Leistungsnachweis DB & DWH

HS 2018

Patrick Lengacher und Raphael Monstein

1. Laden und studieren der bestehenden DB

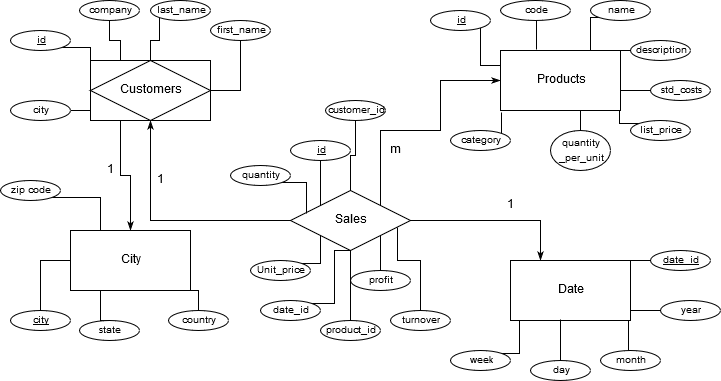
Die Datenbank der Aufgabestellung wurde geladen mit den SQL Skripten geladen und studiert.

1. Festlegen, welche Auswertungen mit dem DWH ermöglicht werden sollen

In einem ersten Schritt wurde festgelegt, welche Auswertungen mindestens gemacht werden sollen. Dabei wurden folgende Auswertungen definiert:

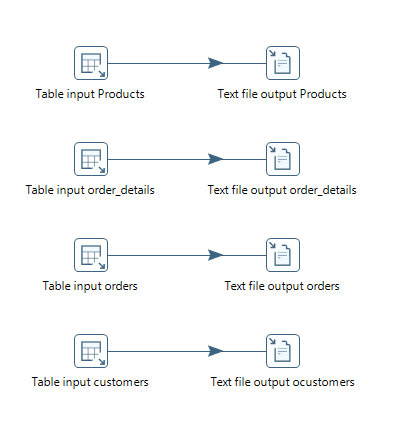
* Welche Produkte wurden am meisten gekauft?
* Welche Produkte hatten den höchsten Umsatz?
* Welche Produkte haben die höchste Marge?
* Woher stammen die Kunden mit dem höchsten Umsatz?

Um diese Auswertungen zu ermöglichen, wurde ein minimalistisches Schneeflockenschema ausgewählt. Das nachfolgende ERM Diagram zeigt dies auf.



1. Exportieren der DB in CSV Dateien

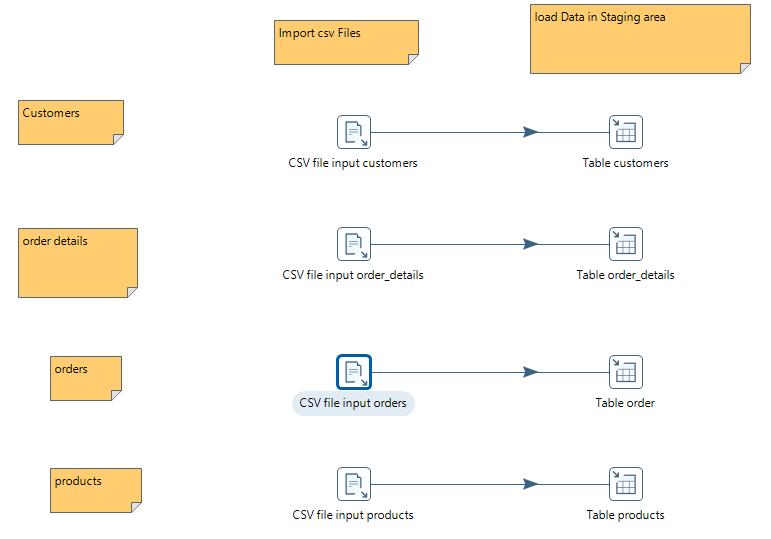
Die Daten in der Datenbank wurden in Pentaho Data Integration (PDI) geladen und von dort aus in CSV exportiert. Ein Screenshot ist nachfolgend zu finden.



Dabei wurden nur die Tabellen der Datenbank in die CSV Dateien exportiert, die anschliessend auch benötigt werden.

