빅 데이터

컴퓨터공학개론

1) IBM 연구 자료 (2012년)

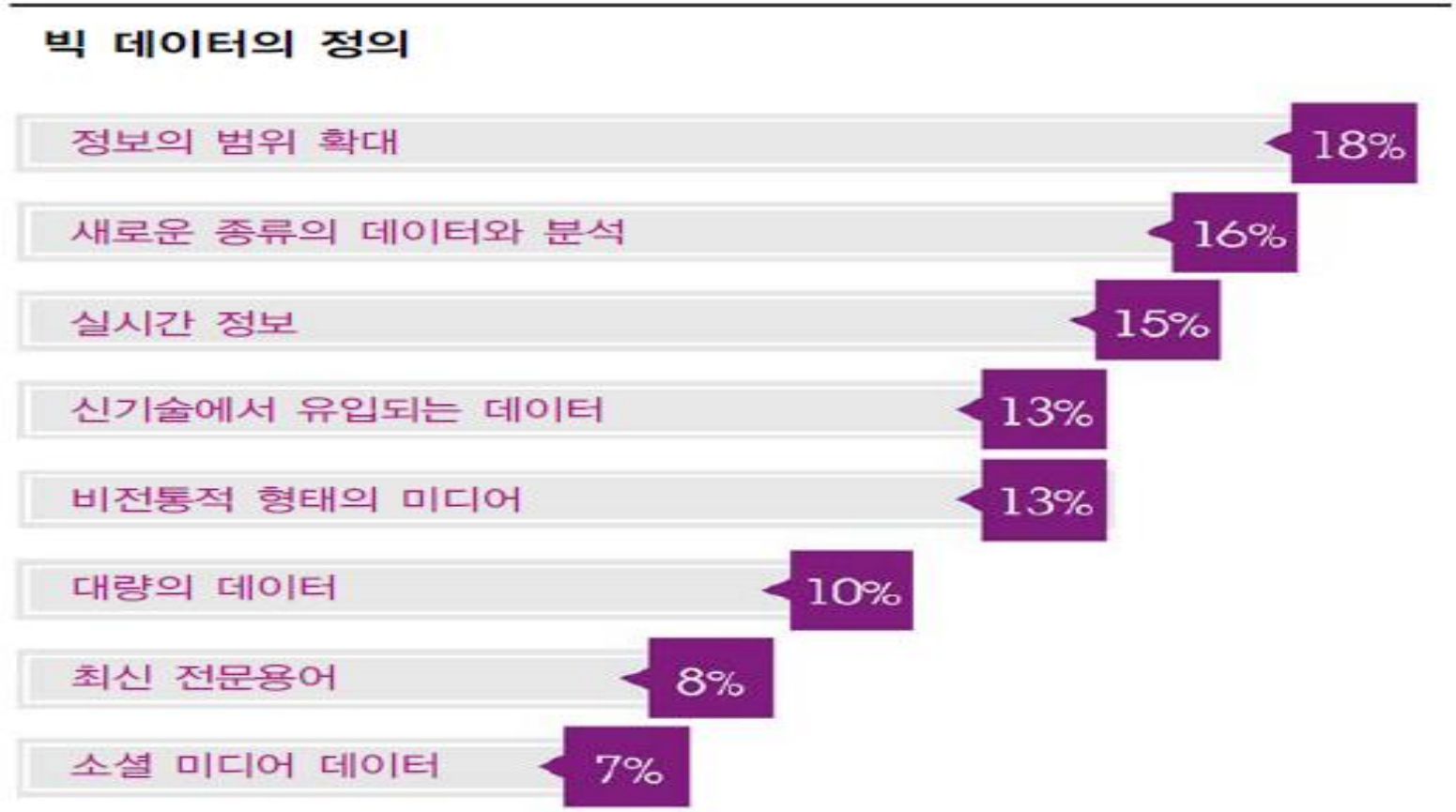
2) KT중앙연구소 자료 (2012년)

분석: 빅 데이터의 현실적인 활용 (IBM, 2012)

“빅 데이터(Big data)”

- 기술 분야뿐 아니라 다양한 분야에서 여러 가지 의미로 사용
- 근래에 비즈니스 분야의 주목
  - 빅 데이터가 글로벌하게 통합된 환경에서 상거래에 막대한 영향을 미칠 수 있기 때문
- 기업들이 오랫동안 고민해온 난제에 대한 해결책을 제공해 줄뿐 아니라, 프로세스와 조직, 산업 전반, 심지어 사회 자체를 변화시킬 수 있는 새로운 방법까지 제시

