Journal of The Korea Society of Computer and Information
Vol. 22 No. 11, pp. 57-63, November 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.11.057

Development of the design methodology for large-scale database based on MongoDB

Jun-Ho Lee*, Kyung-Soo Joo**

Abstract

The recent sudden increase of big data has characteristics such as continuous generation of data, large amount, and unstructured format. The existing relational database technologies are inadequate to handle such big data due to the limited processing speed and the significant storage expansion cost. Thus, big data processing technologies, which are normally based on distributed file systems, distributed database management, and parallel processing technologies, have arisen as a core technology to implement big data repositories. In this paper, we propose a design methodology for large-scale database based on MongoDB by extending the information engineering methodology based on E-R data model.

▸Keyword: Big Data, Data Modeling, NoSQL Database, MongoDB

I. Introduction

급격히 증가하는 정보량으로 기존의 관계형 데이터베이스에 모든 정보를 저장하기 어려운 상황이 되어 빅 데이터 데이터베이스가 등장하게 되었다.

이러한 빅 데이터를 처리하기 위한 데이터베이스는 NoSQL(Not only SQL) 데이터베이스라고 부르며 그 종류는 MongoDB[1], Hadoop[2], Apache의 Casandra[3], HBASE[4] 등 여러 가지가 존재한다. 그 중 MongoDB의 경우 Document방식의 데이터모델을 채택하고 있는데, 이는 데이터베이스를 일종의 문서 형태로 관리한다.

NoSQL 데이터베이스들은 기존의 관계형 데이터베이스와 다르게 대규모의 데이터를 유연하게 처리할 수 있는 것이 강점이다. 한편, 이러한 NoSQL 데이터베이스를 위한, 정보 요구사항 분석부터 데이터베이스 구축까지의 일관적인 설계방법들이 존재하지 않아, 현재 설계방법을 위한 여러 연구가 진행 중에 있다[5, 6, 7].

본 논문에서는 관계형 데이터베이스 설계를 위한 E-R 데이터 모델 기반의 정보공학 방법론(Information Engineering Methodology)을 확장하여 NoSQL 중 하나인 MongoDB 기반의 대규모 데이터베이스를 위한, 정보 요구사항 분석부터 데이터베이스 구축까지의 일관된 설계 방법론을 개발하여 제안한다. 제안한 방법론을 적용하여 MongoDB를 구축하기 위한 데이터 셋으로는 통계청에서 제공하는 교사 및 학교 관리자를 대상으로 한 기초학력 진단-보정 시스템 설문조사 응답 내용을 사용하였다.

구체적인 확장 내용은, 관계형 데이터베이스를 위한 정보공학 방법론 중 PDM 단계를 MongoDB를 위한 PDM 단계로 변경하여 제안한 설계 방법론을 완성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 정보공학 방법론 및 MongoDB에 대해 기술하고, 3장에서는 본 논문에서 제안하는 E-R 데이터 모델을 기반으로 한 정보공학 방법론을 확장하여 MongoDB 기반의 대규모 데이터베이스를 위한, 정보 요구사항 분석부터 데이터베이스 구축까지의 일관적인 설계 방법론을 설명한다. 그리고 제안한 모델로 실제 데이터를 이용해 MongoDB에 적용하고, MongoDB 스키마를 보여준다. 4장에서는 결론 및 향후 연구에 대해 기술한다.

• First Author: Jun-Ho Lee, Corresponding Author: Kyung-Soo Joo
*Jun-Ho Lee (wnsgh461@naver.com), Dept. of Computer Science, Soonchunhyang University
**Kyung-Soo Joo (gsoojoo@naver.com), Dept. of Computer Software Engineering, Soonchunhyang University
• Received: 2017. 09. 27, Revised: 2017. 10. 23, Accepted: 2017. 11. 22.
• This work was supported by the Soonchunhyang University.

