

Computer. Mensch. Vision.

## Inhalt

- Begriffserklärungen
  - Business Intelligence
  - Data Warehouse
- Open Source Business Intelligence heute
- Komponenten einer BI-Lösung
- Data-Warehouse-Technologie
- Vorstellung der wichtigsten Open-Source-BI-Lösungen

## Begriffserklärung: Business Intelligence

- Deutsch: „Geschäftsanalytik“
- „Analytischer Prozess, der Unternehmens- und Wettbewerbsdaten in handlungsgerechtes Wissen für die Unternehmenssteuerung überführt“

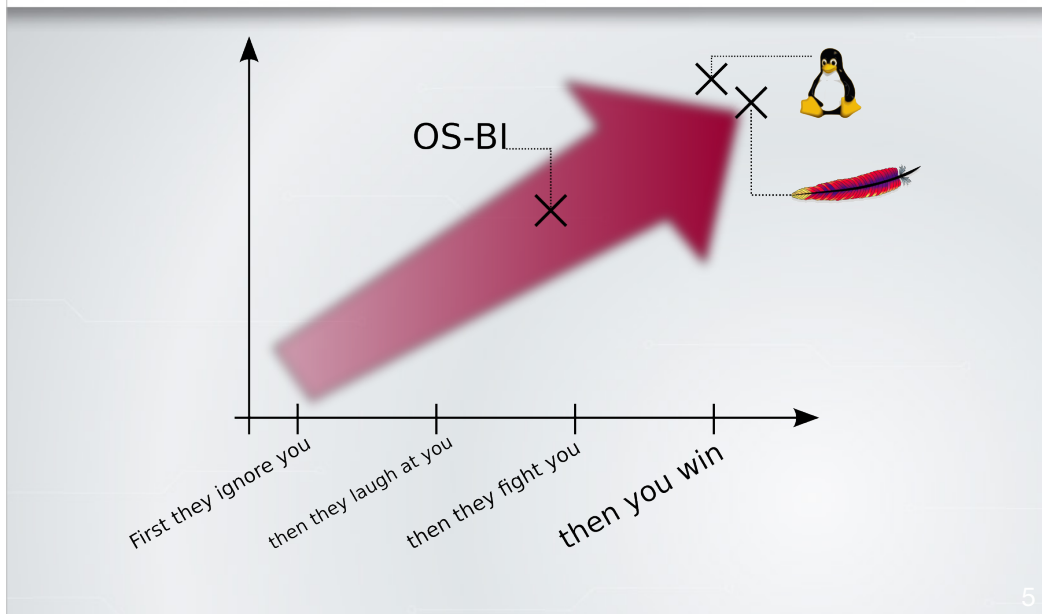
(Prof. Peter Chamoni, 2009)

Verwandter Begriff „Corporate Performance Management“: Methoden, Kennzahlen, Prozesse und Systeme, um die Leistung des Unternehmens zu messen und zu steuern (Gartner Group, 2002)

## Anwendungen von BI

- Darstellung („Was geschieht, was ist geschehen?“):  
Berichte, Dashboards, Diagramme, ...
  - Als Entscheidungsgrundlage, hauptsächlich im operativen Bereich
- Analyse („Warum?“): Entwicklung von Kennzahlen, Verteilungen, Abweichung von Soll oder Vorjahr, ...
- Planung („Was sollten wir tun?“): Festlegung von Zielen (Soll-Zahlen)
- Prognose, Simulation, Risikomanagement, ...

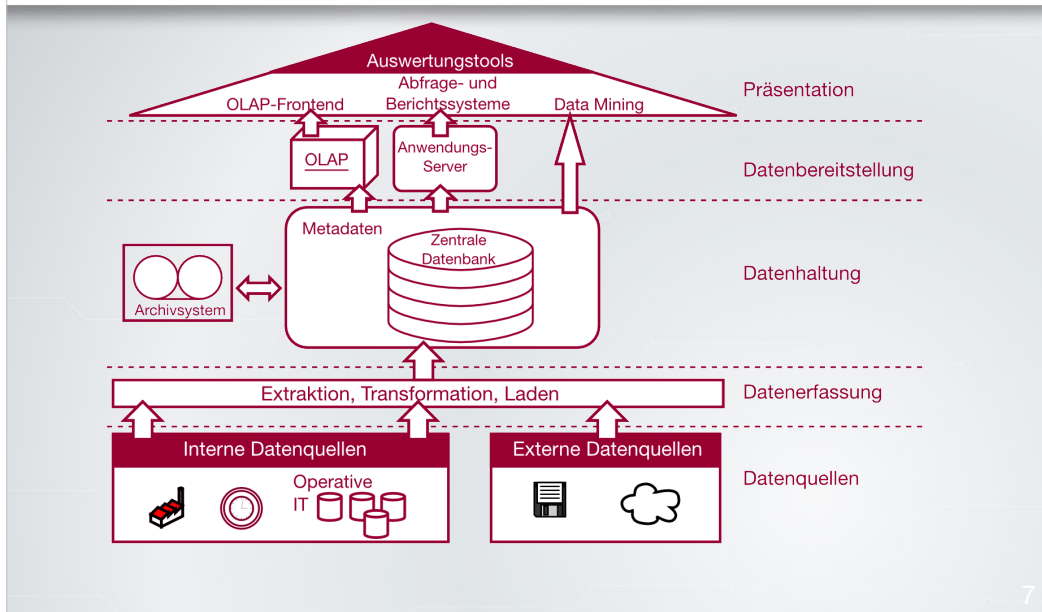
## Open Source Business Intelligence – Aktueller Stand



## Begriffserklärung: Data Warehouse

- Unternehmensweites Konzept
- Einheitliche und konsistente Datenbasis zur Entscheidungsunterstützung
- Getrennt von den operativen Systemen
- An „Dimensionen“ (Themen) ausgerichtet
  - z. B. Kunden, Regionen, Produkte, Zeit
- Dauerhaft
- Zeitlicher Bezug: „versionierte“ Speicherung von Attributen, „Schnappschuß des Unternehmens“
  - z. B. Kunde zieht aus Stadt A in Stadt B um: Umsätze vorm Umzug in Stadt A, danach in B

# Komponenten eines Data Warehouse



## Literatur:

Ralph Kimball, Margy Ross: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling  
Wiley, 2002 (2. Auflage)

## Quelldaten-ETL: Extract, Transform, Load

- Extrahieren aus internen und externen Datenquellen: Datenbanken, Textdateien, Web, ...
- Transform: Standardisierung, Umformung, Säuberung, Berechnung neuer Kennzahlen, ...
- Load: Füllen des Data Warehouse

### Literatur:

Ralph Kimball, Joe Caserta: The Data Warehouse ETL toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data

Wiley, 2004



## Einige ETL-Besonderheiten

- Einfluß auf Quellsysteme ist möglichst zu minimieren
  - Einfache Abfragen
  - Nur das Nötigste
  - Wenn weniger ausgelastet
- „Change Data Capture“: Lückenlose Erfassung gelöschter/geänderter/neuer Datensätze im Quellsystem
  - Trigger, IDs, Änderungsdatum, ...
- Staging- und Präsentationsbereich
  - Staging für interne ETL-Zwecke
  - In den Präsentationsbereich kommen nur aufbereitete Daten

## Datenspeicherung

- Datenbank: Spezielle Anforderungen für Data Warehouse und Business Intelligence
  - Auch spezielle (nicht relationale) Datenbanken in Verwendung
- Metadaten-Schicht
  - Vermittler zwischen technischer Datenbanksicht und Benutzersicht
  - Übersetzt Abfragen auf die technische Ebene, optimiert die Abfragegeschwindigkeit (Cache, Aggregation, ...)

