

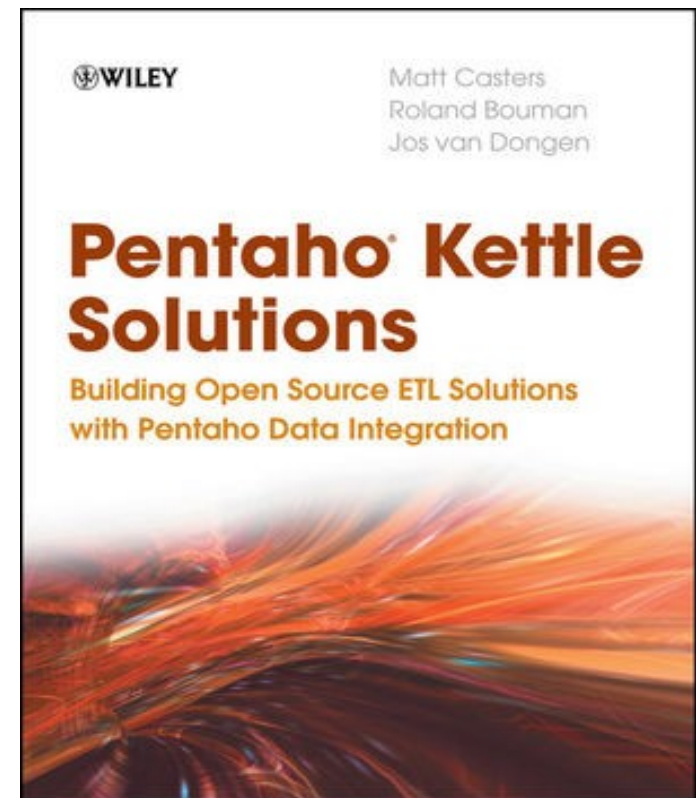


## Integrating the worlds data with Pentaho Data Integration

Matt Casters  
*CodeBits Lisbon November 2011*

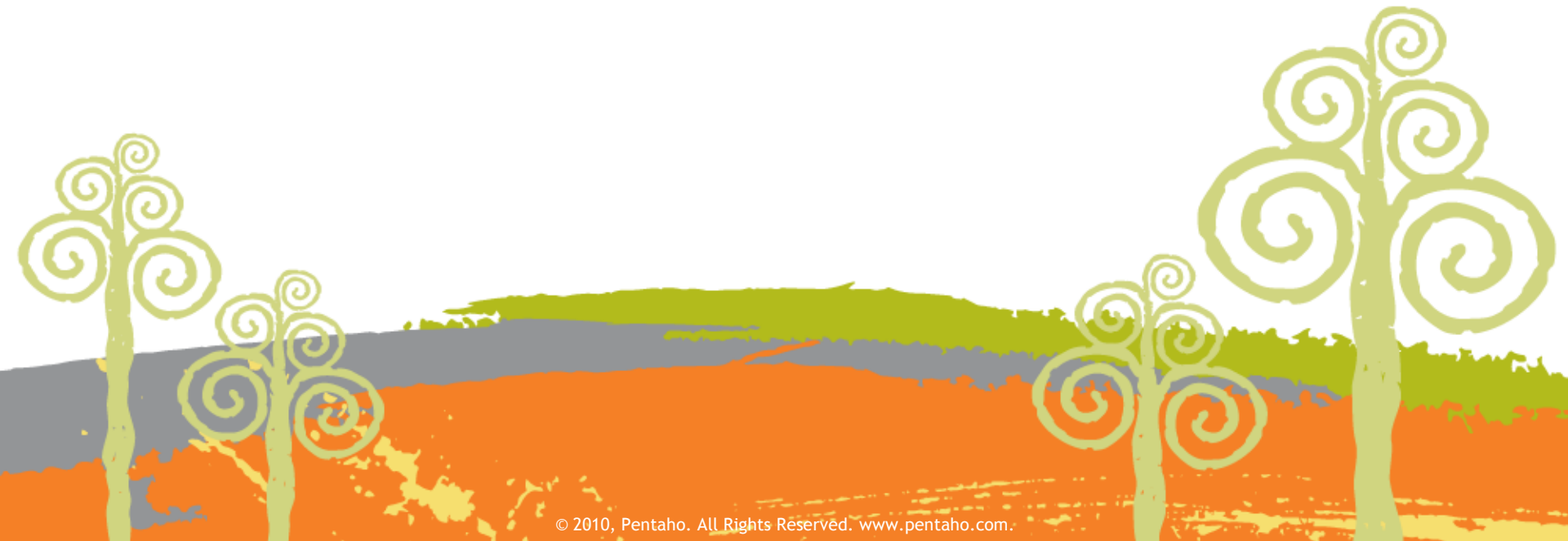
# Matt Casters

- Chief of Data Integration at Pentaho
  - Lead Development
  - Project manager
  - Community contact
- Kettle Project Founder
- Author of Pentaho Kettle Solutions
  - Published by Wiley
  - 650 pages

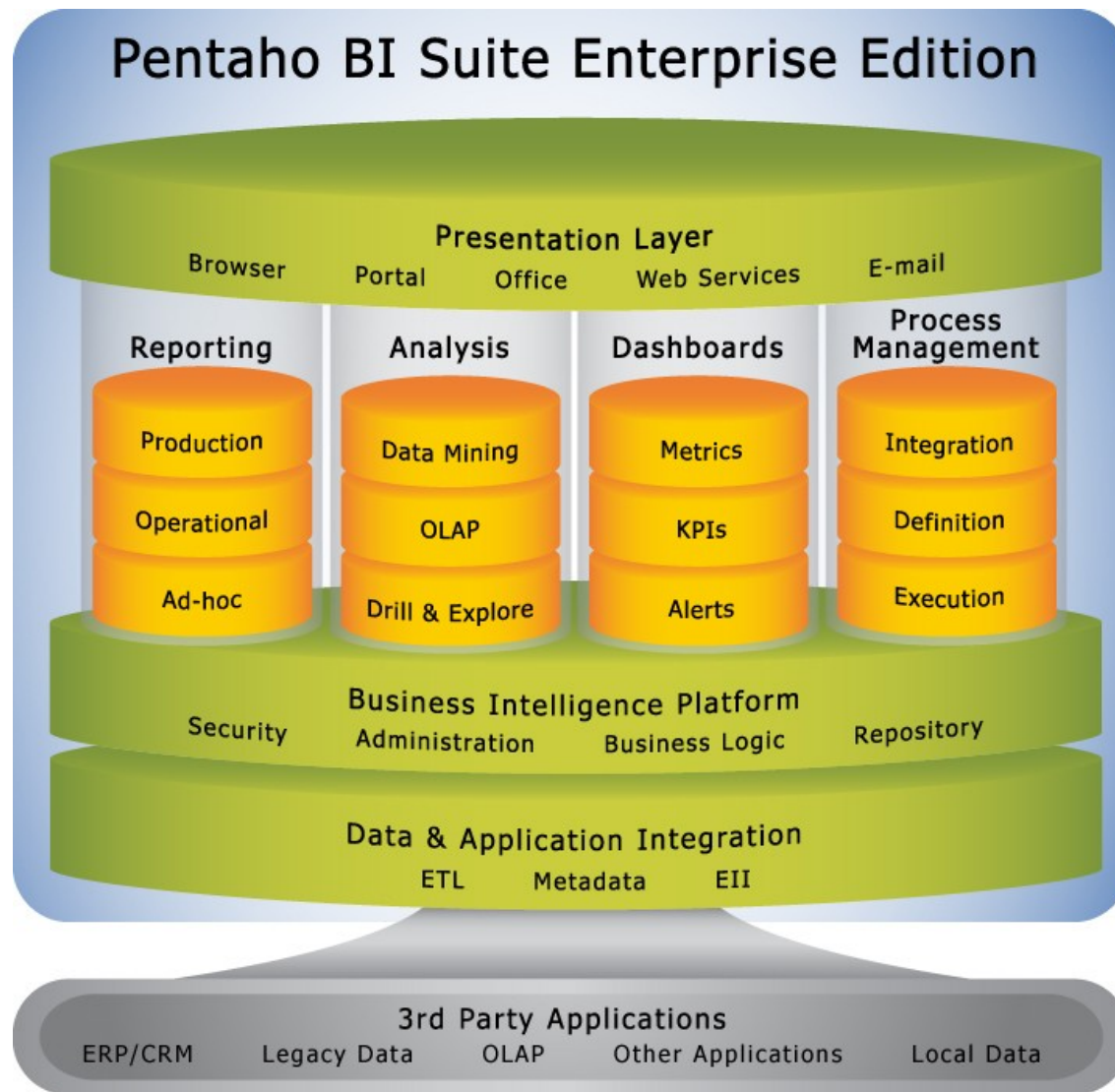





About Pentaho...

