MACQUARIE University

This question paper must be returned. Candidates are not permitted to remove any part of it from the examination room.

SEAT NUMBER: …… ROOM:……

FAMILY NAME:……..……

OTHER NAMES:……..……..……..……..

STUDENT NUMBER:…..……..……..……..

FORMAL EXAMINATION PERIOD: SESSION 2, NOVEMBER 2018

Unit Code:	ITEC874
Unit Name:	Big Data Technologies
Duration of Exam (including reading time if applicable)**:**	3 hours plus 10 minutes reading time
Total No. of Questions:	8
Total No. of Pages (including this cover sheet)**:**	3

GENERAL INSTRUCTIONS TO STUDENTS:

- Students are required to follow directions given by the Final Examination Supervisor and must refrain from communicating in any way with another student once they have entered the final examination venue.
- Students may not write or mark the exam materials in any way during reading time.
- Students may only access authorised materials during this examination. A list of authorised material is available on this cover sheet.
- All watches must be removed and placed at the top of the exam desk and must remain there for the duration of the exam. All alarms, notifications and alerts must be switched off.
- Students are not permitted to leave the exam room during the first hour (excluding reading time) and during the last 15 minutes of the examination.
- If it is alleged you have breached these rules at any time during the examination, the matter may be reported to a University Discipline Committee for determination.

EXAMINATION INSTRUCTIONS:

Answer ALL questions.

This examination consists of two sections, A and B, each worth 25 marks. The answers to each section should be written in a separate answer book – one answer book per question. Write your name, student number and the section name (A or B) on the cover of each answer book.

Write your name and student number at the top of this page. You may do rough working on this question paper, but all answers MUST be submitted as described above. Hand in this question paper at the end of the examination.

AIDS AND MATERIALS PERMITTED/NOT PERMITTED:

Dictionaries: No dictionaries permitted

Calculators: No calculators permitted

Other: Closed book – No notes or textbooks permitted

SECTION A: 3 QUESTIONS, 25 MARKS

Write your answers to this section in a separate answer book and write "Section A" on the first page of the book.

1. (10 marks) Understanding Big Data:
   (a) (2 marks) Using a real-world scenario, explain in 50 words what the Big Data problem is.
   (b) (2 marks) Explain in 50 words what the problem of organizing big data is. Justify your answer using the Four V's (volume, variety, velocity and veracity) of Big Data.
   (c) (2 marks) Explain what a key-value database is and discuss how the data will be stored and queried in such databases.
   (d) (2 marks) Is a distributed algorithm more fault tolerant than a centralized algorithm? Provide an example to support your answer.
   (e) (2 marks) Compare Cluster, Grid and Cloud Computing.

2. (10 marks) Data and Knowledge Lake:
   (a) (2 marks) Explain what a Data Lake is. Name four main components of a Data Lake and explain one of them in 50 words.
   (b) (2 marks) Compare Data Lakes and Data Warehouses.
   (c) (2 marks) Explain what Data ingestion is. Name two Big Data Technologies/Systems that can be used for ingesting streaming data.
   (d) (2 marks) Explain what a Knowledge Lake is. Name four main components of a Knowledge Lake and explain the 'Data and Knowledge Extraction' component in 50 words.
   (e) (2 marks) Explain what Big Data Summarization is and how it can facilitate analyzing the Big Data. Name four techniques for summarizing the Big Data.

3. (5 marks) Processing Big Data
   (a) (1 mark) Explain why transparency is an important property that a distributed system designer should achieve.
   (b) (1 mark) Explain what Apache Hadoop is. Name four components of the Hadoop Ecosystem and explain the role of the Zookeeper component (in the Hadoop Ecosystem) in 50 words.
   (c) (3 marks) Assume that we have a set of Tweet documents, similar to the Tweet dataset presented in assignment 2. Provide the MapReduce algorithm pseudocode for calculating the count of number of occurrences of each word in the text of Tweets.

END OF SECTION A. PLEASE TAKE ANOTHER ANSWER BOOK.

SECTION B: 5 QUESTIONS, 25 MARKS

Write your answers to this section in a separate answer book and write "Section B" on the first page of the book.

1. (5 marks) Data Analytics
   (a) (2 marks) A financial institution asks you to develop a system that would predict the risk of granting a loan to a potential customer. They have a record of past requests of loans and whether the customer defaulted. Answer the following questions.
      1. (1 mark) Is this an example of descriptive, diagnostic, predictive, or prescriptive analytics? Justify your answer.
      2. (1 mark) Given the data that the financial institution has, Would you apply supervised machine learning or unsupervised machine learning? Justify your answer.
   (b) (3 marks) List the overall steps of a Data Mining project. Each step should be explained in one sentence or two.

2. (5 marks) Text Analytics
   (a) (3 marks) Enumerate and briefly explain three characteristics of text that make it particularly challenging for computer processing.
   (b) (2 marks) Explain what Named Entity recognition is and how it could be useful for text analytics.

3. (5 marks) Visual Analytics
   (a) (2 marks) Explain how one can use scatterplots to determine whether a variable can be useful as a predictor.
   (b) (3 marks) A company wishes to analyse the impact of a product recently introduced in the market. Explain three visual analytic techniques that could be useful for this study. For each visual analytic technique, make sure that you specify the visual analytic technique, the reason for its use, and the data source on which it would be applied.

4. (5 marks) Stream Processing
   (a) (2 marks) Explain the characteristics of stream processing with reference to the four V's of Big Data (volume, variety, velocity, and veracity).
   (b) (3 marks) Draw the Stream Model and explain all of its components.

5. (5 marks) Big Data and Society
   (a) (2 marks) Explain the privacy issues that arised when the Netflix Prize was introduced and that lead to a lawsuit.
   (b) (3 marks) Briefly Explain two of the six critical questions for Big Data presented by Danah Boyd and Kate Crawford in their 2012 paper "Critical Questions for Big Data".

END OF SECTION B. PLEASE REVISE YOUR ANSWERS.