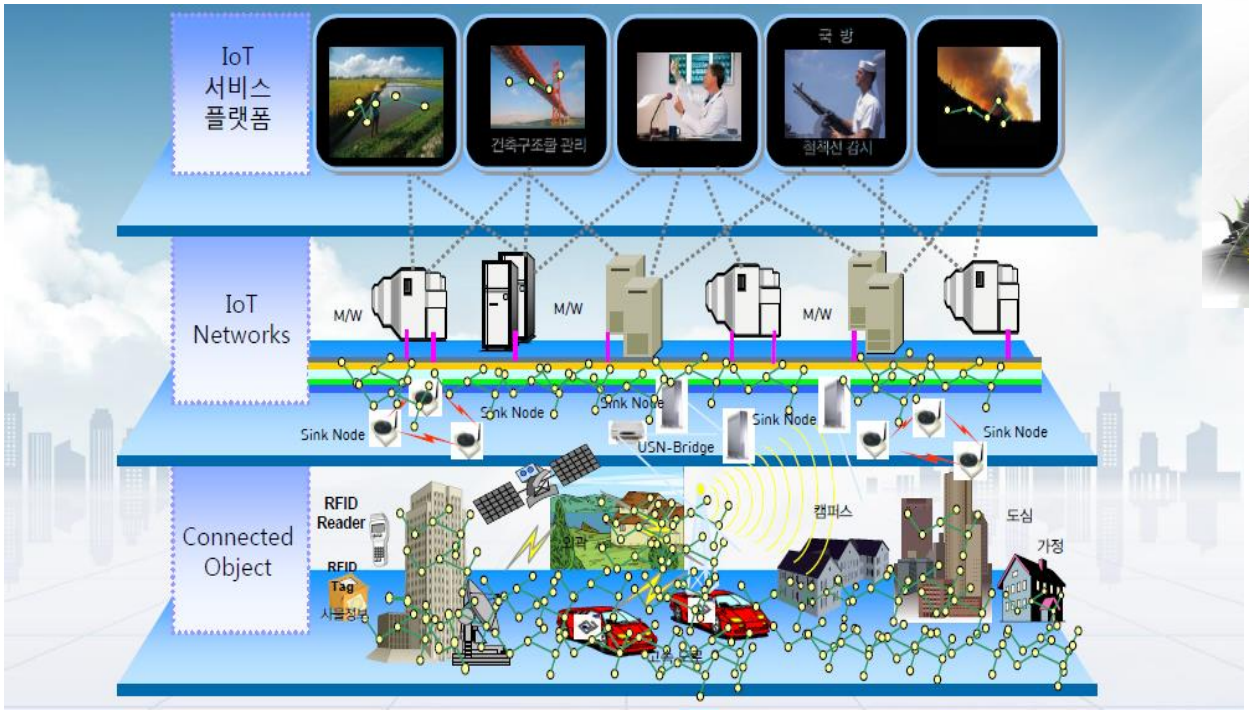
제4장 빅데이터 기술과 방법



Contents

1. 빅데이터 기술
   - 빅데이터 과정
   - 빅데이터 도구

2. 빅데이터 방법
   - 빅데이터 방법 의의
   - 빅데이터 방법 종류

3. 빅데이터 주요 기법
   - 텍스트마이닝
   - 데이터마이닝

4.1 빅데이터 기술

4.1 빅데이터 기술

빅데이터 기술

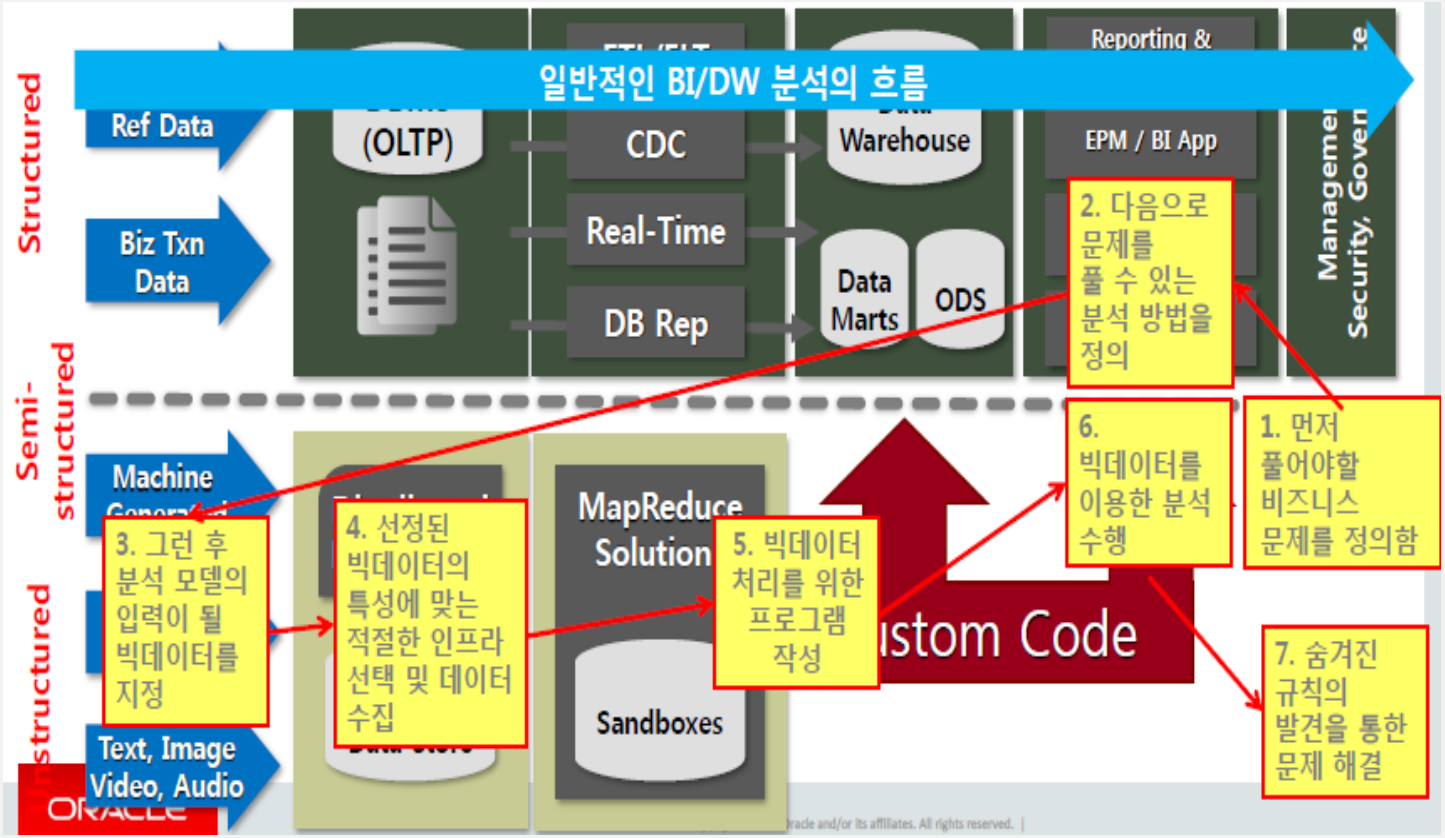
- 빅데이터 분석 기술과 방법들은 기존 통계학과 전산학에서 사용되던 데이터 마이닝, 기계 학습, 자연 언어 처리, 패턴 인식 등

