OLAP and Data Warehousing

* Marked slides courtesy of:
  Surajit Chaudhuri
  Microsoft Research

  Umeshwar Dayal
  Hewlett-Packard Labs.

Additional slides courtesy of:

  Julia Stoyanovich
  Columbia University
What is OLAP?

- **On-Line Analytical Processing**
- Information technology to help the knowledge worker (executive, manager, analyst) make faster and better decisions.
- OLAP is an element of decision support systems (DSS).

Running Example: Car Sales

- Cars: `carId`, make, model, color
- Dealers: `dealerId`, city, state
- Time of Sale: `tid`, year, month, day
- Sales: `carId`, `dealerId`, `tid`, price
OLTP Queries: Examples

- create a new sales record that indicates that a red VW Golf was sold in Boston, MA
- see how many black and silver VW Passats were sold at dealership #123 on April 11, 2016

OLAP Queries: Examples

- Analyze comparative sales of the different colors of VW Golf by state
- See which months are particularly favorable to the sale of different VW models and colors
- Rank VW dealerships by revenue, displaying a ranked list of dealerships and % differences in sales between each dealership and the one ranked 1 place higher
OLAP vs. OLTP

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>Clerk, IT professional</td>
<td>Knowledge worker</td>
</tr>
<tr>
<td>Function</td>
<td>Day to day operations</td>
<td>Decision support</td>
</tr>
<tr>
<td>DB design</td>
<td>Application-oriented</td>
<td>Subject-oriented</td>
</tr>
<tr>
<td></td>
<td>(E-R based)</td>
<td>(Star, snowflake)</td>
</tr>
<tr>
<td>Data</td>
<td>Current, Isolated</td>
<td>Historical, Consolidated</td>
</tr>
<tr>
<td>View</td>
<td>Detailed, Flat relational</td>
<td>Summarized, Multidimensional</td>
</tr>
<tr>
<td>Usage</td>
<td>Structured, Repetitive</td>
<td>Ad hoc</td>
</tr>
<tr>
<td>Unit of work</td>
<td>Short, simple transaction</td>
<td>Complex query</td>
</tr>
<tr>
<td>Access</td>
<td>Read/write</td>
<td>Read mostly</td>
</tr>
<tr>
<td>Operations</td>
<td>Index/hash on prim. key</td>
<td>Lots of scans</td>
</tr>
<tr>
<td># Records accessed</td>
<td>Tens</td>
<td>Millions</td>
</tr>
<tr>
<td># Users</td>
<td>Thousands</td>
<td>Hundreds</td>
</tr>
<tr>
<td>Db size</td>
<td>100 MB - GB</td>
<td>100 GB - TB</td>
</tr>
<tr>
<td>Metric</td>
<td>Trans. throughput</td>
<td>Query throughput, response</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 trans. workloads.
  » Complex OLAP queries would degrade performance for operational transactions.
  » 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 data consolidation (aggregation, summarization) 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 Architecture
OLAP Queries: Challenges

- Many AND, OR in the WHERE clause
- Self-join, nested sub-queries
  - Last year’s sales vs this year’s sales for each product
  - Show reps for whom every sale has been more than $15000
- Extensive use of aggregation, often on related datasets
- Aggregation over time periods
- Ranking
- Use of statistical functions
- Very large datasets
- Expectation of an interactive response time

OLAP Query Tools

- Goal of OLAP is to support ad-hoc querying for the business analyst (Power user)
- 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 amount, budget
- A set of dimensions with respect to 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

Attribute Hierarchies

- Attributes of a dimension may be related
- An m:1 dependency is most common
- Dependency graph may be:
  - Hierarchy: e.g.,
    - city -> state -> country
  - Lattice:
    - date -> month -> year
    - date -> week -> year
- Hierarchies are most common
- Dependencies influence choice of operations and data representation
Multidimensional Data

