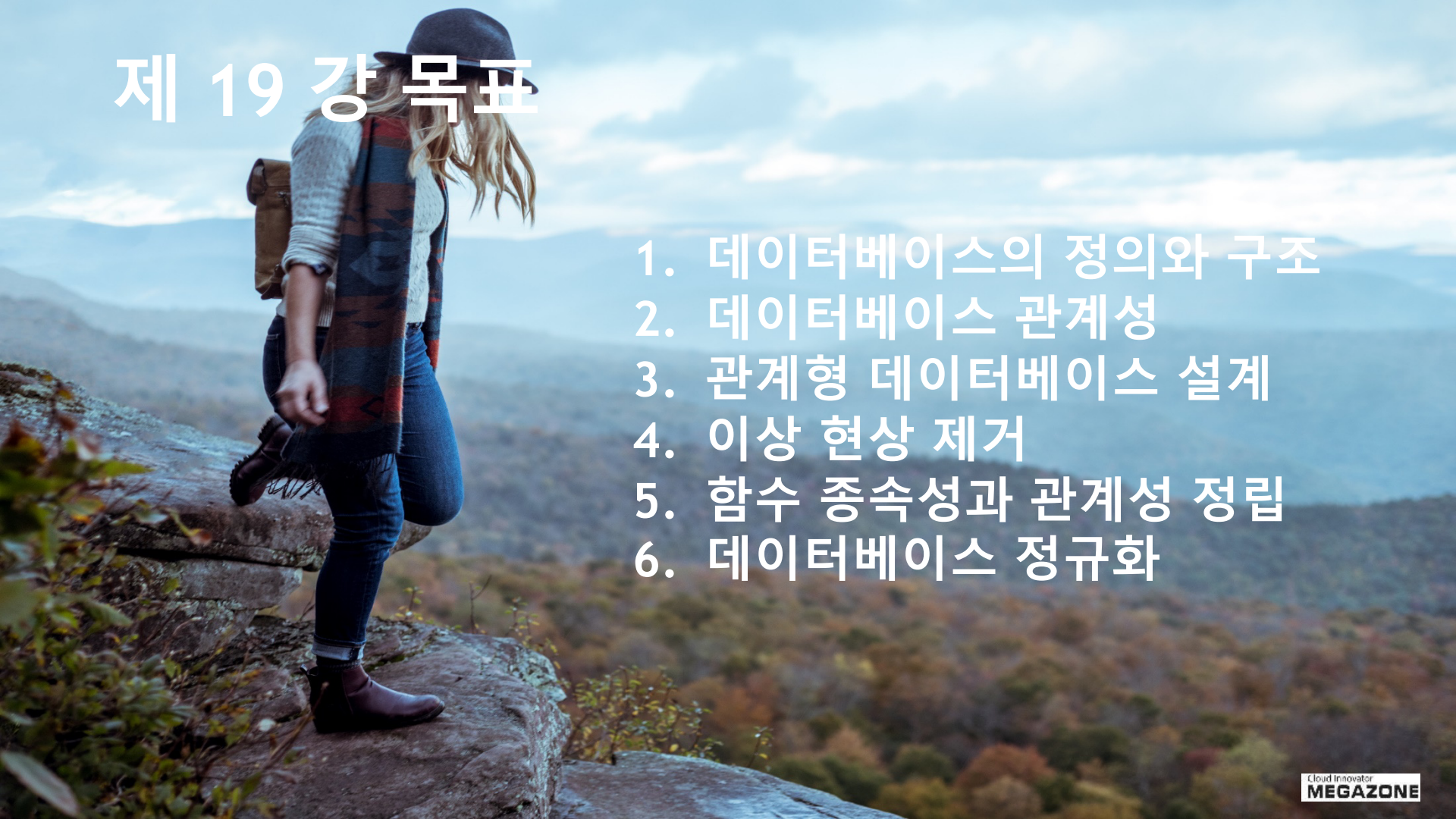
빅쿼리를 활용한 빅데이터 분석 (19)

