‘24.09.03

OT

빅데이터 이론 + 툴 + 사례

중간은 시험 + 기말은 프로젝트 발표

창훈이 정보

소개해준 프로그램 쓰려다 실패해서 엑셀로 빅데이터 정리해서 냈음, 이것도 인정해줌

빅데이터라는게 대량의 데이터가 있고, 프로그램 이용해서 내가 원하는 정보를 뽑아내는거,

예를 들어 수만개의 학생시험이 있는데, 6/7/8월의 달 별로 학생들의 전체 성적 변화 추이 그래프 이런거 보여주고

변화 추이가 하향선이네? 그럼 이 이유를 다른 데이터로 근거를 적어주는거임(축제가 많았다거나?)

원래는 프로그램 써서 특정 키워드 잡아서 그리는건데 쓸줄 몰라서 엑셀로 그림, 사빈이네만 성공함

빅데이터 정리를 도와주는 툴일뿐이지, 실제로 해야 하는건 어차피 내가 정리해야 함(정리해서 분석하는게 목적)

중간 안 보고 교수님이 올려주신 공모전 함(논문 쓰기), GPT로 조짐

→ 기말은 기획력(20, 프로젝트 내용), 구현정도(50, 프로젝트 결과물 실현정도), 팀워크(30, 발표력, 팀원 역할 내용)

'24.09.10

배경 → 4차 산업*혁명 = 빅데이터(핵심은 *데이터=리소스(21세기의 석유)) → 지도앱, 네비앱 : 데이터가 필요
*혁명 : 기존에 있지 않은 새로운 것 → 사회 변화
디지털 대전환시대 → *데이터를 뽑아내기 위해(∴지속적으로 활용이 가능), 출석 디지털화 → 학과별 출석률 등
예) 카카오택시 → 데이터가 있으면, 손님-기사 모두 효율적

빅데이터란? → ①수집②저장③관리④분석+대량의 정형/비정형 데이터 집합 → **가치 추출 → 결과 분석하는 *기술 기본 →
수집 어떻게? 저장 어디에? 관리(데이터를 잘 사용할 수 있도록 정리=데이터품질), 분석(처리해보니 현상에 대한 결론이 나오는 것=결론은 가치가 있어야 함)
+ 정형(변하지 않는 데이터), 비정형(감정, 분위기, 날씨, 풍향 등 변하는 데이터)
→ 가치가 없으면? ①한 번 더 분석해보거나, ②과감히 버리기
효용 → 비즈니스 관점 사례 → *①Innovation(혁신), ②Competition(경쟁력↑), ③Productivity(생산성↑)
*5가지 V → *클라우드 컴퓨팅 등장으로 빅데이터 원활, 다양성(정형/비정형/반정형 등 다양한 종류 분석 가능), 속도(엔비디아), 정확성 중시(통계처럼 오차율 띡이 아닌), *Value 창출 : 단순하고 다양한 데이터를 빠르게 분석해 가치를 찾는 것.
※ 추론은 통계적 관점(과거데이터→미래예상), 예측은 인공지능 관점(과거+형이상학+분석)→빅데이터와 통계는 다름
통계는 모집단을 통해 결과를 내고, 결과에 대한 오차율이 존재 ↔ 빅데이터는 그런거 없음(모든 데이터가 대상)
※ 인생은 돈을 따라가면 안된다. 돈 가지고 일희일비하면 안됨. 있다가도 없는게 돈임. 내 가치를 찾아서 몇 년이 걸려도 스탭바이스탭으로 진행해나가는 것이 현명한 것. 내 주변)대기업은 시간을 갈아 돈을 버는거임(가정이 없음)

클라우드 컴퓨팅 → 빅데이터 각광, 앞으로 더 활성화될 것.
빅데이터 어떻게 구성? 아키텍처 → ①저장 관리, ②데이터베이스, ③처리(*병렬), ④데이터 통합, ⑤분석, ⑥*시각화
→ 시각화 상당히 중요. 분석 아무리 잘 해도, 분석의 가치가 얼마나 있는지 전달되지 않으면 의미 X(*직관적이여야)

①데이터 수집 → 반정형(웹 정보=정형처럼 보여도 내용은 비정형인게 많음, 센서 데이터) → 종류에 따라 수집 기술도 다름, ②데이터 저장/관리, ③데이터 분석, ④데이터 시각화
※API는 약속 같은 것.(A와 B가 서로 데이터 통신 시, 필요한 약속? 같은 것)

'24.09.24

복습

***빅데이터 5가지 특성 → 알고 의미 이해

빅데이터 역사는 70년 이상

***아키텍처 6가지

p.12 *현상 추적 → 예측 검정*(통계적 분석)

정보 구조화 → 예) 링크드인

4개 영역 별로 각각 전문가가 있는 것

→ **1. 데이터 시각화가 젤 수요가 많음, ∵ 분석을 잘 해도 전달이 안되면 의미 X (기업에서 중시하는 기여도 순)

2. 데이터 저장/관리 3. 분석(요샌 툴이 있어서 금방 함) 4. 수집

ppt의 참고는 그냥 넘어가심. 빅데이터 안에 통계가 있는 것

*빅데이터는 클라우드와 뗄 수 없는 관계.

정적인 데이터는 죽은 데이터라고도 함, 동적인 데이터가 중요.(SNS 데이터도 동적 데이터)

인공지능과 빅데이터의 차이? 요즘엔 둘을 구분 짓지 않는 추세, 굳이 나누자면

분석력,예측력(학습->다른 문제 예측, 데이터 부족 시 확장) ↔ 신뢰성, 현실성(실제 데이터, 있는 데이터로만 학습)

즉, 빅데이터는 데이터 관점, 인공지능은 결론적인 부분에 대한 접근

수집은 1. API로 가져오는 방식(실시간 데이터), 2. DB에 연결해서 가져오는 방식

빅데이터의 핵심은 병렬 처리 저장/관리/처리(60~80%) / 분석+시각화(20~40%) → (소요 시간 기준)

→ 유의미한 결론이 나올 때 까지 무한 반복 → 대부분 각 단계들이 하나하나가 업종으로 구분돼있음(대기업 제외)

시뮬레이션, 최적화 → 디지털 트윈 : 현실에 있는 데이터를 통해 가상공간에도 똑같이 구현한 것.

예) 건설에서 건물을 짓기 전 최적의 조건 탐색(최적화 값), 삼전반도체 수율 시뮬레이션 등. 비용 절감 가능

AI를 사용하면 부정탐지 가능, 예)아파트 동대표들 난방비 0원 담합

코로나, 사스와 같은 데이터들은 피크치를 버려야하는 문제로 분석 불가한 데이터로 분류됨

데이터 신뢰성 중요* → 의사 수가와 의평원 사례

p.17 밑에 표만 알면 좋을 듯하다 하심. 어떤 관련이 있는지

(계층) 수집(kafka)/저장(hadoop,redis)/처리(spark,mapreduce)/분석 - 매니지먼트 - 클러스터 플랫폼

p.18 배치 수집(ETL 기능) → ETL을 통해(마이그레이션) 레거시 시스템의 데이터를 하둡에 보냄

실시간은 FLLME -> kafka -> hdps → 클라우드는 마이그레이션이 필요없음. 그래서 안전/버전 관리 쉬움)

p.19 데이터 적재 과정 알아두면 좋음(하둡이 이렇게 작동함)

p.20 통계는 데이터하우스, 빅데이터는 데이터레이크(구성 시 전체에 영역을 둠)

정형, 비정형에 따라 받아들이는 과정이 다름(속성에 맞게 처리해서) 데이터레이크에 저장

p.21 *데이터 거버넌스* → 컴퓨터가 그냥 처리하는게 아니라 이건 가이드라인으로 처리하는 것임

p.23 앞전까지는 분석하는 과정의 내부, 이건 그걸 기반한 툴임(visual P.L), 삼성SDS-브라이틱스AI(잘돼있음)

-> 그렇기 때문에 분석은 어렵지 않음(알고리즘도 다 추천해줌),

정제, 정리 → 1. *모델링이 중요, 2. 수집, EDA 처리가 제일 중요

컨테이너 -> 하나의 독립된 공간(클라우드 컨셉) -> 자기만의 워크스페이스임.

p.25 *spark, python, 텐서플로우, 파이토치 다 수용 가능한게 브라이틱스임*

'24.10.08

빅데이터 이론과 실제 → 10월 22일 화요일 10시

1장 PPT, 주관식(아는 만큼 쓰시오) → *수업 때 한 이야기 위주로 쓰면 점수 잘 줄 것

*p.26 기획재정부 차세대 디브레인 데이터 프레임워크→ 굉장히 중요(이 데이터 아키텍처 구성할 수 있으면 전문가)

