

# DWH Projekt

## Ziele

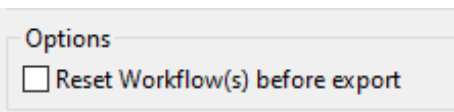
In diesem Projekt erstellen Sie ein Mini-DWH bestehend aus **Staging Area** und **Data Mart**. Danach erstellen Sie unterschiedlichste **Analysen**.

Das Projekt simuliert, wie ein Unternehmen aus mehreren Datensätzen ein DWH inkl. Analyse aufbaut. Sie sollten somit einen sehr guten praktischen Einblick in die DWH-Thematik erhalten, mit der sich die unterschiedlichsten Unternehmen befassen.

Fassen Sie Ihre Ergebnisse in einem **max. 10-seitigen PowerPoint Bericht** (inkl. Screenshots) zusammen. Danach laden Sie Ihren **Bericht + Source Code (KNIME-Workflow-Datei + SQL-Files)** auf Moodle. Pro Gruppe muss nur eine Arbeit abgegeben werden. Bitte geben Sie die Namen aller Gruppenmitglieder auf der Titelfolie des Berichts an.

**Bitte setzen Sie die KNIME Workflows beim Export nicht zurück** (Checkbox, siehe Bild), damit wir den

Output jeder Zelle betrachten können, ohne die entsprechenden Tabellen und Datenbanken lokal nachbauen zu müssen.



Options  
☐ Reset Workflow(s) before export

## Details

Das Praktikum besteht aus 3 Teilen. Voraussetzung ist, dass Sie die Praktika 1 bis 3 durchgearbeitet haben, da wir auf diesen Stoff aufbauen. Im Unterschied zu den Praktika gibt es für das Projekt jedoch keine schrittweise Anleitung, stattdessen müssen Sie das bereits Gelernte anwenden. Falls Sie irgendwann einmal nicht weiterkommen, wenden Sie sich gerne an Ihren Praktikumsleiter. Er hat sicherlich ein paar Tipps für Sie.

## 1 Staging Area

Ausgangsbasis sind unterschiedliche Datensätze des Beispielunternehmens Nordwind. Alle Datensätze liegen in Form mehrerer CSV-Files vor.

Im ersten Schritt erstellen Sie eine Staging Area, die alle Datensätze beinhaltet. Bearbeiten Sie folgende 6 Tabellen:

- Products
- Categories
- Customers
- Orders
- OrderDetails
- Employees

Folgende Punkte sind zu beachten:

- Verwenden Sie das Tool KNIME
- Schreiben Sie die Inhalte jeder CSV-Datei in eine separate PostgreSQL-Tabelle
- Erstellen Sie Tabellen mit sinnvollen Datentypen
- Behandeln Sie Datenqualitätsprobleme *bevor* Sie die Daten in eine PostgreSQL-Tabelle speichern

### **Aufgabe: Fehlerhafte Einträge**

Ebenfalls wurde in der Spalte Discount der Datenquelle *OrderDetails* teilweise der Rabatt als negative Zahl erfasst (beispielsweise -0.15). Es soll keine negative Rabatte geben.

### **Aufgabe: Duplikate**

In der Datenquelle *Products* wurden dieselben Produkte aus Versehen mehrmals erfasst. Finden Sie heraus, inwiefern sich die Einträge der identischen Produkte unterscheiden. Eliminieren Sie dann alle Duplikate, so dass es pro Produkt genau einen Eintrag in der Zieltabelle gibt.

### **Aufgabe: Unterschiedliche Datumsformate**

In den Datenquellen *Orders* und *Employees* haben sich unterschiedliche Datumsformate eingeschlichen, z.B. US-amerikanischen Format *MM-DD-YYYY* vs. internationales Format *DD-MM-YYYY*). Speichern Sie alle Tabellen mit demselben Datenformat.

Überlegen Sie sich, wie sie mit fehlenden (NULL) Datumswerten umgehen.

## **2 Data Mart**

Jetzt sind Sie bereit, einen Data Mart zu erstellt. Verwenden Sie mindestens alle 6 Tabellen, die in der Staging Area enthalten sind. Hierfür müssen Sie zunächst ein **Star-Schema** mit einer Faktentabelle und mehreren Dimensionstabellen entwerfen. Beschreiben Sie das Star-Schema im Bericht.

Bevor Sie den Data Mart designen, überlegen Sie sich hierzu, welche Fragestellungen Sie später genauer analysieren möchten (siehe Aufgabe „3 Analysen“).

Danach erstellen Sie die entsprechenden Tabellen und befüllen Sie mit den Daten aus der Staging Area.

Falls Sie eine Zeitdimension verwenden möchten, nehmen Sie sich ein Beispiel an der Praktikumsübung. Die Daten dafür müssen Sie nicht in KNIME bzw. Postgres generieren.

### 3 Analysen

Nun können Sie basierend auf Ihren Data Mart Analysen durchführen. Überlegen Sie sich fünf interessante Fragestellungen, die Sie gerne bearbeiten möchten. Hierbei sind Ihrer Kreativität keine Grenzen gesetzt. Als Gedankenanstregung betrachten Sie folgende Fragestellungen:

- Aus welchen Ländern kommen die Kunden mit den 5 grössten Aufträgen?
- Was ist der durchschnittliche Discount aller Produkte, die in lateinamerikanische Länder geliefert werden?
- Welche Produkte werden stärker nachgefragt als deren Lagerkapazität?

Welche Tasks und Funktionen Sie dafür verwenden, ist Ihnen freigestellt.  
*Wagen Sie Experimente!*

### Ergebnis

**Hinweise für den max. 10-seitigen PowerPoint Bericht:**

- Beschreiben Sie kurz den Aufbau der Staging Area und des Data Marts. Verwenden Sie Screenshots, um die Datenmodelle und die ETL-Prozesse darzustellen.
- Formulieren Sie Ihre 5 Fragestellungen und beschreiben Sie die Analyseergebnisse (inkl. Screenshots).
- **Wichtig:** Beschreiben Sie Ihre Erfahrungen und ein Fazit zum Projekt auf einer Folie

### Bewertung

Insgesamt können Sie 8 Punkte für das Projekt erlangen. Die genaue Punkteaufteilung ist wie folgt:

- Staging Area: 2 Punkte
- Data Mart: 2 Punkte
- Analyse: 2 Punkte
- Bericht: 2 Punkte