IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2016

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)
MRes in High Performance Embedded and Distributed Systems
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

Monday 21 March 2016, 14:00
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators required

1 In the International Cricket Council 2015 World Cup which will be hosted by Australia and New Zealand together, 14 teams will be competing to win the World Cup Trophy. ICC have divided all the International Cricket Playing nations into two Pools, pool A and pool B, each Pool having 7 teams. Pool A teams will play matches against other teams in Pool A; similarly Pool B teams will play matches among themselves. Each match will be played between two teams at some Venue. In each match three umpires will be there, one of them playing the role of third umpire and other two will be field umpires. The data model is illustrated in Figure 1.

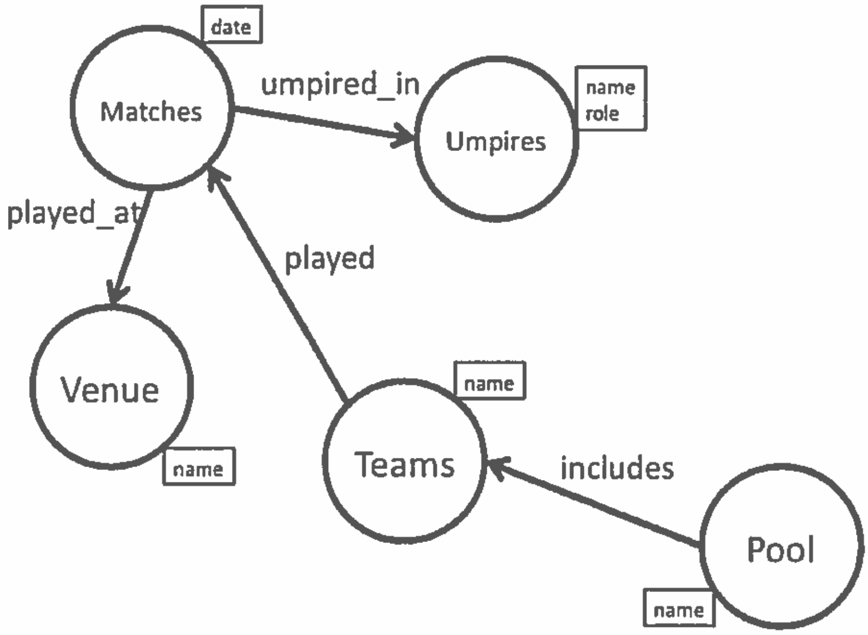

Fig. 1: Data model of the ICC 2015 World Cup.

Given this description, write Cypher queries for the following:

a List all the field umpires.

b Will Ireland play any match on February 25 2015?

c Find all the venues where India will play its matches.

d Find the pool matches that will be played at Hagley Oval, Christchurch.

e Find the venue where the biggest number of matches will be played.

f Which matches will be played on 8 March 2015?

g Find all umpires who will umpire a game from pool A on March 8 2015 at Hagley Oval, Christchurch.

h Will there be any pool match played between India and Pakistan?

i Who will be the field umpires in the India and Pakistan match?

j Find all the matches India has played along with the venues where they have been played as well as the umpires. Use a query with one relationship only.

The ten parts carry equal marks.

2 Given the following structure of the *restaurant* collection:

```
{
   "address": {
      "building": "1007",
      "coord": [ -73.856077, 40.848447 ],
      "street": "Morris Park Ave",
      "zipcode": "10462"
   },
   "borough": "Bronx",
   "cuisine": "Bakery",
   "grades": [
      { "date": { "$date": 1393804800000 }, "grade": "A", "score": 2 },
      { "date": { "$date": 1378857600000 }, "grade": "A", "score": 6 },
      { "date": { "$date": 1358985600000 }, "grade": "A", "score": 10 },
      { "date": { "$date": 1322006400000 }, "grade": "A", "score": 9 },
      { "date": { "$date": 1299715200000 }, "grade": "B", "score": 14 }
   ],
   "name": "Morris Park Bake Shop",
   "restaurant_id": "30075445"
 }
```

Write the following MongoDB queries:

a Write a query to display all the documents in the collection restaurants.

b Write a query to display all the restaurants which are in the borough Bronx.

c Write a query to find the restaurants that achieved a score of more than 90.

d Write a query to find the restaurants which have a latitude value less than -95.754168.

e Write a query to find the restaurant ID, name, borough and cuisine for those restaurants that contain 'Wil' as the first three letters in their name.

f Write a query to find the restaurants which belong to the Bronx borough and prepare either American or Chinese dishes.

g Write a query to find the restaurants with a latitude between 42 and 52. Sort the result by the name of the restaurant, address and latitude in ascending order.

h Write a query to arrange the name of the cuisine of all restaurants in ascending order. Those with the same cuisine should be ordered by the name of their borough in descending order.

i Write a query which will select all documents in the restaurants collection where the coord field value is a double.

j Join all restaurants with information about boroughs from a second document *boroughs*: $\{"borough" : "Bronx", "information" : "..."\}$

The ten parts carry equal marks.

3 Consider a disk with a sector size of 512 bytes, a surface size of 2000, a track size of 50, 5 double-sided platters and average seek time of 10 ms. Assume a transfer time of 1ms per block, an average rotational delay of 5 ms and a block size of 1,024 bytes.

Further assume that a file containing 100,000 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? Explain your answer briefly.

c How many records of 100 bytes each can be stored using this disk? Explain your answer briefly.

d What is the time required to read a file containing 100,000 records of 100 bytes each sequentially? How would your answer change if the disk were capable of reading/writing from all heads in parallel (and the data was arranged optimally)? Explain your answer briefly.

e Discuss the use of SSD/Flash in a database system. Give three possible ways in which SSDs can be used in a database system. Why are random write on SSDs slower than sequential writes? What do random writes mean for the durability of SSD devices?

The five parts carry equal marks.