

산업현장에서 요구하는 빅데이터 필요성과 데이터의 중요성

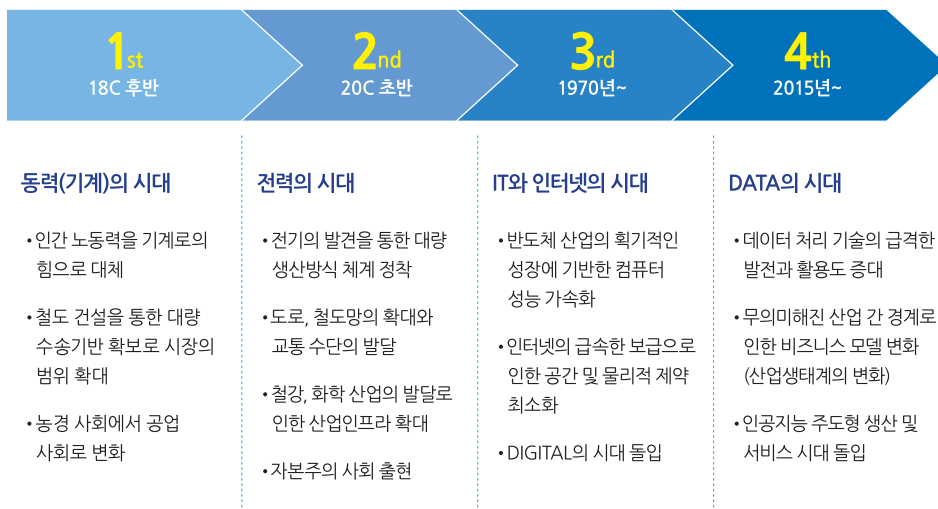


김재현_ 어니컴(주) 빅데이터사업부 부장

1. 머리말

2015년을 전후하여 4차 산업혁명이라는 정보통신기술(ICT) 융합의 새로운 혁명 시대가 등장한다. 이 시대의 핵심은 빅데이터(Big data), 인공지능(Artificial Intelligence), 클라우드(Cloud), 사물인터넷(Internet of Things), 모바일(Mobile) 등으로 분류

되며 이를 ICBMA라고 불리고 있다. 이중 빅데이터는 물리적, 생물학적, 디지털적인 관점에서 경제 및 산업 등 모든 분야에서 발생하는 데이터를 수집하고 분석하여 다양하게 활용하도록 하는 신기술이다. 이러한 이유로, 빅데이터는 4차 산업혁명을 견인하는 핵심 동력이라고 불리며 21세기의 원유로 비교되기도 한다.



[그림 1] 산업혁명의 흐름과 데이터 시대의 도래

지금까지 축적된 데이터를 원유로까지 비유하는 시대에 산업현장에서는 각 분야별로 보유하고 있는 데이터를 어떻게 활용해야 하는지 고민에 빠져 있다. 이는 데이터로부터 새로운 가치창조를 통해 내수시장에 국한되지 않고 글로벌 시장에서도 생존해야 하는 심각한 상황에 직면해 있기 때문이다. 스마트 기기의 확산과 각종 SNS의 활성화를 통한 비정형 데이터의 증가, 그리고 사물인터넷의 보급으로 실시간 수집되고 있는 방대한 데이터 등 2025년에는 전 세계 데이터 생산량이 약 163 제타바이트(ZB) 정도가 될 것으로 예측하고 있다. 참고로 1ZB는 1GB의 동영상 파일을 약 10억 개 정도 저장할 수 있는 공간이다. 이미 헬스케어 시장에서 IBM의 경우, 약 3억 명의 환자 데이터를 보유하고 있으며, 구글은 100만 명의 안구검사 기록을 확보하고 있다. 알리페이는 약 5억 명의 스마트폰 결제정보를 매초 2천 건씩 축적하고 있을 정도로 글로벌 기업들이 수집하고 분석하는 데이터량은 상상을 초월하고 있다. 본고에서는 산업현장에서

