



Project Based Internship

ETL and ELT & System Development Management

Transforming Data for Maximum Impact: Exploring the Differences and Benefits of ETL and ELT & System Development Management

Daftar Isi

A. Introduction	4
B. What is ETL?	4
C. ETL vs ELT	5
D. BAGAIMANA ETL BEKERJA?	6
1. Extract	6
2. Transform	6
3. Load	7
F. System Development Management in Data Warehouse and ETL	9
1. Perencanaan	9
2. Analisis	10
3. Desain	10
4. Development	10
5. Pengujian	11
6. Deployment	11
7. Maintain	11
References	12

A. Introduction

Karena data terus memainkan peran yang semakin penting dalam operasi bisnis modern, kebutuhan untuk mengelola dan mengubah informasi dalam jumlah besar secara efektif menjadi sangat penting. Dua pendekatan untuk tantangan ini adalah ETL (*Extract, Transform, Load*) dan ELT (*Extract, Load, Transform*). Meskipun kedua teknik tersebut bertujuan untuk meningkatkan kualitas dan aksesibilitas data, keduanya berbeda dalam urutan dan pelaksanaannya. Pada artikel ini, kami akan mengeksplorasi perbedaan antara ETL dan ELT dan memeriksa manfaat dari masing-masing pendekatan.

B. What is ETL?

ETL, yang merupakan singkatan dari *extract, transform and load*, adalah proses integrasi data yang menggabungkan data dari berbagai sumber data menjadi satu penyimpanan data yang konsisten yang dimuat ke *data warehouse* atau sistem target lainnya. Ketika database semakin populer di tahun 1970-an, ETL diperkenalkan sebagai proses untuk mengintegrasikan dan memuat data untuk perhitungan dan analisis, akhirnya menjadi metode utama untuk memproses data untuk proyek *data warehousing*.

ETL menyediakan dasar untuk analitik data dan alur kerja pembelajaran mesin. Melalui serangkaian aturan bisnis, ETL membersihkan dan mengatur data dengan cara yang memenuhi kebutuhan *business intelligence* tertentu, seperti pelaporan bulanan, tetapi ETL juga dapat menangani analitik yang lebih canggih,

yang dapat meningkatkan proses back-end atau pengalaman pengguna akhir.

ETL sering digunakan oleh sebuah organisasi untuk:

- Ekstrak data dari sistem lama
- Bersihkan data untuk meningkatkan kualitas data dan membangun konsistensi
- Memuat data ke dalam database target

C. ETL vs ELT

Perbedaan yang paling jelas antara ETL dan ELT adalah perbedaan urutan operasinya. ELT menyalin atau mengekspor data dari lokasi sumber, tetapi alih-alih memuatnya ke area pementasan untuk transformasi, ELT memuat data mentah langsung ke penyimpanan data target untuk diubah sesuai kebutuhan.

Sementara kedua proses memanfaatkan berbagai repositori data, seperti database, *data warehouse*, dan *data lake*, setiap proses memiliki kelebihan dan kekurangan. ELT sangat berguna untuk kumpulan data bervolume tinggi dan tidak terstruktur karena pemuatan dapat terjadi langsung dari sumbernya. ELT bisa lebih ideal untuk manajemen data besar karena tidak memerlukan banyak perencanaan awal untuk ekstraksi dan penyimpanan data. Proses ETL, di sisi lain, membutuhkan lebih banyak definisi di awal. Poin data tertentu perlu diidentifikasi untuk ekstraksi bersama dengan "kunci" potensial apa pun untuk diintegrasikan di seluruh sistem sumber yang berbeda. Bahkan setelah pekerjaan itu selesai, aturan bisnis untuk transformasi data perlu dibangun. Pekerjaan ini biasanya dapat memiliki ketergantungan pada persyaratan data

untuk jenis analisis data tertentu, yang akan menentukan tingkat peringkasan yang perlu dimiliki data. Sementara ELT menjadi semakin populer dengan adopsi *database cloud*, ELT memiliki kelemahannya sendiri karena merupakan proses yang lebih baru, yang berarti bahwa praktik terbaik masih dibangun.

D. BAGAIMANA ETL BEKERJA?

Cara termudah untuk memahami cara kerja ETL adalah dengan memahami apa yang terjadi di setiap langkah proses.

1. *Extract*

Selama ekstraksi data, data mentah disalin atau diekspor dari lokasi sumber ke *staging area*. Tim manajemen data dapat mengekstraksi data dari berbagai sumber data, yang dapat terstruktur atau tidak terstruktur. Sumber-sumber tersebut termasuk tetapi tidak terbatas pada:

- SQL Server atau NoSQL
- Sistem CRM dan ERP
- *Flat files*
- E-mail
- Halaman web

2. *Transform*

Di area pementasan, data mentah mengalami pemrosesan data. Di sini, data diubah dan dikonsolidasikan untuk kasus penggunaan analitik yang dimaksud. Fase ini dapat melibatkan tugas-tugas berikut:

- Memfilter, membersihkan, menghilangkan duplikasi, memvalidasi, dan mengautentikasi data.
- Melakukan perhitungan, penerjemahan, atau peringkasan berdasarkan data mentah. Ini dapat mencakup mengubah tajuk baris dan kolom untuk konsistensi, mengonversi mata uang atau unit pengukuran lainnya, mengedit string teks, dan lainnya.
- Melakukan audit untuk memastikan kualitas dan kepatuhan data
- Menghapus, mengenkripsi, atau melindungi data yang diatur oleh regulator industri atau pemerintah
- Memformat data ke dalam tabel atau tabel gabungan agar sesuai dengan skema gudang data target.

3. Load

Pada langkah terakhir ini, data yang diubah dipindahkan dari staging area ke gudang data (*data warehouse*) target. Biasanya, ini melibatkan pemuatan awal semua data, diikuti dengan pemuatan berkala perubahan data tambahan dan, lebih jarang, penyegaran penuh untuk menghapus dan mengganti data di gudang. Bagi sebagian besar organisasi yang menggunakan ETL, prosesnya otomatis, terdefinisi dengan baik, berkelanjutan, dan didorong oleh batch. Biasanya, ETL berlangsung di luar jam kerja saat lalu lintas pada sistem sumber dan gudang data berada pada titik terendah.

E. ETL and Other Data Integration Methods

ETL dan ELT hanyalah dua metode integrasi data, dan ada pendekatan lain yang juga digunakan untuk memfasilitasi alur kerja integrasi data. Beberapa di antaranya:

- Change Data Capture (CDC) mengidentifikasi dan menangkap hanya sumber data yang telah berubah dan memindahkan data tersebut ke sistem target. CDC dapat digunakan untuk mengurangi sumber daya yang diperlukan selama langkah "ekstrak" ETL; itu juga dapat digunakan secara independen untuk memindahkan data yang telah diubah menjadi data lake atau repositori lainnya secara real time.
- Replikasi data menyalin perubahan dalam sumber data secara real time atau dalam batch ke database pusat. Replikasi data sering terdaftar sebagai metode integrasi data. Faktanya, ini paling sering digunakan untuk membuat cadangan untuk pemulihan bencana.
- Virtualisasi data menggunakan lapisan abstraksi perangkat lunak untuk membuat tampilan data yang terpadu, terintegrasi, dan dapat digunakan sepenuhnya—tanpa menyalin, mengubah, atau memuat data sumber secara fisik ke sistem target. Fungsionalitas virtualisasi data memungkinkan organisasi untuk membuat gudang data virtual, data lake, dan data mart dari sumber data yang sama untuk penyimpanan data tanpa biaya dan kerumitan membangun dan mengelola platform terpisah untuk masing-masingnya. Sementara virtualisasi data dapat digunakan bersama ETL, ini semakin dilihat sebagai alternatif untuk ETL dan metode integrasi data fisik lainnya.
- Stream Data Integration (SDI) persis seperti namanya—ia terus mengkonsumsi aliran data secara real time, mengubahnya, dan memuatnya ke sistem target untuk dianalisis. Kata kuncinya di sini adalah terus menerus.

Alih-alih mengintegrasikan snapshot data yang diambil dari sumber pada waktu tertentu, SDI mengintegrasikan data secara konstan saat tersedia. SDI memungkinkan penyimpanan data untuk mendukung analitik, pembelajaran mesin, dan aplikasi waktu nyata untuk meningkatkan pengalaman pelanggan, deteksi penipuan, dan lainnya.

F. System Development Management in Data Warehouse and ETL

Gudang data (*data warehouse*) adalah tempat penyimpanan besar dan terpusat yang menyimpan data historis dan terkini dari berbagai sumber, dan dirancang untuk mendukung aktivitas *business intelligence* (BI) seperti *reporting*, analisis, dan pengambilan keputusan. *Extract, Transform, and Load* (ETL) adalah proses mengekstraksi data dari berbagai sumber, mengubahnya menjadi format yang konsisten, dan memuatnya ke gudang data. *System Development Management* (SDM) adalah proses pengelolaan pengembangan sistem informasi, termasuk gudang data dan proses ETL.

1. Perencanaan

Tahap pertama SDM adalah perencanaan, yang melibatkan penentuan ruang lingkup dan tujuan proyek, mengidentifikasi pemangku kepentingan, dan membuat rencana proyek. Dalam konteks gudang data dan pengembangan ETL, perencanaan melibatkan pemahaman kebutuhan bisnis dan sumber data, serta menentukan model data dan arsitektur gudang data. Rencana proyek harus mencakup garis waktu, tonggak, dan kebutuhan sumber daya, dan harus dikomunikasikan kepada semua pemangku kepentingan.

2. Analisis

Tahap analisis melibatkan pengumpulan dan analisis persyaratan untuk gudang data dan proses ETL. Ini termasuk mengidentifikasi sumber data, menentukan persyaratan kualitas data, dan menentukan aturan integrasi dan transformasi data. Fase analisis harus menghasilkan dokumen persyaratan terperinci yang menguraikan ruang lingkup proyek dan hasil yang diharapkan.

3. Desain

Tahap desain melibatkan pembuatan desain terperinci dari gudang data dan proses ETL, berdasarkan persyaratan yang dikumpulkan dalam fase analisis. Ini termasuk membuat model data, merancang proses ETL, dan menentukan aturan integrasi dan transformasi data. Fase desain harus menghasilkan dokumen desain terperinci yang menguraikan arsitektur gudang data dan proses ETL.

4. Development

Tahap pengembangan melibatkan membangun gudang data dan menerapkan proses ETL, berdasarkan dokumen desain. Ini termasuk membangun skema gudang data, mengembangkan skrip ETL, dan menguji aturan integrasi dan transformasi data. Fase pengembangan harus menghasilkan gudang data yang berfungsi penuh dan proses ETL, yang memenuhi persyaratan yang diuraikan dalam fase analisis.

5. Pengujian

Tahap pengujian melibatkan pengujian gudang data dan proses ETL untuk memastikan bahwa mereka memenuhi persyaratan dan berfungsi seperti yang diharapkan. Ini termasuk menguji kualitas data, aturan integrasi dan transformasi data, dan kinerja proses ETL. Fase pengujian harus menghasilkan data warehouse dan proses ETL yang sepenuhnya teruji dan tervalidasi.

6. Deployment

Tahap penerapan melibatkan penerapan gudang data dan proses ETL ke dalam produksi. Ini termasuk mengonfigurasi lingkungan produksi, memigrasikan data dari sistem sumber ke gudang data, dan memastikan bahwa proses ETL berjalan dengan benar. Fase penerapan harus menghasilkan gudang data yang beroperasi penuh dan proses ETL, yang siap mendukung aktivitas BI.

7. Maintain

Tahap pemeliharaan melibatkan dukungan berkelanjutan dan pemeliharaan gudang data dan proses ETL. Ini termasuk memantau kinerja proses ETL, pemecahan masalah, dan membuat pembaruan dan penyempurnaan gudang data sesuai kebutuhan. Fase pemeliharaan sangat penting untuk memastikan bahwa gudang data tetap up-to-date dan terus memenuhi kebutuhan bisnis yang terus berkembang.

References

<https://www.geeksforgeeks.org/data-warehouse-development-life-cycle-model>

<https://aws.amazon.com/id/what-is/etl/>