II. Related works

2.1 Information Engineering Methodology

정보공학 방법론은 정보 요구사항과 어플리케이션 요구사항을 병행하여 분석하는데 데이터베이스 설계를 위한 정보 요구사항은 Fig. 1과 같이 데이터 모델링(Conceptual Data Modeling), 데이터베이스 설계(Logical Data Modeling), 데이터베이스 구현(Physical Data Modeling)의 세 단계로 수행 된다.[8]

데이터 모델링 단계는 E-R Data Model을 통해 엔티티를 정의한다. 데이터베이스 설계 단계는 정의된 엔티티를 사용하여 데이터베이스의 테이블을 정의한다. 마지막, 데이터베이스 구현 단계는 실제 사용 가능한 데이터베이스를 구현한다.

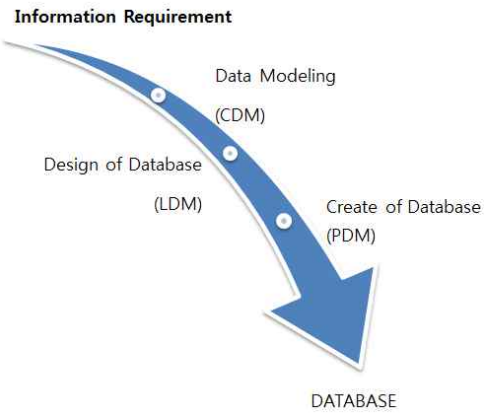

Fig. 1. Information Engineering Methodology

2.2 MongoDB

MongoDB는 10gen이라는 기업에서 2009년 개발한 NoSQL 방식의 빅 데이터 처리에 탁월한 성능을 가진 오픈소스 데이터베이스이다[9]. 필요한 쿼리 및 인덱싱을 통해 원하는 확장성과 유연성을 갖춘 Document 형태의 데이터 모델을 사용한다[10].

MongoDB는 아래 Table 1과 같은 특징을 가진다.

Table 1. Special Feature of MongoDB

1. JSON type data storage structure is provided. 2. Sharding / Replica function provid. 3. MapReduce function provid. 4. CRUD(Create, Read, Update, Delete) centered multiple transaction processing is possible. 5. Based on Memory Mapping technology, it provides excellent performance for Big Data processing.

MongoDB는 JSON 타입의 데이터 표현방식을 사용하여 데이터를 저장한다. 또한, 관계형 데이터베이스의 주요 기능인 CRUD 위주의 다중 트랜잭션 처리도 가능하다. 하지만 관계형 데이터베이스가 데이터를 보다 효율적으로 처리하기 위한 기술이라면, NoSQL은 빅 데이터의 빠른 저장과 추출 및 분석을 위한 기술로, 그 용도가 다르다. MongoDB는 기존의 관계형 데이터베이스와 용어 면에서도 차이를 가지고 있는데, 아래 Table 2에서 그 차이를 확인할 수 있다[11].

Table 2. Term differences between MongoDB and Relational Database

MongoDB	RDB
Collection	Table
BSON Field	Column
BSON Document	Row
Embedded & Linking	Relation Ship

III. Development Data Modeling Methodology for MongoDB and is Application

데이터 모델은 데이터베이스에 필요한 데이터 구조의 개념적 표현이다. 데이터 구조에는 데이터 객체, 데이터 객체 간의 연결 및 객체에 대한 작업을 제어하는 규칙이 포함된다. 데이터 모델의 목표는 데이터베이스에 필요한 모든 데이터 객체가 완전하고 정확하게 표현되도록 하는 것이다. 데이터 모델은 쉽게 이해할 수 있는 표기법과 자연어를 사용하기 때문에 사용자가 정확하게 검토하고 확인할 수 있다[11].

본 논문에서는 MongoDB 데이터베이스를 설계함에 있어, 하향식 접근 방식을 사용하는데 이는 개념적 데이터 모델로 시작하여 논리적 모델, 물리적 모델을 거쳐 MongoDB 데이터베이스를 만드는 것으로 끝난다[12]. 적용할 실제 데이터는 통계청에서 제공하는 기초학력 진단-보정 시스템 설문조사 내용을 사용하였다.

