SOFTWARE PLATFORM

(Big Data)

KYUNGSUNG UNIVERSITY SINCE 1955

빅데이터(Big Data)

디지털 환경에서 생성되는 데이터로 그 규모가 방대하고, 생성 주기도 짧고, 형태도 수치 데이터 뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터 - 네이버 지식백과

기존 데이터보다 너무 방대하여 기존의 방법이나 도구로 수집/저장/분석 등이 어려운 정형 및 비정형 데이터들 - 국립중앙과학관

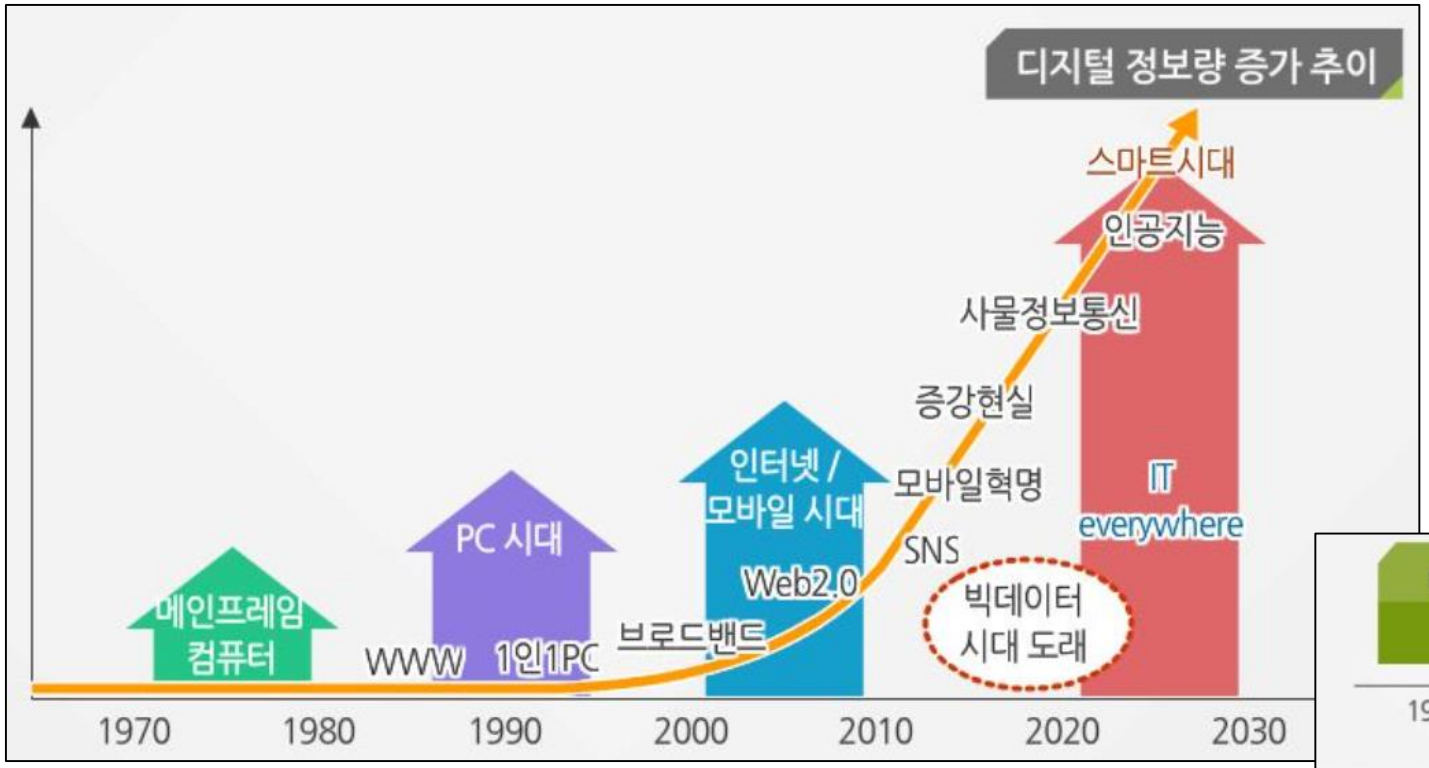
기존 데이터베이스 관리도구의 능력을 넘어서는 대량(수십 테라바이트)의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터 집합조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술 – 위키백과

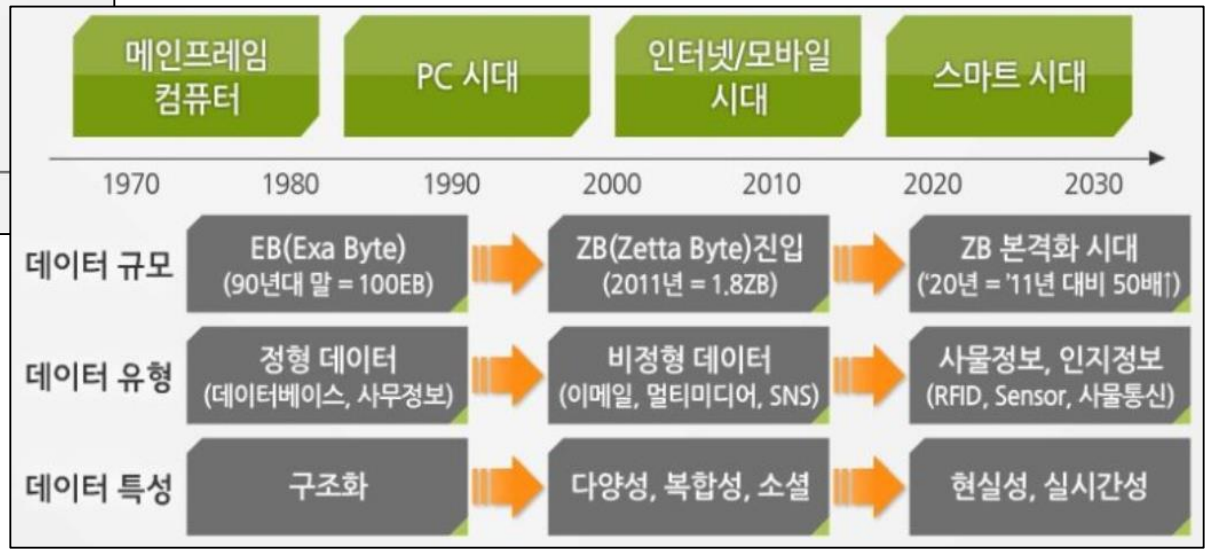
Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software - Wikipedia

Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications. - SearchDataManagement

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. - Gartner

빅데이터 시대의 도래


디지털 정보량 증가 추이
스마트시대
인공지능
사물정보통신
증강현실
모바일혁명
SNS
Web2.0
브로드밴드
WWW
1인1PC
메인프레임
컴퓨터
PC 시대
인터넷 /
모바일 시대
IT
everywhere
빅데이터
시대 도래
1970
1980
1990
2000
2010
2020
2030


메인프레임
컴퓨터
PC 시대
인터넷/모바일
시대
스마트 시대
1970
1980
1990
2000
2010
2020
2030
데이터 규모
EB(Exa Byte)
(90년대 말 = 100EB)
ZB(Zetta Byte)진입
(2011년 = 1.8ZB)
ZB 본격화 시대
('20년 = '11년 대비 50배↑)
데이터 유형
정형 데이터
(데이터베이스, 사무정보)
비정형 데이터
(이메일, 멀티미디어, SNS)
사물정보, 인지정보
(RFID, Sensor, 사물통신)
데이터 특성
구조화
다양성, 복합성, 소셜
현실성, 실시간성

빅데이터 활용 사회 구축 사례

1. **빅데이터 활용을 통한 행복한 사회 구축 사례**
   - ➢ **미국 국세청의 탈세 방지 시스템 개발을 통한 국가 재정 강화**
     - ✓ **탈세 및 사기 방지 시스템 - 범죄 네트워크 발굴**
     - ✓ **지능형 감시 시스템 구축**

2. **빅데이터 활용을 통한 건강한 사회 구축 사례**
   - ➢ 미국 국립보건원의 유전자 데이터 공유를 통한 질병치료체계 마련
     - ✓ 인종별, 국가별 특성에 따른 유전자 특성 정보를 분석하여 질병 치료에 활용하고자 함
     - ✓ 인간 유전체의 다양성을 알아내고자 함

3. **빅데이터 활용을 통한 안전한 사회 구축 사례**
   - ➢ 싱가포르의 국가 위험 관리 시스템을 통한 국가안전관리
     - ✓ 인종별, 국가별 특성에 따른 유전자 특성 정보를 분석하여 질병 치료에 활용하고자 함
     - ✓ 인간 유전체의 다양성을 알아내고자 함

4. **빅데이터 활용을 통한 창의적 사회 구축 사례**
   - ➢ 영국 패치베이의 국민참여형 안전관리 창의 플랫폼 구현
     - ✓ 개방된 소스로서 재난 안전관리 시스템의 상호 연계를 지원함
     - ✓ 공유 데이터를 기반으로 웹 프로그램, 스마트폰 앱 개발 등에 응용 및 활용함

행복한 사회 구축	
미국 국세청	탈세 방지 시스템을 통한 국가 재정 강화
일본	센서 데이터를 활용한 지능형 교통 안내 시스템
이탈리아 밀라노	지능형 교통정보 시스템으로 길안내 서비스 제공
뉴욕주 시라큐스시	데이터 분석 기반으로 스마트 시티 추진
덴마크 베스타드 윈디 시스템	풍력 에너지 관리로 에너지 생산 효과 극대화
구글	실시간 자동 번역 시스템을 통한 의사소통 불편 해소
월마트	데이터 분석을 통한 투자 수익 증대
자라	상품별 데이터 분석을 통한 판매량 증대
마이크론 테크놀로지	제품 생산 시간 분석을 통한 비용 절감
코카콜라	SNS 데이터 활용 을 통한 제품 판매 의사결정 반영
리츠 칼튼 호텔	데이터 관리를 통한 고객맞춤형 서비스 제공
할리우드 영화 시장	SNS 분석을 통해 흥행 수익 예측
넷플릭스	데이터 분석으로 온라인 DVD 판매제고 및 고객 서비스 향상

건강한 사회 구축	
미국 국립보건원(NIH)	유전자 데이터 공유를 통한 질병 치료제 개발 Pillbox 프로젝트를 통한 의료 개혁
미국	미국 퇴역군인의 전자의료기록 분석을 통한 맞춤형 의료 서비스 지원
싱가포르	주민위원회 센터 네트워크를 기반으로 맞춤형 복지사회 구현
캐나다 온타리오 공과대병원	미숙아 모니터링을 통한 감염 예방 및 예측
건강보험회사 웰포인트	슈퍼컴퓨터를 활용한 효율적인 환자치료
구글	검색어 분석을 통한 독감예보 서비스 제공
네덜란드 스파크드	빅데이터를 활용하여 건강한 소 사육 환경 구축

안전한 사회 구축	
싱가포르	국가 위험관리 시스템을 통한 국가안전관리
FBI	유전자 색인 시스템을 활용한 단시간 범인 검거 체계 마련
샌프란시스코	범죄예방 시스템으로 안전 지역사회 구축
싱가포르 출입국관리소	통합적 정보분석으로 출입국 보안 및 국경 통제 강화
일본	다양한 센서 데이터를 활용한 재난대응 능력 강화

창의적 사회 구축	
미국 미시간 주	데이터웨어하우스 구축으로 공공서비스 질적 향상
영국 패치베이	국민참여형 안전관리 플랫폼 구현
케냐 우샤히디	집단지성으로 이루어진 재난관리 오픈소스 플랫폼
IBM 왓슨	인공지능 슈퍼컴퓨터로 인류의 창조성과 혁신 촉진
애플 시리	지능형 음성인식을 통해 더 똑똑해지는 창의적 사고 가능

