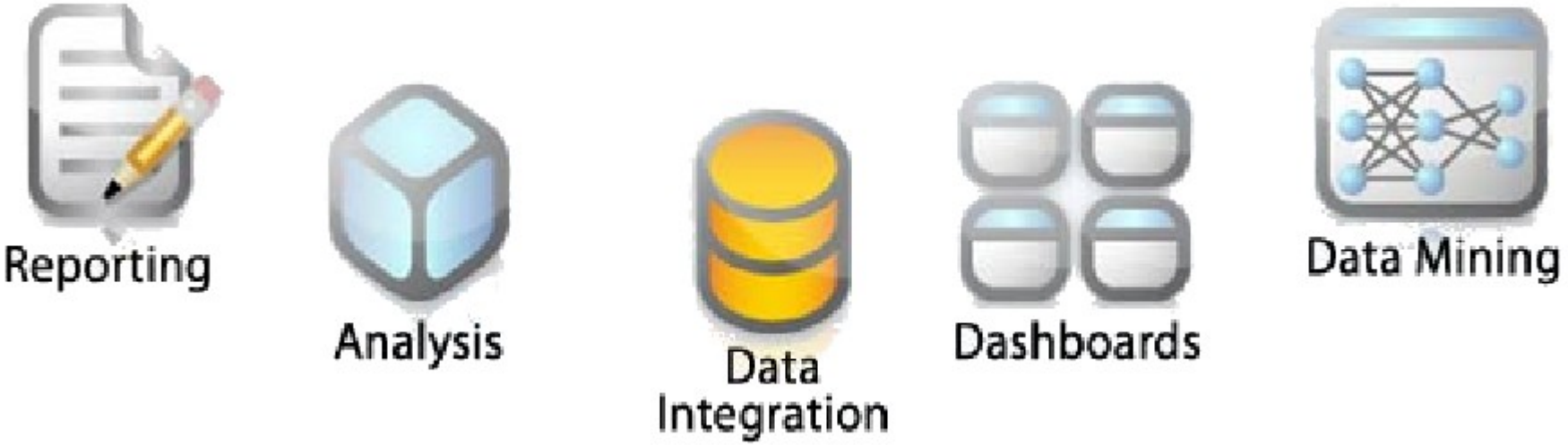
Pentaho BI Suite

Main features and data integration