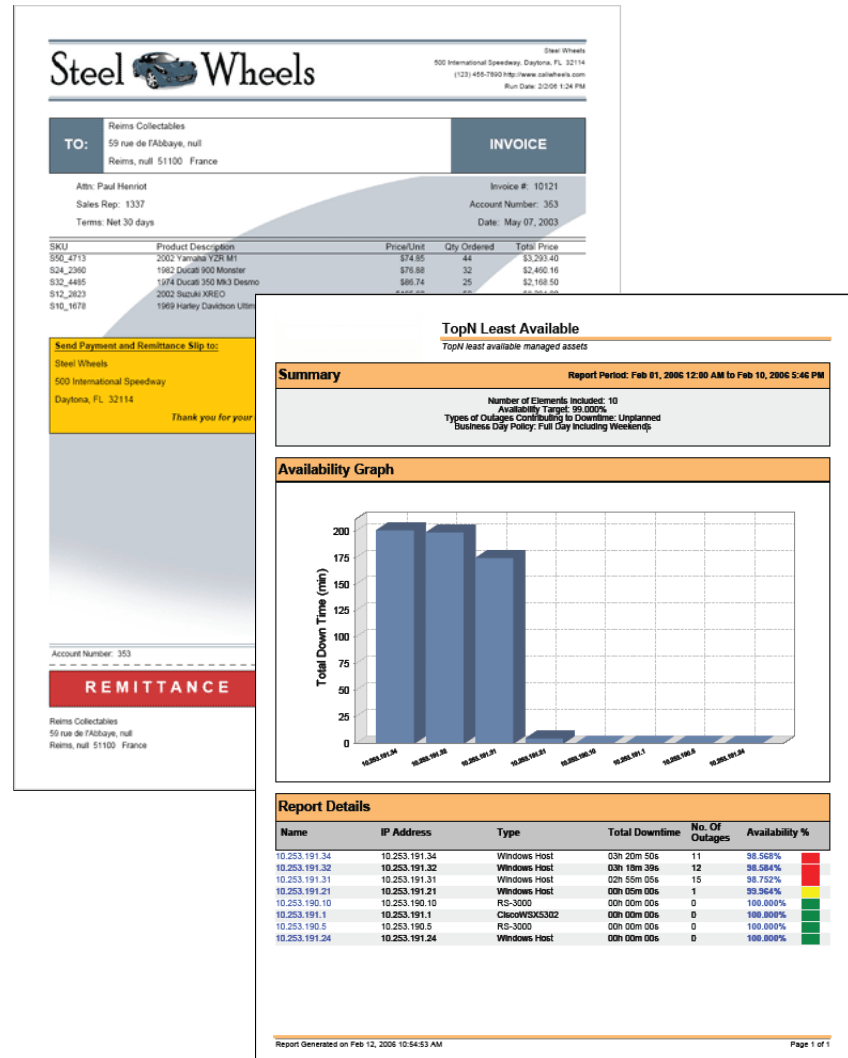


# Pentaho Software Stack Overview



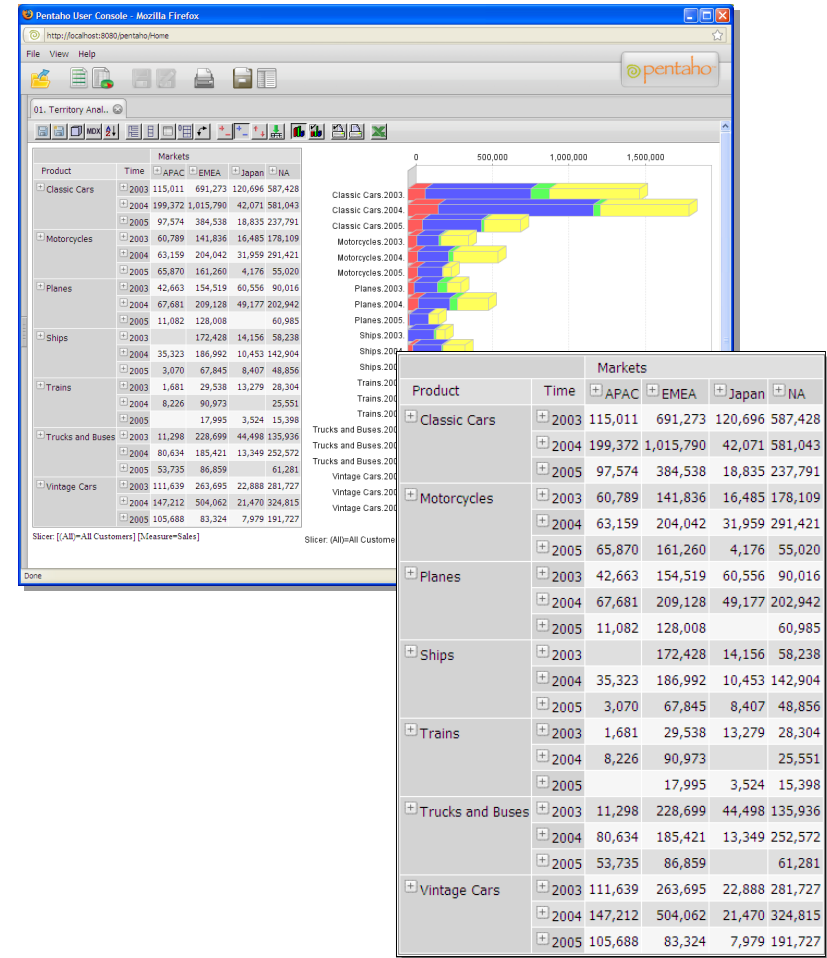
# Pentaho Reporting

- Access and format data from disparate sources
  - RDBMS, XML, OLAP
- Produce in popular formats
 
- Multiple report types
  - Operational
  - Analytical
  - Financial
  - Parametrized
- Go directly against data sources or Pentaho's centralized metadata layer



# Pentaho Analysis

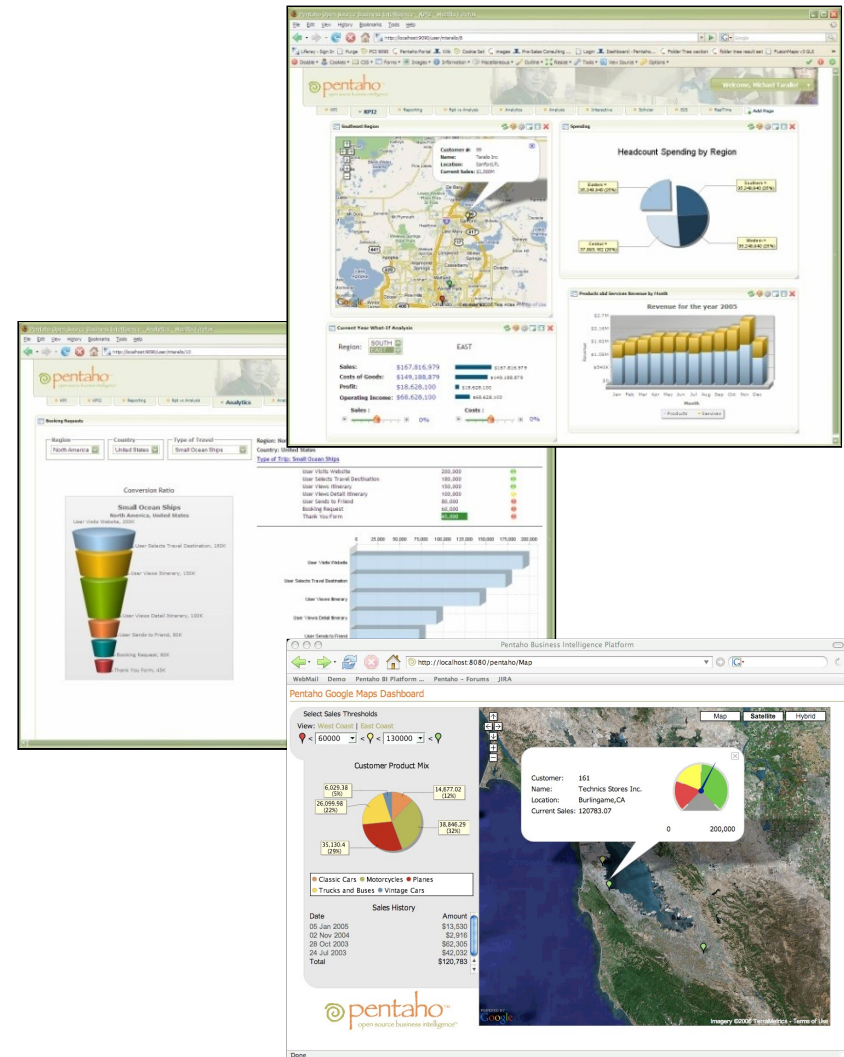
- Navigate and explore
  - Ad hoc, interactive analysis
  - Drill into further detail
  - Select specific members for analysis
- View data “dimensionally”
  - i.e. Sales by region, by channel, by time period
- ROLAP architecture
  - Works with all popular open source and proprietary DBs
  - No intermediate storage
  - Aggregate table “aware” for faster analytic queries
- Design tools to build OLAP schemas and improve query performance





