3 NoSQL, JSON, SQL++

3. (20 points)

   (a) (10 points) We are given a JSON file with noble prize laureates, with the following structure:

```
{"prizes": [
    { "year": "2018",
      "category": "physics",
      "overallMotivation": "\For groundbreaking inventions in the field of laser physics"",
      "laureates": [
        { "id": "960",
          "name": "Arthur Ashkin",
          "motivation": "\"for the optical tweezers and their application to biological systems\"",
          "share": "2"
        },
        { "id": "961",
          "name": "Gérard Mourou",
          "motivation": "\"for their method of generating high-intensity, ultra-short optical pulses\"",
          "share": "4"
        },
        { "id": "962",
          "name": "Donna Strickland",
          "motivation": "\"for their method of generating high-intensity, ultra-short optical pulses\"",
          "share": "4"
        }
      ]
    },
    { "year": "2018",
      "category": "chemistry",
      ...
    },
    { "year": "2018",
      "category": "medicine",
      ...
    }
  ]
```

   Write a SQL++ query that returns each noble prize laureate who has received more than one award, along with a list of the years and categories that each such laureate has received. Your query should return a JSON file with a structure like the following:

```
{ "name": "Frederick Sanger",
  "awards": [ { "year": "1958", "category": "chemistry" },
              { "year": "1980", "category": "chemistry" } ]
}
{ "name": "Marie Curie, née Sklodowska",
  "awards": [ { "year": "1903", "category": "physics" },
              { "year": "1911", "category": "chemistry" } ]
}
...
```

[this page is intentionally left blank]

(b) For each statement below indicate if it is true or false.

i. (2 points) JSON is in First Normal Form.

i. ____________

True or false?

ii. (2 points) JSON represents semistructured data.

ii. ____________

True or false?

iii. (2 points) It is easy to represent a many-to-one relationship in JSON, but it is difficult to represent a many-to-many relationship.

iii. ____________

True or false?

iv. (2 points) SQL++ can express all queries expressible in datalog.

iv. ____________

True or false?

v. (2 points) ACID transactions today are increasingly using JSON data as opposed to relational data.

v. ____________

True or false?

`Sim(version, year, loc, temp)`

4 Query Execution and Optimization

4. (30 points)

   (a) (10 points) Write a logical plan for the following query.

```
select x.year, x.loc, x.temp
from Sim x
where x.version = 22
 and x.temp > (select avg(y.temp)
               from sim y
               where y.version = 23
                 and x.loc = y.loc);
```

You should turn in a relational algebra tree.

(b) In this question we consider three relations $R(A, B), S(B, C), T(C, D)$ and the following statistics:

$$\begin{aligned}
T(R) &= 10^5 & B(R) &= 100\\
T(S) &= 6 \cdot 10^6 & B(S) &= 3000\\
T(T) &= 5 \cdot 10^4 & B(T) &= 40000\\
V(R, A) &= 5 \cdot 10^4 & &\\
V(R, B) &= V(S, B) = 3 \cdot 10^3 & &\\
V(S, C) &= V(T, C) = 2 \cdot 10^4 & &\\
V(T, D) &= 10^4 & &
\end{aligned}$$

i. (5 points) Estimate the number of tuples returned by $\sigma_{A=2432}(R)$. You should turn in an integer number.

ii. (5 points) Estimate number of tuples returned by the following query:

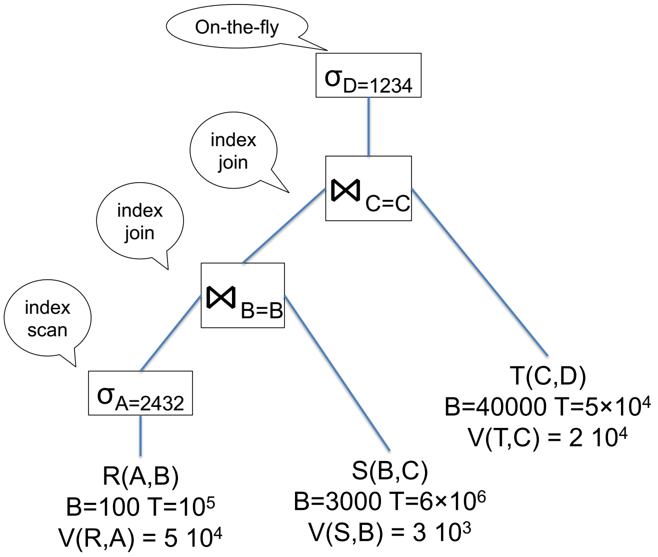
```
SELECT *
FROM R, S, T
WHERE R.A = 2432 and R.B = S.B and S.C = T.C and T.D = 1234
```

