Data Warehouse

Sirojul Munir S.SI, M.KOM – Semester Genap TA 20182

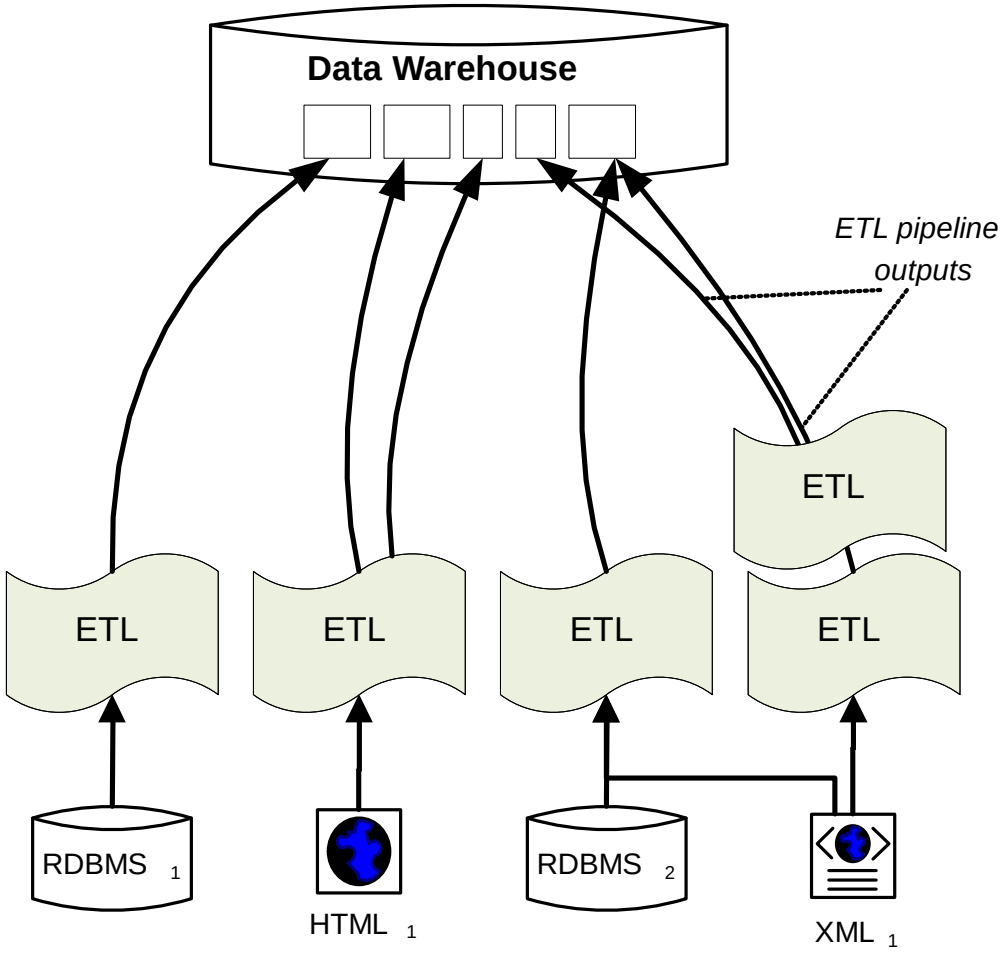
PROSES ETL DATA WAREHOUSE

Definisi : ETL

- ETL merupakan langkah di dalam pemrosesan data pada database yang melibatkan kegiatan peng-ekstrakan (**extracting**) data-data dari sumber data, mempertahankan kualitas data, menerapkan standarisasi data, menyajikan dalam berbagai bentuk (**transformation**) untuk kemudian dialirkan atau diteruskan (**loading**) ke data warehouse (**Ralph Kimball & Joe Caserta**)

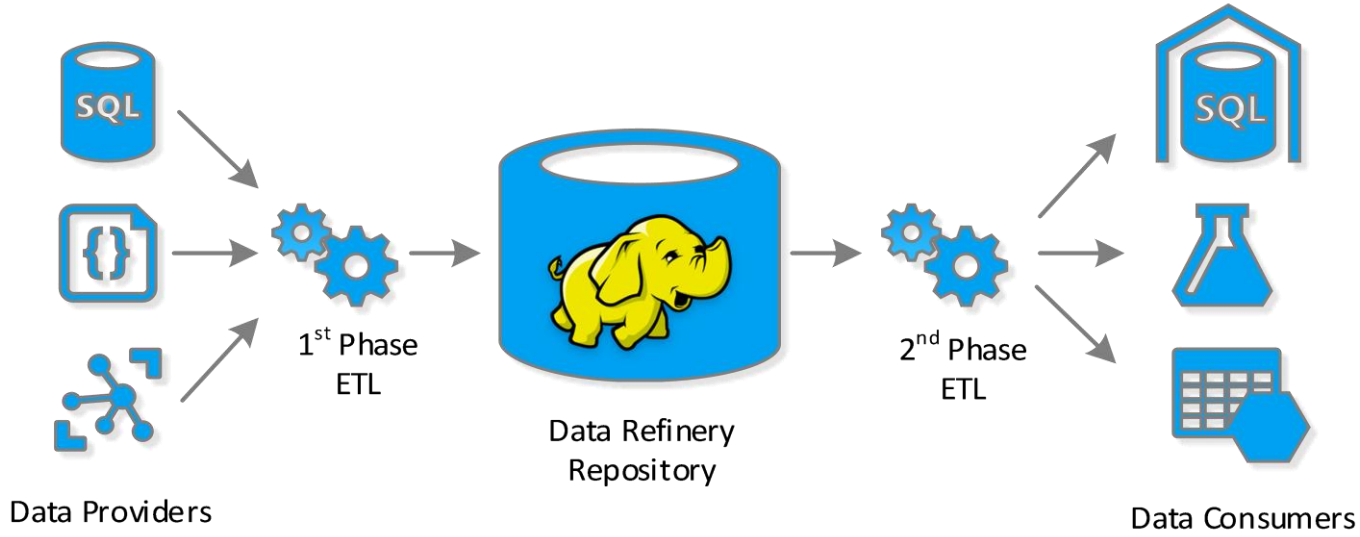
ETL Proses

- At the top – a centralized database
  - Generally configured for queries and appends – not transactions
  - Many indices, materialized views, etc.
- Data di load secara periodic di update ke Data Warehouse



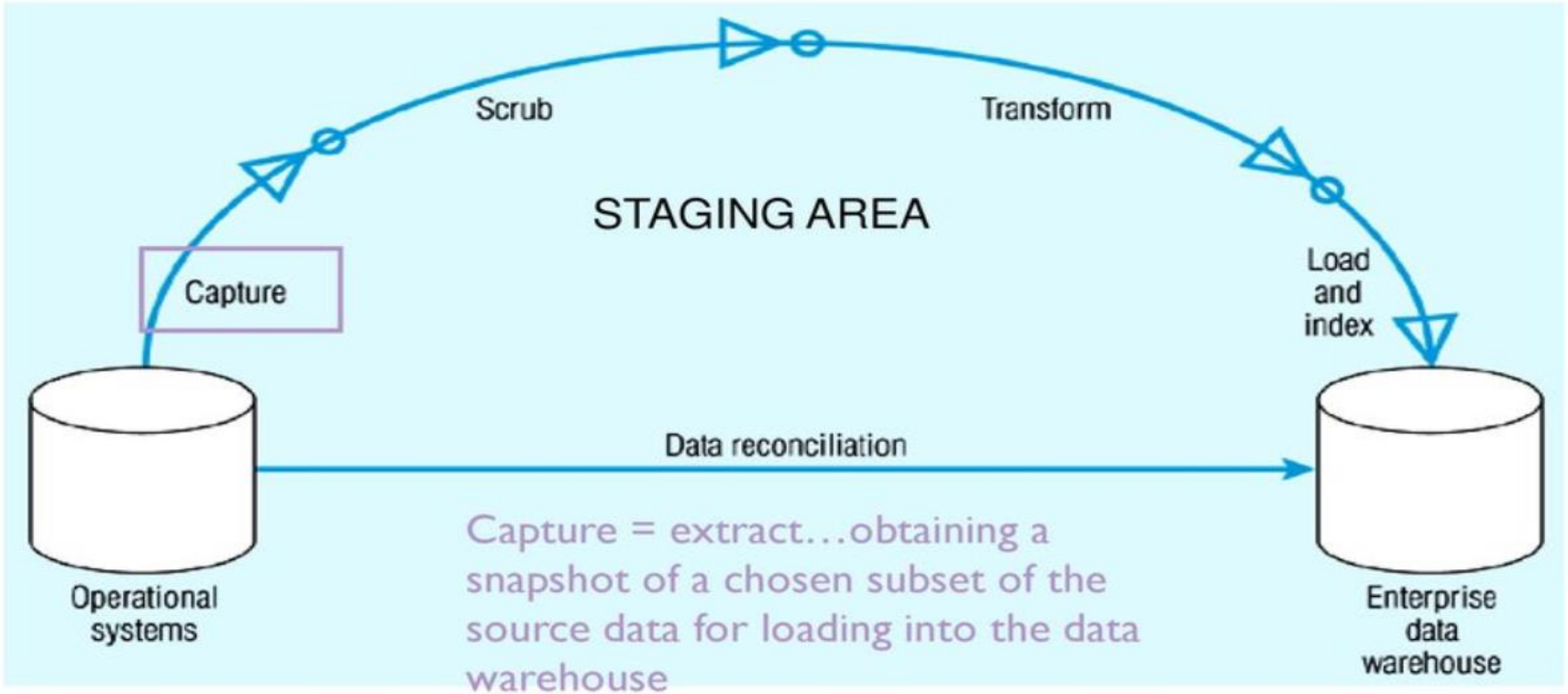
Fungsi : ETL

- **ETL** : Extraction Transformation Loading
- Membantu perangkat lunak komputer (software) yang dibangun untuk memudahkan di dalam mengambil file yang memuat data didalamnya
- Memudahkan proses mendapatkan informasi
- Dapat di integrasikan ke system database terkini seperti teknologi BIG DATA dan komputasi skala besar (Hadoop)



ETL :: Extraction – Rekonsilasi Data

Capture Data



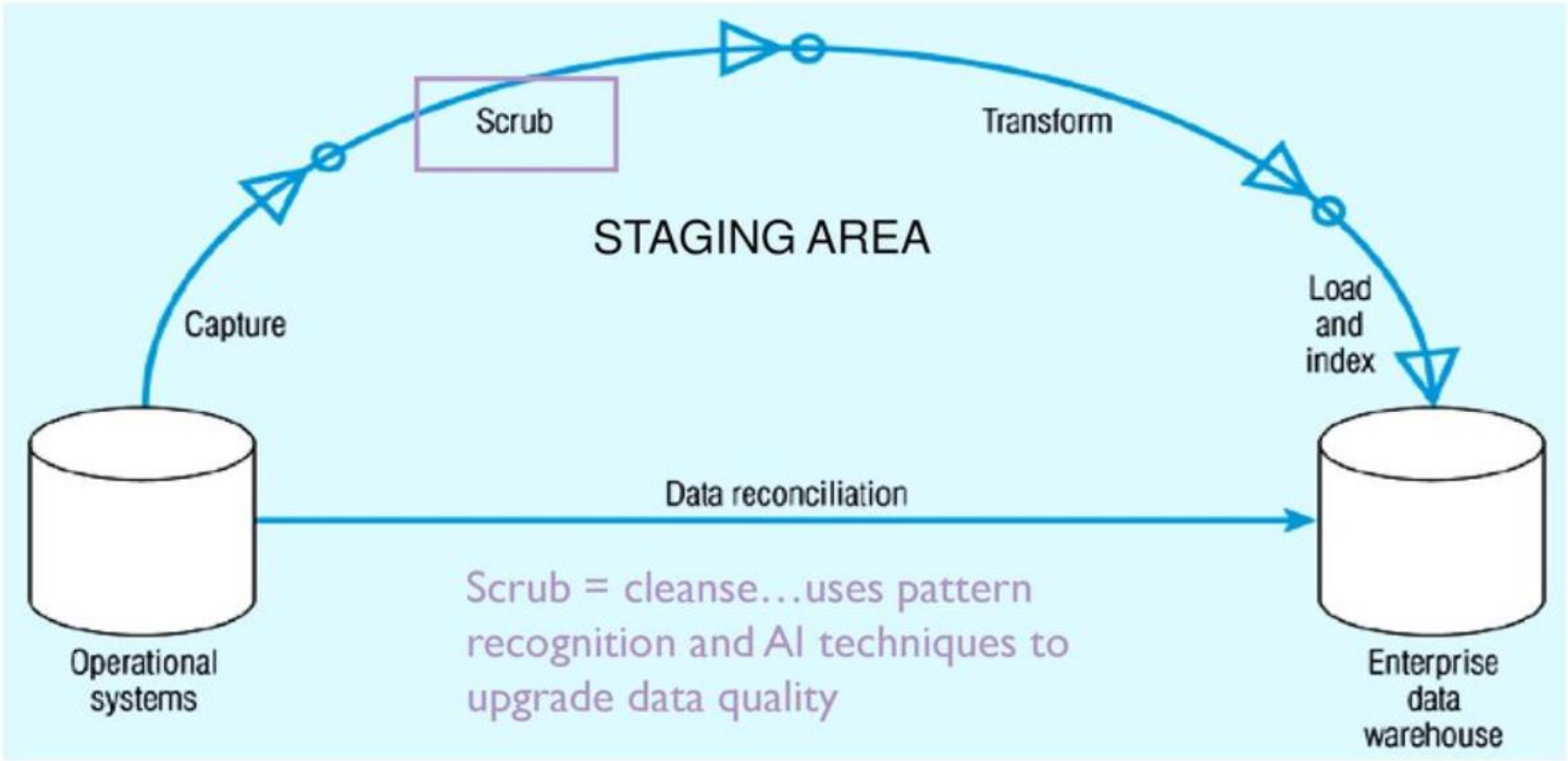
Static extract = mengambil data-data dari sumber pada waktu tertentu, dan biasanya hanya dilakukan sekali di awal proses.

Incremental extract = mengambil hanya data-data yang mengalami perubahan akibat static extract

ETL :: Extraction – Rekonsilasi Data

Scrub & Cleansing Data:

1. Parsing
2. Correcting
3. Standardizing
4. Matching
5. Consolidating



Fixing errors: salah ejaan, tanggal yang salah, penggunaan kolom yang salah, alamat yang tidak cocok, data yang hilang, data ganda, inkonsistensi

juga: decoding, reformatting, time stamping, konversi, key generation, penggabungan, deteksi error, pencarian data hilang

ETL :: Extraction – Step Cleansing Data

Parsing

- Melakukan parsing untuk menempatkan dan meng-identifikasi elemen data individual dalam file2 sumber dan kemudian melakukan isolasi data ini kedalam file target
- Contoh:
  - Parsing data untuk menentukan first name, last name , middle name
  - Parsing data untuk element field alamat : nama jalan, nama kota, nomor rumah

ETL :: Extraction – Step Cleansing Data

Correcting

- Melakukan koreksi hasil parsing komponen data individual dengan menggunakan algoritma data yang canggih dan sumber data sekunder

- Contoh:
  - Perbaikan data field alamat dan kode pos

ETL :: Extraction – Step Cleansing Data

Standardizing

- Penerapan standarisasi format data / konversi data yang konsisten sesuai aturan bisnis proses organisasi

- Contoh:
  - Pemberian prefix field nama (gelar depan) atau gelar belakang, standar data alamat / jalan

ETL :: Extraction – Step Cleansing Data

Matching

- Pencarian dan mencocokkan data dari seluruh data yang diparsing untuk diperbaiki dan memenuhi standar aturan bisnis yang ditentukan untuk menghilangkan duplikasi data

- Contoh:
  - Identifikasi nama atau alamat yang sama (double)

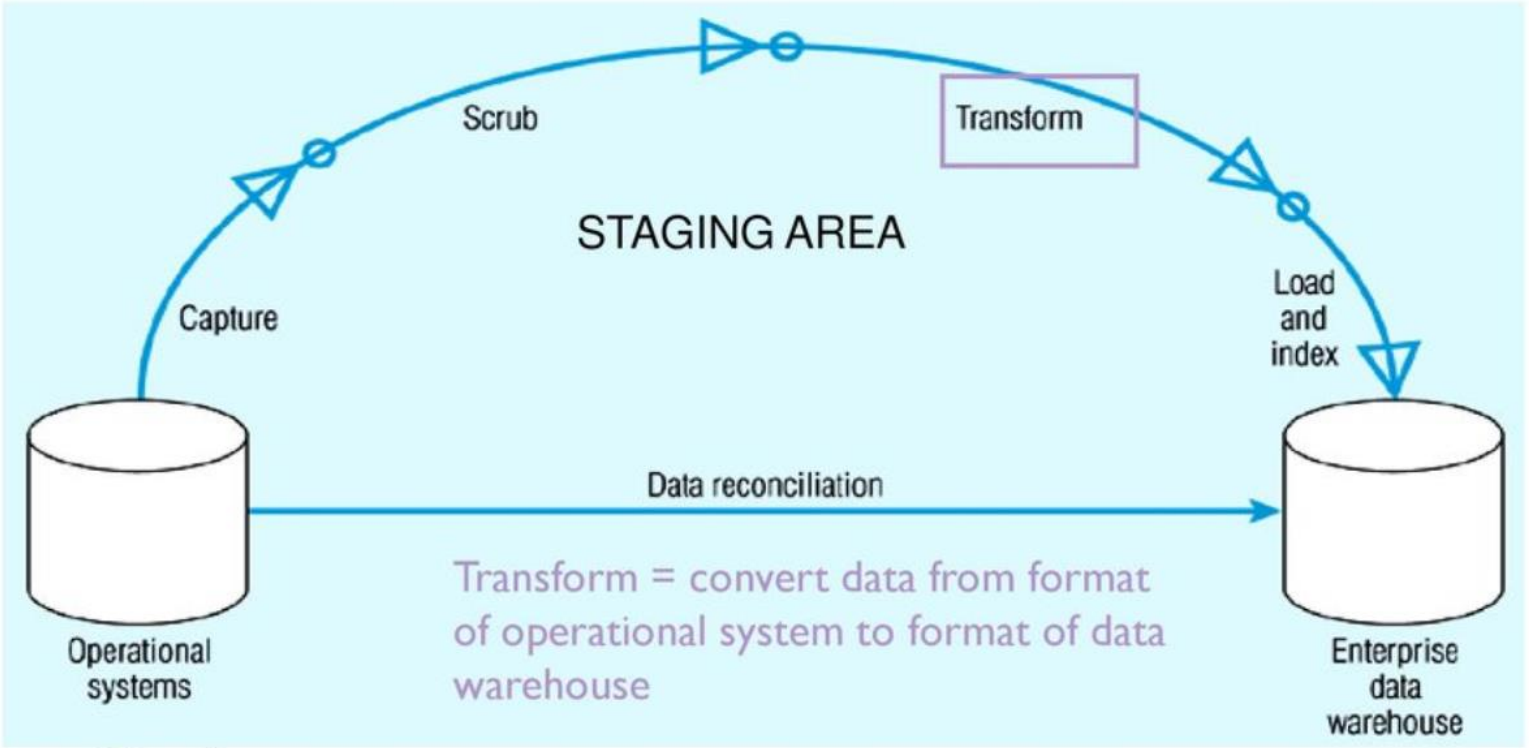
ETL :: Extraction – Step Cleansing Data

Consoliding

- Melakukan analisa dan identifikasi hubungan antara (relation) data dan melakukan konsolidasi/penggabungan menjadi SATU representasi data
- Contoh:
  - Konsolidasi data dari dua sumber OLTP menjadi satu representasi data; misal system CRM & ERP untuk representasi data produk/jasa/customer/employee

ETL :: Transformation

- Proses transformasi data sesuai dengan bisnis proses dan standard yang sudah dijalankan organisasi



Record-level:

Selection – pemisahan data
Joining – penggabungan data
Aggregation – peringkasan data

Field-level:

single-field – dari one field ke one field
multi-field – dari many fields ke one, atau one field ke many

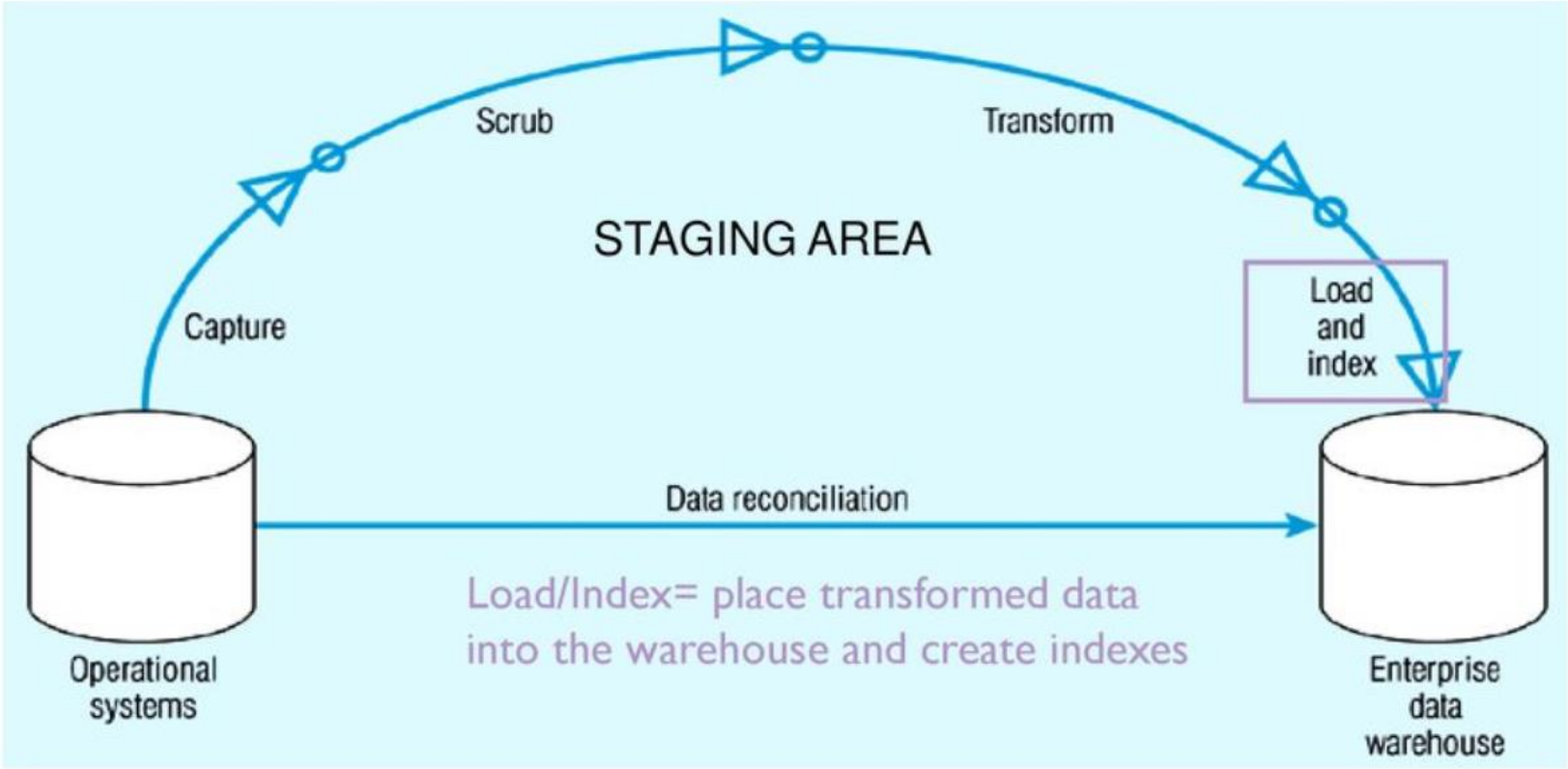
ETL :: Transformation

- Terdapat tiga buah level skema dalam penyeragaman format data pada proses transformation menurut Panos Vassiliadis, Alkis Simitsis : University of Loannina Georgia, yaitu:

1. **Schema Level Problem**
   Penyelesain format penamaan data jika terjadi nama yang sama pada subjek yang berbeda, tujuannya membuat satu representasi data yang sama.
2. **Record Level Problem**
   Penyelesaian masalah redundansi data (data duplikat) yang berasal dari berbagai sumber data.
3. **Value Level Problem**
   Penyelesaian pada nilai (value) data yang berbeda-beda dari setiap data, misalnya format data jenis kelamin Laki-laki - L , Perempuan P , format uang , format tanggal

ETL :: Loading & Index

- Proses data secara fisik berpindah ke data warehouse
- Proses update data ke data warehouse dapat dijalankan secara real time



Refresh mode: penulisan berulang data tujuan secara massal dan berkala

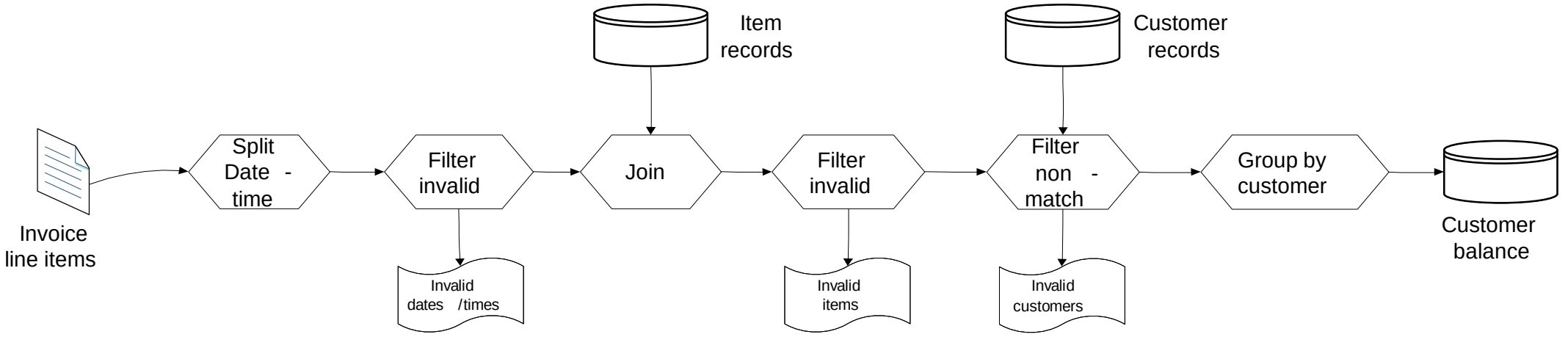
Update mode: hanya perubahan-perubahan pada data sumber yang dimasukkan ke data warehouse

ETL :: Loading

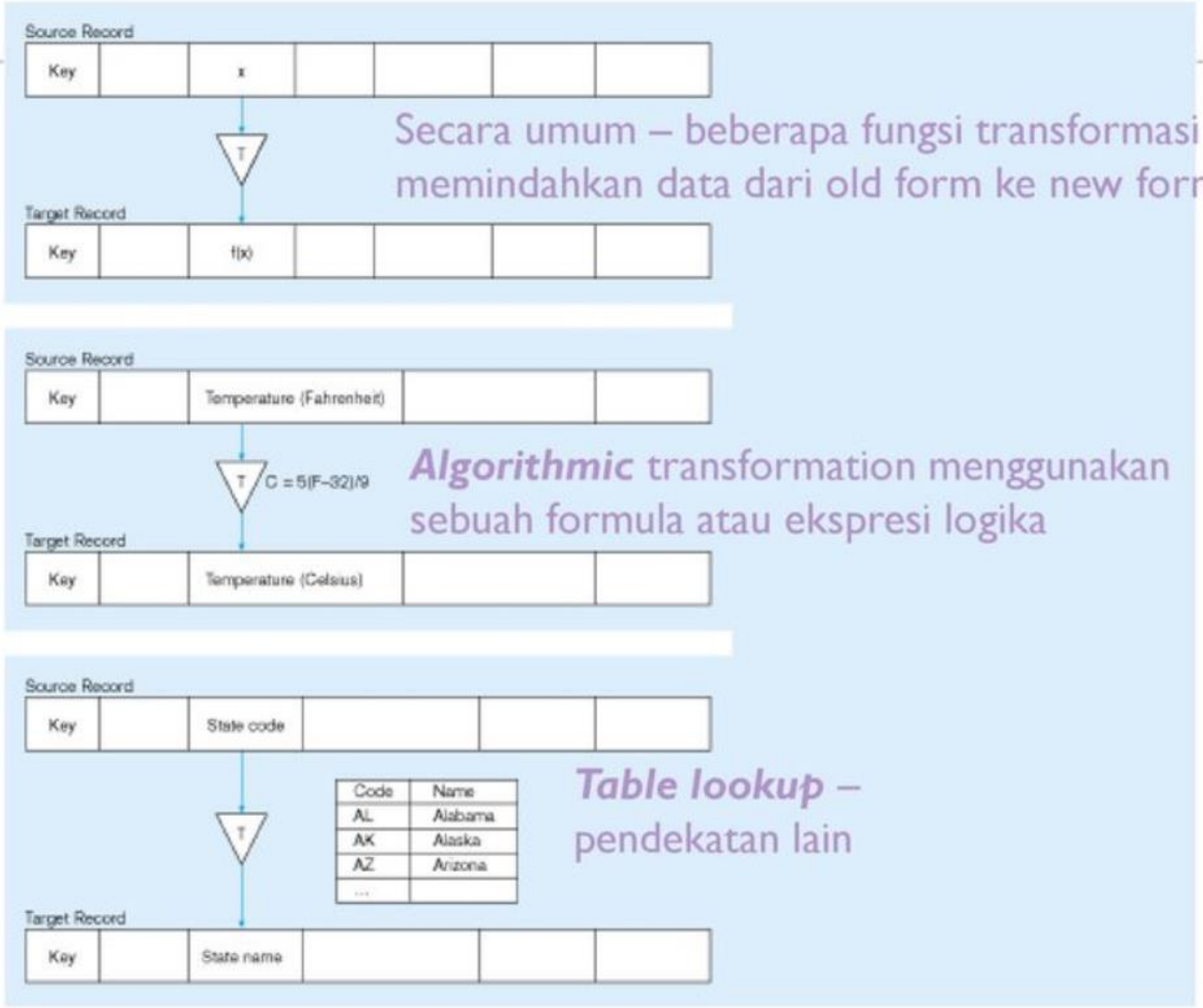
- Terdapat tiga hal yang mungkin dilakukan pada saat proses loading data:

1. **Load Up Data (LUD)**
   Mengalirkan data-data yang telah di integrasikan dan diseragamkan ke sebuah aplikasi pengguna (ditampilkan ke antar muka user).
2. **Load Insert Data (LID)**
   Memasukan data yang diteruskan ke dalam database pada data warehouse, sesuai dengan format metadata yang disepakati
3. **Load Bulk Data (LBD)**
   Meneruskan data dalam bentuk bulk data yang memuat data-data dan informasi di dalamnya, yang merupakan hasil penggabungan (integerasi) yang telah dilakukan pada proses transformation

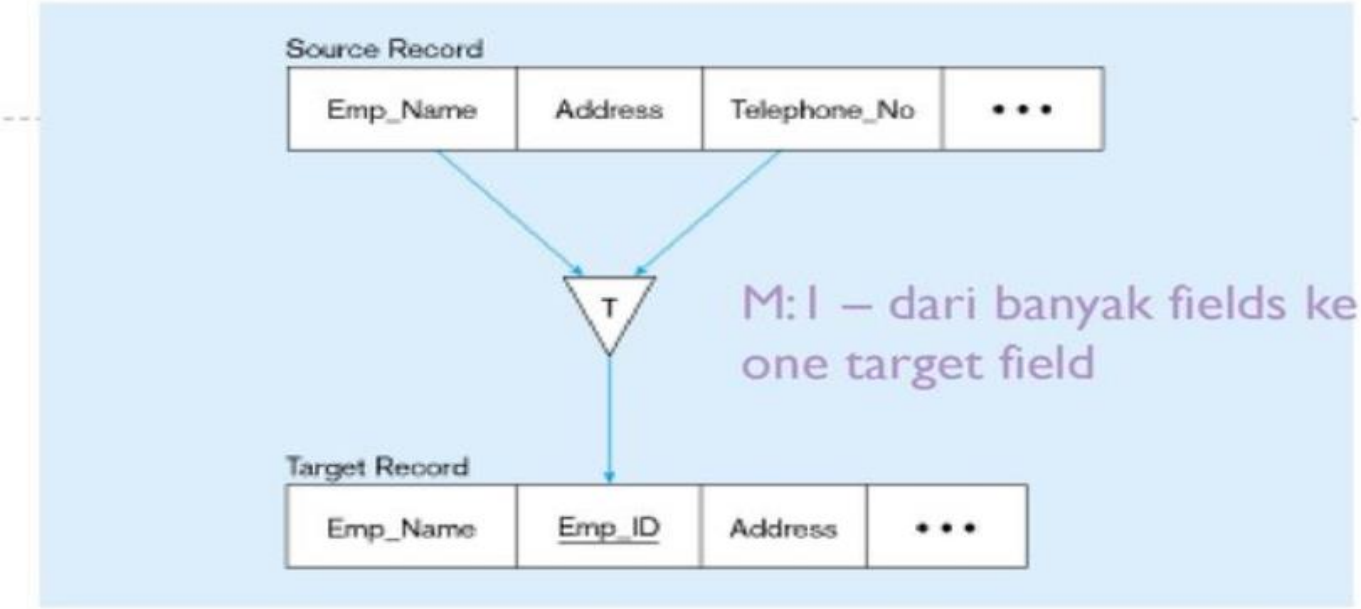
Contoh : e-commerce loading

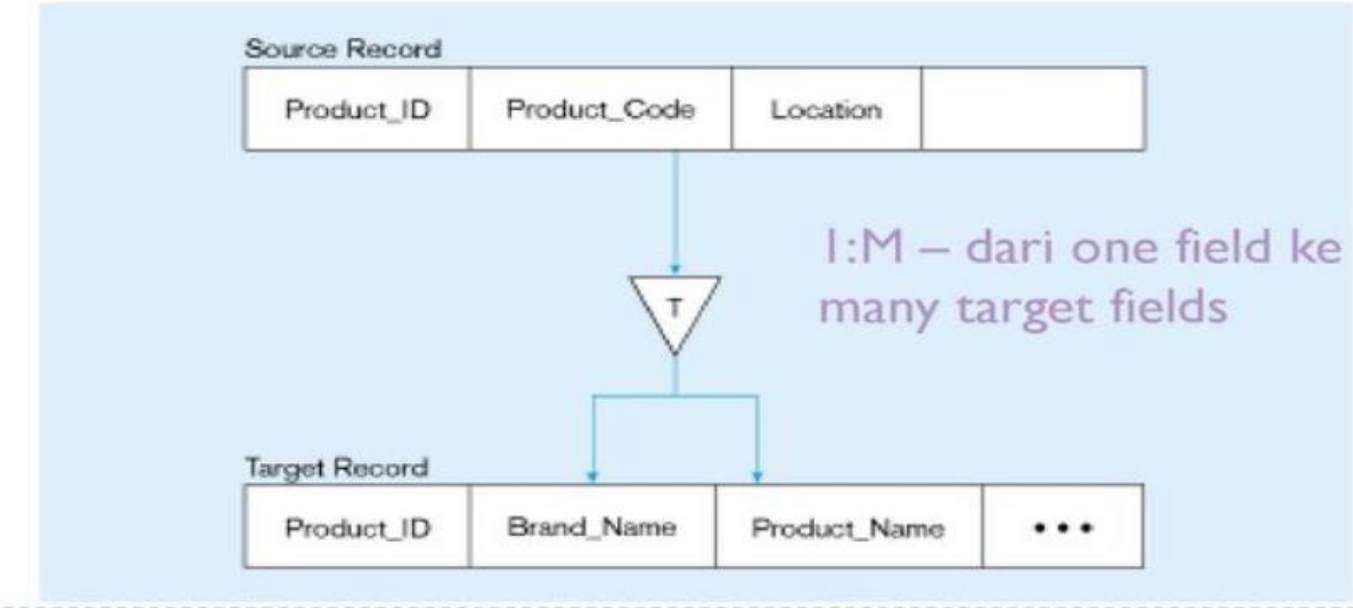


Model Transformation : Single Field Transformation



Model Transformation : MultiField Transformation





Karakteristik Data setelah : ETL

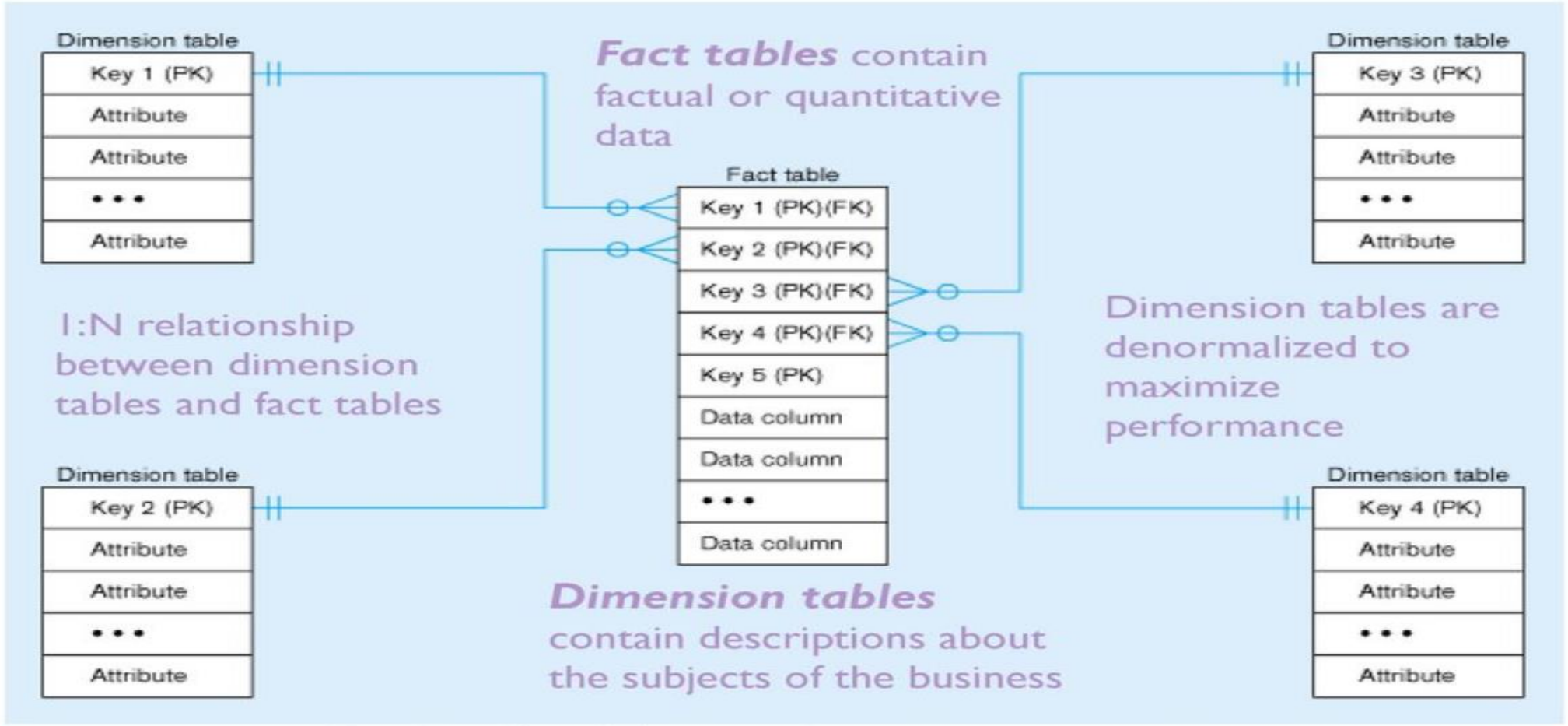
1. Terperinci ; Data terperinci, tidak sekedar peringkasan data
2. Historical; Data secara periodik
3. Ternormalisasi; 3rd NF atau lebih
4. Komprehensif; perpektif dan enterprise
5. Timely; up-to-date (tidak harus real time)
6. Quality Controlled; kualitas yang baik

Derived Data

- **Tujuan:**
  1. Mempermudah penggunaan aplikasi pendukung
  2. Respon yang cepat terhadap permintaan pengguna yang telah ditetapkan
  3. Data yang telah disesuaikan untuk pihak-pihak tertentu
  4. Dukungan untuk permintaan pelaporan
  5. Kemampuan untu data mining
- **Sifat:**
  1. Terperinci (data secara periodic)
  2. Ringkas (untuk penyimpanan)
  3. Terdistribusi (untuk diguanakan layanan bagian-bagian tertentu

