Data Ingestion

Big Data
STF1724

Rini Nuraini Sukmana, M.T
0020087901 - 08882024236
rini.nuraini@usbypkp.ac.id

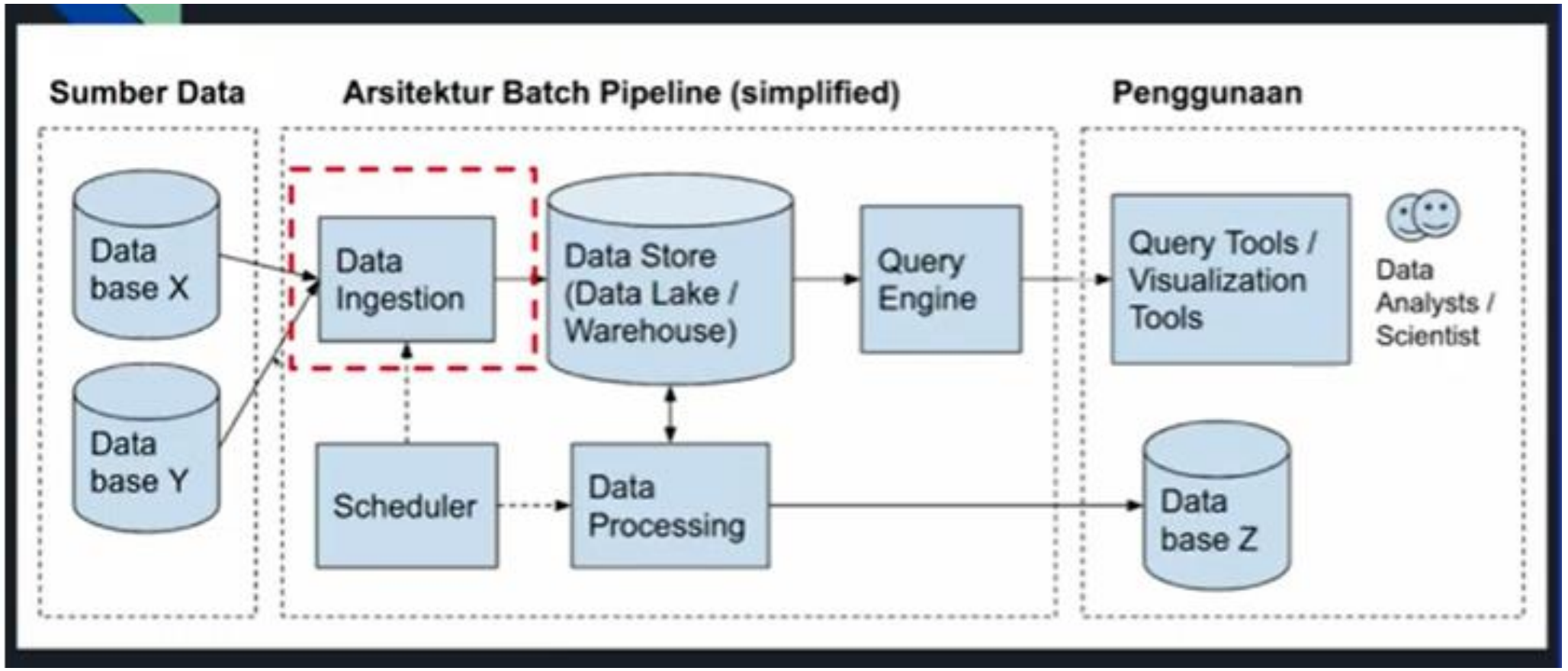
Data Ingestion - Pengantar

- Mengumpulkan data yang berasal dari berbagai sumber memerlukan banyak waktu dan usaha.
- Saat ini data menjadi andalan bagi perusahaan untuk membuat keputusan, memprediksi tren, hingga merencanakan strategi bisnis.
- Penting untuk memastikan kelancaran proses dalam memvisualisasikan dan menganalisis semua data sekaligus.
- Proses tersebut dapat dimudahkan dengan melakukan *data ingestion*

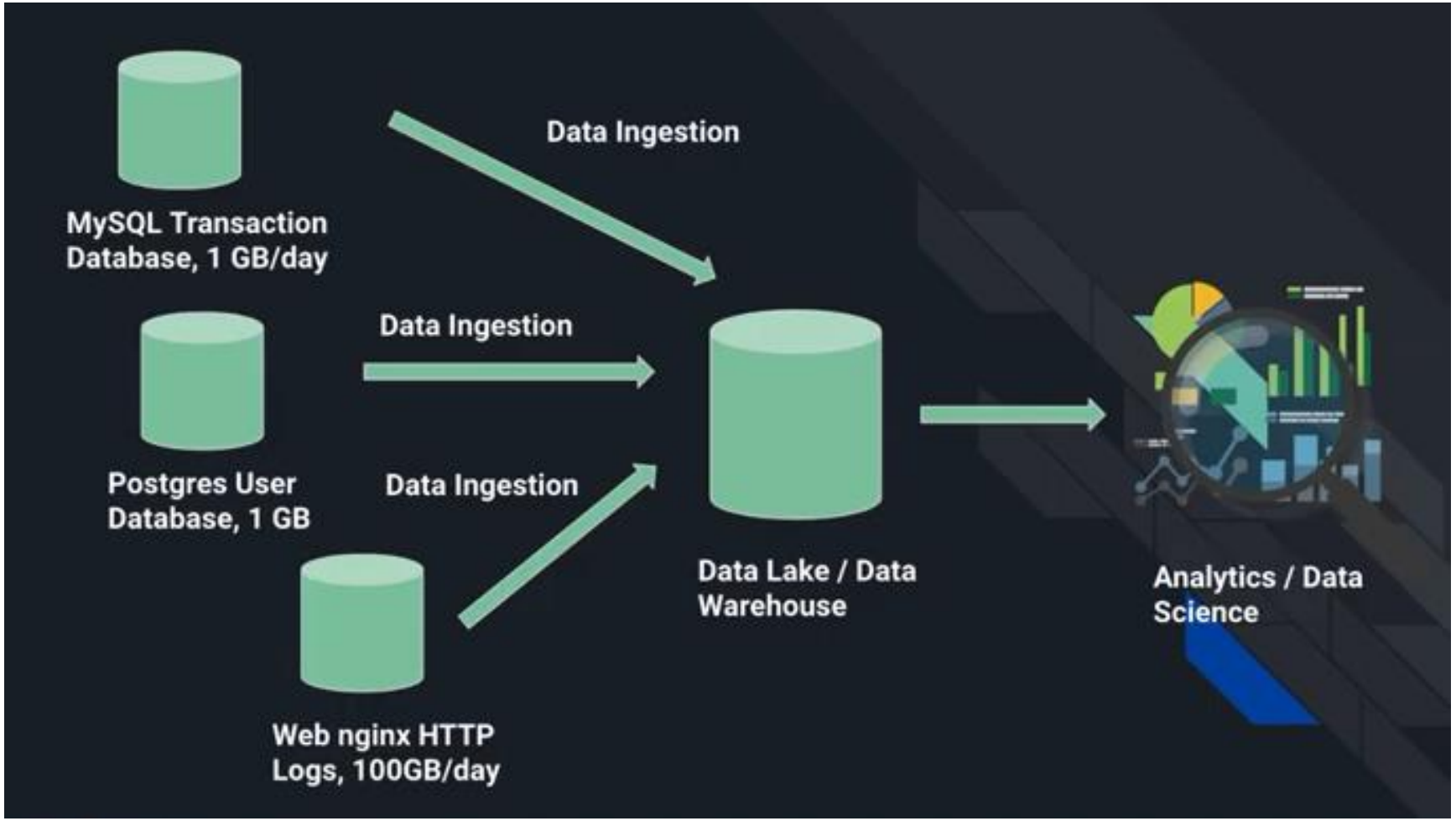
Data Ingestion

- *Data ingestion* adalah proses pemindahan data dari satu atau beberapa sumber ke suatu penyimpanan. Data tersebut nantinya akan disimpan dan dianalisis lebih lanjut, menurut Alooma.
- Ada banyak jenis format data yang dikumpulkan dari berbagai sumber data.
- Data yang tidak cocok satu sama lain bagaikan potongan *puzzle* yang akan sulit untuk dianalisis.
- Karena itu, sebelum menganalisisnya data yang dikumpulkan tersebut perlu dibersihkan dan diubah formatnya.
- Menurut TechTarget, data dalam jumlah besar dan format yang beragam akan memakan waktu dalam proses pengumpulannya.
- Jadi, biasanya perusahaan memilih menggunakan *software* atau aplikasi tertentu untuk mengotomatisasi proses *data ingestion*.

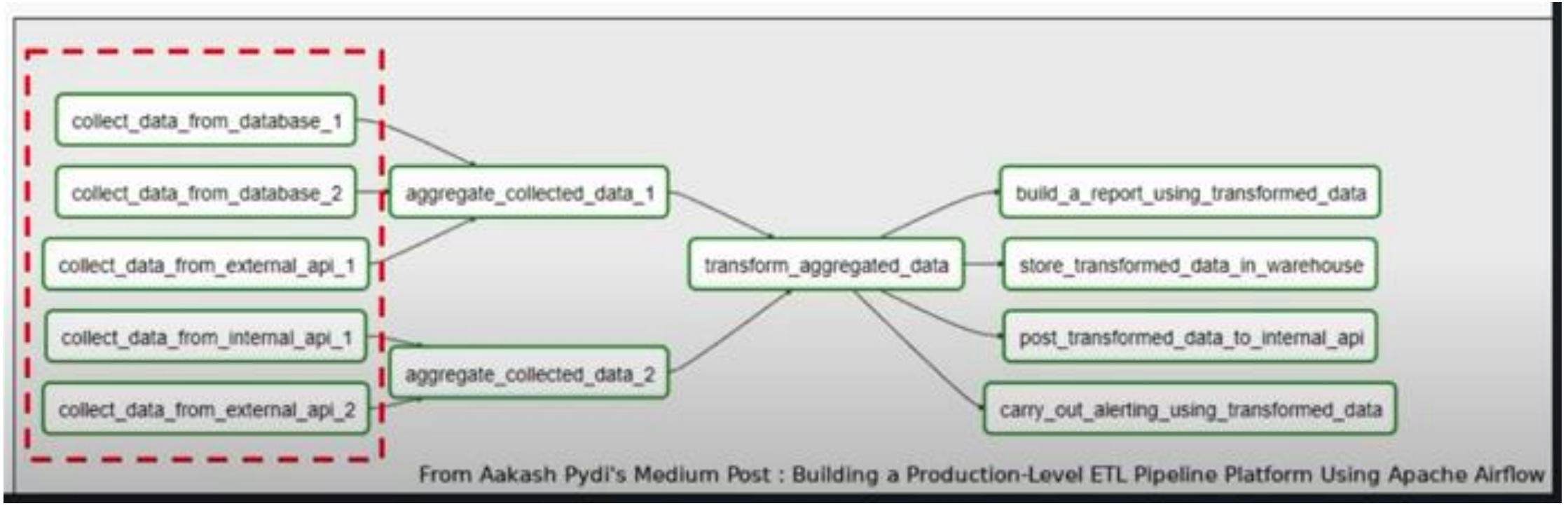
Data Ingestion



Data Ingestion



Data Ingestion



Perbedaan *data ingestion* dan *data integration*

- HubSpot menjelaskan bahwa *data integration* selangkah lebih rumit dari *data ingestion*. Hal itu disebabkan dalam proses *data ingestion* kumpulan data hanya dipindahkan ke lokasi baru.
- Namun, dalam *data integration*, kumpulan data tersebut akan dipastikan kesesuaiannya meskipun dari sumber yang berbeda.
- Dengan begitu, proses menganalisis data bisa dilakukan lebih mudah dan akurat.

Data integration

- *Data integration* adalah salah satu cara yang bisa dilakukan untuk menggabungkan data yang dimiliki oleh perusahaan dari berbagai sumber *database*.
- Saat ini data menjadi salah satu hal terpenting bagi perusahaan. Pasalnya, di era serba digital ini data adalah bagian penting yang bisa digunakan sebagai acuan untuk mengambil keputusan.
- Proses pengelolaan data cukuplah panjang hingga akhirnya bisa dijadikan acuan untuk mengambil keputusan.
- Salah satu tahapan pertama dalam proses pengelolaan data adalah *data integration* atau pengumpulan data.

Data integration

- *Data integration* adalah proses menggabungkan data dari berbagai sumber.
- *Data integration* juga menjadi hal utama yang harus dilakukan terlebih dahulu sebelum mulai proses analisis, membuat laporan, dan membuat perencanaan strategi.
- *Data integration* biasanya akan dimulai dengan proses penyerapan yang mencakup langkah-langkah seperti pembersihan data, pemetaan ETL (Extract, Transform, and Load), dan transformasi.
- Dalam melakukan proses yang satu ini biasanya melibatkan beberapa elemen seperti jaringan sumber data, server master, dan orang yang berkepentingan untuk mengakses data dari server master.
- Umumnya orang yang berkepentingan tersebut terlebih dahulu akan mengirimkan permintaan ke server master untuk memperoleh data yang lengkap.
- Kemudian, server master akan mengambil data yang dibutuhkan dari sumber internal dan eksternal.
- Barulah data akan diekstraksi dari sumber dan digabungkan menjadi kumpulan data yang kohesif.
- Hasil penggabungan data tersebut selanjutnya akan dikirimkan ke orang yang berkepentingan untuk digunakan.

Jenis Data Ingestion

Ada 3 jenis cara untuk melakukan *data ingestion*.

- ***Real-time***
- ***Batch-based***
- ***Lambda architecture-based***

Jenis Data Ingestion

Real-time

- Jenis yang pertama ini mengumpulkan dan mentransfer data dari sistem secara *real-time*.
- *Real-time data ingestion* bermanfaat untuk perusahaan yang harus bereaksi cepat terhadap informasi baru. Misalnya untuk perdagangan pasar saham atau pemantauan jaringan listrik.

Jenis Data Ingestion

Batch-based

- *Batch-based data ingestion* adalah proses mengumpulkan dan mentransfer data dalam sebuah kumpulan sesuai dengan interval yang dijadwalkan.
- Pengumpulan data bisa berdasarkan jadwal, peristiwa, atau urutan yang disesuaikan.

Jenis Data Ingestion

Lambda architecture-based

- Jenis yang satu ini merupakan kombinasi dari proses yang terdiri dari metode *real-time* dan *batch*.
- Pengaturannya sendiri terdiri dari proses pengumpulan, penyajian, dan lapisan kecepatan.
- Dua proses pertama melakukan pengindeksan data dalam sebuah kumpulan.
- Lalu, proses lapisan kecepatan akan secara instan mengindeks data yang belum terambil dari proses pengindeksan yang pertama.

Macam Data Ingestion

Database Ingestion: ambil dari database. Contoh: table transaksi diambil harian

- **Batch** Database Ingestion
  - Ambil harian, atau per jam
  - Menggunakan export table atau SQL
  - Contoh SQL menyusul
- **Stream** Database Ingestion
  - Ambil setiap ada perubahan data, tidak nunggu harian atau per jam
  - CDC (Change Data Capture)
  - Contoh: Maxwell's daemon yang listen ke binlog

Macam Data Ingestion

Event Ingestion: dari aktivitas user yang tidak di-record di database. Ditulis dalam bentuk logs atau dikirim ke messaging platform. Contoh: page view

- **Batch** Event ingestion
  - Batch copy dari logs, misal Nginx HTTP logs
- **Stream** Event ingestion a.k.a. **Tracking**
  - Aplikasi kirim ke messaging / pubsub, data engineers subscribe
  - Contoh: setiap user klik button "Add to Cart", ada event yang dikirim ke messaging seperti Kafka atau Cloud Pub/Sub

Macam Data Ingestion

Event Ingestion: dari aktivitas user yang tidak di-record di database. Ditulis dalam bentuk logs atau dikirim ke messaging platform. Contoh: page view

- **Batch** Event ingestion
  - Batch copy dari logs, misal Nginx HTTP logs
- **Stream** Event ingestion a.k.a. **Tracking**
  - Aplikasi kirim ke messaging / pubsub, data engineers subscribe
  - Contoh: setiap user klik button "Add to Cart", ada event yang dikirim ke messaging seperti Kafka atau Cloud Pub/Sub

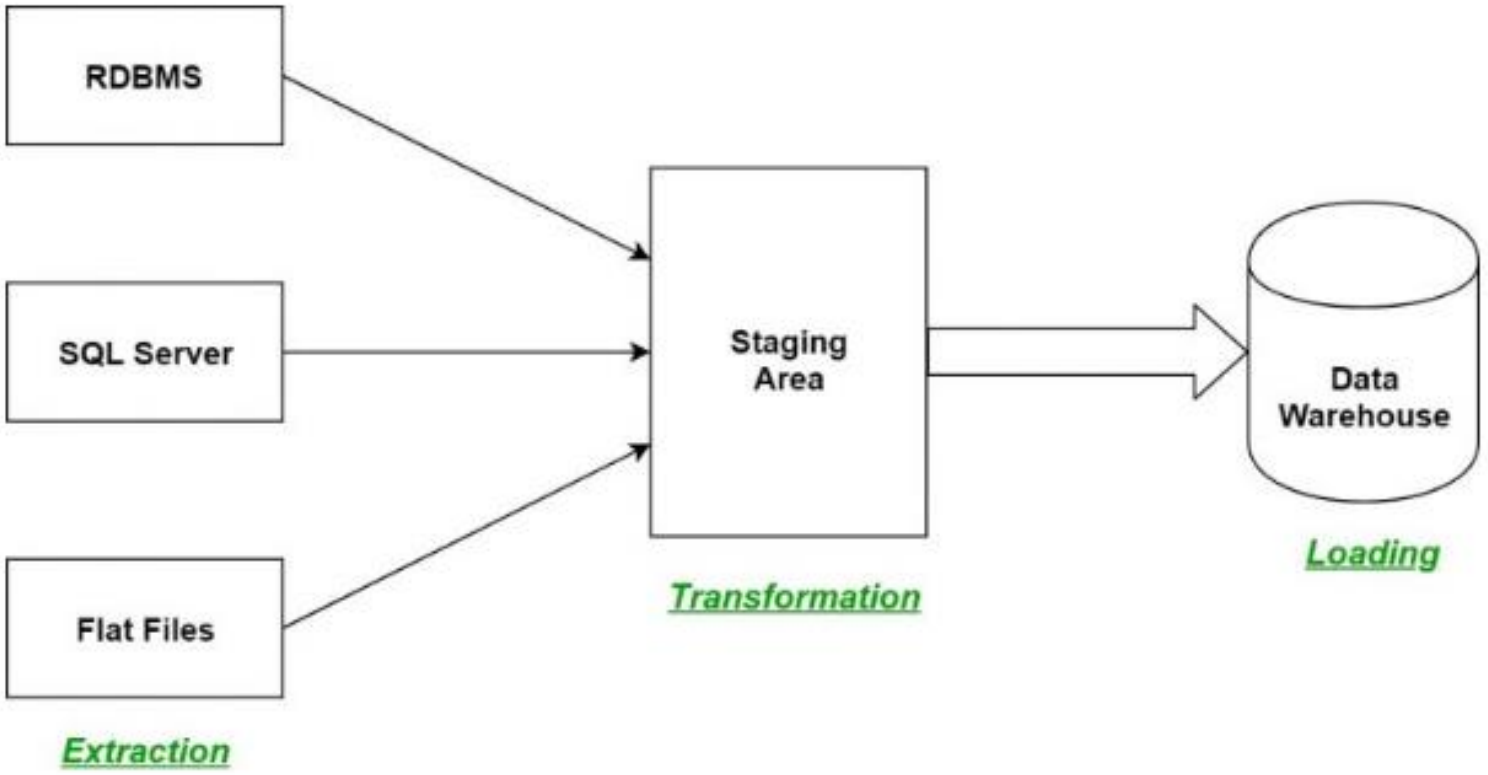
Perbedaan *data ingestion* dan ETL

- *Tools data ingestion* mungkin terlihat memiliki fungsi yang sama dengan platform ETL. Padahal, ada beberapa perbedaan yang perlu diketahui.
- Melansir Striim, *data ingestion* berkaitan dengan proses ekstraksi data dari sumber dan memuatnya ke tempat penyimpanan tujuan.
- Sementara itu, ETL tidak hanya melibatkan ekstraksi dan transfer data, tapi juga transformasi data sebelum mengirimkannya ke penyimpanan tujuan.

ETL

- ETL adalah singkatan dari *extract*, *transform*, dan *load*. ETL merupakan proses integrasi data.
- Di sana, data akan dikombinasikan dari berbagai sumber. Setelah itu, mereka disimpan di tempat bernama *data warehouse*.
- Di perusahaan, orang yang bertanggung jawab atasnya merupakan ETL *developer*.

Proses dalam ETL



Proses dalam ETL

1. *Extraction*

- Langkah pertama bernama *extraction*. Layaknya namanya, dalam proses ini, kamu mengambil data dari berbagai sumber
- Nah, setelah diambil, kamu tak serta-merta menaruhnya di *warehouse*. Tempat untuk data ini adalah *staging area*.
- Format dari data tersebut berbeda-beda. Belum lagi, ada kemungkinan informasi tersebut bersifat *corrupt*.

Proses dalam ETL

2. *Transformation*

- Tahap ETL selanjutnya adalah *transformation*. Pada langkah ini, data akan diolah sehingga punya satu format yang sama.
- Biasanya, ada 5 hal yang dilakukan pada data:
  - *filtering*, menyaring data dengan filter tertentu
  - *cleaning*, menyesuaikan format penulisan, misalnya “Amerika Serikat” diubah jadi “AS”
  - *joining*, ciri data yang serupa menjadi satu
  - *splitting*, memecah ciri data yang berbeda menjadi dua atau lebih
  - *sorting* mengurutkan data berdasarkan ciri tertentu

Proses dalam ETL

3. *Loading*

- Langkah terakhirnya bernama *loading*. Akhirnya, data yang selesai diproses masuk ke *data warehouse*.
- Kadang kala, proses ini terjadi sangat cepat. Tiap data selesai diolah, ia langsung menjalani proses *loading*.
- Akan tetapi, kita bisa mengatur alirannya menjadi beberapa saat sekali.

Manfaat Data Ingestion

- **Data telah tersedia**

Proses ini membantu perusahaan untuk mengumpulkan data yang disimpan di berbagai platform. Lalu, data tersebut dipindahkan ke penyimpanan yang lebih terpadu untuk segera dianalisis.

- **Data tidak terlalu rumit**

Manfaat *data ingestion* yang selanjutnya adalah untuk menyederhanakan data sebelum mengirimkannya ke *data warehouse*.

- **Menghemat waktu dan tenaga**

Mengambil dan memindahkan data bisa dilakukan secara otomatis. Jadi bisa menghemat tenaga karyawan perusahaan.

- **Membuat keputusan lebih baik**

Real-time data ingestion memungkinkan bisnis untuk melihat masalah dan peluang dengan cepat. Karena itu, proses pengambilan keputusan juga jauh lebih efisien.