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. (800 자 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 2 가지를 언급하고 간단히 설명하라.
3. 데이터 처리 시스템 분리: OLTP 와 OLAP 를 분리하여 구성해야 하는 이유를 설명하고, 이 두 시스템 간의 데이터 흐름(예: ETL)을 간략하게 제시하시오.

답변 :

이 서비스의 데이터베이스 아키텍처는 폴리글랏 전략을 기반으로 설계하는 것이 적합하다. 먼저 사용자 프로필, 인증, 결제와 같이 구조가 명확하고 강한 일관성을 요구하는 정형 데이터는 관계형 데이터베이스(RDB)를 활용한다. RDB 는 스키마 기반 관리, 참조 무결성, 트랜잭션(ACID)을 보장하므로 안정적인 사용자 정보 관리에 유리하다. 반면 영상 시청 기록, '좋아요', 댓글 등 비정형적이고 대량으로 발생하는 활동 로그는 NoSQL 을 사용한다. 문서형 또는 와이드컬럼 DB 는 스키마 유연성과 수평 확장성이 뛰어나 초당 수천~수만 건의 로그를 처리하는 데 적합하다. 또한 장기 보관 및 분석을 위해서는 오브젝트 스토리지에 원시 로그를 저장할 수 있다.

시스템 환경은 온프레미스가 아닌 클라우드 기반 분산 시스템이 적합하다. 첫째, 클라우드는 오토스케일링과 멀티 리전 지원을 통해 급격한 트래픽 증가나 장애에도 안정적인 서비스를 제공한다. 둘째, 데이터베이스, 스트리밍, 분석 도구를 관리형 서비스 형태로 활용할 수 있어 운영 부담을 줄이고, 초기 투자 비용 없이 사용량 기반 과금으로 비용 효율성을 확보할 수 있다.

마지막으로 OLTP(온라인 트랜잭션 처리)와 OLAP(온라인 분석 처리)는 분리해야 한다. OLTP 는 사용자 요청을 실시간으로 처리하는 데 최적화되어 있고, OLAP 은 대량의 데이터를 집계·분석하여 추천 모델을 학습하는 데 사용된다. 두 환경을 분리하지 않으면 운영 성능 저하가 발생할 수 있다. 데이터 흐름은 애플리케이션에서 발생한 트랜잭션 데이터와 로그가 OLTP 및 NoSQL 에 저장된 후, ETL 또는 스트리밍 파이프라인을 통해 데이터 레이크로 적재된다. 이후 데이터 웨어하우스에서 분석 및 모델 학습용으로 가공되어 추천 시스템에 반영된다.

이렇게 설계함으로써 서비스는 안정성·확장성·분석 효율성을 모두 갖춘 데이터 아키텍처를 확보할 수 있다.