1. Importieren der Daten in die Staging Area

Die Daten in den CSV Dateien wurden anschliessend eins zu eins in PDI in eine leere Datenbank namens *staging* importiert, wie der nachfolgende Screenshot zeigt.



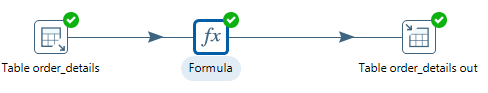
Da die importierten Daten noch, wie in der Aufgabestellung beschrieben, Qualitätsprobleme haben, mussten diese transformiert werden. Dies wurde in nächsten Schritt gemacht.

1. Beheben der Datenqualitätsprobleme

Die Datenqualitätsprobleme betrafen die Tabellen *order\_details* und *products*.

Order\_details

In *order\_details* wurden die alle Discounts als negativ erfasst. Das umwandeln in positive Werte kann relativ einfach in PDI gemacht werden. Dazu kann ein Formula-Block verwendet werden.



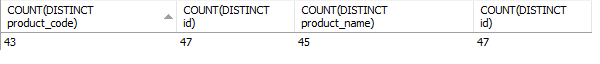
Die verwendete Operation im Formula-Block ist die *ABS* Funktion.

Products

Die Tabelle *products* hatte das Problem, dass irrtümlicherweise gewisse Einträge doppelt erfasst wurden. Um herauszufinden, wie diese Einträge detektiert werden könne, wurden ein paar SQL Abfragen geschrieben. Als erstes wurde geschaut, ob *product\_code* und *product\_name* eindeutig sind.

SELECT COUNT(DISTINCT product\_code), COUNT(DISTINCT id), COUNT(DISTINCT product\_name), COUNT(DISTINCT id)

FROM northwind.products;



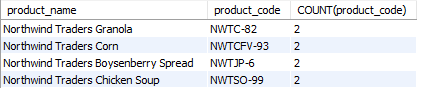
Dies deutet darauf hin, dass es Duplikate in beiden Spalten gibt. Um einen besseren Eindruck zu bekommen, wurden die Duplikate angezeigt.

SELECT product\_name, product\_code, COUNT(product\_code)

FROM northwind.products

GROUP BY product\_code

HAVING COUNT(product\_code) > 1;



SELECT product\_name, product\_code, COUNT(product\_name)

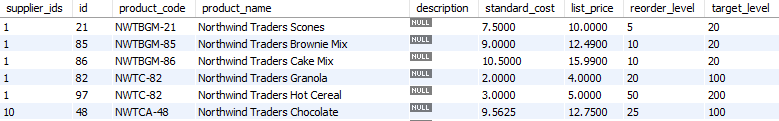
FROM northwind.products

GROUP BY product\_name

HAVING COUNT(product\_name) > 1;



Weitere Abfragen haben ergeben, dass die Duplikate im *Product\_code* unterschiedliche Beschreibungen und auch sonst unterschiedliche angaben haben, wie im nachfolgenden Screenshot dargestellt wird.



Dabei wird ersichtlich, dass z.B. das Produkt NWTC-82, welches zwei Mal mit dem gleichen Produkt Code aufgeführt ist, wahrscheinlich kein Duplikat ist. Aus diesem Grund wurden nur die Tupel entfernt, welche den gleichen Produktenamen haben. Dies wurde mit der folgenden SQL Abfrage gemacht:

SELECT \*

FROM northwind.products

WHERE id IN

(

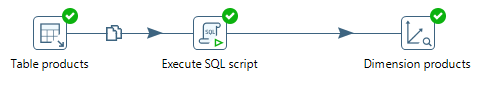
SELECT MIN(id)

FROM northwind.products

GROUP BY product\_name

);

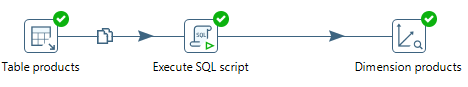
Nachdem die Abfrage in der MySQL Workbench getestet wurde, wurde die Transformation in PDI integriert.



