

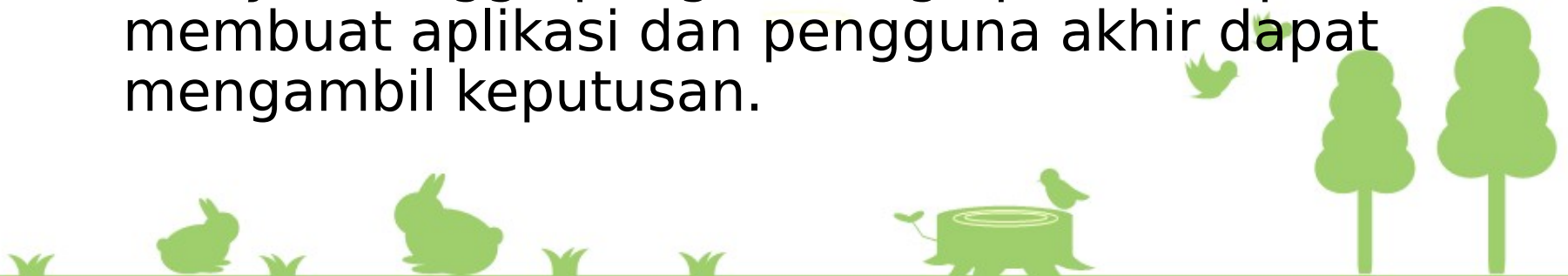
ETL (Extract-Transform-Load)

Big Data Analytics



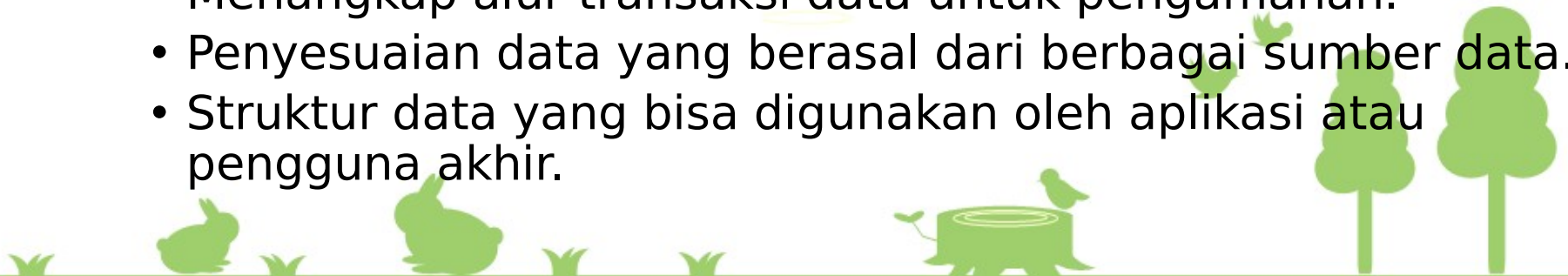
Pengenalan ETL (1)

- Sistem Extract-Transform-Load (ETL) adalah dasar dari pengolahan data khususnya Big Data.
- Sebuah sistem ETL yang baik akan mengekstrak data dari sumber, memberlakukan konsistensi standard dan kualitas data serta adaptif untuk menyesuaikan format data.
- Menyampaikan data dalam bentuk *presentation-ready* sehingga pengembang aplikasi dapat membuat aplikasi dan pengguna akhir dapat mengambil keputusan.

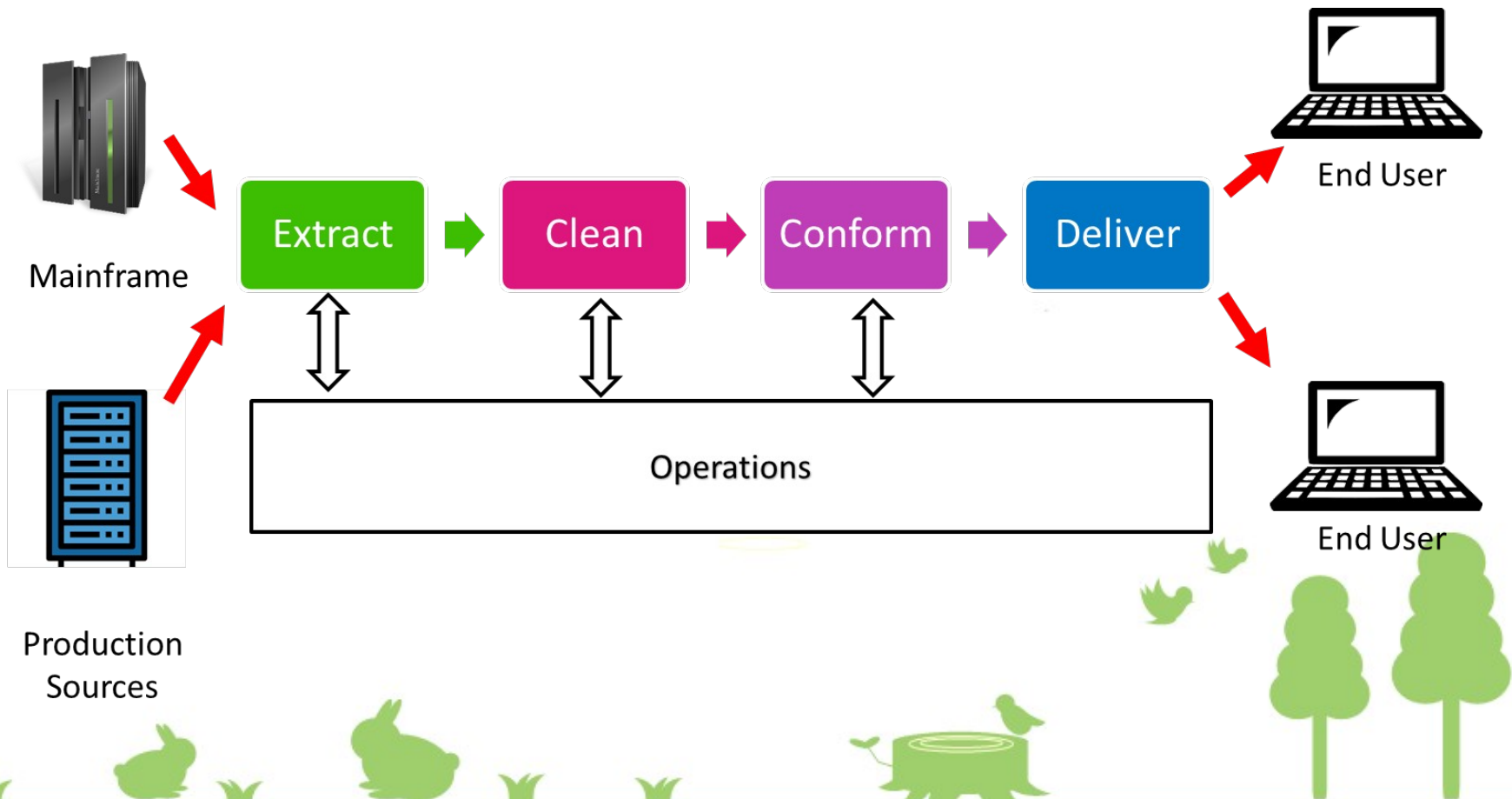


Pengenalan ETL (2)

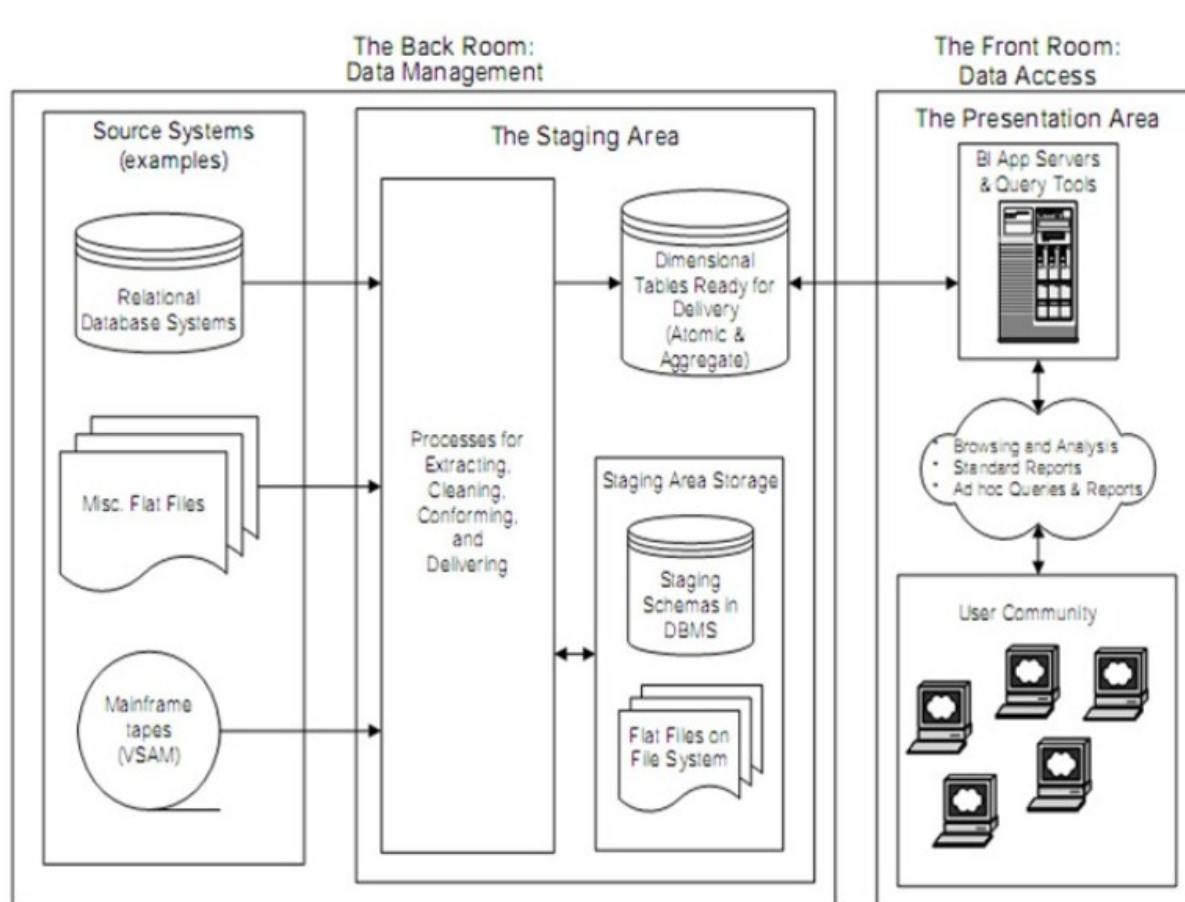
- Membangun system ETL adalah aktifitas di belakang layer yang tidak terlihat oleh pengguna akhir, menggunakan 70% sumber daya yang dibutuhkan untuk implementasi dan pemeliharaan Big Data.
- Sistem ETL terdiri dari:
 - Menghapus kesalahan dan mengkoreksi data yang hilang.
 - Menyediakan kepercayaan data yang terukur dan terdokumentasi.
 - Menangkap alur transaksi data untuk pengamanan.
 - Penyesuaian data yang berasal dari berbagai sumber data.
 - Struktur data yang bisa digunakan oleh aplikasi atau pengguna akhir.



Bagan Alir ETL (1)

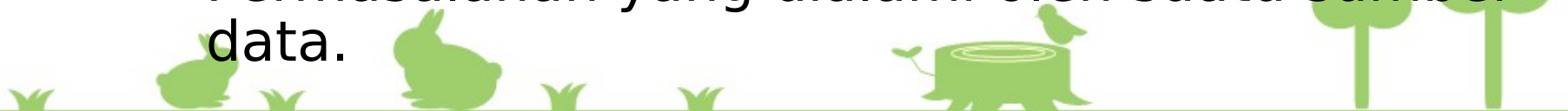


Bagan Alir ETL (2)



Kualitas Data (1)

- Penerapan system ETL adalah untuk mendapatkan data dengan kualitas yang baik.
- Kualitas data dipengaruhi oleh:
 - Heteroginitas sumber data.
 - Perbedaan Teknologi
 - Perbedaan Platform
 - Data dengan ukuran yang besar yang dihasilkan setiap hari oleh suatu sumber data.
 - Permasalahan yang dialami oleh suatu sumber data.



Kualitas Data (2)

- Permasalahan yang dialami:
 - Duplikasi data
 - Inkonsistensi data
 - Data ambigu
 - Data yang tidak lengkap
- Sehingga, dibutuhkan sistem ekstraksi dan pembersihan data untuk menghasilkan kualitas data yang baik.

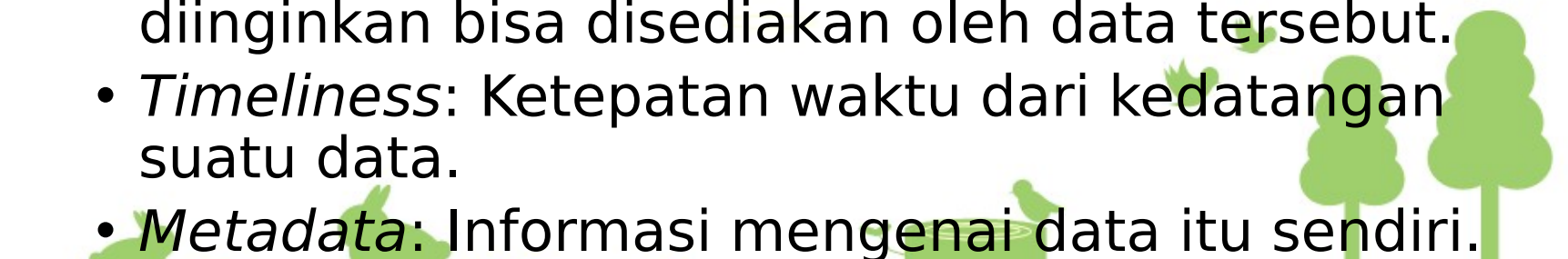


Kualitas Data (3)

- Kualitas data yang baik merupakan asset yang sangat berharga. Sementara kualitas data yang buruk dapat membahayakan kredibilitas dan akurasi hasil pengolahan data.
- Kualitas data adalah sebuah persepsi atau penilaian kelayakan data untuk melayani tujuannya dalam konteks tertentu.

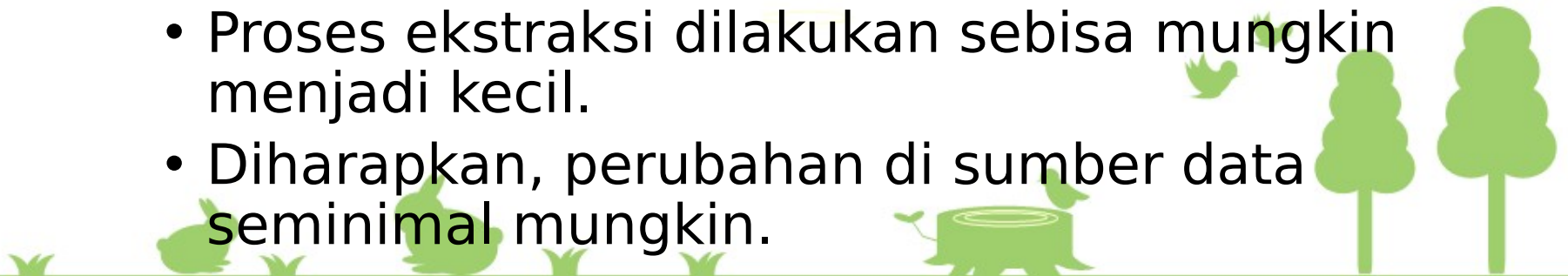


Kualitas Data (4)

- Parameter kualitas data:
 - *Correctness/Accuracy*: sejauh mana data dapat menggambarkan secara benar sebuah entitas nyata.
 - *Consistency*: Data memberikan satu versi kebenaran walau diperlakukan dalam kondisi yang berbeda.
 - *Completeness*: Sejauh mana atribut yang diinginkan bisa disediakan oleh data tersebut.
 - *Timeliness*: Ketepatan waktu dari kedatangan suatu data.
 - *Metadata*: Informasi mengenai data itu sendiri.
- 

Extraction (1)

- Seperti yang ditunjukkan pada bagan alir, tahap pertama dari system ETL adalah *Extraction* (ekstraksi)
- Prinsip-prinsip dasar pada ekstraksi data adalah:
 - Volume data yang diambil berukuran besar.
 - Proses ekstraksi dilakukan secepat mungkin.
 - Proses ekstraksi dilakukan sebisa mungkin menjadi kecil.
 - Diharapkan, perubahan di sumber data seminimal mungkin.



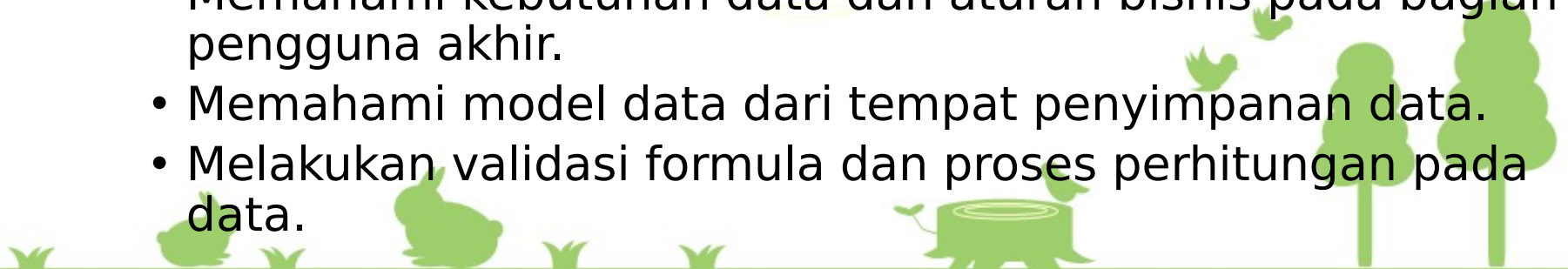
Extraction (2)

- Sebelum melakukan ekstraksi data, diperlukan sebuah peta logika data yang menggambarkan relasi antara *feature* dari sumber data dan *feature* data yang akan diolah atau ditampilkan ke pengguna akhir.
- Dokumen ini menjadi panduan yang mengikat proses ETL dari awal hingga akhir.



Extraction (3)

- Langkah-langkah pembuatan peta logika data:
 - Memiliki perencanaan yang matang – berlandaskan pada metadata.
 - Identifikasi kandidat-kandidat sumber data – identifikasi sumber-sumber data yang dibutuhkan dalam pengambilan keputusan.
 - Analisa sumber data dengan aplikasi *data-profiling* – Anomali data harus dapat dideteksi dan didokumentasi dengan baik.
 - Memahami kebutuhan data dan aturan bisnis pada bagian pengguna akhir.
 - Memahami model data dari tempat penyimpanan data.
 - Melakukan validasi formula dan proses perhitungan pada data.



Extraction (4)

- Komponen-komponen pada peta logika data:
 - Sumber data
 - Parameter-parameter sumber data
 - Parameter-parameter pada keluaran data
 - Transformasi



Transformation (1)

- *Transformation data* pada system ETL melingkupi aktifitas-aktifitas berikut:
 - *Formatting* dan standarisasi.
 - Mengubah ke angka atau teks tertentu atau format tanggal.
 - Terjemahkan data ke bentuk yang lain.
 - Agregasi atau merangkum data pada level yang lebih tinggi.



Transformation (2)

- Transformasi data juga melibatkan prinsip-prinsip:
 - *Leakage* (kebocoran) terjadi ketika proses ETL mengunduh data secara lengkap dari sumber data, namun pada kenyataannya terdapat beberapa *record* yang hilang.
 - *Recoverability* (pemulihan) berarti bahwa proses ETL harus *robust* sehingga jika terjadi kegagalan, ini bisa segera dipulihkan tanpa kehilangan atau merusak data.



Metode *Logical Extraction* (1)

- Terdapat dua metode logical extraction yaitu:
 - Full extraction
 - Incremental extraction



Metode *Logical Extraction* (2)

- *Full Extraction*:
 - Pengambilan keseluruhan data dari sumber data.
 - Ekstraksi ini mereplikasi semua data pada sumber data, sehingga tidak diperlukan proses untuk melacak perubahan pada sumber data sejak ekstraksi sukses terakhir.



Metode *Logical Extraction* (3)

- *Incremental Extraction*:
 - Pada titik tertentu dalam waktu, hanya data yang telah berubah sejak terdefinisi dengan baik yang akan diekstraksi.
 - Dalam kebanyakan kasus, menggunakan metode tertentu untuk menambahkan logika ekstraksi baru ke sumber data.



Load (1)

- Langkah terakhir dari system ETL adalah *Load* yang berupa:
 - Penyimpanan data ke *data warehouse*
 - Menampilkan data ke aplikasi atau pengguna akhir.
- Arsitektur ETL secara keseluruhan hingga tahap load adalah sebagai berikut:

