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







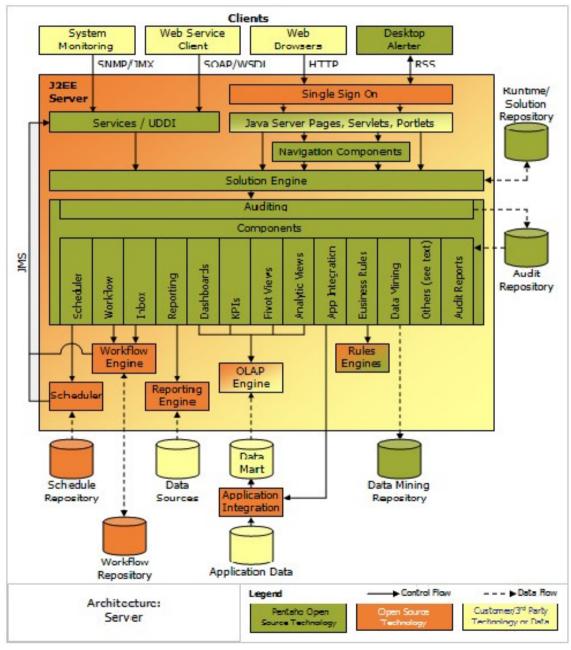


Dashboards



Data Mining

Pentaho Architecture

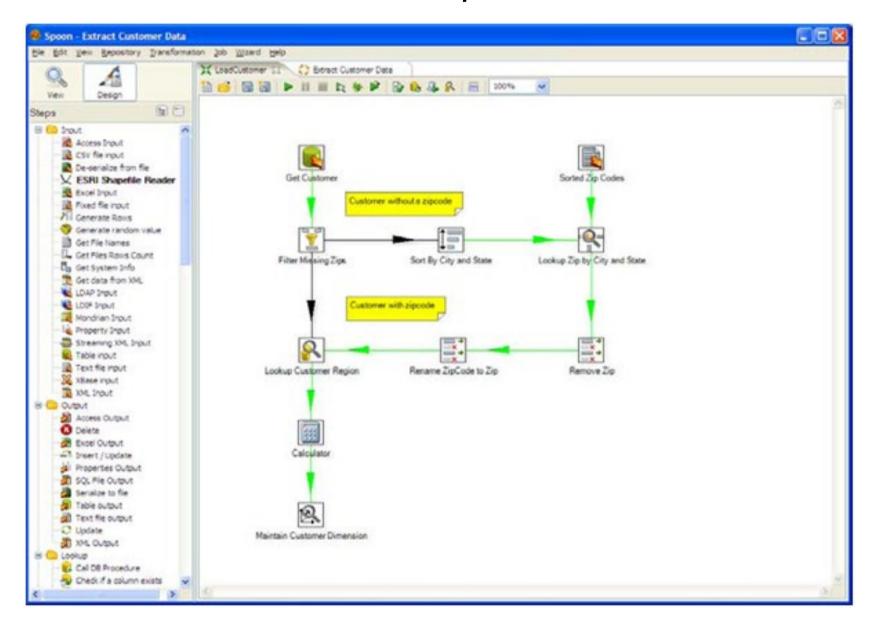


Pentaho Data Integration (PDI)

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

View XML Schema - FoodMart (Food 公子 A Schema Schema Schema Sales Warehouse	Mart.xml)
Schema - FoodMart (Food 公 人 外 内 N Schema	IS UDF CM Sp mid or SQ ▲ SA & & & & & & & & & & & & & & & & & &
 会報 Schema Sales Sales	IS UDF CM Sp mid or SQ ▲ SA & & & & & & & & & & & & & & & & & &
Schema	Cube
Sales	
~	Autor a second 10 store in sector interval an object interval
 Store HR Sales Ragged Sales 2 Store Store Size in SQFT Store Type Time Product Warehouse Warehouse and Sales California manager No HR Cube 	<table name="sales_fact_1997"> <aggexclude ignore<br="" name="agg_c_special_sales_fact_1997"></aggexclude> </table>

Pentaho OLAP Analysis

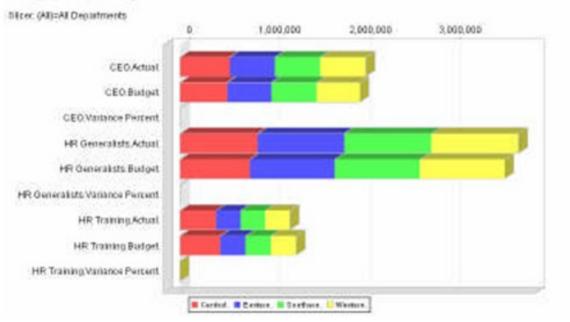
An OLAP Analysis allows us to:

- 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	1549,625.00	+500,000.00	\$500,000.00	+500,000.00				
	Budget	\$522,250.00	488,750.00	498,750.00	4478,750.00				
	Variance Percent	-5 24%	-2.30%	- 25%	-4.44%				
HR Generalists	Actual	856,190.00	4961,000.00	\$961,000.00	4961,000.00				
	Budget	1771,225.00	4940,158.00	4940,158.00	4938,158.00				
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%				
HR Training	Actual	4397,473.00	\$271,200.00	\$271,200.00	4271,200.00				
	Budget	1443,570.00	1279,674.00	1279,674.00	1277,674.00				
	Variance Percent	10.39%	3.03%	3.03%	2.33%				

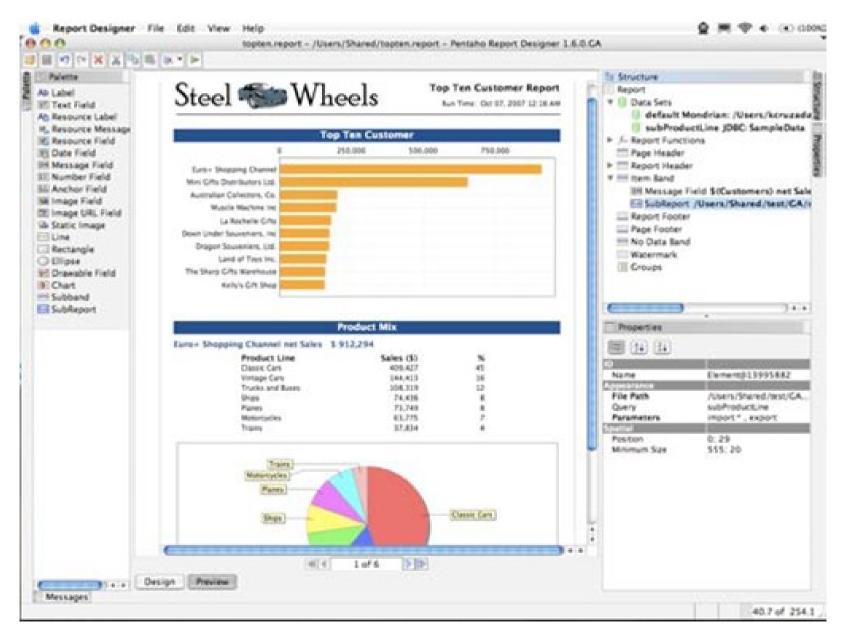
Sken: [(Al)=Al Departments]



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

Artist Taxad					reles										
A CONTRACTOR OF STREET	with the second	aru,hana													
Van Ti		-		_			_		_		_			-	entaho
	00	28	2	Wh -										a lai	ALC: NOT AL
	A STREET				-										
lecutive Sal	n pa. Q														
-							Locative Sal	ies Dashi	ward	_	-				
marty Sales						South Hollong Sol		111-1-11			Top Suppler		-	101102	-
2.434.33	- Total 1	Latere					affi America - ulti America	- berepa	-	100	Reason in Association,		_	Lacates	
3.124.74						2.479.040					Cases, Swa			Sanatan Sanat Yana	\$50.345 \$542.680
	÷.					1.307.545				~	Second Gen			Landon	1224.87
1.821.24	*					2.305.040			1	1	Exets Design Seathers Do			Baging Cat. Angelete	8108.440
1.1.1.1.1	13-					1011300-			1	/	TIDA CALA	a Connecta		Detroit	6304.38
						1.725.230	~	0	0		Red Sign 2x			Sau 194 Dicept	8263.02
-						1.407.518	1)	and the	1		Spinete H	tion Connects		Prosta	\$251.36
101.0	-		· · · · · ·			1.135.220-	1	-	1		Table Local Autorst Date	n Productions In Treasur		Lowper Bastel	1276.00 1277.00
			-	_		162,212	6	1		1	Converting and	CentLeanth		Deatte	5281.89
627.01	624					121.000	0-0-	-0	1		Indu Carl			Natural Vigets Naturality	\$307,72 \$195,25
313.54	a					1.		- Comment	0		Address.	100		Busin	\$185,25
						287,818	0				Eastmit Deal Writers Verla			Crientis . Carles	2108.309
	241	8 2004	- 22	18.		1	es det	These 2	ni 34	et Fall	Australia Des	and the second se		1999	10.10
		Time						Month							
				au sui		24									
	Anutyus Still C	FILE.	1												
		Neters			hard had	-									
						THEA				- NA					
1 <u>51</u>		Hertets - AFSC Product				-EHEA Product				Product					
Mandares	Time 1	Marteta AFAC Product Chessic Caris - M	atorgeles (Paries	Shipt	-EHEA Product - Cheste Care			Shipi	Product Clease Cars		Parat			
19H	Tere 290	Narlats ARAC Product Christ Cara - M 115,011	eturojdes eturoj	Parenti 42,663	Shipt	EMEA Product Chesses Cares 641,273	141,836	134,519	- Shipi 172,428	Product Clease Carsi 387,428	179,109	90,014	54,238		
Manalarian	Tene 2003 2004	Harluts AFAC Product Cases: Care M 113,011 114,372	61,799 63,138	Planeti 42,563 67,681	- 5Pages 16,3423	- EHEA Product Chesk, Cars 461,273 1.015,790	141,834 204,042		Ships 172,428 186,992	Product Clease Cars 387,428 381,843	179,109	90,014 202,942 3	54,238 42,904		
Face ()	Terms 2003 2004 QTH1	Harturs AFAC Product Chessic Care 1 M 115,011 194,372 14,995	eturojdes eturoj	Planeti 42,563 67,681	55kget 35,323 6,339	- EHEA Product - Chesse, Cars 641,273 1,015,790 255,428	141,834 294,043 44,732	134,519 208,129	594ps 172,428 186,992 48,413	Product Consoc Cars 187,428 381,043 76,399	178,109 291,421	90,214 202,942 2 25,908	94,238 42,904 32,295		
Manalarian	Tene 2963 -2004 -2004 -2004 -2004 -2004	Harfarts AFAC Product Chesse Care - M 115,011 194,372 14,995 44,000	61,799 63,138	Planeti 42,563 67,681	- 5Pages 16,3423	- EHEA Product - Cennet Cenn - 641,373 - L015,790 - 235,638 - 186,341	141,834 204,043 44,732 12,776	134,519 208,129 26,533	594pt 172,428 186,992 48,413 4,840	Product Censor Ciers 347,428 381,043 76,299 27,854	178,109 291,401 117,579	90,214 202,942 2 25,909 73,804	94,338 43,904 32,295 46,811		
Messares	Tene 2968 2964 2004 2004 2004 2004 2004 2004 2004 20	Harlars APAC Product Chess Cars M 115,011 194,372 14,995 64,000 51,754	64.799 63.139 53.138 25.668	Pares 42,663 47,681 31,951	- SPige 35,323 6,139 2,429	- gHEA Product - Chemic Carel 491,273 L015,790 235,628 196,241 202,613	141,434 294,043 44,732 12,734 67,439	134,519 208,128 26,503 90,042	594ps 172,428 186,592 41,413 4,340 31,315	Product Connet Ciers 547,428 581,043 76,289 27,854 207,325	178,109 291,401 117,579 46,501	90,014 202,942 1 25,900 73,806 18,672	54,238 42,904 32,295 46,811 41,317		
Manalarian	1000 2003 2004 2004 2004 2004 2005 2004 2005 2006 2006	Horlars APAC Product Chemic Cars M 115,011 194,372 14,995 44,000 51,794 94,393	Marcydes 90,709 53,139 25,868 34,271	Pares 42,643 87,641 31,951 26,720	- SPape 35,323 6,139 2,429 26,335	- 2015.4 Product - Consec Carel 491.273 1.015.790 205.428 196.241 202.913 347.618	141,438 294,043 44,732 12,774 67,439 78,134	134,519 208,128 26,503 90,042 92,553	594pt 172,428 186,992 46,413 4,848 31,819 87,928	Product Chemic Caris 387,438 381,443 361,343 361,344 375,199 27,854 207,323 268,564	178,109 291,421 117,379 60,331 113,310	90,214 202,942 2 25,808 73,806 18,872 84,225	94,238 42,904 32,295 46,811 41,317 22,321		
Manadarea Salar	Tene 2968 2964 2004 2004 2004 2004 2004 2004 2004 20	Harlars APAC Product Chess Cars M 115,011 194,372 14,995 64,000 51,754	64.799 63.139 53.138 25.668	Pares 42,663 47,681 31,951	- SPige 35,323 6,139 2,429	- gHEA Product - Chemic Carel 491,273 L015,790 235,628 196,241 202,613	141,438 294,043 44,732 12,774 67,439 78,134	134,519 208,128 26,503 90,042	594pt 172,428 186,992 46,413 4,848 31,819 87,928	Product Connet Ciers 547,428 581,043 76,289 27,854 207,325	178,109 291,401 117,579 46,501	90,014 202,942 1 25,900 73,806 18,672	94,238 42,904 32,295 46,811 41,317 22,321		
Manalarian	1000 2003 2004 2004 2004 2004 2005 2004 2005 2006 2006	Horlars APAC Product Chemic Cars M 115,011 194,372 14,995 44,000 51,794 94,393	Marcydes 90,709 53,139 25,868 34,271	Pares 42,643 87,641 31,951 26,720	- SPape 35,323 6,139 2,429 26,335	- 2015.4 Product - Consec Carel 491.273 1.015.790 205.428 196.241 202.913 347.618	141,438 294,043 44,732 12,774 67,439 78,134	134,519 208,128 26,503 90,042 92,553	594pt 172,428 186,992 46,413 4,848 31,819 87,928	Product Chemic Caris 387,438 381,443 361,343 361,344 375,199 27,854 207,323 268,564	178,109 291,421 117,379 60,331 113,310	90,214 202,942 2 25,808 73,806 18,872 84,225	94,238 42,904 32,295 46,811 41,317 22,321		
l <u>i mij ĝi</u> Nessuro Sales	1000 2003 2004 2004 2004 2004 2005 2004 2005 2006 2006	Horlars APAC Product Chemic Cars M 115,011 194,372 14,995 44,000 51,794 94,393	Marcydes 90,709 53,139 25,868 34,271	Pares 42,643 87,641 31,951 26,720	- SPape 35,323 6,139 2,429 26,335	- 2015.4 Product - Consec Carel 491.273 1.015.790 205.428 196.241 202.913 347.618	141,438 294,043 44,732 12,774 67,439 78,134	134,519 208,128 26,503 90,042 92,553	594pt 172,428 186,992 46,413 4,848 31,819 87,928	Product Chemic Caris 387,438 381,443 361,343 361,344 375,199 27,854 207,323 268,564	178,109 291,421 117,379 60,331 113,310	90,214 202,942 2 25,808 73,806 18,872 84,225	94,238 42,904 32,295 46,811 41,317 22,321		

