Build a Real-Time Spark Streaming Pipeline on AWS

DATA PIPE LINE

Amazon Lambda function sẽ stream log file vào Amazone Kinesis Data Streams. EMR sẽ chạy một spark job để đọc từ Kinesis Data Streams và load data ở định đạng cần thiết trong Kinesis Firehosue. Firehose sẽ thu thập dữ liệu đã chuyển đổi và ghi vào OpenSearch. Sau đó sẽ sử dụng Kibana để trực quan hóa dữ liệu.

CÁC BƯỚC THỰC HIỆN

- Tạo AWS S3 bucket
- Upload data vào AWS S3 bucket
- Tạo Amazon Kinesix Data Streams
- Tạo Lambda function
- Add trigger event vào Lambda function
- Tạo EC2 key pair
- Tạo Amazon EMR cluster
- Sending data từ AWS S3 đến Amazone Kinesis Data Streams bằng Lambda function
- Đọc data từ Amazone Kinesis Data Streams bằng EMR
- Tạo Amazon Kinesis Firehose delivery stream
- Writing data từ Amazon Kinesis Data Streams vào Amazon Kinesis Firehose bằng EMR
- Tạo OpenSearch domain
- Integrating OpenSearch with Amazon Kinesis Firehose delivery stream
- Tạo index pattern in OpenSearch
- Trực quan hóa dữ liệu bằng OpenSearch

KIẾN TRÚC