Data Warehouse dan ETL Pipeline

Menggunakan DuckDB & Python/Pandas

Oleh:
Sigit Hanafi (24/546999/PPA/06868)
Kadek Gunamulya Sudarma Yasa (24/547500/PPA/06892)
Abdul Razak Aliudin (24/547523/PPA/06898)
Rafa Nafisah (24/537451/PPA/06798)

Pendahuluan

1. Deskripsi Bisnis

Bisnis yang dipilih adalah bisnis perbankan. Bisnis perbankan berfokus pada penyediaan layanan keuangan kepada individu, bisnis atau institusi. Bisnis perbankan memiliki banyak layanan keuangan. Beberapa layanan yang disediakan seperti rekening tabungan, deposito, pinjaman, kartu kredit, transfer dana dan investasi. Pada tugas ini kami berfokus pada data transaksi yang mencakup informasi mengenai aliran uang masuk dan keluar, nasabah, akun, merchan, channel transaksi, tanggal dan lokasi transaksi keuangan.

2. Tujuan Analisis & Pertanyaan Bisnis

Data Warehouse yang dibuat bertujuan untuk menyediakan data yang dapat digunakan untuk proses analisis. Data berasal dari data transaksi yang kemudian dilakukan proses ETL dan dipindahkan ke data warehouse agar proses analisis tidak mengganggu sistem / data yang digunakan dalam proses transaksi perbankan.

Tujuan utama dari data warehouse yang dibuat adalah untuk menyediakan data yang dapat digunakan untuk analisis strategis guna meningkatkan layanan, segmentasi nasabah, analisis volume transaksi, dan beberapa analisis untuk mendukung keputusan strategis bisnis lainnya.

Beberapa pertanyaan bisnis yang harus dijawab oleh data warehouse ini:

a. Analisis Transaksi dan Volume
    - Berapa total transaksi harian/mingguan/bulanan di seluruh channel (ATM, teller, online banking)?
    - Bagaimana tren transaksi berdasarkan wilayah, transaksi dalam dimensi waktu tertentu?
    - Bagaimana perbandingan transaksi debit vs kredit?
b. Segmentasi dan Analisis Nasabah
    - Berapa banyak nasabah aktif dalam periode waktu tertentu?
    - Bagaimana pola transaksi nasabah berdasarkan segmentasi pekerjaan/umur/lokasi?
    - Seperti apa segmentasi nasabah dalam suatu wilayah?
c. Analisis Kinerja Cabang dan Channel Transaksi
    - Lokasi mana yang memiliki volume transaksi tertinggi dan terendah?
    - Channel transaksi apa yang paling sering digunakan oleh pelanggan (ATM/Online/Cabang)?
    - Bagaimana tren penggunaan online transaction, pelanggan yang seperti apa yang menggunakan channel online?
    - Berapa lama rata-rata waktu transaksi yang digunakan nasabah berdasarkan channel? Channel mana yang tercepat dan terendah?

3. Indikator Utama Monitoring

Berikut adalah Indikator utama yang akan dimonitoring melalui data warehouse:

a. Indikator Transaksi
    - Total Volume Transaksi (harian, bulanan, tahunan)
    - Jumlah Transaksi Debit vs Kredit
    - Rata-rata Nilai Transaksi
    - Persentase Penggunaan Layanan untuk setiap channel
b. Indikator Nasabah
    - Jumlah Pelanggan Aktif dalam waktu dan lokasi tertentu

- Segmentasi Pelanggan berdasarkan umur, pekerjaan dan lokasi
- Tingkat Adopsi Mobile Banking & Internet Banking (channel online)

c. Indikator Kinerja Cabang & Channel
   - Total Transaksi per Cabang / Channel
   - Total Transaksi per Channel (ATM, Teller, Online)
   - Tingkat Ketersediaan dan Downtime setiap channel (ATM/Online)
   - Rata-rata waktu yang dibutuhkan untuk setiap transaksi

4. Laporan yang Dibutuhkan

Laporan Transaksi

- Volume transaksi berdasarkan jenis layanan persatuan waktu dan lokasi Laporan Kinerja Cabang & Channel
- Total transaksi per cabang dan wilayah
- Penggunaan channel transaksi digital vs konvensional (online /ATM /cabang)

