Operational Databases
-
- Handles day-to-day operations of an organization
- A.K.A. Online Transaction Processing (OLTP) systems
- Characterized by
  - Content – detailed and current
  - Users – client and employees
  - Access pattern – short, atomic, r/w transactions
  - Design – ER, normalized

Data Warehouse
-
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision making process" – W. H. Inmon

Characteristics of Data Warehouse
-
- Subject-oriented
- Integrated
- Time-variant
- Nonvolatile
- To support decision making

Data Warehouse vs. Operational Database

<table>
<thead>
<tr>
<th>Operational Database</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Detailed and current</td>
</tr>
<tr>
<td>Users</td>
<td>Clients and employees</td>
</tr>
<tr>
<td>Access Patterns</td>
<td>short, atomic, r/w transactions</td>
</tr>
<tr>
<td>Design</td>
<td>ER, normalized</td>
</tr>
</tbody>
</table>

Data Warehouse Architecture

- Query/report
- Analysis
- Data mining

OLAP servers
- Monitoring
- Admin
- OLAP servers
- Data warehouse servers
- Metadata repository
- Operation databases
- External data sources
Why The Multidimensional Model

- Decision support applications are dominated by queries involved aggregations and group-bys
- And such queries often can’t be expressed or executed efficiently by OLTP databases

Standard SQL Aggregation Functions

- Operate on multiple rows and return a single result
  - sum
  - avg
  - count
  - max and min

GROUP BY

select category, count(id) from products group by category;

products

<table>
<thead>
<tr>
<th>id</th>
<th>category</th>
<th>description</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CPU</td>
<td>Intel Core 2 Duo</td>
<td>$200.00</td>
</tr>
<tr>
<td>2</td>
<td>CPU</td>
<td>Intel Pentium D</td>
<td>$198.99</td>
</tr>
<tr>
<td>3</td>
<td>CPU</td>
<td>AMD Athlon 64</td>
<td>$74.49</td>
</tr>
<tr>
<td>4</td>
<td>CPU</td>
<td>AMD Athlon 64x2</td>
<td>$115.98</td>
</tr>
<tr>
<td>5</td>
<td>HD</td>
<td>Seagate 320G</td>
<td>$77.49</td>
</tr>
<tr>
<td>6</td>
<td>HD</td>
<td>Maxtor 250G</td>
<td>$90.89</td>
</tr>
</tbody>
</table>

Understanding GROUP BY ...

- Without aggregation/GROUP BY
  
select category, id from products;

- With aggregation/GROUP BY
  
select category, count(id) from products group by category;

... Understanding GROUP BY

- With aggregation/GROUP BY
  
select category, count(id) from products group by category;

The Multidimensional Model

- Time
- City
- Product
- Sales
Data Cube

- Dimensions
  - Time, product, city ...
- Facts
  - Sales, units sold, expenses ...

Data Cube as a Lattice of Cuboids

Data

<table>
<thead>
<tr>
<th>time</th>
<th>item</th>
<th>location</th>
<th>supplier</th>
<th>sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>TV</td>
<td>LA</td>
<td>Sony</td>
<td>100</td>
</tr>
<tr>
<td>Jan</td>
<td>Laptop</td>
<td>NY</td>
<td>HP</td>
<td>120</td>
</tr>
<tr>
<td>Feb</td>
<td>TV</td>
<td>LA</td>
<td>Sony</td>
<td>110</td>
</tr>
<tr>
<td>Feb</td>
<td>TV</td>
<td>NV</td>
<td>Sony</td>
<td>200</td>
</tr>
<tr>
<td>Feb</td>
<td>Laptop</td>
<td>LA</td>
<td>HP</td>
<td>130</td>
</tr>
<tr>
<td>March</td>
<td>TV</td>
<td>LA</td>
<td>Sony</td>
<td>100</td>
</tr>
</tbody>
</table>

4-D Cuboid

- (time, item, location, supplier)
  - Number of cells??
  - Cells with non-zero values??