3.1 CDM(Conceptual Data Modeling)

핵심 개념과 관계를 토대로 시스템의 범위를 파악하기 위해 개념 데이터모델(Cconceptural Data Model, CDM)을 작성한다. CDM 모델링 과정은 아래 Fig. 2와 같다[13].

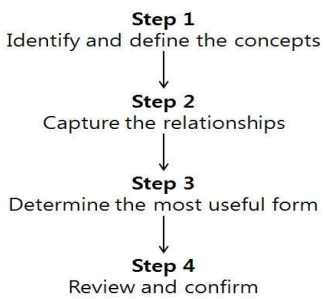

Fig. 2. Flowchart of CDM

개념적 데이터 모델링의 첫 번째 단계로 개념 템플릿을 만든

다. 개념 템플릿이란, 데이터 모델링에 필요한 각각의 개념들을 정리하는 것 이다. 개념적 데이터 모델링의 첫 번째 단계는 모든 개념을 확인하고 정의하는 단계에서 Table 3과 같은 개념 템플릿을 완성 할 수 있다. Table 3에서 결정한 각각의 개념들에 대한 정의는 다음 Table 4와 같은 형식으로 정리한다.

Table 3. Concept Template

Who	What	When	Where	Why	How
Organization	Survey				
Industry	Survey Category	Survey Completion Date			Completed Survey
Survey Respondent	Survey Section				
	Survey Question				

Table 4. Definitions for each of these concepts

Completed Survey	One filled in survey that contains a collection of opinions from a survey respondent in reaction to service.
Industry	The general sector in which an organization operates.
Organization	The company or government agency that needs the survey.
Survey	A questions designed to be completed by an single for improvement.
Survey Category	Category of Survey.
Survey Completion Date	The date that an individual filled in the survey.
Survey Question	The inquiry an organization uses to seek feedback.
Survey Respondent	The respondent who completes the survey.
Survey Section	A logical grouping within the survey.

두 번째 단계에서는 각 개념들 간의 관계를 파악하는 단계로써, 각 개념들의 관계를 식별 한 후 그림으로 정리하여 다음 Fig. 3과 같이 표현할 수 있다.

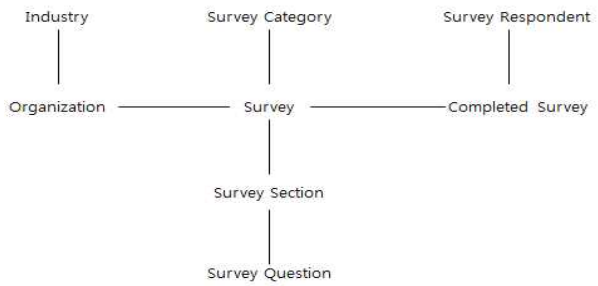

Fig. 3. Relationship of concepts

세 번째 단계는 파악한 개념들의 관계를 토대로 해당 데이터 모델을 표현하기 위한 표기법을 결정해야 한다. 정보공학 방법론에 의해 표기할 수 있고, Axis기법 이라고 하는 비즈니스 친화적 모델링 형식을 사용 할 수도 있다[13]. 본 논문에서는 정보공학 방법론에 의해 표기하고, 다음 Fig. 4의 결과를 얻었다.

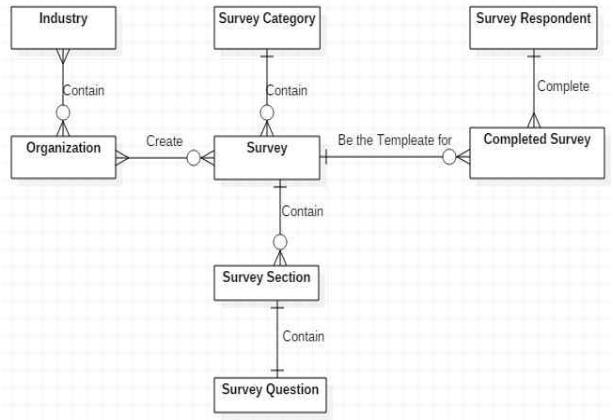

Fig. 4. Relationship Conceptual Data Model

네 번째 확인 및 검토 단계에서는 도출된 모델이 올바른지 확인하는 단계이다. 올바른 모델이라고 판단 될 때까지 위의 모델링 방법을 반복 적용하여 수정할 수 있고, 올바르다고 판단된다면 논리적 모델링 단계로 넘어간다.