서진호

자료 다운로드: https://www.github.com/synabreu/BigQuery



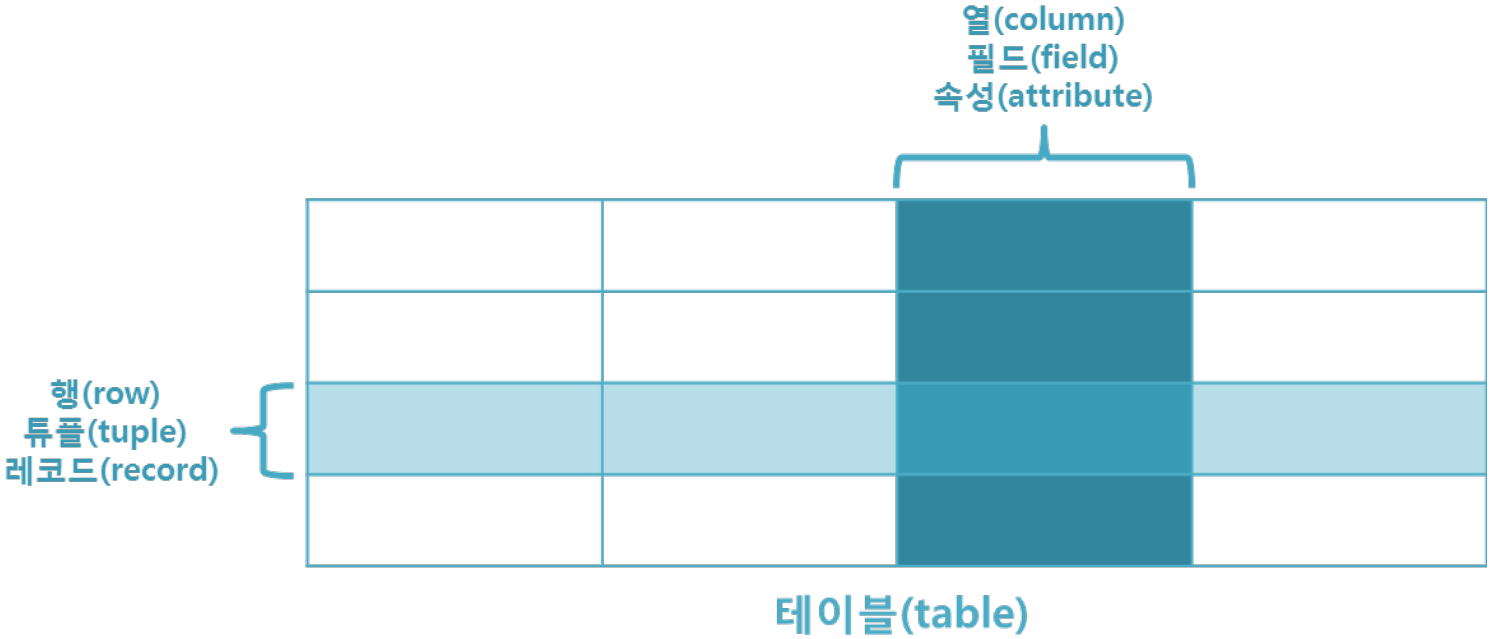
제 19 강 목표

1. 데이터베이스의 정의와 구조
2. 데이터베이스 관계성
3. 관계형 데이터베이스 설계
4. 이상 현상 제거
5. 함수 종속성과 관계성 정립
6. 데이터베이스 정규화

데이터베이스란 무엇인가?

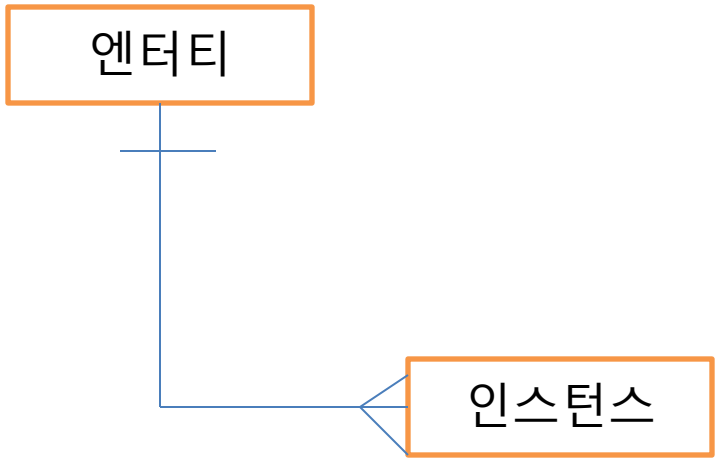
- 여러 사람에 의해 공유되어 사용될 목적으로 통합하여 관리되는 데이터의 집합
- 특정 다수의 사용자들에게 필요한 정보를 제공하거나 조직내에서 필요한 정보를 체계적으로 저장 및 보관하여 사용자들에게 제공
- 사용 예 – 은행 거래에서 입출금, 스마트폰 연락처, 쇼핑몰 고객 관리, 상품 데이터, 배송 관리
- 관계형 데이터베이스 관리 시스템(RDBMS)은 IBM 산호세 연구소의 에드거 F. 커드가 도입한 관계형 모델을 기반으로 하는 데이터베이스 관리 시스템이다.

