An Introduction to Data Warehousing

Prakash Kewalramani
http://www.Evaltech.com
Introduction and Motivation

- What is a Warehouse?
- Data Warehouse Architecture
- Implementing the Data Warehouse
- An Introduction to Decision Support
- References
What is a Data Warehouse?
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
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]
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]
Data Warehousing Market

- In 1996, close to 90% of IT professionals had either created a data warehouse or were planning to create one.
- Average 3 year ROI of 400%.
- Average payback was 2.3 years on costs averaging $2.2 million.
The Data Warehouse Market

Values in Millions of US$

Meta Group
95% of Fortune 1000 Companies are creating Warehouses

<table>
<thead>
<tr>
<th>Total Market Size</th>
<th>1994</th>
<th>1999</th>
<th>Compound Annual Growth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Market Size</td>
<td>$1,568.0</td>
<td>$6,960.0</td>
<td>34.7%</td>
</tr>
<tr>
<td>Data Extraction/Movement</td>
<td>$65.0</td>
<td>$210.0</td>
<td>26.4%</td>
</tr>
<tr>
<td>Administration</td>
<td>$10.0</td>
<td>$450.0</td>
<td>114.1%</td>
</tr>
<tr>
<td>RDBMS</td>
<td>$288.0</td>
<td>$1,100.0</td>
<td>30.7%</td>
</tr>
<tr>
<td>Hardware</td>
<td>$1,075.0</td>
<td>$3,950.0</td>
<td>29.7%</td>
</tr>
<tr>
<td>Consulting Services</td>
<td>$130.0</td>
<td>$1,250.0</td>
<td>57.3%</td>
</tr>
</tbody>
</table>

All revenues are in millions of U.S. dollars and are Gartner Group estimates. Source: Gartner Group, Inc.
Warehouses are Very Large Databases

Source: META Group, Inc.

- 5GB: 35%
- 5-9GB: 30%
- 10-19GB: 25%
- 20-49GB: 20%
- 50-99GB: 15%
- 100-249GB: 10%
- 250-499GB: 5%
- 500GB-1TB: 0%
Very Large Data Bases

- Terabytes -- $10^{12}$ bytes: Walmart -- 24 Terabytes
- Petabytes -- $10^{15}$ bytes:
- Exabytes -- $10^{18}$ bytes:
- Zettabytes -- $10^{21}$ bytes:
- Zettabytes -- $10^{24}$ bytes:

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

- Extraction
- Cleansing
- Optimized Loader
- Data Warehouse Engine
- Analyze
- Query
- Metadata Repository

- Relational Databases
- ERP Systems
- Purchased Data
- Legacy Data
Data Warehouse for Decision Support

- 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
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 is 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
More about 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

<table>
<thead>
<tr>
<th>Data</th>
<th>Industry</th>
<th>Usage</th>
<th>Technology</th>
<th>Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer File</td>
<td>All</td>
<td>Track Customer Details</td>
<td>Legacy application, flat files, main frames</td>
<td>Small-medium</td>
</tr>
<tr>
<td>Account Balance</td>
<td>Finance</td>
<td>Control account activities</td>
<td>Legacy applications, hierarchical databases, mainframe</td>
<td>Large</td>
</tr>
<tr>
<td>Point-of-Sale data</td>
<td>Retail</td>
<td>Generate bills, manage stock</td>
<td>ERP, Client/Server, relational databases</td>
<td>Very Large</td>
</tr>
<tr>
<td>Call Record</td>
<td>Telecommunications</td>
<td>Billing</td>
<td>Legacy application, hierarchical database, mainframe</td>
<td>Very Large</td>
</tr>
<tr>
<td>Production Record</td>
<td>Manufacturing</td>
<td>Control Production</td>
<td>ERP, relational databases, AS/400</td>
<td>Medium</td>
</tr>
</tbody>
</table>
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
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
To summarize...

- OLTP Systems are used to "run" a business

- The Data Warehouse helps to "optimize" the business
Wal*Mart Case Study

- Founded by Sam Walton
- One of 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 Conference.
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;
- 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

24 TB Raw Disk; 700 - 1000 Pentium CPUs
**Data Warehouse Architecture**

- **Relational Databases**
- **ERP Systems**
- **Purchased Data**
- **Legacy Data**

**Flow Process**:
1. **Extraction Cleaning**
2. **Optimized Loader**
3. **Data Warehouse Engine**
4. **Metadata Repository**
5. **Analyze Query**

- **CSI'99**
- **ERP Systems**
- **Purchased Data**
- **Legacy Data**
Components of the Warehouse

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

- Data Mining
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
  - Nielsens, IMRA, Vendors, Partners
Data Quality - The Reality

- Tempting to think that all that is there to creating a data warehouse is 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 corporate plan
    - “And get it done quickly, we do not have time to do it properly.”
Data Integration Across Sources

- Savings
  - Same data, different name

- Loans
  - Different data, same name

- Trust
  - Data found here nowhere else

- Credit card
  - Different keys, same data
Data Transformation Example

Data Warehouse

encoding

appl A - m,f
appl B - 1,0
appl C - x,y
appl D - male, female

unit

appl A - pipeline - cm
appl B - pipeline - in
appl C - pipeline - feet
appl D - pipeline - yds

field

appl A - balance
appl B - bal
appl C - currbal
appl D - balcurr
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 million catalogues at Rs. 50 each costs Rs. 50 million. A 2% savings would save Rs. 1 million.
Data Transformation Terms

**Enrichment**
- Bring data from external sources to augment/enrich operational data. Data sources include Dunn and Bradstreet, Nielson, 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
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
Optimizing the Warehouse for Decision Support
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

- 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

Time is part of key of each table
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

䒜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

Frequently accessed

Rarely accessed

Smaller table and so less I/O

acctno

balance

date opened

....

acctno

balance

acctno

address
date -opened

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

- Time
- Cust
- Fact
- Prod
- City
Snowflake schema

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

- Time
- Customer
- Fact
- Product
- City
- Region
Fact Constellation

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

Diagram:
- Hotels
- Booking
- Checkout
- Promotion
- Room Type
- Customer
- Travel Agents
Denormalization

- 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
- Denormalization 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
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
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 Partitioning?

- 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
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
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 Index**

<table>
<thead>
<tr>
<th>gender</th>
<th>vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Y</td>
</tr>
<tr>
<td>F</td>
<td>Y</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td>F</td>
<td>Y</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gender (f)</th>
<th>vote (y)</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Customer

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

```
Calls  

Time

Location

C+T

C+T+L

Plan

C+T+L+P
```
Optimized Star Join Processing

Time
Location
Plan

Apply Selections

Virtual Cross Product of T, L and P

Calls
Bitmapped Join Processing

Time

Location

Plan

Calls

Bitmaps

1 0 1

0 0 1

1 1 0

AND
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
Technological Requirements

- Managing Large amounts of data
- Managing multiple media -- storage hierarchy
  - cache (L1 and L2)
  - main memory
  - disks
  - optical disks
  - tapes
  - fiche
Technological Requirements

- Ability to index data at will
  - temporary indices, sparse indices
- Ability to monitor data freely and easily
  - to determine whether reorganization is required
  - to determine if index is poorly structured
  - to determine statistical composition of data
- Need to interface to many technologies
  - for both receiving and passing data
Technological Requirements

- Programmer/Designer control of data
- Parallel Storage/Management of data
- Good Metadata management
- Load the warehouse efficiently
- Use indexes efficiently
- Compaction of data
Technological Requirements

- Compound Keys
- Variable Length data
- Lock Management
  - Need to be able to turn the lock manager on and off
- Index Only processing
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
Server Scalability

- Scalability is the #1 IT requirement for Data Warehousing
- Hardware Platform options
  - SMP
  - Clusters (shared disk)
  - MPP
    - Loosely coupled (shared nothing)
  - Hybrid
**SMP Characteristics**

- SMP -- Symmetric multi processing -- shared everything
- Multiple CPUs share same memory
- Workload is balanced across CPUs by OS
- Scalability is limited to bandwidth of internal bus and OS architecture
- Not tolerant to failure in processing node
- Architecture is mostly invisible to applications
SMP Benefits

- Lower entry point -- can start with SMP
- Mature technology
MPP Characteristics

- Each node owns a portion of the database
- Nodes are connected via an interconnection network
- Each node can be a single CPU or SMP
- Load balancing done by application
- High scalability due to local processing isolation
MPP benefits

- High availability
- High scalability
Other Warehouse Related Products

- Connectivity to Sources
  - Apertus
  - Information Builders EDA/SQL
  - Platinum Infohub
  - SAS Connect
  - IBM Data Joiner
  - Oracle Open Connect
  - Informix Express Gateway
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
Other Warehouse Related Products

- Query/Reporting Environments
  - Brio/Query
  - Cognos Impromptu
  - Informix Viewpoint
  - CA Visual Express
  - Business Objects
  - Platinum Forest and Trees
Data Warehouse vs. Data Marts

What comes first
From the Data Warehouse to Data Marts

Information

Individually Structured

Departmentally Structured

Organizationaly Structured

Data Warehouse

Less History Normalized Detailed

More
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
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
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
Viewing the Data with 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)

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

Product Industry Category Region Country
Region

Product
City
Office

Time
Year Quarter Month Week Day
Visualizing Neighbors is simpler

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 |
"Slicing and Dicing"

The Telecomm Slice

Product

Household
Telecomm
Video
Audio

Region

Europe
Far East
India

Sales Channel

Retail
Direct
Special
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
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
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

Data Warehouse

Database Layer

Store atomic data in industry standard RDBMS.

ROLAP Engine

Application Logic Layer

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

Decision Support Client

Presentation Layer

Obtain multi-dimensional reports from the DSS Client.
MD-OLAP: 2 Tier DSS

MDDB Engine

Database Layer

Application Logic Layer

Decision Support Client

Presentation Layer

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 Syndrome

Number of Dimensions

Number of Aggregations

(4 levels in each dimension)

Microsoft TechEd’98
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
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
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, 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

- 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/