Apache Spark SQL - Data Sources

Spark SQL cung cấp nhiều cách để đọc dữ liệu từ các nguồn khác nhau.

Bao gồm:

- Đọc dữ liệu từ một tệp văn bản: bạn có thể đọc dữ liệu từ các tệp văn bản như CSV, TSV, và các tệp khác bằng cách sử dụng phương thức `read()` của đối tượng `SparkSession` và chỉ định định dạng tệp và các tùy chọn cấu hình khác (ví dụ: `spark.read.format("csv").option("header", "true").load("path/to/file.csv")`).
- Đọc dữ liệu từ một cơ sở dữ liệu quan hệ: bạn có thể đọc dữ liệu từ một cơ sở dữ liệu quan hệ như MySQL, PostgreSQL, Oracle, và SQL Server bằng cách sử dụng các thư viện JDBC và chỉ định địa chỉ JDBC của cơ sở dữ liệu, tên người dùng, mật khẩu và các tùy chọn cấu hình khác (ví dụ: `spark.read.format("jdbc").option("url", "jdbc:mysql://localhost:3306/mydatabase").option("user", "myuser").option("password", "mypassword").option("dbtable", "mytable").load()`).
- Đọc dữ liệu từ một tệp Parquet: bạn có thể đọc dữ liệu từ một tệp Parquet bằng cách sử dụng phương thức `read()` của đối tượng `SparkSession` và chỉ định đường dẫn đến tệp (ví dụ: `spark.read.parquet("path/to/file.parquet")`).
- Đọc dữ liệu từ một tệp JSON: bạn có thể đọc dữ liệu từ một tệp JSON bằng cách sử dụng phương thức `read()` của đối tượng `SparkSession` và chỉ định đường dẫn đến tệp (ví dụ: `spark.read.json("path/to/file.json")`).

- Đọc dữ liệu từ các nguồn dữ liệu khác: Spark SQL cũng hỗ trợ đọc dữ liệu từ nhiều nguồn dữ liệu khác nhau như Hive, Cassandra, Elasticsearch, và Kafka. Bạn có thể tìm hiểu thêm về cách đọc dữ liệu từ các nguồn này trong tài liệu chính thức của Spark SQL.

Sau khi đọc dữ liệu vào Spark SQL, bạn có thể sử dụng các API truy vấn và biến đổi dữ liệu của Spark SQL để truy vấn và xử lý dữ liệu.

Revision #2
Created 21 June 2023 16:23:22 by Laptrinh.vn
Updated 21 June 2023 16:24:27 by Laptrinh.vn