Decision Support, Data Warehousing, and OLAP

Anindya Datta
Director, iXL Center for E-Commerce
Georgia Institute of Technology
adatta@cc.gatech.edu

Outline

- Terminology: OLAP vs. OLTP
- Data Warehousing Architecture
- Technologies
- Products
- Research Issues
- References
Decision Support and OLAP

- Information technology to help the knowledge worker (executive, manager, analyst) make faster and better decisions.
  - What were the sales volumes by region and product category for the last year?
  - How did the share price of computer manufacturers correlate with quarterly profits over the past 10 years?
  - Which orders should we fill to maximize revenues?
  - Will a 10% discount increase sales volume sufficiently?
  - Which of two new medications will result in the best outcome: higher recovery rate & shorter hospital stay?

- On-Line Analytical Processing (OLAP) is an element of decision support systems (DSS).

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
## OLTP vs. OLAP

<table>
<thead>
<tr>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User</strong></td>
<td>Clerk, IT Professional</td>
</tr>
<tr>
<td><strong>Function</strong></td>
<td>Day to day operations</td>
</tr>
<tr>
<td><strong>DB Design</strong></td>
<td>Application-oriented (E-R based)</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Current, Isolated</td>
</tr>
<tr>
<td><strong>View</strong></td>
<td>Detailed, Flat relational</td>
</tr>
<tr>
<td><strong>Usage</strong></td>
<td>Structured, Repetitive</td>
</tr>
<tr>
<td><strong>Unit of work</strong></td>
<td>Short, Simple transaction</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Read-write</td>
</tr>
<tr>
<td><strong>Operations</strong></td>
<td>Index/hash on prim. Key</td>
</tr>
<tr>
<td><strong># Records accessed</strong></td>
<td>Tens</td>
</tr>
<tr>
<td><strong>#Users</strong></td>
<td>Thousands</td>
</tr>
<tr>
<td><strong>Db size</strong></td>
<td>100 MB-GB</td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td>Trans. throughput</td>
</tr>
</tbody>
</table>

## Data Warehouse

- A decision support database that is maintained separately from the organization’s operational databases.
- A data warehouse is a
  - subject-oriented,
  - integrated,
  - time-varying,
  - non-volatile
- collection of data that is used primarily in organizational decision making.
**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.

---

**Data Warehousing Market**

- Hardware: servers, storage, clients
- Warehouse DBMs
- Tools
- Market growing from
  - $2B in 1995 to $8 B in 1998 (Meta Group)
  - 1.5B today to $6.9B in 1999 (Gartner Group)
- Systems integration & Consulting
- Already deployed in many industries: manufacturing, retail, financial, insurance, transportation, telecom., utilities, healthcare.
Data Warehousing Architecture

- Monitoring & Administration
- OLAP servers
- Metadata Repository
- Extract, Transform, Load, Refresh
- Data Marts
- Data Warehousing Architecture

Three-Tier Architecture

- Warehouse database server
  - Almost always a relational DBMS; rarely flat files
- OLAP servers
  - Relational OLAP (ROLAP): extended relational DBMS that maps operations on multidimensional data to standard relational operations.
  - Multidimensional OLAP (MOLAP): special purpose server that directly implements multidimensional data and operations.
- Clients
  - Query and reporting tools.
  - Analysis tools
  - Data mining tools (e.g., trend analysis, prediction)
Data Warehouse vs. Data Marts

- Enterprise warehouse: collects all information about subjects (customers, products, sales, assets, personnel) that span the entire organization.
  - Requires extensive business modeling
  - May take years to design and build

- Data Marts: Departmental subsets that focus on selected subjects:
  Marketing data mart: customer, products, sales.
  - Faster roll out, but complex integration in the long run.

- Virtual warehouse: views over operational dbs
  - Materialize some summary views for efficient query processing
  - Easier to build
  - Requisite excess capacity on operational db servers

Design & Operational Process

- Define architecture. Do capacity planning.
- Integrate db and OLAP servers, storage and client tools.
- Design warehouse schema, views.
- Design physical warehouse organization: data placement, partitioning, access methods.
- Connect sources: gateways, ODBC drivers, wrappers.
- Design & implement scripts for data extract, load refresh.
- Define metadata and populate repository.
- Design & implement end-user applications.
- Roll out warehouse and applications.
- Monitor the warehouse.
OLAP for Decision Support

- Goal of OLAP is to support ad-hoc querying for the business analyst
- Business analysts are familiar with spreadsheets
- Extend spreadsheet analysis model to work with warehouse data
  - Large data set
  - Semantically enriched to understand business terms (e.g., time, geography)
  - Combined with reporting features
- Multidimensional view of data is the foundation of OLAP

Multidimensional Data Model

- Database is a set of facts (points) in a multidimensional space
- A fact has a measure dimension
  - quantity that is analyzed, e.g., sale, budget
- A set of dimensions on which data is analyzed
  - e.g., store, product, date associated with a sale amount
- Dimensions form a sparsely populated coordinate system
- Each dimension has a set of attributes
  - e.g., owner city and county of store
- Attributes of a dimension may be related by partial order
  - Hierarchy: e.g., street > county > city
  - Lattice: e.g., date > month > year, date > week > year
Multidimensional Data