빅 데이터의 정의

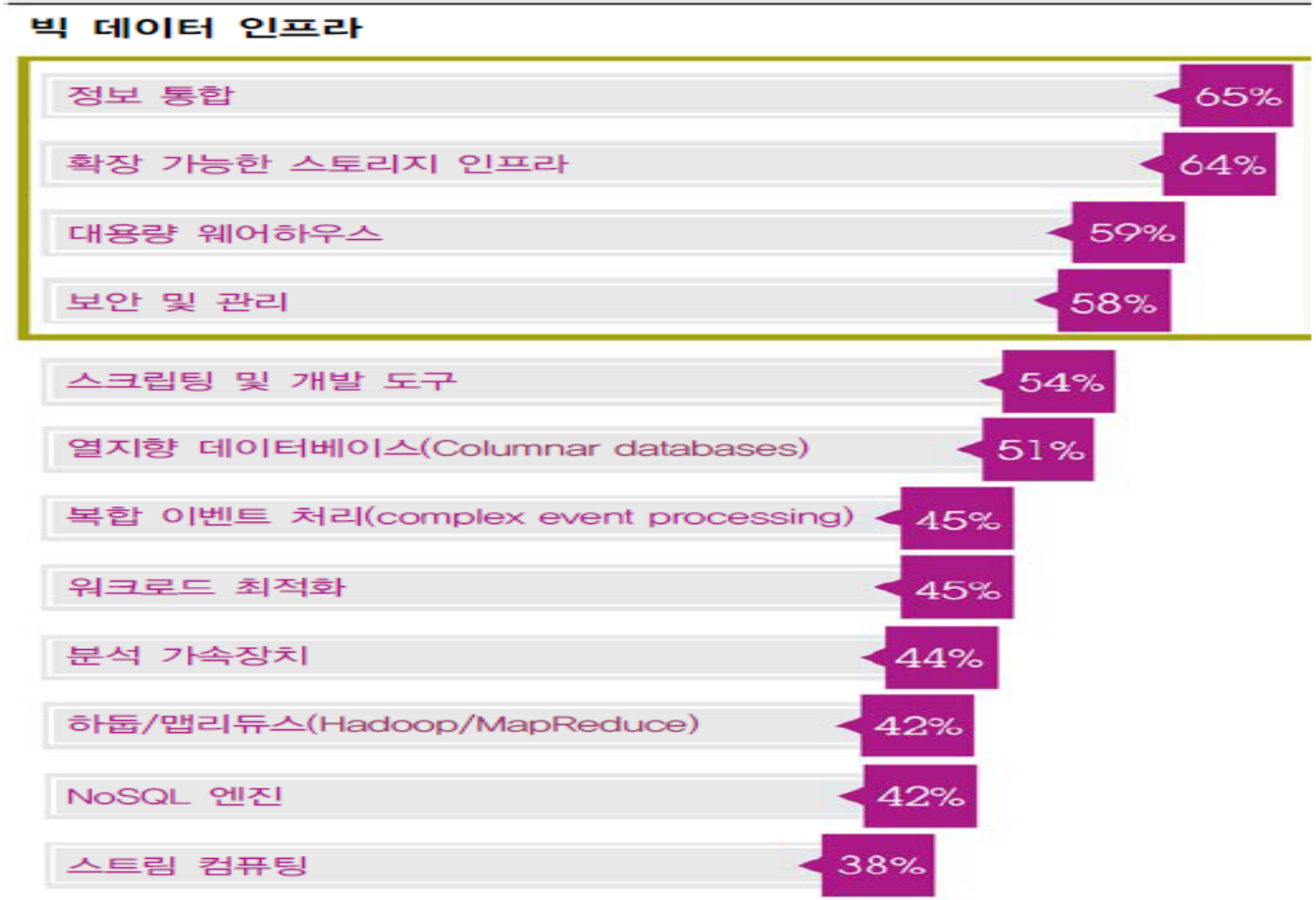


보기 중에서 응답자들이 생각하는 빅 데이터의 정의를 최대 두 가지씩 선택하도록 하였으며 나머지 보기는 생략됨. 응답 비율은 총 합이100% 가 되도록 표준화.
총 응답자 수 = 1.144명.

