COMP 60711

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Thursday 25th January 2018

Time: 09:45 - 11:45

Please answer ONE Question from Section A and ONE Question from Section B

Use a SEPARATE answer book for each SECTION

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

1.

a) Consider the data sample shown in the table below, which describes user actions when accessing the commercial Web site of a company. Each table row describes an action performed by a client of the company when accessing the Web site.

Suggest four data reduction techniques that could be used to avoid the storing of the thousands of actions collected daily by the company and that could facilitate analysis of the data as well as the gathering of Business Intelligence. State any assumptions you make about how the data is to be analysed.

(8 marks)

Date	Time	Username	IPaddress	URL	Action	Object
13-06-2017	14:22:56	sramos	2001:db8:0:1234:0:567:8:1	https://www.AMretailers.com	Login	
13-06-2017	14:23:20	sramos	2001:db8:0:1234:0:567:8:1	https://www.AMretailers.com/b/ref=...	Search	
13-06-2017	14:24:01	dverd	2010:db9:0:3492:0:534:6:2	https://www.AMretailers.com/b/ref=...	choose_product	prod_010
13-06-2017	14:27:12	sramos	2001:db8:0:1234:0:567:8:1	https://www.AMretailers.com/b/ref=...	choose_product_class	prod_class_30
13-06-2017	14:29:44	sramos	2001:db8:0:1234:0:567:8:1	https://www.AMretailers.com/b/ref=...	choose_product	prod_120
13-06-2017	14:27:12	dverd	2010:db9:0:3492:0:534:6:2	https://www.AMretailers.com/b/ref=...	add_to_basked	prod_010
13-06-2017	14:31:21	sramos	2001:db8:0:1234:0:567:8:1	https://www.AMretailers.com/b/ref=...	add_to_basked	prod_120
13-06-2017	14:33:50	sramos	2001:db8:0:1234:0:567:8:1	https://www.AMretailers.com/b/ref=...	pay_basked	

b) Data acquisition is one of the phases in the big data lifecycle where the employed data collection methods produce very large sets of data that can vary in format, speed, and quality (e.g., different types of sensors). In this context, discuss, at least, four issues that typically arise in data acquisition.

(8 marks)

c) Define data profiling and discuss four challenges involved in profiling very large data sets.

(9 marks)

d) Explain the steps involved in the process of cleaning data, discussing the issues that can arise at each step and the limitations of data cleaning as a process.

(3 marks)

e) Explain why and how integration of Big data from heterogeneous sources can impact on the quality of the integrated data and make data cleaning more difficult.

(2 marks)

2.

a) Apply the general lifecycle management principles of *Identification of Objectives in Maintaining Data*, *Minimalism* and *Information Security* in the context of the Amazon.com, Inc. (the largest Internet-based retailer in the world). This will require from you the proposal of specific management procedures for Amazon's data sets. Consider, for example, its commercial website, which collects user behaviour data on a daily basis. You should also make assumptions about other datasets relevant to the Amazon's businesses. Your procedures should dictate how the torrent of data (for example data describing user actions when accessing the Amazon Web site) should be managed by Amazon, as well as other relevant datasets. State any assumptions you make.

(10 marks)

b) Considering parallel processing as a solution to the problem of performance of Big data applications for data profiling, answer the following questions: (i) Explain how the *Weighted Average* of a numerical column in a table can be calculated in a distributed manner (i.e., if the column values are distributed across a number of nodes of a cluster, and the overall *Weighted Average* of the values is to be calculated). (ii) Explain how the *Median* of the values in same column can be calculated in a distributed manner.

(6 marks)

c) Recent technological advances (e.g., network and communication infrastructures, service oriented programming paradigms, etc.) have had an impact on the organization and functionality of modern Information Systems, changing their architecture into network-based structures. In this context, explain how these changes have had an impact on the quality of the data that runs through the processes of these systems, and discuss how data quality can be assessed and improved in these systems.

(4 marks)

d) Describe Data Lakes and discuss how Data Lake architectures can help integrating data from different sources.

(4 marks)

(Question 2 continues on the following page)

(Question 2 continues from the previous page)

e) Consider the following relational database schemas that store information about hotels, room bookings and guests, and which need to be integrated. Aiming at reconciling the conflicts that need to be resolved before integration is possible, answer the following questions. Use the notation *Schema_i.TableName.AttributeName*, to refer to an attribute in any of the two schemas; *Schema_i.TableName*, to refer to a table in any of the two schemas.

Schema_1:

```
Hotel(hotelNo, hotelName, address)
Room(roomNo, hotelNo, type, price, features)
Booking(hotelNo, guestNo, dateFrom, dateTo, roomNo)
HolidayGuest(guestNo, guestName, guestAddress, packageNo)
BusinessGuest(guestNo, guestName, guestAddress)
Package(packageNo, packageName, Company, discount)
```

Schema_2:

```
Hotel(hotelNo, hotelName, classification, streetNo, streetName,
postcode, city, country)
HotelFacilities(hotelNo, facilityNo, numOfUnits)
Facility(facilityNo, facilityName, description)
Bedroom(roomNo, hotelNo, type, additionals, price)
Reservation(roomNo, hotelNo, guestNo, dateFrom, dateTo,
discountGuestType)
Guest(guestNo, guestName, guestAddress, type)
```

i. Describe two different one-to-one table structure conflicts, one of them being the case of a missing, but implicit attribute.

(2 marks)

ii. Describe one table inclusion conflict.

(1 marks)

iii. Produce an SQL view that derives a table with the following structure from the Schema_2 database, where *durationOfStay* can be obtained by simply subtracting attribute *dateFrom* from attribute *dateTo*, discount can be null for some guests, and *totalToPay* can be obtained using *durationOfStay*, *roomPrice* and *discount*.

```
HotelBill(guestNo, roomPrice, durationOfStay,
                                    discount, totalToPay)
```

(3 marks)

Section B

3.

a) In the context of big data and business intelligence:

(i) Characterise the 5 Vs of big data.

(2 marks)

(ii) Consider the differences between OLTP and OLAP? Why might different database systems be used for each type?

(2 marks)

(iii) Compare and contrast the use of row store and column store; in particular, consider the advantages of their use for data warehousing and OLTP. Illustrate your answer.

(4 marks)

b) In the context of classification:

i) Outline a decision tree classification algorithm

(4 marks)

ii) Explain the potential effects of unbalanced data on the usefulness of a classifier. Suggest the advantages of various methods to handle unbalanced data. Give examples to support your answer.

(4 marks)

c) In the context of association rule (itemset) mining:

The following contingency matrix shows a breakdown of transactions for coffee and tea drinkers (assume 1000 transactions).

	coffee	not coffee	Total
tea	150	50	200
not tea	650	150	800
Total	800	200	1000

(i) Calculate support, confidence and lift for the association rule *{tea}* -> *{coffee}*.

(2 marks)

(Question 3 continues on the following page)

(Question 3 continues from the previous page)

(ii) Explain why lift better represents the relationship between tea and coffee drinkers.

(2 marks)

(iii) Consider ways in which association rule (itemset) mining might be extended. Discuss in your answer adding multi-dimensionality, quantitative intervals and hierarchies to the standard approach.

(4 marks)

d) In the context of "the analytics landscape"

Discuss and contextualise the four purposes of analytics in the "analytics landscape" and their role in insight and decision making; in particular, consider the temporal and business value factors. Illustrate your answer.

(6 marks)

4.

a) In the context of association rule (itemset) mining:

(i) Discuss the terms correlation and causation.

(2 marks)

(i) Outline the working of the Apriori algorithm. Explain the importance of the subset property.

(5 marks)

(ii) Using Apriori, suppose that L3 is the list:
{ {a,b,c}, {a,b,d}, {a,c,d}, {b,c,d}, {b,c,w}, {b,c,x}, {p.q.r}, {p,q,s}, {p,q,t}, {p,r,s}, {q,r,s} }.

a. At the join step of the algorithm, which itemsets are placed in C4 (the candidate set)?

(3 marks)

b. Which itemsets are discarded by the prune step of the algorithm?

(3 marks)

Provide your working as appropriate.

b) In the context of classification:

(i) Two classifiers designed to predict patients' susceptibility to allergy are being designed and tested, independently of one another. Each of the classifiers predict that a patient is either positive (allergic) or negative (normal) based on a combination of observable factors. The tests result in the following two confusion matrices, one for each classifier:

Table A:

		predicted		
		allergic	normal	Total
actual	allergic	30	70	100
	normal	20	500	520
	Total	50	570	620

Table B:

		predicted		
		allergic	normal	Total
actual	allergic	70	30	100
	normal	200	320	520
	Total	270	350	620

(Question 4 continues on the following page)

(Question 4 continues from the previous page)

Calculate the accuracy, recall, and F-measure for each classifier based on these tables. Based on their values, can you recommend one classifier over the other, give this type of application? Justify your answer

(4 marks)

ii) Compare and contrast information gain, gain ratio and the GINI index.

(3 marks)

(iii) Some classification algorithms run out of memory in trying to fit all data in memory to create the classification model. Discuss ways in which you might address the issue of memory capacity?

(4 marks)

c) Discuss the trade-offs between privacy of the individual on the one hand and the utility of data on the other. Use well-known case studies of data disclosure (AOL, Netflix, Barabasi) as appropriate to illustrate these trade-offs. How do we better guarantee that data is private?

(6 marks)

END OF EXAMINATION