IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2019

MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C412H

LARGE SCALE DATA MANAGEMENT

Tuesday 19th March 2019, 11:40
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators not required

1 Write queries for graph databases and document stores in the following two parts.

a A simple Neo4J graph database stores information about books, their authors, readers and publishers. Each book has author(s), editor(s) as well as a publisher. Each book also has a number of readers which also frequent shops (presumably where the books are sold) which have a location. The books can also be sold by online shops where readers can buy them. Editors edit books and work for a publisher. Each book is published by a publisher. The data model is shown in Figure 1, possible attributes are listed in the boxes next to the nodes.

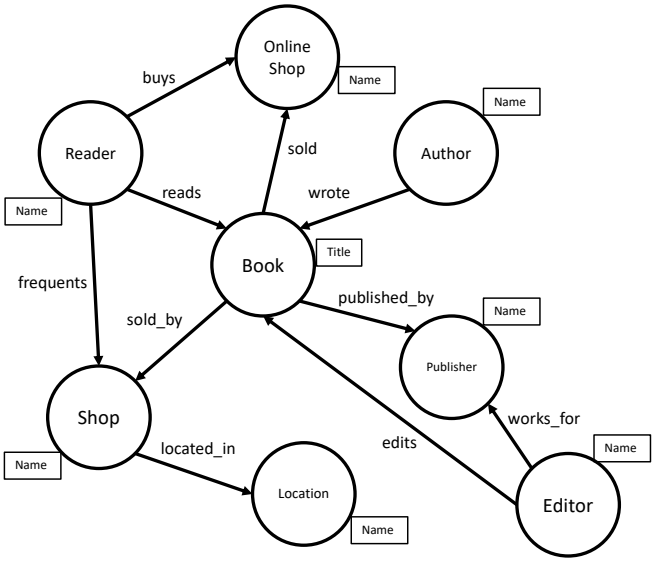

Fig. 1: Data model for books.

Given this description, write Cypher queries for the following:

i) Count all books written by Mark Needham.

ii) Return a list of all shops selling the book ”Introduction to Neo4J”.

iii) Find the names of all shops and locations the reader by the name of ”Mark” frequents/visits.

iv) Find how many books each publisher has published by author and list all editors who have work on books written by each author.

v) Return the names of all editors who work for a publisher which published books that the reader ”Mark” read, but who have not edited Mark’s books.

b You are given the structure of the *people* collection, containing information about people, their location and the cars that they own etc., as shown in Figure 2.

```
{
        first_name: “Paul”,
        surname: “Miller”,
        city: “London”,
        location: [45.123,47.232],
        address: {
                street: “Queens Gate”
                postcode: “SW72AZ”
        },

        cars: [
                { model: “Bentley”, year: 1973, value: 100000, ….},
                { model: “Rolls Royce”, year: 1965, value: 330000, ….}
        ]
}
```

Fig. 2: Example of the people collection in MongoDB.

i) Write a query to display all the people in the collection *people*.

ii) Write a query to find all people at an address with the postcode SW35HF.

iii) Write a query to find all people living within a distance of 1000 meters of the point with coordinates 45, 47.

iv) Write a query to find all people who own a Bentley and a Rolls Royce.

v) Find all people who own at least one car which that is worth more than 100000.

The two parts carry equal marks.

2 Consider a disk with a sector size of 1024 bytes, a platter surface size of 800, a track size of 50 and 5 double-sided platters. Assume a block size of 1,024 bytes.

Further assume that a file containing 10,000 records of 200 bytes each is to be stored on such a disk and that no record is allowed to span two blocks.

a What is the capacity of a track in bytes? What is the capacity of each surface? What is the capacity of the disk? How many blocks does the disk have? Explain your answer briefly.

b How many blocks are required to store the entire file? How much space (bytes) is wasted? Is there potentially space wasted if it is stored in main memory? Given a sequential read of the file retrieving each record separately from memory using a cache with a cache line of 60 bytes, how much memory bandwidth is wasted? Explain your answer briefly.

c Name one storage class technology that may replace or complement SSD in the near future and discuss all its benefits as well as drawbacks compared to main memory? How will its characteristics evolve over time?

d What is the purpose of the flash translation layer? What functions does it provide?

The four parts carry equal marks.

3 Document and graph databases.

a Some document databases, e.g., MongoDB, are schema-free. Discuss how documents (as used and stored in MongoDB) can be stored in a relational database. What are the challenges?

b MongoDB does not support joins. Explain methods that can be used to combine documents along with their upsides and downsides.

c Define the schema on how to store an XML document in a graph database. Is the opposite (storing a graph in an XML document) possible? What are the challenges

d How do graph databases (specifically Neo4J) differ from document databases (specifically MongoDB) regarding joins? What are the consequences for the size of the data on disk? Compare the two approaches.

The four parts carry equal marks.