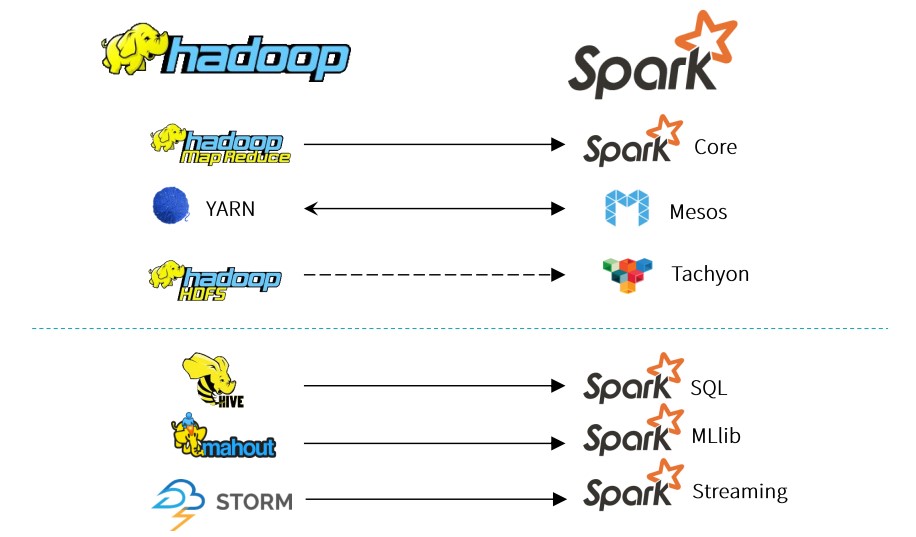
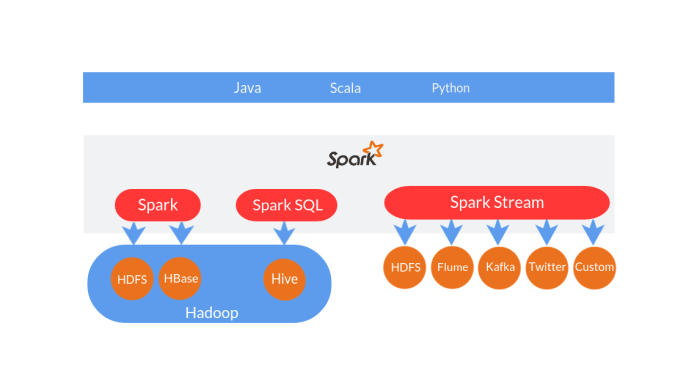
# Hadoop vs Spark



* **Hadoop MapReduce** có thể bị thay thế bởi **Spark Core**. Như những phân tích ở trên, tốc độ là ưu điểm nổi bật của Spark Core điều này khiến Hadoop MapReduce dường như đang nằm ngoài dòng chảy của dữ liệu lớn.
* **YARN** có thể thay thế bởi **Mesos.**
* **Hadoop HDFS** được viết từ những năm 2000 khi bộ nhớ còn rất đắt, ngày nay khi bộ nhớ ngày càng rẻ hơn thì những hệ thống lưu trữ phân tán lấy bộ nhớ làm trung tâm như **Apache Tachyon** (tên mới là **Alluxio**) có thể là xu hướng của tương lai.
* **Spark SQL** có thể dùng thay thế cho **HIVE,** Spark SQL ngày càng được cải tiến để dễ sử dụng hơn và cung cấp nhiều tính năng mới.
* **Mahout:** đang ngày càng đánh mất thị phần của mình, và có thể sẽ bị thay thế bởi **Spark MLLib**

Hadoop đã có thời gian phát triển lâu dài và đã chứng minh sự hữu ích trong nhiều sản phẩm, vì thế có thể một số thành phần của hệ sinh thái Hadoop bị thay thế bởi Spark nhưng không thể nói Spark sẽ thay thế hoàn toàn được Hadoop. Với sự phát triển mạnh mẽ trong vài năm trở lại đây của Spark thì lập trình viên, các nhà khoa học máy tính có thêm công cụ hữu hiệu để phục vụ công việc của mình và người ta sẽ dần quên “Hadoop Stack” mà thay thế vào đó sẽ là “Big data Stack”, với nhiều sự lựa chọn hơn không chỉ là Hadoop.



# Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) dựa trên Google File System (GFS), cung cấp một hệ thống dữ liệu phân tán, được thiết kế để chạy trên các cụm máy tính lớn (gồm hàng ngàn máy tính) có khả năng chịu lỗi cao.

HDFS sử dụng kiến trúc master/slave, trong đó master gồm một NameNode để quản lý hệ thống file metadata và một hay nhiều slave DataNodes để lưu trữ dữ liệu thực tại.

Một tập tin với định dạng HDFS được chia thành nhiều block và những block này được lưu trữ trong một tập các DataNodes. NameNode định nghĩa ánh xạ từ các block đến các DataNode. Các DataNode điều hành các tác vụ đọc và ghi dữ liệu lên hệ thống file. Chúng cũng quản lý việc tạo, huỷ, và nhân rộng các block thông qua các chỉ thị từ NameNode.

HDFS cũng hỗ trợ các câu lệnh shell để tương tác với tập tin như các hệ thống file khác.

# Spack Streaming

