데이터 파이프라인 구축

데이터 파이프라인 구축

데이터 파이프라인

다양한 데이터 소스에서 원시 데이터를 수집한 다음 분석을 위해 데이터 저장소로 이전하기까지의 처리 과정

수집 -> 저장 -> 처리 -> 탐색 -> 분석 -> 운영
의 과정 중 수집, 저장, 처리에 필요한 작업의 흐름(Work Flow)을 의미한다.

대표적인 데이터 파이프라인 방법론으로 ETL과 ELT가 있다.

용어

Data Warehouse : 사용자의 의사 결정에 도움을 주기 위하여 기간시스템의 데이터베이스에 축적된 데이터를 **공통의 형식**으로 변환해서 관리하는 데이터베이스

Data Lake : **정형, 반정형, 비정형 상태**인 대량의 데이터를 저장, 처리, 보호하기 위한 저장소

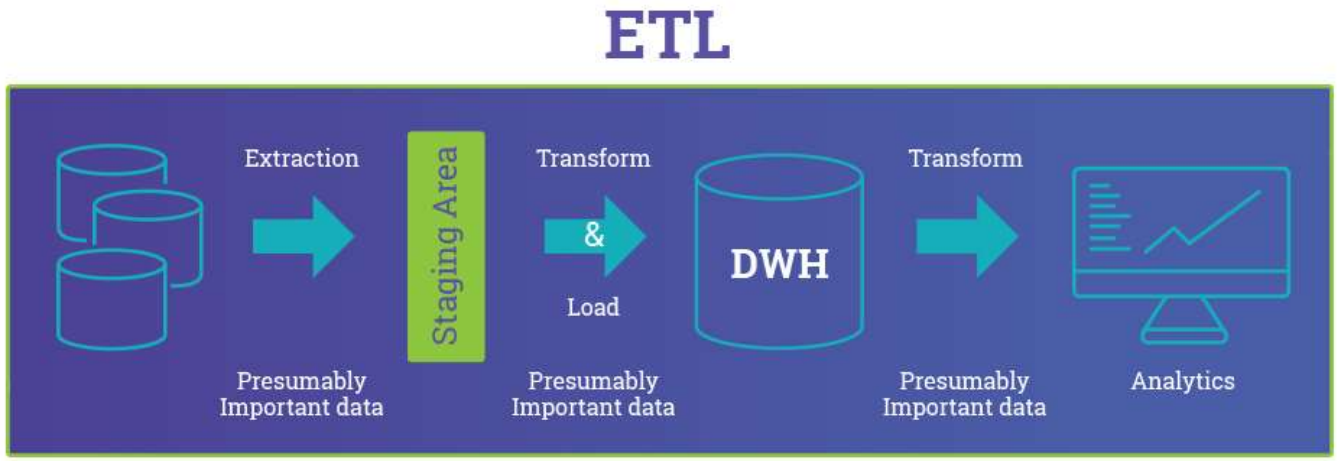
데이터 파이프라인 구축

ETL 파이프라인 : **Extracting(추출) -> Transform(변경) -> Load(적재)**

수집한 데이터를 DWH에서 정의한 형식에 맞춰 변경한 다음 적재하는 파이프라인

DWH의 설계가 매우 중요하다.

수집되는 데이터의 양과 Transform 단계에 필요한 비용이 비례한다.



