빅데이터의 개요

빅데이터 정의

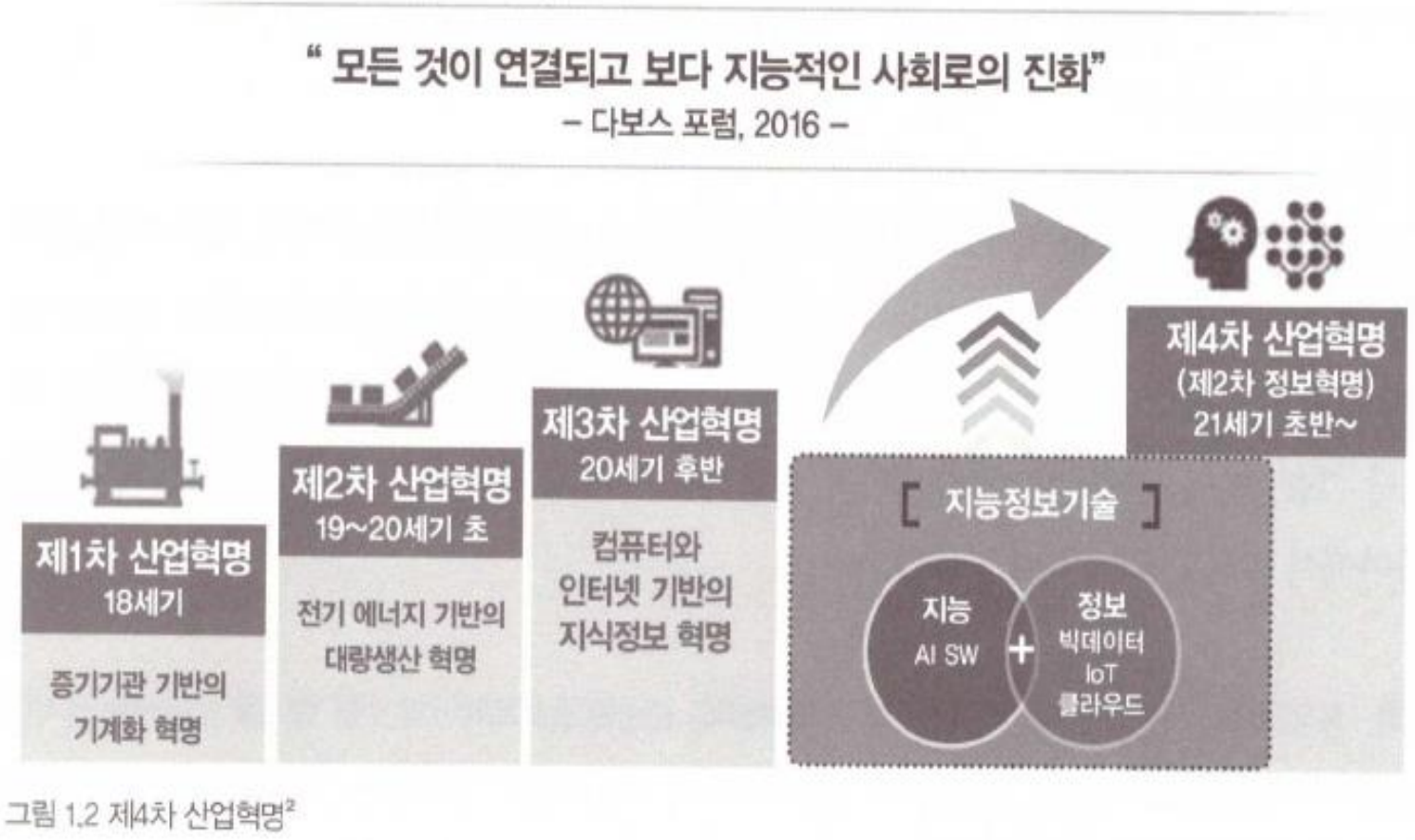


그림 1.2 제4차 산업혁명[2]

