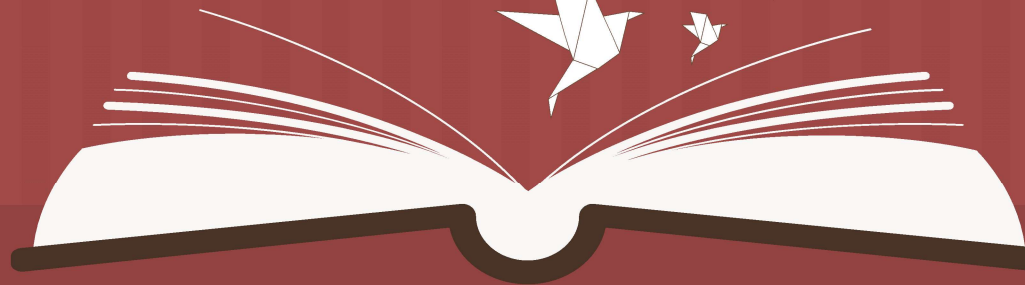


Pertemuan 6

Proses ETL (Extract, Transform
and Loading) dan Teknik
Pemodelan Data
Multidimensional

SRI HERAWATI, S.KOM, M.KOM



PRODI SISTEM INFORMASI
JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS TRUNOJOYO MADURA
2023

BISNIS CERDAS

◎ Tujuan Instruksional Umum

M a h a s i s w a m a m p u m e m a h a m i d a n mengimplementasikan Bisnis cerdas

◎ Tujuan Instruksional Khusus

Mahasiswa dapat menjelaskan Proses ETL (Extract, Transform and Loading) dan Teknik Pemodelan Data Multidimensional

TOPIK BAHASAN

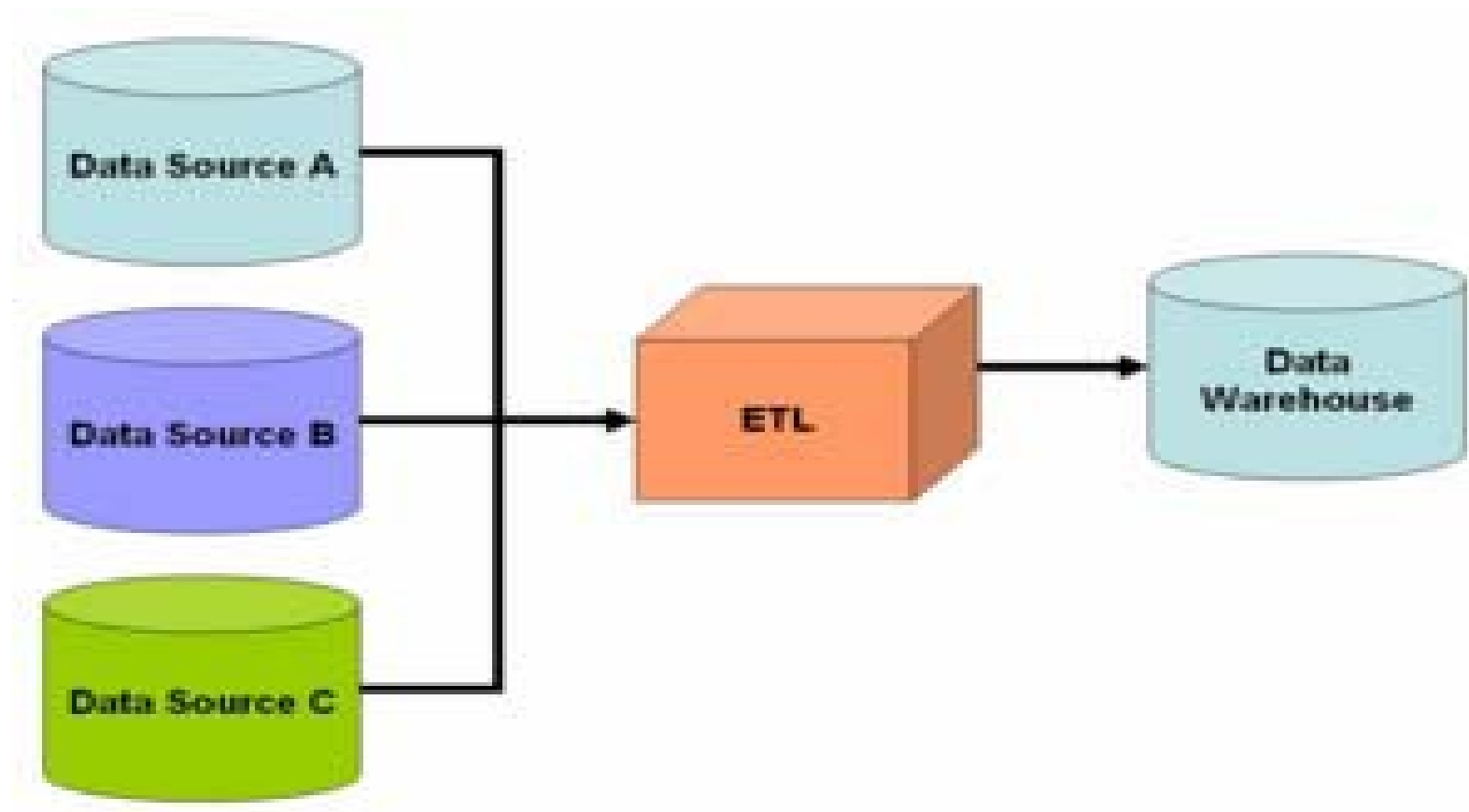
01 Pengertian ETL

02 Proses ETL

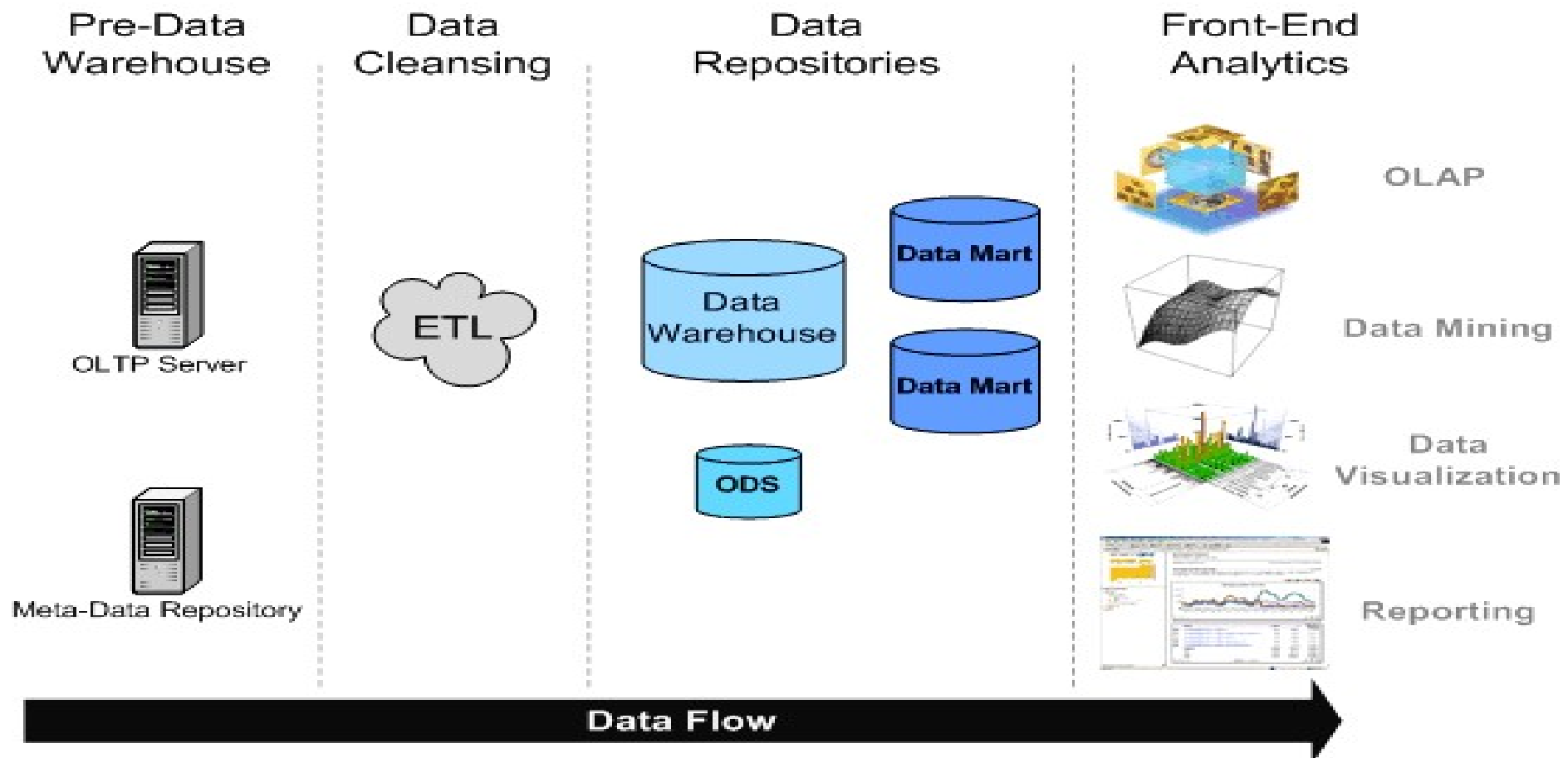
03 Teknik Pemodelan Data
Multidimensional



Arsitektur Data Warehouse (1)



Arsitektur Data Warehouse (2)



ETL (Extract Transform Loading)

- **Data Extraction**

Fungsi ini biasanya berhadapan dengan bermacam *data source*, dan menggunakan teknik yang sesuai dengan setiap *data source*. Sumber data mungkin berasal dari *source machine* yang berbeda dalam format data yang berbeda pula.

ETL (Extract Transform Loading)

■ Data Transformation

Data transformation melibatkan berbagai bentuk dalam mengkombinasikan bagian dari data yang berasal dari sumber yang berbeda. Kombinasi data dilakukan dari sumber *record* tunggal, atau dapat juga dilakukan dari elemen data yang berelasi dengan banyak sumber *record*. Proses *cleaning* mungkin dilakukan dalam *data transformation*, dimana proses *cleaning* memiliki fungsi untuk melakukan koreksi terhadap kesalahan pengejaan, atau untuk melakukan eliminasi terhadap duplikat data.

ETL (Extract Transform Loading)

