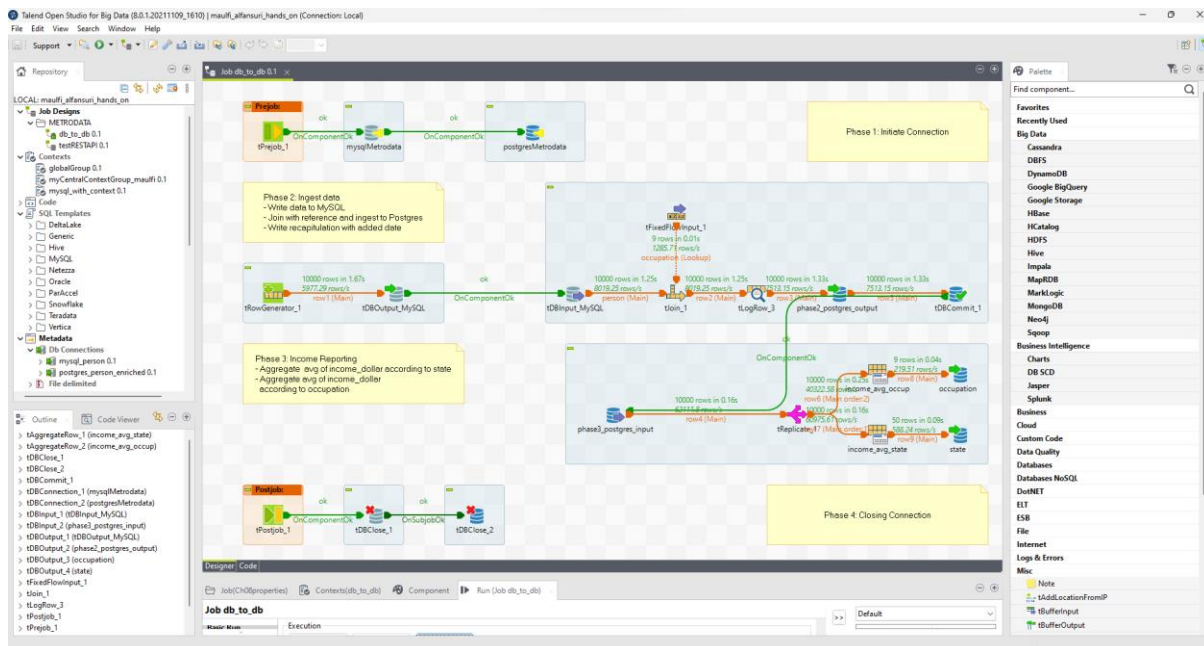


TALEND HANDS ON
MAULFI ALFANSURI

1. Overview



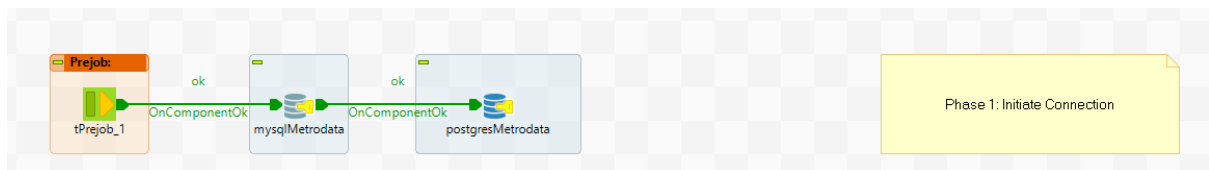
Pada kesempatan ini, akan didesain sebuah *use case* ETL terhadap 10 ribu data penduduk amerika serikta yang memuat *id*(kode unik), *lastName*(nama akhir), *state*(negara bagian), *income*(pendapatan) dan *occupation*(profesi).