# Pentaho Dashboards

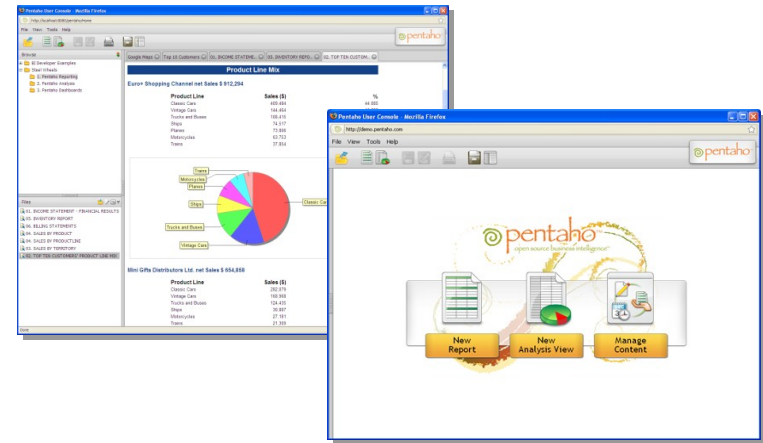
- Gain visibility into your organization's key performance indicators (KPIs)
  - Monitor top-level performance and drill into supporting detail
  - Illuminate metrics for quick insight into business activities
  - Track exceptions and receive alerts
- Leverage the full Pentaho BI Suite
  - Comprehensive auditing of user activity, performance and data access
  - Context-sensitive drilling to reports and analysis views
  - Integrated security, scheduling, alerting, portal integration
- Integrate with 3<sup>rd</sup>-party and custom applications



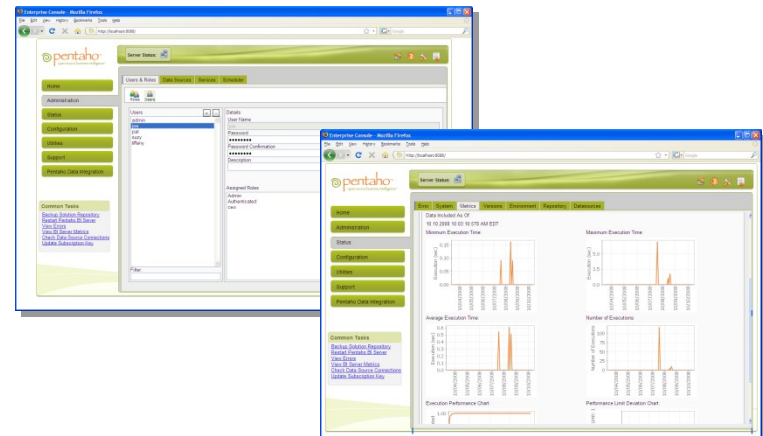
# Pentaho BI Platform

- Provides critical services for end users
  - Easy access to business information
  - Intuitive scheduling
  - Delivery over the web or via email
  - Alerting and notification
- Provides critical services for administrators
  - Centralized thin-client administration
    - Data source and security management
    - Auditing and Performance monitoring
  - Enterprise security integration
  - Definition and execution of business rules
  - Integration points with 3<sup>rd</sup> party applications

