**1. Data engineer (kỹ sư dữ liệu):** Chuyên gia thiết kế, xây dựng, vận hành và duy trì các hệ thống xử lý và phân tích dữ liệu. **2. Manufacturing operation (hoạt động sản xuất):** Quá trình tạo ra sản phẩm bằng cách chuyển đổi nguyên liệu thô thành thành phẩm. **3. Data warehouse (kho lưu trữ dữ liệu):** Hệ thống tập trung dữ liệu từ nhiều nguồn khác nhau trong tổ chức, được sử dụng để phân tích và báo cáo. **4. Data services (dịch vụ dữ liệu):** Các dịch vụ đám mây cung cấp các chức năng liên quan đến dữ liệu, chẳng hạn như lưu trữ, xử lý và phân tích. **5. Combine data (kết hợp dữ liệu):** Trộn dữ liệu từ nhiều nguồn khác nhau thành một tập dữ liệu duy nhất. **6. Reformat data (định dạng lại dữ liệu):** Chuyển đổi dữ liệu từ một định dạng sang định dạng khác phù hợp hơn cho phân tích. **7. Analytical models (mô hình phân tích):** Cấu trúc dữ liệu được tổ chức theo cách hỗ trợ phân tích. **8. Subsequent querying, reporting, and visualization (truy vấn, báo cáo và trực quan hóa sau đó):** Các hoạt động sử dụng dữ liệu trong kho lưu trữ dữ liệu để trả lời câu hỏi, tạo báo cáo và biểu đồ.

**9. Data integration service (dịch vụ tích hợp dữ liệu):** Dịch vụ giúp trích xuất, chuyển đổi và tải dữ liệu từ nhiều nguồn khác nhau vào một hệ thống duy nhất. (Azure Data Factory) **10. Noise (dữ liệu nhiễu):** Dữ liệu không mong muốn hoặc không chính xác cần loại bỏ khỏi tập dữ liệu. **11. Ingested data (dữ liệu đã thu thập):** Dữ liệu được lấy từ các nguồn khác nhau và đưa vào hệ thống.

**12. Data store (kho lưu trữ dữ liệu):** Vị trí lưu trữ dữ liệu, có thể là một cơ sở dữ liệu quan hệ hoặc một hồ dữ liệu.

**13. Pipeline (luồng xử lý):** Trình tự các hoạt động được thực hiện trên dữ liệu, thường được biểu diễn dưới dạng chuỗi các tác vụ. **14. Graphical user interface (giao diện người dùng đồ họa - GUI):** Giao diện người dùng sử dụng các yếu tố hình ảnh như biểu tượng và cửa sổ để người dùng tương tác với hệ thống.

**15. Data lake (hồ dữ liệu):** Kho lưu trữ dữ liệu dung lượng lớn chứa dữ liệu thô, chưa được xử lý. **16. Raw data (dữ liệu thô):** Dữ liệu chưa được xử lý hoặc định dạng. **17. Structured data (dữ liệu có cấu trúc):** Dữ liệu được tổ chức theo một định dạng cụ thể, thường được lưu trữ trong các bảng cơ sở dữ liệu. **18. Staging point (điểm trung gian):** Vị trí tạm thời để lưu trữ dữ liệu trước khi được xử lý và đưa vào kho lưu trữ dữ liệu chính thức.

**19. Massaged data (dữ liệu được chỉnh sửa):** Dữ liệu thô đã được xử lý và định dạng lại để phù hợp cho phân tích.

**20. Hierarchical directory structure (cấu trúc thư mục phân cấp):** Cách tổ chức dữ liệu thành các thư mục và thư mục con. **21. File system semantics (ngữ nghĩa hệ thống file):** Cách thức hệ thống lưu trữ và quản lý các file.

**22. Near infinite file system (hệ thống file dung lượng gần như vô hạn):** Hệ thống file có khả năng lưu trữ một lượng dữ liệu rất lớn.

**23. Granular role-based access control (RBAC) (kiểm soát truy cập dựa trên vai trò chi tiết):** Phương pháp kiểm soát quyền truy cập vào dữ liệu dựa trên vai trò và cấp độ của người dùng.

**24. Portable Operating System Interface (POSIX) (giao diện hệ điều hành di động):** Một tập hợp các tiêu chuẩn cho các hệ điều hành máy tính, giúp đảm bảo tính tương thích giữa các hệ thống khác nhau.

**25. Hadoop Distributed File System (HDFS) (Hệ thống file phân tán Hadoop):** Hệ thống file phân tán được thiết kế để lưu trữ và quản lý các tập dữ liệu lớn trên các cụm máy tính.

