Using open source tools for data warehousing and reporting
Ben Chapman
Assistant Dean for IT
Emory University School of Law
Introduction
Data comes to us from a variety of source systems. We are looking for a way to store, retrieve, and report on this data. This process needs to be scalable and repeatable.
Most of what we’re doing is based on this excellent book:

This book changed my life.
Introduction to Pentaho BI Suite

Pentaho is written in Java and comprises a suite of programs:

• Kettle - Kettle is an “ETL” (Extract, Transform, and Load) tool. It gets data from source systems into MySQL.
• Pentaho Report Designer - This is analogous to Crystal Reports or similar. It allows you to report on your data.
• Pentaho BI Server - This is a Java web application that allows easy reporting and web-based management of BI processes.
A brief non-technical digression before we dive in - I think it's important that we identify areas where law school IT can make unique contributions to the school.

What are some of these areas?

• Creating unique applications
• Doing unique analysis and adding understanding to local data
• Providing law-school specific instructional support
Current data processes - data goals

- Excel spreadsheets
- Powerpoint graphs and charts

Goals:

- Change this process to provide auditable processes with regard to data sources
- Change the data to being **read-only** and being **authoritative**
- Creating processes that do not depend on one or two people only
Data Sources

- Peoplesoft
  - Grades (output measures)
  - Unofficial Transcripts
- LSAC ACES2
  - Demographic data
  - LSAT/UGPA (input measures)
- Symplicity
  - Engagement
  - Participation
  - 1L/2L/Permanent employment
- Manually generated data
  - Soft skills/SBA roles/Leadership, etc
Getting Pentaho BI Suite Community Edition

Visit [http://community.pentaho.com/](http://community.pentaho.com/)

You’ll need Java and a MySQL datasource. I also recommend that you purchase the book mentioned earlier. There are a number of configuration resources on the net, so I’m not going to spend time on them during this presentation.
OK, let’s get started
• Identify data
• Get a sense of what’s in the data
• Prepare a transformation to load the data into a database table
• Data comes from
  • http://www.fakenamegenerator.com/
• Steps
  • Fire up Pentaho Data Integration or PDI or Kettle
  • Connect to database
  • Start playing!
Our goal: create a test import
<table>
<thead>
<tr>
<th>Ge</th>
<th>Giv</th>
<th>Mid</th>
<th>Sur</th>
<th>Str</th>
<th>City</th>
<th>State Zip</th>
<th>Cou</th>
<th>Undergr</th>
<th>Cum</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>li</td>
<td>Lise</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he</td>
<td>Lili</td>
<td>M</td>
<td>Dhu</td>
<td>251</td>
<td>Irvine</td>
<td>CA</td>
<td>92US</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he</td>
<td>Haie</td>
<td>H</td>
<td>Co</td>
<td>173</td>
<td>Madison</td>
<td>UT</td>
<td>84US</td>
<td></td>
<td>85</td>
<td>166</td>
</tr>
<tr>
<td>ke</td>
<td>Rose</td>
<td>H</td>
<td>Ad</td>
<td>399</td>
<td>sac</td>
<td>CA</td>
<td>95US</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mi</td>
<td>Kyle</td>
<td>M</td>
<td></td>
<td>572</td>
<td></td>
<td>TX</td>
<td>78US</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ne</td>
<td>An</td>
<td>S</td>
<td>Short</td>
<td>360</td>
<td>Da</td>
<td>OH</td>
<td>45US</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

File read - separator and quote chars have been autodetected!
Analysis job run in Data Cleaner tells us a good bit about the data:

This helps us tweak and change the data table design.
Returning to Kettle, we do something like this:

```
ALTER TABLE applicants MODIFY GivenName VARCHAR(24)
ALTER TABLE applicants MODIFY Surname VARCHAR(24)
ALTER TABLE applicants MODIFY StreetAddress VARCHAR(30)
ALTER TABLE applicants MODIFY EmailAddress VARCHAR(50)
ALTER TABLE applicants MODIFY Undergraduate_School_Long VARCHAR(50)
ALTER TABLE applicants MODIFY High_LSAT_Score INT
```
Playing with queries in SQL*Architect

select
Undergraduate_School_Long
, avg(cali_applicants.High_LSAT_Score) as LSAT
, count(surname) as Students
from
cali_applicants
group by Undergraduate_School_Long
order by Students DESC, LSAT DESC
SQL Power*Architect provides
• Quick query capability
• Graphical representations of tables
Working with PDI
Reference material for PDI steps
http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps
Pentaho Report Designer

