투빅스 6주차 정규세션 1교시 MLOps&Database 과제

24기 정서윤

시나리오: 당신은 수백만 명의 글로벌 사용자를 목표로 하는 소셜 미디어 스타트업의 데이터 엔지니어입니다. 이 서비스는 사용자의 '프로필 정보(ID, 이메일 등 정형 데이터)'와 '활동 로그(영상 시청 기록, '좋아요', 댓글 등 비정형 데이터)'를 모두 처리해야 합니다. 또한, 서비스가 갑자기 성장하더라도 안정적인 운영이 가능해야 하며, 수집된 데이터를 분석하여 사용자 맞춤형 콘텐츠 추천 모델을 개발해야 합니다.

문제: 위 시나리오를 바탕으로, 이 서비스에 필요한 데이터베이스 아키텍처를 설계하고 그 이유를 아래 요소들을 포함하여 종합적으로 서술하시오. (800자 이내)
1. 데이터베이스 유형 선택: 서비스의 각 기능(예: 사용자 프로필 관리, 활동 로그 수집)에 관계형 데이터베이스(RDB)와 비관계형 데이터베이스(NoSQL) 중 무엇을, 왜 사용해야 하는지 포함하라.
2. 시스템 환경 구성: 온프레미스(On-premise)가 아닌 클라우드(Cloud) 기반의 분산 시스템을 선택해야 하는 이유 2가지를 언급하고 간단히 설명하라.
3. 데이터 처리 시스템 분리: OLTP와 OLAP를 분리하여 구성해야 하는 이유를 설명하고, 이 두 시스템 간의 데이터 흐름(예: ETL)을 간략하게 제시하시오

1. 해당 서비스는 정형 데이터(사용자 프로필)와 비정형 데이터(활동 로그)를 동시에 처리해야 하므로, RDB와 NoSQL의 하이브리드 구조가 적합하다. 사용자 프로필 정보(ID, 이메일 등)는 스키마가 명확하고 무결성이 중요하므로 관계형 데이터베이스(RDB)를 사용하여 트랜잭션 일관성과 보안을 보장한다. 반면, 활동 로그나 영상 시청 기록은 구조가 다양하고 대량으로 축적되므로, 확장성과 유연성이 높은 NoSQL(예: MongoDB, Cassandra)을 사용하여 빠른 쓰기 및 조회 성능을 확보한다.

2. 서비스가 급격히 성장할 가능성이 높으므로 클라우드 기반 분산 시스템을 채택하는 것이 좋을 것이다. 클라우드는 오토스케일링(Auto-scaling)을 통해 트래픽 급증 시 자동으로 자원을 확장해 안정적인 서비스 운영이 가능하고, 글로벌 리전 분산을 통해 지역별 사용자의 latency를 최소화하고 장애 발생 시 빠른 복구를 지원한다.

3. 운영 트랜잭션(회원가입, 로그인 등)을 처리하는 OLTP 시스템과 데이터 분석 및 추천 모델 학습을 위한 OLAP 시스템을 분리해야 한다. 이는 분석 작업이 실시간 서비스 성능에 영향을 주지 않도록 하기 위함이다. 주기적으로 ETL 파이프라인을 구성해 OLTP의 데이터를 정제·적재하여 데이터 웨어하우스(OLAP)로 전달함으로써, 서비스 운영과 분석을 효율적으로 병행할 수 있다.