PART 1 - Fundamental DE

TASK 1

1. Apa peran utama seorang Data Engineer dalam ekosistem data? Bagaimana peran ini berbeda dari Data Scientist dan Data Analyst?
   Peran utama seorang Data Engineer dalam ekosistem data adalah merancang, membangun, dan memelihara infrastruktur data yang memungkinkan organisasi untuk mengumpulkan, menyimpan, dan mengakses data dengan efisien. Tugas-tugas utama seorang Data Engineer meliputi:
   a. **Membangun Pipeline Data:** Membuat dan mengelola alur kerja otomatis untuk mentransfer data dari berbagai sumber ke sistem penyimpanan yang terpusat.
   b. **Mengelola Database dan Penyimpanan:** Mengelola database dan sistem penyimpanan data, memastikan data terstruktur dengan baik dan dapat diakses dengan cepat.
   c. **Optimasi Kinerja:** Mengoptimalkan kinerja sistem penyimpanan data agar dapat menangani volume data yang besar dan permintaan yang tinggi.
   d. **Keamanan Data:** Menjamin keamanan dan integritas data melalui implementasi kontrol akses dan enkripsi.
   e. **Kolaborasi dengan Tim Lain:** Bekerja sama dengan Data Scientist dan Data Analyst untuk memahami kebutuhan data dan menyediakan infrastruktur yang diperlukan untuk analisis data.

   Perbedaan peran Data Engineer dengan Data Scientist dan Data Analyst adalah sebagai berikut:
   a. **Data Engineer vs. Data Scientist:**
      i. Data Engineer fokus pada infrastruktur dan alur kerja data. Mereka memastikan data dapat diakses dan digunakan oleh seluruh organisasi. Mereka lebih banyak bekerja dengan bahasa pemrograman seperti Python, Java, dan alat-alat big data seperti Hadoop, Spark, dan Kafka.
      ii. Data Scientist fokus pada analisis data dan pembuatan model prediktif. Mereka menggunakan data untuk menemukan wawasan, membuat model machine learning, dan melakukan analisis statistik. Mereka lebih banyak bekerja dengan alat-alat seperti R, Python, dan platform machine learning.
   b. **Data Engineer vs. Data Analyst:**
      i. Data Engineer fokus pada penyediaan infrastruktur data. Mereka memastikan data tersedia dan diorganisir dengan baik untuk dianalisis.
      ii. Data Analyst fokus pada interpretasi data dan pelaporan. Mereka menggunakan data yang disediakan oleh Data Engineer untuk membuat laporan, dashboard, dan analisis ad hoc guna mendukung pengambilan keputusan bisnis. Mereka sering menggunakan alat seperti SQL, Excel, dan software BI (Business Intelligence) seperti Tableau atau Power BI.

   Dengan demikian, meskipun ketiga peran ini berhubungan dengan data, masing-masing memiliki fokus dan tanggung jawab yang berbeda dalam ekosistem data.

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

a. Cleaning Data:

1) **Data Engineer**:
   - **Pipeline Data Cleaning**: Membangun dan mengelola pipeline untuk membersihkan data secara otomatis saat data diambil dari berbagai sumber.
   - **Automasi Cleaning**: Menggunakan skrip atau alat ETL untuk mengotomatisasi proses pembersihan data, seperti penghapusan duplikat dan pengisian nilai yang hilang.
2) **Data Scientist**:
   - **Advanced Cleaning**: Melakukan pembersihan lebih lanjut pada data yang disediakan oleh Data Engineer, terutama yang memerlukan pengetahuan domain atau analisis statistik mendalam.
   - **Feature Engineering**: Membuat fitur baru dari data yang telah dibersihkan untuk digunakan dalam model machine learning.
3) **Data Analyst**:
   - **Initial Cleaning**: Membersihkan data yang akan dianalisis, seperti mengoreksi kesalahan tipografi atau menghapus nilai yang tidak relevan.
   - **Manual Adjustments**: Melakukan penyesuaian manual pada dataset berdasarkan kebutuhan analisis atau permintaan khusus dari pemangku kepentingan.

b. Anomaly Detection:

1) **Data Engineer**:
   - **Infrastructure Setup**: Membangun dan mengelola infrastruktur yang memungkinkan deteksi anomali secara real-time atau batch processing.
   - **Data Pipeline for Anomaly Detection**: Menyusun pipeline data yang dapat mengidentifikasi dan memproses anomali dalam data secara otomatis.
2) **Data Scientist**:
   - **Model Development**: Mengembangkan dan melatih model deteksi anomali menggunakan berbagai teknik machine learning atau statistik.
   - **Model Integration**: Bekerja sama dengan Data Engineer untuk mengintegrasikan model deteksi anomali ke dalam pipeline data yang ada.
3) **Data Analyst**:
   - **Monitoring Anomalies**: Memantau hasil deteksi anomali dan menganalisis pola yang muncul untuk memberikan wawasan bisnis.
   - **Reporting**: Membuat laporan atau dashboard yang menampilkan data anomali dan potensi dampaknya terhadap bisnis.

c. Data Preparation:

1) **Data Engineer**:
   - **Data Integration**: Mengintegrasikan data dari berbagai sumber dan memastikan data tersebut terstruktur dengan baik untuk analisis lebih lanjut.
   - **Data Transformation**: Mengubah dan menormalkan data agar konsisten dan sesuai dengan kebutuhan analisis atau model prediktif.
2) **Data Scientist**:
   - **Feature Selection and Engineering**: Memilih dan menciptakan fitur yang relevan dari data yang telah disiapkan untuk digunakan dalam model machine learning.
   - **Data Sampling**: Melakukan sampling data untuk memastikan representasi yang tepat dalam pelatihan model dan validasi.
3) **Data Analyst**:

- **Exploratory Data Analysis (EDA)**: Melakukan analisis eksploratif awal untuk memahami distribusi, pola, dan hubungan dalam data.
- **Data Aggregation**: Mengagregasi data sesuai kebutuhan analisis atau pelaporan, seperti menghitung rata-rata, median, atau total.

3. Jelaskan langkah-langkah proses ETL dan ELT yang berperan dalam pekerjaan seorang data engineer!

   Proses ETL (Extract, Transform, Load) dan ELT (Extract, Load, Transform) adalah dua pendekatan utama dalam integrasi data yang sering digunakan oleh seorang Data Engineer. Berikut penjelasan langkah-langkah dari kedua proses tersebut:

   a. **ETL (Extract, Transform, Load):**

      ETL adalah proses tradisional di mana data diekstraksi dari sumber data, ditransformasi di suatu tempat (biasanya di server ETL atau alat ETL), dan kemudian dimuat ke dalam data warehouse atau sistem penyimpanan lainnya.

      1) **Extract (Ekstraksi)**:
         - **Pengumpulan Data**: Mengambil data dari berbagai sumber seperti database relasional, sistem ERP, file flat, API, atau sumber data lain.
         - **Teknik Ekstraksi**: Menggunakan query SQL, konektor API, atau alat integrasi data untuk menarik data.
         - **Faktor yang Dipertimbangkan**: Memastikan proses ekstraksi tidak mengganggu operasi sumber data dan data yang diambil adalah yang terbaru dan relevan.
      2) **Transform (Transformasi)**:
         - **Pembersihan Data**: Menghapus duplikasi, menangani nilai yang hilang, dan mengoreksi kesalahan dalam data.
         - **Normalisasi dan Denormalisasi**: Mengubah struktur data sesuai dengan skema data warehouse.
         - **Penggabungan Data**: Menggabungkan data dari berbagai sumber menjadi satu set data yang kohesif.
         - **Enkripsi dan Masking**: Melindungi data sensitif dengan mengenkripsi atau menutupi informasi yang diperlukan.
      3) **Load (Pemindahan)**:
         - **Pemindahan Data**: Memuat data yang telah ditransformasi ke dalam data warehouse, data mart, atau sistem penyimpanan lainnya.
         - **Teknik Pemindahan**: Menggunakan metode batch loading atau incremental loading, tergantung pada kebutuhan dan volume data.
         - **Validasi dan Verifikasi**: Memastikan data yang dimuat sesuai dengan yang diharapkan dan dapat digunakan untuk analisis.

   b. **ELT (Extract, Load, Transform):**

      ELT adalah pendekatan yang lebih modern, di mana data diekstraksi dari sumber, dimuat langsung ke dalam data warehouse atau sistem penyimpanan besar (seperti data lake), dan kemudian ditransformasi menggunakan kekuatan pemrosesan sistem tersebut.

      1) **Extract (Ekstraksi)**:
         - **Pengumpulan Data**: Sama seperti pada proses ETL, data diambil dari berbagai sumber.
         - **Teknik Ekstraksi**: Menggunakan query SQL, konektor API, atau alat integrasi data untuk menarik data.

2) **Load (Pemindahan)**:
   - **Pemindahan Data Mentah**: Data mentah dari berbagai sumber langsung dimuat ke dalam data warehouse atau data lake.
   - **Teknik Pemindahan**: Sering menggunakan batch loading untuk mentransfer volume data besar secara efisien.
   - **Validasi Awal**: Memastikan data yang dimuat konsisten dengan data sumber dan tidak rusak selama transfer.
3) **Transform (Transformasi)**:
   - **Transformasi di Tempat**: Menggunakan kekuatan pemrosesan data warehouse atau data lake untuk melakukan transformasi data setelah data dimuat.
   - **Skalabilitas**: Memanfaatkan kemampuan pemrosesan paralel dari sistem penyimpanan besar untuk menangani transformasi data yang kompleks dan besar.
   - **Fleksibilitas**: Memungkinkan transformasi data yang lebih dinamis dan dapat diubah sesuai kebutuhan analisis yang berkembang.

Peran Data Engineer dalam Proses ETL dan ELT:

a. **Desain dan Implementasi**:
   - **Desain Pipeline Data**: Merancang pipeline ETL atau ELT yang efisien dan dapat diskalakan.
   - **Pemilihan Alat**: Memilih alat dan teknologi yang tepat untuk proses ETL/ELT, seperti Talend, Apache Nifi, Informatica, atau alat cloud seperti AWS Glue, Azure Data Factory, dan Google Dataflow.

b. **Otomasi dan Orkestrasi**:
   - **Scheduling**: Mengotomatisasi proses ETL/ELT menggunakan scheduler seperti Apache Airflow atau cron jobs.
   - **Monitoring**: Memantau pipeline data untuk mendeteksi dan memperbaiki kesalahan dengan cepat.

c. **Pemeliharaan dan Optimasi**:
   - **Pemeliharaan Pipeline**: Memastikan pipeline ETL/ELT berjalan dengan lancar dan memperbarui pipeline sesuai kebutuhan.
   - **Optimasi Kinerja**: Mengoptimalkan proses ekstraksi, transformasi, dan pemindahan untuk meningkatkan kinerja dan efisiensi.

d. **Keamanan dan Kepatuhan**:
   - **Keamanan Data**: Melindungi data sensitif selama proses ETL/ELT dengan enkripsi dan kontrol akses.
   - **Kepatuhan Regulasi**: Memastikan proses ETL/ELT mematuhi regulasi dan kebijakan privasi data yang berlaku.

e. **Kolaborasi**:
   - **Kolaborasi dengan Data Scientist dan Data Analyst**: Bekerja sama dengan tim lain untuk memahami kebutuhan data dan menyediakan infrastruktur yang mendukung analisis dan model machine learning.
   - **Komunikasi**: Mengkomunikasikan perubahan dalam pipeline data dan dampaknya terhadap analisis atau pelaporan.

Dengan pemahaman yang kuat tentang proses ETL dan ELT, seorang Data Engineer dapat memastikan bahwa data yang digunakan dalam organisasi adalah akurat, konsisten, dan siap untuk analisis lebih lanjut.