COMP 62421

Two hours - online

This paper version is made available as a backup.

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Querying Data on the Web

Date: Thursday 17th January 2019

Time: 09:45 - 11:45

This is an online examination. Please answer ALL Questions

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) List the three main components in a query processor. (2 marks)

   b) Describe the three main issues in query optimization. (3 marks)

   c) Given two relations $R(a : int, b : str)$ and $S(c : int, d : str)$, let $|R| < |S|$. For each of the algebraic expressions lsited 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. (5 marks)

      i) $R \cap S$

      ii) $\pi_a R$

2. Write a SPARQL query over DBpedia which answers the following query: 'Does Angelina Jolie have more children than Madonna?' Assume the scenario where the range of the predicate 'children' corresponds to the number of children.

   The URIs which you need for issues this query are:

   ```
   <https://dbpedia.org/resource/Angelina_Jolie>
   <https://dbpedia.org/resource/Madonna(entertainer)>
   <http://dbpedia.org/property/children>
   ```

   (5 marks)

3. a) Explain why map-reduce computations can 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. (4 marks)

b) Describe the typical recipe for processing Graph Algorithms using MapReduce (6 marks)

4. Articulate the complementarity (at least two items) of SQL and NOSQL systems.
(5 marks)

5. Contrast ACID guarantees in SQL databases and BASE guarantees in NoSQL databases.
(6 marks)

6. For the RA algebraic expression below: $(R \bowtie S) \cup T$, where the schemas of the input relations are:

$$R(a,b), S(b,c), T(a,b,c)$$

For each non-leaf node in the logical operator tree, sketch the pseudocode of the mapper and reducer functions that would compute the correct value for the given algebraic expression. (10 marks)

7. Compare and contrast MapReduce, DataFlow Engines to Massively Parallel Processing (MPP) Databases. (5 marks)

8. Design the high-level components of a data architecture for the scenarios described below. Include a diagrammatic representation containing the main high level data processing components, their associated data processing paradigms and examples of supporting software frameworks (including database systems) for each component. Please provide a justification for each design choice.

   i) A realtime threat detection tool (e.g. detecting hacking attempts), consuming unstructured log data from 100 distributed firewalls from 3 different vendors, where each firewall produces 106 events (1 event is one log line) per second. The client requires the detection of specific events and also the integration, aggregation of data from different events in order to generate a threat alert. (6 marks)

   ii) Stratified sampling of 10PB of data (100 attributes) from an experiment performed yesterday at the Large Hadron Collider (LHC). (3 marks)

END OF EXAMINATION