The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 18, No. 1, pp.119-127, Feb. 28, 2018. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2018.18.1.119
JIIBC 2018-1-17

연구지원 데이터베이스에서 최적화된 데이터모델링을 통한 데이터 비만도 개선에 관한 연구

A Study on Reducing Data Obesity through Optimized Data Modeling in Research Support Database

김희완*

Hee-Wan Kim*

요 약 현업에서 사용하고 있는 정형 데이터는 데이터모델링에 대한 이해부족 및 적용의 미흡으로 정규화되지 않은 채로 테이블 형태로 관리되고 있는 현실이다. 데이터베이스 설계의 균형이 파괴되면 데이터 질의에 대한 응답속도에 영향을 미치며, 데이터 비만도가 높아지게 된다. 본 논문에서는 최적화된 데이터모델링을 통한 데이터베이스 설계를 통하여 데이터 비만도가 어떻게 개선되었는지를 연구하였다. 데이터 비만도가 과다하게 나타나는 방사형 및 업무 중심의 고립형 설계에서 객체(데이터)와 객체간의 관계 중심의 데이터모델링을 통한 정방형 설계를 함으로 데이터 질의 경로가 선명하게 가시화되었다. 데이터비만도 면에서도 기존의 연구지원 데이터베이스의 비만도는 57.2%였으나, 새로운 연구지원 데이터베이스에서는 16.2%로 나타나 데이터 비만도가 40.5%가 개선되었으며, 데이터의 중복을 최소화함으로써 데이터의 정확성과 무결성이 보장되는 데이터베이스로 개선되었다.

Abstract The formal data used in the business is managed in a table form without normalization due to lack of understanding and application of data modeling. If the balance of the database design is destroyed, it affects the speed of response to the data query, and the data obesity becomes high. In this paper, it is investigated how data obesity improved through database design through optimized data modeling. The data query path was clearly visualized by square design through data modeling based on the relationship between object (data) and object, from the radial and task - oriented isolation design where data obesity is excessive. In terms of data obesity, the obesity degree of the current research support database was 57.2%, but it was 16.2% in the new research support database, and the data obesity degree was reducd by 40.5%. In addition, by minimizing redundancy of data, the database has been improved to ensure the accuracy and integrity of the data.

Key Words : Data Modeling, Data Obesity, Database Design, Data Query, Square Design

Ⅰ. 서 론

인터넷이 일상화된 ICT 시대에 디지털 데이터가 폭증하는 데이터 홍수 시대에 살고 있다. 정보기술의 발전에 따른 데이터 저장 및 처리비용의 하락, 소셜 네트워크 서비스의 확대 등으로 막대한 비정형 데이터가 생성되고 있으나, 현업에서 사용하고 있는 정형 데이터는 데이터 모델링에 대한 이해부족 및 적용의 미흡으로 정규화되지

*정회원, 삼육대학교 컴퓨터.메카트로닉스공학부
접수일자: 2018년 1월 31일, 수정완료: 2018년 2월 7일
게재확정일자: 2018년 2월 9일

Received: 31 January, 2018 / Revised: 7 February, 2018
Accepted: 9 February, 2018
*Corresponding Author: hwkim@syu.ac.kr
Division of Computer-Mechatronics, Shamyook University, Korea

않은 채로 테이블 형태로 관리되고 있는 현실이다. 데이터모델링에 대한 마인드 부족으로 정규화되지 않은 형태로 데이터가 저장, 관리되고 있으므로 데이터의 응답시간이 점점 느려지게 만드는 주요 원인이 되고 있다. 업무를 분석함에 있어서 프로세스와 데이터를 구분하지 않고 파일 형태로 테이블을 만들어 관리함으로 데이터의 불일치[1][2][3], 데이터의 중복 및 데이터 홍수 현상[4][5], 또는 데이터 과적현상[6]의 원인이 되기도 한다.

데이터가 중복 없이 데이터의 정확성과 무결성을 갖는 데이터베이스 품질을 유지하기 위한 개발 방법론[7]에 대한 연구에서는 데이터베이스 설계 과정에서 데이터를 수집하여 데이터베이스를 설계하는 과정을 소개하고 있다. 그러나, 데이터의 중복을 최소화하는 설계에 대한 데이터베이스 품질에 대한 연구는 미흡한 실정이다.

데이터 모델링이란 기업의 '전사적 데이터 지도[8][9]'를 만드는 일련의 방법론을 말한다. 많은 양의 데이터를 효과적으로 운용하기 위해 불가피한 중복은 예외인 상태에서 데이터 중복을 최대한 제거하고 속도와 질을 모두 만족시키는 데이터 지도만이 현업의 데이터 무결성을 보장하는 동시에 정보시스템의 응답 속도 및 서비스 품질을 향상시킬 수 있기 때문이다.

정보들을 수집하고 분석하기 위해서는 체계적으로 데이터베이스를 구성[10]하여야 하며, 데이터베이스의 구성은 논리데이터모델에서 개체(entity)와 개체 간의 관계를 형성해주는 관계(relationship)로 구성되어 있으며, ERD를 구성하는 ERD 편집도구[11]들을 사용하여 설계한다. 데이터모델의 균형이 파괴되면 데이터 질의에 대한 응답속도에 영향을 미치며, 데이터 비만도가 높아지게 된다. 본 논문에서는 개체 및 업무 중심으로 설계된 연구지원 데이터베이스를 설계의 균형을 유지하고 효율적인 성능을 지원하는 정방형 구조로 설계하고자 한다.

데이터베이스 시스템에서 데이터 품질제고를 위해서는 데이터의 정확성, 완전성 및 일관성을 보장하여야 한다. 따라서 본 연구에서는 실제 대학 연구지원 데이터베이스를 대상으로 평균 데이터 중복률을 조사하였다. 또한 최적화된 데이터모델링을 통한 데이터베이스 설계를 통하여 데이터 비만도가 어떻게 개선되었는지를 도출하였다.

현행 연구지원 데이터베이스의 경우 전반적으로 데이터베이스 설계의 품질에 문제가 있었다. 이는 데이터 중복이나 널 값의 입력허용으로 기인하게 되었다. 데이터 비만도가 과다하게 나타나는 방사형 및 업무 중심의 고립형 설계에서 객체(데이터)와 객체간의 관계 중심의 데이터모델링을 통한 정방형 설계를 함으로 데이터 질의 경로가 선명하게 가시화되고, 데이터베이스의 확장성과 정확성을 보장하도록 데이터베이스를 설계하고자 한다.

II. 관련 연구

DB품질은 DB의 바람직한 정도 또는 우수성이라고 정의되며, DB 품질기준은 데이터 품질과 서비스 품질로 양분하여 접근하고 있다. 모두 7개 기준 중 DB데이터 품질을 구성하고 있는 요소는 정확성, 완전성, 현행성, 일관성이다. DB서비스 품질은 검색성, 사용 용이성, 그리고 사용자 지원성으로 구성되어 있다[12].

본 연구에서는 7개 기준 중 표현하고자 하는 실세계의 중요한 객체들과 속성들이 모두 담겨있어야 하는 점이 관건인 데이터 완전성에 대해 살펴보고자 한다. DB데이터의 완전성은 데이터 구조(structure)의 완전성, 데이터 값(value)의 완전성, 데이터 표현(representation)의 완전성 세 가지로 나누어진다[13].

1. 데이터 구조의 완전성

데이터 구조의 완전성이란 데이터베이스의 데이터가 실세계의 중요한 객체그룹을 모두 포함하고 있는지, 또한 객체에 관한 중요한 속성들을 모두 담고 있는지를 분석함으로써 데이터베이스 품질을 평가하는 도구이다. 데이터 구조가 완전하지 못하다 함은 사용자의 정보요구분석단계, 데이터모델링 단계, 혹은 데이터베이스의 논리설계 단계에서 치명적 결함이 발생했음을 의미하며, 이러한 결함은 데이터베이스 품질뿐 아니라 데이터베이스 존재가치에 까지 영향을 줄 수 있다. 그러나 무엇이 중요한 속성이며 그것이 왜 빠졌는가? 라는 문제는 해당 분야 전문가의 심층적 분석과 판단이 요구된다.

2. 데이터 값의 완전성

데이터의 완전성에서는 누락된 행(row)이나 컬럼(column)의 값이 널(null)이 없는지를 평가한다. 이러한 문제는 잘못 설계된 테이블에 데이터를 입력할 때 발생한다. 전체 행중에서 누락된 행의 비율에 의해 데이터의 완전성 정도를 측정할 수 있다. 컬럼의 널 값이 갖는 문

제로는 다음과 같은 것들이 있다.

Student

Student_ID	FirstName	Surname	Tel_No_1	Tel_No_2	Tel_No_3
111	Rosy	Smith	403-738-1234	null	null
222	John	Peter	637-112-1632	909-743-1241	704-766-5134
333	Maria	Jones	504-810-1296	null	null

그림 1. Student 테이블
Fig. 1. Student Table

컬럼 값이 널 인 경우 데이터베이스의 제1정규형에 위배가 된다. 제1정규형이란 테이블 내에 모든 열은 원자값(automic value)을 갖는다는 것을 말한다. 즉 열의 값에 널 값을 허용하지 않는다는 것이다. 예를 들어, 그림 1은 학생별로 3개의 전화번호 행을 가지고 있는 학생 테이블이다. 그림 1의 Student 테이블은 제1정규형이 아니다. Tel_No_1, Tel_No_2, Tel_No_3 행은 동일한 도메인과 데이터타입을 가지고 있다. 학생별 3개의 전화번호 행을 가지게 될 경우에는 다음과 같은 논리적인 문제들이 있다.