5. Sumber Data

Sumber data yang digunakan merupakan open source data yang bersumber dari Kaggle. Data bisa diakses melalui url berikut bank transaction dataset. Data berisi 2.512 data transaksi dan memiliki beberapa atribut detail transaksi, demografi nasabah dan merchant. Dataset “Bank Transaction Dataset for Fraud Detection” dari Kaggle memiliki karakteristik yang relevan dengan data modeling untuk data warehouse yang akan dibuat. Berikut beberapa karakteristik dataset yang digunakan:

a. Struktur sesuai dengan kebutuhan data warehouse

   Dataset ini mengandung informasi transaksi perbankan yang dapat digunakan untuk membangun tabel fakta transaksi dalam data warehouse. Data warehouse perbankan membutuhkan struktur yang terdiri dari:

   - Fakta transaksi (jumlah transaksi, jenis transaksi, nominal, dll)
   - Dimensi customer (ID, status, umur, lokasi, dll)

- Dimensi waktu (tanggal transaksi, waktu transaksi)
- Dimensi channel (ATM, online banking, cabang)
- Dimensi lokasi (ID, status, umur, lokasi, dll)

b. Terdapat indikator utama yang akan dimonitoring

   Sebagai sumber untuk data warehouse, dataset ini bisa digunakan untuk melihat indikator utama dalam bisnis perbankan, misalnya:

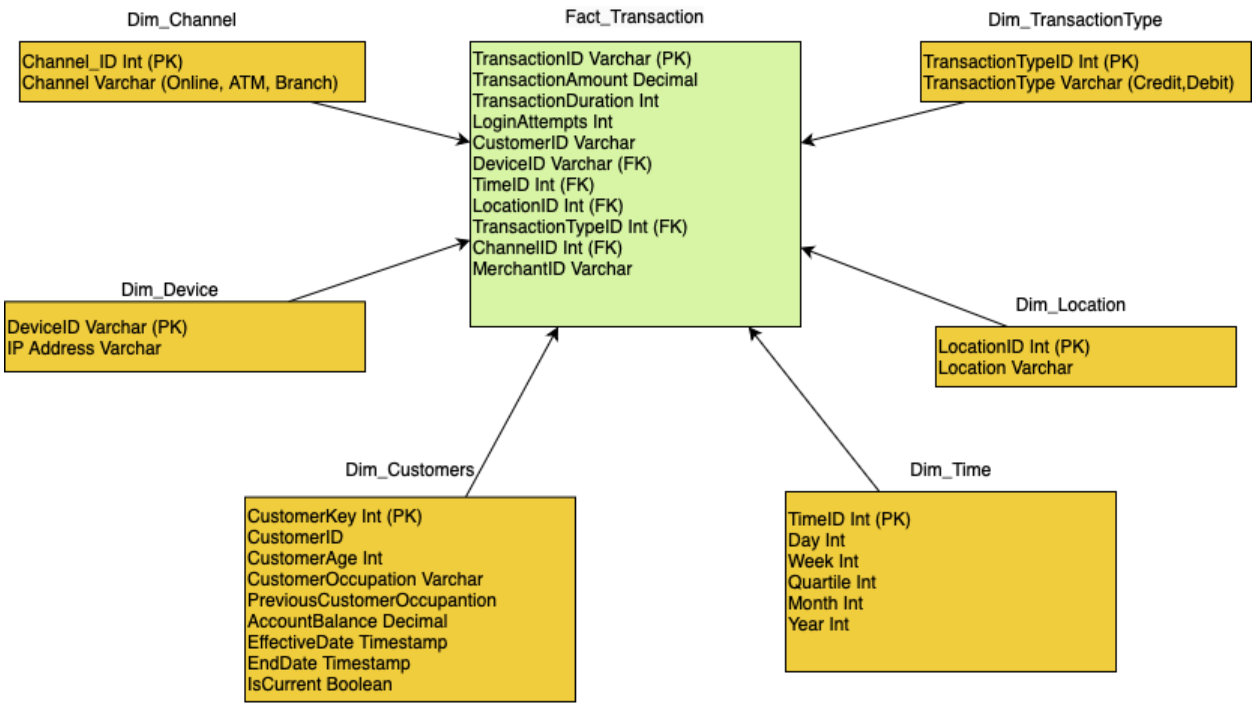
   i. Indikator Transaksi
   ii. Indikator Nasabah
   iii. Indikator Kinerja Cabang & Channel

   Dengan menggunakan dataset ini, kita dapat membangun laporan yang mencerminkan transaksi perbankan secara real-time.

c. Dataset sesuai dengan laporan yang dibutuhkan

Bagian 1: Data Modeling untuk Data Warehouse

Desain Skema Data Warehouse



Pemilihan Star Schema sebagai model skema dalam Data Warehouse didasarkan pada beberapa pertimbangan utama yang sesuai dengan karakteristik dataset yang dimiliki. Star Schema memiliki struktur yang lebih sederhana dibandingkan dengan skema lain seperti Snowflake Schema, sehingga memudahkan proses desain dan implementasi. Struktur ini cocok untuk data yang digunakan yang tidak memiliki kompleksitas tinggi dalam hubungan antar entitas, sehingga tidak memerlukan normalisasi berlebihan yang dapat memperumit query dan analisis data.

Script DDL untuk DuckDB

Script SQL Data Definition Language (DDL) untuk membuat tabel di dalam database duckdb bisa diakses pada repositori berikut https://github.com/sigidhanafi/dwib-etl-pipeline/tree/main/sql.

Bagian 2: ETL (Extract, Transform, Load) Process

Dataset Sumber

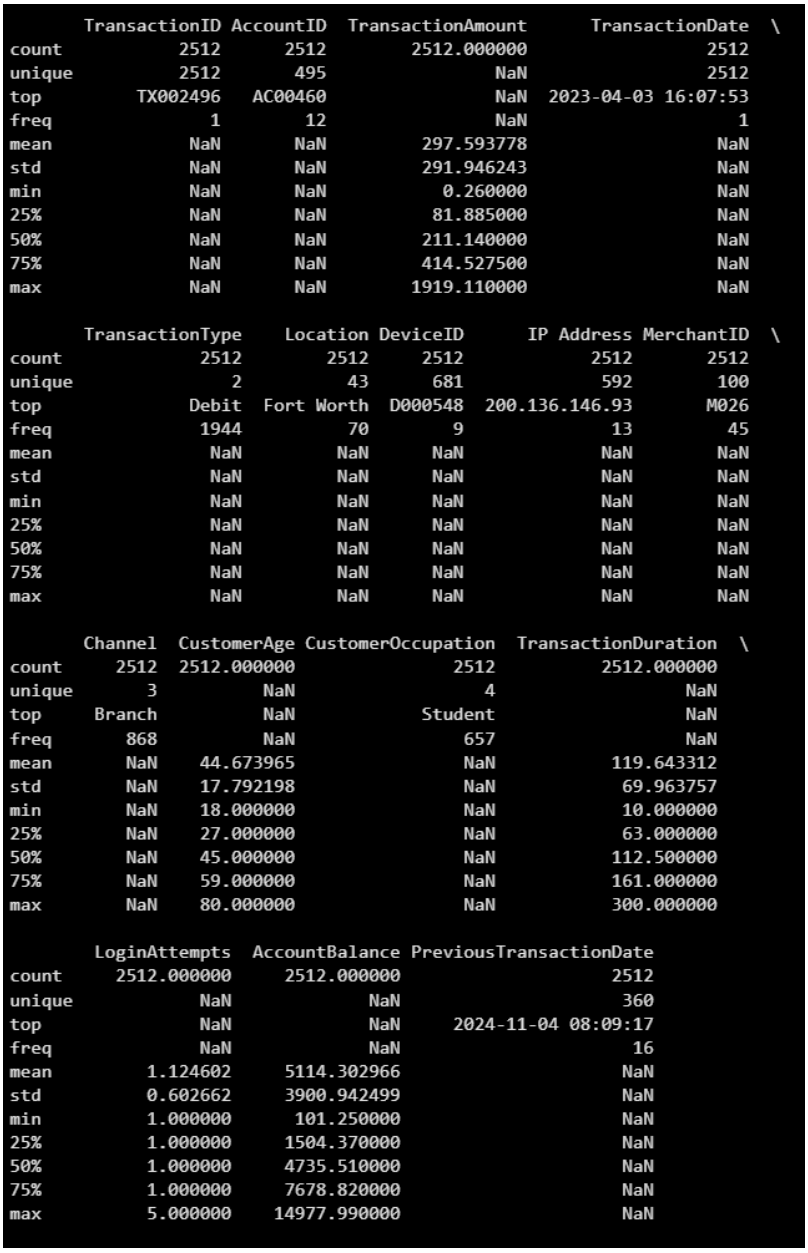
Sumber data yang digunakan merupakan open source data yang bersumber dari Kaggle. Data bisa diakses melalui url berikut bank transaction dataset. Data berisi 2.512 data transaksi dan memiliki beberapa atribut detail transaksi, demografi nasabah dan merchant. Dataset “Bank Transaction Dataset for Fraud Detection” dari Kaggle memiliki karakteristik yang relevan dengan data modeling untuk data warehouse yang akan dibuat.