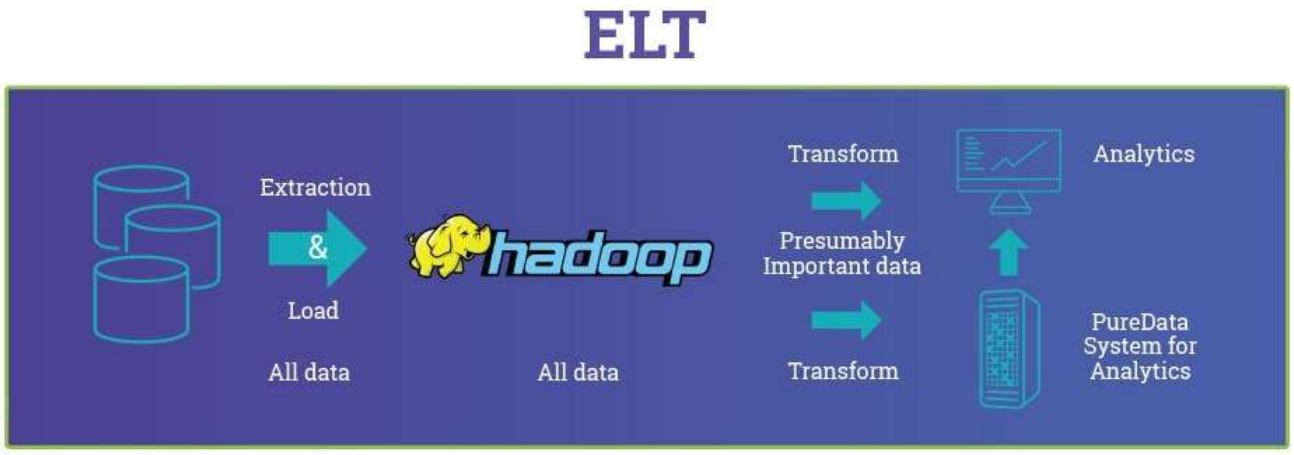
데이터 파이프라인 구축

ELT 파이프라인 : Extracting(추출) -> Load(적재) -> Transform(변경)

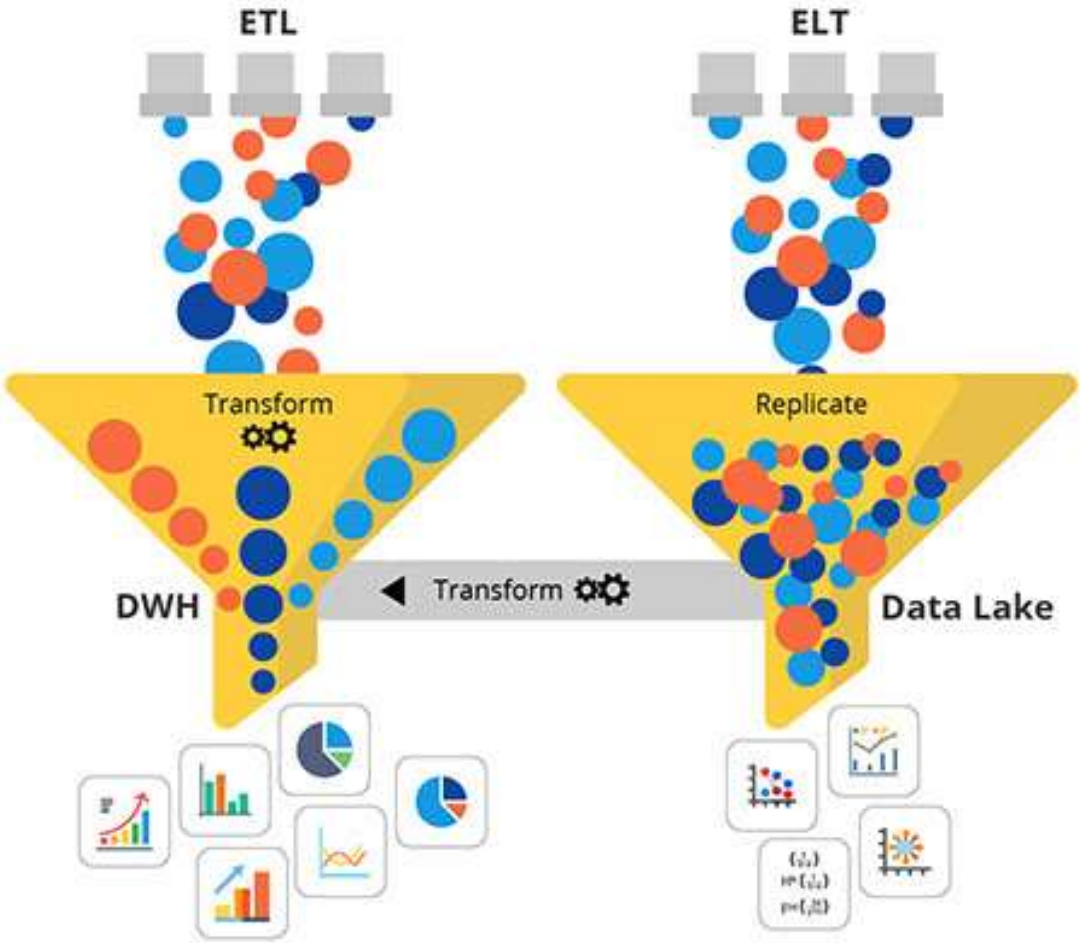
수집한 데이터를 Data Lake에 적재한 다음 필요에 따라 데이터를 Transform 하는 파이프라인

