

데이터베이스의 개념

데이터베이스



데이터베이스

- 데이터베이스란?
- 데이터베이스 존재 이전
- 파일 시스템의 문제점
- 데이터베이스 관리시스템의 등장
- 데이터베이스 관리시스템이란?
- DBMS의 장점
- DBMS의 주요 기능
- 스키마란?
- SQL이란?
- 데이터 구성의 기초

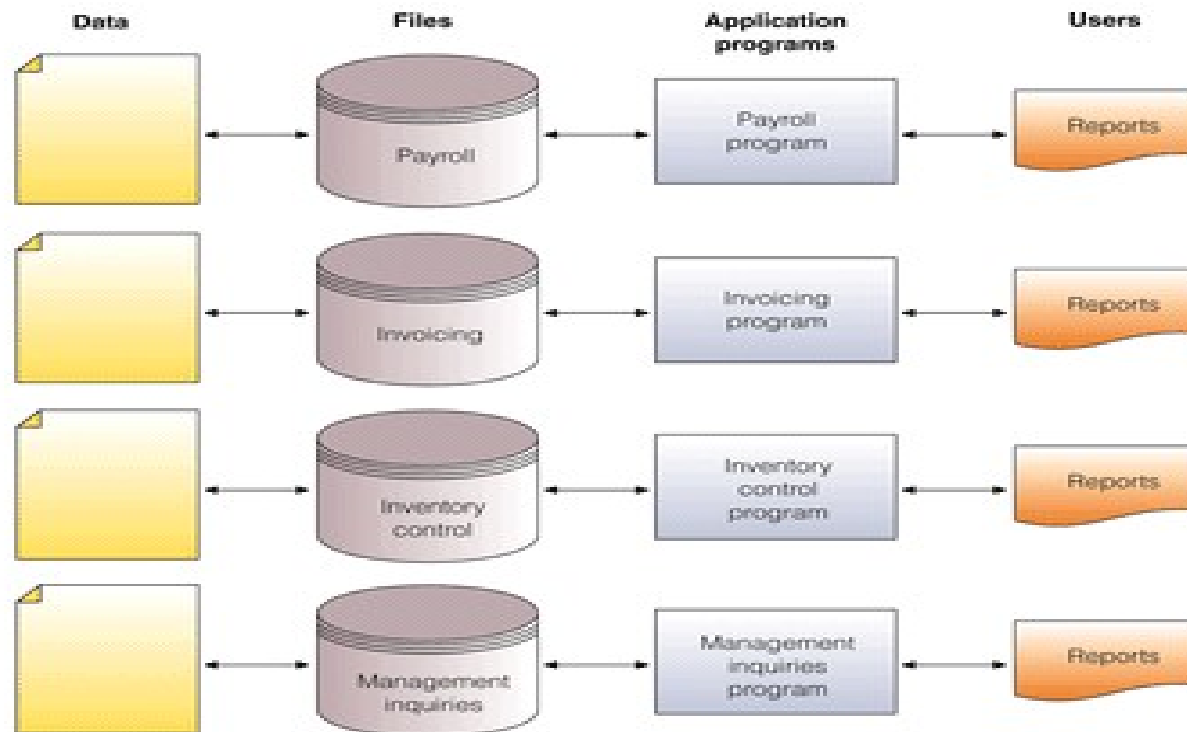
데이터베이스(Database)란?

- 여러 가지 업무에 공동으로 필요한 데이터를 유기적으로 결합하여 저장한 집합체
- 조직에 필요한 정보를 얻기 위해 논리적으로 연관된 데이터를 모아 구조적으로 통합해 놓은 것
 - 데이터를 효율적, 효과적으로 처리하기 위하여 개발
 - 데이터의 중복성, 종속성 문제를 해결
- 데이터, 정보, 지식
 - 데이터: 관찰의 결과로 나타난 정량적 혹은 정성적인 실제 값(에베레스트의 높이는 8,848m)
 - 정보: 데이터에 의미를 부여한 것(에베레스트는 세계에서 가장 높은 산)
 - 지식: 사물이나 현상에 대한 이해(에베레스트 등정 보고서)
- 데이터베이스 관리시스템(Database Management Systems)
 - 데이터베이스를 관리하는 시스템

데이터베이스가 없던 시절

■ 파일 시스템(File system)으로 관리

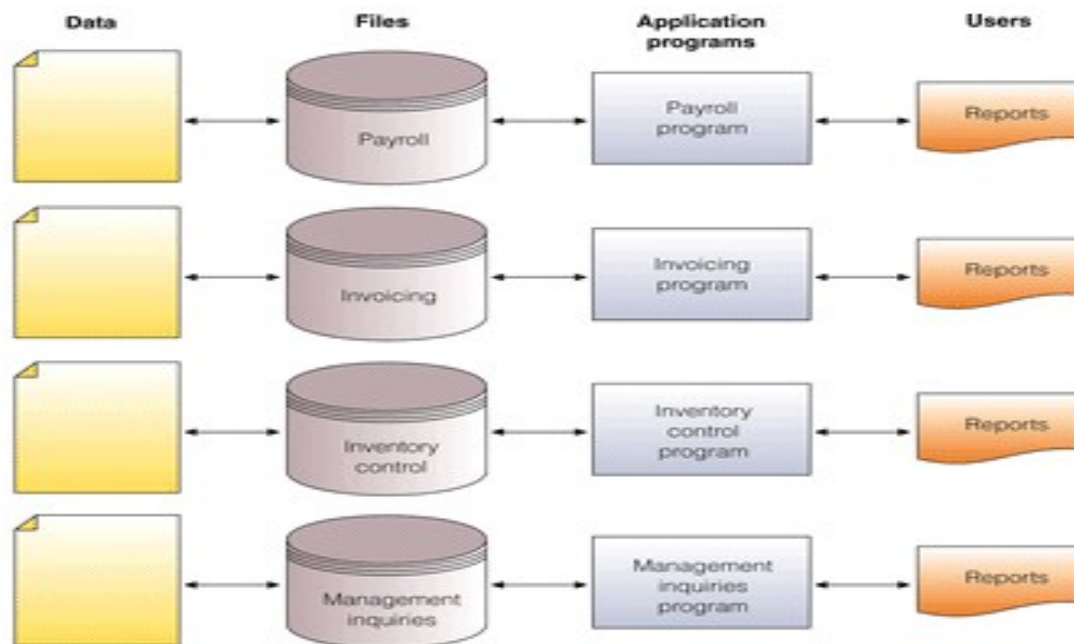
- 응용 프로그램마다 필요한 데이터를 별도의 파일로 관리



파일 시스템(File system)의 문제점

■ 종속성

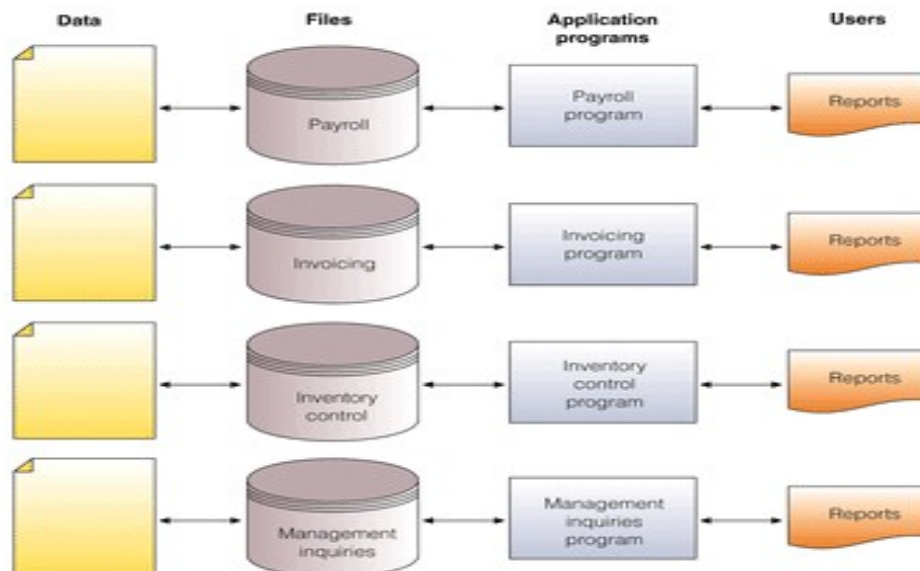
- 데이터 파일은 해당 응용프로그램에서만 사용가능
- 저장된 데이터의 접근 방법을 변경할 때는 응용 프로그램도 같이 변경 필요



파일 시스템(File system)의 문제점...

■ 중복성

- 일관성: 중복된 데이터 간에 내용이 일치하지 않는 상황
- 보안성: 중복되어 있는 모든 데이터에 동등한 보안 수준 유지의 어려움
- 경제성: 저장공간의 낭비, 동일한 데이터의 반복 작업으로 인한 비용 증가
- 무결성: 제어의 분산으로 인한 데이터의 정확성 유지가 어려움



■ 1963

- 데이터베이스라는 용어가 'Development and Management of Computer Center Data Bases' 라는 심포지엄에서 처음 사용
- 최초의 범용 데이터베이스 관리시스템(DBMS) 설계
 - ➔ GE사의 'Integrated Data Store'

■ 1970

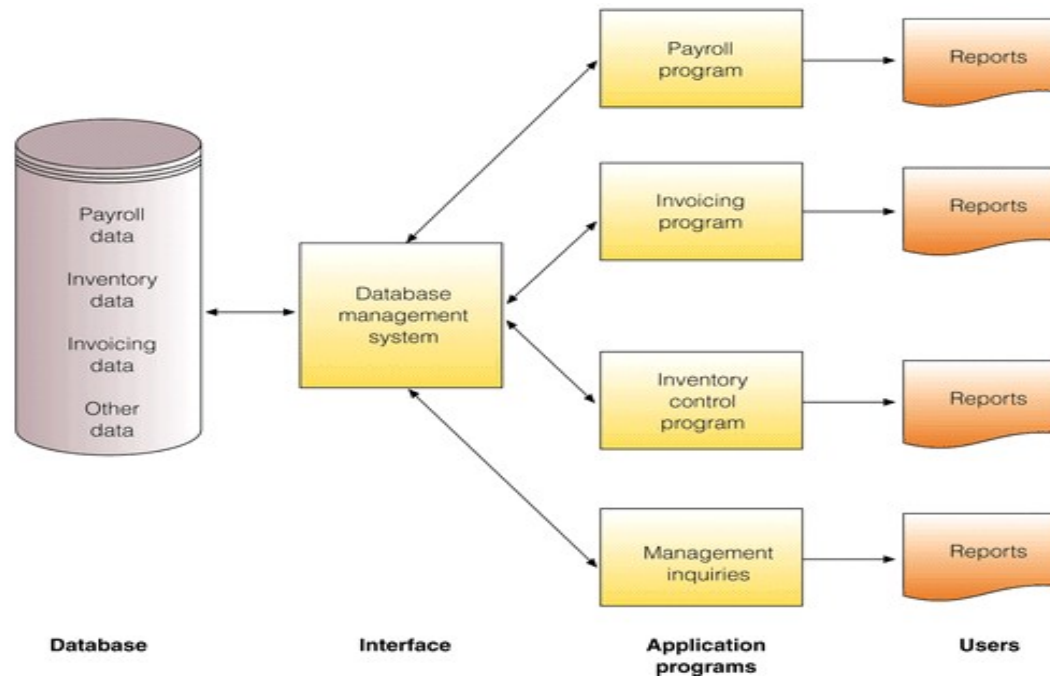
- 에드거 프랭크 커드(Edgar Frank Codd) 관계형 데이터베이스 모델 제안

■ 1980년대

- 상용 관계형 데이터베이스 관리시스템(DBMS) 등장
 - ➔ Oracle, DB2, Sybase 등

데이터베이스 관리시스템이란?

- 컴퓨터에 저장되는 데이터베이스를 관리해주는 소프트웨어 시스템
- 응용프로그램들과 물리적 데이터 파일들 간의 인터페이스를 제공

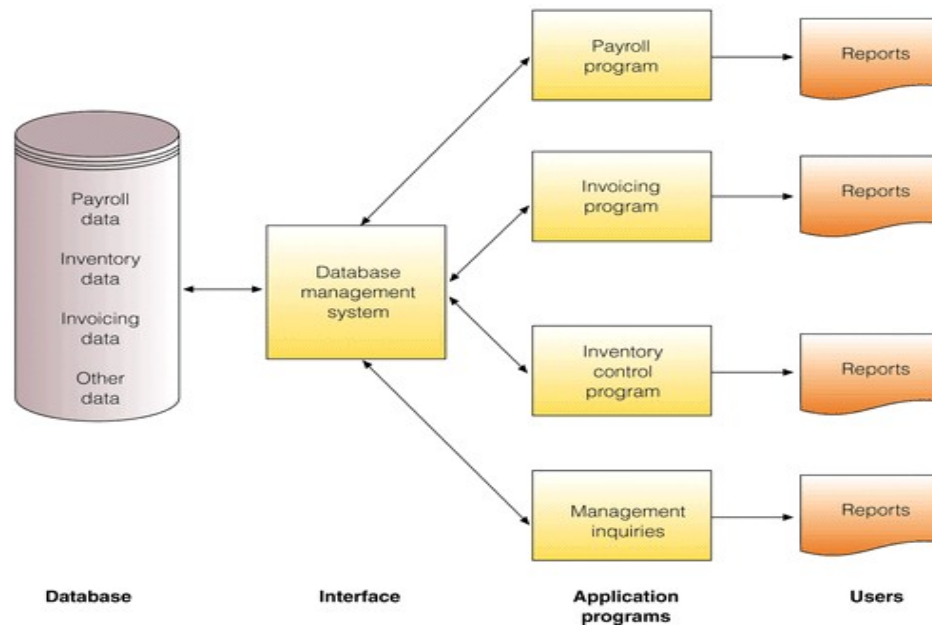


■ 독립성

- 응용프로그램과 데이터간의 독립성 확보

■ 데이터 중복의 제거

- 일관성 확보
- 보안 수준 제고
- 경제성 제고
- 데이터 무결성 제고



- 데이터 정의 기능(Data Definition)
 - 데이터의 내용과 구조, 즉 데이터베이스 스키마(schema)를 정의
- 데이터 조작 기능(Data Manipulation)
 - 데이터베이스 내의 데이터 항목들에 대한 갱신, 대체, 추출, 삽입, 삭제 분류 등과 같은 데이터의 조작에 사용
- 데이터 제어 기능(Data Control)
 - 정당한 사용자가 허가된 데이터만 접근할 수 있도록 보안을 유지하고 권한을 검사하는데 사용

스키마(Schema)란?

- 데이터베이스의 구조와 제약조건에 관한 전반적인 명세를 기술
- 데이터베이스를 구성하는 데이터 개체(Entity), 속성(Attribute), 관계(Relationship) 및 데이터 조작 시 데이터 값들이 갖는 제약 조건 등에 관해 전반적으로 정의
- 스키마는 데이터 사전에 저장되며, 데이터 사전은 다른 이름으로 메타 데이터 (Meta-Data)라고도 함
 - 데이터 사전: 데이터 베이스에 저장되는 각 데이터 항목들에 관한 정보를 모아놓은 것

SQL(Structured Query Language)이란?

- 데이터베이스를 구축하고 활용하기 위하여 사용하는 언어
- 보통 관계형 데이터베이스를 조작하기 위한 표준 언어
- 데이터 정의언어(DDL : Data Definition Language)와 데이터 조작 언어(DML : Data Manipulation Language)로 구분
- DDL은 관계 생성, 관계 삭제, 관계 변경 등에 사용
- DML은 검색, 삽입, 삭제, 갱신 등에 사용

릴레이션

도서 1, 축구의 역사, 굿스포츠, 7000		도서번호	도서이름	출판사	가격
도서 2, 축구아는 여자, 나무수, 13000		1	축구의 역사	굿스포츠	7000
도서 3, 축구의 이해, 대한미디어, 22000		2	축구아는 여자	나무수	13000
도서 4, 골프 바이블, 대한미디어, 35000		3	축구의 이해	대한미디어	22000
도서 5, 피겨 교본, 굿스포츠, 8000		4	골프 바이블	대한미디어	35000
		5	피겨 교본	굿스포츠	8000

그림 2-1 데이터와 테이블(릴레이션)

도서번호 = {1, 2, 3, 4, 5}
 도서이름 = {축구의 역사, 축구아는 여자, 축구의 이해, 골프 바이블, 피겨 교본}
 출판사 = {굿스포츠, 나무수, 대한미디어}
 가격 = {7000, 13000, 22000, 35000, 8000}

→ 첫 번째 행(1, 축구의 역사, 굿스포츠, 7000)의 경우 네 개의 집합에서 각각 원소 한 개씩 선택하여 만들어진 것으로 이 원소들이 관계(relationship)를 맺고 있다.

