Data Warehousing Techniques and Standards

Presented by:
Davin Flatten
Systems Analyst
MassSAFE
University of Massachusetts
Introduction

• The definition of a data warehouse
  – What they are
  – What they are not

• The star schema
  – Dimensions
  – Facts

• RDBMS requirements for warehousing
  – ETL (Extraction, Transformation, Loading)
  – ROLAP (Relational Online Application Processing)

• Keys to success
What a data warehouse is not

• Not a OLTP (Online Transactional Processing) system
• Not necessarily a physical place
• Not a single project with an end
• Not a single product or application
What a data warehouse is

- Modeling technique specific to analysis
- Bill Inmon’s four characteristics of a data warehouse:
  - Subject based design
  - Integrated with your data
  - Nonvolatile ‘picture’ of your states data
  - Time variant view of your data
Where does the warehouse fit?
Why have a data warehouse?

- Supplements transportation data systems
- Avoids conflicting resources
- Increases your organization's understanding of your data
- Most importantly…
It completes the data life cycle.
The Star Schema

- Key concept for data warehousing
- Modeling technique that simplifies joins and tables
- Organizes data into a format that is easy for business users to understand
- Allows application developers to standardize ad-hoc queries
Elements of a Star Schema

• Dimension Tables
  - Easy to understand groupings of subject areas
  - Can be hierarchical used to drill down
  - Denormalized, decoded, and cleaned set of descriptive data elements

• Fact Tables
  - Contains foreign keys referencing dimension records
  - Contain either additive or semi-additive measures for analysis
Sample Citation Schema
The Dimension Table

- Each record contains a single town (grain)
- The associated information is denormalized and hierarchical
- All values are decoded in plain language

<table>
<thead>
<tr>
<th>L_TOWN_NAME</th>
<th>L_COUNTY_NAME</th>
<th>L_REGION_OF_STATE</th>
<th>L_STATE_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hudson</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Marlborough</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Natick</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Pepperell</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Shirley</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Sudbury</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Townsend</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Wayland</td>
<td>Middlesex</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Ashburnham</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Auburn</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Berlin</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Blackstone</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Bolton</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Boylston</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>Charlton</td>
<td>Worcester</td>
<td>Central</td>
<td>Massachusetts</td>
</tr>
</tbody>
</table>
The Fact Table

- Each record contains a single violation (grain)
- Each dimension is reference with a foreign key
- Measures are provided for fines and violations
The Star Query

- Regardless of subject matter, the same type of query can be used
- Results can be easily read and used by analysts
- No complicated outer-joins, sub-selects, or other complex SQL

```sql
select v.v_age_group_nhtsal "Age group",
       d.t_holiday_travel "Holiday Travel",
       sum(f-violations) "Total Violations"
from citation_mart_test.fact_violation f,
     citation_mart_test.dim_violator v,
     citation_mart_test.dim_date d
where f.violator_id = v.violator_id
     and f.date_id = d.date_id
     and d.t_year_4_digit = 2003
group by v.v_age_group_nhtsal,
        d.t_holiday_travel
order by v.v_age_group_nhtsal,
        d.t_holiday_travel;
```

<table>
<thead>
<tr>
<th>Age group</th>
<th>Holiday Travel</th>
<th>Total Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 03:10-14</td>
<td>Yes</td>
<td>17</td>
</tr>
<tr>
<td>6 04:15-19</td>
<td>No</td>
<td>136677</td>
</tr>
<tr>
<td>7 04:15-19</td>
<td>Yes</td>
<td>8370</td>
</tr>
<tr>
<td>8 05:20-24</td>
<td>No</td>
<td>215763</td>
</tr>
<tr>
<td>9 05:20-24</td>
<td>Yes</td>
<td>12372</td>
</tr>
<tr>
<td>10 06:25-34</td>
<td>No</td>
<td>272885</td>
</tr>
<tr>
<td>11 06:25-34</td>
<td>Yes</td>
<td>14135</td>
</tr>
<tr>
<td>12 07:35-44</td>
<td>No</td>
<td>212781</td>
</tr>
<tr>
<td>13 07:35-44</td>
<td>Yes</td>
<td>9848</td>
</tr>
<tr>
<td>14 08:45-54</td>
<td>No</td>
<td>123911</td>
</tr>
</tbody>
</table>
Selecting a Relational Database Management System (RDBMS)

- **ETL (Extraction, Transformation, Loading)**
  - Ability to create stored procedures
  - Job scheduler
  - Logging and security tracking

- **ROLAP**
  - Optimizer capable of performing star queries
  - Table partitioning across time
  - Bitmap index capabilities
Why we use Oracle for our warehouse

- Table partition pruning
- Star query optimizer hint
- Bitmap indexes
- PL/SQL stored procedures
- Transportable table spaces
- Query rewrite
- Materialized views
- Job scheduler
Keys to a successful Warehousing project

- Identified and involved warehouse users
- Strong and committed leadership
- Diversified project team
- Established partnerships with all key source data holders
- Incremental project plan that produces fast results
- Correct design philosophy