PREPARATION

1. Studi Kasus : Membuat Data warehouse dari Database Sakila (sampel database tentang Rental Film dari MySql)
2. Langkah Persiapan:
    a. Install database server: MySql (versi 5)
    b. Install software mysql management: Sql Yog Ultimate (ATAU YG LAIN)
    c. Download and Install software pentaho untuk ETL: Pentaho Data Extraction (PDI) versi community edition
    d. Copy file sampel data: sakila.sql dan whsakila2021.sql
    e. Buat database dengan nama sakila
    f. Import file sakila.sql ke dalam database sakila

TUTORIAL ETL (EXTRACTION TRANSFORMATION LOADING)

Inti dari proses ETL adalah **memasukkan data** ke dalam data warehouse dari sumber data yang dimiliki, baik database operasional/transaksional, maupun sumber-sumber lain yang berbentuk teks file atau yang lain.

Dalam tutorial ini, sumber data untuk data warehouse (whsakila2021) adalah database operasional sakila. Tabel **utama** di database sakila yang menjadi **sumber data** adalah tabel **payment** karena di tabel ini terdapat data **amount** (uang sewa film) yang merupakan **measure** yang dibutuhkan pada data warehouse yang dibangun.

Untuk data-data selain **amount** dan tidak ada di tabel pendukung, maka perlu ditelusuri dengan melihat tabel-tabel yang berelasi dengan **payment** dan begitu seterusnya.

1. Database whsakila2021 mempunyai 5 tabel:
   i. Dimensi **customer** (customer_id, nama)
   ii. Dimensi **store** (store_id, kota)
   iii. Dimensi **film** (film_id, judul, kategori)
   iv. Dimensi **time** (time_id,tahun, bulan,tanggal,tanggallengkap, namahari)
   v. Fakta **fakta_pendapatan** (customer_id,time_d,film_id,store_id,amount, lamapinjam)
2. 3 tabel pertama akan diisi dengan menggunakan SQL (SQL YOG)
3. Tabel time diisi dengan menggunakan Import file CSV (SQL YOG) atau Stored Procedure
4. Tabel fakta_pendapatan diisi dengan menggunakan Pentaho Data Integration

A. Menggunakan File CSV dan SQL

1. Tabel Dimensi Customer
   a. Analisis

Data yang dibutuhkan tabel customer di whsakila2021	Sumber Data di database sakila	Keterangan
Customer_id	Tabel **customer** → customer_id	
Nama	Tabel **customer** → first_name dan last_name	First_name dan last_name harus digabung menggunakan fungsi **CONCAT()**

   b. Penyusunan Query

```
INSERT INTO whsakila2021.customer
SELECT sk.customer_id, CONCAT(sk.first_name,' ',sk.last_name)
FROM sakila sk
```

2. Tabel Dimensi Store
   a. Analisis

Data yang dibutuhkan tabel customer di whsakila2021	Sumber Data di database sakila	Keterangan
Store_id	Tabel **store** → store_id	
Kota	Tabel store → address → city → city	

b. Penyusunan Query

```
INSERT INTO whsakila2021.`store`
SELECT st.store_id , ct.`city`
FROM sakila.store st
JOIN sakila.address ad ON (st.`address_id` = ad.`address_id`)
JOIN sakila.city ct ON(ct.`city_id`=ad.`city_id`)
```

3. Tabel Dimensi Film
   a. Analisis

Data yang dibutuhkan tabel film di whsakila2021	Sumber Data di database sakila	Keterangan
Film_id	Tabel **film** → film_id	
Judul	Tabel film → title	
Kategori	Tabel film → film_category → category → name	

b. Penyusunan Query

```
INSERT INTO whsakila2021.film
SELECT f.film_id, f.title, c.name
FROM sakila.film f
JOIN sakila.film_category fc ON (f.film_id = fc.film_id)
JOIN sakila.category c ON (fc.category_id = c.category_id)
```

4. Tabel Dimensi Time

Analisis:

a. Untuk data pada tabel dimensi time, data satuan waktu terkecil yang dibutuhkan adalah harian, artinya hanya sampai tanggal. Tabel dimensi time adalah tabel acuan yang tidak tergantung pada data transaksi, akan tetapi untuk **baris pertama**-nya perlu dilihat data transaksi pinjam film pertama yang ada di dalam database **sakila**.

b. Dengan perintah Sql berikut ini akan diketahui tanggal paling awal dalam tabel rental yaitu tanggal 24 mei 2005.:

```
SELECT rental_date
FROM sakila.`rental`
```

```
ORDER BY rental_date ASC
LIMIT 1
```