- 소셜 미디어 등 비정형 데이터의 증가로 인해 분석기법들 중에서 텍스트 마이닝, 오피니언 마이닝, 소셜네트워크 분석, 군집분석 등

4.1 빅데이터 기술

4.1.1 빅데이터 과정

- 빅데이터 접근방법

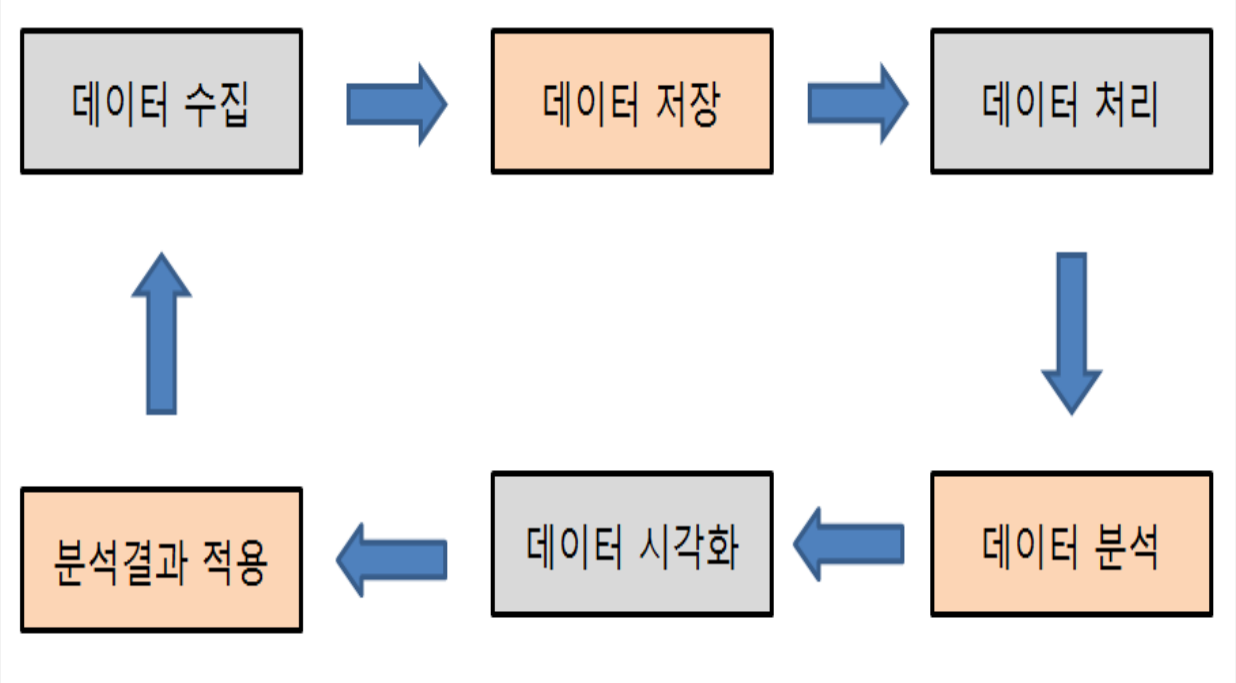


출처 : 한국데이터베이스진흥원

4.1 빅데이터 기술

4.1.1 빅데이터 과정

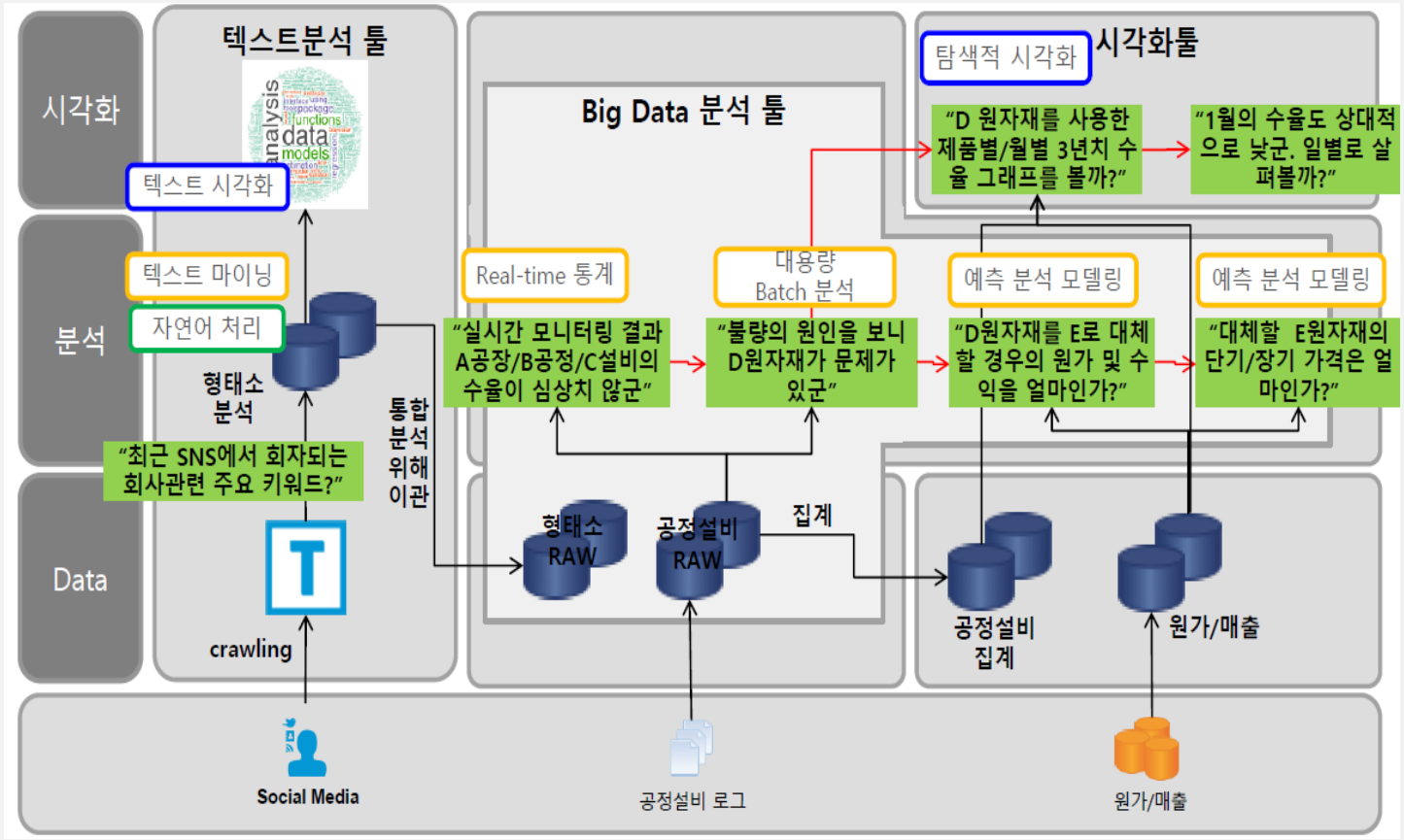
- 빅데이터 과정(1)



4.1 빅데이터 기술

4.1.1 빅데이터 과정

- 빅데이터 과정(2)



출처 : 한국데이터베이스진흥원, 유충현, 2014

4.1 빅데이터 기술

4.1.2 빅데이터 도구

- **R**
  - **통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경**
  - **뉴질랜드 오크랜드 대학** 로버트 젠틀맨(Robert Gentleman)과 로스 이하카(Ross Ihaka)
  - **R- 강점**

    1) 12,086개 이상의 패키지(2018.02 현재)

    2) 그래픽 기능 강화(시각화)

    3) R은 사용자가 제작한 패키지를 추가하여 기능을 확장

    4) 전세계에서 범용적으로 사용

4.1 빅데이터 기술

4.1.2 빅데이터 도구

- **Python**
  - 파이썬(Python)은 1991년 프로그래머인 귀도 반 로섬 (Guido van Rossum) 발표
  - 플랫폼이 독립적이며 인터프리터식, 객체지향적, 동적 타이핑(dynamically typed) 대화형 언어
  - 순수한 프로그램 언어로서의 기능 외에도 다른 언어로 쓰인 모듈들을 연결하는 풀언어(glue language)로써 자주 이용
  - 유니코드 문자열을 지원해서 다양한 언어의 문자 처리에도 강함

4.1 빅데이터 기술

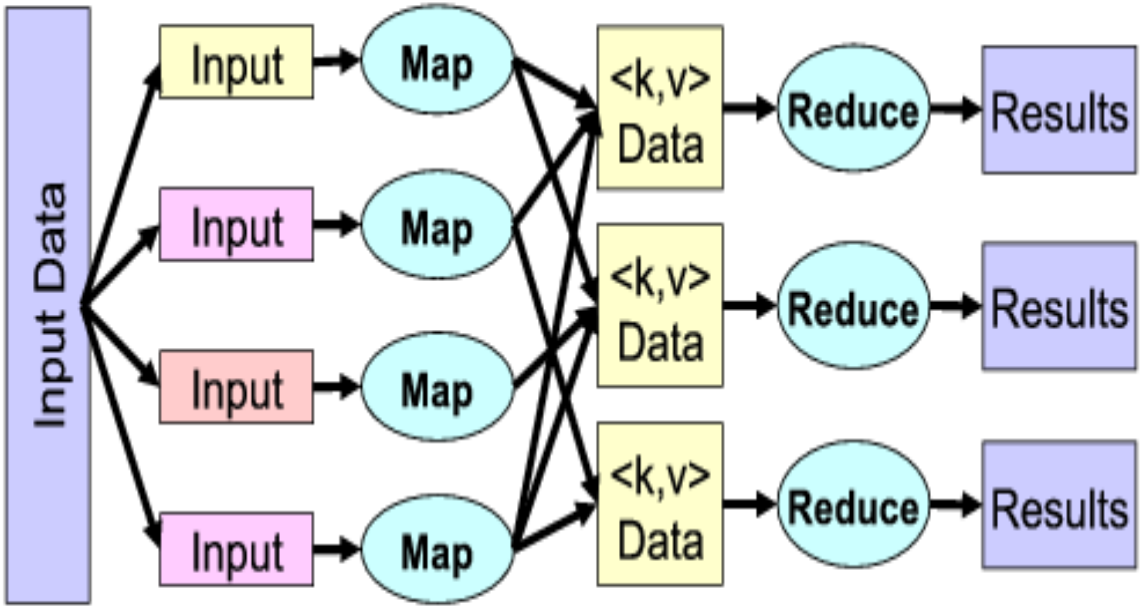
4.1.2 빅데이터 도구

- **Mapreduce**
  - 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크
  - 프레임워크는 페타바이트 이상의 대용량 데이터를 신뢰도가 낮은 컴퓨터로 구성된 클러스터 환경에서 병렬 처리를 지원하기 위해서 개발
  - Map과 Reduce라는 함수 기반으로 구성

