Name	NetID

CS411: Database Systems **Spring 2018**

Midterm 2, May 2

- READ THESE INSTRUCTIONS CAREFULLY BEFORE YOU START. DO NOT turn this page UNTIL the proctor instructs you to.
- Check that your exam booklet consists of 18 printed sides, including this one.
- Write your NetID at the top of all the sheets.
- The exam lasts for 90 minutes, i.e., from 8–9.30am.
- We will not answer any questions during the exam. For questions 2–5, if you need to make any assumptions for any of the questions, please feel free to do so and clarify the assumption in your answer.
- All questions are compulsory.
- If you need more space, feel free to use the back side of each page. We will not provide extra sheets, so please use your space wisely. For questions 2–5, show all necessary steps as part of your calculation to get partial credit.
- The maximum score you can obtain is $40 + 16 + 12 + 22 + 10 = 100$.
- You must stop writing when time is called by the proctors.
- You should submit your cheat sheet at the end of the exam.
- **Cheating: No.**

Question	1	2	3	4	5	Total
Points						

1 Objective Questions - 40 points

Please **circle the right answer(s)**. **No explanation is necessary, and will not be graded.** Please read instructions carefully—some questions have only one right answer, some have multiple right answers. Please use the space provided below the question, along with the back (blank) sides of the exam booklet, as scratch space, and only circle the final answers here.

1. [6] Consider a relation R(A,B,C,D) and the following functional dependencies over R:

   ```
   A -> B
   D -> B
   C -> A
   ```

   Which of the following decompositions is in BCNF? (**Circle all answers that apply.**)

   (a) AB, AC, CD
   (b) AC, BC, CD
   (c) AB, ACD
   (d) AC, AD, BD

   [**Answer**]: a, b

2. [3] Consider a relation R(A,B,C,D,E,F) with the set of functional dependencies (FDs):

   ```
   A -> C
   DE -> F
   B -> D
   ```

   Based on these FDs, there is one minimal key for R. What is it? (**Circle one answer.**)

   (a) ABC
   (b) AB
   (c) ADE
   (d) ABE

   [**Answer**]: (D). ABE

3. [4] Consider a schema with two relations, R(A,B) and S(B,C), where all values are integers. Make no assumptions about keys. Which of the following relational algebra expressions are equivalent (i.e., produce the same answer on all databases) to

$$\pi_{A,C}(R \bowtie \sigma_{B=1}(S))?$$

(**Circle all answers that apply.**)

(a) $\pi_A(\sigma_{B=1}(R)) \times \pi_C(\sigma_{B=1}(S))$
(b) $\pi_{A,C}(\pi_A(R) \times \sigma_{B=1}(S))$
(c) $\pi_{A,C}(\sigma_{B=1}(R) \bowtie S)$

[**Answer**]: a,c

4. [3] Consider the following three relations A(x,y), B(y,z), and C(z,x)

A(x,y)	B(y,z)	C(z,x)
T(A) = 2500	T(B) = 1000	T(C) = 6000
V(A, x) = 50	V(B, y) = 250	V(C, z) = 40
V(A, y) = 200	V(B, z) = 50	V(C, x) = 60

Estimate the size (measured in number of tuples) of

```
SELECT x, y, z
FROM A, B, C
WHERE A.y = B.y and B.z = C.z
```

(**Circle one answer**.)

(a) $(2500 \times 1000 \times 6000)/(50 \times 200 \times 250 \times 50 \times 40 \times 60)$
(b) $(2500 \times 1000 \times 6000)/(200 \times 250 \times 60)$
(c) $(2500 \times 1000 \times 6000)/(250 \times 50)$
(d) $(2500 \times 1000 \times 6000)/(250 \times 50 \times 60)$

[**Answer**]: (c)

5. [3] Use the same table in the previous question. Estimate the size (measured in number of tuples) of ($\sigma_{x\,=\,10}$ A) $\bowtie$ B. (**Circle one answer.**)

   (a) $(2500 \times 1000)/(50 \times 200)$
   (b) $(2500 \times 1000)/(50 \times 250)$
   (c) $(2500 \times 1000)/250$
   (d) $(2500 \times 1000)/(250 \times 200)$

