**Sinh viên: Nguyễn Thị Quỳnh Như**

**MSSV: 51800223**

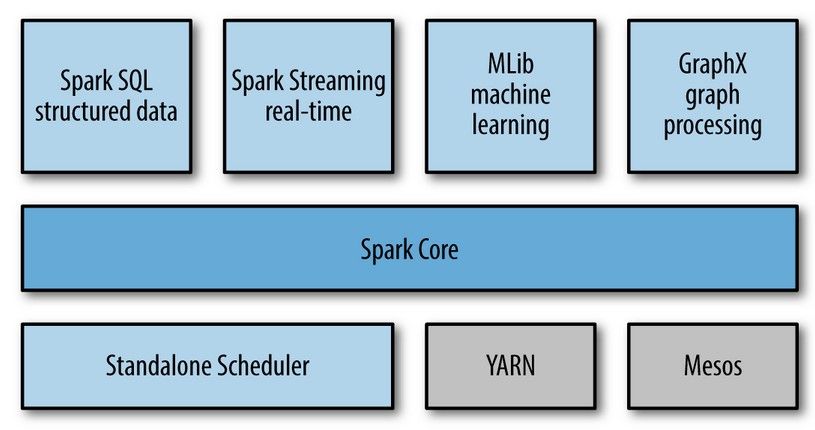
**APACHE SPARK**

Tổng quan về Apache Spark

* Apache Spark là một open source cluser computing framework.
* Cho phép xây dựng các mô hình dự đoán nhanh chóng, có thể tính toán cùng lúc trên toàn bộ tập dữ liệu mà không cần phải trích xuất mẫu tính toán thử nghiệm.
* Tốc độ xử lý của Spark nhanh là do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau.
* Việc tính toán được thực hiện ở bộ nhớ trong hay thực hiện hoàn toàn trên RAM.

Thành phần của Apache Spark

* Spark Core: cung cấp những chức năng cơ bản nhất như lập lịch cho các tác vụ ,quản lý bộ nhớ, fault recovery, tương tác với các hệ thống lưu trữ… Cung cấp API để định nghĩa RDD là tập hợp của các item được phân tán trên các node của cluster và có thể xử lý song song.
* Spark có thể chạy trên nhiều loại Cluster Managers như Hadoop YARN, Apache Mesos hoặc trên chính Cluster Manager được cung cấp bởi Spark là Standalone Scheduler.
* Spark SQL: cho phép truy vấn dữ liệu cấu trúc qua các câu lệnh SQL, có thể thao tác với nhiều nguồn dữ liệu như Hive tables, Parquet và JSON.
* Spark Streaming: cung cấp API để dễ dàng xử lý dữ liệu stream.
* MLlib: cung cấp rất nhiều thuật toán như classification, regression, clustering, collaborative…
* GraphX: thư viện để xử lý đồ thị.



Những tính năng nổi bật

* “Spark as a Service”: giao diện REST để quản lý spark job, spark context.
* Tăng tốc, giảm độ trễ thực thi job xuống mức chỉ tính bằng giây bằng cách tạo sẵn spark context cho các job dùng chung.
* Stop job đang chạy bằng cách stop spark context.
* Bỏ bước upload gói jar lúc start job làm cho job được start nhanh hơn.
* Cung cấp hai cơ chế chạy job đồng bộ và bất đồng bộ.
* Cho phép cache RDD theo tên, tăng tính chia sẻ và sử dụng lại RDD giữa các job.
* Hỗ trợ viết spark job bằng cú pháp SQL.
* Dễ dàng tích hợp với các công cụ như Business Intelligence, Analytics, Data Integration Tools.

**MAPREDUCE**

Tổng quan về Mapreduce

* Mapreduce là mô hình có khả năng lập trình xử lý các tập dữ liệu lớn song song và phân tán thuật toán trên một cụm cái tính.
* Mapreduce gồm các thủ tục: thủ tục Map() và thủ tục Reduce(). Thủ tục Map() gồm lọc và phân loại trên dữ liệu khi thủ tục Reduce() thực hiện.

Các hàm chính của MapReduce

* Hàm Map(): nhận input cho các cặp giá trị/ khóa và output chính là tập những cặp giá trị/ khóa trung gian. Dữ liệu chỉ cần ghi xuống đĩa cứng và tiến hành thông báo cho các hàm Reduce() trực tiếp nhận dữ liệu.
* Hàm Reduce(): tiếp nhận giá trị/ khóa trung gian sau đó ghép chúng lại tạo thành một cặp khóa khác nhau.
* Shuffle: trung gian giữa Map và Reduce, tổng hợp giá trị/ khóa được Map sinh ra và chuyển cho Reduce tiếp tục xử lý.

Các ưu điểm nổi bật

* Xử lý dễ dàng mọi bài toán có lượng dữ liệu lớn nhờ khả năng tác vụ phân tích và tính toán phức tạp.
* Có khả năng chạy song song trên các máy phân tán khác nhau.
* Có khả năng chạy trên nhiều nguồn ngôn ngữ lập trình khác nhau.
* Các ứng dụng Mapreduce hướng đến việc phát hiện các mã độc để xử lý chúng.

Nguyên tắc hoạt động

* Phân chia các dữ liệu cần xử lý thành nhiều phần nhỏ trước khi thực hiện.
* Xử lý các vấn đề nhỏ theo phương thức song song trên các máy tính rồi phân tán hoạt động theo hướng độc lập.
* Tiến hành tổng hợp những kết quả thu được để đề ra được kết quả sau cùng.

Các bước hoạt động

* Tiến hành chuẩn bị các dữ liệu đầu vào để cho Map() có thể xử lý.
* Lập trình viên thực thi các mã Map() để xử lý.
* Tiến hành trộn lẫn các dữ liệu dược xuất ra bởi Map() vào trong Reduce Processor.
* Tiến hành thực thi tiếp mã Reduce() để có thể xử lý tiếp các dữ liệu cần thiết.
* Thực hiện tạo các dữ liệu xuất ra cuối cùng.

Luồng dữ liệu nền tảng

* Input Reader
* Map Function
* Partition Function
* Compare Function
* Reduce Function
* Output Writer

**TÀI LIỆU THAM KHẢO**

<https://blog.itnavi.com.vn/mapreduce-nhung-uu-diem-va-cach-thuc-hoat-dong-cua-nen-tang-nay/>

<https://viblo.asia/p/tong-quan-ve-apache-spark-cho-he-thong-big-data-RQqKLxR6K7z>