Data Warehousing

CSE 6331 / CSE 6362
Data Mining
Fall 1999
Diane J. Cook

Motivation: "Necessity is the Mother of Invention"

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of stored data
- Need to integrate data
  - Grocery store chain
  - Sales database, inventory database, marketing database

Data Warehouse

- A decision support data repository that has been extracted and integrated from heterogeneous distributed sources.
- "A data warehouse is a
  - subject-oriented,
  - integrated,
  - time-varying,
  - non-volatile
- collection of data that is used primarily in organizational decision making" [Chaudhuri and Dayal, 1997]

Data Warehouse - Subject Oriented

- Oriented to the major subject areas of the corporation defined in the data model
  - Insurance company: customer, product, transaction/activity, policy, claim, account, ...
- Operational DBs and applications may be organized differently
  - E.g. based on type of insurance: auto, life, medical, fire, etc.

Data Warehouse - Integrated

- There is no consistency in encoding, naming conventions, etc., among different data sources
- When data is moved to the warehouse, it is transformed to a consistent representation
- Operational DBs can be big
- Data warehouses are usually huge (really big)

HUGE

Data Warehouse - Non-Volatile

- Operational data is regularly accessed, manipulated, and updated (OLTP)
- Warehouse data is loaded and accessed, but not typically updated

Data Warehouse - Time Variance

- The time horizon for the data warehouse is significantly longer than for operational systems
- Operational DB contains current value data
- Data warehouse data is series of snapshots, taken as of some moment in time
- The key structure of operational data may or may not contain some element of time
- The key structure of the data warehouse always contains some element of time

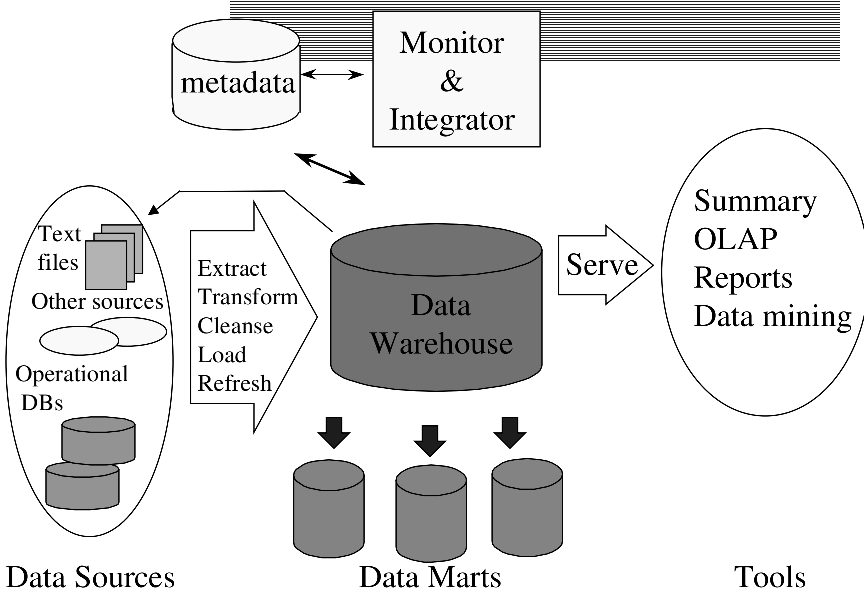
Why Separate Data Warehouse?

- Performance
  – Special data organization, access, and implementation needed to support multidimensional views, OLAP, and data mining
  – Complex OLAP queries would degrade performance for operational transactions

Why Separate Data Warehouse?

- Function
  - Missing data: Decision support requires historical data which operational DBs do not typically maintain
  - Data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - Data quality: different sources typically use inconsistent data representations and formats

Warehouse Architecture



Data Sources

- Data sources contain lowest-level, raw data
- Data sources are designed for operational use
- Multiple data sources often from different systems
- Multiple data sources introduce a large number of issues

Data Extraction

- Retrieve desired data from source
- Convert to uniform format
- Data extractor usually tightly coupled to source

Data Cleaning

- Operational data from multiple sources often dirty
- Three classes of tools
  - Data migration: allow simple data transformation
  - Data scrubbing: use domain-specific knowledge to scrub data
  - Data auditing: discover rules and relationships by scanning data (detect outliers)

Load and Refresh

- Loading the warehouse: check integrity constraints, sort, summarize, build indices
- Refreshing a warehouse: propagate updates to the data stored in the warehouse
  - When to refresh
  - How to refresh
- Data purging