- Sophisticated reporting with output in PDF, HTML, plain text
- Relatively easy to use
- We’re going to look at some sample output
- We’ll also work through a demo
<table>
<thead>
<tr>
<th>OCS Relationship</th>
<th>Financial</th>
<th>Certificate Prog.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationship: B</td>
<td>Scholarships: ☐ ☐</td>
<td>Lamar Inn: Yes</td>
</tr>
<tr>
<td>Engagement:</td>
<td>Loans: ☐</td>
<td>Bar Assn: Yes</td>
</tr>
</tbody>
</table>

**Employment Preferences**

- Geographical Preference: N/A
- Practice Area Preference: N/A
- Bar Exam Location: NY and NJ

**Employment**

- 1L Summer: Jut Clerk
- 2L Summer: Govt
- Perm employment status: Employed Full-time
- US: ☐
- Salary: ☐

<table>
<thead>
<tr>
<th>Transfer</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course name</td>
<td>Faculty member</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Civil Procedure I</td>
<td>Freer, Richard Dale</td>
</tr>
<tr>
<td>Contracts</td>
<td>Abrams, Howard Evan</td>
</tr>
<tr>
<td>Legal Writ., Ranch, &amp; Advoc Prog</td>
<td>Kirk, Aaron</td>
</tr>
<tr>
<td>Torts</td>
<td>Smith, Gary R</td>
</tr>
<tr>
<td>Legal Methods</td>
<td>Buzzbee, William W</td>
</tr>
<tr>
<td>Professionalism Program</td>
<td>Pratt, Janette B</td>
</tr>
<tr>
<td>Civil Procedure II</td>
<td>Freer, Richard Dale</td>
</tr>
<tr>
<td>Criminal Law</td>
<td>Nourse, Victoria F</td>
</tr>
<tr>
<td>Constitutional Law I</td>
<td>Schapiro, Robert A</td>
</tr>
<tr>
<td>Legal Writ., Ranch &amp; Advoc Prog</td>
<td>Kirk, Aaron</td>
</tr>
<tr>
<td>Property</td>
<td>Jackson, Sara S</td>
</tr>
<tr>
<td>Professionalism Program</td>
<td>Pratt, Janette B</td>
</tr>
<tr>
<td>Business Associations</td>
<td>Carney, William J</td>
</tr>
<tr>
<td>Alternative Dispute Resolution</td>
<td>Armstrong, Phillip M.</td>
</tr>
<tr>
<td>Criminal Proc: Investigation</td>
<td>Levine, Kay Leslie</td>
</tr>
<tr>
<td>Fundamentals of IncomeTaxation</td>
<td>Abrams, Howard Evan</td>
</tr>
<tr>
<td>Survey Employee Benefits Law</td>
<td>Hinson, Douglas</td>
</tr>
<tr>
<td>Personal</td>
<td>Before Law School</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Gender: $(gender)$</td>
<td></td>
</tr>
<tr>
<td>Ethnicity: $(ethnicity)$</td>
<td></td>
</tr>
<tr>
<td>DOB: $(date_of_birth)$</td>
<td></td>
</tr>
<tr>
<td>Perm City: $(permanent_city)$</td>
<td></td>
</tr>
<tr>
<td>Perm State: $(permanent_state)$</td>
<td></td>
</tr>
<tr>
<td>Birth City: $(birth_city)$</td>
<td></td>
</tr>
<tr>
<td>Birth State: $(birth_state)$</td>
<td></td>
</tr>
<tr>
<td>Birth Country: $(birth_country)$</td>
<td></td>
</tr>
<tr>
<td>EMPLID: $(emplid)$</td>
<td></td>
</tr>
</tbody>
</table>

Functions:
- $(fullname_short)$
- $(degree_applied_long)$
- $(photo)$
- $(LSAT)$
- $(ROPA)$
- $(GPA)$
- $(cumulative_gpa)$
- $(graduation_year)$
- $(school_long)$
- $(major)$
- $(graduate_major_long)$
- $(other_degrees)$
- $(graduate_degree_long)$
- $(years_of_reg)$
- $(years_off_reg)$
- $(notes)$
Pentaho Report Designer
• Traditional point and click report designer
• Includes the ability to do sub-reports
• Includes banded reports
• Does not seem to include the ability to do gridded reports
Links
• http://www.pentaho.com/
• http://www.sqlpower.ca/page/architect
• http://datacleaner.eobjects.org/

Ben Chapman
ben.chapman@emory.edu

Thank you!