Exploratory Data Analysis (EDA)

1. Deskripsi dataset dan struktur data

   Dataset ini berisi data transaksi bank dengan total 2.512 transaksi dan 16 kolom, termasuk informasi yang tersedia yaitu 'TransactionID', 'AccountID', 'TransactionAmount', 'TransactionDate', 'TransactionType', 'Location', 'DeviceID', 'IP Address', 'MerchantID', 'Channel', 'CustomerAge', 'CustomerOccupation', 'TransactionDuration', 'LoginAttempts', 'AccountBalance', 'PreviousTransactionDate'.
2. Menampilkan ringkasan statistik data

   Berikut hasil ringkasan statistik yang didapat dari proses EDA yang terdapat pada gambar di bawah.



```
       TransactionID AccountID  TransactionAmount      TransactionDate  \
count           2512      2512        2512.000000                 2512
unique          2512       495                NaN                 2512
top         TX002496   AC00460                NaN  2023-04-03 16:07:53
freq               1        12                NaN                    1
mean             NaN       NaN         297.593778                  NaN
std              NaN       NaN         291.946243                  NaN
min              NaN       NaN           0.260000                  NaN
25%              NaN       NaN          81.885000                  NaN
50%              NaN       NaN         211.140000                  NaN
75%              NaN       NaN         414.527500                  NaN
max              NaN       NaN        1919.110000                  NaN

       TransactionType    Location DeviceID      IP Address MerchantID  \
count             2512        2512     2512            2512       2512
unique               2          43      681             592        100
top             Debit  Fort Worth  D000548  200.136.146.93       M026
freq              1944          70        9              13         45
mean               NaN         NaN      NaN             NaN        NaN
std                NaN         NaN      NaN             NaN        NaN
min                NaN         NaN      NaN             NaN        NaN
25%                NaN         NaN      NaN             NaN        NaN
50%                NaN         NaN      NaN             NaN        NaN
75%                NaN         NaN      NaN             NaN        NaN
max                NaN         NaN      NaN             NaN        NaN

       Channel  CustomerAge CustomerOccupation  TransactionDuration  \
count     2512  2512.000000               2512          2512.000000
unique       3          NaN                  4                  NaN
top     Branch          NaN            Student                  NaN
freq       868          NaN                657                  NaN
mean       NaN    44.673965                NaN           119.643312
std        NaN    17.792198                NaN            69.963757
min        NaN    18.000000                NaN            10.000000
25%        NaN    27.000000                NaN            63.000000
50%        NaN    45.000000                NaN           112.500000
75%        NaN    59.000000                NaN           161.000000
max        NaN    80.000000                NaN           300.000000

        LoginAttempts  AccountBalance PreviousTransactionDate
count     2512.000000     2512.000000                    2512
unique            NaN             NaN                     360
top               NaN             NaN     2024-11-04 08:09:17
freq              NaN             NaN                      16
mean         1.124602     5114.302966                     NaN
std          0.602662     3900.942499                     NaN
min          1.000000      101.250000                     NaN
25%          1.000000     1504.370000                     NaN
50%          1.000000     4735.510000                     NaN
75%          1.000000     7678.820000                     NaN
max          5.000000    14977.990000                     NaN
```

a. Identitas dan Informasi Dasar Transaksi yaitu **TransactionID & AccountID** yang berisi **2.512 transaksi unik** dengan **495** akun unik. **TransactionID** bersifat unik (tidak ada duplikasi). Akun dengan ID **AC00460** paling sering muncul (**12 kali**).
b. Informasi Keuangan yaitu **TransactionAmount** memiliki Rata-rata transaksi berkisar **$297.59** dimana **Minimal** transaksi **$0.26** dan **Maksimal** transaksi sama dengan **$1,919.11**. Kuartil menunjukkan bahwa

sebagian besar transaksi berada di bawah **$414.53**. Pada tabel **AccountBalance**, Saldo akun rata-rata didapat **$5,114.30**.dengan minimum **$101.25** dan untuk Saldo maksimum yaitu **$14,977.99**.

