Zukunft des digitalen Erbes?

2.-4. März 2017
Karl-Franzens-Universität Graz

Digitale Bibliothek - 7. Konferenz

Workshop

Datenintegration mit Open Source Werkzeugen

Pentaho Data Integration - Kettle

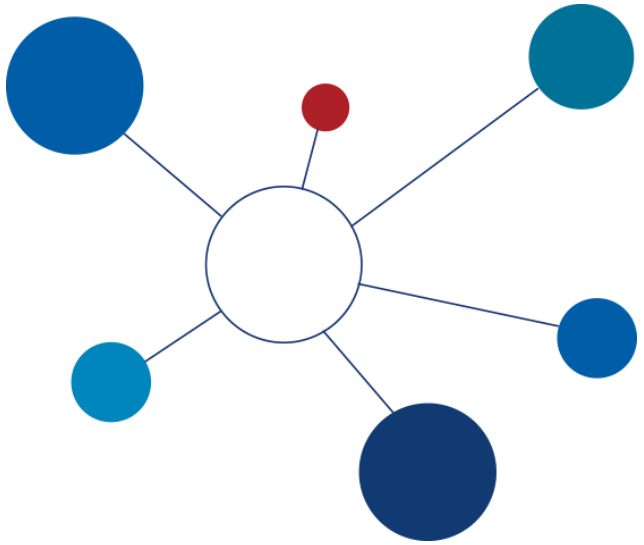
Prof. Dr. Klaus-Georg Deck
Duale Hochschule Baden-Württemberg
D-74821 Mosbach

Inhalt

- Datenintegration
- Pentaho Data Integration (PDI)
  - Installation
  - Komponenten: Steps – Transformationen – Jobs
  - Beispiele – Live Demo
  - Funktionsumfang (Auswahl)
- Diskussion

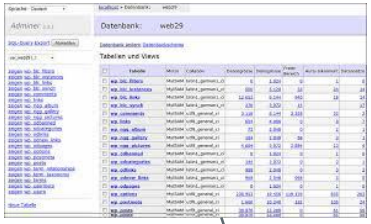
Datenintegration: Zusammenführen von Daten

- Heterogene Herkunft
  - Datenquellen
  - Datenformat
  - Granularität
  - Aggregationsniveau
  - Zeitlicher Bezug
  - Qualität
- Ziele
  - Breite Datenbasis/Umfang
  - Hohe Datenqualität
  - Homogene Datenformate
  - Historisierte Information
  - Unabhängige Verfügbarkeit



Datenintegration: Szenarien

- ETL (Extract-Transform-Load) Prozess in einem Data Warehouse
  - Datenintegration aus verschiedenen Quellsystemen (z.B. operative Systeme)
  - Automatisiertes inkrementelles Laden (wöchentlich/täglich)
  - Qualitätssicherung (Datenbereinigung mit Feedback an Quellsysteme)
- Qualitätssicherung und Datenbereinigung
  - Erkennen von Redundanzen und Duplikaten
  - Vervollständigung von Daten (missings)
  - Ähnlichkeitsanalysen/semantische Integration
- Migration (A $\rightarrow$ B)
- Aufbau eines neuen Systems (A + B + C $\rightarrow$ D)
- Datenextraktion für andere Tools (Excel, RapidMiner, Neo4J)
- Archivierung/Historisierung



ETL: Extrahieren – Transformieren – Laden

PDI

Big Data

Social Media

ETL: Extrahieren – Transformieren – Laden

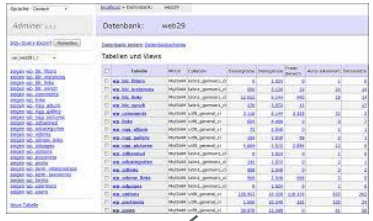
PDI

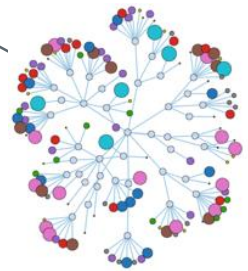
Daten-

- Transformation
- Plausibilitäts- und Konsistenzprüfung
- Merge/Abgleich/Ergänzung
- Aufbereitung/Vorberechnung
- ...

ETL: Extrahieren – Transformieren – Laden

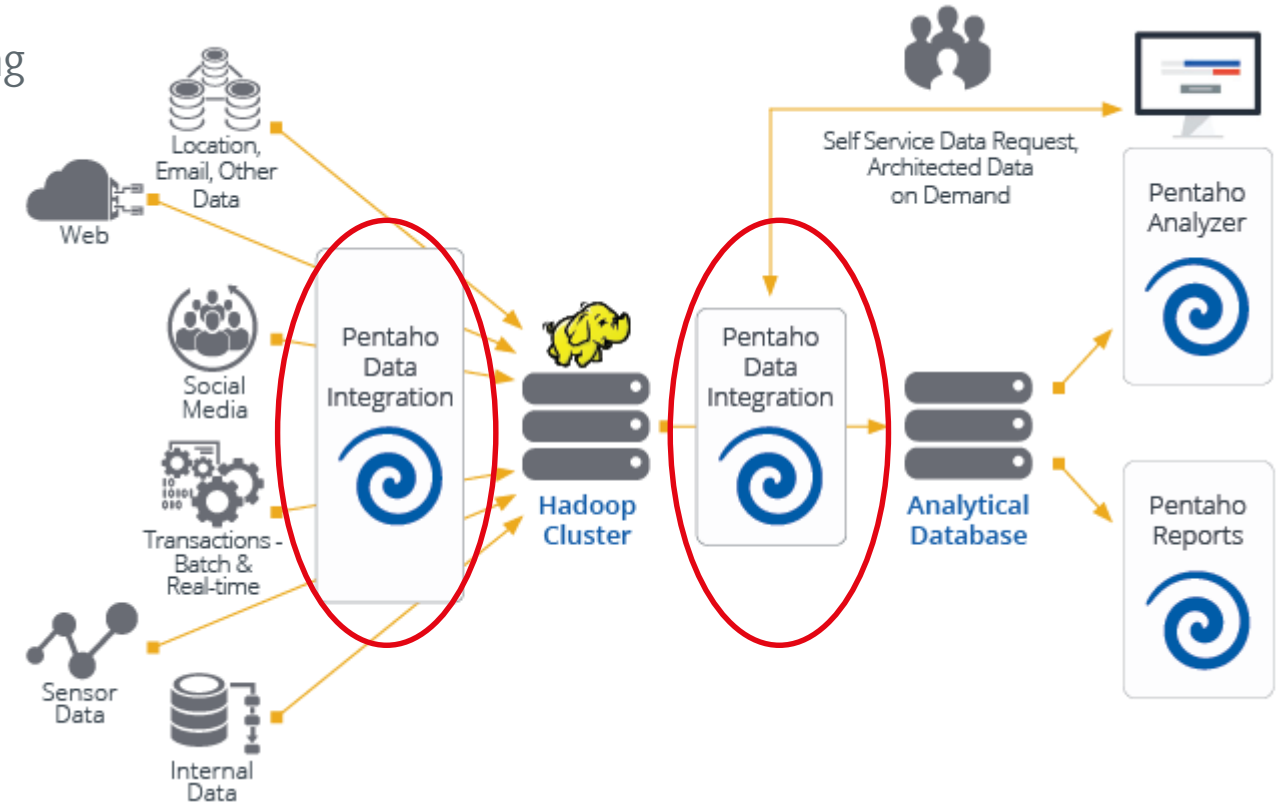
PDI





Pentaho Data Integration (PDI)

- Pentaho: Business Intelligence (BI) Suite
  - Datenintegration / ETL
  - Reporting
  - Analyse
  - Data mining



