CSE 344 Final Examination

March 18, 2014, 2:30pm-4:20pm

Name: _____

Question	Points	Score
1	20	
2	20	
3	20	
4	20	
5	20	
Total:	100	

- This exam is open book and open notes but NO laptops or other portable devices.
- You have 1h:50 minutes; budget time carefully.
- Please read all questions carefully before answering them.
- Some questions are easier, others harder; if a question sounds hard, skip it and return later.
- Even if you cannot fully answer a question, write partial solution for partial credit.
- Good luck!

Reference for SQL Syntax

Common aggregation functions in SQL

```
-- Relation R(A, B)

COUNT(*)
COUNT(A)
COUNT(DISTINCT A)
SUM(A)
AVG(A) == SUM(A)/COUNT(A)
MAX(A)
MIN(A)
```

Common keywords for sub-queries in SQL

```
IN
EXISTS
ANY
ALL
EXCEPT
(NOT) IN/EXISTS/...
```

The WITH Statement

```
WITH T AS (SELECT * FROM R WHERE R.K > 10),
     S AS (SELECT * FROM R WHERE R.a > 50)
SELECT * FROM T, S WHERE T.K<20 AND S.a < 20
```

Reference for the Relational Algebra

Name	Symbol
Selection	σ
Projection	Π
Join	$\bowtie$
Group By	γ
Set difference	$-$

SQL and Relational Languages

1. (20 points)

   Suppose a company stores information about its employees and their hobbies using the following schema:

   - `Emp(eid, name, city)`
   - `Hobbies(eid, hobby)`; `Hobbies.eid` references to `Emp.eid`

   (a) (5 points) Write an SQL query to output the `name` of the employees who live in `city = 'Seattle'` and have at least five (≥ 5) hobbies.

- `Emp(eid, name, city)`
- `Hobbies(eid, hobby)`

(b) The following Relational Calculus (RC) expression finds all cities such that any employee living in such a city has no other hobby than White Water Rafting (`hobby = 'WWR'`) or Philately (`hobby = 'PT'`).

$$Ans(c) = \forall e\ \forall n\ \forall h\ ((Emp(e,n,c)\ \wedge\ Hobbies(e,h))\ \Rightarrow\ ((h =' WWR')\ \vee\ (h =' PT')))$$

Convert this RC expression into equivalent non-recursive Datalog with negation, Relational Algebra, and SQL.

i. (5 points) Write an equivalent **non-recursive Datalog program with negation**.

- `Emp(eid, name, city)`
- `Hobbies(eid, hobby)`

ii. (4 points) Write an equivalent **Relational Algebra (RA) expression or logical query plan**.

$$Ans(c) = \forall e\ \forall n\ \forall h\ ((Emp(e,n,c)\ \wedge\ Hobbies(e,h))\ \Rightarrow\ ((h =' WWR')\ \vee\ (h =' PT')))$$

(**This page is intentionally left blank**)

- `Emp(eid, name, city)`
- `Hobbies(eid, hobby)`

iii. (4 points) Write an equivalent **SQL query**.

$$Ans(c) = \forall e\ \forall n\ \forall h\ ((Emp(e,n,c)\ \wedge\ Hobbies(e,h))\ \Rightarrow\ ((h =' WWR')\ \vee\ (h =' PT')))$$

(c) (2 points) Write a **Relational Calculus (RC)** expression to output the names of employees who are not interested in skiing (i.e. do not have any `hobby = 'ski'`).

Database Design, Views

2. (20 points)

(a) (**Design theory/normalization**)

i. (3 points) Consider Relation $R(ABCD)$.
and functional dependencies (FDs): $BD \rightarrow AC$; $AB \rightarrow D$; $AC \rightarrow B$

- Is this relation in Boyce-Codd Normal Form (BCNF)? Write YES/NO.

i. _____

.

- Identify a key (not a superkey).

i. _____

ii. (3 points) Consider Relation $R(ABCDE)$.
and functional dependencies (FDs): $A \rightarrow C$; $B \rightarrow AE$; $E \rightarrow D$.

- Is this relation in Boyce-Codd Normal Form (BCNF)? Write YES/NO.

ii. _____

.

- Identify a key (not a superkey).

ii. _____

iii. (3 points) If any of the above relations is not in BCNF, decompose it into BCNF. Also underline the keys in the final relations after decomposition. Otherwise write "BOTH IN BCNF".

(b) (**Views**) Given three relations $R(A, B)$, $S(B, C)$, $T(C, D)$, someone has created two views $V1, V2$ as follows:

```
CREATE VIEW V1 AS
  SELECT A, C, sum(B) as bs, count(B) as bc
  FROM R, S
  WHERE R.B = S.B
  GROUP BY A, C
```

```
CREATE VIEW V2 AS
  SELECT DISTINCT B, S.C, D
  FROM S, T
  WHERE S.C = T.C
```

For each of the following SQL queries, write equivalent SQL queries that will only access the views $V1, V2$. without using the original relations R, S, T.

---

i. (3 points) SQL Query:

```
SELECT DISTINCT R.A, T.D
FROM R, S, T
WHERE R.B = S.B AND S.C = T.C
```

---

ii. (3 points) SQL Query:

```
SELECT A, avg(B) as myavg
FROM R, S
WHERE R.B = S.B
GROUP BY A
```

---

iii. (1 point) Your answer in the above question (part b.ii) will not hold if we omit the relation S and the join condition "`WHERE R.B = S.B`". Explain why.

```
(i.e. for the query )
  SELECT A, avg(B) as myavg
  FROM R
  GROUP BY A
```

(c) i. (1 point) Is this statement correct?

"Materialized views are precomputed offline, and therefore are fast at runtime."

i. _____

Write TRUE/FALSE.

ii. (1 point) Is this statement correct?

"Queries that use Virtual views get data that are not always up-to-date."

ii. _____

Write TRUE/FALSE.

iii. (2 points) Insert attribute-level and/or tuple-level constraint(s) to the following `CREATE TABLE` statement to ensure that

1. the value of attribute B is always > 10 and < 20.
2. the value of attribute A is not null.

```
CREATE TABLE R (
  A INT,
  B INT

)
```

Transactions

3. (20 points)

   (a) Consider the following schedule.

   ```
   r1(X), r2(X), r3(X), r1(Y), w2(Z), r3(Y), w3(Z), w3(Y)
   ```

   Recall that a schedule $S1$ is conflict-equivalent to another schedule $S2$ if $S2$ can be obtained from $S1$ by swapping adjacent non-conflicting actions.

   i. (1 point) Is this schedule serial?

   i. _____

   Write YES/NO.

   ii. (2 points) Draw the precedence graph.

   iii. (1 point) Is this schedule conflict-equivalent to (T_2, T_3, T_1)?

   iii. _____

   Write YES/NO.

   iv. (1 point) Is this schedule conflict-equivalent to (T_2, T_1, T_3)?

   iv. _____

   Write YES/NO.

(b) Consider the following isolation levels in SQL Server:
`READ UNCOMMITTED, READ COMMITTED, REPEATABLE READ`.

For each of these isolation levels, complete the tables with three transactions T_1, T_2, T_3 as follows:

1. Insert lock/unlock on the element B. Use
   - $L_i(B), U_i(B)$ for long-term read or write locks by T_i in strict 2PL, and
   - $SL_i(B)$, $SU_i(B)$ for short-term read locks by T_i.
   - Also assume that short-term read locks $SL_i(B)$ can be requested by T_i and granted while another transaction T_j is holding the long-term lock $L_j(B)$.

   (note that this is a simplified locking scenario, e.g., SQL Server uses more complex locking scheme in practice.)
2. You also have extra empty rows in the tables to write any action that will be deferred by the system. Write "$L_i(B)$ `REQUESTED`" and "$L_i(B)$ `GRANTED`" in appropriate cells for any deferred lock requests, and strike out the original action that is deferred.
3. Mention the new value of B on disk if it changes at any point.
4. Mention the value of B read by the transactions in the read statements $r_i(B)$.

Note: $w_i(B, 300)$ means write $B = 300$. A_i means abort/rollback T_i, and C_i means commit T_i.

