Requirements : Build a realtime / near realtime ETL pipeline from DL to DW using PySpark

DL : Cassandra

DW : MySQL

Core : ETL Pipeline from DL to DW using PySpark (1)

Add ons : Real time / near realtime (2)

1. 1. Extract : Đọc được data từ DL (Dùng pyspark connect tới Cassandra và đọc data)

2. Transform : Xử lý thành output mong muốn

3. Load : import dữ liệu sau khi xử lý xuống DW (Dùng pyspark import data xuống MySQL)

1. Thêm tính năng realtime / near realtime