

Pentaho BI Suite

Main features and data integration

edited by Vladan Mijatovic
(vladan.mijatovic@gmail.com)

Pentaho BI Suite

- Open source Business Intelligence tool
- It provides support for:
 - Data Integration
 - Reporting
 - Dashboards
 - OLAP Analysis
 - Data Mining



Reporting



Analysis



Data
Integration

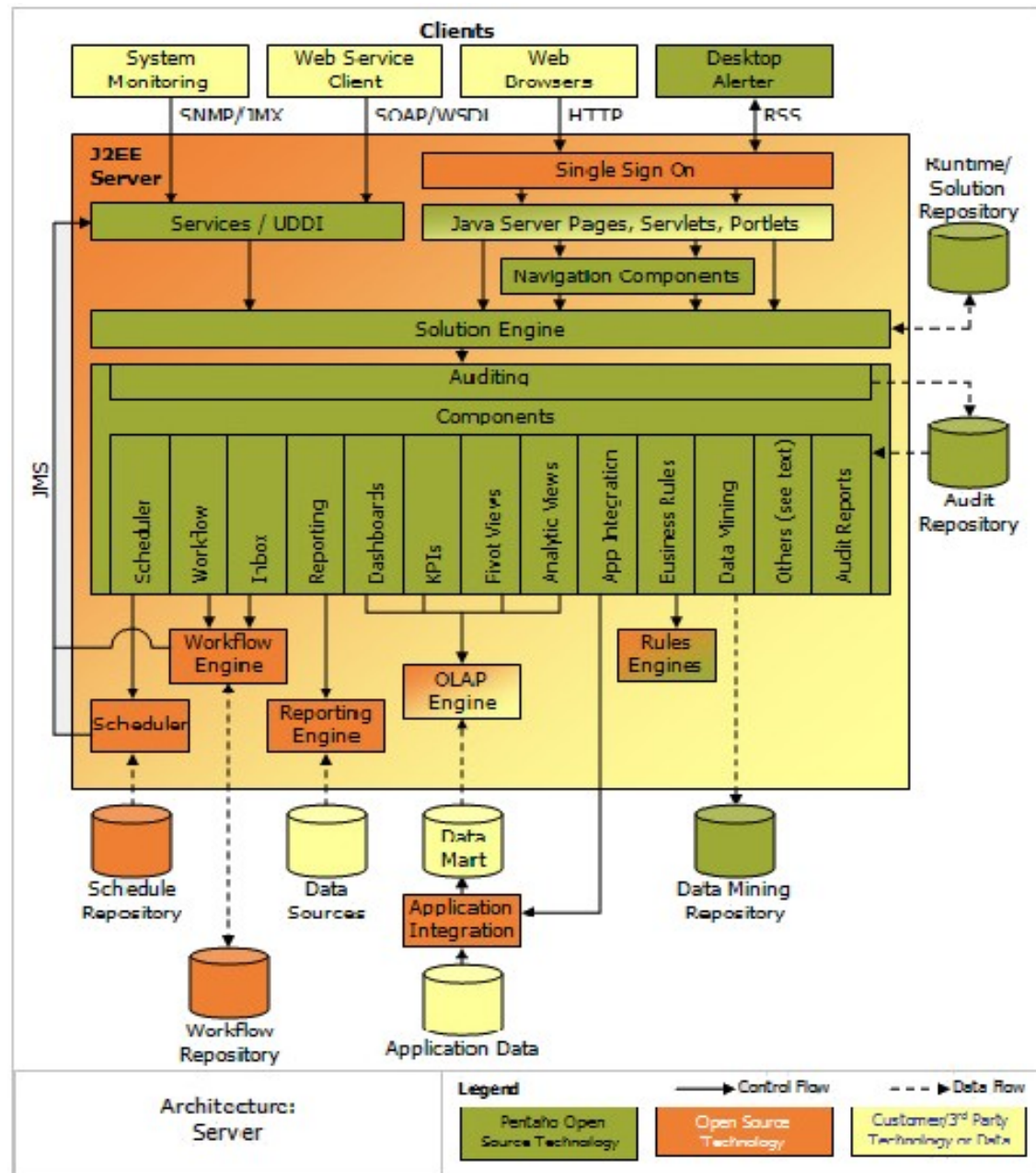


Dashboards



Data Mining

Pentaho Architecture



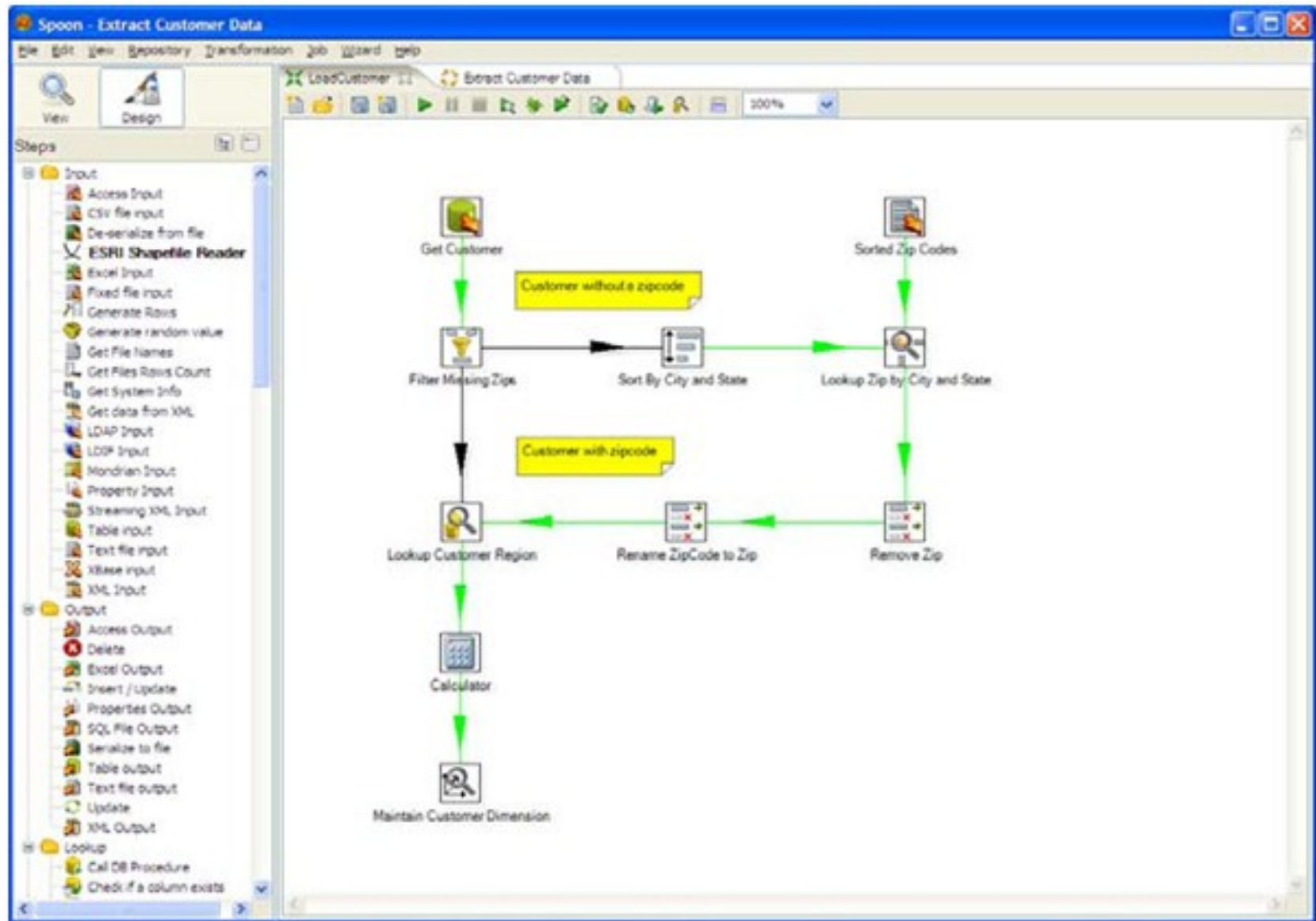
Pentaho Data Integration (PDI)

