투빅스 DB&MLOps 과제

<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.

<답안>
1. 데이터베이스 유형 선택
SNS 서비스는 정형 데이터인 프로필 정보와 비정형 데이터인 활동 로그를 모두 처리해야 하므로, 데이터 특성에 따라 데이터베이스를 분리하는 것이 적절하다. 사용자 ID, 이메일 등 정형 데이터는 ACID 특성을 보장하는 관계형 데이터베이스(RDB)에 저장하고 영상 시청 기록, 좋아요, 댓글 등 비정형 데이터는 확장성이 높은 비관계형 데이터베이스(NoSQL)에 저장한다.

2. 시스템 환경 구성
시스템 환경은 온프레미스가 아닌 클라우드 기반의 분산 시스템을 선택해야 한다. 그 이유는 트래픽 급증 시 유연하게 자원을 늘릴 수 있어 서비스 안정성을 확보할 수 있고, 물리적 서버 설치 없이도 단기간 내에 환경 구축 및 변경이 가능해 높은 가용성과 운영 효율성을 제공하기 때문이다.

3. 데이터 처리 시스템 분리
안정적 서비스 운영과 효율적 분석을 위해 OLTP와 OLAP 시스템을 분리해야 한다. OLTP는 사용자 활동 데이터를 실시간으로 저장/처리하고, OLAP은 ETL 과정을 거쳐 수집된 대규모 데이터를 분석 및 추천 모델 학습에 활용한다. 이 구조를 통해 실시간 서비스 안정성을 유지하면서 분석 성능과 추천 모델용 데이터의 신뢰도를 동시에 확보할 수 있다.