Information and Database Management Systems I

(CIS 4301 UF Online)

Fall 2024

Instructor: Alexander Webber

TA: Kyuseo Park

Homework 2

Printed Name: Kyle Lund
UFID: 275 b to 39
Email Address: Kyle Lund af Led n

Instructions: Please provide your answers to the questions of the following pages in Word or handwritten on separate sheets of paper. Mark clearly to which question each answer belongs. Then convert or scan your work into PDF (the latter by using either a scanner or a suitable scanner app on your smartphone). Note that only the PDF format is allowed and that your submission must be a zipped along with your source code and work for Question 5. In order to enable the graders to fast find the solutions to your questions, it is important that you correctly specify the location of your answer for each question.

Note: All homework assignments are designed for a period of two, three, or even four weeks (see course deadline sheet). This means they cannot be solved in two or three hours but require a considerable amount of time and effort. Therefore, the first recommendation is to start with them as soon as they are posted. The second recommendation is to distribute the work on a homework assignment over the entire available period. The third recommendation is to submit the homework solutions on time before the deadline.

Pledge (Must be signed according to the UF Honor Code);

On my honor, I have neither given nor received unauthorized aid in doing this assignment.

Student signature

¹Each student is obliged to print out this page, fill in the requested information in a handwritten and readable manner, make the handwritten signature, scan this page into PDF, and put this page as the first page of the PDF submission.

```
1)
A) \rho_{E1}(\pi_{PatientID,DepartmentID}(PatientTreatments \bowtie Physicians))
\rho_{E2}(\pi_{PatientID, DepartmentID}(PatientTreatments \bowtie Nurses))
\pi_{FirstName,LastName,DepartmentID}(Patients \bowtie E1 \bowtie E2)
\mathsf{B})\,\rho_{E1}(\pi_{PatientID}(\sigma_{FirstName="PatientA"}(Patients)) \cup \,\,\pi_{PatientID}(\sigma_{FirstName="PatientB"}(Patients)))
\rho_{E2}(\pi_{PhysicianI~,PatientID}(\sigma_{date~"Nov~11"}(Appointments)))
\rho_{E3}(\pi_{PhysicianID}(E1 \bowtie E2))
\pi_{FirstName,LastName}(Physicians \bowtie E3)
C) \rho_{E1}(\pi_{PhysicianI} \ (\sigma_{Name=Gene} \ \ Sugery(Departments \bowtie Physicians)))
\rho_{E2}(\pi_{PhysicianI} \ (\sigma_{FirstName="NurseA}(Nurses \bowtie PatientTreatments \bowtie E1)))
\pi_{FirstName,LastName}(E2 \bowtie Physicians)
\mathsf{D})\,\rho_{E1}(\pi_{PatientID}(\sigma_{Physicians.FirstName="PhysicianA"}(Patients \bowtie Physicians \bowtie Appointments)))
\rho_{E1}(\pi_{PatientID}(PatientTreatments))
\pi_{FirstName,LastName}(E1-E2)
\mathsf{E}) \, \rho_{E1}(\pi_{PatientID,Cost}(PatientTreatments)) \, \mathsf{E}) \, Patients \bowtie Treatments))
\rho_{E4}(\pi_{PatientID}(\pi_{PatientID}(E1) - \pi_{PatientID}(\sigma_{E2,Cost}) = 3.Cost(\rho_{E2}(E1)X \rho_{E3}(E1))))
\pi_{FirstName.LastName.Cost}(E1 \bowtie E4)
2)
A) \pi_{Date,Location,Description}(\sigma_{Name="pilotA"}(Accidents \bowtie FlightPilotAssignments \bowtie Pilots))
B) \rho_{E1}(CrewMembers \div \pi_{CrewID}(\sigma_{Name="PilotA"}(FlightCrewAssignments \bowtie
FlightPulotAssignments \bowtie Pilots)))
\pi_{FirstName,LastName}(CrewMembers \bowtie E1)
C) \rho_{E1}(\pi_{AircraftID}(Accidents \bowtie Flights))
\pi_{ModelName.Manufacturer}(Aircraft - E1)
                                             (CrewMembers)))
D) \rho_{E1}(\pi_{CrewID}(\sigma_{YearsOfExper}))
\rho_{E2}(\pi_{CrewID}(FlightCrewAssignments \bowtie FlightPilotAssignments \bowtie \sigma_{Name} "pilotA"(Pilots)))
\pi_{FirstName,LastName}(CrewMembers \bowtie (E1 \cap E2))
\mathsf{E})\,\rho_{E1}(\pi_{FlightID}(\sigma_{Origin="New York" \, \land Destination \, "London" \, \land Date="December \, 25}(Flights)))
\pi_{FirstName,LastName}(E1 \bowtie Tickets \bowtie Passengers)
3)
A) \sigma_{F1 \wedge F2 \wedge F3 \wedge F4(R)}
This expression runs in O(r), evaluating each tuple in R only once. It is optimal as each tuple in a
select operation needs to be read, evaluated, and possibly written at least once, and this query
evaluate each of the 4 filter functions on the same tuple without needs to iterate through the data
more than once.
```

