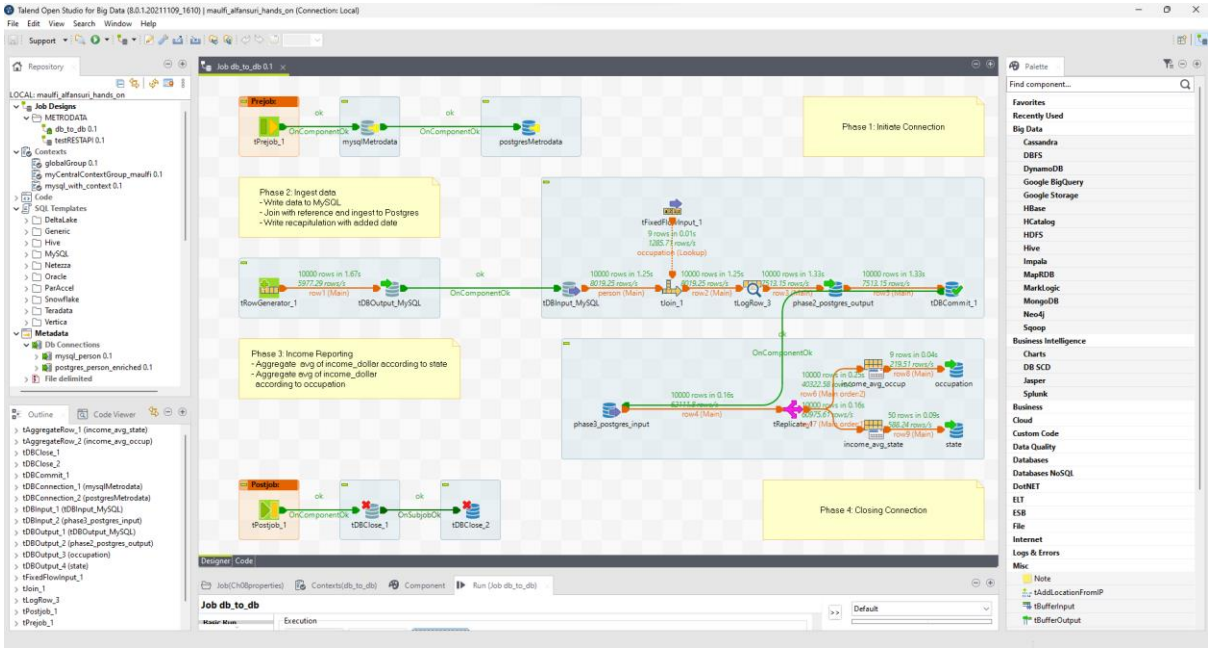
TALEND HANDS ON

MAULFI ALFANSURI

1. Overview



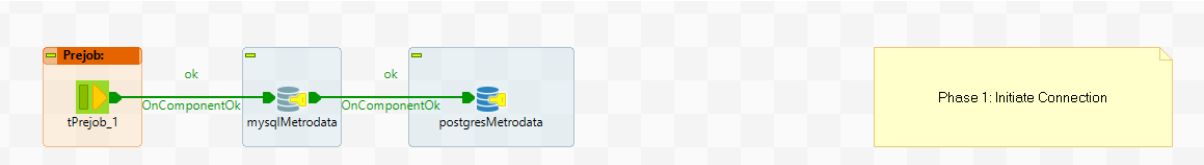
Pada kesempatan ini, akan didesain sebuah *use case* ETL terhadap 10 ribu data penduduk amerika serikta yang memuat *id*(kode unik), *lastName*(nama akhir), *state*(negara bagian), *income*(pendapatan) dan *occupation*(profesi).

ETL *pipeline* yang dikembangkan ditujukan untuk mengambil data agregasi nilai rata rata pendapatan dari 2 sisi -baik dari negara bagian maupun profesi masing masing-, yang masing masing termuat kedalam dua tabel di postgresql (dari container Docker) yaitu INCOME_AVG_STATE dan INCOME_AVG_OCCUPATION.