→ 기획재정부는 국가 예산을 관리하는 부처(국세 수입(국세청이 산하로)과 예산 집행 역할)

→ 국가 예산을 관리하는 시스템, 기존엔 디브레인이라는 시스템 사용

→ 기존엔 국고 관리(쌀, 돈 등)를 엑셀로 함(800조 넘는 예산을) + 예측까지 → 업무 과중(사무관 1명이 혼자)

→ AI 시대에 맞춰 AI/빅데이터 분석을 기반으로 해야 한다는 것이 요구사항 이었음 (3,000억 어마어마한 예산)

크게 3가지 업무 → ①재정 의사결정, ②예산 편성 지원, ③재정 정보 공개

*빅데이터는 목적이 중요함 → 어떤 일을 할 것인지부터 생각해보는 것이 중요

동전의 앞/뒷면 나올 확률은 1/2인데, 빅데이터 사용한다고 더 나은 예측이나 결과? 아니다.

→ 단순히 데이터를 많이 수집하는 것이 아닌, 그 데이터를 통해 해결하고자 하는 문제나 목표를 분명히 하고 접근

①의사결정 → AI와 빅데이터로 자문하는 방식, ※ R&R : 조직 또는 팀 내에서 개인이 맡은 역할과 그에 따른 책임

GDP(국민 총 수익), 금리에 따라 국가 정책이 달라짐

→ 분기마다 발표하는데, 발표를 하려면 해당 분기가 지나야 함. 발표도 2달 뒤에나 됨 → 적절한 정책 결정에 영향

AI로 예측해보자. → 반발이 있긴 했음(여태 우리가 모아놓은 지표들이 간단한거로 보이느냐?)

→ 기존 GDP 자료를 사용하지 않고, GDP 수치를 결정하는 요소 데이터들로만 학습을 해봤음

→ 반도체 지수, 대중교통량 등

→ 한달 반 먼저 계산했고, 심지어 다음 분기 GDP까지 예측이 가능했음

→ 경제/산업/사업 등 경제에 민감한 관련자들의 의사결정이 빨라지게 됨(리스크 줄일 수 있음)

→ 여기서 제일 중요한건 데이터임 → 얼마나 목적에 맞는 데이터로 학습했느냐가 결과를 판가름함

②예산 편성 지원 → 편성이란 목적에 맞게 짜는 것 → 이를 빅데이터의 도움을 받자

기존에는 기재부가 예산에 대한 증거를 남기지 않았음(지웠음) → 현재는 DB에 넣어 분석, 예측의 중요한 자료가 됨

→ *저장/관리의 중요성

③재정 정보 공개 → 오늘날 *투명성(공개)이 중요함(자기 PR시대, 나만 숨기고 갖는게 아니라, 내 능력을 알림)

운영계는 비즈니스 운영을 지원하는 계층(실시간 처리), 정보계는 운영을 위해 기반이 되는 서비스(분석, 예측) 계층

운영계에 있는 데이터들 중 ※ POL : 기술적 검증

1) 정형데이터 → ETL을 통해 받음(ODC로 이동) → DW(웨어하우스), DM(마켓)을 통해 → 통계 기반 분석

2) 반정형, 비정형데이터 → *DL(레이크)를 통해 분석, 예측 → DL, BI(정보 수집), 인공지능 등으로 DB구축

(NOD, OD, SD는 기타 참고 비정형 데이터라고 보면 됨)

점선은 필요시에만 연결되는 것 → 고정적인 것들(일정 주기마다 사용하는 것)은 통계로

→ GDP 예측과 같은 목적에 의한 예측이 필요한 것들은 필요할 때 연결

→용어들 찾아볼 것

KB(Knowledge Base, 지식 기반) → 정보를 체계적으로 저장하고 관리하여 의사결정에 활용하는 시스템

OLTP(Online Transaction Processing) → 실시간으로 DB에 트랜잭션을 처리하는 시스템, 주로 운영계에서 사용

OLAP(Online Analytical Processing) → 빅데이터를 다차원적 분석하여 인사이트를 제공하는 시스템, 주로 정보계

CBR(Case-Based Reasoning) → 사례 기반 추론: 과거의 사례를 활용하여 문제를 해결하는 인공지능 기법

EIS(Executive Information System) → 경영진이 의사결정을 내리는 데 필요한 정보를 제공하는 시스템

SB(Sandbox) → 안전한 환경에서 데이터를 테스트하고 분석할 수 있는 공간