Data Mining - mention Weka

🗮 Veka Explorer						
Preprocess Classify Cluster Associate Select attributes Visualize						
Open Ne Open URL Open DB Ger	nerste Edt Seve					
Ren						
Choose Name	Apply					
Current relation	Selected attribute					
Relation: german_oredit Attributes: 21 Instances: 1000 Sum of weights: 1000	Name: credit, history Type: Nominal Missing: 0 (P%) Detroct: 5 Unique: 0 (0%)					
kebtes	No. Label Court Weight					
	1 ro credits/al paid 40 40.0					
All None Invert Pattern	0.01 00 00 00 00 00 00 00 00 00 00 00 00 0					
	3 existing paid \$30 \$30.0					
No. Nane	4 objeyed previously 80 88.0					
1 decking_status	5 otical/other existing (redit [293 293.0					
2 duration						
3 Credit Hetory						
4 purpose 5 predit amount						
6 saving status						
7 enployment						
8 rutalment_connect						
9 personal_status						
10 other parties	Class: class (Nem) Visualize Al					
11 residence_since						
12 property_magnitude	410					
13 009						
14 other payment plans						
15 houring 16 existing_credits						
17 wb						
16 num_dependents						
19 own_telephone	70					
20 Yoreign_worker						
21 dass						
Renove						
Saus						
CK .	Log 💉 X					

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

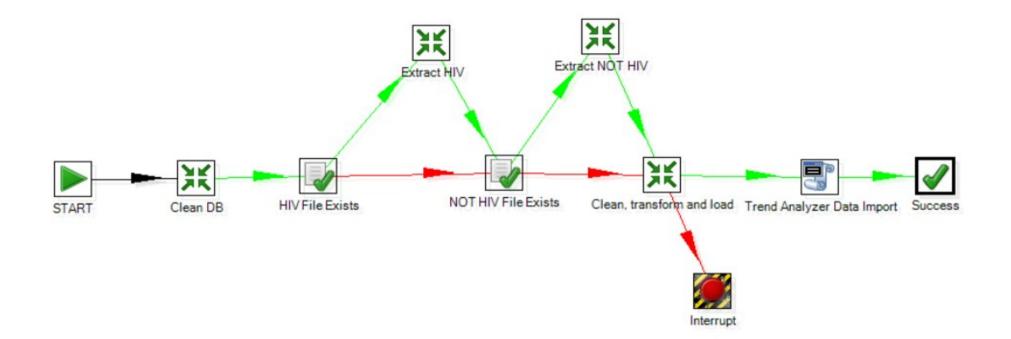
Jobs and Transformations

- 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
Set 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