Terdapat 4 tahap utama yang dilalui diantaranya adalah:

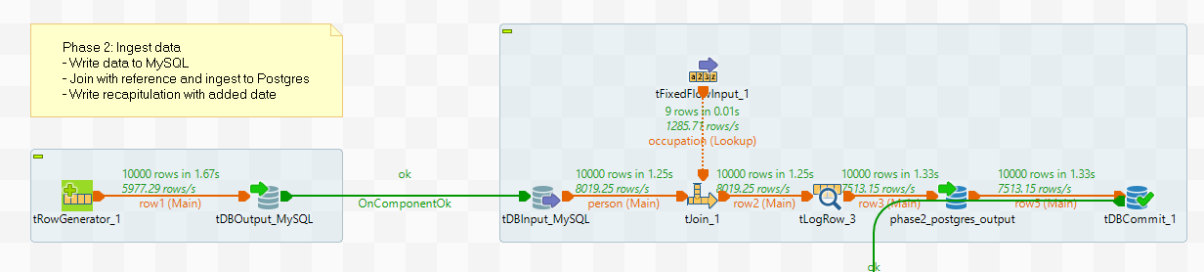
- Phase 1: Inisiasi koneksi terhadap DB MySQL dan PostgreSQL
- Phase 2: Manufaktur data dan *ingestion* kedalam MySQL, transformasi join dan *ingestion* ke dalam PostgreSQL
- Phase 3: Transformasi aggregasi terhadap tabel fase 2 postgres kedalam 2 tabel income_avg_state dan income_avg_occupation
- Phase 4: Penutupan koneksi DB

1. Komponen *Job*:



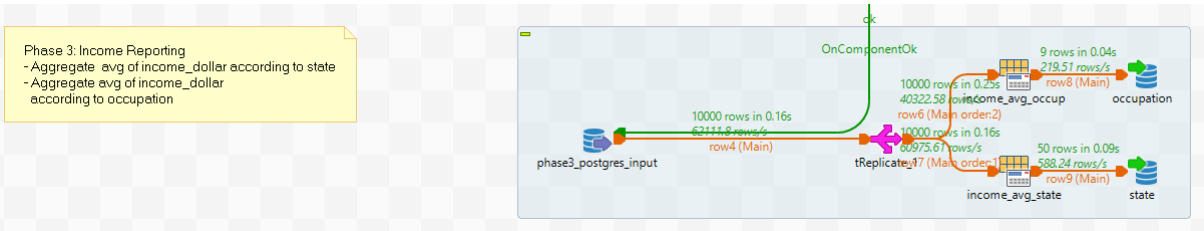
Komponen fase 1:

- tPreJob
- tDBConnection(mysqlMetrodata & postgresqlMetrodata): Untuk inisiasi koneksi MySQL



Komponen fase 2:

- tRowGenerator: Untuk manufaktur data
- tDBOutput_MySQL: Untuk penulisan data kedalam tabel MySQL
- tDBInput_MySQL: Untuk pengambilan data dari tabel MySQL
- tFixedFlowInput: Untuk data referensi *occupation code – occupation*
- tJoin_1: Untuk transformasi join.
- tLogRow_3: Untuk mendapatkan hasil sebelum penulisan kedalam postgres
- tDBCommit_1: Untuk melakukan *commit* terhadap transaksi yang dibikin



Komponen fase 3:

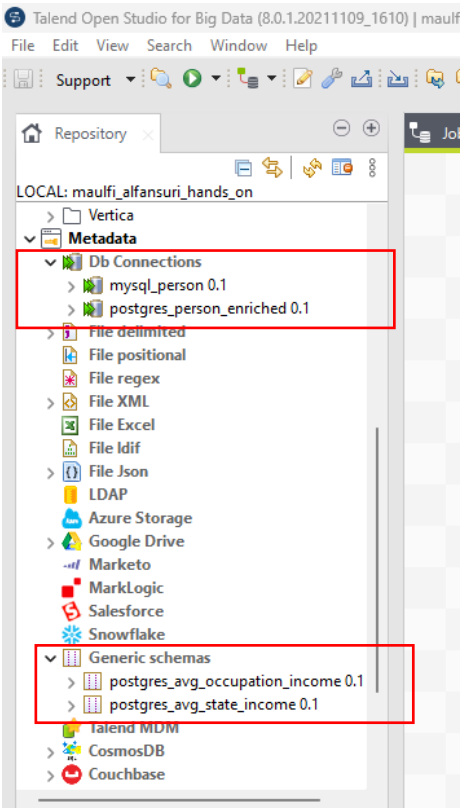
- phase3_postgres_input: Untuk mengambil data dari postgres (tDBInput)
- tReplicate: Untuk memungkinkan agregasi secara simultan kedalam beberapa tabel berbeda
- ‘occupation’ & ‘state’: Untuk penulisan tabel aggregasi kedalam postgres

