ETL (DataWarehouse)

# Talend Open Studio

**Talend starten:**

* File Manager starten
* Verzeichnis **/opt** öffnen
* Verzeichnis **/opt/TOS\_DI-r111943-V5.4.1** öffnen
* Entweder **TOS\_DI-linux-gtk-x86\_64** oder **TOS\_DI-linux-gtk-x86.sh** ausführen

# Pentaho

**Pentaho installieren:**

* Webseite **community.pentaho.com** öffnen
* Pentaho Data Integration downloaden
* Gezippten Ordner entpacken **sudo unzip „Path zu download“ –d /opt/PantahoDataIntegration**
* Verzeichnis **/opt** öffnen
* Verzeichnis **/opt/PentahoDataIntegration/data-integration** öffnen
* Befehl **chmod +x \*.sh** ausführen

**Pentaho starten:**

* Terminal starten
* Verzeichnis **/opt/PentahoDataIntegration/data-integration** öffnen
* **./spoon.sh** startet Pentaho GUI

**Pentaho DB-Connection herstellen:**

* Bei Database Connection auf **Configure** klicken
* Connection Type **Oracle** auswählen
* Connection Name ist frei wählbar (muss aber ausgefüllt werden)
* Host Name: **xubuntu**
* Database Name: **XE**
* Port-Number: **1521**
* User-Name: **tos**
* Password: **tos**

## Aufgabenstellung / Zielsetzung

ETL Software auf Funktionalität untersuchen. Speziell Software in OpenSource (Lowcost) Bereich. Das wird anhand von Talend Open Studio und Pentaho Data Integration vorgenommen.

Aufgrund des Umfangs dieser Arbeit soll die Funktionalitäten anhand von drei Beispieljobs untersucht werden. Diese beziehen sich auf „Fileimports“ „Datenbank-imports“ und „Datenbank-outputs“

In Talend Open Studio sind diese 3 Beispiel Jobs schon gegeben. 🡪 Untersuchung, wie diese umgesetzt wurden.

In Pentaho Data Integration müssen die 3 Beispiel Jobs noch erstellt werden.

Nach der Analyse von Talend und des Job Designs in Pentaho werden die beiden opensource/lowcost ETL Produkte kritisch gegenübergestellt.

# Pentaho Data Integration

Um die Jobs in Pentaho zu erstellen und anschließend die Ergebnisse mit den Jobs aus Talend Open Studio zu vergleichen wird eine neune Tabelle in der Datenbank erstellt. Dazu wird in SQLdeveloper die Datenbank TOS um die Tabelle CUSTOMERPENTAHO erweitert. Diese erhält die gleichen Spalten wie die Tabelle CUSTOMER, die für den Output von Talend Open Studio vorgesehen ist.

**DB Connection**:

Um eine Datenbankverbindung mit Pentaho Data Integration herzustellen muss man eine Transformation erstellen und kann für diese dann eine Datenbankverbindung einrichten. Das Einrichten einer Datenbankverbindung ist durch einen Wizard möglich. Allerdings werden in der Standardinstallation von Pentaho keine JDBC Treiber für Datanbanken mitgeliefert, was wiederum einen manuellen Eingriff in die Software zufolge hat. Man muss den richtigen JDBC Treiber für die eingesetzte Datenbank finden, herunterladen und in den passenden Ordner von Pentaho kopieren. In unserem Fall ist es die „ojdbc-10.2.0.3.0.jar“