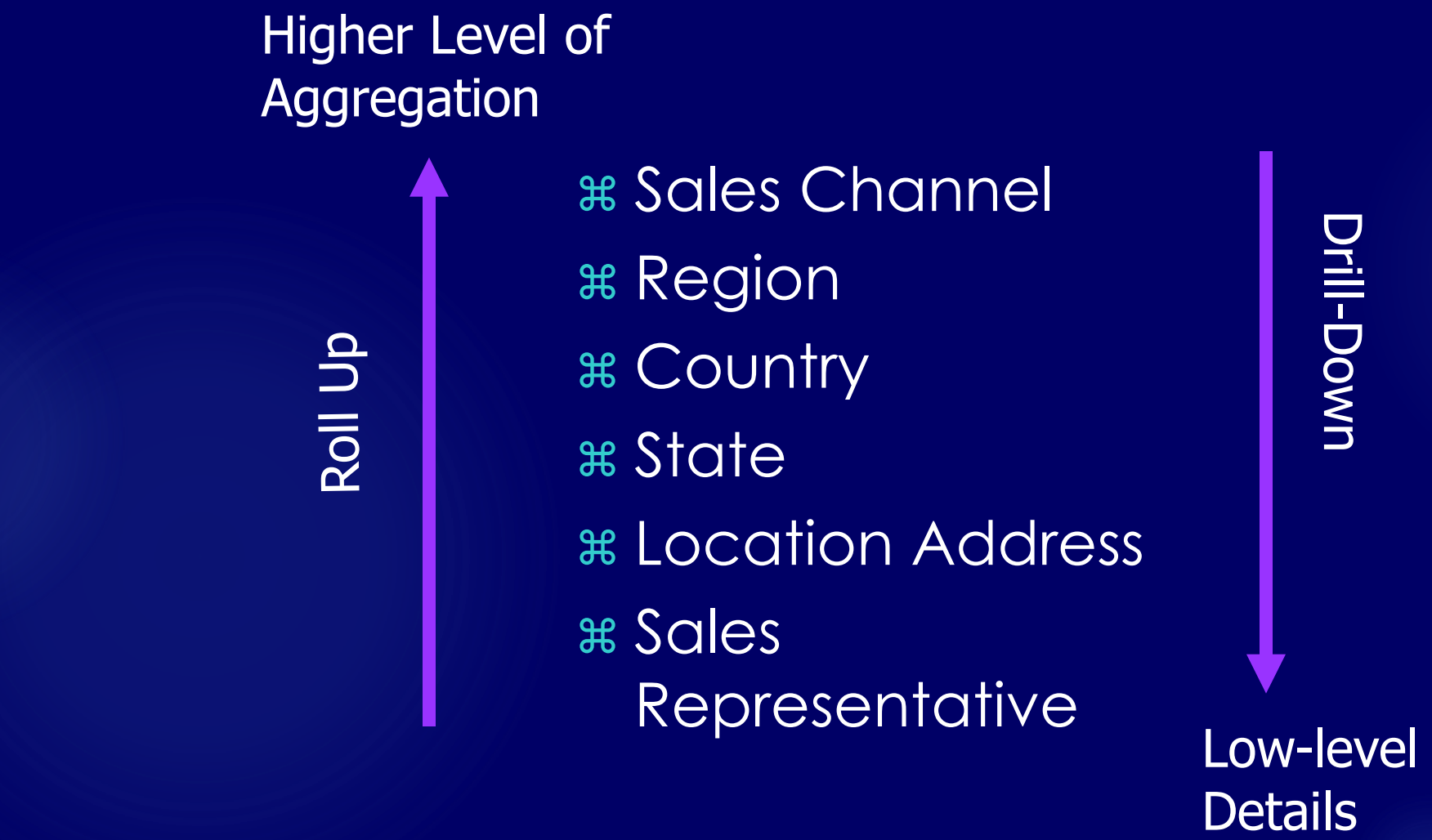


"Slicing and Dicing"