릴레이션

❖관계(relationship)

- ❶ 릴레이션 내에서 생성되는 관계 : 릴레이션 내 데이터들의 관계
- ❷ 릴레이션 간에 생성되는 관계 : 릴레이션 간의 관계

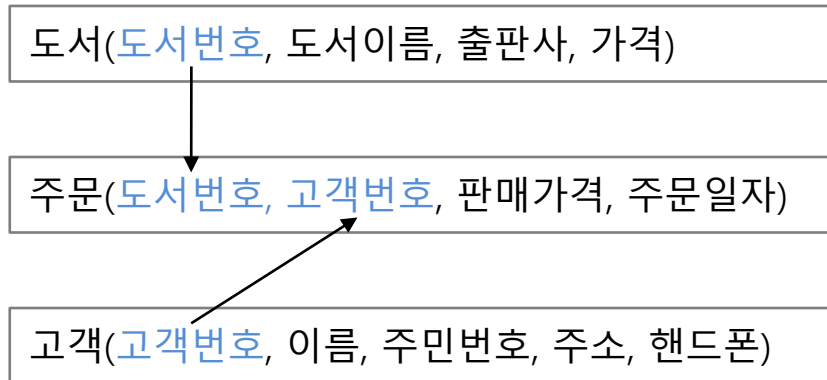


그림 2-2 릴레이션 간의 관계

릴레이션 스키마와 인스턴스

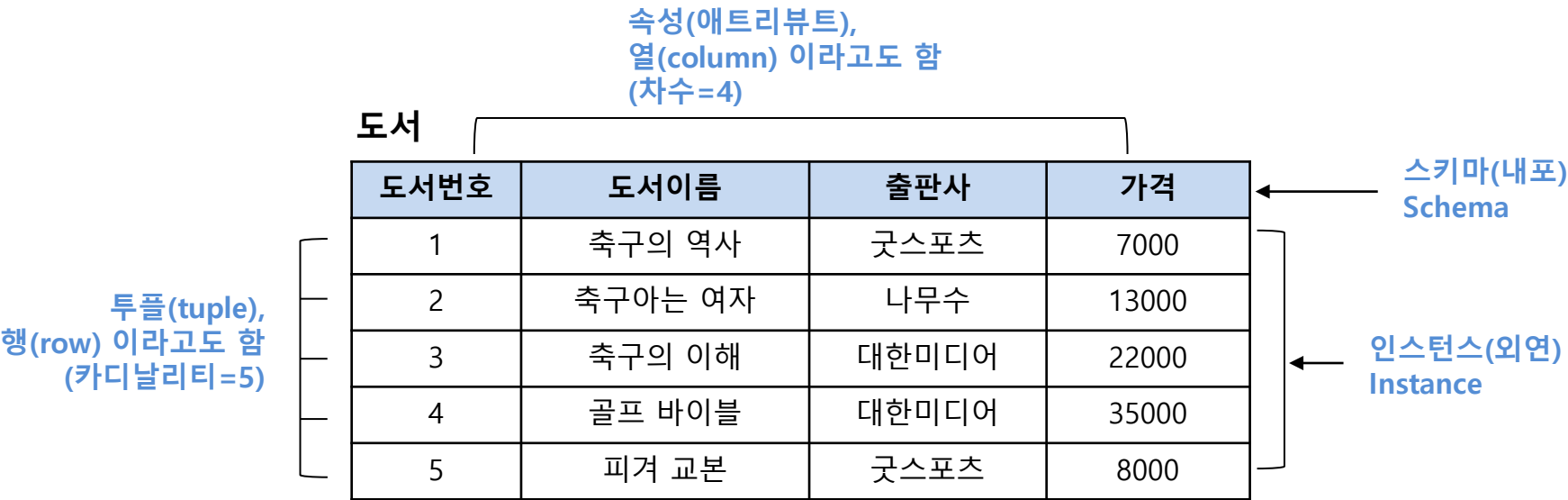


그림 2-3 도서 릴레이션

릴레이션 스키마와 인스턴스

❖ 릴레이션 스키마

■ 스키마의 요소

- 속성(attribute) : 릴레이션 스키마의 열
- 도메인(domain) : 속성이 가질 수 있는 값의 집합
- 차수(degree) : 속성의 개수

■ 스키마의 표현

- 릴레이션 이름(속성1 : 도메인1, 속성2 : 도메인2, 속성3 : 도메인3 ...)
EX) 도서(도서번호, 도서이름, 출판사, 가격)

릴레이션 스키마와 인스턴스

❖ 릴레이션 인스턴스

■ 인스턴스 요소

- 튜플(tuple) : 릴레이션의 행
- 카디널리티(cardinality) : 튜플의 수

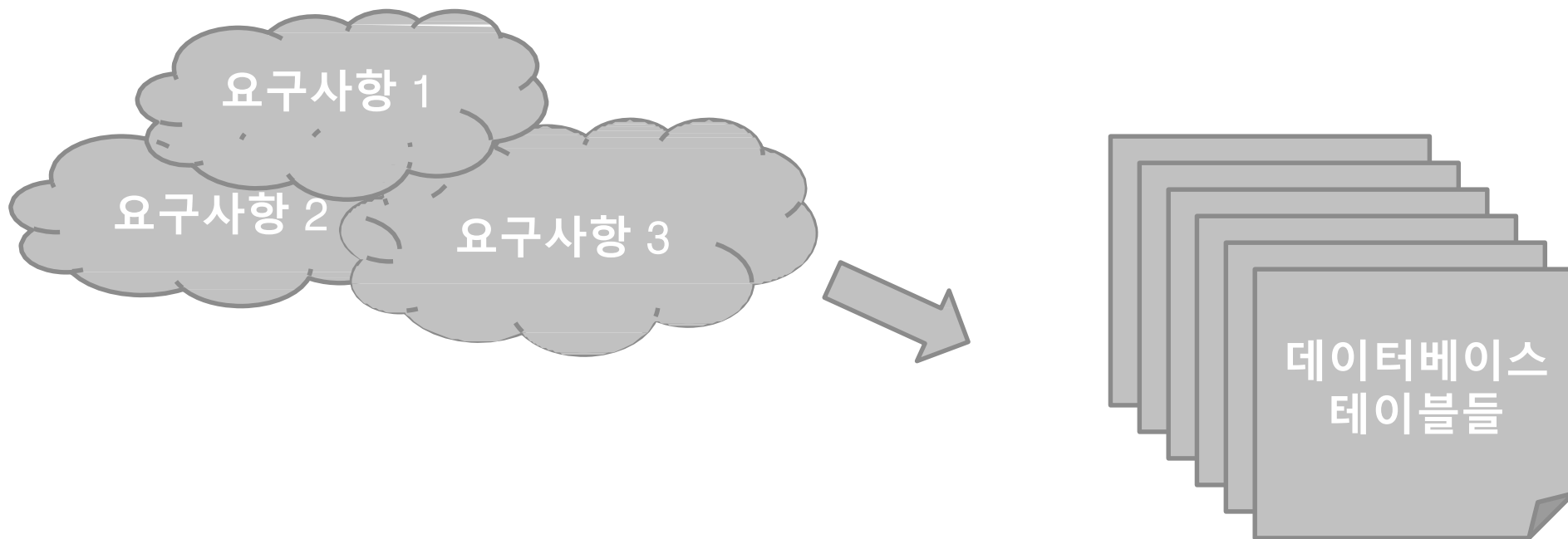
→ 튜플이 가지는 속성의 개수는 릴레이션 스키마의 차수와 동일하고,
릴레이션 내의 모든 튜플들은 서로 중복되지 않아야 함

표 2-2 릴레이션 구조와 관련된 용어

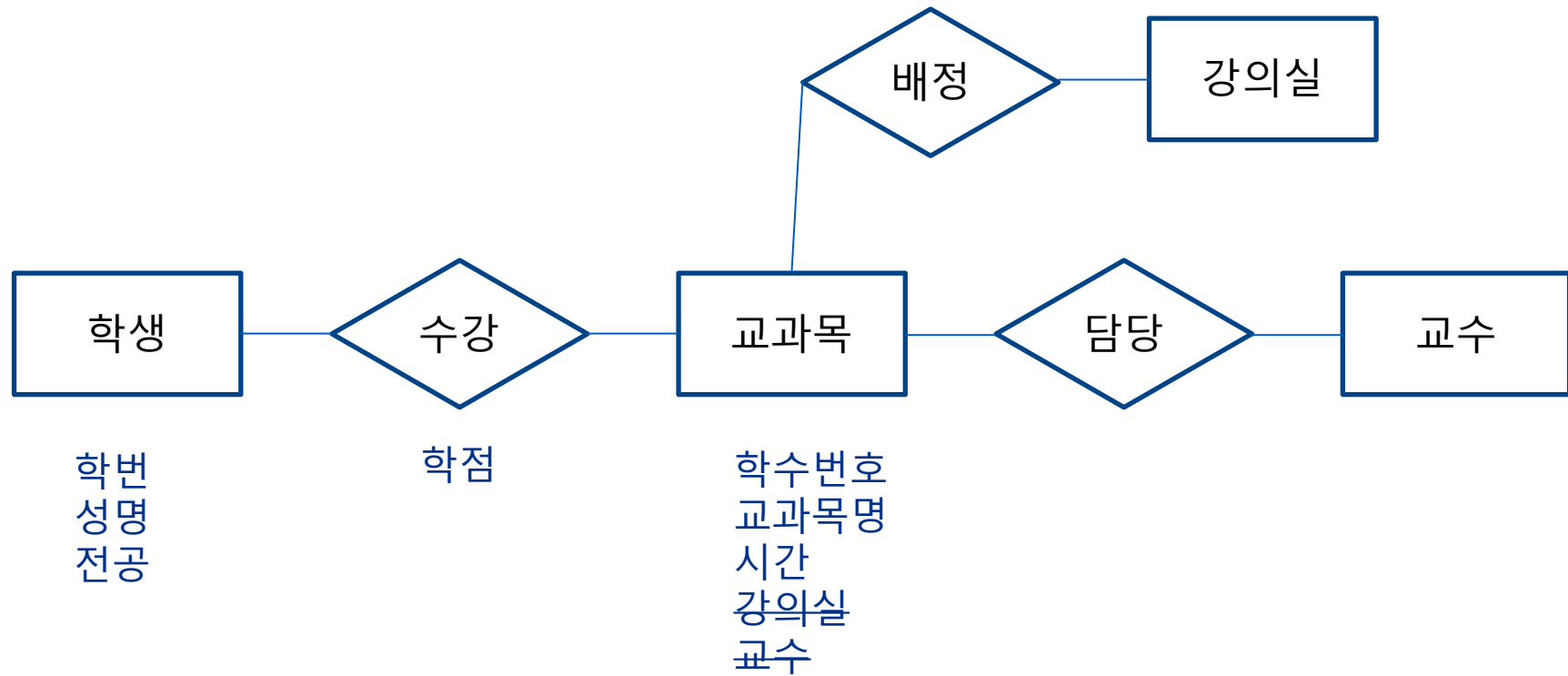
릴레이션 용어	같은 의미로 통용되는 용어	파일 시스템 용어
릴레이션(relation)	테이블(table)	파일(file)
스키마(schema)	내포(intension)	헤더(header)
인스턴스(instance)	외연(extension)	데이터(data)
튜플(tuple)	행(row)	레코드(record)
속성(attribute)	열(column)	필드(field)

데이터베이스 모델링

- 데이터를 체계적으로 저장할 수 있도록 데이터베이스를 설계해야겠는데 어떻게 하지?



데이터베이스 설계 맛보기



데이터 모델링의 개념

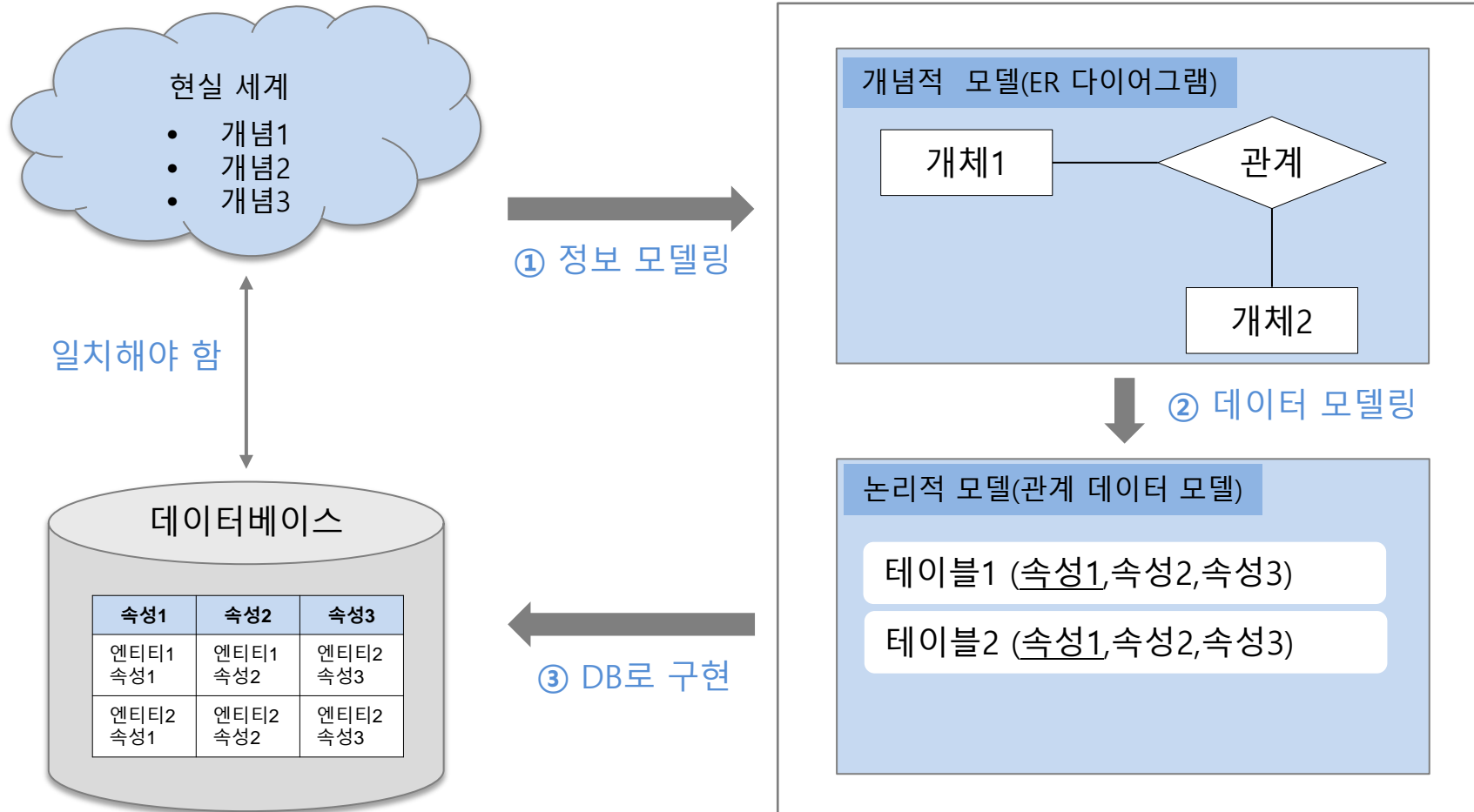


그림 6-2 데이터 모델링의 개념

데이터 모델링 과정

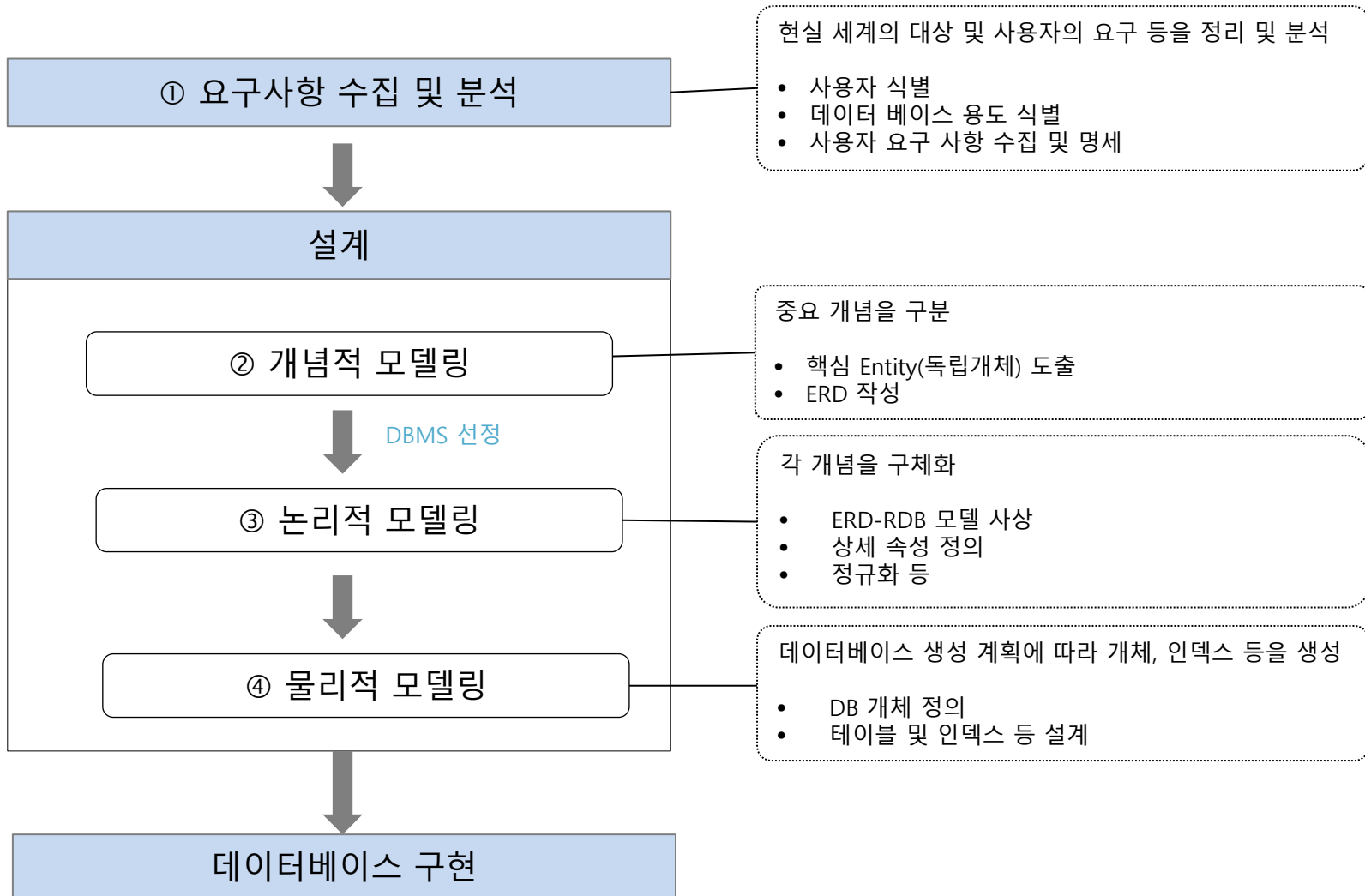


그림 6-4 데이터 모델링 과정

데이터 모델링 과정

❖ 요구사항 수집 및 분석

■ 요구사항 수집 방법

1. 실제 문서를 수집하고 분석함
2. 담당자와의 인터뷰나 설문조사를 통해 요구사항을 직접 수렴함
3. 비슷한 업무를 처리하는 기존의 데이터베이스를 분석함
4. 각 업무와 연관된 모든 부분을 살펴봄

데이터 모델링 과정

❖ 개념적 모델링

- 요구사항을 수집하고 분석한 결과를 토대로 업무의 핵심적인 개념을 구분하고 전체적인 뼈대를 만드는 과정
- 개체(entity)를 추출하고 각 개체들 간의 관계를 정의하여 ER 다이어그램(ERD, Entity Relationship Diagram)을 만드는 과정까지를 말함

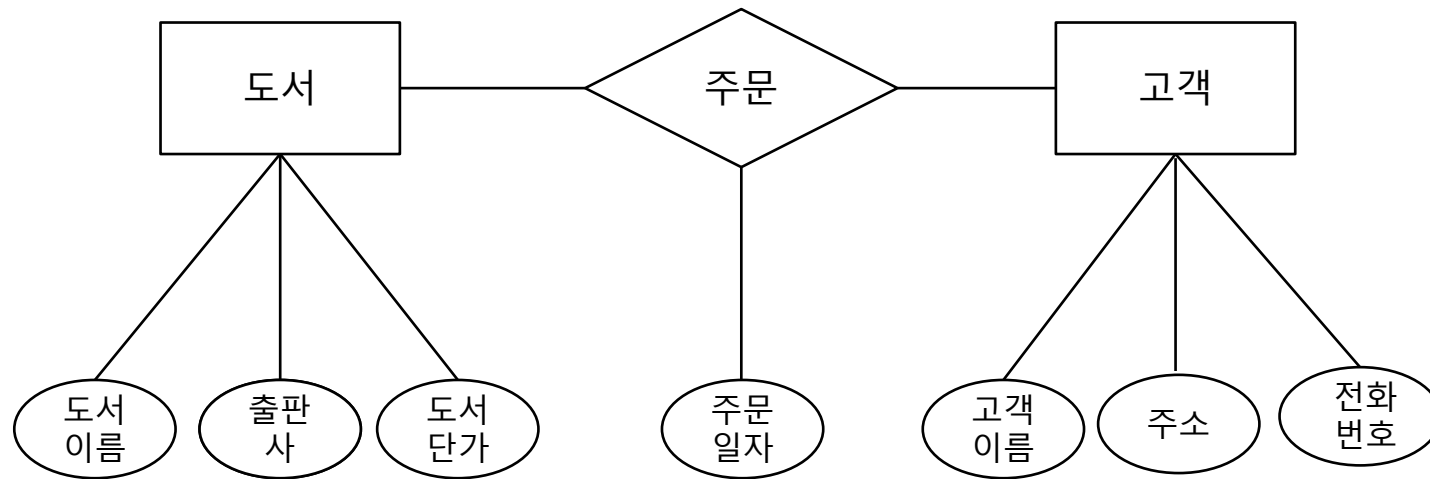


그림 6-5 개념적 모델링의 예

데이터 모델링 과정

❖ 논리적 모델링

- 개념적 모델링에서 만든 ER 다이어그램을 사용하려는 DBMS에 맞게 사상(매핑, mapping)하여 실제 데이터베이스로 구현하기 위한 모델을 만드는 과정

