UNIVERSITY OF LONDON

CM3010

BSc EXAMINATION

COMPUTER SCIENCE

Databases and Advanced Data Techniques

Release date: Tuesday 25 March 2025 at 12:00 midday Greenwich Mean Time

Close date: Wednesday 26 March 2025 by 12:00 midday Greenwich Mean Time

Time allowed: 4 hours to submit

INSTRUCTIONS TO CANDIDATES:

Part A of this assessment consists of a set of **TEN** Multiple Choice Questions (MCQs). You should attempt to answer **ALL** the questions in **Part A.** The maximum mark for Part A is **40**.

Candidates must answer **TWO** out of the **THREE** questions in **Part B**. The maximum mark for Part B is **60**.

Part A and Part B will be completed online together on the Inspera exam platform. You may choose to access either part first upon entering the test area but must complete both parts within **4 hours** of doing so.

Calculators are **NOT** permitted in this examination.

You may use **ONE** A4 page of previously prepared notes in this examination. Please hold up your notes to the camera at the start of the examination.

File upload is **NOT** permitted in this examination.

Do not write your name anywhere in your answers.

PART A

Question 1

Candidates should answer the **TEN** Multiple Choice Questions (MCQs) in Part A.

PART B

Candidates should answer any **TWO** questions from Part B.

Question 2

For several centuries, officials in London published weekly 'bills of mortality' listing deaths in each parish (district) in the city's jurisdiction. Cambridge researchers have gathered data of causes of death from 1644-1849 and ages of death from 1729-1849. They have shared this as a series of CSV files. Weeks are numbered, with the first bill of each year having a week number of 1.

For example, the file `ages.txt` begins as follows:

```
ageID|weekID|agegroup|ageyearmin|ageyearmax|agen
1|1729/01|under 2|0|1|141
2|1729/01|2-5|2|4|52
3|1729/01|5-10|5|9|17
```

A file for deaths, by parish, (and whether they were caused by plague) is called counts.txt, and begins like this:

```
countID|weekID|parcode|counttype|countn
4216|1644/15|GEOS|plague|1
4218|1644/15|OLSW|plague|1
4233|1644/15|WHCL|plague|3
4244|1644/15|ROTH|plague|1
4245|1644/15|STEP|plague|2
```

And a key for decoding the 4-letter parish codes is in ParcodeDict.txt:

```
parcode|parish|alias1|alias2|billsgroupbefore1660|billsgroupafter1660
GEOS|St George Southwark|||without|without
OLSW|St Olave Southwark|||without|without
ROTH|St Mary Rothorithe|Rothorith Parish|St Mary
Rothorith|other|outparishes
STEP|St Dunstan Stepney|Stepney Parish||other|outparishes
WHCL|St Mary Whitechappel|||outparishes|outparishes
```

(a) A separate research group would like to work with this data. They have decided to prepare a MySQL database. List all the tables, keys, and fields you would recommend. Explain any fields that you have removed or added and state what normal forms the result is in.

[12 marks]

(b) How did you decide how to represent date in your database? In 1752, as a correction (and to align with the rest of Europe), the English calendar skipped 11 days – Saturday 2 September was followed by Thursday 14 September. The bills of mortality are numbered sequentially across this period, with no gap (so there are only 51 bills in that year instead of 52). What issues does this raise for your database?

[3 marks]

(c) Give a MySQL query to retrieve from your database the number of deaths in St Dunstan, Stepney, reported in week 2 of 1729 that were attributed to plague.

[2 marks]

(d) Give a MySQL query to retrieve the annual number of deaths for each age group between 1760 and 1790.

[4 marks]

(e) A fourth file gives a more detailed breakdown of causes of death, but for the whole city (i.e. no separate data for each parish). The file begins like this:

```
codID|weekID|cod|codn
4300|1644/15|Aged|18
4307|1644/15|Bloody Flux|2
4319|1644/15|Childbed|2
4322|1644/15|Chrisoms|7
4325|1644/15|Consumption|34
4326|1644/15|Convulsions|10
```

Show how you would add this data to your database and give a query to check whether the number of deaths given for each parish adds up to the number given for the whole city.

[5 marks]

(f) Some researchers want to use this dataset to explore trends in population health over time. What issues can you imagine arising from this weekly dataset of causes of death in London spanning 205 years? What external information would be useful to add to it to give context?

[4 marks]