## Pentaho User Console



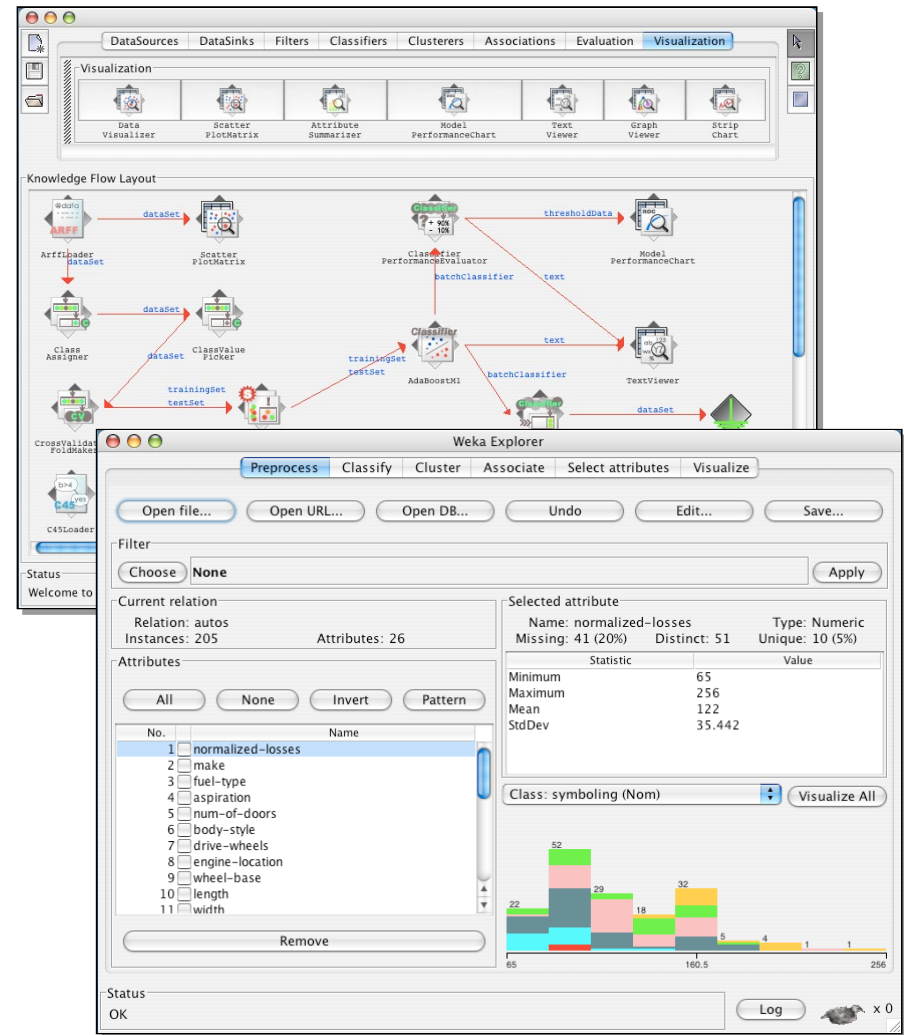
## Pentaho Enterprise Console





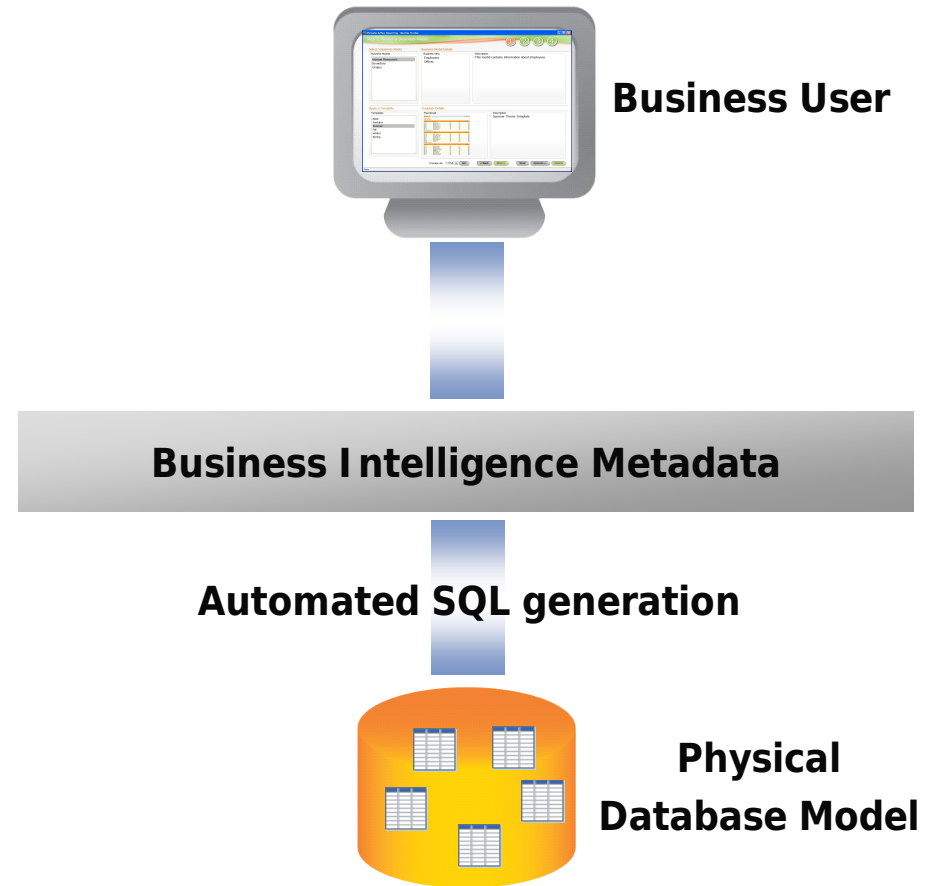
# Pentaho Data Mining

- Take BI to the next level with predictive analytics
- Gain insight into hidden patterns and relationships
- Discover indicators of future performance
- Exploit correlations to improve organizational performance
- Embed recommendations in reports, dashboards, or custom applications



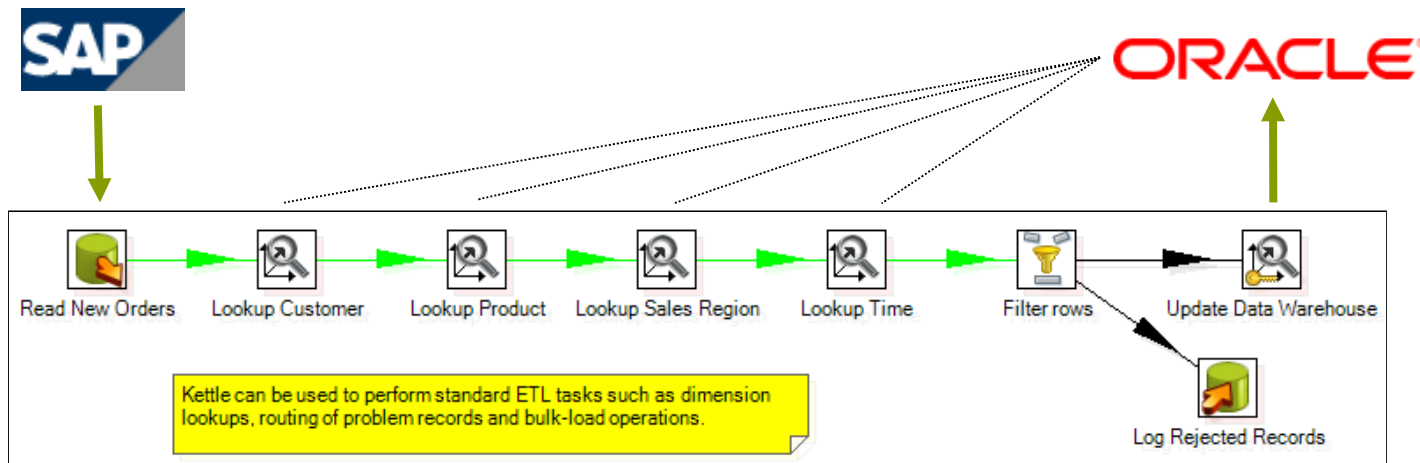
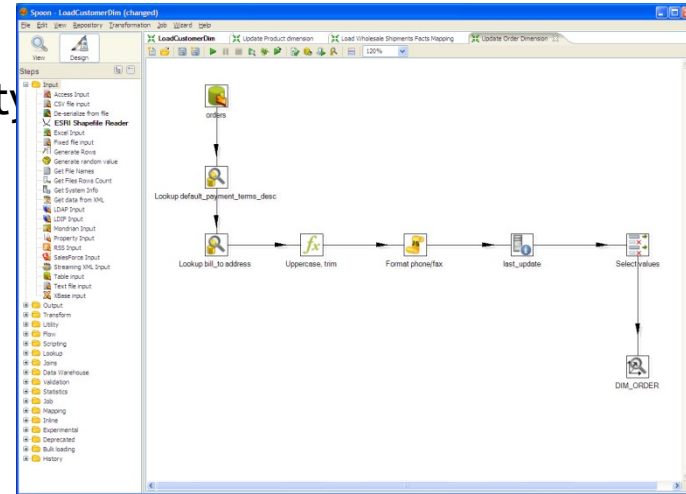
# Pentaho Metadata

- Provides an abstraction layer between source systems and business user concepts
- Graphical design environment for defining metadata model
- Data presented to business users in business terms
- Allows business users to create their own ad hoc reports based on centralized business rules, without any technical skills or knowledge of SQL
- Changes to physical database do not impact reports or analytic views



# Pentaho Data Integration

- Rich Feature Set
- Enterprise-class performance and scalability
- Broad Database Support
- 100% Meta-data Driven
- Graphical, model-driven design
- Mature, 4th generation product





## STORING DATA

The right tool for the right job...

In what is the worlds data stored?



# Relational Data Sources

- Relational database examples:

- *Apache Derby, AS/400, Borland Interbase, Calpont InfiniDB, dBase III-IV-5, ExtenDB, Firebird SQL, Generic database, Greenplum, Gupta SQL Base, H2, Hypersonic, IBM DB2, Infobright, Informix, Ingres, Ingres VectorWise, Intersystems Cache, KingbaseES, LucidDB, MaxDB (SAP DB), MonetDB, MS Access, MS SQL Server, MySQL, Native Mondrian, Neoview, Netezza, Oracle, Oracle RDB, PostgreSQL, Remedy Action Request System, SAP ERP System, SQLite, Sybase, SybaseIQ, Teradata, UniVerse database, Vertica*

- Split up into different categories:

- ISAM like xBase variants, SQLite
- Relational like MySQL, PostgreSQL, Oracle, SQL Server
- Columnar like InfiniDB, InfoBright, LucidDB, VectorWise, Sybase IQ

- All different in support for the SQL standard

- All aiming for various use-cases

- All have specific advantages and disadvantages.

# Text File Data Sources

- CSV
- Fixed Width
- Varied width (Cobol redefines)
- Complicating matters:
  - (binary) delimiters, line separators, enclosures
  - Single, double, triple & quadruple byte encoding
  - Various compression formats (.z, .gz, .zip, .rar, ...)



# Other File like Data Sources

- XML files : any kind of content
- INI files : key-value pairs
- Properties files : key-value pairs
- xBase files : rows of data
- JSON
- SOAP/WSDL
- MS Access : binary ISAM
- Spreadsheets: xls,xlsx, ods
- LDAP / LDIF
- SAS
- YAML

# Non-relational Data Sources

- SAP R/3 application server
- LDAP
- Message Queues (JMS, MQSeries, ...)
- HL7
- EDIFACT
- Windows Messaging
- SNMP
- ...

# Cloud Data Sources

- Salesforce, SugarCRM, ...
- Amazon S3
- Google Analytics
- Twitter
- ...

→ All different in format without a standard

→ Various Web Services and authentication methods

# NoSQL Data Sources (a few popular ones...)

- Large and diverse group of different types of data sources
- Document store
  - Apache CouchDB, Jackrabbit, MongoDB, SimpleDB, ...
- Graph databases
  - Neo4J, HyperGraphDB
- Key Value stores (cont.)