Data Warehouse DBMS

- Often built as RDBMS or MDDB
- Specialized features
- Specialized capabilities
  - Parallel relational database design
  - Methods to speed up RDBMS (index structures)
  - Multidimensional databases (MDDBs)

Monitor

- Detect changes to an information source that are of interest to the warehouse
- Propagate the change in a generic form to the integrator

Integrator

- Receive changes from the monitors
  - make the data conform to the conceptual schema used by the warehouse
- Integrate the changes into the warehouse
  - merge the data with existing data already present
  - resolve possible update anomalies

Metadata Repository

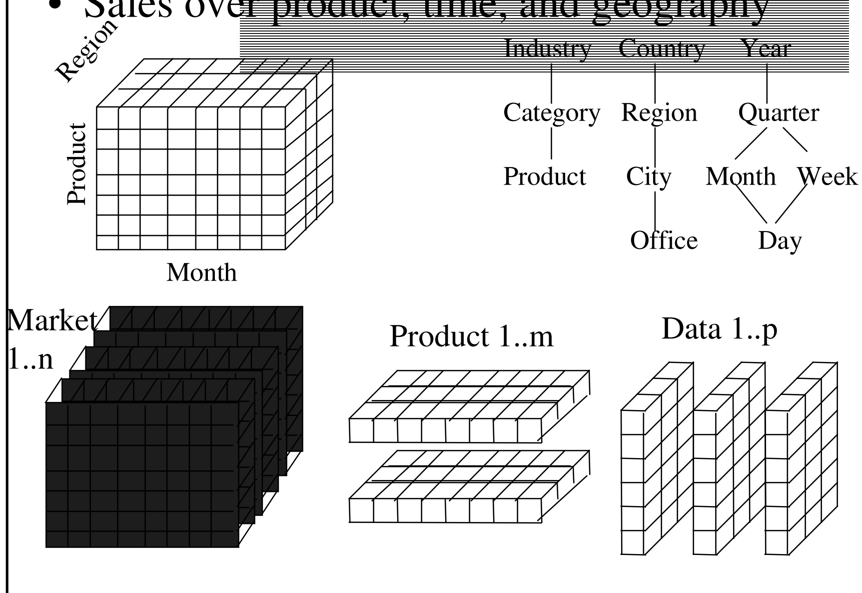
- Source database and their contents
- Dimensions and hierarchies
- Pre-defined queries and reports
- Data mart locations and contents
- Data partitions
- Data extraction, cleansing, transformation rules
- Data refresh and purge rules
- User authorization, access control, ownership
- Currency of data: active, archived, purged
- Monitoring information: warehouse usage statistics, error reports, audit trails

Data Marts

- Specialized system that stores partition of data needed for a department or particular application
- Physically separate or within the data warehouse using application-specific *views*
- Data marts may use different data representations and include their own OLAP engines

Multidimensional Modeling

- Sales over product, time, and geography



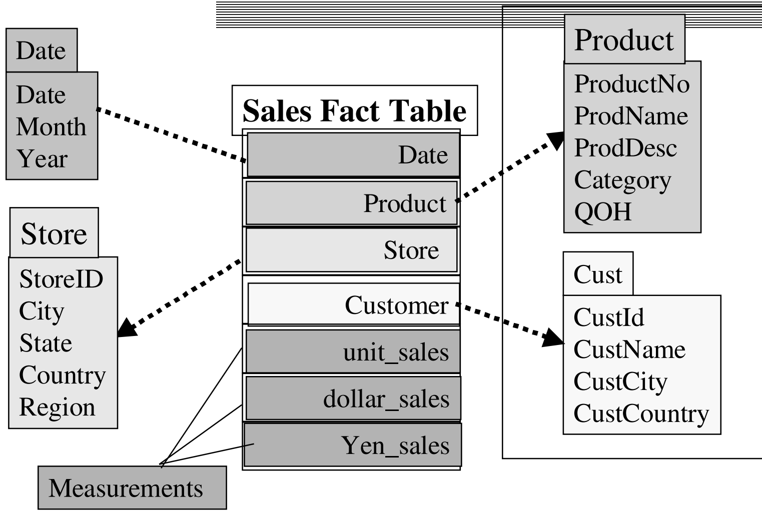
Dimensional Modeling

- We've seen MOLAP description
- Multidimensional storage (useful for data cubes)
- How or ROLAP for multidimensional data?
- Two types of tables
  - dimension table
  - fact table
- Schema (Star and Snowflake)