Sedangkan tanggal paling akhir dalam tabel rental yaitu tanggal 14 Februari 2006:

```
SELECT rental_date
FROM sakila.`rental`
ORDER BY rental_date DESC
LIMIT 1
```

Jadi data dalam tabel time akan dimulai pada 24 mei 2005 sampai setidaknya tanggal 14 Februari 2006. Tentu saja boleh diteruskan sampai tanggal berapapun.

c. Cara Pertama (menggunakan file CSV)
   - Buat file excel baru
   - Di baris pertama kolom pertama tuliskan: 1
   - Di baris pertama kolom kedua tuliskan: =year(E1)
   - Di baris pertama kolom ketiga tuliskan: =month(E1)
   - Di baris pertama kolom keempat tuliskan: =day(E1)
   - Di baris pertama kolom kelima tuliskan: 5/24/2005
   - Di baris pertama kolom keenam tuliskan:
     `=IF(WEEKDAY(E1)=1;"minggu";IF(WEEKDAY(E1)=2;"senin";IF(WEEKDAY(E1)=3;"selasa";IF(WEEKDAY(E1)=4;"rabu";IF(WEEKDAY(E1)=5;"kamis";IF(WEEKDAY(E1)=6;"jumat";"sabtu"))))))`
   - Blok A1:F1 dan drag ke bawah sampai baris ke 342 (30 April 2006) sebagai batasan saja dan tekan Ctrl-C untuk mengcopy
   - Buat File excel baru
   - Klik kanan di A1, pilih paste special → values
   - Hapus kolom F1 (5/24/2005), dan gantikan dengan berikut: =concatenate(B1,"-", C1,"-",D1)
   - Blok dan drag ke bawah sampai 04/30/2006
   - Simpan file excel terbaru sebagai CSV
   - Buka SQL YOG, temukan tabel time di database whsakila2021, klik kanan->import->import csv data using Load Local
   - Pilih file CSV yang dihasilkan di atas
   - Klik tombol **change,** ubah **fields terminated by** dengan koma (\,)
   - Klik OK, Klik Import

d. Cara Kedua (menggunakan Stored Procedure)
   - Buka SQL YOG
   - Pilih database whsakila2021, klik kanan di stored procs
   - Ketikkan nama procedure: "isitimedimens"

- Copy baris berikut

```
DELIMITER $$


USE `whsakila2021`$$


DROP PROCEDURE IF EXISTS `isitimedimens`$$


CREATE DEFINER=`root`@`localhost` PROCEDURE `isitimedimens`(IN
waktuMulai DATE, IN waktuSelesai DATE)

BEGIN

  mulaiLoop: LOOP

    INSERT        INTO        TIME        (TAHUN,BULAN,TANGGAL,
tanggallengkap,NAMAHARI)

               VALUES(YEAR(waktuMulai),

                MONTH(waktuMulai),

                DAY(waktuMulai),

                      waktuMulai,

                      DAYNAME(waktuMulai)

          );

    SET waktuMulai = DATE_ADD(waktuMulai,INTERVAL 1 DAY);

    IF DATEDIFF(waktuMulai,waktuSelesai)!=0 THEN

      ITERATE mulaiLoop;

    END IF;

    LEAVE mulaiLoop;

  END LOOP mulaiLoop;

 END$$
```

-
- `DELIMITER ;`Select all dengan Ctrl+A pada window isitimedimens, dan paste-kan script di atas
- Klik tombol run query
- Klik tombol untuk tambah New Query Editor
- Ketikkan dan jalankan query berikut:

```
CALL isitimedimens('2005-05-24', '2006-04-30');
```

B. Menggunakan Pentaho Data Integration

Untuk tabel **fakta_pendapatan**, diperlukan proses lookup untuk melihat time_id yang sesuai untuk rental_date maupun return_date yang mana itu lebih mudah dilakukan dengan menggunakan Pentaho Data Integration. Berikut adalah langkah-langkah yang harus dilakukan:

1. Langkah Persiapan:
   Dengan asumsi anda tidak mempunyai data warehouse sakila, maka lakukan langkah berikut:
   a. Buat database dengan nama whsakila2021
   b. Import file whsakila2021.sql ke dalam database whsakila2021

2. Analisis :
   a. Periksa kebutuhan data di **table fakta** di warehouse sakila.
      Field yang ada:
      - customer_id (int)
      - time_id (int) → waktu pinjam
      - film_id (int)
      - store_id (int)
      - **amount (decimal)**
      - **lamapinjam (int)**