## Datenbankabfragen im DWH

- Tendenziell wesentlich komplexer als in operativen Systemen
  - Operatives System: „Gib mir die Daten von Kunde ID 17“
  - Analytische Abfrage: „Gib mir die Daten von Kunden aus Wien im März und April, die eine Tastatur gekauft haben“
- Kleinerer Ausschnitt der Daten
  - z. B. 50 Attribute von Kunden gespeichert, nur 5 abgefragt

## Optimierungsmethoden für analytische Datenbanken

- Spaltenbasierte (Column-Store-) relationale Datenbanken: Weniger Attribute zum Abfragen; weniger Unterschiede in den Daten, daher besser komprimierbar
- Indizes für häufig abgefragte Kriterien
- Aggregationen: Vorberechnete Summen, z. B. täglich pro Artikel
- Dimensionale (nichtrelationale) Datenbanken

## Präsentation und Analyse

- Berichtswesen (Reporting)
  - Interaktiv (Ad-Hoc; Parametrisierung)
  - Standard-Reporting: E-Mail, periodische Ablage auf Server, ...
- Dashboards/Cockpits: Übersichtsbildschirm
- Online Analytical Processing (OLAP): „Dimensionale“ Analyse („Datenwürfel“)
- Fortgeschrittene Analysemethoden, Data Mining
- Planung, Soll-Ist-Vergleich, Prognose

## Vorstellung: Open-Source-BI-Lösungen

- Datenbank: PostgreSQL
- Gesamtlösung: Pentaho
- ETL: Pentaho Data Integration (Kettle)
- Online Analytical Processing
  - Mondrian: auf relationaler DB
  - PALO: speicherbasierte OLAP-Datenbank
- Data Mining: RapidMiner
- Analyse: R
- ... und die Konkurrenz

## PostgreSQL im Data Warehouse

- Exzellentes, vielseitiges Datenbanksystem für die meisten Anforderungen
  - Nicht auf Data Warehouse spezialisiert
- Tuning für DWH empfehlenswert
  - RAM  $\geq$  Datenbankgröße, wenn möglich
  - work\_mem vergrößern, z. B. 512 MB, um komplexe Abfragen zu beschleunigen
  - effective\_io\_concurrency = Anzahl der „effektiven Festplatten“ (RAID)
  - random\_page\_cost = kaum größer als seq\_page\_cost (z. B. 1.01 und 1.0)
  - effective\_cache\_size richtig setzen (aus top)

15

PostgreSQL: <http://www.postgresql.org/>

## Pentaho BI Server

- US-Firma hat die Entwickler führender Open-Source-BI-Komponenten (in Java) angestellt
- Fügt die Komponenten zu einer Gesamtlösung zusammen
- Community Edition
- Enterprise Edition mit Support und Zusatzfeatures
  - 30-Tage-Demo verfügbar

Pentaho: <http://www.pentaho.com/>



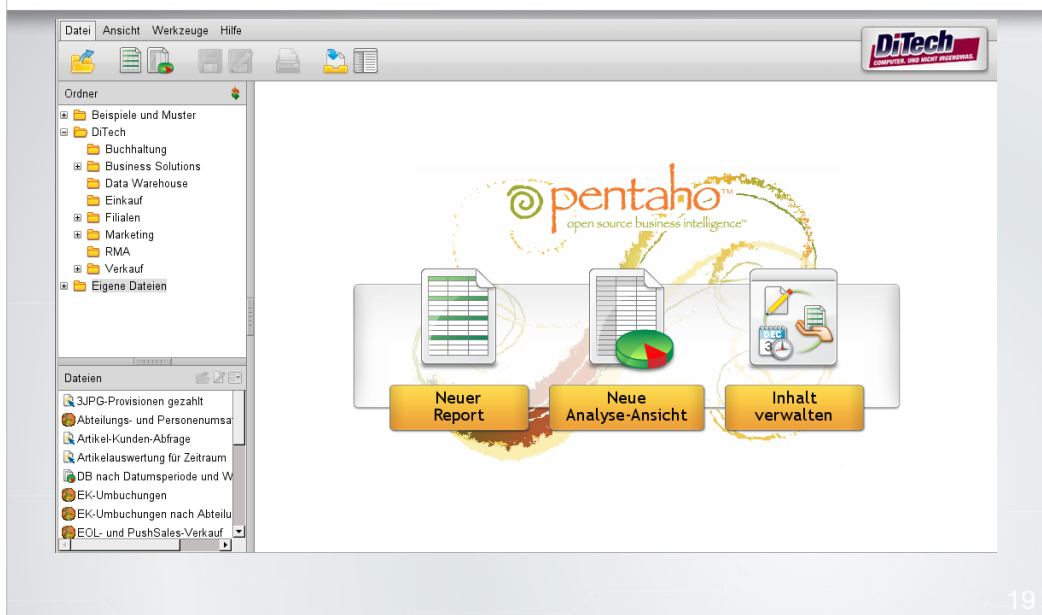
## Pentaho BI Server: Komponenten

- ETL: Pentaho Data Integration (vormals Kettle)
- Berichte und Dashboards: Pentaho Reporting (vormals JFreeReport)
- OLAP: Mondrian (Relationales OLAP), JPivot (Web-Frontend)
- Data Mining: Weka
- Alles zusammengefaßt im Pentaho BI Server unter einer einheitlichen Web-Oberfläche
- Zusätzliche Desktop-Programme für Berichtsdesign, Metadatenerstellung, OLAP-Cube-Gestaltung, Aktionssequenzen

## Pentaho BI Server: Funktionen

- Eigener Arbeitsbereich pro Benutzer
- Explorer-ähnliche Oberfläche
- Webbasierte Ad-Hoc-Abfragen
- Parameterabfrage für Berichte
- Analysen
- Festlegung von Zugriffsrechten
- Festlegung von Zeitplänen für Berichte und Aktionen

## BI Server: Startbildschirm



# Ad-Hoc-Berichterstellung: Geschäftsmodell wählen

Schritt 1: Wählen Sie ein Business Model > Schritt 2: Wählen Sie die Objekte Ihres Berichts > Schritt 3: Passen Sie Ihren Bericht an > Schritt 4: Einstellungen zum Bericht