데이터베이스 구조



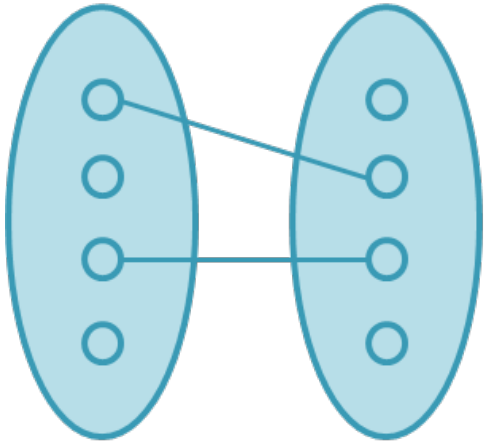
발췌: http://www.tcpschool.com/mysql/mysql_intro_relationalDB

데이터 모델 - 엔터티와 인스턴스

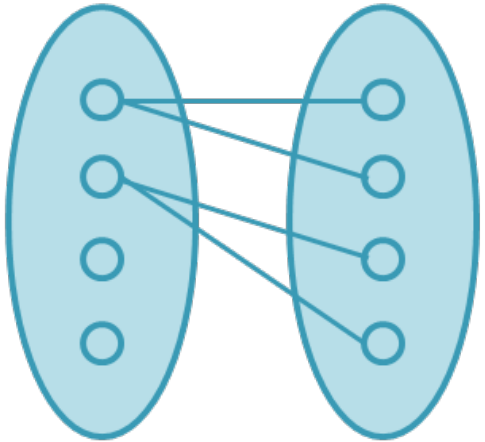


엔터티	인스턴스
과목	국어
	영어
강사	서진호
	홍길동
시간	2022-001
	2022-002

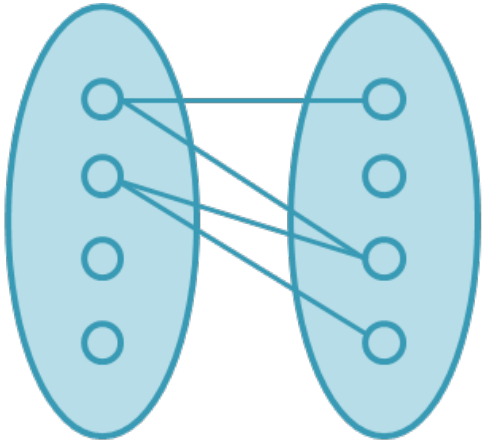
데이터베이스 관계성



일대일(one-to-one)



일대다(one-to-many)



다대다(many-to-many)

관계형 데이터베이스 설계

Original Data

OrderId	CustomerId	CustomerName	timestamp	Location	purchasedItems			
					sku	description	quantity	price
1000001	65401	John Doe	12/18/2017 15:0	Faraway	ABC123456	Redwood 8x4	3	36.3
					TBL535522	Sapient Table	1	878.4
					CHR762222	Cherrywood Ch	6	435.6
					sku	description	quantity	price
1000002	74682	Jane Michaels	12/16/2017 11:3	Nearland	GCH635354	Garden chairs	4	345.7
					GRD828822	Ceramic Pots	2	9.5
					sku	description	quantity	price
1000003	63636	Jose Carlos	12/16/2017 13:4	Nearland				

Transaction Fact

Order Id	timestamp	CustomerId	sku	quantity	price
1000001	12/18/2017 15:02:00	65401	ABC123456	3	36.3
1000001	12/18/2017 15:02:00	65401	TBL535522	1	878.4
1000001	12/18/2017 15:02:00	65401	CHR762222	6	435.6
1000002	12/16/2017 11:34:00	74682	GCH635354	4	345.7
1000002	12/16/2017 11:34:00	74682	GRD828822	2	9.5

Customer Dimension

CustomerId	CustomerName	Location
65401	John Doe	Faraway
74682	Jane Michaels	Nearland
63636	Jose Carlos	Nearland

Product Dimension

sku	description
ABC123456	Redwood 8x4
TBL535522	Sapient Table
CHR762222	Cherrywood Chair
GCH635354	Garden chairs
GRD828822	Ceramic Pots

이상 현상(Anomaly) 제거 - 갱신 이상

Employees' Skills

Employee ID	Employee Address	Skill
426	87 Sycamore Grove	Typing
426	87 Sycamore Grove	Shorthand
519	94 Chestnut Street	Public Speaking
519	96 Walnut Avenue	Carpentry

편집 이상: Employee 519는 다른 레코드에서 다른 주소를 가짐.

이상 현상(Anomaly) 제거 - 추가이상

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201
424	Dr. Newsome	29-Mar-2007	?

추가 이상: 신입 교수인 Dr. Newsome은 아직 수업을 배정받지 않았다는 이유로 교수 정보를 관리하는 이 테이블에 Newsome 교수 레코드를 추가할 수가 없다.

이상 현상(Anomaly) 제거 - 갱신이상

Faculty and Their Courses

Faculty ID	Faculty Name	Faculty Hire Date	Course Code
389	Dr. Giddens	10-Feb-1985	ENG-206
407	Dr. Saperstein	19-Apr-1999	CMP-101
407	Dr. Saperstein	19-Apr-1999	CMP-201

DELETE

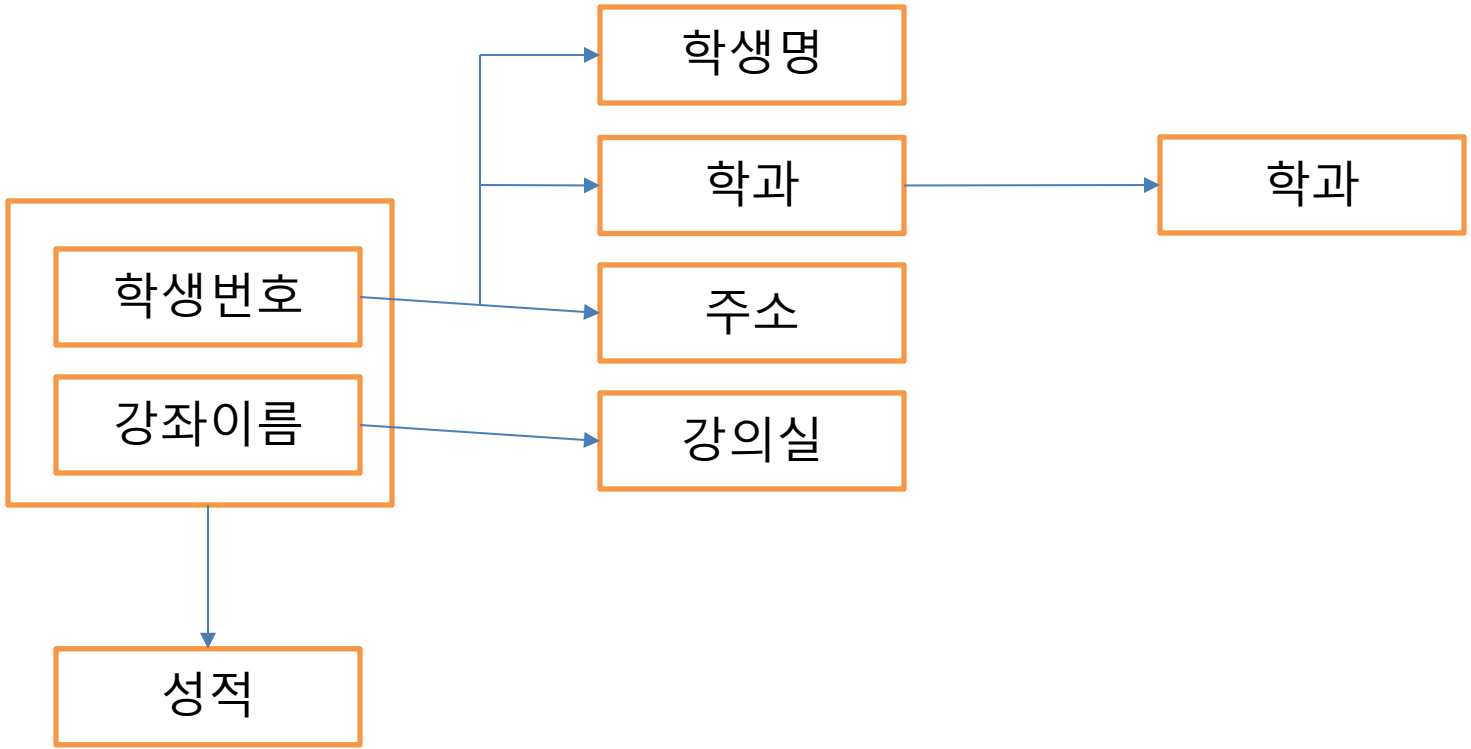
삭제 이상: ENG-206 수업이 끝나 해당 레코드를 삭제하면, Dr. Giddens 교수의 모든 정보가 삭제 됩니다.

