Industry Examples- Session - 04

Segment - 01

Session Introduction

SESSION OVERVIEW

- Kafka integration
- Reading tweets in real time

Segment - 02

Kafka Integration

KAFKA

- State of the art messaging system
- Highly scalable
- Follows the Pub Sub Model
- Integrates well with Spark Streaming
- Topic equivalent to tables in DB systems

LAB FLOW - 1

- Setup Kafka
- Create Topic
- Publish to Kafka Topic
- Setup Spark Job to read from Kafka Topic
- Execute

LAB FLOW - 2

- Read from Streaming file source – HDFS using Spark
- Publish to Kafka Topic
- Verify from kafka consumer console
- Execute

Coding Lab

Segment - 03

Read Tweets in Real Time

TWITTER STREAM API

- Java/ Scala – TwitterUtils
- Python – POST APIs over HTTPS
- Filter API
  - https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview
  - Location based filters
  - KeyWord based filters
- Allows Devs to create streams based on real time Tweets based on filters

TWITTER DEVELOPER APP

- Need to create an App
  - Setup Developer Account
  - Procure Access Keys to connect to Twitter Endpoints
- Code Flow
  - Create a request using Twitter Developer Credentials
  - Connect to Twitter Endpoint and send request
  - Send the Twitter Stream to a localhost Socket
  - Read the socket stream using Spark Structured Streams
  - Do Sentiment Analysis, GeoSpatial Event Analysis, Marketing Campaign Analytics or whatever you want with Twitter Data

Twitter Developer Account Setup

Coding Lab

Segment – 04
Session Summary

SUMMARY

- Industry use case – Kafka integration
- Industry use case – Reading tweets in real time

Segment – 05
Module Summary

STREAMING

- What is streaming data
- Differences and similarities with Batch processing
- Challenges of Stream Processing
- Streaming Data Architecture

STRUCTURED STREAMING

- How Spark Handles Streaming Data
- Different APIs
- Transformations and Actions
- Similarities with Static DataFrames
- Event Time Processing
- Handling Late Arriving Data

INDUSTRY USE CASES

- Kafka Integration with Spark
- Twitter Streaming Analytics

FURTHER READING

- Spark Docs
  - https://spark.apache.org/docs/2.2.0/structured-streaming-programming-guide.html
- Spark Streaming API Cookbook
  - https://spark.apache.org/docs/2.2.0/api/python/pyspark.sql.html#module-pyspark.sql.streaming
- Kafka Integration Guide
  - https://spark.apache.org/docs/2.2.0/structured-streaming-kafka-integration.html

Happy Learning