Pentaho Data Integration

Komponente der ***Pentaho BI Suite***

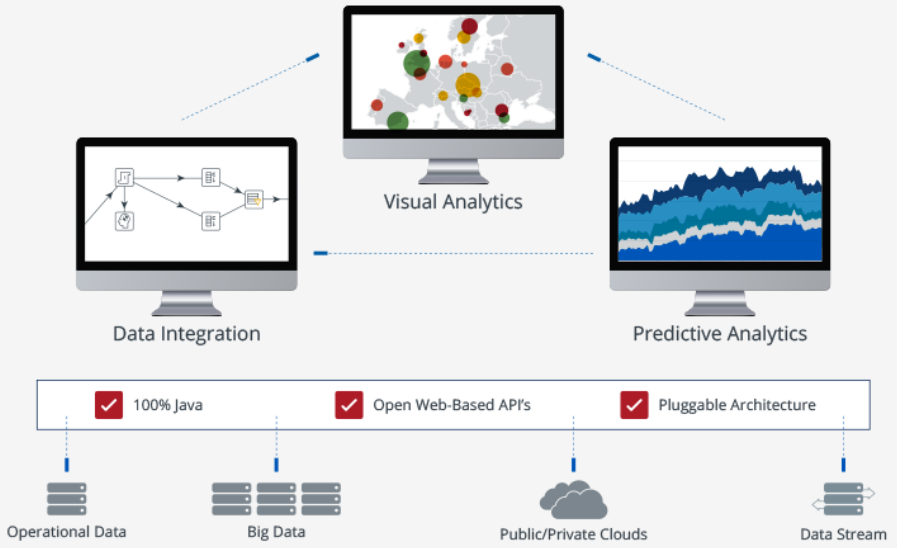
pentaho
A Hitachi Group Company

Download Support Contact Us PentahoWorld Sites English

Product Solutions Services Customers Partners Resources **Try Pentaho**

Any Analytics, Any Data, Simplified

Pentaho addresses the barriers that block your organization's ability to get value from all your data. Our platform simplifies preparing and blending any data and includes a spectrum of tools to easily analyze, visualize, explore, report and predict. Open, embeddable and extensible, Pentaho is architected to ensure that each member of your team — from developers to business users — can easily translate data into value.



How do you want to use the platform?

BIG DATA
Accelerate value with Hadoop, NoSQL, and other big data

DATA INTEGRATION
Access, manage and blend any data from any source

EMBEDDING ANALYTICS
Seamlessly embed a full suite of custom analytics

BUSINESS ANALYTICS
Turn data into insights and make better decisions

Pentaho Data Integration (PDI)

- Lizenzmodelle
  - Community Edition (community.pentaho.com)
    Breite Kernfunktionalität
    Ausreichend für fast alle (auch kommerziellen) PDI-Szenarien
    Freie Dokumentation
    Support via Foren

  - Enterprise Edition (pentaho.com)
    Kostenpflichtig
    Erweiterter Funktionsumfang
    Professioneller Support

Pentaho Data Integration (PDI)

- KETTLE

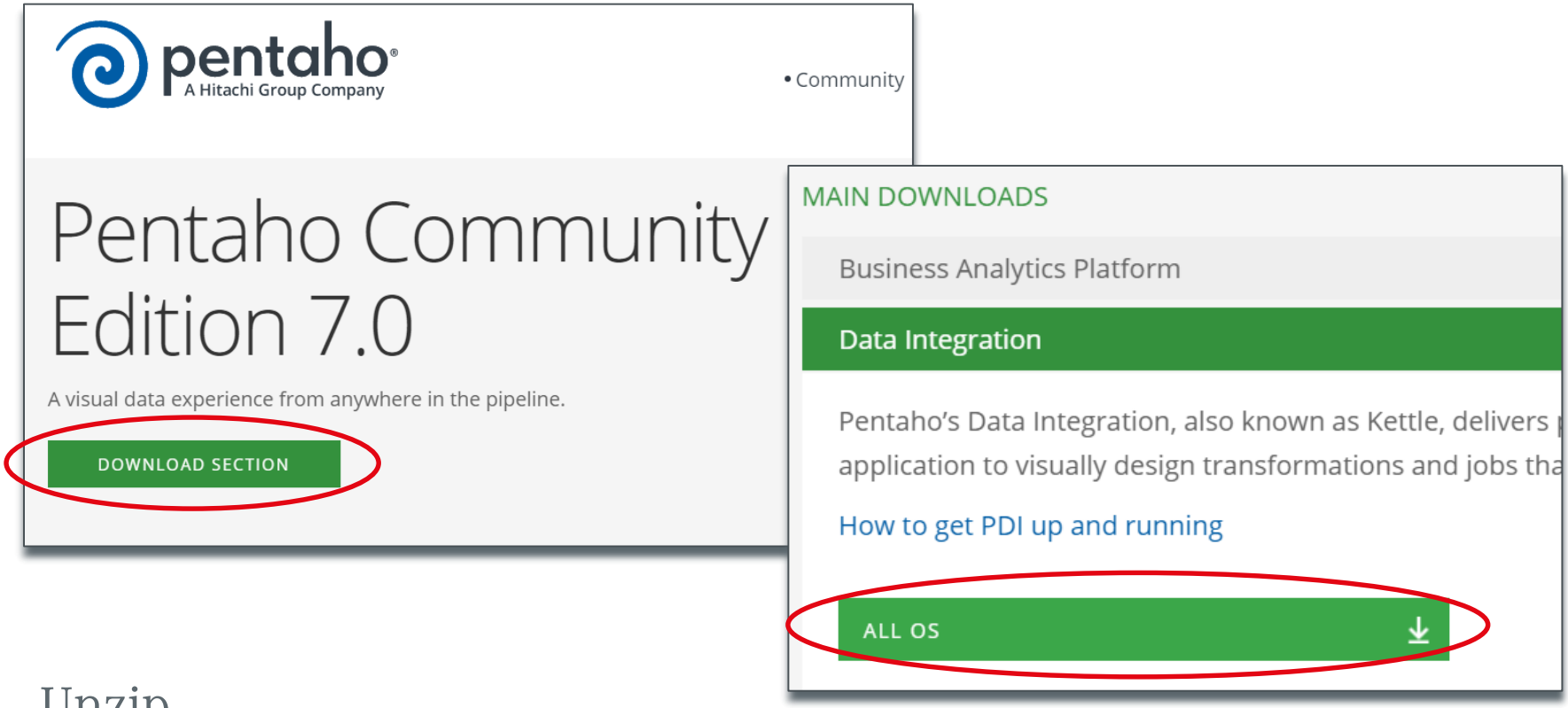
Kettle is an acronym for "Kettle E.T.T.L. Environment." Kettle is designed to help you with your ETTL needs, which include the **E**xtraction, **T**ransformation, **T**ransportation and **L**oading of data.

