ECE30030/ITP30010 Database Systems

Term Project

Charmgil Hong

charmgil@handong.edu

Spring, 2022

Handong Global University

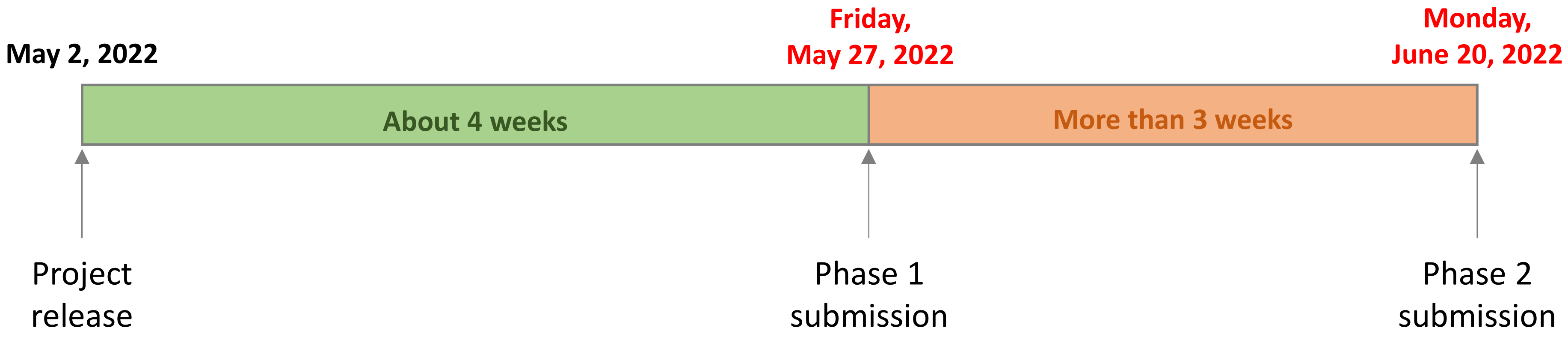
Term Project

- Goals
  - To practice the concepts and underlying mechanisms of database management system with an actual database instance
  - To represent database designs in modeling languages and analyze the designs with respect to given constraints
  - To articulate the relational database language (structured query language)
  - To exercise the optimization and evaluation of the database performance

- In this project, each team will be given a large chunk of data that is completely unnormalized
  - Your objective is to design a "good" database schema that can accommodate the provided data without any loss of information
    - "Good" in that…
      - Efficient in terms of space and time complexity

Term Project Overview

- Planned timeline



- Phase 1 “space” submission: Friday, May 27, 2022
- Phase 2 “time” submission: Monday, June 20, 2022

KUBiC: Korean Unification Bigdata Center

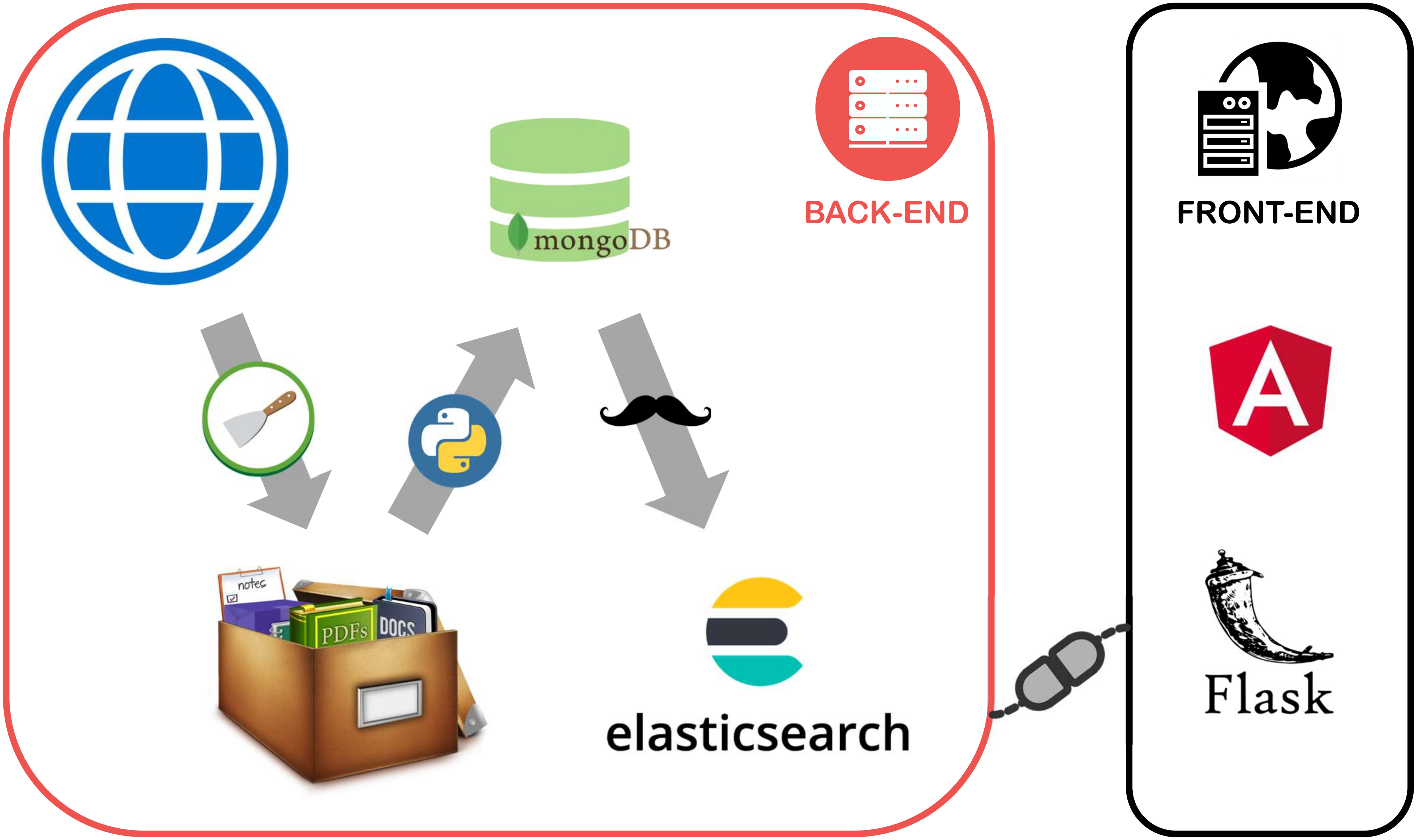
- **Term-Project data is provided by the KUBiC project team**
- A government-funded project on a data-center development focusing on the Korean unification
  - URL: https://kubic.handong.edu/
  - Data archive + search engine + web-based analysis tools, specialized on the Korean unification and North Korea research
  - Contains a lot of academic papers and government reports on the relevant topics

