CSE 544: Homework 5
Some Theory Question

Dan Suciu

Released: February 27, 2021
Hard Deadline: Monday, March 15, 2021

Name: _____

Question	Points	Score
1	40	
2	40	
3	20	
Total:	100	

Write your answers in a file (pdf, or docx, or txt, or md) in the `submission` directory. Write the problem number and answer, for example (not real answers):

```
2.b.(ii)   0.73

3.a:       ABORT ABORT ABORT ABORT
```

Don't spend time trying to make the file look very nice, but make sure your answers are clear and unambiguous.

1. (40points)

   Consider a database schema with three relations and their sizes:

   ```
   Likes(drinker, beer)     // 100,000 tuples
   Frequents(drinker, bar)  // 2,000,000 tuples
   Serves(bar, beer)        // 5,000,000 tuples
   ```

   Consider the following two queries:

   ```
   Q1:
   select *
   from Likes, Frequents, Serves
   where Likes.drinker = Frequents.drinker
     and Frequents.bar = Serves.bar
   ```

   ```
   Q2:
   select *
   from Likes, Frequents, Serves
   where Likes.drinker = Frequents.drinker
     and Frequents.bar = Serves.bar
     and Likes.beer = Serves.beer
   ```

   Answer the questions below:

   (a) (20 points) What is the largest possible output of Q1?

   (b) (20 points) What is the largest possible output of Q2?

2. (40points)

   Consider the relation `Products(pid, name, category, price)`, with
   $N = 1,000,000,000$ tuples and $V(\text{Product}, \text{category}) = 100,000$ distinct categories.

   For the questions below use a calculator and write an approximate answer, or choose
   from the available options (whichever you prefer):

   (a) We hash-partition `Products` by the key `pid` over $P = 1,000$ servers.

       i. (10 points) What is the probability that the load of some server exceeds the
          expected load by more than 50%?
          - ≥ 0.99 (almost certain)
          - $(0.5, 0.99)$
          - $(0.1, 0.5)$
          - $(0.01, 0.1)$
          - $(0.001, 0.01)$
          - < 0.001 (absolutely tiny)

       ii. (10 points) What is the probability that the load of some server exceeds the
           expected load by more than 0.5%?
           - ≥ 0.99 (almost certain)
           - $(0.5, 0.99)$
           - $(0.1, 0.5)$
           - $(0.01, 0.1)$
           - $(0.001, 0.01)$
           - < 0.001 (absolutely tiny)

   (b) Now suppose we hash-partition `Products` by `category` over $P = 1,000$ servers.
       Assume that each category occurs the same number of times.

       i. (10 points) What is the probability that the load of some server exceeds the
          expected load by more than 50%?
          - ≥ 0.99 (almost certain)
          - $(0.5, 0.99)$
          - $(0.1, 0.5)$
          - $(0.01, 0.1)$
          - $(0.001, 0.01)$
          - < 0.001 (absolutely tiny)

       ii. (10 points) What is the probability that the load of some server exceeds the
           expected load by more than 0.5%?
           - ≥ 0.99 (almost certain)
           - $(0.5, 0.99)$
           - $(0.1, 0.5)$
           - $(0.01, 0.1)$
           - $(0.001, 0.01)$
           - < 0.001 (absolutely tiny)

3. (20points)

 (a) (20 points) Consider a multiversion concurrency control manager. Indicate what happens for each question mark below. If it is a write operation say whether it is executed or aborted; if it is a read operation, indicate the value returned (e.g. say "$X = 9$"), or say that is aborted. An instruction like $W_t(X = v)$ means that the transaction with timestamp t writes the element X, and the value of X being written is v.

START_1	(executed OK)
$W_1(X = 1)$	(executed OK)
COMMIT_1	(executed OK)
START_2	(executed OK)
START_3	(executed OK)
START_4	(executed OK)
$W_4(X = 4)$	(executed OK)
COMMIT_4	(executed OK)
START_5	(executed OK)
START_6	(executed OK)
START_7	(executed OK)
$W_7(X = 7)$	(executed OK)
COMMIT_7	(executed OK)
$R_5(X)$	?
$R_3(X)$	?
$W_2(X = 2)$	?
$W_6(X = 6)$	?