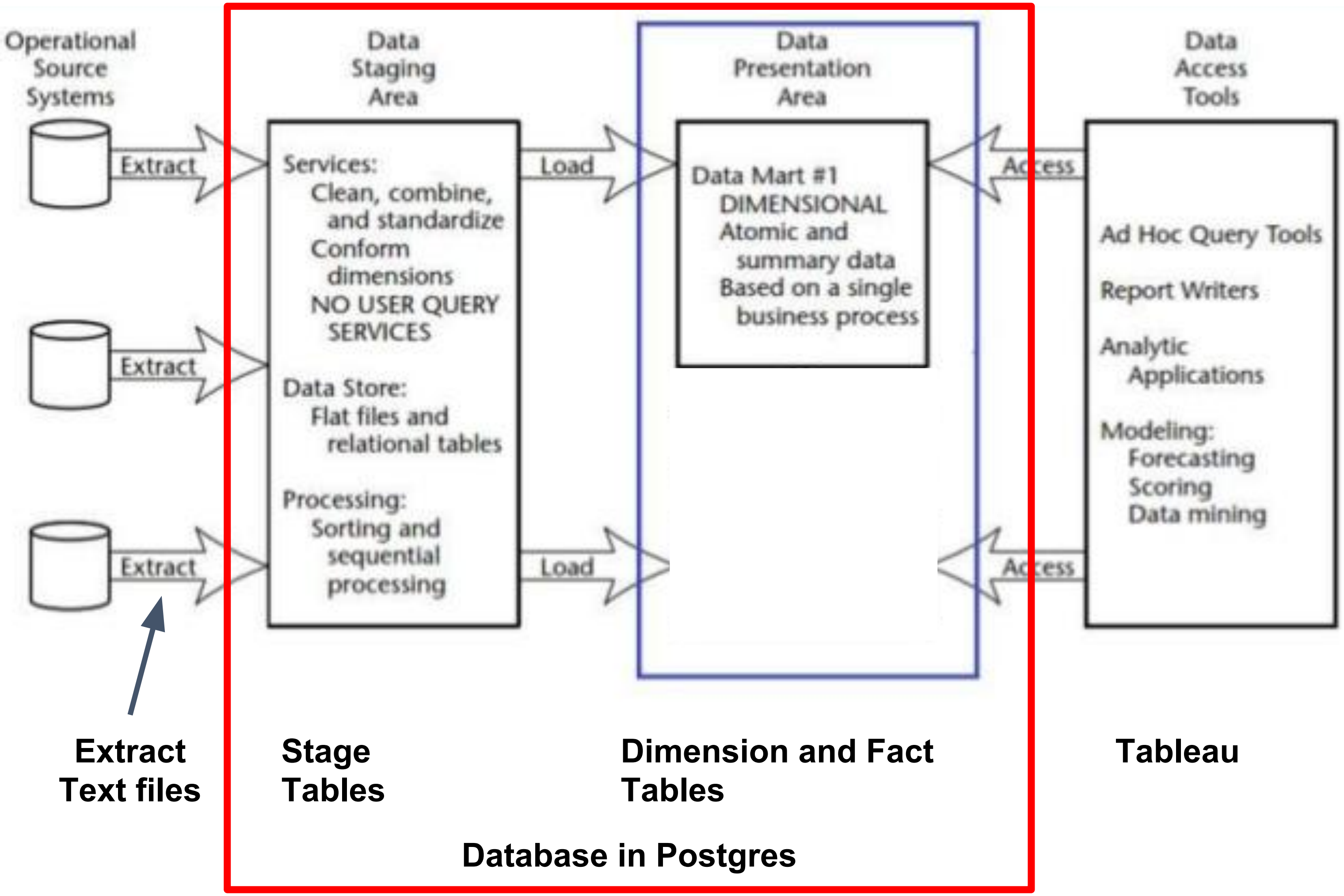
Building Sales Data mart Using Pentaho
Part 2

By Naheed Anjum Arafat

Sales Data mart: High Level View

Data mart Architecture



Extract Text files

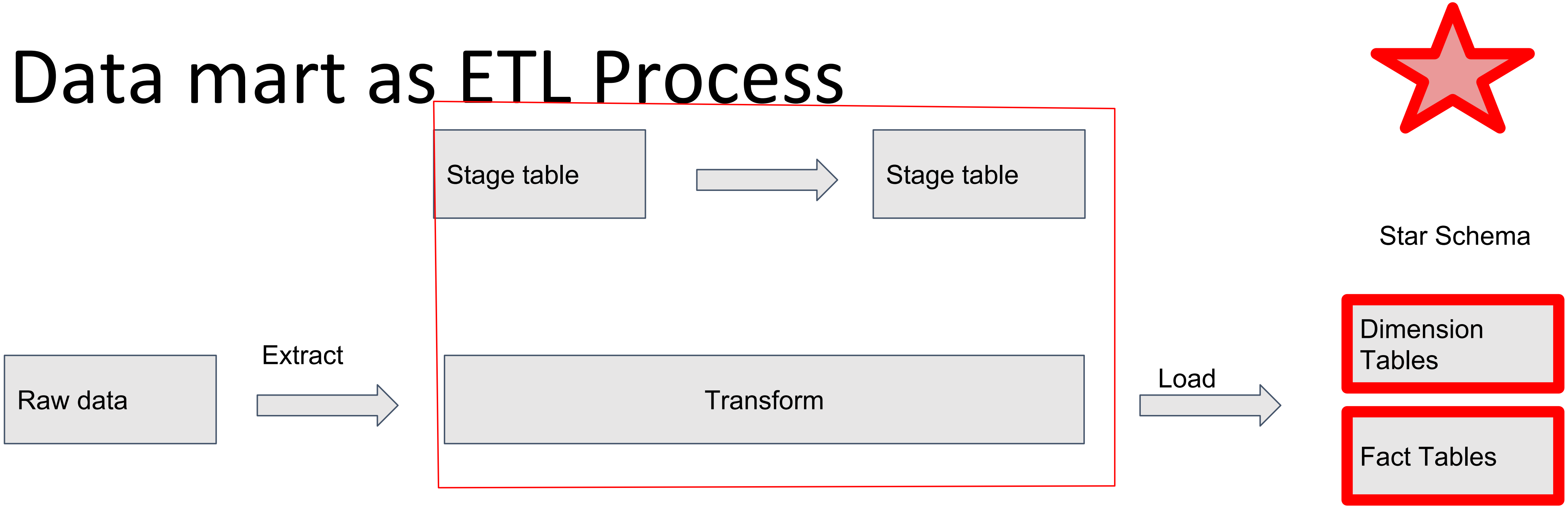
Stage Tables

Dimension and Fact Tables

Tableau

Database in Postgres

Data mart as ETL Process

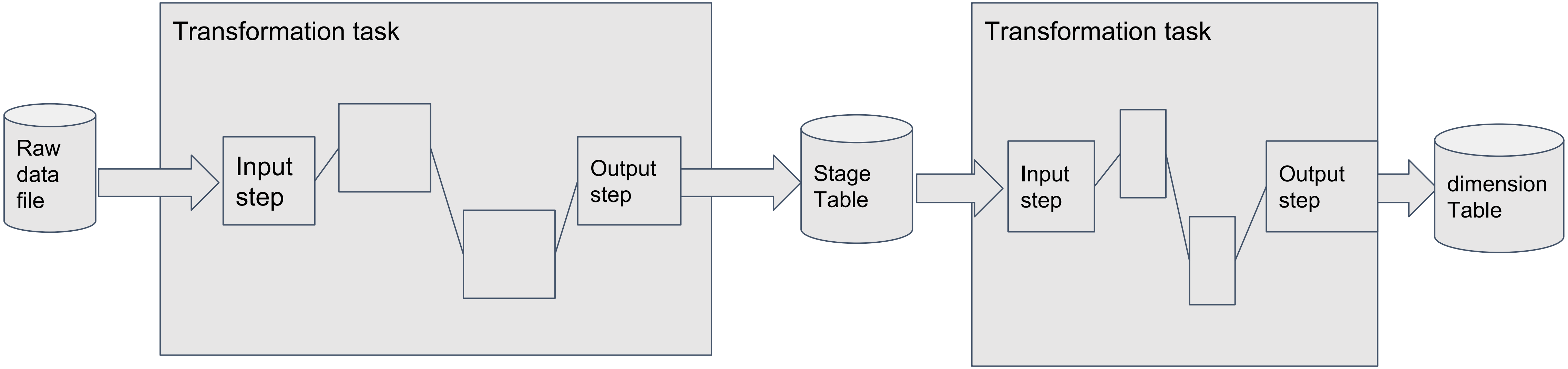


- Extract from source

- **Start fresh**
- Cleanse data
- Align to data type
- Field renaming
- Transformations
- Aggregates
- Splitting into tables

- **Persistent storage**
- Keys and measures
- Performance oriented
- Enforce Grain
- Enforce Unique keys