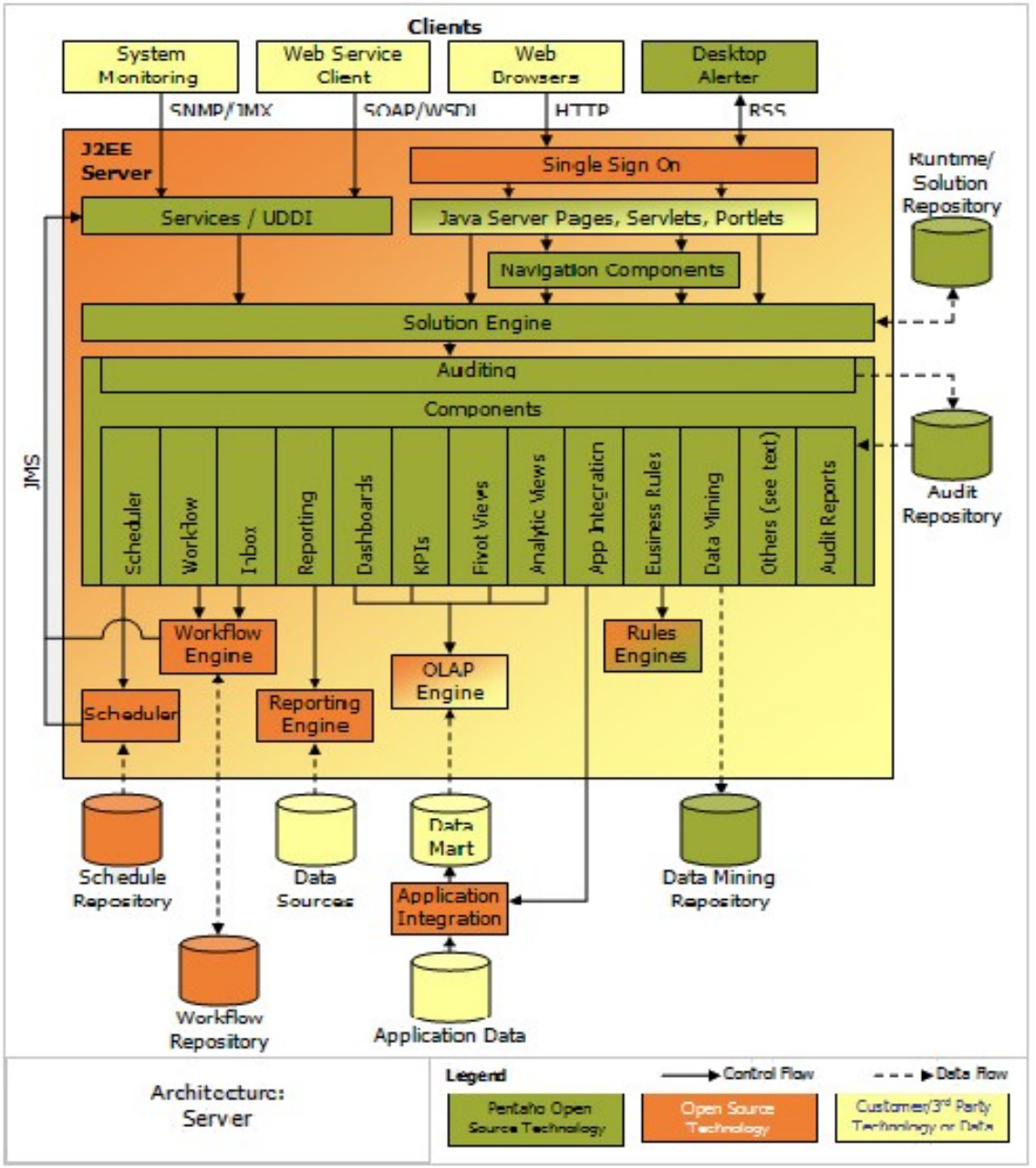
edited by Vladan Mijatovic
(vladan.mijatovic@gmail.com)