3.2 LDM(Logical Data Modeling)

논리적 데이터 모델링(Logical Data Modeling, LDM)은 상세 내용을 파악하는 단계로 CDM 과정에서 나온 데이터 모델을 토대로 만들어지게 된다. LDM 과정은 Fig. 5와 같다.

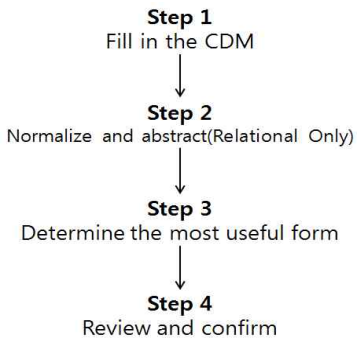

Fig. 5. Flowchart of LDM

논리적 데이터 모델의 첫 번째 단계는 개념의 속성을 식별하고 정의하는 것이다. CDM에서 정의한 개념은 LDM의 엔티티가 되며, 이 단계를 마치게 되면 속성 템플릿과 속성 특성 템플릿을 얻을 수 있다. 각 엔티티의 속성 또는 데이터요소를 파악하여 속성 템플릿을 작성한다. 작성된 템플릿은 각각 Table 5, Table 6과 같다.

Table 5. Properties Template

	Industry	Organization	Survey	Survey Category	Survey Section	Survey Question	Survey Respondent	Completed Survey
Name	Id_name	Org_name		Sc_name	Ss_name	Sq_l_name	Sr_name	
Text					Ss_description	Sq_description, Poss_An_valeue, Poss_An_discription		Comp_S_Free_Text, Comp_S_Fiexed_answer
Date		Org_First_S_date						Comp_S_date
Code	SIC_code		S_code	Sc_code				
Number		Ogr_DUNS_num				Sq_num		
Identifier							Sr_id	
Indicator						Sq_singular_Respon_indicator		

Table 6. Properties Characteristics Template

Property	Definition	Sample Value	Format	Length
Id_name	The common term used to describe the general sector an organization operations within. This is the standard description for the SIC code.	Computer programming services	Char	50
SIC_code	The Standard Industry Classification (SIC) is a system for classifying industries by a four-digit code. it is used by government agencies to classify industry areas.	62010 [Computer programming services]	Char	6
Org_name	The common term used to describe the company or government agency that needs the survey.	Google Naver	Char	50
Org_First_S_date	The date when the organization first started using survey.	sep-08-2015	Date	
Ogr_DUNS_num	Dun & Bradstreet(D&B) provides a DUNS Number, a unique nine digit identification number, for each organization.	123456789	Char	9
S_code	The unique and required short term referring to a survey.	A001-A	Char	6
Sc_name	The common term used to describe the driver for the survey.	Consumer Feedback	Char	50
Sc_code	The unique and required short term to describe the driver for the survey such as employee satisfaction.	AA BB	Char	2
Ss_name	The common term used to describe the logical grouping within the survey.	General	Char	50
Ss_description	The detailed text explaining the logical grouping within the survey. This is not displayed on the survey form.	Contains those questions pertaining to overall using experience.	Char	255
Sq_l_name	What the survey respondent sees on the form. That is, the question that appears.	What about the overall user interface?	Char	255
Sq_description	An explanation or background on the question, which is not displayed on the survey form.	This question lets the respondent rate user interface.	Char	255
Poss_An_valeue	Certain questions have a fixed response such as the Gender question for “Male” or “Female” or the “From 1 to 5. This field stores all of the possible fixed responses.	Male Female	Char	50
Poss_An_discription	This field stores the meaning for each of the Possible Answer Values.	1 means poor 3 means Average	Char	100
Sq_num	A required number assigned to each question. This number is unique within a survey.	1 2	Integer	3
Sq_singular_Respon_indicator	Some questions allow for more than one response.	Y N	Boolean	
Sr_name	The name the survey respondent writes on the completed survey.	JunHo Lee KyungSoo Joo	Char	100
Sr_id	A unique and required value for each survey respondent.	000001	Integer	6
Comp_S_Free_Text	Captures the responses to those questions that do not require a fixed response.	management of students	Char	255
Comp_S_Fiexed_answer	Captures the responses to those questions that require a fixed response.	Male Female	Char	100
Comp_S_date	The date the survey respondent completed the survey.	Sep-09-2017	Date	