Roll-up and Drill Down

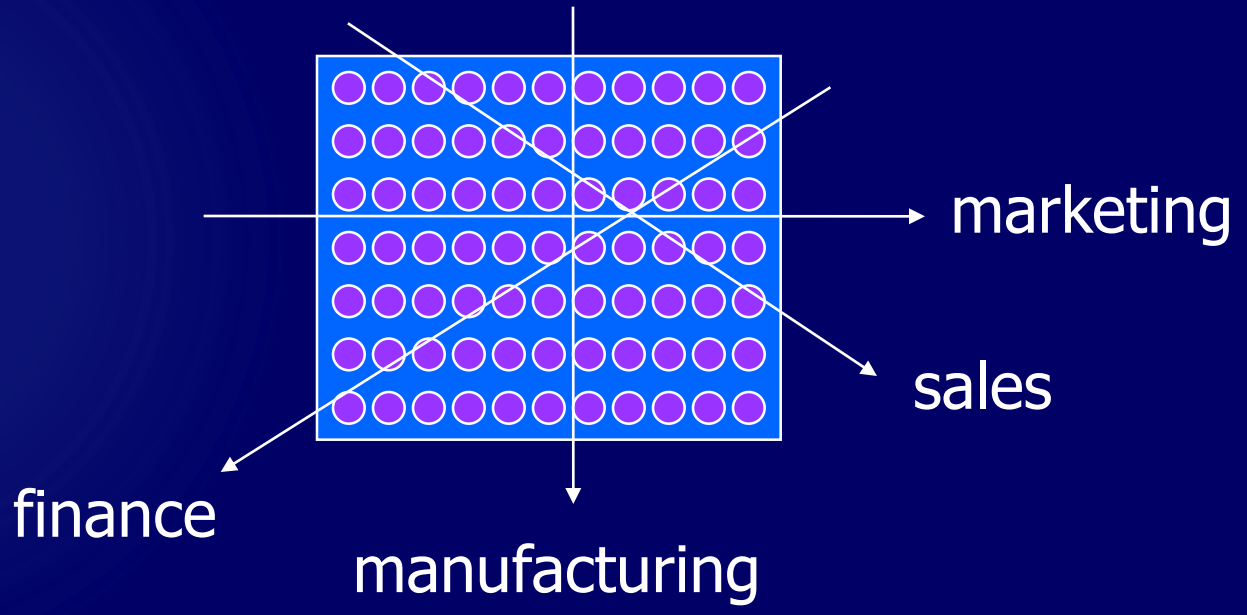


Nature of OLAP Analysis

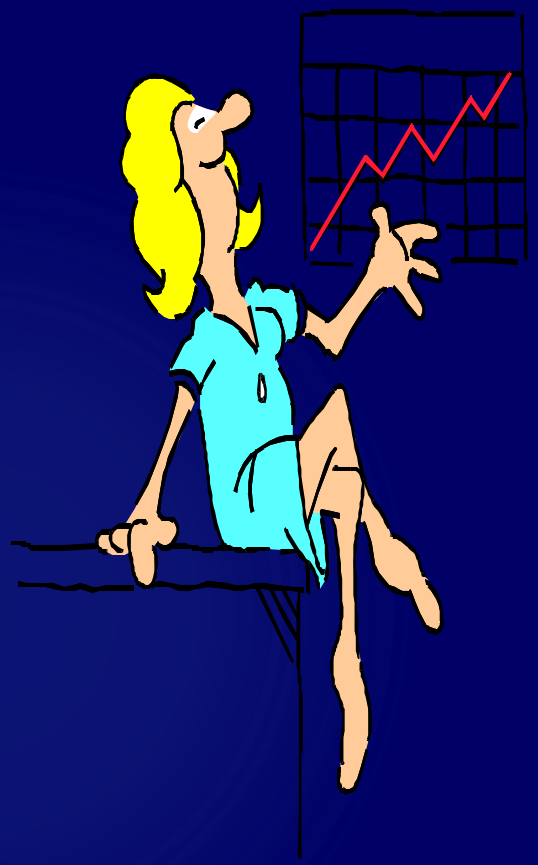
- Aggregation -- (total sales, percent-to-total)
- Comparison -- Budget vs. Expenses
- Ranking -- Top 10, quartile analysis
- Access to detailed and aggregate data
- Complex criteria specification
- Visualization

Organizationally Structured Data

- Different Departments look at the same detailed data in different ways. Without the detailed, organizationally structured data as a foundation, there is no reconcilability of data



Multidimensional Spreadsheets



- Analysts need spreadsheets that support
  - pivot tables (cross-tabs)
  - drill-down and roll-up
  - slice and dice
  - sort
  - selections
  - derived attributes
- Popular in retail domain

