빅 데이터 기술 개요 및 서비스

8조

목차

빅데이터란?

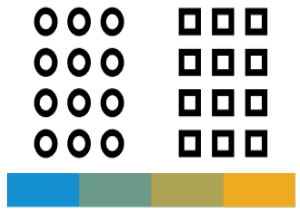
빅데이터(Big Data)

기존 데이터 처리도구(ex. 스프레드시트)로는 관리나 분석이 어려운
매우 크고 복잡한 데이터 집합

-> 소셜 미디어 게시물에서 금융 거래까지 방대하게 사용 / 생성됨

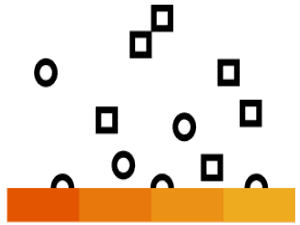
빅데이터 유형

빅데이터의 3가지 유형



Structured data

정형 데이터



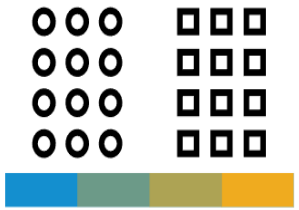
Unstructured data

비정형 데이터

Semi-structured data

반정형 데이터

빅데이터 유형

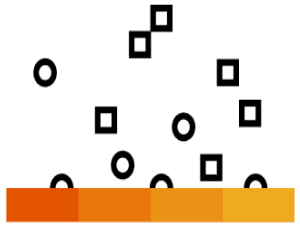


Structured data

정형 데이터

- 데이터가 행렬과 같은 형태 등으로 구조화된 데이터
- 구조화가 되어 검색이나 조직화가 용이함
- SQL(Structured Query Language) 로 정형 데이터를 관리
- 재고 데이터베이스, 금융 거래 목록 등에 사용됨

빅데이터 유형



Unstructured data

비정형 데이터

- 자유로운 형태로 존재하는 데이터
- 데이터에서 얻을 수 있는 가치는 크지만 분석에 많은 비용이 소모
- 주로 NoSQL(Not only SQL)로 데이터 관리
- 오디오, 동영상, 이미지 등 다양한 형태로 존재

빅데이터 유형

Semi-structured data

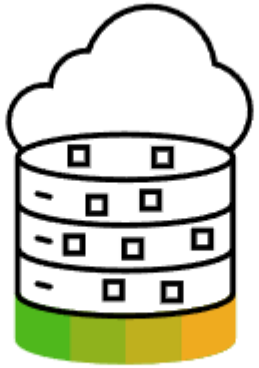
반정형 데이터

- 정형과 비정형의 하이브리드 형태
- 형태는 구조화를 이루지만 내용은 비정형인 데이터(ex. 이메일)
- 정보 수정이 쉽고 전송과 공유가 용이
- html, JSON이 대표적인 반정형 데이터 구조를 가짐

빅데이터 작동 방식

Gather Big Data

소스에서 방대한
데이터를 수집



Store Big Data

데이터를 가공 후
스토리지 업로드

Analyze Big Data

분석하여 패턴,
관계 등의 정보
획득

빅데이터 작동 방식-분석 방법

서술적 분석
기본 특성을 이해하기 위해 과거 데이터 요약과 설명에 중점

진단적 분석
데이터의 심층적인 분석으로 서술적 분석에서 관찰된 패턴과 추세 식별

예측 분석
과거 데이터, 통계 모델링, 머신 러닝을 통해 추세를 예측

규범적 분석
앞선 분석들을 기반으로 향후 작업 최적화를 위한 권장 사항을 제공

빅데이터 사용기술-

Hadoop

- Java 기반 오픈소스 프레임워크, 대규모 데이터 저장과 처리를 관리
- 분산형 스토리지와 병렬처리로 빅데이터 및 분석 작업 처리 함

Hadoop의 모듈

Hadoop 분산 파일 시스템(HDFS)

상용 하드웨어에서 실행 가능한 대규모 데이터 세트 관리 파일 시스템

YARN(Yet Another Resource Negotiator)