Comes with a user friendly interface and provides various tools to:

- Retrieve data from multiple data sources
- Clean, correct and normalize the data
- Filter only valuable data
- Group data (cross DBMS joins)
- Load data
- Possibility of creating a customized tools

PDI – Example

Kettle/Spoon



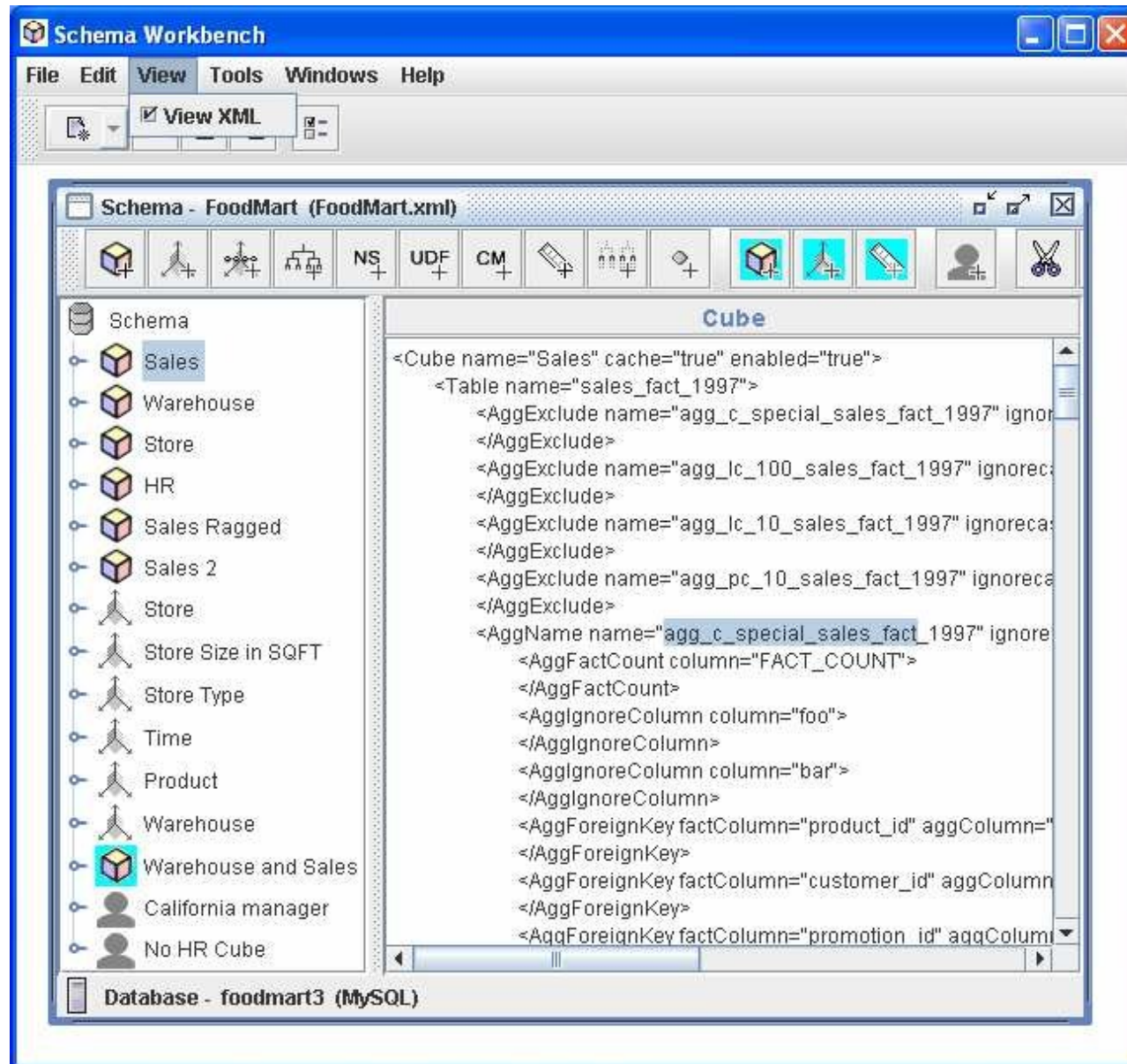
Pentaho Schema Workbench (PSW)