두 번째 단계로는 정규화 및 추상화 작업을 실행한다. 추상화는 모델내의 속성과 엔티티 그리고 관계를 재정의하고 결합하는 과정으로써, 설계에 유연성을 줄 수 있다[13].

Table 5와 6의 템플릿을 작성한 후 정규화를 거친 결과 데이터 모델은 아래 Fig. 6과 같다.

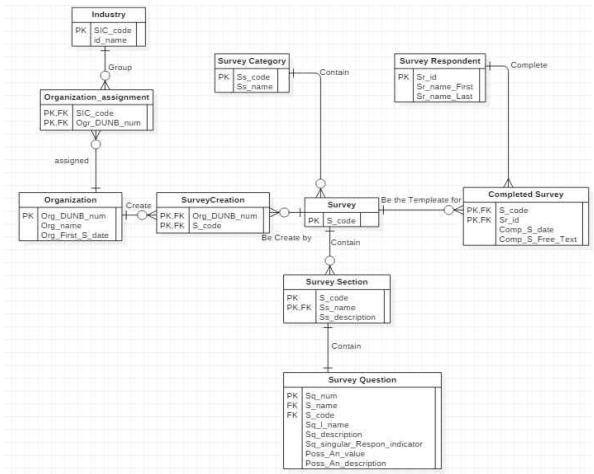

Fig. 6. Logical Data Model

세 번째, 네 번째 단계는 CDM 단계와 동일하게 유용한 표기법을 결정하고 데이터 모델을 확인하는 단계이다.

3.3 PDM(Physical Data Modeling)

물리적 데이터 모델링(Physical Data Modeling, PDM)은 실제 MongoDB 컬렉션을 확정하는 단계로, LDM 과정에서 나온 데이터 모델을 토대로 만들어지게 된다. PDM 과정은 아래 Fig. 7과 같다.

Fig. 7. Flowchart of PDM

물리적 데이터 모델링의 첫 번째 단계로 LDM 단계에서 나온 논리적 데이터 모델을 기초로 임베드(Embed) 또는 참조(Reference) 여부를 결정하여 초기 MongoDB 컬렉션을 준비한다. Industry와 Organization은 비슷한 역할을 하기 때문에 하나의 컬렉션으로 롤업(Roll-Up)한다. 또한, Survey Respondent와 Completed Survey는 비슷한 변동성을 가지기 때문에 롤업한다. 그리고 Survey Category와 Survey, Survey Section은 서로 의존 엔티티이고, 변동성이 비슷하기 때문에 하나의 엔티티로 롤업한다. 최종적으로, Fig. 8과 같은 데이터 모델을 가지게 된다.

