빅데이터 기획/분석 개념 및 실습

경남대학교 전 하 용

강의 목차

1. 빅데이터 기획 및 분석
2. 빅데이터 분석 기술 및 기법
3. 데이터 분석 실습
4. 데이터 분석 결과보기
5. 오픈 마켓 데이터 분석을 통한 제품 추천

빅데이터 기획 및 분석

1. 데이터
2. 데이터 베이스
3. 빅데이터
4. 생산 주체에 따른 데이터 분류
5. 데이터 처리
6. 빅데이터 분석
7. 빅데이터 분석 도구
8. 빅데이터 분석 기법

데이터

- 데이터 정의
  - 어떠한 사실, 개념, 명령 또는 과학적인 실험이나 관측 결과로 얻은 수치나 정상적인 값 등 실체의 속성을 숫자, 문자, 기호 등으로 표현한 것
  - 최근 데이터의 의미가 객관적인 사실 뿐만 아니라 이를 통해 재생산, 추론, 예측, 전망, 추정을 위한 근거로 기능하는, 다른 객체와의 상호관계, 당위성을 포함하는 개념을 발전
  - 데이터 = 객관적 사실 + (다른 객체와의 관계가 맺어질 수 있음을 내포)
  - 데이터 크기 : Bit, Btye, KB, MB, GB, TB, PB, EB, ZB, YB
  - 데이터 분류
    - 정성적 데이터(비정형 데이터) : 언어, 문자 등으로 표현되어 저장, 검색, 분석 등의 활용에 상대적으로 비용과 투자가 수반하는 데이터
    - 정량적 데이터(정형 데이터) : 수치, 도형, 기호 등으로 표현되어 저장, 검색, 분석 등의 활용이 편리한 데이터

데이터베이스

- 데이터베이스 등장 1950s
  - 데이터의 기지라는 뜻으로 데이터베이스 용어 탄생
- 1960s : 데이터베이스의 공식적 사용과 이를 관리하는 데이터베이스 관리시스템의 개념 등장
- 1970s : 데이터베이스 단어의 일반화 데이터베이스 정의 데이터베이스
  - 체계적으로 정리되고 개별적 접근이 가능한 데이터 집합 및 이를 관리하는 시스템을 포함
  - 데이터베이스 특징 : 통합된(integrated), 저장된(stored), 공용(shared), 변화(changable), 가독, 검색, 원격

데이터베이스 정의

- 대량의 데이터를 축적하는 기지
- 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합
- 체계적으로 정렬된 데이터 집합
- 데이터량과 이용이 늘어나면서 데이터를 저장/관리/검색/이용할 수 있는 컴퓨터 기반의 데이터베이스로 진화
- 정보의 집합체
- 데이터베이스의 특징
  - 통합된 데이터 : 중복 x
  - 저장된 데이터 : 저장매체에 저장
  - 공용데이터 : 서로 다른 목적, 공동 데이터 이용
  - 변화되는 데이터 : 계속 변화하면서도 항상 현재의 정확한 데이터 유지
- 데이터베이스의 특성
  - 정보의 축적 및 전달 : 기계가독성, 검색가능성, 원격조작성 = 원거리에서도 즉시 온라인으로 이용
  - 정보 이용 : 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득, 원하는 정보를 정확하고 경제적으로 찾아낼 수 있다.
  - 정보 관리 : 정보를 체계적으로 축적하고 새로운 내용 추가나 갱신이 용이하다.
  - 정보기술 발전 : 정보처리, 검색/관리 소프트웨어, 하드웨어, 정보 전송을 위한 네트워크 기술 등의 발전을 견인할 수 있다.

데이터베이스 활용

- 데이터베이스는 주로 기업에서 활용
- 데이터베이스 변화 과정

1) OLTP(Online Transaction Processing)

- 단순한 정보의 '수집'
- 단순 자동화
- 데이터베이스의 데이터를 수시로 갱신하는 프로세싱
- 데이터 갱신 위주

2) OLAP(Online Analytical Processing)

- 정보 위주의 분석 처리
- OLTP에서 처리된 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파악 등을 프로세싱
- 데이터 조회 위주
- 쉽고 빠르게 다차원적인 데이터에 접근하여 의사 결정에 활용할 수 있는 정보를 얻게 해준다.

- 3) CRM(Consumer Relationship Management)
  - 고객관계관리
  - 고객별 구매이력 데이터베이스를 분석하여 고객에 대한 이해를 돕고 이를 바탕으로 각종 마케팅 전략을 펼치는 것
- 4) SCM(Supply Chain Management)
  - 공급망 관리
  - 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화 시키기 위한 것

데이터베이스 활용

5) ERP(Enterprise Resource Planning)
- 전사적 자원관리, 경영자원을 하나의 통합 시스템으로 재구축

6) RTE(Real Time Enterprise)
- 회사의 주요 경영정보를 통합관리하는 실시간 기업의 새로운 기업경영시스템
- 회사 전 부문의 정보를 하나로 통합

7) BI(Business Intelligence)
- 기업이 보유하고 있는 수많은 데이터를 정리하고 분석해 기업의 의사결정에 활용하는 프로세스
- 질의(query), 보고(reporting), 온라인 분석처리(OLAP), 통계분석, 예측, 데이터 마이닝 등의 결합

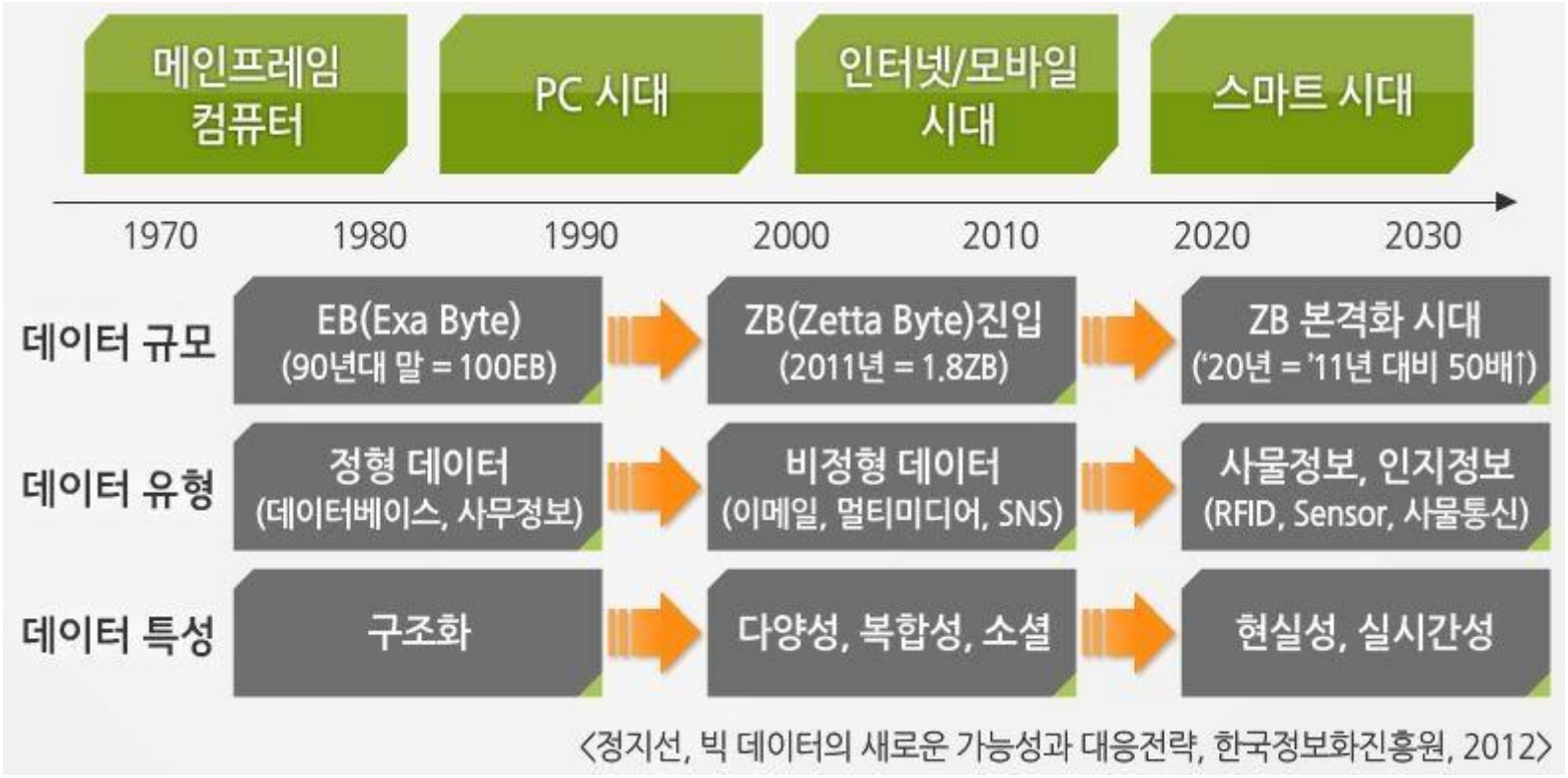
8) EAI(Enterprise Application Integration)
- 기업 내 상호 연관된 모든 애플리케이션을 유기적으로 연동하여 필요한 정보를 중앙 집중적으로 통합, 관리, 사용할 수 있는 환경을 구현하는 것
- 손쉬운 확장 : 새로운 애플리케이션 도입 시 어댑터(Adapter)만 필요

9) KMS(Knowledge Management System)
- 기업 경영을 지식이라는 관점에서 새롭게 조명하는 접근방식
- 객체지향 DBMS : 멀티미디어 등 복잡한 데이터 구조를 관리하는 DBMS
- 데이터웨어하우스 : 방대한 조직내 분산된 데이터베이스 관리시스템을 통합, 운영 시간성을 가지는 비휘발성 데이터의 집합
- SQL : 데이터베이스와 통신을 위해 고안된 언어

빅데이터

- 빅데이터 정의
  - 일반적인 데이터베이스 관리의 범위를 초과, 저비용/초고속 분석 지원이 가능하게 하는 기술
  - 3V(volume규모, variety다양성, velocity속도) + 1V(value, 비즈니스적 가치)
  - 발전요인
    - 저장(하드디스크 용량↑, 가격↓)
    - 효율(CPU, network, cloud)
    - 인재(데이터중심, 조직, 인재)


디지털 정보량 증가 추이
스마트시대
인공지능
사물정보통신
증강현실
모바일혁명
SNS
IT
everywhere
인터넷 /
모바일 시대
Web2.0
PC 시대
메인프레임
컴퓨터
WWW
1인1PC
브로드밴드
빅데이터
시대 도래
1970
1980
1990
2000
2010
2020
2030
출처 정지선, 빅 데이터의 새로운 가능성과 대응전략, 한국정보화진흥원, 2012



〈정지선, 빅 데이터의 새로운 가능성과 대응전략, 한국정보화진흥원, 2012〉

빅데이터 처리 과정


기술
빅데이터 처리 과정
데이터소스
수집
저장
처리
분석
표현
내부데이터
외부데이터
facebook
me2day
미디어
수동
로그 수집기
자동
크롤링
센싱
정형 · 비정형
데이터등급
장치
배치처리
실시간 & 배치
NoSQL
관계형DB
분산병렬처리
hadoop
전처리
분석방법
Machine Learning

데이터사이언스

- 데이터 사이언스
  - 데이터로부터 의미 있는 정보를 추출해내는 학문
  - 빅데이터에 대한 이론적 지식, 숙련된 분석 기술을 바탕으로 통찰, 전달, 협업 능력 보유, 인사이트 도출, 전량 방향 제시 기존 분석, 통계학에서 발전된 정형, 비정형을 포함하여 다양한 유형의 데이터를 대상으로 총체적 접근법을 사용
  - hard skill : 빅데이터 이론적 지식 / 분석 기술에 대한 숙력
  - soft skill : 통찰력 있는 분석 / 설득력 있는 전달 / 다분야간 협력
  - 가져야할 소양
    - 데이터공학, 수학, 최적화, 확률통계학, 컴퓨터공학, ML/DL, 프레젠테이션/시각화, 해당분야 전문지식을 종합 인문학적 사고
    - 공급자 중심의 산출물 중시에서 소비자 중심의 전략 수립과 통찰력

- 빅데이터 테크닉 예시
  - 연관규칙학습 : 변수간 상관관계 파악
  - 유형분석 : 문서, 조직 등을 그룹을 나눌 때
  - 기계학습 : 훈련 데이터로부터 학습한 특성 활용
  - 예측 회귀분석: 독립변수에 대한 종속변수의 관계 파악
  - 감정분석 : 특정 주제에 대한 글쓴이의 감정
  - 소셜네트워크분석 : 특정인과 다른 사람간의 관계, 영향력 있는 사람 파악
  - 유전알고리즘 : 최적화

- 빅데이터가 만들어낸 변화
  - 사전처리 → 사후처리
  - 표본조사 → 전수조사
  - 질 → 양
  - 인과관계 → 상관관계
  - 위기요인과 통제 방안
  - 사생활침해 → 동의에서 책임으로
  - 데이터 오용 → 알고리즘 접근 허용

'Big'데이터

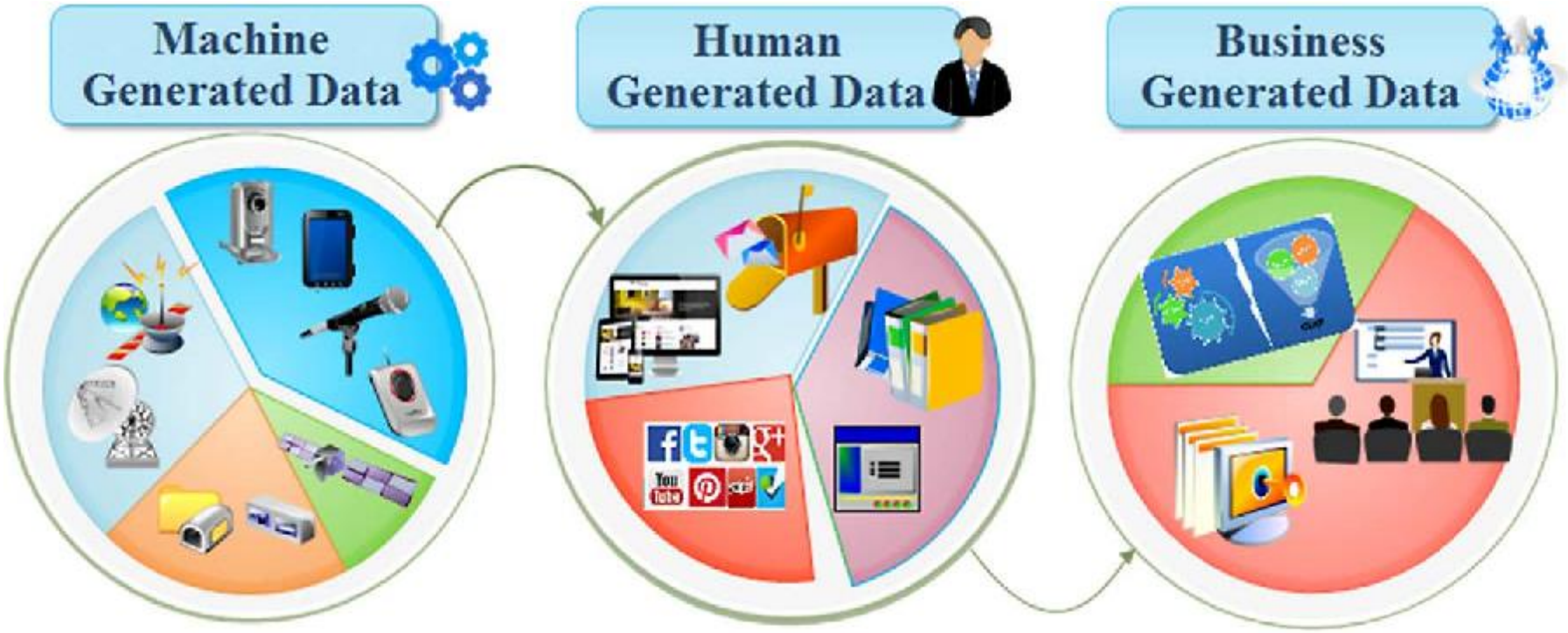
- 직관에 기초한 의사결정보다 데이터에 기초한 의사결정이 중요하다.
- 데이터의 양 대신 다양성에 초점. 새롭고 다양한 정보 원천의 활용
- 무작정 빅데이터를 찾는 것이 아닌, 비즈니스의 핵심에 대해 보다 객관적이고 종합적인 통찰을 줄 수 있는 데이터를 찾는 것이 중요하다.
- 전략과 비즈니스의 핵심 가치에 집중하고 이와 관련된 분석 평가지표를 개발하고 이를 통해 효과적으로 시장과 고객 변화에 대응할 수 있을 때 빅데이터 분석은 가치를 줄 수 있다.

생산 주체에 따른 데이터 분류

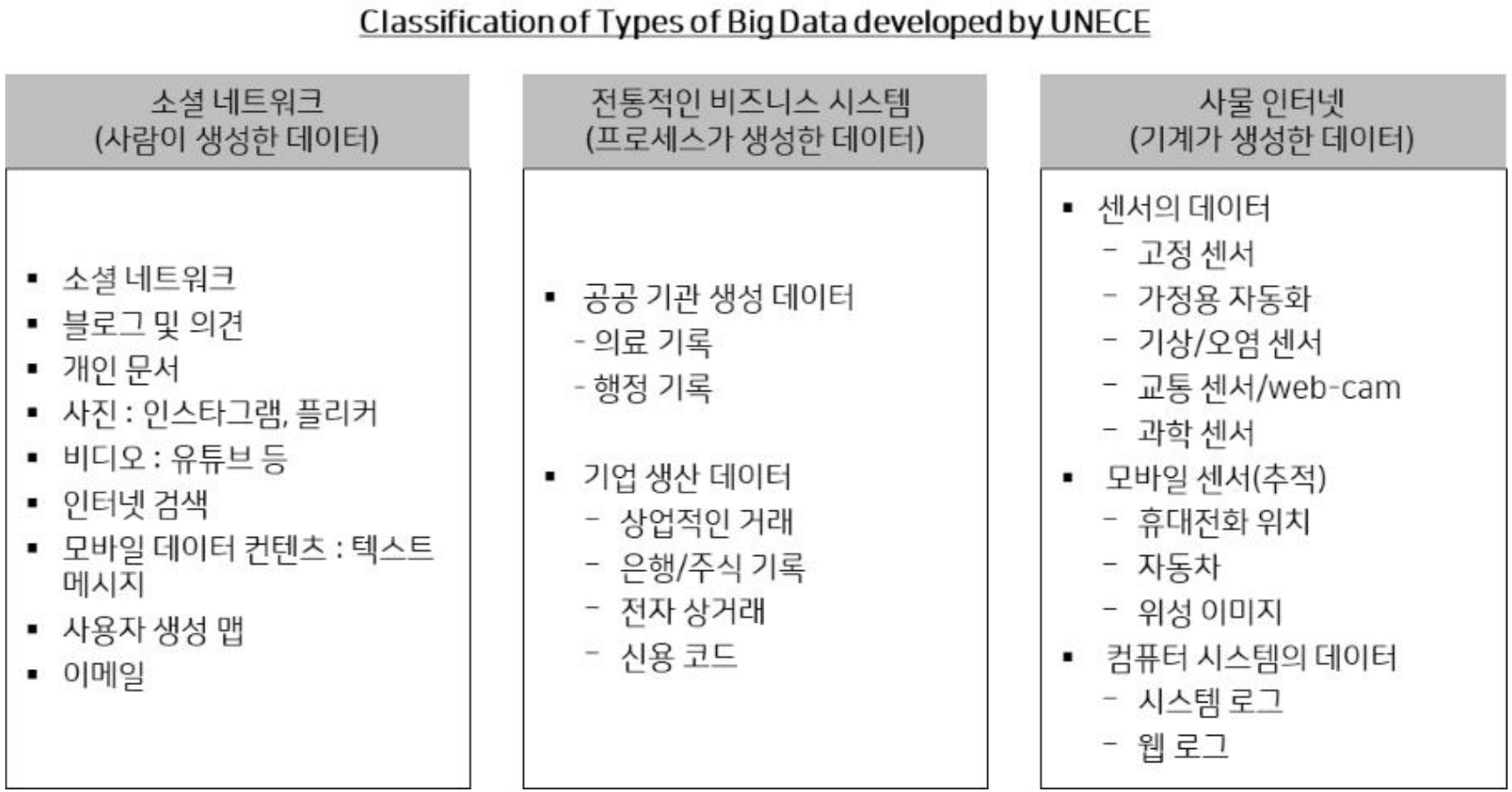
기계 생성 데이터

인간 생성 데이터

조직 생성 데이터



Classification of Types of Big Data developed by 유엔유럽경제위원회(United Nations Economic Commission for Europe, UNECE) (source: De Francisci, 2015, p. 16).