**Source:** <http://en.wikipedia.org/wiki/NoSQL>

# NoSQL Data Sources (more confusion)

- Key-Value stores:

- **Eventually consistent:** Apache Cassandra, Dynamo, Project Voldemort
- **Hierarchical:** GT.M
- **Hosted:** Freebase
- **RAM Cached:** memcached, Velocity, Citrusleaf
- **Disk stored:** BigTable, MemcacheDB, Citrusleaf, MongoDB
- **Ordered:** Berkeley DB, Informix C-ISAM, MemcacheDB, NDBM
- **Multivalued:** Intersystems Caché, ESE/NT, OpenQM
- **Object databases:** Caché, JADE, ObjectDB, ObjectStore, db4o
- **Tabular:** BigTable, Apache Hadoop, Apache Hbase, Hypertable, Mnesia
- **Tuple store:** Apache river

**Source:** <http://en.wikipedia.org/wiki/NoSQL>

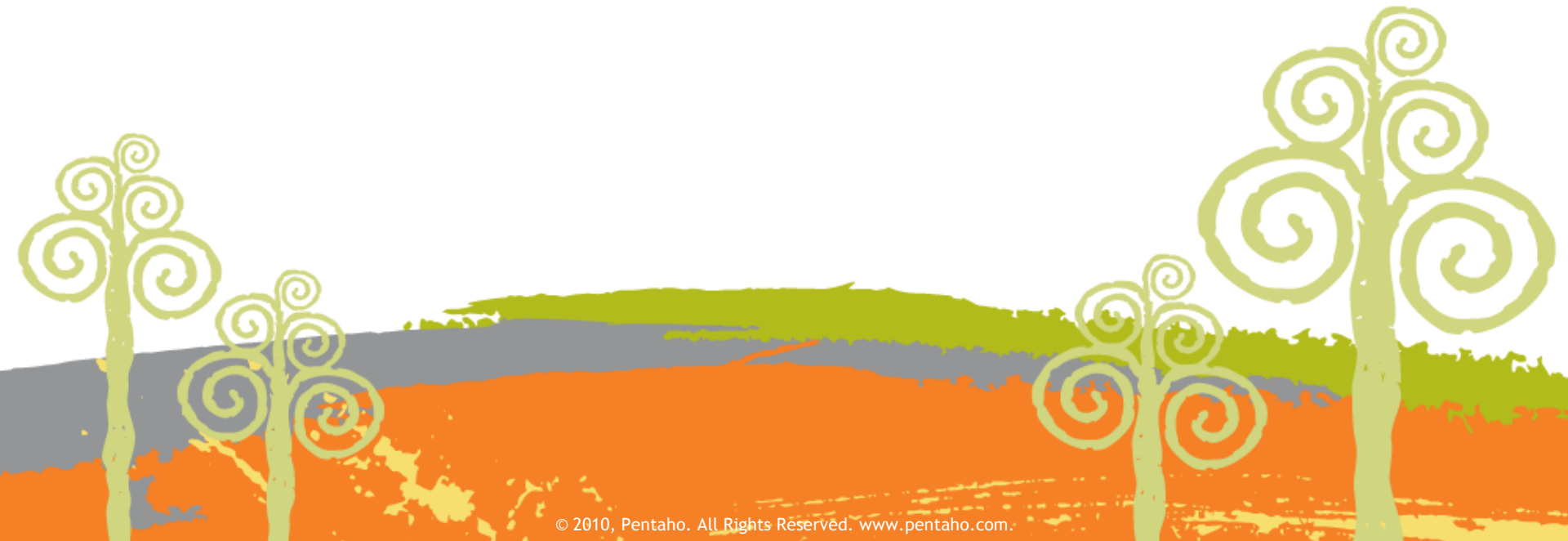
# Other...

- Home brew data formats...
  - Legacy stuff
  - Because there just aren't enough options out there!
- Non-structured data formats
  - PDF
  - Images
  - Video
  - Web content
  - ...





But that's not all...



# Adding complexity by mixing data sources

- Reading from A, writing to B
- Performing lookups
- Joining data
- Calculating
- Scripting
- ...

EXPLODING COMPLEXITY: X systems x Y architectures x Z operations

- Trashing it all when new technology surfaces!

# Classical Solutions...

- Standardize on software stacks
- Limit the number of talking systems

Good advice but...

Perceived as limiting, stifling and preventing innovation.

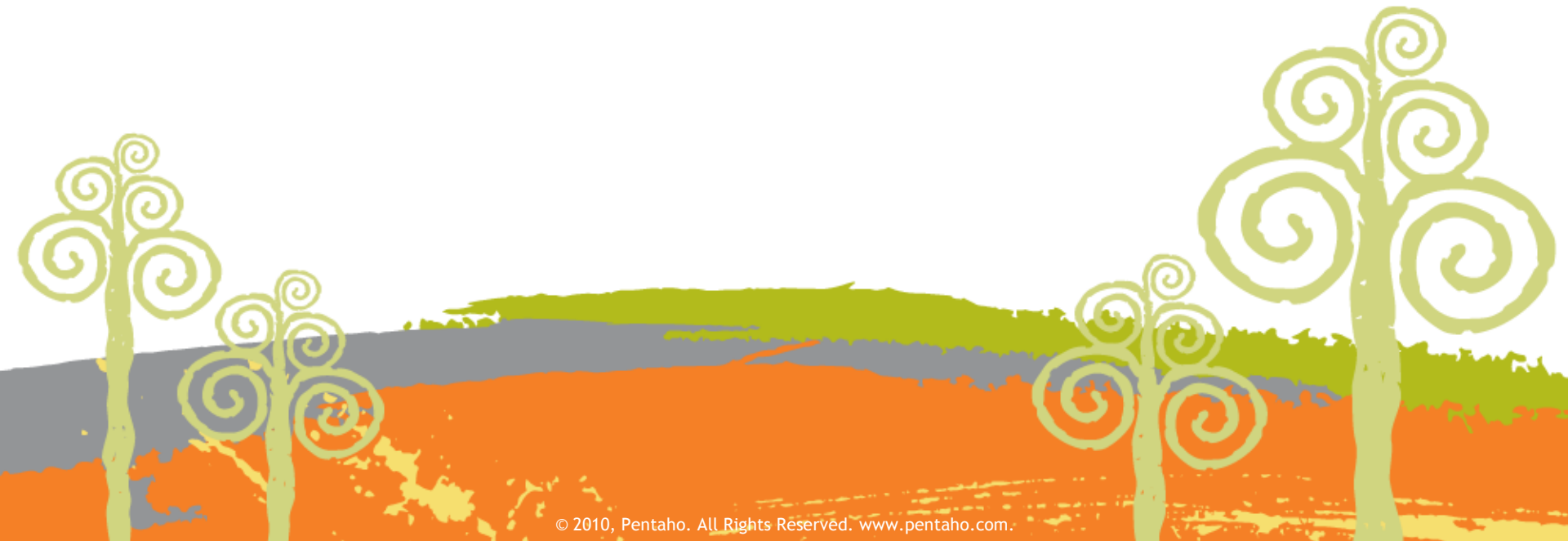
# What does it all mean to me

- Sooner or later you're going to be using any of these data sources
- ... to read from or to write to
- ... with more other data sources coming after it
- Get ready to learn more and more APIs
- Rate of change seems to be increasing





Hand coding you say?  
Perhaps we can think of something easier?

