Proses ETL

Apa perbedaan ETL dan ELT?
Pada Extract ada 2 methode yang digunakan untuk mengambil data source, file dan stream. Jelaskan

ETL (Extraction, Transformation, Loading). ETL adalah kumpulan proses menyiapkan data dari operational source untuk data. Proses ini terdiri dari extracting, transforming, loading, dan beberapa proses yang dilakukan sebelum dipublikasikan ke dalam data warehouse. Jadi, ETL atau extract, transform, loading adalah fase pemrosesan data dari sumber data masuk ke dalam data warehouse. Tujuan ETL adalah mengumpulkan, menyaring, mengolah dan menggabungkan datadata yang relevan dari berbagai sumber untuk disimpan ke dalam data warehouse. ETL juga dapat digunakan untuk mengintegrasikan data.

ELT (Extraction, Loading and Transformation). ELT merupakan kebalikan dari ETL dimana dalam ELT dilakukan proses pengekstrakan data dari sumber data, setelah itu data diload atau dimasukkan ke dalam data warehouse, kemudian data ditransformasikan (data dipilah berdasarkan pengelompokan data).

Pada extract data ada 2 metode untuk mengambil data source yaitu :
File – extracted output from a source system. Useful with 3rd parties / legacy systems.
Stream –initiated data flows out of a system: Middleware query, web service.

File tersebut berguna karena menyediakan titik restart tanpa melakukan query ulang dari sumbernya.

Perbedaan ETL dan ELT

ETL (*Extraction*, *Transformation*, *Loading*)
Proses ETL (*Extraction*, *Transformation*, *Loading*) merupakan proses yang harus dilalui dalam pembentukan *data warehouse*. ETL adalah suatu proses mengambil dan mengirim data dari data sumber ke *data warehouse*. Dalam proses pengambilan data, data harus bersih agar didapat kualitas data yang baik. Contohnya ada nomor telepon yang invalid, ada kode buku yang tidak eksis lagi, ada beberapa data yang *null*, dan lain sebagainya. Pendekatan tradisional pada proses ETL mengambil data dari data sumber, meletakan pada *staging area*, dan kemudian mentransform dan meng-*load* ke *data warehouse*

ELT (*Extraction*, *Loading*, *Transformation*)
ELT merupakan variasi dari ETL (*Extraction*, *Transformation*, *Loading*). ELT memungkinkan data mentah dimuat secara langsung pada *data warehouse* yang kemudian akan transformasi pada *aata warehouse* tersebut. Kemampuan ini sangat berguna untuk memproses set data yang besar yang diperlukan untuk *Business Intelligence* dan analisis data yang besar. Salah satu kemampuan utama ELT adalah pengurangan waktu*loading* jika dibandingkan dengan model ETL

Perbedaan ETL dan ELT

- **Waktu - Beban**
  ETL: Menggunakan area staging dan sistem, waktu tambahan untuk load data
  ELT: Semua dalam satu sistem, hanya muat satu kali

- **Waktu - Transformasi**
  ETL: Perlu menunggu, terutama untuk ukuran data yang besar - seiring pertumbuhan data, waktu transformasi meningkat
  ELT: Semua dalam satu sistem, kecepatan tidak tergantung pada ukuran data

- **Waktu - Pemeliharaan**
  ETL: Pemeliharaan yang tinggi - pilihan data untuk load dan transform dan harus melakukannya lagi jika terhapus atau ingin meningkatkan repositori data utama
  ELT: Pemeliharaan rendah - semua data selalu tersedia

- **Kompleksitas Implementasi**
  ETL: Pada tahap awal, membutuhkan sedikit ruang dan hasilnya bersih
  ELT: Membutuhkan pengetahuan mendalam tentang alat dan desain ahli dari gudang repositori utama

- **Style Analisis dan Pengolahan**
  ETL: Berdasarkan beberapa skrip untuk membuat tampilan - menghapus tampilan berarti menghapus data
  ELT: Menciptakan tampilan adhoc - biaya rendah untuk pembangunan dan pemeliharaan

- **Batasan Data atau Pembatasan dalam Supply**
  ETL: Dengan asumsi dan memilih data yang diprioritaskan
  ELT: Dengan kebijakan hardware (tidak ada batasan data)

- **Dukungan Data Warehouse**
  ETL: Model warisan yang lazim digunakan untuk data lokal dan relasional, terstruktur
  ELT: Disesuaikan untuk menggunakan infrastruktur cloud untuk mendukung big data terstruktur dan tidak terstruktur
  Metode pengambilan data source file dan stream.