Question 3

The following is the beginning of a recipe from a website called BeerSmith:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
  <!-- BeerXML Format - Generated by BeerSmith  - see www.beersmith.com
-->
  <RECIPES>
    <RECIPE>
      <NAME>Burton Ale</NAME>
      <VERSION>1</VERSION>
      <TYPE>All Grain</TYPE>
      <BREWER>Brad Smith</BREWER>
      <ASST_BREWER></ASST_BREWER>
      <BATCH_SIZE>18.92716800</BATCH_SIZE>
      <BOIL_SIZE>20.81988500</BOIL_SIZE>
      <BOIL_TIME>60</BOIL_TIME>
      <EFFICIENCY>72.0</EFFICIENCY>
      <HOPS>
        <HOP>
          <NAME>Goldings, East Kent</NAME>
          <VERSION>1</VERSION>
          <ORIGIN>United Kingdom</ORIGIN>
           <ALPHA>5.50</ALPHA>
          <AMOUNT>0.0283500</AMOUNT>
          <USE>Boil</USE>
          <TIME>60.000</TIME>
          <NOTES>Used For: General purpose hops for bittering/finishing
all British Ales
Aroma: Floral, aromatic, earthy, slightly sweet spicy flavor
Substitutes: Fuggles, BC Goldings
Examples: Bass Pale Ale, Fullers ESB, Samual Smith&amp;apos;s Pale Ale
         </NOTES>
         <TYPE>Aroma</TYPE>...
```

(a) What format is this file in?

[1 mark]

(b) What is the root node?

[1 mark]

(c) What schema, if any, does this document use? How would the file be validated?

[3 marks]

(d) Write an xpath query to return the names of all hops used in the recipe called ‘Burton Ale’.

[4 marks]

(e) A researcher builds a machine-learning system that reads these files and tries to classify the style of beer from a list of 15 beer types. They use 10-fold cross validation. What does that mean?

[3 marks]

(f) The resulting system returns the correct label 50% of the time. What would you want to know to decide whether that was a good result?

[6 marks]

(g) Do you think this system is being used as part of a document database system or as a data interchange format between software? Why do you think that?

[3 marks]

(h) Evaluate the relative merits of a tree-based approach compared with a graph or relational model for this sort of data. Explain what you would need to know to make a recommendation.

[9 marks]

Question 4

I ran the following command:

```
curl -H "Accept: application/ld+json” https://musicbrainz.org/artist/183d6ef6-e161-47ff-9085-063c8b897e97
```

I then used some software to convert the result to the following (edited and shortened for this exam):

```
@prefix schema: <http://schema.org/> .
@prefix mbarea: <http://musicbrainz.org/area/> .
@prefix mbart:  <http://musicbrainz.org/artist/> .
@prefix mbrelease: <http://musicbrainz.org/release-group/>

mbarea:05f68b4c-10f3-49b5-b28c-260a1b707043 a schema:AdministrativeArea ;
    schema:containedIn mbarea:489ce91b-6658-3307-9877-795b68554c98 ;
    schema:name "Massachusetts" .

mbarea:11c4099a-ff61-45a3-ada4-23ac7a25d111 a schema:City ;
    schema:containedIn mbarea:05f68b4c-10f3-49b5-b28c-260a1b707043 ;
    schema:name "Lincoln" .

mbarea:489ce91b-6658-3307-9877-795b68554c98 a schema:Country ;
    schema:name "United States" .

mbart:36248428-08ff-4313-abe6-0ebbcaccb4f7 a schema:MusicGroup,
        schema:Person ;
    schema:name "John Flansburgh" .

mbart:b48f22c6-cab9-436c-a6d0-99839a19ee05 a schema:MusicGroup,
        schema:Person ;
    schema:name "John Linnell" .

mbrelease:b9daa8f6-2641-4e24-9a10-ce205cca1df3 a schema:MusicAlbum ;
    schema:albumProductionType "http://schema.org/StudioAlbum" ;
    schema:albumReleaseType "http://schema.org/AlbumRelease" ;
    schema:byArtist mbartist:183d6ef6-e161-47ff-9085-063c8b897e97 ;
    schema:creditedTo "They Might Be Giants" ;
    schema:name "I Like Fun" .

mbartist:183d6ef6-e161-47ff-9085-063c8b897e97 a schema:MusicGroup ;
    schema:foundingDate "1982"^^schema:Date ;
    schema:groupOrigin mbarea:11c4099a-ff61-45a3-ada4-23ac7a25d111 ;
    schema:member
       [ a schema:OrganizationRole ;
         schema:member mbartist:b48f22c6-cab9-436c-a6d0-99839a19ee05 ;
         schema:startDate "1982"^^schema:Date ],
       [ a schema:OrganizationRole ;
           schema:member mbartist:36248428-08ff-4313-abe6-0ebbcaccb4f7 ;
           schema:startDate "1982"^^schema:Date ] ;
    schema:name "They Might Be Giants" ;
    schema:album mbrelease:b9daa8f6-2641-4e24-9a10-ce205cca1df3 .
```

(a) The original that I downloaded was a JSON-LD file. What did I convert it into, and how do the formats relate?

[2 marks]

(b) Which ontology does this file use? (Note: only **ONE** answer is acceptable).

[1 mark]

(c) Many of the triples that are returned by the HTTP request to `https://musicbrainz.org/artist/183d6ef6-e161-47ff-9085-063c8b897e97` have that URL as their subject. Some have it as their object. In some triples, the requested URL does not occur at all. Give **ONE** example.

[1 mark]

(d) There is a bug in the MusicBrainz API which means that some of the information about the schema:MusicAlbum is exported badly. What is wrong? Why might this have happened?

[2 marks]

(e) One colleague looked at this code and claimed that the band called “They Might Be Giants” has two members, another disagreed, saying that it was impossible to know how many there were. Who is right and why?

[2 marks]

(f) If I took a copy of the knowledge graph from MusicBrainz and loaded it into a triplestore with a SPARQL interface, what query would return a list of all groups that were founded in the United States?

[4 marks]

(g) Since individuals (such as John Linnell) are given class schema:MusicGroup, how could you be sure your answer above only included real groups rather than musicians?

[2 marks]

(h) What query would list all albums made by bands of which John Linnell has been a member?

[4 marks]

(i) What reasons can you think of for MusicBrainz not having a public SPARQL endpoint for their Linked Data?

[2 marks]

(j) The MusicBrainz JSON-LD API is a Linked Data view of the internal data representation, which uses the relational model. Based on the data example above, propose a normalised set of tables, fields, and keys for storing the same data.

[10 marks]

END OF PAPER