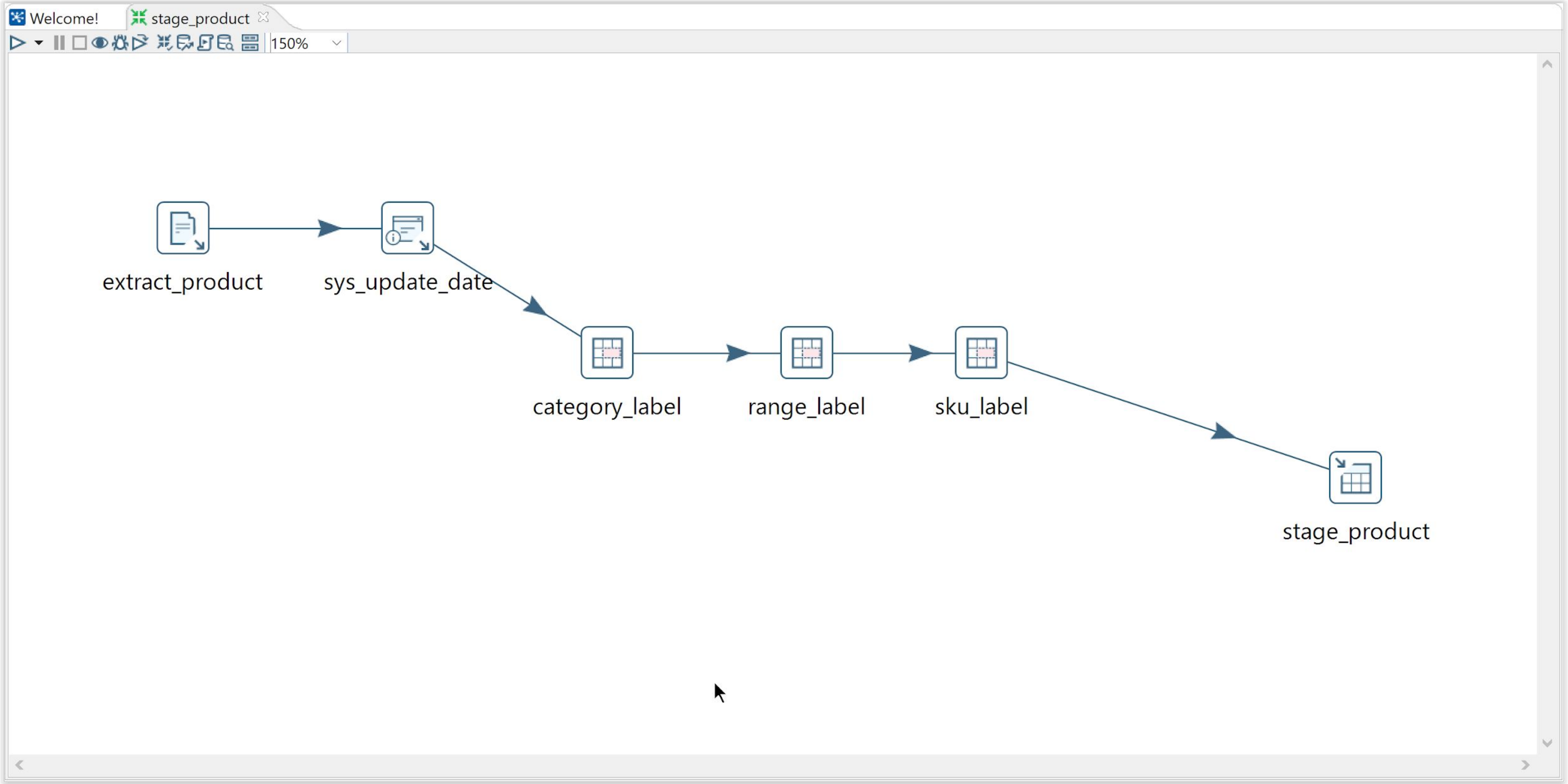
ETL Process (for dimensions) in Spoon



- Cleanse data
- Align to data type
- Field renaming
- Transformations
- Aggregates

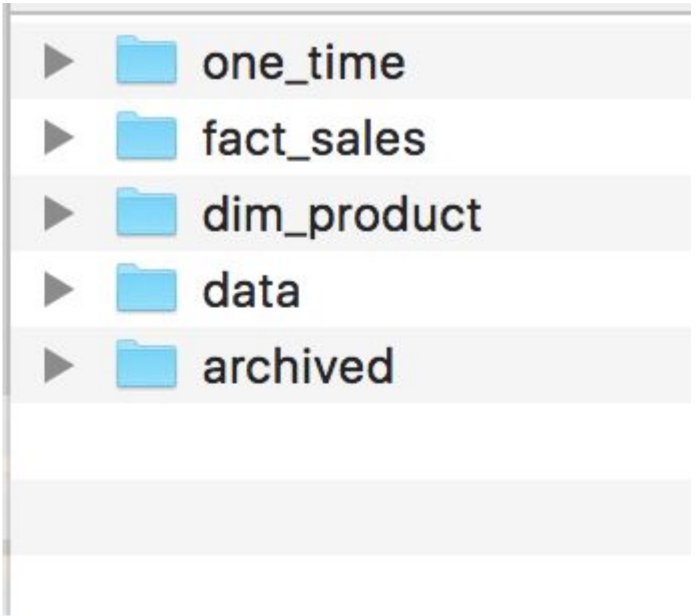
- Enforce Slowly changing Dimension functionality

Example of a transformation



Setup

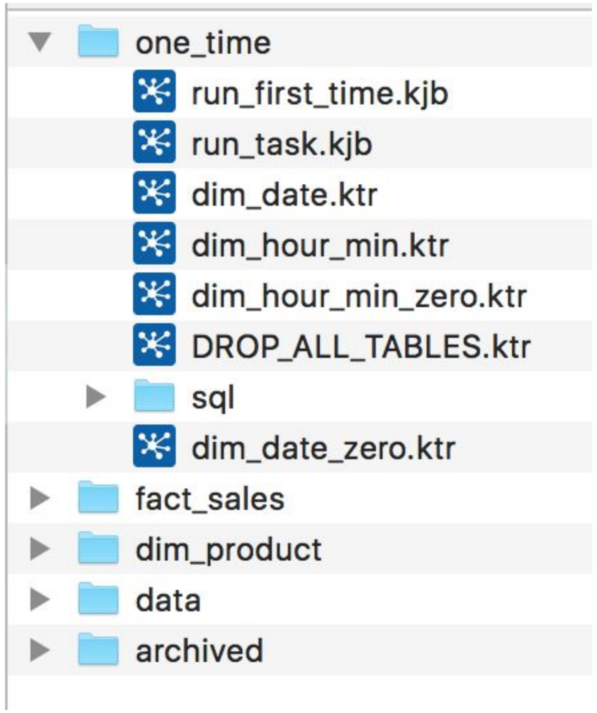
Contents of the lecture Pack



- one_time
- fact_sales
- dim_product
- data
- archived

- Transformation/jobs to be run once
- Task 2 folder
- Task 1 folder
- Datasets (to be used in class)
- Archive directory

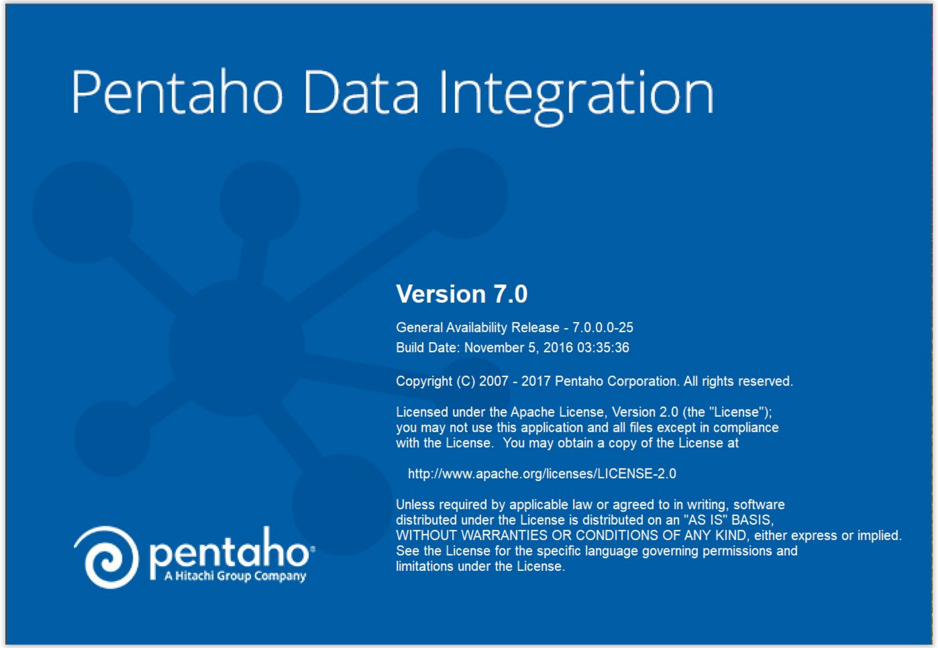
Contents of the lecture Pack



```
▼ one_time
      run_first_time.kjb
      run_task.kjb
      dim_date.ktr
      dim_hour_min.ktr
      dim_hour_min_zero.ktr
      DROP_ALL_TABLES.ktr
    ▶ sql
      dim_date_zero.ktr
▶ fact_sales
▶ dim_product
▶ data
▶ archived
```

Preparation

- Download the Lecture pack from IVLE
- Start up postgres server from manager-osx (Bitnami M/W/LAPP stack)
- Open PDI



Pentaho Data Integration

Version 7.0

General Availability Release - 7.0.0.0-25
Build Date: November 5, 2016 03:35:36

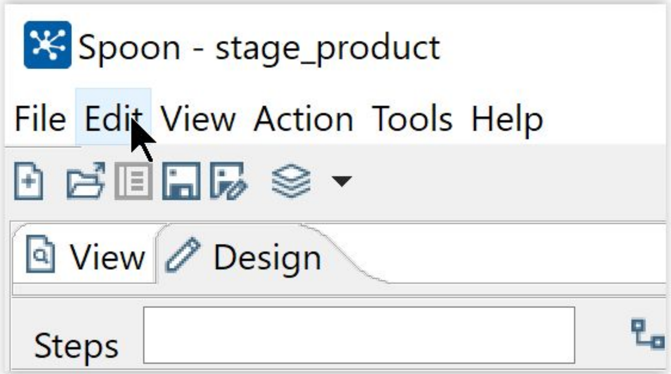
Copyright (C) 2007 - 2017 Pentaho Corporation. All rights reserved.

