Pentaho Data Integration 4 and MySQL

Matt Casters:
Pentaho's Chief Data Integration
Kettle Project Founder

MySQL User Conference, Tuesday April 13th, 2010

Agenda

- ***Pentaho: an introduction***
- Pentaho Data Integration
- Version 4: New features
- MySQL support in PDI
- Q&A

Pentaho Introduction

- Commercial open source alternative for business intelligence (BI)
  - Founded in 2004: Pioneer in commercial open source BI
  - Large referenceable customer base, wide range of BI/DW deployments
- Management - proven BI and open source veterans
  - Business Objects, Cognos, Hyperion, JBoss, Oracle, Red Hat, SAS, SugarCRM
- Board of Directors – deep expertise and proven success in open source
  - Bob Bearden - Executive chairman of the board (former SpringSource)
  - Larry Augustin - founder, VA Software, helped coin the phrase “open source”
  - Zack Urlocker – VP of Products, MySQL/Oracle
  - Benchmark Capital, Index Ventures, New Enterprise Associates
- Widely recognized as the leader in open source BI

pentaho™
open source business intelligence™

Pentaho Introduction

- Complete Business Intelligence Suite
  - End-to-end coverage of all BI needs
  - Standards-based, modular, standalone or embeddable platform
- Open Source Licensing
  - Lower software acquisition costs
  - Lower Total Cost of Ownership (TCO)
- Enterprise Development Methodology
  - Transparent, detailed roadmap
  - Product roadmap and contributions managed by Pentaho
  - Core developers are Pentaho employees
  - Extensive QA
- Expert Services
  - Comprehensive Training, Consulting, Enterprise service offerings
  - Delivered by the Experts

Pentaho Introduction – Enterprise Edition

pentaho™
open source business intelligence™

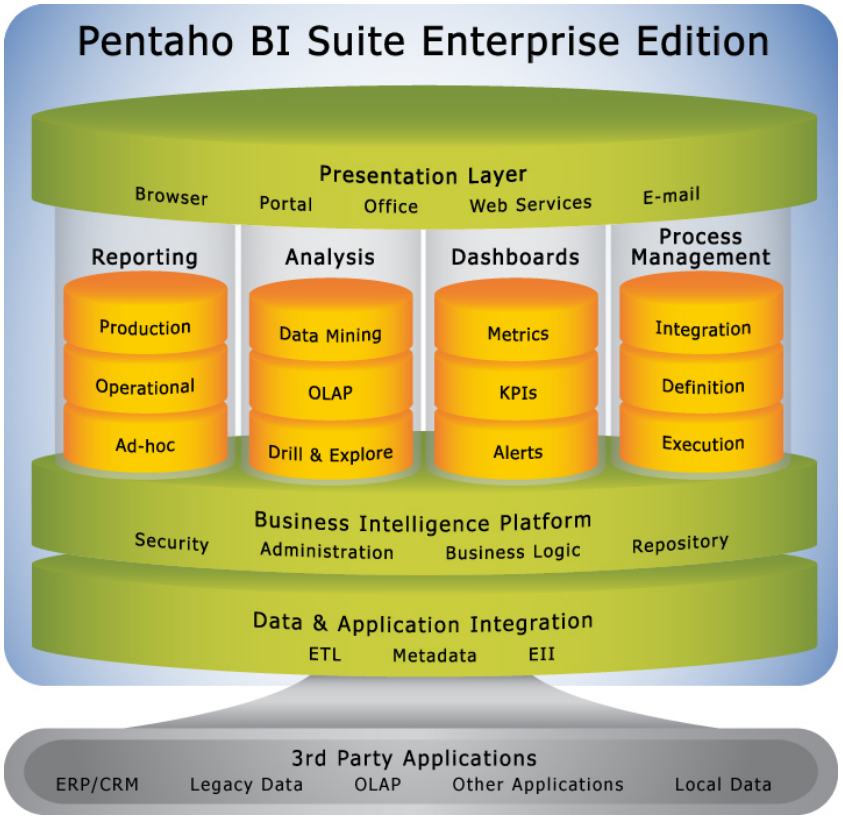
Pentaho Introduction – Deployments

- Wide range of deployments
  - Reporting
  - Data Integration / ETL
  - Dashboards
  - Full BI Suite
- Thousands of users
  - 3,000 on a single server
- Large data volumes
  - Half a terabyte of live interactive OLAP data
  - ETL loading 300K rows/second
- Sophisticated applications
  - Hundreds of dimensions
- Small deployments as well
  - 20 users, MS Access databases

pentaho™
open source business intelligence™

Pentaho Introduction – Technology

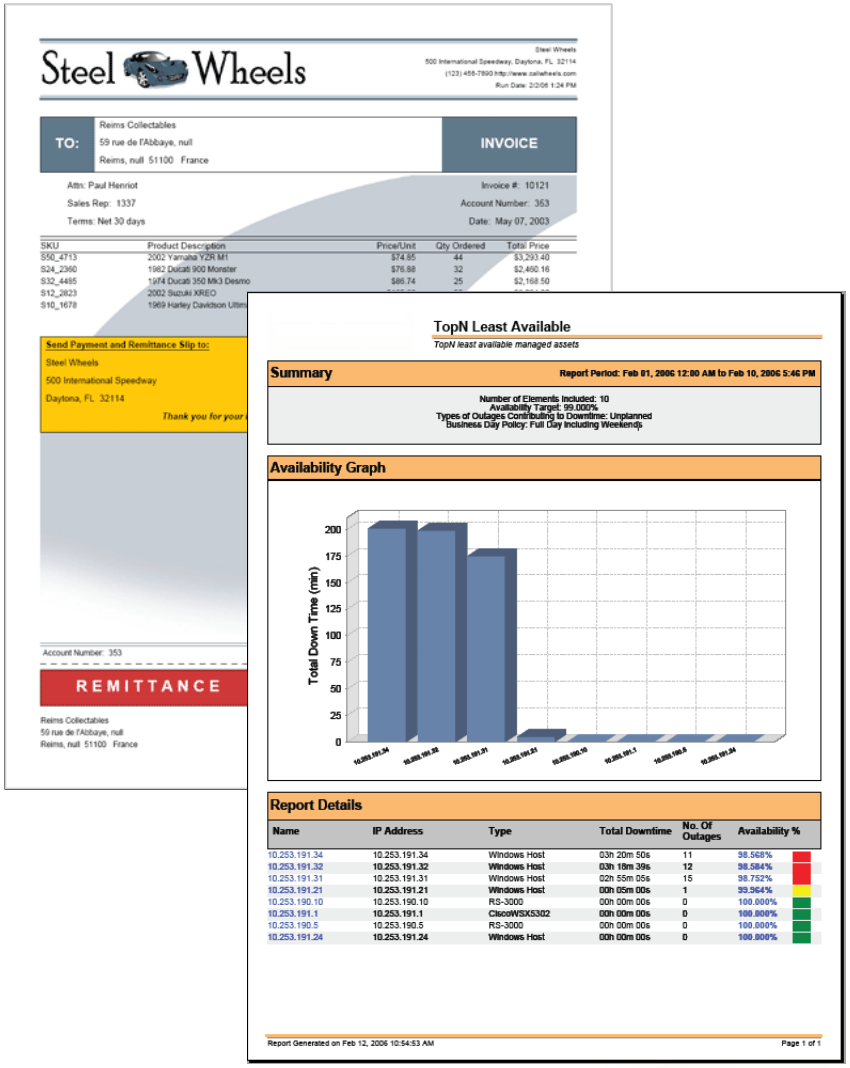
- Componentized and modular
- Service-implemented architecture
  - Built “from the ground up” as a set of services
  - Exposed via AJAX and Web Services
- 100% Java EE server side
  - Scalable, standards-based
- Web-based, thin-client end user interfaces
- Graphical design interfaces
- Embedded process workflow engine