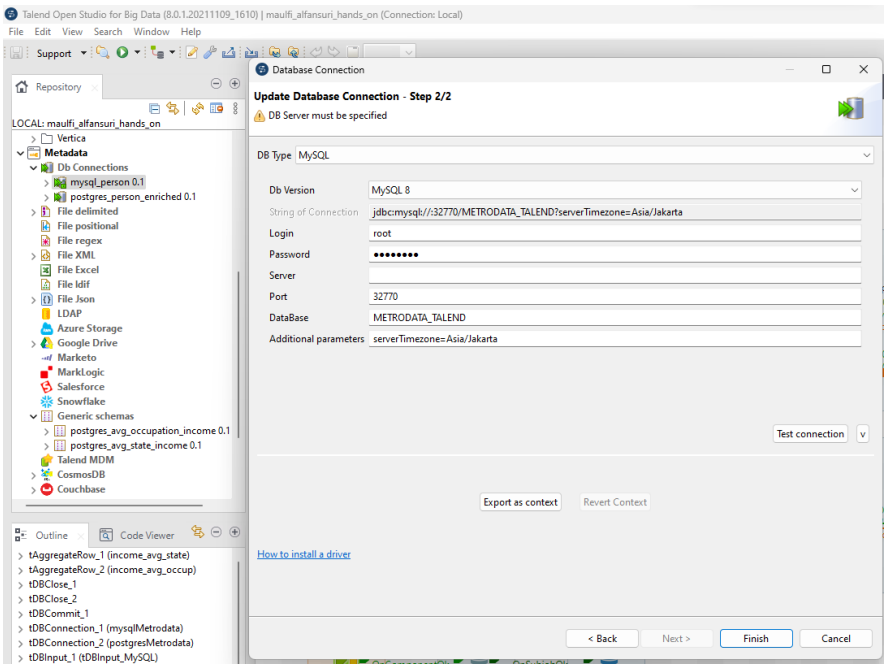
Komponen fase 4:

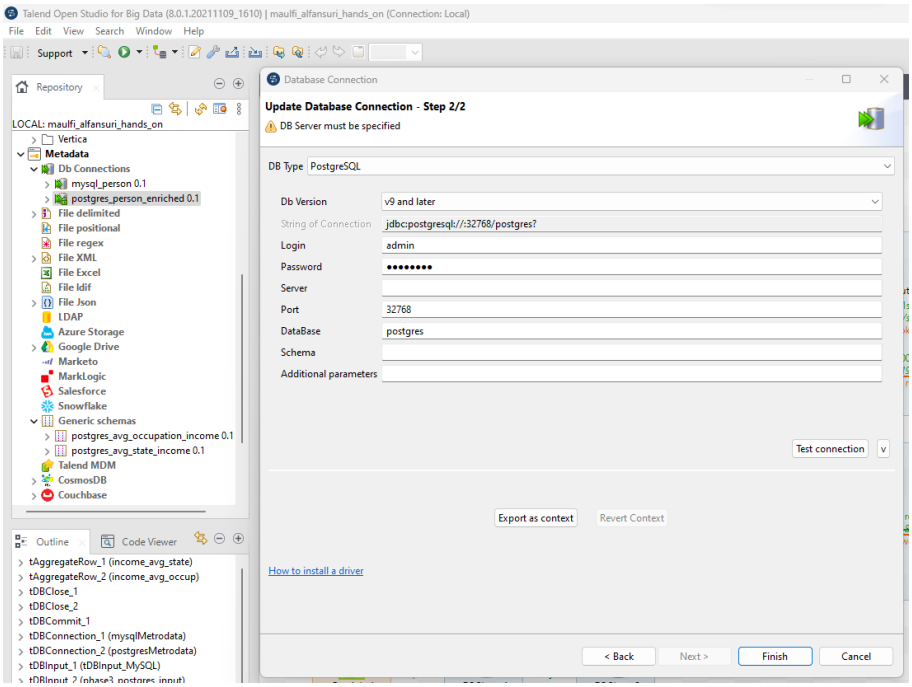
- tPostJob: Untuk finalisasi paska selesainya pekerjaan (*job*)
- tDBClose_1 & tDBClose_2: Untuk menutup koneksi ke 2 DB

