1. 데이터 저장소

* 데이터 모델

1. 데이터 모델

현실 세계의 정보를 인간과 컴퓨터가 이해할 수 있도록 추상화하여 표현한 모델이다.

데이터 모델에 표시해야 할 요소에는 논리적 데이터 구조, 연산, 제약 조건이 있다

1. 구조

데이터베이스에 논리적으로 표현될 대상으로서의 개체 타입과 개체 타입 간의 관계

데이터 구조 및 정적 성질을 표현하는 요소

1. 연산

데이터베이스에 저장된 실제 데이터를 처리하는 작업에 대한 명세

릴레이션을 조작하기 위한 관계 연산을 나타냄 ( SELECT, PROJECT, JOIN, DIVISION )

1. 제약 조건

데이터베이스에 저장될 수 있는 실제 데이터의 논리적인 제약 조건

데이터 무결성 유지를 위한 DB의 보편적

릴레이션의 특정 컬럼에 설정하는 제약을 의미 ( 개체 무결성, 참조 무결성 등 )

1. 데이터 모델 절차
2. 요구조건 분석

도출된 요구사항 간 상충을 해결하고 범위를 파악하여 외부 환경과의 상호작용을 분석을 통해 데이터에 대한 요구 분석

1. 개념적 설계

사용자의 요구에 대한 트랜잭션을 모델링 하는 단계

개념적 데이터 모델은 현실 세계에 대한 인식을 추상적, 개념적으로 표현하여 개념적 구조를 도출하는 데이터 모델

주요 산출물은 개체-관계 다이어그램이 있다

1. 논리적 설계

트랜잭션의 인터페이스를 설계 하는 단계

DBMS에 맞는 논리적 스키마를 설계하는 단계

논리적 데이터 모델은 업무의 모습을 모델링 표기법으로 형상화하여 사람이 이해하기 쉽게 표현한 데이터 모델

논리적 데이터 모델을 통해 관계 데이터 모델, 계층 데이터 모델, 네트워크 데이터 모ㅓ델, 객체 지향 데이터 모델, 객체-관계 데이터 모델 중 하나의 모델에 맞게 설계

