Batch processing with pyspark on aws

DATA PIPELINE

Data được đưa vào S3 bucket. Sau đó sẽ chạy code PySpark trên EMR cluster. Code này sẽ lấy data từ S3 bucket sau đó thực hiện các tác vụ trên data. Sau khi data được xử lý xong sẽ được đẩy vào S3 bucket ở một thư mục khác. Sau đó sẽ sử dụng Athena để truy vấn data vừa xử lý xong này.

CÁC BƯỚC THỰC HIỆN

- Tạo EC2 key-pair
- Tạo AWS EMR cluster
- Tạo AWS S3 bucket và folder
- Upload data vào AWS S3
- Thực hiện các tác vụ trên data
- Phân tích dữ liệu với AWS Athena

KIẾN TRÚC