빅데이터 수집을 위한 다채널 데이터 연계와 실시간 처리 시스템 설계

Multi-channel data connection and Real-time processing system designed for Big Data collection

백 경 석, 오 재 철*, 양 재 혁**
㈜아이온커뮤니케이션즈, ㈜아이온커뮤니케이션즈*,
㈜아이온커뮤니케이션즈**

Paik Kyoung-Seok, Oh Jae-Chel*, Yang Jae-Hyek**
I-ON Communications, I-ON Communications*,
I-ON Communications**

요약

빅데이터 분석을 통한 여러 산업 군과 융합으로 시너지를 발생시키기 위해서, 다양한 유형의 데이터 수집을 통해 빅데이터를 구성하는 것이 첫 번째 단계이며 기상, 교통, 인터넷 활동, 상권 등의 다양한 출처로부터 데이터 연계를 수행하고 사물인터넷과 같은 실시간으로 발생하는 로그 성 데이터 수집을 고려한 실시간 처리 시스템을 설계 하였다. 이를 통해 서로 다른 유형의 데이터가 빅데이터로 수집 되면 여러 산업 군에서 요구되는 인사이트 기반의 빅데이터 분석을 통해 B2B 또는 B2C 서비스에 응용 될 수 있다.

Ⅰ. 서론

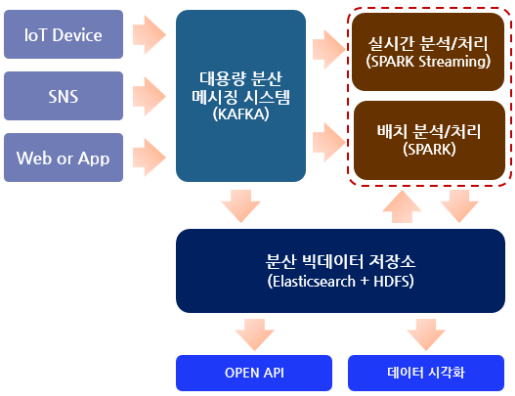
새로운 산업혁명의 핵심 기술로 부각되고 있는 빅데이터는 산업에 도움이 되는 인사이트를 찾기 위한 데이터로의 가치가 날로 상승하고 있다. 이러한 빅데이터를 근간으로 데이터 분석과 인공지능을 선두로 4차 산업혁명이 촉발 될 것으로 기대 된다. 특히, 클릭스트림, 소셜 네트워크, 사물인터넷을 통해 생성되는 데이터를 수집하여 빅데이터로 관리하고, 분석과 인공지능을 통해 여러 산업 군에 접목하는 기술은 앞으로 ICT 융합기술의 핵심 역할을 담당 할 것으로 판단된다. 이를 위해 다 채널로부터 대량의 데이터를 수집하고, 신속한 의사결정 기반의 비즈니스 모델을 처리 할 수 있도록 실시간 분석 및 처리가 가능한 시스템을 설계 하였다.

Ⅱ. 본론

1. 전체 시스템 개요

빅데이터 관련 시스템에서 가장 중요한 요소는 시스템의 확장성과 안정성 이다. 페타바이트(약 100만 GB) 급의 데이터를 저장 및 관리 하려면 데이터를 외부로부터 수집하는 파트와 저장하는 파트, 분석하는 파트 모두 분산 환경을 지원해야 한다. 이러한 조건을 만족함과 동시에 빅데이터 활용을 위한 Open API와 데이터 시각화를 제공하는 목표 시스템을 설계 하였다.

IoT 디바이스와 트위터와 페이스북과 같은 SNS, 웹 사이트나 앱에서의 클릭 스트림과 같은 대용량의 실시간 데이터 처리에 특화된 아키텍처로 구현된 Apache Kafka[1] 기반의 대용량 분산 메세징 시스템과 분산 환경에서 데이터를 실시간과 배치로 모두 처리 할 수 있는 Apache Spark[2] 기반의 데이터 분석 시스템, HDFS와 Elasticsearch 기반의 분산 데이터 저장소로 구성 하였다. 또한 데이터 처리 애플리케이션 개발을 위한 Open API와 데이터 분석 결과 및 데이터 확인을 위한 데이터 시각화를 제공 할 수 있도록 설계 하였다.



▶▶ 그림 1. 전체 시스템의 구성도

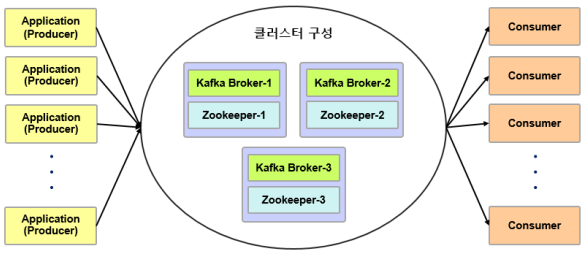
2. 시스템 구조 및 요소 기술

본 시스템의 구조는 대용량 분산 메시징 시스템, 분산 메모리 기반의 실시간 분석 및 처리와 배치 분석 및 처리 시스템, 분산 빅데이터 저장소와 분산 인덱스 기반의

관리 시스템, Open API 제공 시스템과 Kibana 기반의 데이터 시각화 시스템으로 구성 된다.

