70-455
Modern Data Management

Wolfgang Gatterbauer
(gatt@cmu.edu)
Spring 2014
Why Data Management

In today's business world, you will have to use various forms of data to drive decisions.

---

Cloudera, which offers a popular version of the open source database called Hadoop, released software on Wednesday that makes it possible to run queries from a more mainstream SQL programming language interface. SQL, thanks to its adoption by Oracle, Microsoft and others, is known to millions of business analysts.

This enables us to talk to a whole other class of customer," said Mike Olson, the chief executive of Cloudera. "The knock against Hadoop was that it is too complex."

There is a reason for that. Hadoop is one of several so-called unstructured databases that were created at Yahoo and Google, after those two companies found they had previously unimaginable amounts of data about activities like people's Web-surfing habits. Put into databases designed to handle this unstructured behavior, then analyzed, this information was valuable for figuring out things like what advertisement to put in front of each individual Web surfer.
Three Integrated Parts of "Data Management"

Organize Data  Analyze Data  Synthesize Data

- Data & Self-Org.
- E/R diagrams
- Database admin

- Excel
- SQL

- Structured & Visual Communication
Five Corresponding Class Modules

- Data & Self-Org.
- E/R diagrams
- Database admin

- Excel
- SQL
- Structured & Visual Communication

1. Use of Excel
2. "Organization & Synth."
3. Use of SQL
4. Data modeling
5. Database admin
Module 1: Excel

*Basic Knowledge of Excel is assumed! We cover advanced functions such as:*

- Pivot tables
- Excel as database
- Lookup Tables
- Array Formulas
- Advanced formulas
- Basic VBA

*Example Array Formulas*

<table>
<thead>
<tr>
<th>Item</th>
<th>$/piece</th>
<th>quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Item 2</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Item 3</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Item 4</td>
<td>40</td>
<td>4</td>
</tr>
</tbody>
</table>

Total $ = 290

How to create the weighted sum without the intermediate results?

<table>
<thead>
<tr>
<th>Item</th>
<th>$/piece</th>
<th>quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Item 2</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Item 3</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Item 4</td>
<td>40</td>
<td>4</td>
</tr>
</tbody>
</table>

Total $ = 290
Module 2: "Organization & Synth."

- How do you organize yourself and your data?
- How do you synthesize your results into a concise recommendation?
- What is an appropriate data-driven chart?

We have 3 numbers per item. How can we represent all of them in a 2-dimensional plane?
Module 3: SQL

- What is SQL? What can it do that Excel cannot?

- How do I put data into my database? How do I update what is already there?

- How do I query my database to get the information that I am seeking?
Module 4: Data Modeling / Designing Databases

- What is *Data Modeling* and why is it useful?
- How do I create a data model? How can I tell if my data model is "good"?
- What is the *relational data model* and why is it useful?
- What is *normalization*? Why should I care?
- How do I create a database based on my data?
Module 5: Managing Database Systems

• How can I tune and scale my databases?

• What are transactions?

• When is it better to denormalize?

• How can I distribute a database across multiple machines or locations?

• What is the fuss about NOsql databases?
How does this class fit with other Tech classes

70-451: MIS

70-339: IT for Finance

70-465: IT Strategy

70-443: Digital Marketing...

70-453: System Analysis and Design

70-455: Modern Data Management

domain focus

technical skills
Administrative

• Most learning will happen in-class with hands-on exercises and student presentations throughout the course. Bring your laptops with a Windows partition!

• Prerequisites: Basic Excel and programming

• Workload outside of class: ~10 hrs/week
  - Flipped classroom: small preparations before class
  - 5-6 smaller assignments
  - one project in groups of 2-3
  - In-class Midterm and Final exams

• Syllabus: to be posted on Blackboard

• Class: Tue & Thu 1:30-2:50 pm (location TBD)
Excel books

Required

Other relevant books (not required)
• "Excel 2010 Formulas" by John Walkenbach, 2010 (~$30)


• "Excel 2010 Power Programming with VBA" by John Walkenbach (~$30)
"Synthesis" books

Not Required but highly recommended:
• "The Say it with Charts Complete Toolkit"
  by Gene Zelazny, 2006 (~$35)

Other recommended books (not required):
• "The Minto Pyramid Principle: Logic in Writing, Thinking, & Problem Solving"
  by Barbara Minto, 1996 (Used ~$75)

• "Blah Blah Blah: What To Do When Words Don't Work" by Dan Roam, 2011 (~$20)
Database books

Required:
• "Modern database management (10th ed)" by Hoffer, Ramesh, Topi, 2010 (used ~$70)
  www.amazon.com/Modern-Database-Management-Jeffrey-Hoffer/dp/1408264315/

Other relevant book (not required):
• "Fundamentals of Database Management Systems" by Mark Gillenson, 2011 (~$70)