1. Erzeugen der Dimensionen

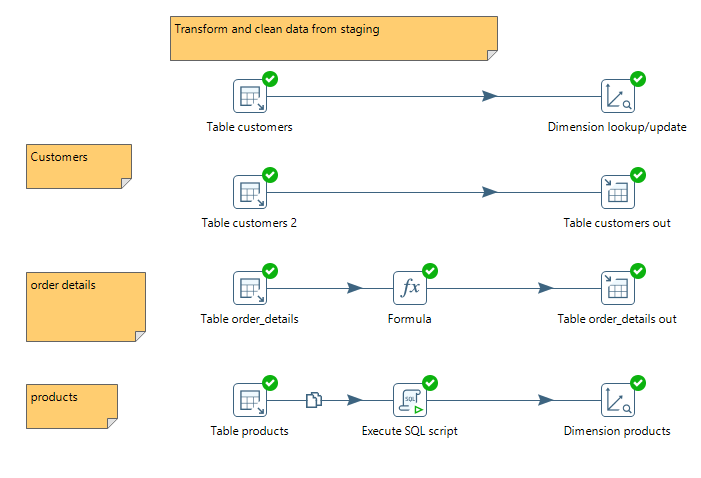
Wie in einer der Übungen gelernt, wurden die Dimensionen in PDI mit Dimension lookup/update Block erstellt.





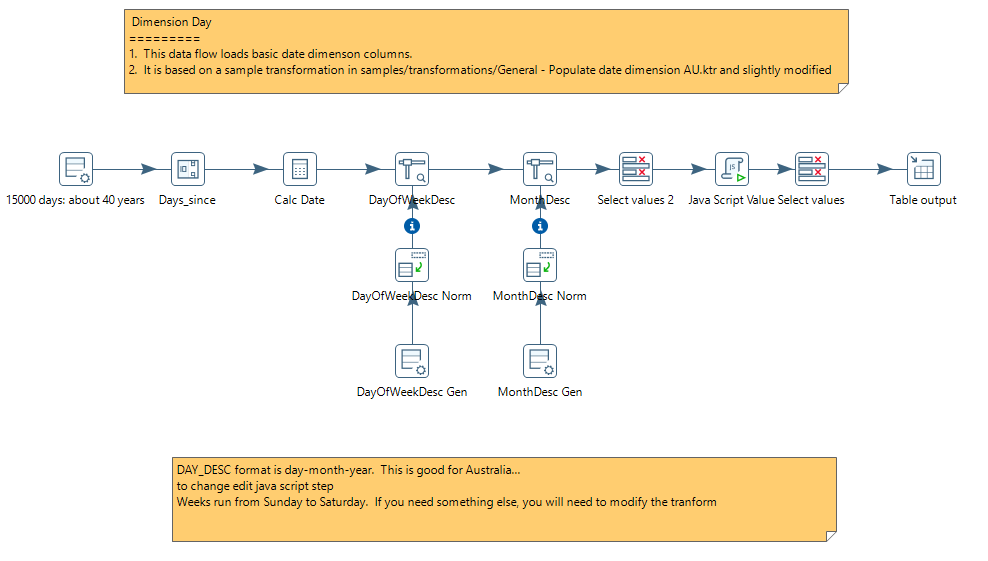
Dabei wurde der Einfachheit halber eine SCD Type 1 angewendet.

Eine Übersicht über alle Transformationen zur Behebung der Qualitätsprobleme und das Erzeugen der Dimensionen ist nachfolgend abgebildet.



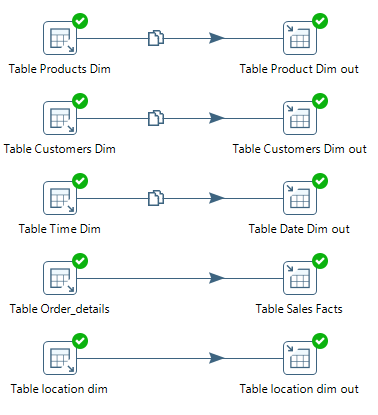
1. Erzeugen einer Datums-Dimension

Eine zusätzliche Dimension für das Datum wurde separat erzeugt. Die Idee ist, dass damit einfacher Abfragen zu gewissen Zeiträumen, z.B. nach Quartal, durchgeführt werden können. Der Block, welcher dazu verwendet wird, ist in den Beispielen von PDI enthalten und wurde leicht modifiziert. Der ursprüngliche Block kann im PDI Installationsverzeichnis unter *samples\transformations\ General - Populate date dimension AU.ktr* gefunden werden.



1. Erzeugen des Schemas im DWH

Alle bisher aufgezeigten Transformationen wurden in der Staging Area durchgeführt. In diesem Schritt werden die Daten in eine neue Datenbank mit dem Namen *DWH* gefüllt.