Sales Volume as a function of time, city and product

Operations in Multidimensional Data Model

- Aggregation (roll-up)
  - dimension reduction: e.g., total sales by city
  - summarization over aggregate hierarchy: e.g., total sales by city and year -> total sales by region and by year

- Selection (slice) defines a subcube
  - e.g., sales where city = Palo Alto and date = 1/15/96

- Navigation to detailed data (drill-down)
  - e.g., (sales - expense) by city, top 3% of cities by average income

- Visualization Operations (e.g., Pivot)
Approaches to OLAP Servers

- **Relational OLAP (ROLAP)**
  - Relational and Specialized Relational DBMS to store and manage warehouse data
  - OLAP middleware to support missing pieces
    - Optimize for each DBMS backend
    - Aggregation Navigation Logic
    - Additional tools and services
  - Example: Microstrategy, MetaCube (Informix)

- **Multidimensional OLAP (MOLAP)**
  - Array-based storage structures
  - Direct access to array data structures
  - Example: Essbase (Arbor), Accumate (Kenan)

- **Domain-specific enrichment**
Relational DBMS as Warehouse Server

- Schema design
- Specialized scan, indexing and join techniques
- Handling of aggregate views (querying and materialization)
- Supporting query language extensions beyond SQL
- Complex query processing and optimization
- Data partitioning and parallelism

Warehouse Database Schema

- ER design techniques not appropriate
- Design should reflect multidimensional view
  - Star Schema
  - Snowflake Schema
  - Fact Constellation Schema
**Example of a Star Schema**

<table>
<thead>
<tr>
<th>Order</th>
<th>Customer</th>
<th>Salesperson</th>
<th>Fact Table</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order No</td>
<td>Customer No</td>
<td>SalespersonID</td>
<td>OrderNO CustomerNO</td>
<td>ProductNo</td>
</tr>
<tr>
<td>Order Date</td>
<td>Customer Name</td>
<td>SalespersonID</td>
<td>ProdNo DateKey CityName</td>
<td>ProdName</td>
</tr>
<tr>
<td></td>
<td>Address</td>
<td>SalespersonName</td>
<td>Quantity</td>
<td>ProdDescr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>City</td>
<td></td>
<td>Category</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CategoryDescription</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Date</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DateKey Date</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>City CityName</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>State Country</td>
</tr>
</tbody>
</table>

**Star Schema**

- A single fact table and a single table for each dimension
- Every fact points to one tuple in each of the dimensions and has additional attributes
- Does not capture hierarchies directly
- Generated keys are used for performance and maintenance reasons
- Fact constellation: Multiple Fact tables that share many dimension tables
  - Example: Projected expense and the actual expense may share dimensional tables
Example of a Snowflake Schema

Snowflake Schema

- Represent dimensional hierarchy directly by normalizing the dimension tables
- Easy to maintain
- Saves storage, but is alleged that it reduces effectiveness of browsing (Kimball)
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

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 $B_v$ 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
Bit Map Index

Customers where Region = W And Rating = 1

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
Issues in Handling of Aggregate Views

- Important component for ROLAP Servers
- Representation in the context of star schema
  - Query Expressions
  - Materialized Views
- Logic for Aggregation Navigation
  - make optimum use of materialized aggregates to answer a query
- Choice of aggregate views to materialize
- HP Intelligent Warehouse pioneered some of the techniques

SQL Extensions for Front End Tools

- Extended Family of Aggregate functions
  - rank (top 10) and N-Tile (“top 30%” of all products)
  - Median, mode…..
- Reporting Features
  - running total, cumulative totals
- Results of multiple group by:
  - total sales by month and total sales by product
- SQL comes in the way of sequential processing and columnar aggregations
  - changes in total sale from 1994 to 1996, aggregated by brand
The storage model is an n-dimensional array.

Front end multidimensional queries map to server capabilities in a straightforward way.

Direct Addressing abilities.

A straightforward array representation has good indexing properties but very poor storage utilization when the data is sparse.

2-dimensional dense arrays indexed by B-Trees.
Population & Refreshing the Warehouse

- Data extraction
- Data cleaning
- Data transformation
  - Convert from legacy/host format to warehouse format
- Load
  - Sort, summarize, consolidate, compute views, check integrity, build indexes, partition
- Refresh
  - Propogate updates from sources to the warehouse

Data Cleaning

- Why?
  - Data warehouse contains data that is analyzed for business decisions
  - More data and multiple sources could mean more errors in the data and harder to trace such errors
  - Results in incorrect analysis
- Detecting data anomalies and rectifying them early has huge payoffs
- Important to identify tools that work together well
- Long Term Solution
  - Change business practices and data entry tools
  - Repository for meta-data
**Data Cleaning Techniques**

- **Transformation Rules**
  - Example: translate “gender” to “sex”
  - Products: Warehouse Manager (Prism), Extract (ETI)