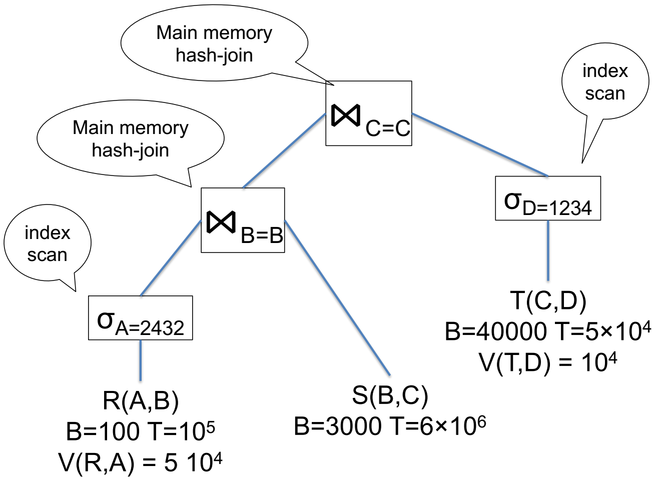
You should turn in an integer number.

iii. (10 points) Assume the following indices:

- Unclustered indexes on $R.A$ and $R.B$
- Clustered index in $S.B$, unclustered index on $S.C$.
- Clustered indexe on $T.C$, unclustered index on $T.D$.

Estimate the I/O cost for two the physical plans below. Use the same statistics as in the previous question (they are shown on the plans, for your convenience).





6 Conceptual Design

6. (35 points)

(a) (5 points) Consider the following table:

A	B	C	D	E
0	0	0	0	0
1	1	1	1	1
2	2	2	2	0
3	3	3	0	1
4	4	0	1	0
5	5	1	2	1
6	0	2	0	0
7	1	3	1	1
8	2	0	2	0
9	3	1	0	1
10	4	2	1	0
11	5	3	2	1
12	0	0	0	0

Find all functional dependencies that hold on this table. You only need to write a minimal set of functional dependencies that logically imply all others. Hint: notice that $B = A \mod 6$ etc; you should be able to find the FD's very fast.

(b) (5 points) Using the functional dependencies you have identified at the previous question, decompose the relation in BCNF. You only need to show your final result: the relation names, their attributes, and their keys (underline the key attributes).

(c) (5 points) The relation $R(A, B, C, D, E)$ satisfies the following functional dependencies:

$$AD \rightarrow E$$
$$CD \rightarrow A$$

Consider the relation returned by the following query:

```
      select distinct R.B, R.D, R.E, 'foo' as F
      from R
      where R.C = 'bar'
```

Find the key in the resulting table.

(e) For each statement below indicate whether it is true or false.

i. (3 points) A superkey is any set of attributes X such that $X = X^+$.

i. ____________

True or false?

ii. (3 points) For any set of attributes X, the set X^+ is a superkey.

ii. ____________

True or false?

iii. (3 points) For any set of attributes X, the following identity holds: $(X^+)^+ = X^+$.

iii. ____________

True or false?

iv. (3 points) If $X \cap Y$ is a superkey, then X is also a superkey.

iv. ____________

True or false?

v. (3 points) If $X \cup Y$ is a superkey, then X is also a superkey.

v. ____________

True or false?

7 Transactions

7. (45 points)

(a) A database consists of the following elements:

$$A_1, \ldots, A_{1000}$$

The system runs a workload of transactions of the following kind:

```
BEGIN
READ(A_i)
/* ... compute... compute... compute ... for 0.1 seconds */
WRITE(A_i)
COMMIT
```

Each transaction reads one element A_i, performs some intensive computation, then writes the same element back. Different transactions may access the same element or different elements.