테이블 질의 시 어려움 발생 : "어느 학생이 전화번호 XXX를 가지고 있는가?", "어떤 학생들끼리 같은 전화번호를 공유하는가?" 등의 질의에 대하여 답하기 어렵다.

학생 테이블에서 전화번호의 유일성을 확보하기 어렵다. 학생 333번에 Tel_No_2 값이 Tel_No_1 값과 동일한 값이 입력되어도 된다.

전화번호 개수의 제한 : 어떤 학생이 4번째 전화번호를 입력하고자 하는 경우에는 그 전화번호를 입력할 수 없다. 이는 데이터베이스 설계 자체가 사용자의 편의를 제한하고 있다는 의미이다.

이러한 문제점들을 해결하기 위해 그림 1의 학생 테이블을 그림 2와 그림 3과 같이 2개의 테이블로 분해하는 것이다. 그림 2의 Student_Name 테이블에서는 Studnet_ID가 기본키이고, 그림 3에서의 Student_TelephoneNumber 테이블에서는 Student_ID와 Serial_Number가 복합키이면서 기본키이다. 하나의 Student 테이블을 두 개의 테이블로 분해한 후에야 널 값의 삽입으로 인한 문제점을 해결하였으며, 테이블내의 모든 컬럼의 값이 원자값을 가지는 제1정규형을 만족하는 데이터베이스가 된다.

Student_Name

Customer_ID	FirstName	Surname
111	Rosy	Smith
222	John	Peter
333	Maria	Jones

그림 2. Student_Name 테이블
Fig. 2. Student_Name Table

Student_TelephoneNumber

Customer_ID	Serial_Number	Telephone_Number
111	1	403-738-1234
222	1	637-112-1632
222	2	909-743-1241
222	3	704-766-5134
333	1	504-810-1296

그림 3. Student_TelephoneNumber 테이블
Fig. 3. Student_TelephoneNumber Table

3. 데이터 표현의 완전성

데이터 표현의 완전성이란 가공된 데이터가 원시 데이터의 내용(정보)을 완전하게 표현하고 있는지를 의미한다. 데이터의 분류나 데이터의 재구성, 혹은 데이터의 재배열된 후 데이터의 내용이 원시 데이터를 일부 누락하고 있는지, 혹은 추출된 키워드는 주제 영역의 개념을 설명할 수 있는지, 또는 가공된 데이터가 원시 데이터의 내용을 전부 포함하고 있는지의 여부를 분석하여 데이터 표현의 완전성을 평가한다[12].

Ⅲ. 표준 업무

1. 가정

본 논문에서는 기존의 데이터베이스 설계에 대해 문제점을 도출하여 새로운 데이터베이스 설계도를 비교하고 시스템 구축 설계를 위해 반드시 필요한 몇 가지 가정을 도입하기로 한다.

가정 1(업무의 종류): 대상 업무는 현재 대학 연구지원 시스템에서 운영하고 있는 업무라고 가정한다.

가정 2(설계방법론): 3단계 데이터모델링 단계 중 논리적 모델링 단계로 개체관계도를 작성하는 것으로 한다.

가정 3(개체관계도의 표현): 정보공학방법론[14]이나 바커 표기법[15]에 의하지 않고 Chen 방식[16]을 사용하여 표현한다.

가정 4(업계관행 수용): 응답시간을 줄이고 성능향상을 위하여 물리적 설계는 테이블 통합으로 된다고 가정한다. 업무기술서 작성에서 데이터모델링 과정을 거쳐 완성된 데이터베이스 설계도에 따라 데이터 비만도를 측정하며, 업계의 관행적 설계방식에 따른다.

가정 5(데이터베이스 관리 시스템의 종류): 설계된 정보 시스템은 가장 보편적으로 사용하고 있는 관계형 데이터베이스관리시스템(RDBMS)으로 구축된다.

2. 연구지원 시스템

실제 대학 연구지원시스템에서의 연구관련 테이블들은 파일형태로 관리되고 있는 실정이었다. 업무기술서조차 먼저 작성하지 않고 연구지원관련 업무에 필요할 만한 데이터만 테이블 형태를 갖춰 모아놓고 사용하고 있었기 때문에 데이터모델링의 결과물인 ERD가 부재했다. 본 논문에서는 현재의 테이블을 구성하고 있는 데이터를 가지고 연구지원팀의 업무를 파악하여 업무기술서를 작성했다.

가. 연구지원 관리 업무기술서

연구지원업무는 연구과제를 수행하는 연구원들이 기자재를 사용하여 연구업적을 주로 관리하는 업무이다. 연구업적을 관리하기 위해서는 예산을 할당받아 업적예산을 관리하고 있으며, 각 연구업적별 업적내역을 관리한다. 각각의 연구원들은 연구소에 소속될 수 있는데, 각 연구소에는 여러 명의 연구원들이 함께 연구를 수행한다. 연구소별 과제를 수행하여 연구실적을 관리하고 있으며, 실적이력도 관리할 뿐만 아니라 연구비에 대한 지출 및 내역을 관리한다.

나. 연구지원 데이터베이스

현행 연구지원 데이터베이스는 연구담당자의 업무위주로 설계되어 있음을 알 수 있었다. 즉 각각의 응용 프로그램에 해당하는 테이블이 존재하고, 테이블들 간의 관계는 관계로만 연결되어 있는 구조로 되어 있다. 연구업적, 업적내역, 연구소실적, 과제내역, 연구소 등의 10개의 테이블들로 구성되어 있다.

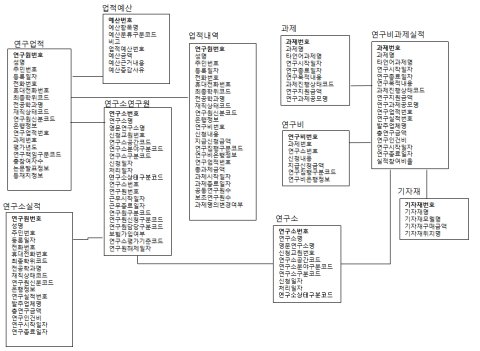

그림 4. 현행 연구지원 데이터베이스
Fig. 4. Current Research Support Database

3. 연구지원 데이터베이스의 문제점

현재 사용 중인 연구지원 데이터베이스를 사용 시 발생할 수 있는 문제점은 필요한 데이터들이 누락되어 널값이 많이 발생하는 문제와 데이터 중복 문제가 있다. 업무 담당자가 데이터 입력 시 발생할 수 있는 중복 입력문제, 각종 통계자료를 추출하기 어려운 문제들을 가지고 있다.

가. 필요데이터 누락 문제

기존의 연구지원 테이블은 업무별로 데이터를 각각의 테이블에 모아놓은 형태로서 하나의 응용에 하나의 파일이 존재하는 것처럼 운영되고 있다. 즉 데이터베이스의 데이터가 중요한 객체그룹들을 포함하고 있지 못하며 객체에 관한 중요한 속성들도 누락하고 있거나 널 값을 가지고 있다.

연구업적 등록 시 연구원정보와 함께 평가년도, 연구책임구분코드, 총참여자수, 논문발표정보, 등재지정보를 동시에 입력되어야 하나, 동시에 입력하기 위한 자료가 부족할 때에는 널 값을 허용할 수밖에 없다. 실제로 많은 속성들에서 널 값이 존재함을 확인하였다.

또한, 연구소별 과제실적을 관리하고자 할 때에도 과제에 대한 내용을 입력하고 연구과제에 대하여 동시에 입력되어야 한다. 이와 같은 중요한 데이터 누락이 발생되도록 테이블이 설계되어 있어서 데이터베이스의 테이

블이라고 할 수 없는 구조를 가지고 있다.

나. 널(null) 값 문제

데이터베이스 테이블에 있는 널 값은 여러 가지 문제들을 가지고 있다. 연구지원 테이블에서는 다음과 같은 문제들을 가지고 있다.

첫째로 연구원이 단순히 연구원 정보만을 등록하고자 할 경우에는 연구원 정보(속성)들만 등록할 수가 없다. 연구원 정보외의 과제정보 및 연구비 관련 정보에 해당하는 속성들에는 일단 널 값을 입력하고 등록할 수 있다. 왜냐하면 연구원 정보만을 등록할 수 있는 테이블이 별도로 존재하고 있지 않고, 다른 속성들과 함께 존재하기 때문이다.

둘째로 연구소별 연구실적 등록 시에 과제정보만을 등록하고자 할 경우에도 연구실적에 해당하는 속성들(과제시작일자, 과제종료일자, 공동연구원수, 보조연구원수, 이하 속성들)은 널 값을 가질 수밖에 없다.

셋째로 테이블 질의 시 처리하는데 문제가 발생할 수 있다. 연구원별 연구업적을 검색하거나, 연구업적별 업적내역을 검색하고자 할 경우에 응용 프로그램에서 연구업적 테이블을 읽어서 조건에 맞는 프로그램이 필요로 한다. 위와 같은 경우에는 잘 설계된 데이터베이스의 경우 한번의 SQL 구문으로 검색할 수 있는 다양한 질의에 대하여 응용 프로그램에서 구현하여야 하는 문제점들을 내포하고 있다.

다. 데이터의 중복

데이터의 중복은 데이터베이스내의 데이터의 정확성과 무결성과 관련이 깊다. 데이터베이스 성능을 향상시키기 위해서는 데이터의 중복을 최소화하도록 하여야 한다. 데이터의 중복으로 데이터의 일관성을 유지하기 어렵고, 데이터베이스 운영 중에 발생하는 데이터 이상현상(anomaly)을 발생시킬 수 있다.