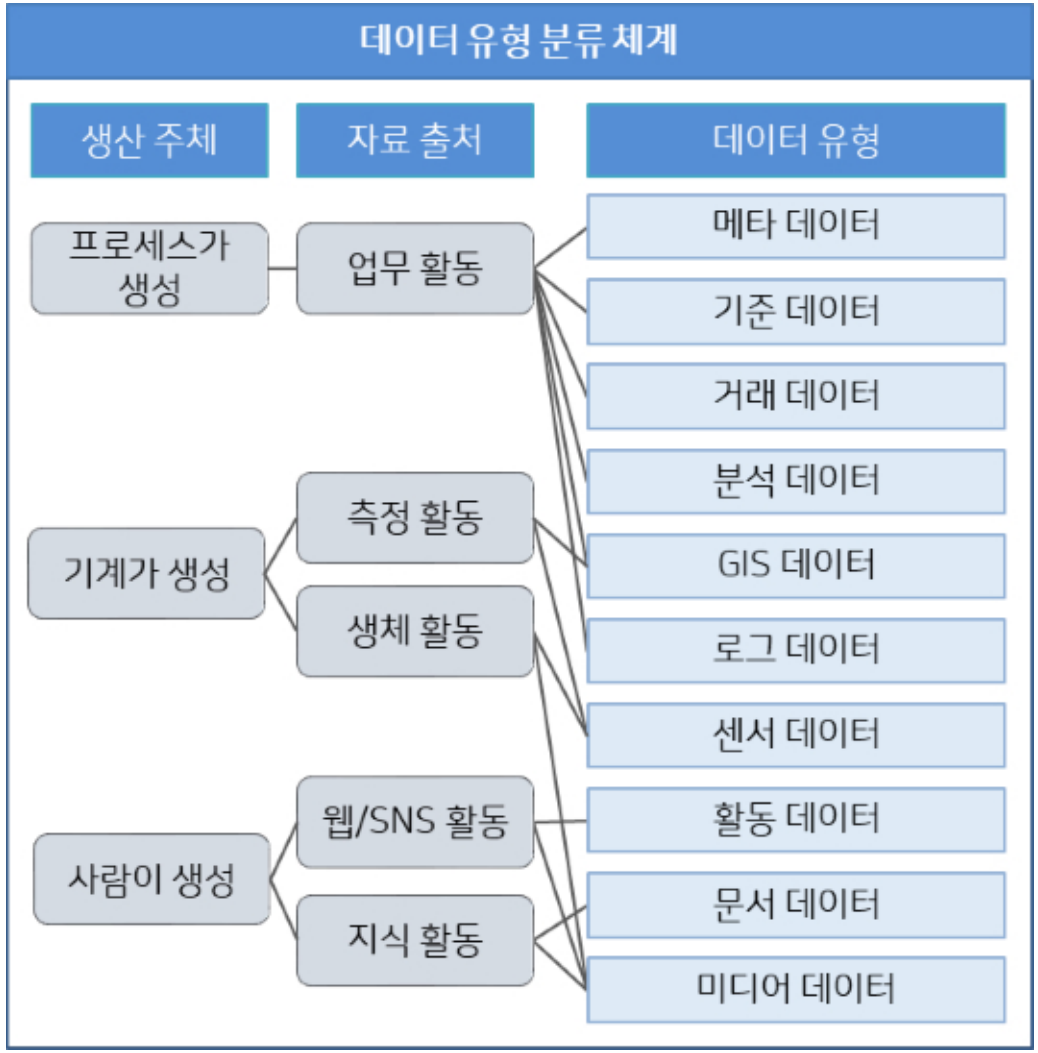


출처: https://www.cisp.or.kr/archives/22015

데이터 유형	설명	예시
메타 데이터 (Meta data)	•테크니컬 메타: 관리를 위해서 파악해야 하거나 통제해야 할 대상이나 항목으로 데이터 구조, 데이터 표준, 데이터 흐름, 데이터 권한 등에 대한 정보	•데이터셋의 물리/논리명, 항목의 물리/논리명, 데이터형식, 업무규칙, 표준사전, 표준도메인 등
	•비스니스 메타: 데이터를 활용하기 위하여 파악해야 할 대상이나 항목으로 데이터를 설명하는 정보로 정보명, 주제영역, 품질수준, 다른 데이터와 연관성 등에 대한 정보	•설명, 생성주기, 출처, 주제영역, 활용영역, 품질수준, 연관정보, 위치정보 등
기준 데이터 (Master data)	•업무 프로세스의 중심이 되는 기준 정보 및 참조 정보(데이터값이 참조하는 코드 정보)	•제품정보, 시설정보, 사업자정보 등•지역코드, 성별코드, 학력코드 등
로그 데이터 (Log data)	•시스템이 생성한 Log 정보 및 웹 크롤링(crawling) Raw file 형태의 정보	•로그기록, WebLog, 웹 크롤링(crawling) 등 Raw file 정보
거래 데이터 (Transaction data)	•기업 또는 기관의 고유한 업무 및 서비스 활동을 처리하는 정보시스템에 의해 생성, 관리되는 트랜잭션 정보	•신용카드 거래 내역 및 금융 거래 내역, 오픈마켓 구매 내역 등
분석 데이터 (Analytics data)	•집계 또는 통계 및 분석을 통하여 결과로 생성된 정보	•업종별 매출현황, 이동인구, 상권분석 결과 등•연관규칙, 분류기준, 상관관계, 공간분석 등
GIS 데이터 (GIS data)	•지형지물에 대한 공간적 정보로서 벡터(Vector), 래스터(Raster) 형태의 공간 정보 및 공간정보의 속성정보, 통상 GIS에 의하여 생성, 관리되는 정보	•행정구역도, 지하매설물도, 수치지형도, 산림도, 정사영상 등
문서 데이터 (Document data)	•문서 작성기로 생성한 문서 정보(hwp, doc, pdf 등 고유의 저장형식으로 생성)	•일반문서, 논문, 보고서 등
미디어 데이터 (Multimedia data)	•다양한 멀티미디어 정보	•사진, 영상, 음성, 엑스레이(x-ray), 초음파, CT, MRI, 위성영상, 항공영상 등
센서 데이터 (Sensor data)	•사물인터넷(IoT), 추적장치(Tracking Device), 공장자동화 기기 등 각종 센서를 통하여 생성되는 정보	•위치, 기상, 수질, 대기, IoT, 차량통행, 생산설비센서 등
활동 데이터 (Online behavior data)	•의견정보(Opinion data), 웹 검색(Web search)정보 등을 포함한 온라인상에서 생성된 것으로 분석을 위하여 전처리(정재, 자연어처리 등)를 수행한 정보	•인터넷 검색 및 페이지뷰 정보 (Web-Log) 및 웹게시글, 카카오톡, 트위터 등 공개 게시글

출처: https://www.cisp.or.kr/archives/22015

생산 주체에 따른 데이터 분류



데이터 생성: 기계 vs 인간 vs 조직

① **기계 생성 데이터:** 산업용 기계나 비행기, 차량 등에 내장된 센서, 온라인의 사용자 행동 추적 기록, 환경 센터, 개인 건강 추적기 등 정형 데이터

② **인간 생성 데이터:** SNS 글, 사진, 그림, 오디오, 미디어 등 비정형 데이터

③ **조직 생성 데이터:** 상업 거래, 신용 카드, 정부 기관, 전자 상거래, 은행 거래, 주식 거래, 의료 기관 등 기록하고 모니터링하기 위해 고도로 구조화된 데이터(Highly-structured organizational data), 관계형 테이블 데이터

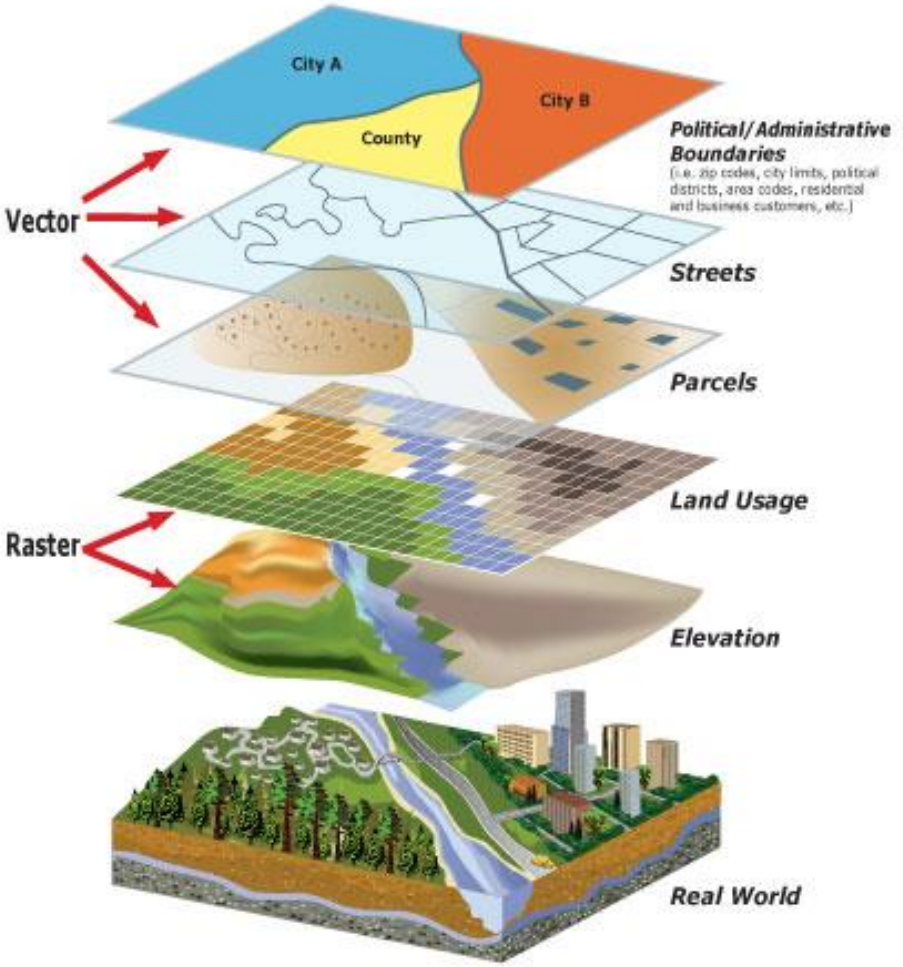
④ **AI생성 합성 데이터**

출처: https://www.cisp.or.kr/archives/22015

기계 생성 데이터(machine-generated data)

- GIS 데이터 / 위성 데이터
- 센서 데이터 / 스캐너 / RFID
- 미디어 데이터 / 사용자 위치 및 타임 로그

GIS 데이터



- GIS 데이터: 현 세계의 공간상에 있는 객체나 현상의 공간적 위치를 지도학상의 좌표체계에 따라 표현한 도형 자료로 정의한다.
- 크게 경계(boundary)데이터와 커버리지(coverage)데이터로 구분된다.
  - **경계 데이터**는 벡터 데이터(vector data)로 불리며, 대부분의 경우 불연속적인(discrete)데이터의 형태를 가지고 있는 지형지물(feature)에 연관된 데이터가 이에 해당된다.
  - **커버리지 데이터**는 일정지역에 걸쳐 연속적인(continuous)데이터의 형태를 가지고 있으며, 공간적인 위치와 직접적으로 연관된 값을 의미한다. 래스터(raster)데이터라고 불리며, 그리드 데이터(grid data)는 커버리지 데이터에 속한다.

센서 데이터

- 센싱(sensing)이란 컴퓨터는 센서(sensor)를 통해 사상을 인식(recognition)한다.
- 센싱(sensing)이란 물리적 사상을 인식하고 크기를 측정하여 전기적 신호(디지털 형태)로 변환하는 것이다.

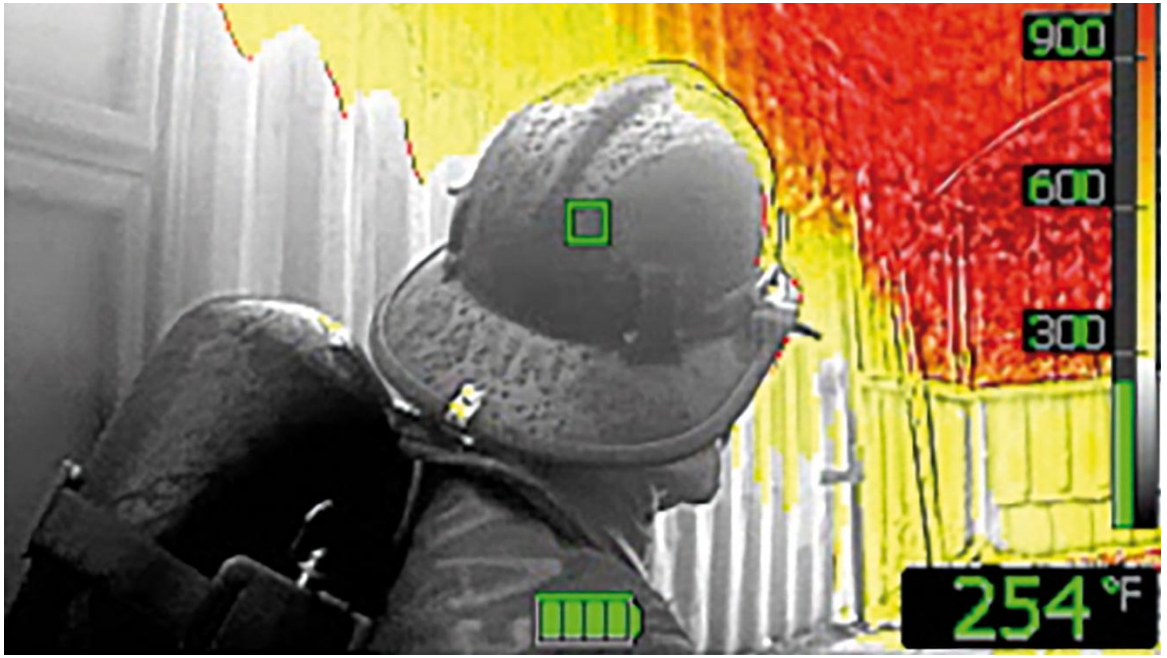
물리적 사상
(physical phenomena)

↓

크기 센서
(magnitude sensor)

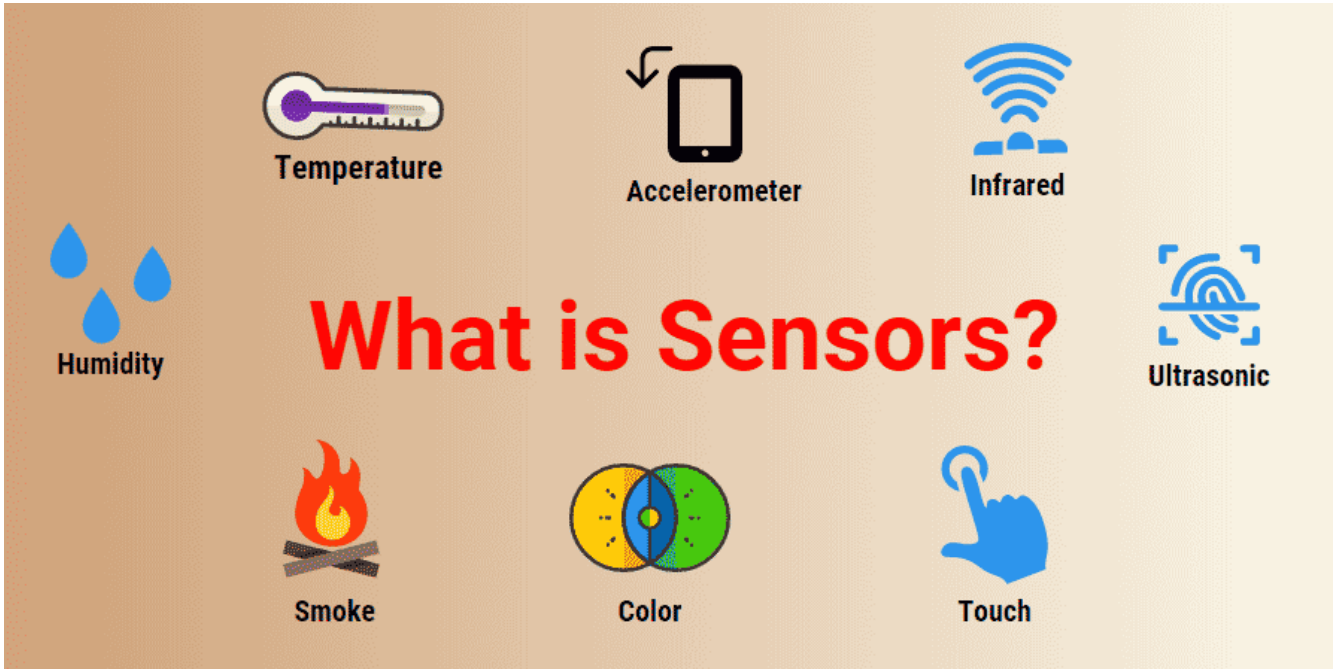
↓

데이터
(data)



컴퓨터의 센서

- 카메라
- 마이크
- 가속도계
- 적외선 센서
- 압력센서
- 온도센서
- 습도센서



인간 생성 데이터

- 인간 생성 데이터는 인간 행동을 통해 사람들이 생성 한 데이터이다.
- 인터넷 검색, 전자 메일, 소셜 미디어 게시물 등 인간이 생성하는 데이터는 대부분 구조화 되지 않은 비정형 데이터(텍스트, 이미지, 비디오, 오디오 등)이다.

