**Lecture 1: The theory  
  
Tóm tắt nội dung tài liệu "Introduction to Data Engineering"**

Tài liệu này giới thiệu về Data Engineering (Kỹ thuật dữ liệu), tập trung vào việc xây dựng **data pipelines** (hệ thống xử lý dữ liệu) và các kỹ năng cần thiết để trở thành một Data Engineer giỏi.

**1. Tổng quan về Data Engineering**

* Data Engineers giúp di chuyển và xử lý dữ liệu để mang lại giá trị cho doanh nghiệp.
* Data pipelines giúp tự động hóa quá trình thu thập, xử lý và lưu trữ dữ liệu.
* Kỹ thuật dữ liệu khác với lập trình phần mềm truyền thống vì kết hợp cả kỹ năng kỹ thuật lẫn quản lý dữ liệu.

**2. Nội dung khóa học**

Khóa học được chia thành nhiều chương, bao gồm:

* **Chương 1: Lý thuyết** – Khái niệm về data pipelines, cách thiết kế hệ thống lưu trữ và truy cập dữ liệu.
* **Chương 2: Cấu trúc Pipeline** – Cách tổ chức code, kiểm thử và tài liệu hóa pipelines.
* **Chương 3: Kiến trúc hệ thống** – Ứng dụng kiến trúc vào quản lý dữ liệu lớn, tối ưu chi phí.
* **Chương 4: Lưu trữ dữ liệu** – So sánh các kiểu lưu trữ (SQL, NoSQL, Parquet, Avro, JSON).
* **Chương 5: Tài nguyên tính toán** – CPU, RAM, lưu trữ và cách tối ưu hệ thống.
* **Chương 6: SQL nâng cao** – Tối ưu truy vấn, indexing và kết hợp Python với SQL.
* **Chương 7: Data Warehousing** – So sánh Data Warehouse, Data Lake và Lakehouse.
* **Chương 8: Mô hình dữ liệu** – Cách thiết kế dữ liệu hiệu quả, tránh trùng lặp, đảm bảo tính toàn vẹn.
* **Chương 9: Chất lượng dữ liệu** – Xác minh tính chính xác và độ tin cậy của dữ liệu.
* **Chương 10: DevOps cho Data Engineering** – CI/CD, kiểm thử và tự động hóa quy trình.

**3. Các yếu tố quan trọng trong Data Engineering**

* **Di chuyển dữ liệu** (Movement) – Lấy dữ liệu từ nguồn và chuyển đến đích.
* **Lưu trữ** (Storage) – Chọn cách lưu trữ dữ liệu phù hợp với nhu cầu hệ thống.
* **Truy cập** (Access) – Batch processing, stream processing, interactive processing.
* **Tính lặp lại** (Repeatable) – Dễ dàng kiểm thử và tái sử dụng pipeline.
* **Tính linh hoạt** (Resilient) – Không dễ bị gián đoạn khi thay đổi yêu cầu.
* **Tính mở rộng** (Scalable) – Khả năng xử lý khối lượng dữ liệu lớn mà không ảnh hưởng đến hiệu suất.

**4. Tóm tắt cuối cùng**

Khi xây dựng data pipeline, luôn cần xem xét các yếu tố: **Di chuyển, Lưu trữ, Truy cập, Tính lặp lại, Linh hoạt và Khả năng mở rộng** để đảm bảo hiệu quả và tối ưu hệ thống dữ liệu.