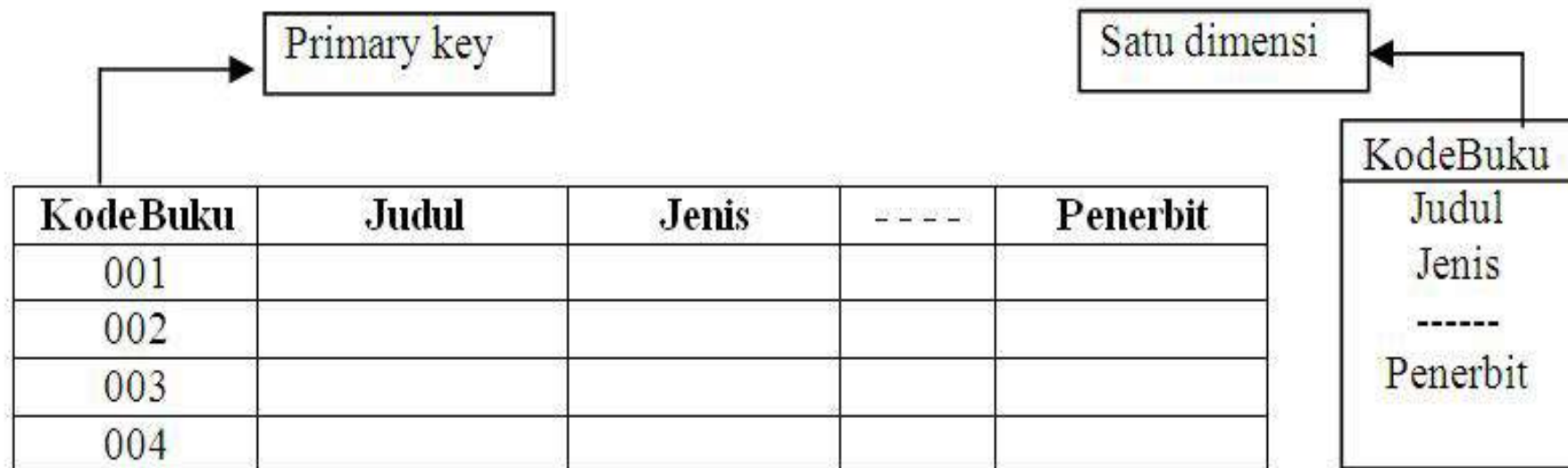
- **Data Loading**

Setelah selesai melakukan desain dan konstruksi dari *data warehouse* dan aplikasi digunakan untuk pertama kalinya, akan dilakukan pengisian awal data ke dalam media penyimpanan *data warehouse*. Dalam pengisian awal, dilakukan pemindahan data dalam jumlah yang besar.

Teknik Pemodelan Data Warehouse

- Pemodelan data warehouse penting karena untuk meyakinkan semua objek data yang diperlukan oleh database telah terpenuhi.
- **Tabel Relasional** → dibangun oleh baris dan kolom. Terdapat 2 sudut pandang: baris sebagai sumbu x dan kolom sebagai sumbu y.
- Tetapi sebenarnya tabel tersebut hanya mempunyai 1 dimensi saja.

Tabel Relasional



Tabel Relasional

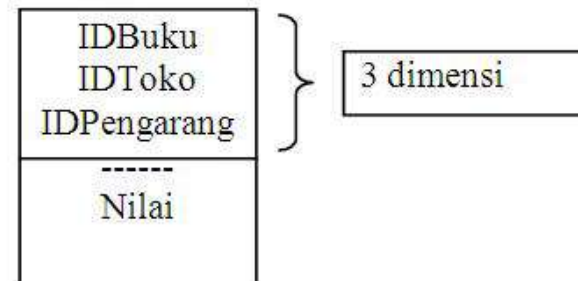
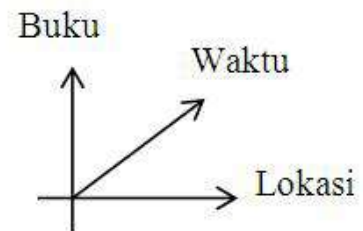
- Setiap record atau baris merepresentasikan data buku yang berbeda-beda.
- Satu baris dengan baris lainnya diidentifikasi dengan sebuah key yaitu primary key.
- Sedangkan bagian kolom, seperti: judul, jenis buku, pengarang menyimpan fakta yang sama atau sejenis, di mana setiap fakta tersebut merujuk pada primary key yaitu Kode Buku.
- Hal inilah yang menunjukkan bahwa tabel relasional hanya mempunyai satu dimensi.

Data Multidimensi

- Data multidimensi adalah ketika kita dapat melihat sebuah data dari berbagai sudut pandang atau dimensi.
- Contoh: penjualan buku dapat dilihat dari segi Buku, waktu, lokasi penjualan / toko dan sebagainya.
- Jika digambarkan, maka akan terdapat tiga koordinat yaitu sumbu x mewakili buku, sumbu y mewakili dimensi waktu dan sumbu z untuk dimensi lokasi.
- Hal inilah yang menjadi perbedaan mendasar antara tabel relasional dan data multidimensi.

...

	Kamal	Peb 01	Peb 02	----	Peb 04
Bangkalan	Jan 01	Jan 02	----	Jan 04	
Buku A			----		
Buku B			----		
Buku C			----		



Pemodelan Data Multidimensi

- Data warehouse dan OLAP dibangun berdasarkan multidimensional data model. Pada model ini diperlukan tabel fakta dan tabel dimensi.
- Berbeda dengan konsep normalisasi (3rd normal form).
- Tabel fakta → berisi measurement atau metric dari proses bisnis dan foreign key dari tabel dimensi.
- Tabel fakta merupakan tabel utama dari cube.

...

- Karakteristik tabel fakta: kumpulan key dimensi dari tabel, ada measure(yang ingin diukur) dan data akan selalu berubah.
- Contoh Measurement:

Jika anda mempunyai bisnis penjualan sepeda motor maka measurement dari bisnis anda adalah “jumlah penjualan motor” atau “rata-rata penjualan sepeda motor merk x”

...

- Tabel dimensi → berisi atribut dari measurement yang disimpan pada tabel fakta.
- Tabel dimensi merupakan hierarki, kategori dan logic yang dapat digunakan untuk menganalisis measurement dari sudut pandang tertentu.
- Tabel dimensi bersifat statis (tidak berubah)

Pemodelan Multidimensi

Dalam dimensional modeling, ada beberapa pendekatan yang digunakan untuk membuat data warehouse, yaitu:

- Skema bintang (*star schema*)
- Skema bola salju (*snowflake Schema*)
- *Fact constellations (galaxy schema)*

Skema Bintang / Star Schema (1)

- Skema ini mengikuti bentuk bintang, di mana terdapat satu tabel fakta (*fact table*) di pusat bintang dengan beberapa tabel dimensi (*dimensional tables*) yang mengelilinginya.
- Semua tabel dimensi berhubungan dengan ke tabel fakta. Tabel fakta memiliki beberapa key yang merupakan kunci indek individual dalam tabel dimensi.

