내부 로그 처리 및 분석 시스템 구조

Q1. 내부 로그 분석 파이프라인의 주요 목적은 무엇인가요?
A1. 셀/단말/서비스에서 발생하는 네트워크 로그와 KPI를 자동으로 수집, 정제, 저장한 후, 장애 감지, 품질 이상 탐지, 트렌드 분석, 예측 모델링 등에 활용하기 위함입니다.

Q2. 로그 수집 단계에서 주로 수집되는 데이터는 어떤 것이 있나요?
A2. OSS 로그 (e.g. PM, CM, FM), CDR/xDR, Probe 데이터, 단말 로그 (e.g. MDT), EMS 알람, Syslog 등 다양한 형식의 로그가 수집됩니다.

Q3. 로그 수집 방식에는 어떤 것들이 있나요?
A3.
- Push 방식: 장비에서 직접 로그를 전송 (예: SNMP, FTP, HTTP Post)
- Pull 방식: 중앙 수집기가 주기적으로 장비에 요청 (예: SFTP, REST API 등)
  → 상황에 따라 혼합 구성이 일반적입니다.

Q4. 로그 수집 주기는 어떻게 설정되나요?
A4. KPI/통계 로그는 일반적으로 5분 또는 15분 주기, CDR은 세션 종료 시점, 알람은 실시간(event-driven), MDT는 사용자 이동 기반으로 수집됩니다.

Q5. 수집된 로그는 어떤 절차로 정제되나요?
A5. 로그 포맷 정규화 → 불필요 필드 제거 → 결측값 처리 → 타입 변환 등의 과정이 있으며,
이 과정에서 데이터 스키마 일관성 확보가 중요합니다.

Q6. 로그 정제 과정에서 사용하는 도구는 무엇인가요?
A6. Logstash, NiFi, Fluentd, Python 스크립트, Kafka Connect 등
스트리밍/배치 기반으로 다양한 처리 도구가 사용됩니다.

Q7. 로그 저장소는 어떤 형태로 구성되나요?
A7.
- 원시 로그: HDFS, S3, Blob Storage
- 정제 로그: Parquet/ORC 포맷 + Hive/Presto 등
- 실시간 지표용: TSDB (InfluxDB), Elasticsearch, Druid 등
→ 사용 목적에 따라 다계층으로 구성됩니다.

Q8. 로그 분석은 어떤 프레임워크를 활용하나요?
A8. PySpark, SQL on Presto/Hive, Pandas, MLflow, Airflow 등의 빅데이터 및 ML 분석 도구를 조합해 사용합니다.

Q9. 실시간 분석은 어떻게 처리되나요?
A9. Kafka + Flink/Spark Streaming을 활용해 실시간 로그 스트림을 처리하며, KPI 임계치 초과, 이상 탐지 등을 실시간 대시보드나 알람으로 전송합니다.

Q10. 로그 파이프라인의 모니터링 방식은 어떻게 되나요?
A10. 로그 지연 시간, 수집률, 처리 성공률 등을 실시간 모니터링하며, Grafana, Prometheus 등을 활용해 시각화합니다.

Q11. 로그 파이프라인과 머신러닝은 어떻게 연동되나요?
A11. 수집된 로그는 정형화되어 Feature Store로 전환되며, 모델 학습용 데이터셋으로 가공되어 이상 탐지, 장애 예측, 사용자 행동 분석 등에 활용됩니다.

Q12. 로그 기반 알람 시스템은 어떻게 작동하나요?
A12. 정해진 KPI Threshold 또는 Rule을 기반으로, 이상 감지 시 Slack, Email, NMS에 자동 알람을 전송하는 구조입니다.
→ 일부는 ML 기반 이상 탐지 모델로 강화됩니다.

Q13. 파이프라인 유지 보수 시 고려해야 할 요소는 무엇인가요?
A13.
- 로그 포맷 변경 감지
- 스키마 자동화 관리
- 장비 추가/변경 대응
- 데이터 적재 지연/누락 감지
  → 운영 자동화 및 가시화 대시보드가 필요합니다.

Q14. 외부 시스템(RAG, 챗봇 등)에서 이 파이프라인을 어떻게 활용하나요?
A14. 정제된 로그 및 KPI를 기반으로 Embedding Index를 생성하거나,
QA 시스템의 Fact Retrieval 소스로 사용됩니다.
→ 특히 텍스트 기반 로그 (CDR, 알람 등)는 RAG의 Contextual Answering에 유리합니다.