실시간 Data Streaming 서비스

개요

최신의 기업 비즈니스는 IoT, 자율 주행, AI 등의 혁신 기술 분야에서 다양한 빅데이터를 실시간으로 생산하고 있습니다. 이들 빅데이터를 가치 있게 활용하려면, 실시간으로 데이터를 전송하고, 처리하고, 분석/의사결정하기 위한 데이터 파이프라인 서비스가 필수적입니다.

이 문서에서는 데이터 대량 배치 처리와 실시간 스트리밍 처리 방법을 모두 활용하여 대량의 데이터 처리에 효율적으로 대응할 수 있는 아키텍처를 설명합니다.

SDS Cloud에서는 데이터 스트리밍 실시간 처리/활용을 위한 **Kafka,** Spark, Hadoop, **Elasticsearch, Object Storage**, **DB Service**(RDB 및 NoSQL DB) 등의 다양한 서비스를 제공합니다.

아키텍처 다이어그램

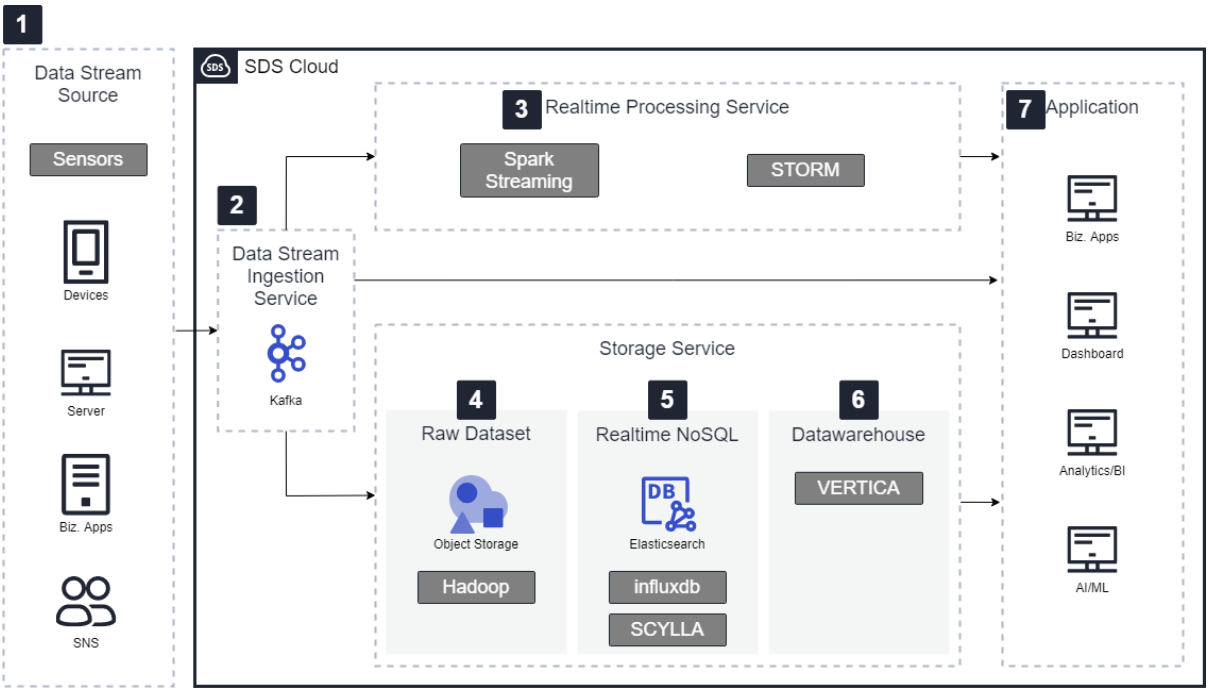

Figure 1. Kafka를 활용한 실시간 Data Streaming 서비스 구조

1. Data Stream Source
   대량의 Data Stream을 발생하는 원천으로는 모바일을 비롯한 IoT Device, 제조 설비의 Sensor, SNS Application, System Server의 Log와 같은 다양한 사례들이 있다.