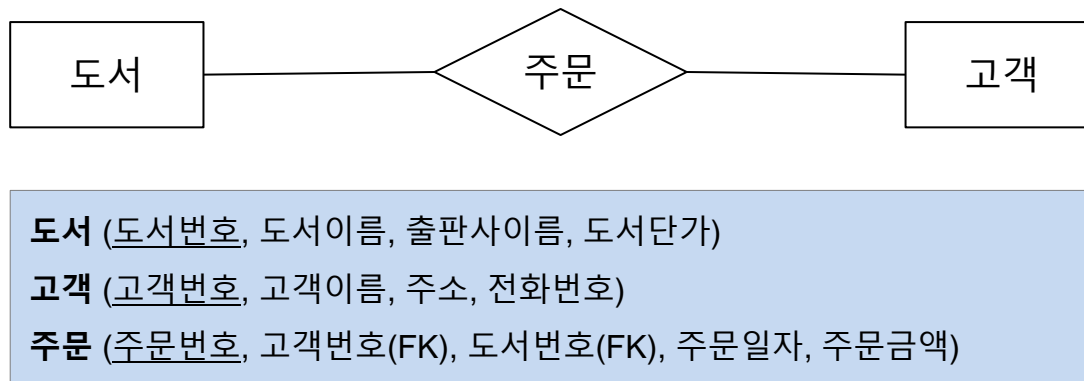


그림 6-6 논리적 모델링의 예

■ 논리적 모델링 과정

- 개념적 모델링에서 추출하지 않았던 상세 속성들을 모두 추출
- 정규화 수행
- 데이터 표준화 수행

데이터 모델링 과정

❖ 물리적 모델링

- 작성된 논리적 모델을 실제 컴퓨터의 저장 장치에 저장하기 위한 물리적 구조를 정의하고 구현하는 과정
- DBMS 특성에 맞게 저장 구조를 정의해야 데이터베이스가 최적의 성능을 낼 수 있음

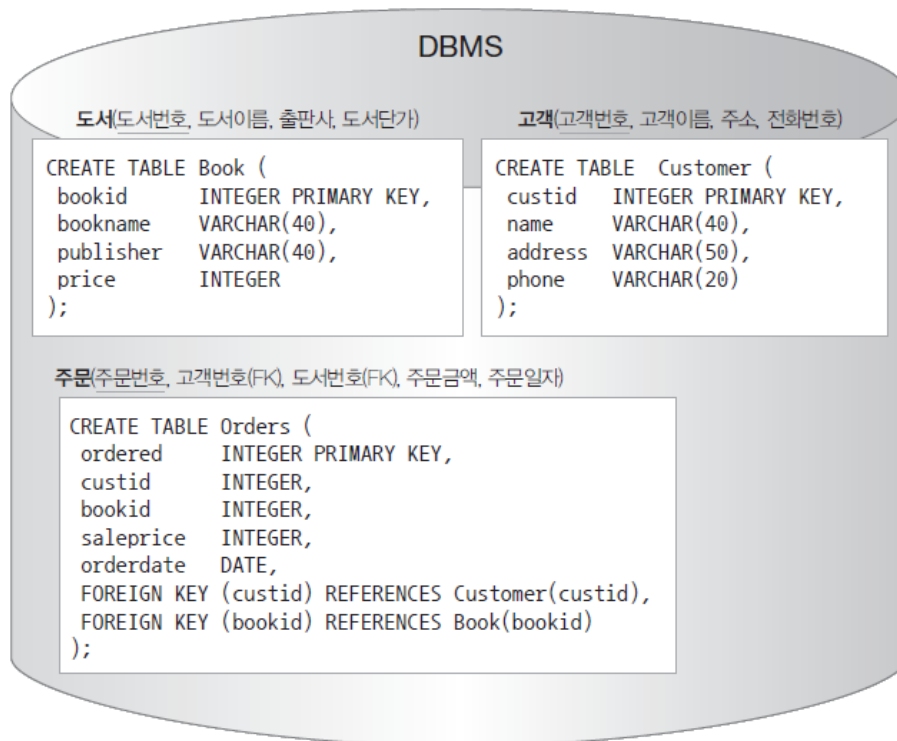


그림 6-7 물리적 모델링의 예

데이터 모델링 과정

❖ 물리적 모델링

- 물리적 모델링 시 트랜잭션, 저장 공간 설계 측면에서 고려할 사항
 1. 응답시간을 최소화해야 한다.
 2. 얼마나 많은 트랜잭션을 동시에 발생시킬 수 있는지 검토해야 한다.
 3. 데이터가 저장될 공간을 효율적으로 배치해야 한다.

■ 개체-관계(Entity-Relationship) 모델

- 개념적 데이터 모델의 가장 대표적인 모델
- 개체 타입(Entity Type)과 이들 간의 관계 타입(Relationship Type)을 이용해 현실 세계를 개념적으로 표현
- 데이터를 개체(Entity), 관계(Relationship), 속성(Attribute)으로 묘사

■ 개체(Entity)

- 데이터베이스에 표현하려는 것으로, 사람이 생각하는 개념이나 정보 단위 같은 현실 세계의 대상체
- 유형, 무형의 정보로서 서로 연관된 몇 개의 속성으로 구성

■ 속성(Attribute)

- 데이터의 가장 작은 논리적 단위, 파일 구조의 데이터 항목 또는 필드에 해당





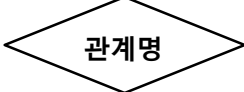
속성들



학생 번호	학생 이름	학생 주소	주민등록번호	전공
1	김병진	서울시 서대문구 연희동 1번지	951002-1234567	경영학
2	윤태형	서울시 강서구 화곡6동 1번지	950226-1234567	심리학
3	성형도	서울시 마포구 창천동 1번지	940910-1234567	사회학
4	박병진	서울시 용산구 한남동 1번지	960101-1234567	심리학
5	은소영	서울시 서대문구 대현동 1번지	951205-1234567	경영학

<개체명: 학생>

■ 개체관계 모델의 도식화 방법

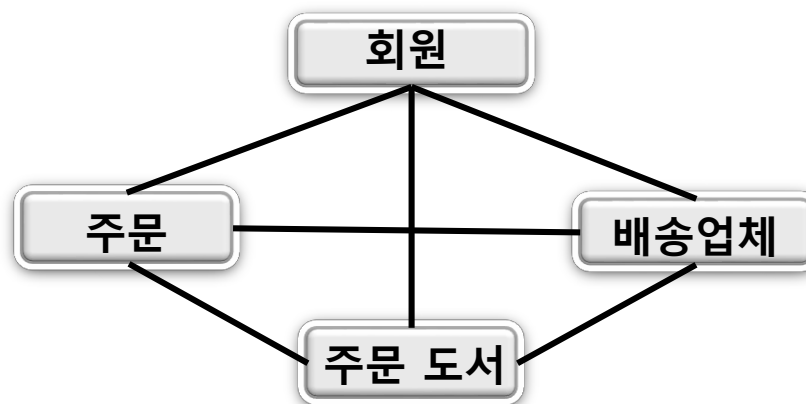
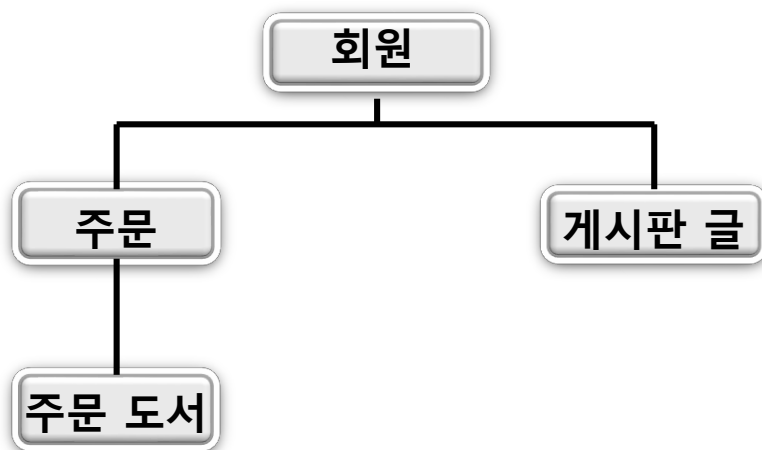
구성요소명	도식화	의미
개체		자료집합 또는 구체적이고 의미 있는 실체
속성		특정 개체를 다른 개체들과 구분하기 위한 고유한 특성 또는 성질로 개체의 구성요소가 된다.
키속성		속성 중 각각의 개별 개체를 구별할 수 있는 고유한 속성이며, 속성명에 밑줄을 그어 사용한다.
연결		연결선이라고도 불리우며 관계 있는 개체와 관계를 연결할 때, 개체와 해당 개체에 속한 속성을 연결할 때 사용한다.
관계		개체와 개체 사이의 관계를 나타낸다.

■ 계층형

- 데이터 상호간의 관계를 계층적으로 나타내어 트리 형태로 구성한 데이터 베이스 모델

■ 네트워크형

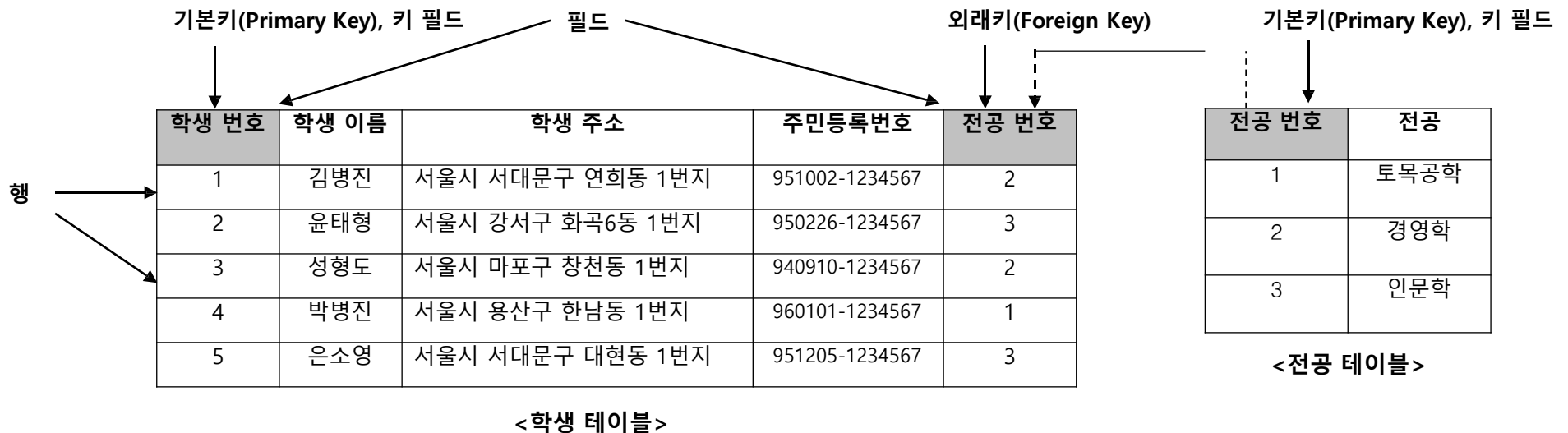
- 서로 연관된 데이터 상호간의 관계를 망 형식으로 하여 레코드를 그래프 형태로 표현한 데이터베이스 모델



관계형 모델(Relational Model)

■ 실세계의 데이터를 누구나 직관적으로 이해할 수 있는 테이블 형식을 이용하여 데이터들을 정의하고 설명한 모델

- 각각의 테이블은 하나의 개체와 속성들의 데이터들로 구성됨
- 데이터 간의 관계를 기본키(Primary Key)와 이를 참조하는 외래키 (Foreign Key)로 표현



다시 수강신청시스템



데이터 웨어하우스

데이터 웨어하우스(Data Warehouse)란?

■ 데이터 웨어하우스(Data Warehouse)란?

- 기업 내에 있는 각종 데이터를 적절히 뽑아내고 조합해 다양한 사업목적에 맞는 정보, 또는 지식으로 바꾸어 주는 기술
- 기존의 데이터베이스처럼 거래처리를 위한 데이터가 아니고, 의사결정 지원을 위한 데이터 베이스

■ 데이터베이스

- 즉각적인 업무 처리를 위한 OLTP(On-Line Transaction Processing) 환경에 적합
- 신속한 데이터 처리에 비해 상대적으로 데이터 분석은 느리고 많은 제한을 가짐

■ 해결 방안 : 업무용 데이터와 분석용 데이터의 저장소를 분리

- 데이터베이스 - 업무 트랜잭션 처리를 위한 운영 저장소로 사용
- 분석을 위한 새로운 저장소를 구축 => 데이터웨어하우스

데이터 웨어하우스(Data Warehouse)란?

■ 몇 가지 정의들

- Inmon (1992)

- ➡ 기업의 의사결정 과정을 지원하기 위한 주제 중심적이고 통합적이며 시간성을 가지는 비휘발성 자료의 집합

- Kelly (1994)

- ➡ 기업 내의 의사결정 지원 어플리케이션들을 위한 정보기반을 제공하는 통합된 데이터 저장공간

- Poe (1994)

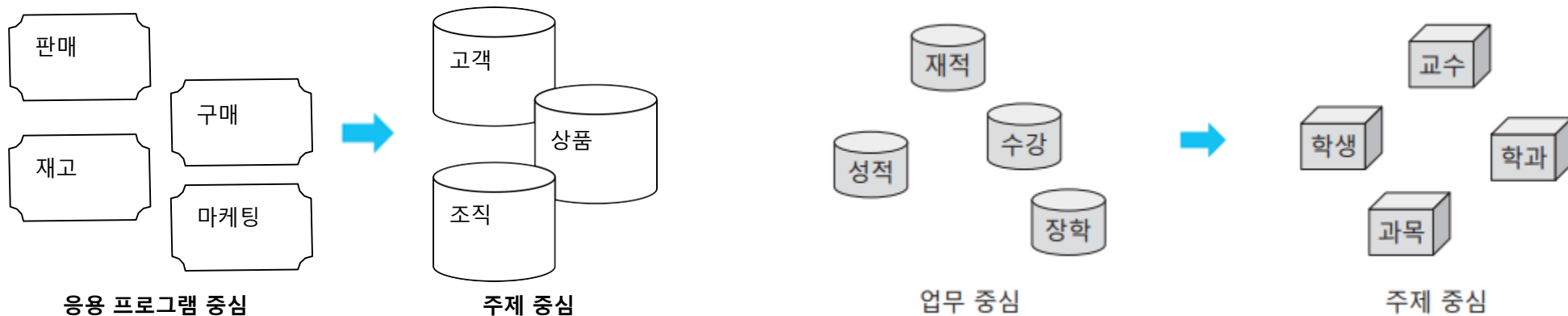
- ➡ 의사결정 지원에 효과적으로 사용될 수 있도록 다양한 운영시스템으로부터 추출, 변환, 통합되고 요약된 읽기 전용 데이터베이스

■ 데이터베이스 Vs. 데이터 웨어하우스

데이터베이스	데이터 웨어하우스
<ul style="list-style-type: none">● 거래처리 중심● 응용프로그램 지원	<ul style="list-style-type: none">● 지식분석 중심● 의사결정지원시스템의 데이터베이스

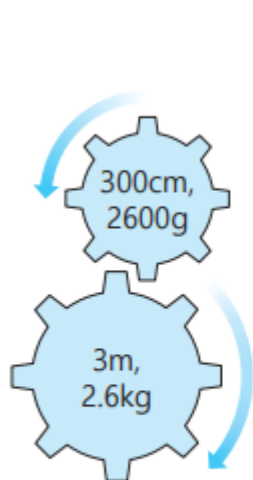
■ 주제 중심적

- 기존의 데이터베이스가 응용프로그램 중심적이었다면
- 데이터베이스는 재고관리, 영업관리, 회계관리 등 기업 운영에 필요한 업무 프로세스 처리를 지원하기 위해 설계
- 데이터 웨어하우스는 기업의 의사결정을 위한 주요 주제 및 그와 관련된 데이터들이 중심

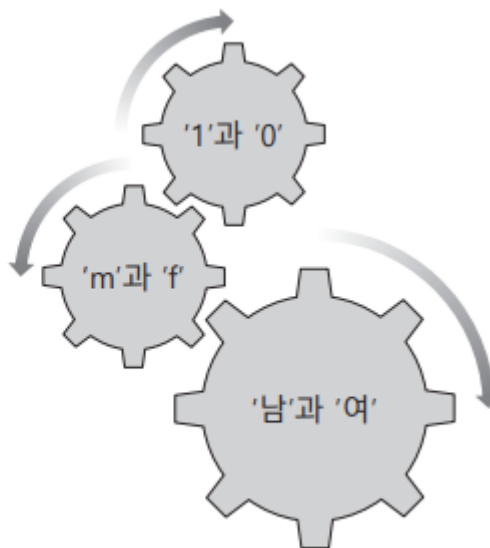


■ 통합적 구조

- 데이터베이스뿐만 아니라 외부 문서(웹 문서, 엑셀 문서, 텍스트 문서 등) 간의 통합을 고려해야
- 분석을 위해서는 하나의 통일된 형식으로 변환이 필요
- 통합 과정에서 데이터 유형과 측정 단위, 데이터의 불일치 문제 등을 해결
- 데이터웨어하우스의 핵심이 되는 가장 중요한 특성으로 많은 시간과 노력이 필요함



(a) 규격 단위 불일치



(b) 성별 형식 불일치

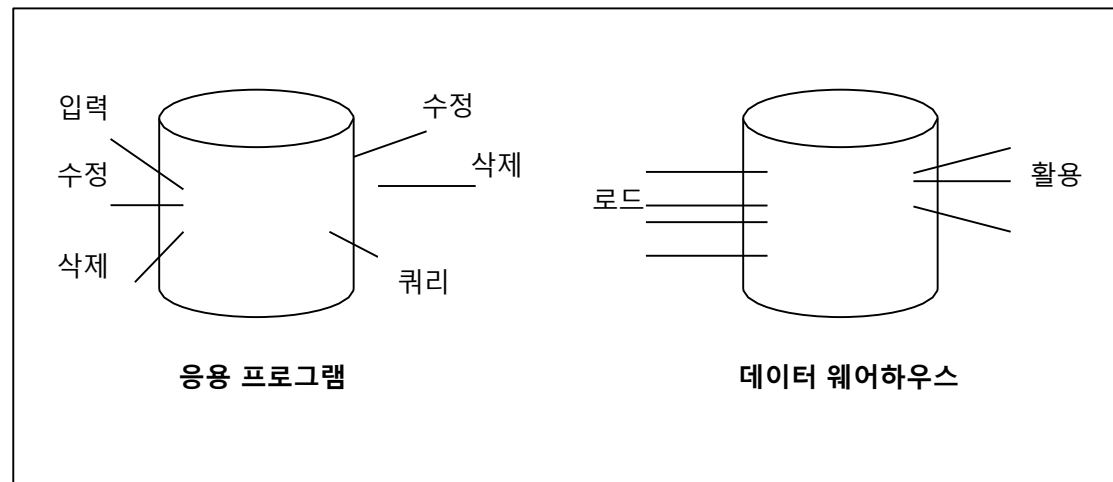
■ 시계열성

- 데이터베이스 - 주로 현 시점까지의 업무 결과를 반영
- 분석 과정 - 과거부터 현재까지 여러 시점의 데이터가 모두 의미가 있음
- 추이 분석을 위해서는 각 시점에서의 데이터 버전 즉, 스냅샷(snapshot)을 저장하는 것이 중요함
- 데이터웨어하우스 - 스냅샷 데이터들이 축적되는 대규모 저장소
- 시간 흐름에 따라 원시 데이터 혹은 요약 데이터 형태로 다양하게 중복 저장