Wählen Sie ein Business Model

Business Models

Verkaufte Artikel

Eingangsrechnungen

Edit

Add

Delete

Business Model Details

Business View

Filiale

Verkäufer

Kunde

Artikel

Hersteller

Gewicht und Volumen

Preis

Geschäftserfolg

Geizhals-Status

Datum und Zeit

Attribute des Verkaufsvorgangs

Rechnungsposition-Eigenschaften

Lieferant

Eingangsrechnungen

Beschreibung

Einzelne Rechnungspositionen von verkauften Artikeln

Wählen Sie ein Design

Designs

Basic

Fall

Summer

Spring

Pentaho

Winter

Design Details

Thumbnail

Product Sales		New Products	
Product	Sales	Product	Sales
01 01 01	1000000	01 01 01	1000000
01 01 02	1000000	01 01 02	1000000
01 01 03	1000000	01 01 03	1000000
01 01 04	1000000	01 01 04	1000000
01 01 05	1000000	01 01 05	1000000
01 01 06	1000000	01 01 06	1000000
01 01 07	1000000	01 01 07	1000000
01 01 08	1000000	01 01 08	1000000
01 01 09	1000000	01 01 09	1000000
01 01 10	1000000	01 01 10	1000000
01 01 11	1000000	01 01 11	1000000
01 01 12	1000000	01 01 12	1000000
01 01 13	1000000	01 01 13	1000000
01 01 14	1000000	01 01 14	1000000
01 01 15	1000000	01 01 15	1000000
01 01 16	1000000	01 01 16	1000000
01 01 17	1000000	01 01 17	1000000
01 01 18	1000000	01 01 18	1000000
01 01 19	1000000	01 01 19	1000000
01 01 20	1000000	01 01 20	1000000
01 01 21	1000000	01 01 21	1000000
01 01 22	1000000	01 01 22	1000000
01 01 23	1000000	01 01 23	1000000
01 01 24	1000000	01 01 24	1000000
01 01 25	1000000	01 01 25	1000000
01 01 26	1000000	01 01 26	1000000
01 01 27	1000000	01 01 27	1000000
01 01 28	1000000	01 01 28	1000000
01 01 29	1000000	01 01 29	1000000
01 01 30	1000000	01 01 30	1000000
01 01 31	1000000	01 01 31	1000000
01 01 32	1000000	01 01 32	1000000
01 01 33	1000000	01 01 33	1000000
01 01 34	1000000	01 01 34	1000000
01 01 35	1000000	01 01 35	1000000
01 01 36	1000000	01 01 36	1000000
01 01 37	1000000	01 01 37	1000000
01 01 38	1000000	01 01 38	1000000
01 01 39	1000000	01 01 39	1000000
01 01 40	1000000	01 01 40	1000000
01 01 41	1000000	01 01 41	1000000
01 01 42	1000000	01 01 42	1000000
01 01 43	1000000	01 01 43	1000000
01 01 44	1000000	01 01 44	1000000
01 01 45	1000000	01 01 45	1000000
01 01 46	1000000	01 01 46	1000000
01 01 47	1000000	01 01 47	1000000
01 01 48	1000000	01 01 48	1000000
01 01 49	1000000	01 01 49	1000000
01 01 50	1000000	01 01 50	1000000
01 01 51	1000000	01 01 51	1000000
01 01 52	1000000	01 01 52	1000000
01 01 53	1000000	01 01 53	1000000
01 01 54	1000000	01 01 54	1000000
01 01 55	1000000	01 01 55	1000000
01 01 56	1000000	01 01 56	1000000
01 01 57	1000000	01 01 57	1000000
01 01 58	1000000	01 01 58	1000000
01 01 59	1000000	01 01 59	1000000
01 01 60	1000000	01 01 60	1000000
01 01 61	1000000	01 01 61	1000000
01 01 62	1000000	01 01 62	1000000
01 01 63	1000000	01 01 63	1000000
01 01 64	1000000	01 01 64	1000000
01 01 65	1000000	01 01 65	1000000
01 01 66	1000000	01 01 66	1000000
01 01 67	1000000	01 01 67	1000000
01 01 68	1000000	01 01 68	1000000
01 01 69	1000000	01 01 69	1000000
01 01 70	1000000	01 01 70	1000000
01 01 71	1000000	01 01 71	1000000
01 01 72	1000000	01 01 72	1000000
01 01 73	1000000	01 01 73	1000000
01 01 74	1000000	01 01 74	1000000
01 01 75	1000000	01 01 75	1000000
01 01 76	1000000	01 01 76	1000000
01 01 77	1000000	01 01 77	1000000
01 01 78	1000000	01 01 78	1000000
01 01 79	1000000	01 01 79	1000000
01 01 80	1000000	01 01 80	1000000
01 01 81	1000000	01 01 81	1000000
01 01 82	1000000	01 01 82	1000000
01 01 83	1000000	01 01 83	1000000
01 01 84	1000000	01 01 84	1000000
01 01 85	1000000	01 01 85	1000000
01 01 86	1000000	01 01 86	1000000
01 01 87	1000000	01 01 87	1000000
01 01 88	1000000	01 01 88	1000000
01 01 89	1000000	01 01 89	1000000
01 01 90	1000000	01 01 90	1000000
01 01 91	1000000	01 01 91	1000000
01 01 92	1000000	01 01 92	1000000
01 01 93	1000000	01 01 93	1000000
01 01 94	1000000	01 01 94	1000000
01 01 95	1000000	01 01 95	1000000
01 01 96	1000000	01 01 96	1000000
01 01 97	1000000	01 01 97	1000000
01 01 98	1000000	01 01 98	1000000
01 01 99	1000000	01 01 99	1000000
01 01 100	1000000	01 01 100	1000000

Beschreibung

Summer Theme Template

Vorschau als: HTML

Bericht anzeigen

< Zurück

Weiter >

# Ad-Hoc-Berichterstellung: Felder und Gruppierungen

Schritt 1: Wählen Sie ein Business Model Schritt 2: Wählen Sie die Objekte Ihres Berichts Schritt 3: Passen Sie Ihren Bericht an Schritt 4: Einstellungen zum Bericht

**Verfügbare Elemente** ☒ Einzelauswahl

**Kunde**

Kunde-Suchbegriff	Kunde-PLZ	Kunde-Zahlungsart
Kunde-Vorname	Kunde-Ort	Kunde-Mehrwertsteuer
Kunde-Nachname	Kunde-Bundesland	Kunde-Kreditlimit
Kunde-Titel	Kunde-Land	Kunde-Konto gesperrt
Kunde-Firma2	Kunde-Telefon	Kunde-Versichert
Kunde-Firma1	Kunde-Mobilteléfono	Kundennummer
Kunde-VIP	Kunde-Telefax	Anmeldedatum
Kundenart	Kunde-E-Mail	
Kunde-Straße	Kunde-Lieferart	

**Artikel**

Artikelname	Artikel im Onlineshop	Artikel-MwSt-Faktor
Artikelnummer	Artikel im Offliner	Artikel-MwSt
Artikelkategorie	End of life	Artikel-Verkaufseinheit
Artikelkategorie-Untergruppe	Artikel gesperrt	
Artikelkategorie-Unter-Untergruppe	Artikel-Garantie	

**Hersteller**

Herstellername	Herstellernummer
----------------	------------------

**Gewicht und Volumen**

Gewicht-Summe	Stückzahl-Summe	Rechnungen-Anzahl
---------------	-----------------	-------------------

**Preis**

Verkaufspreis-Summe	Korrigierter-Verkaufspreis-Summe
---------------------	----------------------------------

**Gewählte Elemente**

**Gruppen**

Level 1  
Kundenart

Level 2  
[Element hier ablegen]

Level 3  
[Element hier ablegen]