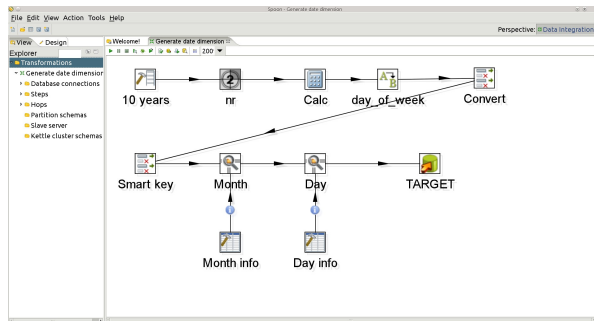


# Engines executing metadata

Task design

Save work

Execution



**Java API**

**XML generation**

**<XLM/>**

**Repositories**

Graphical

Batch

Local

Remote

Java API

M/R

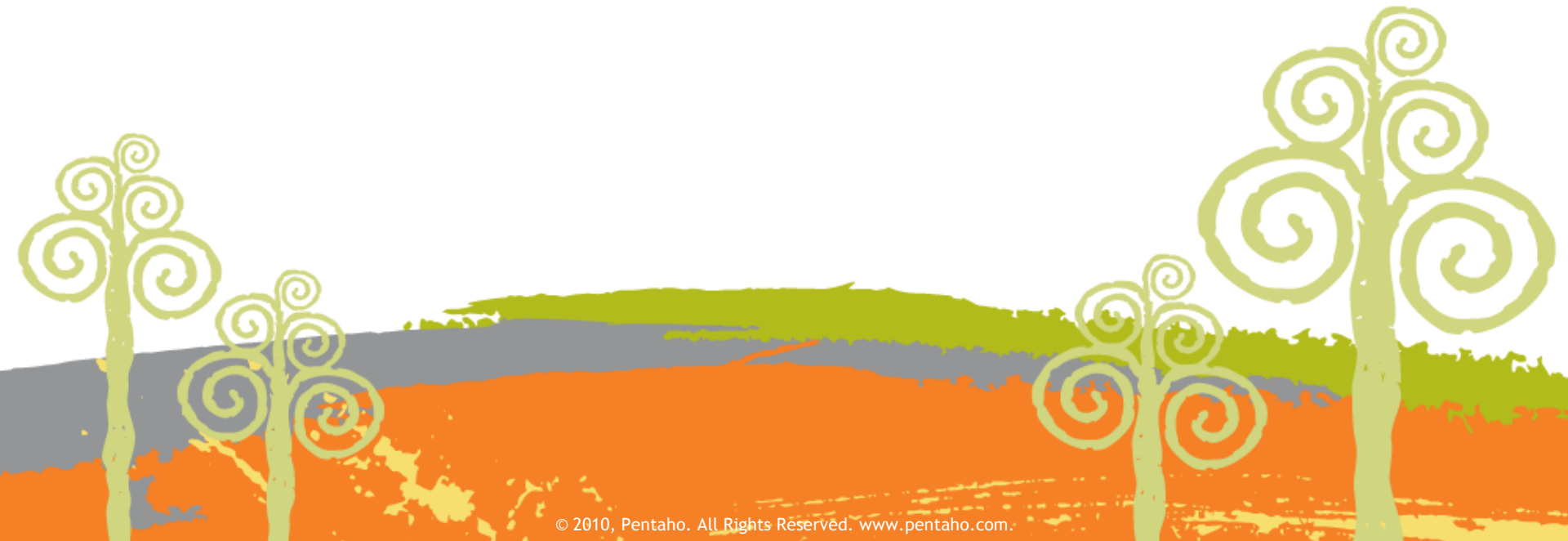
Reporting

Embedded



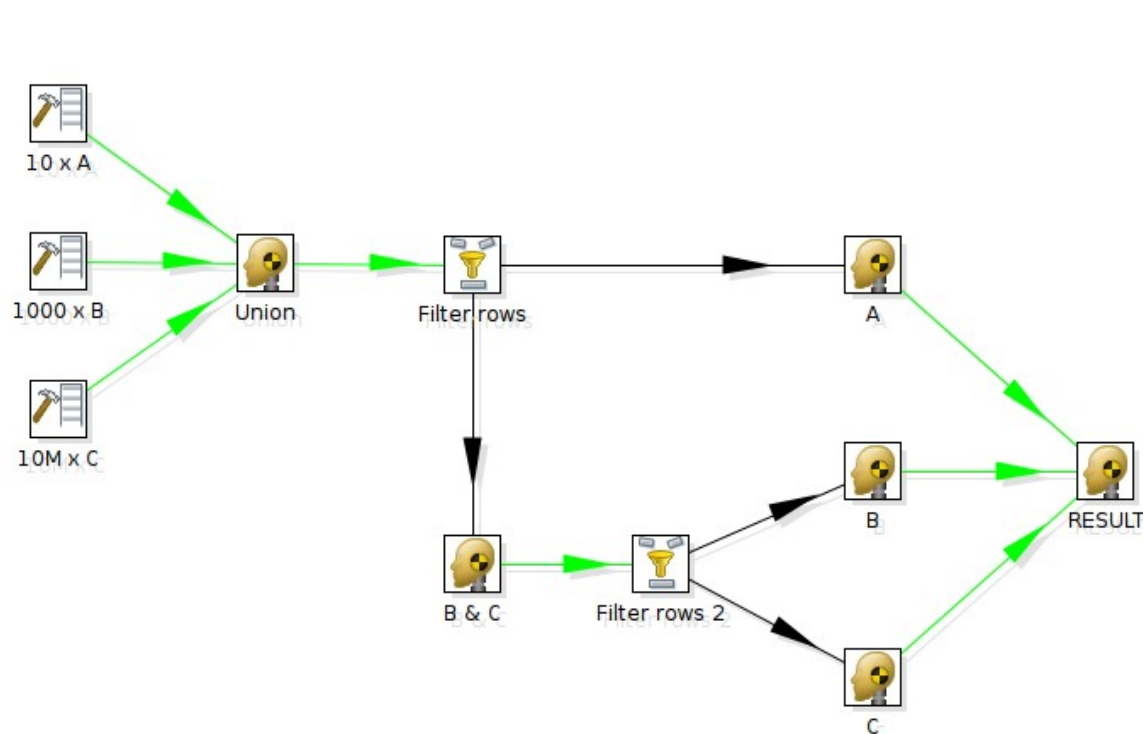


Enter Kettle ...euh.. Pentaho Data Integration



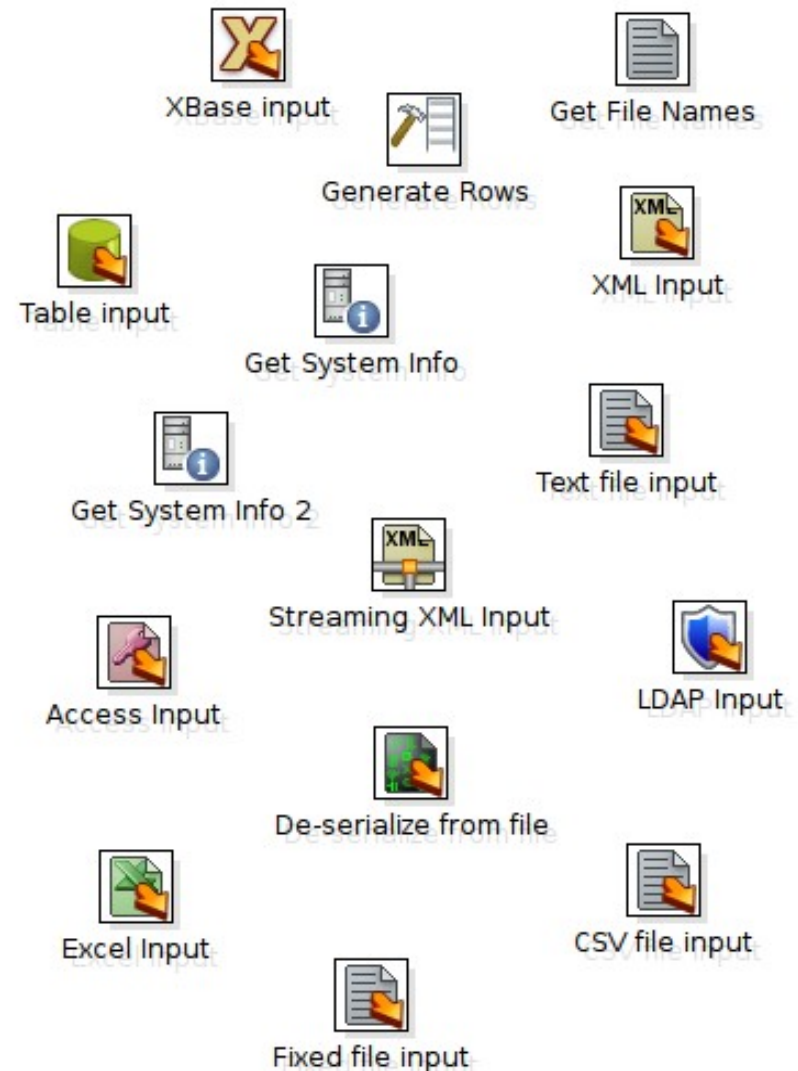
# Pentaho Data Integration = Kettle

**K**ettle  
**E**xtraction  
**T**ransportation  
**T**ransformation  
**L**oading  
**E**nvironment



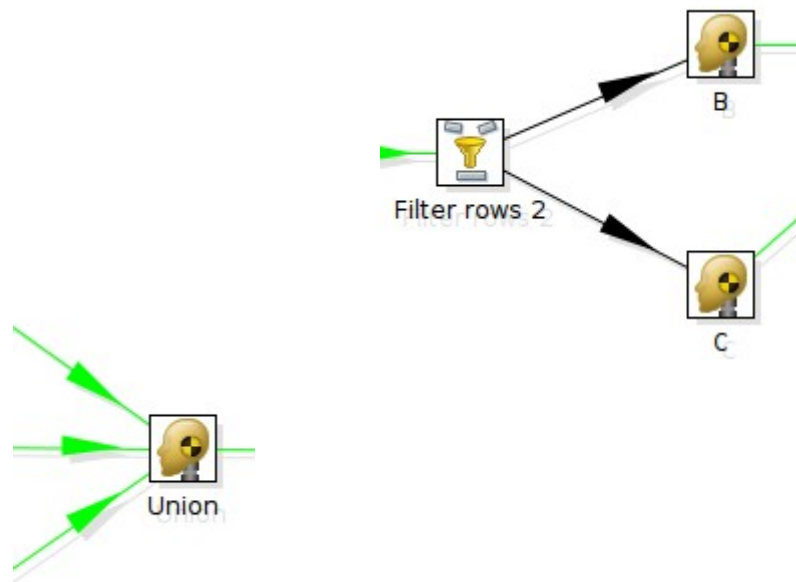
# Pentaho Data Integration - Extraction

- Extract data from :
  - 35+ database types
    - MySQL, PostgreSQL, SQLite, ...
    - Oracle, SQL Server, etc
  - Text files
  - XML files
  - XLS files
  - NoSQL data sources
  - Xbase files (dBase, Foxpro, etc)
  - File systems information
  - Generated data
  - MS Access files
  - LDAP
  - Geo-data
  - ...



