시계열 데이터분석

Timeseries Data Analytics

LG전자 CDO부문

정의석 연구원

01

데이터와 인공지능

데이터 : 전통적인 데이터 vs 빅데이터

LEARNING CREATORS

❖ 전통적인 데이터

*RDBMS: Relational Data Base Management System

- RDBMS(관계형 데이터베이스)에서 주로 관리되는 작은 규모의 데이터 집합
  - ✓ **제한된 크기**: 일반적으로 빅데이터에 비해 데이터의 크기가 작음
  - ✓ **정형화된 구조**: 데이터는 테이블과 로우, 컬럼으로 잘 정리되어 있음
  - ✓ **제한된 인사이트**: 전통적인 데이터는 기본적인 분석과 인사이트 제공에 주로 사용됨

❖ 빅데이터

- 대량의 데이터를 빠르게 수집, 저장, 분석 및 활용할 수 있는 데이터 집합

✓ **볼륨 (Volume)** : 매우 큰 크기의 데이터를 다룸 (ex. 페타바이트, 제타바이트 등)
✓ **다양성 (Variety)** : 다양한 형태의 데이터(정형, 반정형, 비정형)를 처리함
✓ **속도 (Velocity)** : 데이터가 실시간 또는 거의 실시간으로 저장 및 처리될 수 있음 **3V**

✓ **정확성 (Veracity)** : 데이터의 정확성과 신뢰성이 충분히 보장되어야 함 **4V**

데이터 : 전통적인 데이터 vs 빅데이터

❖ 전통적인 데이터와 빅데이터 무엇이 다른가?

특성	전통적인 데이터	빅데이터
크기	기가바이트, 테라바이트	페타바이트, 제타바이트, 엑사바이트
구성	정형 데이터	정형, 비정형, 반정형 데이터
아키텍처	중앙 집중식 아키텍처	분산 아키텍처, 수평 확장 가능, 장애 허용성 높음
데이터 출처	엔터프라이즈 데이터	다양한 춛처 (소셜 미디어, 센서 데이터 등)
데이터 분석	점진적 분석	실시간 분석, 동적, 전체 비즈니스 이해 제공

데이터 : 전통적인 데이터 vs 빅데이터

❖ 전통적인 데이터와 빅데이터 무엇이 다른가?

특성	전통적인 데이터	빅데이터
크기	기가바이트, 테라바이트	페타바이트, 제타바이트, 엑사바이트
구성	정형 데이터	정형, 비정형, 반정형 데이터
아키텍처	중앙 집중식 아키텍처	분산 아키텍처, 수평 확장 가능, 장애 허용성 높음
데이터 출처	엔터프라이즈 데이터	다양한 춛처 (소셜 미디어, 센서 데이터 등)
데이터 분석	점진적 분석	실시간 분석, 동적, 전체 비즈니스 이해 제공

데이터의 유형: 정형 / 반정형 / 비정형

❖ 정형 데이터 (Structured data)

- 정형 데이터는 데이터베이스의 정해진 규칙에 따라 저장된 데이터임.
  - ✓ **예시** : 관계형 데이터베이스의 테이블, 엑셀 스프레드시트, CSV 파일 등

❖ 비정형 데이터 (Unstructured data)

- 비정형 데이터는 정해진 규칙이 없어 의미를 쉽게 파악하기 어려움.
  - ✓ **예시** : 텍스트 문서, 이메일, 소셜 미디어 포스트, 오디오 파일, 비디오 파일 등

❖ 반정형 데이터 (Semi-structured data)

- 반정형 데이터는 완전한 정형 데이터는 아니지만 어느 정도의 구조를 가지고 있음.
  - ✓ **예시** : JSON, XML, HTML, YAML 등

데이터의 유형: 정형 / 반정형 / 비정형

❖ 정형 데이터 (Structured data)

*시계열 데이터는 여기에 해당함

- 정형 데이터는 데이터베이스의 정해진 규칙에 따라 저장된 데이터임.
  이 데이터는 수치나 문자 등으로 의미를 쉽게 파악할 수 있음.
  - ✓ **예시** : 관계형 데이터베이스의 테이블, 엑셀 스프레드시트, CSV 파일 등

❖ 비정형 데이터 (Unstructured data)

- 비정형 데이터는 정해진 규칙이 없어 의미를 쉽게 파악하기 어려움.
  - ✓ **예시** : 텍스트 문서, 이메일, 소셜 미디어 포스트, 오디오 파일, 비디오 파일 등

❖ 반정형 데이터 (Semi-structured data)

- 반정형 데이터는 완전한 정형 데이터는 아니지만 어느 정도의 구조를 가지고 있음.
  - ✓ **예시** : JSON, XML, HTML, YAML 등

인공지능과 데이터 분석

❖ 데이터 분석이란?

