1. Data Engineer memainkan peran penting dalam ekosistem data, dengan fokus pada perancangan, pembangunan, dan pemeliharaan sistem data berskala besar yang menyimpan, memproses, dan mengelola data. Tanggung jawab utama meliputi yaitu arsiktektr data, penyimpanan data, kualitas data, dan keamanan data.

Sedangkan analis data yaitu bertanggung jawab untuk menganalisis dan menafsirkan data, biasanya bekerja pada kumpulan data berskala lebih kecil dan focus pada pelaporan serta visualisasi.

Dan peran dari data scientist ialah yang mengektrasi wawasan, pengetahuan dari kumpulan data besar menggunakan berbagai teknik alat serta bekerja dengan data untuk mengungkap pola, tren, kolarasi dan menggunakan informasi untuk mendorong keputusan bisnis, serta mengoptimalkan proses.

1. Contoh dari peran data engineer yang bersinggungan dengan data scientist dan data analyst ialah dengan menggunakan alat seperti database RDBMS dan NoSQL serta menggunakan database cloud dan gudang data seperti ETL(Microsoft SSIS, XPlenty, Talend, Cognos Data Manager).
2. Proses ETL dibagi ke dalam tiga tahapan: extract, transform, dan load.

a.Extract

Pada tahap ekstraksi, ETL melibatkan pengambilan data dari berbagai sumber seperti *database*, file CSV, API, atau sumber data lainnya. Data ini diambil dan kemudian disimpan ke dalam satu wadah penyimpanan tertentu. Proses ekstraksi dapat dilakukan secara manual maupun melalui proses otomasi yang memanfaatkan tool ETL.

b.Transform

Setelah data diekstraksi, tahap transformasi dilakukan dengan melakukan pemrosesan data. Pada tahap ini, data yang telah diekstraksi akan mengalami perubahan dan pembersihan agar menjadi data yang dapat dianalisis sesuai kebutuhan. Transformasi data melibatkan beberapa proses yaitu:

-Pembersihan data dan deduplikasi: penghapusan data yang tidak relevan atau berganda,

-Standarisasi data: mengubah format data ke dalam format tertentu sesuai kebutuhan.

-Pengelompokan data: menyusun data berdasarkan kategori tertentu.

c.Load

Tahap pemuatan (load) merupakan tahap terakhir dalam proses ETL. Pada tahap ini, data yang telah ditransformasi akan dimuat ke dalam *data warehouse* yang dituju. Data ini akan disimpan dalam struktur yang sesuai sehingga dapat digunakan untuk analisis lebih lanjut. Ada dua metode umum yang digunakan dalam proses pemuatan data:

-Full load yaitu proses pemuatan dilakukan pada semua data sekaligus. Proses ini memakan waktu yang lebih lama dan biasa digunakan sebagai proses awal untuk membangun data warehouse, atau sebagai bagian dari pembaruan periodik yang melibatkan seluruh dataset.

-Incremental load yaitu pemuatan hanya dilakukan pada data yang baru sejak ekstraksi terakhir dengan periode interval tertentu. Dengan menggunakan pendekatan ini, waktu dan sumber daya yang diperlukan untuk pembaruan data dapat dikurangi secara signifikan karena hanya melibatkan data yang relevan saja.