Name:
Section (315/415/615):

FINAL EXAM - 600.315/415/615 - Databases

Date: December 14 2020, 6-9 PM (Baltimore Time)

The total number of points in this exam is approximately 113 for 600.315 students, and 128 for 600.415/615 students. If you work at approximately 0.66 minute per point, you should finish on time.

Relational Algebra Shortcuts:

⋈ -> JOIN	π -> PI
⋈ -> RIGHT JOIN	ρ -> RHO
⋈ -> LEFT JOIN	σ -> SIGMA
∩ -> INTERSECT	∪ -> UNION
— -> MINUS or just -	^ -> AND or just ^

Relational Calculus Shortcuts:

∃ -> EXISTS	∄ -> NEXISTS
¬ -> NOT	∈ -> IN

QBE Shortcut: ¬ -> NOT

Question 1 - Relational Algebra (5 points)

Express the following query in the Relational Algebra. The tables that are used in this (and following) questions are found on your supplementary handout.

• (5 points) Print the name and age of the oldest politician listed in the database (including all ties) You may not use aggregate operators for this question.

PI_(PName,Age)(POLITICIAN JOIN (PI_{Age} SIGMA_{ROWNUM = 1} (POLITICIAN ORDER BY Age DESC)))

Question 2 - Relational Algebra (5 points)

(a) (5 points) **600.315 only**: List the names of all cities which contain more than 50% of the population of their host country.

(b) (5 points) **600.415/615 only**: List the names of all countries where 1 or 2 cities contain a total population of over 50% of the population of their country

Answer:
(b) (PI_{Population} SIGMA_{ROWNUM = 1) (CITY ORDER BY Population DESC) +
PI_{Population} SIGMA_{ROWNUM = 1) (CITY ORDER BY Population DESC))

Question 3 - Tuple Relational Calculus (5 points)

Express the following query in Tuple Relational Calculus:

• (5 points) List the names of all countries who do not have a female politician listed in the database but have a literacy rate greater than 90%.

{T |
EXISTS POLITICIAN AS P1(
EXISTS COUNTRY AS C(
NOT EXISTS POLITICIAN AS P2(

WHERE P2.Gender = 'Female'
AND P1.Country = P2.Country)
AND P1.Country = C.CNBame
AND C.LitRate > 90

T.Country = P1.Country));

Question 4 - Relational Algebra (3 points)

Express the following query in the relational algebra.

(a) (3 points) **600.315 only**: List the names of all female prime ministers who have visited every country in Europe (you can assume there are never two politicians with the same name).

(aa) 3 points) **600.415/615 only**: List the names of all female prime ministers who have visited every country in Europe (you must handle the case where there are two prime ministers with same name but different countries)

PI_(PName)(SIGMA_(Gender = 'Female')(VISITED DIV PI_(CName)(SIGMA_(Continent = Europe) COUNTRY)) JOIN POLITICIAN))

Question 7 - SQL (15 points)

Express the following queries in SQL:

(a) (5 points) **600.315 only**: List the name, GNP and population of the country with the smallest GNP per capita (GNP/population) in the entire database.

(aa) (5 points) **600.415/615 only**: List the country where the listed politicians have the oldest average age.

(aa)
SELECT DISTINCT P.Country
FROM POLITICIAN AS P
GROUP BY P.Country
HAVING AVG(Age) =
SELECT MAX(AvgAge)
FROM (
    SELECT AVG(Age) AS AvgAge
    FROM POLITICIAN
    GROUP BY Country
) AS M

(b) (5 points) List all countries that do not border Mexico or Canada directly, but border a country that borders Mexico or Canada.

```
SELECT DISTINCT B1.Country
FROM BORDERS AS B1,
   (
      SELECT B.Country1
      FROM BORDERS AS B
      WHERE B.Country2 = 'Mexico'
        OR B.Country2 = 'Canada'
   ) AS M
WHERE (B1.Country2 IN M)
   AND (B1.Country1 NOT IN M);
```

(c) (5 points) In a single table, list the names of all continents associated with the names of the country with the smallest population in that continent (e.g.)

SELECT DISTINCT continent,
    MIN(population) AS MinPop
FROM COUNTRY AS C
GROUP BY continent

Question 8 - QBE (10 points)

Express the following queries in QBE. To simplify your work, table shells have been provided. Just fill in the appropriate cells with variables/values.

