Project 2

# Batch Processing Data Pipeline

**Ogi Guntara**

Data Engineering Batch 6

Digital Skola

## Extract

DB SQL :
tb_customer
tb_order

Tools:
PostgreSQL
DBeaver

## Transform

Query
Transform Data

Tools:
Python
Visual Studio Code

## Report in Excel

report_customer.xlsx
report_order.xlsx

# Build Data Warehouse

Langkah pertama dalam project ini adalah menyiapkan database menggunakan PostgreSQL.
Pada project ini data base yang digunakan adalah **db_transaction**

Tabel-tabel yang digunakan di CREATE dengan menggunakan schema pada file **schema_db_transaction.sql**

Setelah kolom tabel telah tersedia kita dapat menambahkan data dengan import sql script lagi pada source file berikut:

1. t**b_customer.sql**
2. **tb_order.sql**

*run menggunakan Execute SQL Script (ALT-X)

## schema_db_transaction.sql

primay key column

```
CREATE TABLE public.tb_customer (
 customer_id text NULL,
 customer_unique_id text NULL,
 customer_zip_code_prefix int8 NULL,
 customer_city text NULL,
 customer_state text NULL
);

CREATE TABLE public.tb_order (
 order_id text NULL,
 customer_id text NULL,
 order_status text NULL,
 order_purchase_timestamp text NULL,
 order_approved_at text NULL,
 order_delivered_carrier_date text NULL,
 order_delivered_customer_date text NULL,
 order_estimated_delivery_date text NULL
);
```

# Build Services

Service utama pada project ini dilakukan menggunakan python.

**Data Base Connection**

Hal yang harus dilakukan pertama kali adalah mengkoneksikan dengan data base yang sudah dibuat.Perintah koneksi terdapat pada file:

**connection.py**

Data yang perlu dimasukkan adalah configurasi database meliputi host, nama db, port, user, dan password.

Agar memudahkan pergantian configurasi serta kerapihan program, maka dibuat file config yang berisikan value tersebut. Pada project ini data tersebut tersimpan pada:

**db.conf**

**Query Data**

Informasi yang diambil dari data base meliputi:

1. order_id, customer_id, customer_city, order_purchase_timestamp.
2. Semua order kecuali order yang dicancel.

File Query terdapat pada folder sql, bernama **dml_query_1.sql**

# Build Services

Peritnah Service Utama terdapat pada **app.py**

Setelah database terkonekasi dan file query disiapkan makan kita lakukan terlebih dahulu read data.

**Read Data**
Data yang didapatkan berupa list, sehingga kita membutuhkan pandas untuk mengubah bentuknya kedalam sebuah data frame. Dalam melakukan read data kita juga meng-*execute* file sql.

**Transformation**
Pada proses transformasi ini kita menggunakan data frame sebelumnya untuk melakukan grouping berdasarkan city dan date, kemudian mengagregatnya jumlah ordernya.

hasil tersebut dibuat dalam bentuk excel dengan perintah to_excel menjadi file berikut:

**report_customer.xlsx**
**report_order.xlsx**

# Hasil

report_customer.xlsx

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | customer | | | | | | | | | | |
| 2 | date | 2016-09-04 | 2016-09-15 | 2016-10-03 | 2016-10-04 | 2016-10-05 | 2016-10-06 | 2016-10-07 | 2016-10-08 | 2016-10-09 | 2016-10-10 | 2016-12-23 |
| 3 | city | | | | | | | | | | | |
| 4 | alem paraiba | | | | 1 | | | | | | | |
| 5 | ananindeua | | | | 1 | | | | | | | |
| 6 | aparecida de goiania | | | | | | 1 | | | | | |
| 7 | apuarema | | | | | | | | 1 | | | |
| 8 | aracaju | | | | | | | | | | 1 | |
| 9 | aracariguama | | | | | | 1 | | | | | |
| 10 | aracatuba | | | | | 1 | | | | | | |
| 11 | atibaia | | | | | 1 | | | | | | |
| 12 | bacaxa | | | | 1 | | | | | | | |
| 13 | bage | | | | | 1 | | | | | | |
| 14 | balneario barra do sul | | | | | | 1 | | | | | |
| 15 | barreirinhas | | | | | | | | | | 1 | |
| 16 | barueri | | | | 1 | | | | | | | |
| 17 | bayeux | | | | | | | | | | 1 | |
| 18 | bebedouro | | | | | | 1 | | | | | |
| 19 | bela cruz | | | | | 1 | | | | | | |
| 20 | belo horizonte | | | | 3 | 5 | | 1 | | 1 | 2 | |
| 21 | belo jardim | | | | | | | | | 1 | | |
| 22 | benedito novo | | | | 1 | | | | | | | |
| 23 | boa vista | 1 | | | | | | | | | 1 | |
| 24 | brasilia | | | | | 1 | 2 | | 2 | | 1 | |
| 25 | cachoeira do sul | | | | | | | | | | 1 | |
| 26 | cachoeiro de itapemirim | | | | 1 | | | | | | | |

report_order.xlsx

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | order | | | | | | | | | | |
| 2 | date | 2016-09-04 | 2016-09-15 | 2016-10-03 | 2016-10-04 | 2016-10-05 | 2016-10-06 | 2016-10-07 | 2016-10-08 | 2016-10-09 | 2016-10-10 | 2016-12-23 |
| 3 | city | | | | | | | | | | | |
| 4 | alem paraiba | | | | 1 | | | | | | | |
| 5 | ananindeua | | | | 1 | | | | | | | |
| 6 | aparecida de goiania | | | | | | 1 | | | | | |
| 7 | apuarema | | | | | | | | 1 | | | |
| 8 | aracaju | | | | | | | | | | 1 | |
| 9 | aracariguama | | | | | | 1 | | | | | |
| 10 | aracatuba | | | | | 1 | | | | | | |
| 11 | atibaia | | | | | 1 | | | | | | |
| 12 | bacaxa | | | | 1 | | | | | | | |
| 13 | bage | | | | | 1 | | | | | | |
| 14 | balneario barra do sul | | | | | | 1 | | | | | |
| 15 | barreirinhas | | | | | | | | | | 1 | |
| 16 | barueri | | | | 1 | | | | | | | |
| 17 | bayeux | | | | | | | | | | 1 | |
| 18 | bebedouro | | | | | | 1 | | | | | |
| 19 | bela cruz | | | | | 1 | | | | | | |
| 20 | belo horizonte | | | | 3 | 5 | | 1 | | 1 | 2 | |
| 21 | belo jardim | | | | | | | | | 1 | | |
| 22 | benedito novo | | | | 1 | | | | | | | |
| 23 | boa vista | 1 | | | | | | | | | 1 | |
| 24 | brasilia | | | | | 1 | 2 | | 2 | | 1 | |
| 25 | cachoeira do sul | | | | | | | | | | 1 | |
| 26 | cachoeiro de itapemirim | | | | 1 | | | | | | | |