pentaho™
open source business intelligence™

Pentaho Introduction – Reporting

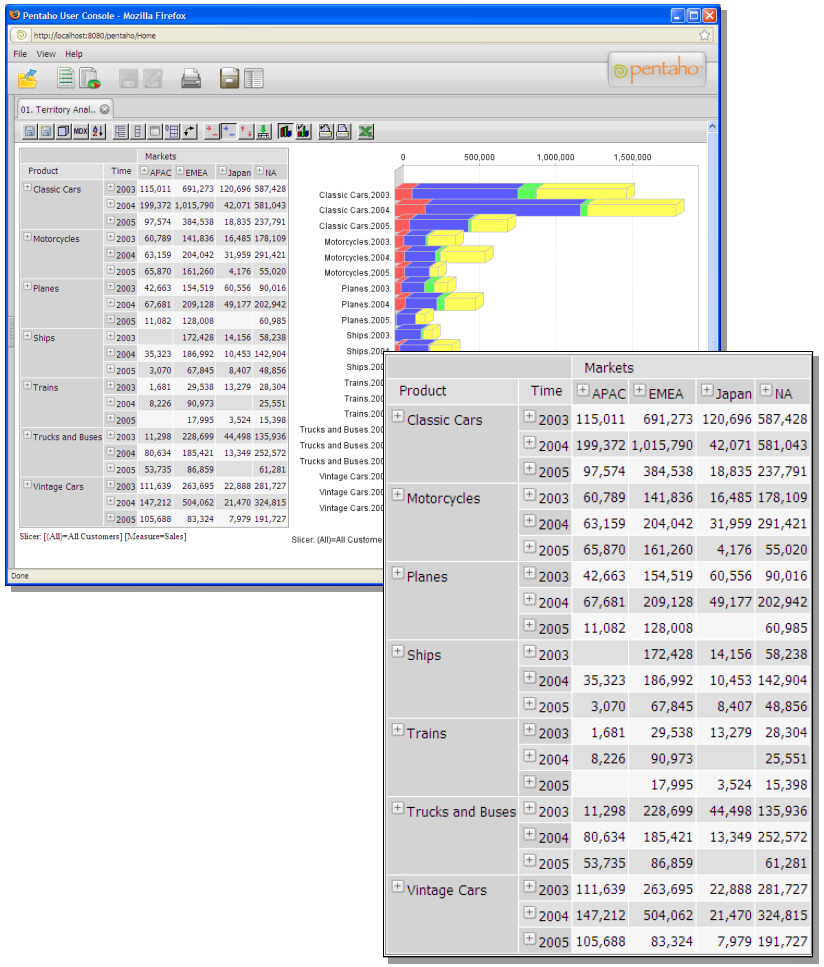
- Access and format data from disparate sources
  - RDBMS, XML, OLAP
- Produce in popular formats
- Multiple report types
  - Operational
  - Analytical
  - Financial
  - Parameterized
- Go directly against data sources or Pentaho's centralized metadata layer