Big Data 3Vs

디지털 데이터 단위

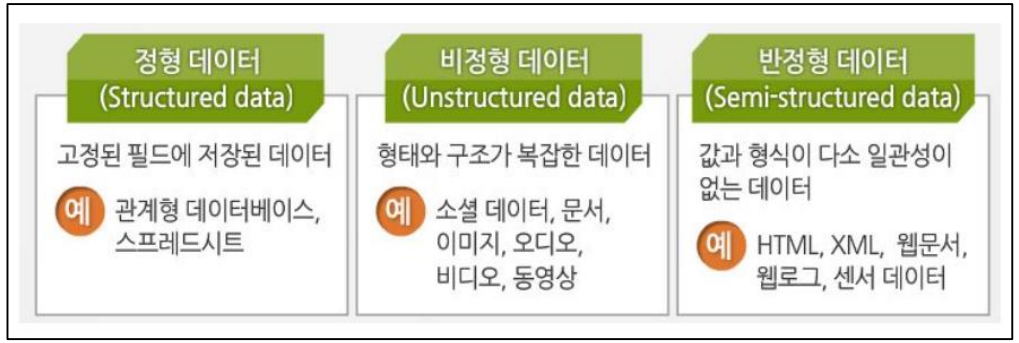
데이터 용량 구분(Memory unit)	크기(Size)
Kilobyte(KB)	10^3 바이트
Megabyte(MB)	10^6 바이트
Gigabyte(GB)	10^9 바이트
Terabyte(TB)	10^{12} 바이트
Petabyte(PB)	10^{15} 바이트
Exabyte(EB)	10^{18} 바이트
Zettabyte(ZB)	10^{21} 바이트
Yottabyte(YB)	10^{24} 바이트

규모(Volume)

- ✓ 규모란 처리해야 할 데이터의 크기를 말하는 속성이다
- ✓ 테라바이트(Terabyte, TB)급 이상의 데이터군을 빅데이터로 통칭

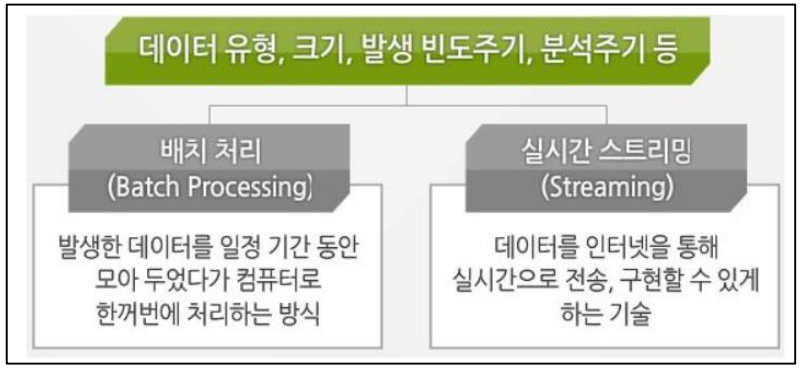
다양성(Variety)

- ✓ 처리해야할 데이터의 유형이 다양함을 말하는 속성
- ✓ 빅데이터는 다양한 데이터 유형을 가짐
- ✓ 데이터의 양과 더불어 유형의 복잡성이 증대함

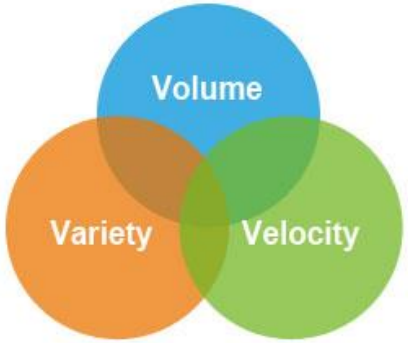


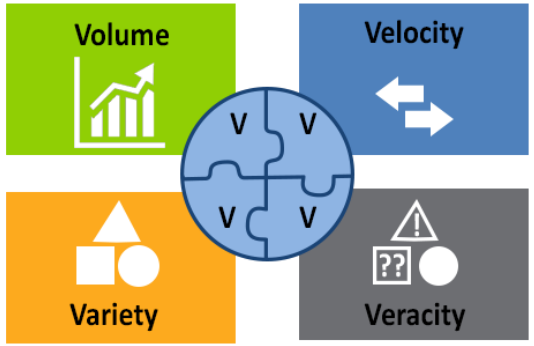
속도(Velocity)

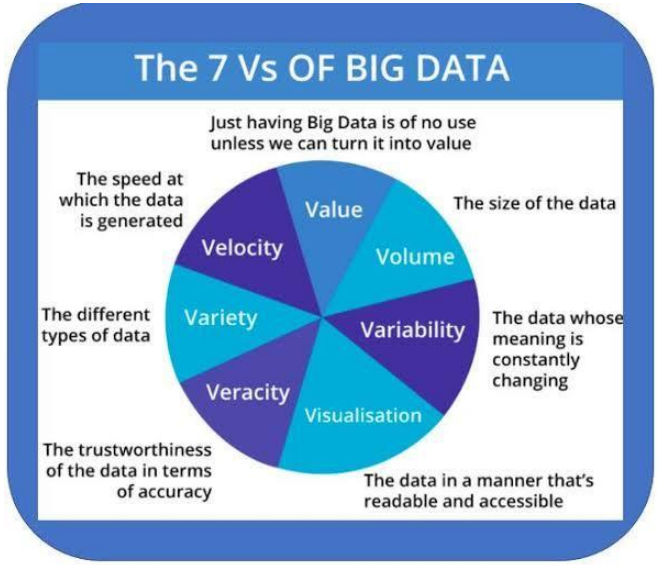
- ✓ 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성
- ✓ 데이터의 빠른 처리 및 분석을 위해 다양한 방식을 적용함

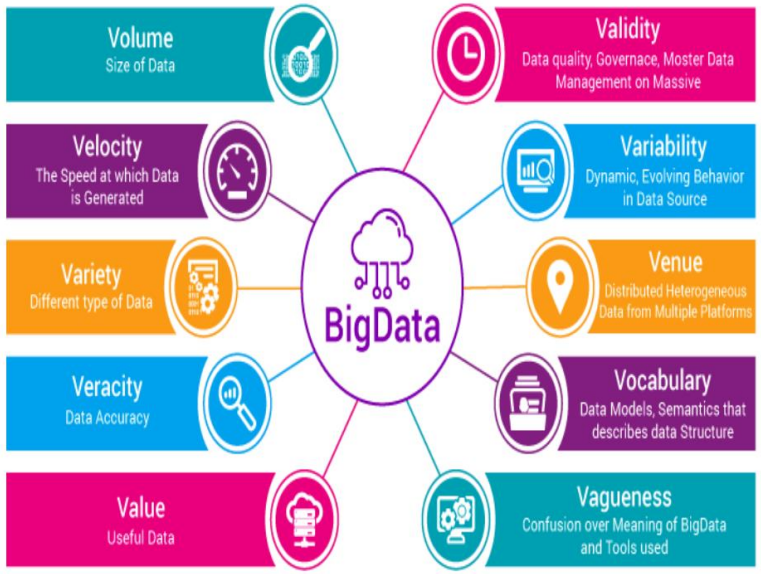


17 'Vs' of Big Data









Volume	**Velocity**	**Variety**	**Value**
Veracity	**Validity**	**Volatility**	**Visualization**
Virality	**Viscosity**	**Variability**	**Venue**
Vocabulary	**Vagueness**	**Verbosity**	**Voluntariness**
Versatility			

빅데이터 활용을 위한 3대 요소

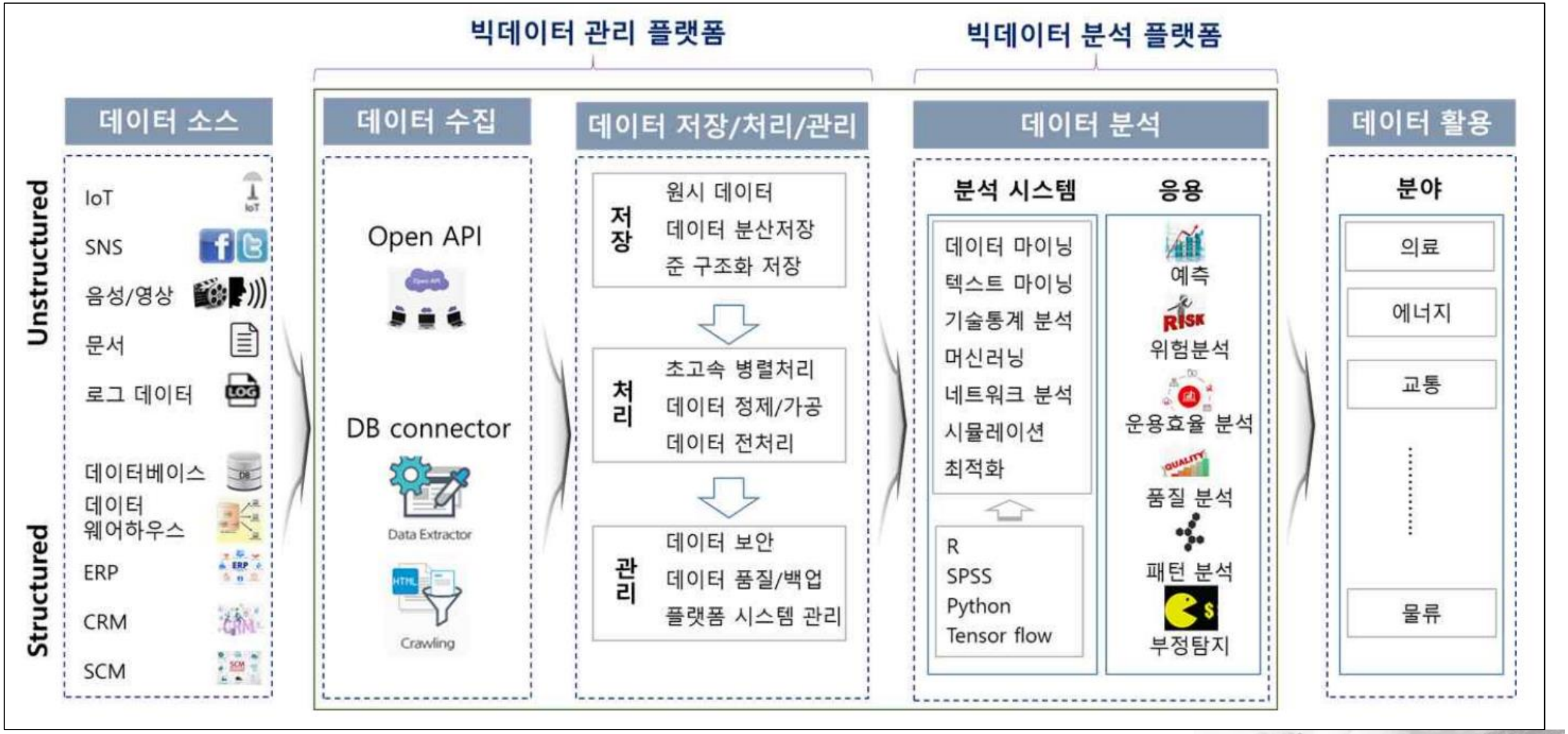
빅데이터(Big Data)
- 데이터 자원 확보
- 데이터 품질 관리

빅데이터 플랫폼(Big Data Platform)
- 빅데이터 기술을 잘 사용할 수 있도록 준비된 환경
- 빅데이터 기술의 집합체
- 데이터 저장, 관리 기술
  - ✓ NoSQL(Not only SQL)
  - ✓ ETL (Extraction, Transformation, and Loading)
- 대용량 데이터 처리
  - ✓ 하둡
  - ✓ 맵리듀스
- 빅데이터 분석
  - ✓ 자연어 처리
  - ✓ 의미 분석
  - ✓ 데이터 마이닝
- 시각화
  - ✓ 빅데이터 표현

데이터자격검정

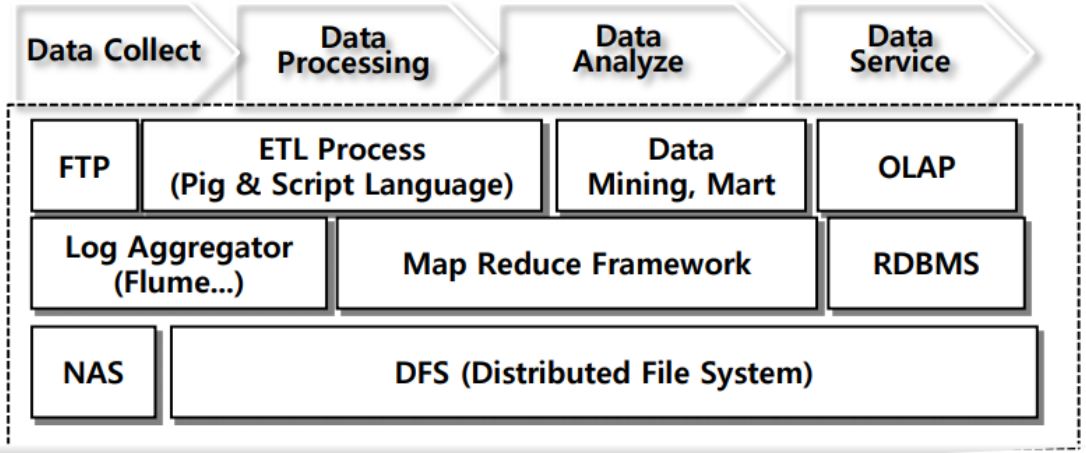
데이터 사이언티스트(Data Scientist)
- 수학, 공학(IT기술과 엔지니어링) 능력
- 경제학, 통계학, 심리학 등 다문확적 이해
- 비판적 시각과 커뮤니케이션 능력
- 스토리텔링 등 시각화 능력

빅 데이터 플랫폼


빅데이터 관리 플랫폼
빅데이터 분석 플랫폼
데이터 소스
데이터 수집
데이터 저장/처리/관리
데이터 분석
데이터 활용
Unstructured
IoT
SNS
음성/영상
문서
로그 데이터
Structured
데이터베이스
데이터 웨어하우스
ERP
CRM
SCM
Open API
DB connector
Data Extractor
Crawling
저장
원시 데이터
데이터 분산저장
준 구조화 저장
처리
초고속 병렬처리
데이터 정제/가공
데이터 전처리
관리
데이터 보안
데이터 품질/백업
플랫폼 시스템 관리
분석 시스템
응용
데이터 마이닝
텍스트 마이닝
기술통계 분석
머신러닝
네트워크 분석
시뮬레이션
최적화
R
SPSS
Python
Tensor flow
예측
위험분석
운용효율 분석
품질 분석
패턴 분석
부정탐지
분야
의료
에너지
교통
물류

OSS 기반 빅 데이터 플랫폼

검증 기술 영역 정의 — OSS 기반 Big Data Platform 구축을 위한 Open Source 및 솔루션 기술 영역



검증 대상 후보 선정 — 기준 : 안정성, Ref.多, 旣 검증 여부, 기능/성능 미흡 여부, Apache main project 여부 등 고려

수집

Open Source
- Flume ☑
- Chukwa ☐
- Scribe ☐

상용 Solution
- CDC ☐
- 복제툴 ☐
- EAI ☐

전송

Open Source
- Sqoop ☑
- Hiho ☐

상용 Solution
- ETL Tool ☐

저장

Open Source
- ✓ 분산파일시스템
  - HDFS ☑
  - GlusterFS ☐
- ✓ NoSQL
  - HBase ☐
  - Cassandra ☐
  - MongoDB ☐
- ✓ In-Memory
  - Redis ☐
  - Membase ☐

처리

Open Source
- ✓ 배치
  - MapReduce ☑
  - Pig ☑
- ✓ 실시간
  - Esper ☐
  - Storm ☐
  - S4 ☐

SQL On Hadoop

Open Source
- ✓ 배치 쿼리
  - Hive ☑
- ✓ 실시간 쿼리
  - Impala ☐
  - Tajo ☐

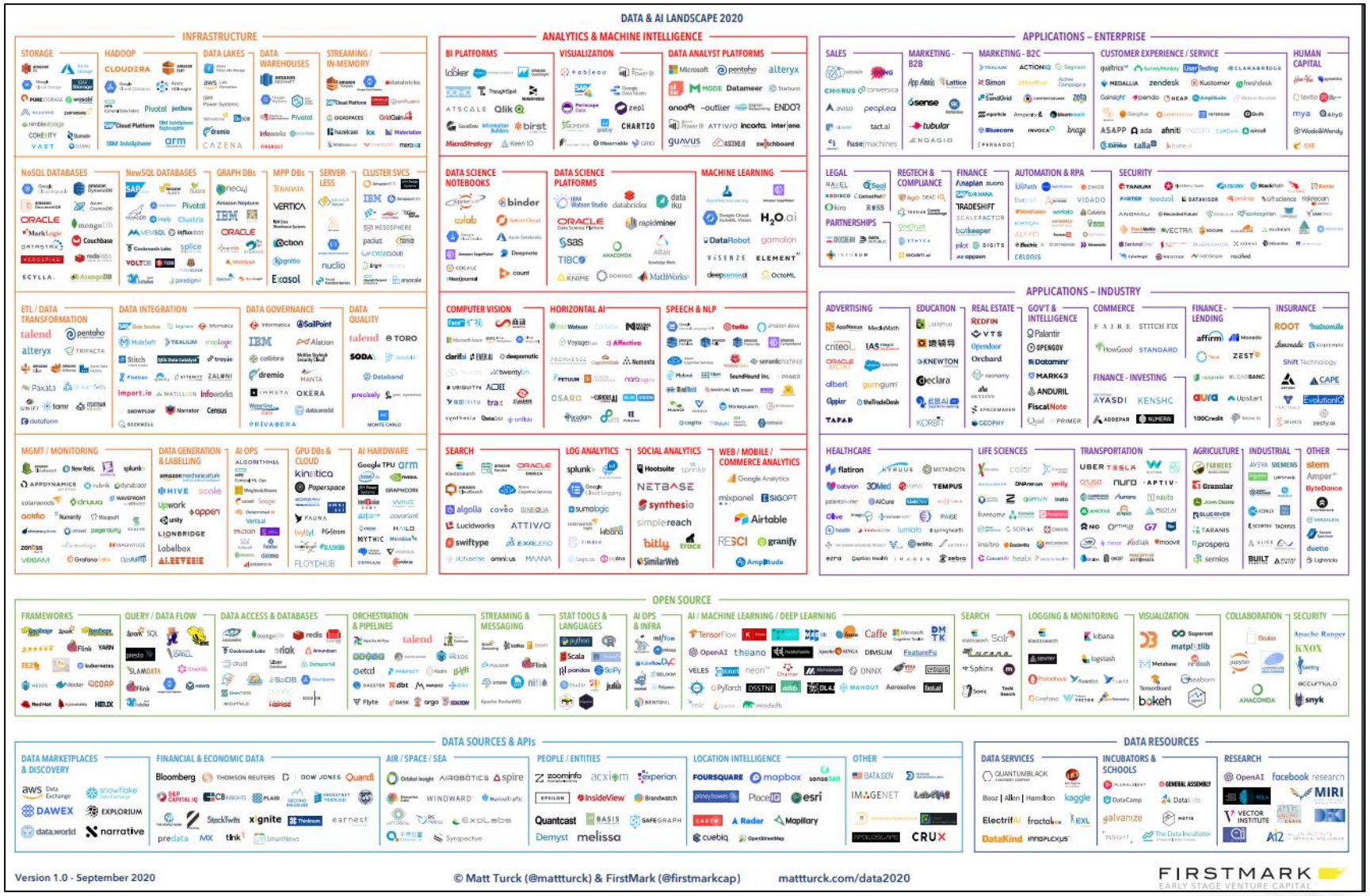
분석

Open Source
- ✓ 통계분석
  - R ☐
- ✓ ETL/OLAP
  - Pentaho ☐
- ✓ 그래프 분석
  - Giraph ☐
- ✓ 기계학습
  - Mahaout ☐

Management

Open Source
- ✓ 모니터링
  - Ambari ☑
- ✓ 워크플로우
  - OOzie ☑
- ✓ 코디네이터
  - Zookeeper ☑

DATA & AI LANDSCAPE 2020



[Underlying list]

빅데이터 기술 분류

과정	설명	해당 기술
생성	조직의 내부와 외부에 존재하는 여러 데이터를 생성하는 기술	데이터베이스, 파일관리시스템, 인터넷으로 연결된 파일, 등
수집	조직 내부와 외부에 존재하는 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술로 단순한 데이터 확보가 아닌 검색, 수집, 변환을 통해 정제된 데이터를 확보하는 기술	로그 수집기, RSS Reader, Open API, 크롤링, ETL, 센싱, 등
저장	작은 데이터라도 모두 저장하고 실시간으로 저렴하게 데이터를 처리하고 처리된 데이터를 더 빠르고 쉽게 분석하도록 효율적으로 저장하는 기술	분산파일시스템, NoSQL, 병렬 DBMS, 등
처리	엄청난 양의 데이터 저장, 수집, 관리, 유통, 분석을 처리하는 일련의 기술	실시간처리, 분산병렬처리, 맵리듀스, 등
분석	데이터를 효율적으로 정확하게 분석하여 비즈니스 등의 영역에 적용하기 위한 기술로 이미 여러 영역에서 활용해온 기술	통계 분석, 평판 분석, 데이터 마이닝, 텍스트 마이닝, 소셜 네트워크 분석, 등
시각화	자료를 시각적으로 묘사하는 기술로, 빅데이터는 기존의 단순 선형적 구조의 방식으로 표현하기 힘들기 때문에 필수	정보 편집 기술, 정보 시각화 기술, 시각화 도구, 등

생성

- 조직의 내부와 외부에 존재하는 **여러 데이터를 생성하는 기술**을 의미함

해당 기술

- 내부 데이터(정형 데이터)
  - **데이터베이스(Database)** 관리시스템
  - 파일관리시스템(File Management system)
- 외부 데이터(비정형 데이터)
  - 인터넷으로 연결된 파일 등

공유되어 사용될 목적으로 통합하여 관리되는 데이터의 집합

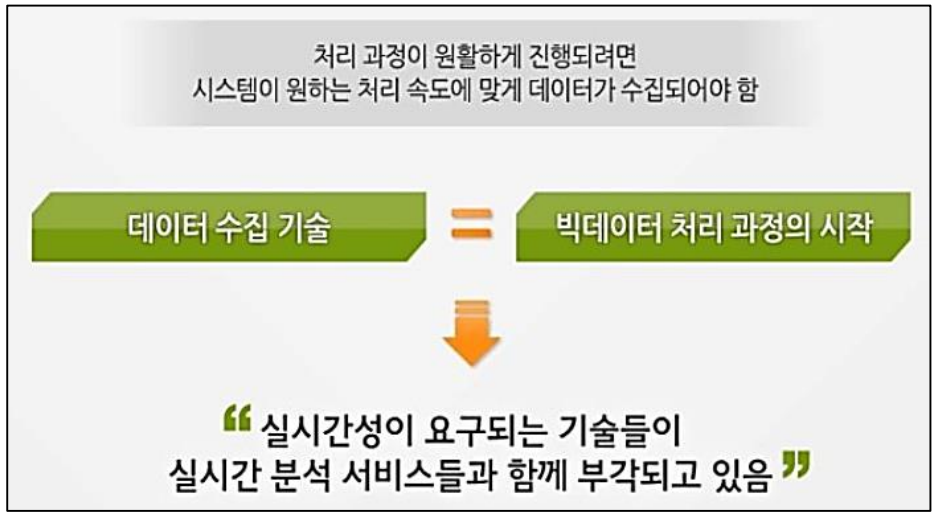
수집

- 조직의 내부와 외부에서 생성되는 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술
- 단순 데이터 확보가 아닌 검색, 수집, 변환을 통해 정제된 데이터를 확보하는 기술을 의미함

해당 기술

- 내부 데이터(정형 데이터)
  - 로그 수집기
- 외부 데이터(비정형 데이터)
  - 크롤링
  - 센싱
  - RSS Reader, Open API
  - ETL(Extraction, Transformation, Loading) 등

Software platform



용 어	뜻
로그 수집기	조직 내부에 존재하는 웹 서버의 로그 수집, 웹 로드, 트랜잭션 로그, 클릭 로그, DB 로그데이터 등을 수집
크롤링	주로 웹 로봇을 이용하여 조직 외부에 존재하는 소셜 데이터 및 인터넷에 공개되어 있는 자료 수집
센싱	각종 센서를 통해 데이터를 수집
RSS Reader	데이터의 생산, 공유, 참여 환경인 웹2.0을 구현하는 기술
스쿱 (Sqoop)	대용량 데이터 전송 솔루션으로 하둡기반 시스템과 통합지원하며 맵리듀스에 사용될 프로그램 코드를 생성
플럼 (Flume)	분산 환경에서 대량의 로그 데이터를 효과적으로 수집해 다른 곳으로 전송하는 서비스로 실시간 로그분석이 가능
척와 (Chukwa)	분산 서버로부터 로그 데이터를 수집하여 하둡 클러스터의 로그나 서버의 상태정보를 관리해 하둡 파일 시스템에 저장하며 실시간 분석이 가능
스플렁크 (Splunk)	업무현장이나 클라우드상에 존재하는 페타비트급의 기록 데이터와 실시간 기계 데이터를 모니터링하고 분석
스크라이브 (Scribe)	facebook이 개발해 공개한 로그수집기술로 대량의 서버에서 실시간으로 오는 로그 데이터를 집약해 하둡 분산 시스템에 로그를 저장
카프카 (kafka)	로그데이터를 수집 할 뿐만 아니라 메시징 시스템을 통해 전송데이터를 압축하고 메시지를 일괄적으로 전송한다.

저장

- 작은 데이터라도 모두 저장하고 실시간으로 저렴하게 데이터를 처리하고 처리된 데이터를 **더 빠르고 쉽게 분석하도록** 효율적으로 저장하는 기술을 의미함
- 빅데이터의 **대용량, 비정형, 실시간성의 속성을 수용**할 수 있는 저장방식이 필요함

해당 기술

- 분산 파일 시스템(Distributed File System)
  - 하둡 분산 파일 시스템 (HDFS : Hadoop Distributed File System)
  - 구글 파일 시스템(GFS : Google File System) 등
- NoSQL
  - H베이스(Hbase), 카산드라(Cassandra), 몽고DB(MongoDB) 등
- 병렬 DBMS
  - 버티카(Vertica), 그린플럼(Greenplum), 넷티자(Netezza) 등

분산 파일 시스템(Distributed File System)

컴퓨터 네트워크로 공유하는 여러 호스트 컴퓨터 파일에 접근할 수 있는 파일 시스템

하둡 분산 파일 시스템(HDFS : Hadoop Distributed File System)

- 하둡(Hadoop)
  - 대량의 자료를 저장하고 처리할 수 있는 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 **자바 소프트웨어 프레임 워크**
  - 7년 간 개발되면서 **개방형 프레임 워크**로 빅데이터 시대를 이끌고 있음
  - 하둡을 중심으로 한 새로운 제품군들이 등장하고 있음
- 하둡 분산 파일 시스템(HDFS : Hadoop Distributed File System)
  - 이기종간의 하드웨어로 구성된 컴퓨터 클러스터에서 **대용량 데이터 처리**를 위하여 개발된 분산 파일 시스템

Google 구글 파일 시스템(GFS : Google File System)

- **구글(Google)**
  - 웹 검색, 클라우드 컴퓨팅, 광고를 주 사업 영역으로 하는 미국의 다국적 회사
- **구글 파일 시스템(GFS : Google File System)**
  - 구글에 의해 자기 회사 사용 목적으로 개발된 분산 파일 시스템
  - 일반 상용 하드웨어를 이용하여 대량의 서버를 연결했기 때문에 데이터에 대한 접근이 효율적이고 안정적임

Software platform

NoSQL

Not only SQL(SQL뿐만 아니라)
SQL은 표준으로 채택하고 있는 특수 목적의 프로그래밍 언어임
➡ NoSQL은 기존과는 다른 새로운 데이터 저장 기술

- 기존 문제를 개선, 보완하기 위해서 사용된 새로운 데이터 저장기술
- 비관계형 데이터베이스를 지칭하는 분산 환경의 데이터 저장소
- NoSQL에서 데이터 저장

▪ 다수의 서버에 분산해서 저장하여 속도가 빠름
- 트랜잭션이 클러스터를 구성하는 전체 서버에 분산되기 때문에 **다수의 클라이언트가 동시에 접속해서 사용하여도 됨**

APACHE HBASE

H베이스(Hbase)

무상으로 공개된 소스 코드

▪ 하둡 분산 파일 시스템(HDFS : Hadoop Distributed File System)에서 동작되는 **오픈 소스** 분산, 비관계형 데이터베이스
- **구글의 '빅테이블'을 참고로 개발됨**
- 파워셋에서 개발했으며, 현재는 아파치 소프트웨어 재단에서 한 프로젝트로 관리함

▪ 데이터를 많이 사용하는 웹사이트에서 사용됨

Cassandra

카산드라(Cassandra)

▪ **분산 시스템에서 대용량 데이터를 처리할 수 있도록 설계**된 오픈 소스 데이터베이스 관리 시스템
- 아마존의 다이나모(Dynamo)와 구글의 빅테이블(BigTable)의 장점만을 수용하여 발전한 형태인 NoSQL의 대표적인 데이터베이스
- 원래 페이스북에서 개발했으며 **지금은 아파치 소프트웨어 재단에서** 한 프로젝트로 관리함

병렬 DBMS

다수의 마이크로프로세서를 사용하여 여러 디스크의 질의, 갱신, 입출력 등 **데이터베이스 처리를 동시에 수행**하는 데이터베이스 시스템

VERTICA

버티카(Vertica)

▪ 휴렛패커드(HP : Hewlett-Packard Company) 자회사

▪ **빠른 분석을 위한** 컬럼 기반의 대규모 병렬처리용 데이터베이스 관리 시스템

Greenplum

그린플럼(Greenplum)

▪ EMC 코퍼레이션 자회사

▪ **관계형 데이터베이스와 하둡을 한 장비에 넣음**
- 데이터웨어하우징(DW)업계 최초

NETEZZA an IBM Company

넷티자(Netezza)

▪ IBM 자회사

▪ 기존 데이터베이스를 이용하는 타사와 달리 **데이터웨어하우징의 처리 고속화를** 위해서 설계된 제품을 제공함

mongoDB

몽고DB(MongoDB)

▪ 신뢰성과 확장성에 기반한 **문서 지향 데이터베이스**
- 방대한 양의 데이터에서 낮은 관리 비용과 사용 편의성을 목표로 함

▪ 가장 유명한 NoSQL 데이터베이스 시스템
- 10gen이 오픈 소스로 개발한 것으로 **상업적인 지원이 가능함**

처리

- 빅데이터에서 유용한 정보 및 숨어있는 지식을 찾아내기 위한 **데이터 가공 및 분석 과정을 지원**하는 기술을 의미함
- 대규모 데이터 처리를 위한 확장성, 데이터 생성 및 처리 속도를 해결하기 위한 처리 시간 단축 및 실시간 처리 지원, 비정형 데이터 처리 지원 등이 필요함

해당 기술

- 하둡(Hadoop)
- NoSQL
- 구글 맵리듀스(MapReduce) 등

맵리듀스(MapReduce)

구글이 발표한 대표적인 빅데이터 병렬처리 모델

- 대용량 데이터를 빠르고 안전하게 처리하기 위한 분산 프로그래밍 모델임
- 맵(Map)함수와 리듀스(Reduce)함수 기반으로 구성되는 데이터를 병렬 처리하는 기술임
- 하둡에서도 구현됨