- Tools
  - Spoon
    - Grafisches Werkzeug (IDE)
    - Modellierung und Definition von
      Transformationen (einzelne Aktivitäten)
      und Jobs (Folge mehrerer Transformationen)
    - Ausführen von Transformationen und Jobs

  - Kitchen, Pan, Cart
    - Command-line Werkzeuge zum Ausführen von Transformationen und Jobs
      z.B. im Batch-Betrieb

Pentaho Data Integration: Installation

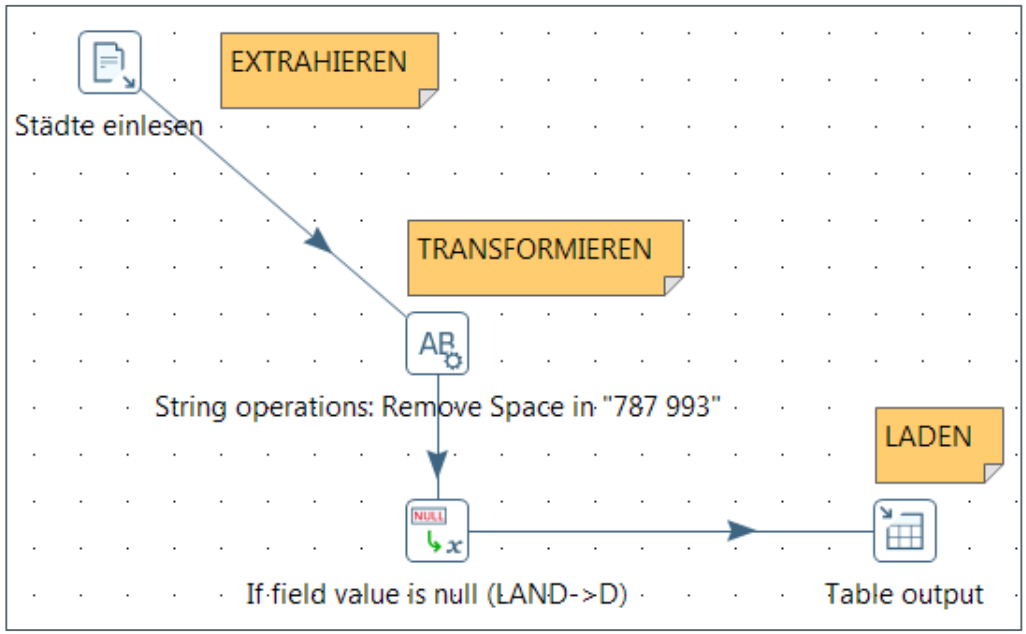
- community.pentaho.com



- Unzip
- Start: Doppelclick `spoon.bat` (Mac OS: *Integration Application*)