pentaho™
open source business intelligence™

Pentaho Introduction – 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

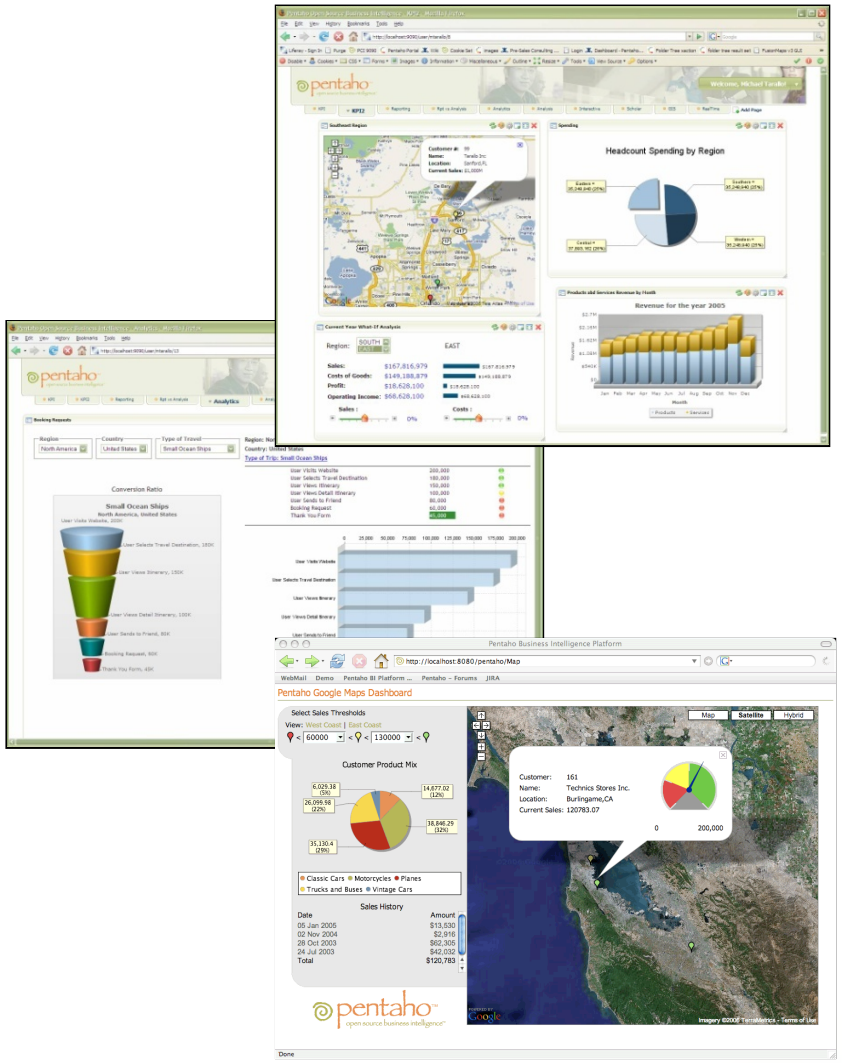


Product	Time	Markets APAC	EMEA	Japan	NA
Classic Cars	2003	115,011	691,273	120,696	587,428
	2004	199,372	1,015,790	42,071	581,043
	2005	97,574	384,538	18,835	237,791
Motorcycles	2003	60,789	141,836	16,485	178,109
	2004	63,159	204,042	31,959	291,421
	2005	65,870	161,260	4,176	55,020
Planes	2003	42,663	154,519	60,556	90,016
	2004	67,681	209,128	49,177	202,942
	2005	11,082	128,008		60,985
Ships	2003		172,428	14,156	58,238
	2004	35,323	186,992	10,453	142,904
	2005	3,070	67,845	8,407	48,856
Trains	2003	1,681	29,538	13,279	28,304
	2004	8,226	90,973		25,551
	2005		17,995	3,524	15,398
Trucks and Buses	2003	11,298	228,699	44,498	135,936
	2004	80,634	185,421	13,349	252,572
	2005	53,735	86,859		61,281
Vintage Cars	2003	111,639	263,695	22,888	281,727
	2004	147,212	504,062	21,470	324,815
	2005	105,688	83,324	7,979	191,727

pentaho™
open source business intelligence™

Pentaho Introduction – Dashboards

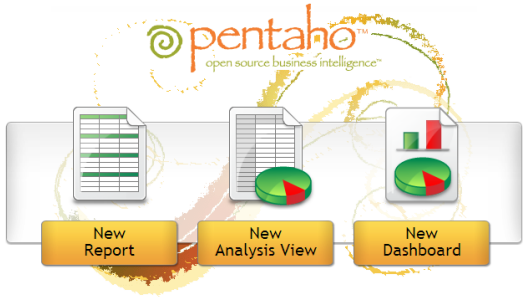
- Gain visibility into your organization's key performance indicators (KPIs)
  - Monitor top-level performance and drill into supporting detail
  - Illuminate metrics for quick insight into business activities
  - Track exceptions and receive alerts
- Leverage the full Pentaho BI Suite
  - Comprehensive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

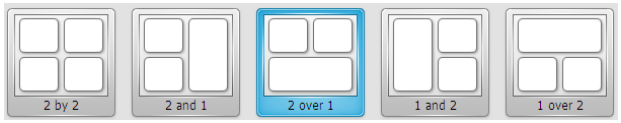


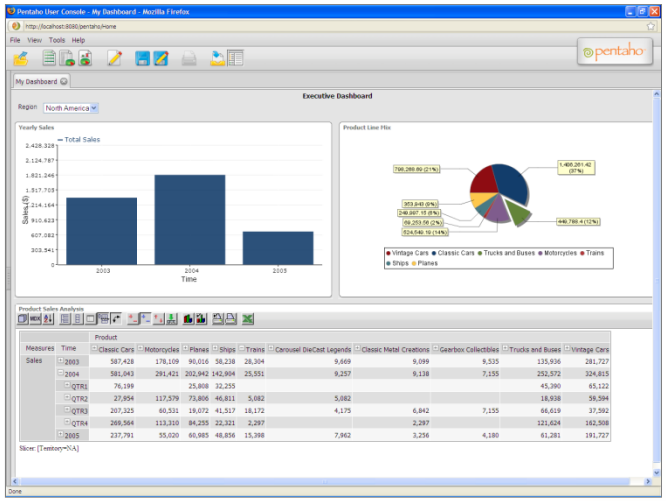
pentaho™
open source business intelligence™

Pentaho Introduction – Dashboard Designer

- Web-based end user dashboard creation
  - From Pentaho User Console
  - “Zero training”
- Template and theme-based creation
- Incorporate reports, analysis views, Adobe Flash-based charts and other Pentaho content
- Create new charts and interactive data grids from scratch
  - Pentaho metadata – no SQL required
- Filter controls