빅데이터의 4가지 차원

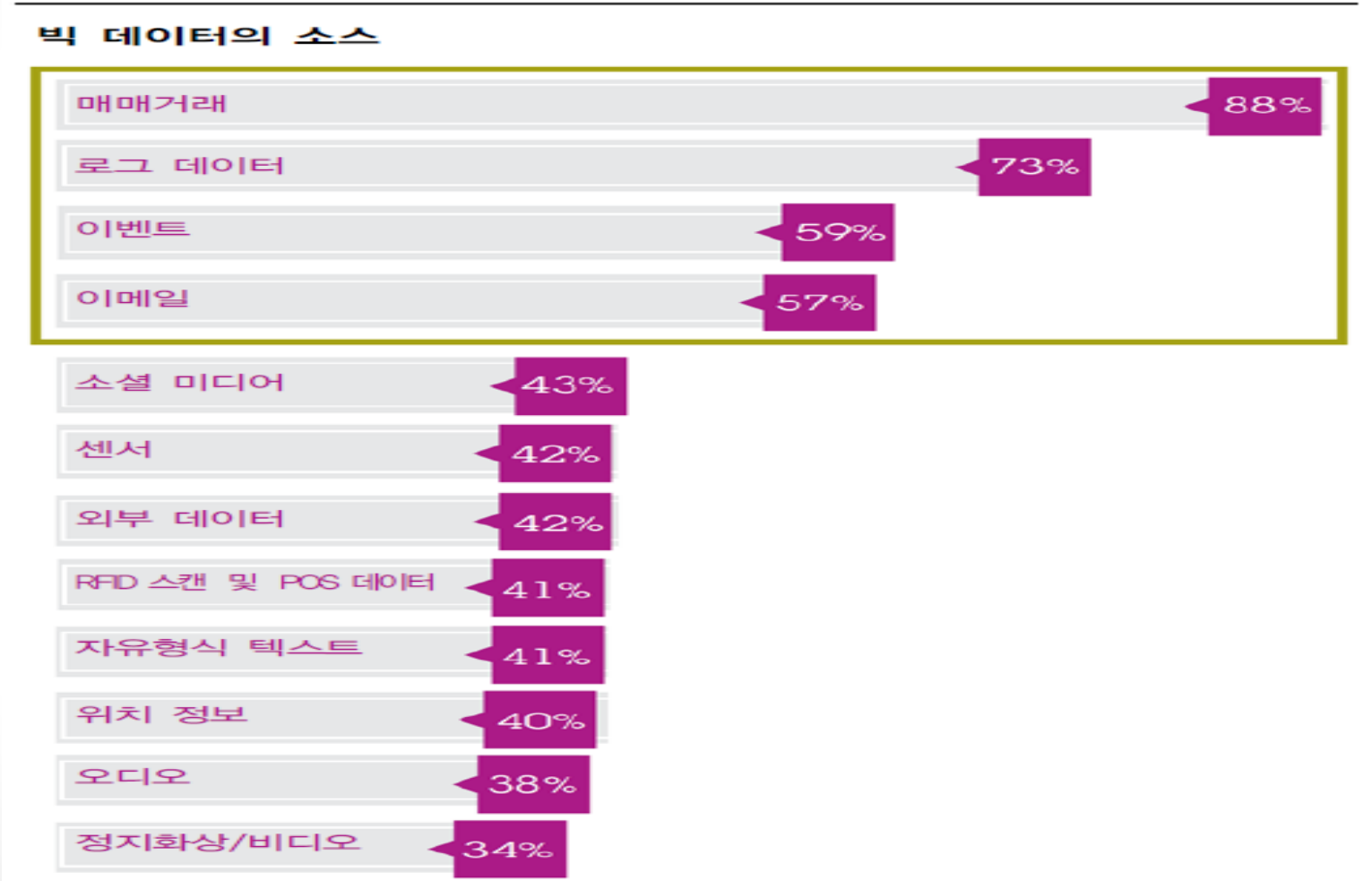

빅 데이터의 4가지 차원
규모
데이터의 양
TB부터 PB
정도의 데이터
다양성
데이터의 다양한 형태
정형/비정형
텍스트, 멀티미디어
속도
데이터의 이동
몇 분의 1초 사이에 의사결정을
가능하게 해주는 스트리밍
데이터 분석
정확성
데이터의 불확실성
본질적으로 불확실한 데이터 유형의 신뢰성
과 예측 가능성 관리

