DEMO

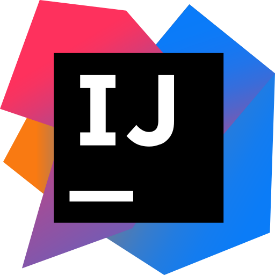
I. Mô tả sơ lược

- Ứng dụng demo Spark Structured Streaming với cửa sổ thời gian trượt là một ví dụ thực hành về cách sử dụng Apache Spark để xử lý dữ liệu streaming trong môi trường thời gian thực. Đoạn mã Scala triển khai một luồng xử lý dữ liệu từ Folder với tập dữ liệu file TXT hai cột chính là "Date" (kiểu thời gian) và "Message" (kiểu chuỗi). Ứng dụng này sử dụng cửa sổ thời gian trượt 30 giây ( tùy chỉnh) để tổng hợp và đếm số lượng tin nhắn trong mỗi cửa sổ.

Bằng cách đọc dữ liệu từ tệp Folder và xác định cấu trúc dữ liệu, ứng dụng tạo một temp view cho phép sử dụng SQL trực tiếp trên DataFrame. Các cửa sổ thời gian trượt và watermark được áp dụng để quản lý thời gian và xử lý dữ liệu trễ. Kết quả được xuất ra console và hiển thị số lượng tin nhắn được đếm trong từng cửa sổ thời gian.

Demo này mang lại cách Spark Structured Streaming có thể được tích hợp để tổng hợp và phân tích dữ liệu streaming, với khả năng quản lý dữ liệu không theo thứ tự một cách hiệu quả.

I. Cách thực hiện  
1. Download và setup IDE để lưu trữ và chạy demo



2. Tạo một object Scala

A screenshot of a computer program

Description automatically generated

3. Thêm Spark Dependency

- Mở tệp built.sbt trong dự án và thêm dependency cho Spark

- A screenshot of a computer program

Description automatically generated

4. Chuẩn bị dữ liệu test

- Tạo một kịch bản data là một chuỗi các tin nhắn được gửi đi theo thời gian, có hai cột là date( đúng định dạng yyyy-mm-dd hh:mm:ss) và message( tin nhắn)

A screenshot of a computer

Description automatically generated

5. Run

A screenshot of a computer

Description automatically generated

II. Kết quả

A screenshot of a computer

Description automatically generated

III. Giải thích

1. Import các thư viện cần thiết

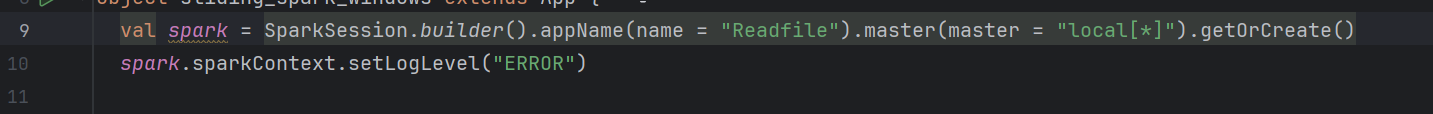
- Import các thư viện cần thiết từ Spark

A computer screen shot of a program

Description automatically generated

2. Tạo SparkSession

- Tạo một phiên SparkSession. setLogLevel("ERROR") được sử dụng để chỉ hiển thị log cấp độ lỗi để giảm sự nhiễu loạn.



3. Định nghĩa Schema cho dữ liệu

- Xác định schema cho dữ liệu streaming, với hai cột là "Date" có kiểu Timestamp và "Message" có kiểu StringType.

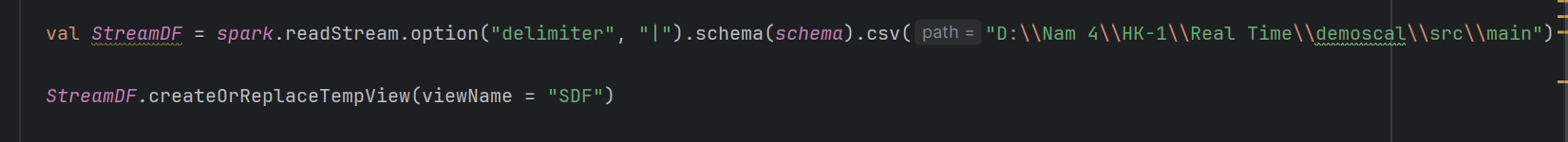
A screen shot of a computer code

Description automatically generated

4. Đọc dữ liệu và tạo View tạm thời

- Đọc dữ liệu streaming từ tệp TXT trong Folder với schema đã được định nghĩa. Cấu hình option("delimiter", "|") cho biết rằng dữ liệu trong tệp TXT trong Folder được phân tách bằng dấu "|".

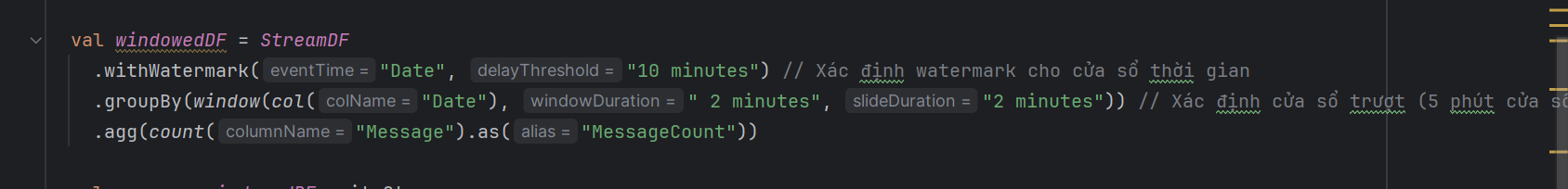
- Tạo một view tạm thời từ DataFrame để sử dụng với các truy vấn SQL



5. Xử lí dữ liệu với cửa sổ thời gian trượt

- Xác định Watermark cho cửa sổ thời gian giúp loại bỏ các sự kiện cũ hơn khỏi cửa sổ thời gian.

- Sử dụng cửa sổ thời gian trượt để tạo các cửa sổ thời gian 30 giây, với mỗi cửa sổ trượt mỗi 30 giây, và đếm số lượng tin nhắn trong mỗi cửa sổ thời gian này. Kết quả cuối cùng là một DataFrame mới có cột "MessageCount" thể hiện số lượng tin nhắn trong từng cửa sổ thời gian.

6. Xuất kết quả

A screenshot of a computer program

Description automatically generated

- Nếu tập dữ liệu mới có thời gian nằm trong khoảng watermark và thuộc về một cửa sổ thời gian mà ứng dụng đang theo dõi, Spark sẽ tính toán lại kết quả cho cửa sổ đó và xuất kết quả mới lên console.  
IV. Nhận xét