

EMORY



**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:

R. Bouman, J van Dongen "Pentaho Solutions, Business Intelligence and Data Warehousing with Pentaho and MySQL" (Wiley, 2009) ISBN: 978-0-470-48432-6



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



Emory University School of Law

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



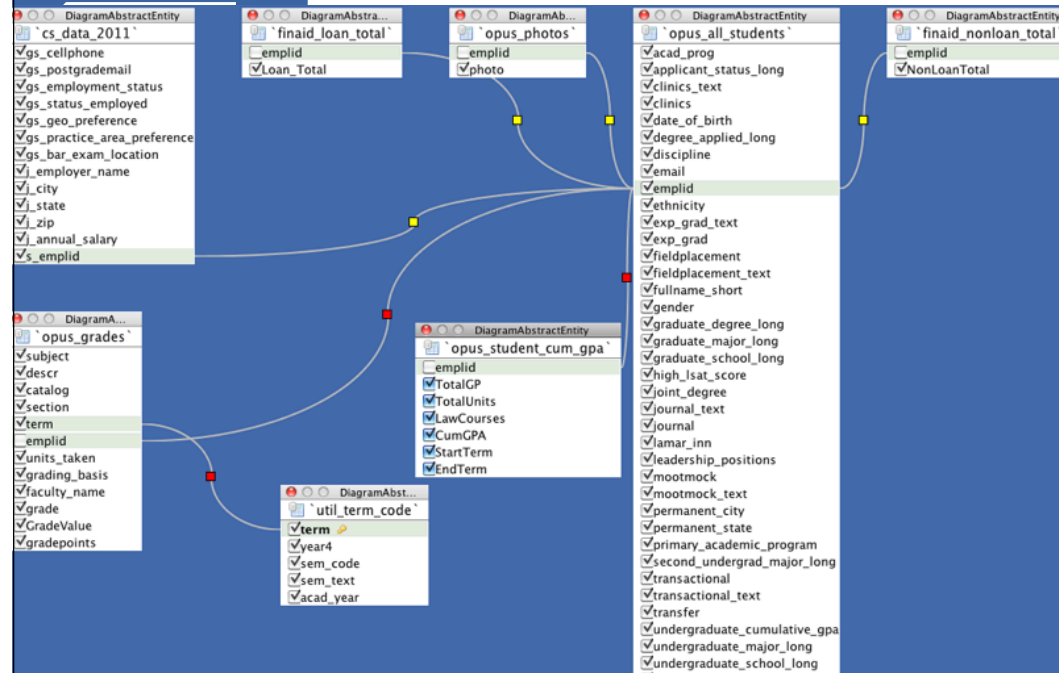
Current data processes - data goals

- 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 to 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/>

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

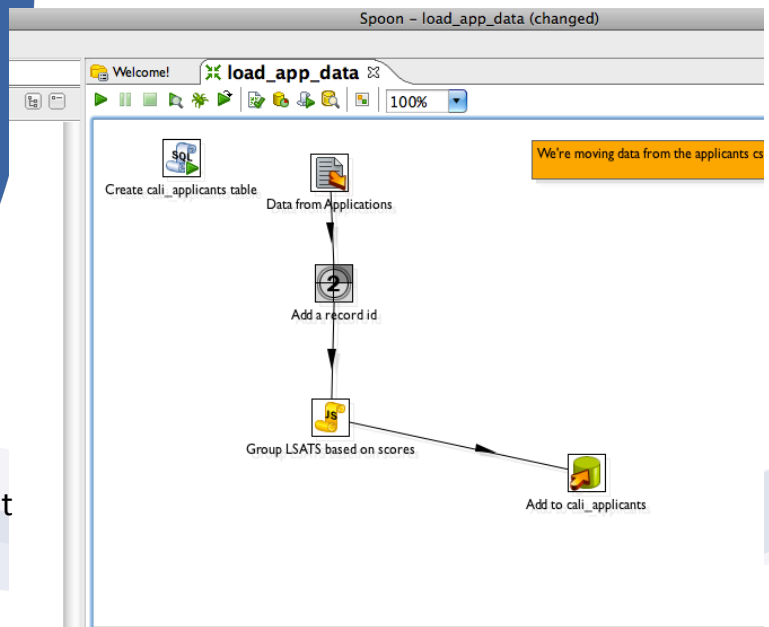
- Identify data
- Get a sense of what's in the data
- Prepare a transformation to load the data into a database table
- Data comes from
 - <http://www.fakenamegenerator.com/>
- Steps
 - Fire up Pentaho Data Integration or PDI or Kettle
 - Connect to database
 - Start playing!

EMORY



Our goal:
create a test
import

Emory University School of Law





CSV file datastore | DataCleaner

Comma-separated file

Datastore name: Admissions_Data

Filename: presentation/names/applicants.csv

Character encoding: ISO-8859-1

Separator: Comma (,)

Quote char: (None)

Ge...	Giv...	Mid...	Sur...	Stre...	City	State	Zip...	Cou...	Ema...	Undergr...	Cum...	High...
fe...	Lisa	D	Neely	282...	Tr...	MI	49...	US	Lisa...	FLORIDA...	3.81	161
fe...	Bra...	C	Smith	413...	Gr...	NC	27...	US	Bra...	GEORGIA...	3.95	166
fe...	Lilli...	M	Du...	251...	Irvine	CA	92...	US	Lilli...	LEONAR...	3.42	166
fe...	Da...	H	Co...	173...	Manti	UT	84...	US	Da...	WEBSTER...	2.97	166
fe...	Rose	H	Aut...	398...	Sac...	CA	95...	US	Ros...	UNIVERSI...	3.57	165
male	Wa...	M	Lyle	572...	Ke...	TX	78...	US	Wal...	WAYNE S...	2.95	134
fe...	An...	S	Short	360...	Da...	OH	45...	US	And...	JOHNS H...	3.11	174

Save datastore

File read - separator and quote chars have been autodetected!



Emory University School of Law

Analysis job run in Data Cleaner tells us a good bit about the data:

	A	B	C	D	E
1		Gender	GivenName	MiddleInitial	Surname
2	Row count	4000	4000	4000	4000
3	Null count	0	0	0	0
4	Entirely uppercase count	0	0	4000	1
5	Entirely lowercase count	4000	0	0	0
6	Total char count	19888	23075	4000	24881
7	Max chars	6	11	1	13
8	Min chars	4	2	1	2
9	Avg chars	4.972	5.769	1	6.22

This helps us tweak and change the data table design.



Returning to Kettle, we do something like this:

```
Simple SQL editor
SQL statements, separated by semicolon ';'
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
;
```



SQL Power*Architect

CALL_TEST

- Information_schema
- ELS_DWH
- ELS_STAGING
 - aces2_complete (TABLE)
 - aces2_matrix (TABLE)
 - cs_data (TABLE)
 - els_students (TABLE)
 - fakenames (TABLE)
 - opus_finaid (TABLE)
 - opus_grades (TABLE)
 - opus_students (TABLE)
 - schools_data (TABLE)
 - util_gpa_value (TABLE)
 - util_term_code (TABLE)
 - applicants (TABLE)
 - call_applicants (TABLE)
- Columns folder for call_applicants
 - Gender: VARCHAR(6)
 - GivenName: VARCHAR(24)
 - MiddleInitial: CHAR(1)
 - Surname: VARCHAR(24)
 - StreetAddress: VARCHAR(40)
 - City: VARCHAR(40)
 - State: VARCHAR(2)
 - ZipCode: BIGINT(19)
 - Country: VARCHAR(2)
 - EmailAddress: VARCHAR(50)
 - Undergraduate_School_Long: VARCHAR(40)
 - Cumulative_GPA: DOUBLE(22)
 - High_LSAT_Score: INT(10)

CALL_TEST

```
SELECT * FROM ELS_STAGING.call_applicants
```

Gender	GivenName	MiddleInitial	Surname	StreetAddress
female	Lisa	D	Neely	2822 Owen Lane
female	Brandi	C	Smith	4131 Edwards Stre
female	Lillian	M	Dugger	2516 Cimmaron Rc
female	Dawna	H	Cooke	1739 Lang Avenue
female	Rose	H	Autrey	398 Byers Lane
male	Walter	M	Lyle	572 Crestview Terr
female	Andrea	S	Short	3603 Boogress Stre
male	Samuel	A	Carver	1951 Ryan Road
female	Lavon	M	Burrows	4257 John Avenue
female	Mildred	J	Walraven	4084 Briarwood Rc
female	Susan	O	Briones	3949 Stonepot Roa
female	Lois	H	Castro	3882 Essex Court
female	Alice	R	Hoover	4549 Francis Mine
female	Natacha	D	Gustavson	870 Golden Street
male	Trina	M	Conner	7788 Turkey Bay P



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
```

EMORY



Emory University School of Law

SQL Power*Architect provides

- Quick query capability
- Graphical representations of tables

EMORY



Emory University School of Law

Working with PDI

Reference material for PDI steps

<http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps>

EMORY



Emory University School of Law

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

Ethnicity: Hispanic/Latino
DOB: [REDACTED]
Perm City: [REDACTED]
Perm State: NJ
Birth City: [REDACTED]
Birth State: NJ
Birth Country: N/A
EMPLID: [REDACTED]

Law School Engagement

Journals: N/A
Moot/Mock: N/A
Clinics: N/A
Field Placement: CI Plcmt:Federal
Defender,CI Plcmt:US Attorney,CI Plcmt:US
Attorney

OCS Relationship

Relationship: **B**
Engagement:

Financial

Scholarships: [REDACTED]
Loans: [REDACTED]

Certificate Prog:

Leadership: No
Research Asst: No
Teaching Fellow: No

Employment Preferences

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

Lamar Inn: Yes

Bar Assn:

Pro Bono: Yes

Advisor: Virginia C [REDACTED]
[REDACTED]
[REDACTED], LLP

Employment

1L Summer: Jud Clerk
2L Summer: Govt
Perm employment status: Employed Full-time
US [REDACTED]
Salary: [REDACTED]

Transfer: No

EMORY



Emory University School of Law

Course name	Faculty member
Civil Procedure I	Freer, Richard Dale
Contracts	Abrams, Howard Evan
Legl Writ, Rsrch, & Advoc Prog	Kirk, Aaron
Torts	Smith, Gary R
Legal Methods	Buzbee, William W
Professionalism Program	Pratt, Janette B
Civil Procedure II	Freer, Richard Dale
Criminal Law	Nourse, Victoria F
Constitutional Law I	Schapiro, Robert A
Legal Writ., Rsrch & Advoc Prog	Kirk, Aaron
Property	Jackson, Sara S
Professionalism Program	Pratt, Janette B
Business Associations	Camey, William J
Alternative Dispute Resolution	Armstrong, Phillip M.
Criminal Proc: Investigation	Levine, Kay Leslie
Fundamentals of Income Taxation	Abrams, Howard Evan
Survey/ Employee Benefits Law	Hinson, H. Douglas

EMORY



Emory University School of Law

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

EMORY



Emory University School of Law

Links

- <http://www.pentaho.com/>
- <http://www.sqlpower.ca/page/architect>
- <http://datacleaner.eobjects.org/>

Ben Chapman

ben.chapman@emory.edu

Thank you!