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this application and all files except in compliance with the License. You may obtain a copy of the License at

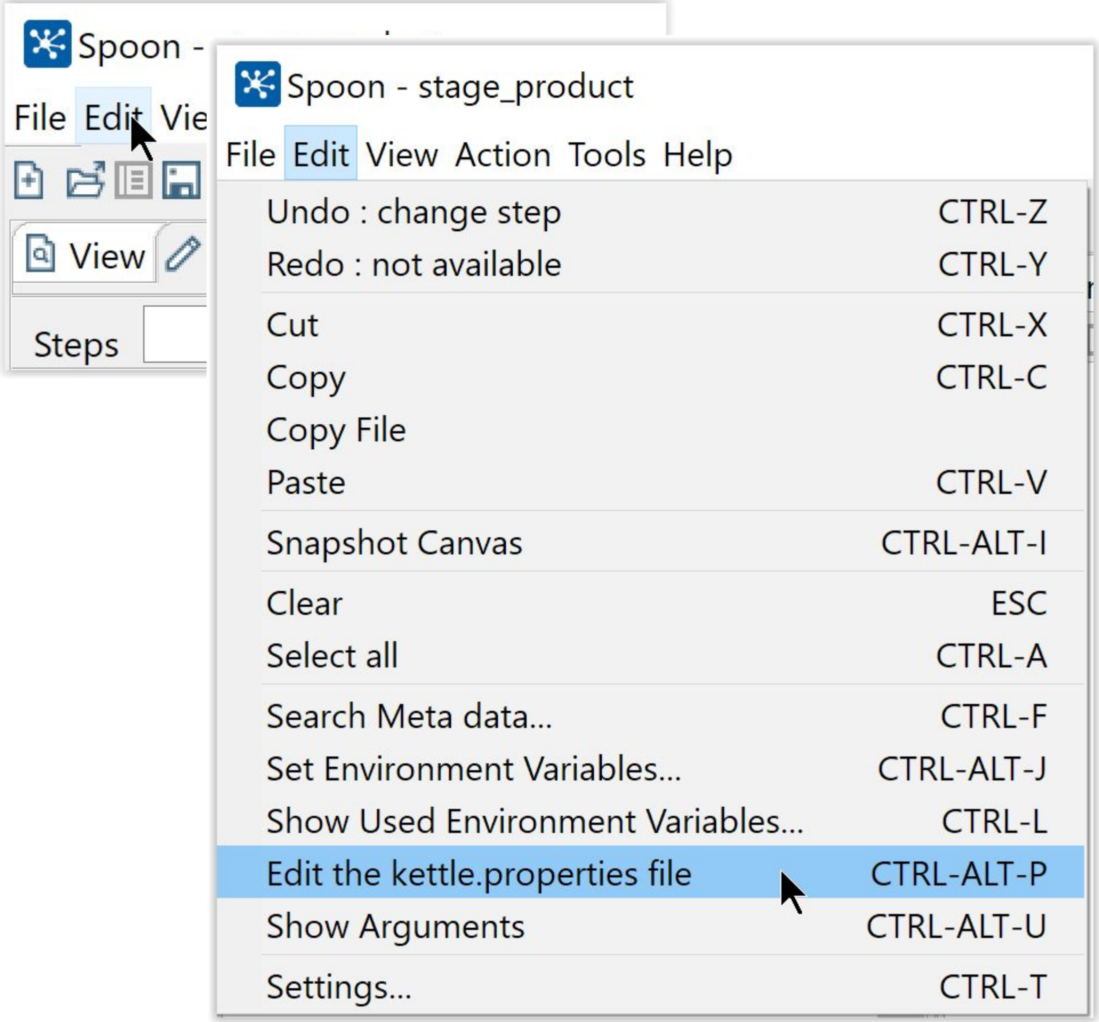
http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Global Variables



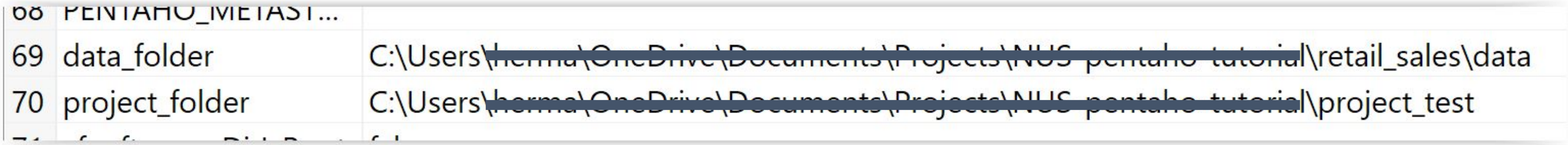
Global Variables


Spoon -
File Edit Vie
View
Steps
Spoon - stage_product
File Edit View Action Tools Help
Undo : change step CTRL-Z
Redo : not available CTRL-Y
Cut CTRL-X
Copy CTRL-C
Copy File
Paste CTRL-V
Snapshot Canvas CTRL-ALT-I
Clear ESC
Select all CTRL-A
Search Meta data... CTRL-F
Set Environment Variables... CTRL-ALT-J
Show Used Environment Variables... CTRL-L
Edit the kettle.properties file CTRL-ALT-P
Show Arguments CTRL-ALT-U
Settings... CTRL-T

Global Variables

`data_folder`

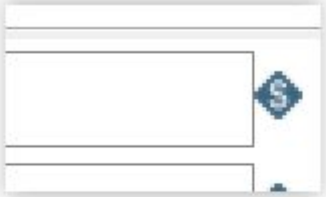
`project_folder`



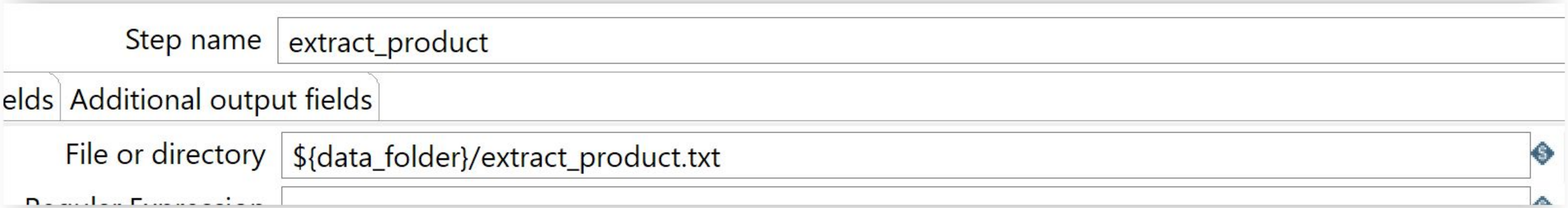
68	PENTAHO_METAST...	
69	data_folder	C:\Users[redacted]\retail_sales\data
70	project_folder	C:\Users[redacted]\project_test

Use variable `${data_folder}` or `${project_folder}`

Any dialog field with icon



How to use Global Variables



Step name: extract_product

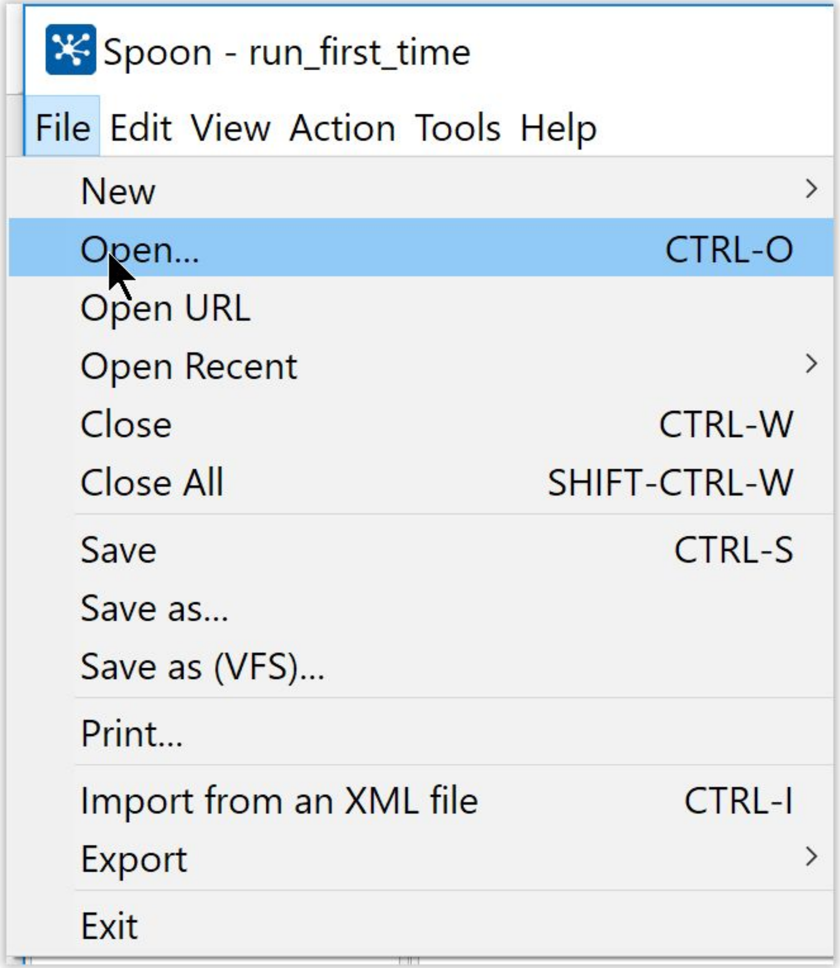
File or directory: ${data_folder}/extract_product.txt