리소스를 관리하고 이를 통해 사용자의 애플리케이션을 예약하는 리소스 관리 플랫폼

빅데이터 사용기술-

MapReduce

대규모 데이터 처리를 위한 프로그래밍 모델

입력 데이터를 분할하여 분산 처리 후(Map) 이를 다시 하나로 합치는 (Reduce) 병렬 처리 기법

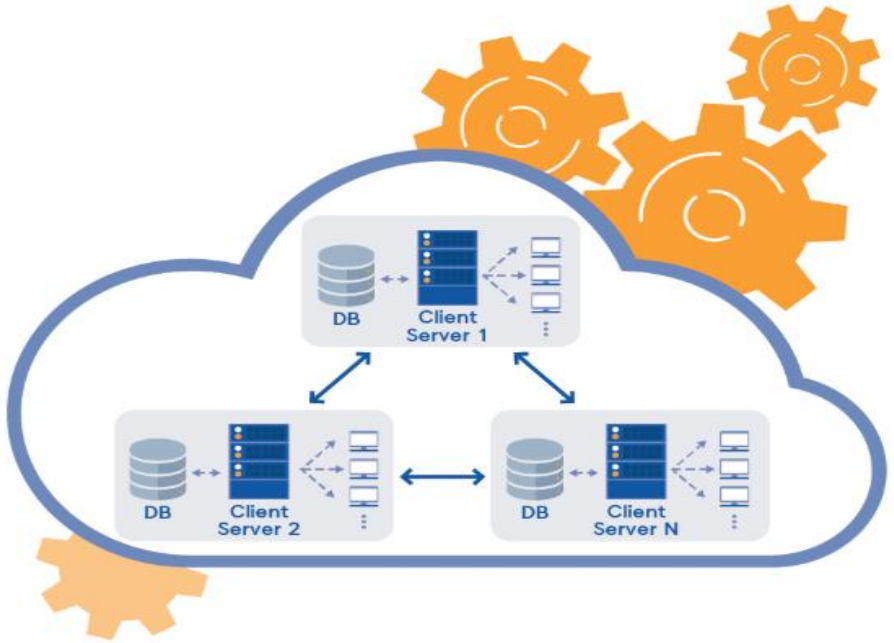
Hadoop Common

Hadoop 모듈이 사용 및 공유하는 라이브러리와 유틸리티가 포함

빅데이터 사용기술

분산 처리(Distributed Processing)

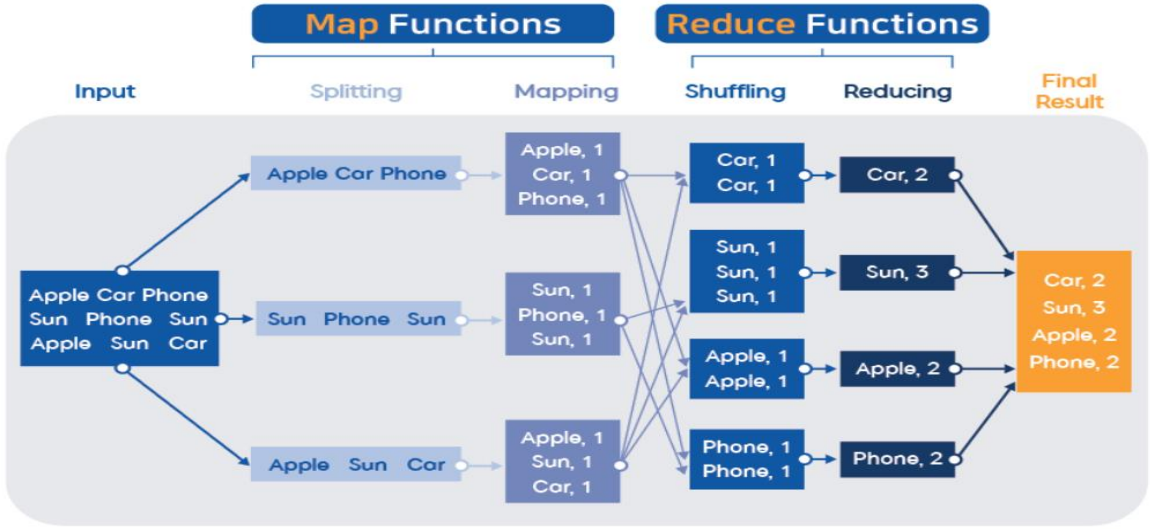
- 하나의 컴퓨터 시스템이 처리나 제어하는 기능을 여러 컴퓨터 시스템에 분산하여 처리하는 기법
- 빅데이터 플랫폼인 Hadoop과 NoSQL DB를 통해 대량의 데이터 저장과 처리를 가능케 함



