텍스트 마이닝과 데이터 마이닝

Part 07. 데이터 마이닝

정 정 민

programmers

Chapter 20. 데이터 웨어하우스

1. 데이터 웨어하우스란?
2. 데이터 웨어하우스의 구조

데이터 웨어하우스란?

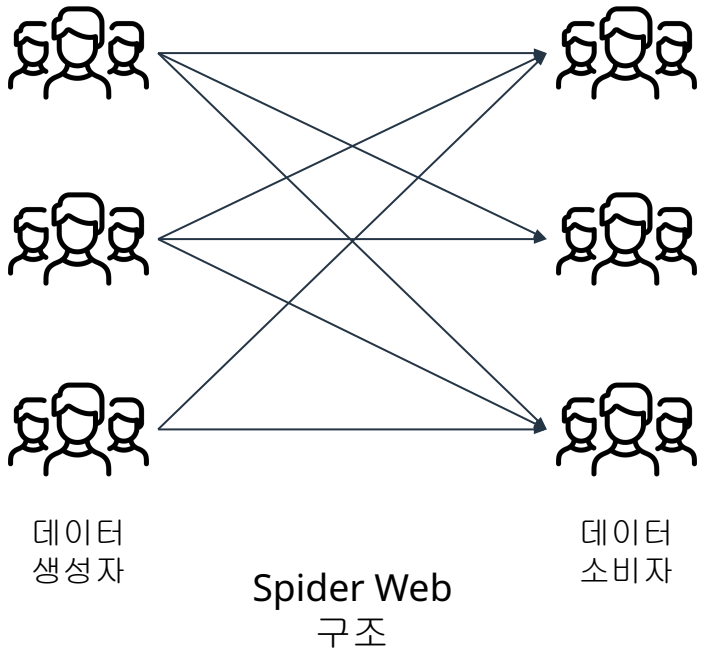
큰 대기업의 데이터 관리자 입장에서 생각해볼까요?

- 다양한 부서가 존재 (마케팅, 영업, CS, 연구 개발 등등)

- 특정 부서는 데이터가 생성되며,
- 임의의 부서는 다른 부서들의 데이터에 접근

- 이때마다 데이터 호출 인터페이스를 각각 따로 만든다면 너무 많은 비용이 필요
  - 만약, 데이터를 생성하는 부서가 10개
  - 데이터를 소비하는 부서가 5개가 있다면
  - 총 50개의 인터페이스가 필요 (곱하기 연산!)

- 데이터의 흐름이 엉켜있는 형태
- 이를 거미집 현상(Spider Web)이라고 함



Spider Web
구조

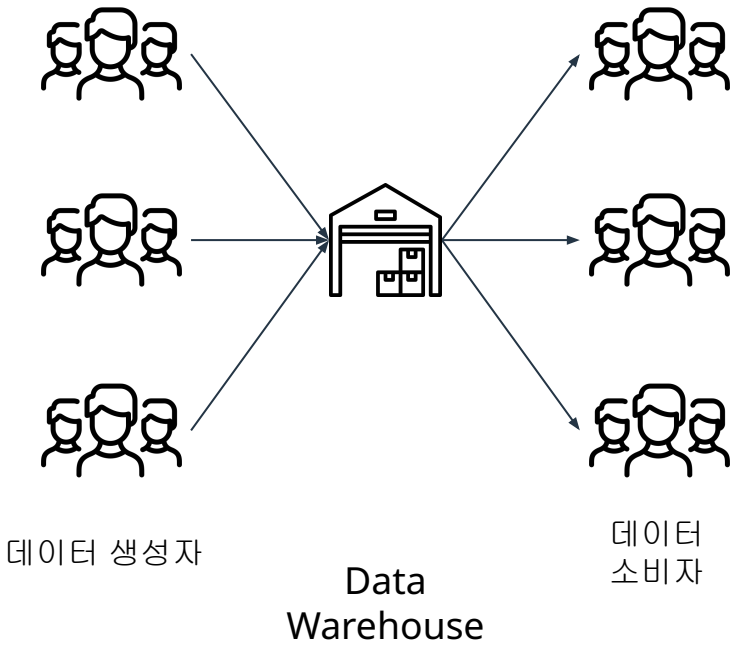
데이터 웨어하우스 (Data Warehouse)