스키마의 평가 및 정제

1. 물리적 설계

논리 데이터 모델을 사용하고자 하는 각 DBMS의 특성을 고려하여 데이터 베이스 저장 구조로 변환 하는 모델

테이블, 인덱스, 뷰, 파티션 등 객체를 생성

반 정규화를 수행

* 논리 데이터 모델

1. 논리 데이터 모델링 개념

업무의 모습을 모델링 표기법으로 형상화하여 사람이 이해하기 쉽게 표현하는 프로세스

개념 모델로부터 업무 영역의 업무 데이터 및 규칙을 구체적으료 표현한 모델

1. 논리적 데이터 모델링 종류
2. 관계 데이터 모델

논리적 구조가 2차원 테이블 형태로 구성되며 기본키와 이를 참조하는 외래키로 관계 표현하고, 1:1 1:N N:M 관계를 자유롭게 표현

1. 계층 데이터 모델

논리적 구조가 트리 형태로 구성된 모델, 상하관계 존재

1:N 관계만 허용

1. 네트워크 데이터 모델

논리적 구조가 그래프 형태로 구성된 모델

CODASYL DBTG 모델이라고 불림

상위와 하위 레코드 사이에 다대다 관계를 만족하는 구조

1. 관계 데이터 모델 ( 중요 )
2. 관계 데이터 모델

데이터를 행과 열로 구성된 2차원 테이블 형태로 구성한 모델이다

E.F.Codd 박사가 제안한 모델

릴레이션, 튜플, 속성, 카디널리티, 차수, 스키마, 인스턴스 로 구성되어있다

1. 관계 대수

관계형 데이터베이스에서 원하는 정보와 그 정보를 어떻게 유도하는가를 기술하는 절차적 정형 언어이다

* 일반 집합 연산자

수학의 집합 개념을 릴레이션에 적용한 연산자이다

합집합 : 합병 가능한 두 릴레이션 R과 S의 합집합

교집합 : 릴레이션 R과 S에 속하는 모든 튜플로 결과 릴레이션 구성

차집합 : R에 존재하고 S에 미 존재하는 튜플로 결과 릴레이션 구성

카티션 프로덕트 : R과 S에 속한 모든 튜플을 연결해 만들어진 새로운 튜플로 릴레이션 구성

* 순수 관계 연산자

관계 데이터베이스에 적용할 수 있도록 특별히 개발한 관계 연산자이다

셀렉트 : 릴레이션 R에서 조건을 만족하는 튜플 반환

프로젝트 : 릴레이션 R에서 주어진 속성들의 값으로만 구성된 튜플 반환

조인 : 공통 속성을 이용해 R과 S의 튜플들을 연결해 만들어진 튜플 반환

디비전 : 릴레이션 S의 모든 튜플과 관련 있는 R의 튜플 반환

1. 관계 해석

튜플 관계 해석과 도메인 관계 해석을 하는 비절차적 언어

프레디킷 해석에 기반한 언어이며 비절차적 언어이다

1. 관계 대수와 관계 해석 비교

관계 대수는 절차적 언어이며, How 어떻게 유도하는가, 순수관계 연산자, 일반집합 연산자

관계 해석은 비절차적 언어이며, What 무엇을 얻을 것인가, 튜플 관계 해석, 도메인 관계 해석

1. 논리 데이터 모델링 속성
2. 개체

데이터베이스에 표현하려는 사물 또는 사건

피터 챈 모델 : 개체를 사각형으로 표기

까마귀발 모델 : 개체를 표 형식으로 표기

1. 속성

개체가 가지고 있는 요소 또는 성질

피터 챈 모델 : 속성을 타원형으로 표기

까마귀발 모델 : 속성을 표 내부에 포기

1. 관계

두 개체 간의 관계를 정의

피터 챈 모델 : 마름모로 표시

까마귀발 모델 : 관계 표시 1:1, 1:N, N:M 관계 일자선 N에 선 3개 둘다 선 3개

1. 개체-관계(E-R) 모델
2. 개념

E-R 모델은 현실 세계에 존재하는 데이터와 그들 간의 관계를 사람이 이해할 수 있는 형태로 명확하게 표현하기 위해서 가장 널리 사용되고 있는 모델

논리 데이터 모델링에서는 모든 이해당사자와 의사소통의 보조 자료로 E-R모델을 활용한다

요구사항으로부터 얻어낸 정보들을 개체, 속성, 관계로 기술한 모델이다

1. 개체-관계 다이어그램 기호

* 개체 집합 : 사각형
* 관계 집합 : 마름모
* 속성 : 타원
* 다중 값 속성 : 이중 타원
* 개체 집합-관계 집합 연결 : 실선
* 개체 집합-속성 연결 : 실선
* 관계 집합-속성 연결 : 점선

1. 정규화
2. 이상 현상

데이터의 중복성으로 인해 릴레이션을 조작할 때 발생하는 비합리적인 현상이다

삽입, 삭제, 갱신 이상

* 삽입 이상 : 정보 저장 시 해당 정보의 불필요한 세부정보를 입력하는 경우
* 삭제 이상 : 정보 삭제 시 원치 않는 다른 정보가 같이 삭제되는 경우
* 갱신 이상 : 중복 데이터 중에서 특정 부분만 수정되어 중복된 값이 모순을 일으킴

1. 함수 종속

릴레이션에서 속성의 의미와 속성 간 상호 관계로부터 발생하는 제약조건

결정자/종속자 X -> Y 관계일 때 X는 결정자, Y는 종속자 이다

함수 종속 종류에는 부분 함수 종속, 완전 함수 종속, 이행 함수 종속이 있다

* 부분 함수 종속 ( Partial Functional Dependency )