It provides the following functionalities:

- Schema editor integrated with the underlying data source for validation
- Test MDX queries against schema and database
- Browse underlying databases structure

PSW – Example

Schema Workbench



Pentaho OLAP Analysis

An OLAP Analysis allows us to:

- Study at once a whole bulk of data
- Observe data from different points of view
- Support decisional processes
- The most common functions are: Slicing, Dicing, Drill-down, Drill-across, Drill-through

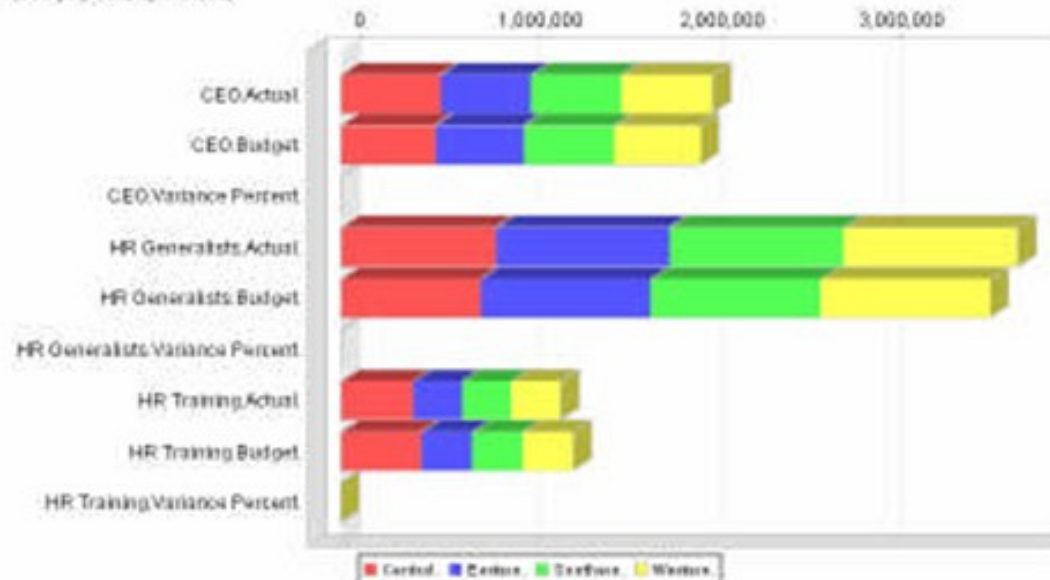
Pentaho Analysis

Mondrian

		Region			
Positions	Measures	Central	Eastern	Southern	Western
CEO	Actual	549,625.00	500,000.00	500,000.00	500,000.00
	Budget	522,250.00	488,750.00	498,750.00	478,750.00
	Variance Percent	-5.24%	-2.30%	-2.5%	-4.44%
HR Generalists	Actual	856,190.00	961,000.00	961,000.00	961,000.00
	Budget	771,225.00	940,158.00	940,158.00	938,158.00
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%
HR Training	Actual	397,473.00	271,200.00	271,200.00	271,200.00
	Budget	443,570.00	279,674.00	279,674.00	277,674.00
	Variance Percent	10.39%	3.03%	3.03%	2.33%

Slcen: [(A)]=All Departments]

Sl|cc: (All)=All Departments



Pentaho Reporting

(vs OLAP analysis)

- OLAP tools are dynamic, they allow users to interact with the system in a simple way while reports are more “static”
- The user does not have to know query languages but a minimum knowledge of the system is required while reports do not require that base knowledge
- They allow operations such as Roll-up, Drill-down, Drill-across, Pivoting, Slice-and-dice directly modifiable while examining the cube; the standard reports don't

Pentaho Reporting

Design Studio, Report Designer

The screenshot displays the Pentaho Report Designer interface. The main workspace shows a report design for 'Steel Wheels' with the following components:

- Report Header:** 'Steel Wheels' logo and 'Top Ten Customer Report' with a run time of 'Oct 07, 2007 12:18:49'.
- Top Ten Customer:** A horizontal bar chart showing sales for various customers. The x-axis ranges from 0 to 750,000.
- Product Mix:** A table showing sales data for different product lines, followed by a pie chart illustrating the distribution.

