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.

답안

1. 사용자의 프로필 정보는 정형 데이터로, 데이터 타입이 지정되고 명확한 스키마를 가지고 정확한 관계를 유지해야 하기 때문에 관계형 데이터 베이스(RDB)를 사용한다. RDB 는 기존의 스키마를 수정하기 어렵고, 복잡한 조인 연산 처리에 강하여 데이터의 일관성과 무결성을 보장할 수 있다.

반면, 활동 로그는 비정형 데이터로, 데이터의 구조가 정해져 있지 않으므로 비관계형 데이터베이스(NoSQL)를 사용한다. NoSQL 은 정해진 스키마가 없고, 유연한 데이터 구조를 제공하기 때문에 대규모의 로그 데이터를 효율적으로 저장하고 확장할 수 있다.

2. 클라우드(Cloud) 기반의 분산 시스템은 다수의 컴퓨터 노드에서 작업을 분산하여 처리한다. 또한, 필요에 따라 서버 자원을 확장할 수 있기 때문에 서비스가 갑자기 성장하더라도 안정적인 운영이 가능하다.

3. OLTP 시스템의 목적은 데이터베이스 트랜잭션을 처리하는 것으로, 개별 정보의 입력, 조회, 삭제, 수정이 효율적으로 이루어지도록 데이터가 정규화되어 저장된다. 따라서 실시간 트랜잭션 처리에 최적화된 시스템이다. 반면, OLAP 시스템은 대량의 데이터를 분석하여 의사결정을 지원하는 것으로, 보고서 및

계획 수립에 초점을 두며, 대량의 데이터에 다양한 패턴으로 접근하고 중복 데이터를 허용한다. 따라서 복잡한 질의와 통계적 분석을 담당한다.

이 두 시스템을 분리하여 구성해야 실시간 서비스의 성능이 분석 연산에 의해 저하되지 않고, 데이터는 ETL 을 통해 OLTP 에서 목적지(데이터 웨어하우스)에 전송되어 분석 및 모델 학습에 활용된다.