PDI: Transformationen

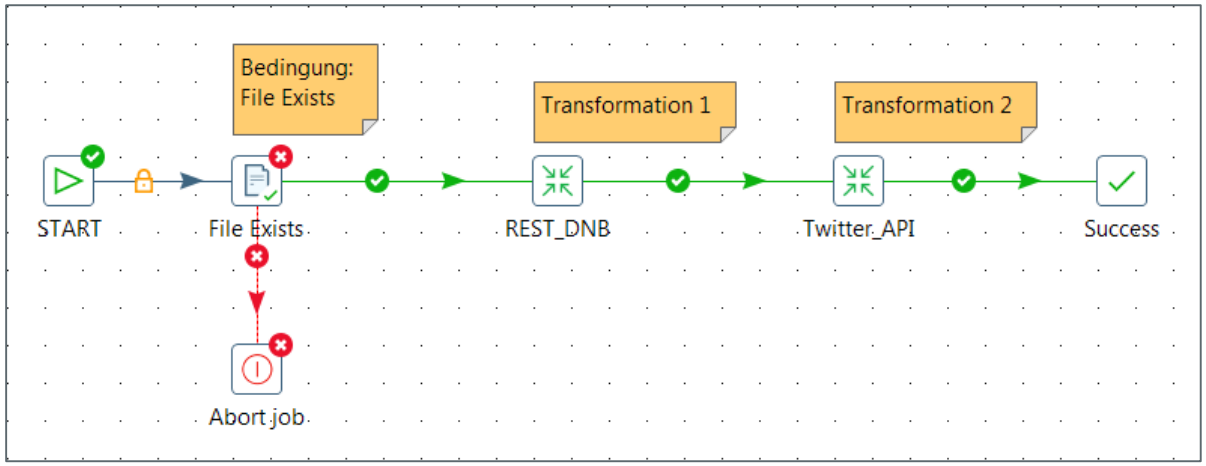
- Folge/Netz von Steps (Step=einzelner Transformationsschritt)

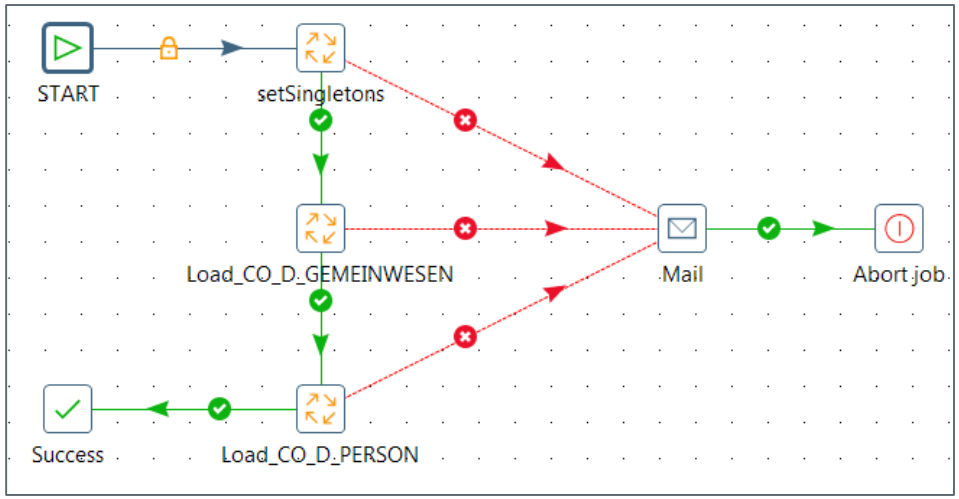


Weitere Beispiele: Live-Demo

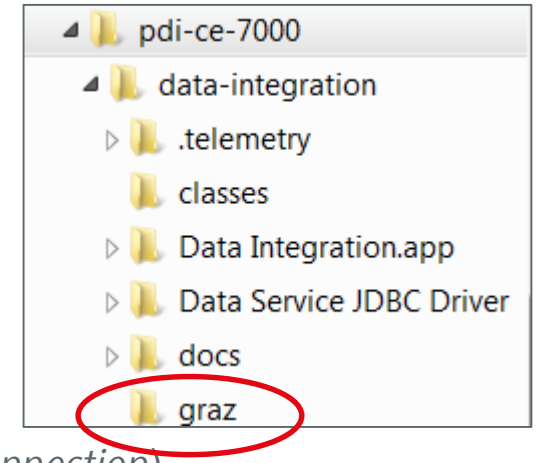
PDI: Jobs

- Sequentielle Steuerung von **Transformationen** und/oder **Jobs**





Pentaho Data Integration: Beispiele



Transformationen

Städteliste

- staedte01.ktr-staedte05.ktr (Entwicklungsvarianten, mit *DB connection*)
- staedte06.ktr (ohne Datenbank, ausschließlich Excel-/Text-Datenquellen)

