**Spark DataFrame**

1. **Spark SQL, DataFrames and Datasets**

* Spark SQL là một Spark module dùng cho xử lí dữ liệu đã cấu trúc. Không giống như basic Spark RDD API, những giao diện được cung cấp bởi Spark SQL thì cung cấp Spark với nhiều thông tin hơn về cấu trúc của cả dữ liệu và sự tính toán được sử dụng. Spark SQL dùng những thông tin mở rộng để thực hiện các tối ưu hóa bổ sung.
* Có nhiều cách để tương tác với Spark SQL bao gồm SQL và Dataset API. Khi tính toán một kết quả, một công cụ thực thi tương tự được sử dụng, không phụ thuộc vào API/ ngôn ngữ nào mà bạn đang sử dụng để biểu diễn sự tính toán. Sự thống nhất này có ý nghĩa rằng các lập trình viên có thể dễ dàng chuyển đổi qua lại giữa các API khác nhau dựa trên cung cấp cách tự nhiên nhất để thể hiện một chuyên đổi nhất định.

1. **SQL**

* Một tính năng của Spark SQL là thực thi các truy vấn SQL. Spark SQL cũng có thể được sử dụng để đọc dữ liệu từ một cài đặt Hive hiện có. Khi chạy SQL từ bên trong một ngôn ngữ lập trình thì kết quả trả về sẽ ở dạng Dataset/DataFrame. Bạn cũng có thể tương tác với giao diện SQL bằng cách sử dụng command-line hay thông qua JDBC/ODBC.

1. **Datasets and DataFrames**

* Một Dataset là một tập dữ liệu phân tán. Dataset là một kiểu mới được thêm vào trong Spark 1.6 nó cung cấp các lợi ích của RDDs với những lỡi ích của bộ công cụ tối ưu hóa thực thi của Spark SQL. Một Dataset có thể được xây dựng từ đối tượng JVM và sau đó được thao tác bằng cách sử dụng các phím biến đổi chức năng (map, flatMap, filter, etc.). Dataset API thì dùng được trong Scala và Java. Python không hỗ trợ Dataset API. Nhưng do sự linh hoạt của Python, nhiều lợi ích của Dataset API đã có sẵn. Tương tự với ngôn ngữ R.
* Một DataFrame là một Dataset được tổ chức vào những cột đã được đặt tên. Về mặt khái niệm, nó tương đương với một bảng trong cơ sở dữ liệu quan hệ hoặc một khung dữ liệu trong R/Python, nhưng với các tối ưu hóa phong phú hơn. DataFrames có thể được cấu trúc từ nhiều nguồn như: những tệp dữ liệu có cấu trúc, các bảng trong Hive, cơ sở dữ liệu bên ngoài hoặc RDD hiện có. DataFrame API sử dụng được trong Scala, Java, Python và R. Trong Scala và Java, một DataFrame thì được biểu diễn bởi các hàng của một tập Dataset. Trong Scala API, DataFrae đơn giản là một kiểu bí danh của Dataset[row]. Trong khi đó, ở Java API, người dùng cần sử dụng Dataset<row> để biểu diễn một DataFrame.