For example, the cells may look like this (this is not a real answer):

1	**$L_1(B)$** $r_1(B) = 50$			
2		**$L_2(B)$** `REQUESTED` ~~$w_2(B)$~~		
3	$w_1(B, 80)$			
4	**$U_1(B)$** A_1			

i. (3 points) **READ UNCOMMITTED**

Time	T_1	T_2	T_3	B **on disk**
0				20
1	r_1(B)			
2		w_2(B, 30)		
3	r_1(B)			
4		A_2		
5			w_3(B, 40)	
6			C_3	
7	r_1(B)			
8	C_1			
9				
10				
11				
12				

ii. (3 points) **READ COMMITTED**

Time	T_1	T_2	T_3	B **on disk**
0				20
1	r_1(B)			
2		w_2(B, 30)		
3	r_1(B)			
4		A_2		
5			w_3(B, 40)	
6			C_3	
7	r_1(B)			
8	C_1			
9				
10				
11				
12				

iii. (3 points) **REPEATABLE READ**

Time	T_1	T_2	T_3	B **on disk**
0				20
1	r_1(B)			
2		w_2(B, 30)		
3	r_1(B)			
4		A_2		
5			w_3(B, 40)	
6			C_3	
7	r_1(B)			
8	C_1			
9				
10				
11				
12				

(c) i. (1 point) What does the isolation level `SERALIZABLE` achieve that is not guaranteed in `REPEATABLE READ`?

ii. (1 point) In SQLite, can a READ lock co-exist with a PENDING lock? Write YES/NO.

ii. _____

iii. (1 point) In SQLite, can a READ lock co-exist with an EXCLUSIVE lock? Write YES/NO.

iii. _____

(d) (3 points) Can the following schedule be produced by a scheduler following two-phase locking protocol (2PL)? Explain briefly why or why not.

```
r1(X), w1(X), r2(X), w2(X), r3(Y), w3(Y), r1(Y), w1(Y)
```

Parallel Databases and MapReduce

4. (20 points)

Consider two relations `R(a,b)` and `S(b,c)`. Both are horizontally partitioned across $N = 3$ nodes as shown in the diagram below. Each node locally stores approximately $\frac{1}{3}$ of the tuples in `R` and $\frac{1}{3}$ of the tuples in `S`.

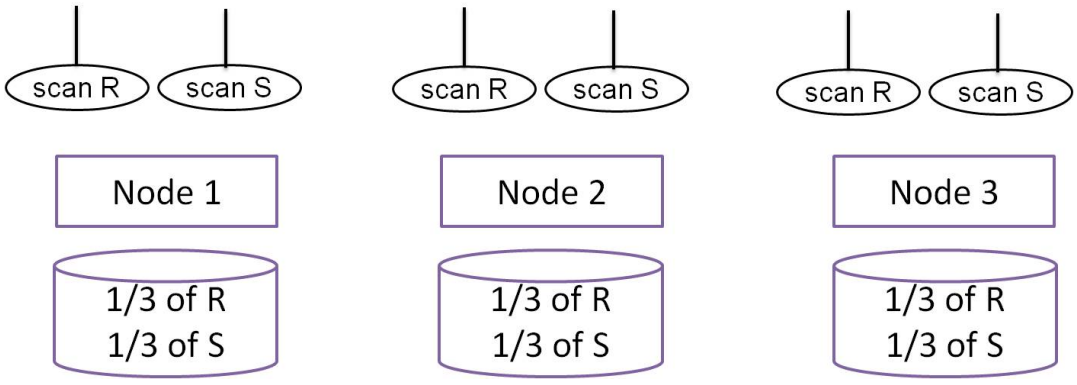
- R is **block-partitioned**.
- S is **hash-partitioned on S.b**.

(a) **Parallel-DB**

i. (8 points) Draw a parallel relational algebra plan for the following SQL query and show how it will be executed across $N = 3$ machines.
You will get more credit for picking an efficient plan that leverages the parallelism as much as possible.
If you need to re-shuffle the data, include required operators and edges to re-shuffle the data across machines.

SQL Query:

```
SELECT R.a, avg(S.c) as myavg
FROM R, S
WHERE R.b = S.b
AND R.a <= 100 and S.c > 200
GROUP BY R.a
```



(**This page is intentionally left blank**)

- R is **block-partitioned**.
- S is **hash-partitioned on S.b**.

```
SELECT R.a, avg(S.c) as myavg
FROM R, S
WHERE R.b = S.b
AND R.a <= 100 and S.c > 200
GROUP BY R.a
```

Answer:

ii. (1 point) Which steps can be removed/have to be added if R is hash-partitioned on R.b?

iii. (1 point) Which steps can be removed/have to be added if S is range-partitioned on S.c?

(b) (2 points) Write one advantage and one disadvantage of map-reduce compared to parallel databases.

(c) Consider two relations $R(a, b)$ and $S(b, c)$. Answer the following questions if the following SQL query is executed by a single **MapReduce job** (**not Pig**). You can describe your answers in English, no pseudocode is necessary.

```
SELECT R.b, max(S.c) as cmax
FROM R, S
WHERE R.b = S.b
AND R.a <= 100
GROUP BY R.b
```

i. (4 points) For the **Map function**, what are the computations performed, and what will be its outputs? Assume that the Map function reads a block of R or S relation as input.

ii. (4 points) For the **Reduce function**, what will be its inputs, what are the computations performed, and what will be its outputs?

Miscellaneous

5. (20 points)

(a) (1 point) Recall the 3-valued logic for evaluating NULL in SQL.
Suppose $A = NULL$, $B = 5$, $C = 10$.
What will be the value of

$$((A < 7) \quad OR \quad (C > 2)) \quad AND \quad (B > 10)$$

(a) _____

Write TRUE/FALSE/UNKNOWN

(b) Consider relations $R(A, B)$ and $S(C, D)$. There are two types of queries in the workload.

```
Type1:
     SELECT A, D
     FROM R, S
     WHERE R.B = S.C
```

```
Type2:
    SELECT *
    FROM R
    WHERE R.A < ? AND R.B > ?
```

i. (1 point) Which index is more useful?
(a) $Index{-}R(A, B)$, or (b) $Index{-}R(B, A)$

i. _____

Write (a) or (b) or "both are same"

ii. (1 point) Which queries can be answered by the covering index $Index{-}R(A, B)$

ii. _____

Write Type1 or Type2 or "None"

(c) Consider the relations

1. `Buyer(bid, bname, city)`
2. `Buys(bid, pname, price)`; bid references `Buyer`

The following SQL query outputs the names of the buyers who did not buy any product with price < 10.

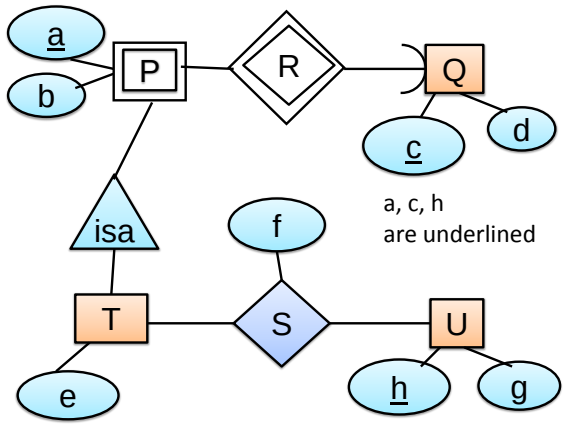
```
SELECT bname
FROM Buyer B1
WHERE 10 <= ALL(SELECT price
                FROM Buys B2
                WHERE B1.bid = B2.bid)
```

i. (2 points) Write an equivalent SQL query that uses `NOT IN` for the subquery. You need to complete the following query.

```
SELECT bname
FROM Buyer B1
WHERE B1.bid NOT IN (...
```

ii. (2 points) Write an equivalent SQL query that does not use a subquery.

(d) (5 points) Consider the following E/R diagram.



Write the schema of the relations corresponding to all the entities and relationships in this diagram. You need to underline the keys of each relation, and mention if there is a foreign key referencing to another relation.

e.g. your answer should look like (this is not a real answer)

1. M1(m, n, o, p), where p references to attribute p of relation $M2$.
2. ....

(e) i. (1 point) Write one difference between Relational data model and semi-structured data model (XML).

ii. (1 point) Is this statement correct?
"Elements in XML are unordered".

ii. _____

Write CORRECT/INCORRECT.

iii. (1 point) Is this statement correct?
"Attributes in XML are unordered".

iii. _____

Write CORRECT/INCORRECT.

iv. (1 point) Is this statement correct?
"Conversion of XML to Relational databases is easier than conversion of Relational databases to XML".

iv. _____

Write CORRECT/INCORRECT.

(f) (2 points) The following relation captures the XOR function X of two bits A, B. Identify all non-trivial functional dependencies in this relation.

A	B	X
0	0	0
0	1	1
1	0	1
1	1	0

(g) (1 point) Suppose you are handling a data integration problem where you do not care much whether you have the most up-to-date data but the queries should run as fast as possible. Which approach would you prefer:

(a) Data Warehouse, (b) Mediator?

(g) _____

Write (a) or (b) or "none"

(h) (1 point) Which of the following properties is not a key feature of a NoSQL system?

(a) Flexible schema, (b) Simpler interface than SQL, (c) Strict ACID concurrency model

(h) _____

Write (a) or (b) or (c) or "none"