릴레이션에서 기본 키가 복합 키일 경우 기본 키를 구성하는 속성 중 일부에게 종속된 경우

* 완전 함수 종속 ( Full Functional Dependency )

릴레이션에서 X -> Y 관계가 있을 때, Y는 X의 전체 속성에 대해 종속하고, 부분 집합 속성에 종속하지 않는 경우

* 이행 함수 종속 ( Transitive Functional Dependency )

릴레이션에서 X -> Y, Y -> Z 종속 관계가 있을 때, X -> Z 가 성립되는 경우

1. 정규화

관계형 데이터 모델에서 데이터의 중복성을 제거하여 이상 현상을 방치하고, 데이터의 일관성과 정확성을 유지하기 위해ㅑ 무손실 분해 하는 과정

* 1NF : 원자값으로 구성
* 2NF : 부분 함수 종속 제거
* 3NF : 이행함수 종속 제거
* BCNF 보이스-코드 정규화 : 결정자 후보 키가 아닌 함수 종속 제거
* 4NF : 다치(다중 값) 종속 제거
* 5NF : 조인 종속 제거

1. 반 정규화
2. 개념

정규화된 엔터티, 속성, 관계에 대해 성능 향상과 개발 운영의 단순화를 위해 중복, 통합, 분리 등을 수행하는 데이터 모델링의 기법이다 비정규화, 역정규화라고도 불린다.

1. 특징

성능 향상과 관리의 효율성이 증가 하지만 데이터의 일관성 및 정합성 저하가 된다

반 정규화를 위해서는 데이터의 일관성과 무결성을 우선으로 할지 데이터베이스의 성능과 단순화에 우선순위를 둘 것인지를 비교하여 조정하는 과정이 중요하다

1. 반 정규화 기법

테이블 : 테이블 병합, 테이블 분할, 중복 테이블 추가

컬럼 : 컬럼 중복화

관계 : 중복관계 추가

* 물리 데이터 모델

1. 데이터베이스 무결성
2. 개념

데이터베이스에 저장된 데이터 값과 그것이 표현하는 현실 세계의 실제 값이 일치하는 성질이다

데이터의 무결성을 유지하는 것은 데이터베이스 관리 시스템( DBMS )의 중요한 기능이며, 주로 데이터에 적용되는 연산에 제한을 두어 데이터의 무결성을 유지한다.

1. 데이터베이스 무결성 종류

* 개체 무결성 : 한 엔티티에서 같은 기본키를 가질 수 없거나, 기본키의 속성이 NULL을 혀용할 수 없는 제약조건, 주요 기법으로는 기본 키 유니크 인덱스가 있다
* 참조 무결성 : 외래 키가 참조하는 다른 개체의 기본 키에 해당하는 값이 기본 키값이나 NULL이 어야 하는 제약 조건, 주요 기법으로는 외래 키가 있다
* 속성 무결성 : 속성의 값은 기본값, NULL 여부, 도메인 이 지정된 규칙을 준수 해야되는 제약 조건, 주요 기법으로는 체크, NULL/NOT NULL, 기본값이 있다
* 사용자 정의 무결성 : 사용자의 의미적 요구사항을 준수해야된다, 주요 기법으로는 트리거, 사용자 정의 데이터 타입이 있다
* 키 무결성 : 한 릴레이션에 같은 키값을 가진 튜플을 허용 할 수 없다, 주요기법으로는 유니크가 있다

1. 키
2. 개념

데이터베이스에서 조건을 만족하는 튜플을 찾거나 순서대로 정렬할 때 다른 튜플들과 구별할 수 있는 기준이 되는 속성이다

1. 키 특성

* 유일성 : 식별자에 의해 엔티티 내에 모든 튜플들을 유일하게 구분하는 특성
* 최소성 : 최소한의 속성으로 식별자를 구성하는 특성

1. 키 종류

* 기본 키 : 테이블의 각 튜플들을 고유하게 식별하는 키
* 대체 키 : 후보 키 중에서 기본 키로 선택되지 않은 키
* 후보 키 : 테이블에서 각 튜플들을 구별하는 기준이 되며 기본 키 + 대체 키
* 슈퍼 키 : 릴레이션을 구성하는 모든 튜플에 대해 유일성은 만족하지만 최소성은 만족 하지 않는 키
* 외래 키 : 한 릴레이션의 컬럼이 다른 릴레이션의 기본 키로 이용되는 키, 테이블 간의 참조 데이터 무결성을 위한 제약 조건