빅데이터 인프라 요소



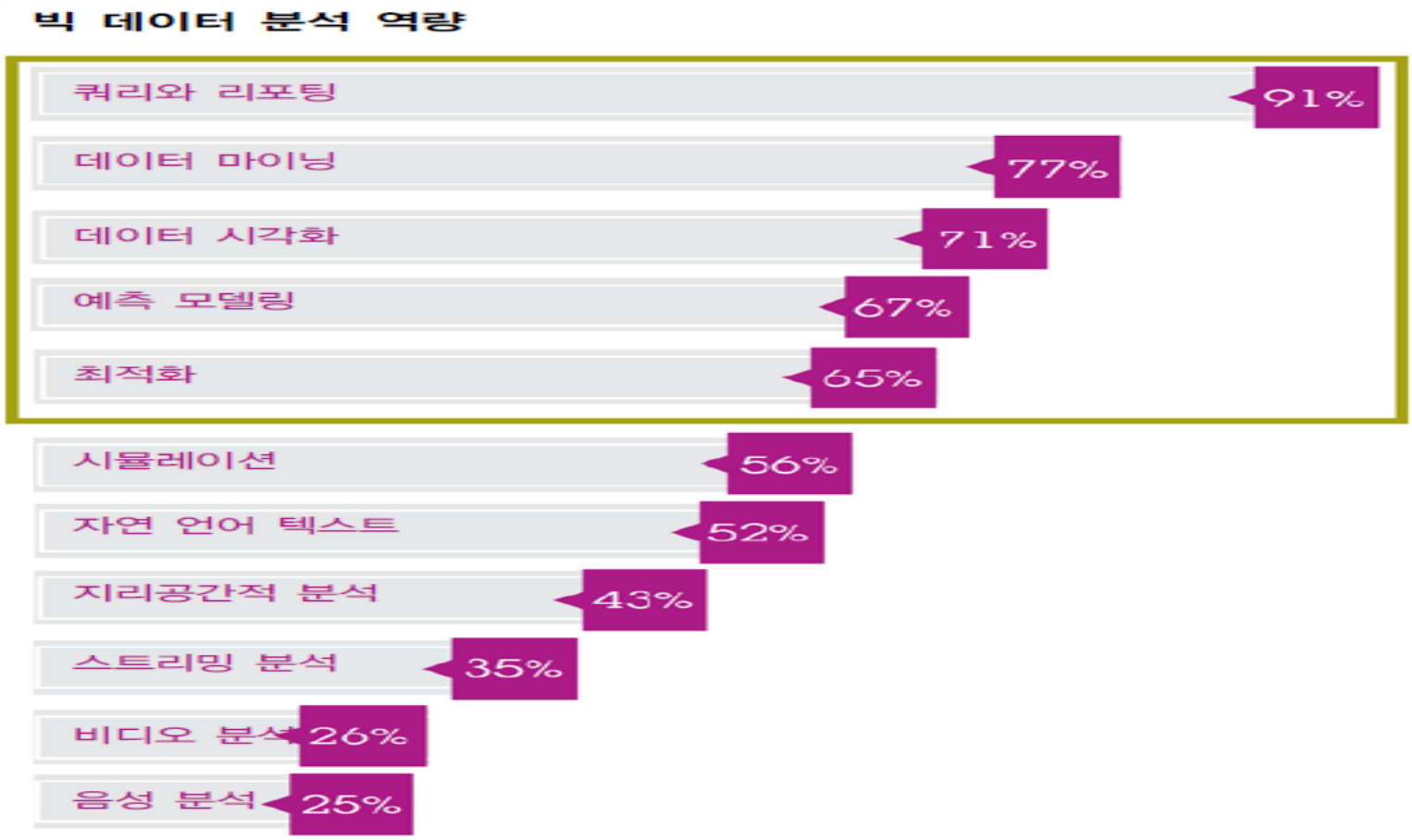
빅 데이터를 활용 중인 기업들이 현재 시범 운영 혹은 아키텍처에 통합 운영하고 있는 플랫폼 요소. 각각의 데이터 포인트는 독립적으로 수집됨. 각 데이터 포인트에 대한 총 응답자 수는 297 명에서 351명 사이.

빅데이터 소스



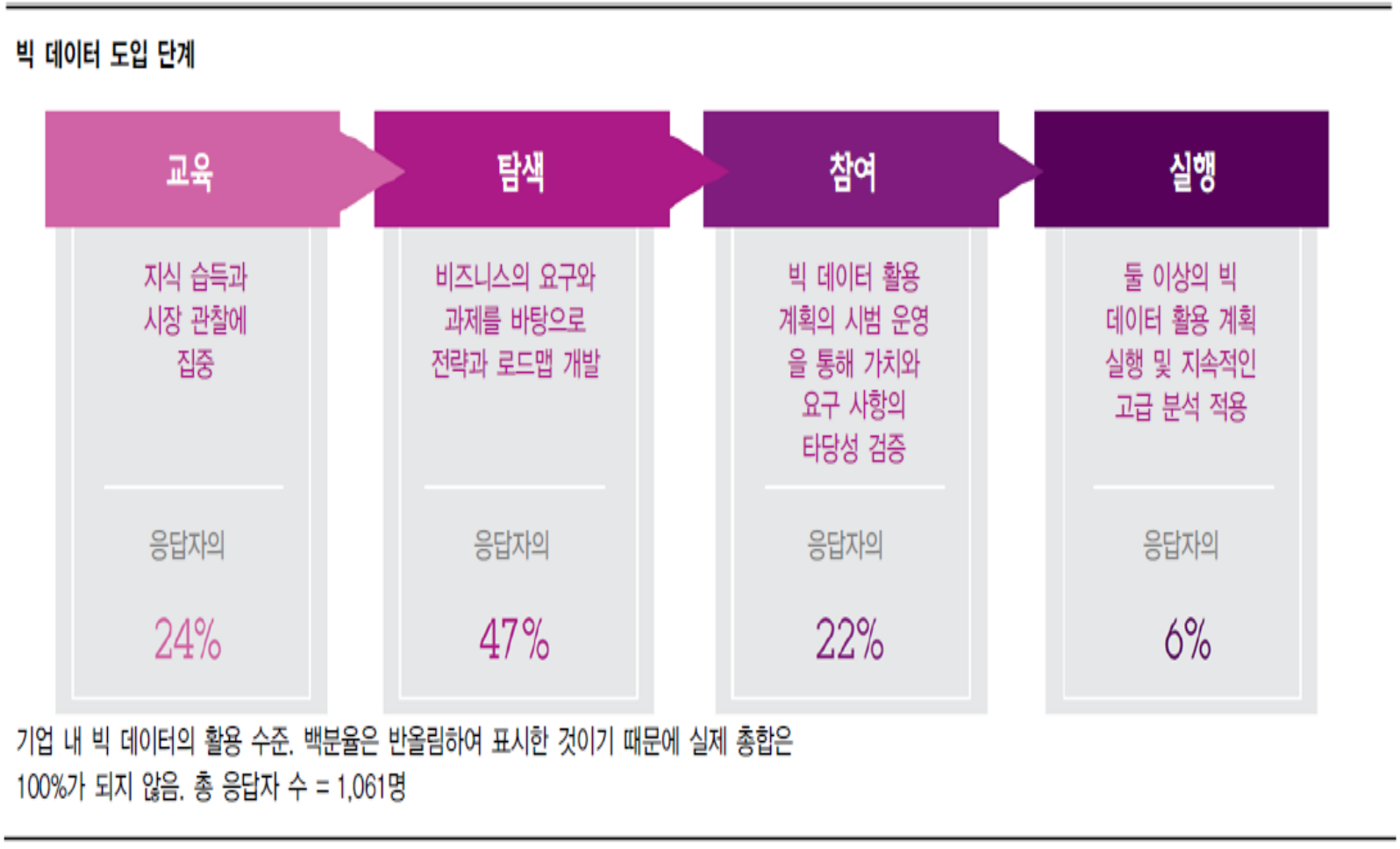
빅 데이터 활용을 위해 현재 수집 및 분석 중인 데이터 소스. 각 데이터 포인트는 독립적으로 수집됨. 각 데이터 포인트에 대한 총 응답자 수는 557명부터 867명 사이.

빅데이터 분석 역량

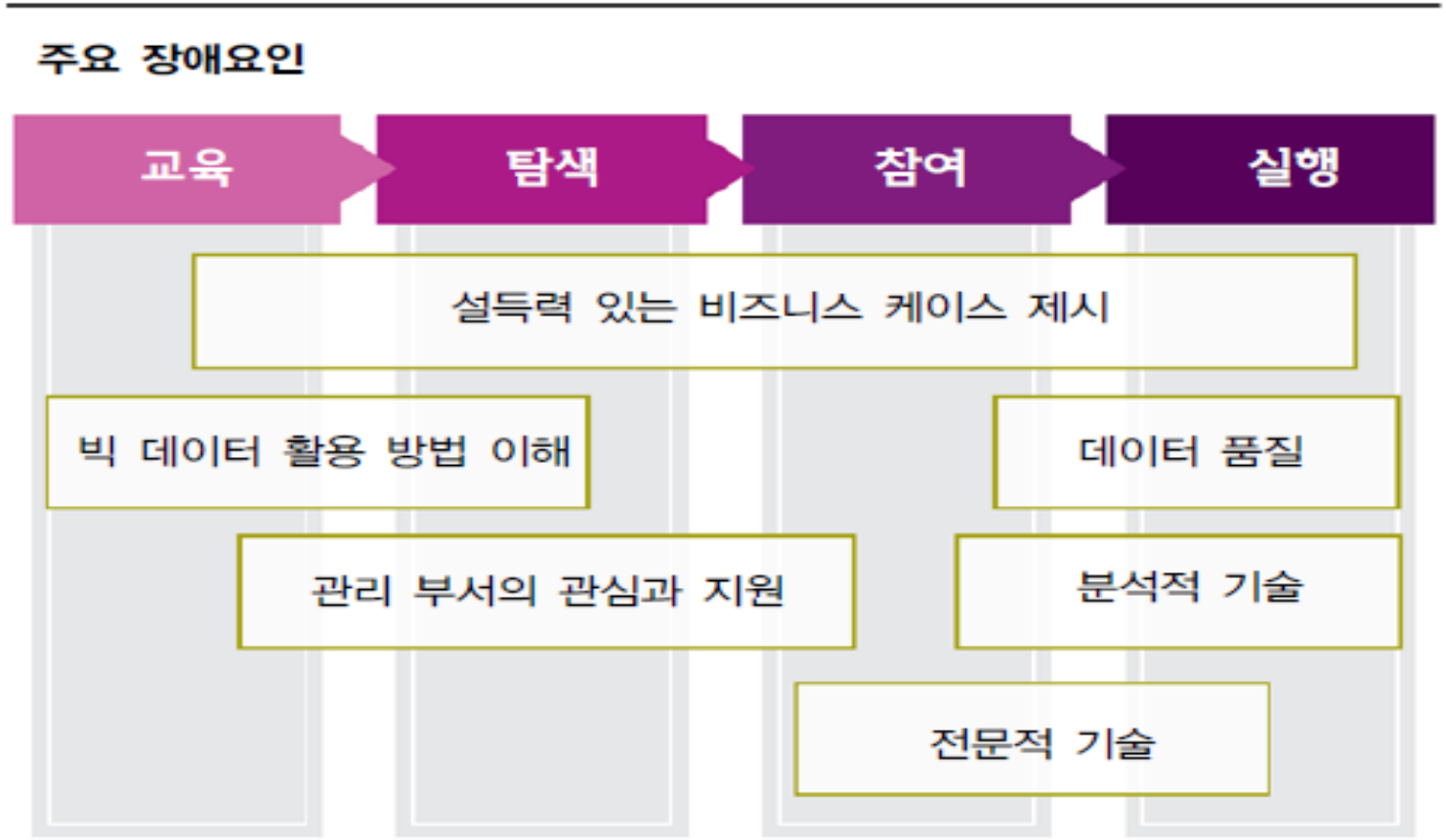


빅 데이터를 활용 중인 기업들이 이용하는 분석 기능. 각 데이터 포인트는 독립적으로 수집됨. 각 데이터 포인트에 대한 총 응답자 수는 508명부터 870명 사이.

빅데이터 도입 단계(2012년)



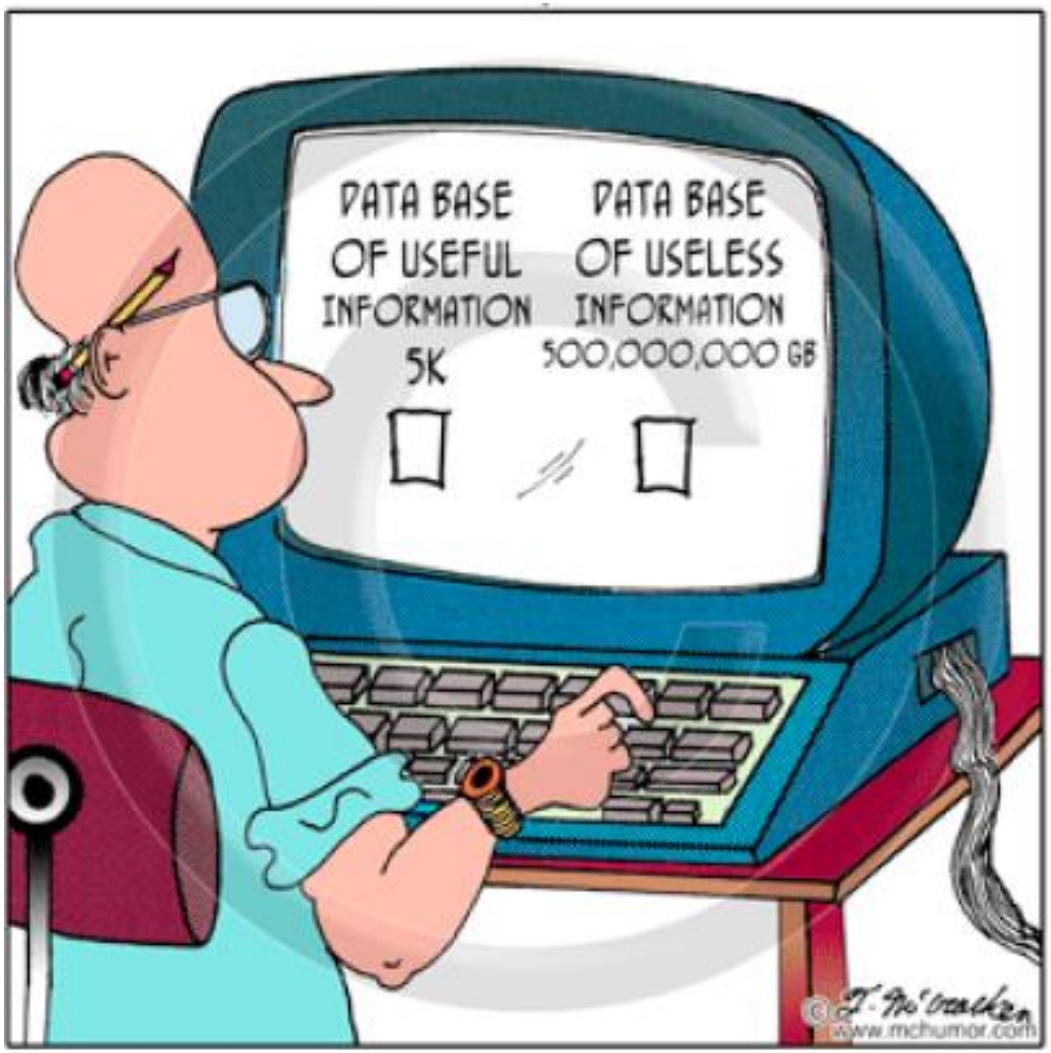
빅데이터 도입의 장애요인



빅 데이터 도입의 가장 큰 과제. 박스의 위치는 각 단계에서 해당 과제가 나타나는 빈도를 나타냄. 총 응답자 수 = 1,062명

빅 데이터의 시대의 기술
(2012년 KT 중앙연구소)

시대의 화두 : 빅데이터


DATA BASE
OF USEFUL
INFORMATION
5K
DATA BASE
OF USELESS
INFORMATION
500,000,000 GB
www.mchumor.com

Big Data란 무엇인가?

- McKinsey (2011년 5월) 보고서 'Big Data : The Next Frontier for Innovation, Competition, and Productivity'에서
  - "빅 데이터의 정의는 기존 데이터베이스 관리 도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 데이터셋(Dataset) 규모로, 그 정의는 주관적이며 앞으로도 계속 변화될 것
  - 데이터량 기준에 대해 산업분야에 따라 상대적이며 현재 기준에서는 몇 십 테라바이트에서 수 페타바이트까지가 그 범위이다"라고 설명

데이터의 종류

- 정형(Structured) :
  - 고정된 필드에 저장된 데이터
  - 관계형 데이터베이스 및 스프레드시트 등
- 반정형(Semi-Structured)
  - 고정된 필드에 저장되어 있지는 않지만, 메타데이터나 스키마 등을 포함하는 데이터
  - XML이나 HTML 텍스트 등
- 비정형(Unstructured)
  - 고정된 필드에 저장되어 있지 않은 데이터.
  - 텍스트 분석이 가능한 텍스트 문서 및 이미지/동영상/음성 데이터 등

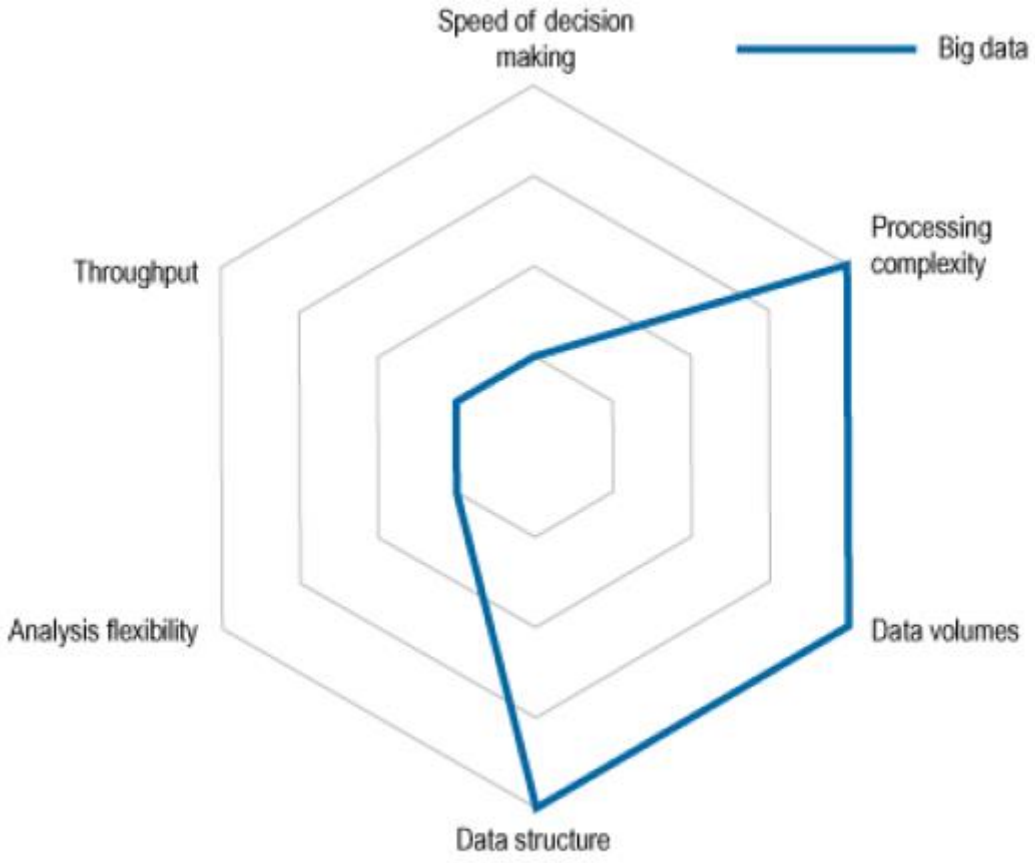
기존 데이터 처리와 빅 데이터 처리의 차이점 (1)

- 빠른 의사결정이 상대적으로 덜 요구된다
  - 대용량 데이터에 기반한 분석 위주로, 장기적/전략적 접근이 필요
  - 기존의 데이터 처리에 요구되는 즉각적인 처리속도와는 달리, 즉각적인 의사결정이 상대적으로 덜 요구됨
- 처리(Processing) 복잡도가 높다
  - 다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리 등으로 인해 처리 복잡도가 매우 높으며, 이를 해결하기 위해 통상적으로 분산 처리 기술 필요
- 처리할 데이터양이 방대하다
  - 클릭스트림Clickstream 데이터를 예로 들면, 고객 정보수집 및 분석을 장기간에 걸쳐 수행해야 하므로 기존 방법과 비교해 처리해야 할 데이터양은 방대하다.

기존 데이터 처리와 빅 데이터 처리의 차이점 (2)

- 비정형 데이터의 비중이 높다
  - 소셜 미디어 데이터, 로그 파일, 클릭스트림 데이터, 콜 센터 로그, 통신 CDR 로그 등
  - 처리의 복잡성을 증대시키는 요인
- 처리/분석 유연성이 높다
  - 잘 정의된 데이터 모델/상관관계/절차 등이 없어, 기존 데이터 처리방법에 비해 처리/분석의 유연성이 높은 편
  - 새롭고 다양한 처리 방법의 수용을 위해, 유연성이 기본적으로 보장되어야 함
- 동시처리량Throughput이 낮다
  - 대용량 및 복잡한 처리를 특징으로 하고 있어, 동시에 처리가 필요한 데이터양은 낮다.
  - 따라서 (준)실시간 처리가 보장되어야 하는 데이터 분석에는 적합하지 않음

기존 데이터 처리와 빅 데이터 처리의 차이점 (3)



Big Data를 위한 분석기법 (1)

- Text Mining
  - 비/반정형 텍스트 데이터에서 자연어 처리(Natural Language Processing) 기술에 기반하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술
  - 텍스트 마이닝 기술을 통해 방대한 텍스트 뭉치에서 의미 있는 정보를 추출해내고, 다른 정보와의 연계성을 파악하며, 텍스트가 가진 카테고리를 찾아내는 등, 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있음
  - 주요 응용분야
    - 문서 분류Document Classification
    - 문서 군집Document Clustering
    - 정보 추출Information Extraction
    - 문서요약Document Summarization 등

Big Data를 위한 분석기법 (2)

- Opinion Mining, 혹은 평판 분석(Sentiment Analysis)
  - 소셜미디어 등의 정형/비정형 텍스트의 긍정Positive, 부정Negative, 중립Neutral의 선호도를 판별하는 기술
  - 특정 서비스 및 상품에 대한 시장규모 예측, 소비자의 반응, 입소문 분석(Viral Analysis) 등에 활용
  - 전문가에 의한 선호도를 나타내는 표현/단어 자원의 축적이 필요
- Social Network Analytics
  - 수학의 그래프 이론Graph Theory에 뿌리
  - 소셜 네트워크 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크 상에서 입소문의 중심이나 허브Hub 역할을 하는 사용자를 찾는데 주로 활용
  - 소셜 네트워크 상에서 영향력이 있는 사용자를 인플루언서Influencer라고 부르는데, 인플루언서의 모니터링 및 관리는 마케팅 관점에서 중요

Big Data를 위한 분석기법 (3)

- Cluster Analysis
  - 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 군Group을 발굴하는데 사용
  - 관심사나 취미에 따른 사용자군을 군집분석을 통해 분류

Big Data 분석 인프라 기술

- Hadoop
  - 오픈 소스 분산처리기술 프로젝트
- R
  - 통계 계산 및 시각화를 위한 언어 및 개발환경을 제공
  - 기본적인 통계기법부터 모델링, 최신 데이터 마이닝 기법까지 구현/개선이 가능
  - 구현한 결과는 그래프 등으로 시각화할 수 있으며, Java나 C, Python 등의 다른 프로그래밍 언어와 연결도 용이
- NoSQL
  - Not-Only SQL, 혹은 No SQL을 의미
  - 전통적인 관계형 데이터베이스(RDBMS)와 다르게 설계된 비관계형 데이터베이스를 의미
  - 대표적인 NoSQL 솔루션
    - Cassandra, Hbase, MongDB