Star Schema (2)

- Setiap tabel dimensi berelasi langsung dengan fact table.
- Tabel dimensi berisikan data tentang informasi atau waktu.
- Relasi antara fact table dengan dimensi-dimensinya adalah 1–N (one to many).
- Tabel dimensi memiliki *primary key* sederhana yang mengandung hanya satu atau dua kolom saja. Namun, tabel fakta akan memiliki sekumpulan *foreign key* yang disusun dari *primary key* komposit dan merupakan gabungan kolom-kolom tabel dimensi yang berelasi.

Contoh Star Schema (1)

ProductDimension

ProductID
ProductCode
ProductName
Category
SubCategory
Brand
Height
Width

SalesFact

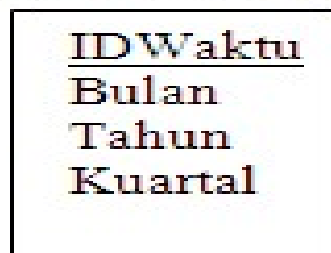
ProductID (FK)
TimeID (FK)
SalesDollars

TimeDimension

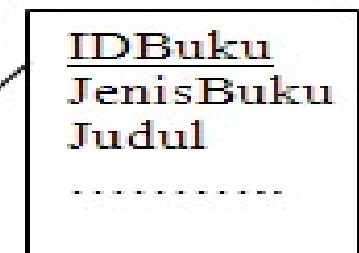
TimeID
DayOfWeek
DayOfMonth
DayOfYear
Month
Quarter
Year
Holiday
Weekend

Contoh Star Schema (2)

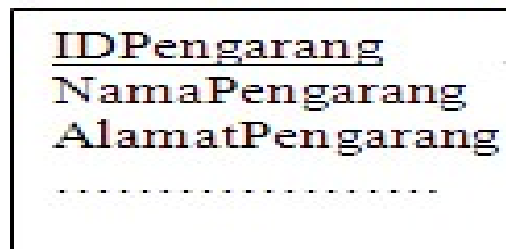
Dimensi Waktu



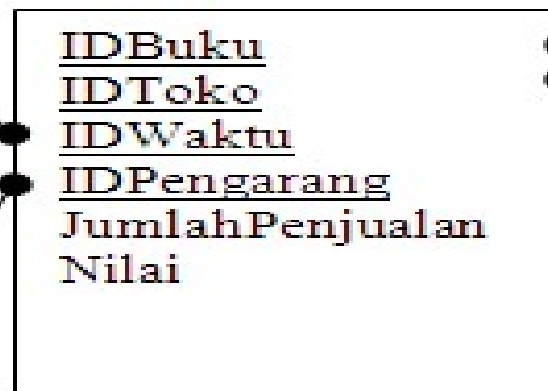
Dimensi Buku



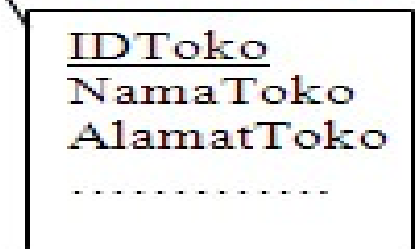
Dimensi Pengarang



Fact Penjualan



Dimensi Toko



■ ■ ■

- Dalam *star schema*, queri yang terbentuk antara tabel fakta dan sejumlah tabel dimensi dinamakan ***star query***.
- Setiap tabel dimensi direlasikan dengan tabel fakta berdasarkan kolom *primary key* dan *foreign key*, namun diantara masing-masing tabel dimensi tidak ada yang saling berelasi (tidak ada hubungan data).
- Qeri yang terbentuk menyebabkan proses eksekusi yang lebih optimal, karena rencana eksekusi queri dalam DBMS akan lebih cepat dengan setiap tabel hanya berelasi dengan satu tabel yang lain.

Kelebihan dan Kekurangan Star Schema

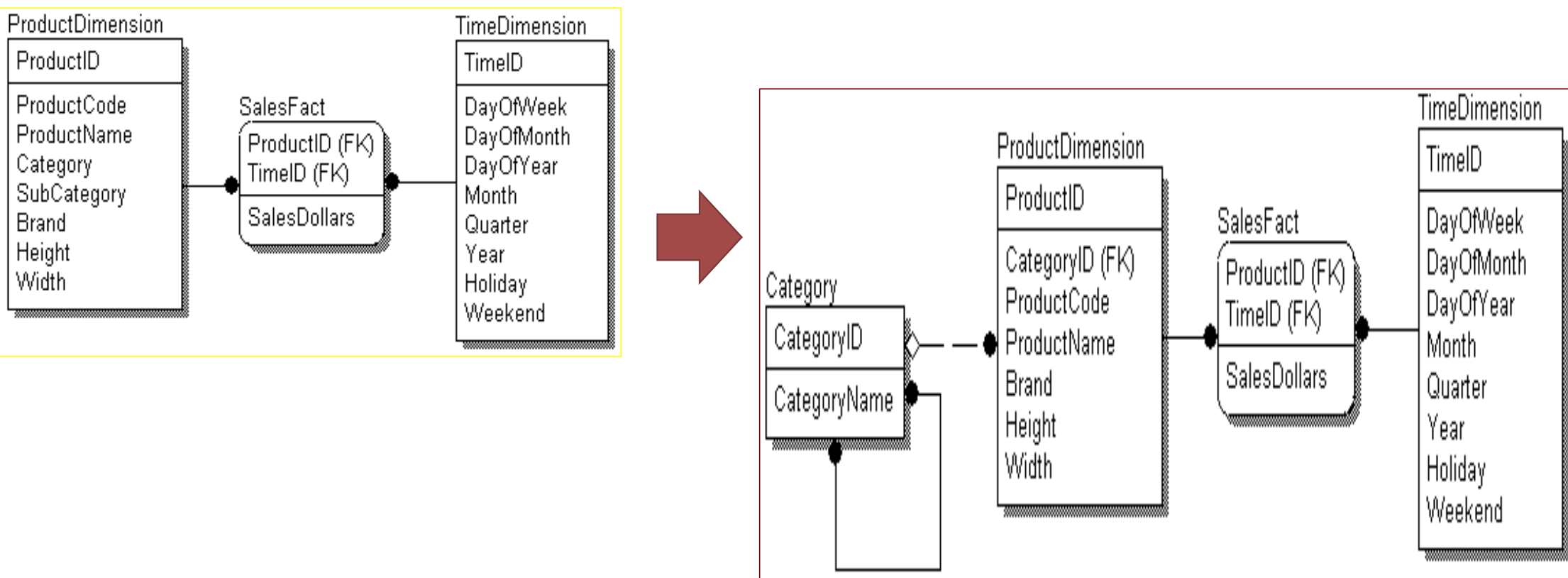
- Kelebihan:
 - Sederhana
 - Mudah dipahami
 - Proses query data lebih cepat
- Kekurangan:
 - Boros dalam space