   [**Answer**]: (b)

6. [4] Consider the following definitions to a view applied to relations R(A, B, C) and S(A, D). In which cases would it be **problematic** to let the view be update-able by a user? That is, it is not straightforward for the database system to translate user-driven changes to the view to the underlying data. (**Circle all answers that apply.**)

   (a) `SELECT A, MAX (B) FROM R GROUP BY A`
   (b) `SELECT A, B, C FROM R WHERE A > 20`
   (c) `SELECT A, B FROM R WHERE C = 10`
   (d) `(SELECT A FROM R) UNION (SELECT A FROM S)`

   [**Answer**]: a, d

7. [4] The memory constraint for a two-pass merge sort based duplicate elimination is the following: $B(R) \leq M(M-1)$. What factors contribute to this constraint? (**Circle all answers that apply.**)

   (a) We need an output buffer in the second pass
   (b) We need an output buffer in the first pass
   (c) There are $B(R)/M$ sorted runs
   (d) We need to have a representative from all of the sorted runs in the first pass
   (e) We need to have a representative from all of the sorted runs in the second pass

[Answer]: a, c, e

8. [4] The memory constraint for a two-pass hashing-based join is the following:

$$\min(B(R), B(S)) \leq (M-1)(M-2).$$

What factors contribute to this constraint? (**Circle all answers that apply.**)

(a) We need an input buffer for R (or S) in the first pass
(b) We need an output buffer for the join of R and S in the second pass
(c) We need one buffer for each of the buckets in the hash table in the first pass.
(d) We need one buffer for each of the buckets in the hash table in the second pass.
(e) The smallest of the buckets from R or S must fit in memory in the second pass.

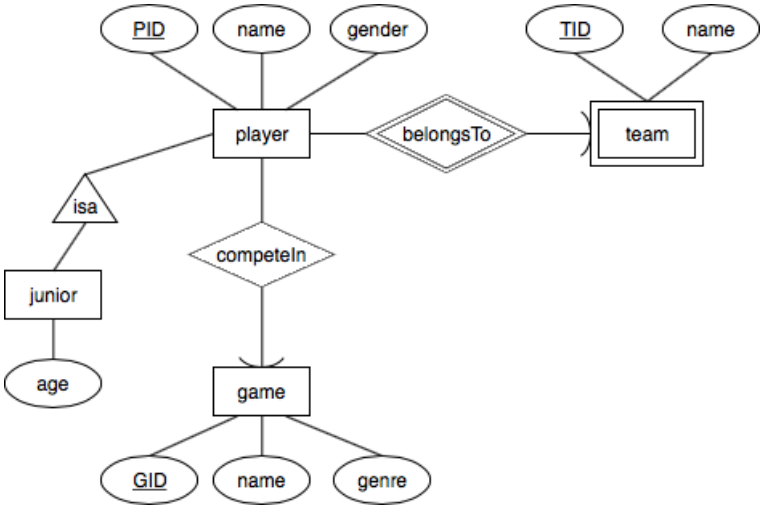
[Answer]: a, b, c, e

9. [3] Consider the following rules, where "$\Longrightarrow$" indicates that the FD on the right hand side can be derived from the ones on the left hand side.
Which rules are **incorrect**? (**Circle all that apply.**)

(a) $A \to B_1B_2B_3, B_2 \to C \Longrightarrow A \to C$
(b) $A \to C$, $B \to C$, $ABC \to D \Longrightarrow A \to D$
(c) $AB \to C \Longrightarrow A \to C$

[Answer]: B and C are incorrect.

10. [6] Consider the following ER diagram for a gaming competition:



Please **circle all** of the expressions that would correctly translate the above ER diagram to a relational model. A correct translation would have keys underlined, and relations merged, as far as possible, to minimize redundancy.

(a) Player(PID, name, gender)
(b) Player(PID, TID, name, gender)
(c) Player(PID, TID, GID, name, gender)
(d) Player(PID, TID, GID, name, gender)
(e) Game(GID, name, genre)
(f) Game(GID, PID, name, genre)
(g) Game(GID, PID, name, genre)
(h) Junior(age)
(i) Junior(PID, age)
(j) Junior(PID, TID, age)
(k) Junior(PID, TID, age)
(l) Team(TID, name)
(m) Team(TID, PID, name)
(n) Team(TID, PID, name)
(o) belongsTo(PID, TID)
(p) belongsTo(PID, TID)
(q) competeIn(PID, GID)
(r) competeIn(PID, GID)

[**Answer**]: The correct answers are: c, e, k, l. However, there's a mistake in the drawing. I intended to have Player as the weak entity set but marked team. Therefore, we will take c or d for Player, e for Game, k or l or j for Junior, and l,m,n for Team.

2 SQL and Constraints/Triggers - 16 points

1. [4] Consider a SQL table T(A int). Assume there are no null values. Make no assumptions about keys. We execute the following SQL query:

```
SELECT SUM(B)
FROM T, (SELECT DISTINCT A, 1 AS B FROM T) AS T2
WHERE  T.A = T2.A AND T2.A = 3
```

   State **in a few words** what this query computes.

   [**Answer**]: Calculate the number of tuples in T where attribute A = 3.

2. [6] Consider a relational database of airline passenger and reservation information:

```
Passenger(passID, name)
Resvn(passID, flight_no, date)
```

   Write a SQL query to find the names of all passengers who have reservations on exactly two different dates.

```
SELECT name FROM Passenger
WHERE passID IN (
    SELECT passID FROM Resvn
    GROUP BY passID
    HAVING COUNT(DISTINCT(date)) = 2;
)
```

3. [6] Consider a relation Employee(ID,salary), where ID is the key. Suppose the following trigger is defined on this relation:

```
CREATE TRIGGER AutoRaise
AFTER INSERT ON Employee
UPDATE Employee
SET salary = salary + (SELECT MIN(salary) FROM Employee)
```

Suppose that the relation contains one tuple: (1, 50).

Now, suppose two new tuples are inserted as the result of *one statement*: (2, 30), (3, 70).

(a) [3] Assume the trigger above is specified as "FOR EACH STATEMENT". After the insertions and trigger execution(s), what are the final salaries for the three employees? **Only list the resulting tuples.**

[Answer]: (1, 80), (2, 60), (3, 100)

(b) [3] Assume the trigger above is specified as "FOR EACH ROW". After the insertions and trigger execution(s), what are the final salaries for the three employees? **Only list the resulting tuples.**

[**Answer**]: (1, 110), (2, 90), (3, 130)

3 Query Execution and Optimization - 12 points

Consider a relational schema with three tables with **no** index:

```
Took (studID, courseNum, year, grade)
Student (ID, name, address, birthdate)
Course (num, dept)
```

Let us make the following assumptions:

- Let $B(T)$, $B(C)$, and $B(S)$ denote the number of blocks of `Took`, `Course`, and `Student` respectively. Assume that $B(T) = 60K$, $B(C) = 2K$, and $B(S) = 5K$.
- Assume $B(\sigma_c R) = B(R) \times 0.1$, where R is any relation and c is any condition. That is, the number of blocks taken by $(\sigma_c R)$ is 0.1 times that of the original relation R.
- Assume $B(R_1 \bowtie R_2) = B(R_1) \times B(R_2) \times 0.01$, where R_1 and R_2 are any two relations. That is, the number of blocks taken by $R_1 \bowtie R_2$ is 0.01 times $B(R_1) \times B(R_2)$.