또한, 데이터의 중복으로 데이터의 일관성을 유지하기 어렵기 때문에 데이터의 무결성을 보장할 수 없게 되는 것이다. 예를 들면, 연구지원 테이블에서 연구원 정보들(성명, 주민번호, 전화번호, 최종학위코드, 전공학과명 등)이 몇 개의 다른 테이블에 중복속성으로 정의되어 있기 때문에, 데이터의 변경이 없이 운영되지 않는 한 언젠가는 서로 다른 데이터 값을 가질 수 있는 구조이다. 동일한 데이터를 중복 입력되어야 하는 상황에는 정확한 값을 입력하지 못하면 데이터의 일관성을 유지할 수 없다. 자주 입력되거나 빈번한 갱신이 일어날 때에는 데이터의 무결성은 유지되기가 어렵다.

데이터가 중복되면 여러 테이블에서 동시에 속성 값의 변경이 이루어져야 한다. 그런데 속성 값의 변경으로 데이터 이상 현상이 유발될 수 있다. 어떤 테이블의 한 속성의 값을 변경하면 일관성을 유지하기 위해 다른 테이블의 같은 속성 값도 변경하여야 한다. 연구업적 테이블에서 연구원 정보가 변경된다면, 업적내역 테이블과 연구소실적 테이블에서 있는 연구원 정보도 동시에 변경되도록 하여야 한다.

라. 통계자료 산출의 어려움

잘 정의된 데이터베이스는 단순한 질의어 몇 개로 테이블에 저장된 여러 정보들을 손쉽게 빨리 검색할 수 있다. 연구지원 테이블은 업무중심으로 설계된 파일 시스템과 같이 테이블들이 정의되어 있어서 다양한 요구조건에 대한 질의를 만족시켜 주지 못하는 형편이다. 그래서 연구지원 테이블에서는 연구와 관련된 여러 통계자료를 산출하기 어려운 구조로 되어 있다. 연구원별 연구업적 내역에 대한 통계, 업적예산에 대한 통계, 연구소별연구실적에 대한 통계자료 등을 구하기 위해서는 별도의 응용 프로그램을 개발할 수밖에 없는 구조로 테이블들이 구성되어 있다. 또한, 여러 테이블들의 정보를 이용하여 SQL의 집합연산자로 처리할 수 있는 구조로 설계되어 있지 않기 때문에 이러한 통계자료들을 검출하기가 어려움이 많은 테이블 구성이라고 할 수 있다.

IV. 새로운 연구지원 데이터모델링

3장에서 살펴본 바와 같이 현재 사용 중인 연구관리 테이블은 문제점들을 가지고 있기 때문에 새로운 데이터베이스를 설계하고자 한다. 우선 현업 담당자의 인터뷰 및 업무기술서를 통하여 업무를 파악하고 분석한 후 업무기술서를 재작성하였다.

1. 새로운 연구지원 업무기술서

새로운 업무기술서는 업무 매뉴얼, 업무 담당자와의 인터뷰를 통해 업무를 분석한 후 작성하였다. 연구지원 업무는 9가지 유형의 행위로 구성되어 있다.

(1. 연구지원 대상) 본 연구지원 시스템에서 관리하는 대상은 연구원 정보, 과제 정보, 연구소 정보, 기자재 정보, 예산 정보에 국한하며, 그 외에 필요한 정보나 코드성 객체는 학사행정시스템에서 공통적으로 사용하는 정보를 사용한다.

(2. 연구업적 등록) 연구원은 등록된 과제에서 연구업적을 등록한다. 이 때 평가년도, 연구책임자여부, 총참여자수, 논문발표정보, 등재지정보 등을 관리한다.

(3. 업적예산) 연구업적과 관련된 업적예산을 등록하고 관리한다. 업적예산에는 예산금액, 예산근거, 예산증감사유 등을 관리한다.

(4. 업적내역) 연구원별 과제에 대한 연구업적에 대한 상세 업적내역을 등록하고 관리한다. 총과제금액, 과제시작일자, 과제종료일자, 공동연구원수, 보조연구원수, 과제명의변경여부를 관리한다.

(5. 연구소연구원 등록) 연구소별 연구원을 등록한다. 근무시작일자, 근무종료일자, 연구원구분코드(연구소장, 연구부장, 출판/홍보부장, 교육/훈련부장, 자문/연구원, 총무부장), 연구원신청구분코드(승인, 부결, 보류), 연구원담당구분코드(인문분야, 자연과학분야, 공학분야, 의약학분야, 사회과학분야, 예체능분야), 보험가입여부, 연구소평가기준코드, 연구원해제일자를 속성으로 둔다.

(6. 연구실적) 연구소별 연구실적을 관리한다. 발주업체명, 총연구금액, 연구인건비, 연구시작일자, 연구종료일자를 관리한다.

