Team 7
Analyzing the relation between #music tweets and Spotify Global 50 playlist

-Vasudha Bangard

-Tanay Saxena

-Ebenezer Ajay Williams

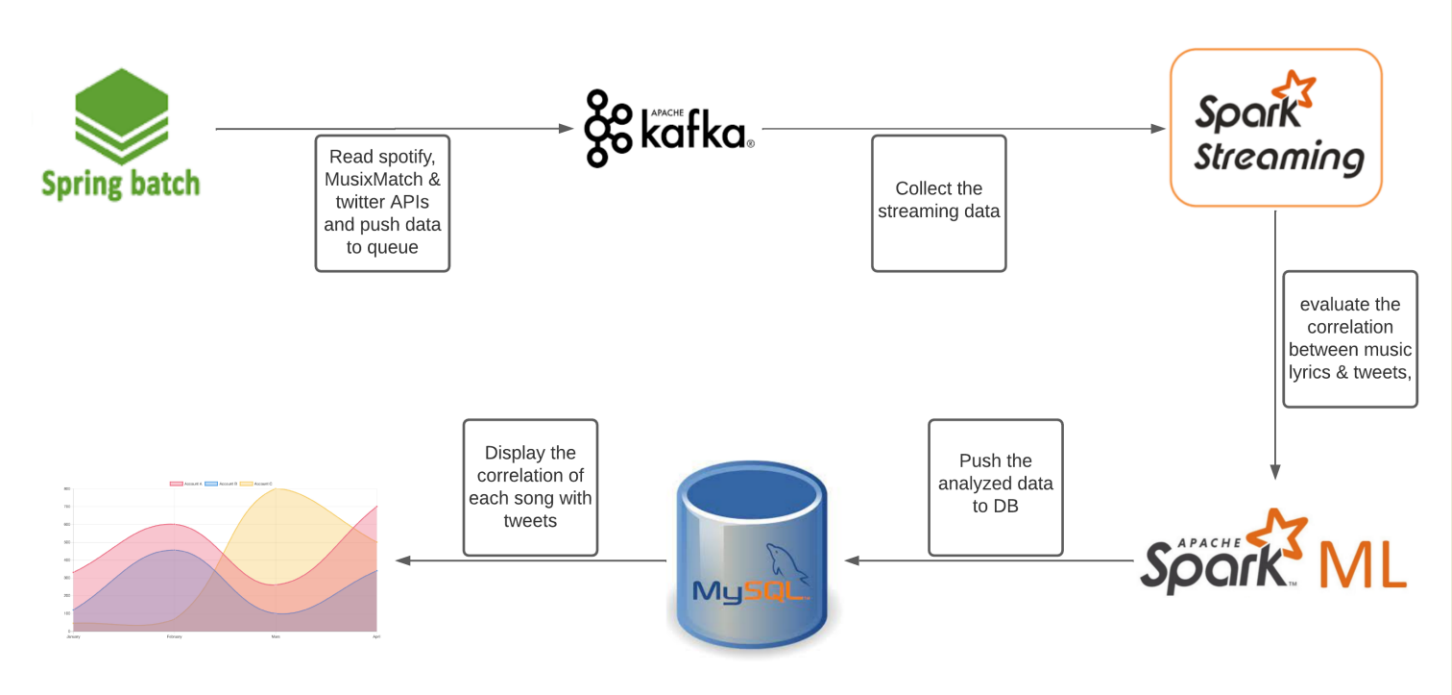
Northeastern University

USE CASES

- The System queries the Spotify developer API to get the list of top 50 songs played globally and the lyrics of those songs from the musixmatch API.
- The system also queries for tweets with the #music every minute.
- The model acts to find a correlation between each song and every tweet.
- A dashboard which shows the number of music tweets inspired for each song on the top 50 playlist is updated every minute. The dashboard also shows the relation between the sentiment of the tweets and the lyrics.
- The actor can view a real-time analysis of the tweets on each song on a dashboard updated every minute.

Northeastern University

Methodology



- Instead of Spring Batch we used Spring boot as a timed executor service as it seemed to be a simpler solution.

Northeastern University

Data sources

Twitter API

Spotify API

Musixmatch API

- Maintain a list of top 50 songs and their lyrics, refreshed as the trend changes
- Live stream of tweets with #music
- Kaggle tweets dataset (**Sentiment140**) for training ML model (160K tweets)
- NOTE: we refrained from training a model of our own and used the pretrained sentiment analysis model from core-NLP Stanford

Northeastern University

Milestones/sprints

Week1 (27 MARCH– 2 APRIL)	Week2 (3 APRIL – 9 APRIL)	Week3 (10 APRIL – 16 APRIL)	Week4 (17 APRIL – 23 APRIL)	Week5 (24 APRIL – 28 APRIL)
Spring Batch & Kafka	Spark Streaming	Spark ML	Database & JS chart	Buffer
Spring Boot & Kafka	Spark Streaming	Spark ML & Spark Streaming	Spark ML & Database	Database & JS chart

Northeastern University

What will you program in Scala ?

- **Spark Streaming:** collecting data from Kafka message queue
- **SPARK ML:** Applying ML model on the streamed data and calculate the feature vectors to be stored in the database
- JAVA: Spring Boot, Kafka
- Javascript: Chart.js

Northeastern University

Acceptance criteria

- Tweets with the exact lyrics must have cosine similarity 0.9
- Every tweet must be processed in real-time within a span of 2secs. When timed we see the time taken to process each tweet is around 2 secs.
- Sentiment analysis of the Tweets must have an accuracy of 60%. We used a pre trained model.

Northeastern University

Goals of the project

Establish a relationship between twitter feeds and top trending songs. (Done)

Find how the tweet sentiment is related to the sentiment of the lyrics. (We have instead computed the sentiment of the tweets and songs and displayed on the dashboard)

Challenges Faced

- The Spark NLP library from JohnSnowLabs had issues in compiling with the version of spark we used.
- Often the Musix match API did not have the song lyrics and so we have a had a few songs with empty lyrics
- Multiple data sources involve multiple Kafka topics which made the comparing of data to be difficult.
- Since the top rated songs do not get updated very frequently, We handle song duplication by setting the title as the primary key

Northeastern University