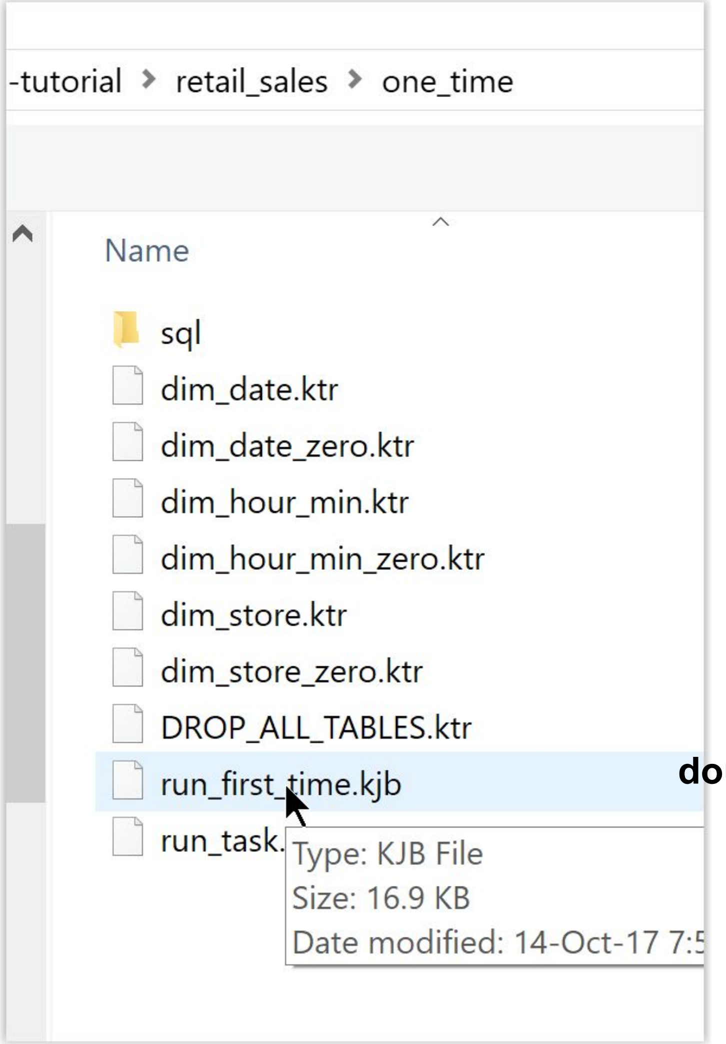
Regular Expression

Preparation

- Make sure you have set up data_folder and project_folder variables
- Open “run_first_time.kjb” Job in Spoon