수집되는 데이터의 양과 Transform 비용이 비례하지 않는다. -> 대량의 데이터가 수집되는 환경일 때 유리

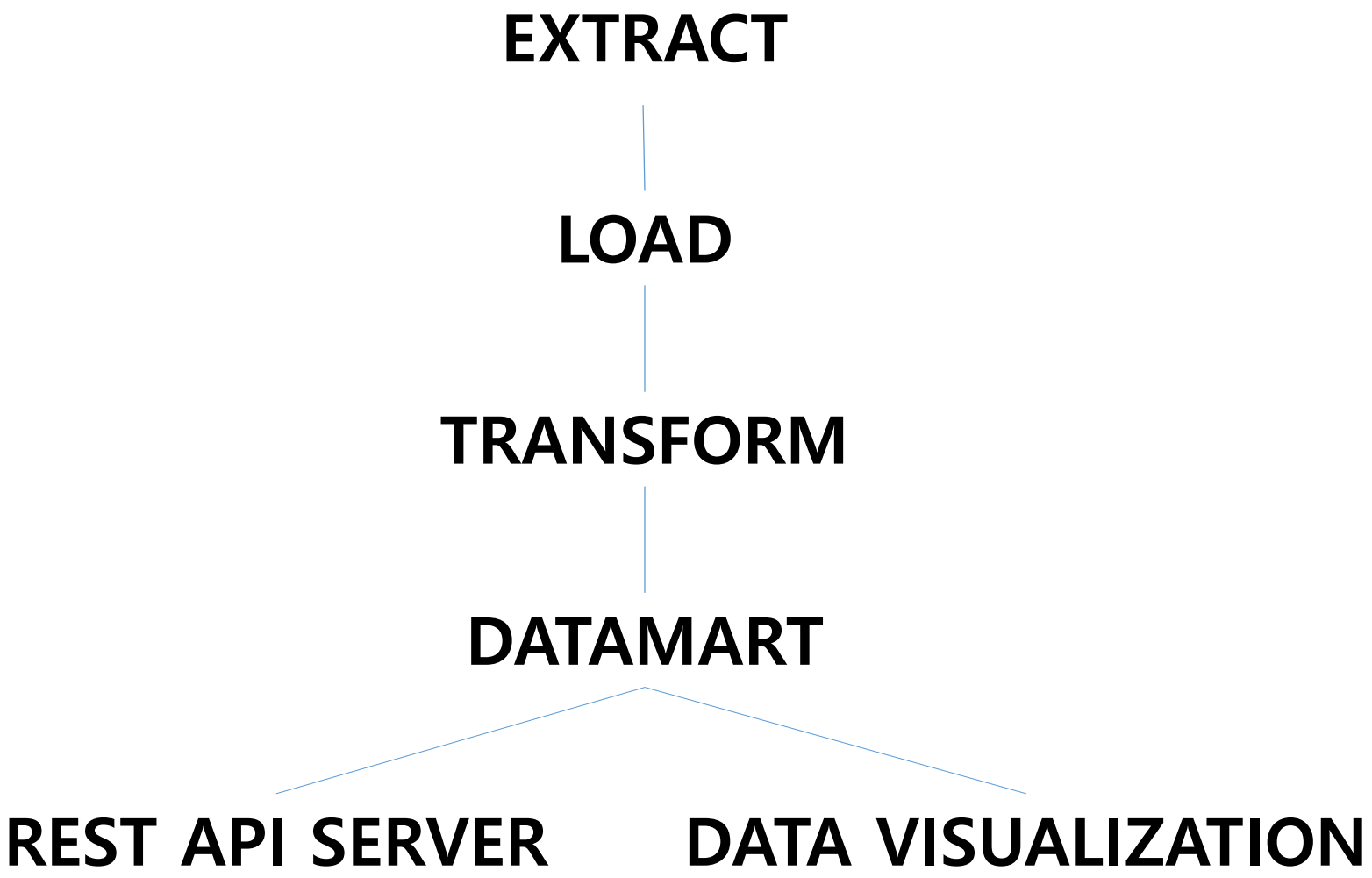
수집되는 모든 데이터를 저장할 수 있는 시스템 리소스가 필요하다 -> 다양한 클라우드 서비스의 등장
ex) google big query, AWS S3, Redshift...