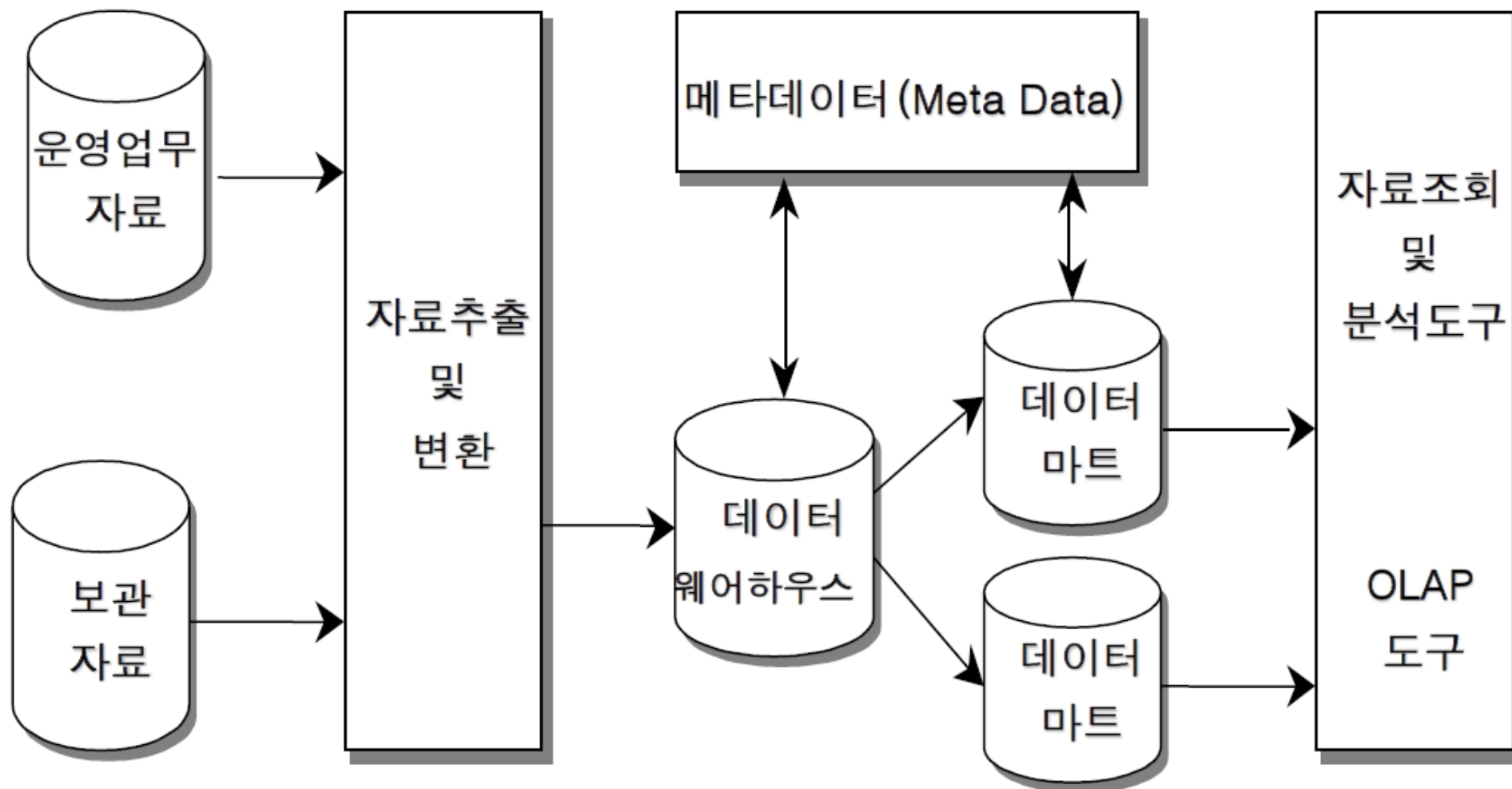
■ 비휘발성

- 각 부서 단위로 운영하고 있는 데이터베이스에서는 추가/삭제/변경과 같은 갱신 작업이 레코드 단위로 지속적으로 발생
- 데이터 웨어하우스에서는 데이터 로드와 활용만이 존재하며, 기존 운영 시스템에서와 같은 삭제와 갱신은 발생하지 않음



- 데이터 웨어하우스와 사용자 사이의 중간층에 위치
- 하나의 주제 또는 하나의 부서 중심의 데이터 웨어하우스
- 데이터 웨어하우스가 도매상이라면 데이터 마트는 소매상
- 데이터 마트의 데이터는 대부분 데이터 웨어하우스로부터 복제

데이터웨어하우스의 구성요소



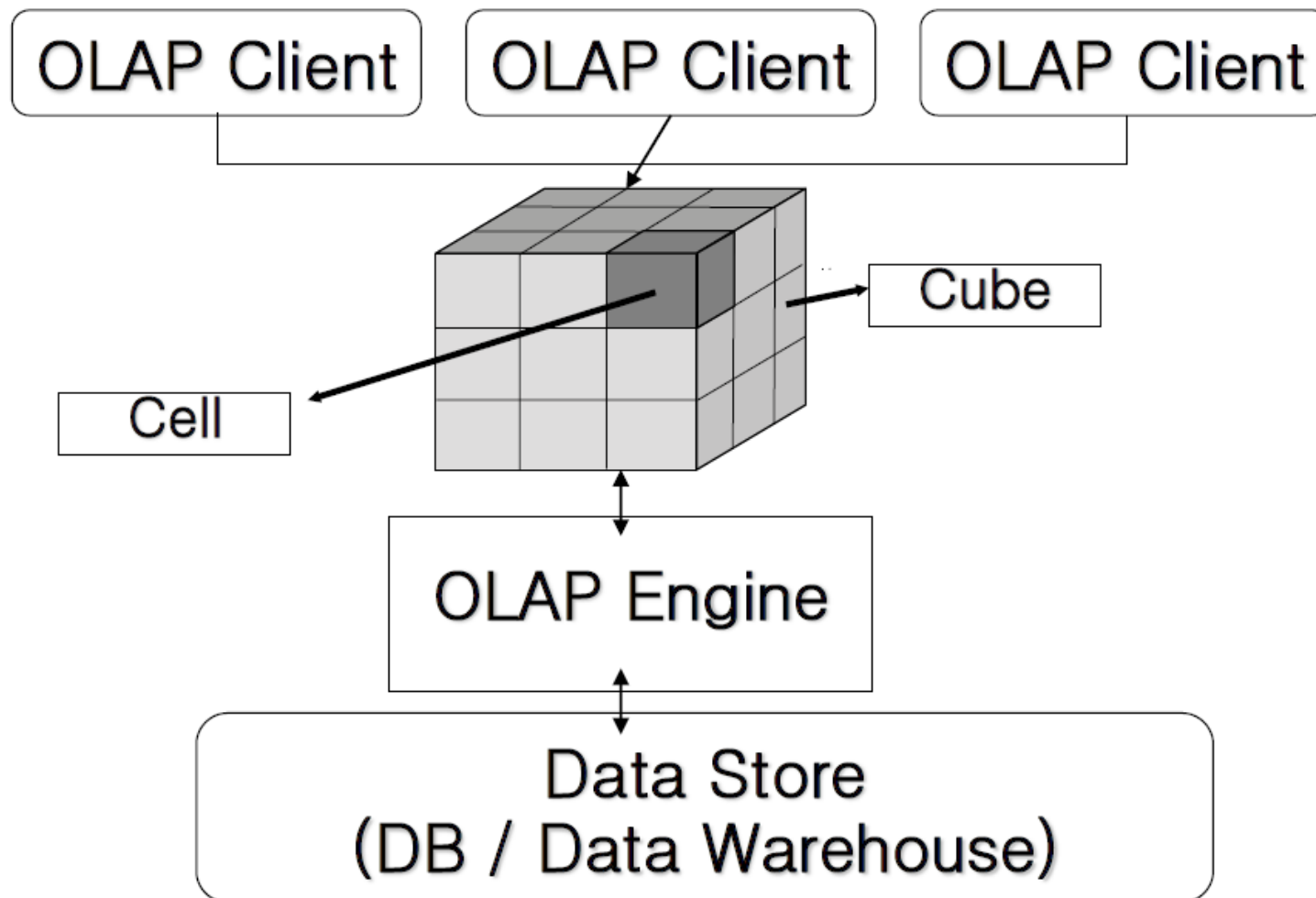
- 데이터웨어하우스에는 단순히 자료가 저장되어 있을뿐만 아니라, 이러한 자료를 추출, 저장, 분류하는 일련의 과정을 포함
- 메타 데이터
 - 데이터의 데이터
 - 데이터웨어하우스의 생성과 유지보수에 관련된 정보를 담고있는 자료
- 데이터 마트
 - 데이터 웨어하우스에 저장된 자료 중에서 일정한 주제나 특정 부서의 자료를 별도의 장소에 중복 저장하여 사용자들이 사용하도록 하게 한 것

OLAP

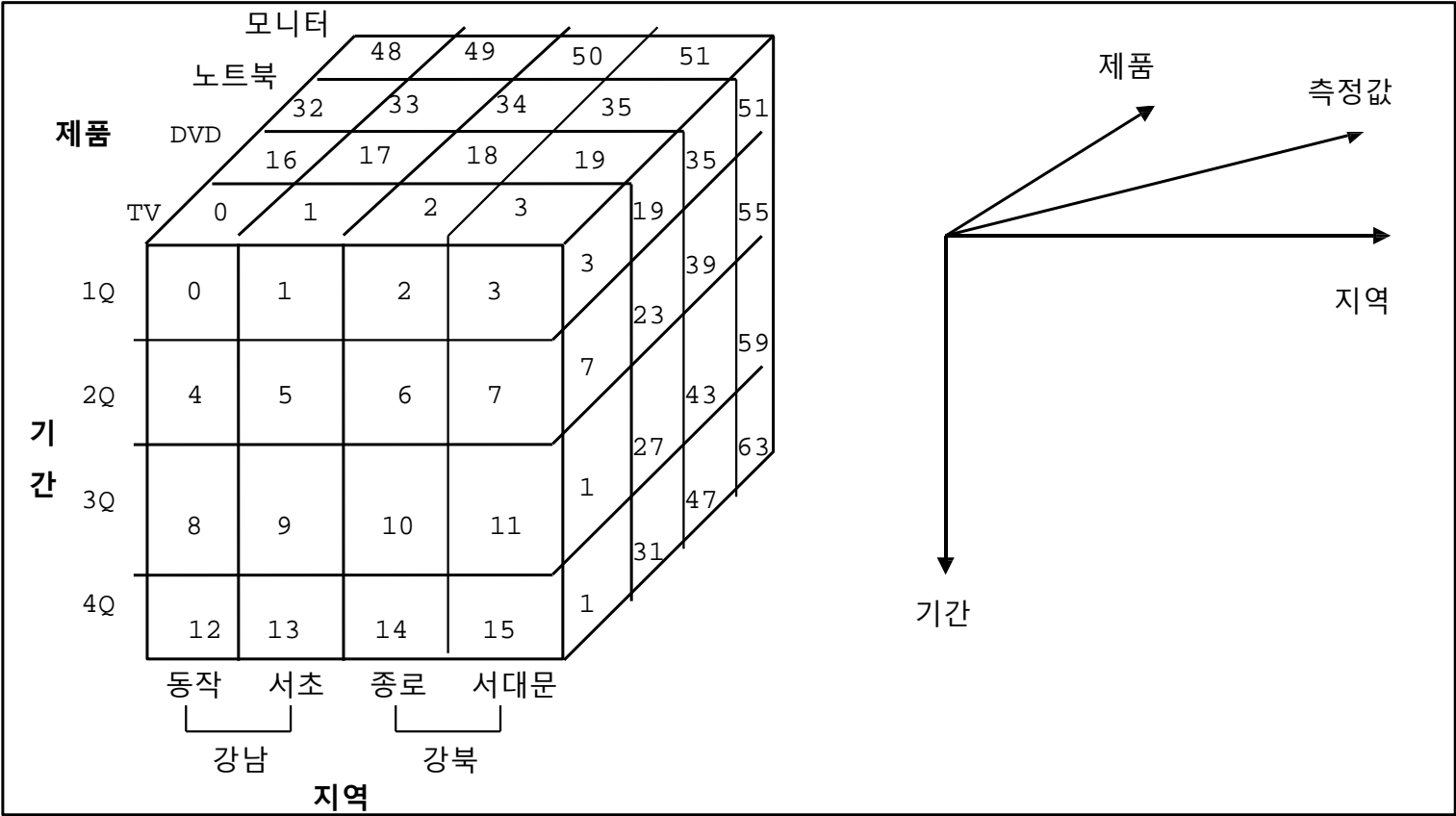
온라인 분석처리(OLAP)란?

- 다차원 데이터 구조를 이용하여 다차원의 복잡한 질의를 고속으로 처리하는 데이터 분석 기술
- 차원(dimensions)과 측정 항목(measure)을 설정하여 관심주제에 대한 분석을 수행
(예: "지역별/분기별/상품별(차원들)" "판매액(측정항목)" 현황분석)
- 일반적으로 최종 사용자가 필요한 정보를 자료원으로부터 직접 가공하여 분석
- 사용자가 분석도중 대화식으로 여러 차원 또는 분석 기법간에 심층분석(Drill-down), 또는 축약분석(Drill-up) 가능

온라인 분석처리의 기본구조



■ 큐브와 셀



■ 드릴 다운 (Drill down)

- 데이터를 어떤 하나의 차원을 기준으로 분석할 때 계층구조상의 가장 상위 수준에 해당하는 집계 데이터부터 먼저 보고, 다음 세부 수준으로 들어가며 데이터를 분석하는 것
- 예를 들어, 년도별 분석 -> 반기별 분석 -> 분기별 분석 -> 월별 분석과 같이 점점 상세 수준으로 데이터를 분석

■ 드릴 업 (Drill up) = 롤 업 (Roll-up)

- 드릴다운의 반대 과정으로, 상세 수준의 데이터로부터 차츰 상위 단계의 데이터를 분석해 가는 것
- 월별 분석 -> 분기별 분석 -> 반기별 분석 -> 년도별 분석으로 분석의 범위를 넓혀 나가면서 데이터를 분석

■ 피벗(pivot)

- 임의의 다차원 뷰(view)를 만들고 검토 및 분석한 후,
차원들의 다양하게 변경하여 새로운 뷰를 만들어 데이터를 분석하는 것
- 다양한 뷰를 만들기 위해 축과 축을 바꾸는 작업

- 다차원화 된 쿼리
- 설정한 디멘션에 따라 측정값(요약값) 산출해서 검토
→ 의사결정에 도움을 받음

비정형 데이터베이스

- 고정된 필드에 저장된 데이터
- 데이터베이스를 설계한 사람에 의해 수집되는 정보의 형태가 정해짐
 - 예를 들어 관계형 데이터 베이스의 테이블들, 스프레드시트 등

	A	B	C	D	E
1	교수번호	성명	전공	소속	내선번호
2	130021	조상진	미생물	서울대	7588
3	132003	김현준	컴퓨터	강원대	1455
4	132002	윤선영	경영	안산대	6585
5	132023	박정미	화학	명인대	2565
6	132004	김민주	컴퓨터	서울대	7855
7	132056	서성현	철학	서운대	9654

<정형 데이터 예시 - 스프레드 시트>

- 미리 정해져서 고정되어 있는 필드에 저장되어 있지 않은 데이터
- 스마트 기기에서 페이스북, 트위터, 유튜브 등으로 생성되는 소셜 데이터
- IoT 환경에서 생성되는 위치 정보나 센서 데이터와 같은 사물 데이터 등
- 문서, 그림, 영상 등이 이에 해당



트위터



인스타그램



유튜브

빅데이터의 유형

■ 정형 데이터(structured data)

- 정해진 형식과 구조에 따라 고정된 필드에 저장되도록 구성된 데이터
- 예) '인사' 테이블이나 '매출' 스프레드시트와 같이 통일된 형식
- 정형화된 형식과 저장 구조를 가지므로 검색, 변경 등의 연산이 쉽다.