빅데이터 정의

- IT 산업의 발달과 함께 모바일기기, 페이스북 / 트위터 등의 소셜 네트워크 성장으로 만들어지는 데이터의 양이 폭발적으로 증가.
- 하둡과 같은 오픈 소스 프로젝트의 등장과 함께 큰 데이터의 처리가 상대적으로, 적은 비용으로 가능해짐.
- 데이터 수집과 데이터 마이닝의 중요성이 부각되기 시작.
  - 데이터 마이닝(Data Mining) :
    - 대용량 데이터에서 어떤 패턴을 찾기 위한 과정.
    - 데이터로부터 어떤 의미를 찾아서 새로운 가치를 생성하는 과정.
    - 컴퓨터 사이언스와 통계/수학 등이 함께 사용되며 흔히 인공지능이나 머신러닝의 기술이 사용.
- 데이터 크기와 관련 없이 그리고 어떤 시스템을 쓰느냐에 관계없이 자신의 시스템에서 수집된 데이터를 잘 가공하고 거기서 새로운 의미를 도출하여 시스템을 더욱 더 향상 시킬 수 있다면 빅데이터 시스템을 갖고 있다고 볼 수 있다.
- 빅데이터를 처리해주는 시스템을 구축(소프트웨어 엔지니어링)하는 것과 이런 시스템을 운영(데이터 과학자)하는 것은 다른 기술이다.

빅데이터 정의

- 정의 1 : '서버 한 대로 처리할 수 없는 규모의 데이터'
  - 존 라우저(John Rauser) : 아마존의 데이터 과학자.
  - 갖고 있는 데이터를 처리하기 위해 분산 환경이 필요하느냐에 포커스를 둔 정의.
- 정의 2 : '기존의 소프트웨어로는 처리할 수 없는 규모의 데이터'
  - 기존 소프트웨어 예 : 오라클이나 MySQL과 같은 관계형 데이터베이스

    -> 데이터 처리 용량을 늘리려면 메모리를 추가하거나 CPU나 디스크를 더 장착하는 방식으로 리소스를 더 추가 : 스케일업(Scale-up) 방식
  - Scale-out 방식

    -> 서버 자체를 더 추가함으로써 전체 시스템의 용량을 키우는 방식.

    -> NoSQL이나 하둡 등의 분산환경의 시스템.

    -> 저가의 장비를 여러 대 사용하는 방식을 주로 택함.
- 정의 3 : 3V(Volume, Velocity, Variety)
  - 컨설팅 회사들이 많이 사용하는 정의.
  - Volume(규모) : 데이터의 크기가 대용량인지 여부
  - Velocity(속도) : 데이터가 얼마나 빠르게 생성되는지 여부
  - Variety(다양성) : 데이터가 구조화/비구조화 된 데이터를 다 포함하는지 여부
  - Variability : 데이터의 형태가 아예 알려져 있지 않은지 아니면 급변하는지 여부

빅데이터 정의 확장

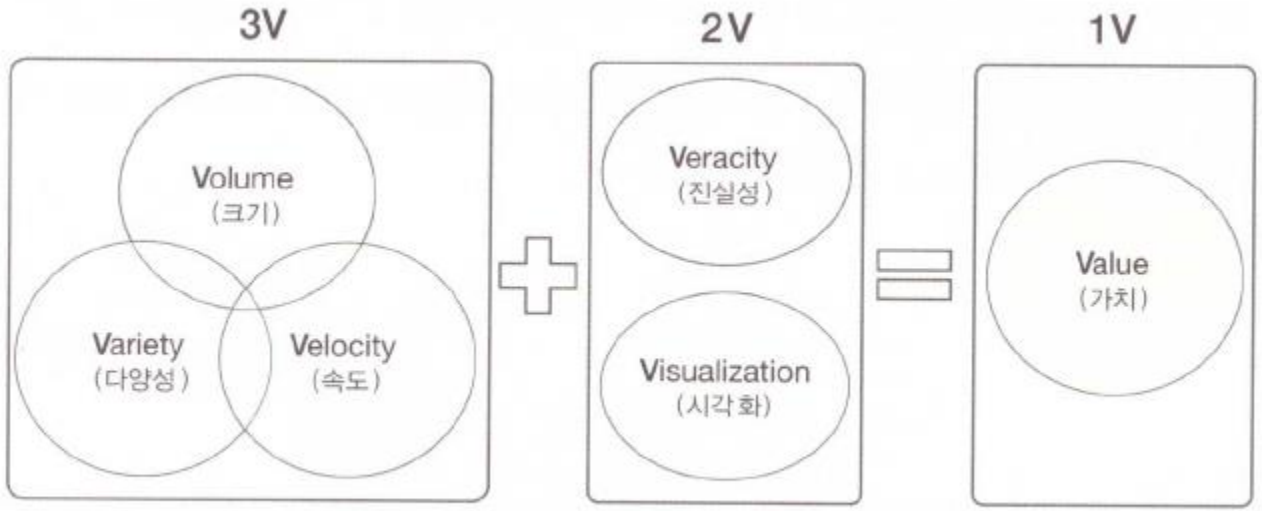


그림 1.4 빅데이터의 정의: 6V

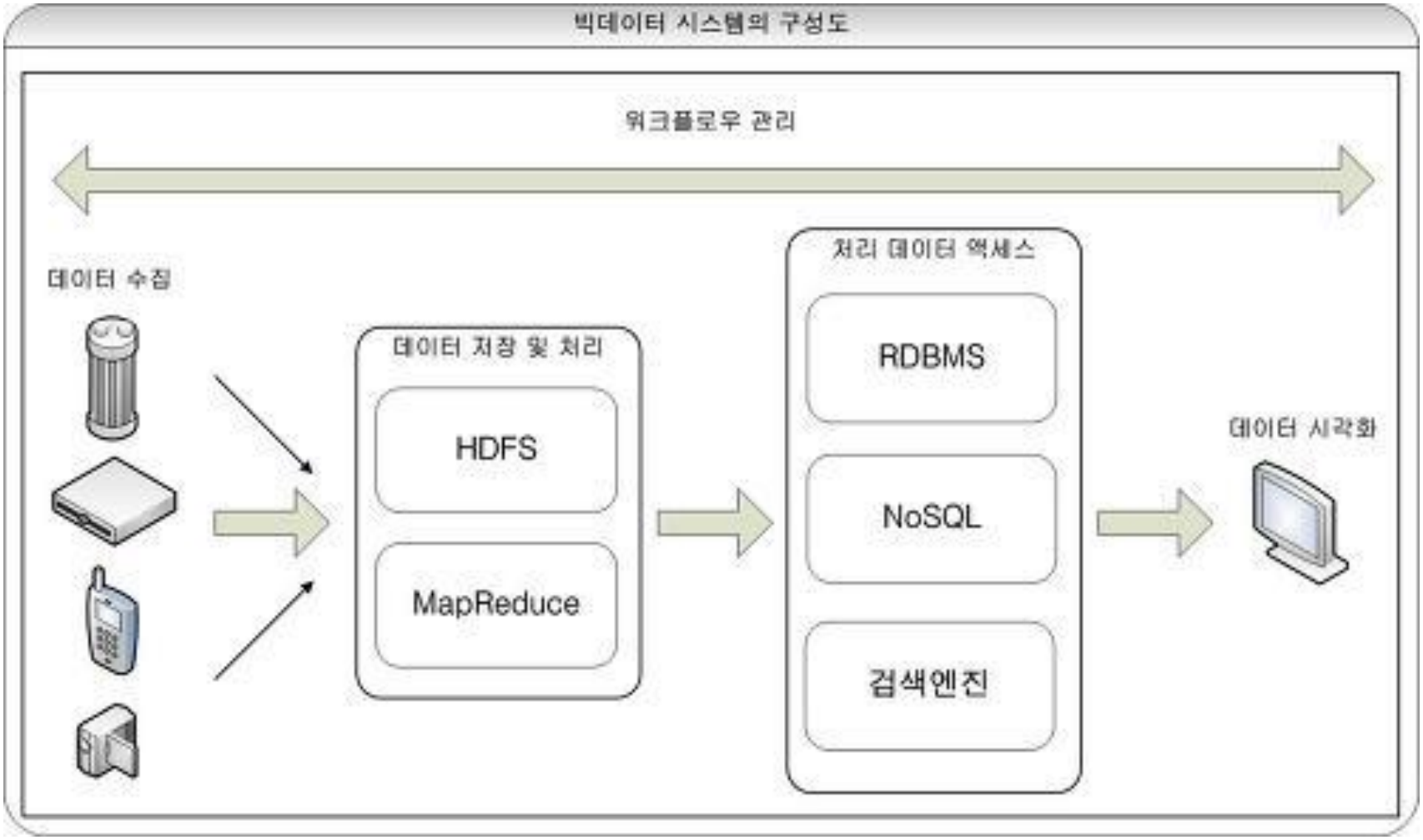
- Veracity(진실성) : 주요 의사결정을 위해 데이터의 품질과 신뢰성 확보
- Visualization(시각화) : 복잡한 대규모 데이터를 시각적으로 표현
- Value(가치) : 비즈니스 효익을 실현하기 위해 궁극적인 가치를 창출