4.1 빅데이터 기술

4.1.2 빅데이터 도구

- **Mapreduce 흐름도**



출저 : http://ko.Wikipedia.org/

4.1 빅데이터 기술

4.1.2 빅데이터 도구

- **Hadoop**
  - 대량의 자료를 처리할 수 있는 대형 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 프리웨어 자바 소프트웨어 프레임워크
  - 하둡은 빅데이터 처리를 위한 분산시스템으로 아파치 하둡(Apache Hadoop)은 하둡 분산 파일 시스템(HDFS: Hadoop Distributed File System)과 맵리듀스를 구현
  - 하둡은 여러 개의 저렴한 컴퓨터를 마치 하나인 것처럼 묶어 대용량 데이터를 처리하는 기술로 값비싼 오라클이나 IBM이 개발한 데이터 분석 솔루션을 대체

4.1 빅데이터 기술

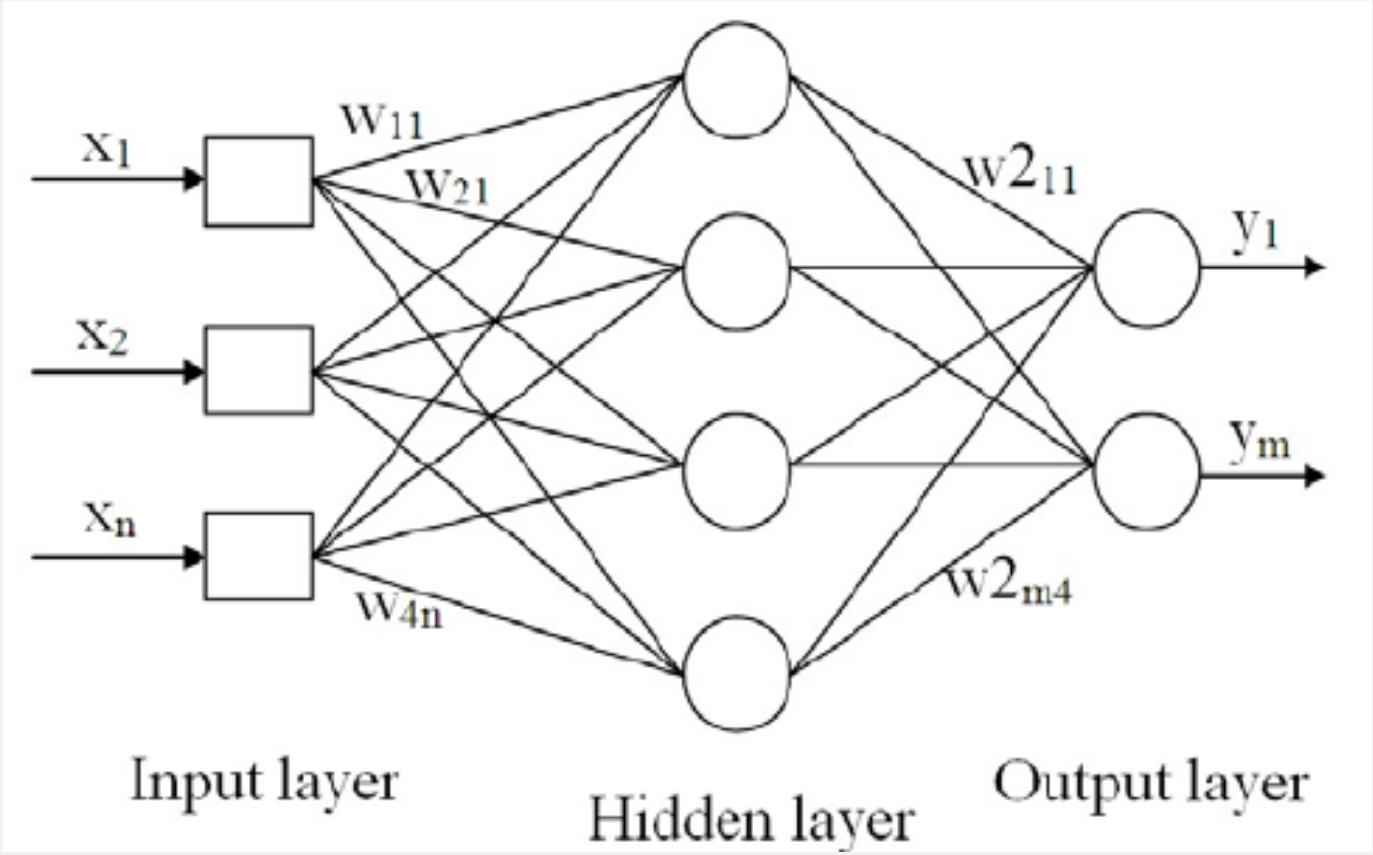
4.1.2 빅데이터 도구

- **Deep Learning**
  - 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화(abstractions, 다량의 데이터나 복잡한 자료들 속에서 핵심적인 내용 또는 기능을 요약하는 작업)를 시도하는 기계학습(machine learning) 알고리즘의 집합
  - 딥 러닝 기법 : 컴퓨터비젼, 음성인식, 자연어처리, 음성/신호처리 등의 분야에 적용
  - 딥 러닝은 인지신경과학자(Cognitive neuroscientist)들이 1990년대 초에 제안한 뇌 발달(Brain development)과 밀접한 관련

