CAS Information Engineering

## Leistungsnachweis für Modul DB & DWH

**Gruppenmitglieder**:

Andreas Fischer

Bernd Novotny

Tobias Schieferdecker

Inhaltsverzeichnis:

Leistungsnachweis für Modul DB & DWH 1

1. Übersetzung des gegeben ERMs in den uns bekannten Dialekt 2

2. Auslagern der relevanten Daten in csv-Dateien 2

3. Erstellung der staging-area 2

4. Reparieren der korrupten Daten 5

5. Konzipierung des Sternschemas für das DWH 8

6. Erstellung und Befüllen des DWH 8

7. Auswertungen auf Basis des DWH 8

8. Fazit und Schlusswort 8

# **Übersetzung des gegeben ERMs in den uns bekannten Dialekt**

# **Auslagern der relevanten Daten in csv-Dateien**

Nachdem die Northwind-Datenbank mit der uns zur Verfügung gestellten Datei “Northwind-Database.sql” mittels MySQL-Workbench erzeugt und im Folgenden im selben Tool mittels der Datei “Northwind-Data.sql” befüllt wurde, haben wir uns der Aufgabe gewidmet, die für unsere späteren Auswertungen relevanten Daten in csv-Dateien auszulagern.



Bild 1 Der Knopf für den Export einer Datentabelle aus MySQL-Workbench ist rot umrandet.

Für diesen Schritt haben wir kein Skript verwendet. Wir haben vielmehr einfach die Funktionalität von MySQL-Workbench verwendet haben, auf Knopfdruck ein csv-File aus der gerade betrachteten Tabelle zu erzeugen (siehe Bild 1). Die Tabellen, die uns für spätere Auswertungen interessant schienen, waren:

* “customers”
* “products”
* “orders”
* “order\_details

Hierbei wurden an dieser Stelle keine Einschränkung der Daten vorgenommen, sondern das dem SQL-Befehl

“SELECT \* FROM ‘TABLE’ ;” (mit ‘TABLE’ gleich einer der vier von uns ausgesuchten Tabellen) entsprechenden Resultat in das csv-Format exportiert. Die entsprechenden csv-Files sind dem zip-File “skripte\_und\_daten\_CAS\_DB\_DWH.zip” beigefügt, das im Email-Anhang mitverschickt wurde.

# **Erstellung der staging-area**

Zunächst ein paar Worte zu unserem Umgang mit der staging-area: Zur effizienteren Arbeitsaufteilung haben wir die Aufgaben der Einrichtung der staging-area und der Erzeugung des multidimensionalen Modells getrennt. Aus Gründen der Einfachheit und der Möglichkeit, schneller mit der Arbeit zu beginnen, haben wir anfangs die Daten für die verschiedenen Transformationen (Reparatur der korrupten Daten, Befüllen des multidimensionalen Modells, ...) aus csv-Dateien eingelesen und teils auch wieder als csv-Dateien rausgeschrieben. Aus zeitlichen Gründen haben wir es dann nicht mehr geschafft, dies noch so umzustellen, so dass das Einlesen – bis auf den einen Schritt zur Befüllen der staging-area Datenbank – immer aus Datenbanktabellen vorgenommen werden würde.

Wir wollen aber gerne versuchen zu demonstrieren, dass wir theoretisch die Integration der Daten über die Datenbank hinbekommen hätten. Wir haben die staging-area Datenbank erstellt und mit den notwendigen Tabellen befüllt. Nur eben die Integration in den Daten-Flow ist nicht gegeben.

Es folgt eine kurze Erläuterung, wie unsere staging-area-Datenbank erzeugt und befüllt wird.