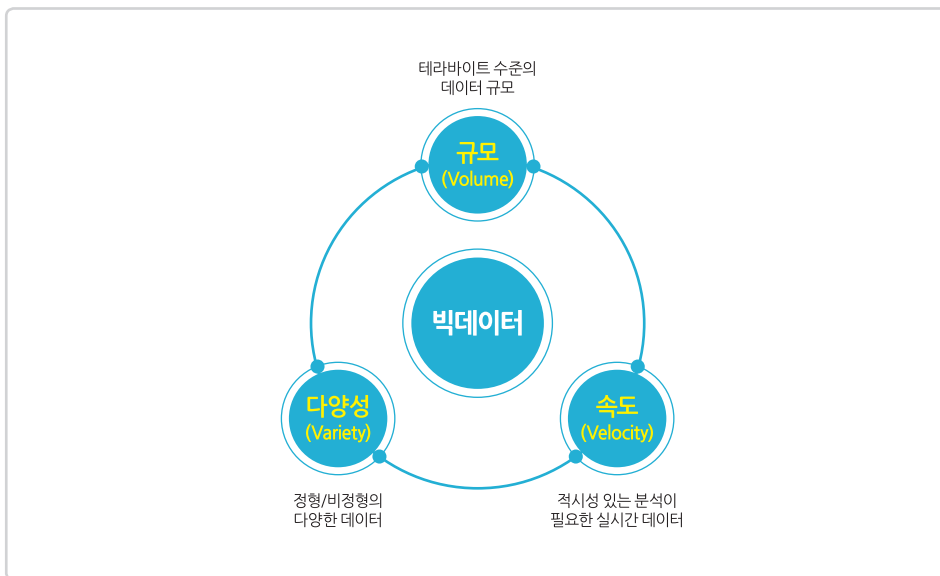
요구하고 있는 빅데이터 활용방안과 정부와 민간에서 진행되고 있는 각종 데이터 관련 사업들이 어떠한 방향으로 진행되고 있는지 전망하고자 한다.

2. 데이터의 가치와 처리 프로세스

2.1 기존 3V와 새로운 3V

빅데이터가 시장에 등장했을 때, 빅데이터의 공통적 속성을 테라바이트(TB) 수준의 대용량 규모(Volume), 정형/비정형의 다양한 데이터라는 다양성(Variety), 적시성 있는 분석이 필요한 실시간 데이터를 처리해야 하는 속도(Velocity)를 중시하며 3V로 규정했다.

여기에 증가하고 있는 방대한 데이터가 분석할 만한 가치가 있는지에 대한 정확성/타당성(Veracity)과 같은 데이터라 해도 사용자에게 따라서는 의미가 달라질 수 있기 때문에 가변성(Variability)이 추가되었다. 마지막으로 수집 및 분석을 통해 가공된 데이터



[그림 2] 빅데이터의 3V

※ 출처: <http://cfile2.uf.tistory.com/image/253E913351ECD08A35D6EC>

<표 1> 데이터 구분

정의	설명	비고
정형 (Structured)	고정된 필드에 저장된 데이터	관계형 데이터베이스 스프레드시트
반정형 (Semi-Structured)	고정된 필드에 저장되지 않지만 메타데이터나 스키마 등을 포함한 데이터	XML 및 HTML 데이터 등
비정형 (UnStructured)	고정된 필드에 저장되지 않은 데이터	텍스트 분석이 가능한 텍스트 문서 이미지/동영상/음성 데이터 등



[그림 3] 빅데이터 처리 프로세스

의 결과를 누구라도 쉽게 이해할 수 있도록 시각화 (Visualization)가 가능해야 한다. 일반적으로 데이터는 <표 1>과 같이 분류하고 있다.

2.2 빅데이터 처리 프로세스

데이터의 유입경로는 크게 내부와 외부로 나뉜다. 앞서 구분에서 이야기한 정형 데이터 대부분은 내부에서 생성되고 활용된다. 고객관계관리 (Customer Relationship Management), 기업자원관리 (Enterprise Resource Planning), 그룹웨어 (Groupware) 등에서 발생하는 데이터는 각 산업현

장의 특성에 맞게 설계되며 기존에는 내부 연동으로만 그친 데이터의 활용이 최근에는 외부와의 연동까지 고려하여 구축하고 있다. 데이터 소스는 대부분 수동으로 수집되나, 최근에는 센서 (Sensor)를 통하거나 사용자의 요구 및 기술 발전으로 로그 수집기 또는 크롤러 (Crawler) 등을 통해 자동 수집되고 있다. 수집된 데이터는 내부 및 외부에 저장되고 이를 실시간 또는 배치 형식으로 처리를 하게 된다. 전처리를 통한 정제된 데이터는 각종 분석기법을 통해 분석 후, 최종 결과물은 다양한 시각화를 통해 활용된다.