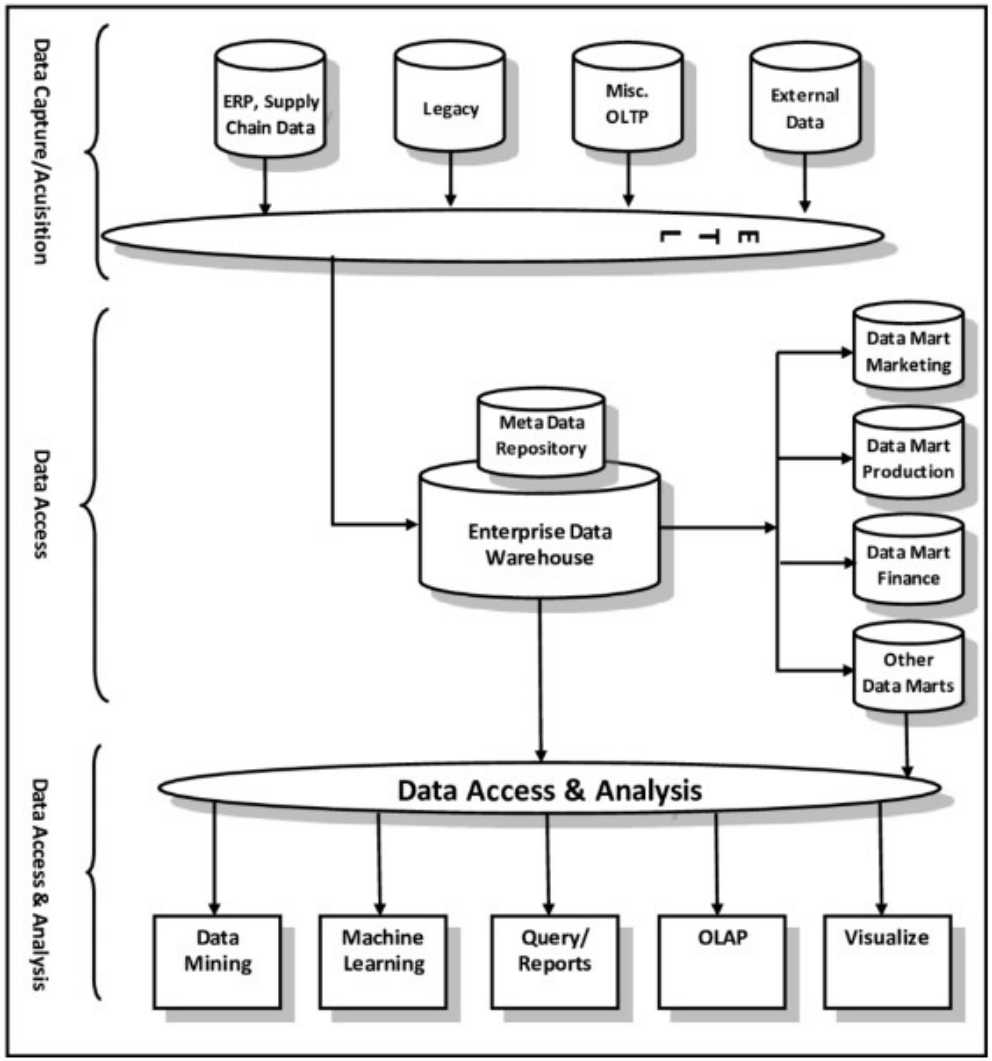
데이터 파이프라인 구축



데이터 파이프라인 구축


EXTRACT
LOAD
TRANSFORM
DATAMART
REST API SERVER
DATA VISUALIZATION

데이터웨어하우스 VS 데이터마트



데이터웨어하우스

- 전사적인 관점에서의 데이터 통합 및 장기 보존
- 대량의 데이터를 저장하는 일에 최적화되도록 설계

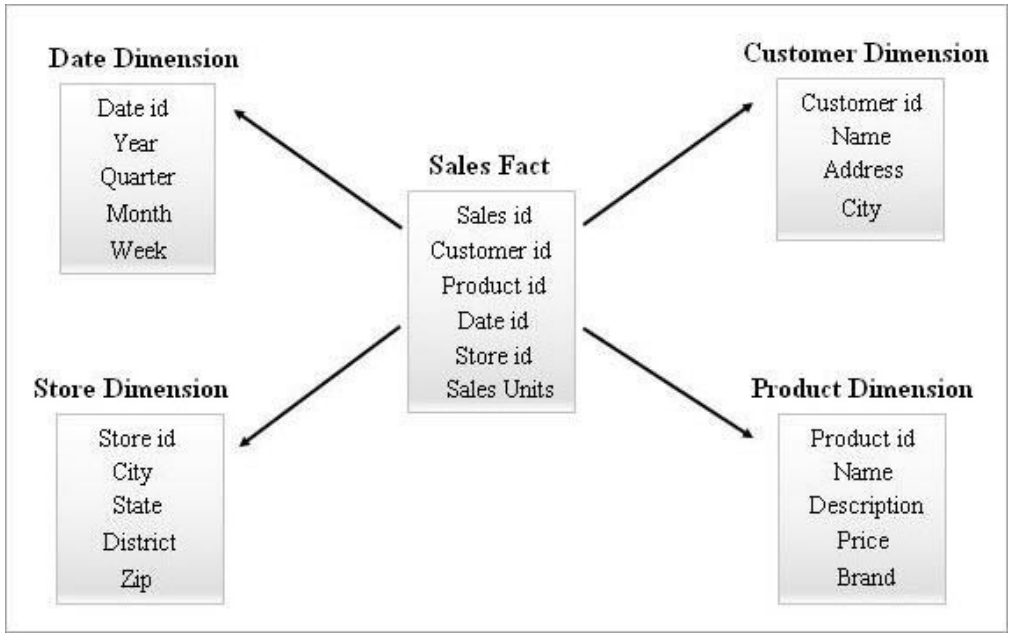
데이터마트

- 전사적인 관점에서의 데이터 통합 및 장기 보존
- 대량의 데이터를 저장하는 일에 최적화되도록 설계

데이터웨어하우스 VS 데이터마트

Fact Table : 비즈니스 측정 값

Dimension Table : Fact Table의 값들을 다양한 관점에서 바라본 값을 저장하는 테이블