빅데이터 사용기술

병렬 처리(Parallel processing)

- 복수의 처리 장치를 사용하여 하나의 프로그램상의 다른 작업을 동시 처리하여 부하를 줄이고 속도를 향상



빅데이터 사용기술-

Spark

- SQL, 스트리밍, 머신 러닝 및 그래프 처리를 위한 기본 제공 모듈이 있는 데이터 처리 엔진
- Hadoop의 맵리듀스 기법보다 100배 빠른 속도로 작업이 가능
- Spark는 인메모리 처리에 주로 사용(Hadoop은 주로 디스크 사용량이 많은 작업에 사용)

빅데이터의 V

빅데이터 활용에 중요한 V

용량(Volume)

데이터를 대량 처리 가능한가?

속도(Velocity)

데이터 수신 및 속도가 빠른가?

다양성(Variety)

다양한 데이터 유형을 활용 가능한가?

진실성(Veracity)

데이터가 신뢰가능한가?

가치(Value)

데이터가 유용한 가치를 지니는가?

빅데이터의 과거

- 2005년 페이스북, 유튜브에서 생성되는 데이터가 주목 받음

- 같은 해 빅데이터 집합을 저장하고 분석하는 오픈 소스 Hadoop 프레임워크가 개발
  + NoSQL 데이터베이스도 각광 받음

빅데이터의 현재

- Hadoop 등장 이후 개발된 Spark를 통해 빅데이터를 간단하고 빠르게 처리함
- 사람뿐만이 아닌 IOT(사물인터넷)과 AI를 통한 머신러닝으로 데이터 생성량 증가
- 25년도에는 175ZB로 늘어날 것으로 예상

빅데이터의 미래

- **AI와 머신러닝 기술의 빅데이터 결합**

대규모 데이터 세트에서 자동으로 패턴을 발견 및 예측하여 고도화된 인사이트 도출

- **사물 인터넷과 연계**

IOT를 통하여 날씨, 교통, 보안 등의 데이터를 빠르게 수집하고 처리

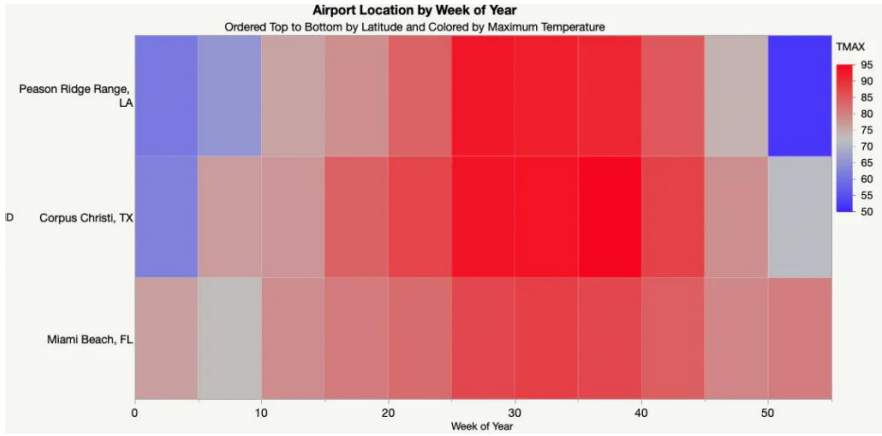
- **상업용 거래 데이터 처리**

대형 온라인 쇼핑몰의 거래량 증가에 따른 거래 데이터 처리량이 더 증가

대표 서비스 사례 - Uber

- 승객요청, 교통 상황 등의 데이터를 실시간 수집하여 수요 증가 파악과 드라이버를 배치
- 히트맵 시각화와 배치 매칭 알고리즘으로 도로 데이터를 효율적으로 파악