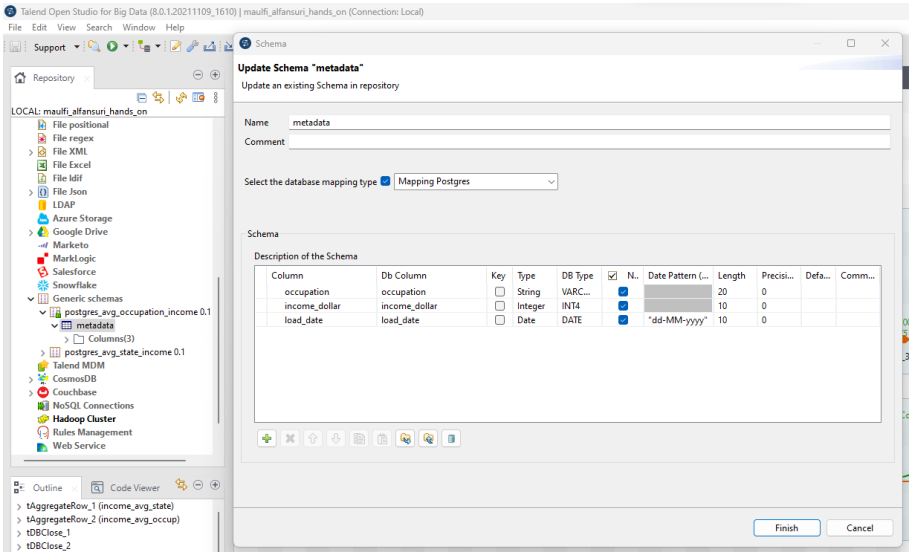
2. Komponen metadata:

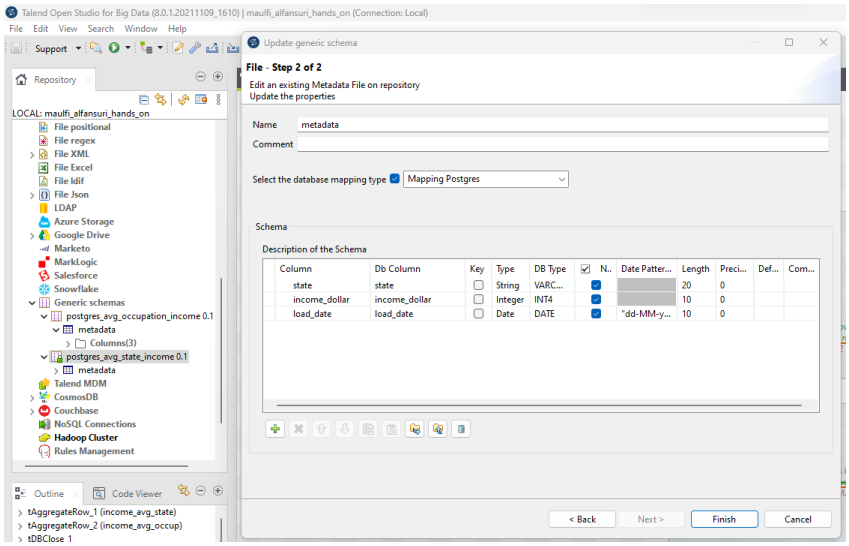
Beberapa konfigurasi seperti koneksi dan schema ditaruh kedalam metadata untuk memudahkan pembuatan ulang komponen dan meminimalisasi *built-in*. Ada 2 metadata koneksi untuk 2 tabel awal -tabel person di MySQL dan tabel person_enriched di PostgreSQL-. Dan 2 metadata schema untuk 2 tabel akhir -tabel income_avg_occupation dan income_avg_state.





