FINAL

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2018

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 20 March 2018, 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 flights (each associated with a ticket) connecting airports. The data model is illustrated in Figure 1, possible attributes are listed in the nodes.

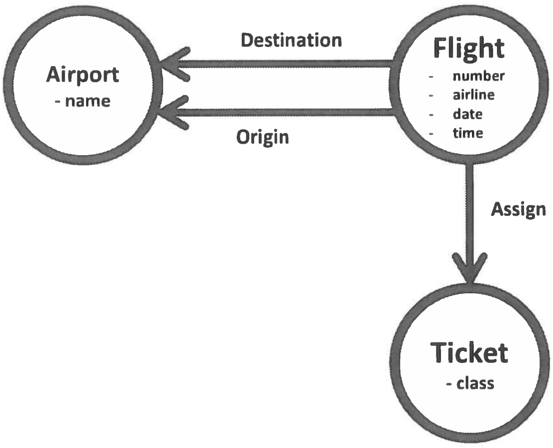


Fig. 1: Data model for flights, tickets and airports.

Given this description, write Cypher queries for the following:

i) Count the number of airports the airline LX serves.

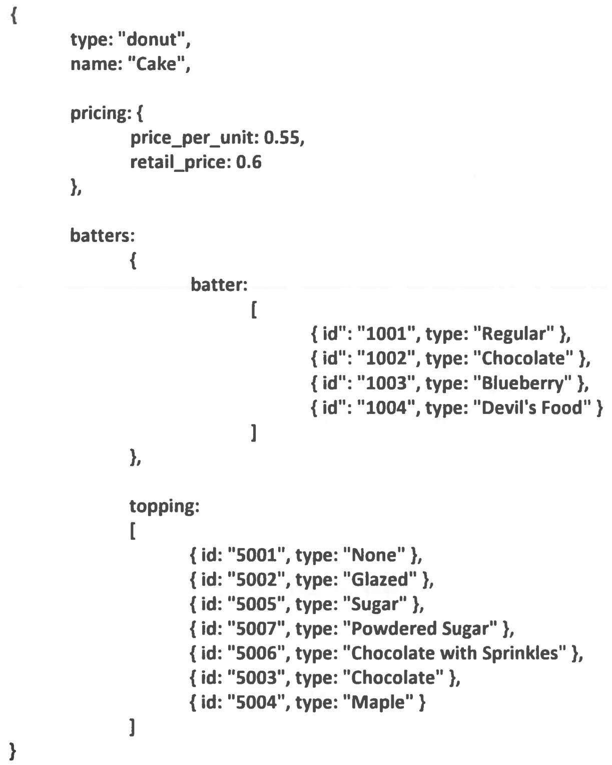
ii) Return a list of all business class tickets for direct flights originating from Seattle.

iii) Find flights with one stop from Seattle to San Francisco.

iv) Find how many flights per airline are originating from each city.

v) Find all the airports which can be reached from Seattle through between 2 and 5 stops.

b You are given the structure of the *bakeware* collection, containing information about bakeware, their topping, category, price etc., as shown in Figure 2.



```
{
        type: "donut",
        name: "Cake",

        pricing: {
                price_per_unit: 0.55,
                retail_price: 0.6
        },

        batters:
                {
                        batter:
                                [
                                        { id": "1001", type: "Regular" },
                                        { id": "1002", type: "Chocolate" },
                                        { id": "1003", type: "Blueberry" },
                                        { id": "1004", type: "Devil's Food" }
                                ]
                },

                topping:
                [
                        { id: "5001", type: "None" },
                        { id: "5002", type: "Glazed" },
                        { id: "5005", type: "Sugar" },
                        { id: "5007", type: "Powdered Sugar" },
                        { id: "5006", type: "Chocolate with Sprinkles" },
                        { id: "5003", type: "Chocolate" },
                        { id: "5004", type: "Maple" }
                ]
}
```

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

i) Write a query to display all the baked goods in the collection bakeware.

ii) Write a query to find all baked goods which have a price per unit bigger than 1.

iii) Write a query to find all baked goods where the name starts with 'C'.

iv) Write a query to find all baked goods of type Donut with chocolate dough but without sugar topping.

v) Find all baked goods that contain chocolate.

The two parts carry equal marks.

2 Consider a disk with a sector size of 1024 bytes, a surface size of 400, 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 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 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. How will its characteristics in terms of density, access time and endurance evolve over time?

d Why are random writes slower than sequential writes on an SSD? Why is the block the smallest unit that can be erased on an SSD device?

The four parts carry equal marks.

3 Differences between XML shredding and document databases.

a Some document databases, e.g., MongoDB, are schema-free. Can a similar schema-free XML shredding approach be implemented on top of a relational database? If so, what are the upsides and what are the downsides?

b MongoDB does not support joins and instead uses embedding and linking. Explain both, with their upsides and downsides.

c Briefly explain the XML shredding approaches XRel, XParent and Edge. Discuss their comparative performance.

d How does MongoDB replicate data? What roles other than the primary are there and what happens if the primary fails? How can one scale out with MongoDB? How are queries routed?

The four parts carry equal marks.