1. 인덱스

검색 연산의 최적화를 위해 데이터베이스 내 열에 대한 정보를 구성한 데이터 구조

1. 파티셔닝

테이블 또는 인덱스 데이터를 파티션 단위로 나누어 저장

1. 레인지 파티셔닝 : 연속적인 숫자나 날짜를 기준으로 하는 파티셔닝 기법
2. 해시 파티셔닝 : 파티션 키의 해시 함수 값에 의한 파티셔닝 기법
3. 리스트 파티셔닝 : 특정 파티션에 저장될 데이터에 대한 명시적 제어가 가능한 기법
4. 컴포지트 파티셔닝 : 레인지, 해시, 리스트 중 2개 이상의 파티셔닝을 결합하는 기법
5. 라운드 로빈 : 라운드 로빈으로 회전하면서 새로운 행을 파티션에 할당하는 기법

2. 데이터베이스 기초 활용

* 데이터베이스 종류

1. 데이터베이스
2. 개념

다수의 인원, 시스템 또는 프로그램이 사용할 목적으로 통합하여 관리되는 데이터의 집합

데이터에 대한 효과적인 관리를 위해 자료의 중복성 제거, 무결성 확보, 일관성 유지, 유용성 보장이 중요하다

데이터베이스는 통합된 데이터, 저장된 데이터, 운영 데이터, 공용 데이터이다

* 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

1. 특성

실시간 접근성, 계속적인 변화, 동시공용, 내용 참조가 있다

* 실시간 접근성 : 쿼리에 대하여 실시간 응답이 가능해야 한다는 특성
* 계속적인 변화 : 새로운 데이터의 삽입, 삭제, 갱신으로 항상 최신의 데이터 유지
* 동시 공용 : 다수의 사용자가 동시에 같은 내용의 데이터를 이용
* 내용 참조 : 데이터를 참조할 때 데이터 레코드의 주소나 위치에 의해서가 아닌 사용자가 요구하는 데이터 내용으로 데이터를 찾아야 한다

1. DBMS

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

1. 데이터베이스 저장 기술
2. 데이터 웨어하우스

사용자의 의사결정에 도움을 주기 위해 기간 시스템의 데이터베이스에 축적된 데이터를 공통 형식으로 변환해서 관리하는 데이터베이스

데이터 웨어 하우스의 특징으로는 주제 지향적, 통합적, 시계열적, 비휘발적이 있다

* 주제 지향적 : 기능이나 업무가 아닌 주제 중심적
* 통합적 : 데이터의 일관성을 유지하며 전시적 관점에서 하나로 통합
* 시 계열적 : 시간에 따른 변경을 항상 반영
* 비휘발적 : 적재가 완료되면 읽기 전용 형태의 스냅 샷 형태로 존재

1. 데이터 마트

전사적으로 구축된 데이터 속의 특정 주체, 부서 중심으로 구축된 소규모 단위 주제의 데이터 웨어하우스이다

1. 빅데이터 특성

데이터의 양, 데이터의 다양성, 데이터의 속도

1. 하둡
2. 개념

하둡은 오픈소스를 기반으로 한 분산 컴퓨팅 플랫폼, 가상화된 대형 스토리지를 형성하고 거대한 데이터 세트를 병렬로 처리할 수 있도록 개발된 자바 소프트웨어 프레임워크

1. 하둡 주요 기술

* 데이터 수집

ETL : 데이터 분석을 위한 데이터를 데이터 저장소인 DW 및 DM 으로 이동시키기 위해 다양한 소스 시스템으로부터 필요한 원본 데이터를 추출하고 변환하여 적재