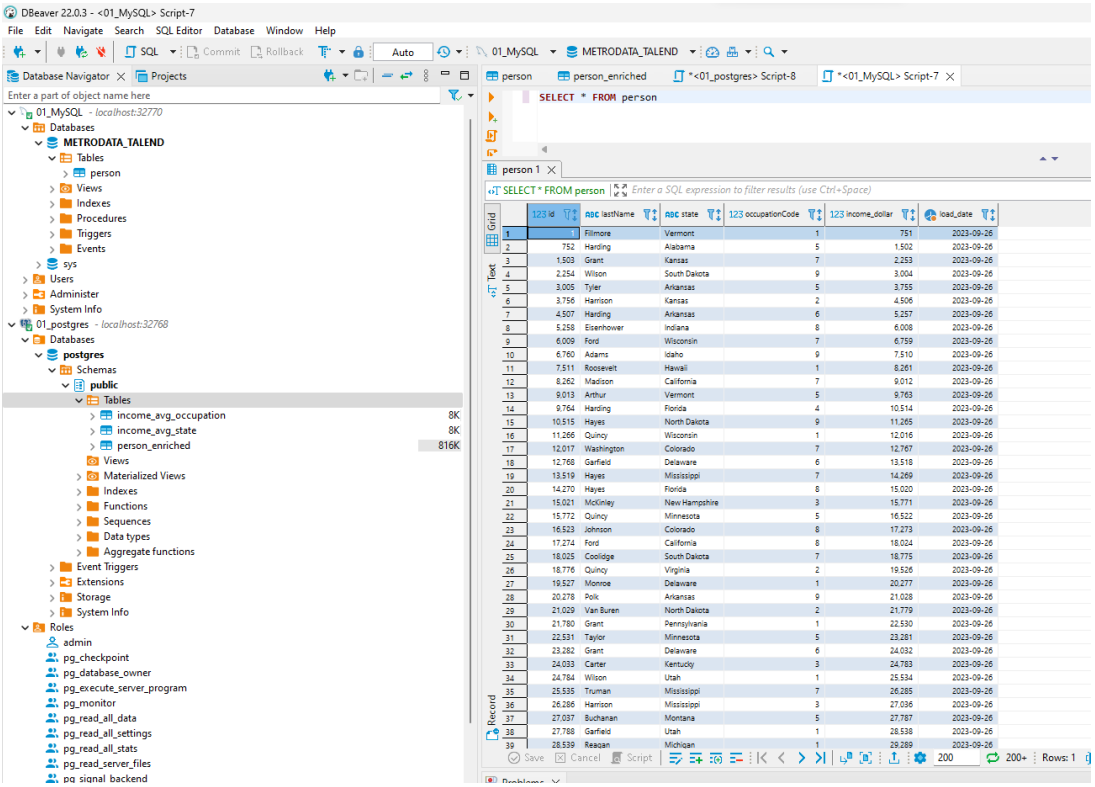

Talend Open Studio for Big Data (8.0.1.20211109_1610) | maulfi_alfansuri_hands_on (Connection: Local)
File Edit View Search Window Help
Support
Repository
LOCAL: maulfi_alfansuri_hands_on
Vertica
Metadata
Db Connections
mysql_person 0.1
postgres_person_enriched 0.1
File delimited
File positional
File regex
File XML
File Excel
File Idif
File Json
LDAP
Azure Storage
Google Drive
Marketo
MarkLogic
Salesforce
Snowflake
Generic schemas
postgres_avg_occupation_income 0.1
postgres_avg_state_income 0.1
Talend MDM
CosmosDB
Couchbase
Outline
Code Viewer
tAggregateRow_1 (income_avg_state)
tAggregateRow_2 (income_avg_occup)
tDBClose_1
tDBClose_2
tDBCommit_1
tDBConnection_1 (mysqlMetrodata)
tDBConnection_2 (postgresMetrodata)
tDBInput_1 (tDBInput_MySQL)
Database Connection
Update Database Connection - Step 2/2
DB Server must be specified
DB Type PostgreSQL
Db Version v9 and later
String of Connection jdbc:postgresql://:32768/postgres?
Login admin
Password ••••••••
Server
Port 32768
DataBase postgres
Schema
Additional parameters
Test connection
Export as context
Revert Context
How to install a driver
< Back
Next >
Finish
Cancel


Talend Open Studio for Big Data (8.0.1.20211109_1610) | maulfi_alfansuri_hands_on (Connection: Local)
File Edit View Search Window Help
Support
Repository
LOCAL: maulfi_alfansuri_hands_on
File positional
File regex
File XML
File Excel
File Idif
File Json
LDAP
Azure Storage
Google Drive
Marketo
MarkLogic
Salesforce
Snowflake
Generic schemas
postgres_avg_occupation_income 0.1
metadata
Columns(3)
postgres_avg_state_income 0.1
Talend MDM
CosmosDB
Couchbase
NoSQL Connections
Hadoop Cluster
Rules Management
Web Service
Outline
Code Viewer
tAggregateRow_1 (income_avg_state)
tAggregateRow_2 (income_avg_occup)
tDBClose_1
tDBClose_2
Schema
Update Schema "metadata"
Update an existing Schema in repository
Name metadata
Comment
Select the database mapping type
Mapping Postgres
Schema
Description of the Schema
Column
Db Column
Key
Type
DB Type
N..
Date Pattern (...
Length
Precisi...
Defa...
Comm...
occupation
occupation
String
VARC...
20
0
income_dollar
income_dollar
Integer
INT4
10
0
load_date
load_date
Date
DATE
"dd-MM-yyyy"
10
0
Finish
Cancel

