**Term Project Phase 1 Report**

Team 14

21400714 조세형

21700147 김은택

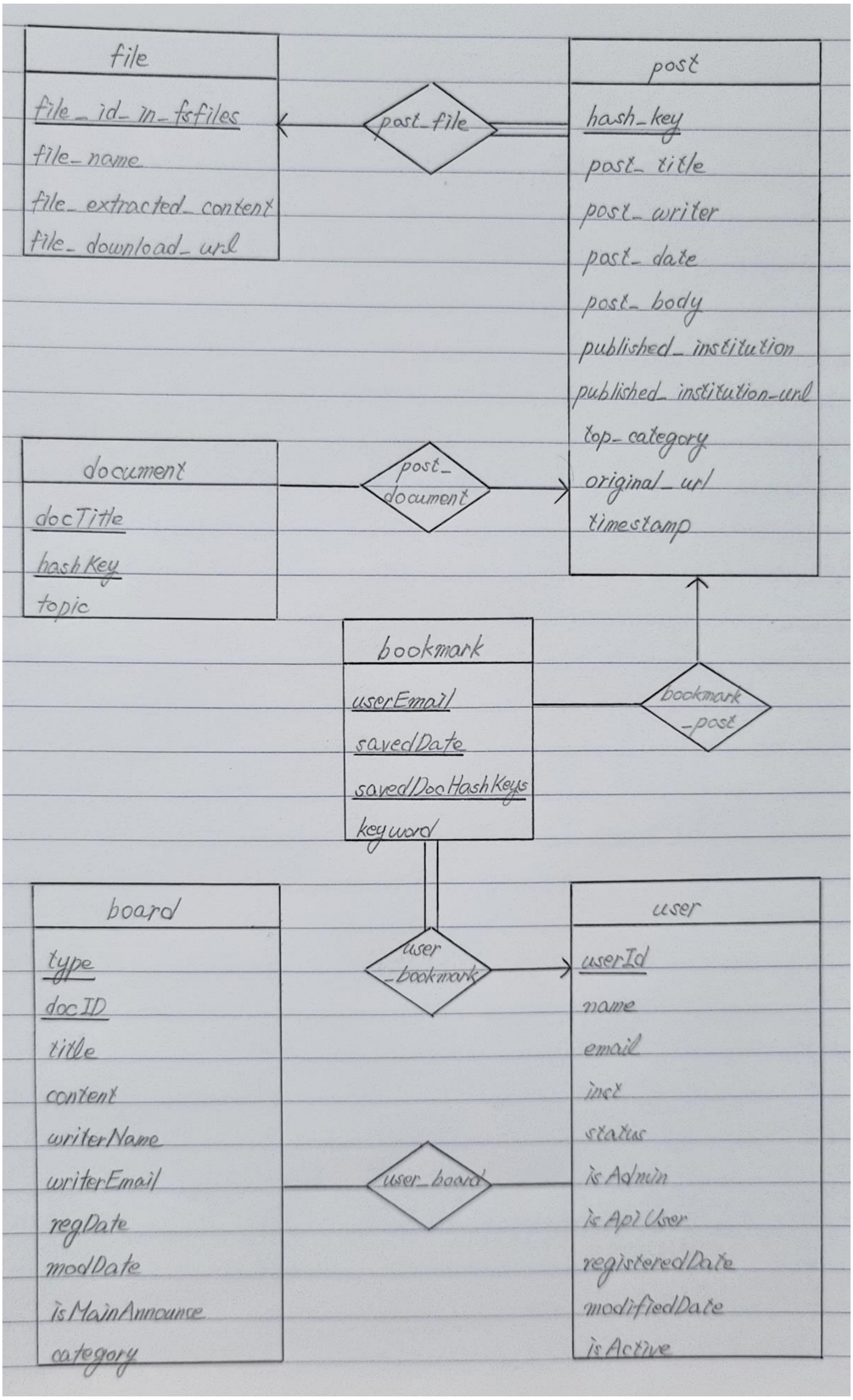
21900395 신소은

The purpose of this report is to describe how to design and implement a database instance that is efficient in space. In the provided data, there are completely unnormalized 116,350 rows and 42 columns. Our group members proceeded with the following process to normalize the data.

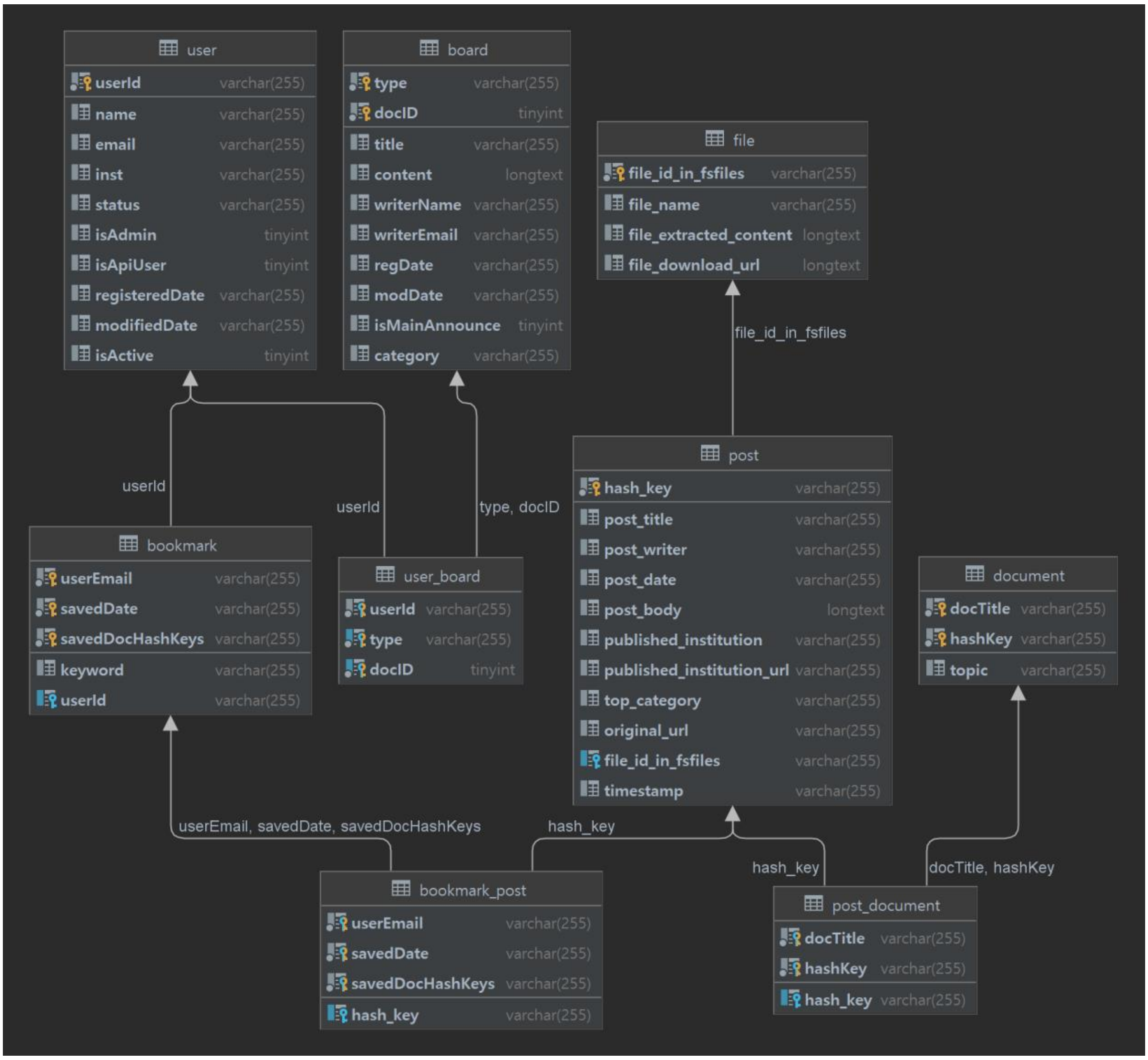
1. Analyze the role of each column.
2. Decompose the provided data and organize several entity sets with related columns using normal forms.
3. Organize several relationship sets among several entity sets using an E-R diagram.
4. Represent entity sets to relation schemas.
5. Analyze each table and use normal forms to reduce data redundancy.
6. Set the appropriate data type for each column to make a database efficiently in space.

Firstly, we analyzed the role of 42 columns. They are a collection of core metadata about web documents and contain user information, bulletin boards, and saved documents. Secondly, we decomposed the provided data *core* and organize several entity sets *user*, *board*, *bookmark*, *post*, and *document* with related columns using the second normal form. Thirdly, we organized several relationship sets among several entity sets using an E-R diagram. We tried to set mapping cardinalities and participations among several entity sets. Fourthly, we represented entity sets to relation schemas. We analyzed mapping cardinalities and set the primary key of each relationship set. Fifthly, we analyzed each table and used the third normal form in the *post* table. We could reduce data redundancy by decomposing the *post* table to *post* and *file* tables. Finally, we set the appropriate data type for each column. In the provided data, the data types of all columns are *char(255)*, *bigint*, and *longtext.* We changed the fixed-length data type *char(255)* to variable-length data type *varchar(255)* and *bigint* to *tinyint* because *tinyint* data type can cover the range of data. We designed and implemented a database by going back and repeating the process as needed. We will introduce the E-R diagram of the implemented database in [Figure 1].

In [Figure 1], 6 entity sets and 5 relationship sets are represented. We tried to identify mapping cardinalities and participation among entity sets. For example, a user can have many bookmarks, and a bookmark can be saved by only one user, then the relationship set from *user* to *bookmark* must be one-to-many. In addition, web browsers may require every *bookmark* to have at least one *user* because only a *user* can save a *bookmark*. Because relationship set *user_bookmark* from entity set *bookmark* to entity set *user* is many-to-one and the participation of *bookmark* in the relationship is total, the schema *user_bookmark* can be combined with the *bookmark* schema later.



**[Figure 1]** E-R diagram of the implemented database by hand



**[Figure 2]** E-R diagram of the implemented database by Tool

In [Figure 2], 9 tables are represented. We will introduce list of all tables and their attributes with precise notions of data types and integrity constraints using DDL query from [Figure 3] to [Figure 11].

```sql
create table board
(
    type           varchar(255) not null,
    docID          tinyint      not null,
    title          varchar(255) null,
    content        longtext     null,
    writerName     varchar(255) null,
    writerEmail    varchar(255) null,
    regDate        varchar(255) null,
    modDate        varchar(255) null,
    isMainAnnounce tinyint      null,
    category       varchar(255) null,
    primary key (type, docID)
);
```

**[Figure 3]** table *board shows* board information

```sql
create table bookmark
(
    userEmail        varchar(255) not null,
    savedDate        varchar(255) not null,
    savedDocHashKeys varchar(255) not null,
    keyword          varchar(255) null,
    userId           varchar(255) null,
    primary key (userEmail, savedDate, savedDocHashKeys),
    constraint z_bookmark_ibfk_1
        foreign key (userId) references user (userId)
);
```

**[Figure 4]** table *bookmark* shows saved document information

```sql
create table bookmark_post
(
    userEmail        varchar(255) not null,
    savedDate        varchar(255) not null,
    savedDocHashKeys varchar(255) not null,
    hash_key         varchar(255) null,
    primary key (userEmail, savedDate, savedDocHashKeys),
    constraint z_bookmark_postInfo_ibfk_1
        foreign key (hash_key) references post (hash_key),
    constraint z_bookmark_postInfo_ibfk_2
        foreign key (userEmail, savedDate, savedDocHashKeys) references bookmark (userEmail,
savedDate, savedDocHashKeys)
);
```

**[Figure 5]** table *bookmark_post* shows a relationship between *bookmark* and *post*

```
create table document
(
    docTitle varchar(255) not null,
    hashKey  varchar(255) not null,
    topic    varchar(255) null,
    primary key (docTitle, hashKey)
);
```

**[Figure 6]** table *document* shows document information in search engine

```
create table file
(
    file_id_in_fsfiles     varchar(255) not null
        primary key,
    file_name              varchar(255) null,
    file_extracted_content longtext     null,
    file_download_url      longtext     null
);
```

**[Figure 7]** table *file* shows file information

```
create table post
(
    hash_key                  varchar(255) not null
        primary key,
    post_title                varchar(255) null,
    post_writer               varchar(255) null,
    post_date                 varchar(255) null,
    post_body                 longtext     null,
    published_institution     varchar(255) null,
    published_institution_url varchar(255) null,
    top_category              varchar(255) null,
    original_url              varchar(255) null,
    file_id_in_fsfiles        varchar(255) null,
    timestamp                 varchar(255) null,
    constraint z_postInfo_ibfk_1
        foreign key (file_id_in_fsfiles) references file (file_id_in_fsfiles)
);
```

**[Figure 8]** table *post* shows post information

```
create table post_document
(
    docTitle varchar(255) not null,
    hashKey  varchar(255) not null,
    hash_key varchar(255) null,
    primary key (docTitle, hashKey),
    constraint z_postInfo_documentInfo_ibfk_1
        foreign key (hash_key) references post (hash_key),
    constraint z_postInfo_documentInfo_ibfk_2
        foreign key (docTitle, hashKey) references document (docTitle, hashKey)
);
```

**[Figure 9]** table *post_document* shows a relationship between *post* and *document*

```
create table user
(
    userId         varchar(255) not null
        primary key,
    name           varchar(255) null,
    email          varchar(255) null,
    inst           varchar(255) null,
    status         varchar(255) null,
    isAdmin        tinyint      null,
    isApiUser      tinyint      null,
    registeredDate varchar(255) null,
    modifiedDate   varchar(255) null,
    isActive       tinyint      null
);
```

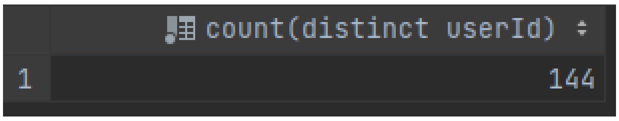
**[Figure 10]** table *user* shows user information

```
create table user_board
(
    userId varchar(255) not null,
    type   varchar(255) not null,
    docID  tinyint      not null,
    primary key (userId, type, docID),
    constraint z_userInfo_boardInfo_ibfk_1
        foreign key (userId) references user (userId),
    constraint z_userInfo_boardInfo_ibfk_2
        foreign key (type, docID) references board (type, docID)
);
```

**[Figure 11]** table *user_board* shows a relationship between *user* and *board*

From [Figure 3] to [Figure 11], the DDL query of each table is represented. We will describe the requested views from [Figure 12] to [Figure 19]. We will represent the SQL query to make the requested views, the size of the resulting table in counts, and the screenshots of the table header and first five records.

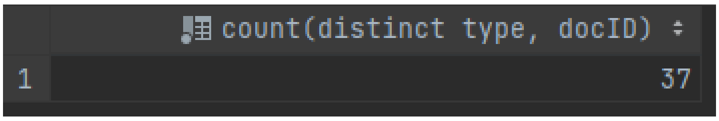
```
create view userCount as
    select count(distinct userId)
    from user
    where isActive = 1;
```



| | count(distinct userId) |
|---|---|
| 1 | 144 |

**[Figure 12]** view 1 (1 row, 1 column)

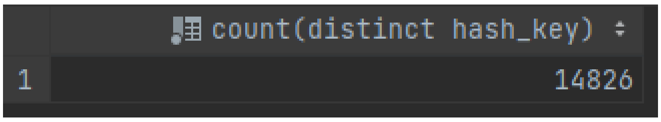
```
create view boardCount as
    select count(distinct type, docID)
    from board;
```



| | count(distinct type, docID) |
|---|---|
| 1 | 37 |

**[Figure 13]** view 2 (1 row, 1 column)

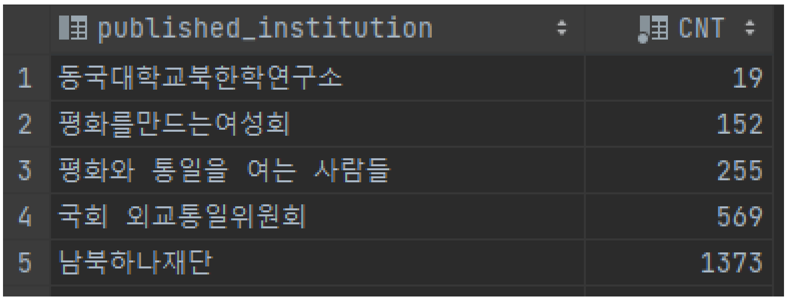
```
create view docCount as
    select count(distinct hash_key)
    from post;
```



| | count(distinct hash_key) |
|---|---|
| 1 | 14826 |

**[Figure 14]** view 3 (1 row, 1 column)

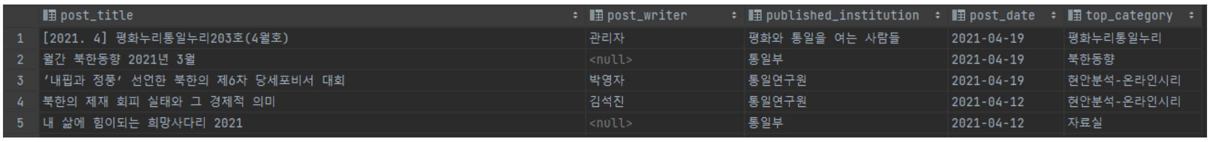
```
create view instPubInfo as
    select published_institution, count(*) as CNT
    from post
    group by published_institution
    order by CNT;
```



| | published_institution | CNT |
|---|---|---|
| 1 | 동국대학교북한학연구소 | 19 |
| 2 | 평화를만드는여성회 | 152 |
| 3 | 평화와 통일을 여는 사람들 | 255 |
| 4 | 국회 외교통일위원회 | 569 |
| 5 | 남북하나재단 | 1373 |

**[Figure 15]** view 4 (7 rows, 2 columns)

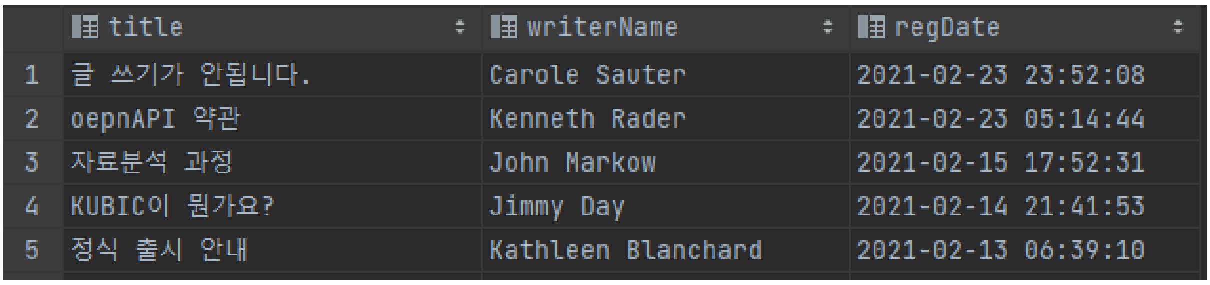
```
create view docInfo as
    select post_title, post_writer, published_institution, post_date, top_category
    from post
    order by post_date desc;
```



|  | post_title | post_writer | published_institution | post_date | top_category |
|---|---|---|---|---|---|
| 1 | [2021. 4] 평화누리통일누리203호(4월호) | 관리자 | 평화와 통일을 여는 사람들 | 2021-04-19 | 평화누리통일누리 |
| 2 | 월간 북한동향 2021년 3월 | <null> | 통일부 | 2021-04-19 | 북한동향 |
| 3 | ‘내핍과 정풍’ 선언한 북한의 제6차 당세포비서 대회 | 박영자 | 통일연구원 | 2021-04-19 | 현안분석-온라인시리 |
| 4 | 북한의 제재 회피 실태와 그 경제적 의미 | 김석진 | 통일연구원 | 2021-04-12 | 현안분석-온라인시리 |
| 5 | 내 삶에 힘이되는 희망사다리 2021 | <null> | 통일부 | 2021-04-12 | 자료실 |

**[Figure 16]** view 5 (14,826 rows, 5 columns)

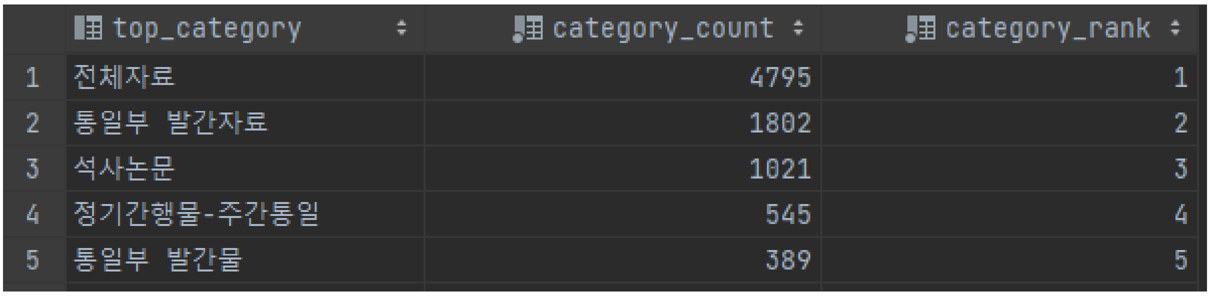
```
create view bulletinSummary as
    select title, writerName, regDate
    from board
    order by regDate desc;
```



|  | title | writerName | regDate |
|---|---|---|---|
| 1 | 글 쓰기가 안됩니다. | Carole Sauter | 2021-02-23 23:52:08 |
| 2 | oepnAPI 약관 | Kenneth Rader | 2021-02-23 05:14:44 |
| 3 | 자료분석 과정 | John Markow | 2021-02-15 17:52:31 |
| 4 | KUBIC이 뭔가요? | Jimmy Day | 2021-02-14 21:41:53 |
| 5 | 정식 출시 안내 | Kathleen Blanchard | 2021-02-13 06:39:10 |

**[Figure 17]** view 6 (37 rows, 3 columns)

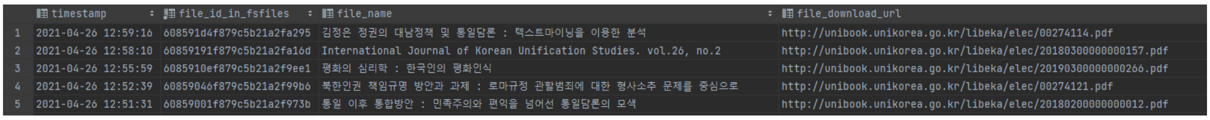
```
create view docSummary as
    select top_category, count(*) as category_count, rank() over (order by count(*) desc)
as category_rank
    from post
    group by top_category;
```



|  | top_category | category_count | category_rank |
|---|---|---|---|
| 1 | 전체자료 | 4795 | 1 |
| 2 | 통일부 발간자료 | 1802 | 2 |
| 3 | 석사논문 | 1021 | 3 |
| 4 | 정기간행물-주간통일 | 545 | 4 |
| 5 | 통일부 발간물 | 389 | 5 |

**[Figure 18]** view 7 (76 rows, 3 columns)

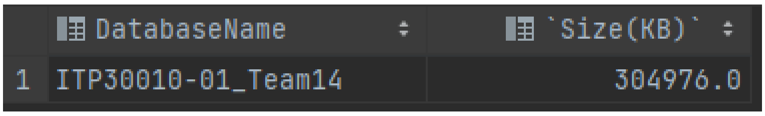
```
create view fileSummary as
    select timestamp, file.file_id_in_fsfiles, file_name, file_download_url
    from file join post on file.file_id_in_fsfiles = post.file_id_in_fsfiles
    order by timestamp desc;
```



| | timestamp | file_id_in_fsfiles | file_name | file_download_url |
|---|---|---|---|---|
| 1 | 2021-04-26 12:59:16 | 608591d4f879c5b21a2fa295 | 김정은 정권의 대남정책 및 통일담론 : 텍스트마이닝을 이용한 분석 | http://unibook.unikorea.go.kr/libeka/elec/00274114.pdf |
| 2 | 2021-04-26 12:58:10 | 60859191f879c5b21a2fa16d | International Journal of Korean Unification Studies. vol.26, no.2 | http://unibook.unikorea.go.kr/libeka/elec/20180300000000157.pdf |
| 3 | 2021-04-26 12:55:59 | 6085910ef879c5b21a2f9ee1 | 평화의 심리학 : 한국인의 평화인식 | http://unibook.unikorea.go.kr/libeka/elec/20190300000000266.pdf |
| 4 | 2021-04-26 12:52:39 | 60859046f879c5b21a2f99b6 | 북한인권 책임규명 방안과 과제 : 로마규정 관할범죄에 대한 형사소추 문제를 중심으로 | http://unibook.unikorea.go.kr/libeka/elec/00274121.pdf |
| 5 | 2021-04-26 12:51:31 | 60859001f879c5b21a2f973b | 통일 이후 통합방안 : 민족주의와 편익을 넘어선 통일담론의 모색 | http://unibook.unikorea.go.kr/libeka/elec/20180200000000012.pdf |

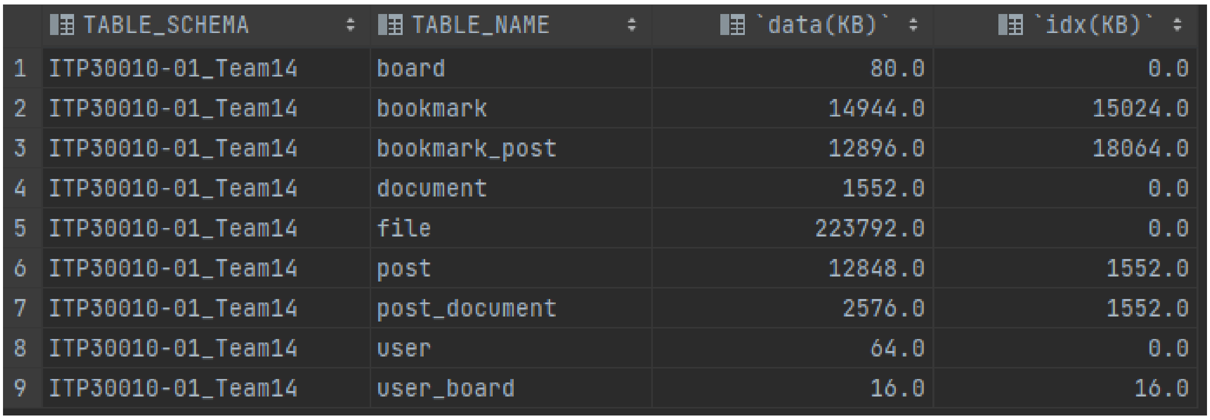
**[Figure 19]** view 8 (9,954 rows, 4 columns)

From [Figure 12] to [Figure 19], the size of the resulting table in counts is represented in parenthesis. Finally, we will introduce the database size and table sizes in kilobytes from [Figure 20] to [Figure 21].



| | DatabaseName | `Size(KB)` |
|---|---|---|
| 1 | ITP30010-01_Team14 | 304976.0 |

**[Figure 20]** The database size in kilobytes



| | TABLE_SCHEMA | TABLE_NAME | `data(KB)` | `idx(KB)` |
|---|---|---|---|---|
| 1 | ITP30010-01_Team14 | board | 80.0 | 0.0 |
| 2 | ITP30010-01_Team14 | bookmark | 14944.0 | 15024.0 |
| 3 | ITP30010-01_Team14 | bookmark_post | 12896.0 | 18064.0 |
| 4 | ITP30010-01_Team14 | document | 1552.0 | 0.0 |
| 5 | ITP30010-01_Team14 | file | 223792.0 | 0.0 |
| 6 | ITP30010-01_Team14 | post | 12848.0 | 1552.0 |
| 7 | ITP30010-01_Team14 | post_document | 2576.0 | 1552.0 |
| 8 | ITP30010-01_Team14 | user | 64.0 | 0.0 |
| 9 | ITP30010-01_Team14 | user_board | 16.0 | 16.0 |

**[Figure 21]** The table sizes in kilobytes

While doing this assignment, we studied as if there was an answer to designing a table, but we realized that reality is different from the study. In the process of designing, we thought countless times about whether what we thought was right, and whether it was right to eliminate all duplication. This is because the table can lose its meaning by removing all duplicate data. We were able to think a lot during this assignment because everything from decomposing tables and setting the primary key and the foreign key depends on our decision. We were able to study more deeply than before by using the E-R diagram and normalization in practice, which we only knew as a concept.