교육 동영상 공유 서비스의 카프카 기반 데이터 공유 방안

이현섭 · 김진덕*

동의대학교

A Kafka-based Data Sharing Method for Educational Video Services

Hyeon sup Lee · Jin-Deog Kim*

Dong-Eui University

E-mail : jdk@deu.ac.kr

요 약

대규모 운영시스템이나 확장성을 고려한 시스템을 구성할 때 마이크로서비스 기법을 도입하는 것이 필요하다. 카프카는 pub/sub 모델을 가지는 메시지 큐로서 분산환경에 잘 적용되는 특징을 가지며, 다양한 데이터 소스를 활용할 수 있다는 점에서 마이크로서비스에 적합하다.

이 논문에서는 아파치의 카프카를 이용한 교육동영상 공유 서비스의 데이터 공유 방안을 제안하고자 한다. 제안하는 시스템은 교육 동영상 공유서비스이 다양한 데이터를 공유하기 위해 카프카 클러스터를 구축하며, 아울러 교육동영상의 유사도를 기반으로 하는 추천 시스템과 연계하기 위해 스파크 클러스터를 이용한다. 그리고 파일, RDBMS의 DB등과 같은 다양한 데이터 소스를 공유하는 방안을 제시한다.

ABSTRACT

It is necessary to introduce micro-service techniques when constructing large-scale operating systems or systems that take into account scalability. Kafka is a message queue with the pub/sub model, which has features that are well applied to distributed environments and is also suitable for microservices in that it can utilize various data sources.

In this paper, we propose a data sharing method for educational video sharing services using Apache's Kafka. The proposed system builds a Kafka cluster for the educational video sharing service to share various data, and also uses a spark cluster to link with recommendation systems based on similarities in educational videos. We also present a way to share various data sources, such as files, various DBMS, etc.

키워드

Micro Service, Kafka, Video Service, Data Sharing

Ⅰ. 서 론

최근 하드웨어의 발달과 함께 소프트웨어의 규모와 데이터 처리의 복잡도는 나날이 증가하고 있다. 그리고 이러한 데이터는 집중되어 제어될 수 있을 때 그 가치를 창출할 수 있다. 만약 다양한 소프트웨어 모듈이 존재해도 , 데이터 처리 측면에서 서로 유기적으로 연계되어 있지 않다면, 해당 분야의 가치있는 정보를 제때에 제공하는 도구로서의 역할을 수행하지만, 더 이상의 역할을 기대할 수 없다. 그리고 서비스의 규모가 커질수록 그 복잡도가 증가하여 규모 확대의 걸림돌이다.

이러한 문제점을 해결하기 위해 마이크로서비스 아키텍처가 도입되고 있다. 이러한 서비스는 모든 세부 컴포넌트의 안정성과 신뢰성을 담보하는 운영이 중요해진다. 그리고 동적 확장이라는 마이크로서비스의 장점을 극대화하기 위해 전체 모듈간의 결합도를 줄이는 노력이 필요하다.

일반적으로 시스템의 결합도는 상당부분 데이터 송수신에 의해 좌우되며, 이에 대한 유연한 대처를 위해 단일 서비스, 분산서비스, 클러스터 기반 등 처리 환경을 고려해야 한다. 따라서 이러한 결합도를 낮추고 마이크로서비스가 가능한 데이터 공유 시스템이 필요하다.

Ⅱ. 데이터 공유 방안

2.1 시스템 구성도

* corresponding author

이 논문에서 타깃으로 삼고 있는 교육 동영상 공유 서비스 시스템[1] 또한 기본적인 추천 서비스를 위해 그림 1과 같이 실시간 처리모듈, 추천 연산 모듈, 동영상 분석 모듈 등으로 구성이 된다.

추천시스템은 협업 필터링을 위해 사용자 정보와 아이템 정보를 동시에 고려해야 한다. 그러므로 사용자 정보를 분석하기 위해 사용자의 로그 기록에 대한 배치 처리, 교육 동영상의 신규 입력 및 삭제 등에 대응하여 빅데이터 클러스터에서의 아이템 특성 분석 처리, 사용자의 동영상 시청에 대응해 유사 동영상 추천 목록에 대한 실시간 도출 처리와 같은 다양한 운영이 필요하다[2].

그리고 이 외에도 사용자 평점부여, 회원 가입, 사용자 클러스터링, 신규 동영상의 카테고리 선정 등 많은 작업이 추가적으로 필요하다[1].