3-D Cuboids

- (time,item,location)
- (time,item,supplier)
- (item,location,supplier)
- (item,location, supplier)

Number of cells in each cuboid??
Cells with non-zero values??

Observations about Data Cubes

- Given a n-dimensional data cube, with each dimension having m values
  - Number of cuboids??
  - Number of cells??
Star Schema ...

<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Branch</th>
<th>Item</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Location</td>
<td>Branch</td>
<td>Item</td>
<td>Sales</td>
</tr>
<tr>
<td>Time</td>
<td>Location</td>
<td>Branch</td>
<td>Item</td>
<td>Sales</td>
</tr>
</tbody>
</table>

... Star Schema

- One Fact Table
  - E.g. sales
- One Dimension Table per dimension
  - E.g. time, item, branch, and location
- Concept Hierarchy
- Redundancy in dimension tables

Concept Hierarchies

- country
  - state
  - city
  - street
- year
  - quarter
  - month
  - day
  - week

- Total order: street < city < state < country
- Partial order: day < {month<quarter, week} < year

Other Schemas for Multidimensional Databases

- Snowflake schema
  - Some dimensions are normalized
  - Normalize the location dimension??
- Fact Constellation schema
  - Dimension tables are shared by more than one fact tables

OLAP Storage Strategies

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)

A ROLAP Data Store

- Summary fact tables

<table>
<thead>
<tr>
<th>RID</th>
<th>Item</th>
<th>Day</th>
<th>Month</th>
<th>Quarter</th>
<th>Year</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>TV</td>
<td>15</td>
<td>10</td>
<td>Q4</td>
<td>2003</td>
<td>250</td>
</tr>
<tr>
<td>1002</td>
<td>TV</td>
<td>23</td>
<td>10</td>
<td>Q4</td>
<td>2003</td>
<td>175</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RID</th>
<th>Item</th>
<th>Day</th>
<th>Month</th>
<th>Quarter</th>
<th>Year</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>5001</td>
<td>TV</td>
<td>10</td>
<td>Q4</td>
<td>2003</td>
<td>45,786</td>
<td></td>
</tr>
</tbody>
</table>
### Aggregation Functions (Measures)

- **Distributive**
  - sum, count, min, max
- **Algebraic**
  - avg = sum / count
- **Holistic**
  - median

### Estimate Median

\[
\text{median} = \frac{L_{\text{median}} + \frac{N}{2} - \sum freq}{freq_{\text{median}}} \cdot \text{width}_{\text{median}}
\]

### More About Aggregation Functions

- **Variance:**
  \[
  \frac{1}{N} \sum_{i=1}^{N} (x_{i} - \overline{x})^2
  \]

- **Incremental maintenance**
  - E.g. max, min, average, median

### OLAP Operations

- **Roll-up**
- **Drill-down**
- **Slice and dice**
- **Pivot (rotate)**

### Roll-up

- Aggregation on a data cube by
  - Going up a concept hierarchy, or
  - Reducing dimension(s)

### Drill-down

- Reverse of roll-up
  - Going down a concept hierarchy, or
  - Adding dimensions

---

**Estimated Median Calculation**

Given the data set: 23,5,12,20,2,11,14,16,18,19,8,1,21,25

\[
\text{median} = \frac{L_{\text{median}} + \frac{N}{2} - \sum freq}{freq_{\text{median}}} \cdot \text{width}_{\text{median}}
\]

**OLAP Operations Diagram**

- Time, Product, City
- Roll-up: Going up a concept hierarchy, or reducing dimension(s)
- Drill-down: Going down a concept hierarchy, or adding dimensions
Slice and Dice

◆ Slice: selection on one dimension
◆ Dice: selection on more than one dimensions
  ◆ E.g. (city = "LA") and (month = "Jan" or "Feb")

Pivot (Rotate)

◆ Rotate the data axes to provide an alternative presentation of the data

Perform OLAP Operations Efficiently

◆ Indexing
◆ Cube pre-computation

Bitmap Indexing ...