Skema Bola Salju / Snowflake Schema

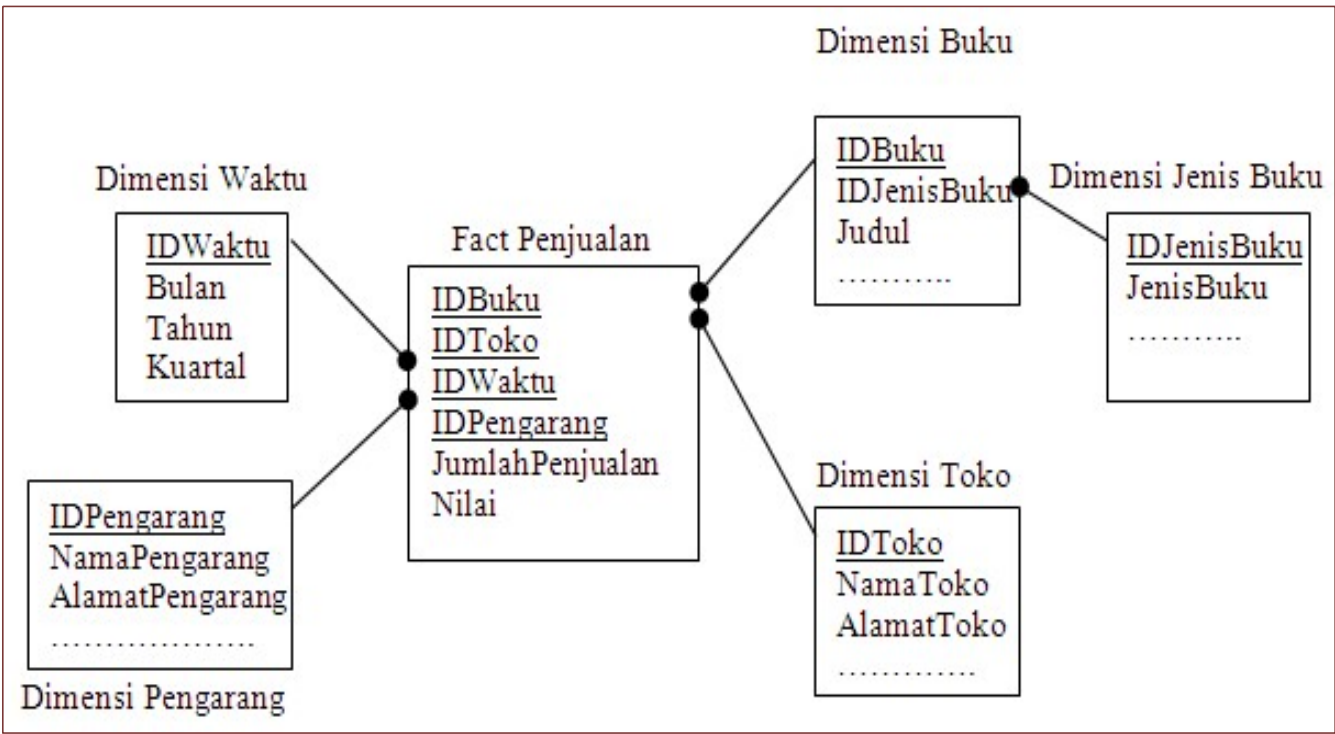
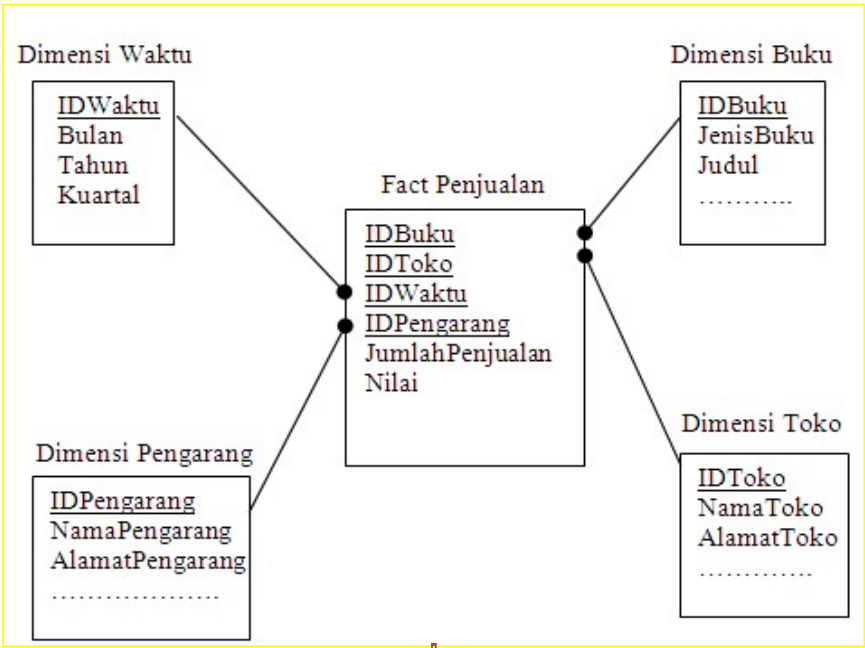
- Struktur basis data ini lebih kompleks dari pada *star schema*, dengan menormalisasi tabel-tabel dimensi yang berukuran besar dengan satu atau lebih kolom yang memiliki duplikasi data.
- Tabel dimensi dinormalisasi dengan cara men-split data pada tabel dimensi ke dalam tabel tambahan.

Contoh Snowflake Schema (1)

- Misal jika tabel dimensi *Product* dinormalisasi maka akan menghasilkan struktur seperti berikut:



Contoh Snowflake Schema (2)



...

- Tabel dimensi dinormalisasi untuk mengurangi redudansi data (duplikasi), sehingga struktur tabelnya akan lebih ramping.
- Dengan pengelompokan ini, data akan lebih mudah dibaca dan membantu pengembang aplikasi untuk menata desain antarmuka sistem dan *filtering* data.
- Struktur ini akan menghemat kapasitas *storage*, namun waktu eksekusi data akan lebih lama mengingat jumlah tabel dimensi yang direlasikan lebih banyak dan membutuhkan tambahan relasi *foreign key*.

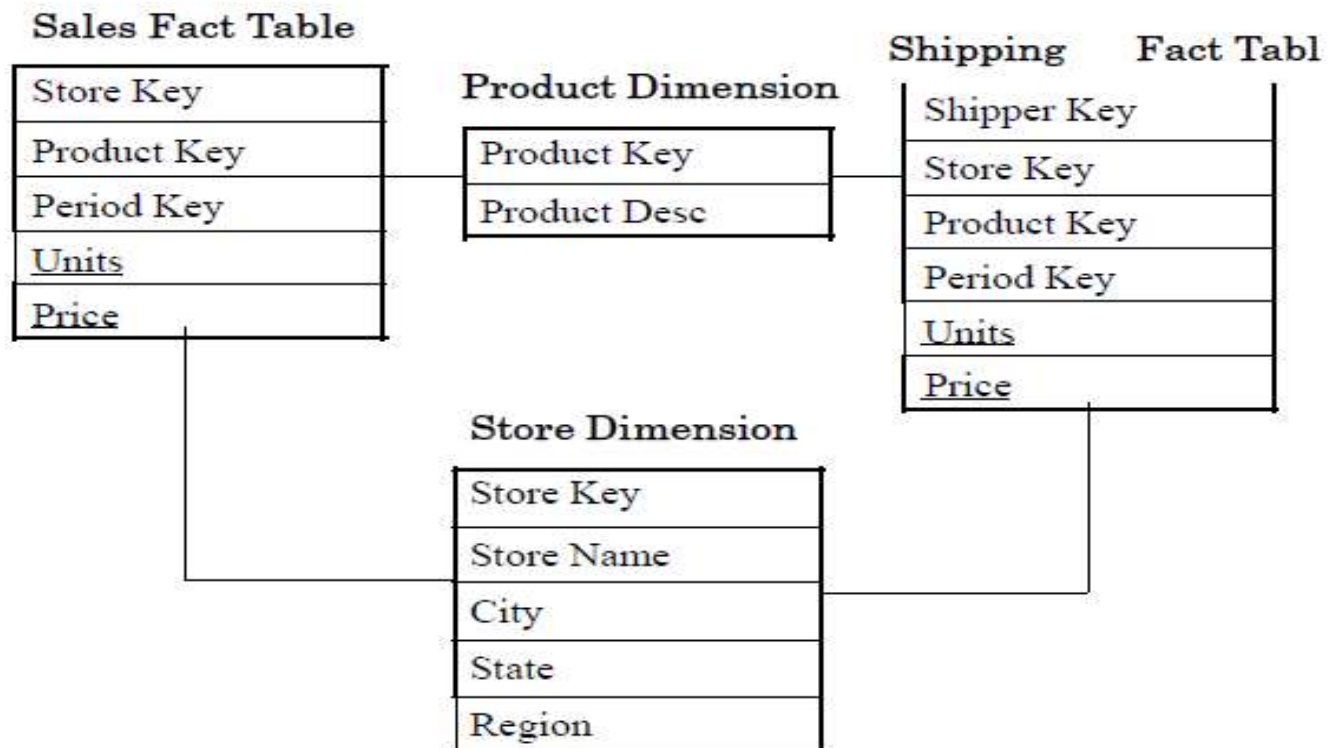
Kelebihan dan Kekurangan Snowflake Schema

- Kelebihan:
 - Pemakaian space lebih sedikit
 - Update dan maintenance lebih mudah
- Kekurangan:
 - Model menjadi kompleks dan rumit
 - Proses query lebih lama
 - Performance kurang bagus

Fact constellations (Galaxy Schema)

- Pada skema ini terdapat beberapa tabel fakta yang menggunakan satu atau beberapa tabel dimensi secara bersama-sama sehingga jika digambarkan akan terlihat seperti sekumpulan bintang.

Contoh Galaxy Schema (1)



Contoh Data Warehouse & ETL

- Data Mahasiswa dalam bentuk Excel

	R	S	T	U	V	W	X	Y	
1	KOTA_ORTU	NAMA_PROPINSI	TELPON_ORTU	KERJA_ORTU	NAMA_SMA	KOTA_SMA	PROPINSI_SMA	JUR_SMA	K
2	Kota Salatiga	Jawa Tengah	0298315604	NULL	SMA NEGERI 3 SALATIGA	SALATIGA	Jawa Tengah	BAHASA	0
3	Kab. Semarang	Jawa Tengah		1	SMA NEGERI 3 SALATIGA	SALATIGA	Jawa Tengah	IPS	0
4	Kota Semarang	Jawa Tengah	0243569709	NULL	SMA KRISTA MITRA SEMARANG	SEMARANG	Jawa Tengah	IPS	0
5	Kab. Banyumas	Jawa Tengah	0811262119	4	SMA IG.SLAMET RIYADI SOLO	SURAKARTA	Jawa Tengah	IPS	0
6	Kab. Halmahera Utara	Maluku Utara	081356401234	1	SMA KRISTEN TOBELO	TOBELO	Maluku Utara	IPA	0
7	Kab. Semarang	Jawa Tengah		5	SMA St. Louis, Semarang	Kota Semarang	Jawa Tengah	IPS	0
8	Kota Salatiga	Jawa Tengah		4	SMA Theresiana, Salatiga	Kota Salatiga	Jawa Tengah	IPS	0
9	Kota Salatiga	Jawa Tengah		1	SMA Kristen 2, Salatiga	SALATIGA	Jawa Tengah	IPA	0
10	Kab. Semarang	Jawa Tengah		1	SMA Negeri 3, Salatiga	SALATIGA	Jawa Tengah	IPA	0
11	Kota Bandung	Jawa Barat		1	SMTA Lain-lain	BANDUNG	Jawa Barat	BAHASA	0
12	Kota Tegal	Jawa Tengah		1	SMA Negeri 2, Slawi	TEGAL	Jawa Tengah	IPS	0

- Data warehouse gunakan database MySQL.