Top Ten Customer Data (Estimated):

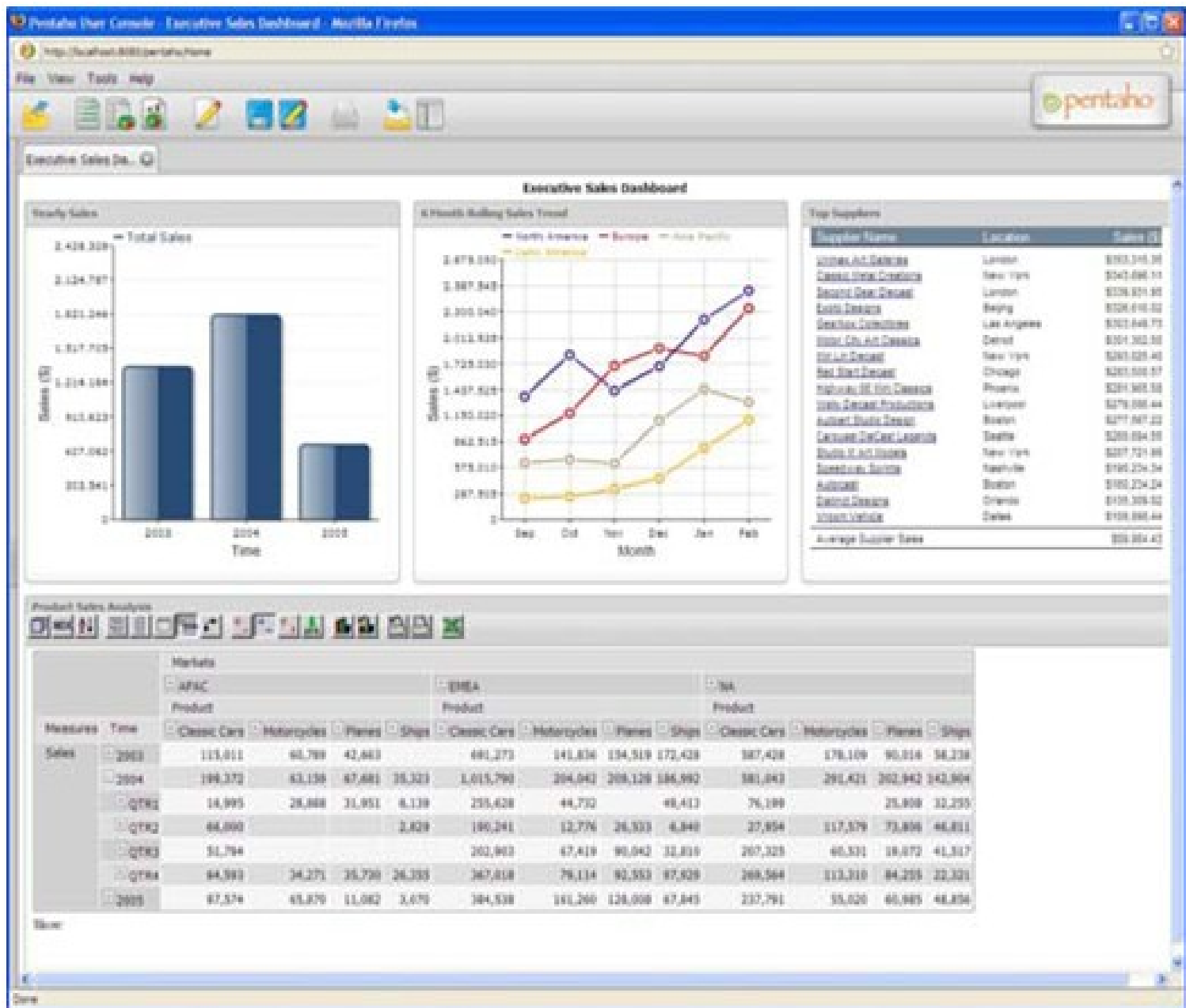
Customer	Sales (Approx.)
Euro+ Shipping Channel	750,000
Max C/O Distributors Ltd.	650,000
Australian Collectors, Co.	450,000
Wycle Machine, Inc.	350,000
La Rochelle C/Os	300,000
Down Under Souvenirs, Inc.	250,000
Dragon Souvenirs, Ltd.	200,000
Land of Toys Inc.	150,000
The Skop C/Os Warehouse	100,000
Kate's C/O Shop	50,000