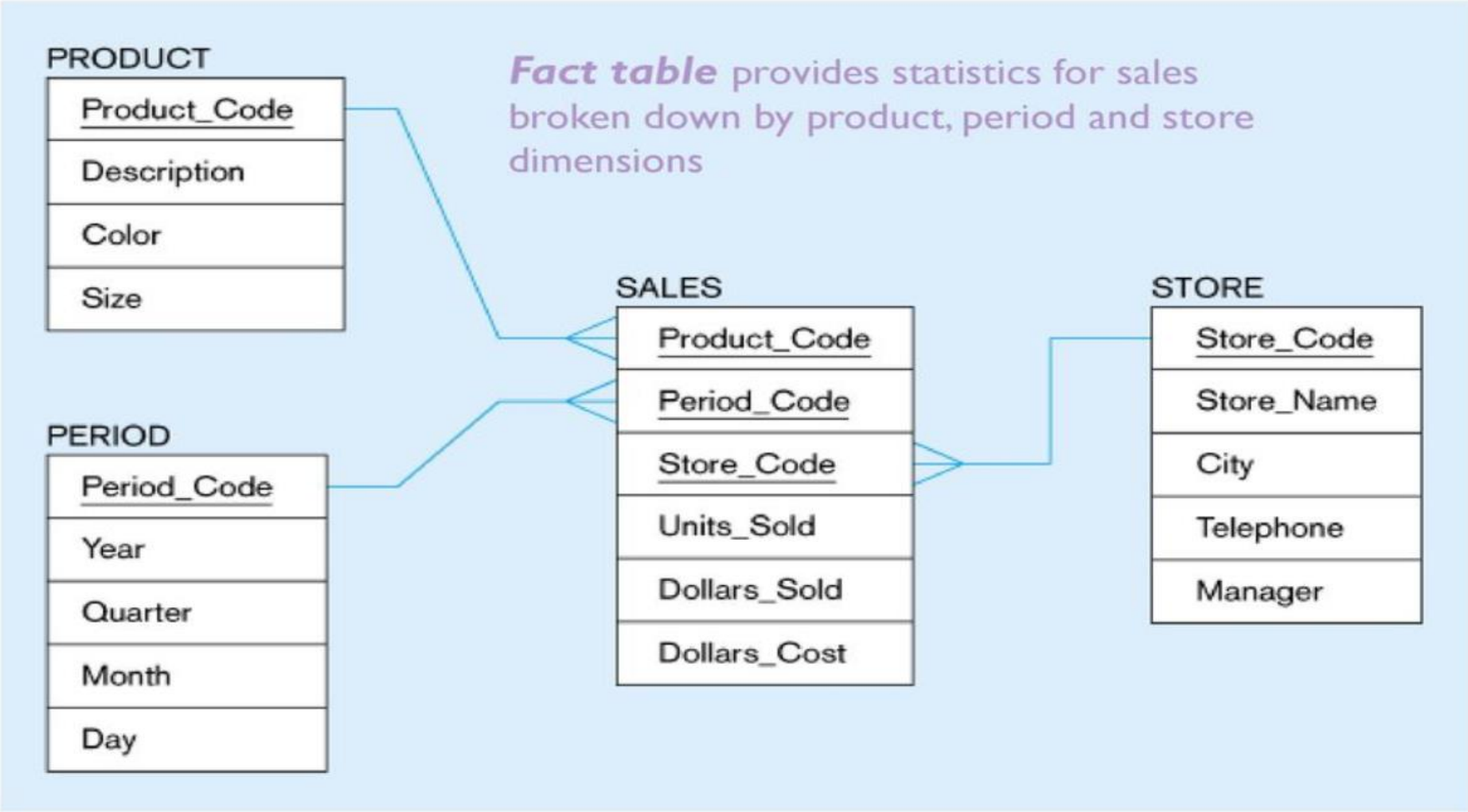
Model yang digunakan biasanya: star schema / dimensional model

Star Schema

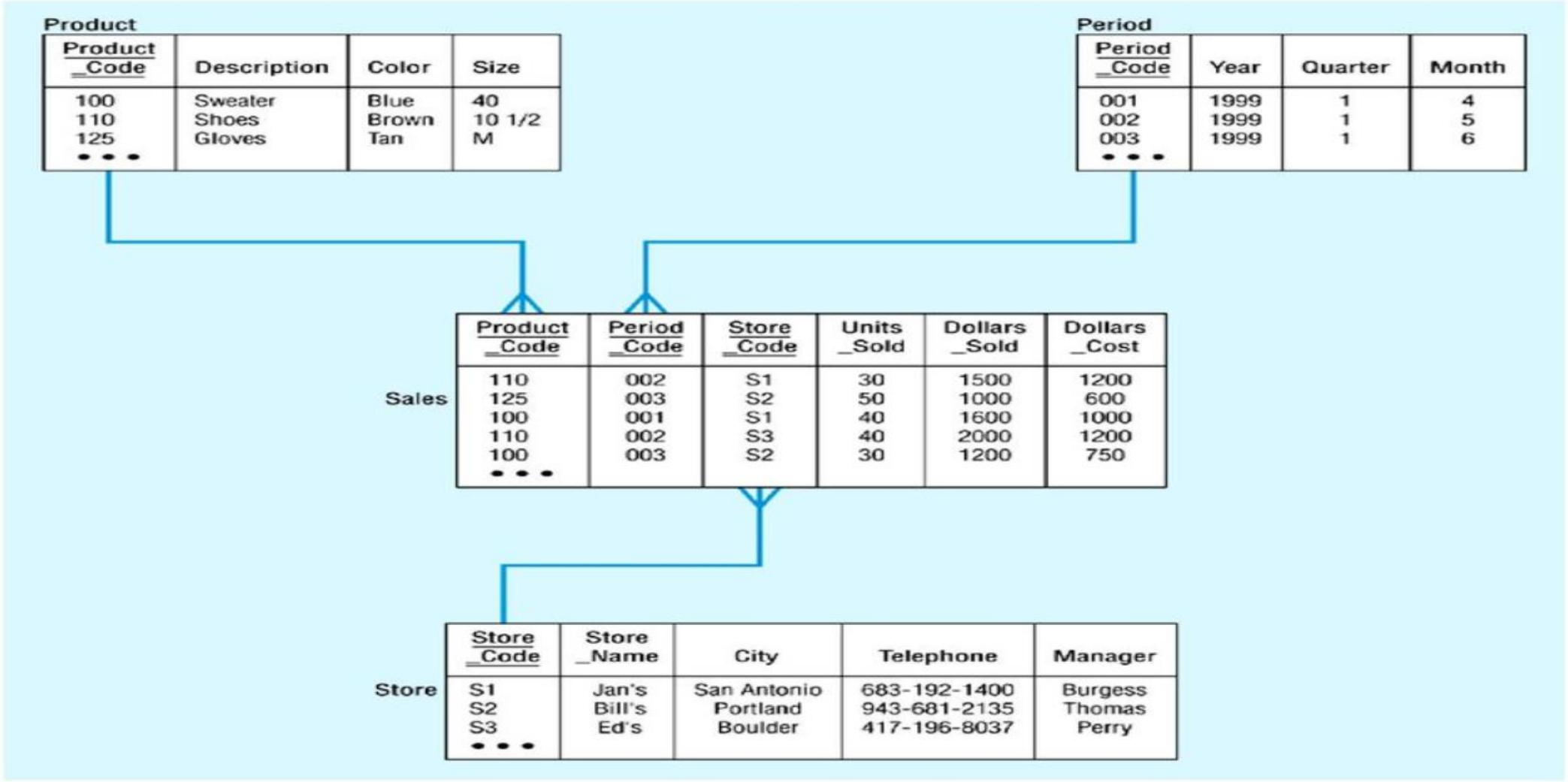


Excellent for ad-hoc queries,
but bad for online transaction processing

Star Schema

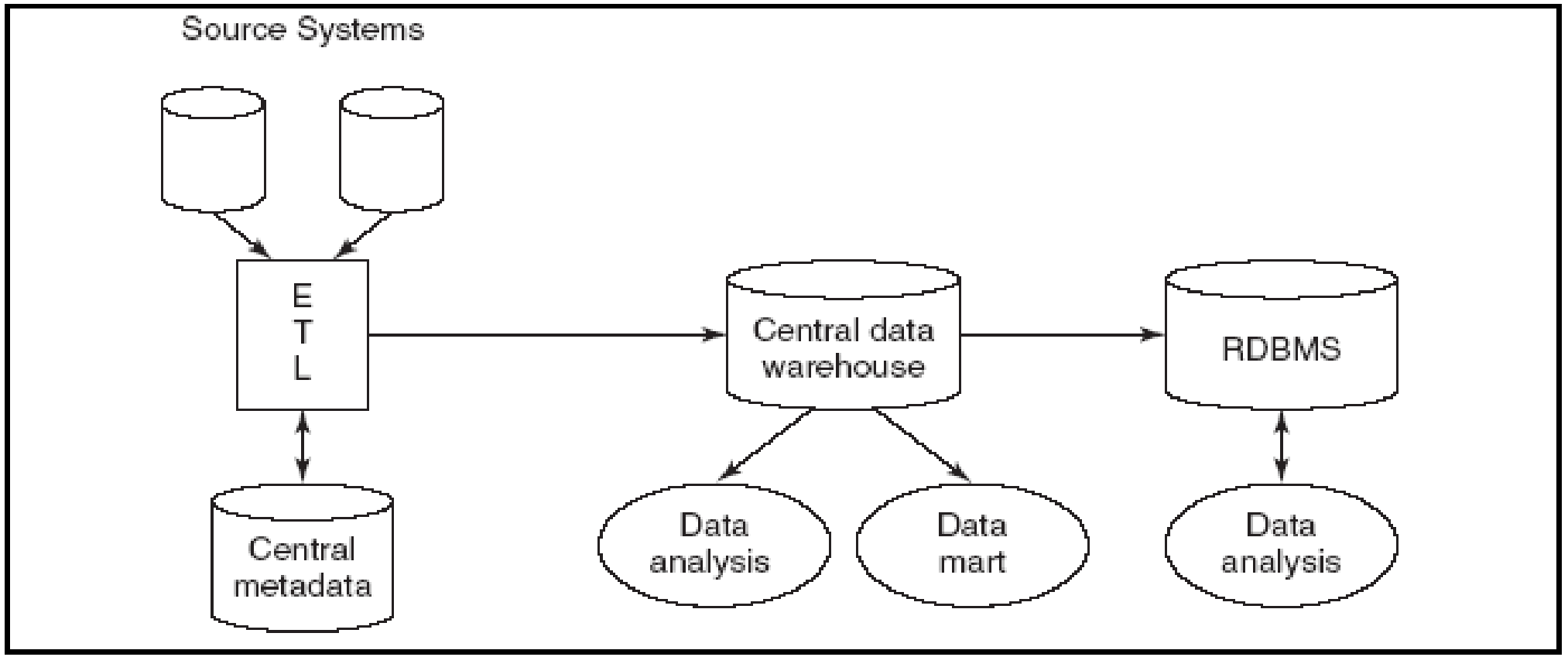

PRODUCT
Product_Code
Description
Color
Size
Fact table provides statistics for sales broken down by product, period and store dimensions
SALES
Product_Code
Period_Code
Store_Code
Units_Sold
Dollars_Sold
Dollars_Cost
STORE
Store_Code
Store_Name
City
Telephone
Manager
PERIOD
Period_Code
Year
Quarter
Month
Day

