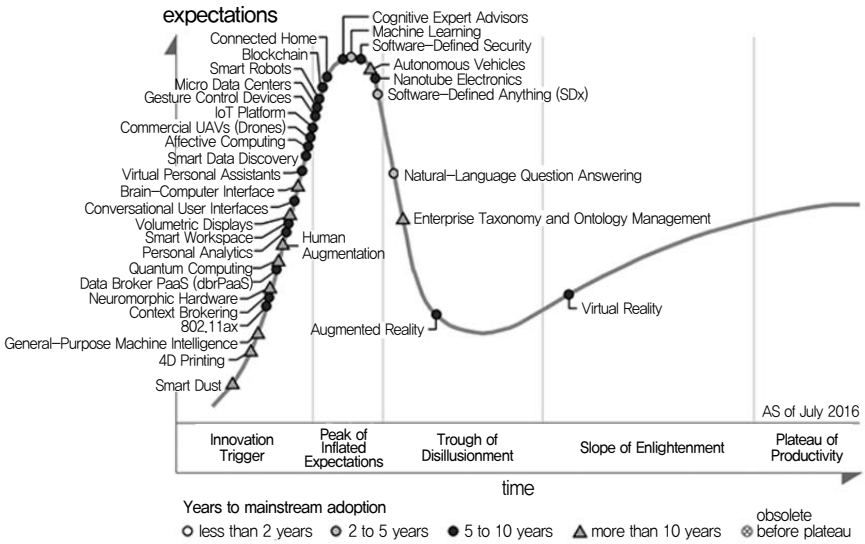
지능형 IoT를 위한 빅데이터 기술 현황

이연희 선임연구원, 유웅식·표철식 책임연구원
/ 한국전자통신연구원, KSB융합연구단
yeonhee@apache.org

서 론

지난해 알파고와 이세돌의 대결을 기점으로 자율주행 자동차, 인공지능비서 등 인공지능에 대한 관심이 한층 높아졌다. 이러한 흐름에 맞춰 IoT 시장에서도 인텔리전트 IoT라는 이름으로 농업, 제조, 에너지 등 다양한 산업 분야에서 모니터링, 판단 및 제어를 위한 지능적인 IoT 기술이 대두되고 있다.



[그림1. 2016년 가트너 '신기술 하이퍼 사이클'] (출처 : 2016년 가트너 '신기술 하이퍼 사이클' 보고서)

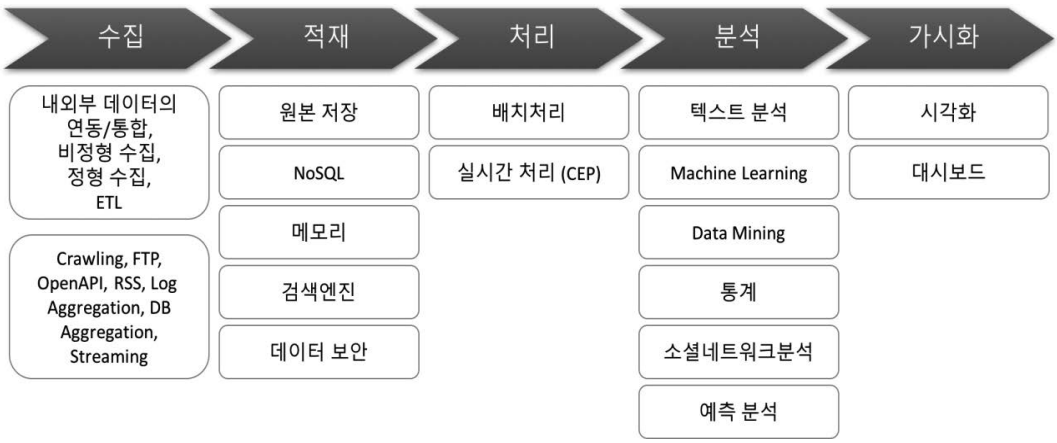
이러한 추세는 2016년 가트너의 '신기술 하이퍼 사이클' 보고서에도 그대로 드러나 있다. 하이퍼사이클 상의 머신러닝은 디지털 비즈니스 혁신을 위한 기술로서의 정점에 있으며, IoT 플랫폼 기술이 그 뒤를 따르고 있다. 빅데이터 기반의 처리 기술의 바탕 위에서 가장 대두되는 어플리케이션이 바로 애널리틱스이고, 그중에서도 딥러닝과 같은 기계학습 기반의 애널리틱스가 중요한 위치를 차지하고 있다. 그리고 이러한 기계학습 기반의 분석이 타켓으로 하고 있는 주요 서비스 분야가 바로 IoT일 것이다.

IoT 분야가 다른 분야의 인공지능 기술과의 가장 큰 차이는 데이터의 발생이 실시간 스트림의 특징을 가지고 있으며, 시간과 공간적인 특성을 가진다는 점이다. 따라서, 실시간 애널리틱스를 위한 스트림 처리 기술과 실시간 예측/분석을 위한 플랫폼 기술의 뒷받침이 필요하다. 이를 통해 궁극적으로 사물인터넷을 통해 감지되는 세상에 대한 인사이트를 빠르게 추출하여 비즈니스와 접목시키는 것이 가능하고, 이와 동시에 사물인터넷을 구성하는 사물들을 효율적으로 제어하기 위한 상위의 기술들을 빠르게 접목시킬 수 있다.

빅데이터 기술 개요

빅데이터 처리 기술은 데이터의 발생지로부터 데이터를 수집하여 원본 그대로 저장하거나 분석에 활용하기 위한 추출, 변환 후 분석이나 다양한 처리에 활용하기 적합하도록 저장소에 적재하는 기술, 수집하여 적재한 데이터를 실시간, 또는 배치 방식으로 처리하여 데이터로부터 인사이트를 추출하기 위한 분석, 그리고 빠르고 정확한 의사결정이 가능하도록 잘 표현하는 기술로 나뉜다. 빅데이터 기술 중 Flume은 로그나 센서 데이터를 수집하기 위한 가장 보편적인 기술이다.

빅데이터 적재 기술로서 Apache Kafka는 대용량의 스트림 데이터를 안정적으로 전달하기 위한 분산 Publish/Subscribe 구조를 갖는다. JDBC, Oracle GoldenGate, MQTT, HDFS, Elasticsearch, MongoDB Cassandra와 같은 다양한 소스와 싱크(Sink) 커넥터를 제공한다. 또 다른 카테고리의 적재 기술로서 대용량 분산 적재 기술인 NoSQL을 들 수 있다. NoSQL은 성능 요구사항과 용도에 따라 다양한 분류와 기술로 구분된다. NoSQL의 선택은 기본적으로 개발하고자 하는 어플리



[그림 2. 2016년 가트너 '신기술 하이퍼 사이클'] (출처 : 2016년 가트너 '신기술 하이퍼 사이클' 보고서)

수집 대상	기 술
DBMS 수집	Sqoop
로그/센서 수집	Flume, Scribe, Fluentd
FTP 수집	ftp
Http 수집	Scroller

[표 1. 빅데이터 수집기술]

케이션의 데이터가 어떤 형태와 요구사항을 가지느냐에 따라 Key/Value store, Column family store, Document store, Graph store로 구분된다. 또한 동일한 구분 내에서도 CAP Theorem 상의 CA나 CP 중 어느 요구사항에 중요성을 두느냐에 따라 달라져야 한다. [그림 3, 4 참고]