**Details**

Stückzahl-Summe

**Filter**

Jahr

Vorschau als: HTML Bericht anzeigen < Zurück Weiter >

## ETL mit Pentaho Data Integration (Kettle)

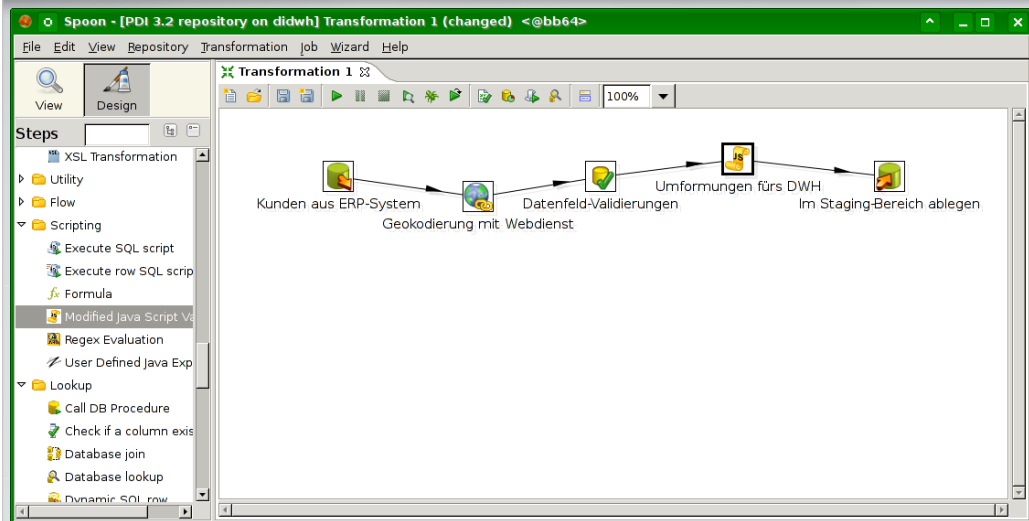
- Organisation der ETL-Aufgaben in „Jobs“ (Aufgabensteuerung) und „Transformationen“ (Arbeitsschritte)
- Java-GUI: Spoon zur Gestaltung der Jobs und Transformationen
- Ausführungskomponenten: Kitchen (für Jobs) und Pan (für Transformationen)
- Webbasierte Steuerungsoberfläche: Carte
- Ablage der ETL-Prozesse in XML-Dateien oder Datenbank-Repository
- Ausführung: Interpretation der erstellten Modelle

22

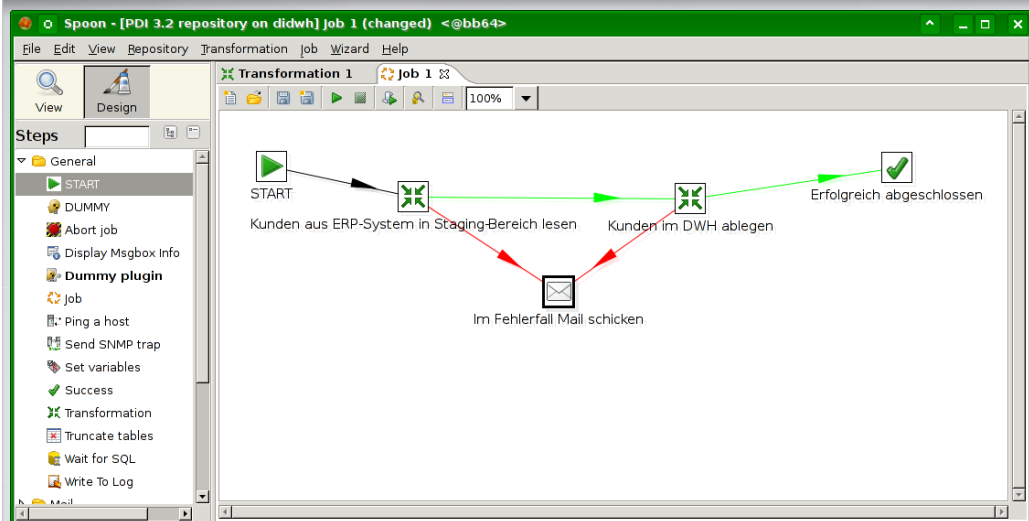
Homepage von Pentaho Data Integration:

<http://kettle.pentaho.org/>

## Beispieltransformation in PDI



## Beispieljob in PDI





## Transformationselemente in PDI

- Input: Datenbank, Textdatei, HTTP, RSS, Systeminfo, LDAP, ...
- Output: Datenbank, OLAP-System, Textdatei, Excel, Löschen (Datei, Datensätze, usw.) ...
- Veränderung: Berechnung, Zähler, Reguläre Ausdrücke, Scripting, ...
- Überprüfung: Filter nach fast beliebigen Kriterien, Gültigkeitsregeln, Formatvalidierungen, ...
- Datenströme verknüpfen, aneinanderhängen, teilen, für Lookups verwenden, ...

## OLAP mit Mondrian und JPivot

- Mondrian: Software-Schicht zwischen Datenbank und Präsentationsebene
  - Interpretiert MDX (MultiDimensional eXpressions)
  - Setzt Abfragen für die relationale Datenbank um
  - Kann Aggregationstabellen verwenden
  - Speichert Abfrageergebnisse für eine gewisse Zeit, die selben Daten müssen nicht in kurzen Abständen hintereinander abgefragt werden
- JPivot: Web-Oberfläche für Abfragegestaltung
- Schema Workbench: GUI für die Erstellung der XML-Datei, die Mondrian steuert

26

Mondrian-Homepage:

<http://mondrian.pentaho.org>

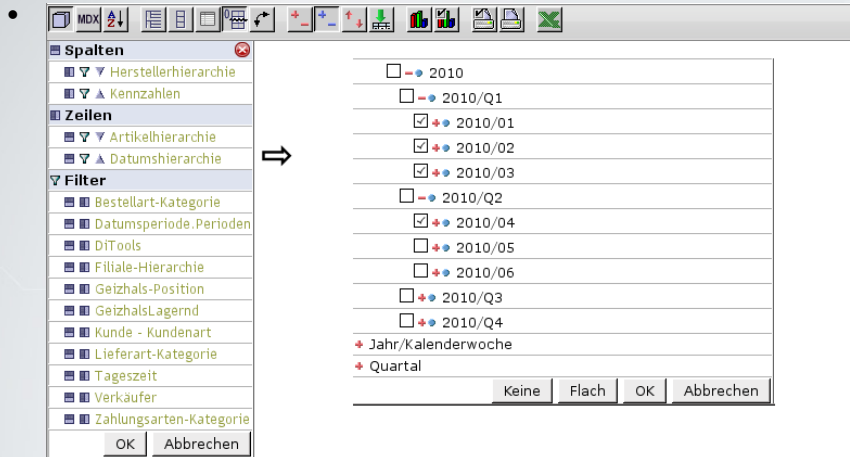
## Mondrian: JPivot-Analyse

- In Pentaho-Oberfläche integriert oder eigenständig
- Anzeige der vorgegebenen Felder und von Berechnungsergebnissen
- Interaktive Auswahl von Dimensionen, Kennzahlen und Filtern
- Zugriff auf die MDX-Abfrage für Power-User
- Simple Diagramme können eingeblendet werden
- „Drill-down“ zu den Originaldaten