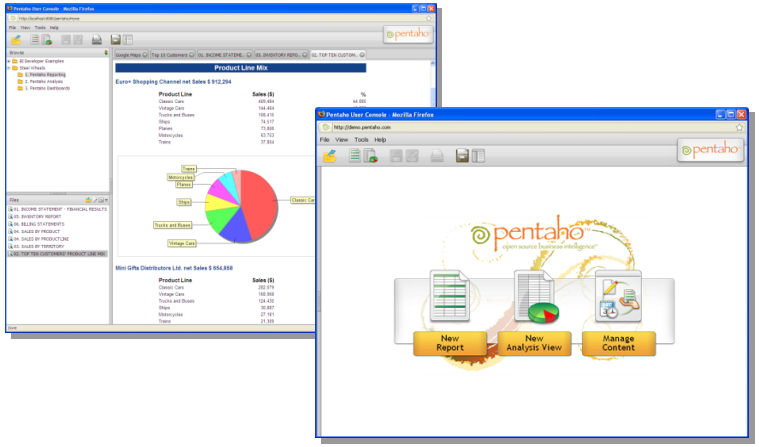


pentaho
open source business intelligence

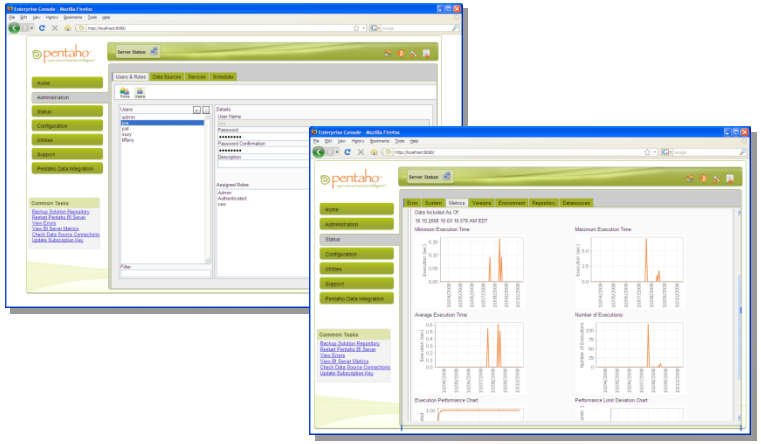
Pentaho Introduction – BI Platform

- 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 3rd party applications

Pentaho User Console



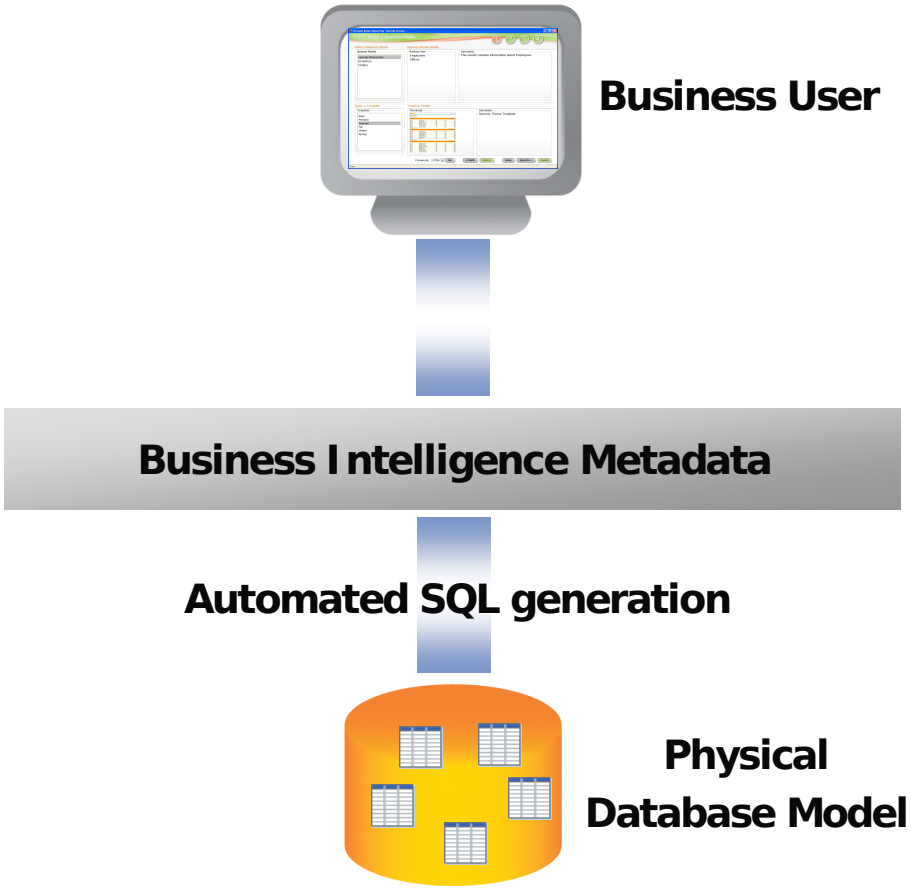
Pentaho Enterprise Console



pentaho™
open source business intelligence™

Pentaho Introduction – 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 Introduction – Data Mining

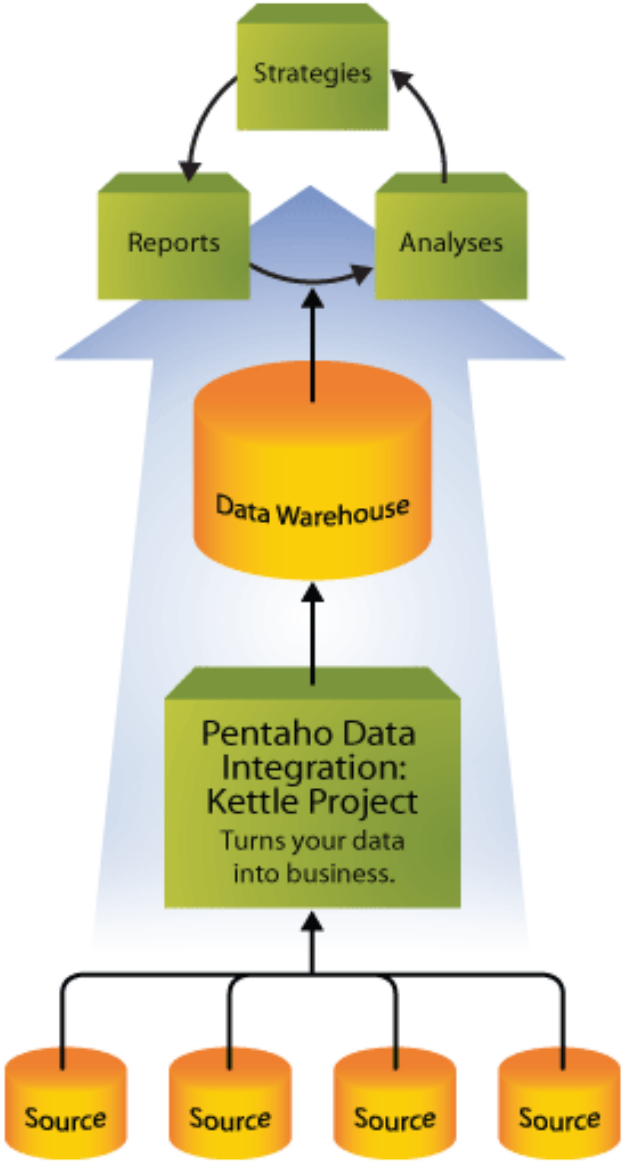
- Take BI to the next level with predictive analytics
- Gain insight into hidden patterns and relationships
- Discover indicators of future performance
- Exploit correlations to improve organizational performance
- Embed recommendations in reports, dashboards, or custom applications

pentaho™
open source business intelligence™

Agenda

- Pentaho: an introduction
- ***Pentaho Data Integration***
- Version 4: New features
- MySQL support in PDI
- Q&A

Pentaho Data Integration for BI



Business Intelligence!

That's what we do.

