CSE4.401: DATA SYSTEMS
ASSIGNMENT 1

DEADLINE: 17:00, 20th January, 2021

In this assignment, you are supposed to develop a mini sql engine which will run a subset of SQL queries using command line interface

Programming Languages Allowed : Java, Python and C/C++

Dataset:

1. csv files for tables.
    a. If a file is : File1.csv , the table name would be File1
    b. There will be no tab separation or space separation, so you are not required to handle it but you have to make sure to take care of both csv file type cases: the one where values are in double quotes and the one where values are without quotes.

2. All the elements in files would be only **INTEGERS**

3. A file named: metadata.txt (note the extension) would be given to you which will have the following structure for each table:

```
<begin_table>
<table_name>
<attribute1>
....
<attributeN>
<end_table>
```

4. Column names are unique among all the tables. So column names **are not** preceded by table names in SQL queries.

Type of Queries: You’ll be presented with the following set of queries:

1. **Project** Columns(could be any number of columns) from one or more tables :

    a. Select * from table_name;
    b. Select col1, col2 from table_name;

2. **Aggregate functions**: Simple aggregate functions on a single column. Sum, average, max, min and count. They will be very trivial given that the data is only integers:

    Select max(col1) from table_name;

3. Select/project with **distinct** from one table: (distinct of a pair of values indicates the pair should be distinct)

   Select distinct col1, col2 from table_name;

4. Select with **WHERE** from one or more tables :

   Select col1,col2 from table1,table2 where col1 = 10 AND col2 = 20;

   a. In the where queries, there would be a maximum of one AND/OR operator with no NOT operators.
   b. Relational operators that are to be handled in the assignment, the operators include "< , >, <=, >=, =".

5. Select/Project Columns(could be any number of columns) from table using "**group by**":

   Select col1, COUNT(col2) from table_name group by col1.

   a. In the group by queries, Sum/Average/Max/Min/Count can be used as aggregate functions.

6. Select/Project Columns(could be any number of columns) from table in ascending/descending order according to a column using "**order by"**:

   Select col1,col2 from table_name order by col1 ASC|DESC.

   a. At max only one column can be used to sort the rows.
   b. Query can have multiple tables in it.

IMPORTANT:

- **ERROR HANDLING**: 10% marks will be for error handling.
- For the above queries, please note all the permutations and combinations of SQL that MySQL permits, specially when it comes to multiple tables. What is mentioned above are examples of what the queries could be.
- You **CAN** use pre built parsers for SQL queries

Format of Input :

Automated evaluation will be done, so include a bash script in your final submission to run the queries. Command to execute a sql query and would be of the form:

<RollNumer.sh> "SQL Query"

Here, the SQL Query would be a command line argument. Example :

"select col(s) from table_name where condition"

Format of Output:

```
<Table1.column1,Table1.column2,...      TableN.columnM>
                    Row1
                    Row2
                    .......
                    RowN
```

Deliverables:

1. Java/Python/C++ Source Code files.
2. Compiled Java/C++ files.
3. Bash script as your roll number.
4. Keep all of this in a folder Roll Number, zip it and upload.

Automated evaluation will be done for this assignment, so strictly stick to the input format described in this document and the output format as per the example files provided.

Copying in the Assignments can lead anywhere from a Zero in the current assignment to an F in the course.