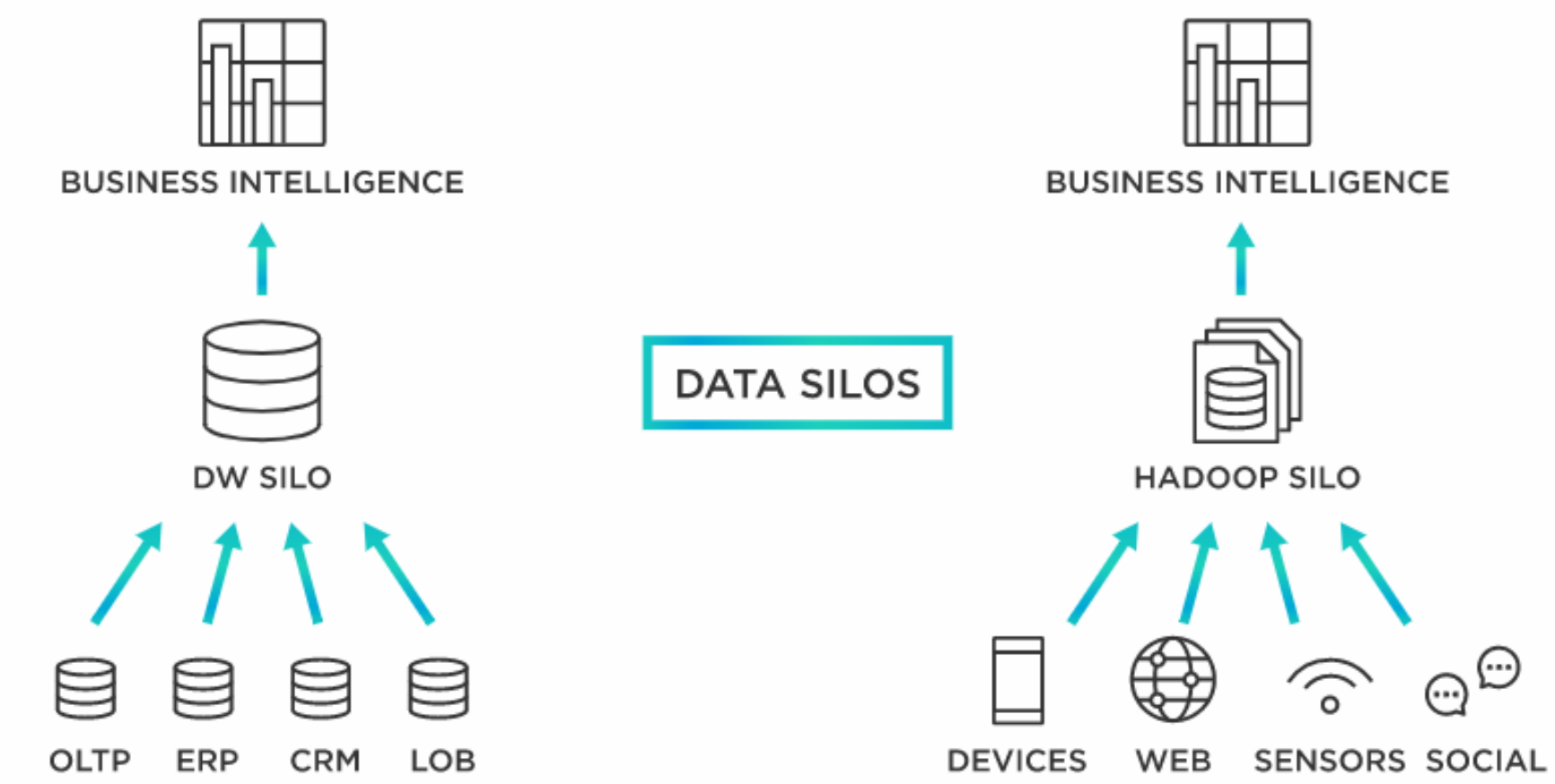
기계 생성 데이터 vs 인간 생성 데이터

데이터 유형		기계 생성	인간 생성
정형	내부	판매 로그, 구매 추적, 진행 제어 측정	검토, 점수, 재능 평가 문서
	외부	트윗용 GPS 업데이트/게시물/트윗의 로그 시간 스마트 기계	트윗 수, 댓글 수 Facebook 좋아요 제품 평가
비정형	내부	감시 비디오 RFID 제품 스캐너	이메일, 편지, 메시지 YouTube 동영상 이미지, 음성 메일, 오디오 스크립트
	외부	위성 CCTV 동영상 과학 데이터 레이더 데이터	Facebook 댓글, 트윗 제품 리뷰 Pinterest 이미지

조직 생성 데이터

- 대부분의 조직에서는 부서와 팀이 분리되어 일하는 경향이 있다. 여러 부서에서 다양한 애플리케이션을 사용함으로 해서 **데이터 사일로(data silo)**가 발생할 수 있다.
  - 데이터 사일로란 데이터가 격리되어 특정 조직/부서/단위에서만 정보 접근 및 공유가 가능하여 다른 조직/부서/단위에서는 데이터가 격리되는 현상을 뜻합니다. 각 부서별로 데이터에 쉽게 접근할 수 없는 분리된 현상을 겪음으로 조직 내 '단절'을 유발시킵니다
  - 데이터 사일로로 인해 비즈니스 비용과 시간이 많이 소요된다. 데이터 사일로를 제거하면 정보 저장 비용과 정보의 중복을 줄이고, 적절한 시기에 올바른 정보에 액세스할 수 있으므로 비즈니스 의사결정에 도움이 된다.
- 조직은 빅데이터 실행을 통해 조직 문화를 통합함으로써 데이터 사일로를 없앰으로서 조직의 운영 효율성, 마케팅 성과 향상, 수익 향상, 고객만족 향상 등 상당한 이득을 얻을 수 있다.

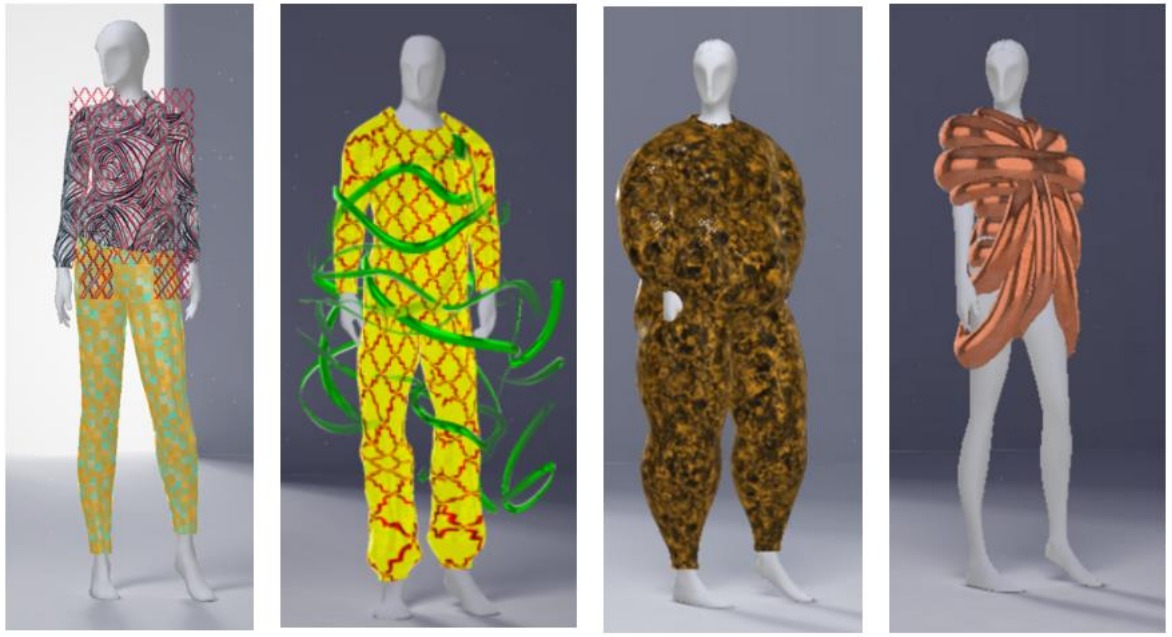
데이터 사일로(data silo)



AI생성 합성데이터

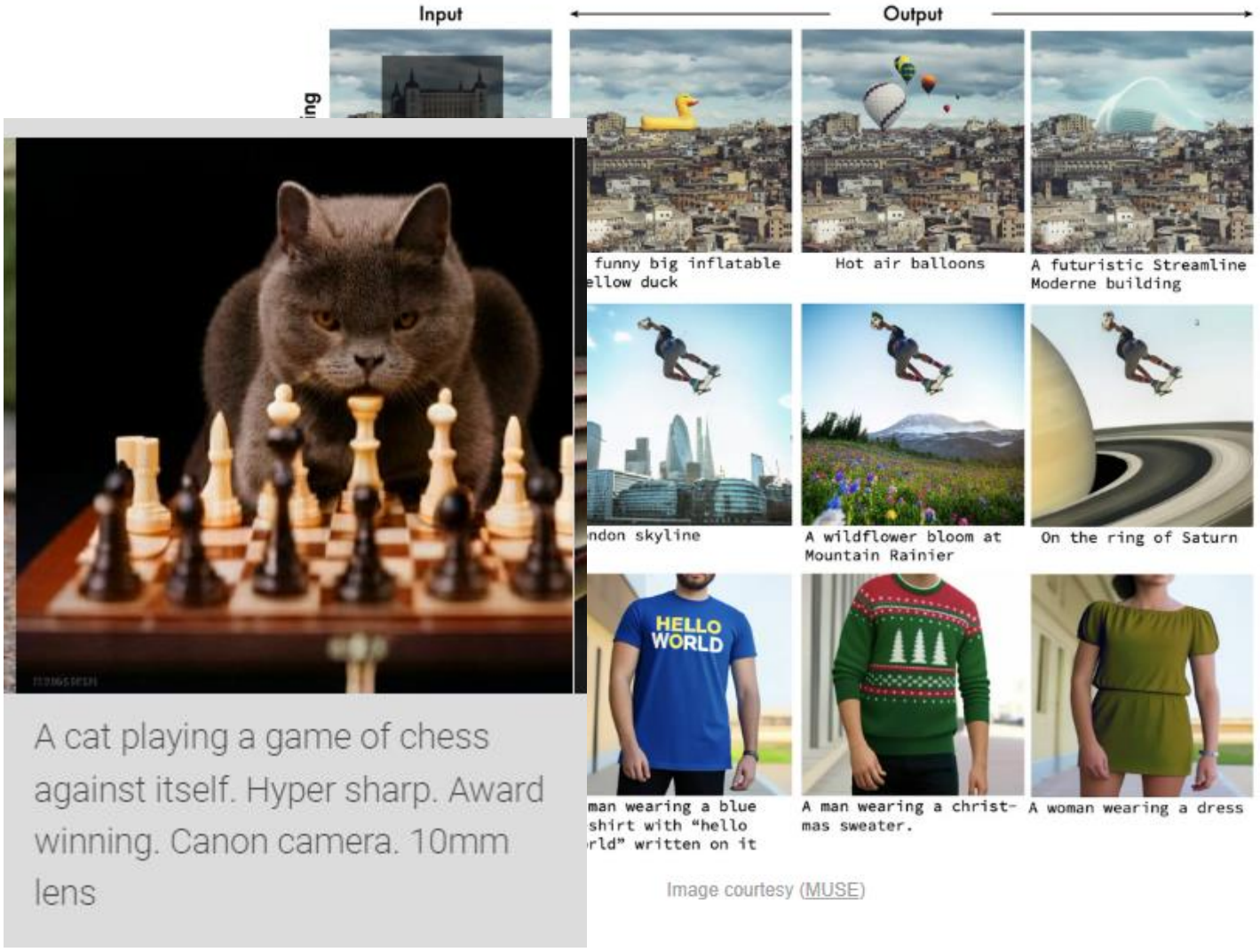
- **더미 데이터 / 모의 데이터**
  - 무작위로 생성된 데이터로, 원본 데이터에 있는 특성, 관계 및 통계 패턴이 생성된 더미 데이터에 보존, 캡처 및 재생되지 않는다.
- **규칙 기반 생성 합성 데이터**
  - 미리 정의된 규칙 집합에 의해 생성된 합성 데이터이다. 미리 정의된 규칙의 예로는 특정 최소값, 최대값 또는 평균값을 가진 합성 데이터가 있다. 규칙 기반으로 생성된 합성 데이터에서 재현하려는 특성, 관계 및 통계 패턴은 미리 정의되어야 한다
- **인공 지능(AI)이 생성한 합성 데이터**
  - 인공지능(AI) 알고리즘이 생성하는 합성 데이터를 뜻한다. AI 모델은 모든 특성, 관계 및 통계 패턴을 학습하기 위해 원본 데이터에 대해 학습된다. AI 알고리즘은 완전히 새로운 데이터 포인트를 생성하고 원래 데이터 세트의 특성, 관계 및 통계 패턴을 재현하는 방식으로 새로운 데이터 포인트를 모델링할 수 있다.

Google의 텍스트-이미지 생성 모델 Muse
text-to-image generation models



패션 디자인 생성 프로젝트

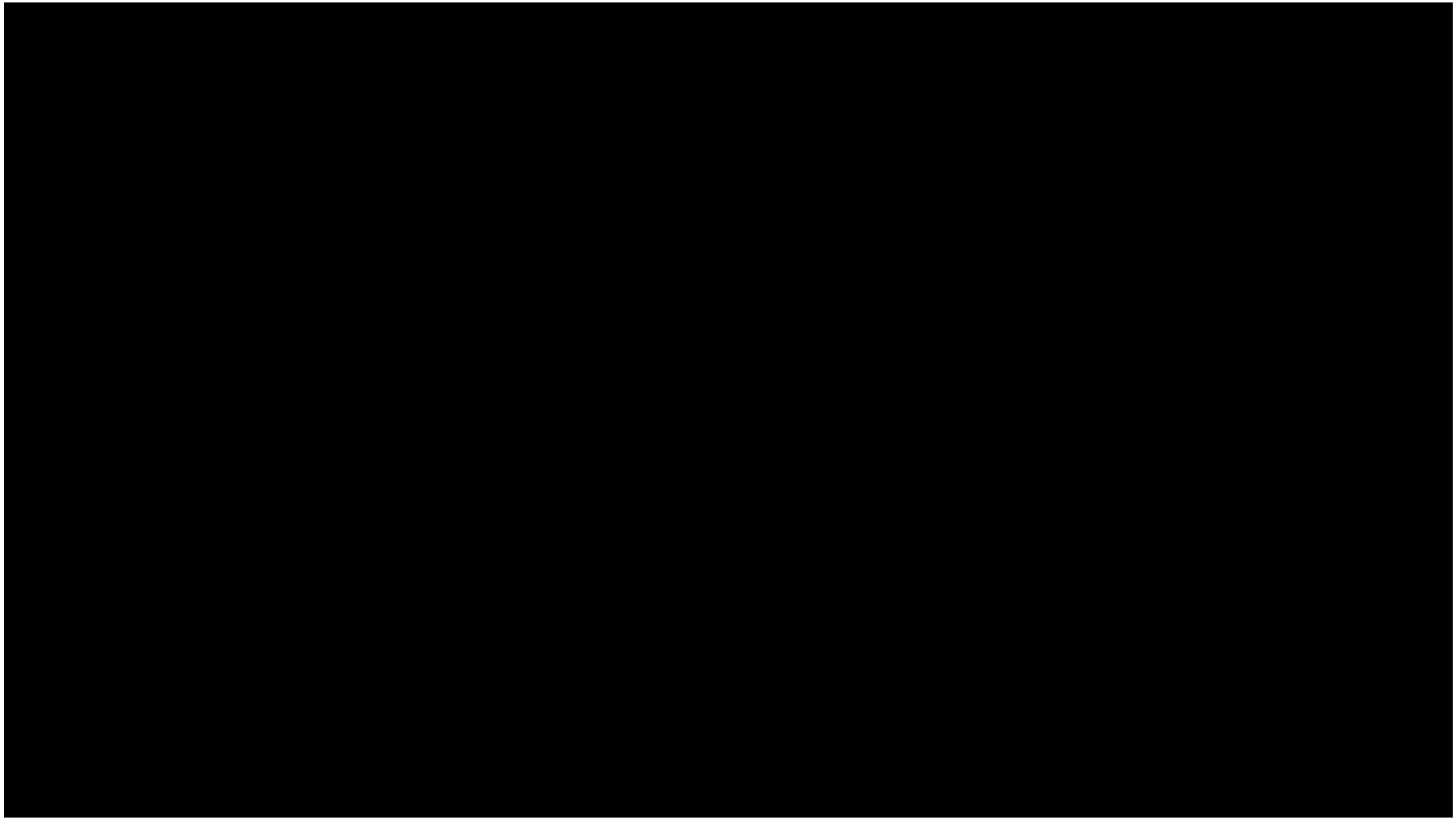
Google의 텍스트-이미지 생성 모델 Muse
text-to-image generation models