"지구상에선 지금 이 순간에도 방대한 크기(Volume)의 다양한(Variety) 데이터들이 빠른 속도(Velocity)로 발생하고 있다. 빅데이터는 3V를 수용하며, 데이터의 진실성(Veracity)을 확보하고, 분석 데이터를 시각화(Visualization)함으로써 새로운 효익을 가져다 줄 가치(Value)를 창출하는 것이다. "

빅데이터 예

- 웹 검색엔진 데이터
  - 웹페이지 데이터
  - 검색어 로그와 클릭 로그 데이터
- 디바이스에서 생성되는 데이터
  - 스마트폰, 스마트TV, 보잉 제트 엔진 등
- 소셜미디어의 데이터
  - 페이스북, 트위터 등

빅데이터 시스템의 구성


빅데이터 시스템의 구성도
워크플로우 관리
데이터 수집
데이터 저장 및 처리
HDFS
MapReduce
처리 데이터 액세스
RDBMS
NoSQL
검색엔진
데이터 시각화

빅데이터 시스템의 구성

- 데이터 수집 모듈
  - 빅데이터 시스템의 구성에 있어 가장 중요.
  - Flume, Chukwa
  - Kafka

- 데이터 저장/처리 모듈(하둡)
  - HDFS(Hadoop Distributed File System) : 분산 파일 시스템(저장)
  - MapReduce : 분산 처리 시스템(처리)

    -> 자바나 스크립트 언어, C++ 사용

    -> Hive 나 Pig 언어들로 프로그래밍 : MapReduce의 개념 없이 하이 레벨에서 프로그래밍
  - 대용량 데이터의 배치 프로세싱에 적합하고, 실시간으로 데이터를 분석하는 용도로 사용하기에는 버겁다.

빅데이터 시스템의 구성

- 처리데이터 액세스 모듈

  : 하둡을 이용해 처리한 결과를 외부에서 실시간으로 액세스해야 한다면 하둡 자체는 적합하지 않고 또 다른 컴포넌트가 필요.

  - 기존 관계형 데이터베이스

    -> 처리된 결과 데이터의 크기가 상대적으로 작거나 별다른 검색이 필요하지 않을 때 적합.

    -> Sqoop : 이런 용도로 사용할 수 있는 오픈소스 프로젝트

  - NoSQL

    -> 스키마 정의없이 동적으로 새로운 필드를 추가하고 삭제하는 것이 가능.

    -> 분산환경을 염두에 두기 때문에 처리할 수 있는 양이나 트래픽이 훨씬 더 거대.

    -> 데이터 중복 저장이 가능하고, 그렇기 때문에 서버가 몇 대 고장나더라도 계속해서 동작 가능.

  - 검색엔진

    -> Lucene : 검색엔진을 만드는데 필요한 기능을 제공해주는 자바 기반 라이브러리.

