Pentaho Data Integration/Kettle

- Dan Moore
- 8z Real Estate
- Kettle user for two years

Questions

- Who has used a relational database?
- Who has written scripts or java code to munge data from one source and load it to another?
  - What did you use?
  - Scripts
  - Custom java code
  - ETL tool

What is Kettle?

- Batch data integration and processing tool written in Java
- Exists to retrieve, process and load data
- ETL
  - Extract, transform and load
- PDI synonomous

What is Kettle good for

- Mirroring data from master to slave
- Syncing two data sources
- Processing data retrieved from multiple sources and pushed to multiple destinations
- Loading data to RDBMS
- Datamart/data warehouse
  - Dimension lookup/update step
- Graphical manipulation of data

Alternatives

- Code
  - Custom java
  - Spring batch
- Scripts
  - perl, python, shell, etc
  - Possibly + db loader tool and cron
- Commercial ETL tools
  - Oracle Warehouse Builder
  - Datastage
  - Informatica
  - SQL Server Integration services
- Open source ETL tools:
  - Talend
  - KETL
  - Clover.ETL
- Special case tools
  - SymmetricDS
  - Db replication

Why Kettle is better

- Higher level than code
  - Graphical interface
  - No connection pooling to worry about
  - No DDL to write
  - Validation/business rules
- Well tested full suite of components
- Data analysis tools
  - Preview
  - Data profiling with data cleaner (add on)
- Free (as in beer and speech)
  - Two editions
  - GPLv2
- Performant?
  - Developer vs computer performant
  - Depends, right?
  - Sitemailsame job 10k rows/second for 125M rows
- Leverage java
  - jvm tuning skills
  - java libraries and logic (in jars)

Data sources

- Files
- Databases
- No SQL
- REST
- XML
- Hadoop/HBase
- JSON
- Excel
- EDI
- RSS
- Google Analytics

Kettle concepts

- Repository
- Rows/Stream
- Steps
- Job
- Transformation

Demo 1: one way sync

- Sync tables

Demo 2: processing

- Process data from one table and replace some values, filter some values
- Lookup table

Demo 3: log file processing

- Load apache logs for analysis

What it is not good for

- User interfaces/user interaction
- Small data sets
  - 500 (from experience)
- Web applications
- One off processes?
  - One off becomes regular

Who uses

- Survey results
  - ~20 people
- Number of downloads: 110K downloads of Kettle 4.4
  - Since Nov 2012
- Our specific use
  - MLS data
    - Different data source formats and types (jdbc, local csv, ftp)
  - Public records data
    - Fixed width files

Larger picture

- Kettle 10 years old
  - joined Pentaho about 7 years ago
- Open source, at version 4.4
  - GPLv2 license
  - EE edition available
- BI suite
  - Reporting
  - Analytics
  - Dashboards
  - Machine Learning (weka)

Kettle tools

- Spoon
- Kitchen
- Pan
- Carte
  - Clustering tool

Advanced topics

- Existing java logic
  - Embedded
  - Polygon example
  - Demo 4
- Deployment
  - Variables Config files are your friend
- Mapping/Parameterization
  - Subroutines of logic

Advanced Topics Continued

- Testing
  - Who tests
- Version control
  - Who uses version control
- Error handling
  - Email
  - Log files

Getting started

- Download
  - sourceforge
    - Includes over 150 example transformations
  - Mysql 3.14 jdbc driver
- Helpful sites
  - Forums: http://forums.pentaho.com/forumdisplay.php?135-Pentaho-Data-Integration-Kettle
  - Wiki: http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps
  - Testing: http://www.mooreds.com/wordpress/pentaho-kettle-testing
- Helpful books
  - Pentaho Kettle Solutions: Casters, Bouman, van Dongen
- Barely scratched surface
- Don't like tools that turn me into a mechanic