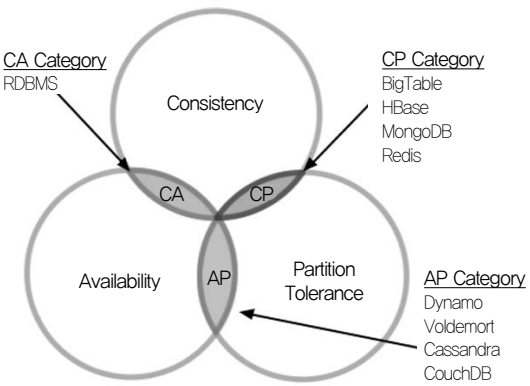
Apache Spark은 대표적인 데이터 처리 기술로서 대용량 배치처리부터, 실시간 스트림처리, 그래프 처리, 머신러닝 기반의 분석까지 지원하는 통합 빅데이터 처리 프레임워크이다.

모든 기술의 근간인 DataFrame API는 데이터의 구조를 테이블 형태로 일원화하여 복잡한 데이터의 처리를 테이블 형태의 데이터를 처리방식으로 추상화한 인터페이스를 제공한다. Spark의 실시간 처리 기술인 Spark stream은 BackType Storm과 비교될 수 있는데, 둘의 처리 모델은 근본적인 차이가 있다. Spark stream은 실시간 처리를 표방하지만, 내부적으로는 작은 미니배치 형태의 처리를 하는 반면, Storm은 스트림 소스로부터의 데이터 출현과 동시에 즉각적인 처리를 지원한다. 이외에도 작년 말 DataTorrent가 아파치재단에 기증한 Apache Apex가 대표적인 빅데이터 기반의 실시간/배치 처리 프레임워크로 볼 수 있다.

빅데이터 분석 및 시각화를 위한 오픈소스 기술을 몇 가지 꼽아보면, Elastic사의 주요한 비즈니스 솔루션으로 발전한 오픈소스 ELK 스택과 Apache Spark stream과 Zepplin, Freeboard.io를 이용한 IoT 분석 및 시각화 기술이 있다. 그중 ELK 스택은 Elasticsearch, Logstash, Kibana로 구성되며, 각각 분산 검색엔진, 로그인덱스, 시각화를 담당하여 다양한 유즈케이스에 쉽게 활용될 수 있도록 에코시스템을 제공한다.

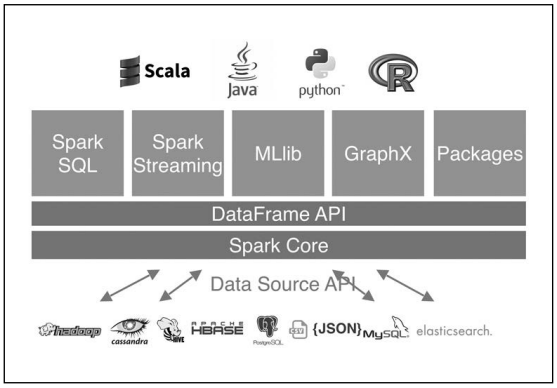
빅데이터 기술은 현재 다양한 유즈케이스를 만족하기 위해 필요한 요소 기술들이 각각의 레이어마다 제공된다. 따라서, 필요에 따라 적절히 기술을 선택하여 구성하는 것이 가능하다.

트위터 스트리밍 컴퓨팅 분야에서 근무하던 Nathan Marz는 일찌기 트위터의 실시간 분석에 대한 요구사항



[그림 3. CAP 이론]

[그림 4. NoSQL Databases]

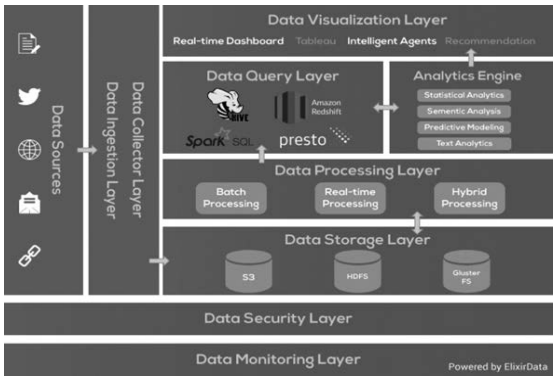


[그림 5] (출처 : Apache Spark)

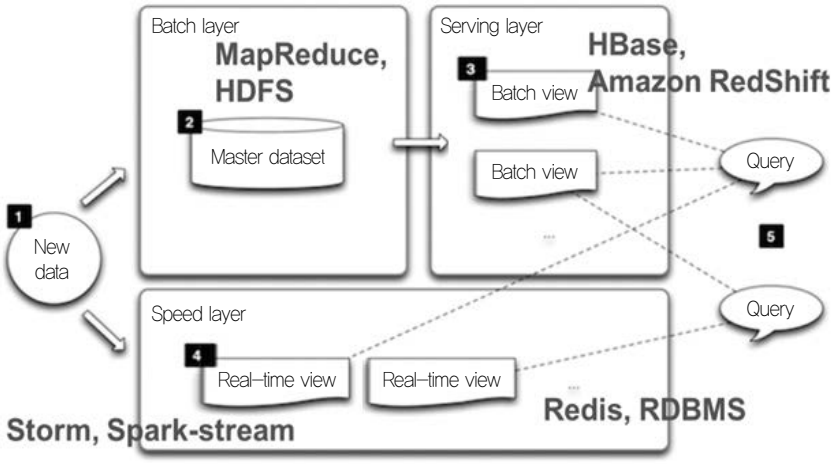
패키지	특 징
SparkCore	• 대용량 배치처리
Spark SQL	• 정형데이터에 대한 SQL 기반 분석
Spark Streaming	• 대용량 Fault-tolerant 스트림 처리
GraphX	• 그래프 병렬 처리를 위한 API
MLlib	• Spark 머신러닝 라이브러리
	• Transformer, Estimator Abstraction
	• Pipeline 기능 제공
DataFrame API	• 다양한 데이터소스 API 제공
	• HBase, JSON, CSV, Parquet, …

[표 2]

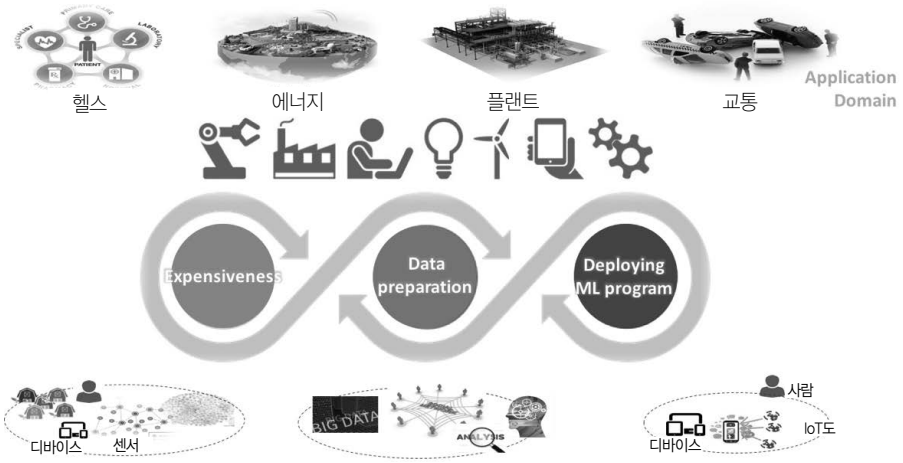
을 만족시키기 위해 대용량 배치처리와 실시간 처리 기술을 결합한 람다 아키텍처를 제안했다. 이는 일반적인 비즈니스에도 요구되는 시장의 수요와 반응을 실시간으로 비즈니스에 반영하기 위한 실시간 분석과 맞아 떨어진다. 또, 람다 아키텍처의 주요한 구성부는 Speed layer, Batch layer, Serving layer로 구성되며, 데이터의 유입과 동시에 실시간 처리와 주기적인 배치처리를 통해 동시에 배치 뷰와 실시간 뷰를 생성하여 사용자의 질의 요청 시에 이 두 뷰를 결합하여 응답함으로써 대용량 데이터에 대한 실시간 분석 서비스를 만족하도록 제공한다.