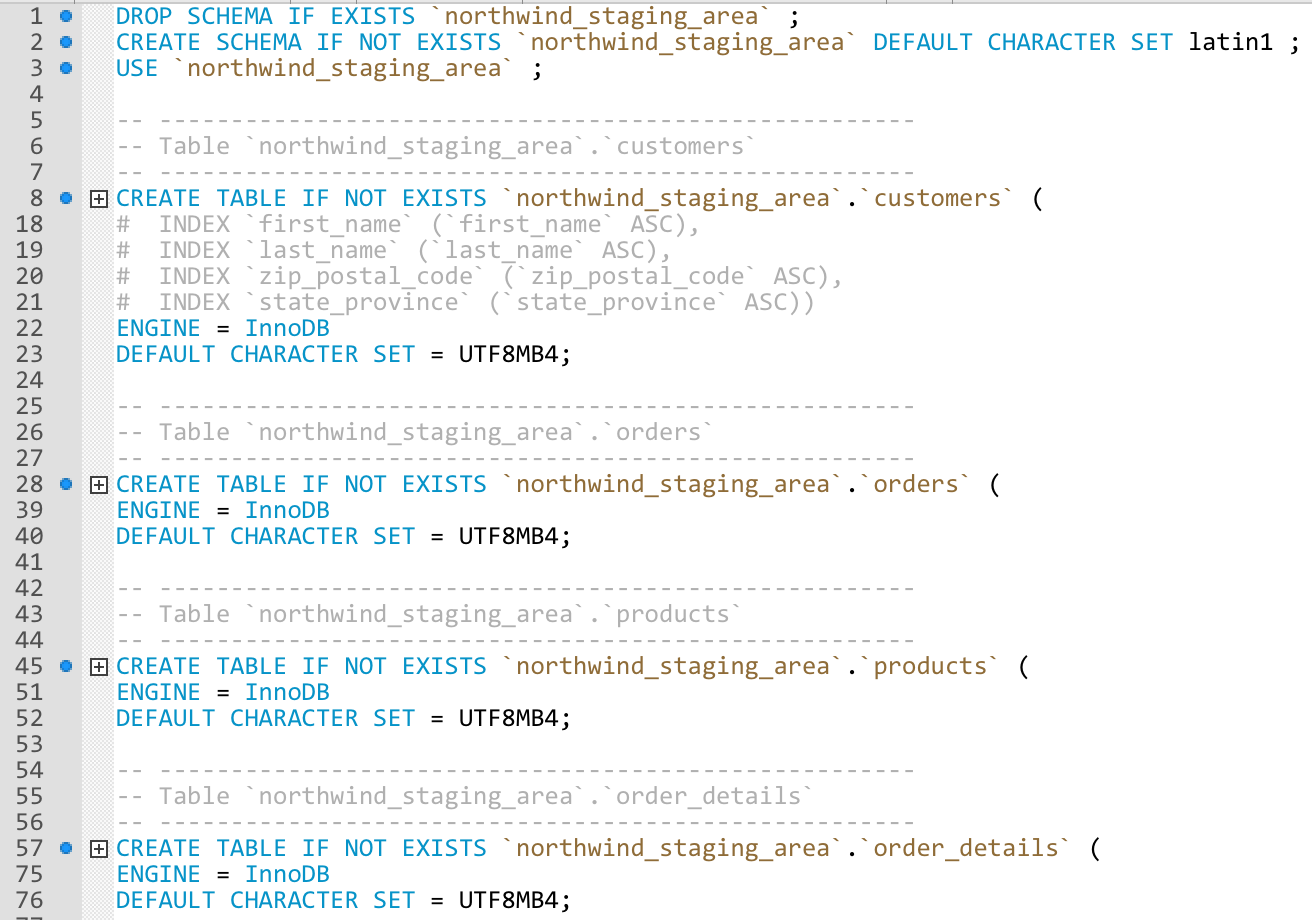


Bild 2: Skript "create\_staging\_area\_schema.sql". Für eine bessere Übersicht sind die Abschnitte für die Erzeugung der einzelnen Spalten zusammengeklappt.