Product Mix Data (Estimated):

Product Line	Sales (\$)	%
Classic Cars	409,427	41
Vintage Cars	244,413	24
Trucks and Buses	208,319	21
Ships	74,438	7
Planes	71,749	7
Motorcycles	63,775	6
Trains	37,834	4

The interface includes a 'Palette' on the left with various report elements like Label, Text Field, and Chart. On the right, there is a 'Structure' pane showing the report's layout and a 'Properties' pane for the selected element.

Pentaho Dashboards - mention



Data Mining - mention Weka

The screenshot shows the Weka Explorer window with the 'german_credit' dataset loaded. The 'credit_history' attribute is selected as the target class. The interface displays a list of 21 attributes, a table of the selected attribute's values, and a bar chart visualization.

Current relation:
Relation: german_credit
Instances: 1000
Attributes: 21
Sum of weights: 1000

Selected attribute:
Name: credit_history
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	no credits/all paid	40	40.0
2	all paid	49	49.0
3	existing paid	520	520.0
4	delayed previously	88	88.0
5	critical/other existing credit	293	293.0

Attributes:

No.	Name
1	checking_status
2	duration
3	credit_history
4	purpose
5	credit_amount
6	savings_status
7	employment
8	instalment_commitment
9	personal_status
10	other_parties
11	residence_since
12	property_magnitude
13	age
14	other_payment_plans
15	housing
16	existing_credits
17	job
18	num_dependents
19	own_telephone
20	foreign_worker
21	class

Class: class (Nom)

The bar chart shows the distribution of the 'class' attribute across the five categories of 'credit_history'. The bars are stacked with blue and red colors. The counts for each category are: 40 (no credits/all paid), 49 (all paid), 520 (existing paid), 88 (delayed previously), and 293 (critical/other existing credit).

ETL – Going into detail

- Pentaho Data Integration (PDI) is a tool used to extract, transform, and load (ETL)

Common uses:

- Data warehouse data loading – from scratch, bulk or incremental loading
- Data migration between different databases and applications
- Data Cleansing with steps ranging from very simple to very complex transformations
- Rapid prototyping of ROLAP schemas