The system is shared-memory, has 100 CPUs, and uses inter-query parallelism (multiple transactions are run in parallel, but each transaction runs on a single CPU). For serializability, the system uses one lock per element (like SQL Server).[1]

How many transactions per second can the system execute in each case below?

i. (5 points) All transactions access the same element A_1. That is, the transactions update elements in this sequence: $A_1, A_1, A_1, \ldots$
Answer Write the number of transactions per second (TPS):

ii. (5 points) The transactions update only elements from the first 50 elements. More precisely, each transaction reads/writes an element A_i chosen uniformly at random from among $A_1, \ldots, A_{50}$. For example, the transactions may update the elements in this sequence: $A_{44}, A_2, A_{13}, A_2, A_{36}, A_{11}, \ldots$
Answer Write the number of transactions per second (TPS):

iii. (5 points) The transactions update all elements. More precisely, each transaction reads/writes an element A_i, choosen uniformly at random from among $A_1, \ldots, A_{1000}$.
Answer Write the number of transactions per second (TPS):

[1] While in a real multicore system speedup is affected dramatically by latch contention, our multicore system has magic latches that do not cause any slowdown.

(b) For each schedule below indicate whether it is conflict serializable and, if it is, indicate the equivalent serial schedule.

i. (5 points)

$$R_1(A), W_2(B), R_3(C), W_3(A), R_4(D), R_4(B), W_1(D), W_2(C)$$

ii. (5 points)

$$R_1(A), W_2(B), R_3(C), W_3(A), R_1(D), R_4(B), W_4(D), W_2(C)$$

(c) For each of the following statements indicate whether it is true or false:

i. (2 points) In a static database, every serializable schedule is conflict serializable.

i. ____________

True or false?

ii. (2 points) In a dynamic database, every serializable schedule is conflict serializable.

ii. ____________

True or false?

iii. (2 points) In a static database, every conflict serializable schedule is serializable.

iii. ____________

True or false?

iv. (2 points) In a dynamic database, every conflict serializable schedule is serializable.

iv. ____________

True or false?

v. (2 points) When a deadlock occurs, then the database system aborts one or more transactions.

v. ____________

True or false?

Transactions

5. (40 points)

   (a) Answer the following questions:

   i. (3 points) Every serializable schedule is also conflict-serializable.

   i. ____________

   Answer Yes/No:

   ii. (3 points) Every conflict-serializable schedule is also serializable.

   ii. ____________

   Answer Yes/No:

   iii. (3 points) SQL Lite uses optimistic concurrency control.

   iii. ____________

   Answer Yes/No:

   iv. (3 points) Strict Two-Phase-Locking is guaranteed to produce a serializable schedule.

   iv. ____________

   Answer Yes/No:

   v. (3 points) Strict Two-Phase-Locking is guaranteed to produce a conflict-serializable schedule.

   v. ____________

   Answer Yes/No:

   vi. (3 points) Strict Two-Phase-Locking is guaranteed to avoid deadlocks.

   vi. ____________

   Answer Yes/No:

(b) For each of the schedules below, indicate whether they are conflict-serializable. If you answer *yes*, then give the equivalent serial order of the transactions. Show your work.

i. (6 points) Is this schedule conflict-serializable? Show your work; if you answer ’yes’, then indicate a serialization order.

```
R1(A), R1(B), W1(A), R2(B), W2(D), R3(C), R3(B), R3(D), W2(B), W1(C), W3(D)
```

ii. (6 points) Is this schedule conflict-serializable? Show your work; if you answer ’yes’, then indicate a serialization order.

```
R1(A), R1(B), W1(A), R2(B), W2(A), R3(C), R3(B), R3(D), W2(B), W1(C), W3(D)
```

(c) A scheduler uses the strict two-phase locking protocol. In each of the cases below, indicate whether the scheduler may result in a deadlock. If you answer *yes*, then give an example of a shedule that results in deadlock.

i. (5 points) Can these transactions result in deadlock?

```
T1: W1(A), W1(C), CO1

T2: W2(B), W2(D), CO2

T3: W3(A), W3(B), CO3

T4: W4(D), W4(A), CO4
```

Answer ’yes’ or ’no’. If you answer ’yes’ then also indicate a schedule that results in deadlock:

ii. (5 points) Can these transactions result in deadlock?

```
T1: W1(A), W1(C), CO1

T2: W2(B), W2(D), CO2

T3: W3(A), W3(B), CO3

T4: W4(B), W4(C), CO4

T5: W5(C), W5(D), CO5
```

Answer ’yes’ or ’no’. If you answer ’yes’ then also indicate a schedule that results in deadlock: