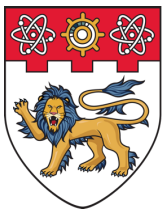

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CE/CZ4123 BIG DATA MANAGEMENT

SEMESTER GROUP PROJECT

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
NANYANG TECHNOLOGICAL UNIVERSITY

1 ASSIGNMENT DESCRIPTION

The goal of this semester project is to conduct a simple analysis on weather time series data to have a flavor of the data management process. You will be given weather data in Singapore in every hour of last 20 years. Here, we provide two types of weather data, temperature and humidity. The data also contains additional information including date (January 2002 – December 2021) and location (Changi and Paya Lebar). Your program needs to search for the respective monthly maximum and minimum values of temperature and humidity columns with given year and location condition.

You are expected to write a program to manage the data in a **column-oriented** manner, including data storage and processing. Your program should first receive queries, then scan the columns of the data, and find out the lines satisfying task requirements.

Your program should input a matriculation number as a query, which scans the weather data corresponding to two years and one location. The year and location are determined by the following rule: the last digit of the required years equals to the last digit of the matriculation number of one of your group members; the location depends on the second last digit of the matriculation number, with even number for Changi and odd number for Paya Lebar. Use the numbers from all your group members to generate corresponding queries.

Example: The student with matriculation number A1234567B should scan the year 2007 and 2017 at Changi (note: because the second last digit 6 is even) and find the four extreme values of each month in these years. In this case, the search task for max humidity in January 2007 should be equivalent to the following SQL query:

Task in SQL

```sql
WITH Tab1 AS (
    SELECT *
    FROM   SingaporeWeather
    WHERE  (YEAR(Timestamp) = 2007)
           AND (MONTH(Timestamp) = 1)
           AND (Station = 'Changi')
)
SELECT DISTINCT DATE(Timestamp), Station, Humidity
FROM   Tab1
WHERE  Humidity = (
    SELECT MAX(Humidity)
    FROM   Tab1
)
```

2 INPUT FORMAT

The input file `SingaporeWeather.csv` is the historical records for weather data in Singapore. The data is extracted from the Singapore ASOS system* located at Changi Climate Station and Paya Lebar Meteorological Station, which automatically measures the weather information every 30 minutes. In this assignment, we focus on the temperature and humidity data with the time span January 1, 2002 – December 31, 2021.

The input data is given in `.csv` format. You can download the data via NTU Learn. The first row is the title row. Each following row contains a line of weather data entries, separated by a comma "`,`". Empty data are marked as "`M`". The columns of the records are listed as follows:

- **`id`**: the increasing index of weather records.
- **`Station`**: "`Changi`" or "`Paya Lebar`", represents the site of the observation.
- **`Timestamp`**: the timestamp of the observation, in format `YYYY-MM-DD hh:mm` of UTC+8 time zone.
- **`Temperature`**: air temperature in degrees celcius (℃).
- **`Humidity`**: relative humidity in %.

3 OUTPUT FORMAT

Each output file `ScanResult_<matriculation_ID>.csv` should contain the results from one matriculation number query. The first row is the title row. Each following row should contain one maximum or minimum value of temperature or humidity and the corresponding date, separated by a comma "`,`". The columns are listed as follows:

- **`Date`**: the corresponding date of the **`Value`**, in format `YYYY-MM-DD` of UTC+8 time zone.
- **`Station`**: "`Changi`" or "`Paya Lebar`", represents the station of the **`Value`**.
- **`Category`**: "`Max Temperature`", "`Min Temperature`", "`Max Humidity`", or "`Min Humidity`", represents the meaning of the **`Value`**.
- **`Value`**: the value of temperature or humidity.

There is no restriction on the order of result rows. If there are multiple dates reaching the same extreme value, output all applicable rows. If there is no data for the whole month, omit the month in the output.

Example: In January 2007, the maximum humidity at Changi is 100% and is found in days including `2007-01-02` and `2007-01-03`. Then the corresponding result rows in the output file should be:

ScanResult_A1234567B.csv (example)

```
Date,Station,Category,Value
2007-01-02,Changi,Max Humidity,100
2007-01-03,Changi,Max Humidity,100
...
```

*Data from: IEM.

4 SUBMISSION

Time: During Week 14 (By April 20 unless otherwise specified)

Method: Via NTULearn

The required files include the output file, the source code of your program, and an assignment report. They should be compressed and submitted in a `.zip` file. Name the `.zip` file with your matriculation number + full name. The requirements of each files are as follows:

- Output Files `ScanResult_<matriculation_ID>.csv`: the scan results following the requirements in **Output Format** Section.
- Source Code `source`: the file or folder containing the source codes that input the file `SingaporeWeather.csv` and the matriculation number, and output the corresponding `ScanResult_<matriculation_ID>.csv`. Source codes should be well-commented and contains essential documentations to help understand the functionalities.
- Report `Report.pdf`: the report exported in `.pdf` format. Your report sections and contents should follow the requirements in **Report Format** in Appendix. The report should be **at most** 5 pages (single column, font size=11, excluding the contribution form).

5 FORMING GROUPS

The expected group size is 3. The group will be allocated by TA and any swap between the groups should be approved by TA.

6 ASSESSMENT

This is a **group project**. Your submission will be evaluated on the comprehensive basis including design sophistication (e.g., what if data go big and cannot be stored in main memory), output accuracy, code quality, and report quality. Late submission will be penalized. The evaluation of an individual is based on the contribution form.

7 GENERAL GUIDELINES

1. If you find it tricky in handling the `.csv` format file, you can change it to the `.txt` format, and regard it as a plain text file.

2. You may assume that the `id` column is monotonically increasing, while the `Timestamp` column may not. Note that there are empty data in the input file marked as "`M`". You can process these empty records freely as long as it does not affect the final output.

3. While we recommend JAVA, you are free to choose any programming language in case you are not familiar with JAVA.

4. We suggest to avoid using high-level tools when storing and processing the data, such as pandas in Python. Please ensure that your program is implemented in the column-store manner. (note: simple implementation based on SQL is not allowed because it is not in the column-store manner.)

Appendix: Report Format

REPORT FORMAT

Name and Matriculation Number

1 Data Storage

In this section, explain how your program handles and stores the data. You may present your design and experience (whether success or failure) related to:

- How to store the data in the column-store approach[†];
- How to design data columns for efficient processing[†];
- How to read and write the input/output files;
- How to handle exceptions (empty entries, absent month, etc.).

2 Data Processing

In this section, explain how your program scans the data and finds the values. You may present contents related to:

- How to scan columns according to task conditions;
- How to decide and record maximum and minimum values;
- How to improve the efficiency in scanning columns[†];

3 Experiment Result

In this section, present the experimental results that your program successfully complete the tasks. The following contents are compulsory:

- Screenshots that your program executes and outputs results successfully;
- Evaluations that the output results are correct.

[†]Exploration and improvements on these aspects are encouraged.

CONTRIBUTION FORM

Group Number in Grouping Form

Name	**Detailed Individual Contribution**	**Percentage (100% in total)**

Name and Signature from all group members:

Name and Signature of Member 1

Name and Signature of Member 2

Name and Signature of Member 3

Name and Signature of Member 4