- **Uses domain-specific knowledge to do scrubbing**

- **Parsing and fuzzy matching**
  - Multiple data sources (can designate a preferred source)
  - Products: Integrity (Vality), Trillum

- **Discover facts that flag unusual patterns (auditing)**
  - Some dealer has never received a single complaint
  - Products: QDB, SBStar, WizRule

**Load**

- **Issues:**
  - huge volumes of data to be loaded
  - small time window (usually at night) when the warehouse can be taken off-line
  - When to build indexes and summary tables
  - allow system administrator to monitor status, cancel suspend, resume load, or change load rate
  - restart after failure with no loss of data integrity

- **Techniques:**
  - batch load utility: sort input records on clustering key and use sequential I/O; build indexes and derived tables
  - sequential loads still too long (~100 days for TB)
  - use parallelism and incremental techniques
Refresh

> Issues:
> 
> - when to refresh
>   - on every update: too expensive, only necessary if OLAP queries need current data (e.g., up-the-minute stock quotes)
>   - periodically (e.g., every 24 hours, every week) or after “significant” events
>   - refresh policy set by administrator based on user needs and traffic
>   - possibly different policies for different sources
> - how to refresh

Refresh Techniques

- Full extract from base tables
  - read entire source table or database: expensive
  - may be the only choice for legacy databases or files.
- Incremental techniques (related to work on active dbs)
  - detect & propagate changes on base tables: replication servers (e.g., Sybase, Oracle, IBM Data Propagator)
    - snapshots & triggers (Oracle)
    - transaction shipping (Sybase)
  - Logical correctness
    - computing changes to star tables
    - computing changes to derived and summary tables
  - optimization: only significant changes
  - transactional correctness: incremental load
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

Metadata 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.
Creating and managing a warehouse is hard.

**Development tools**
- defining & editing metadata repository contents (schemas, scripts, rules).
- Queries and reports
- Shipping metadata to and from RDBMS catalogue (e.g., Prism Warehouse Manager).

**Planning & analysis tools**
- impact of schema changes
- capacity planning
- refresh performance: changing refresh rates or time windows

**Warehouse Management Tools**

- Monitoring and reporting tools (e.g., HP Intelligent Warehouse Advisor)
  - which partitions, summary tables, columns are used
  - query execution times
  - for summary tables, types & frequencies of roll downs
  - warehouse usage over time (detect peak periods)

- Systems and network management tools (e.g., HP OpenView, IBM NetView, Tivoli): traffic, utilization

- Exception reporting/alerting tools (e.g., DB2 Event Alerters, Information Advantage InfoAgents & InfoAlert)
  - runaway queries

- Analysis/Visualization tools: OLAP on metadata
State of Commercial Practice

  - Connectivity to sources
    - Apertus
    - Information Builders EDA-SQL
    - Informix Enterprise Gateway
    - Oracle Open Connect
    - SAS Connect
    - Sybase Enterprise Connect
  - Data extract, clean, transform, refresh
    - CA-Ingres Replicator
    - Evolutionary Tech Inc. ETI-Extract
    - IBM Data Joiner, Data Propagator
    - Platinum InfoRefiner, InfoPump
    - Prism Warehouse Manager
    - SAS Access
    - Sybase Replication Server

State of Commercial Practice..2

- Multidimensional Database Engines
  - Arbor Essbase
  - Oracle IRI Express
  - Comshare Commander OLAP
  - SAS System

- Warehouse Data Servers
  - CA-Ingres
  - Information Builders Focus
  - Oracle
  - Redbrick
  - Sybase MPP
  - Teradata

- ROLAP Servers
  - HP Intelligent Warehouse
  - Informix Metacube
  - Information Advantage Assys
  - MicroStrategy DSS Server
### State of Commercial Practice..3

#### Query/Reporting Environments
- Brio Query
- Cognos Imprompta
- IBM DataGuide
- Informix ViewPoint
- SAS Access

#### Metadata Management
- HP Intelligent Warehouse
- Prism Directory Manager

#### System Management
- CA Unicenter
- IBM DataHub, NetView
- Prism Warehouse Manager
- Tivoli
- Redbrick Enterprise Control and Coordination

#### Process Management
- At& T TOPEND
- IBM FlowMark
- Prism Warehouse Manager

#### Multidimensional Analysis
- Andydne Pablo
- Business Objects
- Dimensional Insight Cross Target
- Information Advantage Decision Suite
- Kenan System Acumate
- Microsoft Excel
- Pilot Lightship
- Prodea Beacon
- Stanford Technology Group Metacube

### State of Commercial Practice..4

#### Metadata Management
- HP Intelligent Warehouse
- IBM Data Guide
- Prism Directory Manager

#### System Management
- CA Unicenter
- IBM DataHub, NetView
- Prism Warehouse Manager
- Tivoli
- Redbrick Enterprise Control and Coordination

#### Process Management
- At& T TOPEND
- IBM FlowMark
- Prism Warehouse Manager

#### Systems integration and consulting
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.

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
- use of workflow technology for process management