**SPARK VÀ MAPREDUCE**

# **Apache Spark**

## ***Khái Niệm :***

Apache Spark là một cụm framework điện toán nhanh, được thiết kế để tính toán nhanh trong xử lý dữ liệu quy mô lớn. Apache Spark là một công cụ xử lý phân tán (distributed processing engine) nhưng nó không đi kèm với trình quản lý tài nguyên cụm (inbuilt cluster resource manager) và hệ thống lưu trữ phân tán sẵn có (distributed storage system) mà phải cắm vào một trình quản lý cụm và hệ thống lưu trữ.

Apache Spark bao gồm Spark Core và Bộ thư viện. Spark core thực thi và quản lý công việc bằng cách cung cấp trải nghiệm liền mạch cho người dùng. Người dùng phải gửi công việc tới Spark core và Spark core đảm nhiệm việc xử lý, thực thi và trả lời lại cho người dùng qua API Spark Core bằng các ngôn ngữ lập trình khác nhau như Scala, Python, Java và R.

Apache Spark là công cụ xử lý dữ liệu cho các chế độ hàng loạt và phát trực tuyến có các truy vấn SQL, Xử lý đồ thị (Graph Processing) và Machine Learning.

## ***Xử lý dữ liệu***

Spark xử lý dữ liệu theo lô và thời gian thực

## ***Tính tương thích***

Có thể tích hợp với tất cả các nguồn dữ liệu và định dạng tệp được hỗ trợ bởi cụm Hadoop.

## ***Hỗ trợ ngôn ngữ***

hỗ trợ Java, Scala, Python và R.

## ***Phân tích thời gian thực***

Apache Spark có thể xử lý dữ liệu thời gian thực tức là dữ liệu đến từ các luồng sự kiện thời gian thực với tốc độ hàng triệu sự kiện mỗi giây. Ví dụ: Data Twitter chẳng hạn hoặc luợt chia sẻ, đăng bài trên Facebook. Sức mạnh Spark là khả năng xử lý luồng trực tiếp hiệu quả.

Apache Spark có thể được sử dụng để xử lý phát hiện gian lận trong khi thực hiện các giao dịch ngân hàng. Đó là bởi vì, tất cả các khoản thanh toán trực tuyến được thực hiện trong thời gian thực và chúng ta cần ngừng giao dịch gian lận trong khi quá trình thanh toán đang diễn ra.

## ***Mục tiêu sử dụng***

Xử lý dữ liệu nhanh và tương tác

- Xử lý đồ thị

- Công việc lặp đi lặp lại

- Xử lý thời gian thực

- joining Dataset

- Machine Learning

Apache Spark là Framework thực thi dữ liệu dựa trên Hadoop HDFS. Apache Spark không thay thế cho Hadoop nhưng nó là một framework ứng dụng. Apache Spark tuy ra đời sau nhưng được nhiều người biết đến hơn Apache Hadoop vì khả năng xử lý hàng loạt và thời gian thực.

# **Apache Hadoop**

## ***Khái Niệm :***

## Apache Hadoop là một framework nguồn mở được thiết kế để mở rộng quy mô từ các máy chủ đơn lẻ đến hàng ngàn máy và chạy các ứng dụng trên các cụm phần cứng.

## Apache Hadoop có thể xử lý nhiều việc khá tốt. Nó đã được phát triển và dần hoàn thiện với các tính năng và khả năng mới để giúp thiết lập và sử dụng dễ dàng hơn. Có một hệ sinh thái lớn các ứng dụng hiện đang tận dụng Hadoop.

## Hadoop của Apache được chia thành hai lớp. Lớp đầu tiên là lớp lưu trữ và được gọi là Hệ thống tệp phân tán Hadoop (Hadoop Distributed File System -HDFS), lớp thứ hai là lớp xử lý và được gọi là MapReduce. Lớp lưu trữ của Hadoop, tức là HDFS chịu trách nhiệm lưu trữ dữ liệu trong khi MapReduce chịu trách nhiệm xử lý dữ liệu trong Hadoop Cluster.

## ***Xử lý dữ liệu***

## Apache Hadoop chỉ xử lý dữ liệu theo lô

## Apache Hadoop chậm hơn Apache Spark

## ***Tính tương thích***

## Apache Hadoop tương thích chủ yếu nhiều các data sources và định dạng tệp

## ***Hỗ trợ ngôn ngữ***

## Apache Hadoop hỗ trợ Ngôn ngữ chính là Java nhưng các ngôn ngữ như C, C ++, ruby, Python, Perl, Groovy cũng được hỗ trợ bằng Hadoop Streaming

## ***Phân tích thời gian thực***

## MapReduce của Apache Hadoop không thể xử lý dữ liệu thời gian thực vì nó được thiết kế để thực hiện xử lý hàng loạt trên lượng dữ liệu khổng lồ, trong khi

## Apache Hadoop được sử dụng cho công việc xử lý hàng loạt như phân tích các tham số khác nhau như nhóm tuổi, địa điểm, thời gian sử dụng video YouTube cụ thể trong 24 giờ hoặc 7 ngày qua.

## ***Mục tiêu sử dụng***

Sử dụng khi cần:

- Xử lý hàng loạt bộ dữ liệu lớn

- Không yêu cầu giải pháp trung gian

Apache Hadoop cung cấp giao diện web để gửi và thực thi công việc. Tất cả các công cụ Business Intelligence (BI) như Qlikview, Tableau, Zoom Data, Zeeplin đều có kết nối với Hadoop và hệ sinh thái của nó.

Có nghĩa là bạn có thể lưu trữ dữ liệu trong HDFS và sau khi xử lý dữ liệu sau khi sử dụng các công cụ Hadoop, bạn có thể trực tiếp hình dung output của mình chạy trên hệ thống lưu trữ.

Nguồn tham khảo : <https://cloudfun.vn/threads/phan-biet-apache-hadoop-va-apache-spark.94/>