1. Das Skript zur Erzeugung der leeren Datenbank ist in MySQL-Workbench auszuführen und lautet „create\_staging\_area\_schema.sql“. Bei der Ausführung des Skripts werden die verschiedenen Tabellen mitsamt ihren Variablen erzeugt, die Reihen bleiben dabei jedoch leer (siehe Bild 2). An dieser Stelle werden schon nicht mehr alle Variablen der ursprünglichen Tabellen übernommen, sondern nur noch jene für die späteren Abfragen notwendigen.
2. Das Befüllen der Datenbank erfolgt mittels des Pentaho Data Integration (PDI) Spoon Skripts „add\_data\_to\_staging\_area.ktr“ (siehe Bild 3). Jedes der csv-Files wird einzeln mit einem „CSV Input“-Schritt eingelesen und jeweils mit einem „Insert / Update“-Schritt in die Datenbank geschrieben. Es hat sich hierbei gezeigt, dass durch die Einschränkung durch die Verlinkung der Tabellen untereinander durch Schlüssel-Fremdschlüssel-Beziehungen, das gleichzeitige Befüllen der Tabellen eine Fehlermeldung provoziert. Um dies zu vermeiden, muss man zunächst die Verbindung zwischen „read orders“ und „Insert orders into staging area“ sowie diejenige zwischen „read order details“ und „Insert order details into staging area“ deaktivieren und die Spoon-Transformation laufen lassen. Nachdem dies erfolgreich geschehen ist, kann man die beiden deaktivierten Verbindungen wieder aktivieren und die Transformation erneut starten. Danach ist die staging-area Datenbank gefüllt.

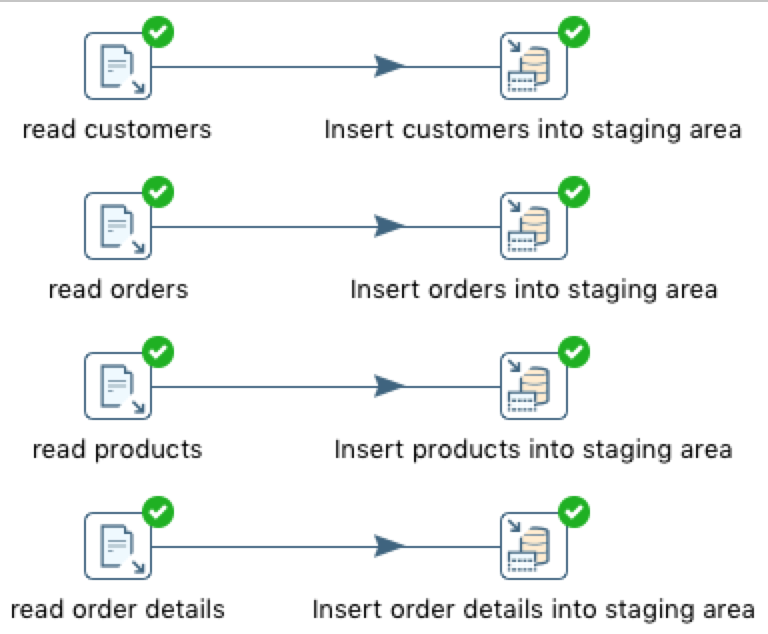


Bild 3: Transformation in Spoon zum Befüllen der staging-area-Datenbank nach erfolgreicher Ausführung.

# **Reparieren der korrupten Daten**

Um – wie in der Aufgabenstellung angegeben – herauszufinden, inwiefern sich die Einträge der identischen Produkte unterscheiden, wurde ein kleines Skript namens “study\_products\_table.R” in der Programmiersprache R verwendet. Dieses ist auch Teil des Lieferumfangs unseres zip-Files. Zu seiner Ausführung ist die Installation des Pakets „dplyr“ nötig. Das Skript tut prinzipiell folgendes:

1. Einlesen der csv-Datei „products.csv“.
2. Es erzeugt einen logischen Vektor mit der Information, wann ein Produktname zum mindestens zweiten Mal auftritt (man sieht hier, dass nur die letzten 3 stellen im Vektor „TRUE“ sind)
3. Da ein Datensatz von allen doppelten Einträgen (inklusive des ersten Auftretens eines später wiederholten Eintrags) erwünscht ist, wird ein filternder „semi-join()“ des ursprünglichen Datensatzes gegen den Datensatz mit den doppelt vorkommenden Produktnamen gemacht. Der Key ist der Produktname. Hierbei entsteht der Datensatz mit den doppelten Einträgen. Dieser hat sechs Zeilen.
4. Mit dem Befehl „View()“ wird der Datensatz visuell dargestellt (siehe Bild 4).
5. Es zeigt sich, dass drei verschiedene Produkte jeweils doppelt vorkommen.
6. Um sicher zu gehen, dass auch wirklich alle Einträge bis auf „id“ doppelt sind, wird für die drei Produkte mit dem Befehl „identical()“ jeweils genau das überprüft – mit Erfolg. Das heisst also, dass diese drei Zeilen bis auf den Eintrag „id“ identisch sind, und die letzten drei Zeilen des ursprünglichen Datensatzes (bzw. allgemeiner, die doppelten) entfernt werden müssen.