플럼 : 많은 양의 로그 데이터를 효율적으로 수집, 집계, 이동하기 위해 이벤트와 에이전트를 활용

스쿱 : 커넥터를 사용하여 관계형 데이터베이스 시스템에서 하둡 파일 시스템으로 데이터를 수집

스크래파이 : 파이썬 언어 기반의 비정형 데이터 수집 기술

* 분산 데이터 저장

HDFS : 대용량 데이터의 집합을 처리하는 응용 프로그램에 적합하도록 설계된 하둡 분산 파일 시스템

* 분산 데이터 처리

맵 리듀스 : 구글에서 대용량 데이터 처리를 분산 병령 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년에 발표한 소프트웨어 프레임워크

1. NoSQL

데이터 저장에 고정된 테이블 스키마가 필요하지 않고 조인 연산을 사용할 수 없으며 수평적으로 확장이 가능한 DBMS

* 특성

Basically Available : 언제든지 데이터는 접근 할 수 있어야 하는 속성

Soft-State : 노드의 상태는 내부에 포함된 정보에 의해 결정되는 것이 아니라 외부에서 전송된 정보를 통해 결정되는 속성

Eventually Consistency : 일정 시간이 지나면 데이터의 일관성이 유지되는 속성

* 유형

Key-Value Store : 유니크한 키 에 하나의 값 redis, dynamoDB

Colum Family Data Store : 키 안에 컬럼, 값 조합으로 된 여러개의 필드를 가진 DB HBase, Cassandra

Document Store : Value의 데이터 타입이 Document라는 타입을 사용하는 db XML, JSON, YAML , MongoDB, CouchBase

Graph Store : 시멘틱 웹과 온톨로지 분야에서 활용되는 그래프로 데이터를 표현하는 DB, Neo4j, AllegroGraph

1. 데이터 마이닝
2. 개념

대규모로 저장된 데이터 안에서 체계적으로 자동적으로 통계적 규칙이나 패턴을 찾아내는 기술

대규모 데이터에서 의미 있는 패턴을 파악하거나 예측하여 의사결정에 활용하는 기법

데이터의 숨겨진 정보를 찾아내어 이를 기반으로 서비스와 제품에 도입하는 과정

통계분석은 가설이나 가정에 따른 분석, 검증을 하지만 데이터 마이닝은 수리 알고리즘을 활용하여 대규모 데이터에서 의미 있는 정보를 찾아낸다

1. 주요 기법

* 분류 규칙 : 과거 데이터로부터 특성을 찾아내어 분류모형을 만들어 이를 토대로 새로운 레코드의 결과 값을 예측
* 연관 규칙 : 데이터 안에 존재하는 항목들 간의 종속관계를 찾아내는 기법
* 연속 규칙 : 연관 규칙에 시간 관련 정보가 포함된 기법
* 데이터 군집화 : 정보가 없는 상태에서 데이터를 분류하는 기법

1. 데이터 관련 용어
2. 텍스트 마이닝

대량의 텍스트 데이터로부터 패턴 또는 관계를 추출 하려 정보 찾는 기법

1. 웹 마이닝

웹으로부터 얻어지는 방대한 양의 정보로부터 유용한 정보를 찾아내기 위해 분석

1. 다크 데이터

수집된 후 저장은 되어 있지만 분석에 활용되지 않는 다량의 데이터

1. 메타 데이터

데이터에 대한 구조적인 데이터로서, 일련의 데이터를 정의 설명 해주는 데이터

1. 디지털 아카이빙

지속적으로 보존할 가치를 가진 디지털 객체를 장기관 관리 이후 이용을 보장할 수 있도록 변환, 압축 저장하여 DB화

1. 마이 데이터

정보 주체가 기관으로부터 자기 정보를 직접 내려받아 이용 제 3자 제공을 허용하느ㅏㄴ 방식으로 정보 주체 중심의 데이터 활용 체계