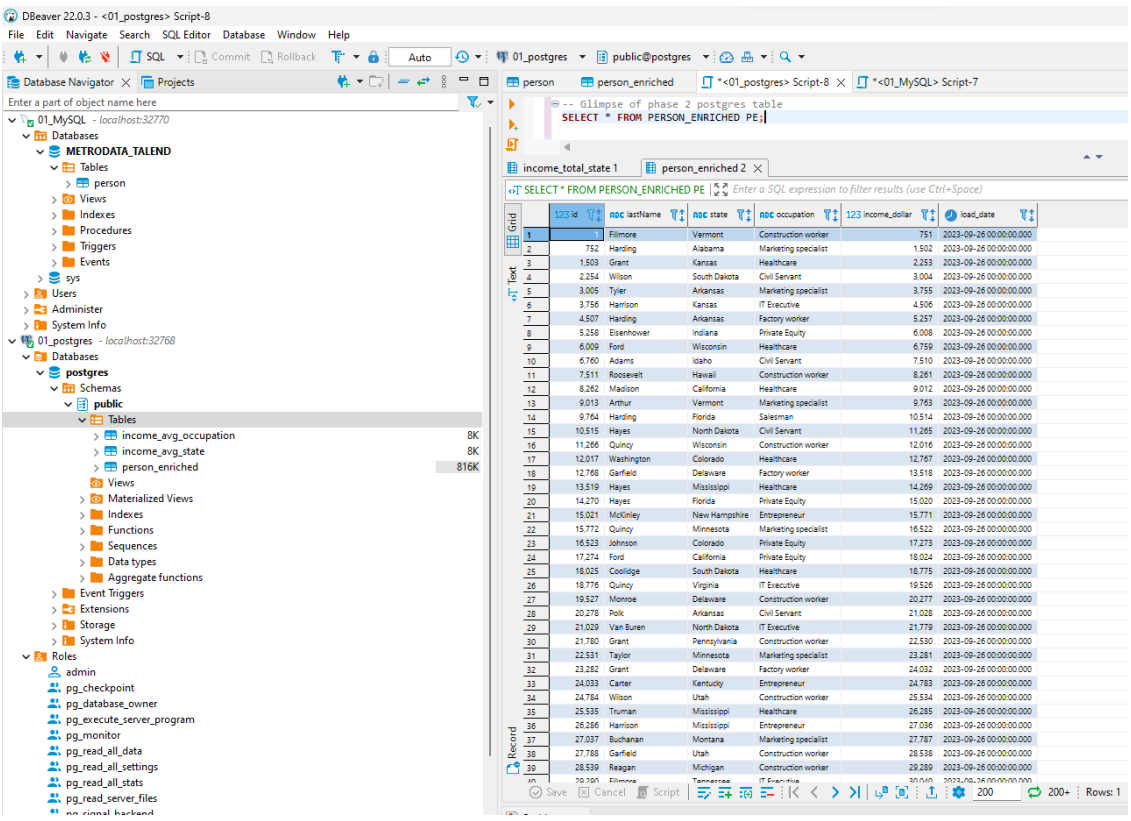

Talend Open Studio for Big Data (8.0.1.20211109_1610) | maulfi_alfansuri_hands_on (Connection: Local)
File Edit View Search Window Help
Support
Repository
LOCAL: maulfi_alfansuri_hands_on
File positional
File regex
File XML
File Excel
File Idif
File Json
LDAP
Azure Storage
Google Drive
Marketo
MarkLogic
Salesforce
Snowflake
Generic schemas
postgres_avg_occupation_income 0.1
metadata
Columns(3)
postgres_avg_state_income 0.1
metadata
Talend MDM
CosmosDB
Couchbase
NoSQL Connections
Hadoop Cluster
Rules Management
Outline
Code Viewer
tAggregateRow_1 (income_avg_state)
tAggregateRow_2 (income_avg_occup)
tDBClose_1
Update generic schema
File - Step 2 of 2
Edit an existing Metadata File on repository
Update the properties
Name metadata
Comment
Select the database mapping type
Mapping Postgres
Schema
Description of the Schema
Column
Db Column
Key
Type
DB Type
N..
Date Patter...
Length
Preci...
Def...
Com...
state
state
String
VARC...
20
0
income_dollar
income_dollar
Integer
INT4
10
0
load_date
load_date
Date
DATE
"dd-MM-y...
10
0
< Back
Next >
Finish
Cancel