히트맵 시각화 : 데이터의 값의 크기나 빈도를 색으로 표현하여 쉽게 확인하는 법



대표 서비스 사례 - NETFLIX

- **방대한 데이터 처리를 위해 사용자 상호작용 토큰화 사용**

사용자 상호작용 토큰화 (Tokenizing User Interactions)

사용자의 클릭 기록을 의미 있는 이벤트로 요약하고 중복을 제거하여 정보밀도를 높이는 방법

-> 중요한 정보를 유지하여 사용자의 진짜 관심도를 파악하게 함

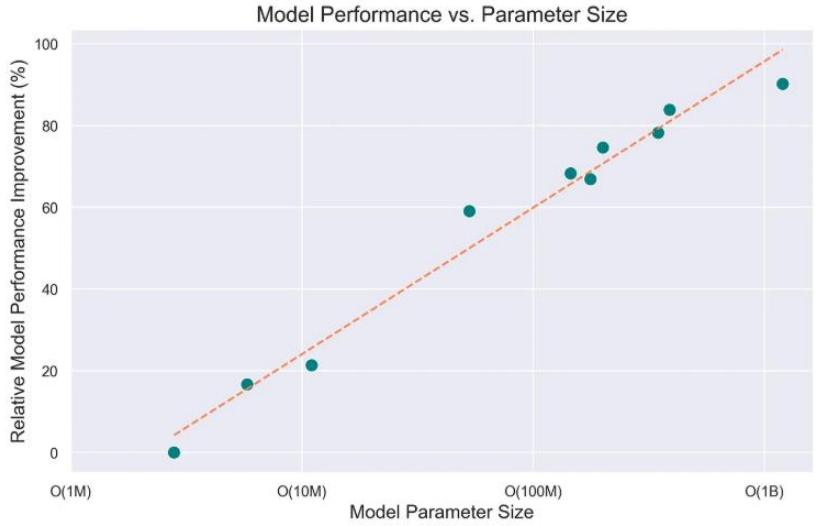
대표 서비스 사례 - NETFLIX

- 대규모 언어 모델(LLM)의 영감을 받아 추천 기능에 스케일링 법칙을 적용

  스케일링 법칙(Scaling Laws)

  데이터와 모델 규모를 확장하면 성능도 비례하는 현상

  -> 데이터 양, 질, 모델 크기 등을 동시확장을 통해 정밀한 추천을 가능케함



대표 서비스 사례 - amazon

- 매일 수백만 건의 주문과 테라바이트(10^{12})급 로그를 처리를 위해 엑사바이트(10^{18})급 인프라 구축

대표 서비스 사례 - amazon

▪ 유통을 담당하는 풀필먼트 센터에는 로봇 이동, 창고 상황, 주문 데이터를 실시간 분석

- 공급망 전반으로 선적 추적 데이터와 머신러닝을 통해 병목 구간 탐지
- 장비 센서 데이터 기반으로 예방 정비를 하여 가동 중단을 줄임

대표 서비스 사례 - amazon

- 빅데이터 기반의 가격 전략
  - 계절별 트렌드, 경쟁사 가격, 수용 예측을 결합한 예측 분석과 처방 분석으로 실시간 가격 조정
  - 판매자에게 순이익, 재고 추세 등 세분화된 지표화, 자동화된 데이터 도구 제공하여 판매 효율을 높임

데이터 준비부터 유통까지 실시간으로 데이터를 예측, 처방 분석하여 수익을 창출함

대표 서비스 사례 - Walmart

- 매일 2억 4,500만 명 이상의 고객 데이터를 수집
  - 매시간 100만 명 이상의 고객으로부터 행동 데이터, 거래 기록, 소셜 정보 등을 분석해 2.5 페타바이트 (10^{15})의 비정형 데이터 수집

