Part 1 – Fundamental DE

1. Peran utama Data Engineer dan perbedaan antara Data Scientist dan Data Analyst
   1) Apa peran utama seorang Data Engineer dalam ekosistem data?

I. 

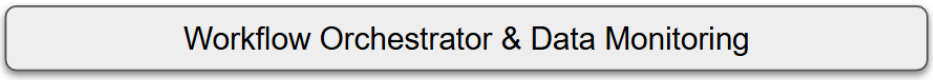
Peran utama seorang Data Engineering adalah mengambil data dari sumber yang masih berupa raw data, yang nantinya akan disimpan dalam sebuah database.

II. 

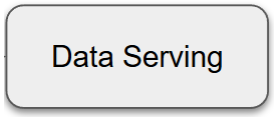
Kemudian dari database tersebut akan dilakukan proses yang bernama ingestion ke database lainnya.

III. 

Lalu dari database tempat proses ingestion tadi akan dilakukan proses lanjutan yaitu transform, dimana proses transform ini adalah normalisasi data, pembersihan data, penggabungan, dan proses – proses lainnya kemudian data tersebut di masukkan kembali ke database.

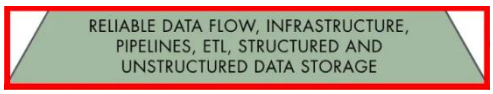
IV. 

Dari proses tersebut juga diantaranya terdapat proses workflow orchestrator dan data monitoring, yang memiliki tujuan untuk memudahkan seorang Data Engineer sehingga ketika proses ini tidak perlu dilakukan berulang kali secara manual, kita bisa menggunakan tools untuk schedule proses ini sehingga kita bisa memantau serta lanjut ke logic logic yang perlu diupdate dan lain sebagainya.

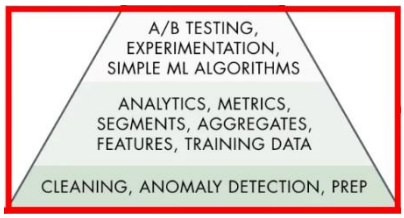
V. 

Kemudian setelah data tersebut sudah stabil atau bisa dikatakan siap, barulah data tersebut disajikan kepada pihak lain seperti Data Analyst, Data Scientist, atau Stakeholder sesuai keperluan mereka masing masing. Berikut ini merupakan gambaran dari penjelasan diatas :

2) Bagaimana peran ini berbeda dari Data Scientist dan Data Analyst?
Perbedaan Data Engineer dengan Data Scientist dan Data Analyst terletak di bagian peran dalam pengerjaan sebuah data, berikut merupakan hierarchy dari masing masing job roles tersebut:

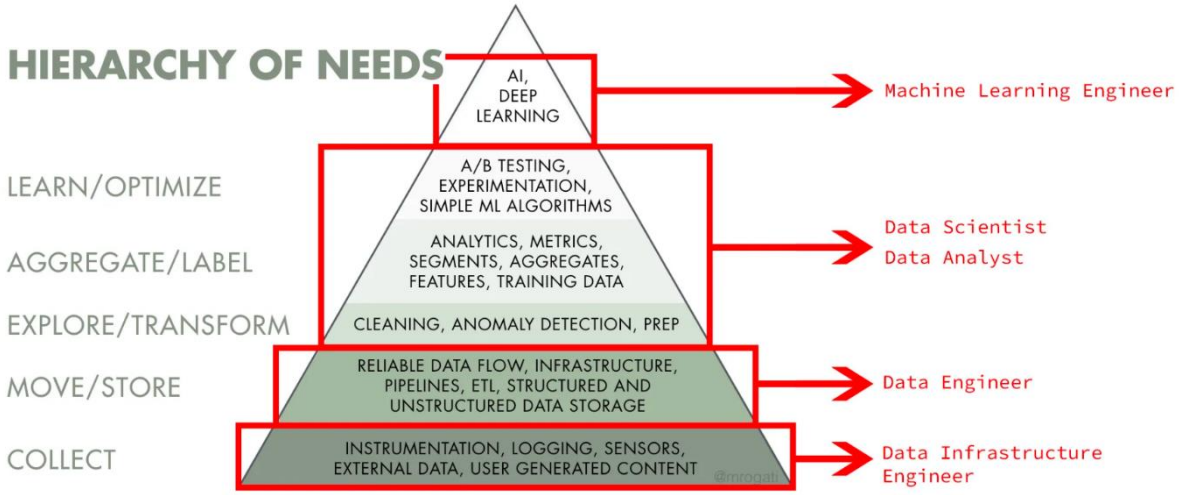
I. 

Data Engineer berurusan di bagian Reliable Data Flow, Infrastructure, Pipelines, ETL, dan Structured and Unstructured Data Storage.

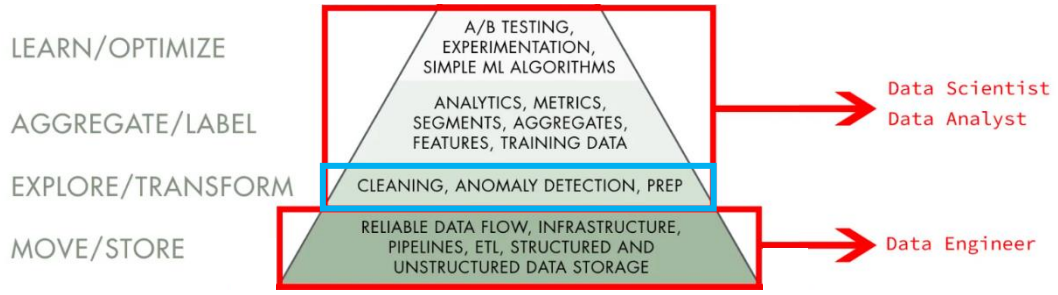
II. 

Sedangkan Data Scientist dan Data Analyst berurusan di bagian A / B Testing, Experimentation, Simple ML Algorithms, Analytics, Metrics, Segments, Aggregates, Features, dan Training Data.

Berikut ini merupakan Gambaran secara utuh dari hierarchy of needs tersebut :



2. Berikan beberapa contoh peran dari seorang Data Engineer yang mungkin bersinggungan atau bahkan sama dengan peran Data Scientist dan Data Analyst!



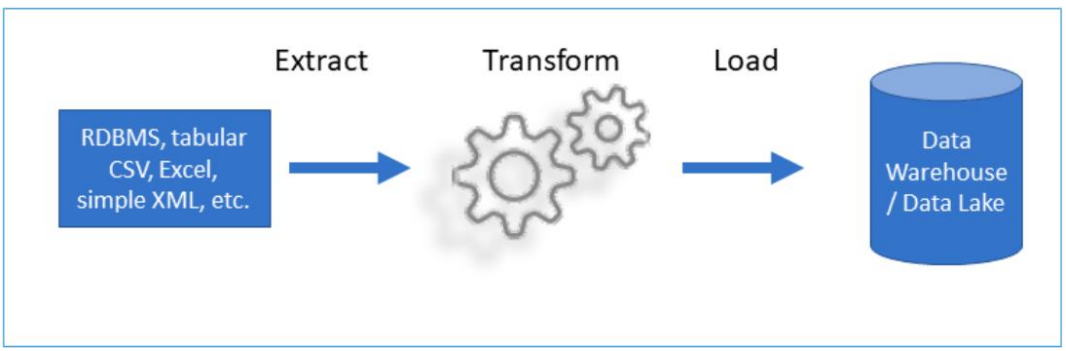
Menurut hierarchy of needs, peran Data Engineer yang mungkin bersinggungan atau bahkan sama dengan peran Data Scientist dan Data Analyst adalah pada bagian "Explore / Transform" yang diberi kotak berwarna biru, dimana di dalamnya berisi proses Cleaning, Anomaly Detection, dan Prep.

3. Jelaskan langkah-langkah proses ETL dan ELT yang berperan dalam pekerjaan seorang Data Engineer!
   1) Proses ETL (Extract, Transform, Load)
      I. Extract
      Mengambil data dari berbagai sumber, seperti RDBMS, CSV, Excel, Simple XML, dan lain sebagainya yang tujuannya untuk mengumpulkan data dari berbagai sumber dan menyimpannya dalam sebuah database. Pengambilan data ini bisa berupa batch atau stream.

      II. Transform
      Raw data sebelumnya yang sudah di ekstraksi, disimpan ke sebuah lokasi temporal yang bisa disebut sebagai staging area. Pada staging area inilah raw data tadi dilakukan proses transform yaitu meliputi pembersihan data, normalisasi data, pemetaan data dan lain sebagainya yang bertujuan untuk merubah data menjadi lebih stabil atau tidak duplikat sesuai yang diperlukan untuk sistem penyimpanan terakhir misalnya ke data warehouse.

      III. Load
      Memuat data yang telah di transformasi ke dalam sistem penyimpanan terakhir seperti data warehouse yang bertujuan untuk menyimpan data yang telah stabil untuk dilakukan analisis, pelaporan, atau pengambilan keputusan.

   Berikut adalah gambaran dari penjelasan diatas :



2) Proses ELT (Extract, Load, Transform)

I. Extract

Mengambil data dari berbagai sumber, seperti RDBMS, CSV, Excel, Simple XML, dan lain sebagainya yang tujuannya untuk mengumpulkan data dari berbagai sumber dan menyimpannya dalam sebuah database. Pengambilan data ini bisa berupa batch atau stream.

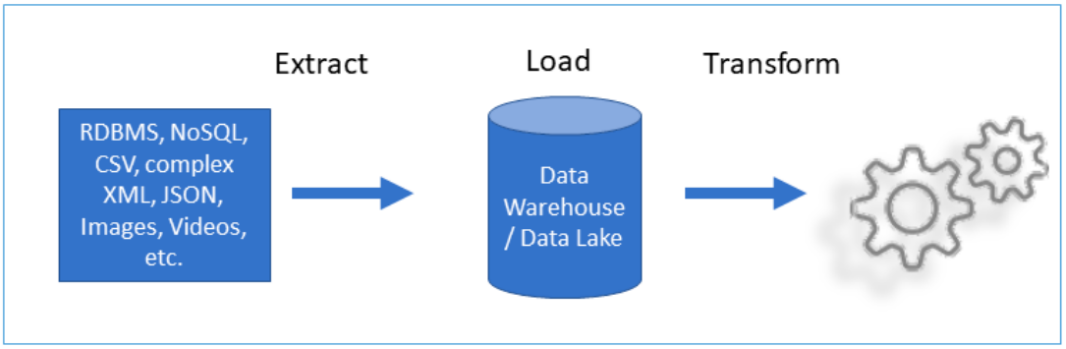
II. Load

Memuat data yang telah di ekstraksi ke dalam sistem penyimpanan terakhir seperti data warehouse yang bertujuan untuk menyimpan data dalam bentuk raw data, sehingga memungkinkan akses lebih cepat dan langsung.

III. Transform

Raw data sebelumnya yang sudah dimuat ke sistem penyimpanan terakhir, kemudian dilakukan proses transform yaitu meliputi pembersihan data, normalisasi data, pemetaan data dan lain sebagainya yang bertujuan untuk merubah data menjadi lebih stabil atau tidak duplikat sesuai yang diperlukan untuk kebutuhan analisis, pelaporan, atau pengambilan keputusan.

Berikut adalah gambaran dari penjelasan diatas :



Untuk proses ETL (Extract, Transform, Load), digunakan ketika data harus diproses dan dibersihkan sebelum dimuat ke dalam sistem penyimpanan. Dalam segi kebutuhan transformasi, ETL (Extract, Transform, Load) lebih cocok karena memungkinkan transformasi dilakukan di luar sistem penyimpanan akhir.

Untuk proses ELT (Extract, Load, Transform), digunakan ketika sistem penyimpanan akhir (seperti data lake atau data warehouse) memiliki kapabilitas pemrosesan yang kuat dan dapat menangani transformasi data secara efisien. Dalam segi volume dan kecepatan data, ELT (Extract, Load, Transform) lebih sesuai untuk volume data besar dan pengolahan real-time jika sistem penyimpanan memiliki kapabilitas pemrosesan yang memadai.