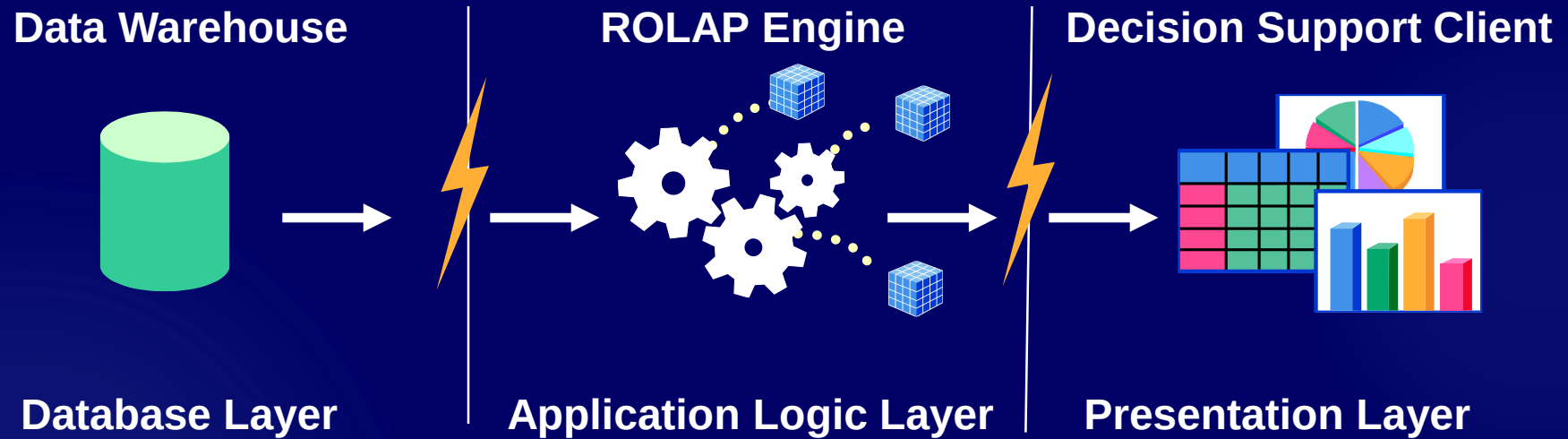
OLAP - Data Cube

- Idea: analysts need to group data in many different ways
  - eg. Sales(region, product, prodtype, prodstyle, date, saleamount)
  - saleamount is a measure attribute, rest are dimension attributes
  - groupby every subset of the other attributes
    - materialize (precompute and store) groupbys to give online response
  - Also: hierarchies on attributes: date -> weekday,
    date -> month -> quarter -> year

SQL Extensions

- Front-end tools require
  - Extended Family of Aggregate Functions
    - rank, median, mode
  - Reporting Features
    - running totals, cumulative totals
  - Results of multiple group by
    - total sales by month and total sales by product
  - Data Cube

