# Apache Spark và MapReduce

## [Apache Spark]

### Sự ra đời

Apache Spark là một open-source clustering computing framework. Spark ra đời để đáp ứng các giới hạn mà MapReduce Hadoop gặp phải trong xử lí dữ liệu lớn. Nó giải quyết được các vấn đề như

* Có thể hoạt động phù hợp với real-time processing, và xử lí được lượng dữ liệu lớn hơn.
* Cung cấp sức mạnh xử lí nhanh hơn khoảng 100 lần trong một số trường hợp.
* Phù hợp với các giải thuật Machine Learning vì xử lí dữ liệu in-memory và có thể cung cấp dữ liệu liên tục, ngay lập tức.

### Thành phần

Apache Spark có nhiều thành phần hổ trợ cho nhiều mục đích xử lí khác nhau. Bao gồm:

* Spark Core và Resilient Distributed Dataséts (RDDs)
* Spark SQL
* Spark Streaming
* Machine Learning library (MLlib)

## [MapReduce]

### Định nghĩa

MapReduce là một một hình lập trình hay mô hình tính toán (programing model) cho phép cấu trúc và tổ chức các tính toán của người sử dụng để tối ưu trên một hệ thống nhiều máy tính. Có thể coi MapReduce là một mô hình “chia để trị” (ý kiến cá nhân).

### Các giai đoạn

* Map Stage : sinh ra các dữ liệu trung gian.
* Shuffle Stage : thu thập tổng hợp các dữ liệu trung gian để chuyển sang cho Reduce xử lí.
* Reduce Stage : nhận dữ liệu tổng hợp, ghép thành các tập cùng khóa nhỏ hơn.

Để dễ diễn đạt công việc mà giai đoạn này thực hiện. Xin được lấy bài tập WordCounting với MapReduce.

Giả sử có nhiều document chứa văn bản. Công việc cần làm là đếm số lần xuất hiện của các từ trong các document đấy.

Map Stage :

Dữ liệu sau khi được tiền xử lí có dạng một list các từ. Được tổng hợp lại thành các cặp (word, counter) với word là từ xuất hiện và counter bằng 1 cho sự xuất hiện của nó trong document. Các cặp (word, 1) này là các dữ liệu trung gian truyền cho giai đoạn kế tiếp.

Shuffle Stage :

Dữ liệu trung gian sau khi vào giai đoạn này được tổng hợp lại thành các tập các dữ liệu cùng key.

(the, 1) (the, 1) (the, 1)

(year, 1) (year, 1) (year, 1) (year, 1)

Reduce Stage:

Ở giai đoạn này các tập dữ liệu được thu gọn thành các tập cùng key với kích thước nhỏ hơn.

(the, 3)

(year, 4)