      (NB: field yang dicetak tebal adalah **measures**)
   b. Periksa sumber data dari database sakila untuk melihat table mana yang bisa digunakan.
      1) Data utama yang dibutuhkan adalah data measures: **amount dan lama pinjam**
      2) Identifikasi table-table yang berisi field-field yang terkait dengan menemukan **table utamanya** dulu, kemudian menelusuri table-table yang terkait melalui relasi yang ada.
      3) Tabel utamanya adalah Tabel **Payment**, dan tabel terkait adalah **rental**, dan **inventory**
      4) Mapping data yang diperlukan dengan sumber data sebagai berikut:

Data yang dibutuhkan di **table fakta** di whsakila2021	Sumber Data di database sakila	Keterangan
Customer_id	Table payment→ customer_id	
Time_id (waktu pinjam)	Table rental → rental_date	Tipe data time_id (int) tidak match dengan rental_date(date). Perlu

		LOOKUP **ke table time** di whsakila2021
Film_id	Table inventory→film_id	
Store_id	Table inventory→store_id	
Amount	Table payment→amount	
Lamapinjam (lama hari pinjam)	Table rental, (return_date - rental_date)	Pakai fungsi SQL **datediff()**

5) Susun query di SQL Yog untuk mendapatkan data yang diperlukan:

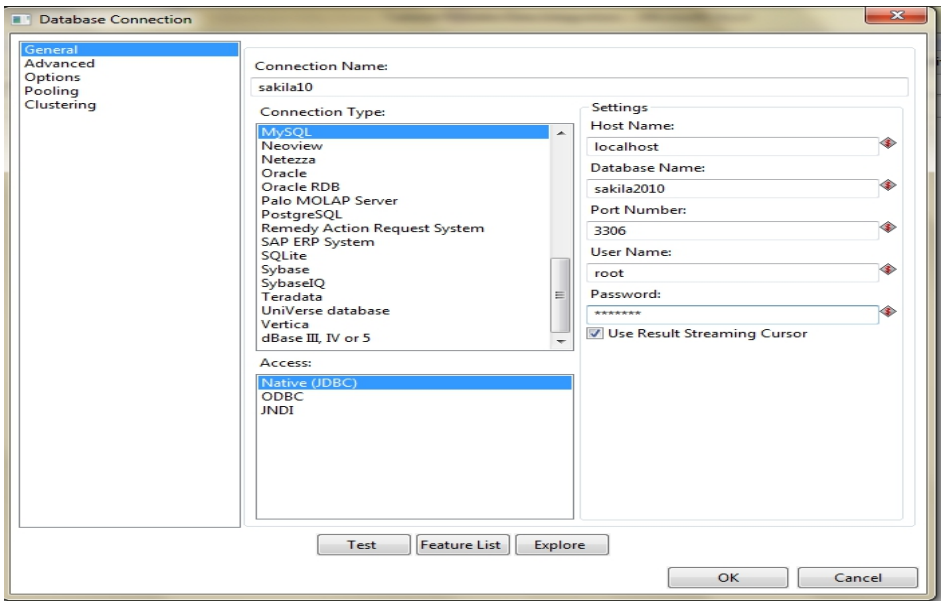
```
SELECT r.`customer_id`,
       DATE(r.`rental_date`) rental_date,
       i.`film_id`,
       i.`store_id`,
       p.`amount`,
       DATEDIFF(r.return_date,r.rental_date) lamapinjam
FROM sakila.payment p
JOIN sakila.rental r ON(p.`rental_id` = r.`rental_id`)
JOIN sakila.inventory i ON(r.`inventory_id`= i.`inventory_id`)
ORDER BY r.`customer_id`, r.rental_date
```

6) Langkah selanjutnya adalah mencari : time_id untuk rental_date dengan menggunakan Pentaho Data Integration

3. Mencari time_id menggunakan Pentaho Data Integration
   a. Double-Click spoon.bat, jika muncul halaman terkait Repository klik Cancel
   b. Klik File→ New→Transformation
   c. Klik di Tab "View" disebelah kiri Tab "Design"

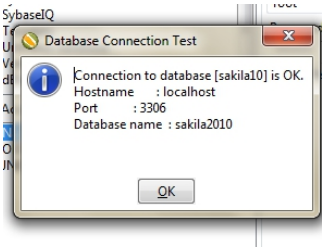
   Membuat koneksi database

   d. Klik Kanan di Database Connections dan Klik New



Gambar 1

e. Isikan data sesuai gambar 1 di atas:
   Note: **nama database, user name dan password** sesuaikan dengan setting database yang ada di komputer anda.
f. Klik tombol Test untuk memastikan koneksi telah berhasil



Gambar 2

   (Note: jika tidak berhasil, cek kembali nama database, username dan password!)
g. Jika koneksi berhasil, klik pada tombol OK di gambar 1.
h. Klik kanan pada sakila10 kemudian klik Share jika anda ingin koneksi ini dipakai berulang-ulang)
i. Ulangi langkah d s/d h di atas untuk koneksi ke database whsakila2021, beri nama koneksinya whsakila10.