Pentaho BI Suite

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



Pentaho Architecture



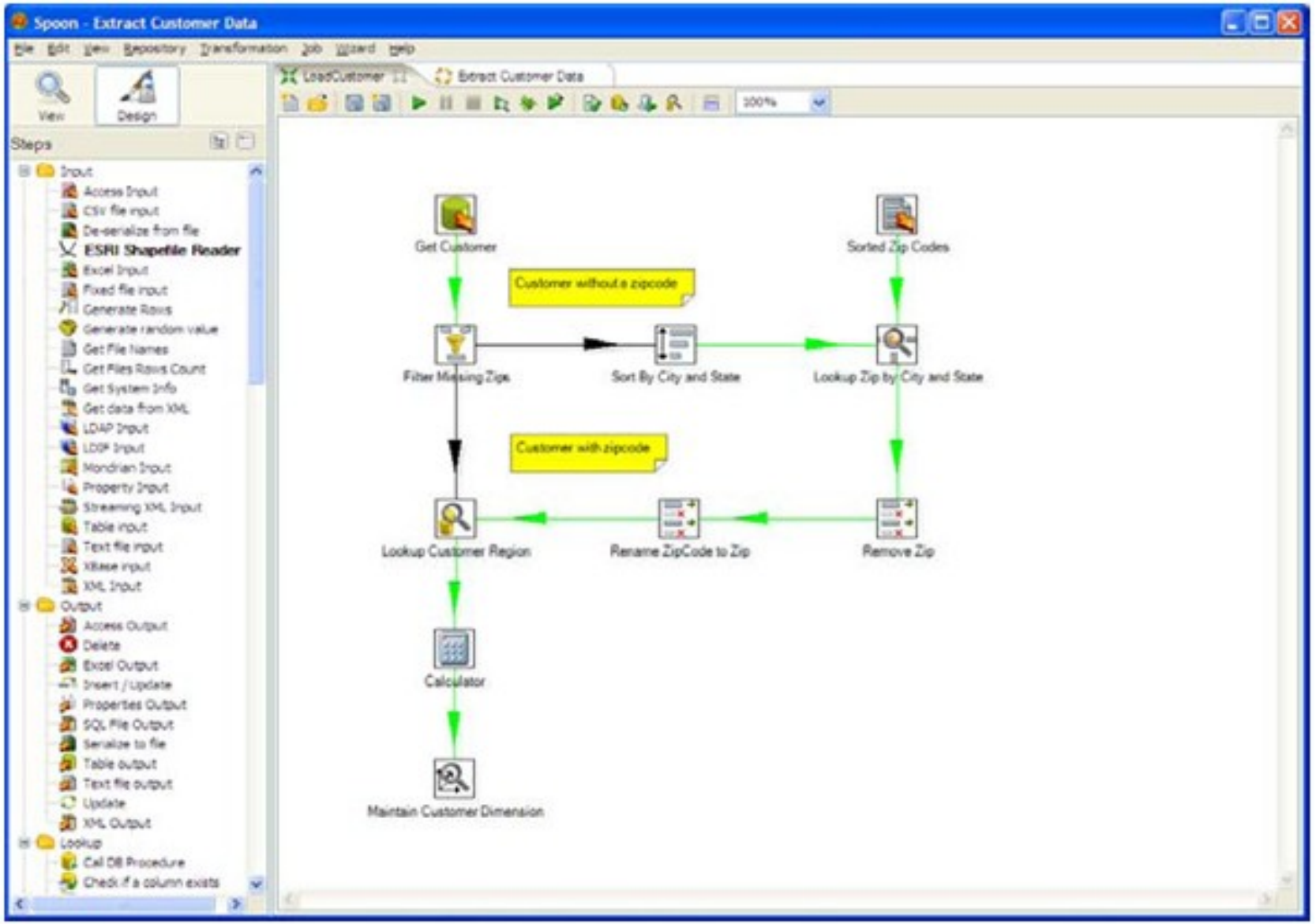
Pentaho Data Integration (PDI)