pentaho™
open source business intelligence™

Pentaho Data Integration – Kettle

Kettle

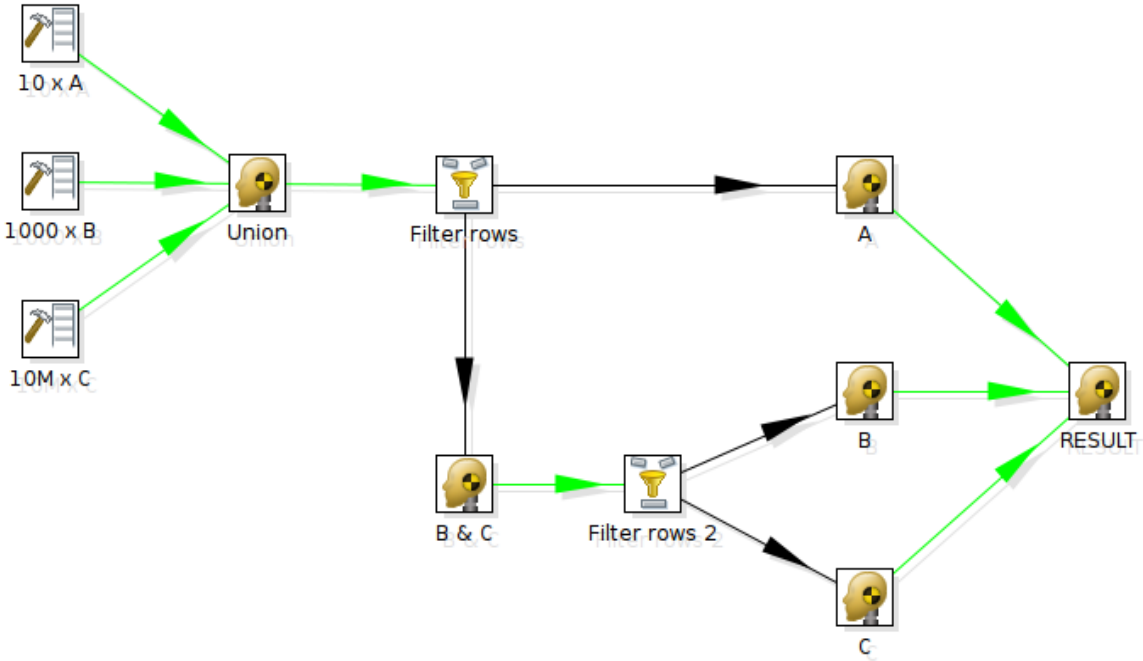
Extraction

Transportation

Transformation

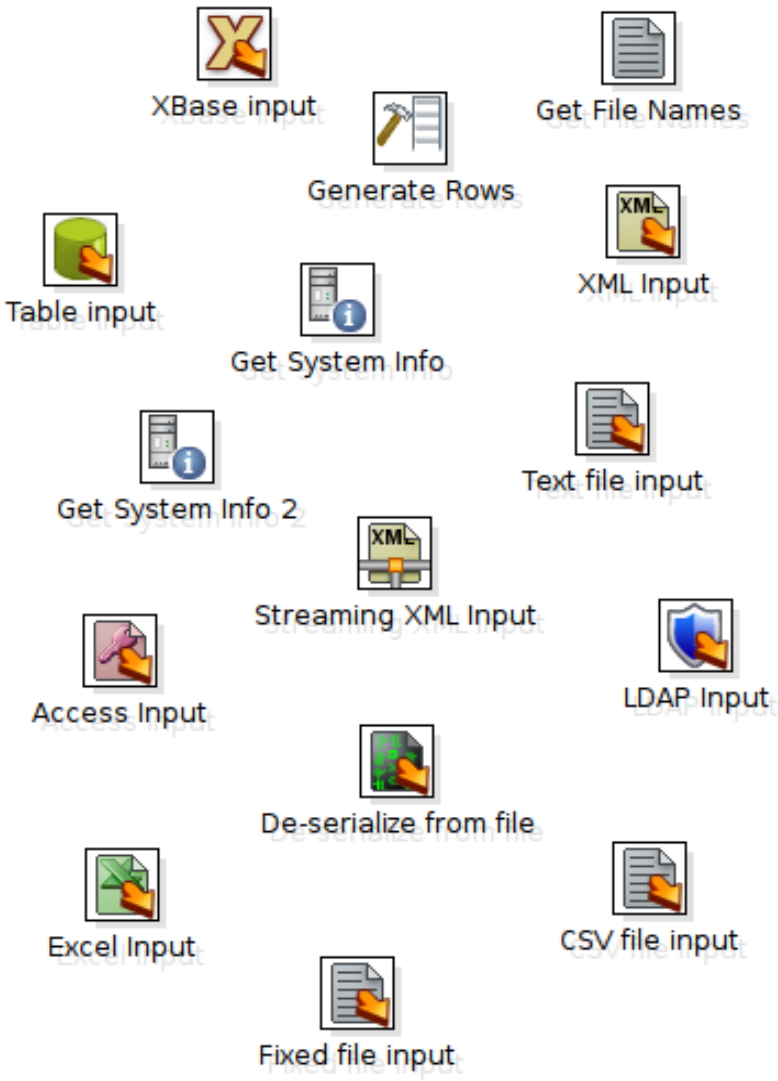
Loading

Environment



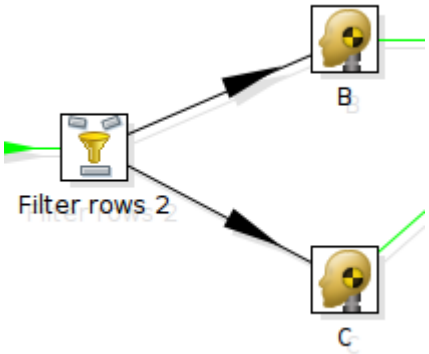
Pentaho Data Integration – Extraction

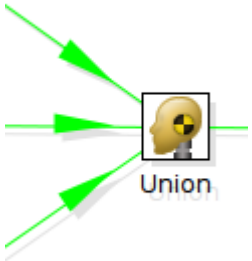
- Extract data from :
  - 35+ database types
    - MySQL, PostgreSQL, SQLite, ...
    - Oracle, SQL Server, etc
  - Text files
  - XML files
  - XLS files
  - Xbase files (dBase, Foxpro, etc)
  - File systems information
  - Generated data
  - MS Access files
  - LDAP
  - Geo-data
  - ...



Pentaho Data Integration – Transportation

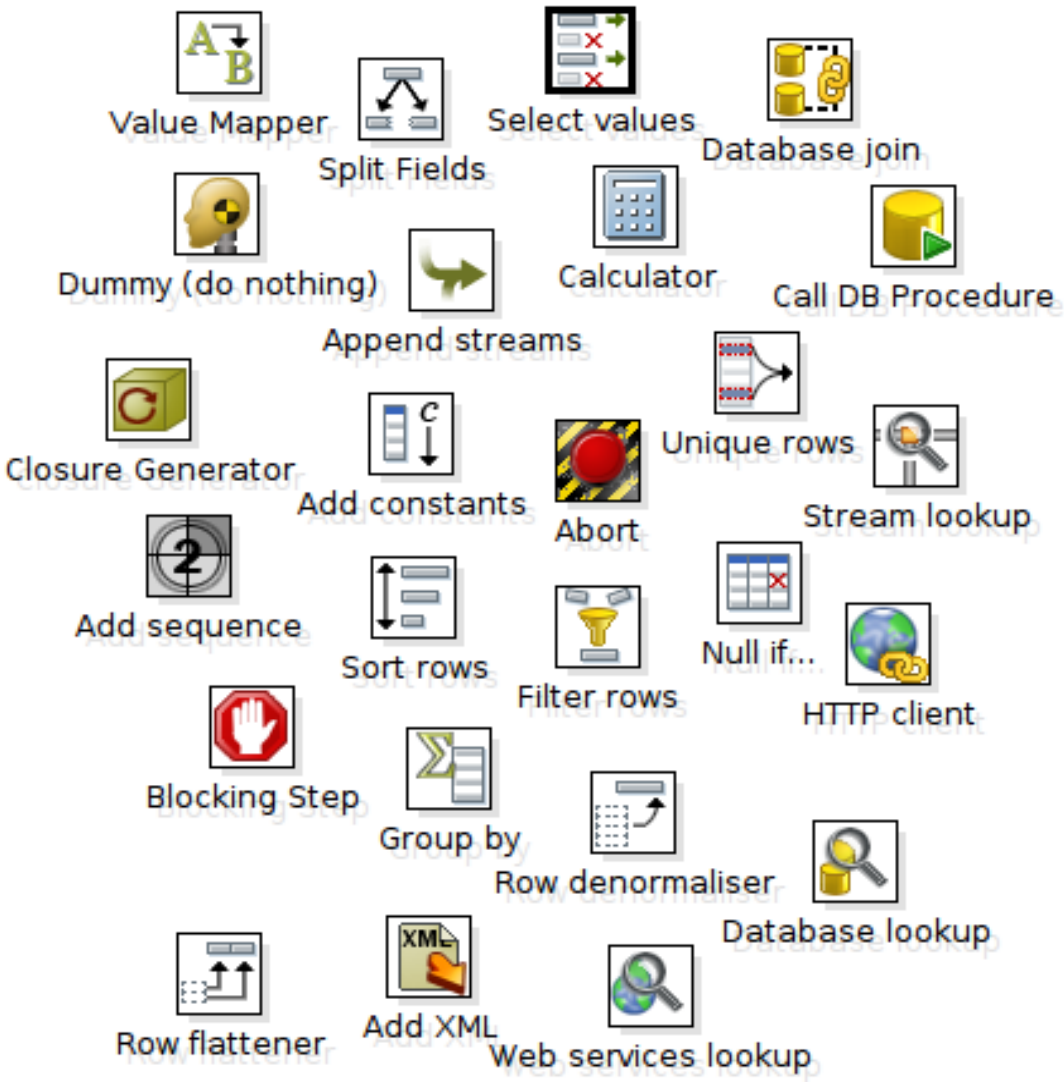
- Transportation of data
  - Engine based data transfer (no code generator)
  - Very flexible pathways:
    - splitting
    - partitioning
    - merging
    - joining
    - duplicating
    - clustering (MPP)





