6주차 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.

답변:
사용자 프로필·인증·권한처럼 정합성이 중요한 영역은 RDB가 적합합니다(ACID 트랜잭션, 스키마 기반, 조인으로 참조 무결성 유지). 반면 활동 로그·좋아요·시청 기록·알림 이벤트는 초당 대량 쓰기와 수평 확장이 핵심이므로 NoSQL(문서/키값/컬럼형)을 택합니다. 파티셔닝·TTL·세컨더리 인덱스로 저비용 대량 보관과 빠른 조회를 동시에 달성하고, 텍스트/댓글 검색은 검색엔진을 보조로 둘 수 있습니다.

클라우드 분산 환경을 선택하는 이유는 두 가지입니다. 첫째, 갑작스런 성장과 지역 확장에 오토스케일·멀티리전으로 탄력 대응해 지연과 혼잡을 줄입니다. 둘째, 관리형 백업·모니터링·보안·멀티AZ/DR로 고가용성과 운영 효율을 높여 인프라 총소유비용을 낮춥니다.

OLTP–OLAP 분리도 필수입니다. 사용자 요청을 처리하는 RDB/NoSQL(OLTP)은 짧은 지연과 높은 가용성이 목표이고, 분석·실험·모델학습(OLAP)은 대량 스캔·집계가 목표라 리소스 특성이 다릅니다. 데이터 흐름은 다음과 같습니다: 앱/서비스 → OLTP(RD

B·NoSQL)와 이벤트 스트림 → 데이터 레이크(원천 로그 적재) → ETL/ELT로 정제·적재 → DWH(집계)와 피처 스토어 생성 → 추천 모델 학습/평가 → 온라인 서빙 캐시(피드/랭킹)에 반영. 이렇게 하면 읽기/쓰기 경합을 줄이면서도 분석·개인화를 빠르게 반복할 수 있습니다.