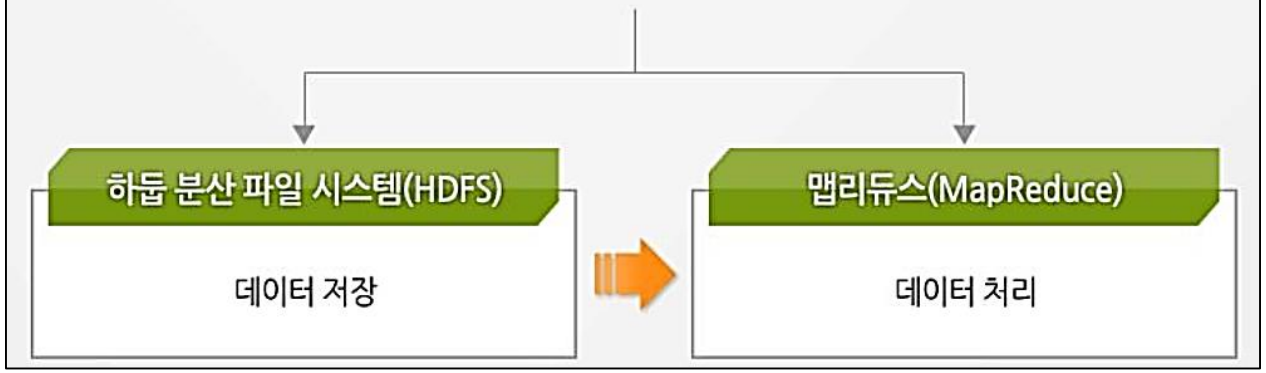
NoSQL

Not only SQL
기존과 다르게 설계된 비관계형 데이터베이스

APACHE HBASE | Cassandra | mongoDB

하둡(Hadoop)

분산 파일 시스템과 맵리듀스를 구현한 빅데이터 처리 기술의 대표적인 프레임 워크



빅데이터 처리 기술의 동향

하둡을 포함한 오픈소스 진영을 중심으로 다양한 기술이 빠르게 진화하고 있음

아직까지 상용 솔루션보다 오픈소스의 비중이 높음

분석

- 데이터를 효율적으로 정확하게 분석하여 비즈니스 등의 영역에 적용하기 위한 기술로 이미 여러 영역에서 활용해온 기술
- 빅데이터로부터 숨어있는 패턴과 지식을 찾아내기 위한 기술을 의미함
- 찾아진 패턴과 지식을 토대로 비즈니스 영역에서는 의사결정을 수행함

해당 기술

- 통계분석
- 데이터 마이닝
- 텍스트 마이닝
- 평판분석
- 소셜 네트워크 분석 등

통계분석

- 다양한 분석에서 활용되는 기술
- 통계적 컴퓨팅에 사용되는 R, SAS 등을 통하여 다양한 통계기법으로 분석할 수 있음

R

- 빅데이터 분석 기술 도구
- 통계계산 및 시각화를 위한 언어 및 개발 환경을 제공할 경우
  - 기본적 통계 기법부터 데이터 마이닝 기법까지 구현이 가능함

SAS(Statistical Analysis System)

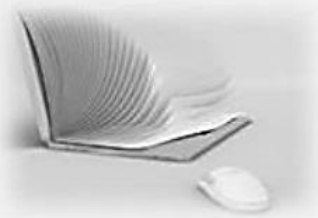
- 미국 노스캐롤라이나 주립 대학교에서 1967년 원형이 개발되고 SAS Institute사에서 기능을 확장함
- 통계 해석을 중심으로 한 소프트웨어 패키지

Software platform

데이터 마이닝

- 통계 및 수학적 기술뿐 아니라 기계학습, 패턴인식, 신경망 등의 기술들을 이용함
- 대용량의 데이터에 숨겨진 의미 있는 패턴, 추세, 지식들을 발견함

텍스트 마이닝



- 구조화되지 않은 대규모의 텍스트 집합으로부터 새로운 지식을 발견하는 기술
- 텍스트 문서로부터 정보 검색, 정보 추출, 체계화 및 분석을 포함함

평판분석

- 소셜 미디어 등의 정형 또는 비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별하는 분석 기술
- 주로 특성 서비스 및 상품에 대한 시장 규모 예측, 소비자의 반응, 입소문 분석 등에 활용됨

소셜 네트워크 분석

- 소셜 네트워크 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 분석하는 기술
- 주로 마케팅을 위하여 소셜 네트워크 상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾는데 활용됨

시각화

- 데이터 분석결과를 쉽게 이해할 수 있도록 시각적인 수단으로 정보를 전달하는 과정
- 데이터 안의 수많은 패턴들을 시각화하여 핵심개념과 아이디어를 직관적이고 명확하게 이해할 수 있는 기술을 의미함

"가장 중요한 기술 분야"

해당 기술

- 정보 편집 기술
- 정보 시각화 기술
- 시각화 도구 등

정보 편집 기술

시각적 매핑, 스토리 텔링 등의 방법이 활용됨

시간 시각화

- 시간에 따른 데이터 변화를 표현하는 방법

스토리 텔링

- Story + Telling, 이야기하다의 의미
- 상대방에게 알리고자 하는 바를 재미있고 생생한 이야기로 설득력 있게 전달하듯이 구성하는 것

Software platform

정보시각화 기술

시간 시각화, 분포 시각화 등의 방법이 활용됨

시간 시각화

- 시간에 따른 데이터 변화를 표현하는 방법

분포 시각화

- 최대, 최소, 전체분포로 분류하여 시각적으로 표현하는 방법

시각화 도구

마이크로소프트 엑셀, 구글 스프레드시트, IBM의 Watson analytics 등의 소프트웨어 도구가 시각화에 사용됨

마이크로소프트 엑셀

- 윈도 환경의 스프레드시트 프로그램

구글 스프레드시트

- 연산 및 표를 작성하고 그래프를 그리는 소프트웨어

IBM의 Watson analytics

- 다양한 범위의 시각 자료를 보고 쉽게 편집할 수도 있는 사이트

Software platform

reference

IT CookBook, 빅데이터 컴퓨팅 기술, 박두순 , 문양세 , 박영호 , 윤찬현 , 정영식 , 장형석, 한빛아카데미
ETRI-데이타플랫폼-사장규모전망.pdf
2-1SK C&C_디지털데일리『오픈 테크넷 서밋 2014』발표 자료 - 오픈소스 기반의 빅데이터 플랫폼 구축_3.pdf