Pentaho Data Integration – Transformation

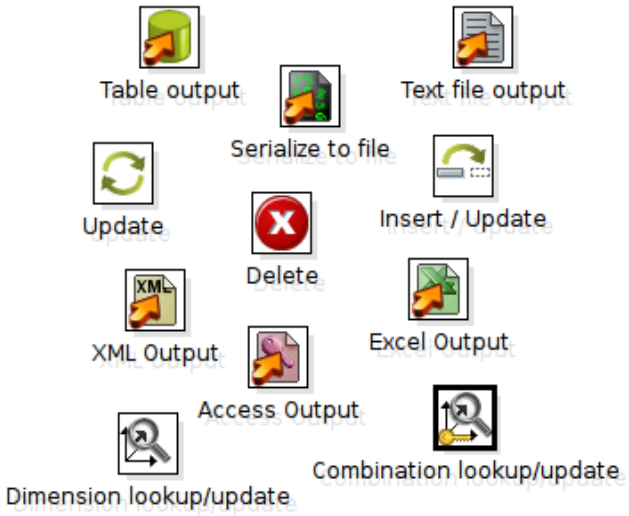
- Flexibly transform data
  - Looking up data
    - databases
    - files
    - memory...
  - Calculating
  - Scripting
    - JavaScript, SQL, RegExp
  - Splitting
  - Mapping
  - Selecting
  - Filtering
  - Pivotting ...

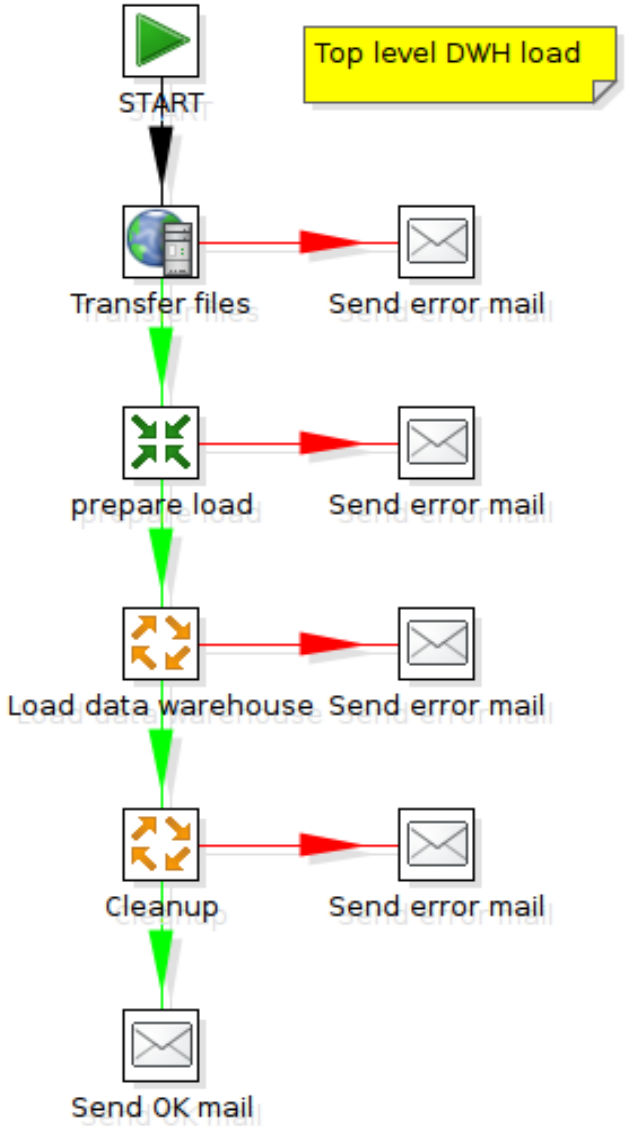


pentaho™
open source business intelligence™

Pentaho Data Integration – Loading

- Load data into a target format
  - Database loads
  - Data warehouse population
  - Partitioned loading
  - Bulk loading
  - Parallel loading
  - Clustering