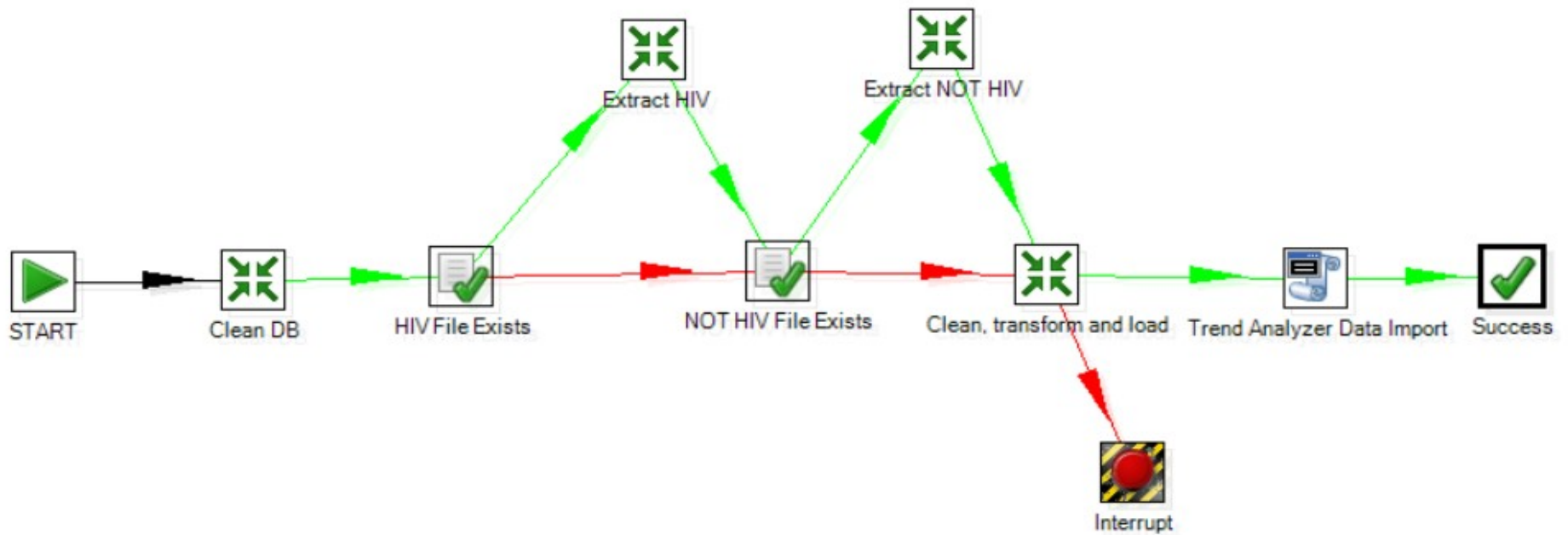
Jobs and Transformations

- All of the data flow is organized in jobs and transformations
- A Transformation is made of Steps linked by Hops. These Steps and Hops form paths through which data flows. Therefore it's said that a Transformation is data-flow oriented.
- A Step is the minimal unit inside a Transformation. A wide variety of Steps are available
- A Hop is a graphical representation of data flowing between two Steps, with an origin and a destination.

How can we create a hop:







- Hold a central mouse button and drag the arrow from one step to another
- Press Shift+click and drag towards the destination step
- Using GUI arrows

ETL Job - Example







ETL – most used steps

input files

 Access Input	Read data from a Microsoft Access File
 CSV file input	Simple CSV file input
 Excel Input	Read data from a Microsoft Excel Workbook
 OLAP Input	Execute and retrieve data using an MDX query
 Table input	Read information from a database table
 Text file input	Read data from a text file in several formats







ETL – most used steps

output files

 Access Output	Stores records into a MS-Access database table
 Excel Output	Stores records into a Excel (XLS) document with a formatted data
 Table output	Write information to a database table
 Text file output	Write rows to a text file






ETL – most used steps

other utils

 Select values	Select or removes fields in a row. Optionally set the field data type
 Split field to rows	Splits a single string field by delimiter and creates a new row for each new string
 Sort rows	Sort rows based upon field values
 Null if...	Sets a field value to null if it is equal to a constant value
 Filter rows	Filter rows using simple equations
 Execute SQL script	Execute a SQL script

ETL – most used steps

other utils - cont.

 <p>Modified Java Script Value</p>	<p>This is a modified plugin for the Scripting values with improved interface and performance</p>
 <p>Database join</p>	<p>Executes a database query using stream values as parameters</p>
 <p>Call DB Procedure</p>	<p>Get back information by calling a database procedure</p>
 <p>File exists</p>	<p>Check if a file exists</p>
 <p>Get Variables</p>	<p>Determines the values of certain (environment or Kettle) variables and put them in field values</p>

Workshop I - ETL

During this workshop your task is to:

- Create a transformation that loads all data from offices.csv, adjust the telephone number (eliminate the “+” sign) and load it to labsia database
- Create a transformation that loads all data from payments.xls to payments table. Pay attention to a “paymentdate” attribute (hint: use “select values”)
- Create a transformation that loads only Sales Reps from employees_aux to employees table (hint: use “filter rows”)
- Create a job that launches all these transformation, and control that the input files/tables exist before completing the job