- 마치 소/도매 업자들이 소비자에게 바로 물건을 보내는 것이 아니라,
- 이들과 소비자 사이의 물류 창고를 두고 물건의 흐름을 컨트롤하는 것을 본뜬다면

- 기업 내부에서 움직이는 데이터의 흐름을 효율적으로 컨트롤 할 수 있음

- 앞선 예를 다시 생각해보면
  - 10개의 데이터 생성 부서와 5개의 소비 부서의 경우
  - 15개의 데이터 흐름만 관리하면 됨 (더하기 연산!!)

- 이를 데이터가 모이는 창고(warehouse)라는 의미로
- 데이터 웨어하우스라고 함



데이터 웨어하우스에는 무엇을 저장하나요?

- 조직이 수집한 데이터를 모두 저장
  - 심지어 과거의 정보까지도!!

- 전통적으로 정형 데이터(structured data)를 저장하고 관리했지만
- 비정형 데이터(unstructured data)를 처리하고 저장하는 기능으로 통합
  - 기술의 발전 및 비지니스 요구의 변화로 인해

- 저의 경우로 보면
  - 회사 내부 자체 데이터 수집 과정에서 모이는 원천 데이터 (비정형)
  - 고객사의 사용 로그 (비정형)
  - 노이즈 및 이상 데이터를 처리하고 레이블링을 진행한 학습 데이터 (정형)
  - 과거 데이터 및 최신 데이터 (과거 및 최신 데이터)
  - 등등

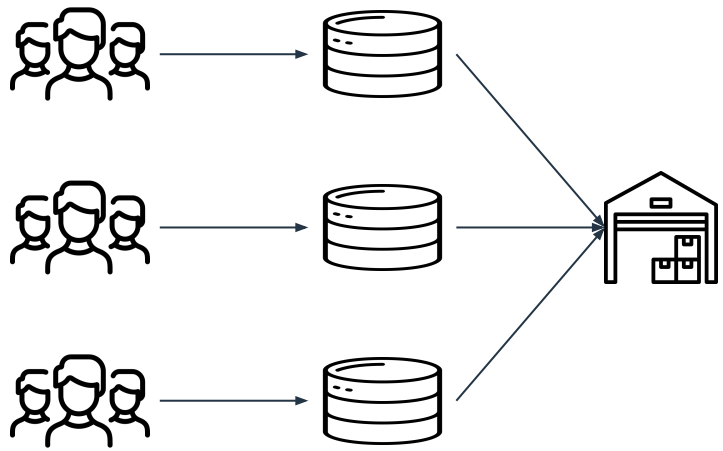
- 이러한 다양한 종류의 데이터를 보관, 관리하고 있습니다.

데이터 웨어하우스(DW) VS 데이터 베이스(DB)

- 두 개념은 정의와 목적부터 차이가 있음
  - **데이터 베이스 (DB)**
    - 실시간 데이터 처리와 트랜잭션 관리에 중점을 둠
    - 일상적인 업무 및 응용 프로그램에 필요한 현재의 데이터를 저장 및 관리
    - 데이터의 신속한 read 와 write의 목적을 갖고 있음
  - **데이터 웨어하우스 (DW)**
    - 대규모 데이터를 통합, 분석, 보고하는데 사용되는 시스템
    - 과거의 데이터도 포함하고 있음

- 또한, 생성과 관리의 차원에서도 차이가 있음
  - **DB**는 데이터 소비처 혹은 생산처에서 만들어지고 관리되는 대상이며
  - **DW**는 DB의 데이터가 주기적으로 모여 만들어지게 됨

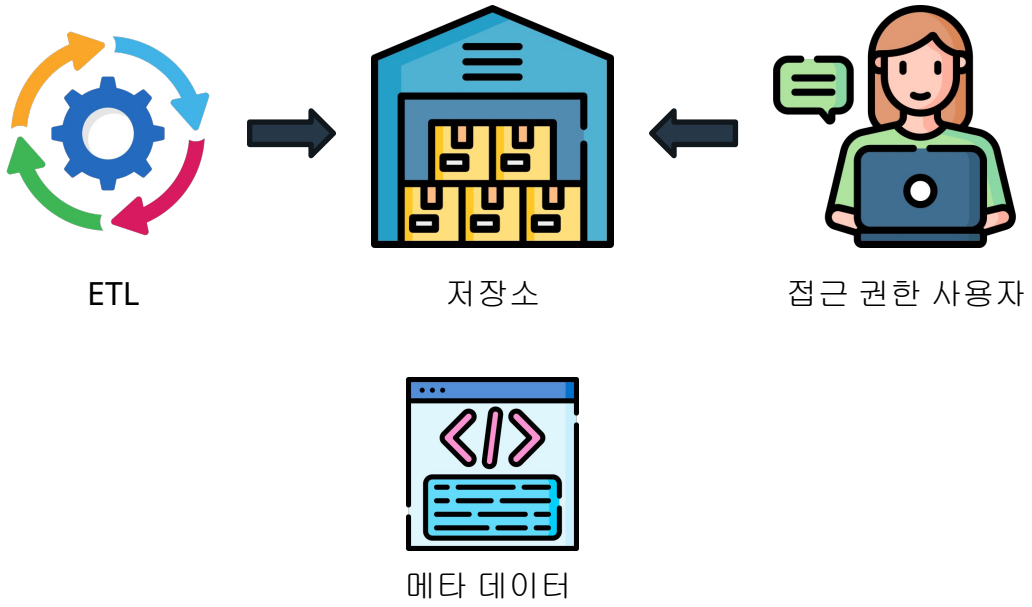
- 그리고, 접근 사용자에 따른 차이도 존재
  - **DB** : 다수의 사용들이 동시에 입력 및 수정 가능
  - **DW** : 조직 내 특정 그룹의 사용자에게만 제한



데이터 웨어하우스의 구조

데이터 웨어하우스의 구성 요소

- **ETL (Extract, Transform, Load)**
  - 원천 데이터 소스에서 데이터를 추출(Extract)하고
  - 저장할 형태에 맞춰 변형(Transform)하고
  - 데이터 웨어하우스 중앙 데이터 저장소로 적제(Load)
- **중앙 데이터 저장소**
  - ETL 처리 된 데이터가 쌓이는 저장소
- **메타 데이터**
  - 데이터가 쌓이면서 만들어지는 추가 정보
  - 원천 데이터의 장소, 중앙 데이터 저장소의 크기 및 구성 방법 등
- **접근**
  - 사용자의 데이터 저장소와의 상호작용 지원



데이터 마트 (Data Mart)

- 특정 부서에서 어떠한 주제로 주기적으로 데이터를 보고자 요청한다고 해볼까요?
  - 예를 들어, 마케팅 부서에서 사용자들의 SNS를 통한 판매 데이터를 보고 싶을 수 있겠죠?

- 이때, 데이터 웨어하우스에서는 요청에 맞는 작은 데이터 집합을 제공해주는데
- 그것을 데이터 마트(Data Mart)라고 함
  - 소비자를 위해 창고에서 물건을 마트에 가져다 두는 것과 비슷한 느낌!