3. Tables Overview:

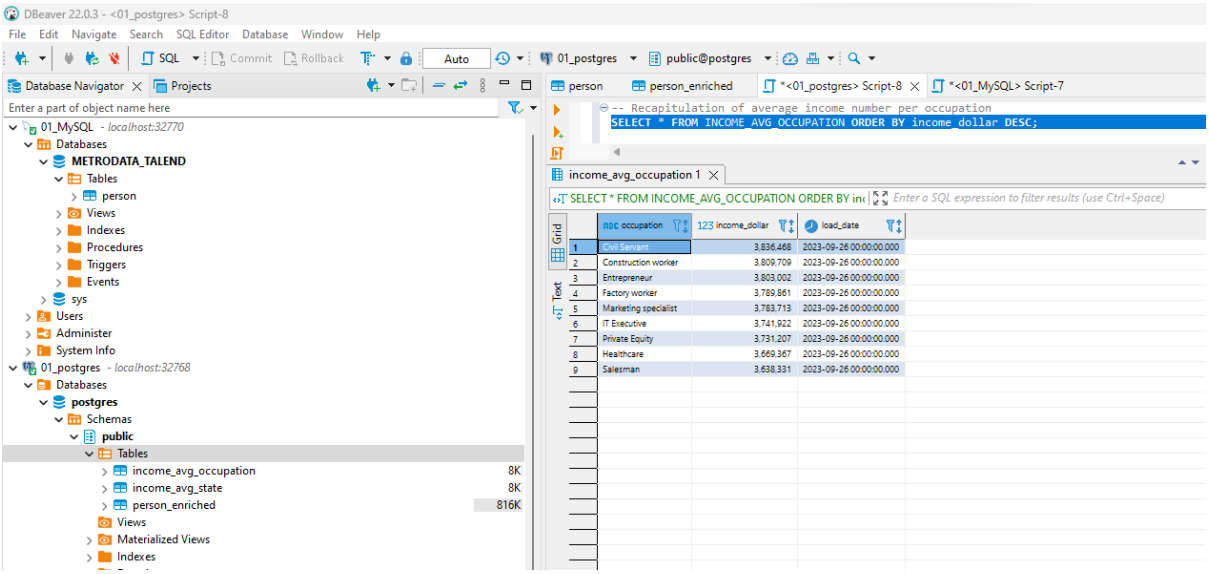
- person Table



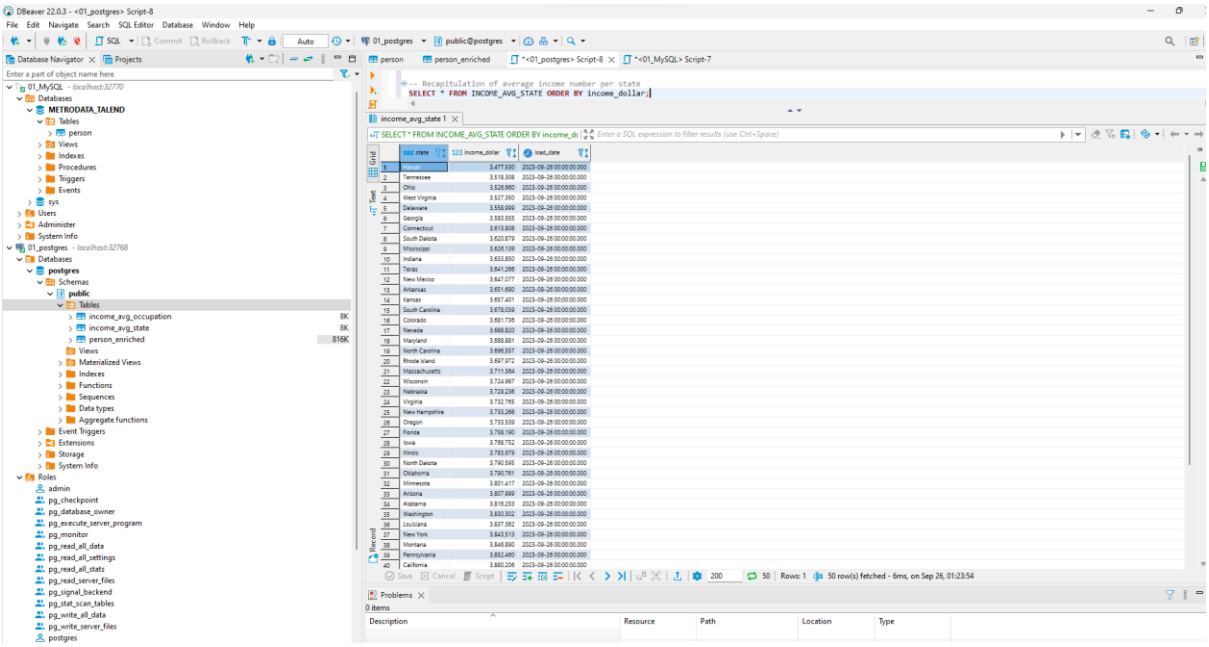
- person_enriched Table



- **income_avg_occupation**

