COMP1702	**Big Data**	**Faculty Header ID**	**Contribution: 30% of course**
Course Leader: Hai Huang	**Coursework_part1**		**Deadline Date: 15 Mar 2021**
Feedback and grades are normally made available within 15 working days of the coursework deadline			
Learning Outcomes: 1 Explain the concept of Big Data and its importance in a modern economy 2 Explain the core architecture and algorithms underpinning big data processing 3 Analyse and visualize large data sets using a range of statistical and big data technologies 4 Critically evaluate, select and employ appropriate tools and technologies for the development of big data applications			

Plagiarism is presenting somebody else's work as your own. It includes copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as your own work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by the University. Please see your student handbook for further details of what is / isn't plagiarism.

All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.

Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.

Coursework Submission Requirements

- An electronic copy of your work for this coursework must be fully uploaded on **the Deadline Date of 15th Mar 2021** using the link on the coursework Moodle page for COMP1702.
- For this coursework **you must submit a single report in PDF format**. In general, any text in the document must not be an image (i.e. must not be scanned) and would normally be generated from other documents (e.g. MS Office using "Save As .. PDF"). An exception to this is handwritten mathematical notation, but when scanning do ensure the file size is not excessive.
- There are limits on the file size (see the relevant course Moodle page).
- Make sure that any files you upload are virus-free and not protected by a password or corrupted otherwise they will be treated as null submissions.
- Your work will not be printed in colour. Please ensure that any pages with colour are acceptable when printed in Black and White.
- You must NOT submit a paper copy of this coursework.
- All courseworks must be submitted as above. Under no circumstances can they be accepted by academic staff

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances,

and penalties for Assessment Offences. See http://www2.gre.ac.uk/current-students/regs

1. Detailed Specification

You are expected to work individually and complete a report that addresses the following tasks. Note: You need to cite all sources you rely on with in-text style. You may include material discussed in the lectures or labs, but additional credit will be given for independent research. References should be in Harvard format.

1. Explain the main features of Big Data.(200 words ±10%) [**10 Marks**]

2. Compare Hadoop and Relational Database system. Give an application scenario that is well suited to Hadoop and explain the reason. (300 words ±10%) [**30 Marks**]

3. Suppose that we have a large housing data file which cannot be stored in a single machine. Each record of this file contains information about a single house: (address, city, state, zip, price). The task is to output the average house price in each zip code.

   Describe how you would solve it using **MapReduce**. You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s)

and value(s) are computed. Then you should explain how the (key, value) pairs produced by the map stage are processed by the reduce stage to get the final answer(s). (300 words ±10%) [**30 Marks**]

4. Compare relational and NoSQL databases and give an example of an application or dataset particularly suited to NoSQL databases. Justify your choice (300 words ±10%) [**30 Marks**]

Grading Criteria

For a distinction (mark 70-79) the following is required:

1. All tasks are completed well.
2. An excellent/very good research demonstrating a very good/ excellent understanding of big data concepts and techniques.

For a mark in the range 60 to 69 the following are required:

1. Most of tasks are completed well.
2. A good research demonstrating a good understanding of big data concepts and techniques.

For a mark in the range 50 to 59 the following are required:

1. Satisfactory answers to the majority of the tasks.
2. A satisfactory research showing some understanding of big data concepts and techniques.

For a mark below 50:

1. A very few or no requirements are met.
2. A poor research showing little understanding of the big data concepts and techniques.