4.1 빅데이터 기술

4.1.2 빅데이터 도구

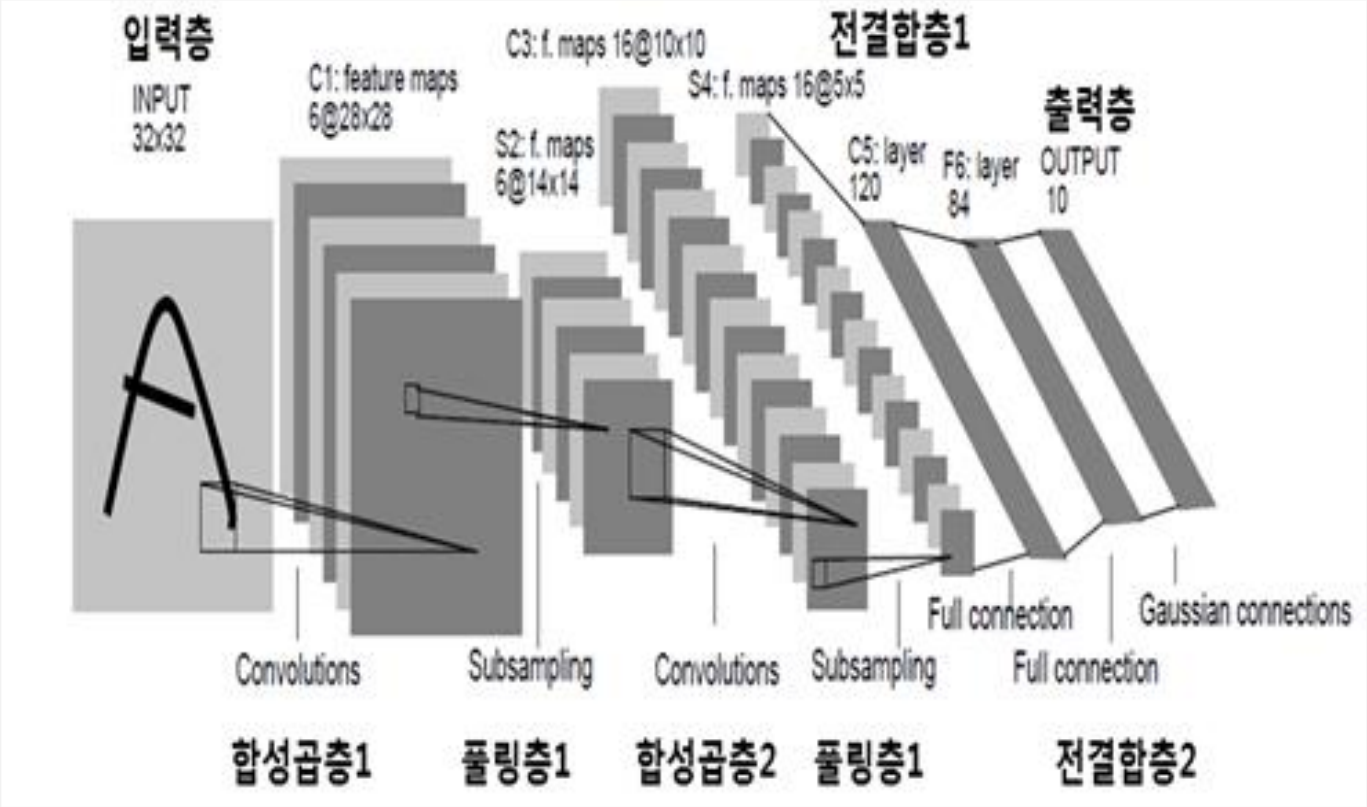
- **Deep Learning 예시**



4.1 빅데이터 기술

4.1.2 빅데이터 도구

- **Deep Learning 예시**



4.2 빅데이터 방법

4.2.1 빅데이터 방법 의의

- 빅데이터 : 기존 데이터베이스 관리도구의 능력을 넘어서는 대량의 정형(수십 TB) 또는 심지어 데이터베이스형태가 아닌 비정형의 데이터 집합조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술(위키피디아)

- 빅데이터 특징 : 크기(Volume), 속도(Velocity), 다양한 유형(Variety), 가치 창출(Value)

- 즉, 빅데이터란 데이터(정형, 비정형)를 이용하여 데이터의 대용량, 속도의 증가, 다양성 등을 통해 새로운 가치 창출을 의미!

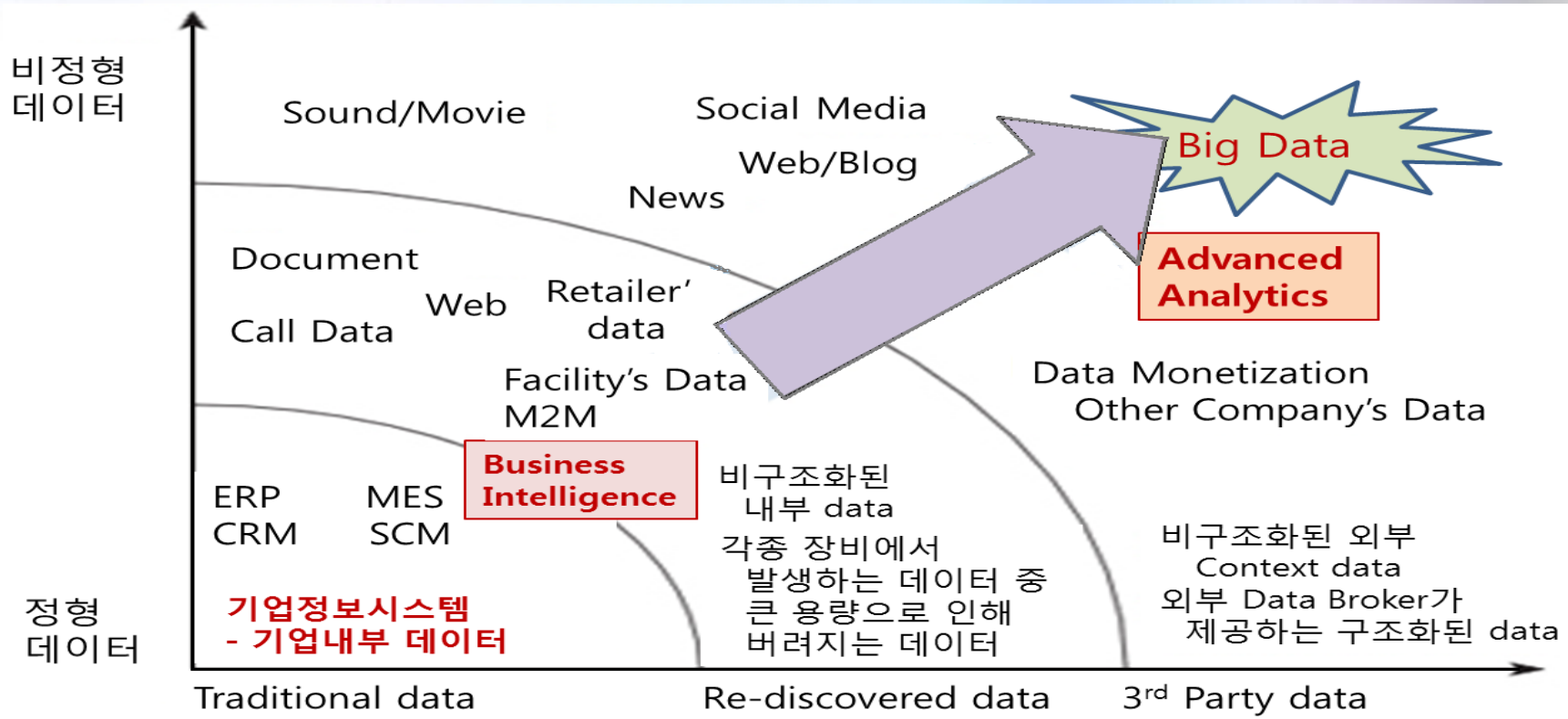
4.2 빅데이터 방법

4.2.1 빅데이터 방법 의의

데이터 수집과 데이터 이해, 데이터 활용에 대한 최적화

4.2 빅데이터 방법

4.2.1 빅데이터 방법 의의



4.2 빅데이터 방법

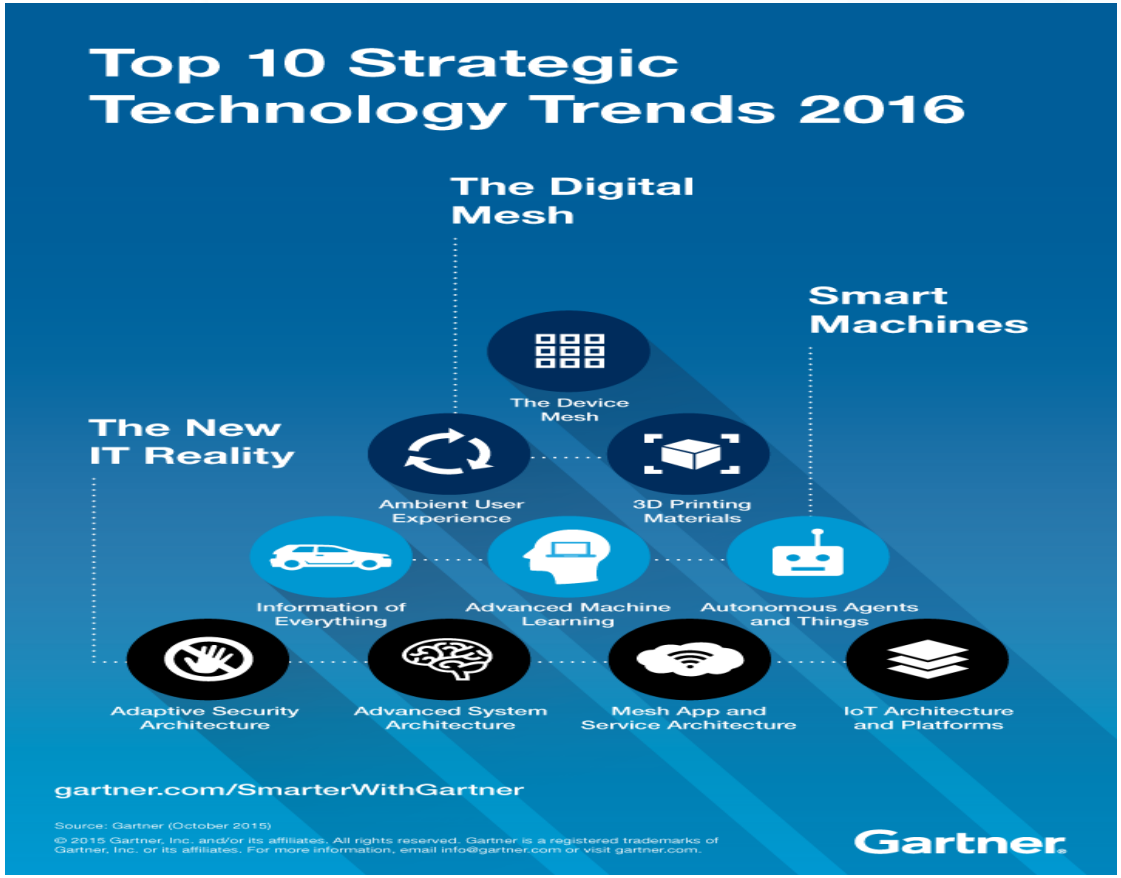
4.2.1 빅데이터 방법 의의

순위	2011년	2012년	2013년	2014년	2015년
1	클라우드 컴퓨팅	미디어 태블릿 그 이후	모바일 대전	다양한 모바일 기기관리	컴퓨팅 에브리 웨어
2	모바일 앱과미디어 태블릿	모바일 중심 애플리케이션과 인터페이스	모바일 앱&HTML 5	모바일 앱과 애플리케이션	사물인터넷
3	소셜커뮤니케이션 및 협업	상황인식과 소셜이 결합된 사용자 경험	퍼스널 클라우드	만물인터넷	3D프린팅
4	비디오	사물인터넷	사물인터넷	하이브리드 클라우드와 서비스브로커로서 IT	보편화된 첨단 분석
5	차세대 분석	앱스토어와 마켓플레이스	하이브리드 IT&클라우드 컴퓨팅	클라우드/클라이언트 아키텍쳐	콘텍스트 리치시스템
6	소셜분석	차세대 분석	전략적 빅데이터	퍼스널 클라우드 시대	스마트 머신
7	상황인식 컴퓨터	빅데이터	실용분석	소프트웨어 정의	클라우드/클라이언트 컴퓨팅
8	스토리지급 메모리	인메모리 컴퓨팅	인메모리 컴퓨팅	웹스케일 IT	소프트웨어 정의 애플리케이션과 인프라
9	유비쿼터스컴퓨팅	저전력 서버	통합생태계	스마트 머신	웹-스케일 IT
10	패브릭기반 컴퓨팅 및 인프라 스트럭처	클라우드 컴퓨팅	엔터프라이즈 앱스토어	3D 프린팅	위험기반 보안과 자가 방어

4.2 빅데이터 방법

4.2.1 빅데이터 방법 의의

2016년 Gartner 10대 전략기술(3가지) : 디지털 메시(The Digital Mesh), 스마트기계(Smart Machines), 새로운 IT세계(The New IT Reality)



4.2 빅데이터 방법

4.2.1 빅데이터 방법 의의

2016년 Gartner 10대 전략기술:

1. 디바이스 메시 - 다양한 기기들의 연결을 의미
2. 앰비언트 사용자 경험 - 위치나 시간에 관계 없는 경험을 의미
3. 3D 프린팅 재료 - 무엇이든 손쉽게 집에서 생산 할 수 있는 도구
4. 만물 정보(IOE) - IOT를 넘어 모든 것에 인터넷이 연결됨
5. 진보된 기계 학습 - 기계가 스스로 학습
6. 지능형 기기 - 기계 학습가 더불어 똑똑한 사람과 같은 서비스가 진화
7. 상황에 따른 보안 - 클라우드 및 개방형 기술발전에 따른 보안 기술
8. 진보된 시스템 아키텍처 - 다양한 연결을 전제로 고도화된 컴퓨팅 자원 필요
9. 메시 앱과 서비스 아키택처 - 클라우드 환경 및 기기와의 연결을 통합하는 기술의 필요
10. 사물인터넷 플랫폼 - 모든 센서, 기기 등을 관리, 통합, 보안이 보장된 플랫폼 전략 필요

4.2 빅데이터 방법

4.2.2 빅데이터 분석방법 종류

■ 기술통계

- 기술통계 : 측정이나 실험에서 수집한 자료의 정리, 표현, 요약, 해석 등을 통해 자료의 특성을 규명하는 통계적 방법

- 기술통계 분석방법 : 대푯값(산술평균, 중앙값, 최빈값), 산포도(분산, 표준편차, 범위, 변동계수 등), 비대칭도(왜도, 첨도), 판별분석, 주성분분석, 군집분석 등

4.2 빅데이터 방법

4.2.2 빅데이터 방법 종류

■ 추론통계

- 추론통계 : 기술통계로 어떤 모집단에서 구한 표본정보를 가지고 그 모집단의 특성 및 가능성 등을 추론해내는 통계적 방법

- 추론통계를 하는 이유

-> 모든 분석대상을 조사 하는 것은 비합리적임

소규모 집단을 가지고 조사하는 것이 경제적이고 효율적임

- 분석방법 : 상관분석, 요인분석, 주성분분석, 회귀분석 등

4.2 빅데이터 방법

4.2.2 빅데이터 방법 종류

- **텍스트마이닝**
  - 텍스트 마이닝 : 대규모의 문서(text)에서 의미 있는 정보를 추출하는 것
  - 텍스트 마이닝은 정보 검색, 데이터 마이닝, 기계 학습, 통계학, 컴퓨터 언어학 등이 결합된 학제적 분야
  - 텍스트 마이닝 분석 : 문서 분류, 문서 군집, 메타데이터 추출, 정보 추출 등

4.2 빅데이터 방법

4.2.2 빅데이터 방법 종류

- **데이터마이닝**

  - 데이터마이닝 : 자료(데이터) 저장소에 저장되어 있는 방대한 양의 데이터로부터 의사결정에 도움이 되는 유용한 정보를 발견하는 일련의 작업들의 집합을 뜻함

  - 데이터 마이닝의 주요 6가지 기법: Classification(분류), Estimation(추정), Prediction(예측) 등이 목표지향(defined) 방법

    Grouping(유사행태 집단화) or Associatioon Rules, Clustering (군집화) 등은 목표불명(undefined) 방법

    Profiling(서술 또는 설명 등의 기초분석) : 목표지향+ 목표불명

4.3 빅데이터 주요 기법

4.3.1 텍스트마이닝

1) 텍스트마이닝 개념

- 비정형, 반정형 데이터에 대하여 자연언어처리(Natural Language Processing) 기술과 문서 처리 기술을 적용하여 유용한 정보를 추출, 가공하는 것을 목적

- 데이터마이닝 방법과 정보 추출, 정보 검색, 자연어처리, 문서 요약 등의 기법들을 결합
- 텍스트 마이닝의 적용의 핵심은 대량의 텍스트로부터 과거에 알려지지 않은 숨겨진 지식을 찾아내는 것

4.3 빅데이터 주요 기법

4.3.1 텍스트마이닝

2) 텍스트마이닝 기법

- 정보 추출(Information Extraction)은 일반적인 텍스트 문서로부터 사용자가 원하는 정보를 추출하는 작업
- 문서 분류 (Text Classification)와 문서 클러스터링 (Text Clustering)은 문서들을 문서의 내용에 따라 구조화하는 전통적인 텍스트 마이닝 기법
- 토픽 트랙킹(Topic Tracking)은 사용자 프로필을 기반으로 사용자가 관심 있어 하는 문서가 어떤 문서일지를 예측하는 시스템
- 웹 마이닝(Web Mining) 텍스트 마이닝 기법을 웹 사이트에 적용하여 사용자들이 좀 더 쉽게 자신이 원하는 정보를 찾게 해주려는 시도

4.3 빅데이터 주요 기법

4.3.1 텍스트마이닝

2) 텍스트마이닝 기법

- 질의응답 시스템(Question Answering)은 사용자가 자연 언어로 질문을 던지면 시스템이 질문에 대한 대답을 제공해주는 시스템
- Concept Linkage 시스템은 각각의 문서들에서 공유되고 있는 의미를 발견하여 사용자에게 제공
- 문서요약(Summarization)은 앞서 언급된 정보 추출이 해당 문서에서 특정 관심 영역만을 문장 또는 단어의 형태로 추출하려는 시도였다면 문서 요약은 좀 더 나아가 문서에서 다룬 중요 내용을 글로 요약하려는 시도
- Duo-mining은 데이터 마이닝과 텍스트 마이닝을 함께 적용하려는 시도

4.3 빅데이터 주요 기법

4.3.1 텍스트마이닝

3) 텍스트마이닝 응용

- CRM은 고객과의 관계를 분석하고 효율적으로 잘 유지하기 위한 기업 차원의 모든 활동을 나타내는 단어
- 안보분야는 9.11 테러 사태 이후 테러리즘에 대한 우려가 커지면서 텍스트 마이닝이 안보 분야에 활발히 응용
- IBM의 WebFountain은 IBM에서 개발한 텍스트마이닝 기법으로 웹 페이지, 이메일, 채팅과 같은 온라인 데이터를 컴퓨터가 분석 할 수있는 포맷으로 변경하여 저장하는 시스템

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

1) 데이터마이닝 정의

- 단순업무 처리를 위해 보관되어 있는 데이터를 분석목적에 적합한 데이터형태로 변화하여 분석하도록 해주는 것
- 일반 업무 목적을 위해 구축된 데이터를 연구나 지식경영을 목적으로 분석할 수 있도록 데이터를 쉽게 정제하고, 통계전문가가 아니더라도 쉽게 정제하고, 분석할 수 있도록 해주는 기법

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

1) 데이터마이닝 정의

- **데이터 마이닝 개념**

- "대량의 데이터 집합으로부터 유용한 정보를 추출하는 것"(Hand et al., 2001)
- "데이터마이닝이란 의미 있는 패턴과 규칙을 발견하기 위해서 자동화되거나 반자동화 된 도구를 이용하여 대량의 데이터를 탐색하고 분석하는 과정(Berry and Linoff, 1997, 2000)."
- "데이터마이닝은 통계 및 수학적 기술뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미 있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정(2004년 1월 가트너 그룹 웹사이트).