(a) (5 points) **600.315 only**: List the names of all countries that border a country that borders a country that borders the United States.

(aa) (5 points) **600.415/615 only**: List the names of all prime ministers of countries which border a country which borders a country which has a life expectancy less than 35.

COUNTRY	CName	Continent	GNP	Population	LitRate	LifeExp
	_a					<35

BORDERS		Country1	Country2
		_b	_a
		_c	_b

POLITICIAN	PName	Gender	Office	Country	District
	P._d			_c	

(b) (5 points) **600.315 only**: List the names of all politicians who have visited at least one country that borders a country where they hold office.

(bb) (5 points) **600.415/615 only**: List the names of all politicians who have visited every country that borders a country where they hold office.

COUNTRY	CName	Continent	GNP	Population	LitRate	LifeExp

BORDERS		Country1	Country2
		_b	_a

POLITICIAN	PName	Gender	Office	Country	Age
	P._c			_a	

VISITED		PName	Country
		_c	_b

Question 9 - Relational Algebra (6 points)

Given the following relations r, s and t:

r

A	B	C
g	20	2
e	40	1
e	50	2
d	10	1
g	10	2
g	10	3
d	30	1
d	30	2
d	30	3

s

C
1
2
3

t

A	B
g	10
d	30

(a) (3 points) Compute r - (t × s)

t × s:

A	B	C
g	10	1
g	10	2
g	10	3
d	30	1
d	30	2
d	30	3

r - (t × s):

A	B	C
g	20	2
e	40	1
e	50	2
d	10	1

(b) (3 points) Compute r ÷ s:

A	B
d	30

Question 10 - Functional Dependencies (12 points)

Question 1 Suppose you are given a relation R = (A,B,C,D,E) with the following functional dependencies: CE → D, D → B, C → A.

a. List all the candidate keys of R.
CE

b. Is R in 3NF? Is it in BCNF? Why?
Answer:
(1)No. It is not 3NF. Because for D -> B, B is not in candidate key since CE is the only candidate key.
(2)No, it is not in BCNF. Because ABD is not super key and it's not trivial. This is conflict with BCNF rule.

c. If the relation is not in BCNF, decompose it until it becomes BCNF. At each step, identify a new relation, decompose and re-compute the keys and the normal forms they satisfy.

Answer:
   (1) Decompose R by CE -> D. Then we have R1 = (C,D,E), R2 = (A,B,C,E). We find that R1 is already a BCNF. Since this is not a dependency reserving decompose, D -> B is lost.
   (2) Decompose R2 by C -> A. Then we have R3 = (A,C), R4 = (B,C,E). In this case,R3 is BCNF, and R4 is 3NF
   (3) Since there is no further rule. This is the final decomposation.

Question 11 - Functional Dependencies (10 points)

Suppose you are given a relation R=(A,B,C,D,E) with the following functional dependencies:
BD → E, A → C.
Show that the decomposition into R1=(A,B,C) and R2=(D,E) is lossy.

Answer:
Because the R1 INTERSECT R2 = {}, and {} -> {}, neither derive R1 nor R2. So the decomposition is lossy.

Find a single dependency from a single attribute X to another attribute Y such that when you add the dependency X → Y to the above dependencies, the decomposition in part a is no longer lossy.

??
I think there's something wrong with the problem and the answer is impossible.

Question 12 - Functional Dependencies (15 points) (600.415/615 only)

You are given the below functional dependencies for relation R(A,B,C,D,E),
F = AB → C, AB → D, D → A, BC → D, BC → E.

a. Is this relation is in BCNF? If not, show all dependencies that violate it.

Answer:
No. D -> A violates the rule because D is not superkey.

b. Is this relation in 3NF? If not, show all dependencies that violate it.

Yes. Because. AB is a candidate key and A is in the candidate key, so D -> A follows the rule.
AB is candidate key(super key) and BC is candidate key(super key). So all of them are
candidate key.

Is the following dependency implied by the above set of dependencies? If so, use Armstrong's
Axioms to demonstrate that it is: ABC → AE.
Answer:
Yes it is implied. Using the augmentation rule in Armstrong's Axions. Since we have BC -> E,
then ABC -> AE.