[그림 6. 빅데이터 기술 스택] (출처: ExlizirData)



[그림 7. 람다 아키텍처]



[그림 8. IoT 분야의 인공지능을 위한 도전과제]

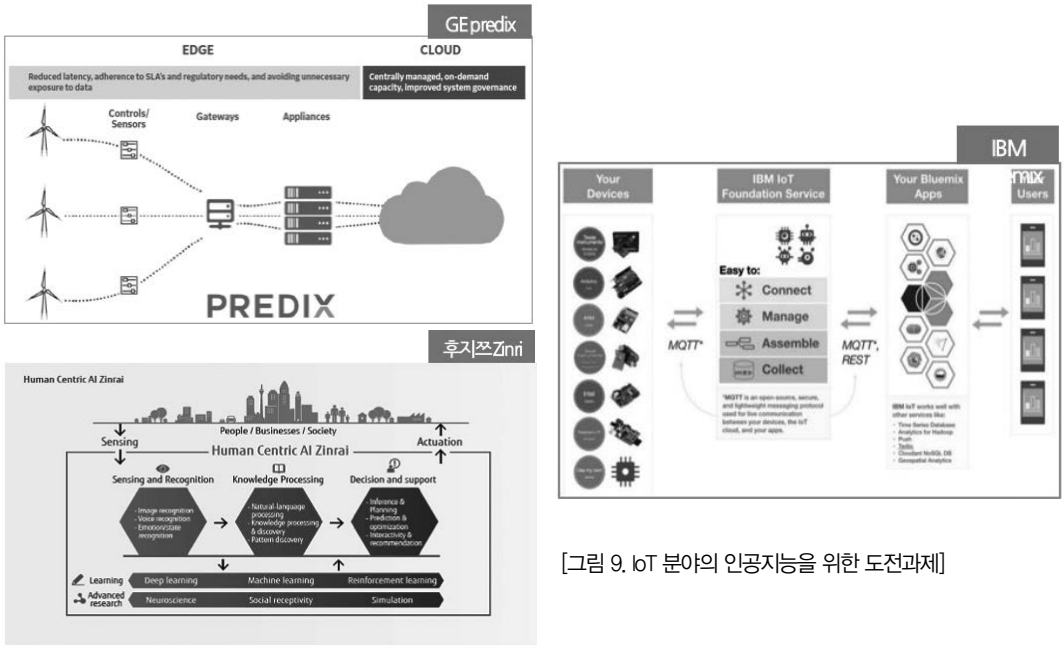
IoT 인공지능 기술의 도전과제와 사례

IoT 분야의 빅데이터에 대한 가장 중요한 요구사항 중의 하나는 시시각각 생성되는 데이터로부터 사용자의 트렌드 분석이나 장애의 판단 및 대응과 같은 적시성에 대한 요구이다. 이러한 영역의 데이터는 다양한 종류의 센서나 소스로부터 발생하므로 그 양과 형태가 다양하다. 또한, 대부분의 데이터 시계열 형태를 띠기 때문에 지속적인 관찰을 통해 예외사항이나 이벤트를 탐지하기 위한 적재나 처리, 분석 기술을 요한다.

최근 인공지능 기술이 급부상하면서 IoT 분야에도 단순한 처리/분석 기술에서 더 나아가 복잡하고 다양한 문제를 해결하기 위한 인공지능이 가미된 서비스를 지향한다. IoT 분야의 특성상 디바이스부터 생성되는 데이터까지 도메인마다 제각기 다른 형태를 띠며 다양한 어플리케이션이 요구되므로 각 어플리케이션별로 개별의 솔루션을 가지는 데는 한계가 있다. 따라서, 인공지능을 구성하기 위한 공통된 기술이나 프레임워크 기술이 절실하다.