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:

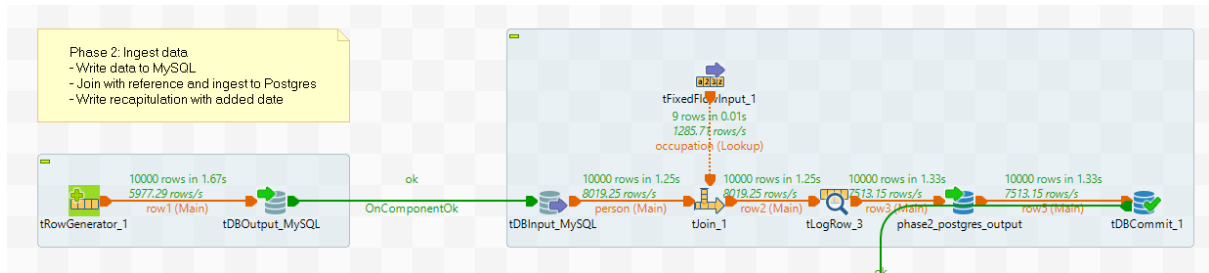
- Phase 1: Inisiasi koneksi terhadap DB MySQL dan PostgreSQL
- Phase 2: Manufaktur data dan *ingestion* kedalam MySQL, transformasi join dan *ingestion* ke dalam PostgreSQL
- Phase 3: Transformasi agregasi terhadap tabel fase 2 postgres kedalam 2 tabel *income_avg_state* dan *income_avg_occupation*
-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

1. Komponen Job:



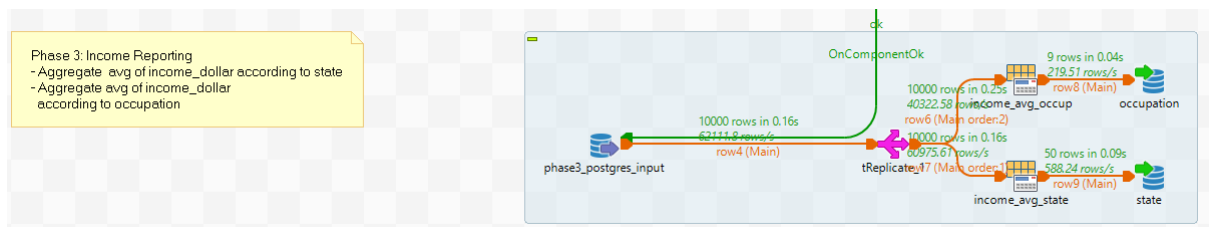
Komponen fase 1:

- 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 agregasi kedalam postgres

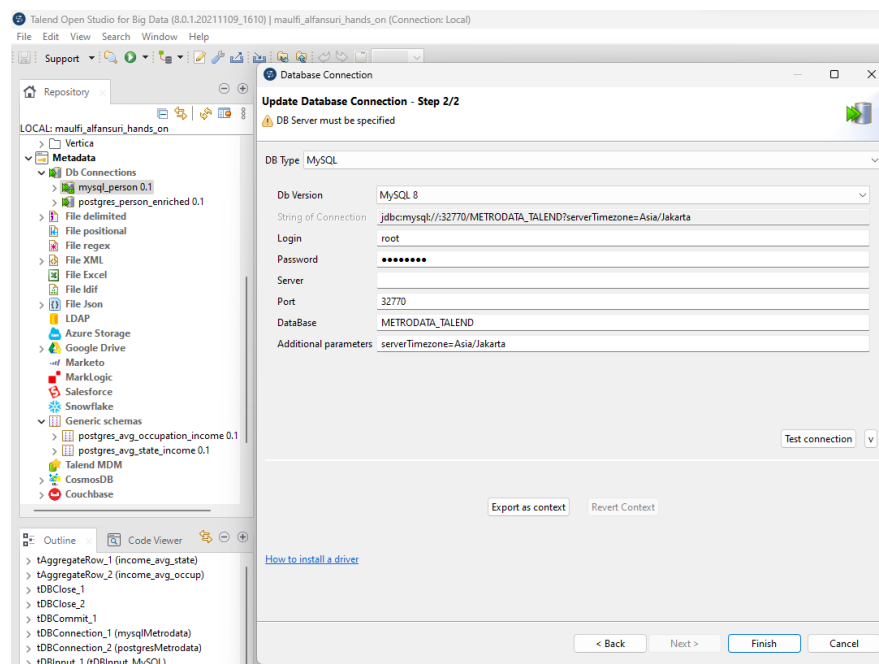
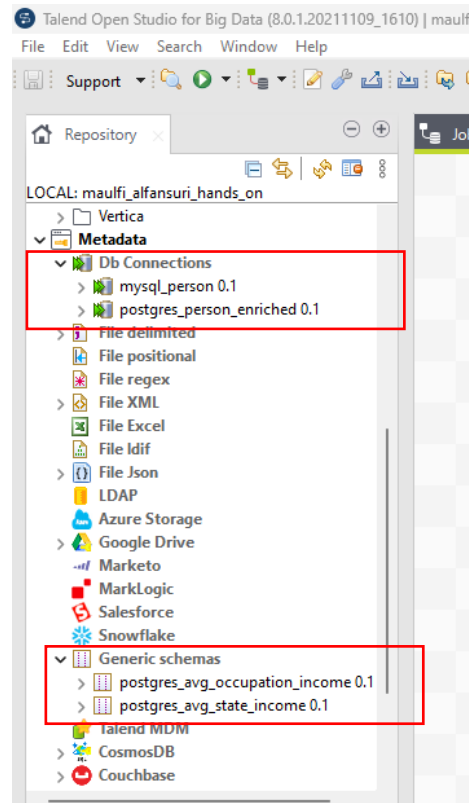


