The Practical Need for Fourth Normal Form

Margaret S. Wu
425 Beldon Ave
Iowa City, Iowa 52246
Phone: (319) 335-0846

ABSTRACT

Many practitioners and academicians believe that data violating fourth normal form is rarely encountered. We report upon a study of forty organizational databases; nine of them contained data violating fourth normal form. Consequently, the need to understand and use fourth normal form is more important than previously believed.

INTRODUCTION

A paramount issue in the design of any database is what data fields should be grouped together into records. In the relational model, the data fields are grouped into logical structures called relations. The determination of which data fields are placed together in a relation is based upon the concept of normal forms; the process is known as normalization. The set of data fields comprising the database is progressively organized into relations in first through fifth normal form (5NF) according to constraints placed upon the relations in each normal form. At any step in the normalization process, we may find that the relations no longer require further reorganization; in that case, we say that final normal form has been achieved. Frequently, a set of data may be in final normal form when it has been normalized only to third normal form (3NF). Because the definition of 3NF does not treat certain instances of data adequately, an additional normal form called Boyce-Codd normal form (BCNF) was created to replace 3NF. In theory, BCNF is placed after 3NF in the normalization process. There is an additional normal form called Domain Key normal form which is not applied to normalize data but serves to verify that a relation has been normalized to its final form by showing that the relation has no modification anomalies.

There is some evidence that academicians view fourth normal form (4NF) as unimportant and thus may neglect the topic in database management courses in MIS. Stamper and Price in [10] state that "fourth and fifth normal forms are so rarely encountered in business applications as to be almost obscure; hence, they are not described in this book." Mittra [8] states that "Although BCNF and 4NF may be less rare than 5NF, they are still highly theoretical in nature." Other database texts [1,6,9] either ignore or do not fully explain 4NF. The view of practitioners regarding 4NF is represented by Edwards in [3] where he states that 4NF is merely an academic issue. Thus, the practical use of normalization theory apparently stops with 3NF or BCNF for some academicians and practitioners. Because the normal forms were defined in order to prevent data inconsistencies and update anomalies, the termination of the normalization process with the database violating 4NF will result in record designs that are ineffective in meeting these objectives.

The purpose of this paper is to present the results of an empirical study which investigated the practical need for 4NF by the examination of real world databases. In the next section, we state the normalization procedures for 4NF and discuss the advantages of 4NF. The second section details the research methodology of this study. In the third section, we present the results of the empirical research. Additional insights into the organization of data gained during this investigation are presented in the fourth section. Finally, a summary of these results is given in the last section.

NORMALIZATION OF DATA

To normalize a set of data items, we may proceed from first normal form (1NF) to second normal form (2NF) to third normal form (3NF) to Boyce-Codd normal form (BCNF) to fourth normal form (4NF). In practice, we can proceed directly from 1NF to BCNF. The use of BCNF rather than 3NF is preferred because normalization to BCNF is easily understood by students and can be quickly implemented; in addition, a relation in BCNF is also in 3NF.

To normalize a record type to 1NF, any repeating fields or groups of fields are repeated for each record. The only attribute values permitted by 1NF are single atomic values. A relation schema R is in Boyce-Codd normal form if whenever a functional dependency X -> A holds in R, then there do not exist two tuples t_1 and t_2 in R such that $t_1[X] = t_2[X]$. X is called a superkey of R. To define fourth normal form, the concept of multivalued dependency (MVD) is required. An MVD X ->-> Y holds for R(X,Y,Z) if and only if whenever (x,y,z) and x,y',z') are tuples of R, then so are (x,y',z) and (x,y,z'). An MVD X ->-> Y is said to be trivial if Y is a subset of X or X∪Y = R.

The definition of 4NF is then given by Fagin [5] as follows:

> A relation schema R^* is in 4NF if, whenever a nontrivial MVD X ->-> Y holds for R^*, then so does the functional dependency X -> A for every column name of R^*.

If X ->-> Y, we also say that Y is a multivalued fact about X. It is possible for Y to be called a multivalued fact about X and yet the relationship between X and Y is not a multivalued dependency. When a multivalued fact is not independent, i.e., a multivalued dependency does not exist, redundancy of data values is necessary. To obtain 4NF, the relation R is decomposed into two or more relations which meets the contraints given by the definition of 4NF.

For the purpose of normalizing real world data, we may restate the definition of 4NF as follows:

> A relation R is in fourth normal form if R is in BCNF and does not contain any nontrivial multivalued dependencies.

We note that this simpler statement does not conform fully to the original definition given by Fagin [4]. However, the theoretical differences do not affect the normalization of real world databases.