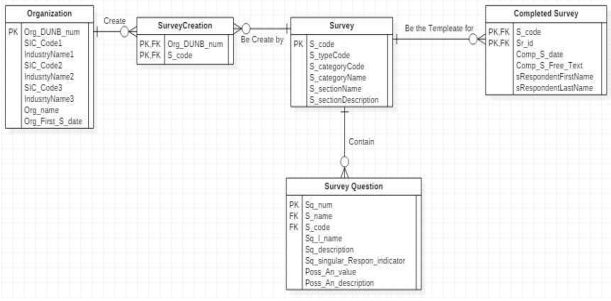

Fig. 8. Physical Data Model

두 번째 단계는 Accommodate History로 필드 값이 시간에 따라 어떻게 변하는지에 대한 옵션을 설정한다. 필드 값의 변화를 다루기 위한 옵션은 SCD(Slowly Changing Dimension)로 불리며, 숫자(0, 1, 2, 3)로 구성된다.

① SCD 0 : 원 상태를 저장하고 변경사항은 저장하지 않는다.

② SCD 1 : 가장 최근의 상태를 저장한다.

③ SCD 2 : 데이터의 변경사항이 있을 때마다, 모든 변경사항을 컬렉션에 저장한다.

④ SCD 3 : 일부 기록에 대해 요구사항(최신보기, 이전보기 등...)이 있다.

각 컬렉션에 필요한 옵션을 결정해 Table 7의 컬렉션 기록 템플릿을 만든다[11].

Table 7. Collection History Template for the Case

	Type 0	Type 1	Type 2	Type 3
Organization		✔		
SurveyCreation		✔		
Survey			✔	
Completed Survey	✔			
Survey Question		✔		

세 번째 단계는 인덱싱(Indexing)과 샤딩(Sharding)이다. 인덱싱은 논리적 데이터 모델의 기본키와 대체키를 고유 인덱스(Unique Index)로 변환하는 것이고, 성능을 위한 인덱스를 추가 할 수 있다. 샤딩은 컬렉션이 두 개 이상의 부분으로 분리되는 경우를 말하며 수평, 수직 두 가지의 분할 방식이 있다[11].

네 번째 단계는 확인 및 검토 단계로 해당 모델이 올바르게 작성된 모델인지 검토한다.

다섯 번째 단계는 설계된 데이터 모델을 토대로 MongoDB 컬렉션을 정의한다[13].

3.4 MongoDB Collection

3.3절의 물리적 모델을 토대로 설계된 MongoDB 컬렉션을 정의할 수 있다. 물리적 모델의 엔티티를 기반으로 설계되었으며, 다음 Table 8, 9, 10, 11, 12와 같이 MongoDB의 컬렉션을 정의한다.

Table 8. Organization Collection

```
Organization:
{
    Org_DUNS_num : “123456789”,
    Industry : [
    {
        SIC_code : “000013”,
        id_name : “Elementary School” ,
        SIC_code : “000014”,
        id_name : “Middle School”
    ] },
    Org_name : “Seoul Middle School”,
    Org_First_S_date : ISODate(“2017-08-01”)
}
```

Table 9. Survey Collection

```
Survey:
{
    S_code: “A001-A”,
    S_typeCode : “AA”,
    S_categoryCode : “CF”,
    S_categoryName : “Consumer Feedback”,
    S_sectionName : “Program Experience”,
    S_sectionDescription : “Contains those questions pertaining
to overall using experience.”
}
```

Table 10. SurveyCreation Collection

```
SurveyCreation:
{
    Org_DUNS_num : “123456789”,
    S_Code : “A001-A”,
}
```

Table 11. Survey Question Collection

```
Survey Question:
{
    Sq_num : “1”,
    S_code : “A001-A”,
    Sq_l_name : “What about the overall user interface?”,
    Sq_Description : “This question lets the respondent rate user
interface.”,
    Sq_Singular_Respons_Indicator : “Y”
}
```

Table 12. Completed Survey Collection

```
Completed Survey:
{
    S_code : “A001-A”,
    Sr_id : “123456”,
    Comp_S_date : ISODate(“2017-09-05”),
    sRespondentFirstName : “JunHo”,
    sRespondentLastName : “Lee”
}
```