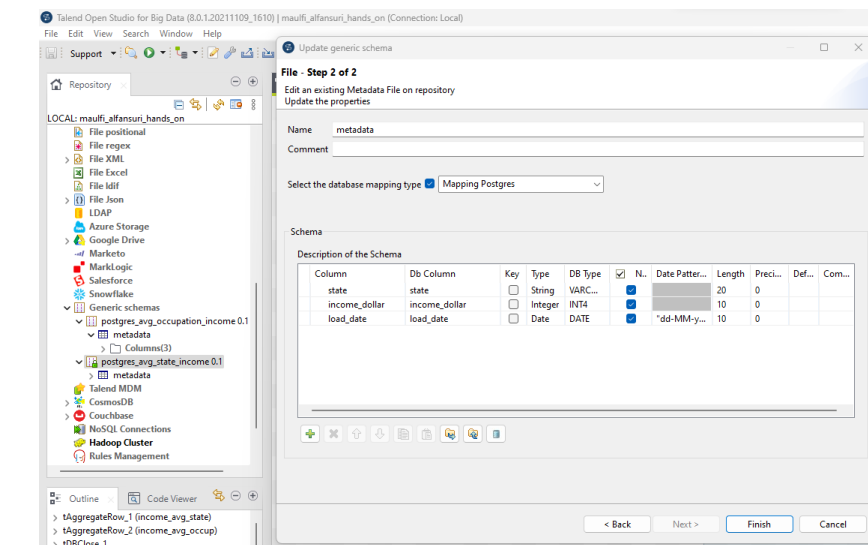
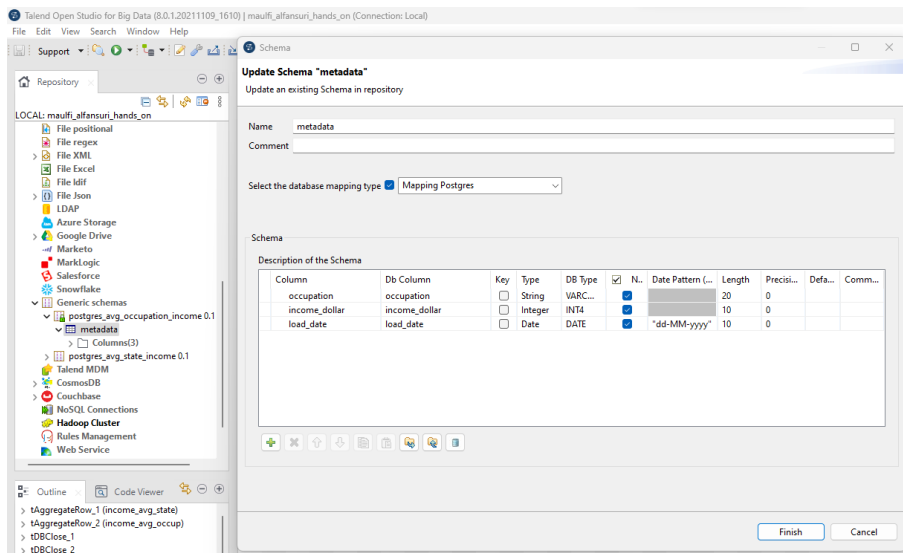
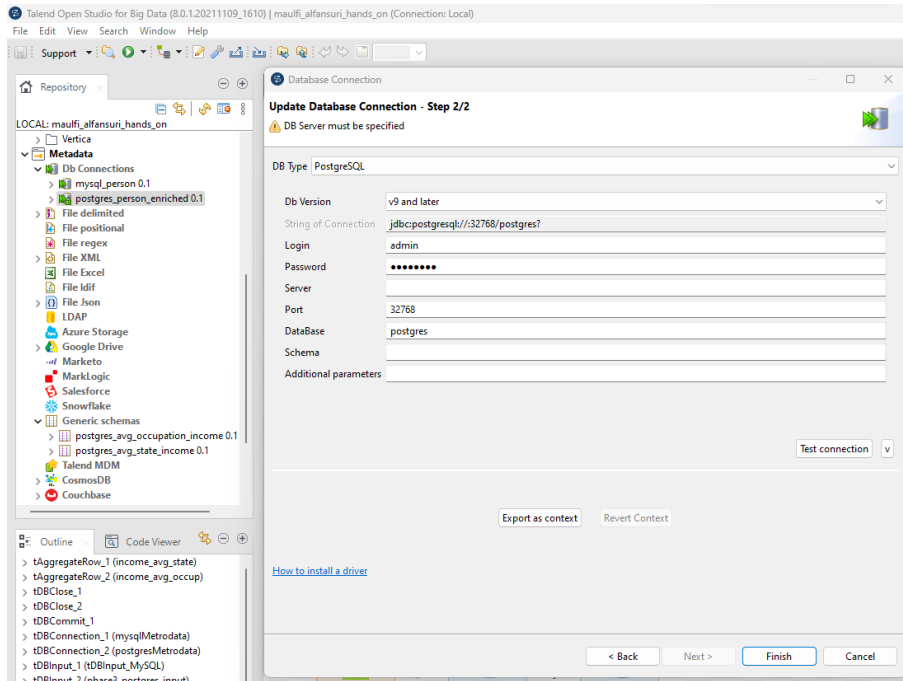
Komponen fase 4:

- 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.





- income_avg_occupation

DBeaver 22.0.3 - <01_postgres> Script-8

File Edit Navigate Search SQL Editor Database Window Help

Database Navigator x Projects

Enter a part of object name here

01_MySQL - localhost:32770

- Databases
 - METRODATA_TALEND
 - Tables
 - person
 - Views
 - Indexes
 - Procedures
 - Triggers
 - Events
 - sys
 - Users
 - Administer
 - System Info
 - 01_postgres - localhost:32768
 - Databases
 - postgres
 - Schemas
 - public
 - Tables
 - income_avg_occupation 8K
 - income_avg_state 8K
 - person_enriched 816K
 - Views
 - Materialized Views
 - Indexes

-- Recapitulation of average income number per occupation
SELECT * FROM INCOME_AVG_OCCUPATION ORDER BY income_dollar DESC;

income_avg_occupation 1 x

+T SELECT * FROM INCOME_AVG_OCCUPATION ORDER BY income_dollar DESC

Grid	occupation	income_dollar	load_date
1	Civil Servant	3.836.468	2023-09-26 00:00:00.000
2	Construction worker	3.809.709	2023-09-26 00:00:00.000
3	Entrepreneur	3.803.002	2023-09-26 00:00:00.000
4	Factory worker	3.789.861	2023-09-26 00:00:00.000
5	Marketing specialist	3.783.713	2023-09-26 00:00:00.000
6	IT Executive	3.741.922	2023-09-26 00:00:00.000
7	Private Equity	3.731.207	2023-09-26 00:00:00.000
8	Healthcare	3.669.347	2023-09-26 00:00:00.000
9	Salesman	3.638.331	2023-09-26 00:00:00.000

- income_avg_state

DBeaver 22.0.3 - <01_postgres> Script-8

File Edit Navigate Search SQL Editor Database Window Help

Database Navigator x Projects

Enter a part of object name here

01_MySQL - localhost:32770

- Databases
 - METRODATA_TALEND
 - Tables
 - person
 - Views
 - Indexes
 - Procedures
 - Triggers
 - Events
 - sys
 - Users
 - Administer
 - System Info
 - 01_postgres - localhost:32768
 - Databases
 - postgres
 - Schemas
 - public
 - Tables
 - income_avg_occupation 8K
 - income_avg_state 8K
 - person_enriched 816K
 - Views
 - Materialized Views
 - Indexes
 - Functions
 - Sequences
 - Data types
 - Aggregate functions
 - Event Triggers
 - Extensions
 - Storage
 - System Info