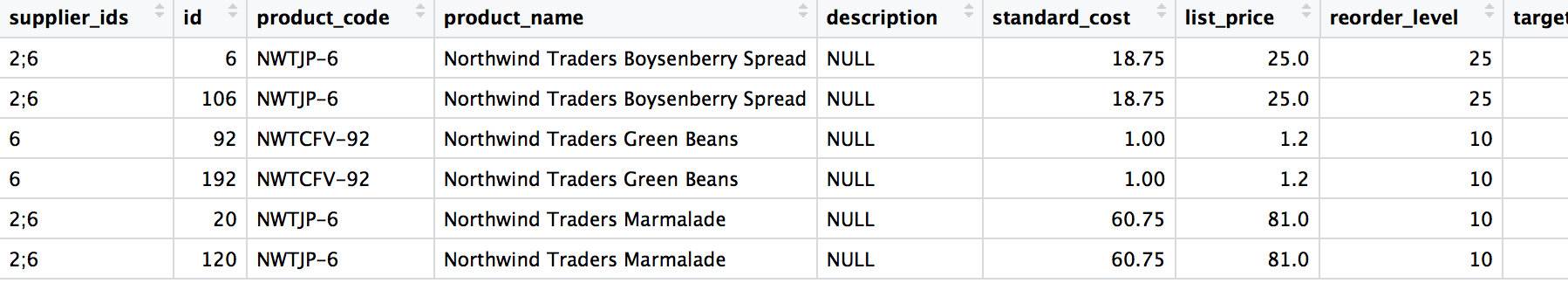


Bild : Ausschnitt (Reihen vollständig, Spalten unvollständig) aus Datensatz doppelter Einträge, extrahiert mittels "R" aus der Tabelle Produkte der Northwind-Datenbank.

Wie schon erwähnt, passiert das Reparieren der fehlerhaften Daten bei unserer Lösung ausserhalb der staging-area Datenbank. Die Daten werden von dem Spoon-Skript “staging\_area-v2.ktr” aus den csv-Files “products.csv” und “order\_details.csv” eingelesen und die reparierten – und zumindest im Falle der “order details”-Tabelle auch um weitere Variablen aus den anderen csv-Files ergänzten – Daten werden in die csv-Files “products\_corrected.csv” und “order\_details\_corrected.csv” geschrieben. Das Hinzufügen von Spalten zur Tabelle der Bestellungsdetails erfolgt im Hinblick auf eine Vereinfachung der Überführung der Daten in das Sternschema.

Die Transformationen zur Behebung der Datenprobleme in der ursprünglichen Northwind-Datenbank laufen folgendermassen ab (siehe auch Bild 5):

1. Produkt-Tabelle
   1. Einlesen der csv-Datei “products.csv”
   2. Sortierung der Zeilen der Tabelle nach Produktnamen mit einem “Sort rows”-Schritt. Dies ist notwendig, da der folgende Schritt nicht funktioniert, falls doppelte Reihen nicht nacheinander kommen.
   3. Entfernen der doppelten Einträge mit einem „Unique rows“-Schritt. Hierbei werden alle Variablen bis auf die Variable „id“ verglichen.
   4. Sortierung der Zeilen der Tabelle nach „id“, so wie sie ursprünglich geordnet waren.
   5. Schreiben der Daten in das csv-File „products\_corrected.csv“
2. Bestelldetails-Tabelle
   1. Einlesen der csv-Datei “order\_details.csv”.
   2. Hinzufügen der Variablen „customer\_id“, „employee\_id“ (diese Variable ist in der staging-area Datenbank nicht vorhanden. Sie wird aber auch für die Auswertungen nicht benötigt), „order\_date“, „payment\_type“ aus der Tabelle „orders“ mittels eines Schrittes vom Typ „Stream Value Lookup“.
   3. Hinzufügen der Variable „standard\_cost“ aus der Tabelle „products“
   4. Hinzufügen der Variablen „city“, „state\_province“ und „country\_region“ aus der Tabelle „customer“
   5. Kreieren einer neuen Spalte „discount\_corr“, die den Betragswert der Variable „discount“ enthält. Dies geschieht durch einen Schritt vom Typ „Calculator“.
   6. Mit einem Schritt „Select / Rename values“ werden 15 Variablen ausgewählt und gleichzeitig die Variable „discount“ durch „discount\_corr“ ersetzt.
   7. Im letzten Schritt werden die Daten in das csv-File „order\_details\_corrected.csv“ geschrieben.