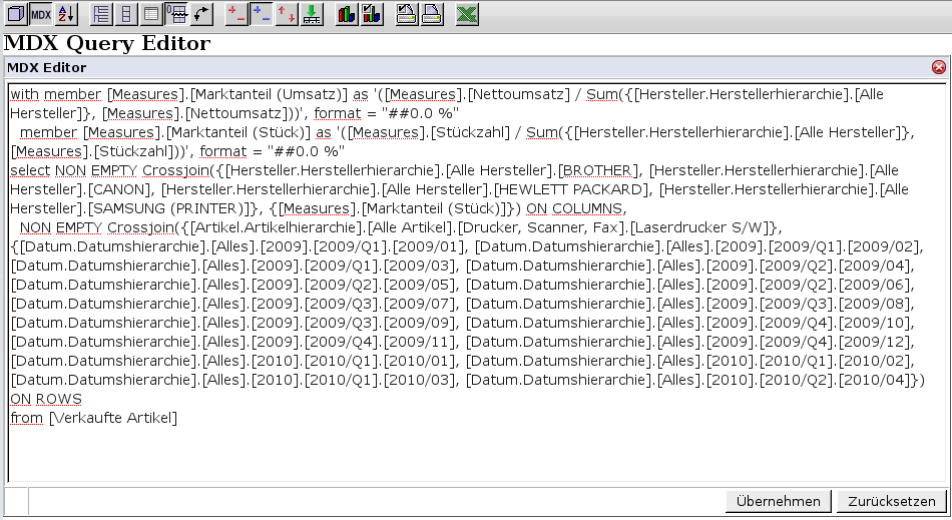
## JPivot-Beispiel: Berechnete Felder

		Herstellerhierarchie			
		+	+	+	+
		BROTHER	CANON	HEWLETT PACKARD	SAMSUNG (PRINTER)
		Kennzahlen	Kennzahlen	Kennzahlen	Kennzahlen
Artikelhierarchie	Datumshierarchie	● Marktanteil (Stück)	● Marktanteil (Stück)	● Marktanteil (Stück)	● Marktanteil (Stück)
+ Laserdrucker S/W	+ 2009/01	43,0 %	8,9 %	48,0 %	
	+ 2009/02	52,7 %	3,2 %	44,1 %	
	+ 2009/03	48,1 %	2,7 %	49,2 %	
	+ 2009/04	51,9 %	4,9 %	43,2 %	
	+ 2009/05	48,6 %	3,1 %	46,7 %	0,4 %
	+ 2009/06	55,2 %	2,8 %	37,5 %	4,5 %
	+ 2009/07	43,2 %	3,9 %	51,4 %	1,6 %
	+ 2009/08	47,7 %	4,3 %	47,7 %	0,4 %
	+ 2009/09	48,3 %	4,2 %	46,1 %	1,4 %
	+ 2009/10	57,0 %	4,7 %	37,4 %	0,9 %
	+ 2009/11	49,2 %	5,1 %	44,1 %	1,7 %
	+ 2009/12	50,5 %	5,1 %	40,3 %	4,0 %

## JPivot: Auswahl von Dimensionselementen



## JPivot: MDX-Ansicht eingeschaltet



**MDX Query Editor**

**MDX Editor**

```
with member [Measures].[Marktanteil (Umsatz)] as '([Measures].[Nettoumsatz] / Sum({[Hersteller.Herstellerhierarchie].[Alle Hersteller]}, [Measures].[Nettoumsatz]))', format = "##0.0 %"
member [Measures].[Marktanteil (Stück)] as '([Measures].[Stückzahl] / Sum({[Hersteller.Herstellerhierarchie].[Alle Hersteller]}, [Measures].[Stückzahl]))', format = "##0.0 %"
select NON EMPTY Crossjoin({[Hersteller.Herstellerhierarchie].[Alle Hersteller].[BROTHER], [Hersteller.Herstellerhierarchie].[Alle Hersteller].[CANON], [Hersteller.Herstellerhierarchie].[Alle Hersteller].[HEWLETT PACKARD], [Hersteller.Herstellerhierarchie].[Alle Hersteller].[SAMSUNG (PRINTER)]}, {[Measures].[Marktanteil (Stück)]}) ON COLUMNS,
NON EMPTY Crossjoin({[Artikel.Artikelhierarchie].[Alle Artikel].[Drucker, Scanner, Fax].[Laserdrucker S/W]},
{[Datum.Datumshierarchie].[Alles].[2009].[2009/Q1].[2009/01], [Datum.Datumshierarchie].[Alles].[2009].[2009/Q1].[2009/02],
[Datum.Datumshierarchie].[Alles].[2009].[2009/Q1].[2009/03], [Datum.Datumshierarchie].[Alles].[2009].[2009/Q2].[2009/04],
[Datum.Datumshierarchie].[Alles].[2009].[2009/Q2].[2009/05], [Datum.Datumshierarchie].[Alles].[2009].[2009/Q2].[2009/06],
[Datum.Datumshierarchie].[Alles].[2009].[2009/Q3].[2009/07], [Datum.Datumshierarchie].[Alles].[2009].[2009/Q3].[2009/08],
[Datum.Datumshierarchie].[Alles].[2009].[2009/Q3].[2009/09], [Datum.Datumshierarchie].[Alles].[2009].[2009/Q4].[2009/10],
[Datum.Datumshierarchie].[Alles].[2009].[2009/Q4].[2009/11], [Datum.Datumshierarchie].[Alles].[2009].[2009/Q4].[2009/12],
[Datum.Datumshierarchie].[Alles].[2010].[2010/Q1].[2010/01], [Datum.Datumshierarchie].[Alles].[2010].[2010/Q1].[2010/02],
[Datum.Datumshierarchie].[Alles].[2010].[2010/Q1].[2010/03], [Datum.Datumshierarchie].[Alles].[2010].[2010/Q2].[2010/04]})
ON ROWS
from [Verkaufte Artikel]
```

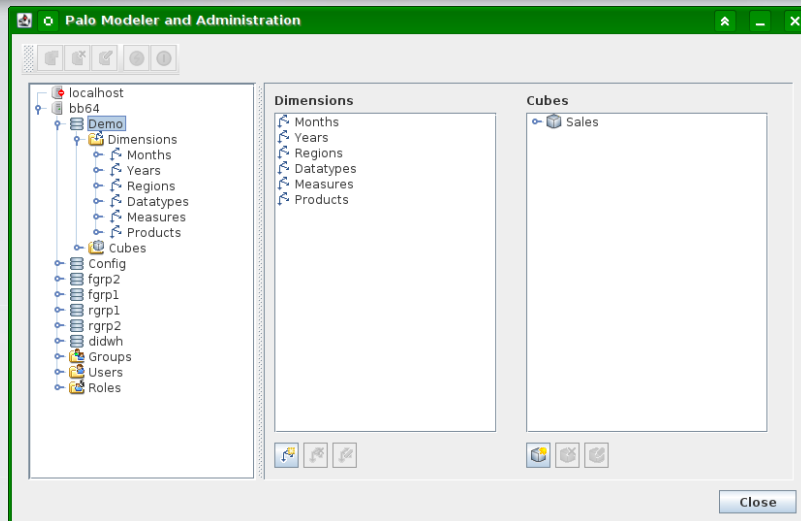
Übernehmen Zurücksetzen

## OLAP mit PALO und OpenOffice.org (oder Excel)

- PALO: In-Memory-OLAP-Datenbank von Jedox
- Läuft eigenständig oder eingebettet in die webbasierte Gesamtlösung „Palo Suite“
  - Web-Frontend für Tabellenkalkulationsdokumente
- Zellen in der Tabellenkalkulation mit dem Server verbunden, Änderungen in Echtzeit
- Gut geeignet für Planung, Analyse, Berichtswesen, Dashboards
- Für die meisten Anwender intuitiv bedienbar, Office-Kenntnisse reichen

PALO-Homepage: <http://www.palo.net/de/>

## Datendefinition aus OoO oder Excel heraus





## Kreuztabelle aus den gewünschten Daten

Palo-Demo.ods - OpenOffice.org Calc

Datei Bearbeiten Ansicht Einfügen Format Extras Daten Palo Fenster Hilfe

Arial 10 B I U

A9 =PALO.ENAME(\$B\$1:"Products";"All Products";3;"All Products";"")

	A	B	C	D	E	F	G	H
1	Database	bb64/Demo						
2	Cube	Sales						
3	Months	Year						
4	Years	All Years						
5	Datatypes	All Datatypes						
6								
7	Edit	Europe	West	East	South	North		
8		Units	Units	Units	Units	Units		
9	All Products	45.367.531,86	22.347.211,87	8.421.296,72	9.183.122,18	5.415.901,08		
10	Stationary PC's	17.412.677,92	8.651.530,23	3.194.631,23	3.482.943,30	2.083.573,16		
11	Portable PC's	13.791.648,88	6.836.360,34	2.447.396,35	2.894.334,93	1.613.557,26		
12	Monitors	12.483.056,99	6.000.591,20	2.463.630,22	2.471.080,72	1.547.754,85		
13	Peripherals	1.680.148,06	858.730,10	315.638,91	334.763,24	171.015,81		
14								
15								
16								
17								
18								

Tabelle1 / Tabelle2 / Tabelle3

Tabelle 1 / 3 Standard STD Summe=0 110%

## Hierarchiestufe aufgeklappt

Palo-Demo.ods - OpenOffice.org Calc

Datei Bearbeiten Ansicht Einfügen Format Extras Daten Palo Fenster Hilfe

Arial 10

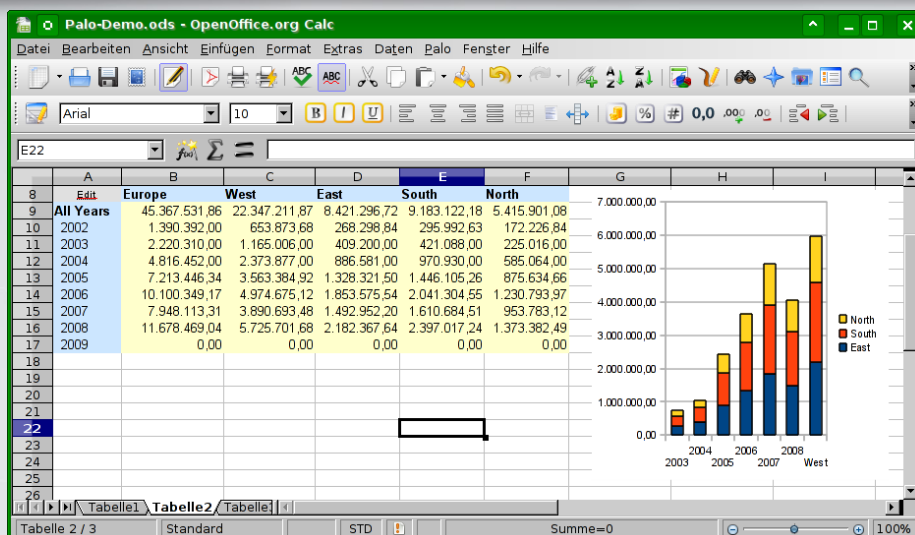
A10 =PALO.ENAME(\$B\$1;"Products";"Stationary PC's";3;"All Products(Stationary PC's";"")

	A	B	C	D	E	F	G
7		Europe	West	East	South	North	
8	Edit	Units	Units	Units	Units	Units	
9	All Products	45.367.531,86	22.347.211,87	8.421.296,72	9.183.122,18	5.415.901,08	
10	Stationary PC's	17.412.677,92	8.651.530,23	3.194.631,23	3.482.943,30	2.083.573,16	
11	Desktop L	3.001.555,99	1.596.125,51	496.608,81	553.679,09	355.142,58	
12	Desktop Pro	2.875.615,00	1.417.507,02	523.892,83	582.185,90	352.029,25	
13	Desktop Pro XL	2.272.589,09	1.097.280,47	448.307,49	415.223,92	311.777,21	
14	Desktop High XL	1.434.992,59	681.191,78	272.915,64	289.900,66	190.984,51	
15	Desktop High XQ	1.080.593,04	526.581,41	222.503,78	190.515,41	140.992,43	
16	Server Power XC	1.698.287,39	787.250,84	330.044,18	380.437,85	200.554,52	
17	Server Power TT	1.274.864,68	661.671,31	231.364,46	271.375,35	110.453,55	
18	Server Dual C	1.419.022,63	695.117,01	249.794,98	284.721,39	189.389,24	
19	Server Dual XC	1.154.865,65	612.267,49	215.092,36	231.903,33	95.602,47	
20	Server Lion RX	1.200.291,87	576.537,38	204.106,70	283.000,39	136.647,40	
21	Portable PC's	13.791.648,88	6.836.360,34	2.447.396,35	2.894.334,93	1.613.557,26	
22	Monitors	12.483.056,99	6.000.591,20	2.463.630,22	2.471.080,72	1.547.754,85	
23	Peripherals	1.680.148,06	858.730,10	315.638,91	334.763,24	171.015,81	

Tabelle1 / Tabelle2 / Tabelle3

Tabelle 1 / 3 Standard STD Summe=0 110%

# Diagramm



35

RapidMiner-Homepage

## Data Mining ○

- „Maschinelles Lernen“
- „Knowledge Discovery in Databases“
- Idee: Auf Basis bisheriger Beobachtungen ein Modell für zukünftige Ereignisse erstellen
- Viele Anwendungsmöglichkeiten
  - Kundenverhalten: Abwanderung, Betrug
  - Automatische Gruppierung von Artikeln, Kunden, ...
  - Textklassifizierung (Spam/Nicht-Spam; Themengebiet)
- Morgen um 10:00 Vortrag von Dr. Alexander K. Seewald

## Data Mining mit RapidMiner

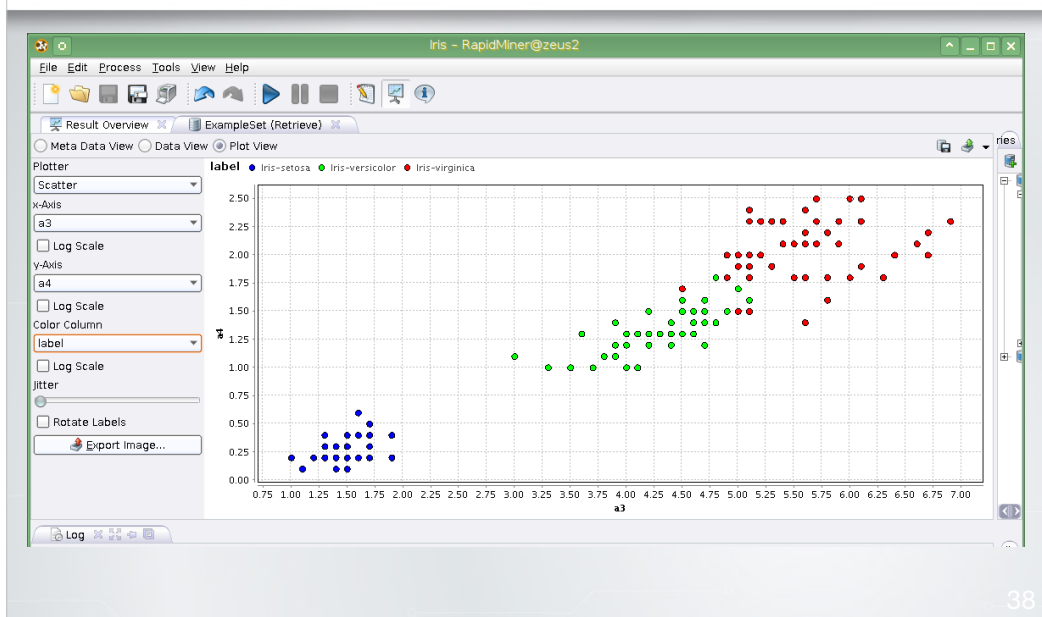
- Vormals „YALE“, aus Uni-Projekt entstanden
- Open-Source- und erweiterte kommerzielle Version erhältlich
- Viele Lernalgorithmen enthalten, weitere nachrüstbar
- Erweiterungen für Berichte, Text-Mining usw.
- Viele grafische Darstellungsformen zur Erkennung von Zusammenhängen in den Daten
- Prozesse in XML-Dateien oder Repository gespeichert, dadurch leicht automatisierbar

37

RapidMiner-Homepage:

<http://www.rapidminer.com/>

## Visualisierung eines Datensatzes

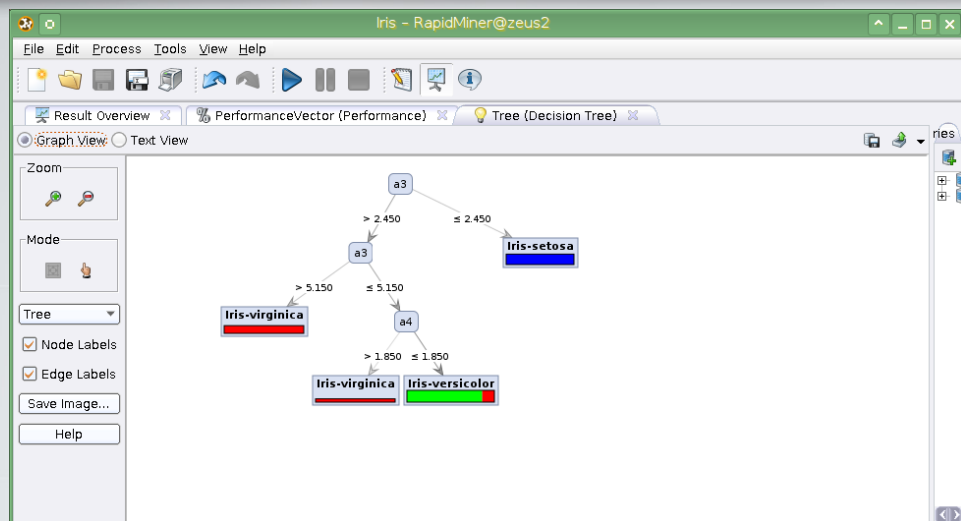


„Iris“-Datensatz von R. A. Fischer, 1936

- a1: Kelchblatt-Länge in cm
- a2: Kelchblatt-Breite in cm
- a3: Blütenblatt-Länge in cm
- a4: Blütenblatt-Breite in cm

Jeweils 50 Blüten der Arten Iris Setosa, Iris Versicolor, Iris Virginica

## Ermittelter Entscheidungsbaum



# Modellqualität

Iris - RapidMiner@zeus2

File Edit Process Tools View Help

Result Overview PerformanceVector (Performance) Tree (Decision Tree)

Table / Plot View Text View

Criterion Selector

- accuracy
- kappa

Table View Plot View

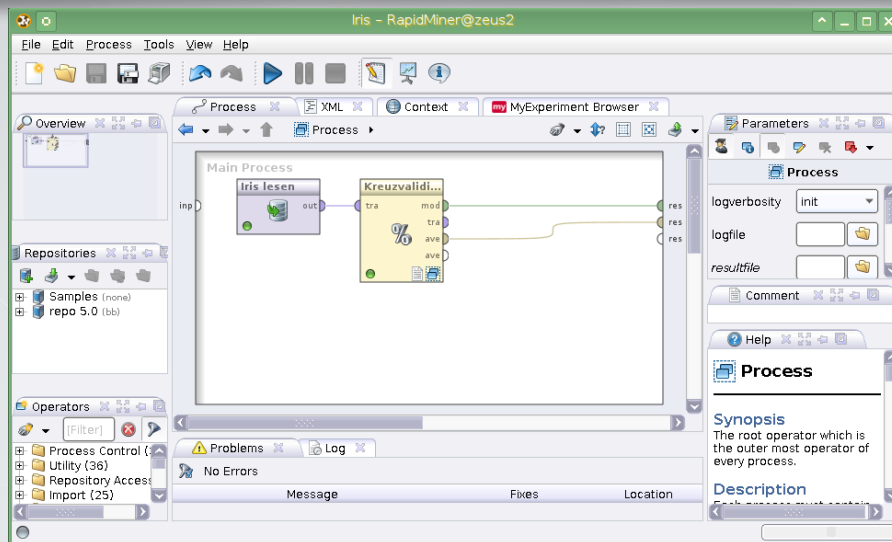
**accuracy: 93.33% +/- 5.16% (mikro: 93.33%)**

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	50	0	0	100.00%
pred. Iris-versicolor	0	47	7	87.04%
pred. Iris-virginica	0	3	43	93.48%
class recall	100.00%	94.00%	86.00%	

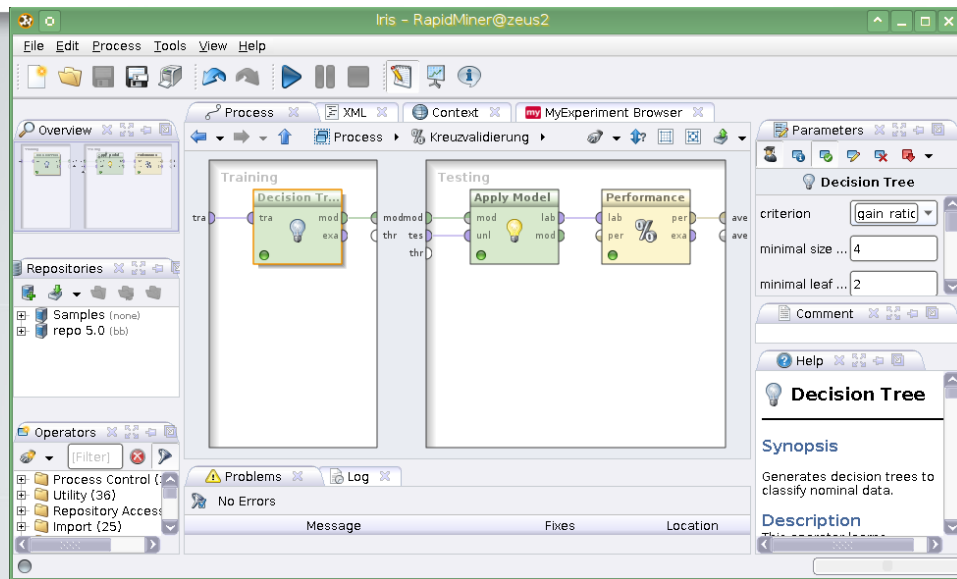
40



# Data-Mining-Prozess in RapidMiner – 1



## Data-Mining-Prozess in RapidMiner – 2



## Data Mining in Pentaho mit Weka

- Vorgehensweise bei der Modellerstellung ähnlich wie bei RapidMiner
- Gespeichertes Modell kann in Pentaho Data Integration angewendet werden
  - z. B. Automatische Kundensegmentierung
- Verschiedene Oberflächen:
  - Explorer: Datensatz untersuchen, visualisieren, klassifizieren
  - Experimenter: Prozesse mit unterschiedlichen Verfahren und Parametern ausführen, testen und vergleichen
  - Knowledge Flow: Komplexe Prozesse erstellen, ähnlich wie in RapidMiner

43

Weka-Homepage bei Pentaho:

<http://weka.pentaho.org/>

Literatur:

Ian H. Witten, Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)

Morgan Kaufmann, 2005

# Weka: Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Filter

Choose

None

Apply

Current relation

Relation: weather

Instances: 14

Attributes: 5

Attributes

All

None

Invert

Pattern

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Remove

Selected attribute

Name: temperature

Missing: 0 (0%)