2.3 빅데이터 처리에 필요한 솔루션 및 활용방안

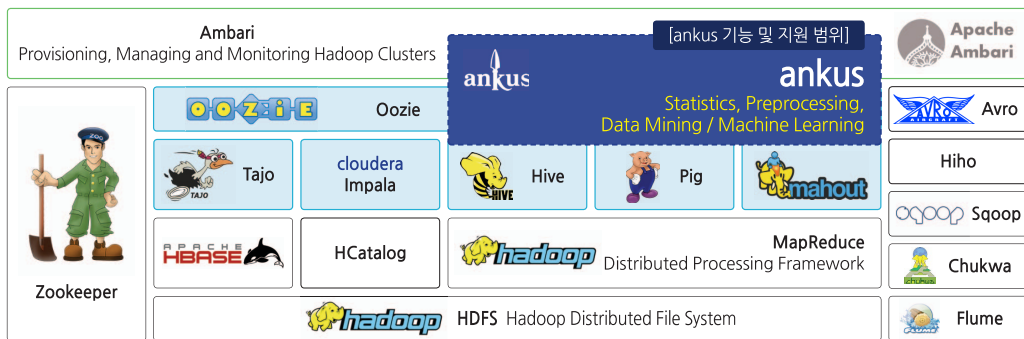
일반적으로 시장에서 유통되고 있는 소프트웨어는 상용소스 기반의 상용 소프트웨어(Commercial Software, 이하 CSS)와 오픈소스 기반의 오픈 소스 소프트웨어(Open Source Software, 이하 OSS)로 구분된다. OSS는 CSS에 비하여 저비용으로 특정 소프트웨어의 종속성에 구애받지 않고 개발 및 구축이 가능하다는 장점을 가지고 있다. 이미 글로벌 시장에서는 다양한 산업현장에서 OSS로 구축된 환경에서 제공되는 서비스를 이용하고 있으며, 국내도 연평균 약 12.3%라는 성장률을 보이고 있다. 빅데이터 처리

에 필요한 솔루션도 거의 대부분이 OSS를 중심으로 제공되고 있다. 국내에서는 OSS의 경우, “이해 및 구축이 어렵다”, “책임소지가 불분명하다”, “커스터마이징(Customizing)이 쉽지 않다” 등의 이유로 도입을 꺼려하는 경우가 있으나, 이는 OSS 커뮤니티를 통하여거나 OSS를 기반으로 개발하여 제공하는 기업(파트나 포함) 또는 단체와의 서비스(기술) 계약 등을 통해 충분한 대응이 가능하다.

빅데이터 처리에 필요한 솔루션과 각 프로세스 구분에 따른 처리내용은 <표 2>와 같다.