Bei diesem Schritt wurden nicht alle Attribute ins DWH kopiert, sondern nur noch diese, welche auch (potentiell) für spätere Abfragen benötigt werden. Besonders hervorzuheben ist, wie die Tabelle *sales\_facts* erzeugt wird. Im Block *Table Order\_details* wird folgende SQL Abfrage durchgeführt:

SELECT ord\_d.id AS id,

ord\_d.quantity AS quantity,

ord\_d.unit\_price AS unit\_price,

d.DATE\_SK AS date\_id,

cust.id AS customer\_id,

prod.id AS product\_id,

ord\_d.unit\_price\*ord\_d.quantity AS turnover,

(ord\_d.unit\_price\*ord\_d.quantity) - (prod.standard\_cost\*ord\_d.quantity) AS profit

FROM staging.order\_details\_clean AS ord\_d,

staging.date\_dim AS d,

staging.orders AS ord,

staging.customers AS cust,

staging.products\_dim AS prod

WHERE ord\_d.product\_id = prod.id AND

ord\_d.order\_id = ord.id AND

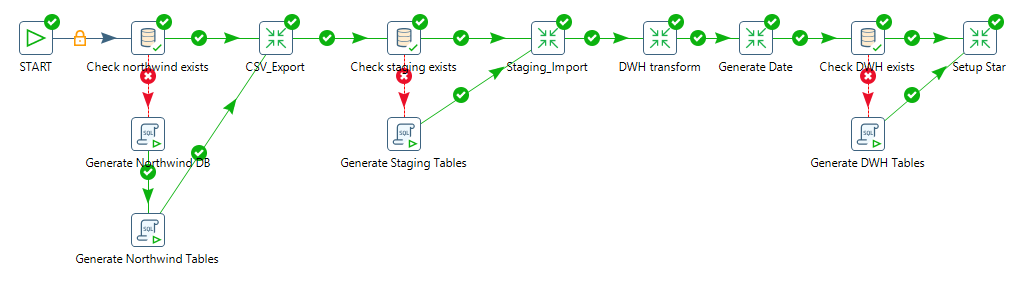
ord.order\_date = d.day\_date AND

ord.customer\_id = cust.id;

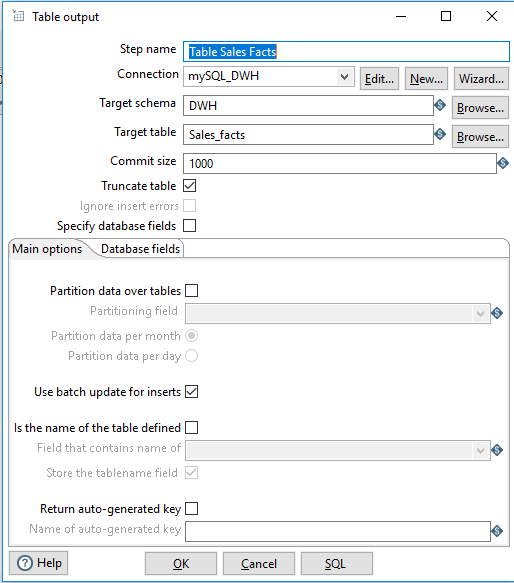
Um spätere Abfragen zu vereinfachen, werden die beiden Attribute *turnover* und *profit* erzeugt.

1. Erstellen eines Jobs

Damit die verschiedenen Transformationen nicht einzeln von Hand ausgeführt werden müssen, wurde ein Job erstellt.



Dieser Job führt die Transformationen in der richtigen Reihenfolge aus. Dabei wurde auch versucht, die Tabellen automatisch zu erzeugen, falls diese nicht vorhanden sind. Leider konnte dies nicht innerhalb von vernünftiger Zeit zum Laufen gebracht werden. Stattdessen müssen, falls die Datenbanken und Tabellen nicht existieren, die beiden SQL Skripts *staging.sql* und *DWH.sql* vorgängig laufen gelassen werden. Diese beide Skripts wurden mehr oder weniger mit Copy-Paste aus den PDI Befehlen erstellt, welche erzeugt werden, wenn eine Tabelle mit dem *Table output* erzeugt wird (beim drücken der Schaltfläche *SQL* in den Eigenschaften des Blocks).



1. Abfragen