Table 1

The nine databases with data violating 4NF

Organization	Functional Area	Total * Fields	Record Types
Financial Firm	Sales Management System	307	24
Physician Database	Data on Individual Physician	68	6
Publishing Firm	Job Scheduling	63	9
College of Nursing	Student Skills Database	55	8
Apartment Complex	Maintenance Tracking	47	17
Division of Sponsored Research	Grants Database	42	7
Animal Laboratory	Inventory of Animals	33	5
Athletic Tutoring Program	Assignment of Tutors	18	7
Automotive Parts Retailer	Inventory	16	4

RESEARCH METHODS

The databases from forty organizations were studied for this project. Each organization was a small firm, a division of a large firm, or a division of a major university. The database for each organization was limited to the data required for the functional area of the organization with the exception of one business firm which was studied in its entirety. Eighteen organizations were functional units within a large university, twenty were private firms, and the remaining two were government organizations. The study included the entire database for a reprographics firm, the manufacturing database for a Fortune 500 firm, and the sales management system for a major financial firm. These three large databases were examined for data that violated 4NF. The remainder of the organizations were studied by student teams for undergraduate classes in the Systems Analysis and Design course. Each team was required to present the data as a set of record types in 3NF. These reports were then carefully reviewed and revised by the author as necessary to obtain record types in 4NF. Additional contact with the organizations was made to resolve any ambiguities in the data and the relationships between the data.

RESULTS OF THE STUDY

This study has determined that data violating 4NF occurred at least once in nine organizational databases constituting over twenty percent of the databases in this study. Twenty of the 350 record types comprising the forty databases resulted from normalization of BCNF relations to 4NF. These record types represent six percent of the normalized record types in this study. Because of this incidence, 4NF is not so rare.

Table 1 provides information regarding the data found in the organizational databases which contained data violating 4NF. The number of data fields for each organizational database has been tabulated. The number of normalized record types for each organization is also given. The tabulation of the number of data fields found at each organization yields a general picture of the magnitude of these databases. Both large and small databases were included in the study. The database represented by the 307 data fields required for the sales management system of a major financial firm was found to violate 4NF. The other eight databases contain an average of over 40 data fields.

AVOIDANCE OF FOURTH NORMAL FORM

It is sometimes possible to avoid the use of fourth normal form by representing the data differently. As an example of this approach, let us consider the database of the Financial Organization. This database in BCNF contains the following relation violating fourth normal form:

Codes (ACCOUNT ID, HOME PENDING CODE, FIELD PENDING CODE)

To obtain 4NF, this relation is decomposed to form two relations:

Home-Codes (ACCOUNT ID, HOME PENDING CODE)
Pending-Reqts (ACCOUNT ID, FIELD PENDING REQUIREMENT)

It is possible to represent the attributes of HOME PENDING CODE and FIELD PENDING REQUIREMENT so that the data does not violate 4NF. The value of these fields is given by a numeric value that indicates a particular item of information has not been received by the organization. For example, if the value of 01 is present for HOME PENDING CODE for a particular value of ACCOUNT ID, we know that the document identified as 01 has not been received. To represent this information without violating fourth normal form, we will require that all values for HOME PENDING CODE be given and an additional field called HOME-PENDING-CODE-Y-N be included. This latter field will contain either a Y or N value to indicate whether a particular document identified by its corresponding value in the HOME PENDING CODE field has been received or not. Similarly, we add the field P-REQTS-Y-N to contain a Y or N value corresponding to a particular value of FIELD PENDING REQUIREMENT. We then have two relations as shown below:

Home-Code-YN (ACCOUNT ID, HOME PENDING CODE, HOME-PENDING-CODE-Y-N)
Pending-Reqts-YN (ACCOUNT ID, FIELD PENDING REQUIREMENT, P-REQTS-Y-N)

These two relations do not violate fourth normal form and require no further normalization.

Another representation for this data is possible by the use of buckets. In actuality, the HOME PENDING CODE field has only 19 possible codes while the FIELD PENDING REQUIREMENT field has only five possible codes.

Table 2

The Violations of Fourth Normal Form

In this research study, nine databases were found to contain data violating fourth normal form. The relevant data for these nine organizations is shown below in relations normalized to 4NF.

1. Financial Organization - Sales Management System

   Home-Codes (ACCOUNT ID, HOME PENDING CODE)
   Pending-Reqts (ACCOUNT ID, FIELD PENDING REQUIREMENT)

2. Physician Identification System

   Is-Member (PHYSICIAN ID, GROUP ID)
   Works-At (PHYSICIAN ID, LOCATION ID)

3. Publishing Firm - Scheduling System

   Associated-Job (JOB NUMBER, ASSOCIATED JOB NUMBER)
   Colors (JOB NUMBER, COLOR REQUIRED)

4. College of Nursing - Student Skills Database

   Has-Prerequisite (SKILL CODE, PREREQUISITE SKILL CODE)
   Course (COURSE, SKILL CODE)

5. Family Housing Office - Maintenance Tracking System

   Adjoining-Apt (APT CODE, ADJOINING APT CODE)
   Maintenance (APT CODE, MAINTENANCE CODE, MAINTENANCE DATE)
   Spraying (APT CODE, SPRAY DATE, AREA SPRAYED)

6. Division of Sponsored Research - Grants Database

   Grant-Area (GRANT CODE, DISCIPLINE)
   Grant-Keyword (GRANT CODE, KEYWORD)
   Faculty-Interests (UNIVERSITY ID, KEYWORD)

7. Animal Laboratory - Inventory of Animals

   Account-Investigators (ACCOUNT NUMBER, INVESTIGATOR NUMBER)
   Account-Animals (ACCOUNT NUMBER, ANIMAL ID)

8. Athletic Tutoring System - Tutor Assignment

   Note: The Athletic Tutoring Service did not wish to classify individual courses by subject area(s) but did wish to track both courses and subject areas for which a student was deemed qualified to serve as a tutor.

   Tutor-Subjects (TUTOR ID, SUBJECT AREA)
   Tutor-Courses (TUTOR ID, COURSE ID)

9. Automotive Parts Retailer - Inventory

   Vehicle-Engine-Part (VEHICLE, MODEL, YEAR, ENGINE TYPE, ENGINE FEATURE, ENGINE-PART-NUMBER)
   Vehicle-Feature (VEHICLE, MODEL, YEAR, FEATURE, F-PART NUMBER)

   Note: The field FEATURE consists of several subfields.

Consequently, we may choose to store this same information in one relation as shown below:

All-Codes (ACCOUNT ID, HOME-PENDING-CODE-1, ..., HOME-PENDING-CODE-19, FIELD-PENDING-REQUIREMENT-1, ..., FIELD-PENDING-REQUIREMENT-5)

Although the relation All-Codes is technically in 4NF, the use of the bucket violates the principles of the relational model. The organization may choose the bucket representation because these data fields have been in use for several years and it is unlikely that change would occur. The bucket representation presents potential problems due to its inflexibility. We note that the proper choice for data representation is dependent on the individual installation's preference for data representation, the requirements for flexibility, the need to minimize storage space, and the trade-off of data structure for ease of programming.

CONCLUSION

The study of forty real world databases showed that 4NF occurs more often than we previously believed. Data violating 5NF did not occur in any of the databases which may indicate that 5NF is only an academic issue. Because data violating 4NF was found to occur in over twenty percent of the organizations in the study, it is advisable that practitioners understand and apply the constraints for 4NF. It is also imperative that the rules for 4NF be emphasized in the teaching of normalization theory.

ACKNOWLEDGEMENT

The author wishes to thank Jeff Hoffer for suggesting this research project.

REFERENCES

1. Courtney, James F. and David B. Paradice. Database Systems for Management. St. Louis: Times Mirror/Mosby College Publishing Co., 1988.
2. Date, C. J. An Introduction to Data Base Systems. 5th ed. Reading, MA: Addison-Wesley Publishing Co., 1990.
3. Edwards, Joe B. "High Performance Without Compromise," Datamation, 36, 13, (July 1, 1990), 53-58.
4. Elmasri, Ramez and Shamkant B. Navathe. Fundamentals of Database Systems. Redwood City, CA: Benjamin/Cummings Publishing Co. 1989.
5. Fagin, Ronald. "Multivalued Dependencies and a New Normal Form for Relational Databases," ACM Transactions on Database Systems, 2, 3, (Sept. 1977), 262-278.
6. Fife, Dennis, W. Terry Hardgrave, and Donald R. Deuth. Database Concepts. Cincinnati: Southwestern Publishing Co., 1986.
7. Kent, W. "A Simple Guide to Five Normal Forms in Relational Data base Theory, Communications of the ACM, 26, 2, (Feb. 1983), 120-125.
8. Mittra, Sitansu S. Principles of Relational Database Systems. Englewood Cliffs, N. Y.: Prentice-Hall, Inc., 1991.
9. Pratt, Philip J. Microcomputer Data Base Management Using dBASE III Plus. Boston, MA: Boyd & Fraser Publishing Co., 1988.
10. Stamper, David and Wilson Price. Database Design and Management: An Applied Approach. New York: McGraw-Hill, Inc., 1990.