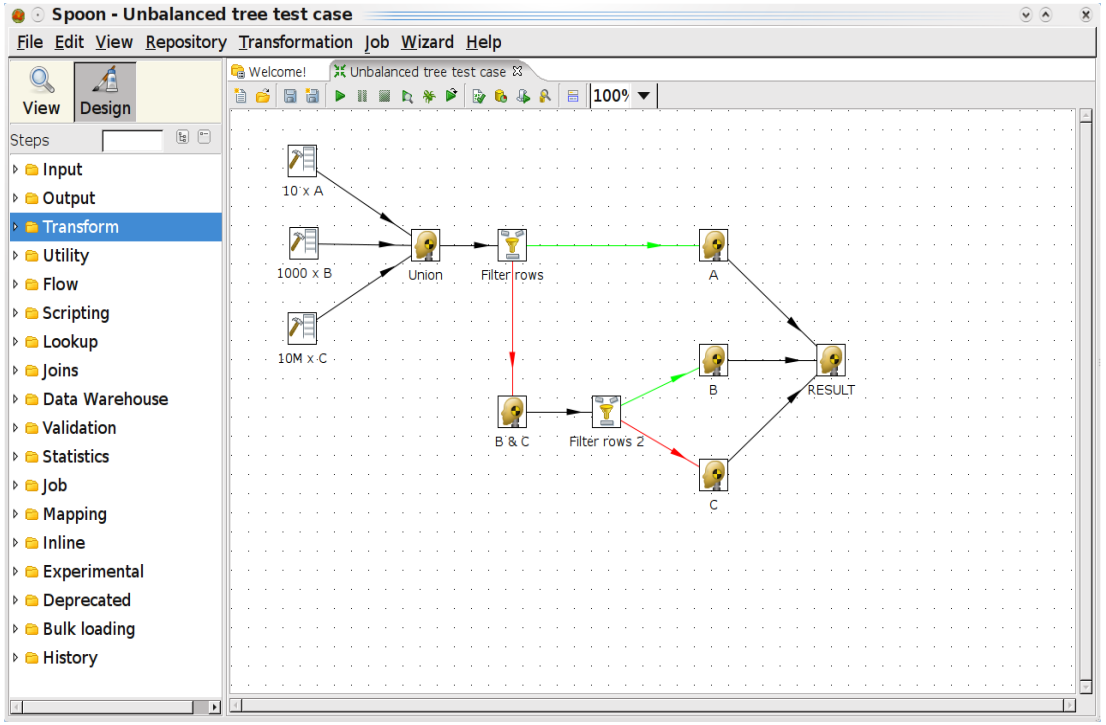




pentaho™
open source business intelligence™

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™
open source business intelligence™

Pentaho Data Integration – Community

- Paying Pentaho customers
- Large and small corporations
  - All possible sectors
- Lone rangers & Hobbiests
- All regions on Earth
- Meet on our Forum : +30,000 posts in 3 years
- Use our JIRA case tracking systems
- Download more than 10,000 copies of Kettle per month

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

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

Pentaho Data Integration – use-cases

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

pentaho™
open source business intelligence™

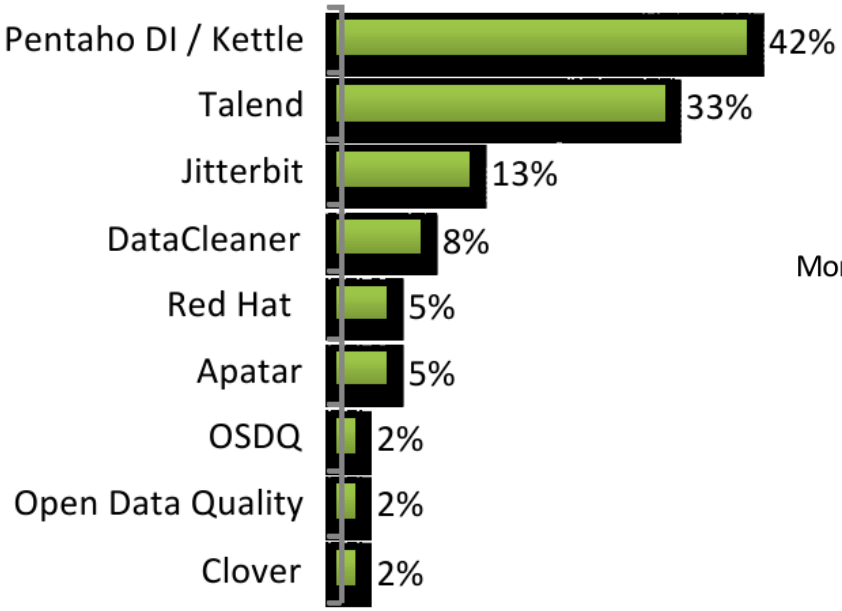
Pentaho Data Integration – Adoption

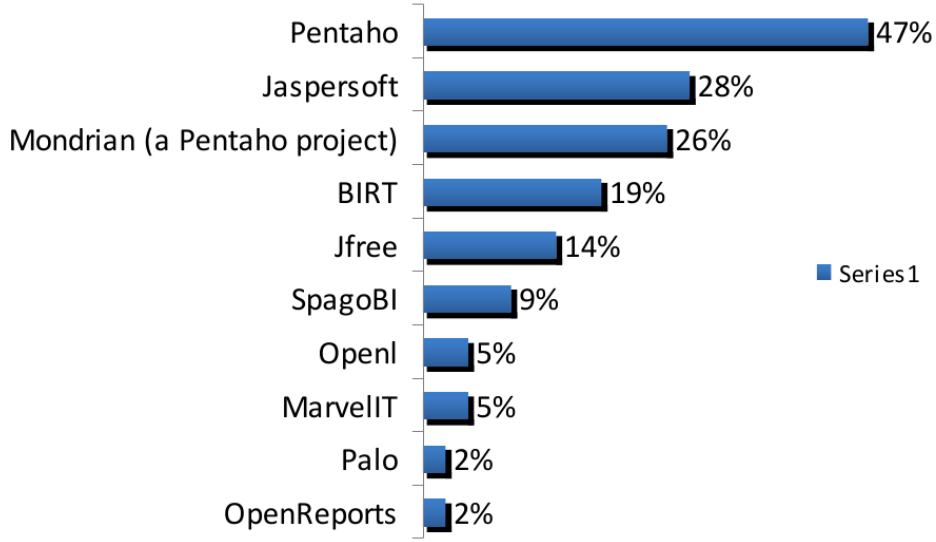
- Wide range of production deployments
  - Small and medium-sized companies
  - Large enterprises
- Rapid product evolution
  - Driven by Pentaho investment
  - Includes significant community contributions
    - "Contribution-friendly" architecture
    - Natural fit for additional data sources, targets and transformations

Pentaho Data Integration – Adoption

- Most deployed open source data integration solution. Independent study by Mark Madsen of Third Nature and the BeyeNETWORK
- Download free study at pentaho.com

BeyeNETWORK
Global coverage of the business intelligence ecosystem





pentaho™
open source business intelligence™

Pentaho Data Integration – Links

- 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

Agenda

- Pentaho: an introduction
- Pentaho Data Integration
- ***Version 4 : New features***
- MySQL support in PDI
- Q&A

Version 4: New features - Visualisation

Demo

- New welcome screen
- Mouse-over slide-outs for icons
- Hop creation
- Improved error handling configuration
- New perspectives support for Agile BI visualisations, modelling, scheduling, etc.