**26. Big data processing (xử lý dữ liệu lớn):** Quy trình xử lý và phân tích các tập dữ liệu có kích thước rất lớn

**Giải thích kỹ hơn các khái niệm:**

**1. Combine data (kết hợp dữ liệu):**

* **Khái niệm:** Kết hợp dữ liệu là quá trình **trộn dữ liệu từ nhiều nguồn khác nhau** thành một tập dữ liệu duy nhất.
* **Mục đích:** Việc kết hợp dữ liệu giúp:
  + **Cung cấp cái nhìn tổng quan** về hoạt động kinh doanh.
  + **Hỗ trợ phân tích dữ liệu** hiệu quả hơn.
  + **Phát hiện các mối tương quan** mà không thể nhìn thấy từ dữ liệu của từng nguồn riêng lẻ.
* **Ví dụ:**
  + Một công ty bán lẻ có thể kết hợp dữ liệu bán hàng từ các cửa hàng khác nhau để phân tích xu hướng bán hàng theo khu vực.
  + Một ngân hàng có thể kết hợp dữ liệu giao dịch từ các kênh khác nhau (ATM, chi nhánh, ứng dụng di động) để phân tích hành vi của khách hàng.

**2. Reformat data (định dạng lại dữ liệu):**

* **Khái niệm:** Định dạng lại dữ liệu là quá trình **chuyển đổi dữ liệu từ một định dạng sang định dạng khác phù hợp hơn cho phân tích**.
* **Mục đích:** Việc định dạng lại dữ liệu giúp:
  + **Tăng hiệu quả lưu trữ**.
  + **Tăng hiệu quả xử lý**.
  + **Tăng khả năng tương thích** với các công cụ phân tích.
* **Ví dụ:**
  + Chuyển đổi dữ liệu từ định dạng CSV sang định dạng JSON để dễ dàng phân tích bằng các công cụ NoSQL.
  + Chuyển đổi dữ liệu từ định dạng text sang định dạng bảng để dễ dàng phân tích bằng các công cụ BI.

**3. Analytical models (mô hình phân tích):**

* **Khái niệm:** Mô hình phân tích là **cấu trúc dữ liệu được tổ chức theo cách hỗ trợ phân tích**.
* **Mục đích:** Việc sử dụng mô hình phân tích giúp:
  + **Đơn giản hóa việc truy vấn dữ liệu**.
  + **Tăng hiệu quả phân tích**.
  + **Thu được các thông tin chi tiết** từ dữ liệu.
* **Ví dụ:**
  + Mô hình dữ liệu khách hàng bao gồm thông tin cá nhân, lịch sử mua hàng và sở thích.
  + Mô hình dữ liệu sản phẩm bao gồm thông tin sản phẩm, giá cả và danh mục.

**4. Pipeline (luồng xử lý):**

* **Khái niệm:** Luồng xử lý là **trình tự các hoạt động được thực hiện trên dữ liệu**, thường được biểu diễn dưới dạng chuỗi các tác vụ.
* **Mục đích:** Việc sử dụng luồng xử lý giúp:
  + **Tự động hóa các tác vụ xử lý dữ liệu**.
  + **Tăng hiệu quả xử lý dữ liệu**.
  + **Đảm bảo tính nhất quán** trong việc xử lý dữ liệu.
* **Ví dụ:**
  + Luồng xử lý ETL (Extract, Transform, Load) để trích xuất dữ liệu từ các nguồn khác nhau, chuyển đổi dữ liệu sang định dạng phù hợp và tải dữ liệu vào kho lưu trữ dữ liệu.
  + Luồng xử lý phân tích dữ liệu để thực hiện các thao tác tính toán, thống kê và trực quan hóa dữ liệu.

**5. Staging point (điểm trung gian):**

* **Khái niệm:** Điểm trung gian là **vị trí tạm thời để lưu trữ dữ liệu trước khi được xử lý và đưa vào kho lưu trữ dữ liệu chính thức**.
* **Mục đích:** Việc sử dụng điểm trung gian giúp:
  + **Giảm tải cho kho lưu trữ dữ liệu**.
  + **Cải thiện hiệu suất xử lý dữ liệu**.
  + **Đảm bảo tính toàn vẹn** của dữ liệu.
* **Ví dụ:**
  + Lưu trữ dữ liệu thô từ các nguồn khác nhau vào điểm trung gian trước khi chuyển đổi sang định dạng phù hợp và tải vào kho lưu trữ dữ liệu.
  + Lưu trữ dữ liệu kết quả sau khi phân tích vào điểm trung gian trước khi