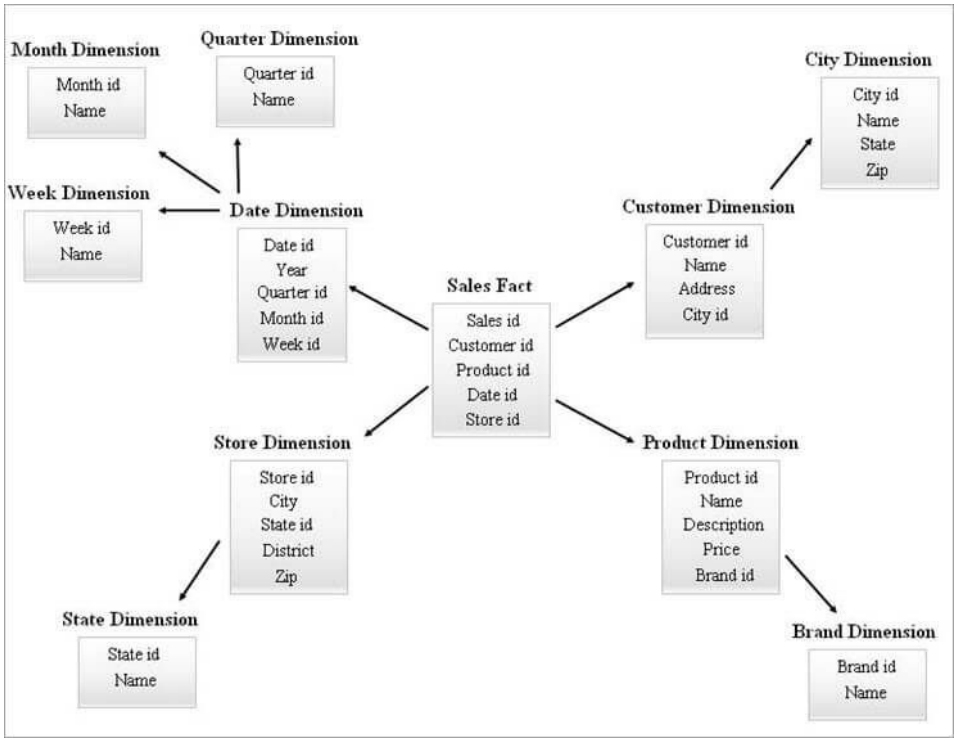


데이터웨어하우스 VS 데이터마트

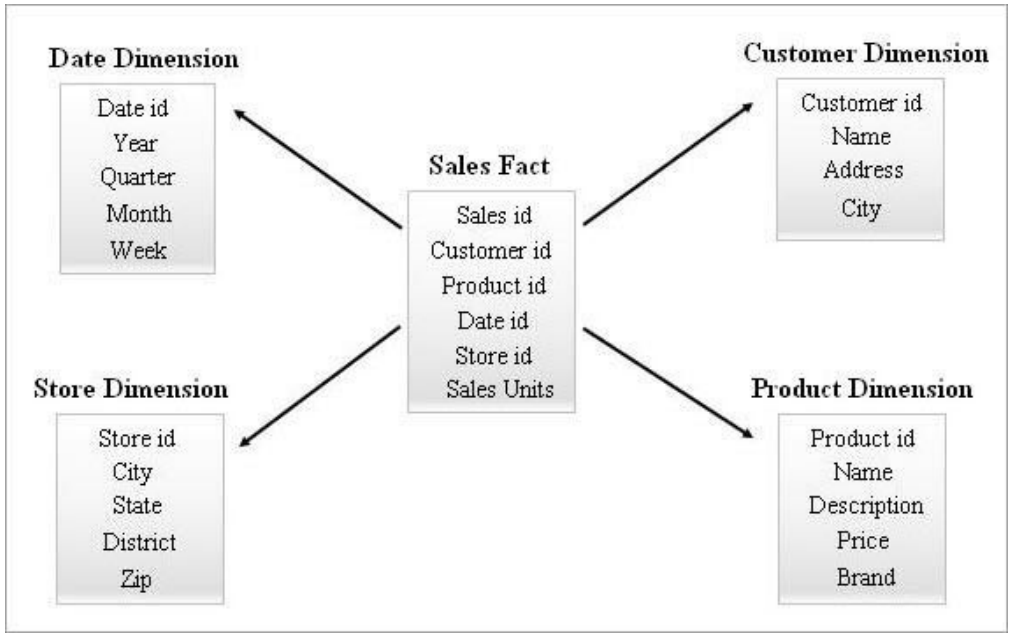
데이터웨어하우스
- SnowFlake Schema에 따라 설계
- 3정규화 까지 수행

데이터마트
- Star Schema에 따라 설계
- 분석에 용이하도록 요약데이터 생성
- 중복데이터가 발생하더라도 사용자 중심으로 설계

SnowFlake Schema



Star Schema



정규화

정규화 : 데이터 중복을 최소화 하고 이상현상이 발생하지 않도록 데이터베이스를 설계하는 과정

이상현상(Anomaly)

삽입이상(Insertion Anomaly)
- 새 데이터를 추가하기 위해 불필요한 데이터도 함께 추가해야 하는 경우

갱신이상(Update Anomaly)
- 중복된 데이터들 중 일부만 변경할 경우 데이터가 불일치가 발생

삭제이상(Deletion Anomaly)
- 데이터를 삭제할 경우 반드시 필요한 데이터가 함께 삭제되는 경우

정규화

정규화 : 데이터 중복을 최소화 하고 이상현상이 발생하지 않도록 데이터베이스를 설계하는 과정

정규화 하는 법

부분함수적 종속을 제거해 완전함수적 종속이 되도록 수정하고

완전함수적 종속일 때 이행적 함수 종속이 발생하지 않도록 한다.

정규화

함수적 종속

A(결정자) -> B (종속자)

B는 A에 함수적 종속이다. A가 B를 결정한다.

A이면 B이고 동시에 A이면 C일 수 없지만, B이면 반드시 A인 것은 아니다.

EX) 아이디가 DE이면 일반회원이고, 아이디가 DE이면 우수회원일 수는 없지만
일반회원이면 반드시 아이디가 DE인 것은 아니다.