KUBiC
Korea Unification Bigdata Center

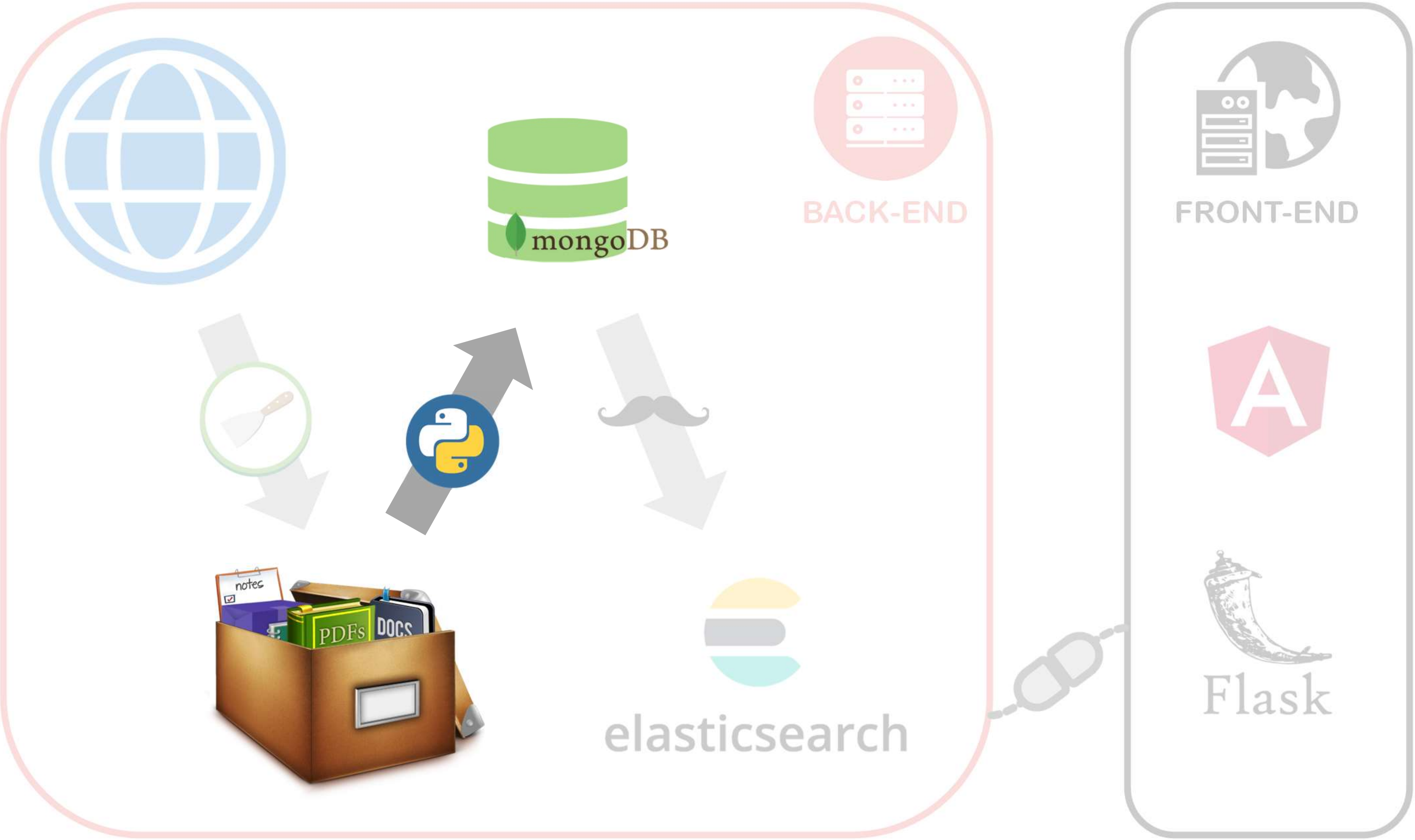
한동대학교
HANDONG GLOBAL UNIVERSITY

통일부
Ministry of Unification

KUBiC: Korean Unification Bigdata Center



KUBiC: Korean Unification Bigdata Center

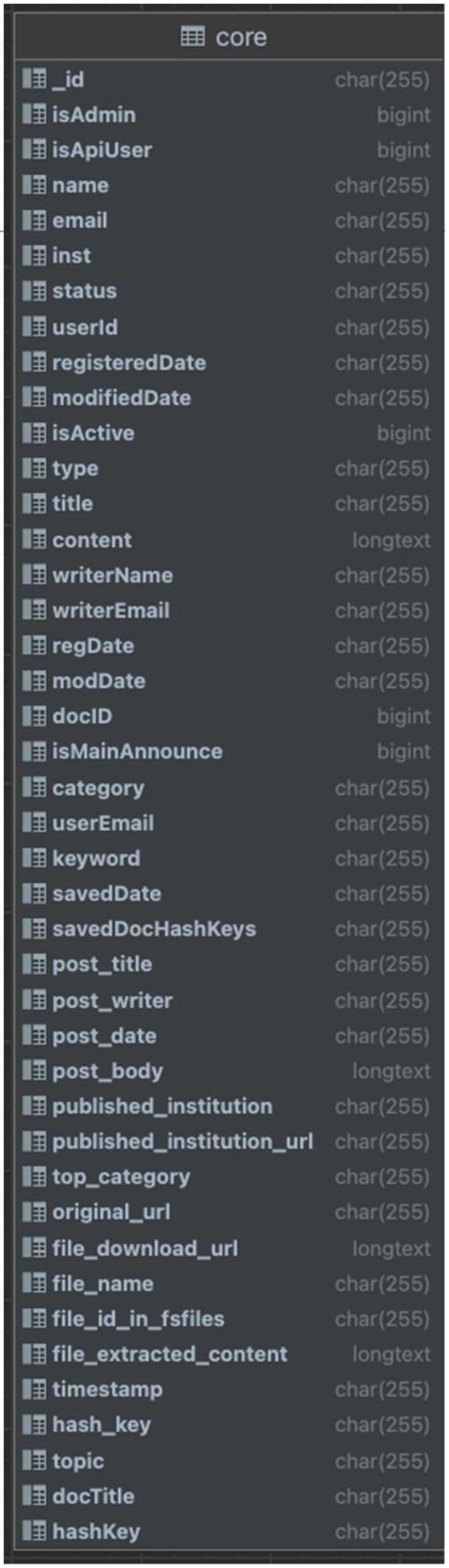


Term Project

- Background
  - You will be given large chunks of data snapshot from the KUBIC database, that consist of one SQL dump file and two csv files
    - core.sql
      - 116,320 records, 42 columns (approx. 2.45 GB)
      - Completely unnormalized
    - tfidf.csv
      - TF-IDF analysis of the service documents
      - 877,490 records, 4 columns (approx. 170.6 MB)
    - rcmd.csv
      - Cosince similarity analysis of the service documents
      - 1,000,000 records, 3 columns (approx. 126.8 MB)
  - SQL dump file: Ordinary text file, written in the SQL syntax
    - Contains a record of the table structure and/or the data from a database
    - Often used for backing up a database so that its contents can be restored in the event of data loss

Provided Data

- Core
  - Collection of core meta-data about the web-documents that KUBIC contains
  - Also contains the bulletin boards, user information, saved documents of each user
  - 116,320 records, 42 columns (2.45GB)
  - Completely unnormalized

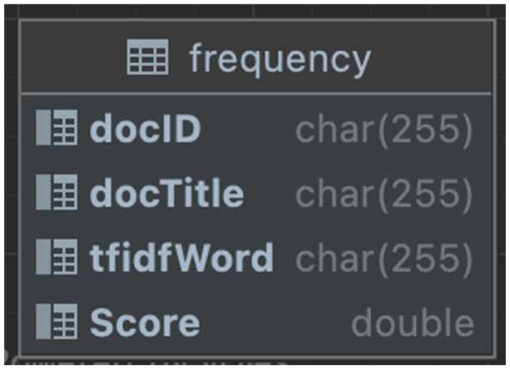


core

_id	char(255)
isAdmin	bigint
isApiUser	bigint
name	char(255)
email	char(255)
inst	char(255)
status	char(255)
userId	char(255)
registeredDate	char(255)
modifiedDate	char(255)
isActive	bigint
type	char(255)
title	char(255)
content	longtext
writerName	char(255)
writerEmail	char(255)
regDate	char(255)
modDate	char(255)
docID	bigint
isMainAnnounce	bigint
category	char(255)
userEmail	char(255)
keyword	char(255)
savedDate	char(255)
savedDocHashKeys	char(255)
post_title	char(255)
post_writer	char(255)
post_date	char(255)
post_body	longtext
published_institution	char(255)
published_institution_url	char(255)
top_category	char(255)
original_url	char(255)
file_download_url	longtext
file_name	char(255)
file_id_in_fsfiles	char(255)
file_extracted_content	longtext
timestamp	char(255)
hash_key	char(255)
topic	char(255)
docTitle	char(255)
hashKey	char(255)

Provided Data

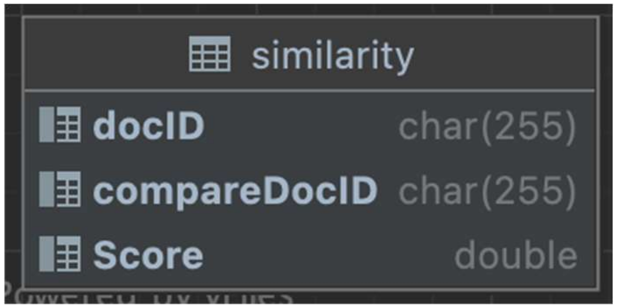
- Tfidf
  - TF-IDF analysis of the service documents
  - 877,490 records, 4 columns (170.6 MB)



frequency	
docID	char(255)
docTitle	char(255)
tfidfWord	char(255)
Score	double

Provided Data

- Rcmds
  - Cosince similarity analysis of the service documents
  - 1,000,000 records, 3 columns (126.8 MB)



similarity	
docID	char(255)
compareDocID	char(255)
Score	double

Term Project

- Phase 1 requirements
  - Design and implment a database that can effectively accommodate the entire data without any loss
    - You and your team will need to draw E-R diagrams and conduct a number of normalization processes
  - Import the data; there should be no missing portion
    - You will be asked to create and submit views
  - Make the database size as small as possible!
- Phase 2 requirements
  - Optimize the database using
    - Denormalization
    - Indexing

Data Files

- Core
  - https://drive.google.com/file/d/1BUTHZv0AgZPUEaOna3IoxUklSXO8VfZ5/view?usp=sharing

- Tfidf
  - https://drive.google.com/file/d/1MUNteBF58NZHNLOf31ZN90BkE0MSUS8H/view?usp=sharing

- Rcmds
  - https://drive.google.com/file/d/14QpCNHPQEucieDK6iWBYjY_Xflz2DWKW/view?usp=sharing

Technical Resources

- Upon completion, submit your result to LMS. Each submission should have the following items:
  - Dump of the database (in `.sql`)
    - How to create a SQL dump?
      - https://dev.mysql.com/doc/refman/8.0/en/mysqldump.html
      - https://dev.mysql.com/doc/refman/8.0/en/mysqldump-sql-format.html
  - Report Documents (in `.pdf`)
    - How to attack this problem?
    - DDL query and result for View instruction
    - ER Diagram of your database
  - Submission should be one `.zip` file

TA’s are up for help

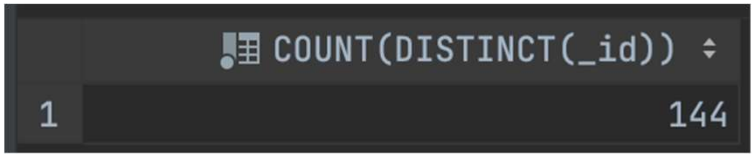
- Jihyung Jang (장지형): Data-specific questions
- Geonyoung Choi (최건영): SQL and DBMS functionalities-related questions
- Juwon Baek (백주원), Dulguun Dorjkham: General inquiries

Phase 1: Database Modeling

- Goal: Design and implement a database instance that is efficient in space
  - You are expected to conduct a database design using ERD and apply the normalization theory
  - We will check the correctness and completeness of your data by examining the output of the views suggested in next slides
  - The database size on the physical storage will be estimated; the smallest 10% teams will earn bonus points (maximum +7%)
  - Before the submission, each team is expected to run several iterations of design, implement, data import, and internal evaluation

Phase 1: Database Modeling

- Views to create (and submit)
  1. View: userCount
     - Count the number of users in the database
     - **SELECT** * **FROM** userCount

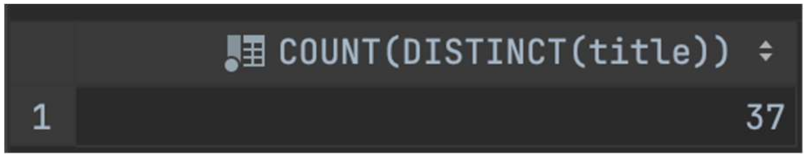


	COUNT(DISTINCT(_id))
1	144

- The column name may vary

Phase 1: Database Modeling

- Views to create (and submit)
  2. View: boardCount
     - Count the number of bulletins on the board
     - **SELECT** * **FROM** boardCount



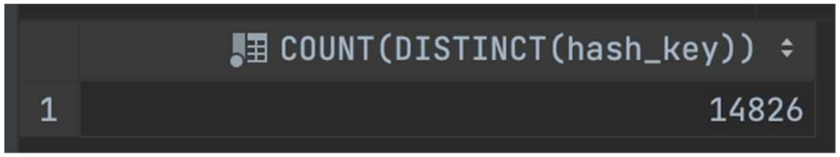
	COUNT(DISTINCT(title))
1	37

       - The column name may vary

Phase 1: Database Modeling

- Views to create (and submit)

  3. View: docCount
     - Count the number of documents that are stored
     - **SELECT** * **FROM** docCount



	COUNT(DISTINCT(hash_key))
1	14826

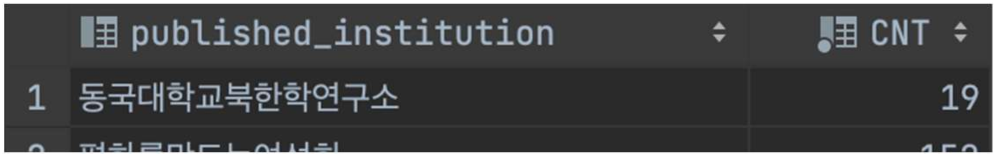
- The column name may vary

Phase 1: Database Modeling

- Views to create (and submit)

  4. View: instPubInfo

     - List the names of publisher institutes and their numbers of publications (sort the results in ascending order of the number of publications)
     - **SELECT** * **FROM** instPubInfo



	published_institution	CNT
1	동국대학교북한학연구소	19

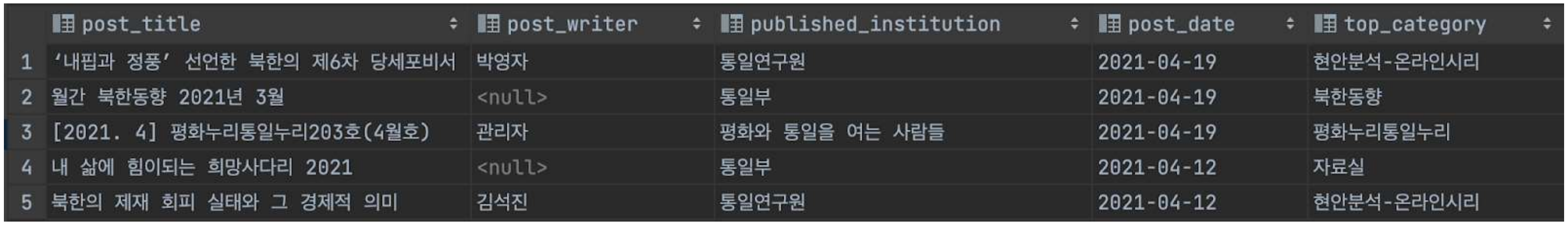
⋮ ⋮

Phase 1: Database Modeling

- Views to create (and submit)

  5. View: docInfo

     - List the posting title, post author name and affiliation, posted date, and top cartegory tag
     - **SELECT** * **FROM** docInfo



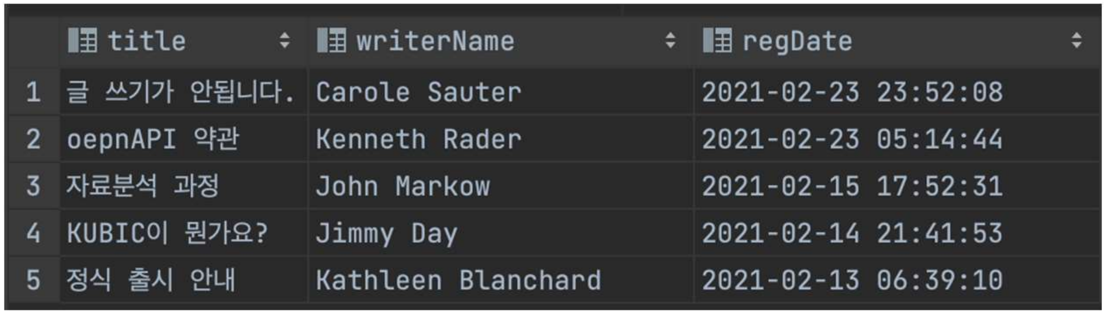
	post_title	post_writer	published_institution	post_date	top_category
1	‘내핍과 정풍’ 선언한 북한의 제6차 당세포비서	박영자	통일연구원	2021-04-19	현안분석-온라인시리
2	월간 북한동향 2021년 3월	<null>	통일부	2021-04-19	북한동향
3	[2021. 4] 평화누리통일누리203호(4월호)	관리자	평화와 통일을 여는 사람들	2021-04-19	평화누리통일누리
4	내 삶에 힘이되는 희망사다리 2021	<null>	통일부	2021-04-12	자료실
5	북한의 제재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

:   :

:   :

Phase 1: Database Modeling

- Views to create (and submit)
  6. View: bulletinSummary
     - List all bulletin titles, author names (writer names), and posted dates
     - **SELECT** * **FROM** bulletinSummary

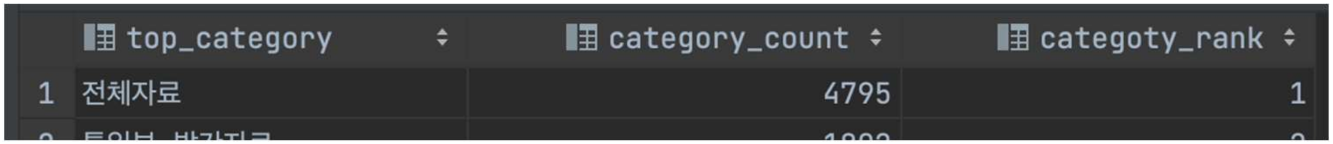


	title	writerName	regDate
1	글 쓰기가 안됩니다.	Carole Sauter	2021-02-23 23:52:08
2	oepnAPI 약관	Kenneth Rader	2021-02-23 05:14:44
3	자료분석 과정	John Markow	2021-02-15 17:52:31
4	KUBIC이 뭔가요?	Jimmy Day	2021-02-14 21:41:53
5	정식 출시 안내	Kathleen Blanchard	2021-02-13 06:39:10

⋮ ⋮

Phase 1: Database Modeling

- Views to create (and submit)
  7. View: docSummary
     - Count the number of documents per each of top category values; show the results in descending order of the counts and put their ranks
     - **SELECT** * **FROM** docSummary



	top_category	category_count	categoty_rank
1	전체자료	4795	1

:　　　　　　　　:

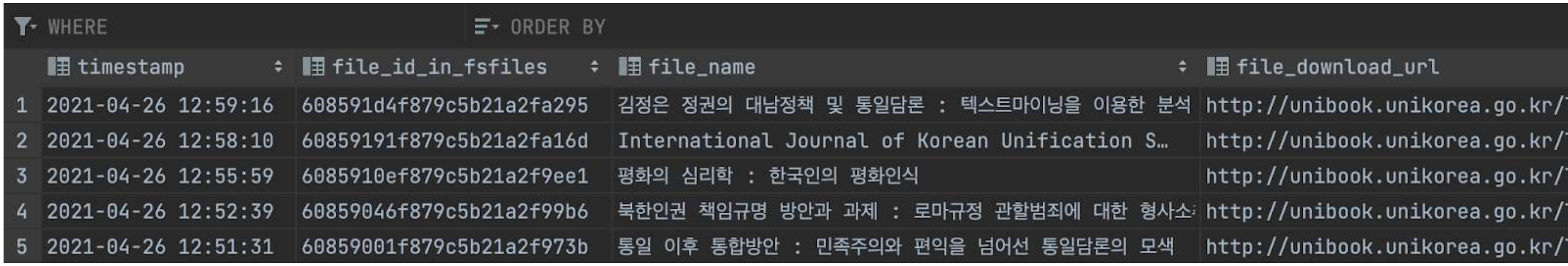
:　　　　　　　　:

Phase 1: Database Modeling

- Views to create (and submit)

  8. View: fileSummary

     - Show the attached file information by summarizing their timestamp, file ID, filename, and download url
     - **SELECT** * **FROM** fileSummary



WHERE ORDER BY

	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/
2	2021-04-26 12:58:10	60859191f879c5b21a2fa16d	International Journal of Korean Unification S…	http://unibook.unikorea.go.kr/
3	2021-04-26 12:55:59	6085910ef879c5b21a2f9ee1	평화의 심리학 : 한국인의 평화인식	http://unibook.unikorea.go.kr/
4	2021-04-26 12:52:39	60859046f879c5b21a2f99b6	북한인권 책임규명 방안과 과제 : 로마규정 관할범죄에 대한 형사소	http://unibook.unikorea.go.kr/
5	2021-04-26 12:51:31	60859001f879c5b21a2f973b	통일 이후 통합방안 : 민족주의와 편익을 넘어선 통일담론의 모색	http://unibook.unikorea.go.kr/

:

:

Phase 1: Database Modeling

- A query to check the size of your database instance

  -
    ```
    SELECT table_schema AS 'DatabaseName’,
              ROUND(SUM(data_length+index_length)/1024, 1) AS 'Size(KB)’
    FROM information_schema.tables
    WHERE table_schema = 'YOUR DATABASE NAME’
    GROUP BY table_schema;
    ```

- A query to check each table size from your database

  -
    ```
    SELECT TABLE_SCHEMA, TABLE_NAME,
              ROUND(DATA_LENGTH/(1024), 1) AS 'data(KB)’,
              ROUND(INDEX_LENGTH/(1024), 1) AS 'idx(KB)’
    FROM information_schema.tables
    WHERE TABLE_TYPE = 'BASE TABLE’
              AND TABLE_SCHEMA = 'YOUR DATABASE SIZE';
    ```

Phase 1: Database Modeling

- What to submit
  - A report including
    - ER diagram of the implemented database
    - List of all tables and their attributes with precise notions of data types and integrity constraints
    - Description of the requested views
      - Size of the resulting table (in counts)
      - The screenshots of the table header and first five records
    - Summary of the database size and table sizes (in Kilobytes)
  - A zipped MySQL dump file containing all the database implementations including the database schema, records, views, *etc*.

Phase 1: Database Modeling

- Resources
  - How to create a dump file
    - MySQL Workbench https://dev.mysql.com/doc/workbench/en/wb-admin-export-import-management.html
    - DataGrip https://www.jetbrains.com/help/datagrip/export-data-in-ide.html
    - HeidiSQL https://www.heidisql.com/screenshots.php?which=export_sql
    - SequelAce https://sequelpro.com/docs/ref/working-with-data

Phase 2: Database Optimization

- Goal: Design and implement a database instance that is efficient in time
  - You are expected to go through the denormalization process and add indexes to the database instance from Phase 1