MIS2502: Data Analytics

Dimensional Data Modeling

Jing Gong

gong@temple.edu

http://community.mis.temple.edu/gong
Where we are...

- **Transactional Database**: Stores real-time transactional data.
- **Analytical Data Store**: Stores historical transactional and summary data.

Data entry flow:
- Data entry → Transactional Database
- Data extraction → Analytical Data Store
- Data analysis

Now we’re here...
What do we know so far?

Why are relational databases good for storing transaction data?

Why are they bad for analytical processing?

What’s the solution?
Some terminology

Data Warehouse
- Takes many forms
- Really is just a repository for data

Data Mart
- More focused
- Specially designed for analysis

Data Cube
- Organization of data as a “multidimensional matrix”
- Implementation of a Data Mart
How they all relate

The data in the operational database... 

...is put into a data warehouse...

...which feeds the data mart...

...and is analyzed as a cube.

We’ll start here.
The Data Cube

- Core component of Online Analytical Processing and Multidimensional Data Analysis
- Made up of “facts” and “dimensions”

Quantity sold and total price are measured facts.

**Why isn’t product price a measured fact?**
The Data Cube

The highlighted element represents all the M&Ms sold in Ardmore, PA in January, 2011

A single summary record representing a business event (monthly sales).
The Data Cube

The highlighted elements represent Famous Amos cookies sold on Temple’s Main campus from January to March, 2013.

This is called “slicing the data.”
The Data Cube

What do the orange highlighted elements represent?

What do the purple highlighted elements represent?
Could you have a data mart with five dimensions?

Then why does our example (and most others) only have three?
Transactional databases aren’t built around dimensions
- They don’t map well to cubes
- They aren’t set up for summarization

So we build a star schema
- Built around “dimensions” and “facts”
- Simplified relational model

The star schema facilitates
- Aggregating individual transactions
- Creation of cubes
A join to make the cube?

Conceptually yes, but storing the join would create many, many, many rows!
So summaries get stored in a “multidimensional matrix”

Periodically summarize the data and store it in the cube

Retrieve only the summary, not the raw data

Much more efficient, but can’t be changed (non-volatile)
It adds up fast...

1000 products

300 stores

365 days

=109,500,000 records per year!
Designing the Star Schema

Kimball’s Four Step Process for Data Cube Design
(Kimball et al., 2008)

1. Choose the business process
2. Identify the fact
3. Decide on the level of granularity
4. Identify the dimensions
Choose the business process

• What your data cube is “about”
• Determined by the questions you want to answer about your organization

<table>
<thead>
<tr>
<th>Question</th>
<th>Business Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who is my best customer?</td>
<td>Sales</td>
</tr>
<tr>
<td>What are my highest selling products?</td>
<td>Sales</td>
</tr>
<tr>
<td>Which teachers have the best student performance?</td>
<td>Standardized testing</td>
</tr>
<tr>
<td>Which supplier is offering us the best deals?</td>
<td>Purchasing</td>
</tr>
</tbody>
</table>

Note that a “business process” is not always about business.
Identify the fact

The data associated with the business event

<table>
<thead>
<tr>
<th>Keys</th>
<th>Measured, numeric data</th>
</tr>
</thead>
</table>
| • Unique identifiers for each event  
• For the event itself and the associated dimensions  
• Associates a combination of the dimensions to a unique business event  
• Example: Sales has Product_ID, Store_ID, and Time_ID | • Quantifiable information for each business event  
• Does not describe any particular dimension  
• Describes a particular combination of dimensional data  
• Example: Sales has quantity_sold and total_price. |

Try it for the “student performance” example.
Decide on the level of granularity

• Level of detail for each event (row in the table)
• Will determine the data in the dimensions

• Example: Who is my best customer?
  – The “event” is a sales transaction
  – Choices for time: yearly, quarterly, monthly, daily
  – Choices for store: store, city, state

How would you select the right granularity?
Identify the dimensions

• The key elements of the process needed to answer the question ("fact")
  – who, what, where, when, why, and how

• Example: Sales transaction
  – A “sale” is the fact
  – Occurs for a particular product, store, and time
  – Could this data mart tell you
    • What is the best selling product?
    • Who is the best customer?

Try it for the “student performance” example.
Data cube caveats

• The cube is “non volatile,” so you’re locked in
  – Measured facts
  – Dimensions
  – Granularity

• So choose wisely!
  – For example: You can’t track daily sales if “date” is monthly
  – So why not include every single sale and do no aggregation?

“In memory” analytics is changing all of this, but not quite yet...
Takeways

• Data warehouse vs. data mart vs. data cube

• Data Cube

• Star schema

• Kimball’s four step process for data mart design