

## 1. 빅데이터 개념

"Big data란 대량의 다양한 정보뿐만 아니라 신속한 정보의 수집 및 처리속도를 갖추는 것, 나아가 데이터의 불확실성을 인지하고 분석을 통한 비즈니스 가치를 찾아내는 것이라 할 수 있다."

표 1-4 전통적 데이터와 빅데이터의 특징 비교 [07]

구분	전통적 데이터	빅데이터
데이터 원천	전통적 정보 서비스	일상화된 정보 서비스
목적	업무와 효율성	사회적 소통, 자기표현, 사회 기반 서비스
생성 주체	정부 및 기업 등 조직	개인 및 시스템
데이터 유형	<ul style="list-style-type: none"> <li>정형 데이터</li> <li>조직 내부 데이터(고객 정보, 거래 정보 등)</li> <li>주로 비공개 데이터</li> </ul>	<ul style="list-style-type: none"> <li>비정형 데이터(비디오 스트림, 이미지, 오디오, 소셜 네트워크 등 사용자 데이터, 센서 데이터, 응용 프로그램 데이터 등)</li> <li>조직 외부 데이터</li> <li>일부 공개 데이터</li> </ul>
데이터 특징	<ul style="list-style-type: none"> <li>데이터 증가량 관리 가능</li> <li>신뢰성 높은 핵심 데이터</li> </ul>	<ul style="list-style-type: none"> <li>기하급수로 양적 증가</li> <li>쓰레기(Garbage) 데이터 비중 높음</li> <li>문맥 정보 등 다양한 데이터</li> </ul>
데이터 보유	정부, 기업 등 대부분 조직	<ul style="list-style-type: none"> <li>인터넷 서비스 기업(구글, 아마존 등)</li> <li>포털(네이버, 다음 등)</li> <li>이동 통신 회사(SKT, KTF 등)</li> <li>디바이스 생산 회사(애플, 삼성전자 등)</li> </ul>
데이터 플랫폼	정형 데이터를 생산·저장·분석·처리할 수 있는 전통적 플랫폼 예) 분산 DBMS, 다중처리기, 중앙 집중 처리	비정형 대량 데이터를 생산·저장·분석·처리할 수 있는 새로운 플랫폼 예) 대용량 비정형 데이터 분산 병렬 처리

## 2. 빅데이터의 구성

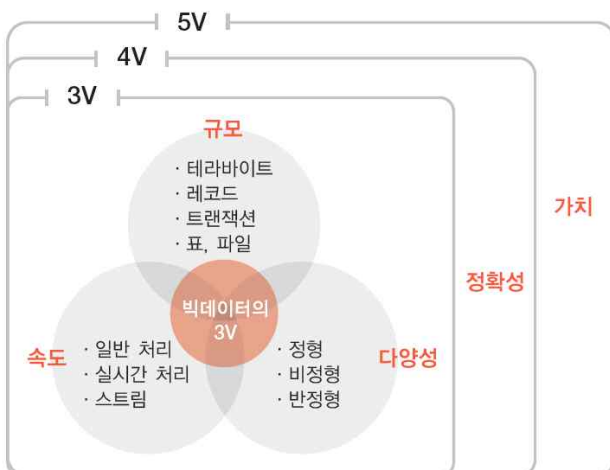


그림 1-2 빅데이터의 속성 [02]

- ① Volume(양) : 물리적인 크기와 개념적인 범위까지 대규모인 데이터의 양
- ② Velocity(속도) : 실시간으로 생산되며 유통 속도 또한 매우 빠른 데이터

- ③ **Variety(다양성)** : 기존의 구조화된 정형 데이터는 물론 사진, 동영상 등의 비정형 데이터가 포함
- ④ **Veracity(정확성)** : 데이터의 원천 및 형태 다양화에도 불구하고 신뢰성이 보장된 데이터
- ⑤ **Value(가치)** : 새로운 가치를 창출할 수 있는 데이터