c. Informasi Teknis Transaksi, pada tabel **TransactionDate** dan **PreviousTransactionDate** yang menyimpan informasi **tanggal transaksi** dan **tanggal transaksi sebelumnya**. Waktu transaksi terbaru tercatat **4 November 2024**.Pada tabel **TransactionType,** jenis transaksi hanya ada **2 kategori** (Debit dan Kredit). **Debit** adalah jenis transaksi yang paling umum (**1.944 kali**). Tabel **TransactionDuration** menampilkan rata-rata durasi transaksi adalah **119 detik**. Transaksi tercepat yaitu **10 detik** dan yang terlama dengan **300 detik**.

d. Informasi Pengguna dan Perangkat, yaitu pada **CustomerAge** terdapat usia rata-rata pelanggan sama dengan **44.67 tahun**. Bagi pelanggan termuda yaitu **18 tahun** dan tertua adalah **80 tahun**. pada **CustomerOccupation**, terdapat **4 jenis pekerjaan** dan yang paling sering muncul adalah "Student" (**657 kali**). Pada tabel **DeviceID, IP Address, Location** berisi **43 lokasi unik**, **681 perangkat unik**, **592 alamat IP unik**, yang bisa menjadi faktor risiko dalam analisis fraud. Tabel **LoginAttempts** menampilkan rata-rata percobaan login sebelum transaksi terhitung **1.12 kali**. Maksimum percobaan login yaitu **5 kali**, yang bisa mengindikasikan upaya peretasan.

3. Missing values dan duplicate data
    - Missing Values: Tidak ditemukan missing values dalam dataset ini. Semua kolom memiliki data yang lengkap.
    - Data Duplikat: Tidak ditemukan data duplikat dalam dataset ini. Setiap transaksi bersifat unik.

Script ETL dengan Python/pandas

Script ETL dan setup database dapat diakses pada repository
https://github.com/sigidhanafi/dwib-etl-pipeline/tree/main/etl

Implementasi Fitur ETL Lanjutan

Fitur ETL lanjutan yang di implementasi adalah

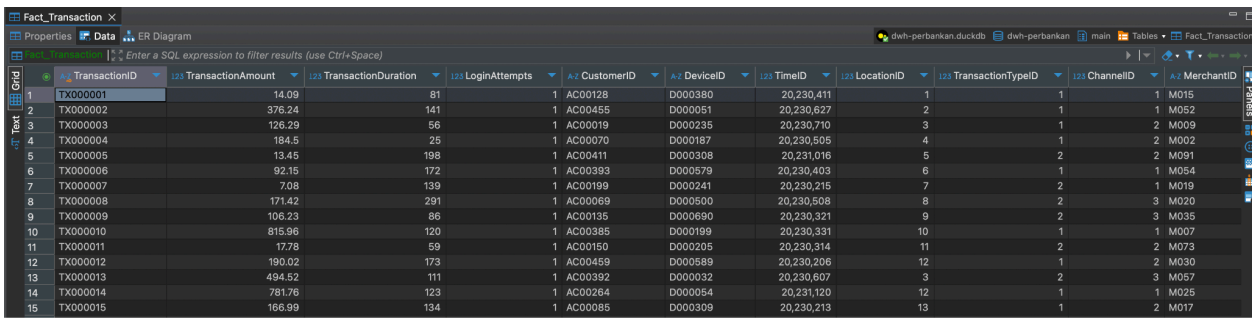
a. ETL incremental (hanya memproses data baru)
b. Implementasi SCD Type 2 untuk menangani perubahan historis pada dimensi data customer