-- Recapitulation of average income number per state
SELECT * FROM INCOME_AVG_STATE ORDER BY income_dollar DESC;

income_avg_state 1 x

+T SELECT * FROM INCOME_AVG_STATE ORDER BY income_dollar DESC

Grid	state	income_dollar	load_date
1	Tennessee	3.477.530	2023-09-26 00:00:00.000
2	Ohio	3.326.602	2023-09-26 00:00:00.000
3	West Virginia	3.327.380	2023-09-26 00:00:00.000
4	Delaware	3.259.099	2023-09-26 00:00:00.000
5	Georgia	3.183.335	2023-09-26 00:00:00.000
6	Connecticut	3.113.888	2023-09-26 00:00:00.000
7	South Dakota	3.420.979	2023-09-26 00:00:00.000
8	Mississippi	3.628.139	2023-09-26 00:00:00.000
9	Indiana	3.633.820	2023-09-26 00:00:00.000
10	Texas	3.641.298	2023-09-26 00:00:00.000
11	New Mexico	3.647.077	2023-09-26 00:00:00.000
12	Arizona	3.651.690	2023-09-26 00:00:00.000
13	Vermont	3.671.421	2023-09-26 00:00:00.000
14	South Carolina	3.678.039	2023-09-26 00:00:00.000
15	Colorado	3.681.158	2023-09-26 00:00:00.000
16	Montana	3.688.820	2023-09-26 00:00:00.000
17	Marshall	3.688.881	2023-09-26 00:00:00.000
18	North Carolina	3.688.881	2023-09-26 00:00:00.000
19	Idaho	3.688.881	2023-09-26 00:00:00.000
20	Rhode Island	3.687.972	2023-09-26 00:00:00.000
21	Massachusetts	3.711.384	2023-09-26 00:00:00.000
22	Wisconsin	3.714.887	2023-09-26 00:00:00.000
23	Nebraska	3.729.236	2023-09-26 00:00:00.000
24	Virginia	3.732.769	2023-09-26 00:00:00.000
25	New Hampshire	3.732.246	2023-09-26 00:00:00.000
26	Oregon	3.733.039	2023-09-26 00:00:00.000
27	Florida	3.768.190	2023-09-26 00:00:00.000
28	Iowa	3.768.752	2023-09-26 00:00:00.000
29	Illinois	3.769.879	2023-09-26 00:00:00.000
30	North Dakota	3.769.879	2023-09-26 00:00:00.000
31	Oklahoma	3.769.781	2023-09-26 00:00:00.000
32	Minnesota	3.801.417	2023-09-26 00:00:00.000
33	Alaska	3.807.889	2023-09-26 00:00:00.000
34	Alabama	3.814.233	2023-09-26 00:00:00.000
35	Washington	3.833.352	2023-09-26 00:00:00.000
36	Louisiana	3.837.362	2023-09-26 00:00:00.000
37	New York	3.843.113	2023-09-26 00:00:00.000
38	Nebraska	3.848.003	2023-09-26 00:00:00.000
39	Pennsylvania	3.852.480	2023-09-26 00:00:00.000
40	California	3.882.208	2023-09-26 00:00:00.000

0 Items

Problems x

Description	Resource	Path	Location	Type
-------------	----------	------	----------	------

4. Verification:

- Kolom id unik:

The screenshot shows the DBeaver 22.0.3 interface with a PostgreSQL database connection named '01_postgres' at 'localhost:32768'. The 'Database Navigator' on the left shows the 'public' schema containing a table named 'person'. The 'SQL Editor' on the right contains the following SQL script:

```
-- Proof of unique ID per citizen
SELECT COUNT (*) FROM
(SELECT DISTINCT id FROM PERSON ENRICHED) PE;
```

The 'Results' pane shows the execution of the script, resulting in a single row with a count of 123.

count
123

The screenshot shows the DBeaver 22.0.3 interface with a MySQL database connection named '01_MySQL' at 'localhost:32770'. The 'Database Navigator' on the left shows the 'METRODATA_TALEND' database containing a table named 'person'. The 'SQL Editor' on the right contains the following SQL script:

```
SELECT COUNT(*) FROM
(SELECT DISTINCT id FROM person) P;
```

The 'Results' pane shows the execution of the script, resulting in a single row with a count of 123.

COUNT(*)
123

- Output total sama:

SQL Script: 01_postgres - public@postgres

```
-- Comparing Talend aggregate vs direct query aggregate for state based: CHECKED
SELECT A.state, A.income_dollar AS talend_sum_income_dollar, B.income_dollar AS true_income_dollar FROM
(SELECT state, income_dollar FROM INCOME_AVG_STATE) A
LEFT JOIN
(SELECT TRUNC(AVG(income_dollar),0) AS income_dollar, state, load_date FROM PERSON_ENRICHED GROUP BY state, load_date) B
ON A.state = B.state;
```

income_avg_state 1 X

SQL: SELECT A.state, A.income_dollar AS talend_sum_income_dollar, B.income_dollar AS true_income_dollar

id	state	talend_sum_income_dollar	true_income_dollar
1	Alabama	3,683,895	3,683,895
2	Missouri	4,014,846	4,014,846
3	West Virginia	3,637,330	3,637,330
4	Connecticut	3,913,838	3,913,838
5	Washington	3,830,302	3,830,302
6	Nevada	3,729,236	3,729,236
7	North Carolina	3,668,357	3,668,357
8	South Carolina	3,678,039	3,678,039
9	Texas	3,641,266	3,641,266
10	Nevada	3,688,820	3,688,820
11	Arkansas	3,651,690	3,651,690
12	Ohio	3,526,660	3,526,660
13	California	3,660,296	3,660,296
14	Tennessee	3,518,308	3,518,308
15	Massachusetts	3,711,384	3,711,384
16	Vermont	3,562,581	3,562,581
17	New York	3,843,513	3,843,513
18	Maine	4,023,645	4,023,645
19	Colorado	3,681,796	3,681,796
20	Wyoming	4,091,682	4,091,682
21	Wisconsin	3,724,987	3,724,987
22	Iowa	3,765,712	3,765,712
23	Idaho	3,657,401	3,657,401
24	New Mexico	3,647,077	3,647,077
25	Vermont	3,623,662	3,623,662
26	Georgia	3,583,555	3,583,555
27	Virginia	3,732,765	3,732,765
28	Louisiana	3,637,262	3,637,262
29	Utah	3,922,830	3,922,830
30	Alabama	3,616,253	3,616,253
31	Alaska	3,906,295	3,906,295
32	New Jersey	3,655,157	3,655,157
33	Pennsylvania	3,852,480	3,852,480
34	North	3,902,765	3,902,765
35	Oregon	3,733,539	3,733,539
36	Minnesota	3,679,217	3,679,217

load_date: timestamp(0)

Rows: 1 | 50 row(s) fetched - 29ms, on Sep 26, 01:27:05

SQL Script: 01_postgres - public@postgres

```
-- Comparing Talend aggregate vs direct query aggregate for occupation based: CHECKED
SELECT A.occupation, A.income_dollar AS talend_avg_income_dollar, B.income_dollar AS true_income_dollar FROM
(SELECT occupation, AVG(income_dollar) AS income_dollar FROM INCOME_AVG_OCCUPATION GROUP BY occupation) A
LEFT JOIN
(SELECT TRUNC(AVG(income_dollar),0) AS income_dollar, occupation FROM PERSON_ENRICHED GROUP BY occupation) B
ON A.occupation = B.occupation;
```

income_avg_occupation 1 X

SQL: SELECT A.occupation, A.income_dollar AS talend_avg_income_dollar, B.income_dollar AS true_income_dollar

id	occupation	talend_avg_income_dollar	true_income_dollar
1	Salesman	3,789,951	3,789,951
2	Salesman	3,638,331	3,638,331
3	Marketing specialist	3,783,713	3,783,713
4	Entrepreneur	3,803,002	3,803,002
5	Healthcare	3,656,367	3,656,367
6	Construction worker	3,800,709	3,800,709
7	IT Executive	3,741,922	3,741,922
8	Private Equity	3,731,207	3,731,207
9	Civil Servant	3,836,458	3,836,458