Version 4: New features - Running jobs

- Drill down into running job entries
- Visual indicators of running and completed job entries
- Success and failure mini-icons
- Mouse over completion mini-icons shows details of execution results
- Log capturing of completed job entries

Version 4: New features - Running transformations

- Drill down into running transformation job entries and mappings
- Row input/output sniff testing: see what rows are passing (demo)
- Remote input/output sniff testing on a Carte server

Version 4: New features - Better logging

- Reduced memory consumption
- Incremental log updates
- Global log buffer size limit for long running jobs/transformations
- Interval logging
- Auto clean-up of old log records
- Log record time-outs & execution lineage
- Log record colour coding in Spoon (blue and red for error lines)
- Step and job entry level Logging
- Execution lineage logging
- Renaming individual columns
- Global configuration options for all log tables

Version 4: New features - Plugins

- Unified plug-in architecture
- Easier deployment and packaging
- Step, job entry, partitioner, database type, spoon perspective, life-cycle, ... : all pluggable
- --> MySQL 5.1 plugin

Version 4: New features - Repositories

- Allowing for 3rd party repositories like the Pentaho Unified Enterprise Repository
- Removed dependencies to relational database repository (still supported though)
- Added support for repositories capable of team-development (file locking)
- Added support for repositories capable of fine-grained security repositories
- Added support for repositories capable of storing and retrieving revision history

Version 4: New features – New steps

- SAP Input
- Data Grid
- OLAP Input (Mondrian, Palo, SSAS, SAP B/W)
- Palo Cell Input/Output, Dimension Input/Output
- Salesforce Delete, Insert, Update, Upsert
- Add fields changing sequence (group sequence)
- User Defined Java Class: create your own plugin in Java on the fly in a step
- Send information using Syslog: Send a message to a Syslog server.
- Java Filter
- Memory Group By
- Farrage streaming bulk loader
- Teradata Fastload Bulk loader
- Experimental steps like Get table names, Email messages input, ...

Agenda

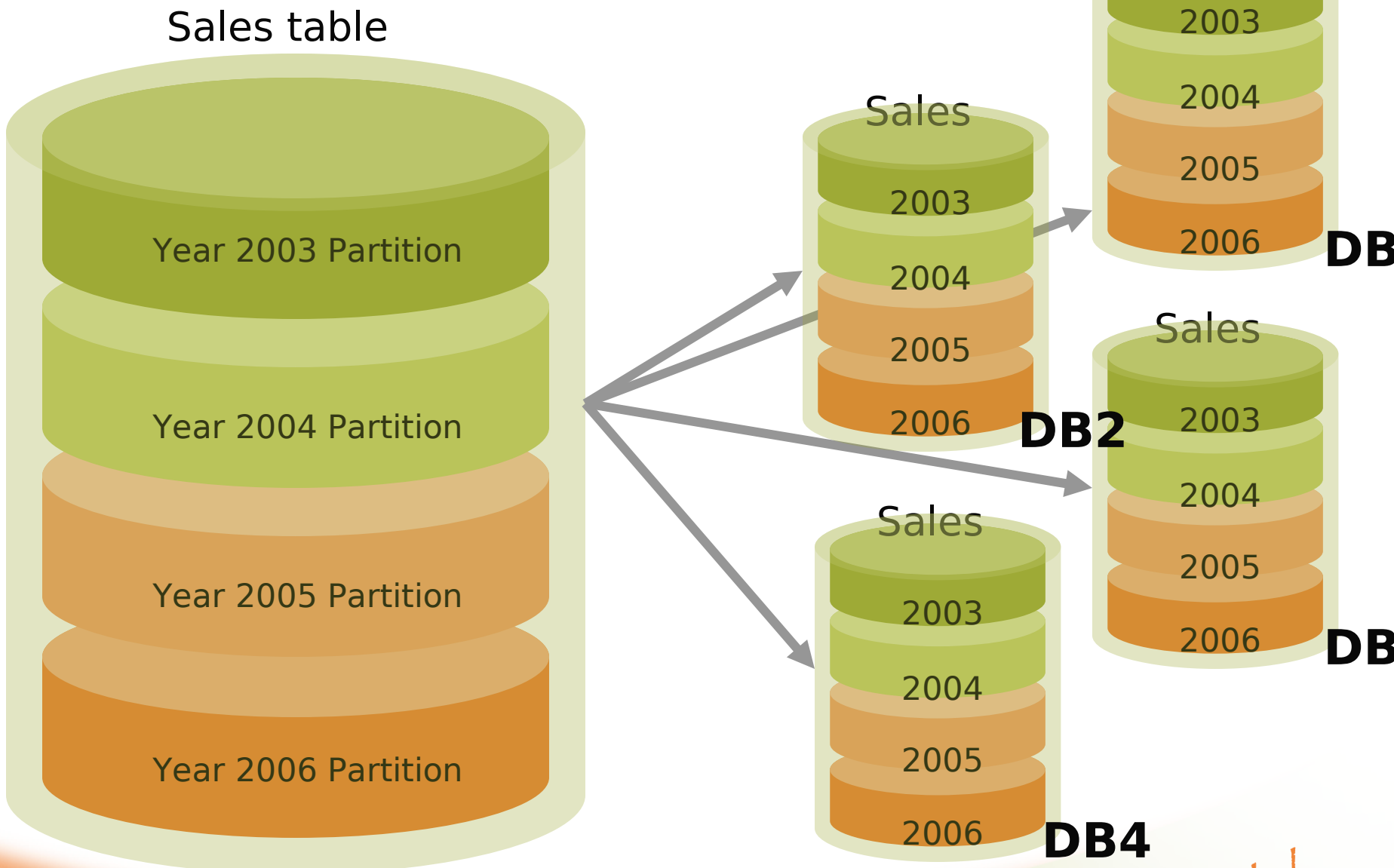
- Pentaho: an introduction
- Pentaho Data Integration
- Version 4 : New features
- ***MySQL support in PDI***
- Q&A

MySQL Support in PDI

- JDBC/ODBC Driver Integration
- Reading: MySQL Result Streaming (cursor emulation) support
- Writing: MySQL dialects for data types
- Job entry: Bulk Loader of text files for MySQL
- Job entry: Bulk writer to a text file for MySQL

Database Partitioning (Sharding) *Demo*

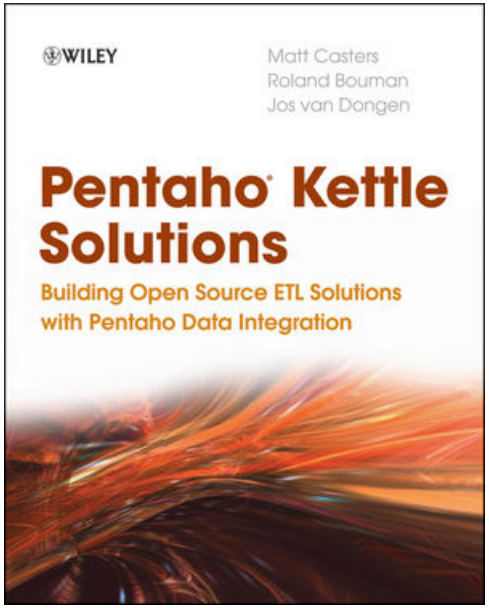
Database partitioning



pentaho™
open source business intelligence™

Questions and Closing

- Other Pentaho related User Conference information:
- Collapsing BI from Months to Minutes (Agile BI)
  - Jared Cornelius
  - Ballroom H
  - 11:55am Tuesday April 13th
- MySQL Binary Log Analysis With Pentaho BI
  - Robert Booth
  - Ballroom B
  - 5:15pm Wednesday April 14th
- The Pentaho Booth 516 in the Exibition Hall



ETA: September 2010

pentaho™
open source business intelligence™