  Metode file; Merupakan metode yang biasa dilakukan dan didukung oleh kebanyakan tool ETL, prinsipnya adalah mengambil sumber data tidak secara real-time, baik dari file-file database oltp, spreadsheet, file txt dan sebagainya.

  Sedangkan metode stream lebih kearah Real-time Analisis karena langsung mengakses database oltp yang sedang berjalan dengan mengakses raw data dan mengekstrak record yang sudah terstruktur. sehingga memungkinkan Analisis real-time data warehouse dengan sumber data yang up to date.

  Contoh Tool ETL yang mendukung Stream:

- Apache Spark
- Apache Storm

  Metodenya adalah mendeteksi baris/row pada tabel-tabel yang bertambah atau berubah dan memasukkannya melalui ETL ke data warehouse.

  2 Metode Extract

ETL merupakan sebuah konsep yang dibutuhkan pada Data Warehouse yang berfungsi mentransformasikan data pada basis data OLTP ke dalam Data Warehouse.

ETL didalam pelaksanaannya membutuhkan proses ETL, Pilihan penerapan algoritma ETL yang akan digunakan dapat disesuaikan dengan desain model Data Warehouse yang digunakan, teknologi manajemen basis data dan aplikasi perangkat lunak yang digunakan yang akan lebih baik jika sama dengan aplikasi perangkat lunak untuk menampilkan laporan dan query dari Data Warehouse dan teknologi manajemen basis data yang digunakan untuk menyimpan data pada Data Warehouse.

Ekstraksi (Extraction) adalah operasi ekstraksi data dari sebuah sistem sumber untuk digunakan lebih jauh dalam lingkungan Data Warehouse. Tahapan ini adalah yang paling pertama dalam proses ETL. Setelah Ekstraksi, data ini akan ditransformasikan dan di-load ke dalam Data Warehouse.

Ada dua bentuk **Metode Ekstraksi logical** :

1. Ekstraksi Full(Full Extraction)
Data diekstrak secara lengkap dari sistem sumber (OLTP). Ekstraksi ini melibatkan seluruh data yang sedang tersedia dalam sistem sumber. Sebuah contoh ekstraksi penuh adalah ekspor file dari sebuah tabel yang berbeda atau kueri remote SQL yang membaca sumber data lengkap.

2. Ekstraksi Inkremental (Incremental Extraction)
Pada poin waktu tertentu, hanya data yang memiliki histori data akan diekstrak. Event ini adalah proses ekstraksi yang dilakukan paling akhir atau sebagai contoh sebuah event bisnis yang komplek seperti hari booking terakhir dari suatu periode fiskal. Informasi ini juga dapat disediakan oleh data sumber itu sendiri seperti sebuah kolom aplikasi, merefleksikan time-stamp perubahan yang paling akhir atau sebuah perubahan tabel dimana sebuah mekanisme tambahan yang sesuai menjaga track perubahan selain transaksi yang permulaan. Dalam banyak hal, menggunakan metode ini berarti menambah logika ekstraksi ke dalam sistem sumber.

Ada dua metode **ekstraksi fisik (physical extraction)**

1. Online Extraction
Data diekstrak langsung dari sistem sumber itu sendiri. Proses ekstraksi dapat berhubungan secara langsung dengan sistem sumber untuk mengakses tabel sumber atau ke sebuah sistem perantara yang menyimpan data dengan sebuah cara yang dikonfigurasikan terlebih dahulu (sebagai contoh log atau tabel perubahan). Dengan catatan bahwa sistem perantara secara fisik tidak berbeda dari sistem sumber.

2. Offline Extraction
Data tidak diekstrak secara langsung dari sistem sumber namun dibatasi secara eksplisit diluar sistem sumber orisinil. Data telah memiliki struktur atau telah dibuat melalui prosedur ekstraksi.

Beberapa **struktur** yang harus dipertimbangkan antara lain :

- Flat file
- Dump File, informasi mengenai objek yang dimasukkan atau tidak dimasukkan, bergantung pada utility yang dipilih.

- Log Archive dan Redo
- Transportable Tablespaces, cara ekstrak dan memindahkan data bervolume besar antar Database.