Sales volume as a function of product, time, geography

Operations on Multidimensional Data Model

- Aggregation (roll-up) of detailed data to create summary data
- Navigation to detailed data (drill-down) from summary
- Selection (slice) defines a subcube
  - Project the cube on fewer dimensions by specifying coordinates of remaining dimensions
  - e.g., sales where state = NY and month = Jan
- Calculation
  - Within a dimension, e.g., (sales - expense) by state
  - Across dimensions
- Ranking
  - top 3% of states by average sales
- Window Queries
Roll-up and Drill-Down

◆ Roll-Up: Use of aggregation
  » *dimension reduction:*
    – e.g., total sales by state by color
    – e.g., total sales by state
  » *navigating attribute hierarchy:*
    – e.g., sales by `city` -> total sales by `state` -> total sales by `country`
    – e.g., total sales by `city` and year -> total sales by `state` and year
      -> total sales by `country`

◆ Drill-Down: Inverse operation of roll-up
  » Provides the data set that was aggregated
    – e.g., show “base” data for total sales figure for CA state

Slice and Dice

◆ What colors of Golf are not doing so well?

  ```
  Select color, sum(price)
  From SALES
  Where model = 'Golf'
  Group By color
  ```

◆ Keep slicing if results are uniform
Multiple Aggregations

- Create a 2-dimensional spreadsheet that shows sum of sales by year as well as by state
- Each subtotal requires a separate aggregate query

<table>
<thead>
<tr>
<th>STATE</th>
<th>Sum by Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEAR</td>
<td>Sum By State</td>
</tr>
</tbody>
</table>

Example:
Multiple Aggregations

<table>
<thead>
<tr>
<th></th>
<th>WI</th>
<th>CA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>63</td>
<td>81</td>
<td>144</td>
</tr>
<tr>
<td>2014</td>
<td>38</td>
<td>107</td>
<td>145</td>
</tr>
<tr>
<td>2015</td>
<td>75</td>
<td>35</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>176</td>
<td>223</td>
<td>399</td>
</tr>
</tbody>
</table>
Generalization: The Data Cube

◆ Base tuples
◆ Aggregate tuples:
  » one aggregation for each subset of dimensions (powerset)
  » exponential number of subsets, but can optimize the computation
◆ Example
  » N = 3 dimensions
    – model = {Golf, Jetta}
    – color = {red, black, white}
    – state = {NY, CA, WI}
  » How many aggregate tuples in the data cube?
    – face – 1D agg; edge – 2D agg; corner – 3D agg

ROLAP and MOLAP

◆ 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
◆ Multidimensional OLAP (MOLAP)
  » Array-based storage structures
  » Direct access to array data structures
Warehouse Database Schema

- Entity-Relationship design techniques not appropriate
- Design should reflect multidimensional view
- Typical schemas:
  » Star Schema
  » Snowflake Schema
  » Fact Constellation Schema

Example of a Star Schema
Star Schema and Variants

- A single fact table and a single table for each dimension
- Generated keys are used for performance and maintenance reasons
- Fact constellation: Multiple Fact tables that share common dimension tables
  - Example: ProjectedExpense and ActualExpense may share dimensional tables
- Snowflake Schema: Represents dimensional hierarchy by normalization

Example of a Snowflake Schema
Performance Considerations

- Normalization for dimension tables
  » Read-only data, so no update anomalies
  » Fewer joins – better performance
- Pre-computation of summary tables
  » Re-use can speed up performance
  » How can we use pre-computed results effectively?
- Data is very large, dimension data often sparse
  » Crucial to use indexes effectively
  » Need for new indexing techniques: bitmap indexes, join indexes

Bit Map Index

- An alternative representation of RID-list
- Comparison, join and aggregation operations are reduced to bit arithmetic
- Specially advantageous for low-cardinality domains
  » Significant reduction in space and I/O (30:1)
  » Adapted for higher cardinality domains
  » Compression (e.g., run-length encoding) exploited
Join Index

- Traditional index maps the value in a column to a list of rows with that value
- Join index maintain relationships between attribute value of a dimension and the matching rows in the fact table
- Join index may span multiple dimensions (composite join index)
  - Use join index to identify regions of cartesian product that are of interest
  - Few people in Southern California may buy umbrellas
Join Index over Star Schema