■ 반정형 데이터(semi-structured data)

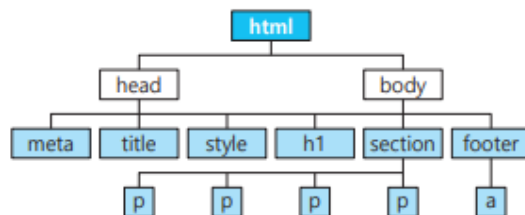
- 고정된 필드로 저장되지만 데이터의 형식과 구조가 변경될 수 있는 데이터
- 보통 구조 정보를 데이터와 함께 제공하는 XML, JSON 등의 파일 형식의 데이터
- 메타 데이터를 포함하고 있어 파싱을 통해 정형 데이터로의 변환이 가능

■ 비정형 데이터(unstructured data)

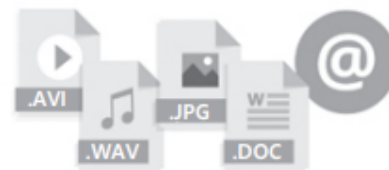
- 음성이나 동영상, 블로그, SNS 메시지 등과 같이 정해진 형식이나 구조가 없는 정형화되지 않은 데이터
- 대부분 크기와 형식을 미리 정의하기 어려운 데이터들을 포함
- 최근 비중이 가장 빠르게 증가하는 데이터 유형, 잠재 가치가 가장 높음

번호	이름	나이	성별
s001	홍길동	25	남
s002	김나리	22	여
s003	이승훈	27	남

정형 데이터

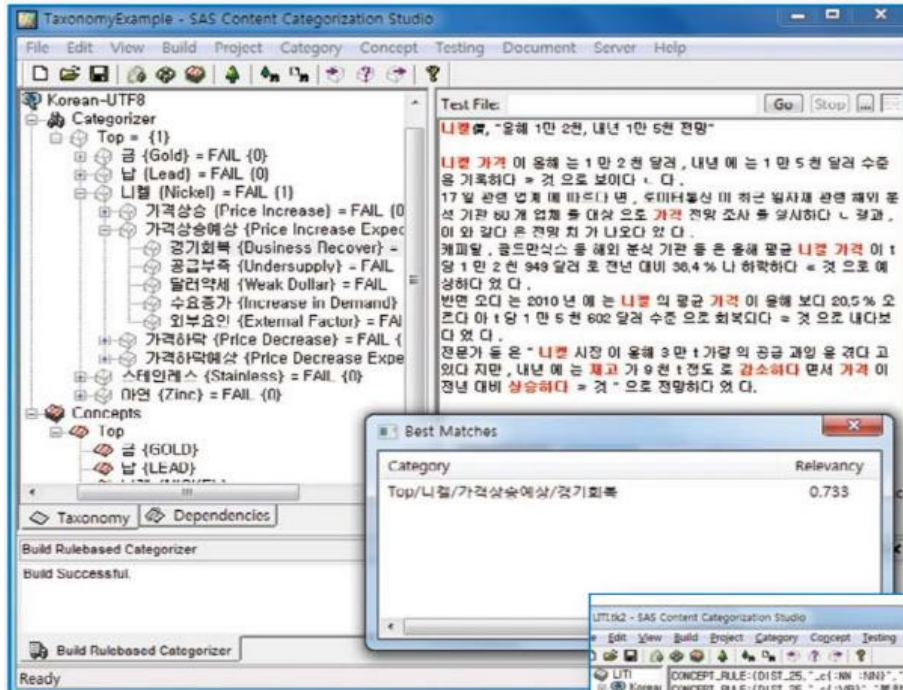


반정형 데이터



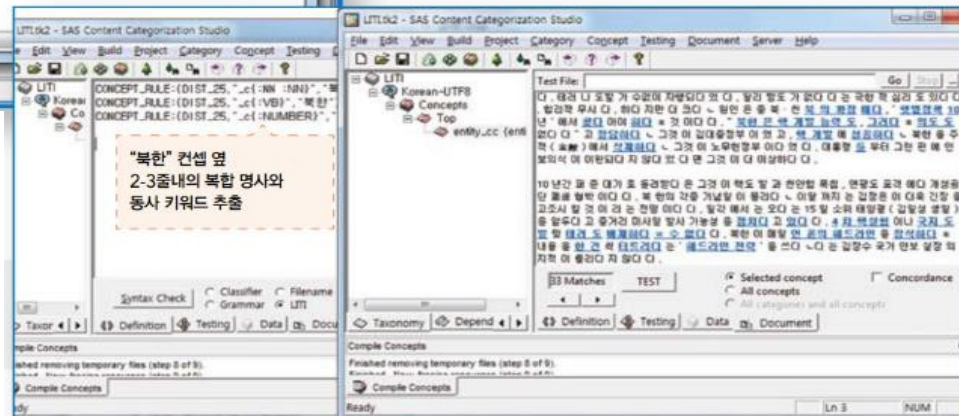
비정형 데이터

분석대상 데이터의 예: 일반문서



■ 텍스트 의미 기반의 정확한 분류

■ 연관 키워드 추출

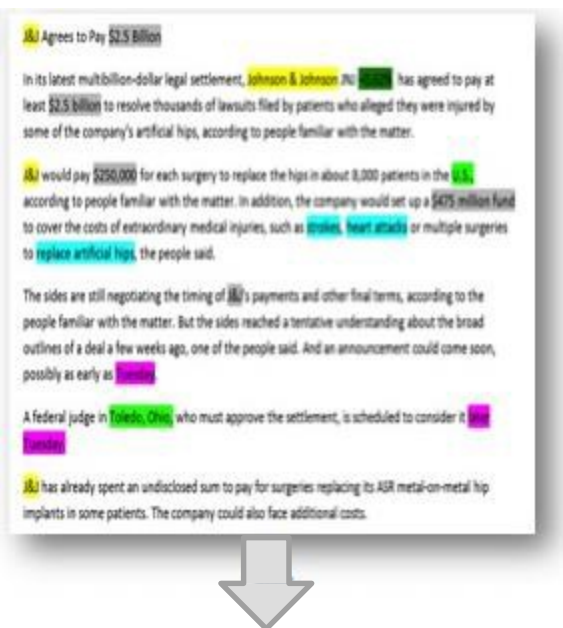


분석대상 데이터의 예: 웹

TaggingText



HTML to Text



Now, Structured Data For Analysis

	A	B	C	D	E	F	G
1	DocMasterID	HarvestDate	Domain	Companies	Places	Money	DocumentText
2		1	11/19/2013 14:37	wsj.com	[United States, Toledo, OH, New Brunswick, N.J.]	[\$2.5 billion, \$250,000, \$475 Million]	in its latest multibillion-dollar legal settlement, Johnson & Johnson (JNJ) +0.62% has agreed to pay at least \$2.5 billion to resolve thousands of lawsuits filed by

분석대상 데이터의 예: SNS

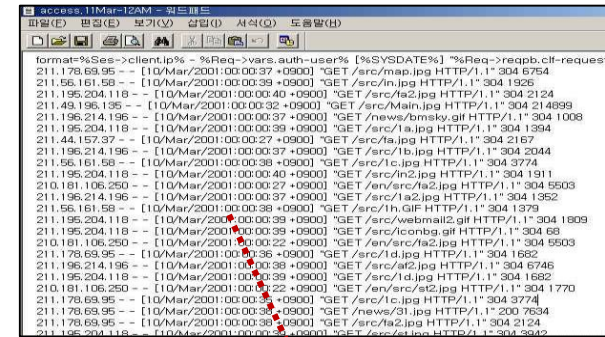


	A	B	C	D	E	F	G	H	I	J	K
1		from_id	from_name	message	created_time	type	link	id	likes_count	comments_count	shares_count
2	1	2.4026e+14	JTBC 뉴스	국속방치된 20	2016-03-25T06:00:00+0000	link	http://new.240263402699918.1023288141064103	11	1	1	
3	2	2.4026e+14	JTBC 뉴스	출근을 늦추거나 퇴근을 당기는 등	2016-03-25T04:36:31+0000	link	http://new.240263402699918.1023261664400084	126	7	14	
4	3	2.4026e+14	JTBC 뉴스	JTBC 뉴스 x Facebook	2016-03-25T03:34:41+0000	video	https://www.240263402699918.1023207377738846	151	2	38	
5	4	2.4026e+14	JTBC 뉴스	JTBC 뉴스 x Facebook	2016-03-25T03:04:34+0000	photo	https://www.240263402699918.1023189777746606	582	11	26	
6	5	2.4026e+14	JTBC 뉴스	NA	2016-03-25T03:00:15+0000	photo	https://www.240263402699918.1023189821073935	49	0	1	
7	6	2.4026e+14	JTBC 뉴스	"형사의 초치는 부당하지 않다"	2016-03-25T02:30:00+0000	video	https://www.240263402699918.1023064791086438	145	9	39	
8	7	2.4026e+14	JTBC 뉴스	손대면 '죽'하고	2016-03-25T00:52:53+0000	link	http://new.240263402699918.1023045144421736	17	0	1	
9	8	2.4026e+14	JTBC 뉴스	#부산국제영화제의 연금 강동이 파국으로 치닫고	2016-03-25T00:30:00+0000	video	https://www.240263402699918.1022506234475627	358	5	38	
10	9	2.4026e+14	JTBC 뉴스	중국계 박동의 끝은 어디일까요	2016-03-24T23:30:00+0000	link	http://new.240263402699918.10224292724483278	74	3	9	
11	10	2.4026e+14	JTBC 뉴스	3월 25일 아침& 주요뉴스입니다.	2016-03-24T22:49:21+0000	video	https://www.240263402699918.102239274432323	36	3	2	
12	11	2.4026e+14	JTBC 뉴스	없었다는 연좌제는	2016-03-24T22:00:00+0000	video	https://www.240263402699918.1022457567813827	184	3	35	
13	12	2.4026e+14	JTBC 뉴스	일찍일찍! 당신의 #할 일 색칠해봐요~	2016-03-24T21:30:00+0000	video	https://www.240263402699918.1022399237819660	66	3	15	
14	13	2.4026e+14	JTBC 뉴스	가정폭력이나 성폭력 피해를 당한 여성들을 상담해	2016-03-24T21:00:00+0000	video	https://www.240263402699918.1022509037808680	118	9	11	
15	14	2.4026e+14	JTBC 뉴스	열~~~~	2016-03-24T15:26:28+0000	video	https://www.240263402699918.1022578831135034	1491	74	0	
16	15	2.4026e+14	JTBC 뉴스	성공한 한국 영화들의 한 가지 공통점 꼽으라!	2016-03-24T14:30:00+0000	photo	https://www.240263402699918.1022407431152174	798	21	28	
17	16	2.4026e+14	JTBC 뉴스	2009년 일 해리 #수입 중고기 자의 #수입비 가 2000	2016-03-24T14:20:54+0000	photo	https://www.240263402699918.1022487097810874	361	42	68	
18	17	2.4026e+14	JTBC 뉴스	신앙성 한영회에서 새내기 대학생이 괴도한 음주로	2016-03-24T14:10:36+0000	video	https://www.240263402699918.1022486094471641	286	82	37	
19	18	2.4026e+14	JTBC 뉴스	"문제는 경제"	2016-03-24T13:00:01+0000	link	http://new.240263402699918.1022428994483351	107	3	17	
20	19	2.4026e+14	JTBC 뉴스	집권 여당인 새누리당의 공천 갈등은	2016-03-24T12:37:20+0000	link	http://new.240263402699918.1022437724482478	50	3	9	
21	20	2.4026e+14	JTBC 뉴스	"아린양자" 이야기들 다시 한번 펼쳐봅니다.	2016-03-24T12:21:29+0000	video	https://www.240263402699918.1022418637817720	674	14	229	
22	21	2.4026e+14	JTBC 뉴스	● 김우성, 유승민 지역 등 5곳 공천 거부	2016-03-24T11:10:58+0000	video	https://www.240263402699918.1022386194487631	45	3	5	
23	22	2.4026e+14	JTBC 뉴스	베이징의 황태자 시한폭탄 하루 아침에 시골로 쫓	2016-03-24T10:39:52+0000	video	https://www.240263402699918.1022366444489606	128	6	11	
24	23	2.4026e+14	JTBC 뉴스	"강릉을 놀라게 했는데, 결국 23시간 끝났어요"	2016-03-24T10:00:00+0000	link	http://new.240263402699918.1022296347829949	149	6	18	
25	24	2.4026e+14	JTBC 뉴스	이명박만 '고쳐줄' 파급과 영향	2016-03-24T09:30:00+0000	video	https://www.240263402699918.1022220841710833	842	49	38	
26	25	2.4026e+14	JTBC 뉴스	오늘(24일)의 한미대는 '참재'입니다.	2016-03-24T09:01:04+0000	video	https://www.240263402699918.1022320341160803	46	4	3	
27	26	2.4026e+14	JTBC 뉴스	장원준의 별(가)구	2016-03-24T08:42:31+0000	video	https://www.240263402699918.1022307651162152	401	23	21	
28	27	2.4026e+14	JTBC 뉴스	"간담하고 백백하게 자라라"	2016-03-24T08:30:00+0000	video	https://www.240263402699918.1022285121164405	521	15	19	
29	28	2.4026e+14	JTBC 뉴스	김 대표는 후보등록 마감일까지 버티자 진박우보	2016-03-24T07:58:49+0000	photo	https://www.240263402699918.1022282324497928	50	3	0	
30	29	2.4026e+14	JTBC 뉴스	기자회견 김우성 대표	2016-03-24T07:15:20+0000	video	https://www.240263402699918.1022234494502801	138	35	18	
31	30	2.4026e+14	JTBC 뉴스	기자회견 이원구 공천관리위원장	2016-03-24T06:40:48+0000	video	https://www.240263402699918.102223574502673	65	93	6	
32	31	2.4026e+14	JTBC 뉴스	그들을 쫓습니다.	2016-03-24T06:00:00+0000	photo	https://www.240263402699918.1022212797838304	661	20	104	
33	32	2.4026e+14	JTBC 뉴스	발이 폭발해 눈부신 성광을 내는 진귀한 장면이 공개	2016-03-24T05:00:00+0000	video	https://www.240263402699918.1022175864508664	507	9	113	
34	33	2.4026e+14	JTBC 뉴스	하지만 우리나라는 부도도 없습니다.	2016-03-24T04:30:00+0000	link	http://new.240263402699918.1022021417857442	49	3	12	
35	34	2.4026e+14	JTBC 뉴스	그들이 몰려온다	2016-03-24T03:00:00+0000	photo	https://www.240263402699918.1022053618187381	97	5	12	
36	35	2.4026e+14	JTBC 뉴스	3월 24일 뉴스브리핑 주요뉴스입니다.	2016-03-24T02:42:21+0000	video	https://www.240263402699918.1022087391818478	62	3	1	
37	36	2.4026e+14	JTBC 뉴스	경찰은 승원 장 배의 음주운전 여부 등 정확한 사고	2016-03-24T01:47:36+0000	link	https://www.240263402699918.1022046767854907	108	19	8	
38	37	2.4026e+14	JTBC 뉴스	군 고위 관계자들이 압제와 갈취에 벌어진 일입니다	2016-03-24T00:29:41+0000	link	http://new.240263402699918.1021967367862847	853	116	122	
39	38	2.4026e+14	JTBC 뉴스	"국내에서 유령할 가능성 희박"	2016-03-23T23:30:00+0000	link	http://new.240263402699918.1021448791248038	47	9	3	
40	39	2.4026e+14	JTBC 뉴스	3월 24일 아침& 주요뉴스입니다.	2016-03-23T22:50:51+0000	video	https://www.240263402699918.1021888491204068	57	2	4	
41	40	2.4026e+14	JTBC 뉴스	#오늘 '세계 일대의 날'	2016-03-23T22:00:52+0000	video	https://www.240263402699918.1021361991256718	159	6	18	
42	41	2.4026e+14	JTBC 뉴스	네티즌들이 현글처럼 주는 이른바 '별풍선'을 받으려	2016-03-23T21:30:12+0000	video	https://www.240263402699918.1021445991248318	198	34	20	
43	42	2.4026e+14	JTBC 뉴스	최 위지조르게 보시자. 대장, 이 전투에 자히. 그제, 사 2016-03-23T21:00:00+0000	video	https://www.240263402699918.102134182141684430	166	10	10	13	

페이스북의 게시물 크롤링 결과

분석대상 데이터의 예: 로그 데이터

- 로그 데이터 수집이란 사용자가 처음 사이트를 방문하는 순간부터 각 웹 페이지를 액세스할 때마다 기록되는 액세스 로그 데이터를 목표 시스템에 저장하는 것을 의미함
- 사용자의 IP주소와 액세스한 파일, 액세스한 시간 등의 정보는 물론 사용자가 요청하는 해당 웹 페이지와 관련된 이미지 파일, 이미지 데이터 등의 모든 연관 파일에 대한 정보가 수집됨



- 211.178.69.95 - - [10/Mar/2001:00:00:38 +0900] "GET /news/31.jpg HTTP/1.1" 200 7634
- ① 누가: : IP, ID
 - ② 언제 : 방문 일자/시간
 - ③ 무엇을 : 방문 페이지
 - ④ 어떻게 : 응답 상태
 - ⑤ 어디로부터: 참조한 페이지

- 센서가 부착된 장치로부터 생성된 데이터를 수집하는 것을 의미함
- 거주 환경에 따른 건강 모니터링을 위한 센서 데이터 수집 예시

환경정보

건강정보

조명 제어 정보	
침실 조명 :	OFF
거실 조명 :	OFF
화장실 조명 :	OFF
주방 조명 :	OFF
가전 제어 정보	
선풍기 :	OFF
가스밸브 :	CLOSE
현관문 :	OPEN
환경 정보	
온도 :	21.70
습도 :	36.200

관상정보

최고혈압 (mmHg)

최저혈압 (mmHg)

정상

맥 박 (회/분)

정상

혈 당 (mg/dl)

체지방률 (%)

배만

체 질 량 (kg)

정상

관장체중 (kg)

[illegible]

CURRENT_TIME_2	SENSORNODE	TEMPERATUR	HUMIDIT
2008-04-29 오후 8:47:40	1	26.6	49.9
2008-04-29 오후 8:48:21	1	26.6	49.9
2008-04-29 오후 8:48:39	1	26.6	49.9
2008-04-29 오후 8:53:06	1	26.6	49.9
2008-04-29 오후 8:53:37	1	26.6	49.9
2008-04-29 오후 8:54:04	1	26.6	49.9

CURRENT_TIME_2	PERSON_ID	SENSOR_ID	BLOOD_PRESSURE_MIN	BLOOD_PRESSURE_MAX
2008-04-01 오후 4:15:02	2	Health	47	11
2008-04-01 오후 4:15:54	2	Health	59	12
2008-04-01 오후 4:18:12	3	Health	64	10
2008-04-01 오후 5:21:07	2	Health	74	11
2008-04-02 오후 3:52:12	2	Health	73	11

센서 데이터 생성

센서 데이터 수집

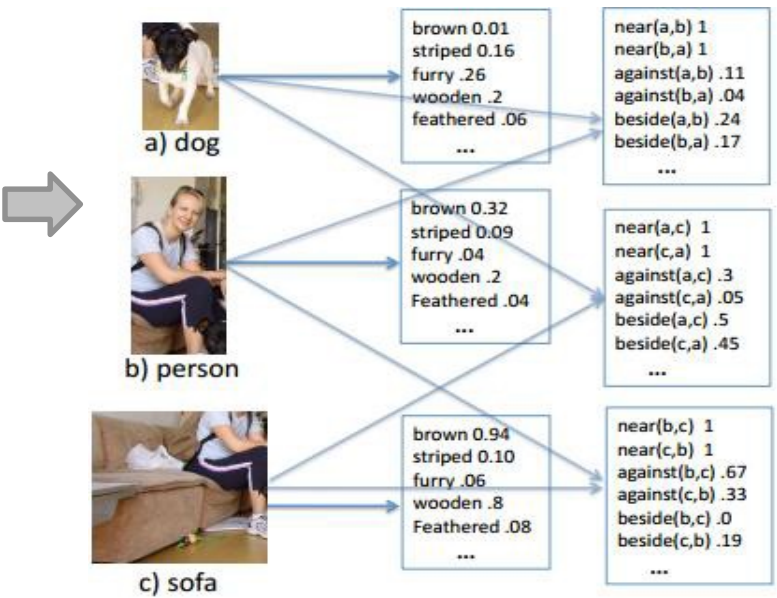
센서 데이터 저장

출처: 윤영민 외(2008), 상황정보 서비스를 위한 이기종 센서정보 관리, 한국인터넷정보학회 춘계학술대회

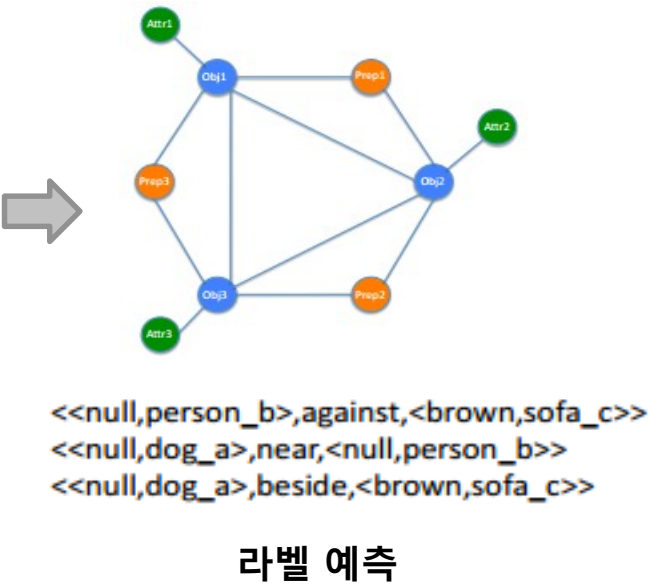
분석대상 데이터의 예: 이미지



이미지 입력



물체의 특성 인식 및 변환



분석대상 데이터의 예: 이미지



This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.



There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.



There are one cow and one sky. The golden cow is by the blue sky.



There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.



Here we see one person and one train. The black person is by the train.



This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.



Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.



Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.



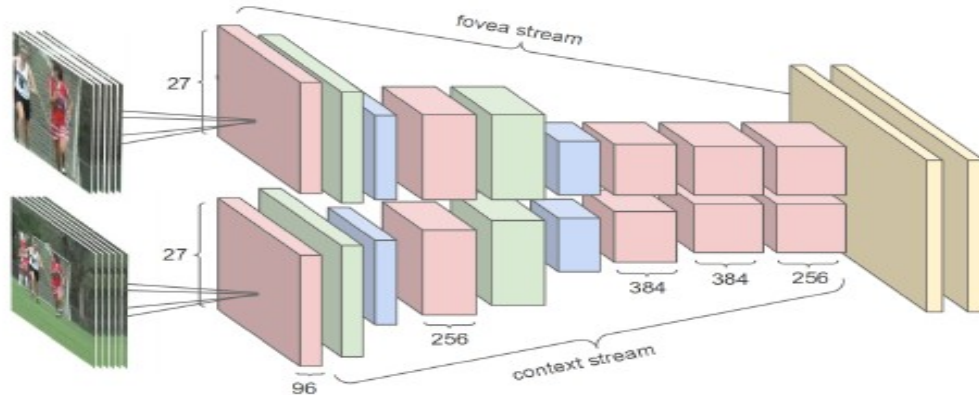
This is a picture of two dogs. The first dog is near the second furry dog.



This is a photograph of two buses. The first rectangular bus is near the second rectangular bus.

이미지에 대한 문장 생성

분석대상 데이터의 예: 동영상



track cycling
cycling
track cycling
road bicycle racing
marathon
ultramarathon



ultramarathon
ultramarathon
half marathon
running
marathon
inline speed skating

<알고리즘을 통한 영상 분류>

- 빅데이터 시대의 주요 특징
 - 데이터의 크기가 엄청나게 크다
 - 데이터의 형태가 비정형적이고, 다양하다
 - 빠른 처리속도가 요구된다
- 기존의 RDB나 DW와 같은 정형 데이터 베이스만으로는 해결이 어려움
- 분산 시스템으로 갈 수 밖에 없다
- 분산환경 하에서 대용량의 데이터를 신속하게 처리할 비정형 데이터베이스의 등장
 - NoSQL, Hadoop 등

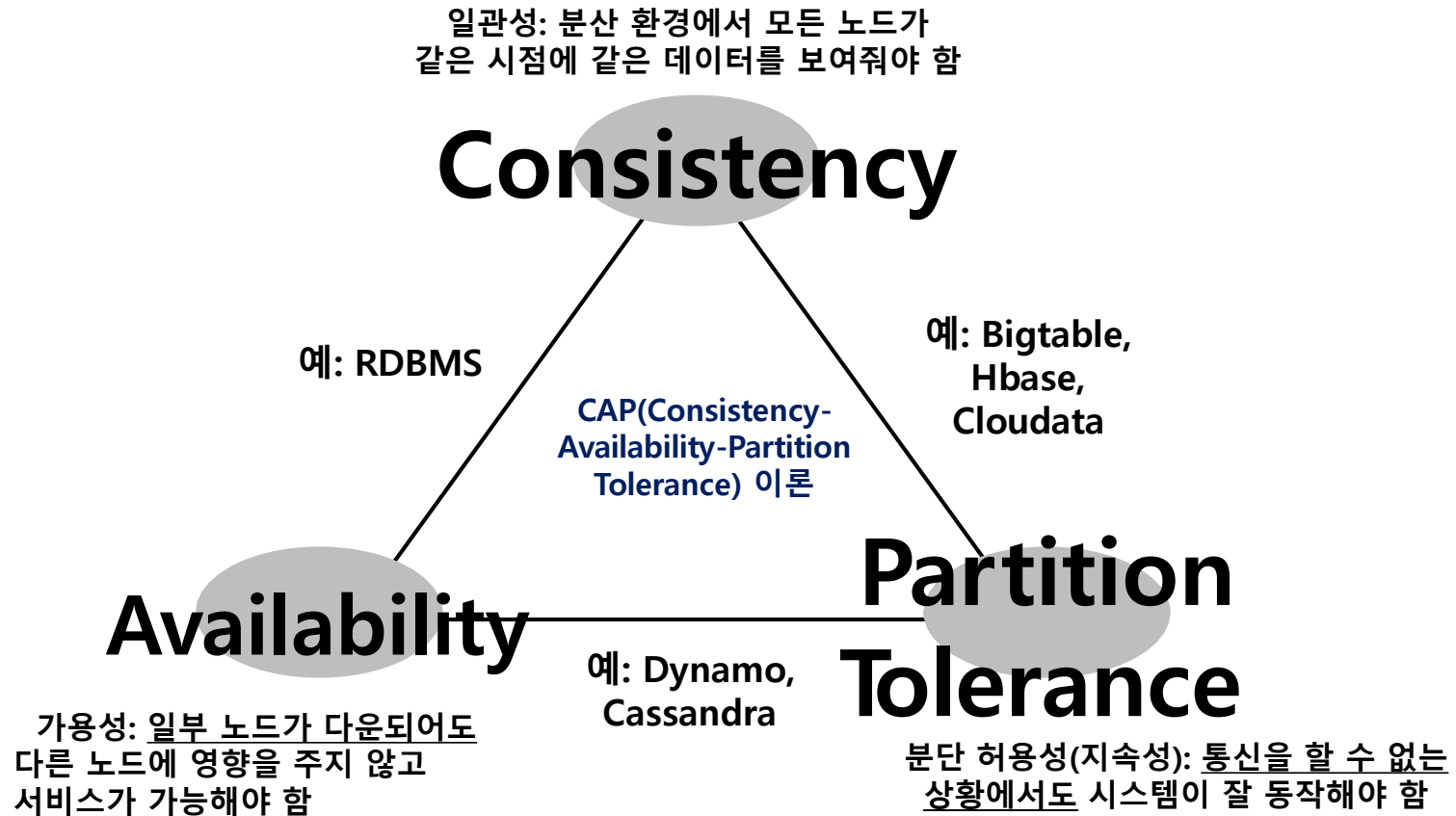
분산 시스템이란?

- 작업이나 데이터를
- 여러 대의 컴퓨터에 나누어서
- 처리, 저장하여
- 그 내용이나 결과가 통신망을 통해 상호교환 되도록 연결되어 있는 시스템

CAP 이론이란?

- 2000년경 버클리 대학 에릭 브루어(Eric Brewer)가 주창한 이론으로, 2002년경 증명
- CAP 정리(CAP Theorem), 혹은 브루어 정리(Brewer's theorem) 등으로 불리움
- 분산 컴퓨터 시스템에서, CAP라고 하는 세 가지 특성을 동시에 충족시키는 것은 불가능하다

CAP 이론이란?



Consistency (일관성)

- 분산 환경에서 모든 노드가 같은 시점에 같은 데이터를 보여 줘야 함
- 즉, 한쪽이 업데이트 되면 즉시 같은 정보가 보여져야 한다

Availability (가용성)

- 일부 노드가 다운되어도 다른 노드에 영향을 주지 않아야 함
- "특정 노드가 장애가나도 서비스가 가능해야 한다"라는 의미

Partition Tolerance (분단 허용성, 지속성)

- 클러스터 사이에 접속이 단절되어 서로 통신을 할 수 없는 상황에서도 시스템이 잘 동작해야 한다.
- 노드(컴퓨터) 간에 통신 문제로 인해 일부 데이터가 손실되더라도 시스템은 정상적으로 동작해야 한다.

- 분산시스템이 이 세가지를 모두 만족시키는 것은 불가능하다
- 두 가지 만족이 가능하다
- C-A C-P A-P
- 최근 CAP 이론이 주목받는 이유
 - 빅데이터 저장문제

CAP 이론과 비정형 데이터베이스

- 기존의 RDB가 C(일관성)와 A(가용성) 중심이라면
- 비정형 데이터베이스는 P(지속성)을 중시

- Not-Only SQL 혹은 No SQL을 의미
- 전통적인 관계형 데이터베이스(RDBMS)와 다르게 설계된 비관계형 데이터베이스
- 스키마가 없는 데이터베이스
- 덜 제한적인 데이터 저장 및 검색 메커니즘 제공

- 용이한 데이터 규모 확장성 – 저가 서버 사용
 - 즉 데이터를 다수의 하드웨어에 분산해서 저장
 - 대용량의 구조적, 반구조적 데이터들을 저장/분석(웹, 소셜 미디어, 그래픽 등)
- ✓ 유연성, 확장성, 경제성, 가용성

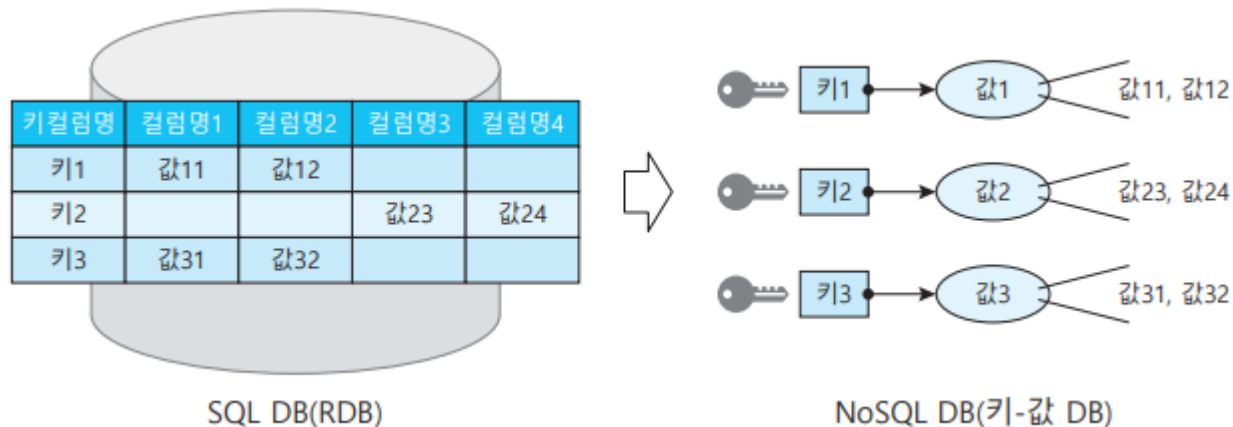
NoSQL의 분류: 데이터 저장 방식에 따라

- 키-값: 다이나모, 리악, 레디스, 캐시, 프로젝트 볼드모트
- 컬럼: H베이스, 아큐물로, 카산드라
- 도큐먼트: 몽고DB, 카우치베이스
- 그래프: Neo4J, 알레그로그래프, 버투오소

NoSQL 데이터베이스의 유형

1) 키-값 데이터베이스(key-value database)

- NoSQL 데이터베이스 중 가장 단순하고 기본적인 형태
- 모든 데이터를 '키'와 '값'의 쌍으로 매핑하여 저장
- 조회를 위한 유일한 키와 하나의 데이터 값을 매핑해 정해진 스키마없이 저장하는 방식
- '키(key)'는 속성 이름을, '값(value)'은 속성에 연결된 데이터 값을 의미
- 예) 수하물 태그는 '키'에, 수하물 짐은 '값'에 비유
- 장점 : 데이터 분할이 가능, 다른 데이터베이스로는 불가능한 수준까지 수평 확장이 가능
- 단점 : 특정값 검색에는 효율적이지만 데이터 정렬, 그룹화, 범위 검색 등이 어려움



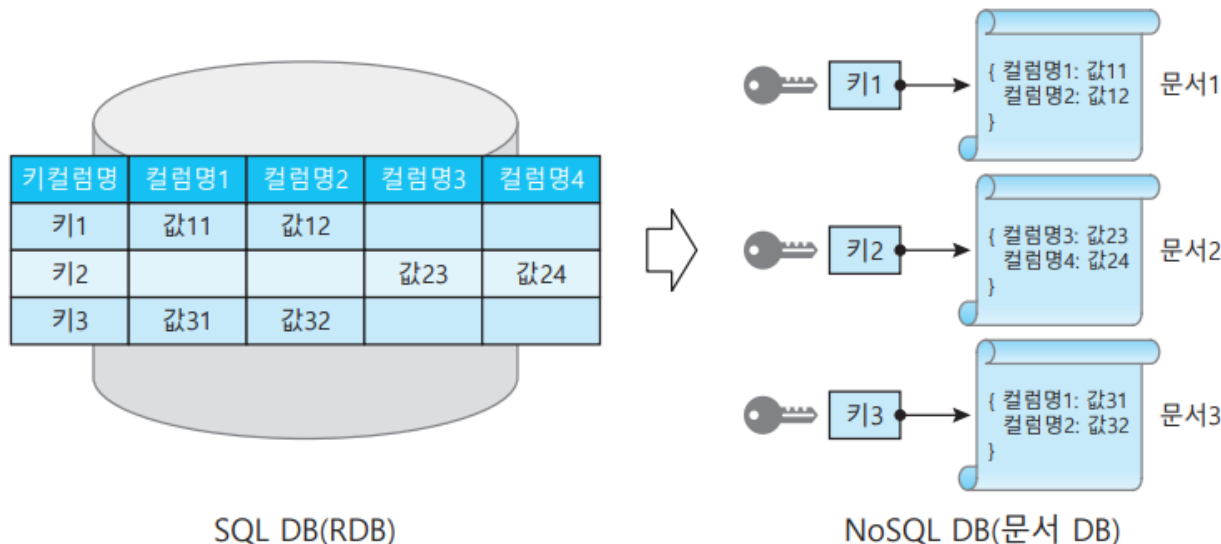
NoSQL의 분류: 데이터 저장 방식에 따라(키-값)

데이터 모델		설명	제품 예
<키, 값> 저장 구조			

NoSQL 데이터베이스의 유형

2) 문서 데이터베이스(document database)

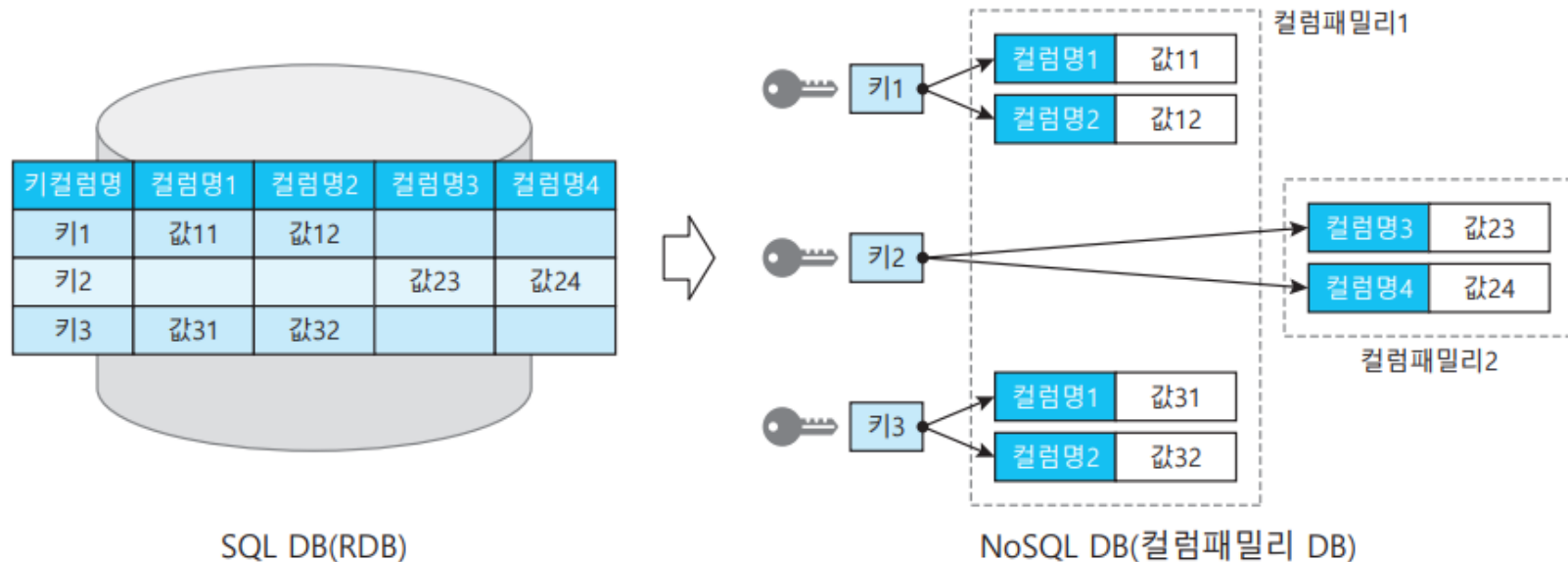
- 키-값 데이터베이스의 발전된 형태
- 반구조적 데이터의 저장과 검색에 사용
- '키-문서' 데이터베이스 형태(키에 대응하는 값이 문서)로 키에 대응하는 각각의 문서는 속성과 속성에 대응하는 데이터를 가짐
- 문서는 반구조화된(semi-structured) 데이터 형태로 계층적 구조를 가지며 객체와 유사하게 하나의 단위로 취급(JSON이나 XML 문서 등)
- 스키마가 자주 바뀌거나 저장할 데이터가 복잡한 계층 구조를 가질 경우 또 문서간의 비교보다는 문서 자체의 검색과 변경이 대부분일 경우 효과적
- NoSQL 데이터베이스 중에서 가장 인기있는 유형



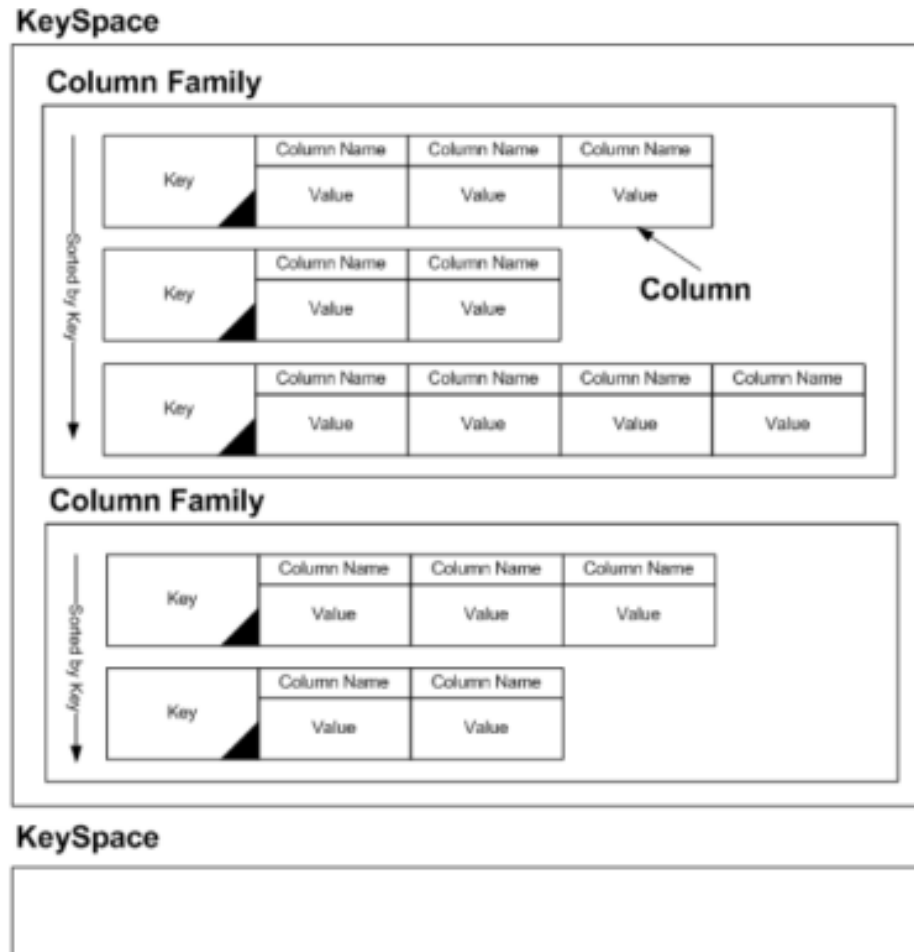
NoSQL 데이터베이스의 유형

3) 컬럼 패밀리 데이터베이스(column family database)

