COMP 62421

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Querying Data on the Web

Date: Wednesday 24th January 2018

Time: 14:00 - 16:00

Please answer all FIVE Questions provided. They amount to a total of 50 marks.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

1. a) Given two relations $R(a : int, b : str)$ and $S(c : int, d : str)$, let $|R| > |S|$. For each of the algebraic expressions below, write down an expression that characterizes its minimum cardinality and another that characterizes its maximum cardinality, stating any assumptions you have made in your answer. (4 marks)

   i) $R \cap S$

   ii) $\pi_a R$

   b) Let *BBD* be a database schema on beers (made by some brewer) that are served in bars (at some address and for a certain price), and on drinkers (at some address) that frequent bars a certain number of times per week, where the beers that drinkers like are also recorded, as follows:

```
Beer(name,brewer)
Serves(bar[fk Bar.name],beer[fk Beer.name],price)
Bar(name,address)
Drinker(name,address)
Frequents(drinker[fk Drinker.name],bar[fk Bar.name],time_a_week)
Likes(drinker[fk Drinker.name],beer[fk Beer.name])
```

   i) Code a solution to the following problem as a TRC expression over the *BBD* database schema: *Return a unary relation with column* `regulars` *where* `regulars` *contains the* `name` *of a* `drinker` *that* `frequents` *at least one* `bar` *more than once a week.* (2 marks)

   ii) Code a solution to the following problem as a SQL query over the *BBD* database schema: *Return a ternary relation with columns* `barName`, `barAddress` *and* `bestValue`, *where* `barName` *contains the* `name` *of the* `bar`, `barAddress` *contains the* `address` *of the* `bar`, *and* `bestValue` *contains the* `price` *of the cheapest* `beer` *that the* `bar serves`, *provided that that* `price` *is less than 3.* Note that the result will have one row per bar (assuming each bar only exists in one address). (4 marks)

2. a) Assume that a query execution plan contains the subtree $R \bowtie S$ over intermediate results R and S. Further assume that the DBMS has been configured to use B bytes of buffer space. Briefly explain how large B must be in terms of the size in bytes of R and S for it to make sense for the optimizer to select a hash join as the physical algorithm with which to evaluate $R \bowtie S$. (2 marks)

   b) Assume that a query execution plan contains the subtree $\sigma_{a>3}(R \bowtie_c \pi_b(S))$ over intermediate results R and S, where $a \in schema(R)$ and $b \in schema(S)$ and $c \in schema(R) \cap schema(S)$. Further assume that execution is pipelined and that the optimizer selects a hash join to evaluate the join in this query plan fragment. Briefly explain how this choice impacts on operator-level multithreading of the right child of the join (i.e., the projection) and, likewise, of the parent of the join (i.e., the selection). (2 marks)

   c) Consider the following query plan fragment: $R \bowtie S \bowtie T \bowtie U$. Assume that the average width of a tuple is the same for all the leaf nodes (i.e., R, S, T, U) and that that width is 10. Further assume that the cardinalities of the leaf nodes are as follows: $|R| = 20$, $|S| = 10$, $|T| = 12$, $|U| = 15$. Finally, assume the worst-case cardinality for the result of all the joins in the plan fragment except when they involve S or T, in which case empirical evidence shows that the resulting size in bytes is, respectively, 10% and 20% smaller than the worst case would have been. State which join order would be selected by the greedy algorithm taught in this course unit indicating the outcome of each pass in the algorithm's execution on the given inputs and stating any assumptions you have made in your answer.
   (6 marks)

3. a) In the context of XQuery Core, briefly describe what is meant by a static semantics. (1 mark)

   b) Consider the following FLWOR expression F:

```
for $i in (<item>C</item>,
           <item>A</item>,
           <item>B</item>)
return <creator>{data($i)}</creator>
```

   Now, consider the following trace of the evaluation of F using the `sequence`, `right unit`, `let` and `data` equivalence laws, where `E1`, `E2`, `E3` and `E4` act as place-holders. Using your knowledge of those equivalence laws, write down the XQuery expressions that instantiate `E1`, `E2`, `E3` and `E4`.

```
  for $i in (<item>C</item>,
             <item>A</item>,
             <item>B</item>)
  return <creator>{data($i)}</creator>
= (sequence)
  E1
= (right unit)
  E2
= (let)
  E3
= (data)
  E4
```

(4 marks)

   c) Write out the expression that results from mapping the following XPath expression into XQuery Core:

```
$root/X/B/y[@z > 0]
```

(5 marks)

4. a) State any <u>two</u> of the various serialization syntaxes for storing and exchanging RDF. (2 marks)

   b) Use the triple binary tables strategy to map the following RDF triples into relational tables. You only need to show the `s` and `p` tables. (8 marks)

```
(uk, capital, london)
(uk, area, 242495)
(usa, capital, washington)
(usa, area, n)
```

5. a) Briefly explain why map-reduce computations so directly express queries of the form

$$\texttt{SELECT } \lambda, \Gamma(\alpha) \texttt{ FROM } \Phi \texttt{ WHERE } \theta \texttt{ GROUP BY } \lambda$$

where Φ is a relation with schema $\lambda \cup \{\alpha\}$, Γ is an element of {`COUNT`, `SUM`, `MAX`, `MIN`, `AVG`}, and θ is a predicate over the schema of Φ. (2 marks)

b) Consider, in scientific computing, the volume of data generated by the Large Hadron Collider (LHC) at CERN, i.e., some 30 petabytes or so, annually. Given the four characteristics of so-called *big data* known as *the four Vs*, briefly comment on how each one applies or not in the case of the LHC at CERN . (4 marks)

c) Consider the following RA algebraic expression over the database in Q1b above:

$$\texttt{Bar} \bowtie \texttt{Frequents}$$

Sketch the pseudocode of the mapper and reducer functions that would compute the correct value for the given algebraic expression. (4 marks)

END OF EXAMINATION