<음악 추천 시스템 아키텍처 설계 아이디어 – 60201561 김아영>

1. 데이터 수집 및 저장

- Spotipy 데이터 수집
- Spotify API 사용해서 데이터 수집
- Flume 설정 및 HDFS 저장
- Flume 통해서 Spotipy에서 수집한 데이터를 HDFS에 저장

Flume 사용 vs 사용 X, 직접 HDFS에 저장하는 방법

- 사용할 경우: 데이터 수집과 전송 자동화 가능, 스트리밍 데이터를 실시간으로 HDFS로 보내는 데 유용
- 사용 X: 데이터 실시간으로 수집 X, 수집한 데이터를 바로 HDFS에 저장

2. 데이터 처리 및 분석

Spark로 데이터 처리
HDFS 데이터를 Spark로 읽어와 필요한 데이터 전처리 수행
콘텐츠 기반 필터링 모델을 생성
Hive를 사용한 SQL 분석

3. 추천 시스템 개발

Spark MLlib에서 코사인 유사도를 계산하여 상위 10개의 유사 곡 추출.

pyspark.ml.feature.BucketedRandomProjectionLSH
-> 대규모 데이터에서 특정 데이터 포인트와 가장 유사한 데이터 검색
-> 근사 최근접 이웃 검색 방식, 코사인 유사도 계산

4단계: API 구현

Flask로 API를 구축해 사용자 요청에 따라 Spark 분석 결과 반환

input: 음악 제목 (+아티스트?)
output: 입력 받은 음악과 유사한 상위 10개 음악 제공