Distributed Database Systems Course - Group Project

1. Project Aim

Students of this course are required to form a team of two members for the completion of a project. The purpose of the group project is

- To help the students gain deep and insight understanding on the knowledge of the advanced big data management techniques in a distributed environment through a hands-on software design and implementation experiment.
- To grasp the very latest big data management technologies and apply them to solve real-world problems.
- To nurture the team-work spirit through cooperative work on a joint project.

2. Distributed Databases

Data to be managed and processed include structured data (5 relational tables) and unstructured data (text, images, and video). Their inter-relations are illustrated in Fig. 2.

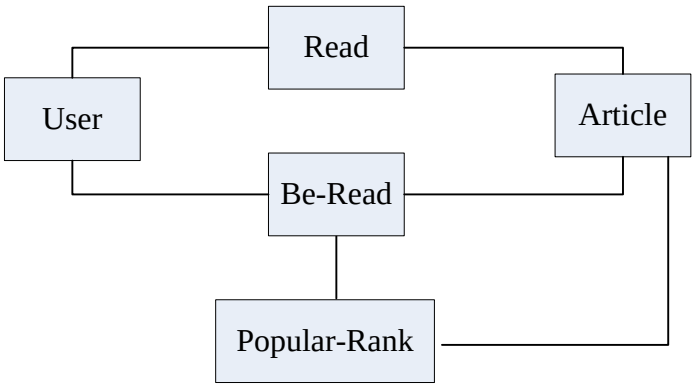

Fig. 2. Entity-Relation diagram.

(1) User table

id, timestamp, uid, name, gender, email, phone, dept. grade, language, **region,** role, preferTags, obtainedCredits

User table is fragmented based on region attribute, where region=”Beijing” allocated in DBMS1 and region= “HongKong” allocated in DBMS2.

(2) Article table

id, timestamp, aid, title, **category,** abstract, articleTags, authors, language, text, image, video

Article table is fragmented based on article category attribute, where
category=”science” allocated in DBMS1 and DBMS2,
category=“technology” allocated in DBMS2.

(3) Read table

id, timestamp, uid, aid, readTimeLength, aggreeOrNot, commentOrNot, commentDetail, shareOrNot

Read table is fragmented based on User table without replica, and with the same allocation schema as User table.

(4) Be-Read table

id, timestamp, aid, readNum, readUidList, commentNum, commentUidList, agreeNum, agreeUidList, shareNum, shareUidList

Be-Read table contains two fragments based on Article table with duplication, where category=”science” allocated to DBMS1 and DBMS2, category=“technology” allocated to DBMS2.

(5) Popular-Rank table

id, timestamp, temporalGranularity, articleAidList

// *temporalGranularity= “daily”, “weekly”, or “monthly”*

Popular-Rank table contains two fragments based on temporalGranularity，where temporalGranularity= “daily” allocated to DBMS1, temporalGranularity= “weekly” or emporalGranularity= “monthly” allocated to DBMS2.

3. Databased Insert and Query Operations

1) Bulk load User table, Article table, and Read table into the data center
2) Query users, articles, users’ read tables (involving the join of User table and Article table) with and without query conditions
3) Populate the empty Be-Read table by inserting newly computed records into the Be-Read table.
4) Query the top-5 daily/weekly/monthly popular articles with articles details (text, image, and video if existing) (involving the join of Be-Read table and Article table)

4. Data Center

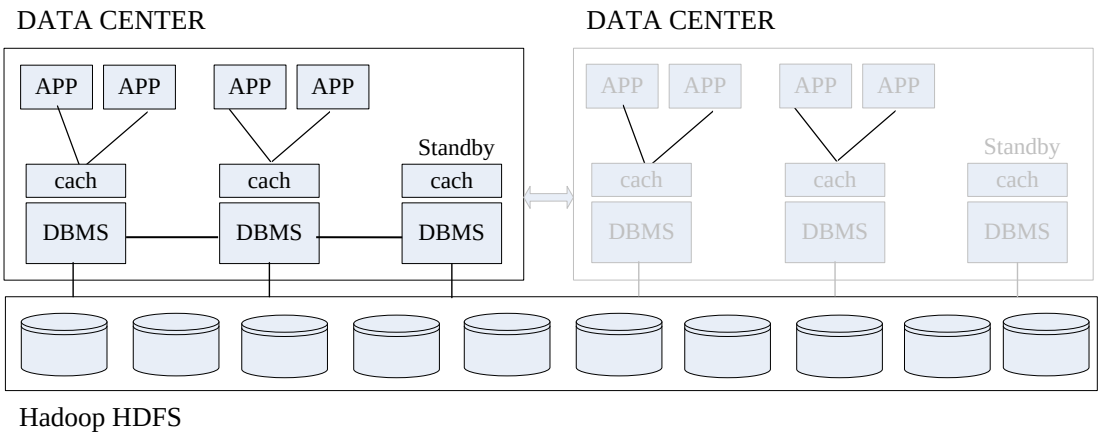

Figure 1. A data center in a distributed context

5. Project Requirements

Implement a data center in a distributed context (Fig. 1) with the following functionalities.

1) Bulk data loading with data partitioning and replica consideration
2) Efficient execution of data insert, update, and queries
3) Monitoring the running status of DBMS servers, including its managed data (amount and location), workload, etc.
4) (Optional) advanced functions
   a) Hot / Cold Standby DBMSs for fault tolerance
   b) Expansion at the DBMS-level allowing a new DBMS server to join
   c) Dropping a DBMS server at will
   d) Data migration from one data center to others

6. Project Evaluation

The group project weighs 60% of the credited mark for the course. The overall assessment of the project will be based on the following items:

- Report and manual (30%)
- Demo (30%)

This report shall include *title, abstract, problem background and motivation, existing solutions, problem definition, proposed solutions, solution evaluation, conclusion including future work,* and *reference*. Write it like a research paper. Hand-written reports are not acceptable.

The system manual is the operation specification, including the instruction of installation, configuration, and operation of the system. The report shall be submitted together with the system executive program.

The project shall be completed by a team of **two** students. One student shall take the leading role of the project. The whole project load should be allocated to members properly. This load allocation must be specified in the final project report.