Corpus via REST-Service

- rest_corpus_leipzig.ktr (Ähnliche Worte - Levenstein-Distanz)

DNB (Deutsche Nationalbibliothek) via REST-Service

- (erfordert Access Token)

Twitter-API

Beispiel: Städteliste

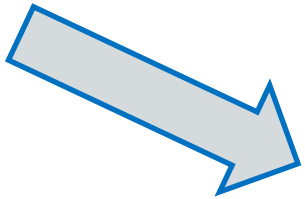
- Szenario
  - Datenquelle: Information über Städte (Einwohnerzahl, Land) aus Excel-Liste/CSV-Datei
  - Ziel: Ermittlung aller Großstädte (Einwohner: mehr als 2 Promille der Landesbevölkerung)

LAND	NAME	EINWOHNER
D	Berlin, Stadt	3 520 031
	Hamburg, Freie und Hansestadt	1 787 408
	München, Landeshauptstadt	1 450 381
	Köln, Stadt	1 060 582
CH	Bern	140 000
CH	Zürich	400 000
CH	Chur	35 000
CH	Werdenberg	57
F	Paris	2 500 000
F	Annecy	50 000
F	Marseille	850 000
F	Avignon	90 000
A	Wien	1 840 000
A	Graz	280200
A	Linz	200800
A	Salzburg	150800
A	Innsbruck	130850
A	Steyr	38395
A	Kufstein	18700

Beispiel: Städteliste

- Steps
  - Einlesen der Daten aus CSV-Datei
  - Konvertierung von Daten: z.B. "2 340 000" → 2340000
  - Ergänzen von "missing information": z.B. "keine Landesangabe" → D
  - Ermitteln der Einwohnerzahlen der Länder: Lookup in anderen Datenquellen
  - Berechnen des landesspezifischen Bevölkerungsanteils pro Stadt
  - Filtern der Städte mit Anteil > 2 Promille
  - Speichern der Großstädte in Textdate, Excel-Datei, Datenbank

	München, Landeshauptstadt	1 450 381
	Köln, Stadt	1 060 582
CH	Bern	140 000
CH	Zürich	400 000
CH	Chur	35 000

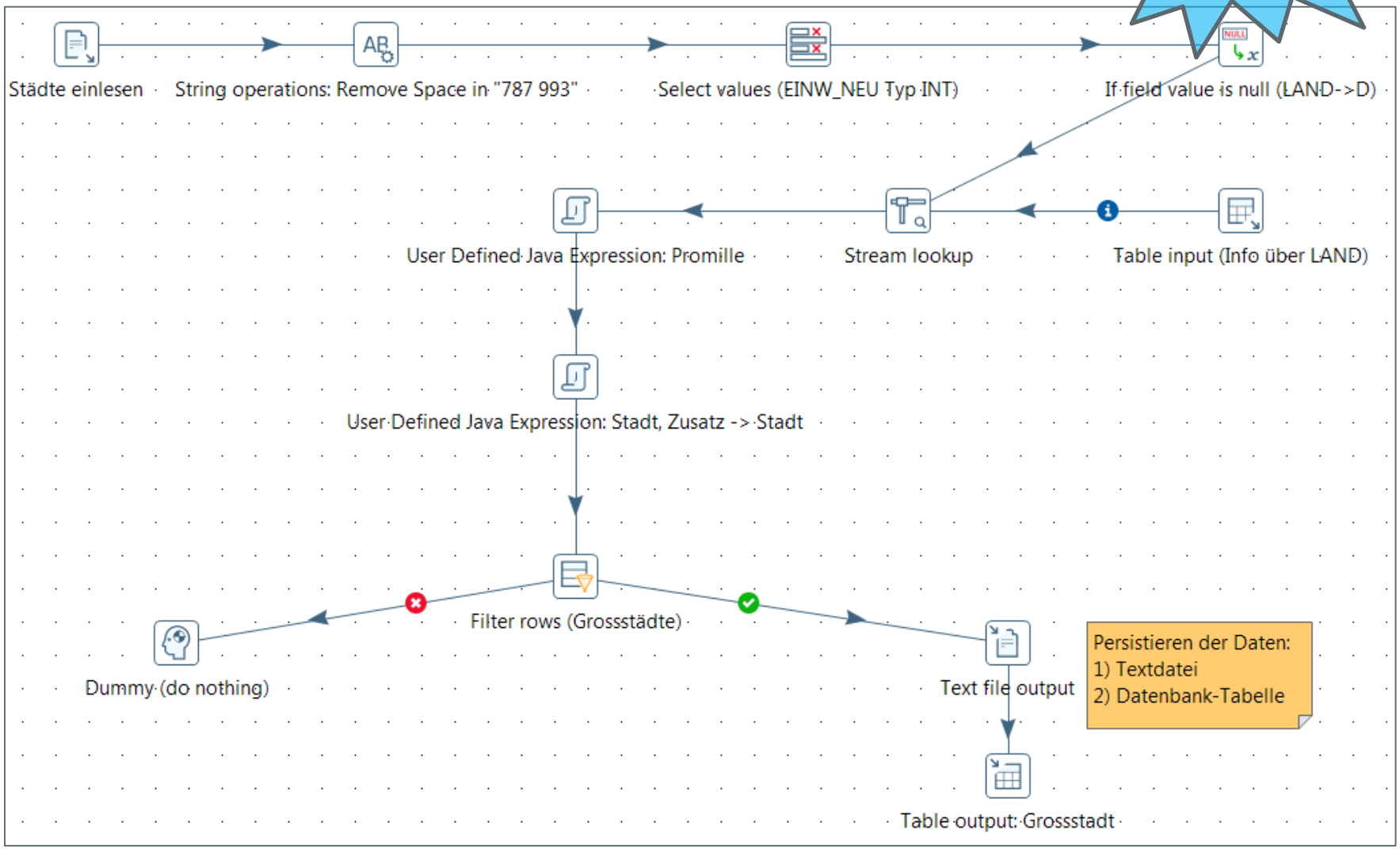


Anteil an Gesamt

München, Landeshauptstadt	1 450 381	1450381	81000000	17,9	München
Köln, Stadt	1 060 582	1060582	81000000	13,1	Köln
Bern	140 000	140000	8123000	17,2	Bern
Zürich	400 000	400000	8123000	49,2	Zürich
Chur	35 000	35000	8123000	4,3	Chur

Beispiel: Städteliste - Transformation

Live Demo



Beispiel: Städteliste - Transformation

- Step-Metrik (bei Ausführung)