Menyusun Diagram Transformasi

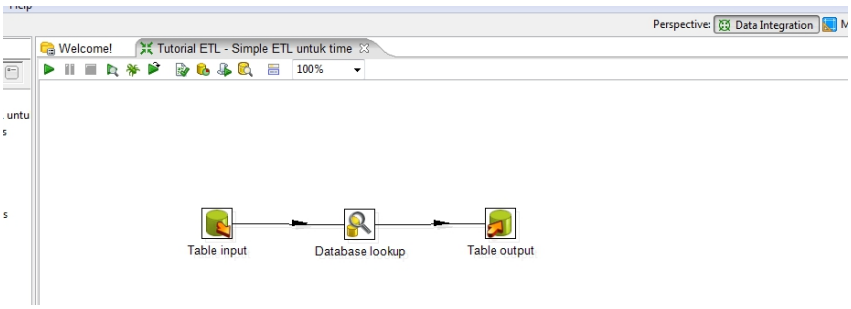
j. Klik tab Design di sebelah kanan tab View
k. Klik di Input→Table Input , tahan dan drag ke window sebelah kanan
   (Ini komponen untuk mengambil data dari Sakila)

l. Klik di Lookup→Database Lookup, tahan dan drag ke sebelah kanan Table Input. *Ini komponen untuk mencari* ***Time_Id*** *untuk* ***rental_date*** .

m. Klik di Output→Table Output, tahan dan drag ke sebelah kanan Database Lookup

n. Letakkan kursor pada komponen Table Input, tunggu sebentar, klik pada tanda panah, tahan dan drag ke komponen Database Lookup di sebelah kanannya. (jika ada pilihan, pilih Main Output Step)

o. Ulangi pada komponen selanjutnya sampai dengan Table Output sehingga terlihat seperti Gambar 3.



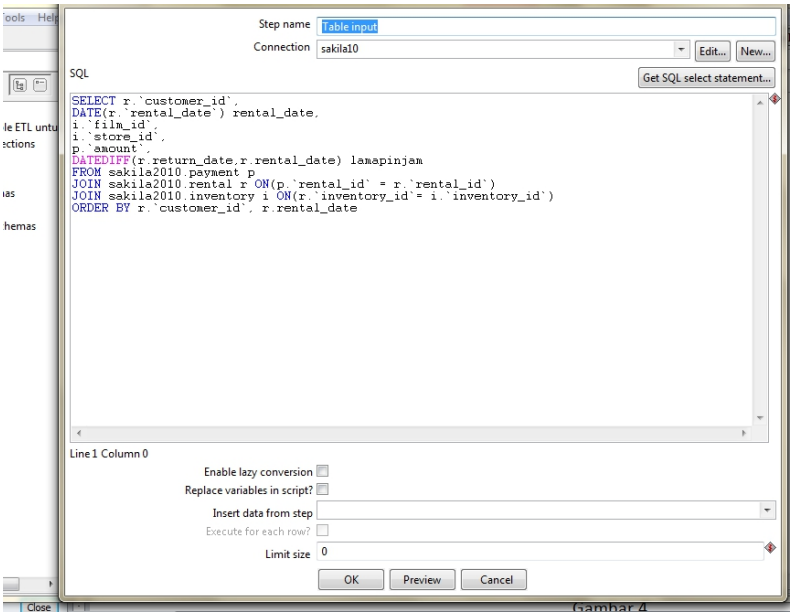
Gambar 3

Menyusun Prosesnya

p. Dobel Klik pada komponen Table Input

q. Pilih connection yang sesuai: sakila10

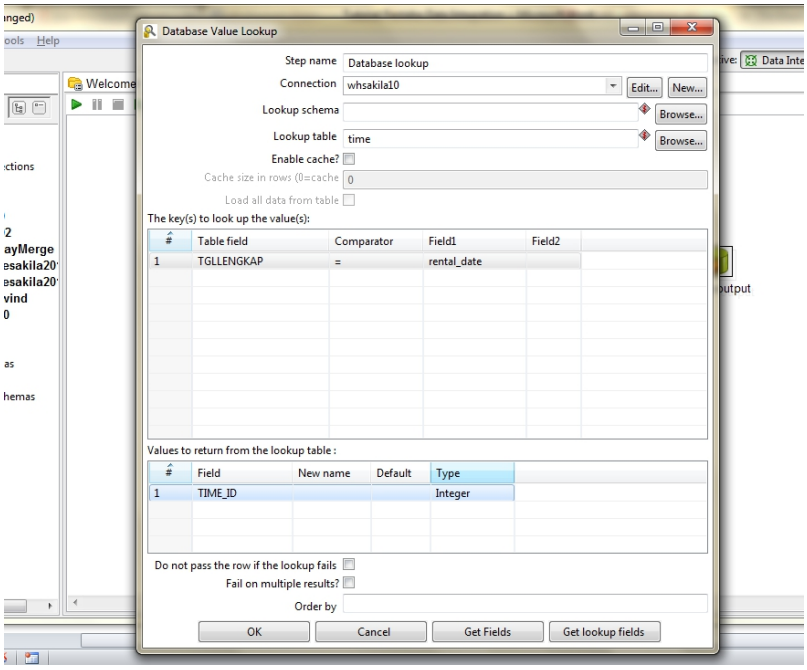
r. Copy query yang disusun pada langkah b(5) dan paste ke dalam box



Gambar 4

s. Klik tombol Preview, dan ketikkan 10 (untuk 10 data saja), untuk melihat apakah query bisa memberikan hasil yang dinginkan

t. Klik tombol Close di halaman preview, dan klik tombol OK
u. Selanjutnya adalah mencari time_id untuk rental_date, dobel klik di komponen database lookup pertama, pilih koneksi whsakila10, dan untuk lookup table cari melalui tombol browse untuk tabel time
v. Isikan di bagian Key(s) to Look up dan bagian Values to return seperti gambar 5 berikut ini:



Gambar 5

w. Klik tombol OK
x. Langkah terakhir adalah mengisikan data ke dalam tabel fakta_pendapatan di whsakila2021
y. Klik pada komponen table output, isikan/pilih koneksi (whsakila10), target table (fakta_pendapatan) dan juga Klik di Specify Database Fields dan mapping-kan Table Field dan Stream Field seperti gambar 6, kemudian klik OK
z. Simpan transformation dengan nama “insert-into-fakta_pendapatan”
aa. Klik tombol Play/Run (segitiga warna hijau) untuk menjalankan proses transformasi.
ab. Selesai

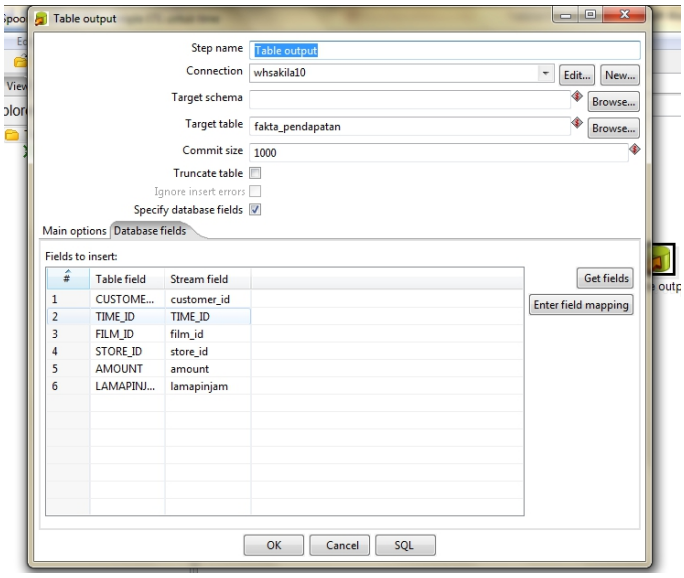


Table output
Step name
Table output
Connection
whsakila10
Edit...
New...
Target schema
Browse...
Target table
fakta_pendapatan
Browse...
Commit size
1000
Truncate table
Ignore insert errors
Specify database fields
Main options
Database fields
Fields to insert:
Table field Stream field
1 CUSTOME... customer_id
2 TIME_ID TIME_ID
3 FILM_ID film_id
4 STORE_ID store_id
5 AMOUNT amount
6 LAMAPINJ... lamapinjam
Get fields
Enter field mapping
OK
Cancel
SQL

Gambar 6