MIS2502: Data Analytics

Extract, Transform, Load
Where we are...

**Transactional Database**
- Stores real-time transactional data

**Analytical Data Store**
- Stores historical transactional and summary data

Data entry

Data extraction

Data analysis
The data in the operational database... is put into a data warehouse... which feeds the data mart... and is analyzed as a cube.

Now let’s address this part...
Extract, Transform, Load (ETL)

Extract data from the operational data store

Transform data into an analysis-ready format

Load it into the analytical data store
The Actual Process

Transactional Database 1

Query

Data conversion

Data Mart

Query

Transactional Database 2

Query

Data conversion

Query

Relational database

Dimensional database
ETL’s Not That Easy!

Data Consistency
- What if the data is in different formats?

Data Quality
- How do we know it’s correct?
- What if there is missing data?
- What if the data we need isn’t there?
Data Consistency:
The Problem with Legacy Systems

• An IT infrastructure evolves over time
• Systems are created and acquired by different people using different specifications

This can happen through:
• Changes in management
• Mergers & Acquisitions
• Externally mandated standards
• Generally poor planning
This leads to many issues

Redundant data across the organization
• Customer record maintained by accounts receivable and marketing

The same data element stored in different formats
• Social Security number (123-45-6789 versus 123456789)

Different naming conventions
• “Doritos” versus “Frito-Lay’s Doritos” versus “Regular Doritos”

Different unique identifiers used
• Account_Number versus Customer_ID

What are the problems with each of these?
What’s the big deal?

This is a fundamental problem for creating data cubes.

We often need to combine information from several transactional databases.

How do we know if we’re talking about the same customer or product?
Now think about this scenario

What are the differences between a “guest” and a “customer”?

Is there any way to know if a customer of the café is staying at the hotel?
Solution: “Single view” of data

• The entire organization understands a unit of data in the same way

• It’s both a business goal and a technology goal

but it’s really more this...

...than this
Closer look at the Guest/Customer

<table>
<thead>
<tr>
<th>Guests</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guest_number</td>
<td>Customer_number</td>
</tr>
<tr>
<td>Guest_firstname</td>
<td>Customer_name</td>
</tr>
<tr>
<td>Guest_lastname</td>
<td>Customer_address</td>
</tr>
<tr>
<td>Guest_address</td>
<td>Customer_city</td>
</tr>
<tr>
<td>Guest_city</td>
<td>Customer_zipcode</td>
</tr>
<tr>
<td>Guest_zipcode</td>
<td></td>
</tr>
<tr>
<td>Guest_email</td>
<td></td>
</tr>
</tbody>
</table>

Getting to a “single view” of data:

<table>
<thead>
<tr>
<th>How would you represent “name?”</th>
<th>What would you use to uniquely identify a guest/customer?</th>
<th>Would you include email address?</th>
<th>How do you figure out if you’re talking about the same person?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Organizational issues

Why might there be resistance to data standardization?

Is it an option to just “fix” the transactional databases?

If two data elements conflict, who’s standard “wins?”
Data Quality

The degree to which the data reflects the actual environment

- Do we have the right data?
- Is the data accurate?
- Is the collection process reliable?
Finding the right data

Choose data consistent with the goals of analysis

Verify that the data really measures what it claims to measure

Include the analysts in the design process

Adapted from http://www2.ed.gov/about/offices/list/os/technology/plan/2004/site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf
Ensuring accuracy

Know where the data comes from

Manual verification through sampling

Use of knowledge experts

Verify calculations for derived measures

Adapted from http://www2.ed.gov/about/offices/list/os/technology/plan/2004/site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf
Reliability of the collection process

Build fault tolerance into the process

Periodically run reports, check logs, and verify results

Keep up with (and communicate) changes

Adapted from http://www2.ed.gov/about/offices/list/os/technology/plan/2004/site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf