FUNDAMENTAL OF DE

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?
   Jawab: Seorang Data Engineer bertanggung jawab untuk mendesain, membangun, dan memelihara infrastruktur data yang memungkinkan perusahaan untuk mengumpulkan, menyimpan, dan mengolah data dengan efisien. Ini meliputi:
   - Membangun Pipeline Data: Mengembangkan pipeline ETL (Extract, Transform, Load) untuk memastikan data dari berbagai sumber dapat diintegrasikan dan diolah.
   - Mendesain dan Mengelola Database: Memastikan database yang digunakan dapat menampung dan mengelola data dalam volume besar dengan performa yang baik.
   - Keamanan dan Integritas Data: Menjaga keamanan, integritas, dan kualitas data.
   - Otomatisasi Proses: Mengotomatisasi proses pengolahan data untuk meningkatkan efisiensi.
     Perbedaan dengan Data Scientist dan Data Analyst:
   - Data Engineer: Fokus pada infrastruktur dan pipeline data.
   - Data Scientist: Fokus pada analisis data yang kompleks menggunakan teknik statistik dan machine learning untuk menghasilkan insight.
   - Data Analyst: Fokus pada analisis data untuk menghasilkan laporan dan visualisasi yang mendukung pengambilan keputusan bisnis.

2. Berikan beberapa contoh peran dari seorang Data Engineer yang mungkin bersinggungan atau bahkan sama dengan peran Data Scientist dan Data Analyst!
   Jawab:
   - Membangun Pipeline Data: Data Engineer dan Data Scientist mungkin bekerja sama untuk membangun pipeline yang menyiapkan data untuk analisis.
   - Data Cleaning: Data Engineer dan Data Analyst mungkin melakukan proses pembersihan data untuk memastikan data yang digunakan adalah data berkualitas.
   - Data Modeling: Data Engineer mungkin membangun model data yang kemudian digunakan oleh Data Scientist untuk analisis lebih lanjut.

- **Implementasi Algoritma Machine Learning:** Data Engineer dapat bekerja sama dengan Data Scientist untuk mengimplementasikan dan mengoptimalkan model machine learning di lingkungan produksi.

3. Jelaskan langkah-langkah proses ETL dan ELT yang berperan dalam pekerjaan seorang data engineer!
   Jawab: ETL (Extract, Transform, Load)
   1. Extract: Mengambil data dari berbagai sumber data seperti database, file CSV, API, dll.
   2. Transform: Membersihkan dan mengubah data sesuai dengan kebutuhan analisis atau penyimpanan.
   3. Load: Memuat data yang sudah diubah ke dalam sistem penyimpanan seperti data warehouse.

   ELT (Extract, Load, Transform)
   1. Extract: Mengambil data dari berbagai sumber data.
   2. Load: Memuat data mentah langsung ke sistem penyimpanan.
   3. Transform: Mengubah data sesuai kebutuhan di dalam sistem penyimpanan tersebut.

PART 2 - Fundamental DE

TASK 1

1. Kapan kita harus menggunakan relational database atau nosql database?
   Jawab: Relational Database digunakan ketika data memiliki struktur yang jelas dan membutuhkan konsistensi tinggi, serta mendukung transaksi ACID. Contoh: MySQL, PostgreSQL. Sementara NoSQL Database digunakan ketika data tidak memiliki struktur yang tetap, skalabilitas tinggi, dan performa cepat untuk operasi baca/tulis. Contoh: MongoDB, Cassandra.

2. Apa perbedaan antara database, data lake, data warehouse, dan data mart?
   Jawab:
   - Database: Sistem penyimpanan data yang terstruktur yang mendukung operasi CRUD (Create, Read, Update, Delete).
   - Data Lake: Penyimpanan besar yang dapat menampung data dalam berbagai bentuk (struktur, semi-struktur, tidak terstruktur).
   - Data Warehouse: Sistem penyimpanan yang dioptimalkan untuk query dan analisis data dalam jumlah besar. Data di sini biasanya sudah diproses dan diorganisir.

- Data Mart: Subset dari data warehouse yang dioptimalkan untuk kebutuhan analisis departemen tertentu.

3. Jelaskan apa itu normalisasi database, dan normalisasikan tabel dibawah!

Employee id	Employee name	Job code	Job	City code	City name	Province code	Province name
1	John Smith	101	Software Engineer	201	New York	301	New York
2	Alice Johnson	102	Data Analyst	202	Los Angeles	302	California
3	Bob Davis	103	Data Engineer	203	Chicago	303	Illinois
4	Emily Wilson	101	Software Engineer	204	Houston	304	Texas
5	Michael Lee	102	Data Analyst	205	Miami	305	Florida
6	Sarah Brown	103	Data Engineer	206	Boston	306	Massachusetts
7	James Clark	101	Software Engineer	207	San Fransisco	307	California
8	Laura Taylor	102	Data Analyst	208	Seattle	308	Washington
9	Daniel White	103	Data Engineer	209	Denver	309	Colorado
10	Olivia Martin	101	Software Engineer	210	Atlanta	310	Georgia

Jawab:

Normalisasi adalah proses mengatur data dalam database untuk mengurangi redundansi dan meningkatkan integritas data. Proses normalisasi biasanya melalui beberapa bentuk normal (Normal Forms):

1. First Normal Form (1NF): Menghilangkan duplikasi data dalam tabel, memastikan setiap kolom hanya berisi satu nilai.
2. Second Normal Form (2NF): Memastikan bahwa data non-primer sepenuhnya bergantung pada kunci utama.
3. Third Normal Form (3NF): Menghapus ketergantungan transitif, yaitu data non-primer tidak boleh bergantung pada kolom non-primer lainnya.

Normalisasi Tabel

Tabel emloyees:

employee_id	employee_name
1	Joh Smith
2	Alice Johnson

3	Bob Davis
4	Emily Wilson
5	Michael Lee
6	Sarah Brown
7	James Clark
8	Laura Taylor
9	Daniel White
10	Olivia Martin

Tabel cities:

city_code	city_name	province_code
201	New York	301
202	Los Angeles	302
203	Chicago	303
204	Houston	304
205	Miami	305
206	Boston	306
207	San Fransisco	307
208	Seattle	308
209	Denver	309
210	Atlanta	310

Tabel provinces:

province_code	province_name
301	New York
302	California
303	Illinois
304	Texas
305	Florida
306	Massachusetts
307	California
308	Washington
309	Colorado
310	Georgia

Tabel employees_jobs:

employee_id	job_code
1	101
2	102
3	103
4	104
5	105
6	106
7	107

8	108
9	109
10	101

Tabel employees_cities:

employee_id	city_code
1	201
2	202
3	203
4	204
5	205
6	206
7	207
8	208
9	209
10	210