Case Study: Tycho Data Warehouse

Vladimir Zadorozhny, DINS, University of Pittsburgh

Challenge

38000 excel sheets
Heterogeneous
Semi-structured
missing “dirty”

Tycho Data Processing Architecture

Clean
Well-structured
Accessible
Example of a source document

Tycho Data Processing Architecture

GUI

Query Processor

Indices, Materialized Views, Replicated & Distributed Data

Performance Optimization & Load Balancing

Data Warehouse

Data Validation and Loading

Preloaded DB

Data Cleansing & Preloading

Source Data Files
Conceptual View of Data Cleansing & Preloading

SSIS Implementation
Using Fuzzy Logic

Fuzzy Group

Places
Louisiana
Louisianat
Louisians
Douisiana:
Louisiana

Reference
Louisiana
Indiana
Virginia

Fuzzy Lookup

Places
Louisiana
Louisianat
Louisians
Douisiana:
Louisiana

Places
Similarity
Confidence
Louisiana 0.989783 0.99766
Indiana 0.63542 0.77856
Virginia 0.57863 0.69832

Data Warehouse Schema
### Loading Statistics

<table>
<thead>
<tr>
<th>Batch</th>
<th>num_files</th>
<th>num_sheets</th>
<th>num_data_rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>337</td>
<td>206000</td>
</tr>
<tr>
<td>2</td>
<td>406</td>
<td>856</td>
<td>455902</td>
</tr>
<tr>
<td>3</td>
<td>521</td>
<td>4388</td>
<td>1560211</td>
</tr>
<tr>
<td>4</td>
<td>508</td>
<td>3922</td>
<td>1293274</td>
</tr>
<tr>
<td>5</td>
<td>522</td>
<td>6131</td>
<td>2670207</td>
</tr>
<tr>
<td>6</td>
<td>522</td>
<td>5243</td>
<td>2173204</td>
</tr>
<tr>
<td>7</td>
<td>517</td>
<td>3473</td>
<td>3960916</td>
</tr>
<tr>
<td>8</td>
<td>517</td>
<td>1561</td>
<td>1151614</td>
</tr>
<tr>
<td>9</td>
<td>406</td>
<td>1317</td>
<td>1548759</td>
</tr>
<tr>
<td>10</td>
<td>462</td>
<td>1837</td>
<td>1439066</td>
</tr>
<tr>
<td>11</td>
<td>295</td>
<td>1191</td>
<td>1044489</td>
</tr>
<tr>
<td>12</td>
<td>420</td>
<td>1721</td>
<td>1861490</td>
</tr>
<tr>
<td>13</td>
<td>478</td>
<td>3949</td>
<td>3238394</td>
</tr>
<tr>
<td>14</td>
<td>105</td>
<td>1050</td>
<td>1111946</td>
</tr>
<tr>
<td>15</td>
<td>181</td>
<td>648</td>
<td>499625</td>
</tr>
<tr>
<td>16</td>
<td>353</td>
<td>1308</td>
<td>1167299</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td><strong>6350</strong></td>
<td><strong>38932</strong></td>
<td><strong>25382386</strong></td>
</tr>
</tbody>
</table>
Dashboard
You can add/remove features using menu at the bottom. Double click a gauge to remove it. Select a city under "More cities" to add a one. Move mouse over the figures to explore more features. The next screen shows a motion chart (click there).

Motion Chart. You can explore animation in different formats. You can also scroll time both ways manually. Move mouse around to learn more. The next screen shows map and timeline exhibit (click there).
Problem of Data Fusion

- Data fusion is a process of resolving data conflicts due to redundancy and inconsistency in data extracted from multiple pre-existing data sources.
- Data fusion begins after the data heterogeneity is resolved.
  - mapping of data items in a homogenous schema with proper syntactic data cleaning (eliminating typos, misspellings, and formatting errors),
  - standardization of names, conversion of data types;
  - enforcing trivial integrity constraints (e.g., obvious temporal relationships, such as From <= To  (event cannot end before it begins)
    - To <= DRD (data cannot be recorded before the event of interest occurred)
    - DRD <= SPD (data cannot be published before it was recorded, etc.)
An Example of Data Conflict

<table>
<thead>
<tr>
<th>emp_name</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>1000</td>
</tr>
<tr>
<td>Brown</td>
<td>2000</td>
</tr>
<tr>
<td>Jones</td>
<td>3000</td>
</tr>
<tr>
<td>Smith</td>
<td>4000</td>
</tr>
</tbody>
</table>

Integrity Constraint: emp_name \(\rightarrow\) salary  
(employee can receive only one salary)

Tuples t1 and t2 are conflicting. DB is inconsistent.

Query: 
\[
\text{select sum(salary) from Emp}
\]

Consistent answer?

Temporal & Spatial Conflicts

| Measles | NYC | 10/10/1900 | 10/10/1920 | 700
|---------|-----|------------|------------|---
| Smallpox| NY  | 10/20/1900 | 10/20/1920 | 500
| Smallpox| NYC | 10/30/1920 | 10/30/1930 | 600
| Measles | NYC | 10/20/1910 | 10/30/1930 | 300

Measles reports: 700
Smallpox reports: 500 (NY) 600 (NYC)

Total number of Measles cases in New York City from 1900 to 1930: 
700+300 = 1000 ??? Temporal conflict between t1 and t4

Total number of Smallpox cases in New York State from 1900 to 1930: 
500+600 = 1100 ??? Spatial conflict between t2 and t3
**Naming Conflicts**

- t1 | source_ref1 | Yellow fever | NY | 10/10/1900 | 10/10/1920 | 700
- t2 | source_ref2 | Hepatitis | NY | 10/10/1900 | 10/10/1920 | 700
- t3 | source_ref4 | Hepatitis B | NY | 10/20/1910 | 10/30/1930 | 300

**Redundant and Inconsistent Database**

**Redundant:**

Smallpox reports:

- 600
- 300
- 150
- 100

**Redundant and Inconsistent:**

Measles reports in NYC:

- 200
- 400
- 200
- 300

<table>
<thead>
<tr>
<th>R1: 700</th>
<th>R2: 500</th>
</tr>
</thead>
</table>

More about it later ....