출처: https://muse-model.github.io/

Google의 텍스트-이미지 생성 모델 Muse
text-to-image generation models

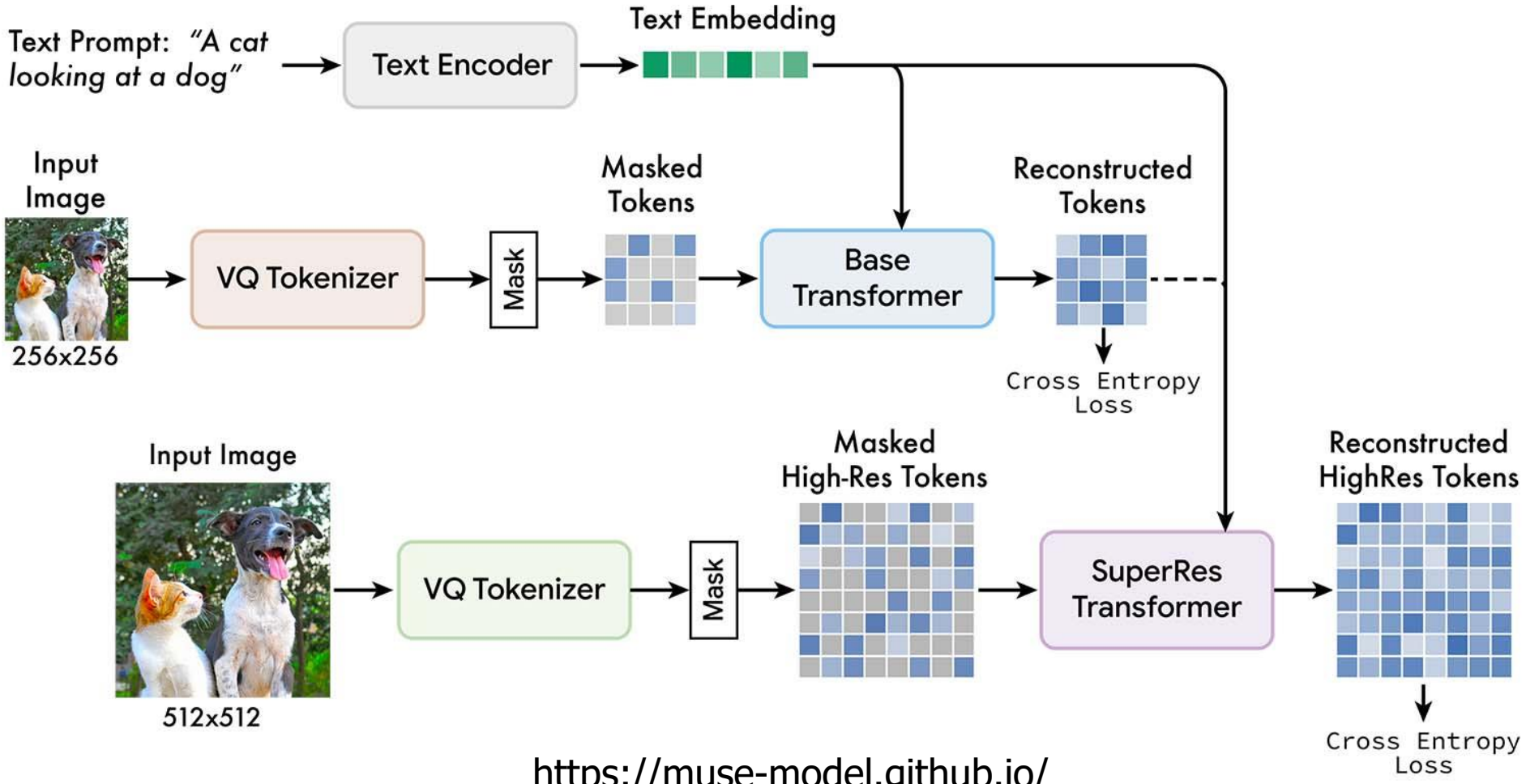
Interactive editing



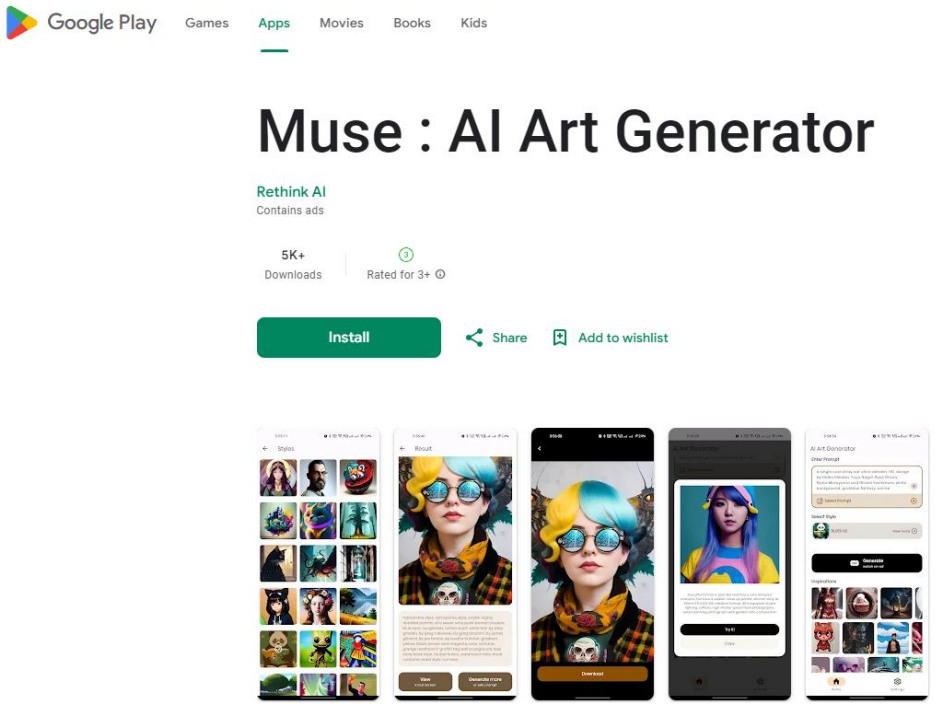
https://youtu.be/2ITNOvp5oJU

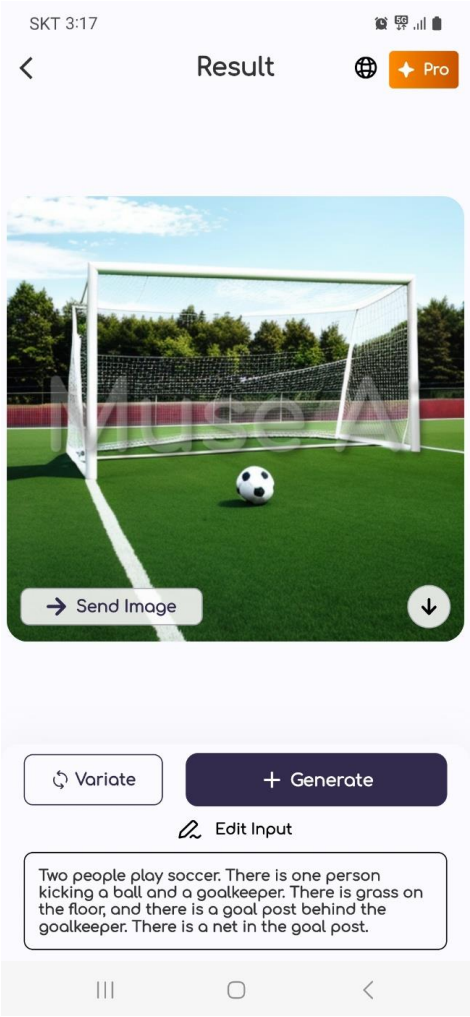
Google의 텍스트-이미지 생성 모델 Muse

text-to-image generation models



https://muse-model.github.io/





Two people play soccer. There is one person kicking a ball and a goalkeeper. There is grass on the floor, and there is a goal post behind the goalkeeper. There is a net in the goal post.

빅데이터를 생성 및 처리할 때 무엇이 중요한가?

데이터 통합(integration)

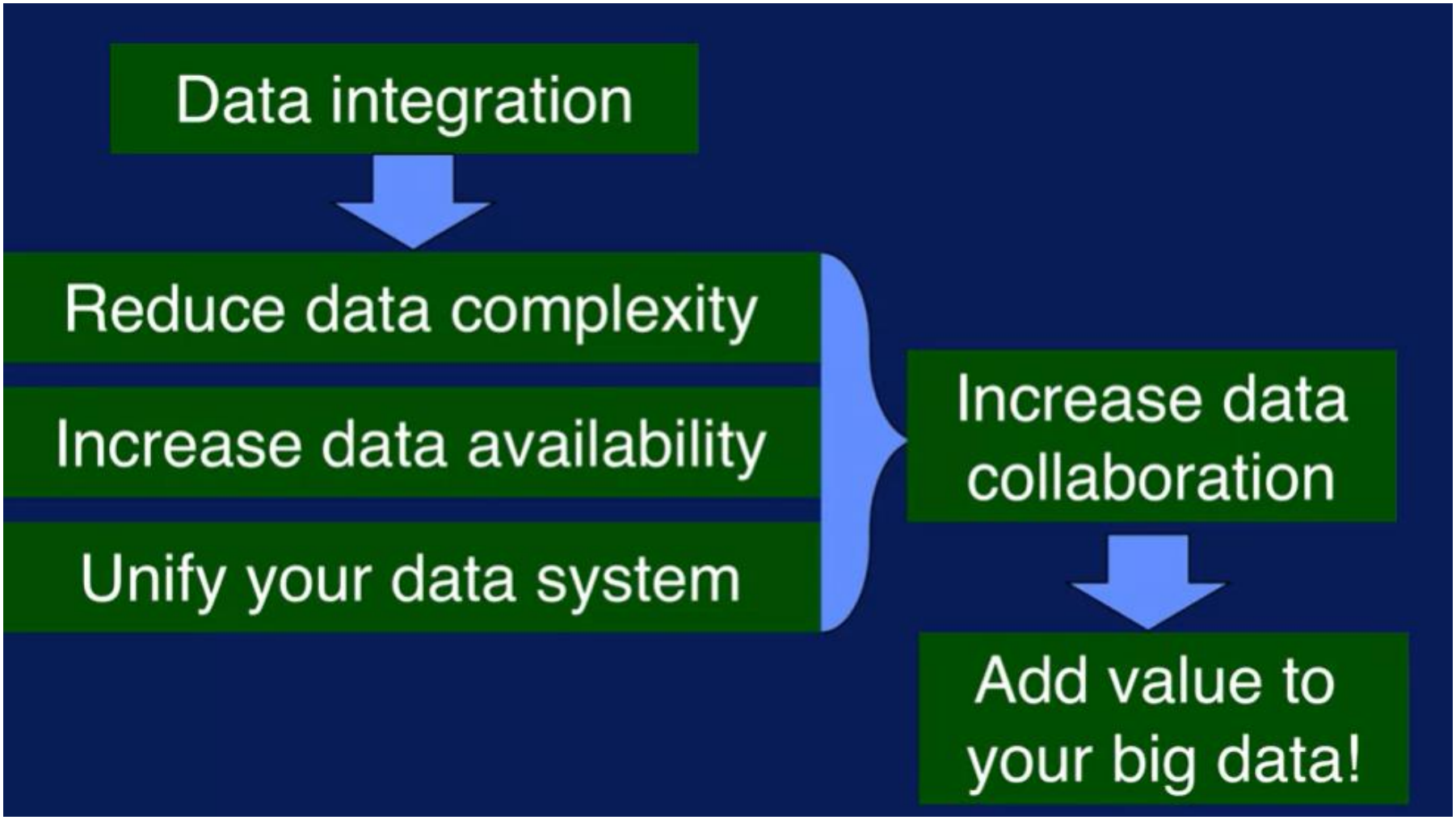
개인맞춤(personalization)

정밀도(precision)

데이터 통합(data integration)

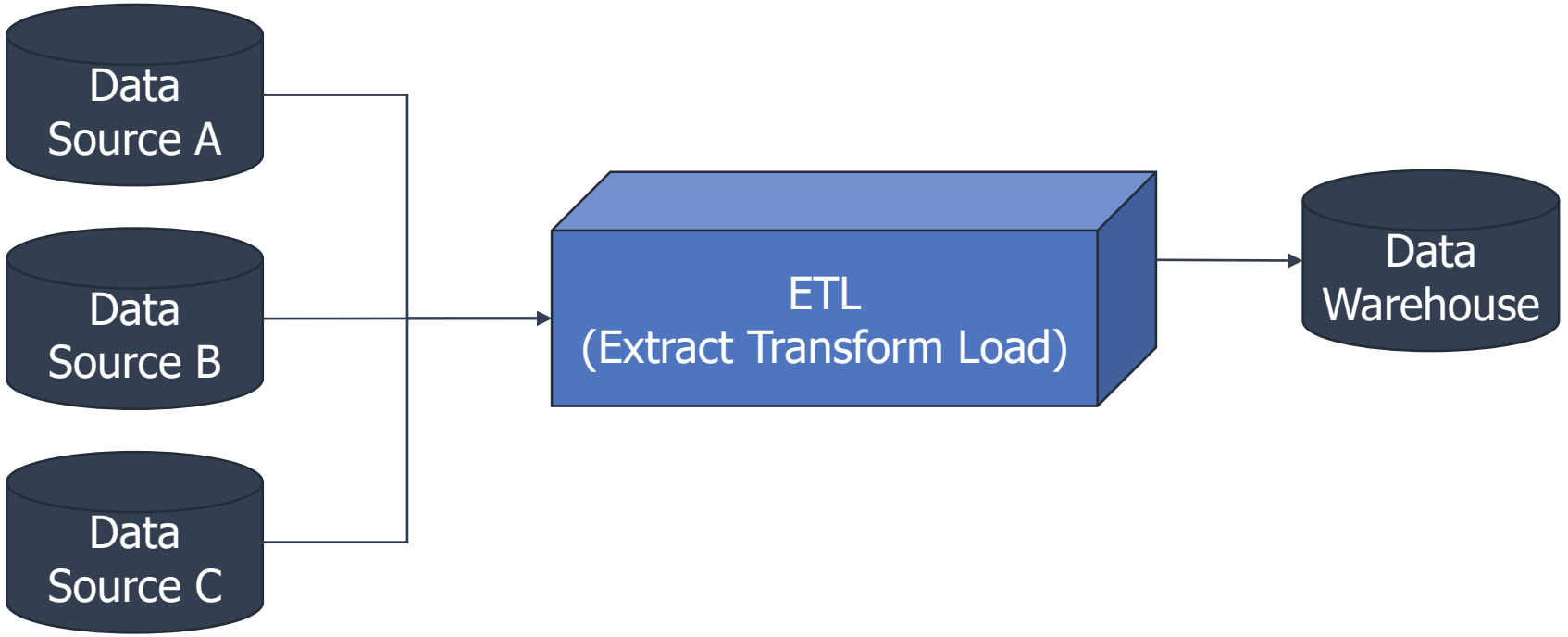
- 데이터 통합(data integration)이란 다양한 출처의 데이터를 결합하여 일관성 있고 유용한 정보로 바꾸는 것을 의미한다. 이것을 지식(knowledge)이라고 부른다.
- 데이터 통합의 목적은 더 기술적으로 데이터를 관리하고 프로그래밍 방식으로 사용할 수 있는 데이터로 변환하는 것으로 데이터 검색, 접근, 모니터링, 협업이 가능하도록 데이터를 통합하는 것이다.

데이터 통합의 효과

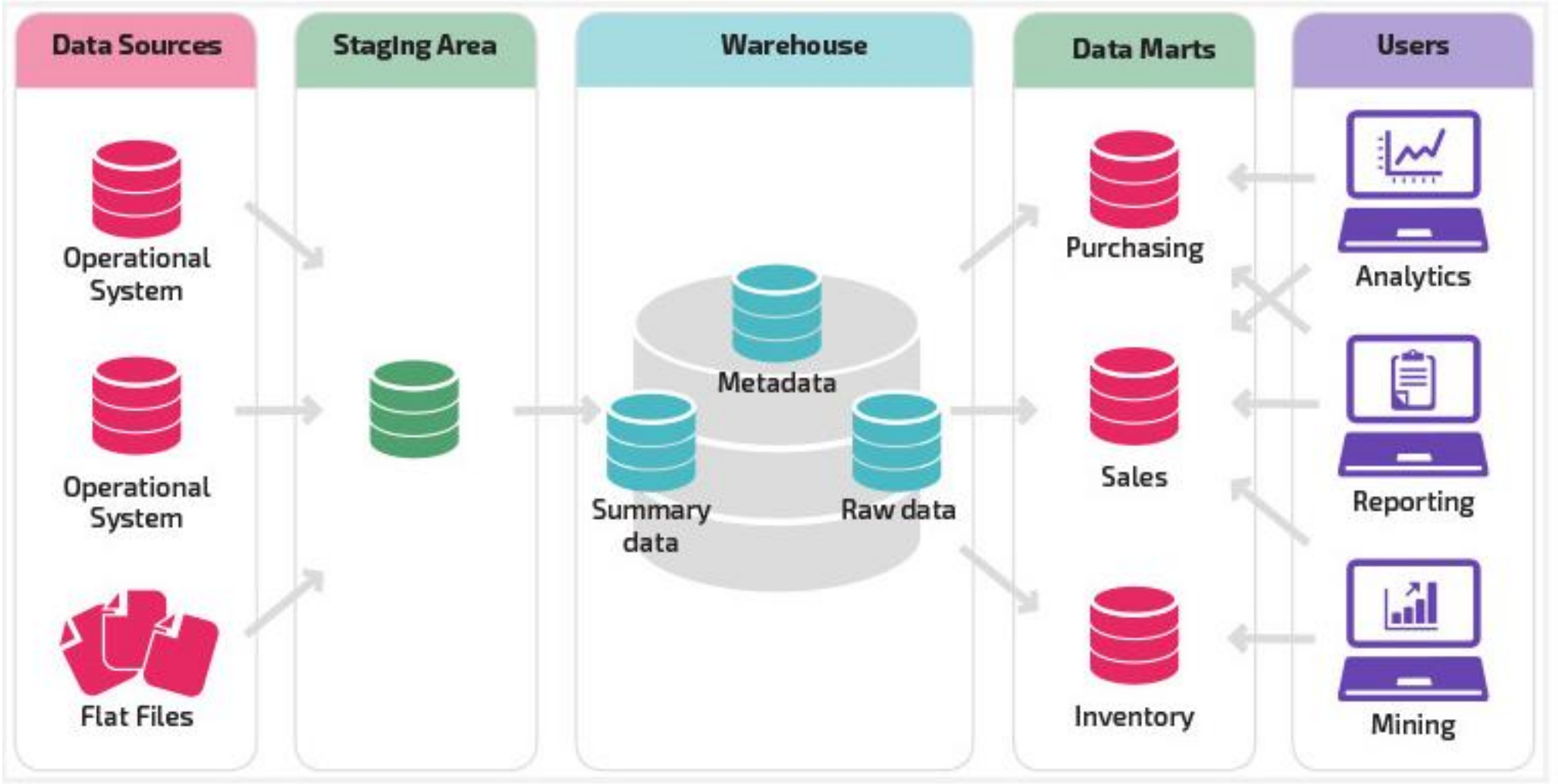

Data integration
Reduce data complexity
Increase data availability
Unify your data system
Increase data collaboration
Add value to your big data!

전통적인 데이터 웨어하우스

데이터 웨어하우스(data warehouse)란 사용자의 의사 결정에 도움을 주기 위하여 기간시스템의 데이터베이스에 축적된 데이터를 **공통의 형식으로 변환**해서 관리하는 데이터베이스를 말한다. 줄여서 DW로도 불린다.



데이터 웨어하우스 vs 데이터 마트



출처: https://majjangjjang.tistory.com/161

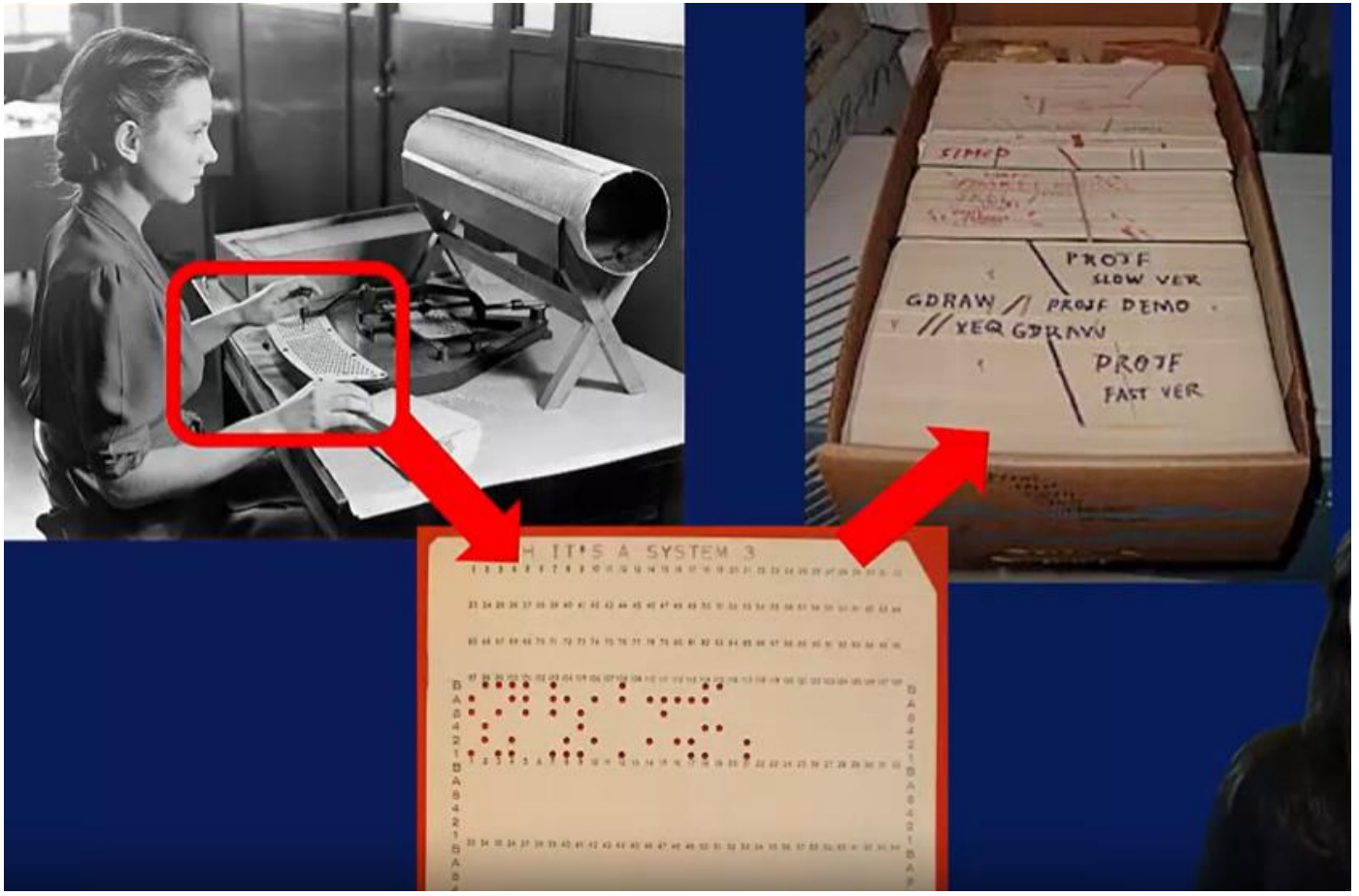
데이터 웨어하우스 vs 데이터 마트

- **데이터 웨어하우스(data warehouse, DW)**는 데이터를 쌓아 놓는 곳이다. 데이터를 통합하고 정제해서 가장 필요한 데이터들만 최소 비용으로 최적화되어 있지 않고 비효율적으로 배치된 상태이다.
- **데이터 마트(data mart, DM)**는 데이터 웨어하우스에 있는 일부 데이터를 가지고 특정 사용자를 대상으로 한다. 그래서 사용하기 쉽게 시스템에 최적화하고 사람이 알기 쉽게 변환하고 성능 면에서 효율적으로 모아놓는 곳이 데이터 마트이다.

빅데이터 웨어하우스


BIG DATA
Database
Cloud Computing
Filtering
Data Share
Data Analysis
Cloud Network

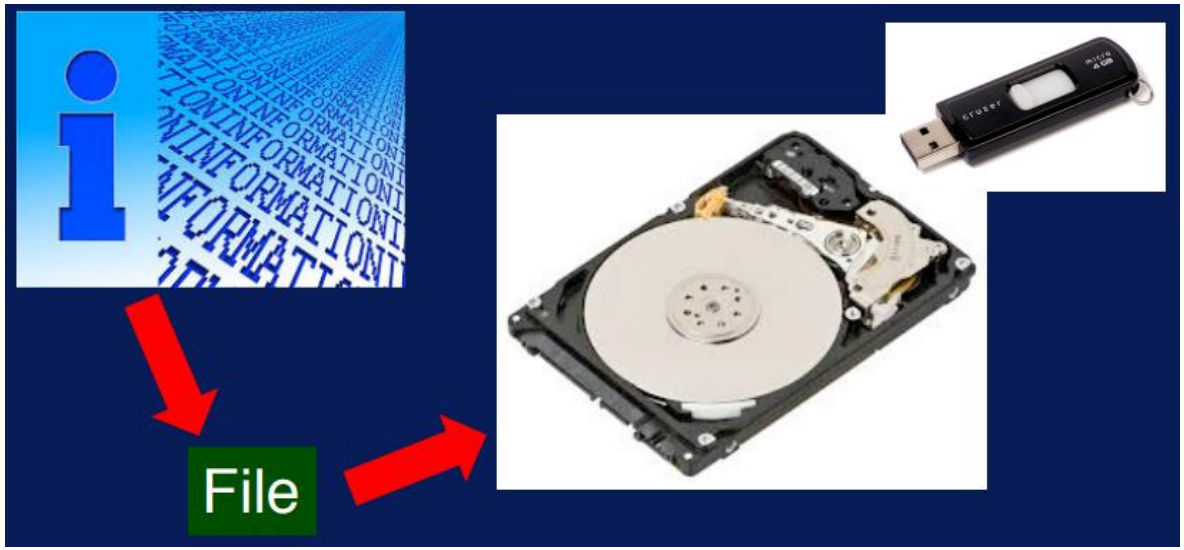
데이터 저장



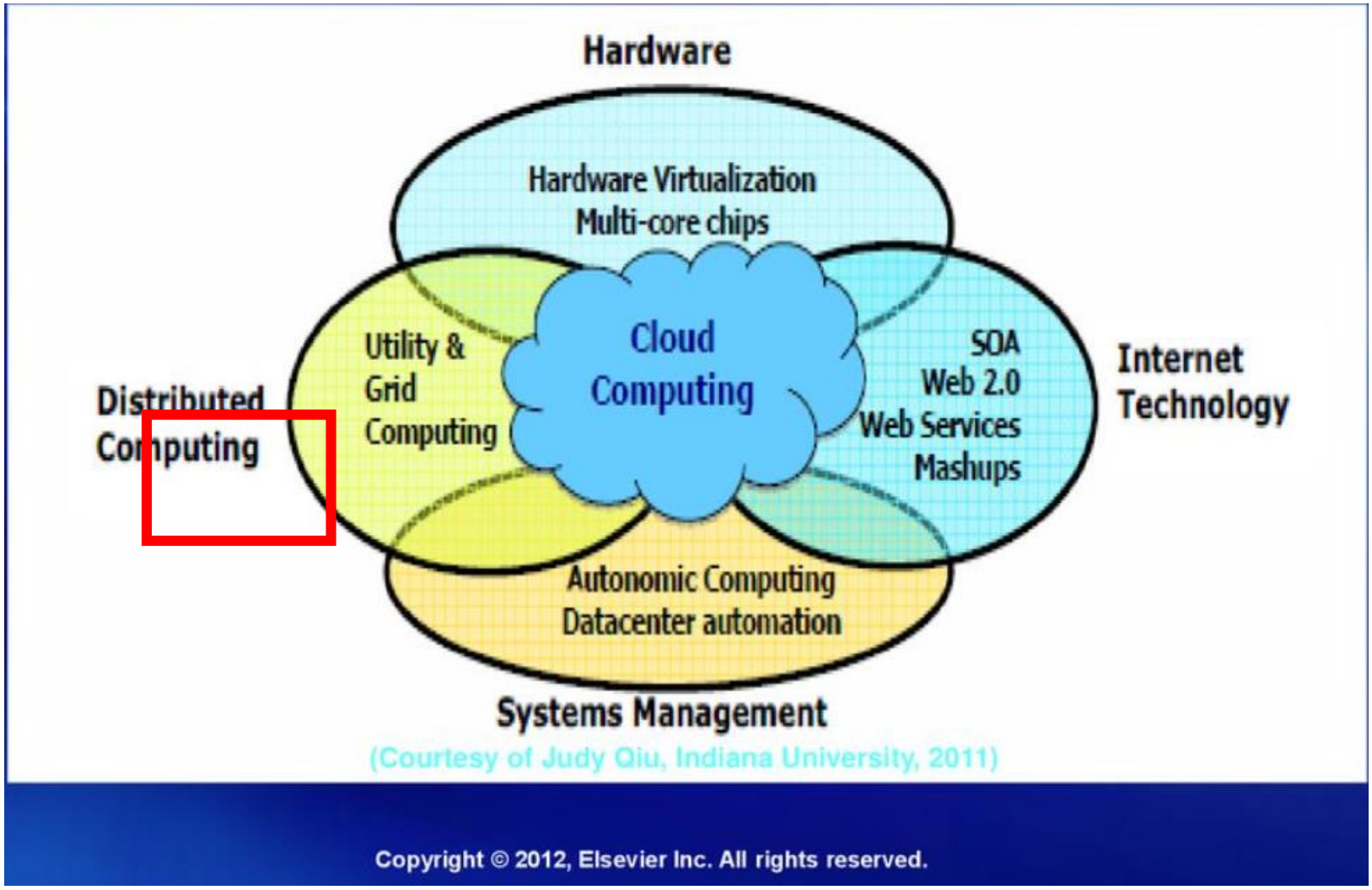
정보저장(펀치 카드)

디스크 or 하드 or 외장 하드 드라이브 저장

- 데이터에 대한 접근의 효율성과 속도에 큰 영향을 줌
- 파일에는 드라이브의 위치에 대한 정확한 주소가 있음, 이것을 플랫 구조(flat structure)라고 부름
- 확장자는 어떤 종류의 파일인지 알려줌

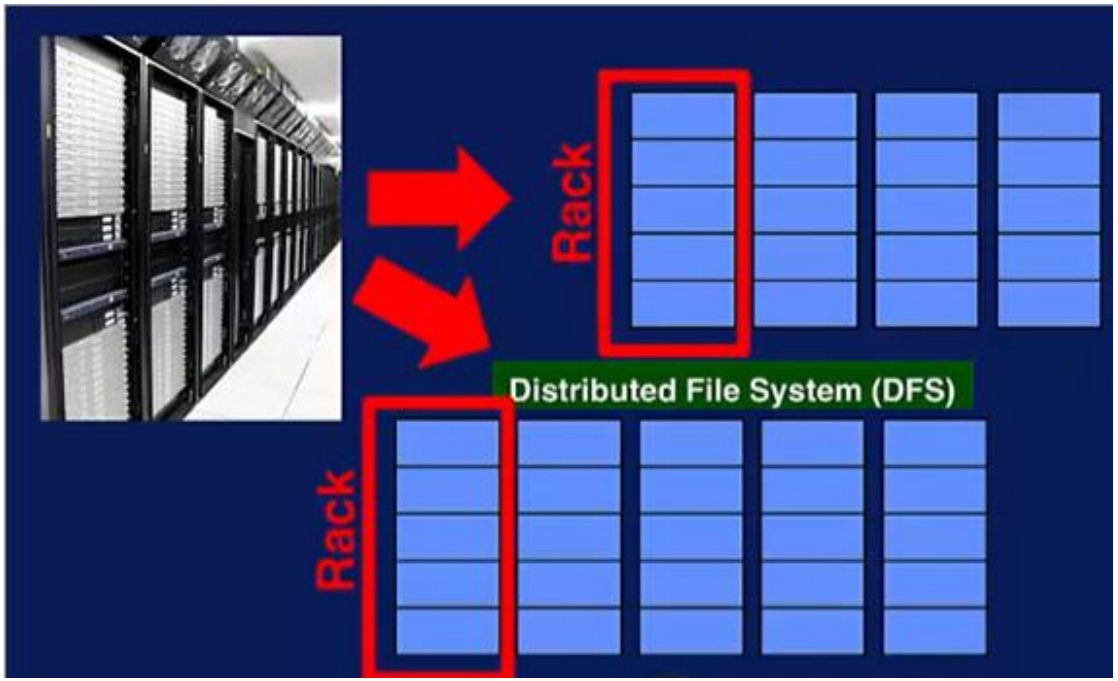


데이터 홍수 시대: 인터넷을 통한 데이터 확장



분산 파일 시스템(Distributed File System)

- 대량의 데이터를 저장할 때 데이터의 접근을 처리하고 이를 수행할 수 있는 시스템



데이터 분할(Data Partitioning)

↓

데이터 복제(Data Replication)

↓

데이터 확장성(Data Scalability)

↓

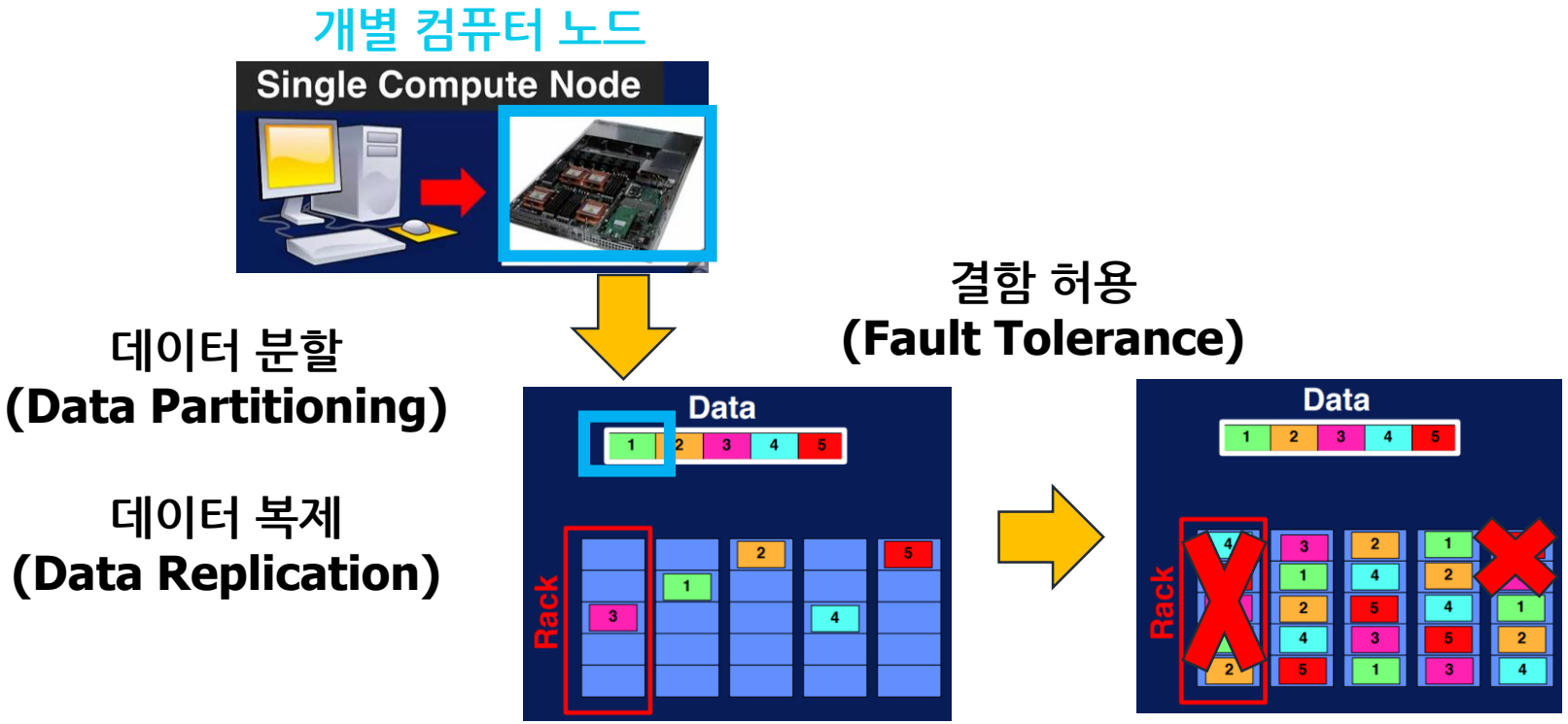
결함 허용(Fault Tolerance)

↓

높은 동시성(High concurrency)

분산 파일 시스템(Distributed File System)

- **랙(Rack)**이란 전산센터를 구성할 때 작은 공간을 효율적으로 사용하고 장비들을 안정적으로 보호하기 위해 사용하는 도구로써, 랙 안에는 여러 대의 서버들로 구성되어 있다.



결함 허용(Fault Tolerance)

- 단일 시스템에서 실패(failure)는 모든 요소에 영향을 미치므로 전체 시스템을 붕괴시킨다.
- 분산 시스템에서는 하나의 시스템에 부분 실패(partial failure)가 일어나더라도 다른 부분에서 작업을 계속 할 수 있다.
- 분산 시스템에서 가장 중요한 목표는 결함 허용(fault tolerant)이다.
  - 전반적인 수행에 심각한 영향을 미치기 전에 부분 실패로부터 자동 복구(recovery)시키는 것
  - 실패(failure)가 발생해도, 시스템을 지속적으로 작동하면서 수리하는 것

용어 정리

- **Failure**: A system fails when it cannot provide one or more of its services
  - 더 이상 서비스를 제공하지 못하여 작동하지 않는 시스템
  - ex) 웹 서버가 더이상 웹페이지를 주지 못함
- **Error**: The part of a system state that can lead to a failure
  - failure가 일어나게 된 어떤 특정한 시스템 상태
  - ex) 에러가 있는 패킷을 받게됨
- **Fault**: The cause of an error
  - ex) 패킷 에러가 난 원인으로 노이즈나 interference
  - ex) 프로그램 버그(programming error), 크래쉬 프로그램(failure))
- **Fault tolerance**: a system can provide its services in the presence of faults
  - fault가 있어도 서비스를 제공할 수 있는 시스템
  - 결과적으로 system fault가 error을 발생시키고, error는 failure로 이어진다. 이런 경우가 발생하더라도 시스템이 문제 없이 작동할 수 있다면 fault tolerance이다.

확장성과 신뢰성

- **Scalability(확장성)**
  - 사용자의 요구를 충족시키기 위해 컴퓨터 응용 프로그램 또는 제품(하드웨어 또는 소프트웨어)의 크기나 볼륨을 변경했을 때 계속 작동하는 기능
  - 중앙 집중식 컴퓨터를 사용하는 대신, 병렬 및 분산 컴퓨팅 시스템을 사용함으로써 인터넷을 통해 대규모 데이터 계산문제를 해결함.
  - 분산 컴퓨팅은 데이터 집양적이고 네트워크 중심이 된다.
- **신뢰성(Reliability)**
  - 문제 없이 연속적으로 사용될 수 있는 시간 (time)

빅데이터 모델링의 요구사항

① 빅데이터 운영 지원

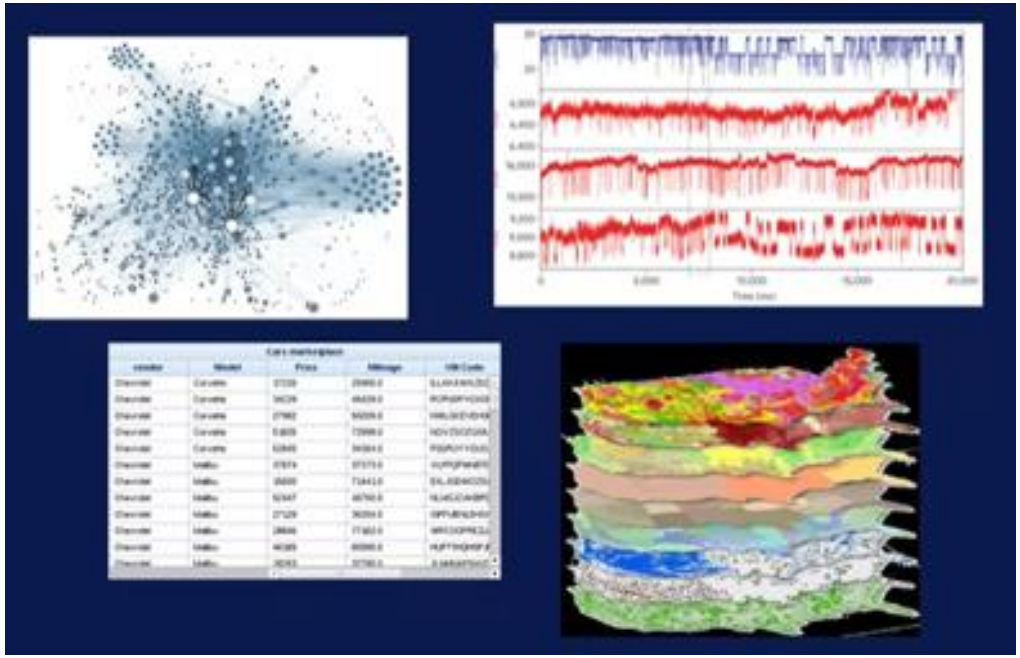
② 결함 허용

③ 더 많은 랙의 추가 가능

④ 특정 데이터 유형에 최적화

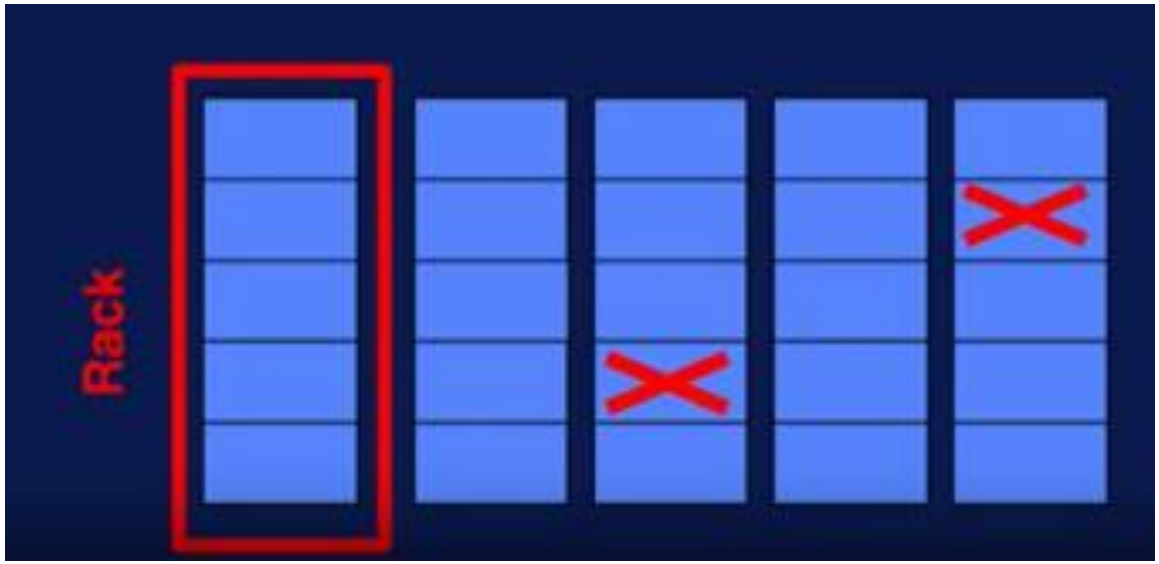
빅데이터 운영 지원

- Split volumes of data : 대용량 데이터의 분할
- Access data fast: 빠른 속도로 데이터에 접근
- Distribute computations to nodes: 빠른 속도로 랙 내의 노드로 배포



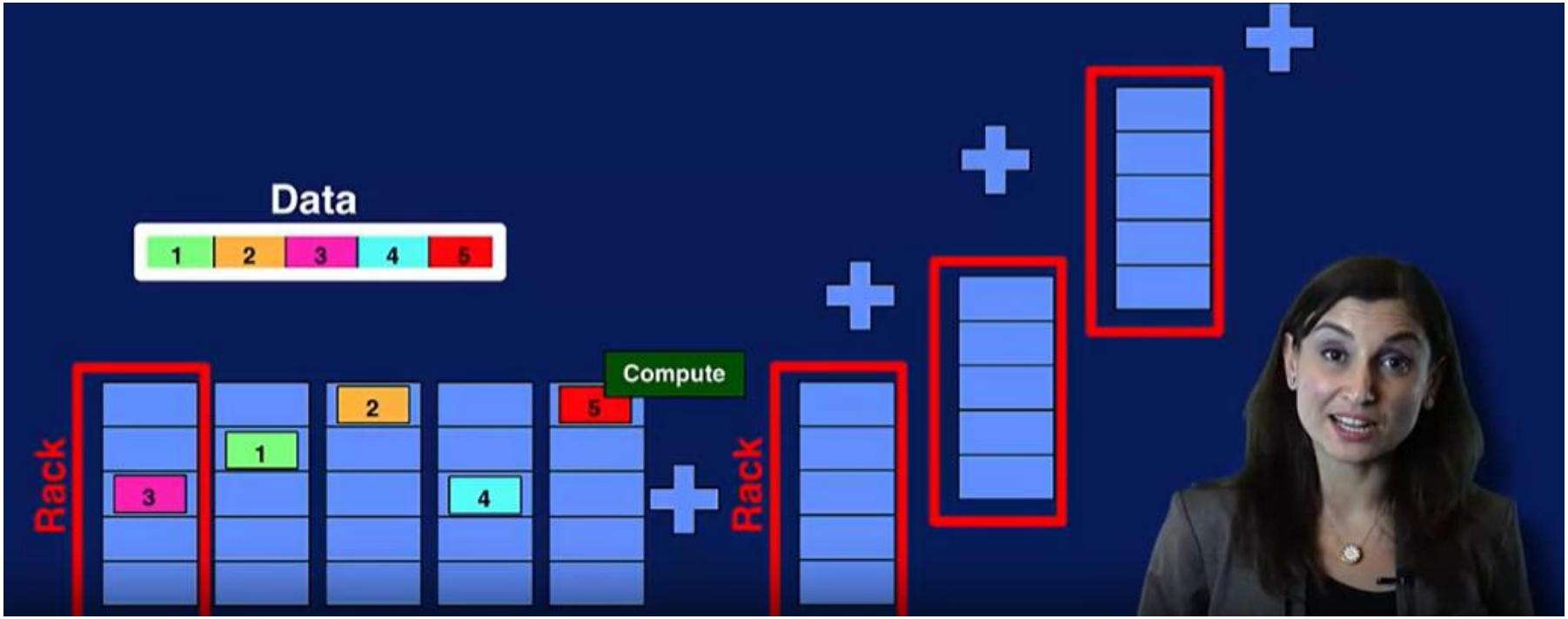
결함 허용

- Replicate data partitions: 데이터 파티션 복제
- Recover files when needed: 필요에 따라 파일 복제 및 복구가 가능해야 함.

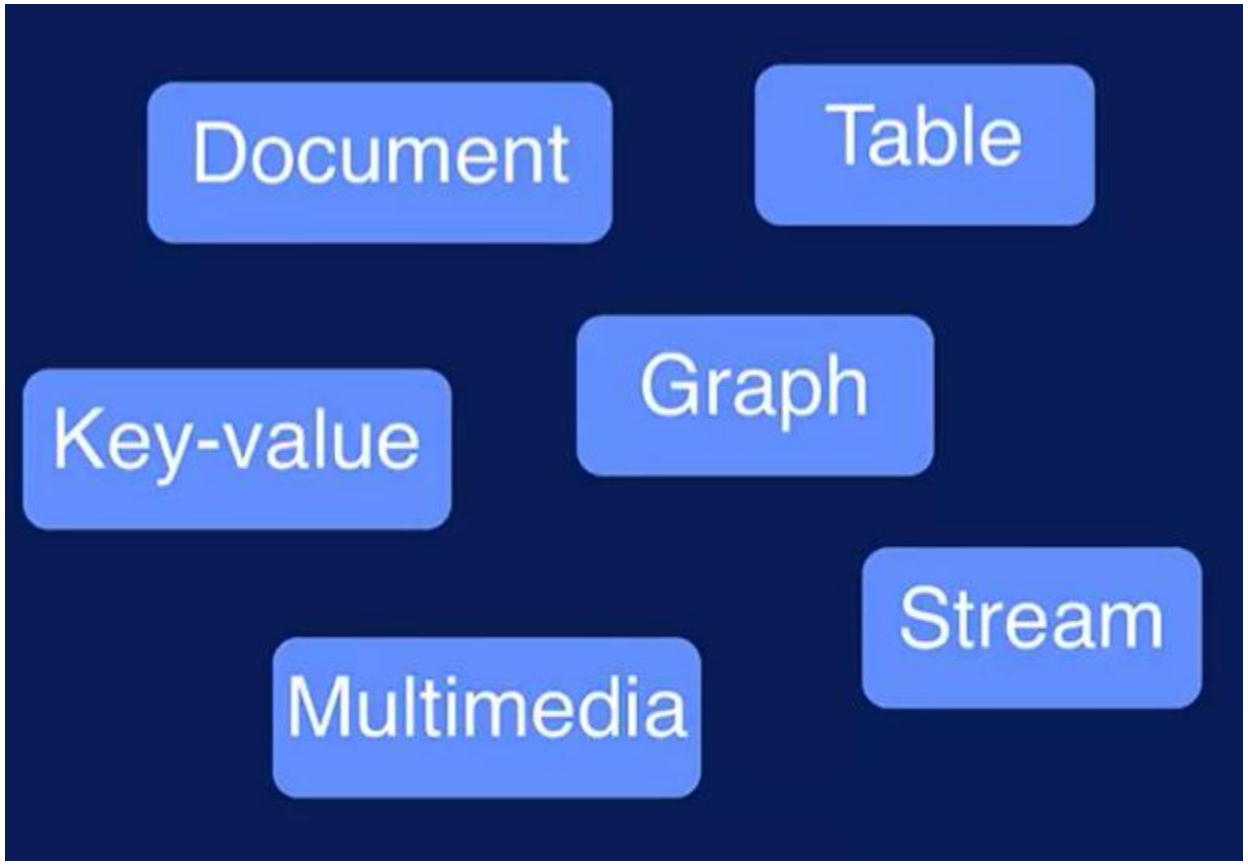


더 많은 랙의 추가 가능

확장가능성(**scalability**)



특정 데이터 유형에 최적화



빅데이터 모델링을 위한 시스템

- 하둡 분산 파일 시스템(HDFS)
- 맵리듀스(MapReduce) 스케일링 프로그래밍 모델

하둡(Hadoop) 특징

1. 오픈소스 자바 소프트웨어 프레임워크임
   - 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원함

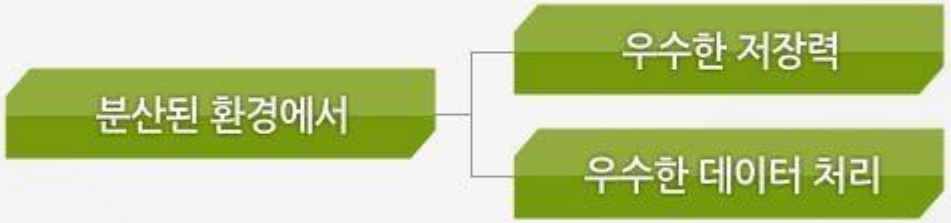
2. 오픈소스 하둡
   - 빅데이터 활용을 가능하게 만든 빅데이터 플랫폼의 핵심 기술임

3. 저비용으로 방대한 양의 데이터를 저장 및 처리할 수 있음
   - 저가 장비 및 스토리지(저장장치)를 활용함

하둡 분산파일시스템 특징

하둡이 사용하는 분산 저장소

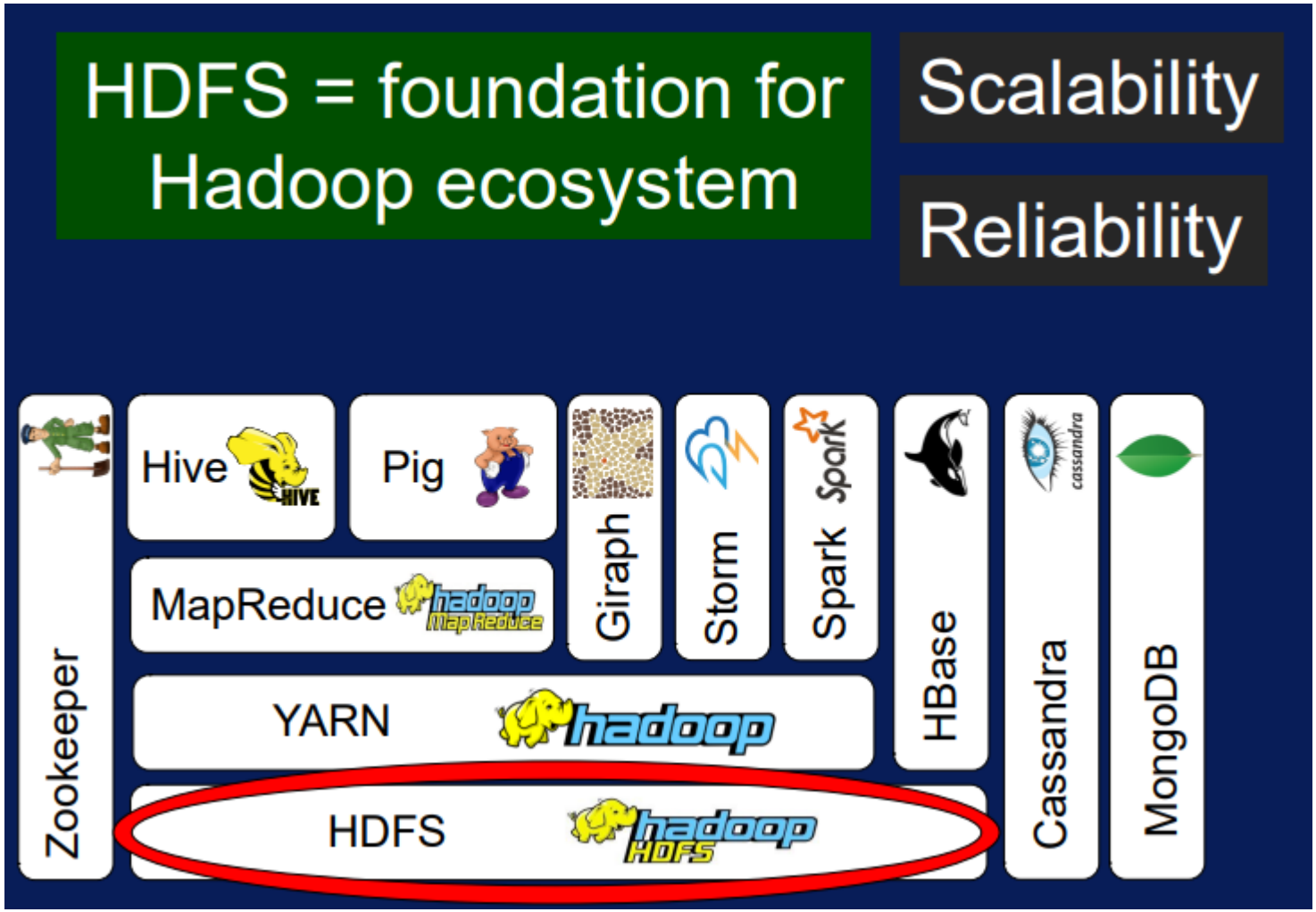
분산된 환경에서 다양한 형태, 초대용량의 데이터를 안전하게 저장할 수 있을 뿐만 아니라 저장되어 있는 데이터를 빠르게 처리할 수 있도록 설계됨



전체 성능이나 용량을 늘리기 위해 많은 서버를 이용하여 구축함

- 값싼 서버들 이용
- 높은 수준의 고장방지기능 이용

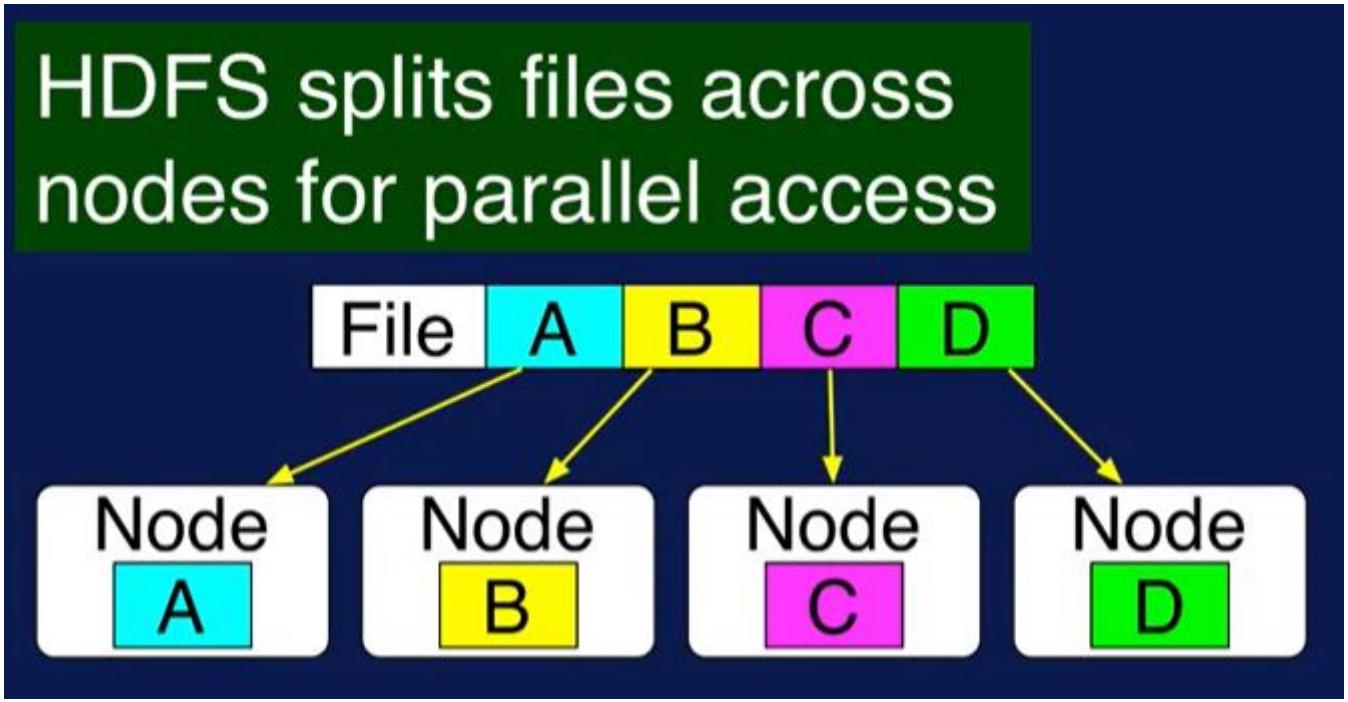
하둡 분산 파일 시스템(HDFS)