- 유용한 정보를 추출하고 의사 결정을 지원하기 위해 데이터를 정리 및 모델링하는 과정임

❖데이터 분석의 중요성

- 데이터 분석은 유용한 정보를 발굴하고 의사 결정을 지원하는 과정임.
- 다양한 종류와 양의 데이터를 효과적으로 활용하여 비즈니스 프로세스를 개선하고 작업자의 의사 결정을 도우는 데에 사용됨.

❖ 빅데이터 분석(Big Data Analytics) 영역

- **서술적 분석(Descriptive Analytics)**:
  - ✓ 과거 데이터를 수집하고 정리하여 쉽게 이해할 수 있는 방식으로 설명함.
- **예측 분석(Predictive Analytics)**:
  - ✓ 과거 데이터와 패턴을 분석하여 미래의 상황을 예측함.
- **처방적 분석(Prescriptive Analytics)**:
  - ✓ 서술적 분석과 예측 분석을 통해 얻은 정보를 바탕으로 최선의 행동 방안을 제시함.

5

Introduction

현업에서의 데이터 분석 프로세스

LEARNING CREATORS

❖ **문제 정의**

- 업무의 큰 방향성과 전반적인 프레임은 설정하는 단계임.
- 유관자와 업무의 목적, 이유, 비즈니스에 미치는 영향 등은 협의함.

❖ **데이터 수집**

- 로그 설계와 검증 부분은 담당하며, 데이터 수집 및 처리는 주로 데이터 엔지니어가 맡음.

❖ **데이터 처리**

- 데이터 추출, 필터링 등은 SQL로 진행하며, 이상치 제거나 분포 변환은 분석 툴(Python, R 등)은 사용함.

❖ **데이터 분석**

- 지표 정의, 탐색적 데이터 분석, 통계분석, 머신러닝 등 다양한 분석은 진행함.

❖ **리포팅/피드백**

- 분석 결과를 설득력 있게 정리하고 전달하는 단계임. 내용의 초점은 상대방에 맞추고, 명확한 메시지로 전달함.

데이터 분석 툴(Tool)

❖스프레드시트 (Excel, Google Sheets)

- **장점**: 사용하기 쉬움, 기본적인 데이터 분석과 시각화 가능
- **단점**: 대용량 데이터 처리 붇가, 고급 분석 기능 제한

❖통계 및 프로그래밍 언어 (R, Python)

- **장점**: 고급 분석 가능, 라이브러리와 패키지가 풍부
- **단점**: 코드 작성 필요, 데이터 처리와 분석에 시간이 오래 걸릴 수 있음

❖BI 툴 (Tableau, Power BI)

- **장점**: 드래그 앤 드돕으도 쉬운 시각화, 대용댱 데이터 처리
- **단점**: 라이선스 비용이 받생핟 수 있음, 고급 분석 기능 제한

❖클라우드 기반 툴 (AWS, Google Cloud)

- **장점**: 확장성이 좋음, 다양한 데이터 서비스 제공
- **단점**: 비용이 받생핟 수 있음, 보안 이슈륻 고려해야 함

데이터 분석 툴(Tool)

LEARNING CREATORS

❖스프레드시트 (Excel, Google Sheets)

- **장점**: 사용하기 쉬움, 기본적인 데이터 분석과 시각화 가능
- **단점**: 대용량 데이터 처리 붇가, 고급 분석 기능 제한

❖통계 및 프로그래밍 언어 (R, Python)

- **장점**: 고급 분석 가능, 라이브러리와 패키지가 풍부
- **단점**: 코드 작성 필요, 데이터 처리와 분석에 시간이 오래 걸릴 수 있음

❖BI 툴 (Tableau, Power BI)

- **장점**: 드래그 앤 드돕으도 쉬운 시각화, 대용댱 데이터 처리
- **단점**: 라이선스 비용이 받생핟 수 있음, 고급 분석 기능 제한

❖클라우드 기반 툴 (AWS, Google Cloud)

- **장점**: 확장성이 좋음, 다양한 데이터 서비스 제공
- **단점**: 비용이 받생핟 수 있음, 보안 이슈륻 고려해야 함

7

Introduction

강의 사용 분석 툴 : Google Colab

LEARNING CREATORS

❖주요 특징

- 클라우드 기반으로 돌도, Python 코드를 웹 브라우저에서 쉽게 작성하고 실행할 수 있음.

❖사용법

- 주피터 노트북을 기반으로 Python 코드를 셀에 입력하고 실행함.
- 라이브러리 설치도 간단하게 할 수 있으며, 결과는 바로 확인 가능함.

❖주요 특징

- **접근성**: 웹 브라우저만 있으면 어디서든 접근 가능함.
- **GPU**: 무료로 GPU를 사용할 수 있어, 머신러닝이나 데이터 분석 작업에 유용함. (무료는 제한 있음)
- **공유성**: 구글 드라이브와 연동되어 쉽게 공유하고 협업할 수 있음.

❖장단점

- **장점**: 사용하기 쉬우며, 무료로 고성능 컴퓨팅 자원을 사용할 수 있음.
- **단점**: 인터넷 연결이 필요하며, 데이터 보안에 신경을 써야 함.

감사합니다