(7. 연구실적이력) 연구소별 과제에 대한 연구실적이력을 관리한다. 연구실적에 대한 과제별 실적참여비율을 관리한다.

(8. 연구비 관리) 연구소별 과제별, 기자재별 연구비에 대하여 관리한다. 신청내용, 지급신청금액, 연구집행구분코드, 연구비은행정보를 속성으로 둔다.

(9. 연구지원 대상 등록) 연구지원 업무에서 사용되는 대상(객체)들에 대하여는 각 객체별 속성을 도출하여 관리하고, 관련 속성들은 추가하거나 삭제할 수 있도록 한다.

2. 연구지원 데이터베이스 설계모형

작성된 연구지원 업무기술서를 바탕으로 다음과 같은 개체관계도 그림 5를 도출하였다.

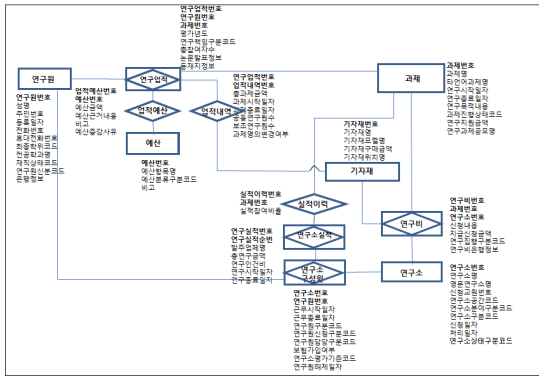

그림 5. 새로운 연구지원 데이터모델링
Fig. 5. New Research Support Data Modelling

3. 문제점이 해결된 새로운 데이터모델링

3장에서 기존 연구지원 데이터베이스의 문제점을 4가지 유형으로 분석하였는데, 이러한 문제점들을 새로운 연구지원 데이터모델링에서 다음과 같이 해결했다.

가. 필요데이터 누락문제 해결

기존의 업무별로 데이터를 각각의 테이블에 모아놓은 형태로 설계된 데이터베이스를 데이터 중심으로 설계함으로 필요데이터에 대한 누락문제는 근원적으로 해결하였다.

연구업적 등록 시 연구원정보는 별도의 연구원 테이블에 입력하고, 연구업적 등록 시에 평가년도, 연구책임구분코드, 총참여자수, 논문발표정보, 등재지정보를 입력하도록 설계하였다. 연구원 테이블과 연구업적 테이블의 관계를 1:다 관계로 설정하여 필요데이터는 누락되거나 널 값을 최소화하도록 데이터베이스를 설계하여 필요데이터 누락문제를 해결하였다.

또한, 연구소별 과제실적을 관리하고자 할 때에도 연구소 정보는 연구소 테이블에 입력하고, 과제에 대한 내용은 연구업적 테이블이나 연구소실적 테이블에 별도로 입력하게 함으로 필요데이터의 누락 문제는 해결되었다.

데이터 중심으로 데이터모델링을 설계함으로 객체는 별도의 테이블로 구분하였고, 객체와 객체간의 관계는 1:다 관계로 맺어서 관계 테이블을 구성함으로 필요데이터의 누락문제는 완전하게 해결되었다고 할 수 있다.

나. 널(null)값 해결 및 데이터 중복율 최소화

기존의 연구지원 테이블은 많은 널 값을 내포할 수밖

에 없었다. 예를 들어, 연구원이 단순히 연구원 정보만을 등록하고자 할 경우에는 연구원 정보(속성)들만 등록할 수가 없었다. 연구원 정보외의 과제정보 및 연구비 관련 정보에 해당하는 속성들에는 일단 널 값을 입력하여야 하였다. 그러나, 새로운 연구지원 테이블에서는 연구원 테이블과 과제 테이블, 연구비 테이블을 별도로 정의하여 구성함으로 각 객체별 데이터를 별도로 관리하게 함으로 널 값 문제는 해결되었다.

둘째로 연구소별 연구실적 등록 시에도 연구소 테이블과 연구소연구실적 테이블을 별도의 테이블로 설계함으로 연구실적에 해당하는 속성들(과제시작일자, 과제종료일자, 공동연구원수, 보조연구원수, 이하 속성들)은 필요할 때 데이터를 입력할 수 있으므로 널 값 문제를 해결하였다.

셋째로 테이블 질의 시 처리하는데 문제가 발생할 수 있었다. 연구원별 연구업적을 검색하거나, 연구업적별 업적내역을 검색하고자 할 경우에 응용 프로그램을 별도로 만들지 않고도 한번의 SQL 구문으로 다양한 질의에 대하여 처리할 수 있도록 설계되었다.