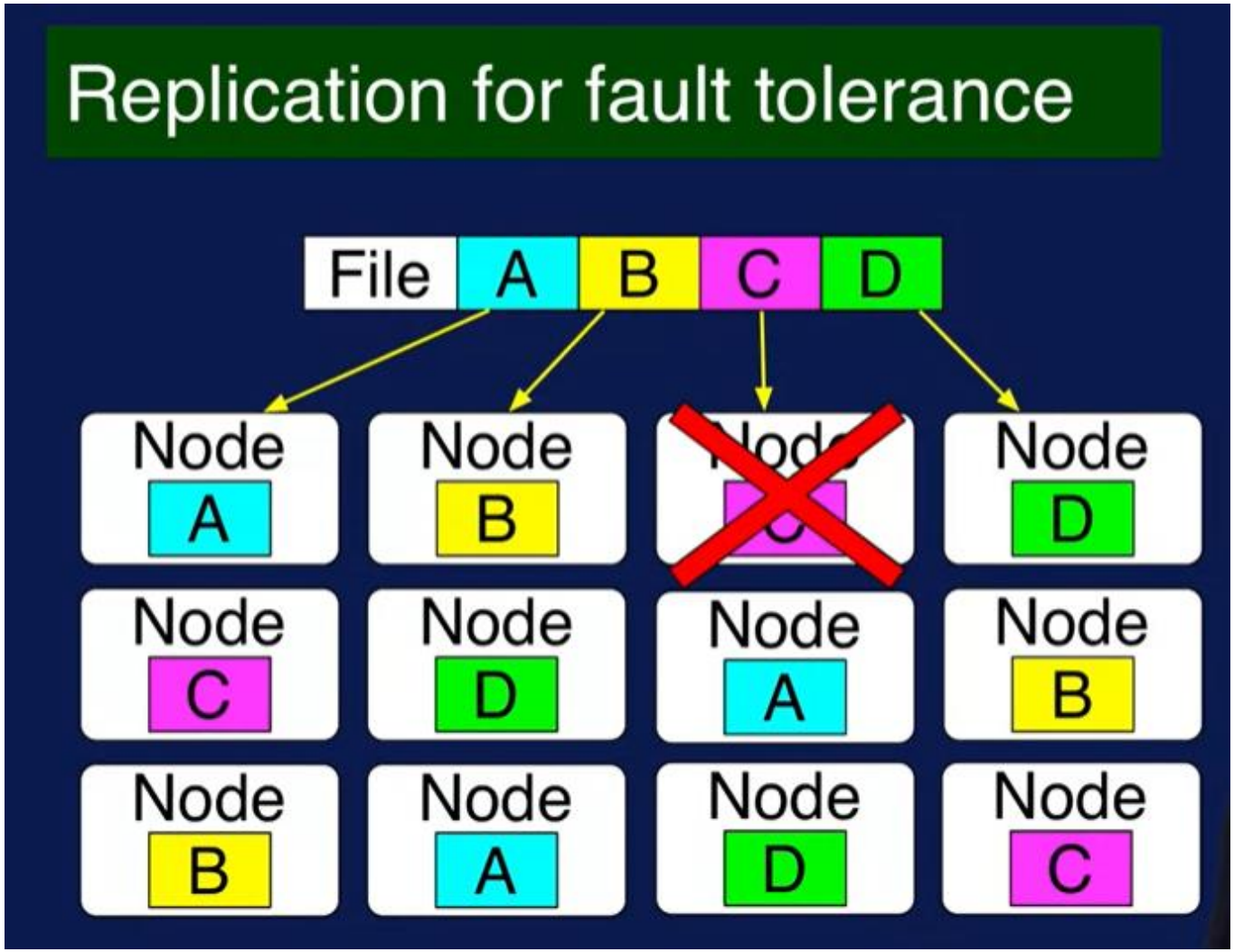
확장성

신뢰성

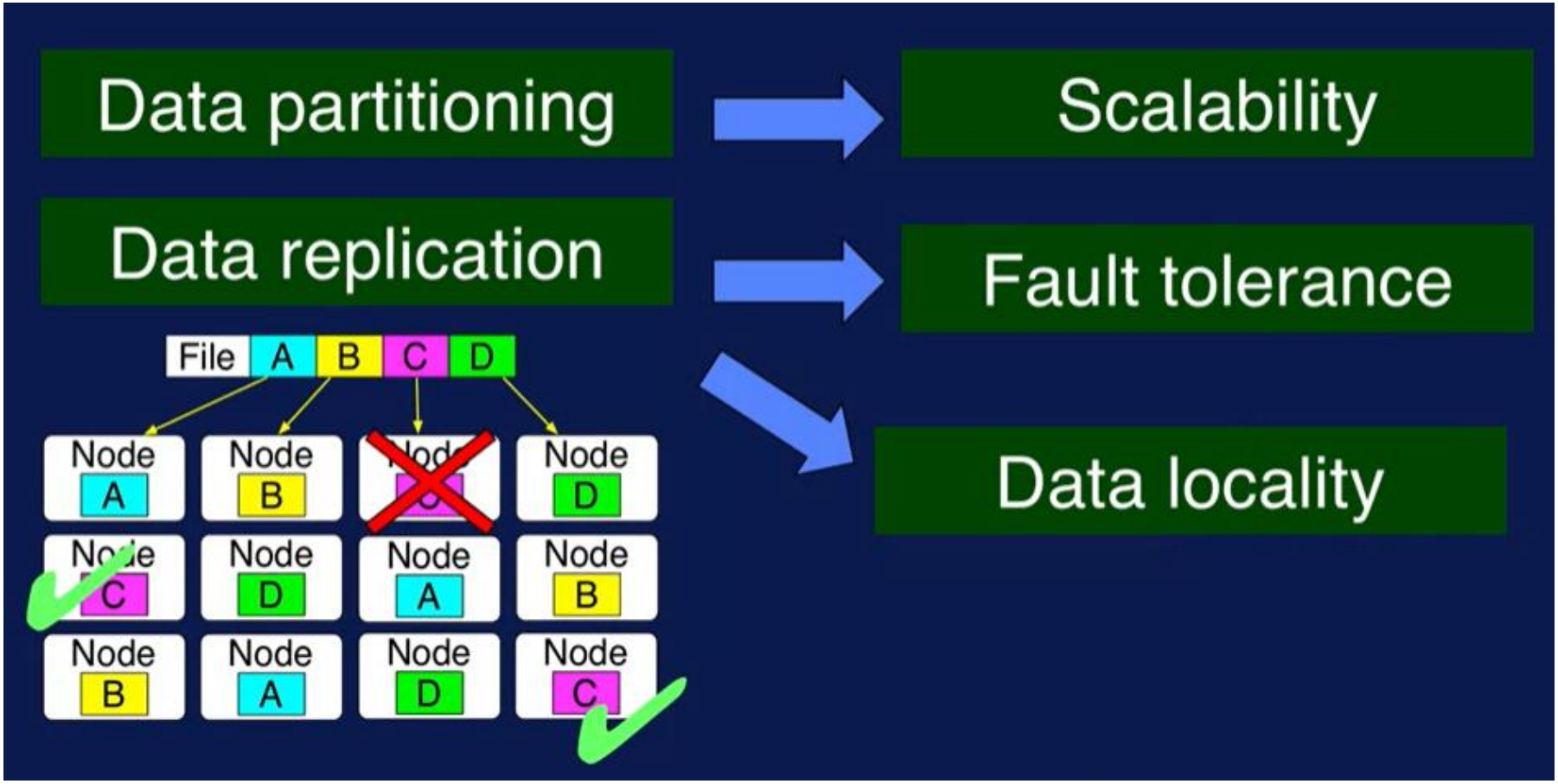
하둡 분산 파일 시스템(HDFS)



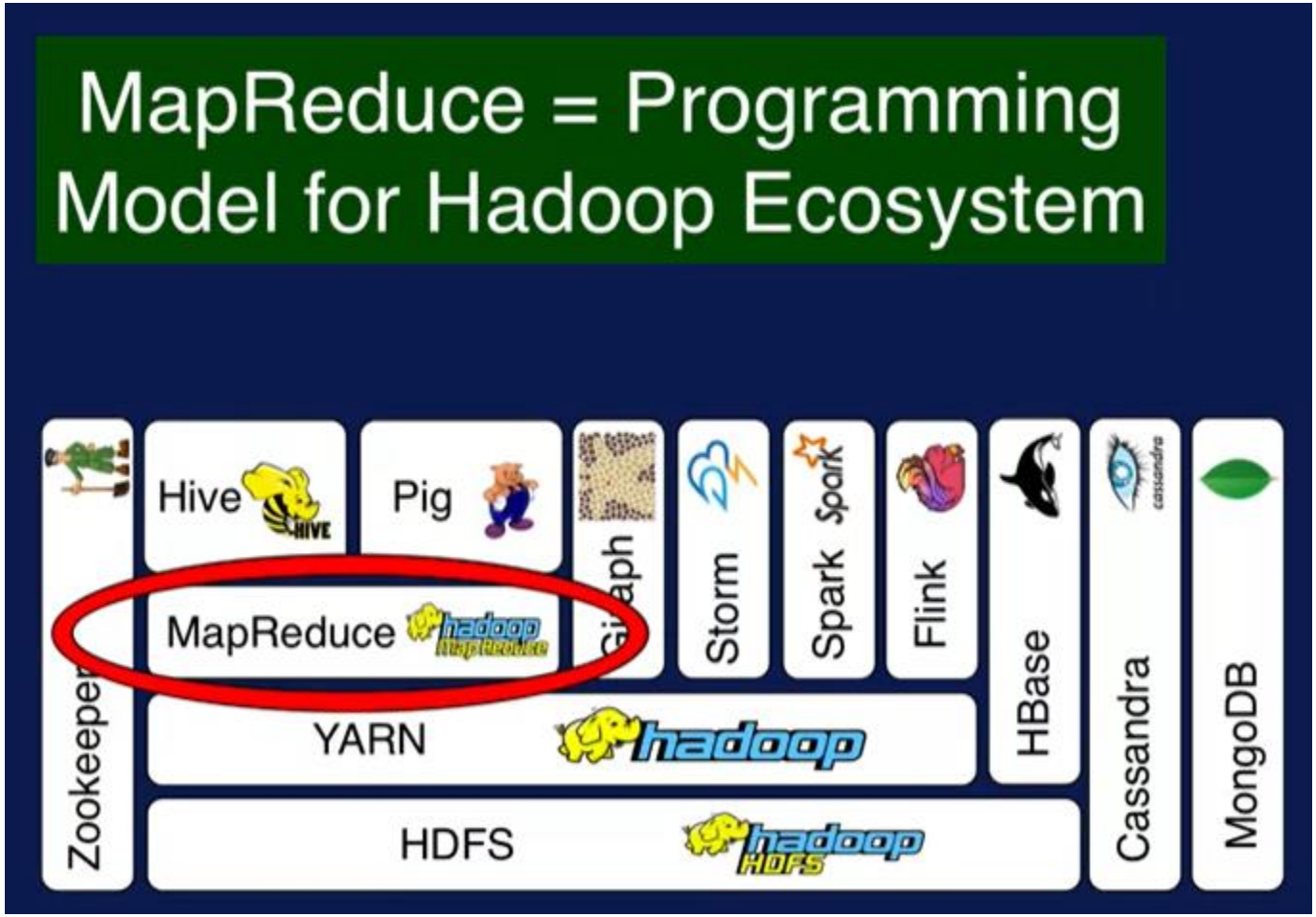
하둡 분산 파일 시스템(HDFS)



하둡 분산 파일 시스템(HDFS)

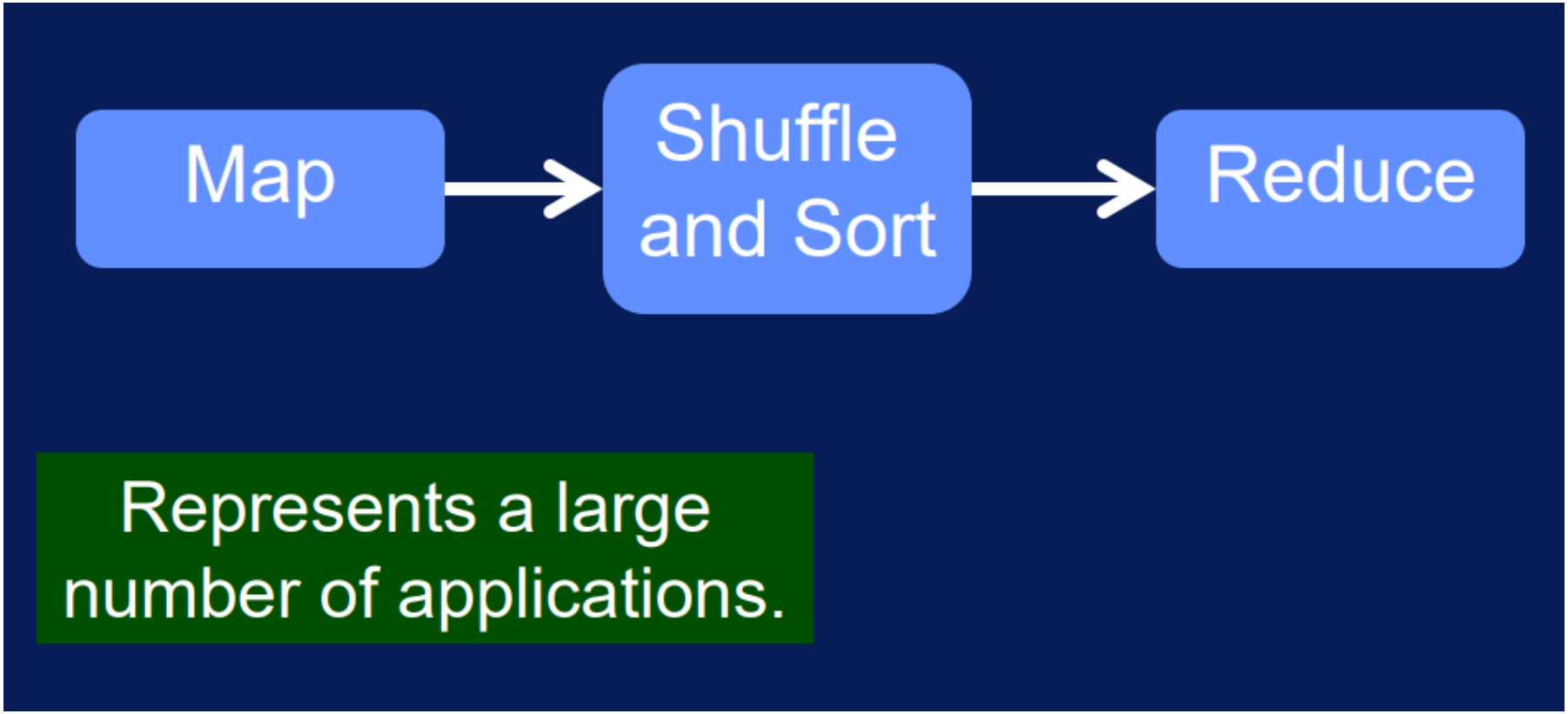


MapReduce 스케일러블 프로그래밍 모델

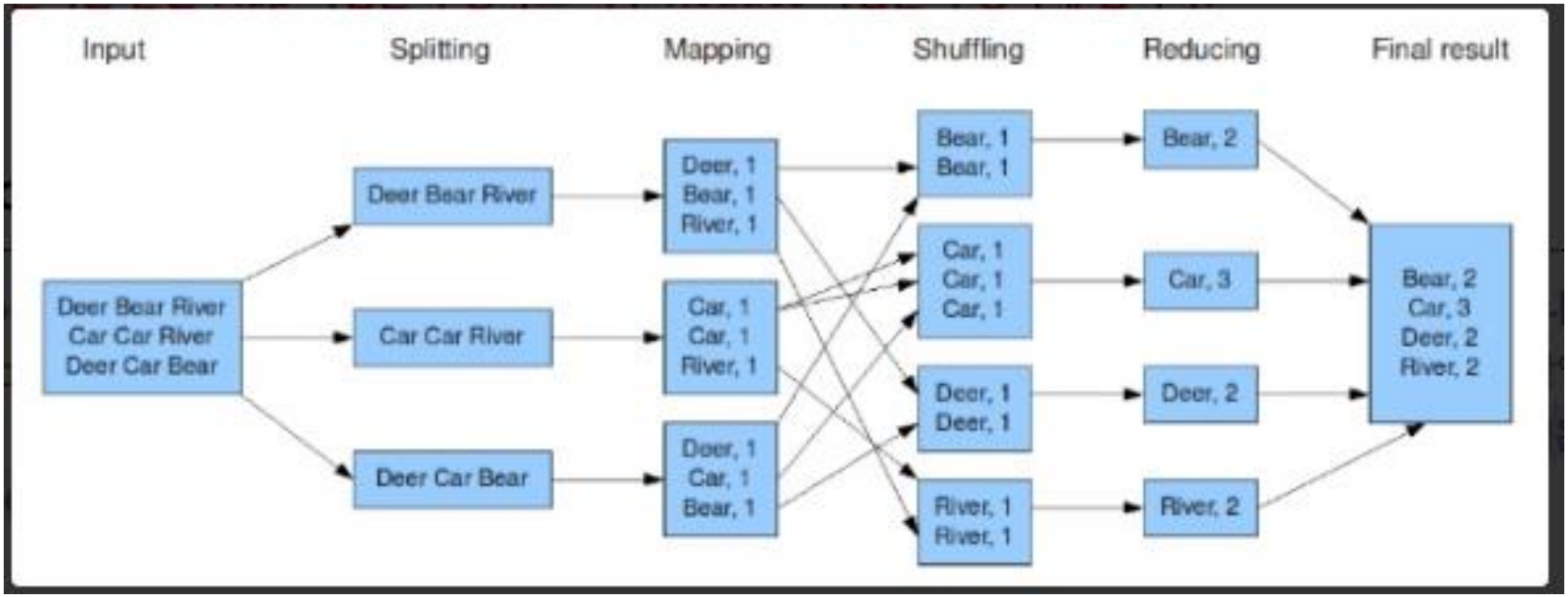


MapReduce 스케일러블 프로그래밍 모델

- Map-Reduce는 매핑(mapping), 셔플링(shuffle) 및 정렬(sort), 축소(reduce)의 세 가지 주요 단계로 구성된다.

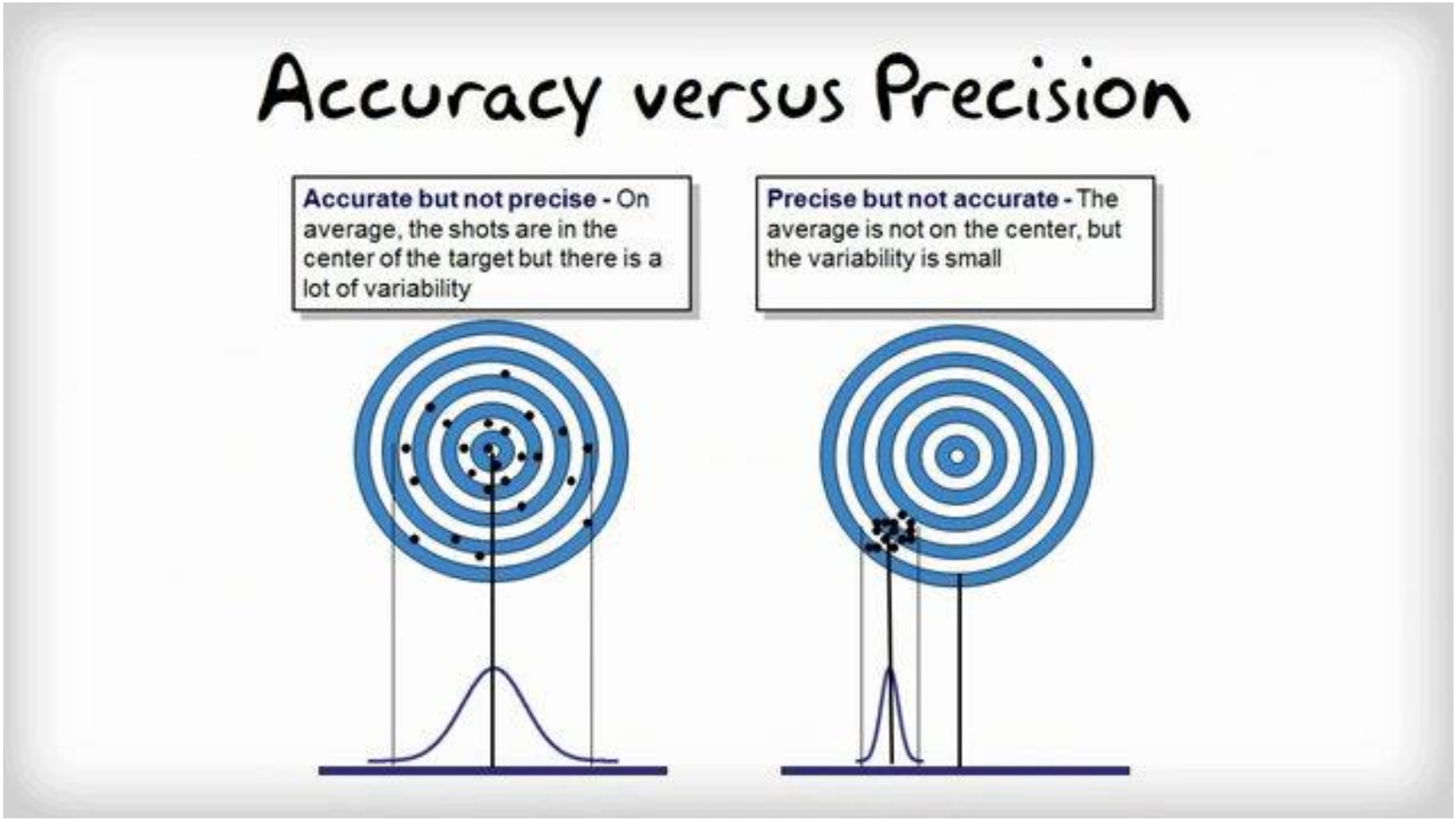


MapReduce 스케일러블 프로그래밍 모델

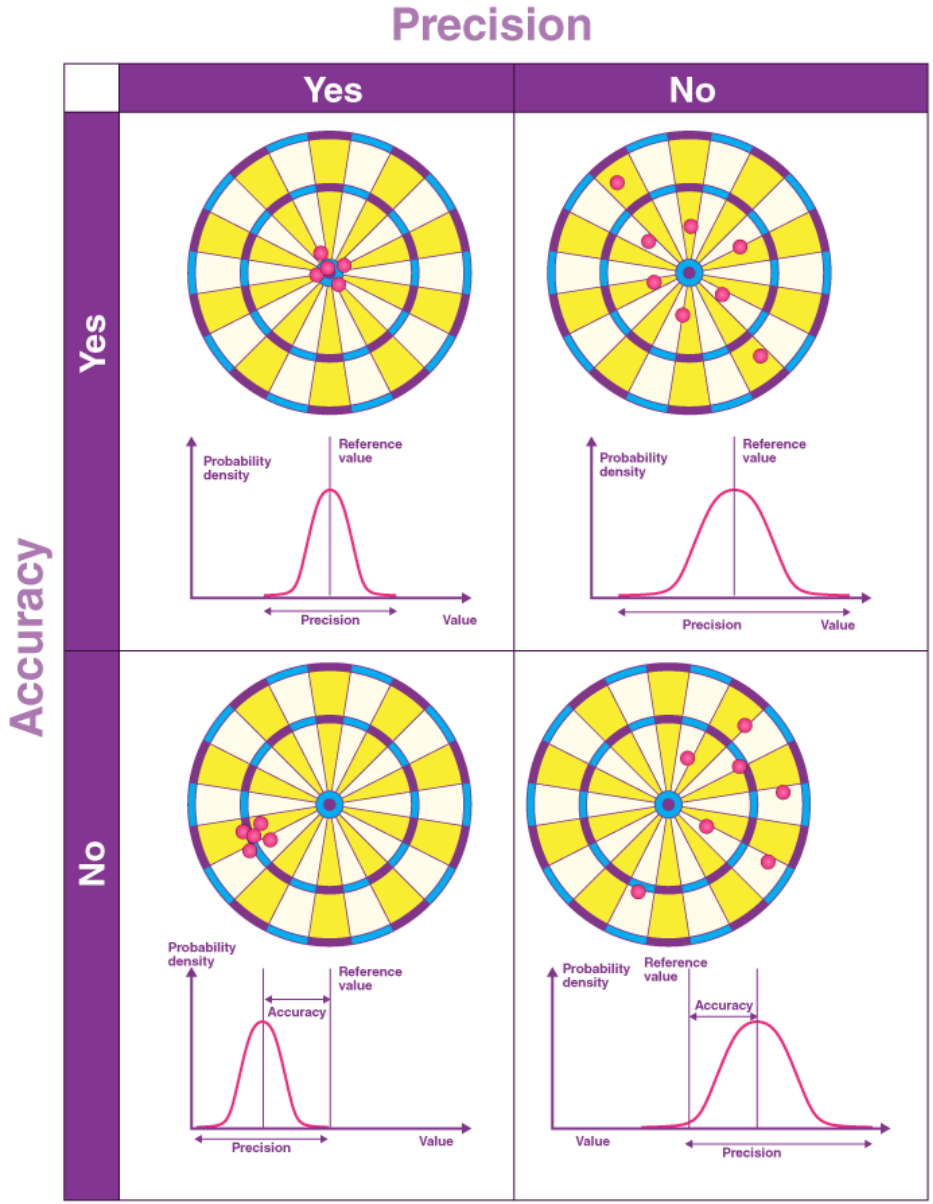


빅데이터의 필요 충분 조건

- Accuracy(정확도) & Precision(정밀도)



