Inhalt

Inhalt

[Installation Pentaho 1](#_Toc40045560)

[Auswahl und Kurzbeschreibung Datenset 1](#_Toc40045561)

[Transformation Datenset in Sternschema (Python) 2](#_Toc40045562)

[Import in Pentaho und Schreiben der Tabellen in postgres Datenbank 2](#_Toc40045563)

[Mehrdimensionale Analyse Datenset 6](#_Toc40045564)

# Installation Pentaho

Betriebssystem Windows 10

Download Pentaho von: (Peter)

Java 8 inkl. JRE und JDK

Neustart -> Systempfad update

Vor dem ersten Ausführen:

set-pentaho-env.bat ausführen

Zum Starten: spool.bat

Zusätzlich: Postgres 12.2, pgAdmin 4.20

Zum Importieren der CSV-Dateien waren noch weitere Schritte erforderlich:

1. Plugin installieren: Import from txt

2. Neustart Pentaho

Eigentlicher Import

3. New - Transformation - Import - Dateiauswahl

# Auswahl und Kurzbeschreibung Datenset

Typische Verkaufsdaten.

Link: <https://www.kaggle.com/kyanyoga/sample-sales-data#sales_data_sample.csv>

Das Datenset liegt als csv vor, und es gibt mehrere „Daten-Dimensionen“ (wie im Sternschema).

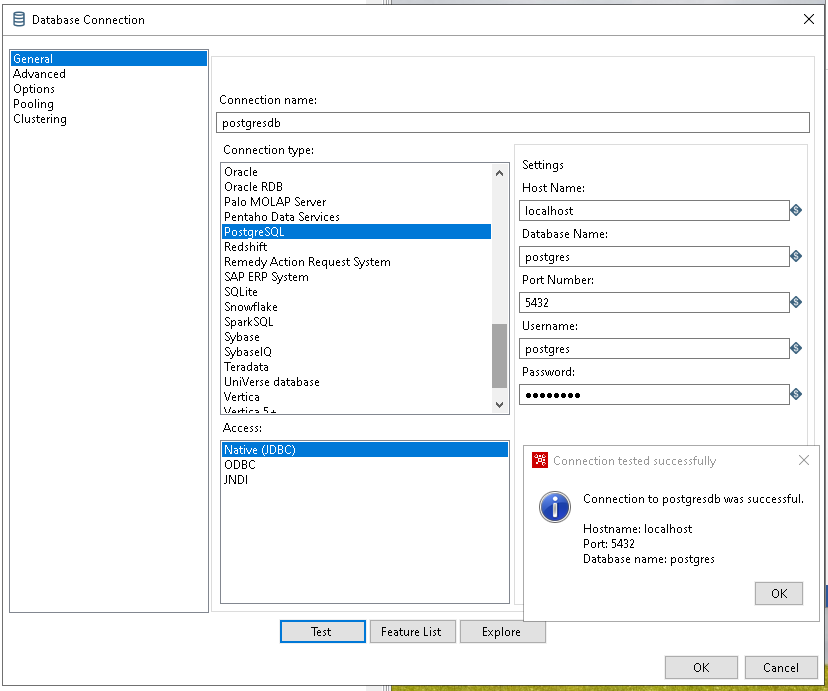
Kurzbeschreibung: Dimension; genauere Daten-Details

# Transformation Datenset in Sternschema (Python)

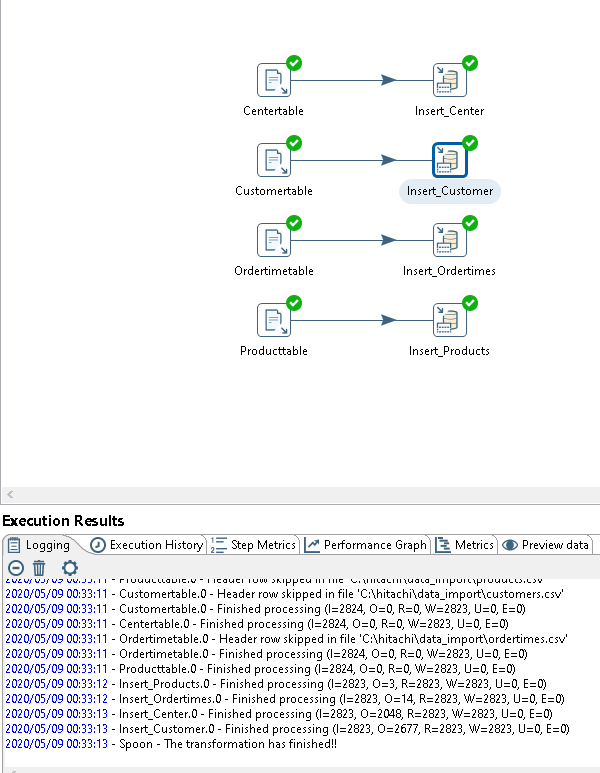
Der Datensatz enthielt ein paar Fehler, wichtig war die Auswahl des richtigen Encodings beim Import (latin-1). Die Daten wurden in Python vorweg aufbereitet, da erste Gehversuche in Pentaho aufgrund von Datenfehlern kläglich scheiterten. Dabei wurde der Datensatz in mehrere Tabellen aufgegliedert und diese jeweils über Keys verbunden.

# Import in Pentaho und Schreiben der Tabellen in postgres Datenbank

Durch die Vorbereitung in Python, war der Datenimport und die Übertragung in die Datenbank relativ einfach möglich.

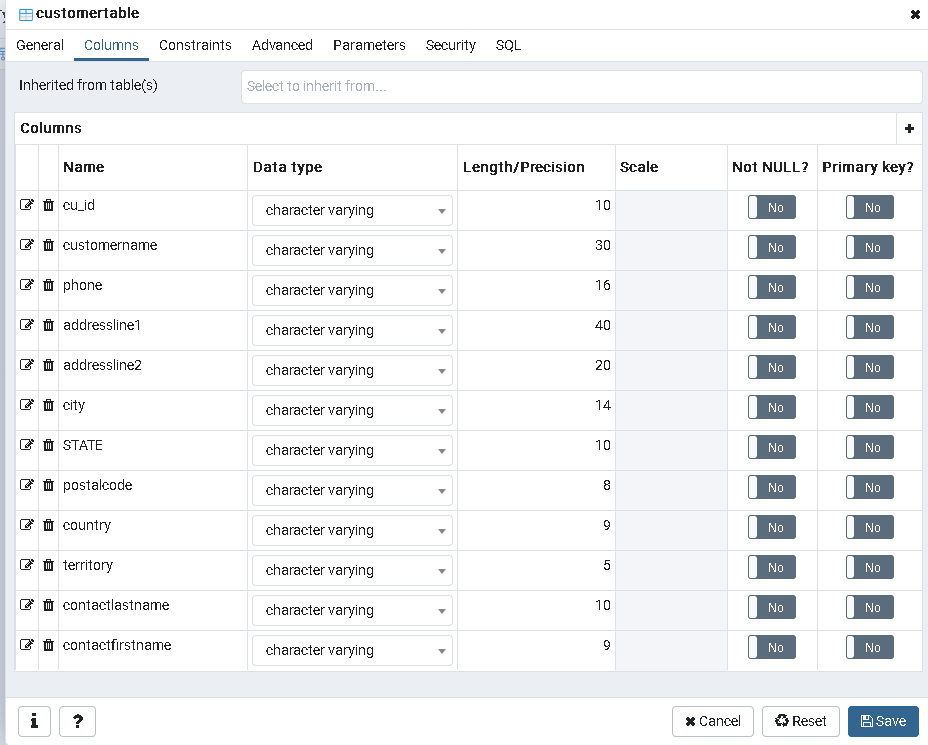


Der Übertragungsprozess hat nach der Anpassung der Spaltengrößen funktioniert.



Bei der Übertragung in die postgres-Datenbank waren viele Datentypen von Pentaho als zu klein spezifiziert.

Das musste in pgAdmin nachkorrigiert werden.



Sternschema

Starshape:

Orderdetails zentral

Satellit: Customer Data

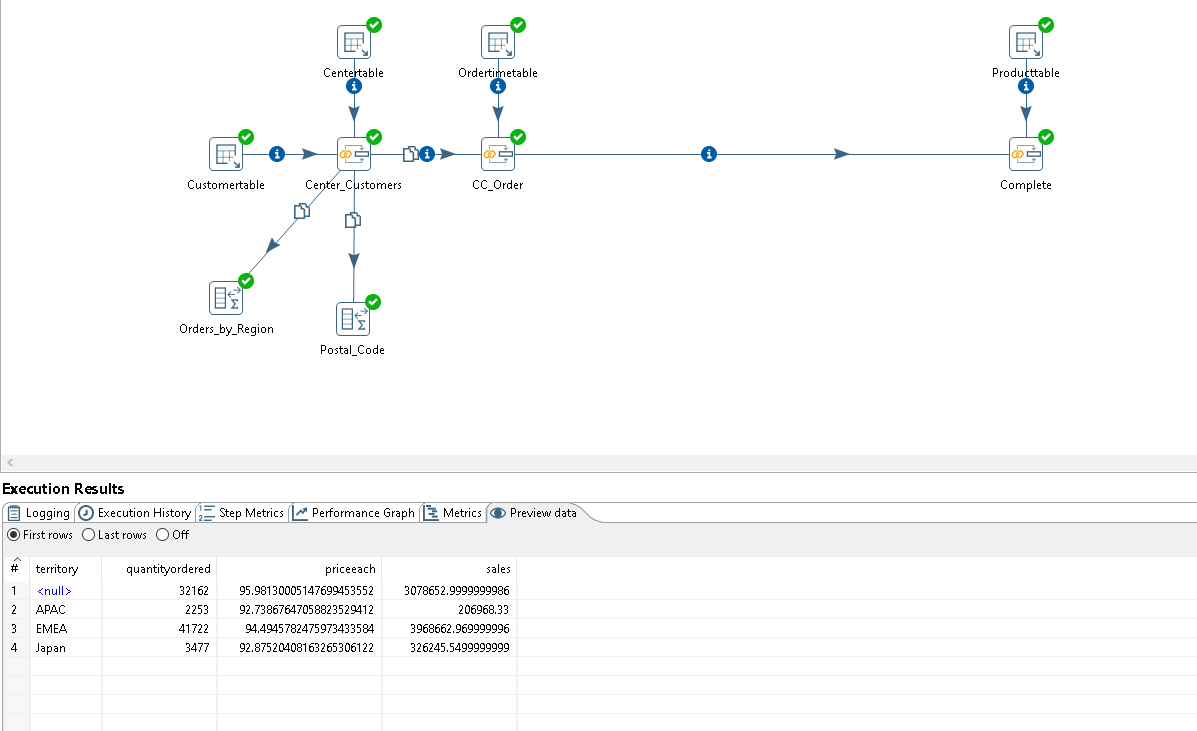
Satellit: Product Data

Satellit: Time; PK Orderdate

Gemäß der OLAP-Theorie wurde der Datensatz in mehrere Dimensionen in einem Sternschema aufgeteilt. Dabei wurden die wesentlichen Daten zu den Bestellungen (Orders) in einer Zentralen Tabelle gehalten. Es wurden drei zusätzliche Tabellen erzeugt: eine mit der zeitlichen Dimension, eine mit einer räumlichen Dimension (inklusive der Kundendaten, da die räumliche Verknüpfung über die Kundendaten erfolgte), und einer Detail-Ansicht zu den einzelnen Produkten.

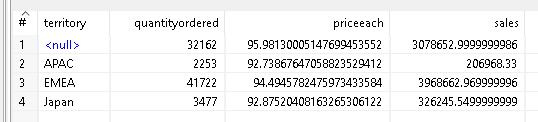
# Mehrdimensionale Analyse Datenset

Der Olap-Prozess wurde mehrstufig abgebildet. Zuerst wurden die Kundendaten zur Zentraltabelle gejoined, dann die zeitliche Dimension und abschließend die Produktdetails. Das ermögliche die mehrdimensionale Betrachtung im Prozessverlauf.

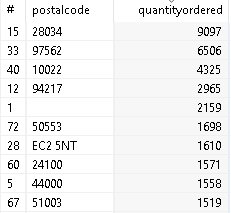


Mögliche Fragen:

Bestellungen/Durchschnittspreise und Verkäufe nach Region



Top 10 PLZS



Zeitverlauf

Welches Jahr generell am ertragreichsten?

Das war 2004.

Wo sind die Kunden?

Man könnte das auf die Postleitzahl herunterbrechen. Aber generell gab es am meisten KundInnen in der Region EMEA, gefolgt von Japan und APAC.

Welche Produkte sind am teuersten ?

Wo sind die Premium Customers?