    -> 생성된 데이터는 Hadoop과 Spark 기반 분산 처리 인프라로 저장 분석

    -> 매출 회전율과 고객 경험을 개선

대표 서비스 사례 – Walmart

- 월마트의 빅데이터 활용
  - 소셜 미디어 분석 : 실시간으로 트렌드 상품을 파악하여 매장에 입점 시킴
  - 예측 분석 : 데이터 기반으로 운영 전략 최적화와 상품을 동적 조정
  - 장바구니 분석 : 특정 상품 간 연관성을 통해 진열과 추천 시스템 반영
  - 고객별 행동 추적 : 오프라인/ 온라인 구매 이력, 상품 선호도를 분석해 맞춤형 쇼핑 제공
  - 모바일 앱 : 실시간 데이터 분석으로 쇼핑 목록, 쿠폰, 매장 위치 등을 안내함

대표 서비스 사례 - 게임산업

- 데이터 과학을 통해 게임 플레이 중 생성되는 방대한 양의 데이터를 분석해 개발

EA(Electronic Arts)

- 플레이어의 모든 버튼 조작, 이동, 결정에 대한 데이터를 수집하여 게임 참여도, 버그, 밸런스 최적화를 함
- 머신러닝을 통해 플레이어 행동 예측과 새로운 게임 기능을 개발

대표 서비스 사례 - 게임산업

▪ 게임 내 플레이어의 지출 패턴을 분석하고 효과적인 수익 창출 전략 수립

King Digital Entertaimnent

- 플레이어의 지출 행동을 예측하고 혜택을 최적화
  (ex. 인게임 구매 가능성이 높은 플레이어에게 맞춤형 혜택 제시)
- 이탈율을 분석하여 참여 유지와 장기적 지출 유도

대표 서비스 사례 - 게임산업

- 플레이어 행동 및 거래 패턴의 이상 징후를 분석하여 사기를 감지하고 예방

Valve

- VAC(Valve Anti-Cheat) 시스템은 게임 플레이 데이터를 분석하여 부정행위 플레이어를 자동 차단
- VAC는 합법 및 불법적인 플레이의 방대한 데이터 세트를 기반으로 학습된 머신러닝 모델

출처

https://www.oracle.com/kr/big-data/what-is-big-data/

https://www.ibm.com/kr-ko/think/topics/big-data-analytics

https://www.sap.com/korea/products/technology-platform/what-is-big-data.html

https://cloud.google.com/learn/what-is-hadoop?hl=ko

https://cloud.google.com/learn/what-is-apache-spark?hl=ko

https://terms.tta.or.kr/dictionary/dictionaryView.do?subject=%EB%B3%91%EB%A0%AC+%EC%B2%98%EB%A6%AC

https://terms.tta.or.kr/dictionary/dictionaryView.do?subject=%EB%A7%B5%EB%A6%AC%EB%93%80%EC%8A%A4

출처

https://terms.tta.or.kr/dictionary/dictionaryView.do?word_seq=041665-3

https://aws.amazon.com/ko/what-is/apache-spark/

https://www.techtarget.com/searchdatamanagement/feature/Top-trends-in-big-data-for-2021-and-beyond#:~:text=We%27re%20also%20seeing%20the%20convergence,term%20strategic%20planning

https://www.acceldata.io/blog/top-8-big-data-trends-shaping-2025#:~:text=1,AI%29%20Integration

https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/

https://velog.io/@leesh0567/Amazon-EMR%EC%9D%B4%EB%9E%80

출처

https://www.dataideology.com/data/by-2025-idc-predicts-that-the-total-amount-of-digital-data-created-worldwide-will-rise-to-163-zettabytes-ballooned-by-the-growing-number-of-devices-and-sensors/#:~:text=digital%20data%20created%20worldwide%20will,Data%20Ideology

https://www.pickl.ai/blog/use-of-data-analytics-by-uber-to-enhance-supplyefficiency-and-service-quality/?utm_source=chatgpt.com

https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39

https://tacticallogistic.com/logistics/amazon-big-data/?utm_source=chatgpt.com

https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109?utm_source=chatgpt.com

https://ioaglobal.org/blog/role-of-data-science-in-gaming-industry/