Accuracy(정확도) & Precision(정밀도)



Accuracy(정확도) & Precision(정밀도)

Accuracy(정확도)	Precision(정밀도)
정확도는 실제 측정과 절대 측정 사이의 일치 수준을 나타낸다.	정밀도는 동일한 요인의 여러 측정 값에 있는 변동 수준을 의미한다.
결과가 표준 값과 얼마나 일치하는지를 나타낸다.	결과가 서로 얼마나 밀접하게 가까이 있는지를 나타낸다.
단일 요소 또는 측정이 필요하다.	정밀도에 대한 의견을 제시하려면 여러 측정 또는 요인이 필요하다.
때때로 측정 값이 정확할 수 있다. 측정이 일관되게 정확하기 위해서는 정밀해야한다.	결과는 정확하지 않고 정밀할 수 있다. 또는 결과는 정확하면서 정밀할 수 있다.

빅데이터 분석 목적과 역할

빅데이터 분석

대량의 데이터로부터
숨겨진 패턴과 알려지지 않은
정보를 찾아내기 위한 과정

빅데이터 분석 목적

데이터 사이언티스트들에 의해
분석된 정보를 토대로
각 분야의 의사결정을 수행

- 의사결정 시 최선의 대안을 선택할 수 있도록 근거를 제시하는 중요한 역할을 함
- 불확실성이 높고 의사결정이 초래하는 파급효과가 큰 의사결정일수록 실제 데이터 분석을 바탕으로 의사결정을 해야 함
- 많은 기업에서 빅데이터를 활용하여 주요 의사결정을 내리고 있음
- 효과적인 빅데이터 분석을 위해서 일반적으로 빅데이터 분석 플랫폼을 구축함

빅데이터 분석 목표와 기술

더 짧은 시간 안에 보다 많은 정보를 빅데이터로부터 추출하는 것

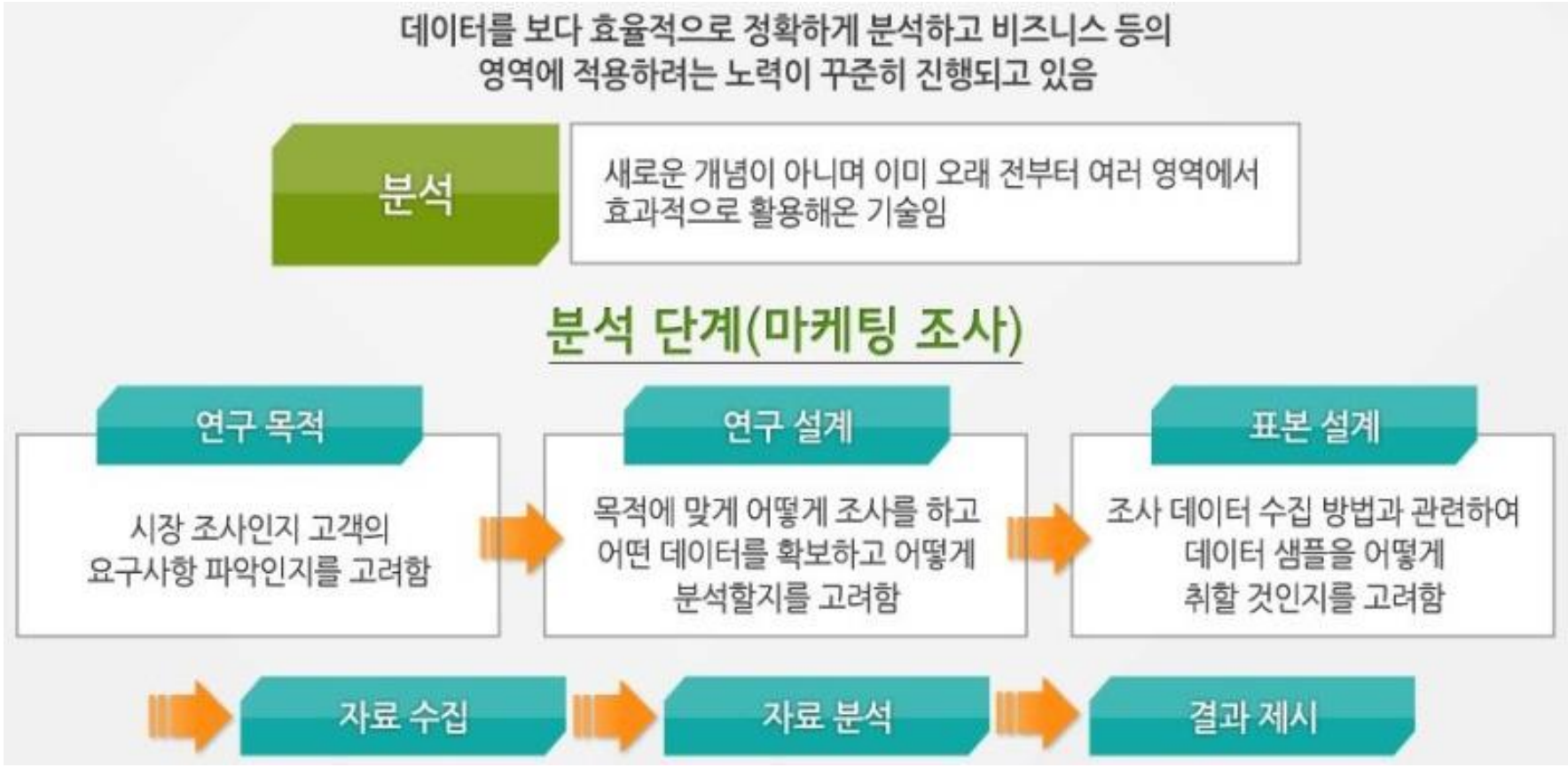
빅데이터 분석

- 데이터 마이닝
  - 대용량의 데이터베이스에 저장된 데이터에 숨겨진 중요한 정보와 지식을 추출하는 기술
- 예측 분석
  - 현황 정보 대신 예측 정보를 제공할 수 있는 분석

빅데이터 분석 관련 기술

- NoSQL
- 데이터베이스
- 하둡과 맵리듀스 등

빅데이터 분석 단계


데이터를 보다 효율적으로 정확하게 분석하고 비즈니스 등의
영역에 적용하려는 노력이 꾸준히 진행되고 있음
분석
새로운 개념이 아니며 이미 오래 전부터 여러 영역에서
효과적으로 활용해온 기술임
분석 단계(마케팅 조사)
연구 목적
시장 조사인지 고객의
요구사항 파악인지를 고려함
연구 설계
목적에 맞게 어떻게 조사를 하고
어떤 데이터를 확보하고 어떻게
분석할지를 고려함
표본 설계
조사 데이터 수집 방법과 관련하여
데이터 샘플을 어떻게
취할 것인지를 고려함
자료 수집
자료 분석
결과 제시

분석 수행 단계

1 문제인식

- 문제가 무엇인지, 왜 이 문제를 해결해야 하는지, 문제 해결을 통해 무엇을 달성할 것인지를 명확히 하는 단계

2 관련 연구 조사 단계

- 문제와 직간접적으로 관련된 **지식을 각종 문헌**(예 잡지, 책, 보고서, 논문 등)을 조사하면 문제를 더욱 명확히 할 수 있을 뿐 만 아니라 **문제와 관련된 주요 요소(변수)들을 파악할 수 있는 단계**

3 모형화(변수 선정) 단계

- 모형은 **문제(연구 대상)를 의도적으로 단순화**한 것을 말하며, 모형화는 **문제와 본질적으로 관련된 변수**만을 추려서 재구성하는 단계

4 자료 수집(변수 측정) 단계

- 인식된 문제는 모형화를 통하여 주요 변수로 재구성되고 측정이라는 과정을 거치면서 자료가 되는 단계

1차 자료	조사자가 관찰, 설문조사, 실험을 통하여 직접 자료를 수집하는 것
2차 자료	다른 사람에 의해 이미 수집, 정리되어 있는 자료

5 자료 분석 단계

- 나열된 숫자에서 변수 간의 규칙적인 패턴, 즉 **변수간의 관련성**을 파악

6 결과 제시 단계

- 자료 분석 결과가 의미하는 바를 해석하여 **의사결정자에게 구체적인 조언**을 하는 단계

빅데이터 분석 분류


빅데이터 분석
비즈니스 분석
(Business Analytics)
현재의 현상과 결과적인 관점을 제시하는데 초점을 두는 것
기업의 품질부서에서 현재의 품질수준 실적에 대해 집중적으로 분석
고급 분석
(Advanced Analytics)
대용량의 데이터로부터 숨겨진 패턴을 발견하고 상황을 예측하는 것
구조화 및 비구조화된 복잡한 형태의 데이터 요인들 간의 상관관계 확인
의미 있는 데이터의 패턴을 식별하고 예측하기 위한 모든 기법과 기술들을 포괄
미래의 품질 수준이 어떻게 될지 예측할 수 있는 모델에 근거하여 예측

빅데이터 분석 도구

R 프로그래밍 언어 | 빅쿼리(BigQuery) | 프레스토(Presto)

R 프로그래밍 언어 개요

오픈소스 프로젝트로 통계 계산 및 시각화를 위한 언어 및 개발 환경을 제공함

- 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현이 가능함
- 통계적 컴퓨팅 언어로 다양한 통계 분석에 용이함

"현재 R 프로그래밍 언어를 이용하여 다양한 빅데이터 분석 및 예측 분석 등을 포함한 고급 분석 기술들이 연구 및 개발되고 있음"

R 프로그래밍 언어 장점

- 사용자가 제작한 패키지를 추가하여 기능을 확장 가능
  - 핵심적인 패키지는 R 프로그래밍 언어와 함께 설치됨
  - CRAN(the Comprehensive R Archive Network)을 통해 700개 이상의 다양한 기능을 가지는 패키지를 내려 받을 수 있음
- 그래픽 기능
  - 수학 기호를 포함할 수 있는 출판물 수준의 그래프를 제공함

R 프로그래밍 언어 | 빅쿼리(BigQuery) | 프레스토(Presto)

빅쿼리(BigQuery) 개요

- 구글의 대용량 데이터를 처리할 수 있도록 개발된 쌍방향 서비스
- 사용자 혹은 개발자 등은 SQL과 같은 익숙한 쿼리문 등을 이용해 인사이트를 전달할 수 있음
  - 일반적으로 SQL문이라고도 불리는 쿼리문이 작성됨

쿼리문

- 데이터베이스에 저장된 값을 불러내기 위함
- 절차적 언어로 작성된 프로그램 문장

SQL

- Structured Query Language 의 약자
- 구조화된 절차적인 데이터베이스 언어

빅쿼리(BigQuery)를 이용하는 방법

먼저 이용자가 데이터 세트를 구글 시스템에 업로드 함

빅쿼리 API를 이용하여 이에 대한 쿼리를 던지는 방식으로 이용함

빅쿼리(BigQuery)를 출시한 목적

구글이 자체 데이터센터가 없는 기업도 쉽게 데이터를 분석할 수 있는 환경을 만들어주기 위해 출시함

- 웹 광고나 실시간 관리 시스템, 온라인 게임의 데이터 현황을 쉽게 관리할 수 있음
  - 예 제약회사
    - 전세계 판매량과 광고 데이터를 바탕으로 일일 마케팅 최적화 전략을 세울 수 있게 됨
    - 사용자 클릭을 바탕으로 제품 권고 사항을 만드는 일도 쉬워짐

R 프로그래밍 언어

빅쿼리(BigQuery)

프레스토(Presto)

페이스북에서 개발한 빅데이터 분석 도구

- 페이스북이 300페타바이트에 달하는 엄청난 내부 데이터를 분석하기 위해 만듦

하둡을 위한 SQL 처리 엔진

- 데이터 분석가가 기존의 SQL 언어로 대용량의 데이터를 대화형 분석을 수행할 수 있도록 해줌

"기존에 많이 쓰는 하이브/맵리듀스 보다 CPU 효율성과 대기 시간이 10배 빠름

빅데이터 분석 활용 기법

1 통계적 분석

- 전통적인 분석 방법으로 주로 수치형 데이터에 대하여 확률을 기반으로 어떤 현상의 추정, 예측을 검정하는 기법

기술통계량 | 상관분석 | 회귀분석 | 분산분석 | 주성분분석

- 대표적으로 평균(산술평균, 중앙값, 최빈값), 분산, 표준편차 등을 구하는 것
- 전체 데이터 그룹이 주로 어디에 위치하고 있으며 이를 중심으로 얼마나 산포를 가지는지를 확인 가능함

평균

- 데이터 집합의 중심적인 경향을 표현하는 값
- 전체 데이터의 합을 전체 데이터 개수로 나누어 산출

분산

- 평균을 중심으로 각각의 데이터 들의 편차를 구하고 편차의 제곱을 모두 더한 후 전체 데이터 개수에서 하나를 뺀 값으로 나눈 값

표준편차

- 분산 값의 제곱근으로 산포를 의미

기술통계량 | **상관분석** | 회귀분석 | 분산분석 | 주성분분석

- 두 변수간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법
  - ✓ 하나의 변수가 증가할 때 비례 또는 반비례적으로 다른 한 변수가 증가 또는 감소하는 정도를 규명
- 분석 시 서로 관계를 가지는 변수들을 찾아 낼 수 있음

기술통계량 | 상관분석 | 회귀분석 | 분산분석 | 주성분분석

- 연속형 변수들에 대해 독립변수와 종속변수 사이의 상관관계에 따른 수학적 관계식을 구하여 어떤 독립변수가 주어졌을 때 이에 따른 종속변수를 예측하는 방법

✓ 종속변수 값을 예측할 수 있는 수학적 모델식을 구성

✓ 특정한 독립변수의 값을 가지는 경우

✓ 종속변수의 값을 예측 가능함

독립변수

- 종속변수에 영향을 주는 요인을 가지는 변수

종속변수

- 독립변수의 값에 의해 종속적으로 영향을 받는 변수

연속형 변수

- 독립변수와 종속변수가 일반적으로 연속형의 값을 가지는 경우

기술통계량 | 상관분석 | 회귀분석 | **분산분석** | 주성분분석

- 3개 이상의 집단에 있어서 평균치 차이가 존재하는지를 검증함
- F분포를 이용하여 가설검정을 하는 방법
- F분포
  - ✓ 두 개 이상 다수의 집단을 비교하고자 할 때 집단 내의 분산, 총평균과 각 집단의 평균의 차이에 의해 생긴 집단 간 분산의 비교를 통해 만들어짐
- 다수의 집단에 있어서 평균치가 차이가 있는지 유의성을 판정할 수 있음

기술통계량 | 상관분석 | 회귀분석 | 분산분석 | **주성분분석**

- 다양한 변수들에 대해 분석하는 다변량(Multivariate) 분석으로 많은 변수들로부터 몇 개의 주성분들을 추출하는 방법
- 많은 변수들을 관리할 수 있는 관리의 로드가 줄어들 수 있음

분석 용도에 따른 데이터 분석기법

2 데이터 마이닝

- 대용량 데이터로부터 **패턴인식, 인공지능 기법 등을 이용**하여 숨겨져 있는 데이터간의 상호 관련성 및 유용한 정보를 추출하는 기술
- 기존 데이터베이스에 마이닝 기술을 적용하여 이들 데이터 간에 숨은 의미 있는 관계성을 다양한 방법으로 발견한 후 이를 현실에 효과적으로 적용하는 방법론으로 사용됨

3 텍스트 마이닝

- 텍스트 기반의 데이터로부터 새로운 정보를 발견할 수 있도록 정보 검색, 추출, 체계화, 분석을 모두 포함하는 Text-processing 기술 및 처리 과정
- 텍스트 내에 존재하는 단어의 등장횟수 등을 평가하여 **문서간의 유사성을 수치화** 하는 텍스트 데이터를 분석하는 방법
- 유사 문서 분류 및 문서 내 정보 추출과 같은 결과를 산출이 가능함

4 소셜 네트워크 분석

- 대용량 소셜 미디어를 언어분석 기반 정보 추출로 탐지함
- 시간의 경과에 따라 유통되는 이슈의 전체 과정을 모니터링하고 향후 추이를 분석함
- 소셜 네트워크 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 분석함
- 활용
  - ✓ 주로 마케팅을 위하여 소셜 네트워크 상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾음
  - ✓ 수학의 그래프 이론을 이용하여 소셜 네트워크의 연결 구조와 연결 강도 등을 바탕으로 사용자의 영향력을 측정함
- 텍스트 마이닝 기법에 의해 주로 이루어짐
- 확산된 내용과 함께 연결의 맥락을 파악하여 분석하는 기법

5 평판 분석(Sentiment Analysis)

- 오피니언 마이닝이라고도 불림
- 소셜미디어 등의 정형/비정형 텍스트의 긍정(Positive), 부정(Negative), 중립(Neutral) 의 선호도를 판별하는 기술
- 활용
  - ✓ 특정 서비스 및 상품에 대한 시장규모 예측
  - ✓ 소비자의 반응
  - ✓ 입소문 분석(Viral Analysis) 등
- 정확한 오피니언 마이닝을 위해서는 전문가에 의한 선호도를 나타내는 표현/단어 자원의 축적이 필요함

6 군집 분석(Cluster Analysis)

- 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 군(Group)을 발굴하는데 사용함

예
- ✓ 트위터 상에서 주로 사진/카메라에 대해 이야기하는 사용자 군
- ✓ 자동차에 대해 관심 있는 사용자 군

- 관심사나 취미에 따른 사용자 군을 군집 분석을 통해 분류 가능

경청 해 주셔서 감사합니다.

경남대학교 전하용
hayongj@kyungnam.ac.kr