3. 데이터 입출력

> [1] 논리 데이터

데이터 모델

현실 세계의 정보를 인간과 컴퓨터가 이해할 수 있도록 추상화하여 표현한 모델

데이터 모델 절차

요구사항 분석 > 개념적 > 논리적(정규화) > 물리적(반정규화)

논리 데이터 모델 종류

- 관계 데이터 모델 : 1:1, 테이블
- 계층 데이터 모델 : 1:N, 트리
- 네트워크 데이터 모델 : N:M, 그래프

논리 데이터 모델링 속성

개체(entity), 속성(attribute), 관계(relationship)

관계 데이터 모델

- 튜플(tuple), 행(row), 카디널리티(cardinality)
- 속성(attribute), 열(column), 차수(degree)

관계 대수

절차적 언어

- 일반 집합 연산자 : 합집합(∪), 교집합(∩), 차집합(—), 카티션 프로덕트(X)
- 순수 관계 연산자 :
  - 셀렉트(σ) : R에서 조건을 만족하는
  - 프로젝트(π) : R에서 주어진 속성들의 값으로만 구성된
  - 조인(⋈) : 공통 속성을 이용
  - 디비전(÷) : 릴레이션 S의 모든 튜플과 관련 있는 R의 튜플 반환

관계 해석

튜플 관계 해석과 도메인 관계해석을 하는 비절차적 언어

개체-관계(E-R) 모델

현실 세계에 존재하는 데이터와 그들 간의 관계를 사람이 이해할 수 있는 형태로 명확하게 표현하기 위해 사용되는 모델

- 개체 □ 관계 ◇ 속성 ○ 다중 값 속성 ◉ 관계-속성 —

정규화(Nomalization)

데이터의 중복성을 제거해 이상현상을 방지하고, 데이터의 일관성과 정확성을 유지하기 위해 무손실 분해하는 과정

- 1NF : 도메인이 원자값
- 2NF : 부분함수 종속 제거
- 3NF : 이행함수 종속 제거(A→ B , B→C 이면 A→C)
- BCNF : 결정자 후보 키가 아닌 함수 종속 제거
- 4NF : 다치(다중 값) 종속 제거
- 5NF : 조인 종속 제거

이상 현상(Anomaly)

데이터의 중복성으로 인해 릴레이션을 조작할 때 발생하는 비합리적인 현상

- 삽입 이상, 삭제 이상, 갱신 이상

반 정규화(De-Normalization)

정규화 된 엔티티, 속성, 관계에 대해 성능 향상과 개발 운영의 단순화를 위해 중복, 통합, 분리 등을 수행하는 과정

> [2] 물리 데이터

물리 데이터 모델링

논리모델을 적용하고자 하는 기술에 맞도록 상세화해가는 과정

참조무결성 제약조건

참조하는 외래키의 값은 항상 참조되는 릴레이션에 기본키로 존재해야한다.

- 제한(RESTRICT), 연쇄(CASCADE), 널 값(SET NULL)

인덱스

전체 데이터 검색 없이 필요한 정보에 대해 신속한 조회 가능

뷰

접근이 허용된 자료만을 제한적으로 보여주기 위해 하나 이상의 기본 테이블로 구성된 가상 테이블

클러스터

데이터 액세스 효율을 향상시키기 위해 동일한 성격의 데이터를 동일한 데이터 블록에 저장하는 물리적 저장 방법

파티션(Partition)의 종류

- 레인지(Range) 파티셔닝 : 연속적인 숫자나 날짜 기준
- 해시(Hash) 파티셔닝 : 파티션 키의 해시 함수 값
- 리스트(List) 파티셔닝 : 특정 파티션에 저장 될 데이터에 대한 명시적 제어 가능
- 컴포지트(Composite) 파티셔닝 : 레인지, 해시, 리스트 중 2개 이상의 파티셔닝 결합

> [3] 데이터베이스

데이터베이스 정의

- 통합된 데이터 : 자료의 중복을 배제한 데이터의 모임
- 저장된 데이터 : 저장 매체에 저장된 데이터
- 운영 데이터 : 조직의 업무를 수행하는 데 필요한 데이터
- 공용 데이터 : 여러 애플리케이션, 시스템들이 공동으로 사용하는 데이터

데이터베이스 특성

실시간 접근성, 계속적인 변화, 동시 공용, 내용 참조

DBMS

데이터 관리의 복잡성을 해결하는 동시에 데이터 추가, 변경, 검색, 삭제 및 백업, 복구 보안 등의 기능을 지원하는 SW

DBMS 유형

- 키-값 DBMS
- 컬럼 기반 데이터 저장(Column Family Data Store)
- 문서 저장(Document Store)
- 그래프(Graph Store) : 시맨틱 웹과 온톨로지 분야

빅데이터

시스템, 서비스, 조직 등에서 주어진 비용, 시간 내에 처리가 가능한 수십 페타바이트 크기의 비정형 데이터

- HDFS : 대용량의 데이터의 집합을 처리하는 응용 프로그램에 적합하도록 설계된 하둡 분산 파일 시스템
- 맵 리듀스(Map Reduce) : 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅 처리하기 위한 목적으로 제작해 2004년에 발표한 소프트 프레임 워크

NoSQL

전통적인 RDBMS와 다른 DBMS를 지칭하기 위한 용어, 데이터 저장에 고정된 테이블 스키마가 필요하지 않고 조인 연산을 사용할 수 없으며, 수평적으로 확장이 가능한 DBMS

NoSQL의 특성(BASE)

- Basically Available : 언제든지 데이터는 접근할 수 있어야 하는 속성
- Soft-State : 노드의 상태는 외부에서 전송된 정보를 통해 결정되는 속성
- Eventually Consistency : 일정 시간이 지나면 데이터의 일관성이 유지

시맨틱 웹(Semantic Web)

기계가 이해할 수 있는 온톨로지 형태로 표현하고 자동화된 기계가 처리하도록 하는 지능형 웹

온톨로지(Ontology)

실세계에 존재하는 모든 개념들과 개념들의 속성, 개념들 간의 관계 정보를 컴퓨터가 이해할 수 있도록 서술해 놓은 지식베이스

데이터 마이닝(Data Minning)

대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아내는 기술

데이터 마이닝 주요기법

- 분류 규칙(Classification) : 과거 데이터로부터 특성을 찾아내어 분류모형을 만들어 결과 값 예측
- 연관 규칙(Association) : 데이터 안에 존재하는 항목들 간의 종속관계를 찾아내는 기법
- 연속 규칙(Sequence) : 연관 규칙에 시간 관련 정보가 포함된 형태의 기법
- 데이터 군집화(Clustering) : 대상 레코드들을 유사한 특성을 지는 몇 개의 소그룹으로 분할하는 작업