데이터의 중복 문제는 새로운 연구지원 데이터베이스는 제3정규형을 만족하도록 설계되었기 때문에 데이터의 중복을 최소화할 수 있다. 이는 데이터의 정확성과 무결성을 보장하고, 데이터베이스 성능을 향상시킨다. 데이터 중복의 최소화로 데이터의 일관성을 유지하고, 데이터베이스 운영 중에 발생하는 데이터 갱신 이상현상(update anomaly)을 줄일 수 있다.

또한, 데이터 중복의 최소화로 데이터의 일관성을 유지할 뿐 아니라 데이터의 무결성을 보장한다. 예를 들면, 기존의 연구지원 테이블에서 연구원 정보들(성명, 주민번호, 전화번호, 최종학위코드, 전공학과명 등)이 몇 개의 다른 테이블에 중복속성으로 정의되어 있었으나. 연구원 테이블과 연구업적 및 연구소실적 테이블을 별도의 테이블로 정의하고 관계를 맺음으로 연구원 정보들에 대한 중복은 허용되지 않는다.

새로운 연구지원 데이터베이스에서는 여러 테이블에서 동시에 속성 값을 관리하지 않고, 객체 중심으로 테이블을 정의하였기 때문에 한 정보(예를 들어, 연구원 정보, 과제 정보, 연구소 정보 등)은 하나의 테이블에서 관리함으로 데이터의 중복은 허용되지 않는다. 또한, 제3정규형을 만족하도록 설계되었기 때문에 속성 값의 변경으로 데이터 이상 현상이 유발되지 않도록 설계하였다. 그러므로, 기존의 연구지원 데이터베이스에서 야기된 널 값 문제와 데이터의 중복 문제는 해결되었다.

다. 통계자료 산출의 어려움 해결

새로운 연구지원 데이터베이스는 단순한 질의어 몇 개로 테이블에 저장된 여러 정보들을 손쉽게 빨리 검색할 수 있다. 여러 테이블들의 정보를 이용하여 SQL의 집합연산자로 각종 통계를 처리할 수 있는 구조로 설계되어 있다. 새로운 연구지원 데이터베이스 설계는 경사진(skewed) 설계와 고립 형 설계를 지양하고, 제3 정규형의 조건에 맞추어 설계된 정방형 설계 혹은 사각 경로 형 설계의 형태를 지니고 있다. 이러한 새로운 연구지원 시스템에서는 통계자료를 다양한 측면에서 검색할 수 있는 구조이기 때문에 필요한 통계자료를 언제든지 추출해 낼 수 있다. 사각정방형 설계로 인해 빠른 응답시간을 보장하며, 필요한 통계자료를 언제든지 검색할 수 있는 구조로 설계되었다.

4. 데이터 비만도 개선

새로 설계한 그림 5의 연구지원 데이터베이스와 사용중인 그림 4의 연구지원 데이터베이스와의 데이터 중복률을 비교하고자 한다.

데이터 중복률은 데이터베이스 스키마에 포함된 총 컬럼(속성)의 수 중에서 부모-자식 테이블 관계에서 기본키(Primary Key)가 자식 테이블(개체)에 상속되어 참조하는 외래키(Foreign Key)와 같이 데이터베이스 스키마 내에서 서로 중복되는 컬럼들의 개수의 비율을 말한다.

그림 4의 현업에서 운영 중인 연구지원 데이터베이스의 데이터 중복률(data redundancy ratio)은

$$\frac{\text{중복된 속성들의 총 빈도수}}{DB\,Schema\text{에 포함된 총 속성의 갯수}} = \frac{71}{124} = 57.2\%$$

이며, 그림 5의 새로운 연구지원 데이터베이스의 데이터 중복률은

$$\frac{\text{중복된 속성들의 총 빈도수}}{DB\,Schema\text{에 포함된 총 속성의 갯수}} = \frac{14}{89} = 16.7\%$$

로 나타났다.

이와 같이 두 데이터베이스간의 중복률이 40.5 % 차이가 남을 알 수 있었다. 즉, 새로운 연구지원 데이터베이스의 중복률이 현저히 감소되었다. 운영 중인 제1 정규형인 데이터베이스를 제3정규형을 만족하는 데이터베이스로 변환함으로써 데이터의 중복율을 최소화하고 데이터의 저장공간을 최적화할 수 있었다. 또한 외래키를 사용함으로써 갱신 이상현상(update anomalies)을 최소화할 수 있었다.

중복율이 최적 중복율 15%를 기준일 때, 현행 연구지원 데이터베이스가 새로운 연구지원 데이터베이스보다 42% 과다하고 있었다는 것은 현행 연구지원 데이터베이스가 보유하고 있는 데이터 중 42%를 줄이더라도 연구지원 데이터베이스는 문제없이 더 효율적으로 운영될 수 있다고 할 수 있다. 또한, 불필요 중복 데이터의 제거함으로써 질의에 대한 응답속도의 향상을 기대할 수 있고, 최적의 데이터베이스 상태를 유지할 수 있게 된다.

V. 결 론

본 논문에서는 정방형 구조를 가진 최적화된 데이터모델링을 통하여 품질 좋은 데이터베이스를 설계하였다. 또한, 현행 데이터베이스와 새롭게 설계된 데이터베이스의 평균 데이터 비만도를 비교하고 분석하였다. 현행 연구지원 데이터베이스에서 업무위주로 설계된 데이터베이스가 얼마나 많은 데이터의 중복이 존재할 수밖에 없었는지를 알 수 있었다. 이러한 설계의 폐단으로 데이터의 과다한 중복과 더불어 많은 컬럼에서 널 값이 존재할 수 있음을 알았다. 데이터 비만도가 과다하게 나타나는 방사형 설계나 업무중심의 고립형 설계에서 객체(데이터)와 객체간의 관계 중심의 정방형 구조로 설계함으로써 데이터 검색 경로가 선명하게 되었다. 그 결과 기존의 데이터베이스에서의 비만도는 57.2 %였으나, 제안한 새로운 데이터베이스에서는 16.7 %로 나타나 데이터 비만도에서 40.5%나 개선되었다.

본 연구에서의 하나의 연구지원 데이터베이스를 선택하여 제안한 연구지원 데이터베이스의 비만도를 비교 분석하여 데이터 비만도로 일반화하기에는 한계점이 있다. 본 연구를 통하여 데이터모델링은 담당자의 업무 중심이 아니라 데이터 중심으로 설계되고 개발되어야 할 것이다. 데이터베이스 품질은 최적화된 데이터모델링을 통한 정방형 구조로 데이터 비만도를 최소화하고, 데이터베이스의 확장성과 정확성을 동시에 보장하도록 설계되어야 할 것이다.

References

[1] C. W. Fisher, B. R. Kingma, “Criticality of data quality as exemplified in two disasters”, Information Systems, Vol. 39, pp.109-116, 2010.

[2] [DB] Elements of a good data model, http://blog.daum.net/fmddn/1787002, 2012.12.17.

[3] C. B. Cinzia Cappiello, C. Francalanci, A. Maurino, “Methodologies for data quality assessment and improvement”, ACM Computing Surveys Vol. 41, No. 3, p. 52, 2009.

[4] D. Katz, M. Bommaroti, J. Zelner, “The data deluge”, The Economist, Mar 1, 2010

[5] Min-Kyu Lee, “Data Performance Case Study through Removing of Data Duplicate Relationships”, Soongsil University, 2010.

[6] T. Shanker, M. Richtel, “Data overload can be deadly”, The New York Times, Jan 16, 2011.

[7] Richard Y. Wang, Henry B. Kon, and Stuart E. Madnick, “Data quality requirements analysis and modeling”, Proceedings of IEEE Ninth International Conference on Data Engineering, pp. 670 - 677. April 1993.

[8] I. Davies, P. Green, M. Rosemann, M. Indulska, and S. Galo, “How do practitioners use conceptual modeling in practice?”, Data and Knowledge Engineering, Vol 58, pp. 358-380, 2006.

[9] “Practical Projects Application of Data Model Normalization / De-normalization”, http://blog.naver.com/jooyong3/40035951092.

[10] Ji-Ho So, Young-Ju Jeon, “Design and Construction of Integrated Database for Contents Development of Pulse Analysis System”, The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 17, No. 5, pp.

137-142, Oct 2017.
DOI: https://doi.org/10.7236/JIIBC.2005.5.2.56.

[11] In-Hwan Jung, Young-Ung Kim, "ER_Modeler: A Logical Database Design Tool based on Entity-Relationship Model", The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 11, No. 5, pp. 11-17, Oct 2011.
DOI: https://doi.org/10.7236/JIIBC.2005.5.2.56.

[12] Hye-Kyung Rhee, Hee-Wan Kim, "A Study on Negligence of Data Modeling Fundamentals at the University Job Information System", Journal of the Korea Society of Computer and Information, Vol. 19, No. 8, pp.139-150, Aug 2014.

[13] Kook-Hee Lee, "A Study on the Database Quality Assessment", Korea Database Agency, Dec, 1995.

[14] James Martin, Information Engineering, Vol. 1, Vol.2, Vol.3, Prentice-Hall, 1989.

[15] Barker, R, CASE *Method: Entity Relationship Modelling, Wokingham: Addison-Wesley, 1990.

[16] Peter Chen, "The Entity Relationship Model-Toward a Unified View of Data", ACM, Vol.1, No.1, 1976.

저자 소개

김 희 완(정회원)

- 2002년 2월 : 성균관대학교 컴퓨터공학 공학박사
- 1996년 5월 : 정보관리기술사 취득
- 1996년~2000년 : 삼육의명대학 컴퓨터정보과 조교수
- 2000년~현재 : 삼육대학교 컴퓨터메카트로닉스공학부 교수

<관심분야> : 데이터베이스, 정보시스템 감리, 데이터보안