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



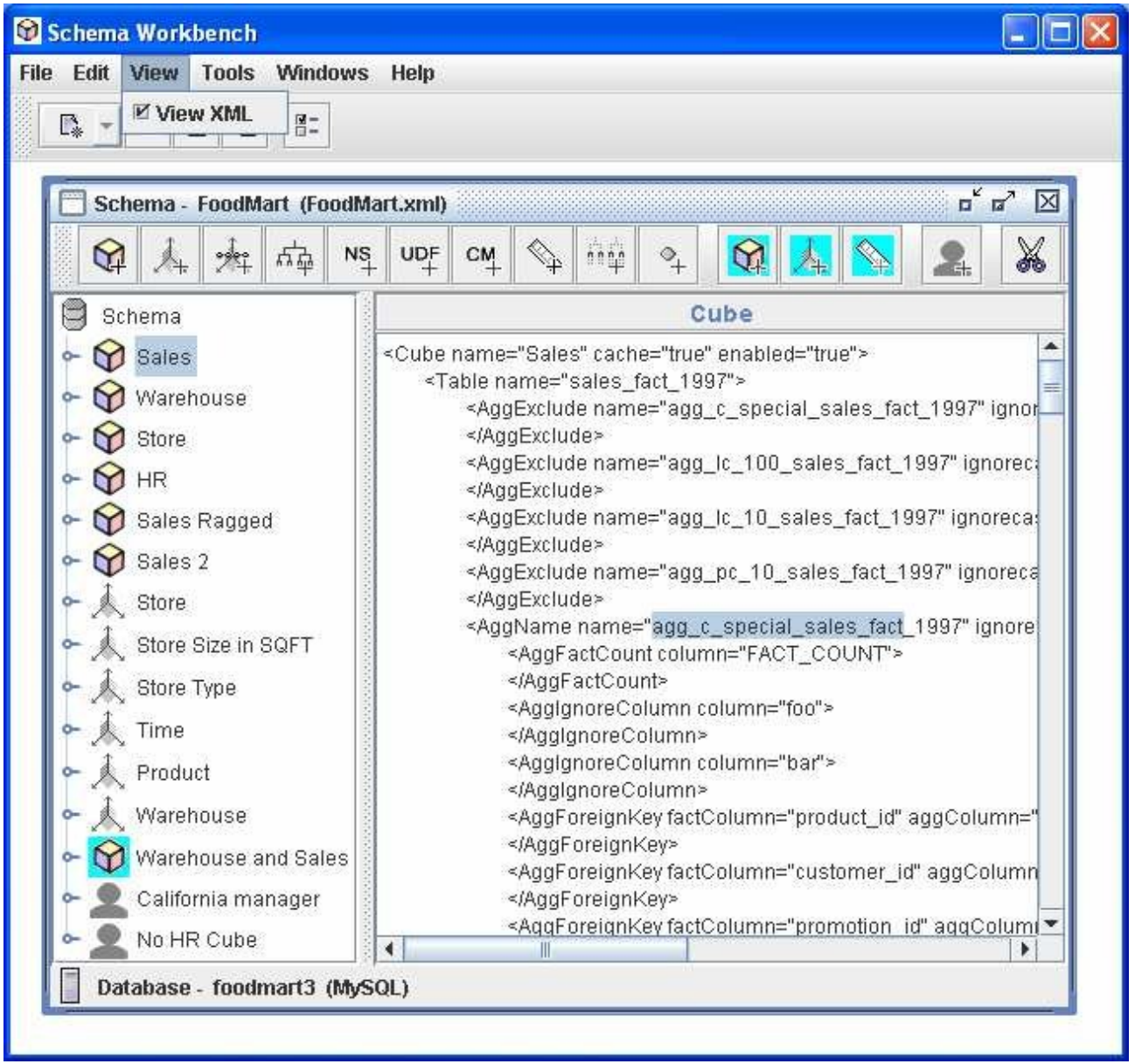
Pentaho Schema Workbench (PSW)

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:

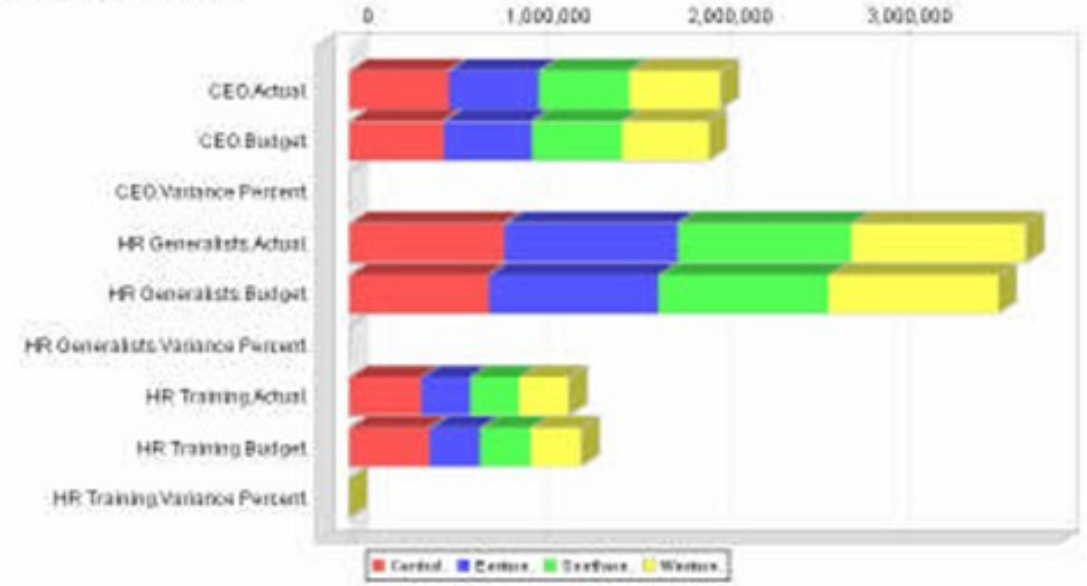
- Study at once a whole bulk of data
- Observe data from different points of view
- Support decisional processes
- The most common functions are: Slicing, Dicing, Drill-down, Drill-accross, Drill-through