2. Data Stream Ingestion
   대량의 Data Stream을 실시간으로 처리하기 위해서는 적정한 성능과(Latency, Throughput) 안정성, 확장성 등이 요구된다.
   Kafka는 오픈소스 Messaging 솔루션으로 확장성과 안정성이 매우 뛰어나며, 대량의 Data Streaming처리에 적합하다. **Kafka** 서비스는 다양한 Data connector를 제공하기 때문에 쉽게 다양한 원천 Data Source와 연결 할 수 있고, 하나의 Data Source로부터 수집된 Data를 여러 서비스에서 활용 가능하다.

3. Real-time Processing
   비즈니스 요건에 따라 Data Streaming 상태에서 실시간으로 데이터 값을 확인하여 처리가 필요할 수 있다. Spark Streaming과 Storm은 In-Memory 기반의 데이터 처리 오픈소스들로 실시간 처리에 적합한 솔루션이다.

4. Raw Dataset Storage
   저비용으로 확장(Scale Out) 가능하고, 안정적인 저장소로 **Object Storage, Hadoop(HDFS)**서비스를 제공한다. 이들 서비스들은 데이터 유실을 방지하도록 설계되어 데이터에 대한 높은 내구성을 제공한다. 데이터 저장 공간은 GB 단위에서 시작하여 PB 단위까지 사업 규모에 따라 확장할 수 있으므로, 비즈니스 민첩성과 유연성이 향상될 수 있다.

5. Real-time NoSQL
   Elasticsearch 서비스를 활용해서 다양한 형식의 로그 및 Metric을 수집하고, 저장(색인)할 수 있다. 특히 다양한 검색 기능(Full-text Search, 한글형태소분석)이 제공되어 까다로운 다양한 검색 조건을 신속하게 충족시킬 수 있으며, Kibana를 통해 실시간으로 데이터를 분석하고 시각화 할 수 있다.
   또한 시계열 데이터 처리에 특화된 Time series DB인 InfluxDB를 활용할 수도 있고, Cassandra와 호환성 있는 고성능 데이터처리 NoSQL인 ScyllaDB를 저장소로 활용할 수 있다.

6. Data Warehouse(DW)
   클라우드 환경에서 대용량의 정형 데이터를 분석하기 위한 서비스로 상용솔루션인 Vertica서비스를 제공한다. Native Column Store로 다른 DW DB에 비해 성능이 우수하며, In-DB Machine Learning 기능 제공으로 데이터의 이동 없이 데이터베이스에서 ML을 수행할 수 있다.

7. 활용 Application
   실시간 Data Stream, 또는 가공된 Batch Data를 최종적으로 활용하는 솔루션에는 실시간 Dashboard, Monitoring, Analytics, BI Tool, AI/ML 솔루션 등이 있다.

사용 사례

A. 제조업 설비관리 실시간 이상 모니터링

제조 현장의 설비에서는 다양한 대량의 센서 데이터가 발생합니다. 실시간으로 데이터를 끊임없이 수집하고 분석하여, 공정 제어에 필요한 적절한 정보를 언제 어디서든지 접근 가능하도록 하는 것이 중요합니다. 제조 공장 설비에 설치된 각종 센서에서 지속적으로 생성되는 데이터를 Kafka에서 수집 및 분류하고, Spark Streaming을 이용해서 실시간으로 처리해서 이상 알림을 사용자에게 실시간으로 전달할 수 있습니다. Raw data는 Elasticsearch에 저장해서 다양한 조건으로 검색 가능하고, 쉽게 데이터를 시각화해서 분석할 수 있습니다.

B. 이상 금융거래 실시간 탐지(Real-time Fraud Detection)

금융업무에서 소비자의 금융거래 패턴 등을 사전 파악해서, 일상 생활에서 해당 고객의 이상징후가 발견될 경우, 추가로 본인 확인 인증을 요청하거나, 고객에게 알림 정보 제공을 하는 등의 서비스가 필수적입니다. SDS Cloud에서 제공되는 실시간 Data Streaming 서비스들을 활용하여 이상 금융거래 탐지시스템(FDS: Fraud Detective System)을 손쉽게 구축할 수 있습니다.

선결 사항

Vertica 서비스 구축을 위해서는 고객 구매 라이선스 사용(BYOL)이 필요합니다.

제약 사항

없음

고려 사항

없음

관련 상품

- Object Storage
- DB Service (Elasticsearch)
- Kafka
- Vertica
- Cloud Hadoop ('21년 출시 예정)
- Timeseries DB (추후 제공 예정)

관련 문서

- GSLB를 활용한 서버 부하부산