Contoh Star Schema dengan Data



ETL dalam Arsitektur Data Warehouse

FIGURE 5.5 Alternative Data Warehouse Architectures



5.5a Enterprise Data Warehousing Architecture

ETL dalam Arsitektur Data Warehouse

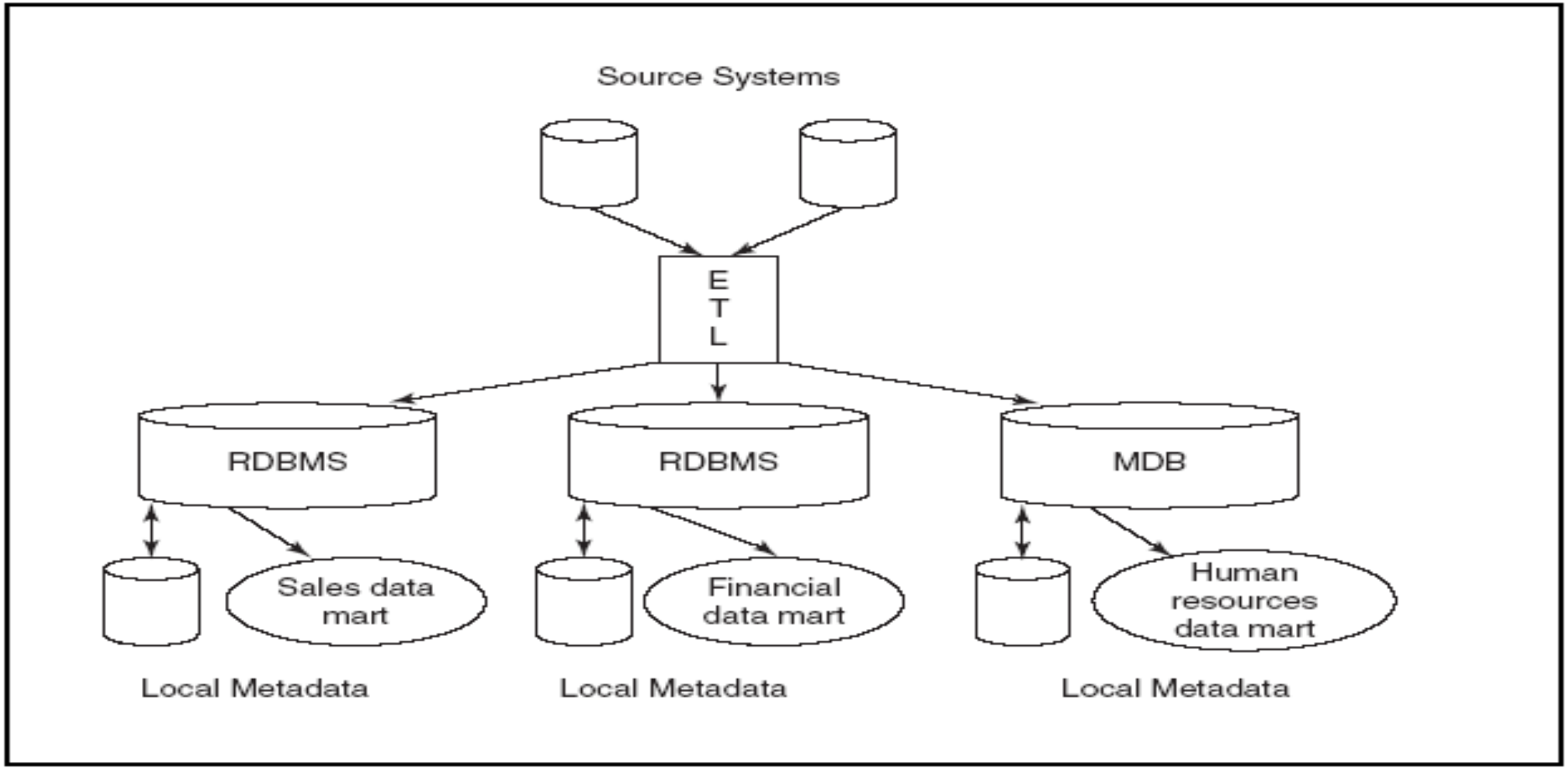


FIGURE 5.5b Data Mart Architecture

ETL dalam Arsitektur Data Warehouse

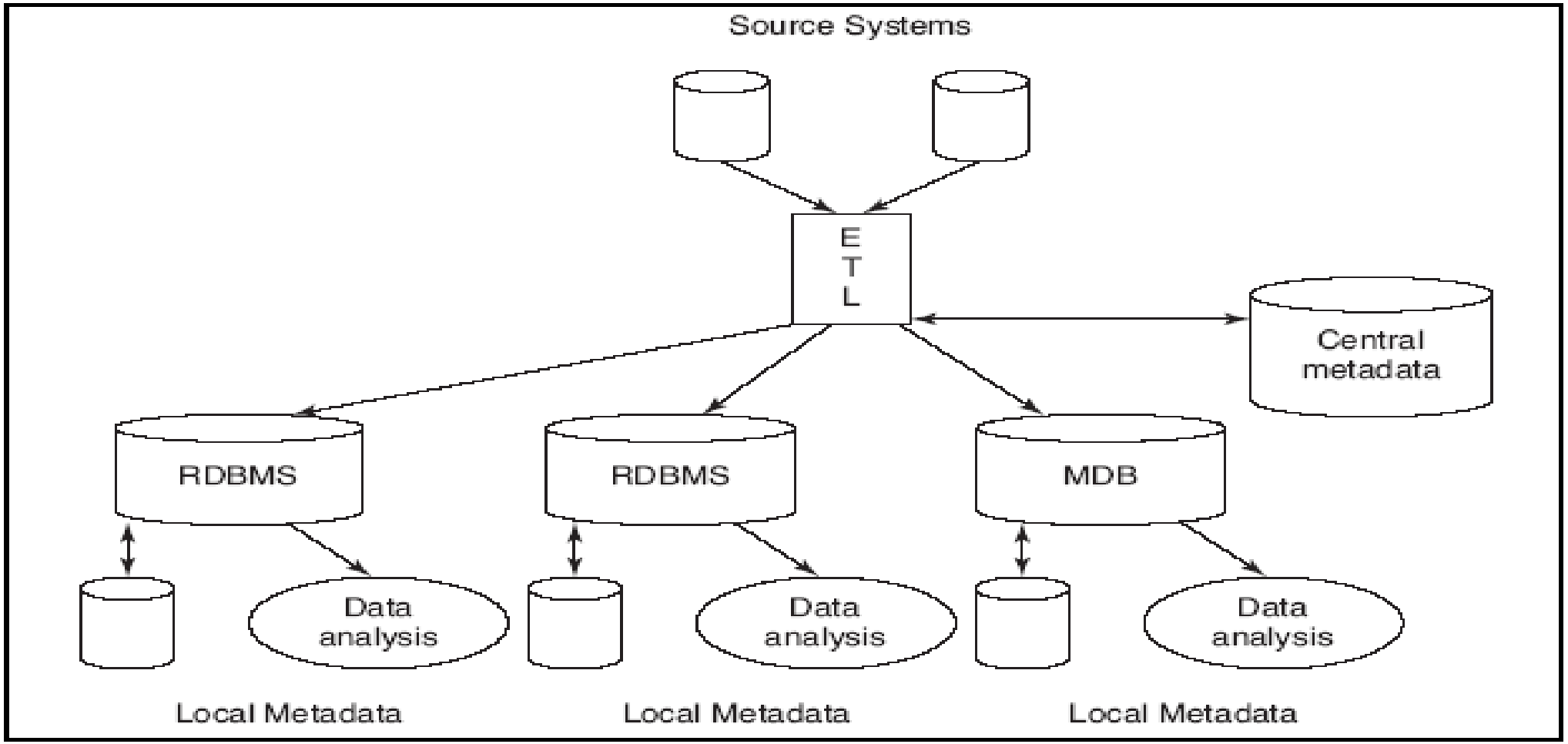


FIGURE 5.5c Hub-and-Spoke Data Mart Architecture