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	488,750.00	478,750.00
	Variance Percent	-5.24%	-2.30%	-.25%	-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%

Slicer: [(All)=All Departments]

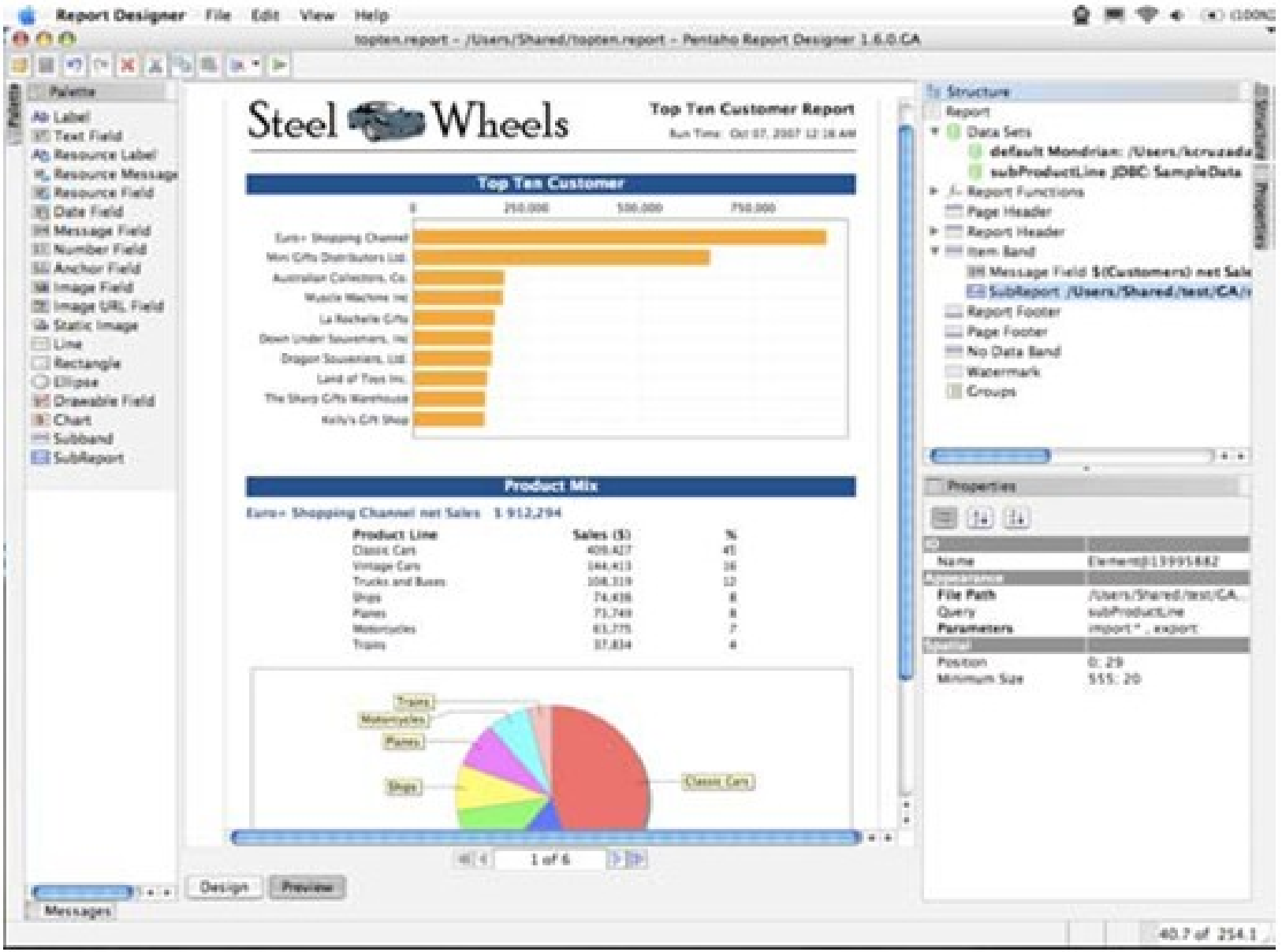


Pentaho Reporting
(vs OLAP analysis)

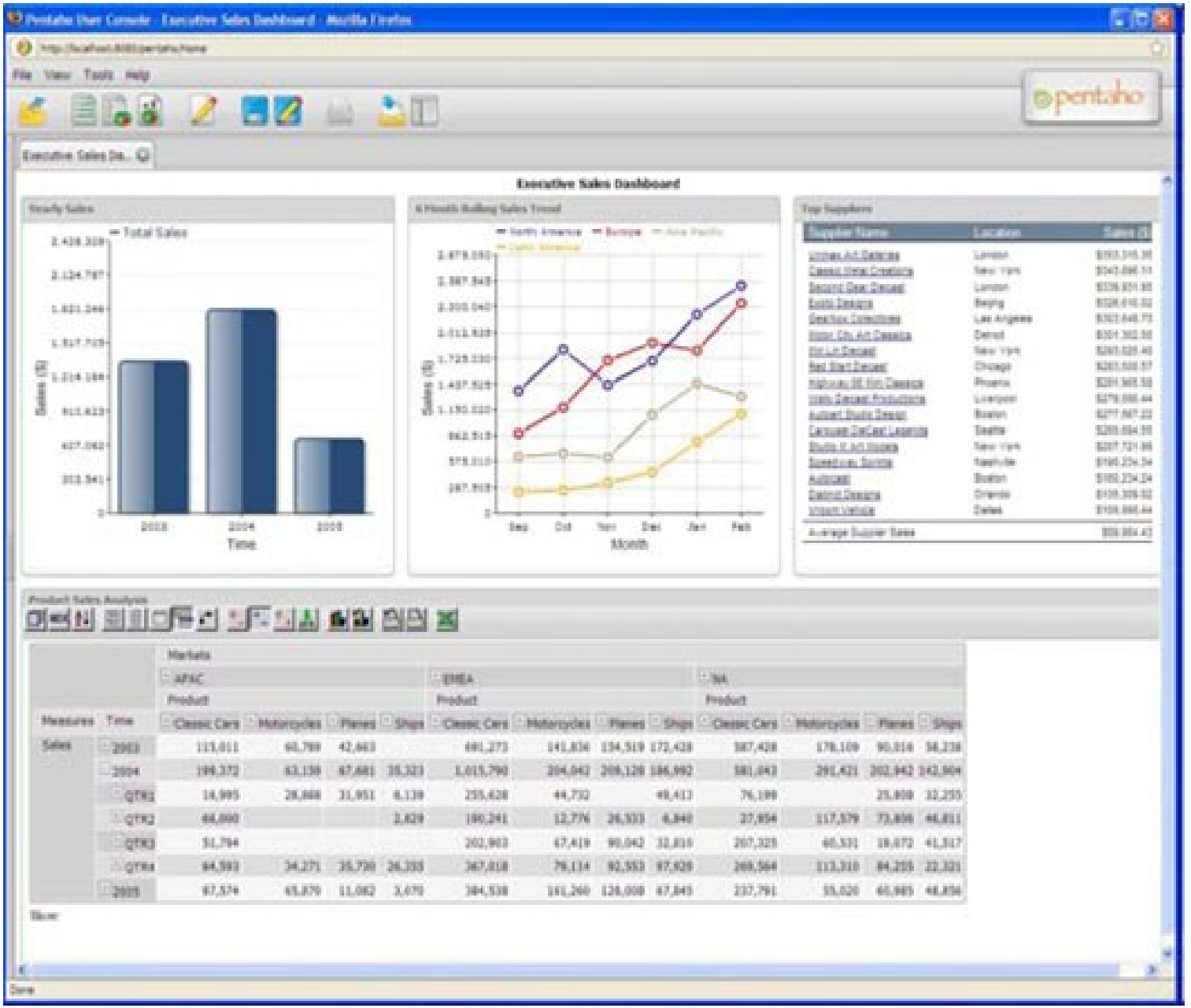
- OLAP tools are dynamic, they allow users to interact with the system in a simple way while reports are more “static”
- The user does not have to know query languages but a minimum knowledge of the system is required while reports do not require that base knowledge
- They allow operations such as Roll-up, 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