Relational OLAP: 3 Tier DSS

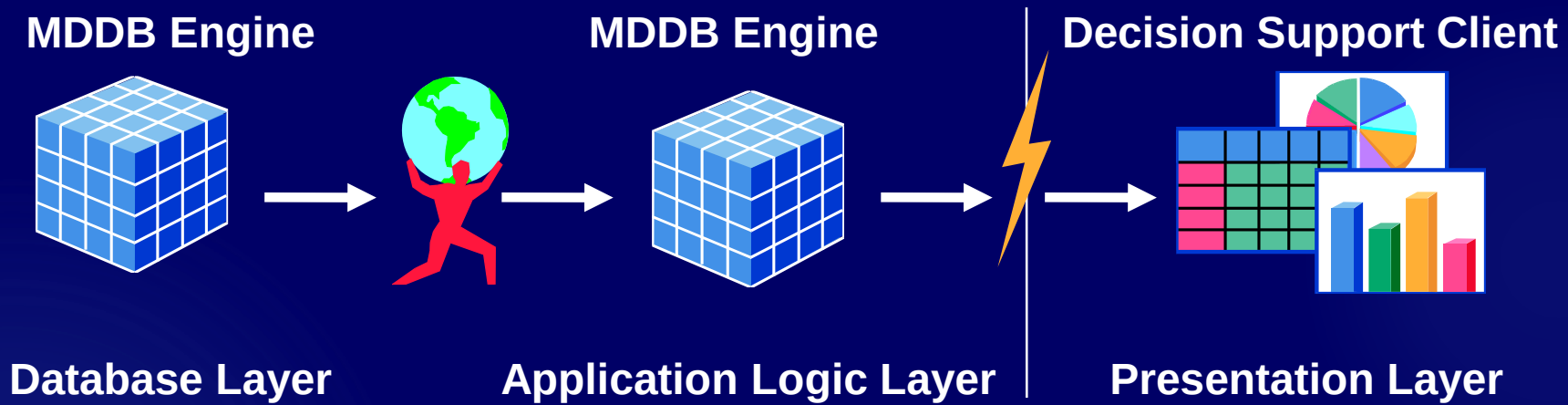


Store atomic data in industry standard RDBMS.

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

Obtain multi-dimensional reports from the DSS Client.

MD-OLAP: 2 Tier DSS

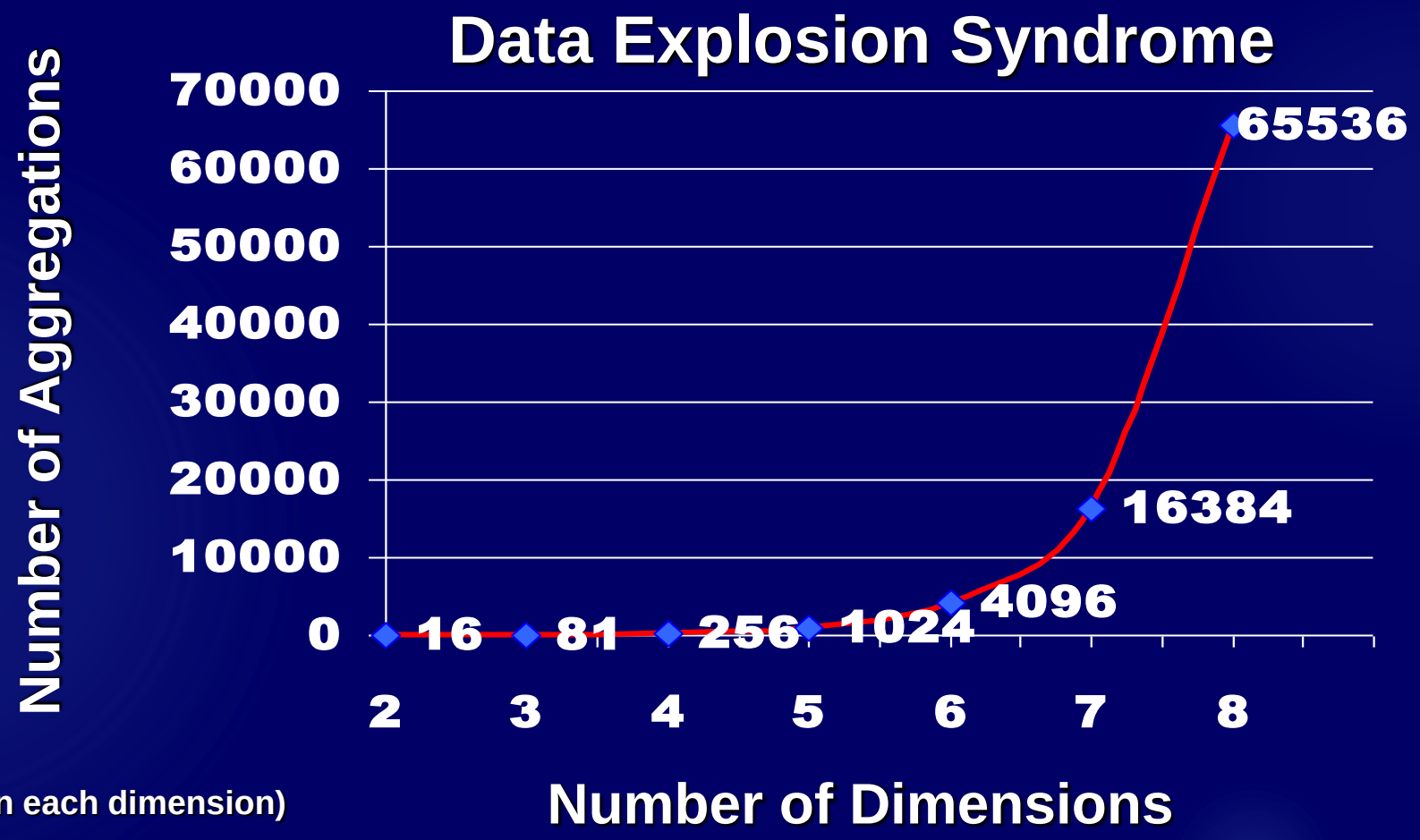


Store atomic data in a proprietary data structure (MDDB), pre-calculate as many outcomes as possible, obtain OLAP functionality via proprietary algorithms running against this data.