Preparation



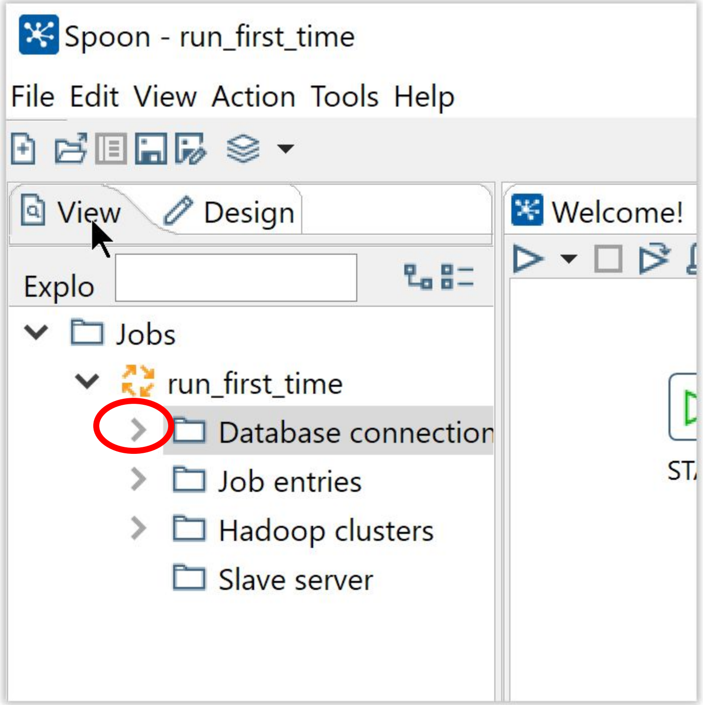


Type: KJB File
Size: 16.9 KB
Date modified: 14-Oct-17 7:5

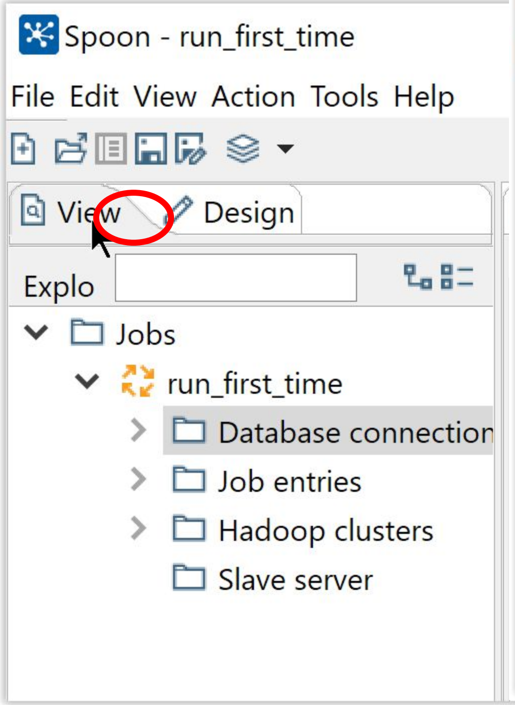
double click

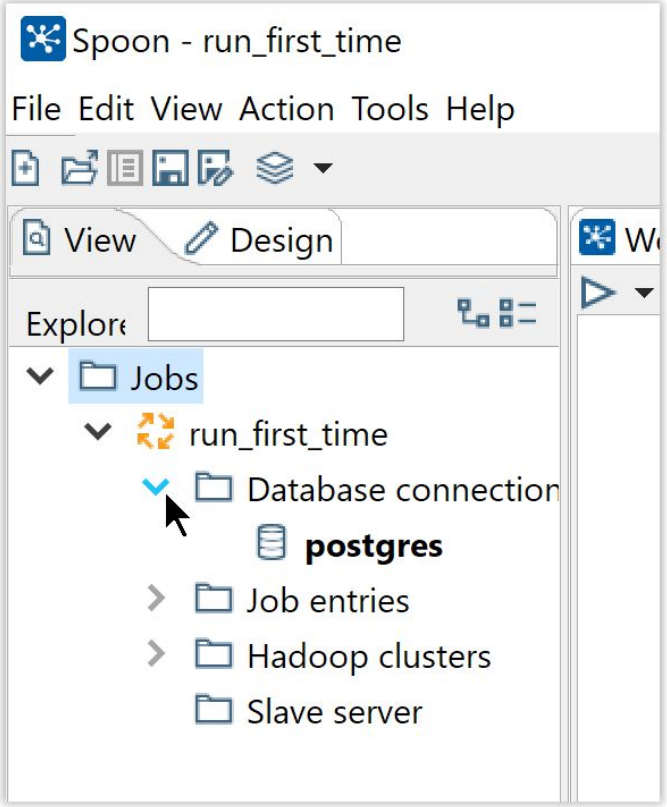
Connect to Postgres

Set up Database Connection

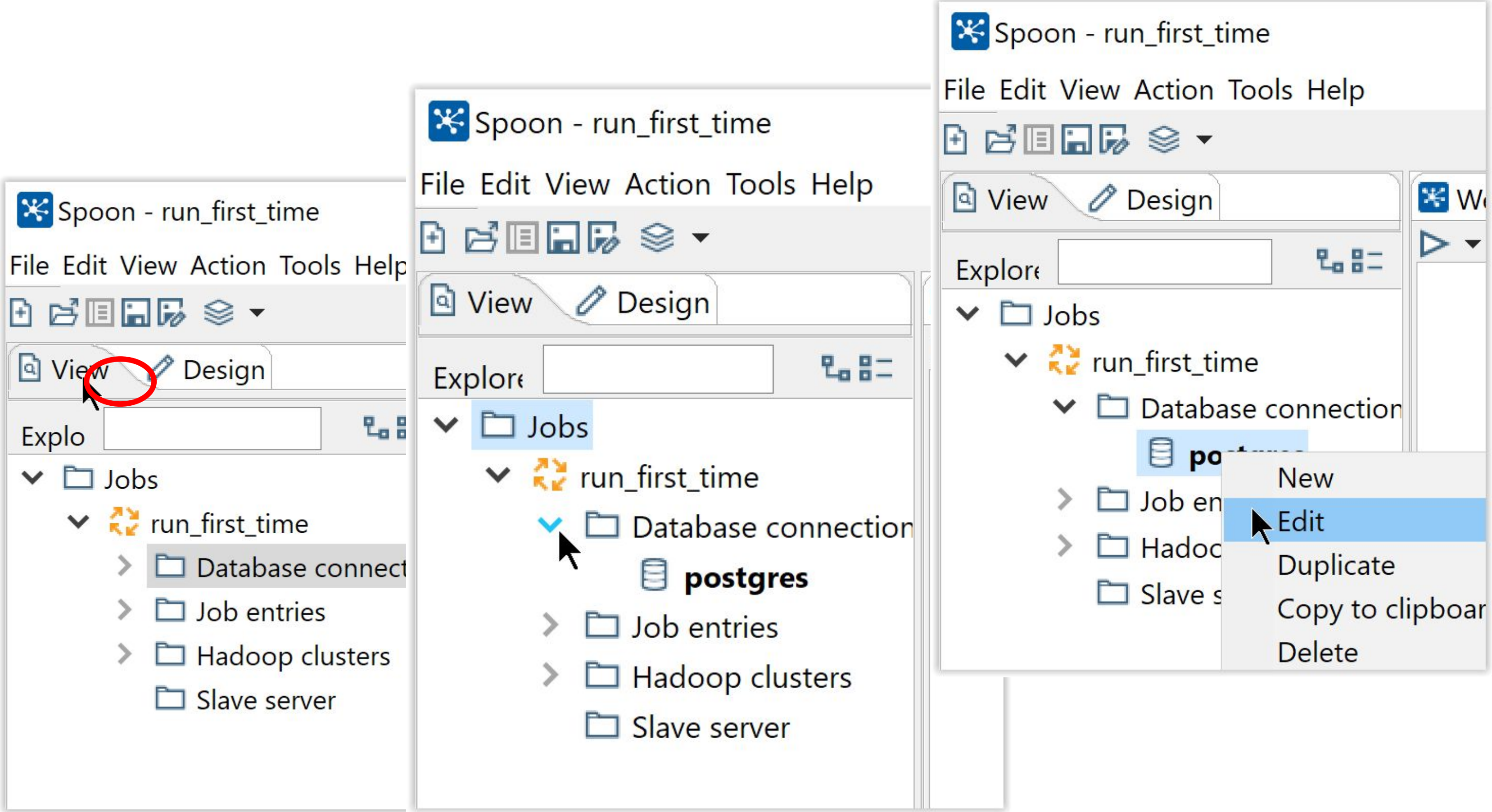


Set up Database Connection





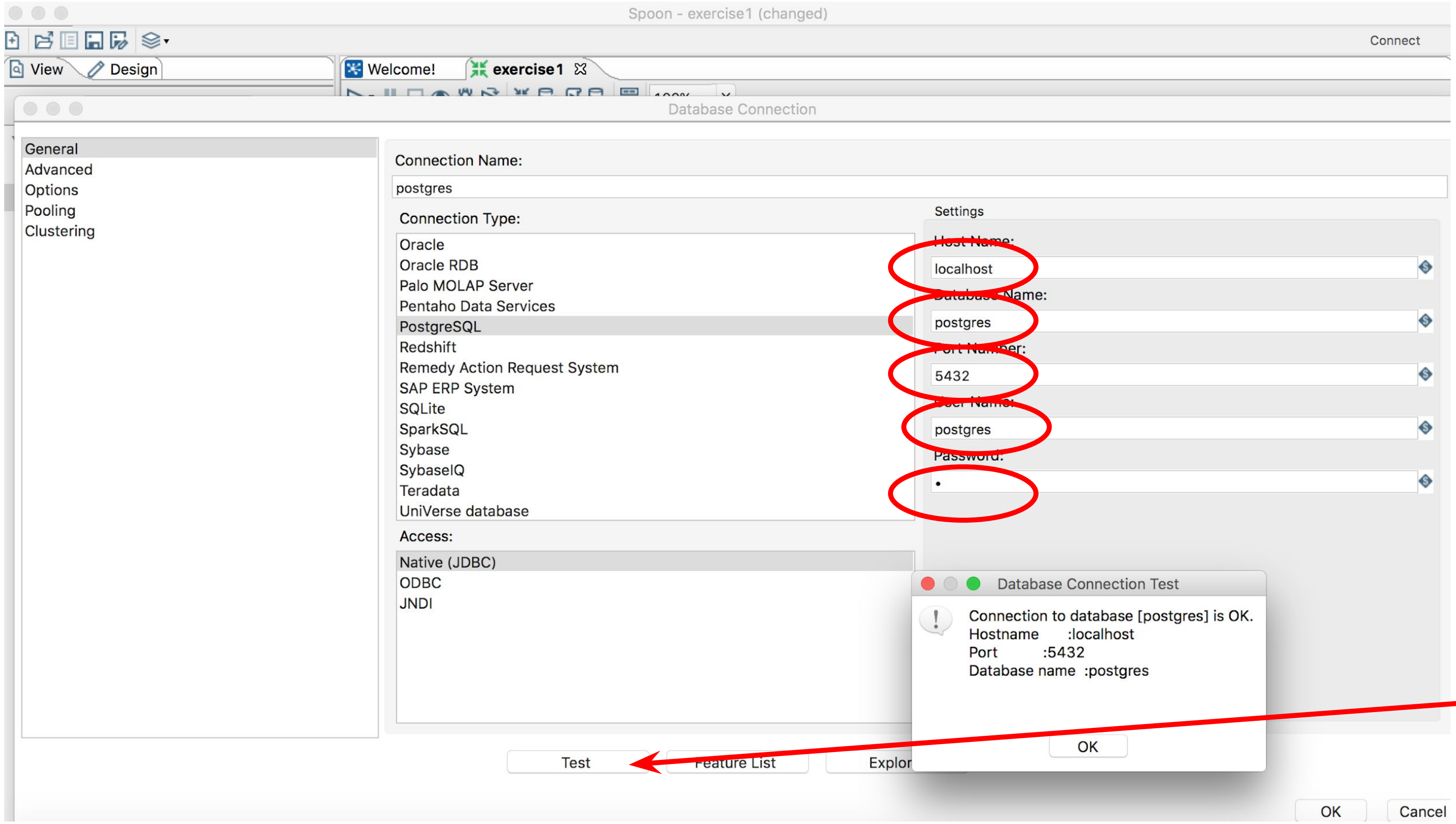
Set up Database Connection



If you don’t see any entry under “Database connection”:-

Double Click on “Database connection” to create one

Edit/Test Database Connection



Connection Name:
postgres

Connection Type:
Oracle
Oracle RDB
Palo MOLAP Server
Pentaho Data Services
PostgreSQL
Redshift
Remedy Action Request System
SAP ERP System
SQLite
SparkSQL
Sybase
SybaseIQ
Teradata
UniVerse database

Access:
Native (JDBC)
ODBC
JNDI

Settings
Host Name: localhost
Database Name: postgres
Port Number: 5432
User Name: postgres
Password: •

Database Connection Test

Connection to database [postgres] is OK.
Hostname :localhost
Port :5432
Database name :postgres

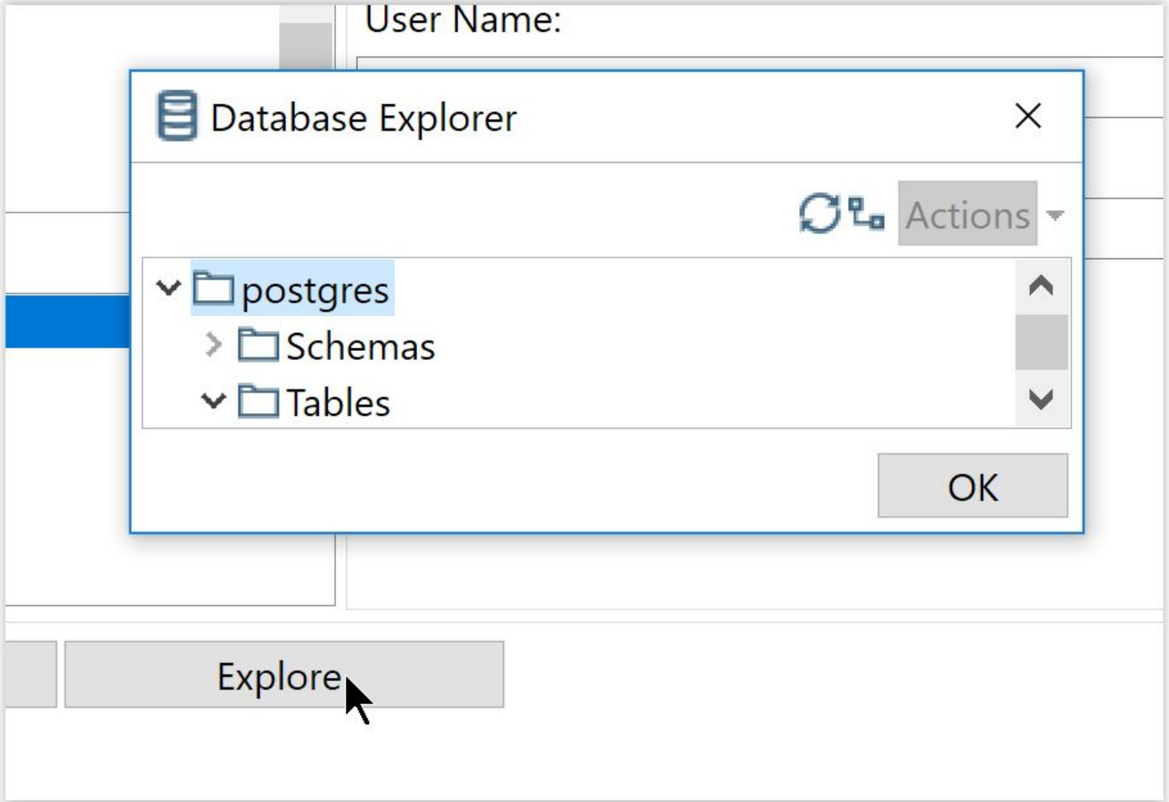
Set the values of the following configuration fields:

Host Name:
Database Name:
Port Number:
User Name:
Password:

Set it according to your configuration of the Postgres installation

Test the connection

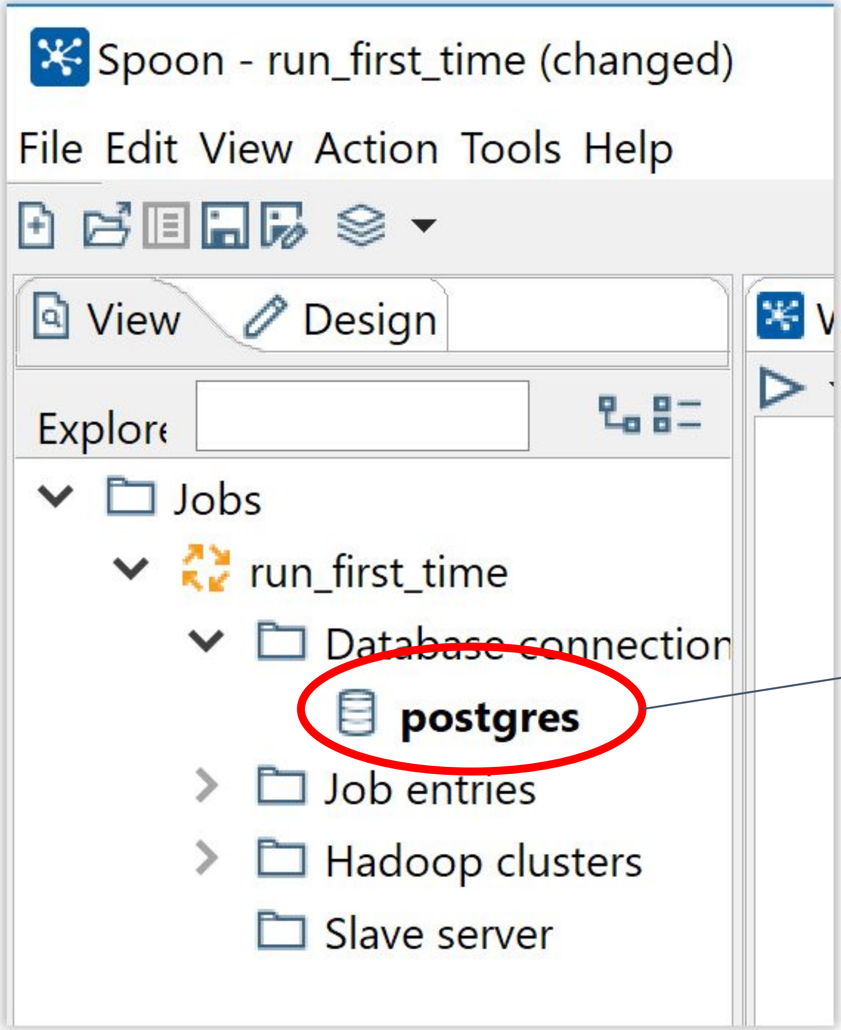
Check the contents of Database



Navigate to Tables
See that the database is where you want
the data mart tables to be

Click on Explore

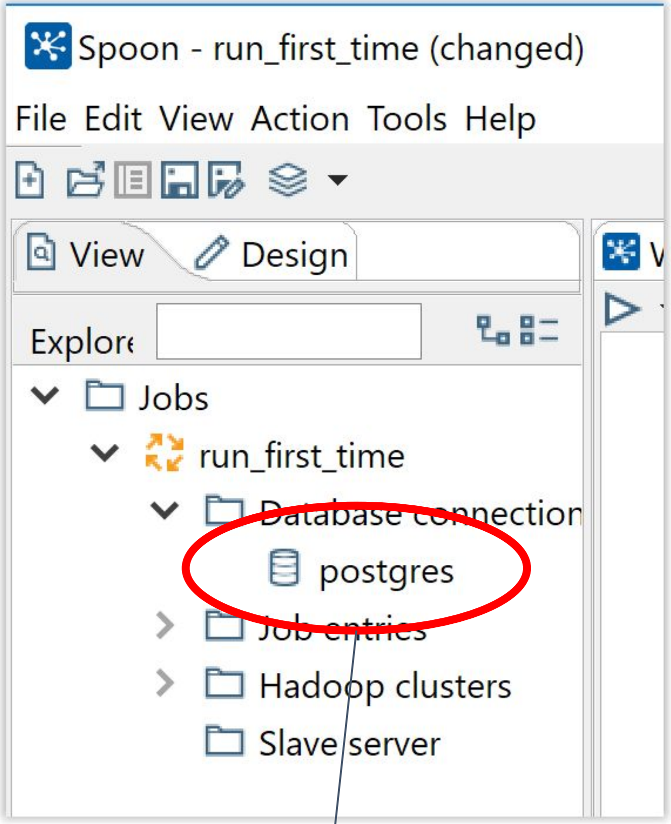
Share Database Connection



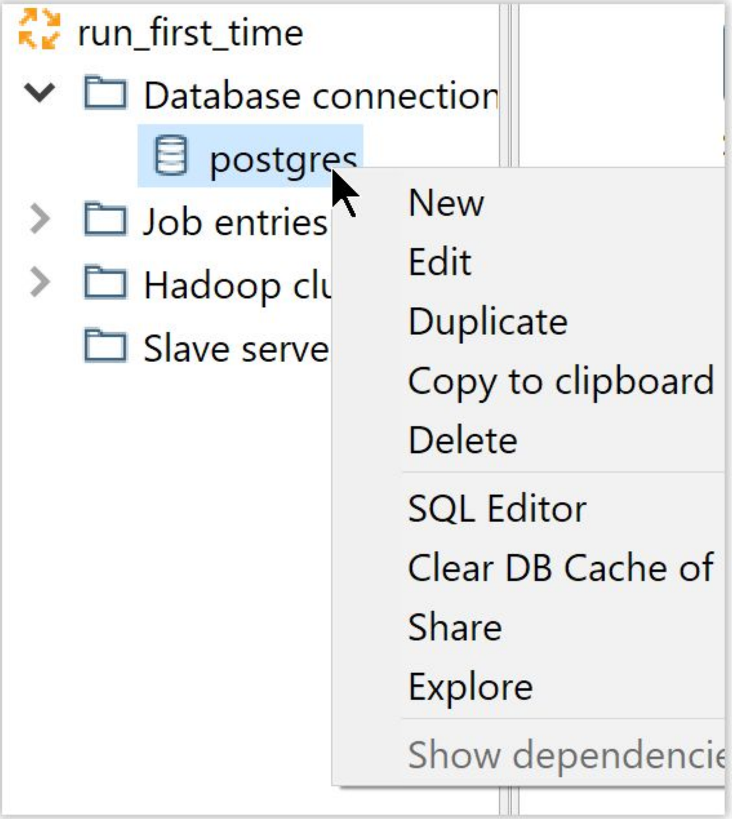
Sharing a connection enables
all transformations and jobs to use the same database.

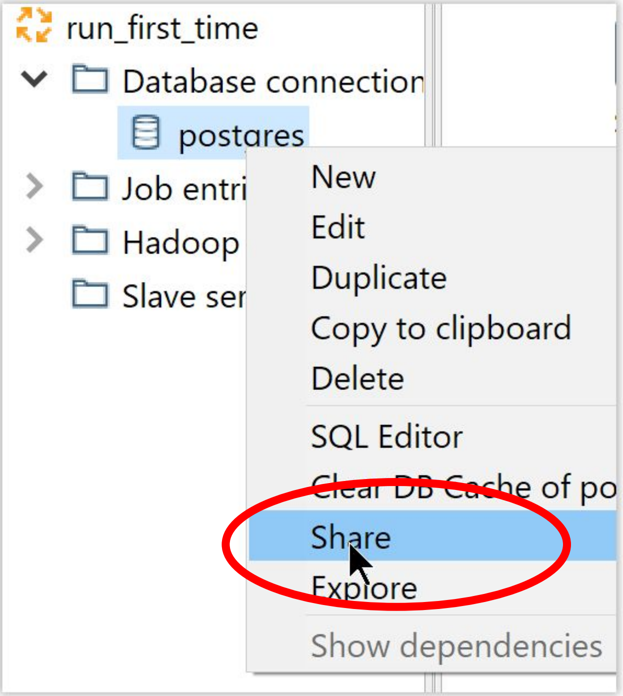
If the Database Connection is shared, it will be in **BOLD**