Menentukan Schema Data Warehouse

- Star Schema
- Snowflake Schema
- Galaxy Schema

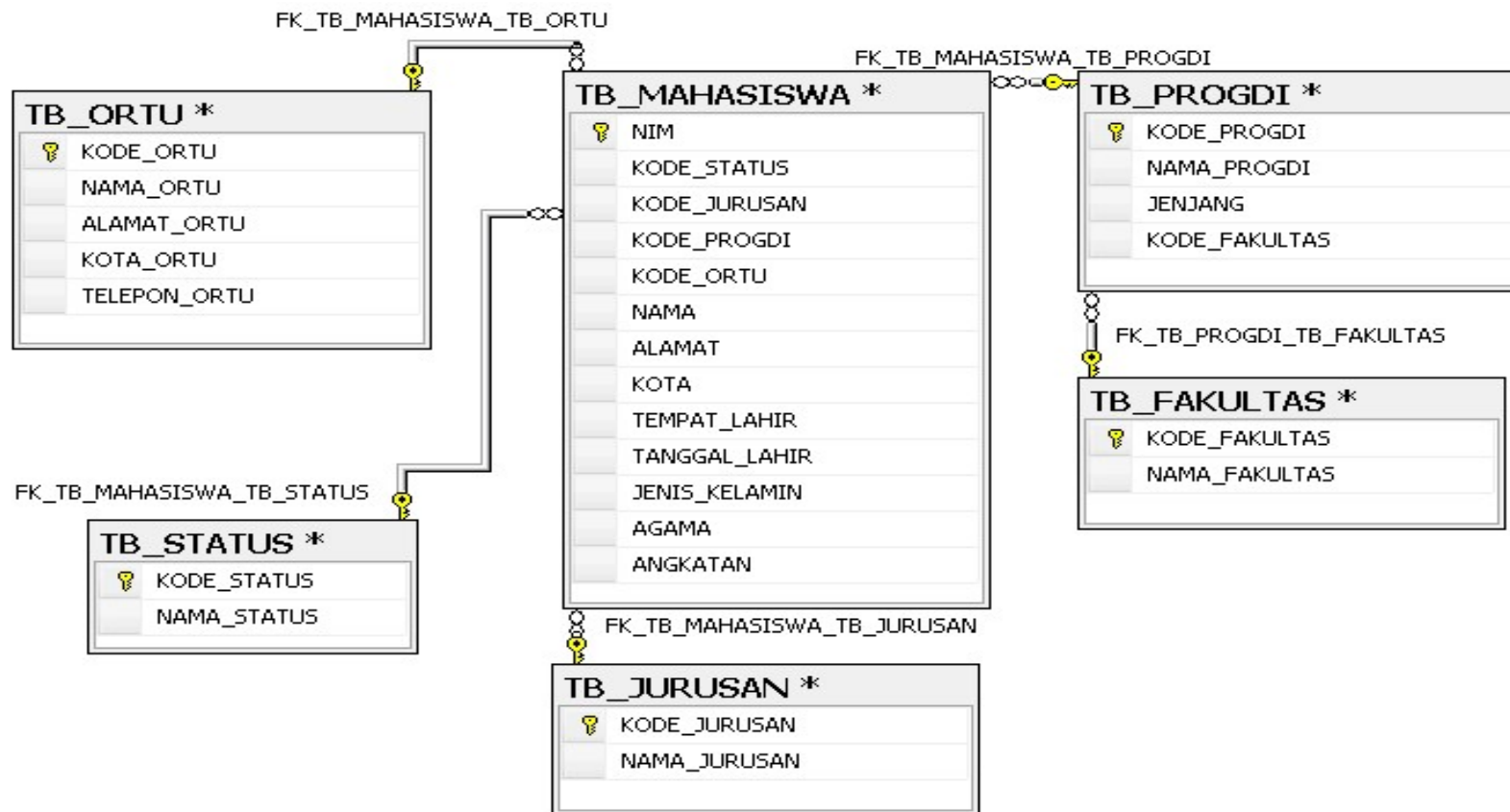
Contoh Data Warehouse & ETL

- Data Warehouse Data Mahasiswa

Tabel yang dibutuhkan:

- Tabel Mahasiswa
- Tabel Program Studi
- Tabel Status
- Tabel Jurusan
- Tabel Orang Tua
- Tabel Fakultas


Snowflake Schema




Tabel TB_MAHASISWA

	Column Name	Data Type	Allow Nulls
▶🔑	NIM	char(10)	<input type="checkbox"/>
	KODE_STATUS	char(10)	<input checked="" type="checkbox"/>
	KODE_JURUSAN	varchar(50)	<input checked="" type="checkbox"/>
	KODE_PROGDI	char(10)	<input checked="" type="checkbox"/>
	KODE_ORTU	int	<input checked="" type="checkbox"/>
	NAMA	varchar(50)	<input checked="" type="checkbox"/>
	ALAMAT	varchar(50)	<input checked="" type="checkbox"/>
	KOTA	varchar(50)	<input checked="" type="checkbox"/>
	TEMPAT_LAHIR	varchar(50)	<input checked="" type="checkbox"/>
	TANGGAL_LAHIR	varchar(50)	<input checked="" type="checkbox"/>
	JENIS_KELAMIN	varchar(50)	<input checked="" type="checkbox"/>
	AGAMA	varchar(50)	<input checked="" type="checkbox"/>
	ANGKATAN	int	<input checked="" type="checkbox"/>

