COMP62421

Two hours - online

UNIVERSITY OF MANCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Querying Data on the Web

Date: Thursday 16th January 2020

Time: 09:45 - 11:45

This is an online examination. Please answer ALL Questions
The examination is worth a total of 50 marks

This is a CLOSED book examination

The use of electronic calculators is NOT permitted

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

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

2. Given two relations $R(a : int, b : str)$ and $S(c : int, d : str)$, let $|R| < |S|$. For each of the algebraic expressions listed 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. Represent: $|R \cap S|$ as $|R$ int $S|$ and $|\pi_a R|$ as $|pi(a)R|$.

   (5 marks)

   i) $R \cap S$

   ii) $\pi_a R$

3. 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 to issue this query are:

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

   (5 marks)

4. a) Briefly explain why the triple store strategy for mapping RDF data into relational data creates obstacles for efficient evaluation. (2 marks)

   b) Use the property table strategy to map the following RDF triples into relational tables: (3 marks)

```
(a1, name, john)
(a1, address, fallowfield)
(a2, name, helen)
(a2, address, chorlton)
(a3, name, mary)
(a3, address, mossside)
(a1, likes, a3)
(a3, likes, a2)
```

5. 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.

Represent Γ as Gamma, θ as theta, α as alpha and Φ and Phi.

(4 marks)

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

7. Considering the RA algebraic expression: $(R \bowtie S) \cup T$, 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. Represent $(R \bowtie S) \cup T$ as $(R >< S)UT$.

(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.

   Create your diagrams using the following form:

   $--|COMPONENT1|-->|COMPONENT2|-->...$

   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. (5 marks)

   ii) A query system over a knowledge graph (KG) extracted from Wikipedia text (51 GB). *Users need to be abstracted away from the schema*. Include all the text processing pipeline components. Include the data model(s) used and describe the scalability strategies for querying the large-scale KG.

   (10 marks)

END OF EXAMINATION