Data Warehousing

Overview, Terminology, and Research Issues

Joachim Hammer
Heterogeneous Database Integration

- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

Integration System

- Digital Libraries
- Scientific Databases
- Personal Databases

World Wide Web
The Traditional Research Approach

- Query-driven (lazy, on-demand)
Disadvantages of Query-Driven Approach

• Delay in query processing
  – Slow or unavailable information sources
  – Complex filtering and integration
• Inefficient and potentially expensive for frequent queries
• Competes with local processing at sources
• Hasn’t caught on in industry
The Warehousing Approach

- Information integrated in advance
- Stored in wh for direct querying and analysis
Advantages of Warehousing Approach

• High query performance
  – But not necessarily most current information

• Doesn’t interfere with local processing at sources
  – Complex queries at warehouse
  – OLTP at information sources

• Information copied at warehouse
  – Can modify, annotate, summarize, restructure, etc.
  – Can store historical information
  – Security, no auditing

• *Has* caught on in industry
Not Either-Or Decision

- Query-driven approach still better for
  - Rapidly changing information
  - Rapidly changing information sources
  - Truly vast amounts of data from large numbers of sources
  - Clients with unpredictable needs
Data Warehousing: Two Distinct Issues

(1) How to get information into warehouse
   "Data warehousing"

(2) What to do with data once it’s in warehouse
   "Warehouse DBMS"
   • Terms coined by Jennifer Widom (WHIPS)
   • Both rich research areas
   • Industry has focused on (2)
What is a Warehouse?

• Stored collection of diverse data
  – A solution to data integration problem
  – Single repository of information

• Subject-oriented
  – Organized by subject, not by application
  – Used for analysis, data mining, etc.

• Optimized differently from transaction-oriented db

• User interface aimed at executive
What is a Warehouse (2)?

- Large volume of data (Gb, Tb)
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at WalMart
  - Complete client histories at insurance firm
  - Stockbroker financial information and portfolios
Warehouse is a Specialized DB

<table>
<thead>
<tr>
<th>Standard DB</th>
<th>Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly updates</td>
<td>Mostly reads</td>
</tr>
<tr>
<td>Many small transactions</td>
<td>Queries are long and complex</td>
</tr>
<tr>
<td>Mb - Gb of data</td>
<td>Gb - Tb of data</td>
</tr>
<tr>
<td>Current snapshot</td>
<td>History</td>
</tr>
<tr>
<td>Index/hash on p.k.</td>
<td>Lots of scans</td>
</tr>
<tr>
<td>Raw data</td>
<td>Summarized, consol. data</td>
</tr>
<tr>
<td>Thousands of users (e.g., clerical users)</td>
<td>Hundreds of users (e.g., decision-makers, analysts)</td>
</tr>
</tbody>
</table>
Warehousing and Industry

• Warehousing is big business
  – $2 billion in 1995
  – $3.5 billion in early 1997
  – Predicted: $8 billion in 1998 [Metagroup]

• WalMart has largest warehouse
  – 4Tb
  – 200-300Mb per day
Warehouse DBMS—Buzzwords

• Used primarily for decision support (DSS)
  – A.K.A. On-Line Analytical Processing (OLAP)
  – Complex queries, substantial aggregation
  – TPC-D benchmark

• May support multidimensional database (MDDB)
  – A.K.A. Data Cube
  – View of relational data: all possible groupings and aggregations
Warehouse DBMS — Buzzwords (2)

• **ROLAP** vs. **MOLAP**
• Special purpose OLAP servers that directly implement multidimensional data and operations
  – *Roll-up* = aggregate on some dimension
  – *Drill-down* = deaggregate on some dimension
• ROLAP: Oracle, Sybase IQ, RedBrick
• MOLAP: Pilot, Essbase, Gentia
Warehouse DBMS - Buzzwords (3)

• Clients:
  – Query and reporting tools
  – Analysis tools
  – Data mining: discovering patterns of various forms

• Poses many new research issues in:
  – Query processing and optimization
  – Database design
  – View management
Data Warehousing: Basic Architecture

![Diagram of Data Warehouse Architecture]

- Data Warehouse
- Integration System
- Metadata
- Extractor/Monitor
- Source
Data Warehousing: Current Practice

- Warehouse is standard or specialized DBMS
- Everything is relational
- Extraction and integration are done in batch, usually off-line, often “by hand”
- Integrator never queries sources
  - Everything is replicated at warehouse
- Generalizing introduce many research issues
Research Issues in Data Warehousing

- Extraction
- Integration
- Warehousing specification
- Optimizations
- Miscellaneous
Data Extraction

• Translation
  – Translate information to warehouse data model
  – Similar to wrapper in query-driven approach [Stanford Tsimmis project]

• Change detection
  – For on-line data propagation and incremental warehouse refresh
  – Different classes of information sources
    • Cooperative
    • Logged
    • Queryable
    • Snapshot
Data Extraction (2)

- Efficient change detection algorithms for snapshot sources
  - Record-oriented
  - Hierarchical data
- Data scrubbing (or cleansing)
  - Discard or correct erroneous data
  - Insert default values
  - Eliminate duplicates and inconsistencies
  - Aggregate, summarize, sample
- Implementing extractors
  - Specification-based or toolkit approach
Data Integration

• Warehouse data ≈ materialized view
  – Initial loading
  – View maintenance

• But: differs from conventional materialized view maintenance
  – Warehouse views may be highly aggregated
  – Warehouse views may be over history of base data
  – Schema may evolve
  – Base data doesn’t participate in view maintenance
    • Simply reports changes
    • Loosely coupled
    • Absence of locking, global transactions
    • May not be queriable
Warehouse Maintenance Anomalies

- Materialized view maintenance in loosely coupled, anon-transactional environment
- Simple example

![Diagram]

\[
\text{Data Warehouse} \\
\text{Integrator} \\
\text{Sales} \rightarrow \text{Sale}(\text{item, clerk}) \\
\text{Comp.} \rightarrow \text{Emp}(\text{clerk, age})
\]

\[
\text{Sold} (\text{item, clerk, age}) \\
\text{Sold} = \text{Sale} \bowtie \text{Emp}
\]
Warehouse Maintenance Anomalies

1. Insert into `Emp(Mary, 25)`, notify integrator
2. Insert into `Sale(Computer, Mary)`, notify integrator
3. (1) $\rightarrow$ integrator adds `Sale(Mary, 25)`
4. (2) $\rightarrow$ integrator adds `(Computer, Mary)`
5. View incorrect (duplicate tuple)
Warehouse Self-Maintainability

• Goal: No queries back to sources
• Research issues:
  – What views are self-maintainable
  – Store auxiliary views so original + auxiliary views are self-maintainable
• Examples:
  – Keep keys
  – Inserts into Sale, maintain auxiliary view:
    \[ \text{Emp} - \Pi_{\text{clerk}, \text{age}}(\text{Sold}) \]
• Lots of issues...
Warehouse Specification

- Ideal scenario:
Optimizations

• Update filtering at extractor
• Multiple view optimization
  – If warehouse contains several views
  – Exploit shared sub-views
• Others?
Additional Research Issues

• Warehouse management
  – Schema design
  – Initial loading
  – Metadata management