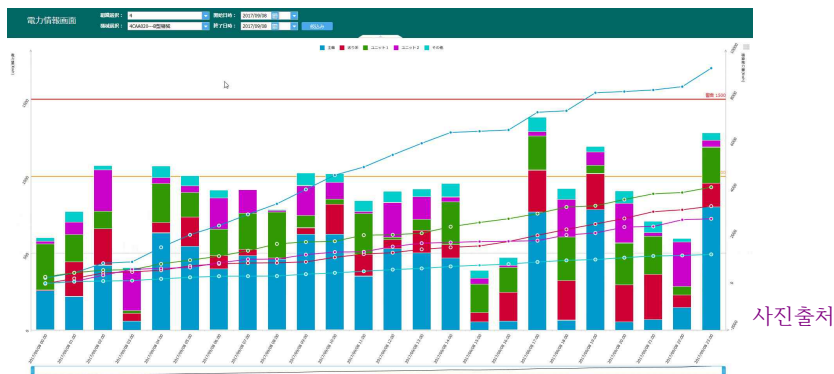
### 3. 빅데이터 기술

1.데이터소스	2.수집	3.저장	4.처리	5.분석	6.표현
<b>-내부데이터</b> 데이터베이스, 파일 관리 시스템  <b>-외부데이터</b> 인터넷으로 연결된 파일, 멀티미디어	<b>-크롤링</b> 검색엔진의 로봇사용 데이터 수집  <b>-ETL</b> 소스 데이터의 추출·전송·변환·적재	<b>-NoSQL데이터베이스</b> 비정형 데이터 관리  <b>-스토리지</b> 빅데이터 저장  <b>-서버</b> 초경량 서버	<b>-맵리듀스</b> 데이터 추출  <b>-프로세싱</b> 다중업무처리	<b>-NLP</b> 자연어처리  <b>-기계학습</b> 머신러닝으로 데이터 패턴 발견  <b>-직렬화</b> 데이터 간의 순서화	<b>-가시화</b> 데이터를 도표나 그래픽적으로 표현  <b>-획득</b> 데이터의 획득 및 재해석

### 4. 빅데이터 활용

#### ① 제조업

- 설비보전: 센서 데이터의 가시화를 통하여 공장전체, 라인별 설비의 운전상황, 고장이 많은 설비의 파악, 설비 문제를 조기 발견 가능합니다.
- 제품 트래킹: 바코드 스캐너와 무선장비를 이용하여 원자재 조달에서 생산, 소비 또는 폐기에 이르기까지 추적할 수 있습니다.



#### ② 의료업

- 유행병의 발생을 예측하고 영향을 최소화하기 위한 예방책을 강구
- 수백만 명의 환자로부터 수집된 엑셀 데이터를 근거에 따라 진단, 웨어러블 디바이스를 사용하여 빅데이터가 환자의 건강상태를 감시하고 의사에게 보고



코로나 현황 대시보드

### ③ 금융업

- 클라우드 컴퓨팅으로 리스크를 관리하여 효율 향상
- 고객 데이터 수집, 분석을 통해 서비스 customizing
- Clustering + Association 데이터 분석 기법을 사용하여 지점 장소 선정 등 중요의사결정

### ④ 소매업

- 판매 데이터를 패턴화해 고객을 분류하고 marketing customizing
- 예측 분석을 통해 공급과 수요 예측의 정확도를 높임
- 히트 상품과 데드 셀러를 분석해서 재고 예측 -> 재고 보유 비용의 영향 최소화
- 

## 5. 빅데이터 해결과제

### [기술관점]

- 데이터 스토리지 : 약 2년마다 2배가량 증가하는 데이터 양의 증가를 따라잡을 데이터 스트리지 기술의 개발 필요
- 데이터 큐레이션 : 클라이언트에게 의미있는 데이터로 정제하기 위해 작업 시간의 50~80%를 사용

### [사용자관점]

- 개인 정보 보안 : 보호 조치 없이 정보활용에만 초점이 맞춰져 가명처리된 고객 정보를 정보 주체의 동의 없이 판매, 공유, 결합할 가능성이 증가 (ex. 카카오맵 어플리케이션 신상 정보 노출, 데이터 3법 논란)
- 국내 기업의 낮은 기술 도입률 : 빅데이터라고 부를 만한 데이터가 없음, 중견 중소기업 일수록 활용률 저조 ( ↓ barplot. 빅데이터 분석 기술을 도입하지 않은 이유)