Obtain multi-dimensional reports from the DSS Client.

Typical OLAP Problems

Data Explosion



(4 levels in each dimension)

Metadata Repository

- Administrative metadata
  - source databases and their contents
  - gateway descriptions
  - warehouse schema, view & derived data definitions
  - dimensions, hierarchies
  - pre-defined queries and reports
  - data mart locations and contents
  - data partitions
  - data extraction, cleansing, transformation rules, defaults
  - data refresh and purging rules
  - user profiles, user groups
  - security: user authorization, access control

Metdata Repository .. 2

- Business data
  - business terms and definitions
  - ownership of data
  - charging policies
- operational metadata
  - data lineage: history of migrated data and sequence of transformations applied
  - currency of data: active, archived, purged
  - monitoring information: warehouse usage statistics, error reports, audit trails.

Recipe for a Successful Warehouse

For a Successful Warehouse

From Larry Greenfield, http://pwp.starnetinc.com/larryg/index.html

- From day one establish that warehousing is a joint user/builder project
- Establish that maintaining data quality will be an *ONGOING* joint user/builder responsibility
- Train the users one step at a time
- Consider doing a high level corporate data model in no more than three weeks

For a Successful Warehouse

- Look closely at the data extracting, cleaning, and loading tools
- Implement a user accessible automated directory to information stored in the warehouse
- Determine a plan to test the integrity of the data in the warehouse
- From the start get warehouse users in the habit of 'testing' complex queries

For a Successful Warehouse

- Coordinate system roll-out with network administration personnel
- When in a bind, ask others who have done the same thing for advice
- Be on the lookout for small, but strategic, projects
- Market and sell your data warehousing systems

Data Warehouse Pitfalls

- You are going to spend much time extracting, cleaning, and loading data
- Despite best efforts at project management, data warehousing project scope will increase
- You are going to find problems with systems feeding the data warehouse
- You will find the need to store data not being captured by any existing system
- You will need to validate data not being validated by transaction processing systems

Data Warehouse Pitfalls

- Some transaction processing systems feeding the warehousing system will not contain detail
- Many warehouse end users will be trained and never or seldom apply their training
- After end users receive query and report tools, requests for IS written reports may increase
- Your warehouse users will develop conflicting business rules
- Large scale data warehousing can become an exercise in data homogenizing

Data Warehouse Pitfalls

- 'Overhead' can eat up great amounts of disk space
- The time it takes to load the warehouse will expand to the amount of the time in the available window... and then some
- Assigning security cannot be done with a transaction processing system mindset
- You are building a HIGH maintenance system
- You will fail if you concentrate on resource optimization to the neglect of project, data, and customer management issues and an understanding of what adds value to the customer

DW and OLAP Research Issues

- Data cleaning
  - focus on data inconsistencies, not schema differences
  - data mining techniques
- Physical Design
  - design of summary tables, partitions, indexes
  - tradeoffs in use of different indexes
- Query processing
  - selecting appropriate summary tables
  - dynamic optimization with feedback
  - acid test for query optimization: cost estimation, use of transformations, search strategies
  - partitioning query processing between OLAP server and backend server.

DW and OLAP Research Issues .. 2

- Warehouse Management
  - detecting runaway queries
  - resource management
  - incremental refresh techniques
  - computing summary tables during load
  - failure recovery during load and refresh
  - process management: scheduling queries, load and refresh
  - Query processing, caching
  - use of workflow technology for process management

Products, References, Useful Links

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

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
- Microsoft -- Plato
- Oracle -- Express
- Pilot -- LightShip
- Planning Sciences -- Gentium
- Platinum Technology -- ProdeaBeacon, Forest & Trees
- SAS Institute -- SAS/EIS, OLAP++
- Speedware -- Media

Other Warehouse Related Products

- Data extract, clean, transform, refresh
  - CA-Ingres replicator
  - Carleton Passport
  - Prism Warehouse Manager
  - SAS Access
  - Sybase Replication Server
  - Platinum Inforefiner, Infopump