함수적 종속 분류

완전함수적 종속 : 종속자가 기본키에만 종속되는 경우.
기본키가 여러 개의 속성으로 이루어져 있다면 모든 속성에 종속되는 경우

부분함수적 종속 : 기본키를 구성하는 여러 속성 중 일부 속성에 종속되는 경우

이행적 함수 종속 : X→Y, Y→Z :
X를 통해 Y를 알 수 있고, Y를 통해 Z를 알 수 있는 경우

정규화

정규화 할 테이블

제 1정규화 : 속성의 원자성을 확보한다. 기본키를 설정한다.

학생번호	이름	학과	학과전화번호	과목번호	강사	학점
11002	김애란	문예창작학과	02-1111-1111	A101, B201	하명도, 이동헌	A,B
97654	헤밍웨이	문예창작학과	02-1111-1111	B201	이동헌	A
24516	빌게이츠	컴퓨터공학과	02-2222-2222	A101	하명도	B
42555	프래디머큐리	실용음악학과	02-3233-3633	D441	이창진	B

정규화

제 1정규화 : 속성의 원자성을 확보한다. 기본키를 설정한다.

학생번호	이름	학과	학과전화번호	과목번호	강사	학점
11002	김애란	문예창작학과	02-1111-1111	A101	하명도	A
11002	김애란	문예창작학과	02-1111-1111	B201	이동헌	B
97654	헤밍웨이	문예창작학과	02-1111-1111	B201	이동헌	A
24516	빌게이츠	컴퓨터공학과	02-2222-2222	A101	하명도	B
42555	프래디머큐리	실용음악학과	02-3233-3633	D441	이창진	B

정규화

제 2정규화 : 기본키가 2개 이상의 속성으로 구성된 경우 부분함수종속성을 제거한다.

학생번호	이름	학과	학과전화번호	과목번호	강사	학점
11002	김애란	문예창작학과	02-1111-1111	A101	하명도	A
11002	김애란	문예창작학과	02-1111-1111	B201	이동헌	B
97654	헤밍웨이	문예창작학과	02-1111-1111	B201	이동헌	A
24516	빌게이츠	컴퓨터공학과	02-2222-2222	A101	하명도	B
42555	프래디머큐리	실용음악학과	02-3233-3633	D441	이창진	B

정규화

제 2정규화 : 기본키가 2개 이상의 속성으로 이루어진 경우 부분함수종속성을 제거한다.

학생번호	과목번호	학점	과목번호	강사	학생번호	이름	학과	학과전화번호
11002	A101	A	A101	하명도	11002	김애란	문예창작학과	02-1111-1111
11002	B201	A	B201	이동헌	97654	헤밍웨이	문예창작학과	02-1111-1111
24516	A101	B	D441	이창진	24516	빌게이츠	컴퓨터공학과	02-2222-2222
42555	D441	B			42555	프래디머큐리	실용음악학과	02-3233-3633

정규화

제 3정규화 : 기본키를 제외한 컬럼 간의 종속성을 제거한다. 즉 이행함수종속성을 제거한다.

학생번호	이름	학과	학과전화번호
11002	김애란	문예창작학과	02-1111-1111
97654	헤밍웨이	문예창작학과	02-1111-1111
24516	빌게이츠	컴퓨터공학과	02-2222-2222
42555	프래디머큐리	실용음악학과	02-3233-3633

정규화

제 3정규화 : 기본키를 제외한 컬럼 간의 종속성을 제거한다. 즉 이행함수종속성을 제거한다.

학생 테이블

학생번호	이름	학과
11002	김애란	문예창작학과
97654	헤밍웨이	문예창작학과
24516	빌게이츠	컴퓨터공학과
42555	프래디머큐리	실용음악학과

학과 테이블

학과	학과전화번호
문예창작학과	02-1111-1111
컴퓨터공학과	02-2222-2222
실용음악학과	02-3233-3633

정규화

정규화 완료

학생 테이블

학생번호	이름	학과
11002	김애란	문예창작학과
97654	헤밍웨이	문예창작학과
24516	빌게이츠	컴퓨터공학과
42555	프래디머큐리	실용음악학과

학과 테이블

학과	학과전화번호
문예창작학과	02-1111-1111
컴퓨터공학과	02-2222-2222
실용음악학과	02-3233-3633

학점 테이블

학생번호	과목번호	학점
11002	A101	A
11002	B201	A
24516	A101	B
42555	D441	B

과목테이블

과목번호	강사
A101	하명도
B201	이동헌
D441	이창진