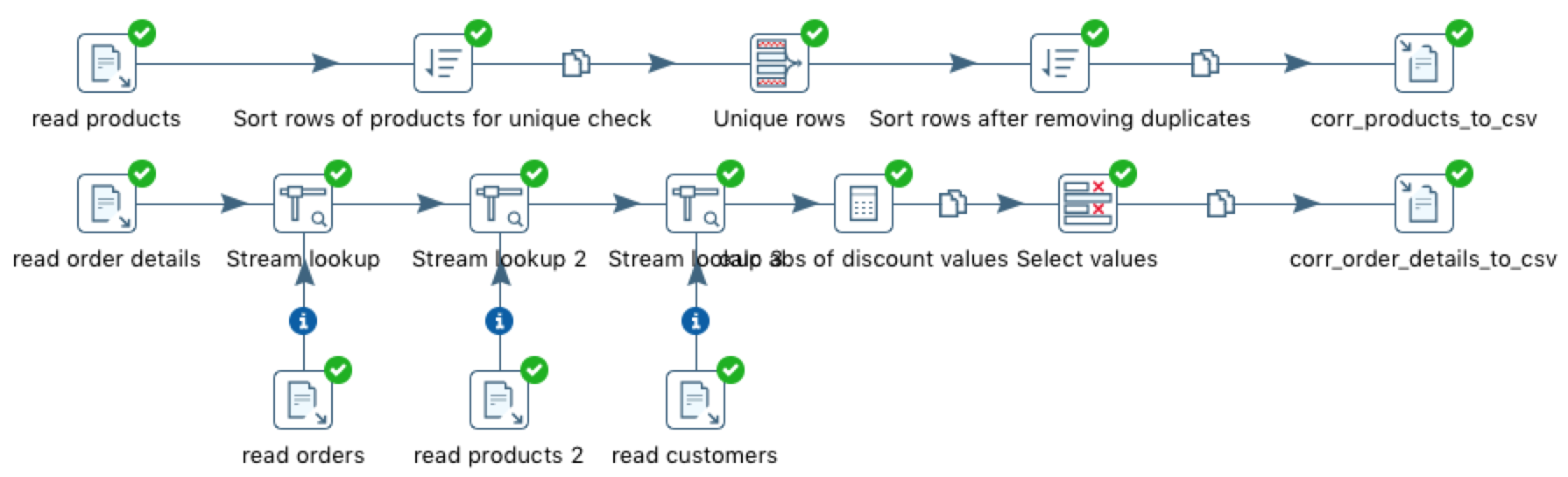


Bild 5: "staging\_area-v2.ktr" nach erfolgreichem Durchlauf. Hier werden fehlerhafte Daten korrigiert und im Fall der Bestelldetails auch weitere Spalten hinzugefügt. Am Ende werden die Daten in die zwei im Text erwähnten csv-Files geschrieben.

Wie in Abschnitt 3 erwähnt, sollte das Einlesen von sowohl der „products“- als auch der „order\_details“-Tabellen eigentlich aus der staging-area Datenbank durch einen „Table input“-Schritt geschehen. Das Schreiben der beiden Tabellen sollte eigentlich ebenso in die staging-area Datenbank geschehen, entweder mit einem „Dimension lookup / update“-Schritt oder mit einem „Table output“-Schritt.

# **Konzipierung des Sternschemas für das DWH**

# **Erstellung und Befüllen des DWH**

# **Auswertungen auf Basis des DWH**

# **Fazit und Schlusswort**