Project Report for 23-1 Semester Database Class 2

Phase 1. Database Modeling

21300024 Hyeon-myeong Kang
21400659 Jae-geun Jang
21900543 Won-bin Lee

1. Design Objectives

This project is divided into two phases. Phase 1 focuses on modeling the given data, while Phase 2 aims to derive results for the given task as quickly as possible based on the model implemented in Phase 1.

2. Data Implementation Process

a. Provided Data Information

The data provided for this project is a collection of articles and other materials related to unification, managed by Handong Global University and the Ministry of Unification. These materials were gathered through data crawling using Python and are made available to users via KUBIC for storage and data API provision.

In this project, we received 36,531 records made up of 42 columns. These are part of the data stored in the database, including collected posts and related files, and website users.

The data provided in Phase 1 is normalized, with many duplicate values depending on the values of each column. Phase 1 of the project aims to reduce the overall size of the data by normalizing it.

b. Data Analysis

Initially, we classified the data based on column names and information in the records into four categories and created new tables.

1. doc_type ~ file_id: Posts collected through crawling (post)
2. userId ~ isAdmin: Unique user information and authority (userInfo)
3. title ~ category: Board posts not collected via crawling (board)
4. savedUser ~ savedDocHashKey: Information on bookmarked crawling posts by users (savedPost)

After confirming that the 'board' table could be classified into four records by removing duplicate data based on docID and title, we organized and saved the data.

In the post table, we found that hash_key is a unique value with no duplicates for each post, so we set this column as the primary key of the post table.

We noticed that several records with identical values from doc_type to collection_time were appearing due to differing file-related attributes (file_name, file_download_url, file_content, file_id). To resolve this, we separated these columns into a separate table named fileList. To link the fileList with the posts, we saved the hash_key of the post as a foreign key. We configured the table with five columns in this way, enabling us to restore the original table by performing a left join on the hash_key between post and fileList when we unnormalize the data in the future.

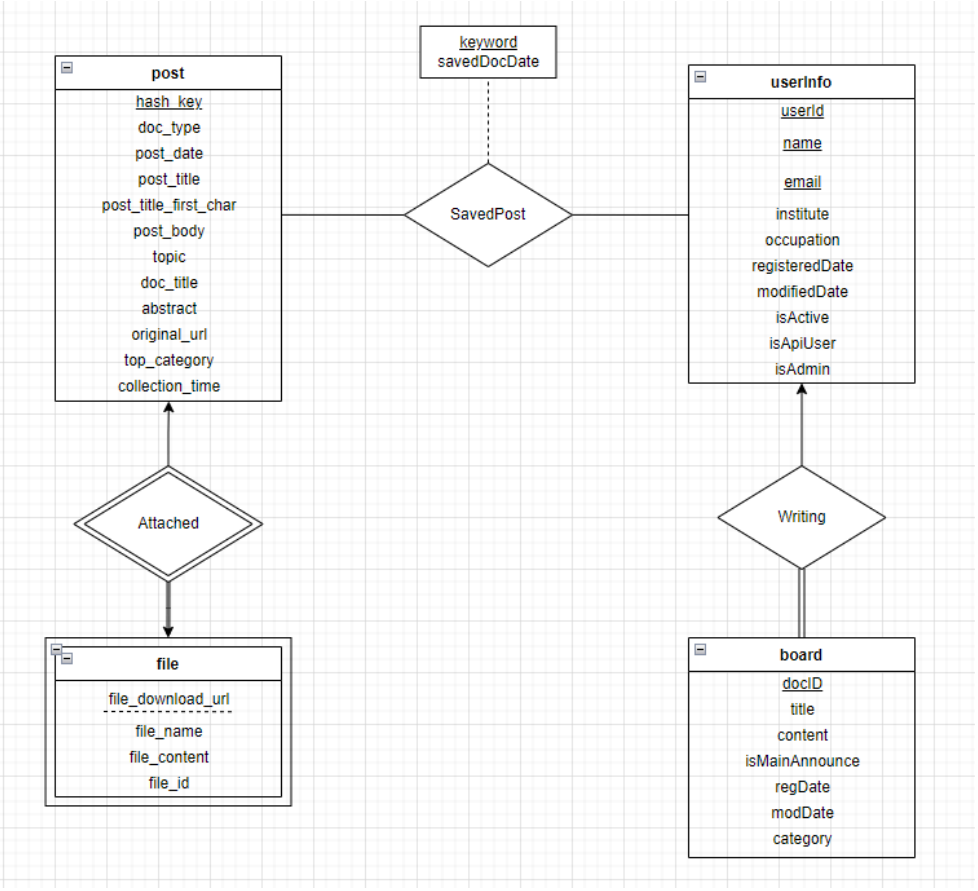
Similarly, we needed to find a standard to unnormalize the savedPost and post tables. By comparing the values of each data, we found that savedDocHashKey is not a unique value assigned to each user, but rather identical to the hash_key of the corresponding post (existing in the post). We decided to use this for unnormalization.

c. Issues and Solutions

When we completed creating the five tables in DB34 from the original data, we encountered an issue where the total data size of DB34 was larger than the original data size. Upon further investigation, we identified the cause. Initially, we did not perfectly import the table but roughly brought in the post table. Due to the oversight of the database characteristic that the size of the initially created table is fixed, we were continually trimming unnecessary content from the first post table we created in DB34. However, when we created a new post table by extracting only the truly necessary values from the original data, the size was significantly reduced.
During the construction of the savedPost table, we faced two options. The first option focused on reducing data size, while the second sacrificed some data size reduction in favor of creating a table structure that allowed for denormalization, thereby preventing data loss. The former approach was feasible for reducing data size immediately, but we lacked ideas for implementing denormalization without incurring data loss. The latter approach, as the name suggests, involved forgoing data size reduction, which later benefited subsequent data processing and denormalization stages. After applying and comparing both approaches, we concluded that preventing data loss was more important than the benefits gained from reducing data size, and thus proceeded with the second direction.

3. Result

a. ER-Diagram



b. View instructions

i. View: total_Volume

```
CREATE VIEW total_Volume AS
    SELECT table_schema AS 'DatabaseName',
    ROUND(SUM(data_length+index_length)/1024, 1) AS 'Size(KB)'
    FROM information_schema.tables
    WHERE TABLE_SCHEMA = 'DB34'
    GROUP BY TABLE_SCHEMA;
```

Result

1 row

	DatabaseName	`Size(KB)`
1	DB34	239392.0

ii. View: userCount

```
CREATE VIEW userCount AS
    SELECT COUNT(userID)
    FROM userInfo
```

Result

	`COUNT(userID)`
1	100

iii. View: boardCount

```
CREATE VIEW boardCount AS
    SELECT COUNT(docID)
    FROM board
```

Result

	`COUNT(docID)`
1	4

iv. View: docCount

```
CREATE VIEW docCount AS
    SELECT COUNT(hash_key)
    FROM post
```

Result

	`COUNT(hash_key)`
1	30782

v. View: instInfo

```
CREATE VIEW instInfo AS
    SELECT published_institution, count(*) CNT
    FROM post
    GROUP BY published_institution
    ORDER BY CNT DESC
    LIMIT 1;
```

Result

1 row

	published_institution	CNT
1	코리아정책연구원	11861

vi. View: my_docs_2019_keyword_count

```
CREATE VIEW my_docs_2019_keyword_count AS
    WITH 2019_keywords AS(
        SELECT keyword, count(*) CNT
        FROM savedPost
        WHERE savedDocDate BETWEEN '2019-01-01' AND '2019-12-31'
        GROUP BY keyword
        ORDER BY CNT DESC)
    SELECT keyword, CNT keyword_count
    FROM 2019_keywords
    WHERE CNT = (
        SELECT MAX(CNT)
        FROM 2019_keywords
        WHERE CNT NOT IN (SELECT MAX(CNT) FROM 2019_keywords)
        );
```

Result

1 row

	keyword	keyword_count
1	토론	188

vii. View: policy_writer_count

```
CREATE VIEW policy_writer_count AS
    WITH writers AS(
        SELECT post_writer, count(*) data_count
        FROM post JOIN savedPost sP on post.hash_key = sP.savedDocHashKey
        WHERE keyword in (SELECT keyword FROM my_docs_2019_keyword_count)
        GROUP BY post_writer
        ORDER BY data_count DESC)
    SELECT post_writer,data_count
    FROM writers
    WHERE data_count = (SELECT MAX(data_count) FROM writers)
```

Result

	post_writer	data_count
1	재나이스피플	76

viii. View: inst_data_max_status

```
CREATE VIEW inst_data_max_status AS
    WITH second_inst AS (SELECT institute, data_count
                    FROM (SELECT institute, count(*) as data_count
                    FROM userInfo JOIN savedPost ON userInfo.email = savedPost.savedUser
                        GROUP BY institute
                        ORDER BY count(*) DESC
                        LIMIT 2) A
                    ORDER BY A.data_count
                    LIMIT 1)

    SELECT userInfo.institute, occupation max_status, count(*) max_status_count,data_count
    FROM userInfo
    JOIN savedPost ON userInfo.email = savedPost.savedUser
    JOIN second_inst ON userInfo.institute = second_inst.institute
    GROUP BY userInfo.institute, occupation
        ORDER BY count(*) DESC
        LIMIT 1;
```

Result

	institute	max_status	max_status_count	data_count
1	서울대학교	대학생	638	694

ix. View: checkDoc

```
CREATE VIEW checkDoc AS
    SELECT post_title, post_writer, published_institution, post_date, top_category
    FROM post
    ORDER BY post_date DESC
    LIMIT 5;
```

Result

	post_title	post_writer	published_institution	post_date	top_category
1	인류사 3500년 중 전쟁 없었던 해는 270년에 불과 전쟁 가능성에 대비해야만 평	통일과나눔	통일과나눔	2023-03-31	통일 스터디
2	논평 성평등 삭제하고 살만하지 않은 사회를 공고하게 만드는 저출산 정책 전면	adminwmp	평화를만드는여성회	2023-03-31	성명서및보도자료
3	정전협정 70년을 맞이하여 한반도 평화를 바라는 여성들의 입장	adminwmp	평화를만드는여성회	2023-03-30	성명서및보도자료
4	자료집 강제동원 정부 해법 관련 긴급 국회 토론회	겨레하나	겨레하나	2023-03-28	자료실/일반자료실
5	성명서 북 연이은 단거리탄도미사일 발사를 규탄한다	관리자	한국자유총연맹	2023-03-27	보도자료 성명서/ 북한

x. View: category_Count

```
CREATE VIEW category_Count AS
    SELECT top_category, category_count, rank() OVER (ORDER BY category_count DESC) category_rank
    FROM (
        SELECT top_category, count(*) AS category_count
        FROM post
        GROUP BY top_category
        ) A
    ORDER BY category_rank;
```

Result

158 rows | Tx: Auto | DDL

WHERE | ORDER BY

	top_category	category_count	category_rank
1	연구자료	7982	1
2	전체자료	5396	2
3	정부자료	2868	3
4	언론자료	1549	4
5	통일부 발간자료	1333	5
6	통일부 발간물	908	6
7	현안분석-온라인시리즈	586	7
8	정기간행물-주간통일정세	545	8
9	통일문제 이해	541	9
10	참고자료	500	10
11	북한소식	453	11
12	북한이해	383	12
13	자료실	357	13
14	남북관계	328	14
15	신진연구자 논문집	300	15
16	평화누리통일누리	299	16
17	기타	285	17
18	통일백서	285	17
19	현안분석-KINU Insight	265	19
20	일반자료실	256	20
21	보도자료 성명서/ 북한	250	21
22	탈북자 증언수기	246	22
23	북한동향	242	23
24	자료실/일반자료실	232	24
25	문화예술	216	25
26	현안분석	196	26
27	인물 정보	167	27
28	북한방송	142	28
29	국내자료	139	29
30	연구보고서-연구총서	136	30
31	통일과 평화	131	31
32	발간자료	117	32
33	홍보출판물	106	33