CSE 232A Fall 2018 Midterm Exam 2

Full Name:

Student ID:

INSTRUCTIONS

1. You have 50 minutes to complete this exam.

2. You can have up to one letter/A4-sized sheet of notes, formulae, etc. Apart from this, the exam is closed book/notes/electronics/peers.

3. Please wait until being told to start reading and working on the exam.

4. If you think a question is ambiguous, write down your assumptions, argue that they are reasonable, and then work on the problem using those assumptions.

5. Please ensure that your writing is clear and legible!

<table>
<thead>
<tr>
<th></th>
<th>Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 1</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Q 2</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Q 3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>
Test:
1. When did the Pilgrims land at Plymouth Rock?

1620.

As you can see, I've memorized this utterly useless fact long enough to pass a test question. I now intend to forget it forever. I've taught me nothing except how to cynically manipulate the system. Congratulations.

They say the satisfaction of teaching makes up for the lousy pay.
Q 1. [15pts] For the following questions, clearly circle True or False.

1. A given physical query plan can correspond to exactly one logical query plan.
   True  False

2. The RDBMS catalog is never useful for query optimization.
   True  False

3. Pipeline parallelism usually helps increase parallelism on multicore machines.
   True  False

4. It is sometimes possible to execute an entire SQL query on a table using only a given index structure on that table.
   True  False

5. Sort-merge join is always a blocking physical operator in the context of pipelined query execution.
   True  False

6. The selectivity of a conjunctive predicate is never higher than what is yielded by the independence heuristic assumption for costing plans.
   True  False

7. Materialized views typically increase runtimes for query execution but help reduce storage space requirements for the database.
   True  False

8. The network often becomes a bottleneck for data access in shared-disk parallel RDBMSs.
   True  False
9. Parallel RDBMSs often help reduce query runtimes even when the database can fit entirely on a single node’s disk.

   True    False

10. Superlinear speedups are impossible in parallel RDBMSs.

    True    False

11. It is impossible to modify a parallel RDBMS to make it fault tolerant.

    True    False

12. Hive on Hadoop+MapReduce is effectively another RDBMS.

    True    False

13. Spark better exploits distributed memory than Hadoop+MapReduce.

    True    False


    True    False

15. NoSQL systems are called so because they offer more expressive declarative query languages than SQL.

    True    False
Q 2. [15pts] Clearly circle the correct answer for each of the following questions (only one option is correct).

1. [3pts] Which of the following data partitioning schemes is the most common in parallel RDBMSs?
   
   (a) Round-robin        (b) Hash-based        (c) Range-based        (d) Random

2. [3pts] Which of the following SQL aggregates cannot always be computed in just one pass over the dataset in a typical parallel RDBMS?

   (a) AVG         (b) SUM         (c) VARIANCE        (d) MIN         (e) MEDIAN

3. [3pts] How many distinct hash functions does the “improved” parallel hash join implementation discussed in class need for the join phase?

   (a) 1       (b) 2       (c) 3       (d) 4       (e) 5       (f) 6

4. [3pts] Which relational operator’s implementation in a parallel RDBMS does the implementation of Hadoop+MapReduce most closely resemble?

   (a) σ        (b) π        (c) ⋈       (d) γ       (e) ρ       (f) ×

5. [3pts] Which of the following SQL aggregates cannot be recomputed incrementally with just algebraic rewrites during materialized view maintenance?

   (a) SUM       (b) COUNT       (c) AVG       (d) MIN       (e) MAX
Q 3. [10pts] Are you still on social media? Given the following relational database schema, for each of the given relational algebra query over this database, which of the relational algebra queries listed under it are logically equivalent to it? Clearly circle the correct answer (only one is correct).

Person (ID, Name, Age)
Friends (ID1, ID2)

Person.ID is the primary key of Person. Both Friends.ID1 and Friends.ID2 are foreign keys referring to Person.ID.

1. [3pts] \( \pi_{Name}(\sigma_{Age>20}(Person)) \)

(a) \( \sigma_{Age>20}(\pi_{Name}(Person)) \)
(b) \( \pi_{Name,ID}(\sigma_{Age>20}(Person)) \)
(c) \( \pi_{Name}(\sigma_{Age>20}(\pi_{Name,Age}(Person))) \)
(d) \( \sigma_{Age>20}(\pi_{Name,ID}(Person)) \)
(e) \( \sigma_{Name}(\pi_{Age>20}(Person)) \)
(f) \( \pi_{Name}(\pi_{Age}(\sigma_{Age>20}(Person))) \)

2. [3pts] \( \sigma_{(Age>20)\land(ID2=1234)\land(Name="Thanos")\land(ID=ID1)}(Person \times Friends) \)

(a) \( \sigma_{(Age>20)\land(Name="Thanos")}(Person) \times \sigma_{(ID2=1234)\land(ID=ID1)}(Friends) \)
(b) \( \sigma_{ID=ID1}(\sigma_{ID1=1234}(Friends) \times \sigma_{(Age>20)\land(Name="Thanos")}(Person)) \)
(c) \( \sigma_{(Age>20)\land(Name="Thanos")}(Person) \bowtie_{ID=ID1} \sigma_{ID1=1234}(Friends) \)
(d) \( \sigma_{ID2=1234}(Friends) \bowtie_{ID=ID1} \sigma_{(Age>20)\land(Name="Thanos")}(Person) \)
3. [4pts] $\pi_{Name,ID}(Person \bowtie_{ID=ID2} \pi_{ID2}(\sigma_{Name="Thanos"}(Person \bowtie_{ID=ID1} Friends)))$

(a) $\pi_{Name}(\pi_{ID1,ID2}(\pi_{ID}(\sigma_{Name="Thanos"}(Person)) \bowtie_{ID=ID1} Friends) \bowtie Person)$

(b) $\pi_{Name,ID}(Person) \bowtie_{ID=ID2} \pi_{ID2}(\sigma_{Name="Thanos"}(Person) \bowtie_{ID=ID1} Friends)$

(c) $\pi_{Name,ID}(Person) \bowtie_{ID=ID1} \pi_{ID1}(\sigma_{Name="Thanos"}(Person) \bowtie_{ID=ID1} Friends)$

(d) $\pi_{Name,ID2}(Person) \bowtie_{ID=ID2} \pi_{ID2}(\sigma_{Name="Thanos"}(Person) \bowtie_{ID=ID1} Friends)$