- 구조면에서 가장 복잡한 유형, 관계형 데이터베이스와 비슷함
- 기본 저장 단위는 컬럼으로 키에 대응되는 '(컬럼명, 값)' 조합으로 된 여러 필드를 갖음
- 컬럼 단위로 묶어서 저장하는 컬럼-지향(column-oriented) 데이터베이스
- 하나의 행에 많은 컬럼을 포함할 수 있어서 유연성이 높고 분산 저장과 확장이 가능하여 빅데이터에 유리한 구조
- 연관된 컬럼끼리 묶은 컬럼의 그룹인 '컬럼 패밀리'가 모여서 하나의 객체를 표현
- 미리 정의된 고정 스키마를 사용하지 않으므로 얼마든지 컬럼을 추가할 수 있음



NoSQL의 분류: 데이터 저장 방식에 따라(컬럼)

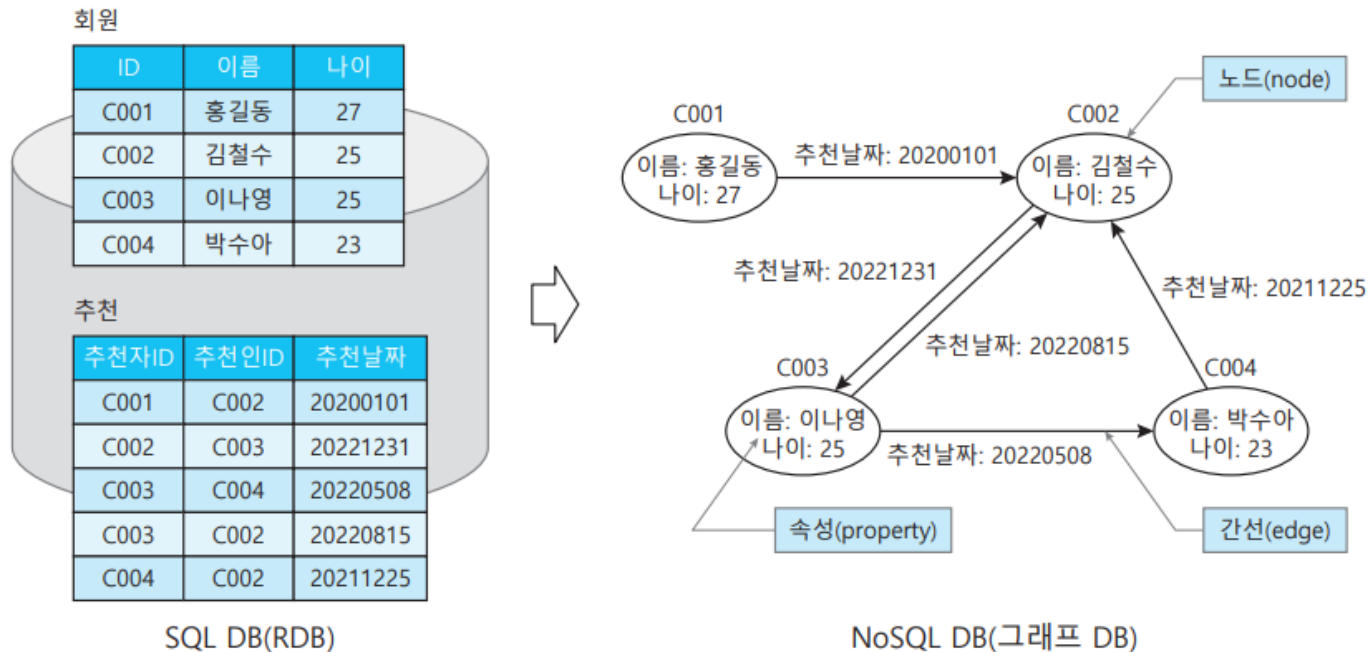


CassandraDB

NoSQL 데이터베이스의 유형

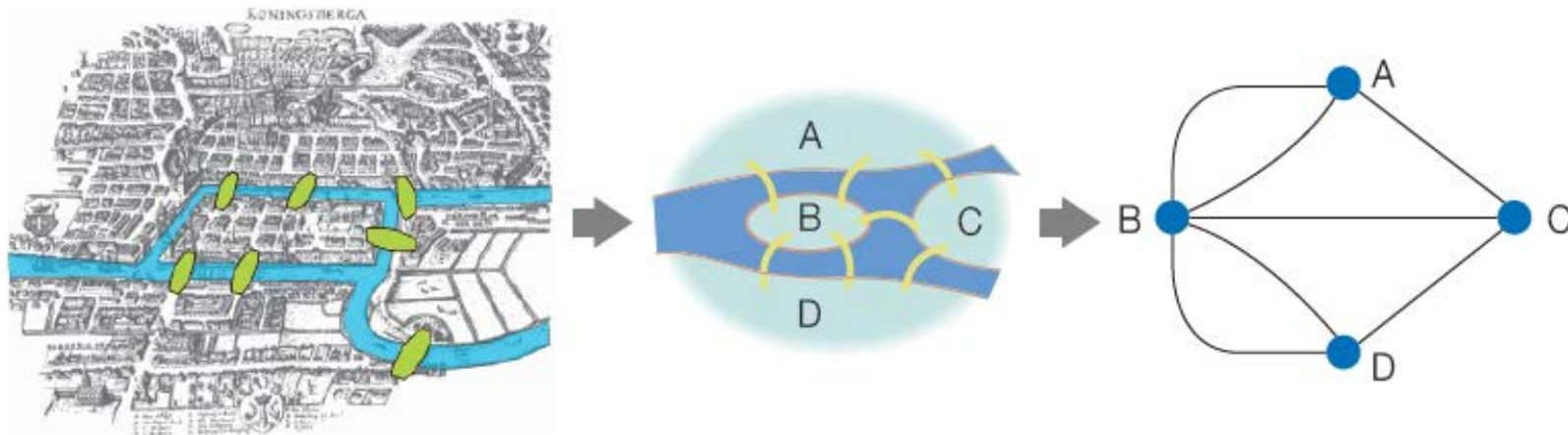
4) 그래프 데이터베이스(graph database)

- 데이터를 데이터 간의 관계와 함께 표현하는 특수한 유형
- 소셜미디어, 교통망, 전력망처럼 많은 객체 간의 연결을 표현하는데 적합
- 객체를 표현하는 '노드'(node)와 두 객체 사이의 연결 관계를 보여주는 '링크'(link) 두 요소를 사용
- 노드에는 식별자와 객체 속성을, 링크에는 노드간의 연관 속성을 저장



■ 그래프 이론

- 점(vertex, node 또는 point)과 점들을 이은 가지(edge, arc 또는 line)로 표현
- 1736년, 수학자인 오일러가 당시 프로이센 쾨니히스베르크(지금의 러시아 칼라닌그라드)에 있는 7개의 모든 다리를 한 번씩만 건너 제자리로 돌아올 수 있는가에 대한 '쾨니히스베르크 다리 문제'를 푸는 데서 시작
- 오일러는 홀수점의 개수가 0이거나 단 두 개의 홀수점을 가지는 경우에 한붓그리기가 가능함을 증명



쾨니히스베르크 다리 문제와 그래프

■ 無 스키마

- 고정된 스키마 없이 키(Key) 값을 이용하여 다양한 형태의 데이터 저장 및 접근 기능
- 데이터 저장 방식은 크게 값(Value), 열(Column), 문서(Document), 그래프(Graph) 등의 네 가지를 기반으로 구분

■ 탄력성(Elasticity)

- 시스템 일부에 장애가 발생해도 클라이언트가 시스템에 접근 가능
- 응용 시스템의 다운 타임이 없도록 하는 동시에 대용량 데이터의 생성 및 갱신
- 시스템 규모와 성능 확장이 용이하며, 입출력의 부하를 분산시키는 데 용이한 구조

■ 쿼리(Query) 기능

- 수십 대에서 수천 대 규모로 구성된 시스템에서도 데이터의 특성에 맞게 효율적으로 데이터를 검색 · 처리 가능

RDBMS vs. NoSQL

- RDBMS는 대용량 데이터 처리 및 다양한 유형의 데이터 처리를 하는데 어려움이 존재하였음
- 강력한 수평적 확장성이 있는 NoSQL을 사용함으로써 데이터 분산 처리 및 다양한 유형의 데이터 관리가 가능해짐

RDBMS vs. NoSQL

구분	관계형 데이터베이스(RDBMS)	NoSQL
설명	일관성(C)과 가용성(A)을 선택	일관성이나 가용성 중 하나를 포기하고, 지속성(P)를 보장
장점	데이터 무결성, 정확성 보장, 정규화된 테이블과 소규모 트랜잭션이 있음	웹 환경의 다양한 정보를 검색 및 저장 가능
단점	확장성에 한계가 있음 클라우드 분산 환경에 부적합	데이터의 무결성과 정확성을 보장하지 않음

ACID: Atomicity(원자성), Consistency(일관성), Isolation(독립성), Durability(지속성)

하둡(Hadoop)이란?

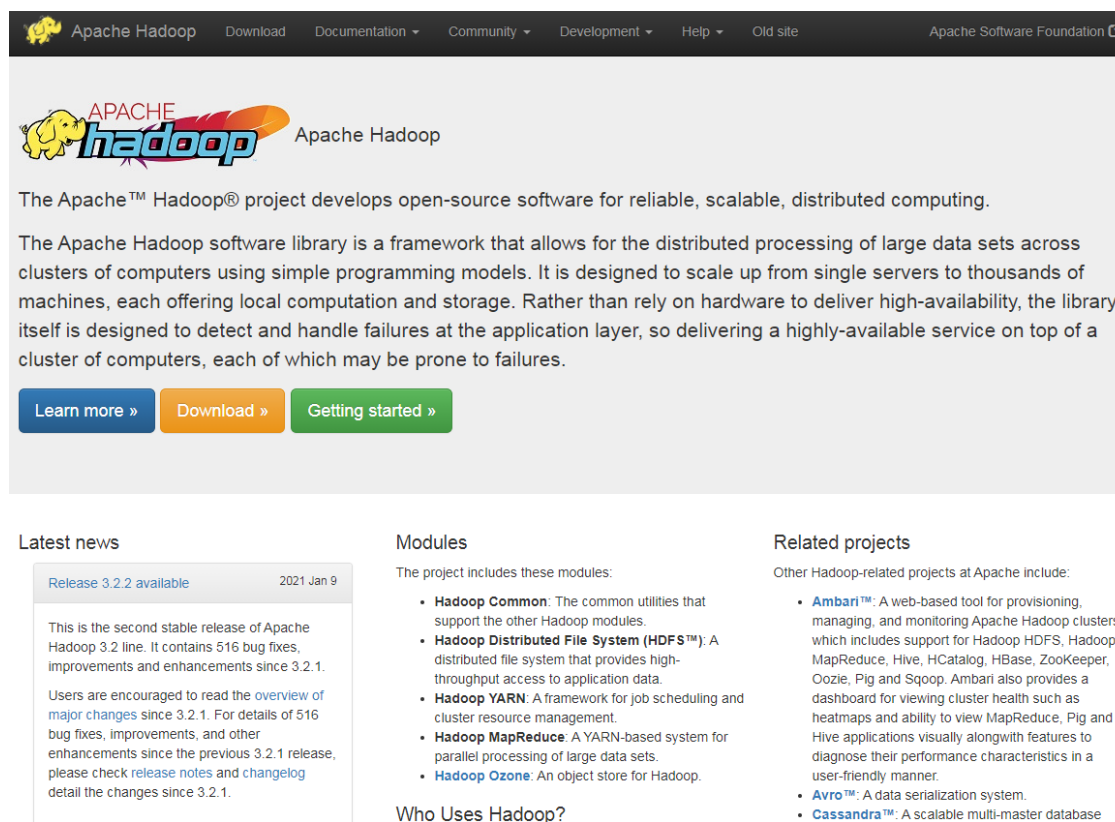
- 대용량 데이터의 분산 저장과 처리가 가능한 자바(Java)기반의 오픈소스 프레임워크이자 패키지들의 집합
- 2006년 더그 커팅과 마이크 캐퍼렐라(Mike Cafarella)가 개발



- 더그 커팅 아들의 코끼리 장난감 이름

하둡(Hadoop)이란?

<http://hadoop.apache.org/>



The screenshot shows the Apache Hadoop website. At the top is a navigation bar with links: Apache Hadoop, Download, Documentation, Community, Development, Help, Old site, and Apache Software Foundation. Below the navigation bar is the Apache Hadoop logo and the text "Apache Hadoop". The main content area describes the project as open-source software for reliable, scalable, distributed computing. It mentions that the software library is a framework for distributed processing of large data sets across clusters of computers. At the bottom of the main content area are three buttons: "Learn more", "Download", and "Getting started". Below the main content area are three sections: "Latest news", "Modules", and "Related projects".

Latest news

Release 3.2.2 available 2021 Jan 9

This is the second stable release of Apache Hadoop 3.2 line. It contains 516 bug fixes, improvements and enhancements since 3.2.1. Users are encouraged to read the [overview of major changes](#) since 3.2.1. For details of 516 bug fixes, improvements, and other enhancements since the previous 3.2.1 release, please check [release notes](#) and [changelog](#) detail the changes since 3.2.1.

Modules

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.
- **Hadoop Ozone**: An object store for Hadoop.

Who Uses Hadoop?

Related projects

Other Hadoop-related projects at Apache include:

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database

하둡(Hadoop) = 분산 데이터 처리 프레임워크

- 하둡은 여러 개의 저렴한 컴퓨터를 마치 하나인 것처럼 묶어 대용량 데이터를 처리하는 기술

■ 하둡의 구성

- 수천대의 분산된 장비에 대용량 파일을 저장할 수 있는 기능을 제공하는 분산파일 시스템(하둡 파일 시스템(HDFS))과
- 저장된 파일 데이터를 분산된 서버의 CPU와 메모리 자원을 이용해서 쉽고 빠르게 분석할 수 있는 컴퓨팅 프레임워크인 맵리듀스로 구성



■ 대용량의 데이터 처리에 최적화

- 분산컴퓨팅, 클라우드 환경

■ 장애의 대비

- 데이터의 복제본을 저장하므로 데이터의 유실이나 장애가 발생했을 때에도 복구가 용이함

■ 저렴한 구축비용

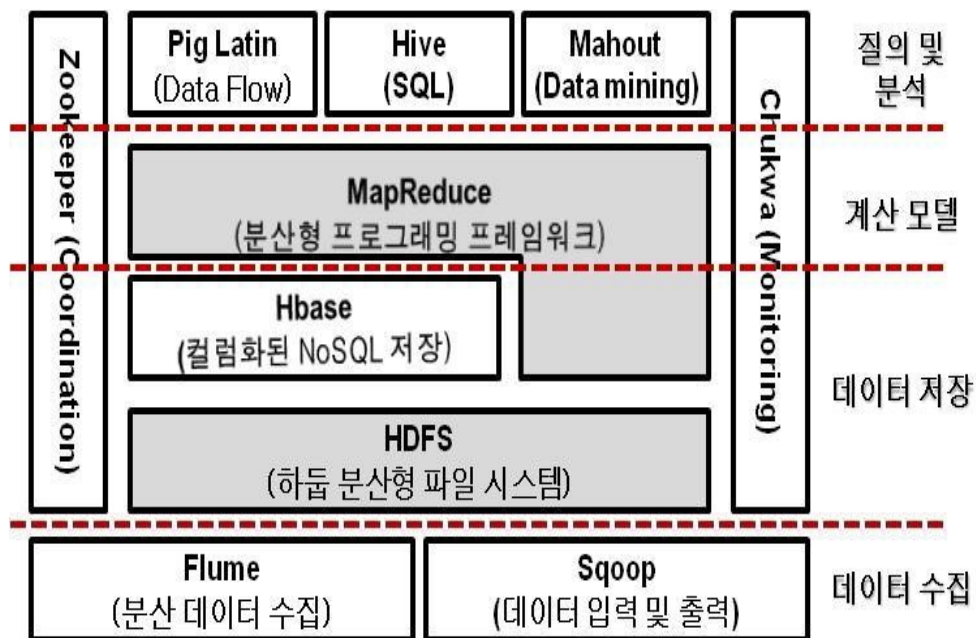
- 오픈소스 프로젝트이므로 소프트웨어 라이선스 비용에 대한 부담이 없음

(참고) 클라우드 컴퓨팅이란?

- 소프트웨어와 데이터를 인터넷과 연결된 중앙 컴퓨터에 저장, 인터넷에 접속하기만 하면 언제 어디서든 데이터를 이용할 수 있도록 하는 것
- 구름(Cloud)과 같이 무형의 형태로 존재하는 하드웨어·소프트웨어 등의 컴퓨팅 자원을 필요한만큼 빌려쓰고, 이에 대한 사용요금을 지급하는 방식
- 서로 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술로 통합해 제공하는 기술

- 주요 구성요소로 하둡 분산 파일시스템(HDFS)과 Map Reduce 가 있음
- 하둡 분산 파일시스템(HDFS)
 - 64MB~128MB 단위로 파일을 나누어 분산 저장
- Map Reduce
 - Map Reduce를 통해 데이터에 대한 분산 처리(계산) 수행

하둡의 구성요소



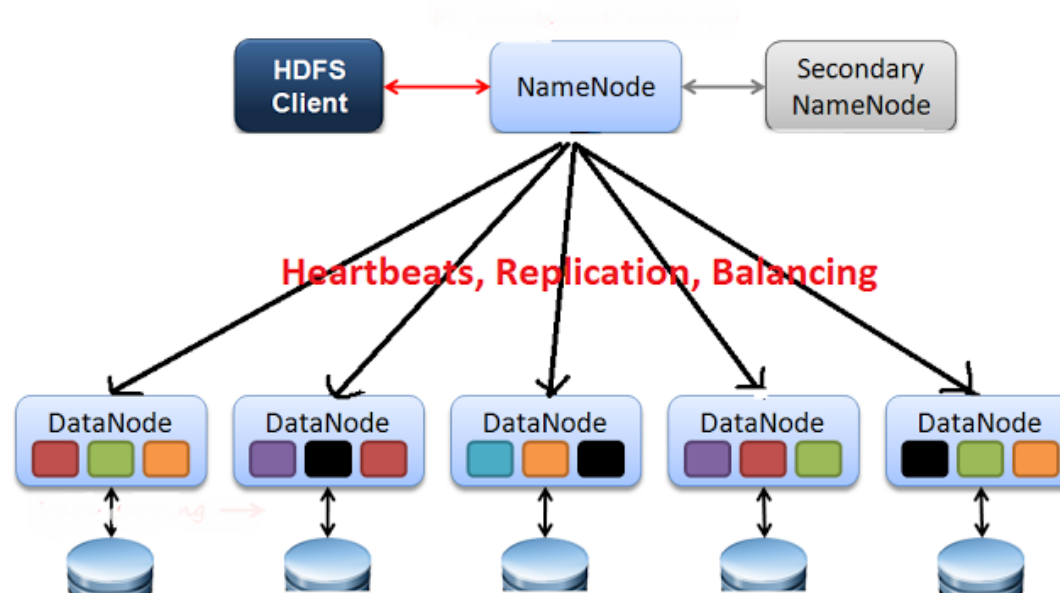
출처: Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: a technology tutorial. Access, IEEE, 2, 652-687.