Consider the query that finds the student's information for all students who took a class in the EE department and got an A.

```
SELECT name, address, birthdate
FROM Student AS s, Took AS t, Course AS c
WHERE t.studID = s.ID and t.courseNum = c.num
and c.dept = 'EE' and t.grade = 'A'
```

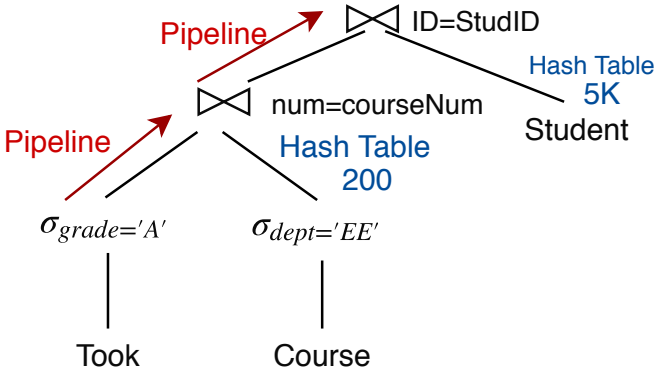
Under each of the following memory constraints, draw the optimal plan tree (**ignore the projection operators**) with **specified physical plan**, which has the most efficient execution in terms of I/O cost. Physical plan refers to the choice of join operators, the placement of selection conditions, and pipelining/materialization, taking memory consideration into account. You can ignore the projection operations.

1. [6] If the number of available memory blocks is $M = 6K$:

   [**Answer**]:

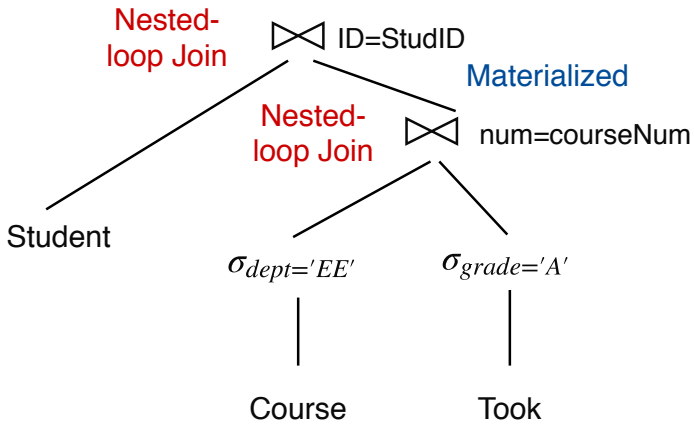
   No materialization is needed.

   Remark: The size of Course after selection is 200, while the size of Student is 5K. Since 200 + 5K < 6K, we can hold them simultaneously in memory. Please refer to slide 20 in part 2 of Query Optimization from lecture note (pdf).



2. [6] If the number of available memory blocks is $M = 10$:

   [**Answer**]:

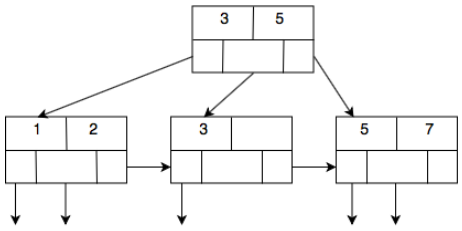


   Since there is only 10 blocks available in memory, we can only use nested-loop join.

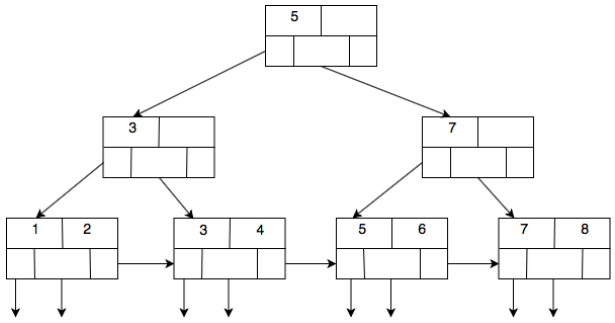
   Remark: Student has no common attribute with Course, thus if we first join Student with Course after selection, the join size is 200*5K = 1000K. While if we join Course after selection and Took after selection, the join size if 0.01*200*6k=12K, which is much smaller than 1000K. Thus, we should first join Course after selection and Took after selection.

4 B-Trees and Indexing - 22 points

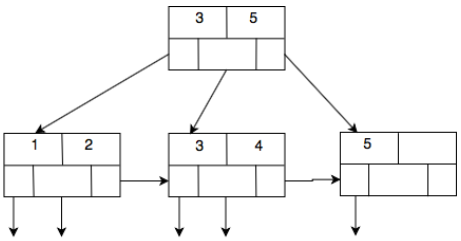
1. [12] Consider a B+tree of degree 1 as below:



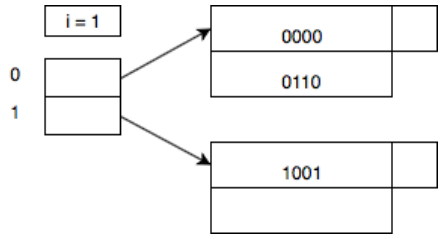
   (a) [6] Insert 4,8,6 incrementally, and draw the final tree. "Incrementally" means that you should insert the values one at a time, check the condition(s), and move to the next one. Make sure that all the values in the intermediate nodes are also carefully updated according to the values in nodes below—that is, do not be lazy about updating those values.



(b) [6] Using your solution from part (a), now delete 8,7,6 incrementally, and draw the final tree. "Incrementally" means that you should insert the values one at a time, check the condition, and move to the next one. Make sure that all the values in the intermediate nodes are also carefully updated according to the values in nodes below—that is, do not be lazy about updating those values.

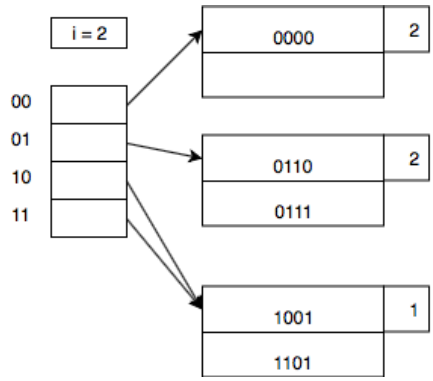


2. [10] Consider a hash table with each data block capable of holding 2 records, drawn as follows:

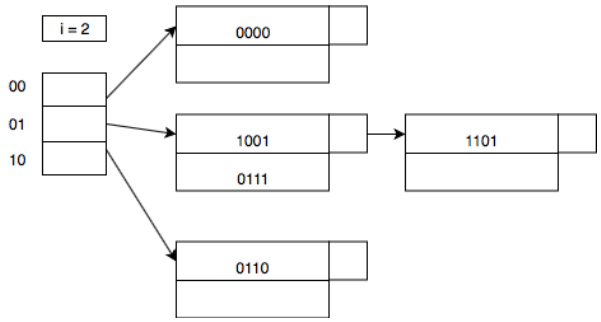


Now, insert 0111 and 1101 incrementally to the table. "Incrementally" means that you should insert the values one at a time, check the condition, and move to the next one.

(a) [5] Draw the resulting final hash table if it is a extensible hash table. Please remember to indicate i and the bit counter for each block accordingly.



(b) [5] Draw the resulting final hash table if it is a linear hash table with $r \leq 1.7 \times n$ (that is, the average number of records per bucket should not exceed 85% of the total number of records per block). Please remember to indicate i.



5 Transaction Management - 10 points

Consider the following **UNDO** log, and use it to answer the following questions.

LogID	Log
1	<START T1>
2	<T1, A, 20>
3	<START T2>
4	<START T3>
5	<T2, C, 7>
6	<T1, B, 15>
7	<T2, D, 9>
8	<COMMIT T1>
9	<START T4>
10	<T3, E, 2>
11	<T4, A, 6>
12	<ABORT T2>
13	<T3, B, 9>
14	<COMMIT T3>
15	<T4, B, 21>
16	<START T5>
17	<START T6>
18	<T6, D, 8>
19	<COMMIT T6>
20	<T5, E, 6>

1. [4] Given this log, what can you infer about how T1 changes variable A? (1–2 lines are sufficient—just mention the initial value and the final value and how you inferred it.).

   [**Answer**]: A = 6 after T1

2. [6] Suppose the system has crashed. Given this log, list which actions (e.g.: <T1, A, 15>) will have to be undone to restore the database to the correct state and in what order. **[Answer]**: logIDs 20, 15, 11, 7, 5