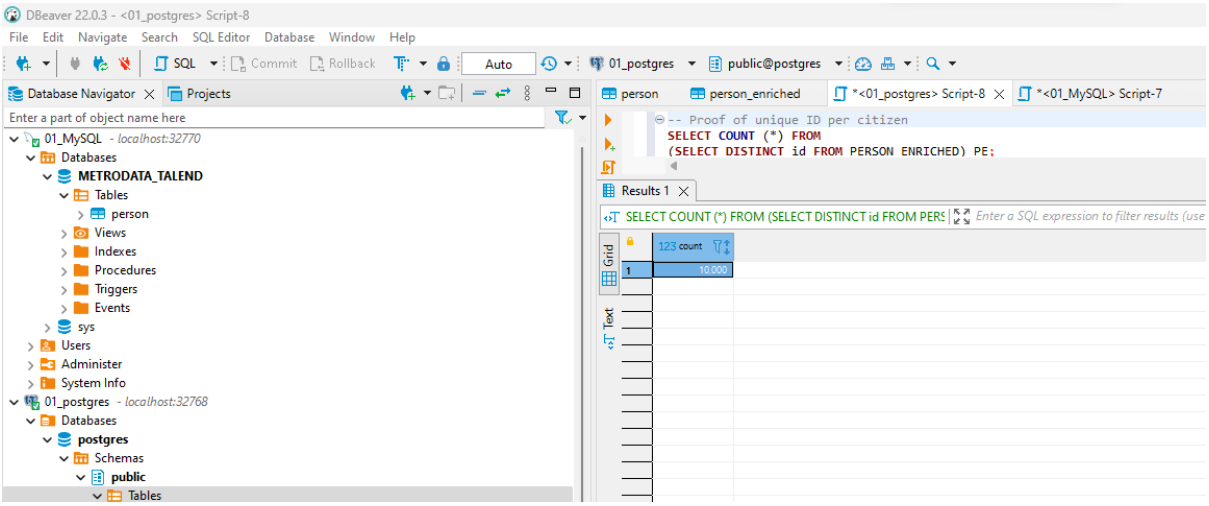


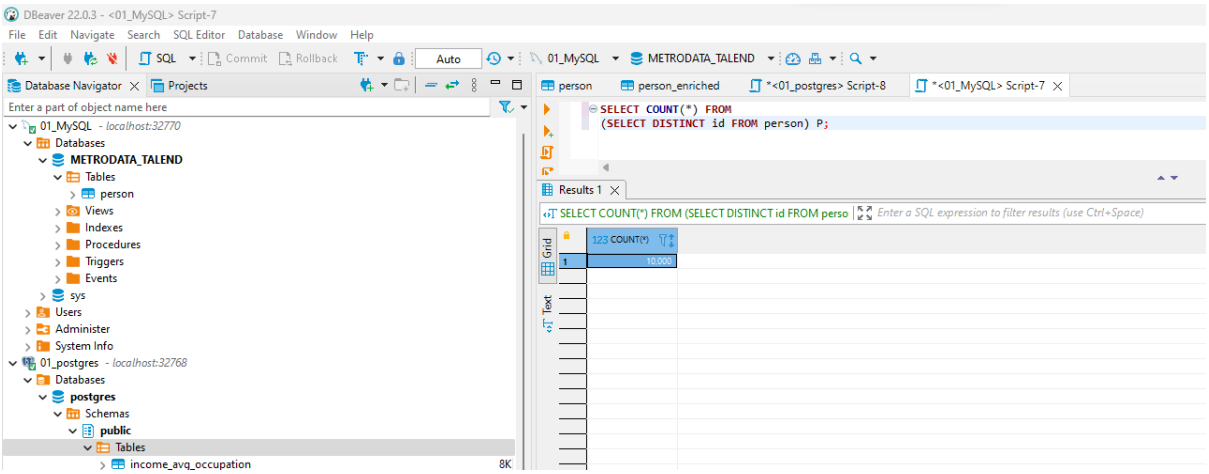
- **income_avg_state**



4. Verification:

- **Kolom id unik:**





- **Output total sama:**