V. Conclusions

본 논문에서는 E-R 데이터 모델을 기반으로 한 정보공학 방법론을 확장하여 MongoDB 기반의 대규모 데이터베이스를 위한, 정보 요구사항 분석부터 데이터베이스 구축까지의 일관된 설계 방법론을 개발하여 제안하였다. CDM 단계에서 질문을 통해 개념을 정의하고, 관계를 식별하여 개념적 데이터 모델링을 한 후, LDM 단계에서 각 엔티티에 대한 속성을 정의하고, 정규화를 통하여 논리적 데이터 모델을 완성한다. 마지막으로, PDM 단계를 거쳐 인덱싱과 샤딩을 통해 MongoDB 기반의 대규모 데이터베이스에 적합한 물리적 모델을 완성하고 MongoDB 스키마를 정의하게 된다. 모든 과정은 질의응답 과정을 거쳐 만들어지며 각 단계별 확인 및 검토를 통해 완료된다.

본 논문에서 제안한 설계 방법론을 적용하여 구축한 MongoDB 기반의 대규모 데이터베이스는 샤딩을 통한 분산시스템의 구축이 가능하여 기존의 관계형 데이터베이스보다 확장성면에서 이점을 가진다.

4.1 Future Work

향후 연구에서는 스타 스키마를 이용한 관계형 데이터베이스 기반의 데이터웨어하우스 모델링 방법론을 확장하여 MongoDB 기반의 대규모 데이터웨어하우스 모델링을 위한 방법론에 대하여 연구할 예정이다.

REFERENCES

[1] MongoDB, https://www.mongodb.org

[2] Hadoop, http://hadoop.apache.org/

[3] Casandra, http://cassndra.apache.org/

[4] HBASE, http://hbase.apache.org/

[5] S. Wen, Y. Xue, J. Xu, H. Yang, X. Li, W. Song, and G. Si, “Toward exploiting access control vulnerabilities within mongodb backend web applications,” in IEEE Annual Computer Software and Applications Conference, vol. 1, pp. 143-153, 2016.

[6] Zhao, Gansen, et al. "Modeling MongoDB with relational model," Emerging Intelligent Data and Web Technologies(EIDWT), Fourth International Conference on IEEE, 2013.

[7] Wei-Ping, Zhu, L. I. Ming-Xin, and Chen Huan. "Using MongoDB to implement textbook management system instead of MySQL," Communication Software and Networks(ICCSN), 2011 IEEE 3rd International Conference on. IEEE, 2011.

[8] TEOREY, Toby J. “Database modeling & design,” Morgan Kaufmann, 1999.

[9] Győrödi, Cornelia, et al. "A comparative study: MongoDB vs. MySQL," Engineering of Modern Electric Systems (EMES), 13th International Conference on IEEE, pp. 1-6, 2015.

[10] S. H. Aboutorabi, M. Rezapour, M. Moradi, and N. Ghadiri, “Performance evaluation of sql and mongodb databases for big e-commerce data,” in International Symposium

on Computer Science and Software Engineering, pp. 1-7, 2015.
[11] Mamenko, J. "Introduction to data modeling and msaccess," Lecture Notes on Information Resources, 2004.
[12] Baxter, Ira D., and Michael Mehlich., "Reverse engineering is reverse forward engineering," Proceedings of the Fourth Working Conference on. IEEE, 1997.
[13] Hoberman, S., "DataModeling for MongoDB," Technics Publications, 2014.

Authors

Jun Ho Lee received the B.S. degrees in Computer Software Engineering from Soonchunhyang University, Korea, in 2015 respectively. Lee received the BS degrees in Computer Software Engineering from Soonchunhyaung University in 2015. Now he is undertaking a master degree of computer engineering courses as a member of the database lab at Soonchunhyang University. He is interested in database and bigdata database.

Kyung Soo Joo received the Ph.D. degrees in Computer Science from Korea University, Korea, in 1993 respectively. Dr. Joo joined the faculty of the Department of Computer Science at Korea University, Seoul, Korea, in 1993. He is currently a Professor in the Department of Computer Software Engineering, Soonchunhyang University. He is interested in data base and bigdata database.