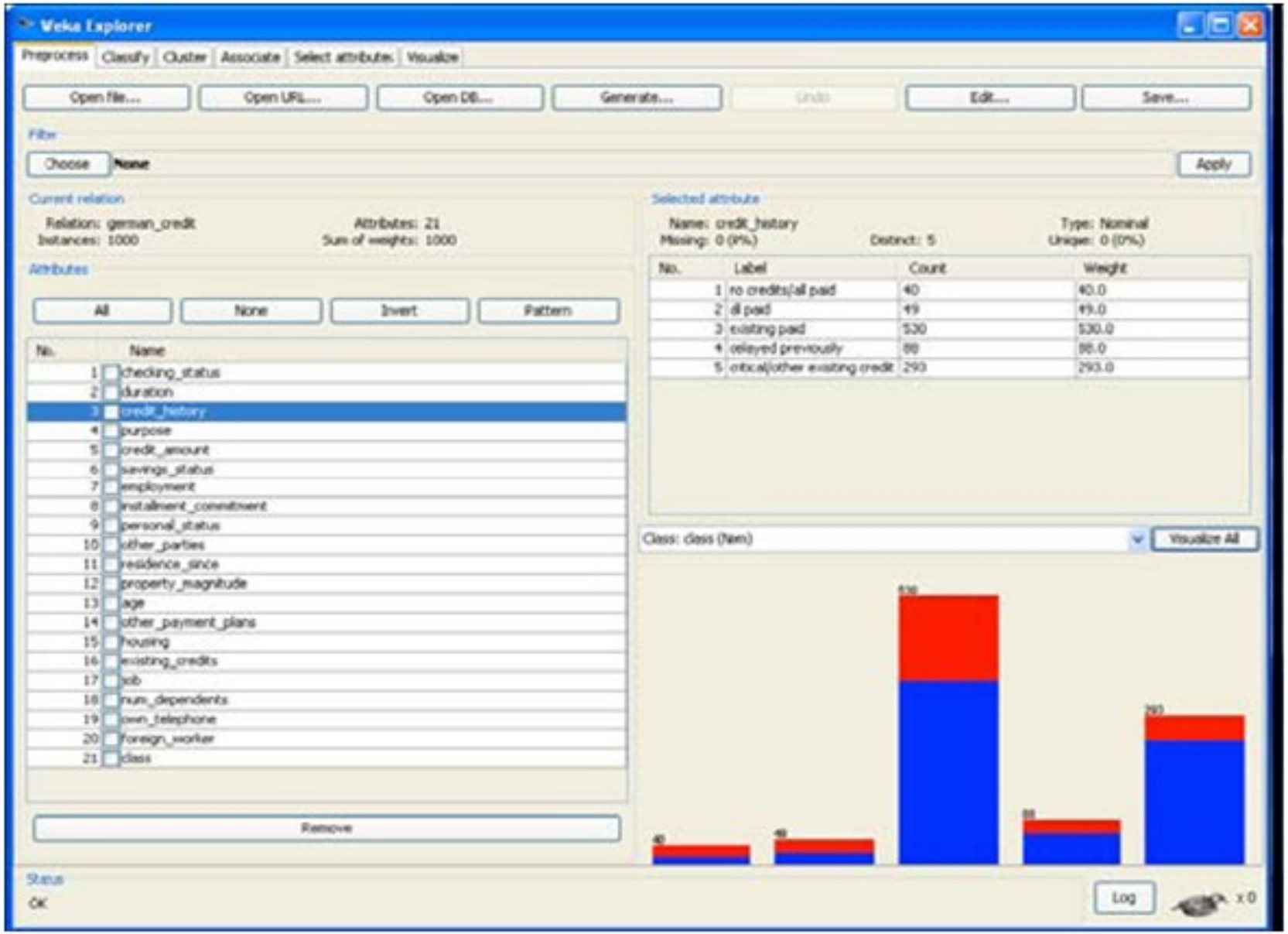


Pentaho Dashboards - mention



Data Mining - mention

Weka



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

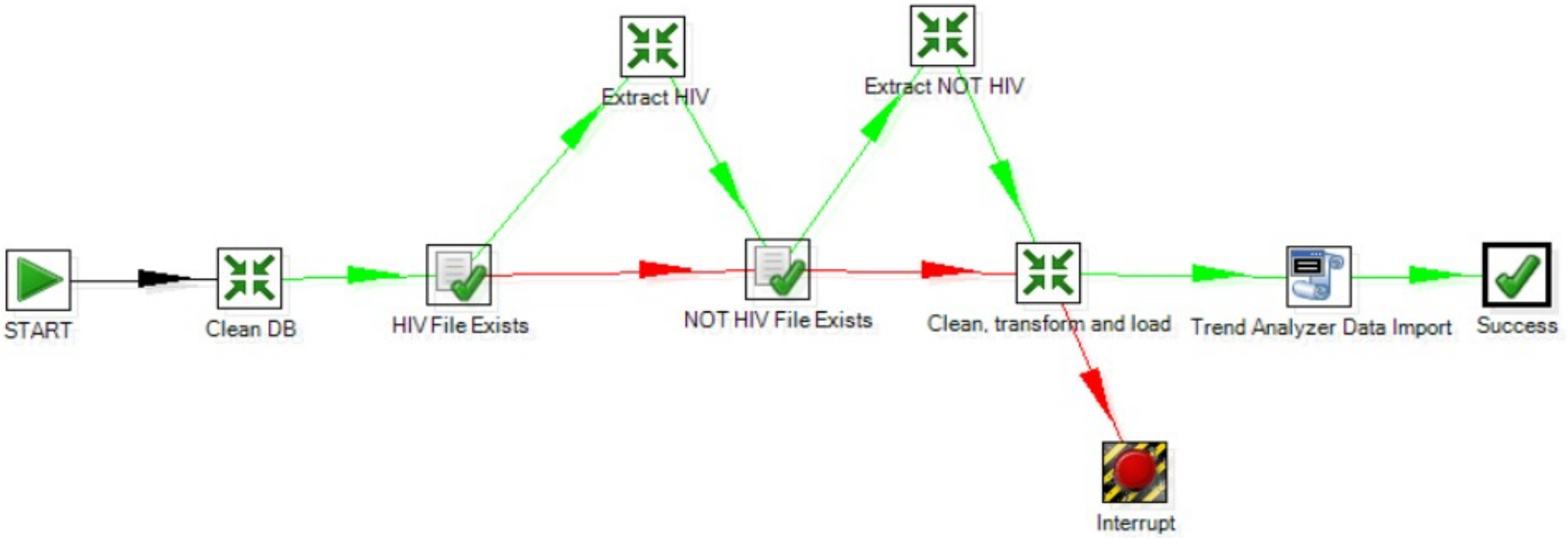
Jobs and Transformations

- 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.

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

Workshop I - ETL

During this workshop your task is to:

- Create a trasformation 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