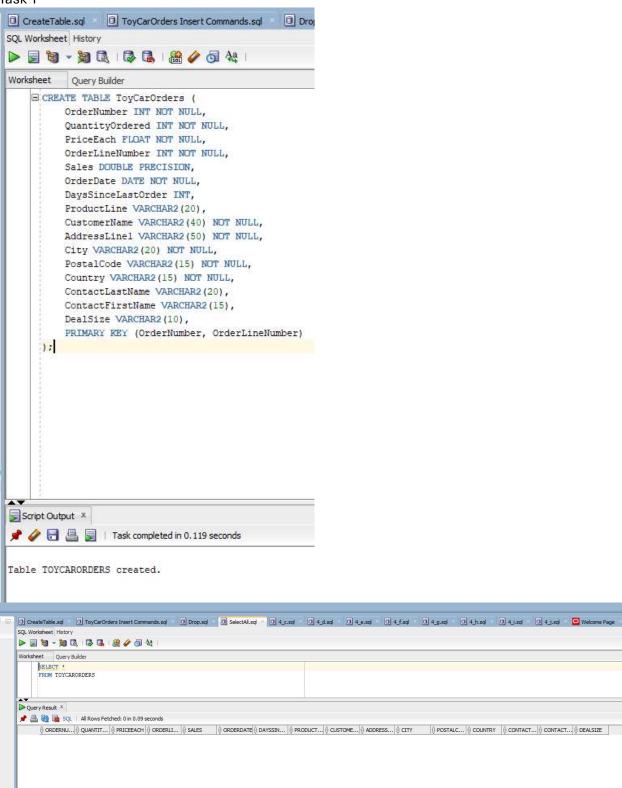
B) $L1 \subseteq L2 \subseteq L3$

This condition ensures that subsequent project operations do not attempt to store previously eliminated attributes. The expression could be optimized by eliminating the L2 and L3 projections, as either way, we still will only end up with the L1 projection.

- C) Possibly but not always. In order for $\sigma_F(\pi_A)$ to be equivalent to $\pi_F(\sigma_A)$, all of the attributes that are being used in F must be present in A. This is not the case for $\pi_F(\sigma_A)$.
- D) This is not a valid operation. For difference operations, R1 and R2 must be schema compliant.
- 4)
- A) Minimum 0, occurs where no tuples are similar between R and T (R intersect T yields null set). Maximum is the smaller of r and t (min(r, t)), occurs when one (or both) of r or t is a subset of the other.
- B) Minimum 0, occurs where no tuples are similar between R and T (R intersect T yields null set). Maximum is the smaller of r and t (min(r, t)), occurs when one (or both) of r or t is a subset of the other.
- C) The minimum is max(r,s), occurs when one of R or S is a subset of the other. The maximum is r + s, occurs when R and S are disjoint (R union S yields null set).
- D) Minimum 0, occurs where no tuples are similar between R and T (R intersect T yields null set). Maximum r, occurs when all tuples in R have a match in T.

5)

Task 1

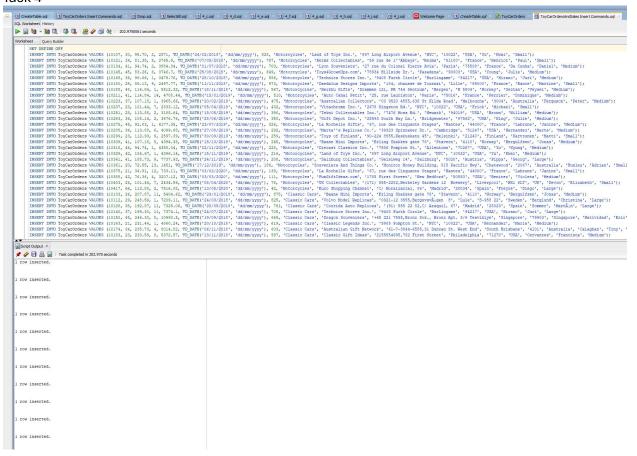


Task 3 I utilized a python script to transform the data as specified. Code is shown here:

```
from operator import itemgetter
USECOLS = [0, 1, 2, 3, 4, 5, 6, 8, 11, 13, -6, -5, -4, -3, -2, -1]
with open("UF\CIS4301\m4\ToyCarOrdersAndSales Insert Commands.sql", "w") as write file:
    # Ignore '&'
    write_file.write("SET DEFINE OFF\n")
    with open("UF\CIS4301\m4\Auto Sales data.csv", "r") as read_file:
        for line in read file.readlines()[1:]:
            data = line.replace('"', "").replace("'", "''").strip().split(",")
            # Check if adress got split
            if len(data) > 20:
                data[13] = ",".join(data[13 : 13 + len(data) - 19])
            data = list(itemgetter(*USECOLS)(data))
            data[5] = f"TO_DATE('{data[5]}', 'dd/mm/yyyy')"
            for col in range(7, len(data)):
                data[col] = f"'{data[col]}'"
            line_data = ", ".join(data)
            write_str = f"INSERT INTO ToyCarOrders VALUES ({line_data});"
            # Write to file
            write file.write(write str + "\n")
```

Starting from the top, I first open the write file, and before anything else, add the command "SET DEFINE OFF". This command prevents SQL Developer from asking for an input variable whenever a "&" is discovered in a string. Next the read file is opened. Starting from the first line of data, each line is read, any random quotation marks are removed, single quotes are escaped with double single quotes (to prevent the program from thinking the string has ended prematurely), '\n''s are stripped from the end of the line, and finally split into a list by commas. Then the script checks whether the address block had any commas which caused an undesired split, and if this occurred, reconsolidates the address data. Then, the data from non-desired columns is eliminated. After that, the date is converted to the desired SQL format using f-strings and single quotes are added around the remaining string data. Finally, the list of data is recombined with commas, wrapped with the necessary SQL INSERT commands, and written into the desired file.

Task 4



Task 5