Extraction and Transformation Tools

- Carleton Corporation -- Passport
- Evolutionary Technologies Inc. -- Extract
- Informatica -- OpenBridge
- Information Builders Inc. -- EDA Copy Manager
- Platinum Technology -- InfoRefiner
- Prism Solutions -- Prism Warehouse Manager
- Red Brick Systems -- DecisionScape Formation

Scrubbing Tools

- Apertus -- Enterprise/Integrator
- Vality -- IPE
- Postal Soft

Warehouse Products

- Computer Associates -- CA-Ingres
- Hewlett-Packard -- Allbase/SQL
- Informix -- Informix, Informix XPS
- Microsoft -- SQL Server
- Oracle -- Oracle7, Oracle Parallel Server
- Red Brick -- Red Brick Warehouse
- SAS Institute -- SAS
- Software AG -- ADABAS
- Sybase -- SQL Server, IQ, MPP

Warehouse Server Products

- Oracle 8
- Informix
  - Online Dynamic Server
  - XPS --Extended Parallel Server
  - Universal Server for object relational applications
- Sybase
  - Adaptive Server 11.5
  - Sybase MPP
  - Sybase IQ

Warehouse Server Products

- Red Brick Warehouse
- Tandem Nonstop
- IBM
  - DB2 MVS
  - Universal Server
  - DB2 400
- Teradata

Other Warehouse Related Products

- Connectivity to Sources
  - Apertus
  - Information Builders EDA/SQL
  - Platimum Infohub
  - SAS Connect
  - IBM Data Joiner
  - Oracle Open Connect
  - Informix Express Gateway

Other Warehouse Related Products

- Query/Reporting Environments
  - Brio/Query
  - Cognos Impromptu
  - Informix Viewpoint
  - CA Visual Express
  - Business Objects
  - Platinum Forest and Trees

4GL's, GUI Builders, and PC Databases

- Information Builders -- Focus
- Lotus -- Approach
- Microsoft -- Access, Visual Basic
- MITI -- SQR/Workbench
- PowerSoft --PowerBuilder
- SAS Institute -- SAS/AF

Data Mining Products

- DataMind -- neurOagent
- Information Discovery -- IDIS
- SAS Institute -- SAS/Neuronets

Data Warehouse

- W.H. Inmon, Building the Data Warehouse, Second Edition, John Wiley and Sons, 1996
- W.H. Inmon, J. D. Welch, Katherine L. Glassey, Managing the Data Warehouse, John Wiley and Sons, 1997
- Barry Devlin, Data Warehouse from Architecture to Implementation, Addison Wesley Longman, Inc 1997

Data Warehouse

- W.H. Inmon, John A. Zachman, Jonathan G. Geiger, Data Stores Data Warehousing and the Zachman Framework, McGraw Hill Series on Data Warehousing and Data Management, 1997
- Ralph Kimball, The Data Warehouse Toolkit, John Wiley and Sons, 1996

OLAP and DSS

- Erik Thomsen, OLAP Solutions, John Wiley and Sons 1997
- Microsoft TechEd Transparencies from Microsoft TechEd 98
- Essbase Product Literature
- Oracle Express Product Literature
- Microsoft Plato Web Site
- Microstrategy Web Site

Data Mining

- Michael J.A. Berry and Gordon Linoff, Data Mining Techniques, John Wiley and Sons 1997
- Peter Adriaans and Dolf Zantinge, Data Mining, Addison Wesley Longman Ltd. 1996
- KDD Conferences

Other Tutorials

- Donovan Schneider, Data Warehousing Tutorial, Tutorial at International Conference for Management of Data (SIGMOD 1996) and International Conference on Very Large Data Bases 97
- Umeshwar Dayal and Surajit Chaudhuri, Data Warehousing Tutorial at International Conference on Very Large Data Bases 1996
- Anand Deshpande and S. Seshadri, Tutorial on Datawarehousing and Data Mining, CSI-97

Useful URLs

- Ralph Kimball's home page
  - http://www.rkimball.com
- Larry Greenfield's Data Warehouse Information Center
  - http://pwp.starnetinc.com/larryg/
- Data Warehousing Institute
  - http://www.dw-institute.com/
- OLAP Council
  - http://www.olapcouncil.com/