Distinct: 12

Type: Numeric

Unique: 10 (71%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

Class: play (Nom)

Visualize All

Status

OK

Log

x 0

44



## Analyse mit R

- R ist eine Programmiersprache und Umgebung für statistische Aufgaben
  - Ursprünglich eine Nachimplementierung der Sprache „S“ (kommerziell: S-Plus)
  - Heute Standard in vielen Forschungsfeldern, Unis, ...
- Befehlszeilenorientiert, verschiedene GUIs für häufige Aufgaben verfügbar
- Vielfältige Visualisierungsmöglichkeiten
- R-Schnittstelle in PostgreSQL ermöglicht Ausführung direkt in der Datenbank

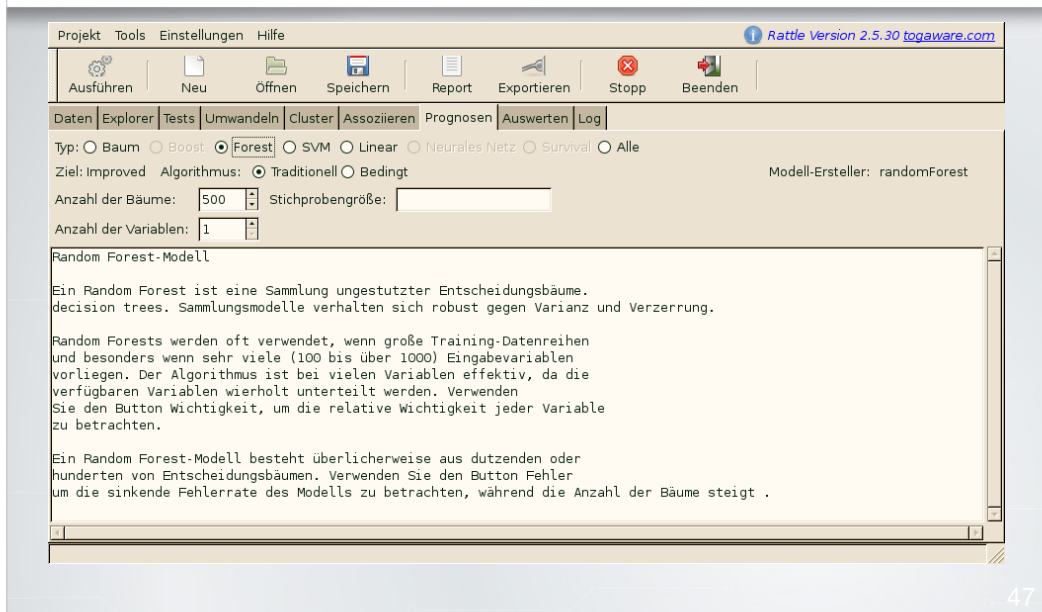
46

Homepage des R-Projekts: <http://www.r-project.org/>

PL/R – R Procedural Language for PostgreSQL:

<http://www.joeconway.com/plr/>

# Data-Mining-Oberfläche für R: Rattle

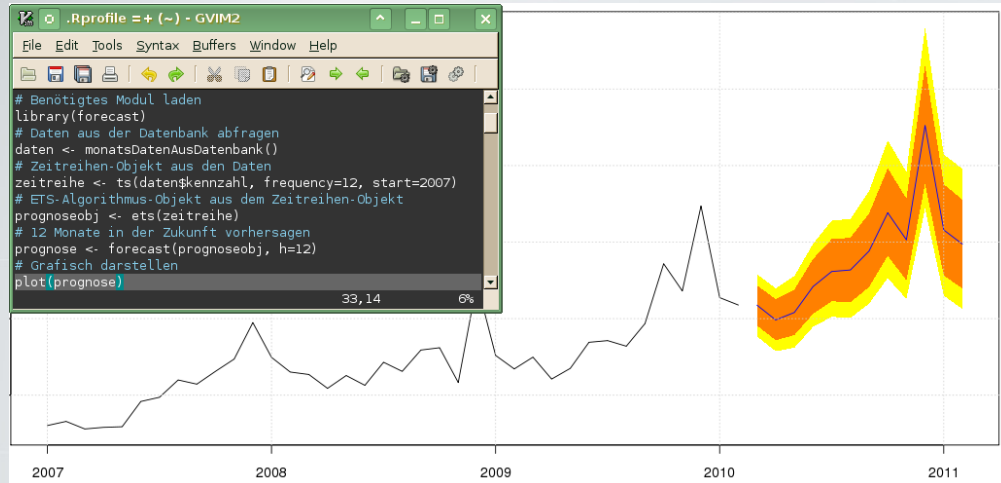


47

Rattle: Cross Platform GUI for Data Mining using R

<http://rattle.togaware.com/>

## Zeitreihenprognose in R mit dem Modul „forecast“



48

Forecast-Package für R:

<http://robjhyndman.com/software/forecast>



## Alternative relationale Datenbanken

- Spaltenbasierte Varianten verbreiteter Datenbanken (alle kommerziell, eingeschränkte Version gratis)
  - Greenplum (basiert auf PostgreSQL)
  - InfoBright (basiert auf MySQL)
  - InfiniDB (basiert auf MySQL)
- Eigenständige Column-Store-Datenbanken
  - LucidDB
  - MonetDB
- Geschwindigkeitsvorteile bei Abfragen
- Nachteile bei Flexibilität, Funktionalität

49

<http://www.greenplum.com/>

<http://www.infobright.com/>

<http://www.infinidb.org/>

<http://www.luciddb.org/>

<http://monetdb.cwi.nl/>

## Alternative ETL-Lösung: Talend Open Studio

- Eclipse-basierte Oberfläche
- Codegenerierung in Perl oder Java für Flexibilität bei der Installation am Zielsystem
- Andere Lösungen von Talend:
  - Talend Open Profiler: Übersicht über Daten gewinnen, Datenqualität überprüfen usw.
  - Talend Master Data Management: Modellierung aller Datenbanken des Unternehmens, Erkennung von Redundanzen, Inkohärenzen, Datenqualitätsproblemen usw.
- Kommerzielle Version erhältlich

50

Talend Open Studio: <http://www.talend.com/>

## Alternativer BI-Server: Jaspersoft BI Suite

- ETL ist gebrandetes Talend
- Webbasierte Portallösung für Dashboards, Berichte, Zeitplanung etc.
- OLAP ist Mondrian
- Kein integriertes Data Mining
- Enterprise- und Professional-Versionen verfügbar;  
Open-Source-Lösung mit deutlich weniger Features

<http://www.jaspersoft.com/>

## Alternativer BI-Server: SpagoBI

- Komplette Open Source, keine kostenpflichtige Enterprise-Version
- Integriert wie JasperSoft Talend für ETL, Mondrian für OLAP, zusätzlich PALO
- Stark im Bereich geografischer Auswertungen
- Service, Support von der Entwicklerfirma

<http://www.spagoworld.org/>

## Der Weg zur eigenen Business-Intelligence-Lösung

- Selber machen
- BI-Appliance von DiTech
  - Als betreuter Server in Ihrem Unternehmen  
oder
  - als virtueller Server im DiTech-Data-Center
  - Halbautomatischer Datenimport aus Ihren Daten
  - Systembetreuung, Performance-Überwachung durch DiTech
  - BI-Beratung, -Schulung und -Entwicklung als Zusatzpaket



COMPUTER. UND NICHT IRGENDWAS.

Fragen?

Balázs Bárány  
DiTech GmbH  
Data Warehouse

E-Mail: [bb@ditech.at](mailto:bb@ditech.at)