# Pentaho Data Integration - Transportation

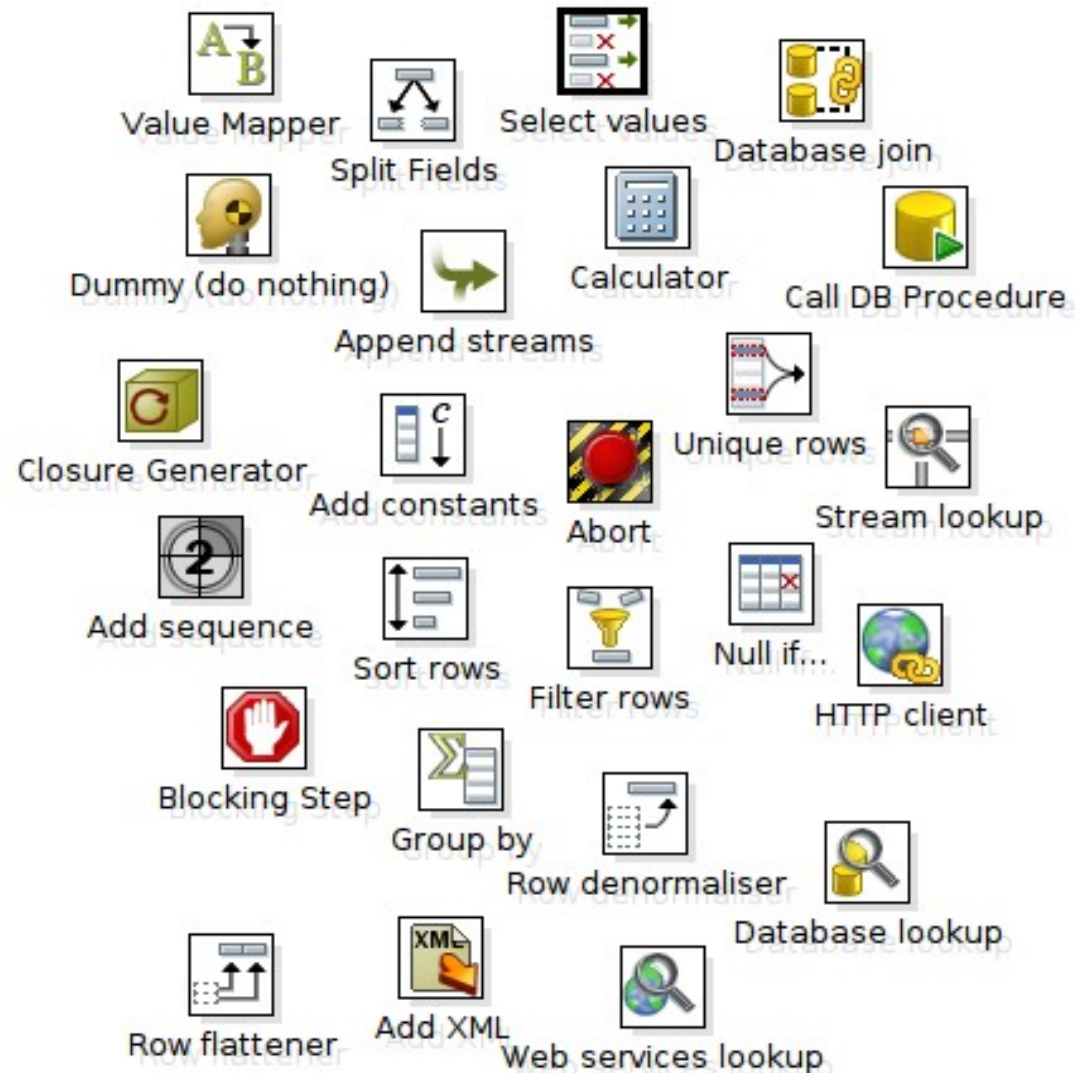
- Transportation of data
  - Engine based data transfer (no code generator)
  - Very flexible pathways:
    - splitting
    - partitioning
    - merging
    - joining
    - duplicating
    - clustering (MPP)
    - Map/Reduce
  - Streaming!



# Pentaho Data Integration - Transformation

- Flexibly transform data

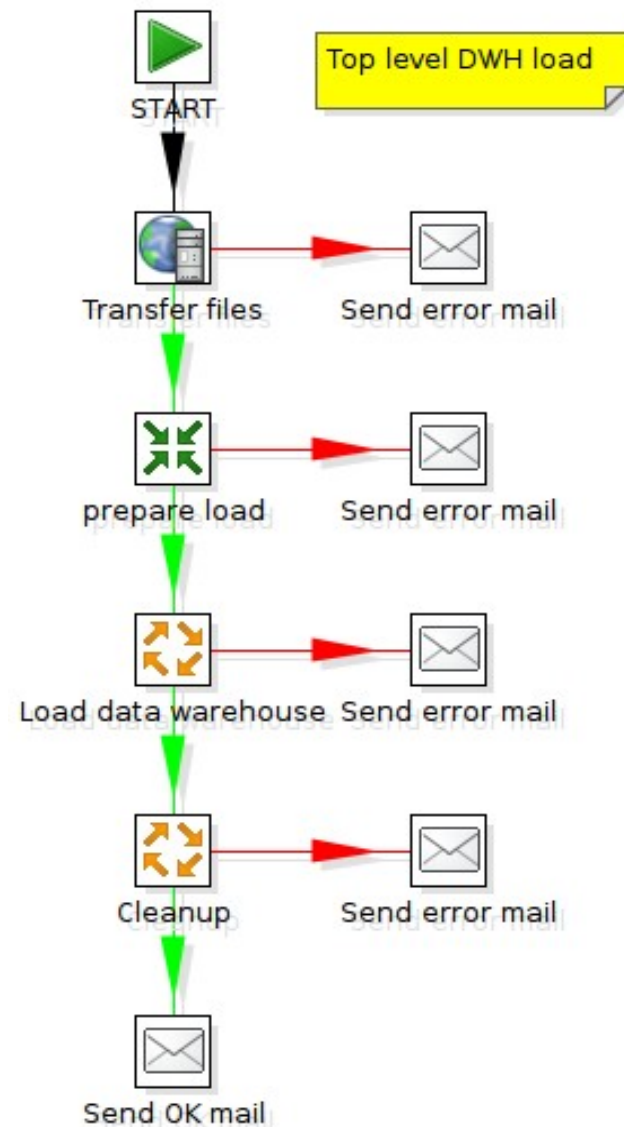
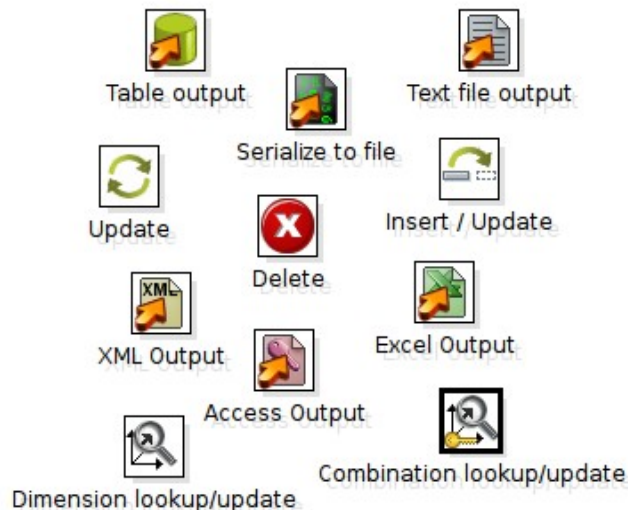
- Looking up data
  - databases
  - files
  - memory...
- Calculating
- Scripting
  - JavaScript, SQL, RegExp
- Splitting
- Mapping
- Selecting
- Filtering
- Pivotting ...



