FINAL

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017

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 21 March 2017, 10:00
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators required

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

a A simple Neo4J graph database stores information about customers who order products supplied by a supplier. Each product belongs to a category. The customer purchases an order, each order contains one or more products. Each supplier supplies one or more products, and each product belongs to one category. Further assume that each node type also has an attribute name. The data model is illustrated in Figure 1.

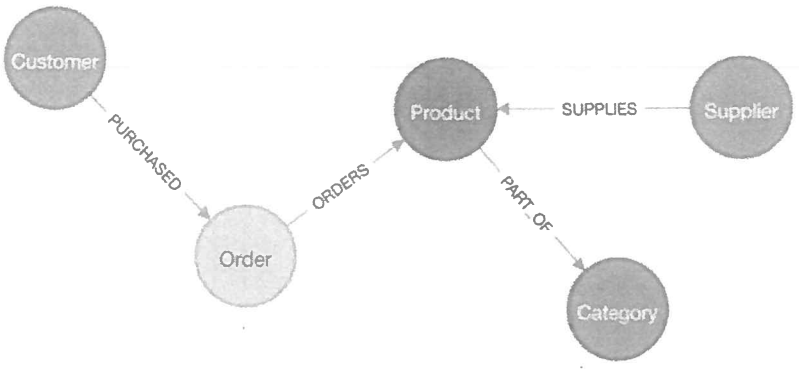

Fig. 1: Data model to manage orders, customers, producers etc.

Given this description, write Cypher queries for the following:

i) List all products that have been ordered by customers.

ii) Has the customer with name C ordered a product with name P produced by a supplier with name S ?

iii) Find all the the categories and suppliers that the customer with the name C ordered from.

iv) Count the different categories from which a customer with the name C has ordered. Return the count and the name of the category.

v) Find all the products, orders as well as suppliers that are related to category C. Use a query with one relationship only.

b You are gven the structure of the *albums* collection, containing information about albums, their content, category, price etc. shown in Figure 2.

```
{
  type: "Audio Album",
  title: "A Love Supreme",
  description: "by John Coltrane",

  shipping: {
    weight: 6,
    dimensions: {
      width: 10,
      height: 10,
      depth: 1
    },
  },

  pricing: {
    list_price: 1200,
    retail_price: 1100,
    savings: 100,
  },

  details: {
    title: "A Love Supreme [Original Recording Reissued]",
    artist: "John Coltrane",
    genre: [ "Jazz", "General" ],
    tracks: [
      "A Love Supreme Part I: Acknowledgement",
      "A Love Supreme Part II - Resolution"
    ],
  },
}
```

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

i) Write a query to display all the documents in the collection albums.

ii) Write a query to find the albums which have a list price bigger than 1200.

iii) Write a query to find the albums where the title starts with 'A'.

iv) Write a query to find the albums by 'John Coltrane' with the weight either 5 or 8.

v) Join all albums with information about sales from a second document *sales*: $\{"title" : "A\ Love\ Supreme", "sales" : "..."\}$

The two parts carry equal marks.

2 Consider a disk with a sector size of 512 bytes, a surface size of 200, a track size of 50, 5 double-sided platters and average seek time of 10 ms. Assume a transfer time of 1 ms per block, an average rotational delay of 5 ms, and a block size of 1,024 bytes.

Further assume that a file containing 10,000 records of 100 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 75 bytes, how much memory bandwidth is wasted? Explain your answer briefly.

c Given these characteristics and a query that retrieves 100 records, (uniformly randomly distributed in the database) is using an index worth it (assuming the index lookup is free)?

d What is the time required to read a file stored on disk containing 10,000 records of 100 bytes each sequentially? What if each record is retrieved using random access? How would your answer change for sequential reading if the seek time was 5ms? How would it change for random reading is the seek time was 5ms?

The four parts carry equal marks.

3 Assume a regular Flash/SSD device.

a What is the purpose of the FTL (flash translation layer)? What functions does it provide?

b How does random and sequential access in an SSD compare to a magnetic disk? Briefly explain why. Why is a random write slower than a random read? Why is an erase block bigger than a read block?

c Name one storage class technology that may replace or complement SSD in the near future and discuss its benefits as well as drawbacks compared to SSD. What is its main benefit compared to main memory?

d Why are random writes slower than sequential writes on an SSD? Discuss how we can address the issue of slow random writes on an SSD device.

The four parts carry equal marks.