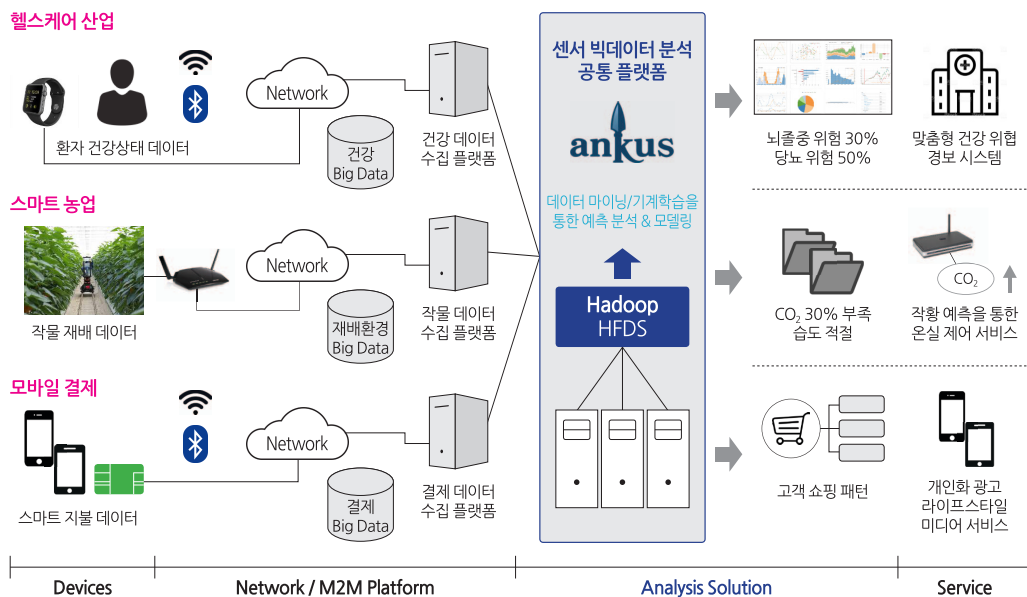
<표 2> 빅데이터 활용에 필요한 솔루션

구분	내용	솔루션
수집	<ul style="list-style-type: none"> · 외부 데이터(SNS, 뉴스, 웹) 검색/수집 및 인덱스 · RDBMS에서 관리되는 정형 데이터 수집 · XML, HTML 등 반정형 데이터 수집 · 파일 문서 등 비정형 데이터 수집 · 센서 데이터 / 실시간 스트림 데이터 수집 	Nutch, Lucene Sqoop, Flume, Chukwa, Kafka, Storm
저장	<ul style="list-style-type: none"> · 대규모 파일 데이터 분산 저장 · NoSQL(Key, Value 저장) · 데이터베이스 저장 	HDFS, Hbase, Cassandra, MongoDB
가공	<ul style="list-style-type: none"> · 대규모 데이터 분석을 위한 분산 처리 · 데이터 추출, 변환, 적재 · 로그 데이터 처리 	Sqoop, MapReduce, Spark, Flume, Hbase, Hive, Pig, Tajo
분석	<ul style="list-style-type: none"> · 전체 또는 부분 데이터에 대해 복잡하고 다양한 분석 · 데이터 수집과 동시에 분석 수행 · 대용량 실시간 스트리밍 분석 · 메모리 기반 데이터 분석 	R, Weka, MLlib, Mahout, anks Library, TensorFlow, Impala, Presto, Phoneix
시각화	<ul style="list-style-type: none"> · 데이터 분석 결과 시각화 · 관리현황 대쉬보드(모니터링) · 이미지 데이터 분석 결과 시각화 	D3, Zeppelin, Plotly, Tensorboard, N3N, Tableau(상용)
활용	<ul style="list-style-type: none"> · OpenAPI를 통해 데이터를 제공받아 서비스 적용 · 시각화면 연계 화면 페이지단위 외부 연계제공 	Open API(REST) In house 개발
운영	<ul style="list-style-type: none"> · 여러 단계에 걸쳐 처리되는 분석 작업의 흐름(Workflow)을 관리하고 정기/비정기적 분석 작업 스케줄링 · 보안 및 접근권한 관리 · 분산 서버 리소스 및 작업들을 관리하는 모니터링 	Oozie, Zookeeper, Hcatalog, Ranger, Ambari



※ 출처: ankus Community(<http://www.openankus.org>)