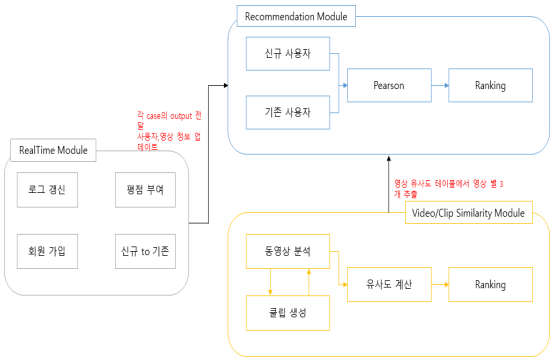

그림 1. 교육동영상 추천 시스템 구조도

따라서 강결합된 현재 시스템 구성을 감안하면 부분 모듈의 업그레이드가 타 시스템에 영향을 미치게 된다. 특히, 시스템의 중단이 허용되지 않는 상용 서비스일 경우 문제의 심각성이 증대된다.

2.2 카프카 기반 데이터 공유 시스템

카프카[5]는 마이크로서비스에 사용가능한 메시징 시스템이다. 카프카는 대용량 데이터 스트림 처리를 위해 메시지 버스에 최적화되어 있다.

제안하고자 하는 데이터 공유 시스템의 전체 아키텍처는 그림 2와 같다. 모든 데이터 공유를 위해 고가용성과 확장성의 Kafka Cluster를 이용한다.

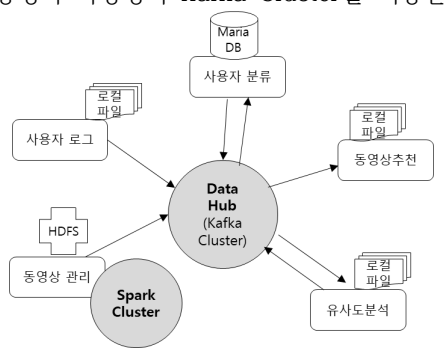

그림 2. 카프카기반 데이터 공유 시스템

제안한 데이터 공유 시스템은 다음과 같다.

(1) '사용자 로그' → '사용자 분류' : 사용자의 시스템 사용에 관한 로그 기록데이터가 사용자 분류 모듈에 전달된다. 이 때 mariaDB 플러그인이 Kafka Connect의 sink connector로 사용된다.

(2) '사용자 분류' → '동영상 추천' : 사용자의 로그 기록을 토대로 기계학습 병렬 분산 처리 프레임워크인 Spark MLlib를 이용하여 클러스터링을 수행한 결과 데이터를 협업 필터링의 일부데이터로 활용하는 '동영상 추천' 모듈로 전송한다.

(3) '동영상 관리' → '유사도분석' : 빅데이터처리 플랫폼이 스파크 클러스터에서 동영상의 자막 정보를 활용하여 형태소를 추출하고 빈도순으로 정렬한 데이터를 전송한다. 이 때 동영상 자막의 실시간 처리를 위해 실시간 병렬 분산 처리 프레임워크인 스파크 클러스터의 Structured Streaming 모듈과 연계[4]한다.

(4) '유사도분석' → '동영상' : 빈도순으로 정렬된 형태소를 활용하여 동영상의 빈도수 벡터 테이블과 유사도 테이블 데이터를 생성[3]하여 협업 필터링으로 활용하는 '동영상 추천' 모듈로 전송한다.

이렇게 생성된 동영상 추천 모듈은 사용자가 동영상을 시청하고 있는 동안 처리되어 사용자의 성향과 동영상 아이템의 특성을 동시에 고려한 협업 필터링을 통해 최적의 동영상을 추천하기 위한 리스트를 생성한다.

III. 결 론

카프카는 분산된 서버에서 대용량의 데이터를 다라는 분산 메시치 처리 시스템이다. 따라서 데이터를 받고, 그 데이터를 타 시스템에 중앙 통제된 방식으로 보내는 용도로 사용되며, 이는 카프카가 여러 시스템간의 연결 고리 역할을 하게 된다.

이 논문에서는 교육용 동영상 공유 서비스의 확장성과 무중단 서비스를 원활하게 하기 위한 마이크로서비스를 위해 카프카를 활용한 데이터 공유 시스템을 설계하였다. 설계한 시스템은 카프카 클러스터와 스파크 클러스터에서 동작하며, 로컬파일, HDFS, MariaDB에서 운용되는 각종 데이터를 원활히 관리할 수 있을 것으로 기대된다.

References

[1] Hyoun-Sup Lee, Jun-Ho Kim, Jae-Chul Lee, Bo-Ah Na, Jin-Deog Kim, "Design of Word and Stemming Extraction System for Keyword Analysis", Proceedings of the Korean Society for Information and Communication Sciences Conference23(2), 2019.10, 538-539

[2] Hyun-Sup Lee, Jindeog Kim, "A Design of Similar

Video Recommendation System using Extracted Words in Big Data Cluster" Journal of Academic Presentation of the Korean Society of Information Sciences, Vol. 24, No.2, (2020): 172-178.
[3] Minjae Kim, Sangjin Lee, "Measures of Abnormal User Activities in Online Comments Based on Cosine Similarity," Journal of KIISE, Vol.24, No.2 (2014):335-343.
[4] http://spark.apache.org
[5] http://kafka.apache.org