Bagian 3: Hasil Akhir

Gambar 3.1 menampilkan hasil proses setup dan ETL yang dilakukan pada terminal yang menampilkan status setiap proses setup dan ETL.

```
(dwh-env) sigithanafi@Sigits-MacBook-Pro dwh-perbankan % python3 main.py
Setup:
        Memulai proses setup database & table in DuckDB!
        Menjalankan dim_channel.sql...
        Menjalankan dim_customer.sql...
        Menjalankan dim_device.sql...
        Menjalankan dim_location.sql...
        Menjalankan dim_time.sql...
        Menjalankan dim_type.sql...
        Menjalankan fact_transaction.sql...
        ✅ DDL Script sukses dijalankan!
        Koneksi database ditutup!
        Setup database & table Selesai!
Proses ETL:
        Memulai proses ETL!
        ✅ Proses extract data berhasil!
        ✅ Koneksi ke database berhasil!
                ETL Dim_Customer
                ✅ Dim_Customer berhasil diproses! Jumlah baris di Dim_Customer: 2513
                Proses Dim_Channel Selesai!


                ETL Dim_Channel
                ✅ Transform ChannelID berhasil!
                ✅ Dim_Channel berhasil diproses! Jumlah baris di Dim_Channel: 3
                Proses Dim_Channel Selesai!


                ETL Dim_Time
                ✅ Transform date format berhasil!
                ✅ Dim_Time berhasil diproses! Jumlah baris di Dim_Time: 261
                Proses Dim_Time Selesai!


                ETL Dim_Location
                ✅ Transform LocationID berhasil!
                ✅ Dim_Location berhasil diproses! Jumlah baris di Dim_Location: 43
                Proses Dim_Location Selesai!


                ETL Dim_Device
                ✅ Dim_Device berhasil diproses! Jumlah baris di Dim_Device: 681
                Proses Dim_Device Selesai!


                ETL Dim_Transaction_Type
                ✅ Dim_Transaction_Type berhasil diproses! Jumlah baris di Dim_Transaction_Type: 2
                Proses Dim_Time Selesai!


                ETL Fact_Transactions
                ✅ Mapping TimeID berhasil!
                ✅ Mapping Location berhasil!
                ✅ Mapping DeviceID berhasil!
                ✅ Mapping ChannleID berhasil!
                ✅ Mapping TransactionTypeID berhasil!
                ✅ Fact_Transaction berhasil diproses! Jumlah baris di Fact_Transaction: 2518
                Proses Fact_Transaction Selesai!


        Koneksi database ditutup!
        Proses ETL Selesai!

Program Selesai
```

Gambar 3.1. Hasil running program



Fact_Transaction

Properties | Data | ER Diagram

dwh-perbankan.duckdb | dwh-perbankan | main | Tables | Fact_Transaction

Enter a SQL expression to filter results (use Ctrl+Space)

	TransactionID	TransactionAmount	TransactionDuration	LoginAttempts	CustomerID	DeviceID	TimeID	LocationID	TransactionTypeID	ChannelID	MerchantID
1	TX000001	14.09	81	1	AC00128	D000380	20,230,411	1	1	1	M015
2	TX000002	376.24	141	1	AC00455	D000051	20,230,627	2	1	1	M052
3	TX000003	126.29	56	1	AC00019	D000235	20,230,710	3	1	2	M009
4	TX000004	184.5	25	1	AC00070	D000187	20,230,505	4	1	2	M002
5	TX000005	13.45	198	1	AC00411	D000308	20,231,016	5	2	2	M091
6	TX000006	92.15	172	1	AC00393	D000579	20,230,403	6	1	1	M054
7	TX000007	7.08	139	1	AC00199	D000241	20,230,215	7	2	1	M019
8	TX000008	171.42	291	1	AC00069	D000500	20,230,508	8	2	3	M020
9	TX000009	106.23	86	1	AC00135	D000690	20,230,321	9	2	3	M035
10	TX000010	815.96	120	1	AC00385	D000199	20,230,331	10	1	1	M007
11	TX000011	17.78	59	1	AC00150	D000205	20,230,314	11	2	2	M073
12	TX000012	190.02	173	1	AC00459	D000589	20,230,206	12	1	2	M030
13	TX000013	494.52	111	1	AC00392	D000032	20,230,607	3	2	3	M057
14	TX000014	781.76	123	1	AC00264	D000054	20,231,120	12	1	1	M025
15	TX000015	166.99	134	1	AC00085	D000309	20,230,213	13	1	2	M017

Gambar 3.2. Contoh data pada fact_transaksi