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.

답 :
이 서비스의 데이터베이스 아키텍처는 관계형 DB와 비관계형 DB를 혼합한 구조로 설계하는 것이 적합합니다. 먼저, 사용자 프로필 정보는 스키마가 명확하고 일관성이 중요한 정형 데이터이므로 관계형 DB를 활용해야 합니다. 이를 통해서 트랜잭션 안정성과 데이터 무결성을 확보할 수 있습니다.

반면에 활동 로그는 초당 대량으로 발생하고 형태가 다양하며 스키마가 자주 변하므로 비관계형 DB가 적합합니다. 문서형 DB나 와이드컬럼 스토어를 이요하면 대규모 로그를 빠르게 기록/조회 할 수 있고 Redis 같은 인메모리 DB를 캐싱 계층에 두어서 성능을 보완할 수 있습니다.

또한 시스템은 온프레미스가 아니라 클라우드 기반 분산 환경을 선택하는 것이 유리합니다.
첫째, 클라우드는 탄력적인 확장성을 제공해서 사용자가 급격하게 늘어나도 자동으로 리소스를 확장할 수 있습니다.
둘째, 글로벌 가용성을 지원해서 여러 리전에 걸친 배포와 관리형 백업 및 재해 복구를 통해 안정적인 서비스 운영이 가능합니다.
이는 스타트업이 빠르게 성장하는 상황에서 비용과 관리 부담을 줄여줍니다.

마지막으로 운영 데이터 처리와 분석 처리는 반드시 분리해야합니다. 운영 데이터 처리 시스템은 사용자 로그인, 댓글 작성 등 실시간 트랜잭션을 안정적으로 처리하는 역할을 하고, 분석처리 시스템은 대규모 로그 데이터를 기반으로 추천 모델 학습이나 통계 분석을 담당합니다.
만약 두 시스템을 통합하면 분석 쿼리가 운영 성능에 영향을 주어서 사용자 경험을 저하시킬 수 있습니다.