Question 15 - Transaction/Concurrency Control (9 points)

(9 points) Consider the following 4 transactions, consisting of a sequence of read and write operations, with (omitted) intermediate code:

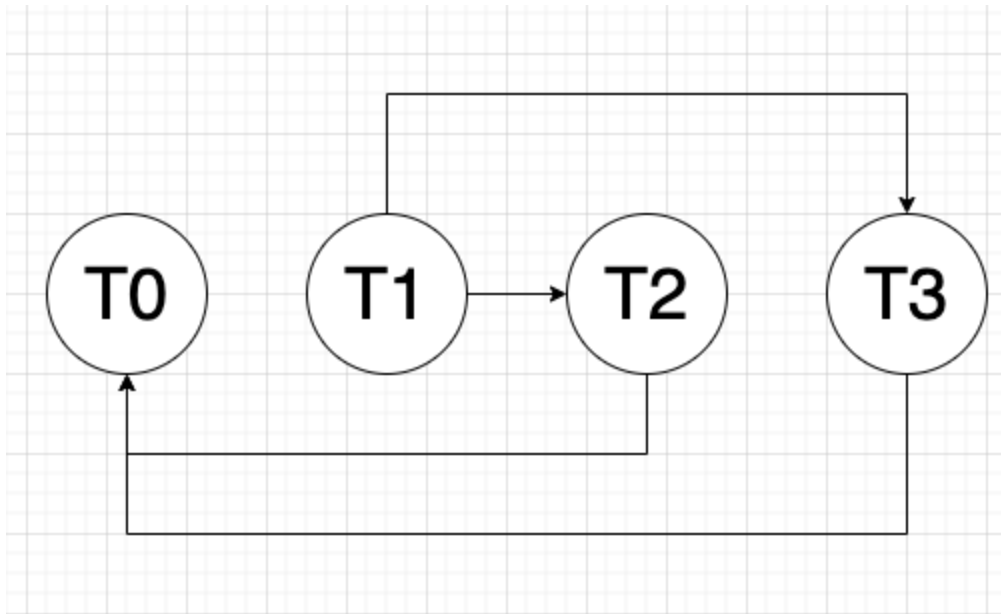| T0	T1	T2	T3
read item(W)	read item(X)	read item(Z)	write item(Q)
write item(W)	read item(Y)	read item(W)	read item(Y)
read item(Q)	write item(X)	write item(Z)	
	write item(Y)		

Also consider the following concurrent schedule of the above transactions:

T0	T1	T2	T3
	read item(X)	read item(Z)	write item(Q)
	read item(Y)	read item(W)	
read item(W)			
	write item(Y)	write item(Z)	
	write item(X)		read item(Y)
write item(W)			
read item(Q)			

Is this concurrent schedule conflict serializable? While you will not need to show the precedence graph in your answer, you should create one at home on paper and you should justify in writing if you do not believe that the schedule is serializable what is the cycle and over what attributes is each arc in the cycle based. If it is conflict serializable, give an equivalent serial schedule (an ordered list of transactions is fine) along with a topological sort.

Answer:

Yes, this schedule is conflist serialzable. You can draw the graph as below

So the equivelant schedule should be like

T0	T1	T2	T3
	read item(X)		
	read item(Y)		
	write item(Y)		
	write item(X)		
		read item(Z)	
		read item(W)	
		write item(Z)	
			write item(Q)
			read item(Y)
read item(W)			
write item(W)			
read item(Q)			

Question 16 - Query Processing (7 points)

Consider a corporate database with two relations: EMPLOYEE (with 10,000 tuples of one block each) and a cardreader ACCESS log (10,000,000 tuples of one block each). Their shared attributed is the employee ID (EID).

For each of the 3 sets of conditions, decide whether **merge join, hash join, index join** or **simple iteration** is more efficient, and briefly say *WHY*. Consider the query in isolation, and any data structures you create for this join will be destroyed before any other queries.

(a) (4 points) Both relations are already sorted by EID, memory is very limited, and no indexes currently exist.

Answer:
In this case ,the merge join is more efficient.
Because the merge join requires sorted data, which is given. And it doesn't need many memory or index. Besides, it's very efficient

(b) (3 points) Neither relation is sorted, no indexes currently exist, but there is a huge amount of fast memory available.

Answer: In this case, I'd prefer the hash join.
Since the relations is neither sorted nor indexed, it could take a lot of time to sort of index it since the date may not be clustering. The hash join does not require index, and sorted , and we have sufficient memory.

Question 19 - Security (4 points)

What are the advantages and disadvantages of using views to restrict access to information in databases?

Advantages:
(1)
A view support multiple user perspectives on the databse corresponding to different information organizations, avoiding the need for **data duplication** or information **consistency problems**.
(2)
Security. It helps with the privacy concerns. Users need only access/modify selected attributes in the data without being able to access the other attributes.

Disadvantages:
    (1) It will be complicate and takes additional resources and reduce the efficiency. It may be inefficient if the view involves complex calculation like aggregate function
    (2) The temporary table must be updates(recomputed) if updating happens. It is costly and make minimal update difficult to determine. It also fail to avoid relational count of eliminating data duplication.

Question 20 - Security (5 points)

Briefly list and describe some problems with using encryption for database security (in particular, how does this negatively impact query processing).

(1)it's hard for sorting and indexing the data since all the data is encrypted and sorting/indexing needs the original data.

Question 21a - Recovery (4 points) Explain how the buffer manager may cause the database to become inconsistent if some log records pertaining to a block are not output to stable storage before the block is output to disk.

As we know, the log in buffer must be output to stable storage before the data is output to the stable storage. If the buffer management has some problem and some of the log are not able to get into the stable storage. Then case1: the system detects the missing of the log and prevents the output of the data, so all the system is stuck. Case 2:The system does not detect the missing of the log, then the log is inconsistent with the data because only part of the log are output. When are looking back at the log, there is a inconsistent problem.

Question 21b - Recovery (4 points) Describe at least 3 key strategies employed to create stable storage for recovery logs.

? ? ? ?

Question 22 - Distributed Databases (12 points)

A stock brokerage database is distributed between 10 local brokerage offices (where BrokerageId=B1...B10) and a central Wall Street site (W).

Given the relations:
      Stock(symbol, name, price)
      Owns(accountID, symbol, numShares, brokerageID)

Assume the Owns relation is fragmented horizontally by brokerageID and each fragment is stored locally at the brokerage. Assume the Stock relation is stored in its entirety at the Wall Street site. Describe a good strategy for processing each of the following. Be as formal as possible in your answer (e.g. *SendQuery(From=*B1,*To=*W,*Query1=*"Select ..);., including full RA/SQL expressions for Query$_i$ =... and a precise statement about exactly what resulting relations are shipped from where to where.

Both queries below are issued from Brokerage B9.

(a) (6 points) What is the total price of shares held by accounts at the broker whose brokerageID is B1?
Answer:
      (1)Senddata(From= WallStreet, to=B1, Data=Stock)
      (2)Result = DoQuery(At = B1, Query = 'SELECT numShares * price FROM Stock as S, Owns as O WHERE O.brokerageID = 'B1' AND S.symbol = O.symbol')
    (3) Returndata(FROM=B1, TO=WallStreet, Data=result)

(b) (6 points) Which brokerage (B1..B10) owns the most of the stock whose price is highest in the database? (you can assume that there is only 1 highest-priced stock).

    (1) At WallStreet, Do Query (HIghestSymbol' = 'SELECT * FROM (SELECT symbol FROM Stock ORDER BY price DESC) WHERE rownum = 1')

(2) Create table Shares = SendQuery(FROM = WALL STREET, TO B1~B10, query = 'SELECT brokeragelID, SUM(numShares) as numShares FROM Owns WHERE symbol = HIghestSymbol'

(3) SELECT brokeragelID FROM Shares WHERE numShares = (SELECT MAX(numShares) FROM Shares);

Question 24 - Stored Procedures (5 points)

Create a table that has 1 number as a field. Write a stored procedure to populate that table with the first 100 fibonacci numbers. Example: 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, ...

$F(n) =$ 0, if n=0;
1, if n=1;
F(n-1) + F(n-2), if n>1

Answer:
delimiter//
Drop procedure if fibonacci exists;
Create procedure fibonacci()
Begin
        Create table Fib;
        Declare a int, b int, c int, i int;
        Set a = 0
        Set b = 1;
        Set c = a + b;
        Set i = 0;
        While i < 100 Do
                Insert into Fib values (a);
                Set a = b;
                Set b = a + b;

```
                    Set i = i + 1
        End While;
End
```