Tabel TB_PROGDI

	Column Name	Data Type	Allow Nulls
	KODE_PROGDI	char(10)	<input type="checkbox"/>
	NAMA_PROGDI	varchar(50)	<input type="checkbox"/>
	JENJANG	char(10)	<input type="checkbox"/>
	KODE_FAKULTAS	varchar(50)	<input type="checkbox"/>

Tabel TB_ORTU

	Column Name	Data Type	Allow Nulls
	KODE_ORTU	int	<input type="checkbox"/>
	NAMA_ORTU	varchar(50)	<input checked="" type="checkbox"/>
	ALAMAT_ORTU	varchar(50)	<input checked="" type="checkbox"/>
	KOTA_ORTU	varchar(50)	<input checked="" type="checkbox"/>
	TELEPON_ORTU	varchar(50)	<input checked="" type="checkbox"/>

Tabel TB_STATUS

	Column Name	Data Type	Allow Nulls
▶ 🔑	KODE_STATUS	char(10)	<input type="checkbox"/>
	NAMA_STATUS	varchar(50)	<input checked="" type="checkbox"/>

Tabel TB_JURUSAN & TB_FAKULTAS

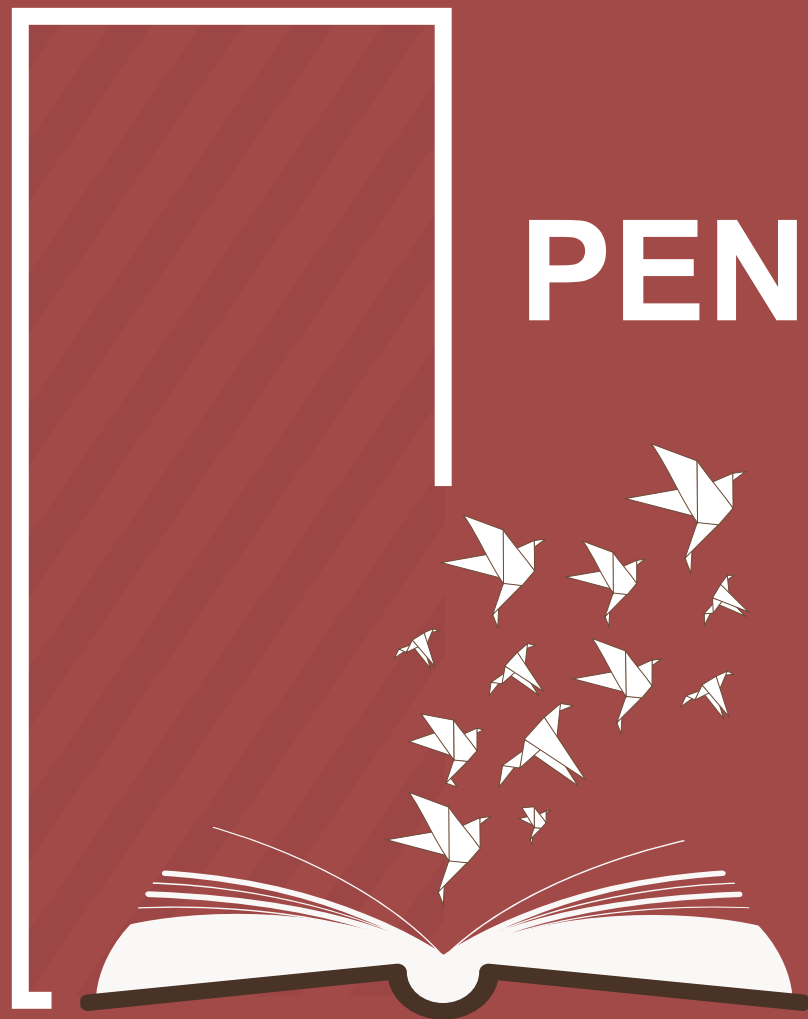
	Column Name	Data Type	Allow Nulls
PK	KODE_JURUSAN	varchar(50)	<input type="checkbox"/>
	NAMA_JURUSAN	varchar(50)	<input type="checkbox"/>

	Column Name	Data Type	Allow Nulls
PK	KODE_FAKULTAS	varchar(50)	<input type="checkbox"/>
	NAMA_FAKULTAS	varchar(50)	<input type="checkbox"/>

Proses ETL

- Proses extract sudah dilakukan, yaitu mengambil data dari sumber data berupa data mahasiswa dalam bentuk .xls.
- Lakukan proses transformasi di mana dilakukan proses cleansing, standarisasi dan penyesuaian dengan schema data warehouse yang dipilih.
- Setelah itu loading-kan data ke dalam data warehouse.

PENTAHO



Pentaho

- Pentaho adalah sebuah perusahaan commercial open source BI yang berpusat di Orlando, Amerika Serikat.

Pentaho Data Integration / Kettle

- Utilitas ETL (Extract, Transform and Load) open source paling populer.
- Designer GUI yang intuitif dan sangat mudah digunakan.
- Multi Platform.
- Script ETL dapat disimpan dalam bentuk filesystem maupun repository.
- Mendukung clustering (master-slave) engine ETL
- Terdiri atas lebih dari 200 step yang mencakup job (workflow kontrol) dan transformation (data workflow).
- Mendukung Apache Virtual Filesystem (Apache VFS) sehingga filesystem seperti HTTP Webdav, FTP, SFTP, dan lain sebagainya dapat dengan mudah diakses dengan konfigurasi yang minimal.

Tugas

Buatlah pemodelan salah satu schema datawarehouse pada studi kasus menggunakan Kettle? (Data berkaitan dengan kegiatan Pariwisata dan UMKM)