# Pentaho Data Integration - Loading

- Load data into a target format

- Database loads
- Data warehouse population
- Partitioned loading
- Bulk loading
- Parallel loading
- Clustering





# Pentaho Data Integration - Environment

- Full GUI called “Spoon” to edit every option in Kettle
  - Drag & Drop
  - Debugger
  - Rich GUI

- Command line tools

- execute jobs
- execute transformations

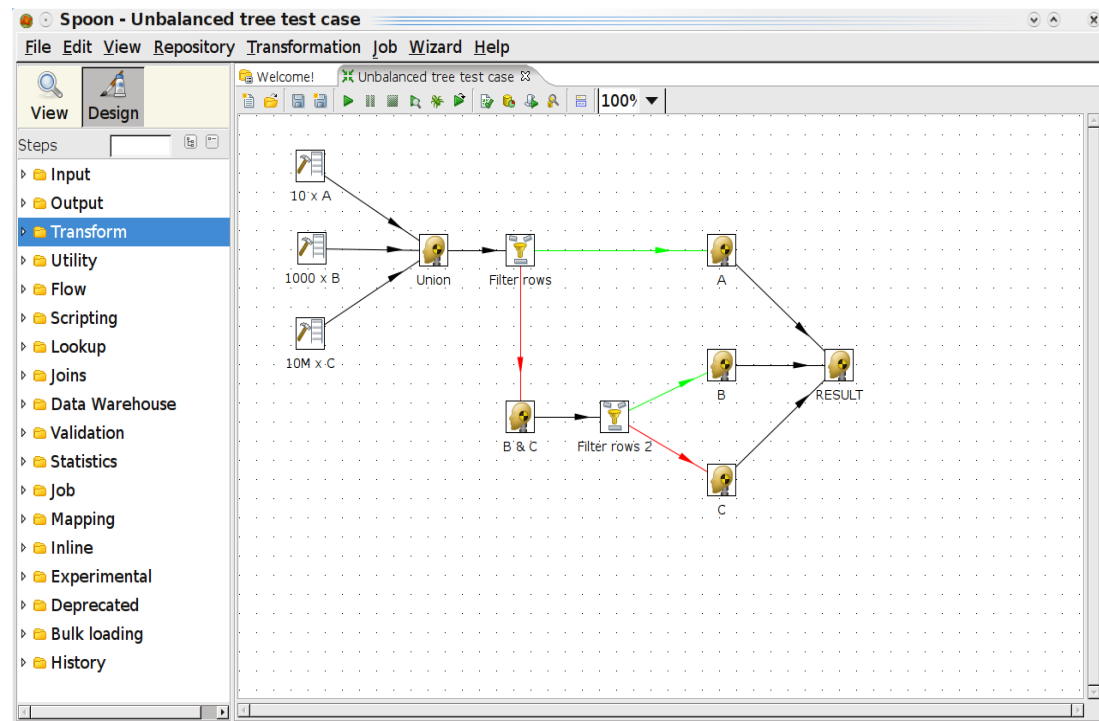
- Web server

- clustering
- remote execution

- Programming API for Java

- Plugin eco-system

- ...



# Pentaho Data Integration - Community

ohloh popular!

<http://www.ohloh.net/projects/3624?p=Kettle>



<http://www.softpedia.com/progClean/Kettle-Clean-80094.html>

- BI/DSS professionals
- Consultants
- Paying Pentaho customers
- All types of corporations
- Lone rangers & Hobbyists
- All regions on Earth
- Meet on our Forum : +51,000 posts in 12,000 threads
- Use our JIRA case tracking systems
- Download over 10,000 copies of Kettle per month

<http://community.pentaho.com>

# Pentaho Data Integration - use-cases

- Load data from text files and store it into a database
- Export data from database to text-file or more other databases
- Data migration between database applications
- Exploration of data in existing databases (tables, views, etc.)
- Information improvement using lookups
- Data cleaning
- Application integration
- Data warehouse population
- Application integration
- Report data generation
- Move data around in the cloud

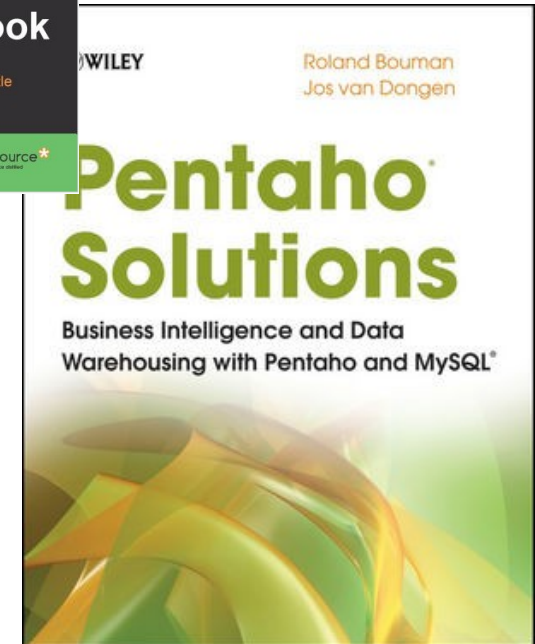
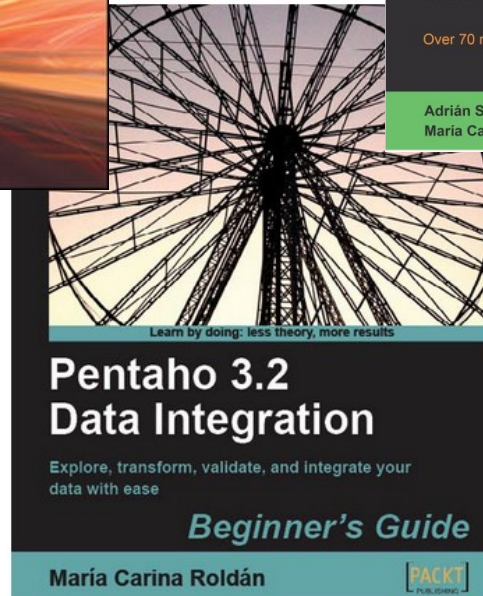
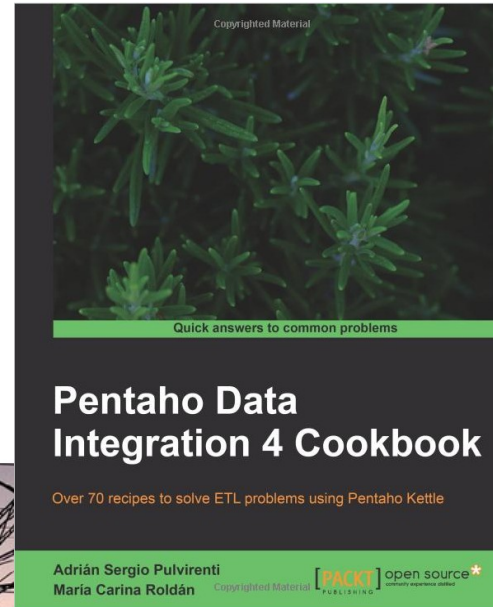
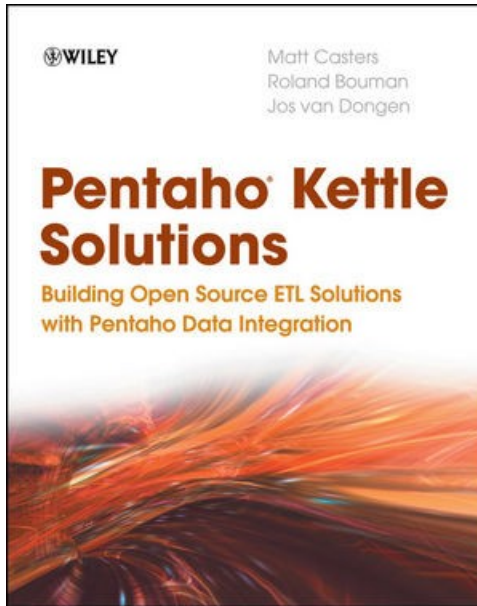


# Examples please!!!

- Load CSV file into a database
- Get some JSON from MongoDB
- Execute these tasks in Java



# Pentaho Books



## To conclude

- Complete stack of BI and Data Integration software
- Software under Open source license(s)
- Free to download and use by each and everyone
- Projects can be integrated or separately used

**Don't re-invent the wheel!**

**Spent time on the fun stuff!**



# Questions and Answers

- Homepage: <http://kettle.pentaho.org>
- Forum: <http://forums.pentaho.org/forumdisplay.php?f=69>
- Case tracker: <http://jira.pentaho.org/browse/PDI>
- Continuous Integration Server: <http://ci.pentaho.com/job/Kettle>
- Wiki : <http://wiki.pentaho.org/display/EAI>
- IRC Channel: ##pentaho (on Freenode)
- Mailing list: <http://groups.google.com/group/kettle-developers>
- My blog: <http://www.ibridge.be>
- My coordinates: mcasters at pentaho dot org