[그림 4] Apache EcoSystem

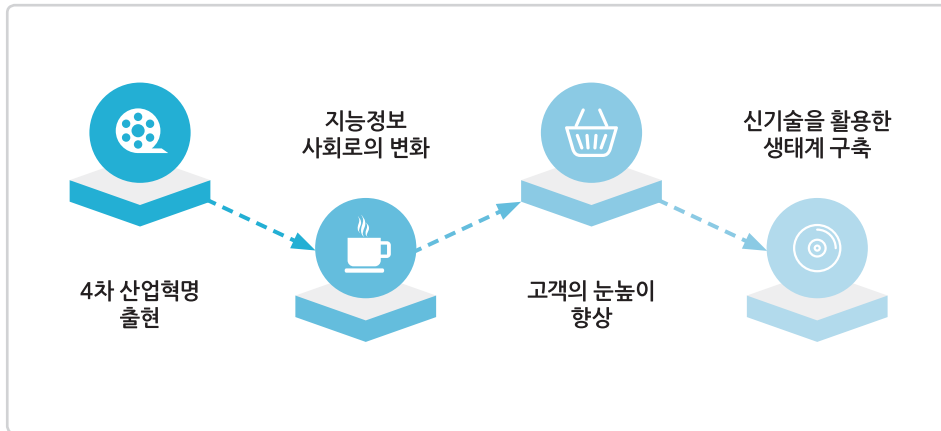


[그림 5] 빅데이터 구현방안 예시

[그림 4]의 아파치 에코시스템(Apache Eco-System)에서는 <표 2>의 솔루션들이 빅데이터 개발 및 구축분야에서 어떠한 위치를 차지하고 있는지 확인할 수 있다.

가령, 헬스케어 분야에서 발생하는 환자의 건강상태 데이터를 실시간과 배치로 수집/저장한 후 데이터 마이닝 및 기계학습을 통해 분석하여 환자의 발

병 위험도를 예측한다. 이를 통해 맞춤형 건강위험경보 시스템 등을 통해 이용자 및 의료기관에서 관리를 한다면 보건·복지와 관련한 산업현장에서는 향상된 의료서비스 제공이 가능해진다. 이러한 프로세스는 다양한 산업현장에서 해당 분야에 특화된 모델로 적용이 가능하다.



[그림 6] 4차 산업혁명 출현에 따른 시장의 변화

3. 시장의 변화와 대응방향

3.1 ‘스마트’라는 시장과 고객의 눈높이

4차 산업혁명이 출현하면서 가장 눈에 띄게 달라진 점은 고객의 눈높이이다. 지금까지는 내·외부 업무 프로세스에 가장 적합한 솔루션을 도입한 후, 해당 분야의 서비스 내지 제품을 소비자에게 제공하여 이익을 창출하면 되었다. 하지만 지능정보 사회로의 발전은 소비자의 눈높이뿐만이 아닌, 공급자의 능력에 대한 평가도 도마 위에 오르게 되었다. 더 이상 특정 분야에 국한된 기술과 정보만으로는 생존하기가 어렵다. 이러한 이유로 신기술을 활용한 생태계 구축이 항상 화두가 되고 있다.

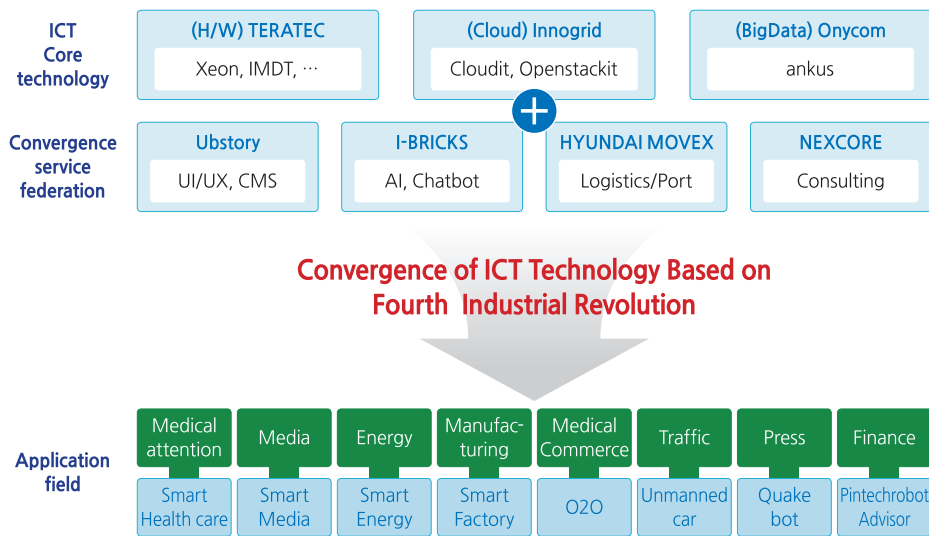
시장에서는 이미 ‘스마트시티’, ‘스마트홈’, ‘스마트팩토리’, ‘스마트팜’ 등 ‘스마트’로 시작되는 다양한 프로젝트가 진행되고 있다. 2019년 2월 부산에서 열린 ‘스마트시티 혁신전략 보고회’에서 문재인 대통령도 언급했듯이 교통, 치안, 재난 방지, 행정, 의료 등의 다양한 분야의 유기적이며 효율적인 연계를 위해서는 기존 기술의 업그레이드와 신기술의 융합이 선행되어야 하며, 데이터는 기본이다.