    -> Solr

    -> ElasticSearch

빅데이터 시스템의 구성

- 작업 워크플로우 관리 정의 모듈
  - 작업 실행 순서 등을 정의하고 관리해 주는 모듈.

- 데이터 시각화 모듈
  - 상용 패키지 : 데이터미어(Datameer), 펜타호(Pentaho)

빅데이터 기술 영역

표 1.1 빅데이터 전문 기술 영역

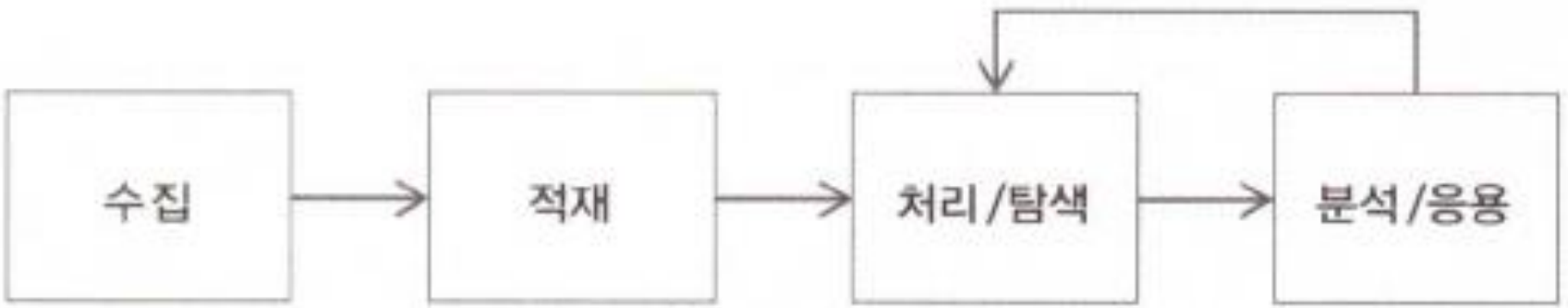
빅데이터 전문 영역		설 명	국내외 사업자
인프라스트럭처	서버	▪ x86급의 CPU, 메모리, 디스크 등을 장착한 서버 ▪ 리눅스 운영체제가 설치된 서버(RedHat, CentOS 등)	HP IBM
	네트워크	▪ 대규모 빅데이터 서버 및 스토리지 지원을 위한 대용량(10G) 네트워크	Cisco
	스토리지	▪ 대규모 데이터를 저장하기 위한 내외부 스토리지 장치	Dell RedHat 등
소프트웨어 플랫폼		▪ 빅데이터의 전방위 기술을 포괄하는 스택 구성 (순수 오픈소스 스택 또는 기업 배포판 스택) ▪ 빅데이터 수집/적재/처리/분석 등의 지원 솔루션 ▪ 빅데이터 시스템 관리 및 모니터링 툴 제공	Cloudera MapR HortonWorks KT넥스알 그루터 클라우다인 등
IT 서비스		▪ 빅데이터 컨설팅 및 구축 이행 ▪ 빅데이터 전문 운영 및 유지보수 ▪ 빅데이터 데이터/분석 서비스 ▪ 빅데이터 교육센터 운영 및 인력 양성	KT DS LG CNS 삼성 SDS SK C&C 다음소프트 등

몇 가지 성공 스토리들

- 넷플릭스의 영화 추천 서비스
- 이베이의 쿼리로그 마이닝
- 트위터의 대용량 머신 러닝 시스템
- 페이스북 – 하둡 구현은 아니지만, HDFS 기반으로 구현된 NoSQL인 HBase 위에서 동작.
- 월마트 – 금요일 저녁 기저귀와 맥주를 같이 배치

빅데이터 아키텍처 레이어 및 역할

빅데이터 구축 단계



빅데이터 시스템 도입에서 얻은 교훈과 문제점

- 교훈
  - 대용량 데이터 중앙 수집의 어려움
  - 성공 스토리의 필요성
  - 데이터 접근 민주화의 중요성

- 문제점
  - ROI(Return On Investment:투자수익률) 고려
  - 개인 정보의 노출
  - 오픈 소스로 구성된 시스템