What Do You Do With Data?

Gather

Store

Retrieve

Interpret
The Information Architecture of an Organization

Data entry → Transactional Database → Data extraction → Analytical Data Store → Data analysis

Transactional Database: Stores real-time transactional data

Analytical Data Store: Stores historical transactional and summary data

Called OLTP: Online transaction processing

Called OLAP: Online analytical processing
The Transactional Database

- Stores real-time, transactional data

In business, a transaction is the exchange of information, goods, or services.

For databases, a transaction is an action performed in a database management system.

Operational databases deal with both: they store information about business transactions using database transactions.

- Examples of transactions
  - Purchase a product
  - Enroll in a course
  - Hire an employee

- Data is in real-time
  - Reflects current state
  - How things are “now”
The Relational Paradigm

• How transactional data is collected and stored

• Primary Goal: Minimize redundancy
  – Reduce errors
  – Less space required

• Most database management systems are based on the relational paradigm
  – Oracle, Microsoft Access, SQL Server
The Relational Database
Online Retailer Example

• A series of tables with logical associations between them

• The associations (relationships) allow the data to be combined
Why more than one table?

- Every review has an associated product
- Every product can have a review
- Products and reviews have a unique ID number
- Split the details off into separate tables

<table>
<thead>
<tr>
<th>Product</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProductID</td>
<td>ReviewID</td>
</tr>
<tr>
<td>Name</td>
<td>ProductID</td>
</tr>
<tr>
<td>Description</td>
<td>StarRating</td>
</tr>
<tr>
<td>Price</td>
<td>Text</td>
</tr>
<tr>
<td>Shipping</td>
<td>ReviewerName</td>
</tr>
<tr>
<td>SalesRank</td>
<td>Likes</td>
</tr>
</tbody>
</table>

This is good because:

- Information is entered and stored once
- Minimizes redundancy
Analyzing transactional data

• Can be difficult to do from a relational database

• Having multiple tables is good for storage and data integrity, but bad for analysis
  – Tables must be “joined” together before analysis can be done

• The solution is the Analytical Data Store

Operational databases are optimized for storage efficiency, not retrieval

Analytical databases are optimized for retrieval and analysis, not storage efficiency and data integrity
The Analytical Data Store

• Stores historical and summarized data
  – “Historical” means we keep everything
• Data is extracted from the operational database and reformatted for the analytical database

We’ll discuss this in much more detail later in the course!!
The Dimensional Paradigm

Data is stored like this around a business event...

...and can be summarized like this for analysis...

Store
- Store_ID
- Store_Address
- Store_City
- Store_State
- Store_Type

Sales
- Sales_ID
- Product_ID
- Store_ID
- Time_ID
- Quantity Sold
- Total Price

Product
- Product_ID
- Product_Name
- Product_Price
- Product_Weight

Time
- Time_ID
- Day
- Month
- Year

Product
- M&Ms
- Diet Coke
- Doritos
- Famous Amos

Store
- Ardmore, PA
- Temple Main
- Cherry Hill, NJ
- King of Prussia, PA

Time
- Mar. 2012
- Feb. 2012
- Jan. 2012
Dimensional Data and the Data Cube

...or it can be expanded in detail like this so that data mining (complex statistical analysis) can be done.

<table>
<thead>
<tr>
<th>Sales ID</th>
<th>Qty. Sold</th>
<th>Total Price</th>
<th>Prod. ID</th>
<th>Prod. Name</th>
<th>Prod. Price</th>
<th>Prod. Weight</th>
<th>Store ID</th>
<th>Store Address</th>
<th>Store City</th>
<th>Store State</th>
<th>Store Type</th>
<th>Time ID</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparing Operational and Analytical Data Stores

<table>
<thead>
<tr>
<th>Operational Data Store</th>
<th>Analytical Data Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on Relational paradigm</td>
<td>Based on Dimensional paradigm</td>
</tr>
<tr>
<td>Storage of real-time transactional data</td>
<td>Storage of historical transactional data</td>
</tr>
<tr>
<td>Optimized for storage efficiency and data integrity</td>
<td>Optimized for data retrieval and summarization</td>
</tr>
<tr>
<td>Supports day-to-day operations</td>
<td>Supports periodic and on-demand analysis</td>
</tr>
</tbody>
</table>
The agenda for the course

Weeks 1 through 5
- Transactional Database
  - Stores real-time transactional data

Weeks 6 through 9
- Analytical Data Store
  - Stores historical transactional and summary data

Weeks 10 through 14
- Data extraction
- Data analysis

Data interpretation, visualization, communication