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

1) 데이터마이닝 정의

- 데이터 마이닝 특징

① 대용량의 자료로 데이터마이닝은 대용량의 자료(Observational Data) 사용

② 컴퓨터 중심적 기법(Computer-Intensive Algorithms)으로 데이터마이닝은 기존의 표본에 의한 통계적 추론에서 정보기술의 발전과 함께 방대한 데이터를 처리할 수 있는 컴퓨터의 능력을 활용

③ 다차원적 계보(Multidisciplinary Lineage)는 데이터마이닝은 통계학, 데이터베이스, 패턴인식, 기계학습, 인공지능, KDD의 여러 분야를 망라

④ 경험적 방법(Adhockery Method)은 이론적 원리보다는 경험에 기초하였기 때문에 데이터마이닝 기법은 수리적으로 밝혀지지 않은 것

⑤ 일반화(Generalization)란 예측모형(Prediction Model)이 새로운 자료(New Data)에 얼마나 잘 적용되도록 하는가

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

2) 데이터마이닝 기법

- 데이터마이닝은 지식, 경제, 경영에 관한 대부분의 문제들이 다음과 같은 6개의 영역으로 구분
- 분류(Classification)는 데이터마이닝 작업에서 가장 보편적

  : 의사결정나무분석 등
- 추정은 목표변수가 이산형인 것을 주로 다룸

  : 추정은 분류작업을 하는데 주로 사용
- 예측은 “어떤 개체가 장차 어디로 분류될까?” 혹은 “장차 우리의 고객이 될 확률은 얼마일까?”처럼 예측은 단지 미래에 대한 것이라는 것만 제외하면 분류나 추정과 같음

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

2) 데이터마이닝의 기법

- 유사 집단화(Affinity Grouping)란 유사한 성격을 갖는 사물이나 물건들을 함께 묶어주는 작업

  : 장바구니분석

- 군집화(Clustering)란 이질적인 원소로 구성되어 있는 모집단을 여러 개의 동질적인 하위그룹 혹은 군집(Cluster)로 나누어 세분화하는 것
- 서술은 아주 복잡한 데이터베이스에 내재하는 사람이나 제품 혹은 데이터를 생성한 과정에 대하여 우리의 이해를 증진시킬 수 있는 뭔가를 찾아 서술하는 것

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

3) 데이터마이닝 관련분야

- 데이터마이닝은 KDD(Knowledge Discovery In Database), 즉 지식발견의 전 과정인 데이터 선정, 정제, 코딩 및 여러 가지 패턴 발견 중 '탐사' 단계에 해당되는 방법
- 패턴인식(Pattern Recognition) 기술은 다양한 형태의 데이터에 포함되어 있는 패턴정보를 자동으로 추출하여 응용하는 기술로서 크게는 컴퓨터 응용 소프트웨어 기술, 작게는 인공지능 기술

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

3) 데이터마이닝의 관련분야

- 기계학습 알고리즘(Machine Learning Algorithm)은 자연 언어 처리 시스템에 인간의 사고능력에 해당하는 추론 기능을 추가하여, 이용하기 편리하면서도 수준 높은 처리가 가능한 지적 시스템을 만들기 위해 연구하여 만든 알고리즘이 기계·학습 알고리즘

- 전문가 시스템(Expert System)은 질병의 진단이나 지질 자료 등을 분류하는데 사용되는 특정된 분야에 국한된 전문가의 지식을 매우 간단한 'IF-THEN'규칙을 사용하여 표현될 수 있는데 기계·학습 알고리즘을 근간

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

3) 데이터마이닝의 관련분야

- 통계학(Statistic)은 회귀분석, 판별분석, 군집분석 등 주로 세 가지 기법
- 데이터 웨어하우징(Data Warehousing)은 운영 데이터로부터 추출된 데이터를 중앙으로 집중하여 저장하는 것
- OLAP(Online Analytical Processing)는 다차원 분석 시 사용되는 도구

4.3 빅데이터 주요 기법

4.3.2 데이터마이닝

4) 빅데이터 주요 분석기법 분류

- 기술통계분석 기법 : 빈도분석, 교차분석, 요인분석, 신뢰도분석 등
- 추론통계분석 기법 : T-test분석, 분산분석, 상관분석, 선형회귀분석, 로지스틱회귀분석, 경로분석, 구조방정식모형분석, 메타분석, 시계열분석 등
- 데이터마이닝 기법 : 분류분석, 군집분석, 판별분석, 연관성분석, 대응분석, 다차원척도법, 생존분석, 신경망분석 등
- 텍스트마이닝 기법 : 워드클라우드분석, 네트워크분석, 감성분석, 시각화분석, SNS 텍스트분석, 패턴분석, 매핑과 공간분석 등

4.4 빅데이터 분석 주요기법

빅데이터 분석방법 분류

- 기술통계 : 대표값, 산포도, 빈도분석, 교차분석, 신뢰도분석 등
- 추론통계 : T-test분석, 분산분석, 상관관계분석, 선형회귀분석, 로지스틱회귀분석, 경로분석, 구조방정식 모형분석, 메타분석, 시계열분석 등
- 데이터마이닝 : 분류분석, 판별분석, 군집분석, 요인분석, 주성분분석, 대응분석, 다차원척도법, 생존분석, 신경망분석 등
- 텍스트마이닝 : 워드클라우드분석, 네트워크분석, 감성분석, 연관성분석, SNS텍스트분석, 패턴분석, 매핑과 공간분석 등