**Câu 1: Trình bày các thành phần của Hadoop**

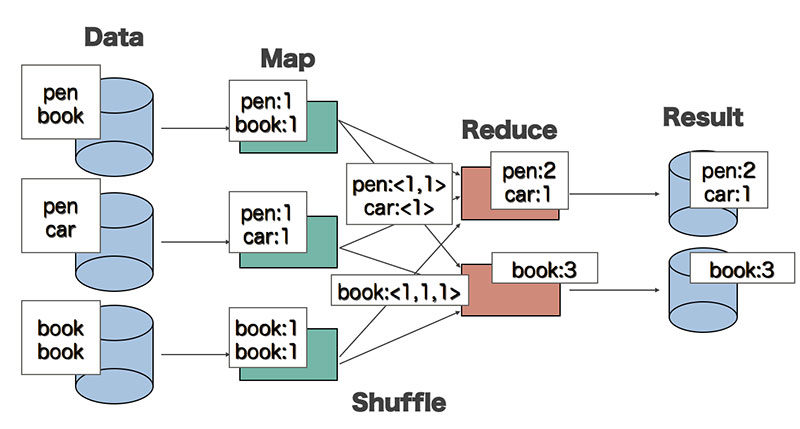
**Khái nhiệm về hadoop:**

* Apache framework nguồn mở viết bằng java cho phép phát triển các ứng dụng phân tán có cường độ dữ liệu lớn miễn phí.
* Được thiết kế để mở rộng quy mô từ một máy chủ đơn sang nhiều máy tính khác có tính toán và lưu trữ cục bộ (local computation and storage)
* Phát triển dựa trên mô hình Map-Reduce và hệ thống file phân tan Google File System (GFS)
* Nhờ cơ chế streaming hadoop có thể phát triển trên các ứng dụng phân tán bằng cả java và các ngôn ngữ lập trình khác như C++, Python, …

**Kiến trúc Hadoop**

Kiến trúc hadoop gồm ba lớp chính:

* **HDFS (Hadoop Distributed File System):** là một hệ thống tập tin phân tán được sử dụng để lữu trữ và quản lý dữ liệu lớn trên mạng
  + Đặc trưng
    - Phân tán
    - Tính toàn vẹn
    - Hỗ trợ lưu trứ lớn
    - Tốc độ truy xuất cao
    - Tính linh hoạt
    - Tính chính xác
  + Thiết kế HDFS
    - HDFS có cấu trúc phân cấp cho các file và thư mục trong hệ thống
    - Mỗi file trong HDFS được chia thành nhiều Block, mỗi Block có một BlockID duy nhất
    - Tất cả Block trữ Block cuối mỗi file có độ lớn 128MB
    - Có kiến trúc Master/slave:
      * NameNode duy nhất cho toàn cluster
      * DataNode không giới hạn, tối thiểu 2
  + Cơ chế đọc ghi dữ liệu của HDFS
    - HDFS được thiết kế để hỗ trợ đọc nhiều lần và ghi một lần (read-many, write-once)
* **Yarn**
  + Yarn có 2 chức năng cơ bản là quản lý tài nguyên và theo dõi lập lịch công việc
  + Thực hiện 2 nhiệm vụ này tương ứng là ResourceManager và ApplicationMaster
  + Các thành phần chính của Yarn
    - ResouceManager
    - NodeManager
    - ApplicationMaster
* **Mapreduce**
  + Định nghĩa
    - Là mô hình xử lý tính toán song song và phân tán trên hệ thống phân tán
    - Phân rã nghiệp vụ chính thành các công việc con để thực hiện xử lý song song, sau đó thu lại kết quả



* + Đối với hàm **Map**:
    - Input là 1 đoạn văn bản
    - Output là các cặp <word, 1>

Hàm Map được thực hiện song song để xử lý các tập dữ liệu khác nhau.

* + Đối Với hàm **Reduce**:
    - Input có dạng <word, [list]>, trong đó list là tập hợp các giá trị đếm được của mỗi từ
    - Output: <word, tổng số lần xuất hiện của từ>

Hàm Reduce cũng được chạy song song để xử lý các tập từ khoá khác nhau.

Giữa hàm Map và Reduce có một giai đoạn xử lý trung gian gọi là hàm Shuffle. Hàm này có nhiệm vụ sắp xếp các từ và tổng hợp dữ liệu đầu vào cho Reduce từ các kết quả đầu ra của hàm Map.

**Câu 2 Trình bày về Dstream Window**

**Khái niệm:**

* Dstream Window là một khung thời gian trượt trên Dstream (Distributed Stream) trong Spark Streaming.
* Cho phép xử lý dữ liệu trong một khoảng thời gian xác định và tạo ra kết quả

**Một số Khái niệm trong Dstream Window:**

* ***Kích thước cửa sổ (Window Size):*** xác định khoảng thời gian muốn xử lý dữ liệu trong đó
  + Ví dụ: nếu thiết lập cửa sổ 5 phút, Spark Streaming sẽ tự động xử lý dữ liệu trong cửa sổ 5 phút liên tiếp
* ***Kích thước bước (Slide Interval):*** là khoảng thời gian giữa các cửa sổ liên tiếp. Nó xác định tần suất mà Dstream Window được cập nhật
  + Ví dụ: nếu thiết lập kích thước bước là 1 phút trong cửa sổ 5 phút. Thì cửa sổ mới sẽ được tạo mỗi 1 phút và xử lý dữ liệu trong 5 phút gần nhất
* ***Xử lý trong cửa sổ:*** Trong mỗi cửa sổ, có thể thực hiện các hoạt động xử lý dữ liệu như áp dụng hàm map, reduce, filter hoặc bất kì phép biến đổi nào khác. Dữ liệu trong cửa sổ sẽ được tự động chuyển đến các phép biến đổi này để tính toán kết quả

**Ứng dụng:**

* Dstream Window thường được sử dụng để thực hiện các tác vụ như tính toán tổng của dữ liệu trong khoảng thời gian, xác định giá trị trung bình, tối đa, tối thiểu hoặc xác định sự thay đổi của dữ liệu theo thời gian.
* Giúp xác định xu hướng và thông tin quan trọng trong dữ liệu thời gian thực một cách hiệu quả.

**Ví Dụ:**

* Nếu muốn theo dõi tổng số lượng truy cập vào 1 trang web trong cửa sổ 1 giờ, có thể sử dụng Dstream Window với kích thước cửa sổ là 1 giờ và kích thước bước là 10 phút. Spark streaming sẽ tự động tính tổng lượng truy cập trong mỗi cửa sổ 1 giờ và cập nhật kết quả sau mỗi 10 phút