3.2 정보화사업 확대 및 민간 협의체의 움직임

2018년 12월을 전후하여 2022년까지 AI·빅데이터 정보화사업의 비중을 기존 21%에서 35%로 확대하기로 발표했다. 이는 데이터 기반의 정보화사업의 중요성을 정부가 심각하게 고려하고 있다는 증거이다. 국내 빅데이터 시장은 2014년부터 2016년까지 연평균 27.9%의 성장률을 기록했다. 해외 성장률의 경우, 2026년까지 총 922억 달러(한화 약 103조)로 예상하고 있다. 국내 빅데이터 시장의 세부성장률은 다음과 같다.

시장 구분	성장률
정부/공공시장	43.1%
기업시스템 투자	27.2%
분석서비스 시장	21.3%

민간에서도 이러한 흐름에 따라 글로벌 벤더사와 국내 우수 기술보유 회사의 솔루션을 융합하여 G-DataHub(가칭) 프로젝트(인텔코리아와 우수 중소기업 연합프로젝트)가 진행되고 있다. 대용량 데이터를 실시간으로 처리할 수 있는 하드웨어 기반에 인프라는 클라우드를 도입한다. 이 위에 AI기반의 빅데



[그림 7] G-DataHub(가칭) 프로젝트 구성안

이더 플랫폼을 올린 후, 다양한 산업분야에서 운용이 가능한 서비스 플랫폼을 구축할 예정이다.

4. 맺음말

빅데이터와 관련된 사업을 진행하면서 가장 힘든 점은 데이터 확보다. 이는 데이터를 제공해야 하는 입장에서는 민감한 정보가 포함된 데이터를 제공할 때 발생 가능한 정보유출의 두려움과 어떠한 데이터를 제공해야 만족할 수 있는 결과물을 얻을 수 있는지에 대한 불확실성 때문이라고 사료된다. 예로 콜센터를 운영하고 있는 기업에서 상담사가 제대로 고객 대응을 하고 있는지에 대한 평가를 기존 수동평가에서, AI기반의 빅데이터 기술과 음성을 텍스트로 전환(Speech-To-Text)하는 기술 등을 이용하여 자동 평가로 전환하고자 한다. 이 경우, 민감한 개인정보를 제외하고 화자구분을 통해 고객의 반응과 상담사의 대응만을 추출한 텍스트 파일만을 저장한다. 다음은 내부평가에서 사용한 지표를 분석모델로 적용

후, 실시간으로 저장되는 데이터와 기존 평가결과를 반복적으로 학습한다. 상담사는 실시간으로 본인이 대응한 이력데이터를 참고하여 서비스 향상에 노력하며, 관리자는 각 상담사의 평가데이터를 토대로 지도한다. 이와 같이 산업현장에서는 기존 프로세스와 신기술을 융합한 업그레이드 된 서비스를 요구하고 있으며, 데이터는 가장 기본적인 단계에서 취급되고 있다. **TTA**

[참고문헌]

- [1] The Digitization of the World From Edge to Core(2018.11., IDC - Seagate)
- [2] 스마트과학과 과학학습콘텐츠 빅데이터의 속성 3V, 4V
- [3] 'Big Data 구축기술과 사례를 중심으로' 재구성(문혜정, 2012.)
- [4] 인공지능·빅데이터·클라우드 정보화사업 비중 35%로 확대 (2019.02.19., 디지털타임스)
- [5] 2019년 공공부문 SW·ICT장비·정보보호 수요예보(예정) (2018.11.29., SWIT)
- [6] 2016년 빅데이터 시장현황조사(한국정보화진흥원)