일찍이 GE사는 자신들의 주 영역인 산업용 IoT 플랫폼 기술을 시작으로, 최초의 산업용 사물인터넷 PaaS 플랫폼인 Predix를 개발하여 서비스를 시작했다. 특히 이들은 Predix Edge, Data management, Analytics, Visualization, Security를 통합한 End-to-End 전주기를 지원하는 클라우드와 엣지를 통합한 산업인터넷 서비스를 주창한다.

IBM은 IBM Bluemix 클라우드 기술을 시작으로 IoT 플랫폼 서비스를 개발하고, 클라우드 서비스로 탑재하여 사물들에 대한 Predictive Analytics 제공한다. 이와는 달리 FUJITSU는 그래프 구조 데이터를 위한 새로운 기계학습 기술인 딥텐서(Deep Tensor)를 개발하여 지식처리를 중심으로 딥러닝과 기계학습, 강화학습을 통해 센싱된 정보로부터 제어를 추론하여 제공하는 클라우드 서비스를 시작했다. 이외에도 많은 기업들이 IoT를 위한 클라우드 기반 솔루션들에 집중하고 있어 클라우드 지능형 IoT 시장의 경쟁은 점점 가속화되고 있다.



[그림 9. IoT 분야의 인공지능을 위한 도전과제]

결론 및 시사점

Storm, Spark-stream, Samza와 같은 오픈소스 분산 병렬 기반 스트림 엔진들이 다양한 비즈니스에 활용되고, 오픈소스 기반 분산 스트림 처리 스택인 ELK 고도화한 Elastic 스택을 출시하여 머신러닝 기술과 접목하는 방향으로 실시간 대용량 처리 기술이 발전해 가고 있다. 대표적인 빅데이터 처리 엔진인 Apache Spark은 데이터와 처리에 대한 추상화 개념 도입하여 Spark MLib 기반 기계학습 파이프라인 기능을 탑재하여 처리를 고도화하고 있다.

한편, 구글 주도의 딥러닝 플랫폼인 Tensorflow는 딥러닝 플랫폼인 Tensorflow의 분산 버전을 출시하여 대용량 학습 기능을 스케일 아웃방식으로 해결하고 있다. 또한, CNN에 최적화된 Caffe 프레임워크는 페이스북의 지원을 받아 분산 딥러닝 기능을 탑재한 Caffe2를 내놓으면서 분산 딥러닝 플랫폼의 뒤를 잇고 있다.

또한 GE사의 최초의 산업용 사물인터넷인 Predix, IBM은 IBM Bluemix기반 IoT 플랫폼 서비스인 IBM Bluemix IoT, 후지쯔 진라이와 같은 지능형 IoT 클라우드 플랫폼 기술이 대두되고 있다.

실시간 대용량 데이터전처리를 위한 빅데이터 기술의 성숙과 다양한 기계학습/딥러닝 플랫폼 고도화 및 오픈소스화, IoT 를 위한 클라우드 기반 솔루션의 출시는 지능형 IoT 기술이 우리 눈앞에 다가왔음을 말한다. 이러한 시점에 인프라를 기반으로 어떤 지능적 모델을 창출하여사용자들에게 한발 더 다가선 IoT서비스를 제공할 것인가에 대한 고민과, 더불어 다양한 도메인의 비즈니스를 만족하는 vertical 솔루션을 적응적으로 생성할 수 있는 공통 프레임워크에 대한 기술개발이 절실한 시점이다.

〈사사 문구〉 본고는 2016년 정부(미래창조과학부)의 재원으로 과학기술연구회 융합연구단 사업(No. CRC-15-05-ETRI)의 지원을 받아 수행된 연구다.