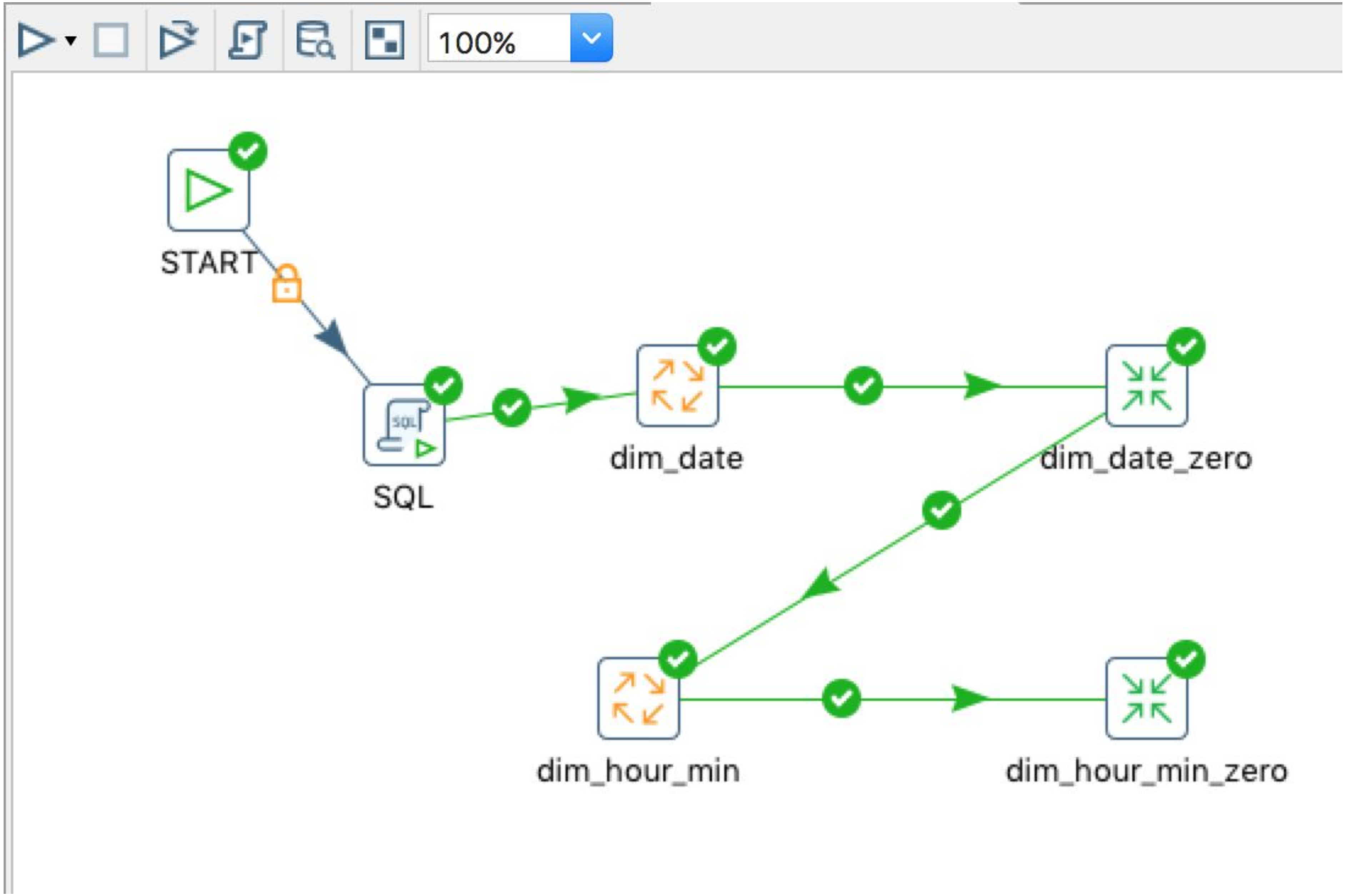
How to set Connection to “Shared”



The database connection is not in BOLD so it is not shared and will not be accessed by other transformations and jobs