ETL dalam Arsitektur Data Warehouse

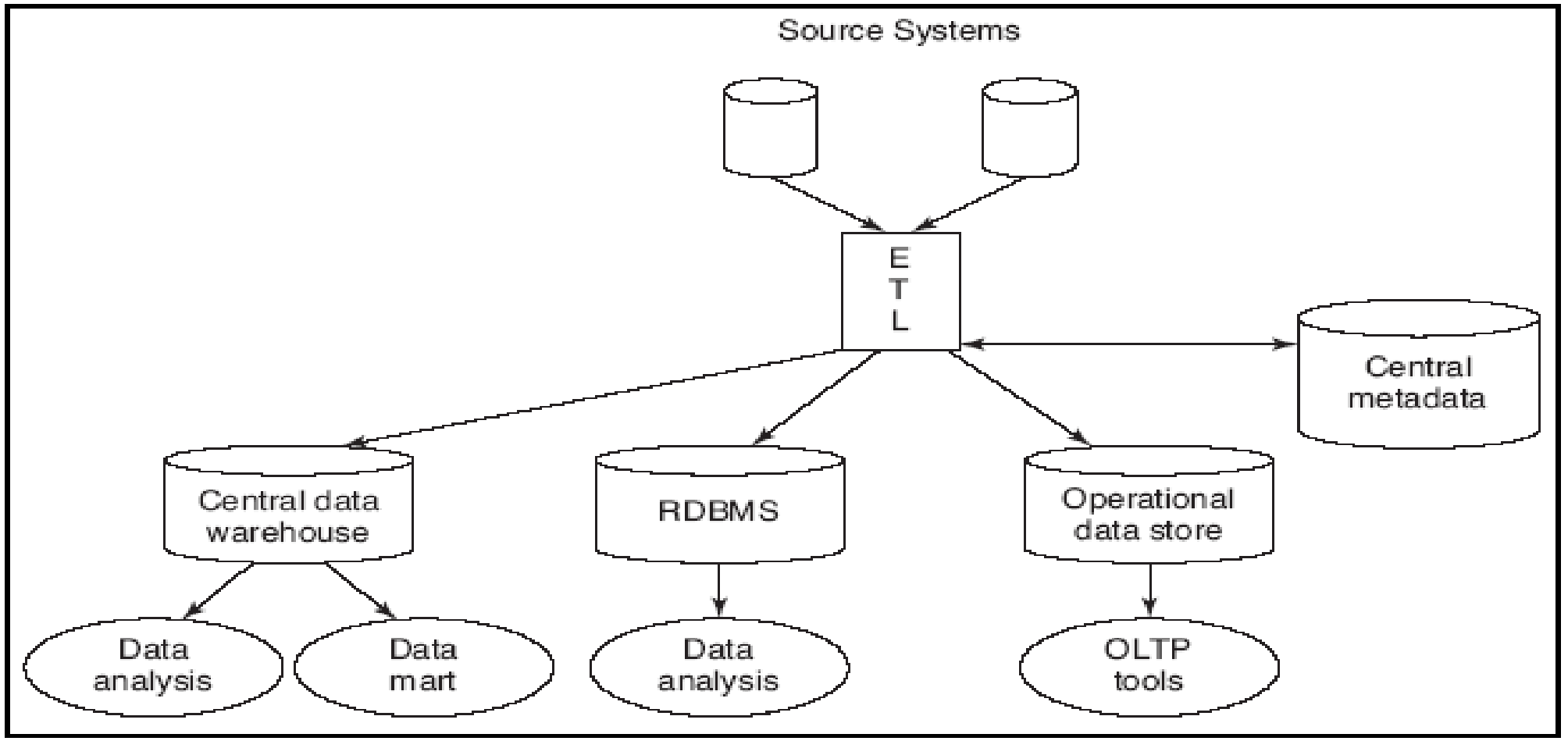


FIGURE 5.5d Enterprise Warehouse and Operational Data Store

ETL dalam Arsitektur Data Warehouse

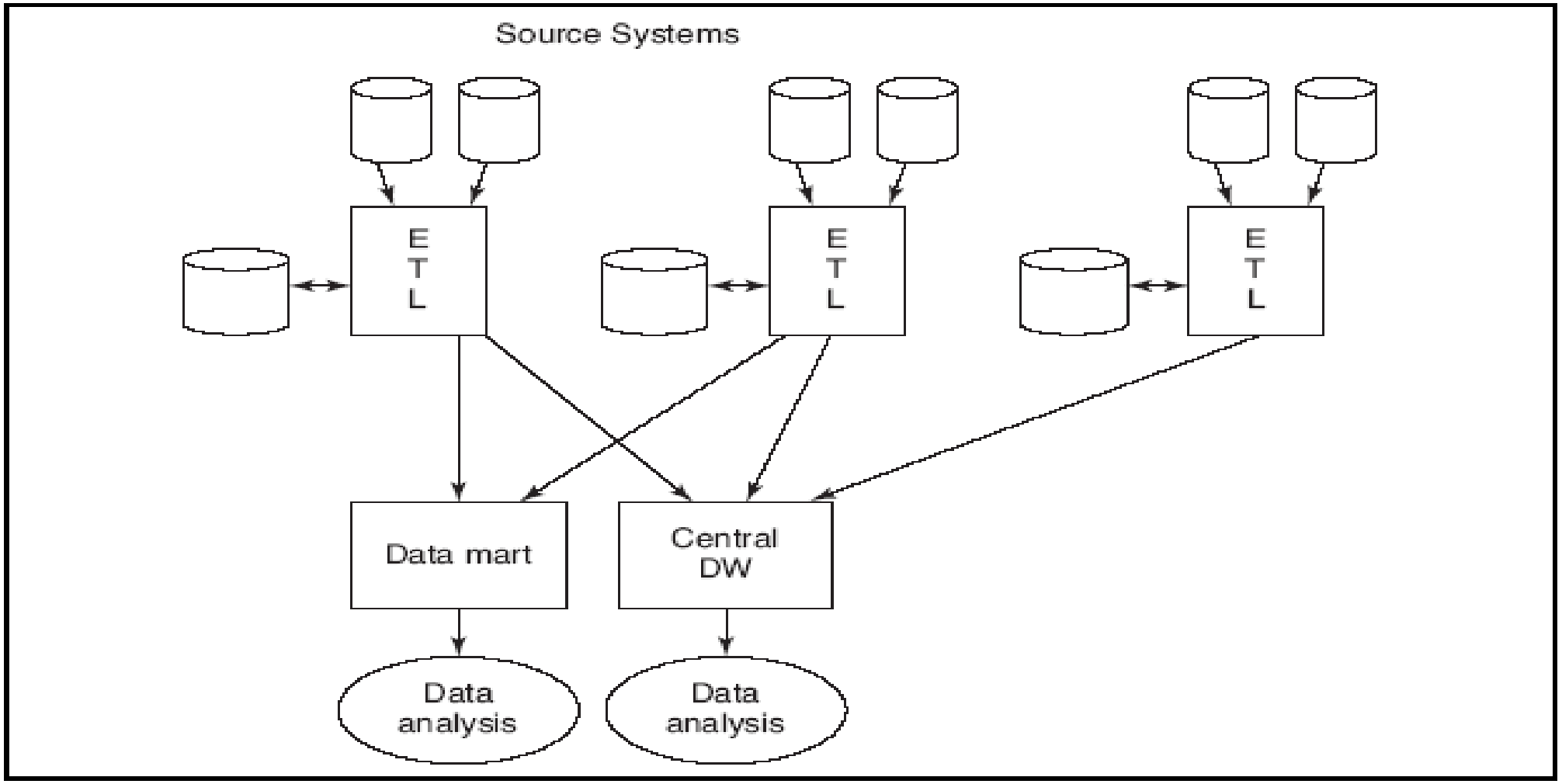


FIGURE 5.5e Distributed Data Warehouse Architecture

Quiz

1. Apa yang dimaksud dengan ETL ?
2. Jelaskan dan berikan lima contoh proses cleansing data pada proses transformation !!
3. Apa yang dimaksud skema bintang !