하둡의 구성요소

구분	기술	주요 기능 및 특징
운영 관리	주키퍼(ZooKeeper)	여러 종류의 하둡 기반 시스템 총체적 관리
	척와(Chukwa)	시스템 상황을 모니터링하고 수집된 데이터의 표시, 모니터, 분석이 가능함
데이터 수집	플럼(Flume)	대용량 로그데이터의 수집 및 집계, 전송을 담당하는 분산 시스템
	스쿱(Sqoop)	HDFS, RDBMS, DW, NoSQL 등 저장소의 대용량 데이터를 전송하는 시스템
데이터 저장	H베이스(Hbase)	HDFS를 지원하기 위한 데이터베이스 모델, NoSQL형식
질의 및 분석	피그(Pig Latin)	대규모 데이터 셋에 대한 분석을 위한 관계형 대수 쿼리 언어 인터페이스
	하이브(Hive)	SQL 프로그램 구현 인프라로 데이터 요약, 쿼리를 수행하고 분석할 수 있는 데이터 웨어하우징 솔루션
	머하웃(Mahout)	하둡 기반으로 데이터 마이닝 알고리즘을 구현한 오픈소스 프로젝트

출처: Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: a technology tutorial. Access, IEEE, 2, 652-687.

HDFS – Hadoop Distributed File System



- 네임 노드(마스터 역할)
 - HDFS의 모든 메타데이터와 데이터 노드를 관리 – 하트비트(3초, 데이터 노드의 동작여부 판단)와 블록리포트(6시간, HDFS에 저장된 파일에 대한 최신 정보 유지) 이용
 - 클라이언트가 이를 이용하여 HDFS에 저장된 파일에 접근할 수 있음
- 데이터 노드(슬레이브 역할)
 - 파일 저장
 - 네임 노드에 하트비트와 블록리ports를 주기적으로 전달
- 클라이언트는 네임 노드에서 원하는 파일이 저장된 블록의 위치를 확인하고, 해당 블록이 저장된 데이터 노드에서 직접 데이터를 조회함

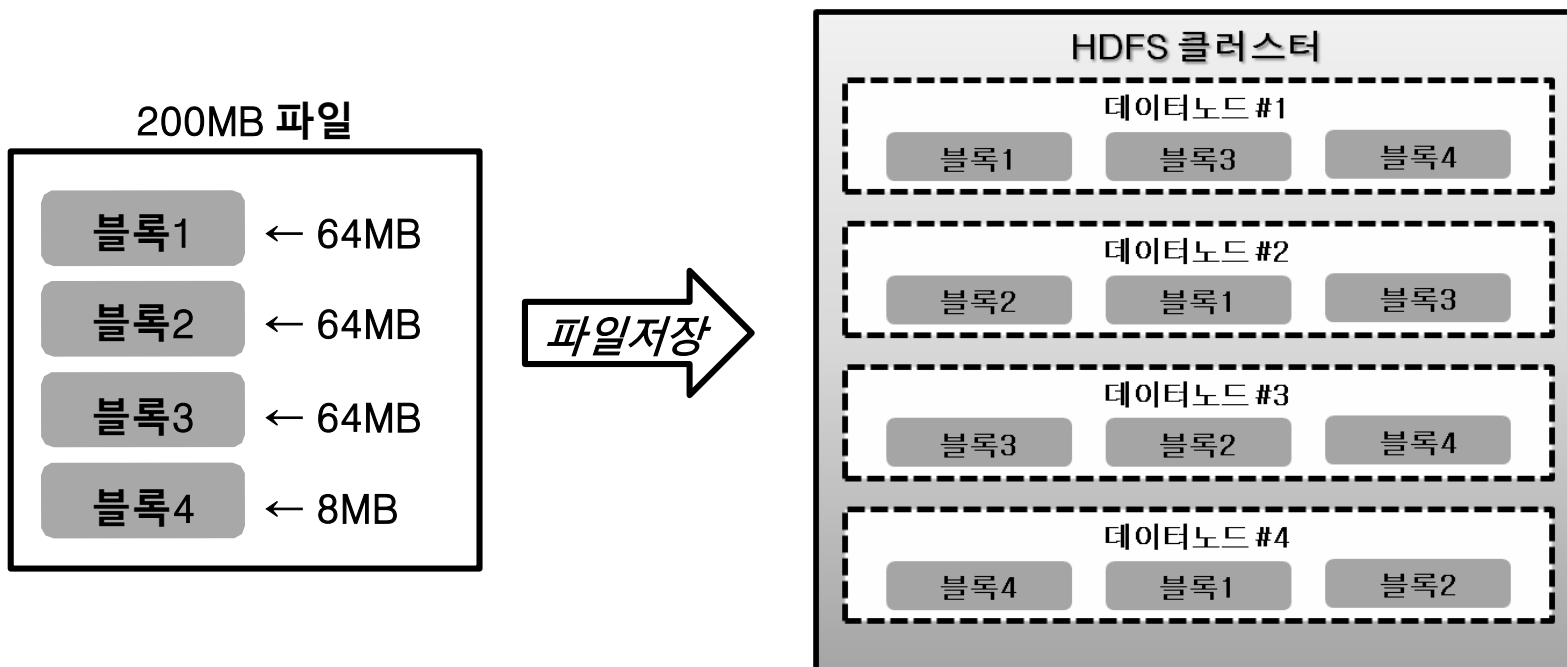
하둡 분산 파일시스템(HDFS)

■ 하둡 어플리케이션이 사용하는 분산 저장소의 역할 수행

- 범용 하드웨어로 구성된 클러스터로 대용량의 데이터를 저장하는 분산 파일시스템. 큰 파일을 작은 블록으로 나누어 개별 컴퓨터에 저장
- 각 블록의 유실, 고장 등의 위험으로 신뢰성 향상을 위해 복제물을 생성

■ HDFS의 파일 저장 방식(예시)

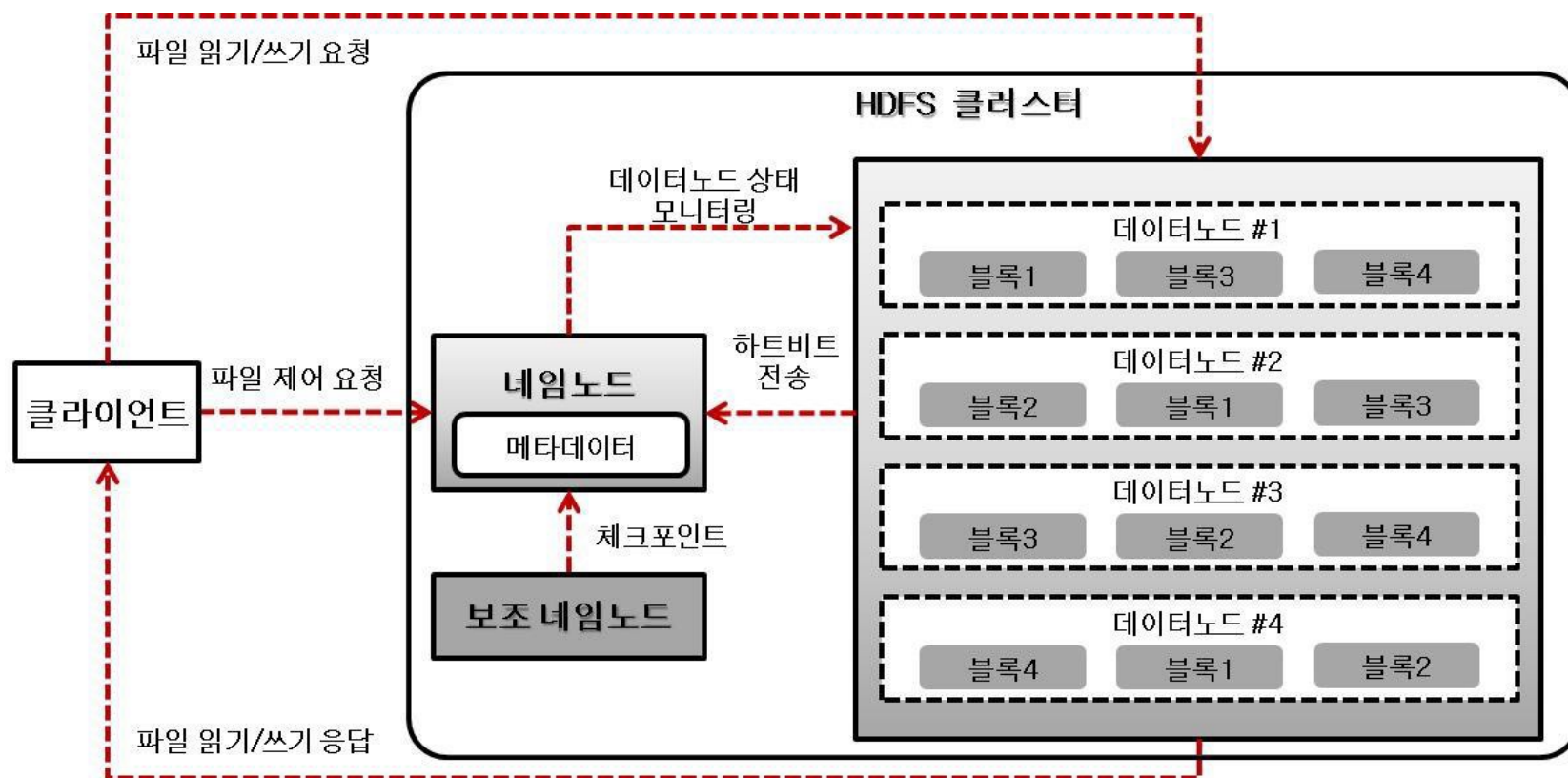
- HDFS의 기본 블록 크기는 64MB이고, HDFS는 블록을 저장할 때 기본 적으로 3개씩 블록의 복제본을 저장



- HDFS클러스터는 크게 하나의 마스터(Master)와 여러 개의 슬레이브(Slave) 노드로 구성
- 마스터 노드(네임 노드): 슬레이브 노드에 대한 메타데이터 관리 및 모니터링 시스템. 파일 및 디렉터리 읽기(open), 닫기(close), 이름 바꾸기(rename)의 기능 수행
- 슬레이브 노드(데이터 노드): 데이터 블록을 분산 처리. 읽기 (read), 쓰기 (write)의 기능 수행

하둡 분산 파일시스템(HDFS)

■ HDFS 아키텍처: 마스터-슬레이브(Master-Slave) 아키텍처



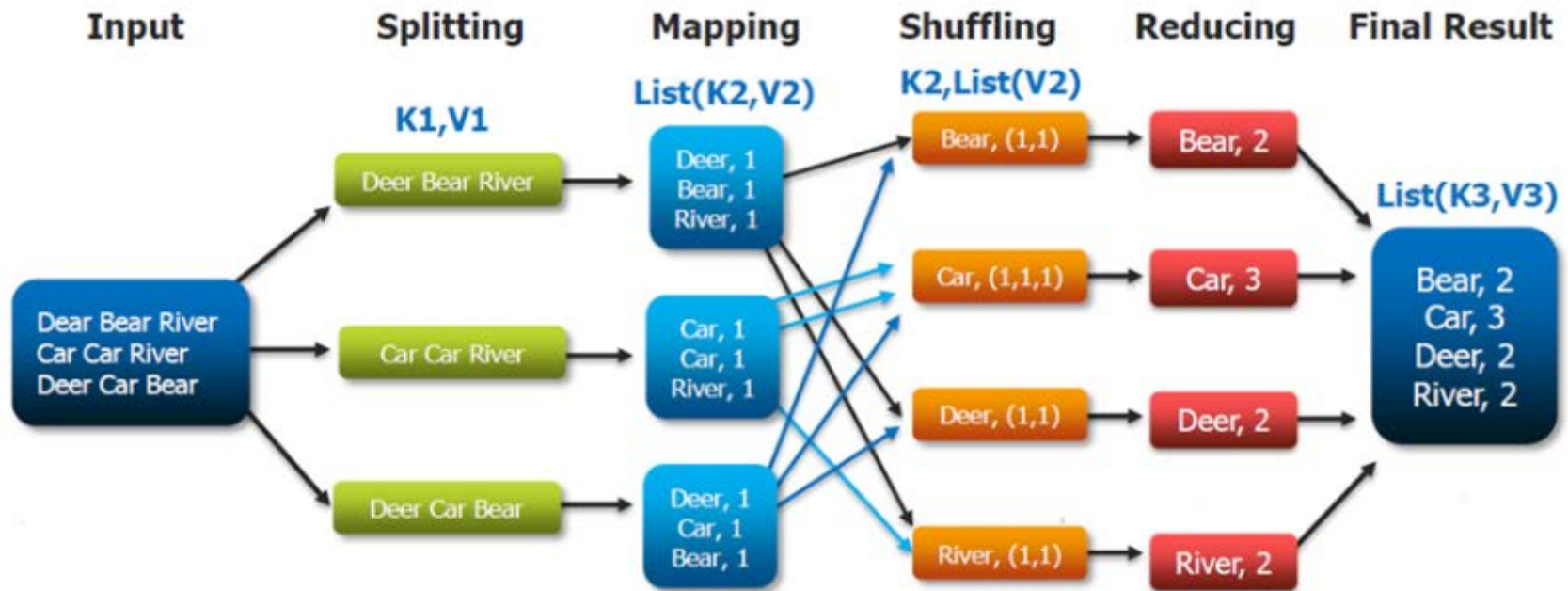
RDBMS vs. Hadoop

구분	관계형 데이터베이스	하둡
데이터 크기	기가바이트 ~ 수 테라바이트	테라바이트 ~ 페타바이트
데이터 조작	작은 데이터 참조, 변경	큰 데이터 삽입, 참조(변경은 없음)
응답 시간	빠름	느림
서버 대수와 성능 향상	여러 대의 서버로 스케일 업	수백 대~수천 대 서버로 스케일 아웃
데이터 구조	구조화 데이터	준 구조화, 비 구조화 데이터

맵리듀스

- 맵(Map) – 흩어져 있는 데이터를 연관성 있는 데이터들로 분류하는 작업(Key, Value의 형태)
- 리듀스(Reduce) – Map에서 출력된 데이터를 중복 데이터를 제거하고 원하는 데이터를 추출하는 작업

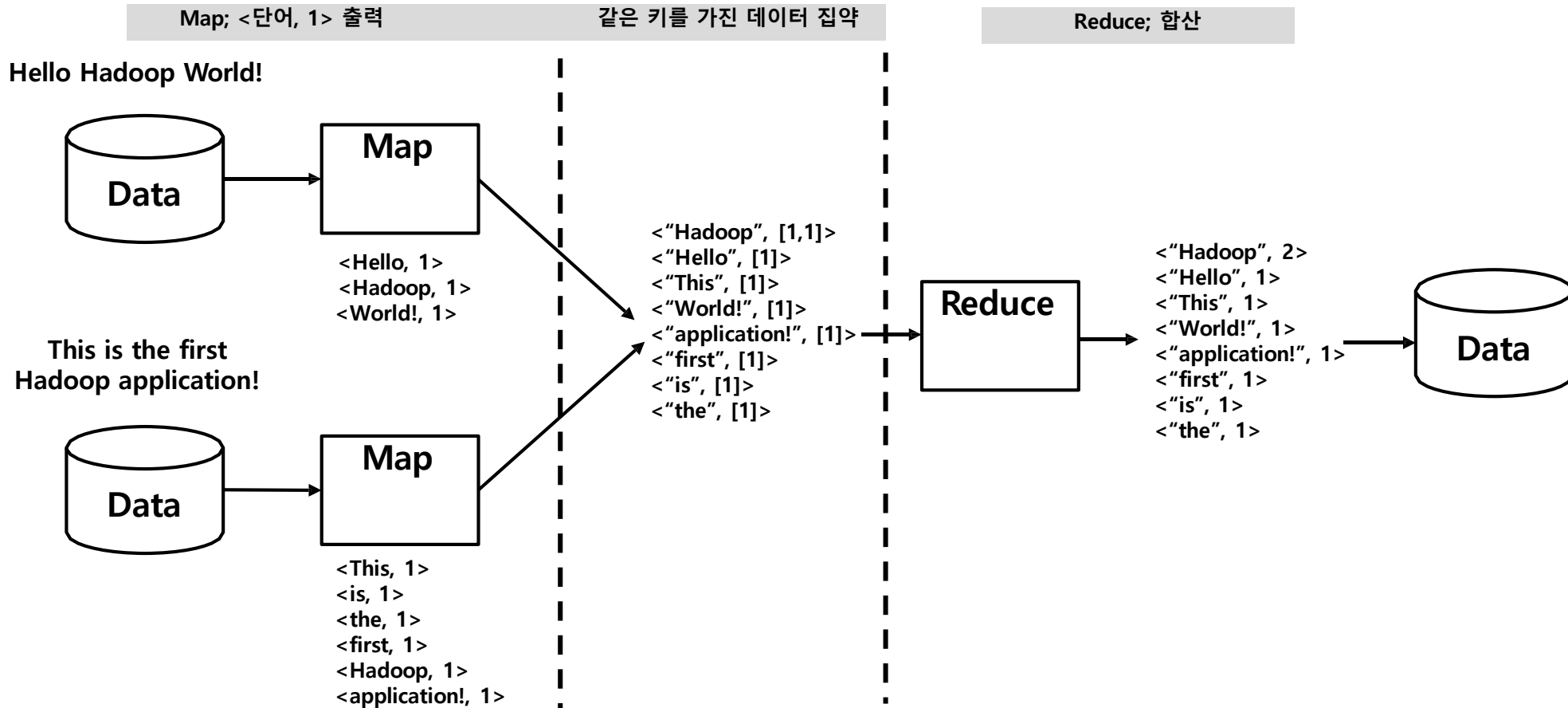
The Overall MapReduce Word Count Process



- 대용량 데이터를 빠르고 안전하게 처리하기 위한 분산 프로그래밍 모델
- 함수형 프로그래밍에서 일반적으로 사용하는 맵(Map)과 리듀스(Reduce) 함수 기반으로 구성되어 단계별로 처리 작업
- 맵(Map): 입력파일을 한 줄씩 읽어 필터링(filtering)하거나 다른 값으로 변환하는 데이터 변형 작업 수행
- 리듀스(Reduce): 맵 함수를 통해 출력된 결과 값을 새로운 키 기준으로 중복 데이터 제거 후 그룹화 한 후 집계연산을 수행한 결과 추출

맵리듀스(Map Reduce)

■ Word Count 애플리케이션



감사합니다.