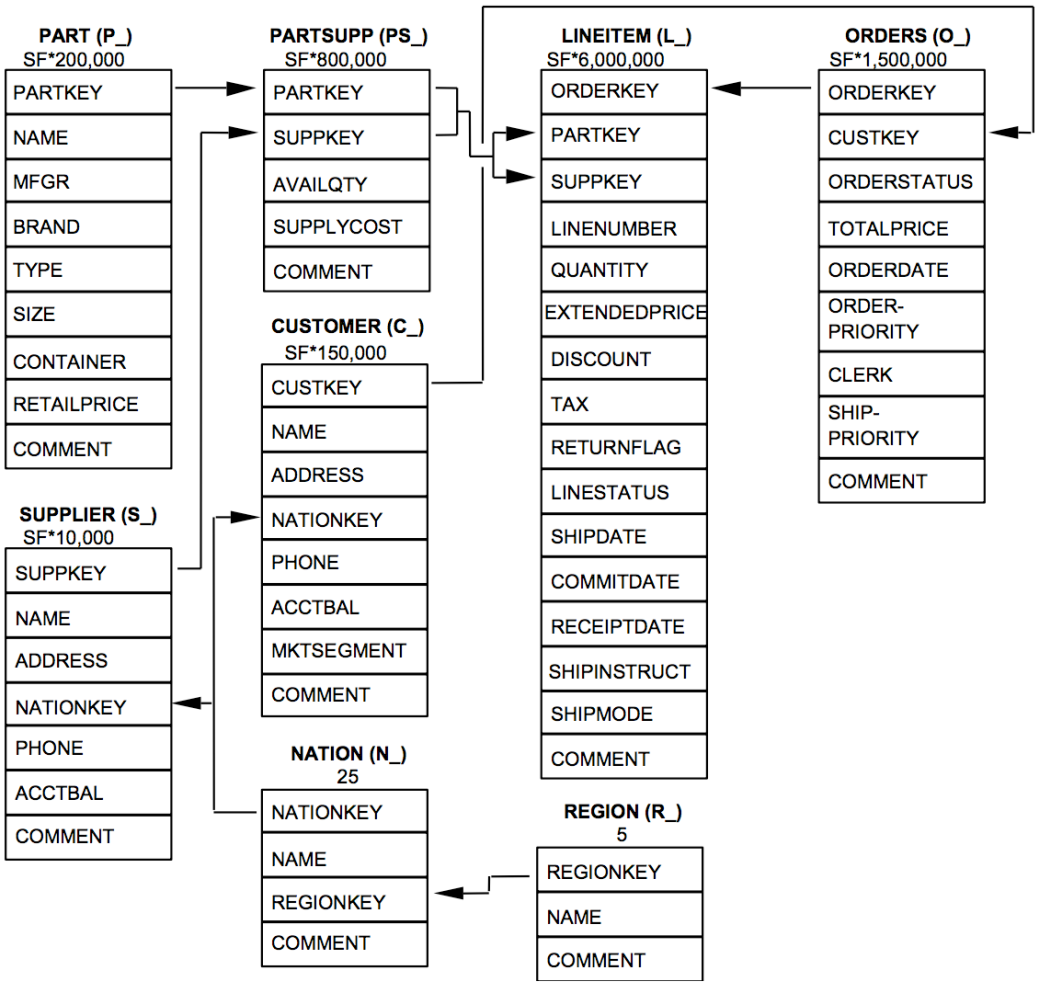
Programming Assignment 4 - 100 points

Due Date: October 03, 2025, 11:59 pm

Objective:

The goal of this assignment is to continue writing SQL queries. We will continue to use the same TPC-H E-R diagram for this assignment. The goal for this assignment is to execute SQL queries through python and measure the performance of the queries on your laptop.

Figure 2: The TPC-H Schema



Legend:

- The parentheses following each table name contain the prefix of the column names for that table;
- The arrows point in the direction of the one-to-many relationships between tables;
- The number/formula below each table name represents the cardinality (number of rows) of the table. Some are factored by SF, the Scale Factor, to obtain the chosen database size. The cardinality for the LINEITEM table is approximate (see Clause 4.2.5).

Create the tpc_h schema and create the tables. Then load the data into those tables. The database script is provided as part of the assignment. The script file is named: tpch_database.sql There is a zip file named: tpc_h_data.zip. This zip file contains the data files for each of the tables. These files are | delimited.

Write SQL queries to get the following answers:

1. Find all customers whose phone numbers start with a prefix of '1' and format their account balance to 2 decimal places. Return only TOP 5 customers based on the decreasing order of the balance.
   Return: (customer_name, formatted_balance, phone_prefix) [10 points]

2. Find all distinct part sizes and calculate what the size would be if increased by 10%. Only include parts with size greater than 5. Return the results in the increasing order of the original size.
   Return: (original_size, increased_size) [10 points]

3. Categorize all orders based on their total price and priority. Create categories for price ranges and show order priority distribution. Return the results in the increasing order of order_priority followed by decreasing order of the total order_count.

   If totalprice >=300000 then the price_category is 'HIGH_VALUE'
   If totalprice >=150000 and totalprice <300000 then the price_category is 'MEDIUM_VALUE'
   If totalprice >=50000 and totalprice <150000 then the price_category is 'LOW_VALUE'
   If totalprice >0 and totalprice <50000 then the price_category is 'MINIMAL_VALUE'
   Else for all other totalprice the price_category is ‘ZERO_VALUE’
   Return: (order_priority, price_category, order_count) [20 points]

4. Find all orders placed between July 1, 1995 and December 31, 1995 with order priority '1-URGENT' or '2-HIGH'. Return the orders in the order of orderdate and orderpriority.
   Return: (orderkey, orderdate, order_priority, customer_name) [10 points]

5. Find top 10 customers by comment length and identify complaints. The complaint keywords are “complaint” or “problem”. Return results in the decreasing order of length of the comments.
   Return: (customer_name, comment_length, has_complaint) [10 points]

6. Find the most expensive part supplied by each supplier. Return suppliers in the decreasing order of supply cost.
   Return: (suppkey, supplier_name, partkey, part_name, max_supplycost) [20 points]

7. Find suppliers who supply more than 5 different parts. Return suppliers in the decreasing order of number of parts supplied.

Return: (suppkey, supplier_name, parts_supplied) [20 points]

Tasks:

1. We are providing you a template .py file that you have to fill in the functions for the 7 queries above. This template file has functions that measures the performance of the query execution on your machine. Along with the template file, there is user_input.py file that contains parameters for your database connection.
2. Rename the template file as pa-4.py file and write in all the 7 queries in the 7 function placeholders.
3. Along with the pa-4.py file, also save the output of the execution and submit that on canvas. The output file name should be pa-4-output.txt

*** There is no partial grading of the queries. If the query returns the correct answer, you get all the points allocated for that query, else you get 0 points for that query. Your Query will be tested against the same exact data from the tpc_h_data.zip file.***