- 해당 부서에서 사용하는 데이터 베이스와는 다르게,
- 과거 데이터를 포함해, 분석과 보고를 목적으로 함

데이터 마트 특징

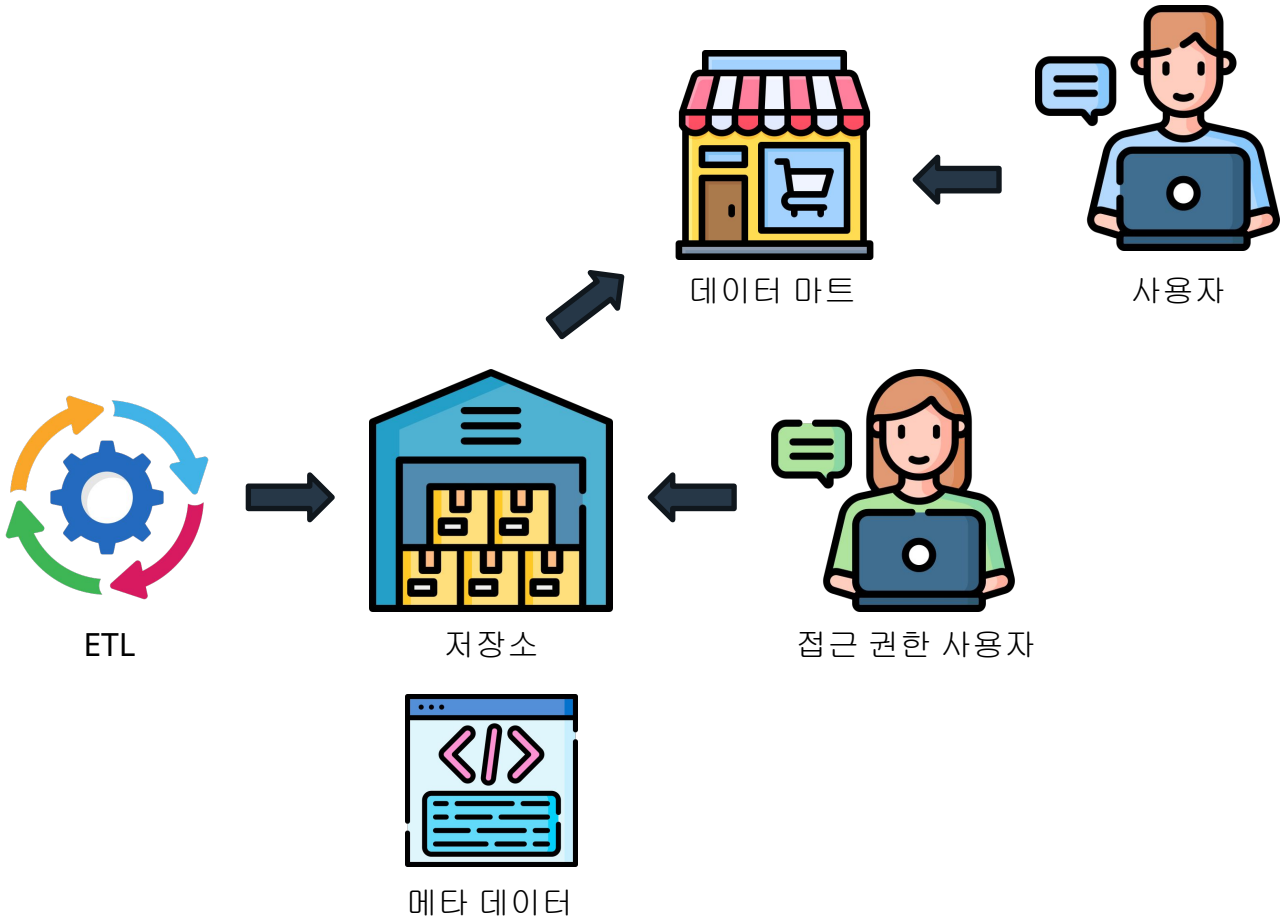
- **부서 중심적 & 주제 중심적**
  - 데이터 마트는 특정 부서나 특정 주제에 맞춰서 설계됨
  - 항상 준비된 것이 아니라 주제에 맞는 부서의 요청이 있을 때 만들어 짐

- **데이터 집중도 ↑**
  - 관련 있는 데이터만 집중적으로 포함하고 있음
  - 사용자 그룹이 필요로 하는 정보를 빠르고 쉽게 확인 가능

- **효율적 운영 및 사용자 친화성**
  - 큰 데이터 웨어하우스 시스템의 일부로 존재
  - 집중도 있는 데이터의 최적화된 집합
  - 필요한 데이터에 대한 간단한 쿼리와 간단한 분석 진행 가능

데이터 마트를 포함한 구조도

- 데이터 마트는 필수 사항은 아니지만
- 조직 내부에서 사업적 분석을 통해 인사이트를 얻고자 많이 사용



E.O.D