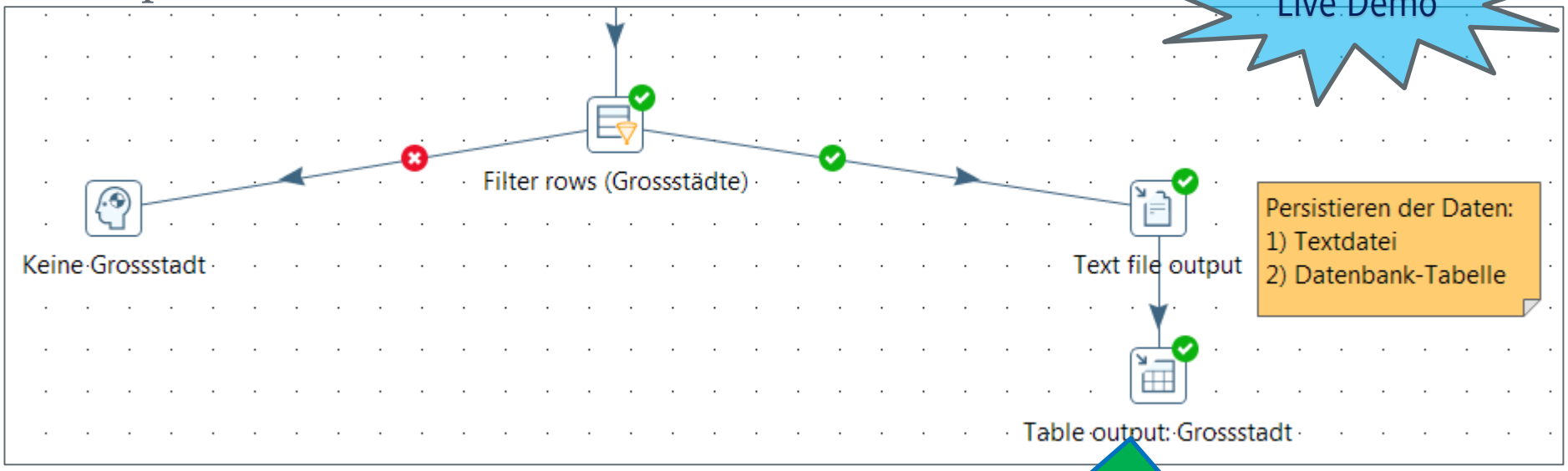
Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	Städte einlesen	0	0	2077	2078	0	1	0	0	Finished	0.2s	9.895
2	String operations: Remove Space in "787 993"	0	2077	2077	0	0	0	0	0	Finished	0.2s	9.751
3	Table input (Info über LAND)	0	0	5	5	0	0	0	0	Finished	0.1s	34
4	Select values (EINW_NEU Typ INT)	0	2077	2077	0	0	0	0	0	Finished	0.2s	9.571
5	If field value is null (LAND->D)	0	2077	2077	0	0	0	0	0	Finished	0.2s	8.801
6	Stream lookup	0	2082	2077	0	0	0	0	0	Finished	0.4s	5.422
7	User Defined Java Expression: Promille	0	2077	2077	0	0	0	0	0	Finished	0.4s	5.258
8	User Defined Java Expression: Stadt, Zusatz -> Stadt	0	2077	2077	0	0	0	0	0	Finished	0.4s	5.206
9	Filter rows (Grossstädte)	0	2077	2077	0	0	0	0	0	Finished	0.4s	4.945
1..	Text file output	0	61	61	0	61	0	0	0	Finished	0.4s	145
1..	Table output: Grossstadt	0	61	61	0	61	0	0	0	Finished	0.5s	130
1..	Dummy (do nothing)	0	2016	2016	0	0	0	0	0	Finished	0.4s	4.582

Beispiel: Städteliste – Transformation - Preview

Live Demo



Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview

First rows | Last rows | Off

#	LAND	NAME	EINW_NEU	GESAMT	ratio	STADT_BASE
1	D	Berlin, Stadt	3520031,0	81000000	43,45717284	Berlin
2	D	Hamburg, Freie und Hansestadt	1787408,0	81000000	22,066765432	Hamburg
3	D	München, Landeshauptstadt	1450381,0	81000000	17,905938272	München
4	D	Köln, Stadt	1060582,0	81000000	13,093604938	Köln
5	CH	Bern	140000,0	8123000	17,235011695	Bern
6	CH	Zürich	400000,0	8123000	49,242890558	Zürich
7	CH	Chur	35000,0	8123000	4,308752924	Chur

Demo - Ende

PDI: Funktionsumfang – Einblick – Transformationen

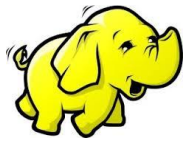
Daten-Input / Daten-Output
- CSV-/Textdatei
- JSON/XML-Datei
- Google Analytics
- Datenbanken: Tabelle / Views / SELECT-Anweisung
- OLAP
- SAP-System
- ...

Transformation
- Calculator
- String-Verarbeitung
- Split / Sortieren / Uniqueness

Sonstiges
- Datenkompression
- Filterung
- Lookup: Database / HTTP / REST / Webservice
- Merge/Join
- Statistik
- Big Data (Cassandra, CouchDB, Hadoop, MapReduce, MongoDB)

PDI: Funktionsumfang – Einblick – Jobs

- Sequentieller Ablauf
- Rekursion und Iteration möglich (nicht ideal)
- Benachrichtigung via Mail integrierbar
- Timer/Scheduler
- Scripting: Shell / SQL …
- BigData: Hadoop/Amazon/Spark
- Datentransfer: (S)FTP



Jobs und Transformationen

- Integrierbar in externe Programme (Kommandozeile, Cron etc.)
- Remote ausführbar
- …

Vielen Dank!

Fragen – Anregungen – Bemerkungen – Erfahrungen - Feedback

Kontakt

Prof. Dr. Klaus-Georg Deck
Duale Hochschule Baden-Württemberg
Lohrtalweg 10
74821 Mosbach/Baden
Deutschland

klaus-georg.deck@mosbach.dhbw.de