<table>
<thead>
<tr>
<th>rid</th>
<th>item</th>
<th>city</th>
<th>month</th>
<th>sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>TV</td>
<td>LA</td>
<td>Jan</td>
<td>100</td>
</tr>
<tr>
<td>1002</td>
<td>PC</td>
<td>LA</td>
<td>Jan</td>
<td>200</td>
</tr>
<tr>
<td>1003</td>
<td>PC</td>
<td>NY</td>
<td>Jan</td>
<td>150</td>
</tr>
<tr>
<td>1004</td>
<td>PC</td>
<td>NY</td>
<td>Feb</td>
<td>100</td>
</tr>
<tr>
<td>1005</td>
<td>Phone</td>
<td>NY</td>
<td>Jan</td>
<td>175</td>
</tr>
<tr>
<td>1006</td>
<td>TV</td>
<td>NY</td>
<td>Feb</td>
<td>200</td>
</tr>
<tr>
<td>1007</td>
<td>Phone</td>
<td>LA</td>
<td>Jan</td>
<td>300</td>
</tr>
<tr>
<td>1008</td>
<td>Phone</td>
<td>LA</td>
<td>Feb</td>
<td>120</td>
</tr>
</tbody>
</table>

… Bitmap Indexing

Bitmap Index on Item:  Bitmap Index on City ??

1 0 0
0 1 0
0 1 0
0 1 0
0 0 1
1 0 0
0 0 1
0 0 1
TV  PC  Phone

Using Bitmap Indexing

◆ List total sales in LA by item

```sql
select sum(sales), item
from sales_table
where city = 'LA'
group by item;
```
Join Indexing ... 

<table>
<thead>
<tr>
<th>Location</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location_key</td>
<td>Item_key</td>
</tr>
<tr>
<td>Street</td>
<td>Item_name</td>
</tr>
<tr>
<td>City</td>
<td>Brand</td>
</tr>
<tr>
<td>State</td>
<td>Type</td>
</tr>
<tr>
<td>Country</td>
<td>Item_type</td>
</tr>
</tbody>
</table>

... Join Indexing ...

<table>
<thead>
<tr>
<th>Location</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location_key</td>
<td>Street</td>
</tr>
<tr>
<td>1</td>
<td>123 Main St.</td>
</tr>
<tr>
<td>2</td>
<td>456 Wall St.</td>
</tr>
<tr>
<td>3</td>
<td>789 State St.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Item_key</th>
<th>Brand</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bravia 42in</td>
<td>Sony</td>
<td>TV</td>
</tr>
<tr>
<td>2</td>
<td>Bravia 46in</td>
<td>Sony</td>
<td>TV</td>
</tr>
<tr>
<td>3</td>
<td>Pavilion A100</td>
<td>HP</td>
<td>PC</td>
</tr>
<tr>
<td>4</td>
<td>Pavilion A200</td>
<td>HP</td>
<td>PC</td>
</tr>
<tr>
<td>5</td>
<td>iPhone</td>
<td>Apple</td>
<td>Phone</td>
</tr>
</tbody>
</table>

Using Pre-computed Cuboids ...

Consider data cube sales_cube:

\[ \text{sales_cube} \left[ \text{time, item, location} \right] = \text{sum}(\text{sales}) \]

- Time: day < month < quarter < year
- Item: item_name < brand < type
- Location: street < city < state < country

... Using Pre-computed Cuboids ...

- Pre-computed cuboids
  - Cuboid 1: \{year, item_name, city\}
  - Cuboid 2: \{year, brand, country\}
  - Cuboid 3: \{year, brand, state\}
  - Cuboid 4: \{item_name, state\} where year = 2004

Query
- \{brand, state\} where year = 2004 ??

Summary

- Architecture
- Data
  - Multidimensional data model – Data Cube
  - Logical and physical data organization
- Operations
  - Aggregation functions
  - OLAP operations
  - Efficient execution
- Readings: Chapter 3 of textbook