함수 종속성

학생번호	학생명	주소	전공	학과사무실	강좌이름	강의실	성적
12890	홍길동	서울시 강남구	컴퓨터 과학	공학관101	C/C++ 프로그래밍	공학관 220	4.0
12094	서진호	서울시 서초구	경영학과	인문관101	데이터베이스 이론	인문관 501	4.3
12898	최철수	서울시 강북구	경영학과	인문관101	데이터베이스 이론	인문관 501	3.9
12907	한영미	서울시 송파구	컴퓨터 과학	공학관101	C/C++ 프로그래밍	공학관 308	4.4

- 관계 스키마 중에서 어느 속성군의 값이 정해지면 다른 속성군의 값이 정해지는 것
- A→B로 표시하고, B는 A에 함수 종속이라고 부른다.

함수 종속성 다이어그램



관계성

학생번호	학생명	주소	전공	취득학점
12890	홍길동	서울시 강남구	컴퓨터 과학	99
12094	서진호	서울시 서초구	경영학과	98
12898	최철수	서울시 강북구	경영학과	95
12907	한영미	서울시 송파구	컴퓨터 과학	92

- 관계 (Relations): 객체 사이의 연관성을 나타냄
- 관계 타입: 객체 타입과 객체 타입 간의 연결 가능한 관계를 정의
- 관계 집합(Relationship set): 관계로 연결된 집합

정규화

고객	사용내역
홍길동	거래번호 / 일자 / 잔고 12890 / 2010-10-14 / –87 12904 / 2010-10-15 / –50
최철수	거래번호 / 일자 / 잔고 12898 / 2010-10-14 / –21
한영미	거래번호 / 일자 / 잔고 12907 / 2010-10-15 / –18 14920 / 2010-11-20 / –70 15003 / 2010-11-27 / –60

정규화 과정

고객	거래번호	일자	잔고
홍길동	12890	2010-10-14	–87
홍길동	12904	2010-10-15	–50
최철수	12898	2010-10-14	–21
한영미	12907	2010-10-15	–18
한영미	14920	2010-11-20	–70
한영미	15003	2010-11-27	–60

- 정규화(Normalization): 이상현상이 발생하는 릴레이션을 분해하여 이상현상을 없애는 과정

정규화 목적 - 다양한 쿼리 지원

SQL 데이터 입출력 수행

```
SELECT * FROM BUYS
WHERE Buys.BuyDate
> '2021/07/01'
```

참조 관계의 여러 테이블로 관리

PK 사용자 테이블(Users)

ID	이름(Name)	성별(sex)
1	진호	남
2	제이	여

PK 상품 테이블(Products)

ID	상품(Product)	단가(UnitPrice)
1	맥북에어	$1199
2	아이폰13	$799
3	아이패드 에어	$999

FK FK

ID	상품(Product)	구매일(BuyDate)
1	1	2021/07/02
1	2	2021/07/02
2	1	2022/07/07
2	2	2022/07/07
2	3	2022/07/07

구매 테이블(Buys)