2.1 대용량 분산 메세징 시스템

Publisher-Subscribe 모델 기반으로 동작하는 메시징 시스템으로 Producer와 Consumer, Broker로 구성되어 있으며, Broker는 Zookeeper를 매개로 클러스터 구성이 가능하여 확장 및 장애 대응이 용이하다. 단순한 메시지 헤더를 지닌 TCP 기반의 프로토콜을 사용하여 오버헤드를 줄이고, 데이터 디스크를 활용하여 시스템의 안정성 향상과 처리량이 높은 구조로 되어 있으므로 대량의 데이터 수집에 적합하다.



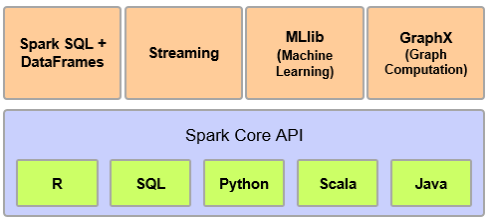
▶▶ 그림 2. 대용량 분산 메시징 시스템 구성

IoT 디바이스와 SNS, Web or APP이 그림 2의 Application(Producer)에 해당 하고 이를 통해 외부의 다채널 연계 포인트를 통해 데이터를 수집하고, Consumer에 해당하는 분석 시스템과 빅데이터 저장소에 실시간 분석이 필요한 데이터 스트림과 즉시 저장이 필요한 원본 데이터로 각각 전달 한다.

2.2 분산 메모리 기반의 분석 시스템

하둡의 MapReduce가 빅데이터 분석을 용이하게 해준 점이 있으나, 복잡하고 다단계 처리가 필요한 업무나 상호적이고 신속한 대응이 필요한 처리 작업에는 적합하지 않은 상황에서 분산 메모리 기반의 Apache Spark가 탄생 했다. 하둡에 비해 10배 이상 빠른 성능을 보여 주고 있다. 특히, 기계학습 알고리즘과 같이 반복 횟수가 많은 연산에서 더욱 강점을 보이고 있다.

Spark는 기존의 HDFS와 NoSQL DB 등 다양한 데이터 스토리지와 호환되고, R/SQL/Python/Scala/Java API를 지원하면서 메모리 기반 분산 환경을 제공 한다. 특히 스트리밍 데이터를 실시간으로 분석 및 처리 하는 기능과 머신러닝 라이브러리, 그래프 분석 작업을 위한 연산 집합 제공이 특징 이다.



▶▶그림 3. Spark Ecosystem

Spark Ecosystem에서 본 시스템에 주로 활용하게 되는 파트는 실시간 스트림 데이터 분석 및 처리를 위한 Streaming과 배치 데이터 분석을 위한 Spark SQL과 MLlib이 대상 이다. 실시간 분석과 배치 분석은 상호 보완적인 성격이 강하므로 유입되는 데이터 유형과 분석하고자 하는 비즈니스에 따라 데이터 연계를 적절하게 처리 한다.

표 1. MLlib에서 제공하는 알고리즘과 기능

제목	설명
기본 통계	요약통계, 상관관계, 커널밀도추정 등
분류 및 회귀	이항적분류, 다중분류, 회귀
협업 필터링	관심 및 선호도로 사용자 필터링
클러스터링	K-means, Gaussian mixture 등
특성 추출	Word2Vec, Normalizer 등
빈발패턴마이닝	데이터 집합에서 빈번한 패턴

2.3 분산 빅데이터 저장소와 관리 시스템

하둡 파일 분산 시스템(HDFS)을 빅데이터 저장소로 사용하고, 저장 된 데이터를 인덱싱하고 검색과 분석을 위해 Elasticsearch를 채용 하였다. 데이터의 모든 필드를 인덱싱 후에 Json 구조로 저장하므로, 데이터 접근이 쉽고 빠른 검색이 가능하도록 지원 한다. 또한, 데이터를 외부로 제공하는 Open API와 연계가 용이하고, 데이터 시각화 시스템 연계가 편리 하다.

Ⅲ. 결론

본 논문은 사물인터넷과 소셜 네트워크와 같은 실시간으로 대량의 데이터가 생성되는 산업에서 빅데이터로 관리하고 인사이트 기반의 데이터 분석을 통해 기존 산업에 새로운 사업 활성화를 창출 할 수 있는 시스템을 구현 하도록 설계 하였다. 향후 목표 시스템 구축으로 다양한 채널의 데이터를 수집하고, 전력 데이터와 비전력 데이터를 통합 관리 및 융합하여 전력 산업의 신규 비즈니스 창출을 목표로 하고 있다.

감사의 글

본 연구는 산업통상자원부 에너지기술개발사업 “10만 호급 전력 빅데이터 관리 체계 및 서비스 개발(과제번호: 20151210200240)” 1차년도 과제에 의해 이루어진 연구로서, 관계부처에 감사드립니다.

■ 참 고 문 헌 ■

[1] Apache Kafka, http://kafka.apache.org.
[2] Apache Spark, http://spark.apache.org