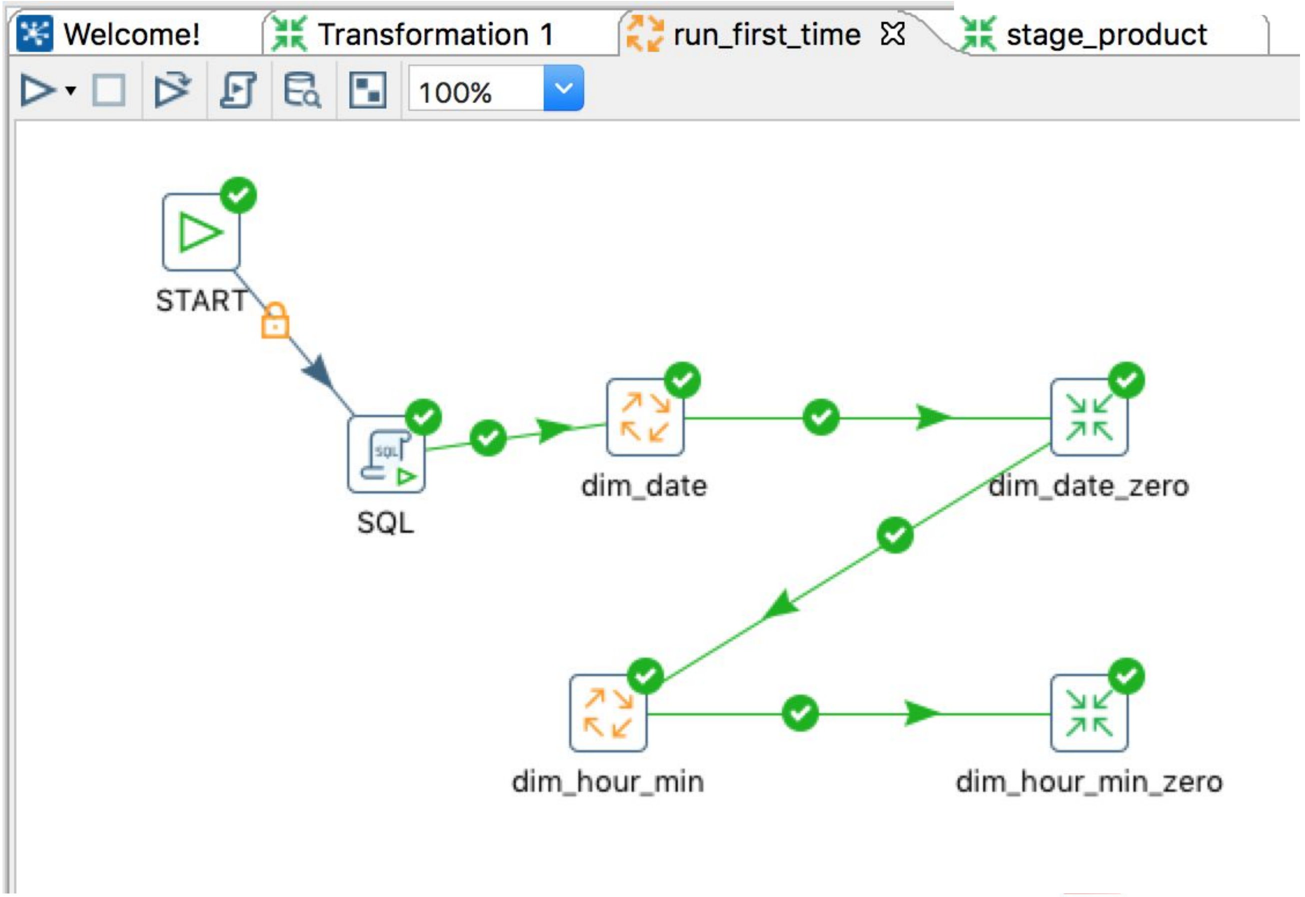




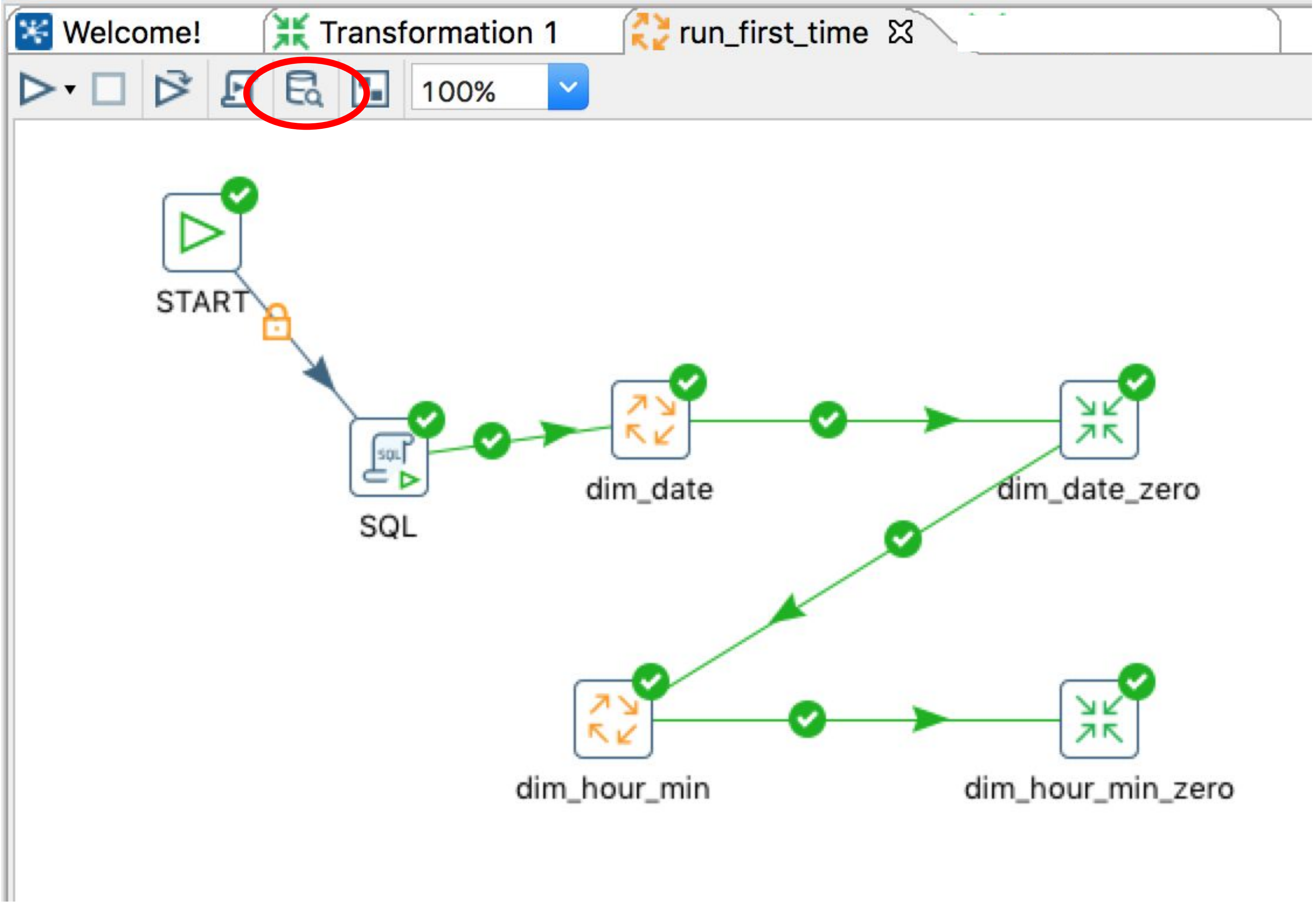

100%
START
SQL
dim_date
dim_date_zero
dim_hour_min
dim_hour_min_zero

Run First Time Job

- Creates 2 dimensions for you
  - date dimension
  - hour-minute dimension



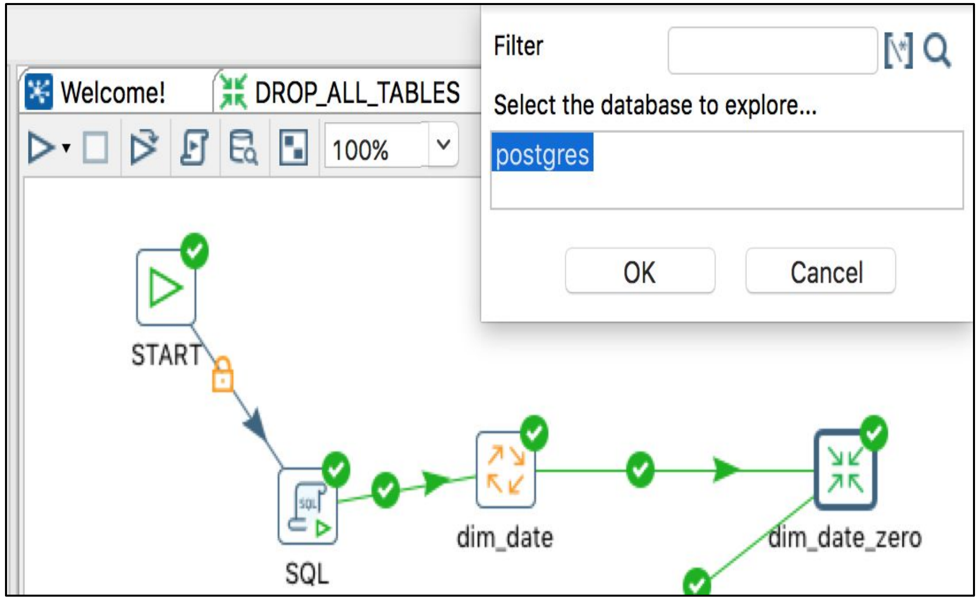
Run the Job



Check the database using the Explore option.

You should see 2 tables.

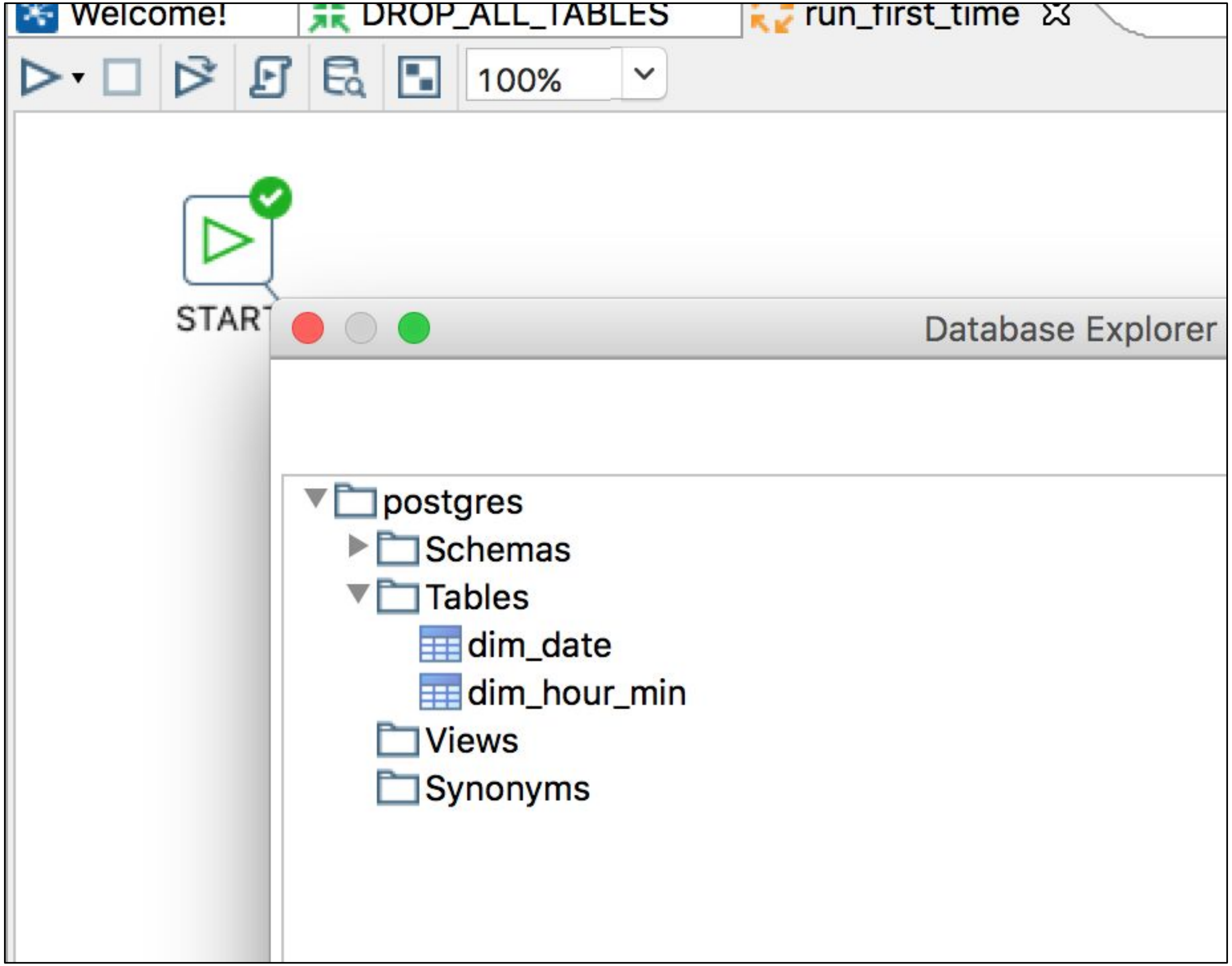
Check Database



Filter

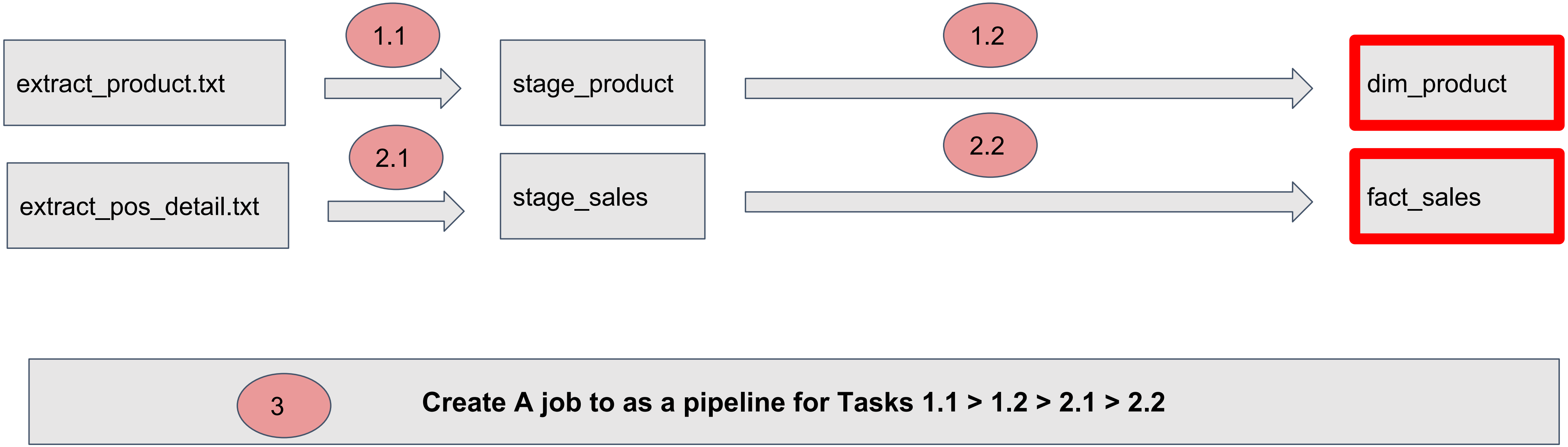
Select the database to explore...

postgres



- postgres
  - Schemas
  - Tables
    - dim_date
    - dim_hour_min
  - Views
  - Synonyms

Todays tasks



Create A job to as a pipeline for Tasks 1.1 > 1.2 > 2.1 > 2.2