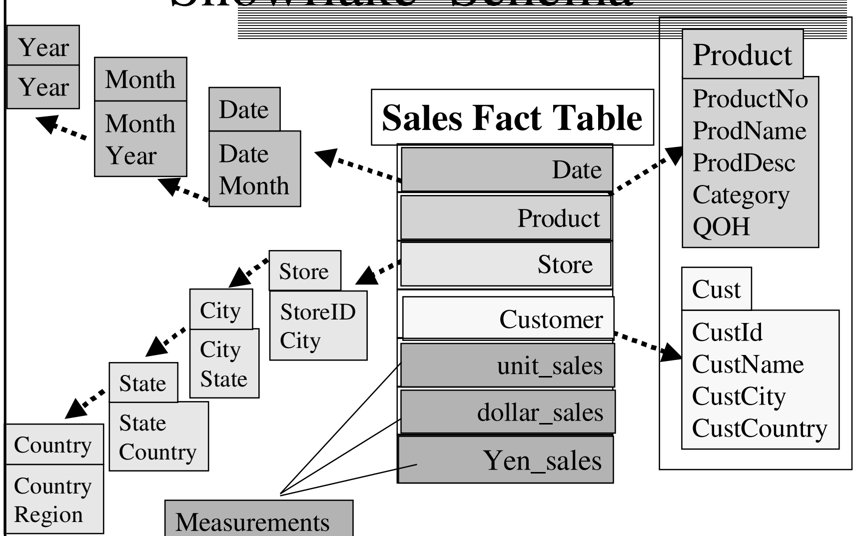
Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measurements
  - Star schema: A single object (fact table) in the middle connected to objects (dimension tables)
  - Snowflake schema: Dimensional hierarchy is normalizes dimension tables.
  - Fact constellations: Multiple fact tables share dimension tables.

Star Schema



Snowflake Schema



Summary Tables

- Data warehouse may store some selected summary data, the pre-aggregated data.
- Summary data stored as *separate fact tables* sharing the same dimension tables with the base fact table.
- Summary data can be encoded in the original fact table and dimension tables.

id	level	date	month	year
0	1	1	1	1998
1	2	NULL	1	1998
2	2	NULL	2	1998
3	3	NULL	NULL	1998

DateID	ProdID	Sales
0	1	1000
1	1	20000
1	2	40000
3	1	300000

Multidimensional Indexing

- Indexing principle: mapping key values to records for associative direct access
- Indexing in data warehouse is still an open problem
  - for large data tables, size of multidimensional index could be very large

Bitmap Indexes

- Indexing for high-performance access in multi-dimensional databases
- The Bitmap index

Base table

Cust	Region	Type
C1	Asia	Retail
C2	Europe	Dealer
C3	Asia	Dealer
C4	America	Retail
C5	Europe	Dealer

Index on Region

RecID	Asia	Europe	America
1	1	0	0
2	0	1	0
3	1	0	0
4	0	1	0
5	0	0	1

Index on Type

RecID	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0
5	0	1

Bitmap Indexes

- With efficient hardware support for bitmap operations (AND, OR, XOR, NOT), bitmap index offers better access methods for certain queries
  - e.g., selection on two attributes
- Some commercial products have implemented bitmap index
- Challenging for high cardinality domains since the number of bit vectors increases

Join Indexes

- Traditional indexes map values to list of record ids
- Join indexes map join result tuples to source tables
- In data warehouse cases, join indices relate the values of the dimensions of a start schema to rows in the fact table.

Using SQL with Data Warehouses

- Parallel query processing
- Piggybacking scans of multiple queries
- SQL Extensions
  - Aggregates (rank, percentile, mean, mode, median)
  - Reporting Features (moving average)
  - Rollup, Cube

Data Warehousing: Summary

- There are forces driving the need for data warehousing
- OLAP provides powerful and fast tools for reporting on data
- There are a number issues associated to both
  - the basic techniques (multidimensional database, query optimization, indexing, …);
  - data cleaning, partitioning, indexing
  - Automated data acquisition, maintenance, performance optimization

Data Warehousing Systems

- Arbor Software: Essbase (M)
- Pilot (M)
- Oracle (IRI): Express (R)
- DB2 (R)
- Sybase (R)
- Cognos: PowerPlay
- Redbrick Systems: Redbrick Warehouse (R)
- Microstrategy: DSS/Server
- Warehouse-related products: http://pwp.starnetinc.com/larryg