> Main Contributors:
>
> Leader and Overall Design: ZHU Yueming
>
> Data Preparation: WANG LISHUANG
>
> Tester: HE Zean, WANG Zihang
>
> Extended from the project of Spring 2022 and 2023

General Requirement:

- It is a group project with **only 2 teammates** who are **in the same lab session**. Each group should finish the project independently and submit only one report written by the teammates.
    - The teammate you select for Project 1 will also be your teammate for Project 2. It is not allowed to change teammates once paired.
- You should submit the report before the deadline. All late submissions after the deadline will receive a score of zero.
- DO NOT copy ANY sentences and figures from the Internet and your classmates. Plagiarism is strictly prohibited in this course.
- The text description should be rigorous, the overall design should be logical organised, the report structure and the layout of diagram should be clear and easy to read, otherwise, you will receive a penalty in the scoring stage.
- The number of pages for your report should be between **6** and **11**. Reports **only or less than 6 pages** and **more than 11 pages** will receive a penalty in the scoring stage.

DBMS can help us manage data in a convenient manner and improve the efficiency of data retrieval. Your work of Project 1 is mainly divided into three parts below:

1. Design an E-R diagram based on the provided data file and data relationships.
2. Design a relational database using PostgreSQL according to the provided data file.
3. Import all data into the database.

Background

Data Description

- lines.json

```
"lines": {
        LINE_NAME: {
            "stations": The stations that the subway line passes through,
arranged in order,
            "start_time": Time of the first train,
            "end_time": Time of the end train,
            "intro": Introduction of the line, including Chinese and
English name of line, and the infroduction.
            "mileage": Length of line(KM),
            "color": Color of line,
            "first_opening": The opening time of line,
            "url": URL of Baidu Encyclopedia
        }
    }
```

- stations.json

```
"stations": {
        STATION_English_NAME: {
            "district": District of Station such as Nanshan, Futian,
            "bus_info": The bus line info around the entrains of stations,
            "out_info": The buildings around the entrains of stations,
            "intro": Introduction of stations,
            "chinese_name": English name of Station,
        }
    }
```

- passenger.json

```
{
    "name": Name of passenger,
    "id_number": ID number of passenger(It is unique),
    "phone_number": phone number of passenger,
    "gender": gender of passenger,
    "district":  Identity of passenger including: "Chinese Mainland",
"Chinese Taiwan", "Chinese Hong Kong", "Chinese Macao"
}
```

- cards.json

```
{
    "code": card number (It is unique),
    "money": balance of the card,
    "create_time": opening time of card
}
```

- ride.json

```
{
      "user": id_number of passenger or the code of card number,
      "start_station": the depart station of passenger,
      "end_station": the arrive station of passenger,
      "price": the price of current journey,
      "start_time": the time of entering into the station,
      "end_time": the time of leaving the station
}
```

The Report and your Tasks

Basic Information of Your Group

1. Names, student IDs, and the lab session of the group members
2. You are required to write down the contributions and the percentages of contributions for each group member. **Please clearly state which task(s)/part of the task(s) is/are done by which member in the group.**
    - **If you failed to link a task/part of a task to one of the group members, we will not count the score for the task** (since we don't know who accomplished this task; maybe it was done by an elf while you were sleeping at night?).

About one page

Task 1: E-R Diagram (30%)

Make an E-R Diagram of your database design with any diagram software. Hand-drawn results will not be accepted. Please follow the standard of E-R diagrams.

In the report, you are required to provide a snapshot of the E-R diagram. Also, please specify the name of the software/online service you use for drawing the diagram.

About one page

Task 2: Relational Database Design (30%)

Design the tables and columns based on the background provided above. Generate the E-R diagram via the "Show Visualization" feature. Briefly describe the design of the tables and columns including (but not limited to) the meanings of tables and columns.

In the report, you are required to provide the following content:

1. Attach the snapshot of the E-R diagram generated by DataGrip.
2. Briefly describe the table designs and the meanings of each table and column.

In addition, please submit an SQL file as an attachment that contains the DDLs (`create table` statements) for all the tables you created. **Please make it into a separate file but not copy and paste the statements into the report.**

About one or two pages

Notes for the database design:

1. All data items should base on five files `lines.json`, `stations.json` , `pessenger.json`, `cards.json` and `ride.json` .
2. Your design needs to follow the requirements of the three normal forms.
3. Every row in each table should be uniquely identified by its primary key. (You may use a simple or a composite primary key).
4. Every table should be involved in a foreign key. No isolated table is allowed. (每个表要有外键，或者有其他表的外键指向。)
5. Your design should contain no circular foreign-key links. (对于表之间的外键方向，不能有环。例如：A表有外键关联B表，B表有外健关联C表，C表有外健关联A表)
6. Each table should contain at least one mandatory (“Not Null”) column (including the primary key but not the id column).
7. Other than the system-generated self-increment ID column, there should be at least one column with the "unique" constraint. (除了主键自增的id之外，需要有其他unique约束的列)
8. You should use appropriate data types for different fields.
9. Your design should be easy to expand when requirements change.

Task 3: Data Import (40%)

In this task, you should write scripts to import the content in those five json files into the database you have designed before. After importing the data, you should also make sure all data is successfully imported.

Task 3.1 Basic Requirements: 10%

导入过程要写，流程图/列step说清楚

1. Introduce the scripts that being used to import data. You can list a similar table below to describe the functions of all the scripts you submit in the attachment.

Script name	Author	Description
Script1.py	xxx	Run this script to complete the import of passenger data.

2. A description of how you use the script to import each json file. In this part, You should clearly state the steps.
    - You can create an intermediate file based on the original json file, and then import the intermediate file through a script.
    - You can also directly parse the json file in a programming language and import it line by line.
    - If you manually modified the file, please also explain the specific modifications.

About one or two pages

Task 3.2 Data Accuracy checking: 15%

According to the data in those five json files, we will **give several questions** in presentation week to check whether all data have been correctly imported into your database. Please prepare SQL queries for the following requirements **in the report**, and you will run the queries by **datagrip** in this part **during your presentation week**.

1. The numbrer of stations, in each district, on each line or in total.
2. Number of female passengers and male passengers respectively.
3. List the number of passengers from Mainland China, Hong Kong, Macau, and Taiwan.
4. List the buses, buildings, or landmarks near a specific **station exit**.
5. List all information about a specific passenger's journey, including passenger name, entry station, exit station, date, and time.
6. List all journey records for a specific travel card, including card number, entry station, exit station, date, and time.
7. Query information about a specific subway station, including Chinese name, English name, number of exits, the district it is located in, and the subway line it belongs to.
8. Query information about a specific subway line, including start time, end time, first opening time, number of stations, and an introduction.

About one page

Task 3.3 Advanced requirements: 15%

You may also need to finish the following advanced requirement to get the remaining points.

prepare statement, batch... 还要+：多线程。效率（导入用时，技术）：用ride.json测试---10w条

分比较高

1. Try to **optimize your script**, and find **more than one ways** to import data, and provide a comparative analysis of the computational **efficiencies** between these ways.
2. Try to import data across multiple systems (e.g., Windows, MacOS, Linux). 同样代码换系统，比较效率
3. Try to import data using various programming languages (e.g., Java, Python, C++). 语言
4. Experiment with other databases; we recommend use **OpenGauss**. 换数据库（华为数据库）...要安装等
5. Try to import data with different data volumes.

另外找数据？ ride是10w条数据，passenger是1w条

10w（给出）
可以针对1w，5w，20w等。用图表展示结果会有report页面等加分

For the advanced points, please make sure to describe your test environment, procedures, and actual time costs. It is required to write a paragraph or two to analyze the experiment results.

About 0 to 4 pages

How to Submit Your Report

1. A report in PDF format
2. A .sql file about your DDL queries.
3. Script files in part3 about importing data

   Submit all files above on the bb website before **13:30 on April 30th, 2024, Beijing Time (UTC+8)**, and compress them into a `.zip` archive.

Disclaimer

The names, characters, and events in the background of this project are purely fictional. The data of lines are based on Baidu Encyclopedia. The items in the names or cards are generated by chatGPT. Any resemblance to actual events, entities or persons is entirely coincidental and should not be interpreted as views or implications of the teaching group of CS307.