COMP1702	**Big Data**	**Faculty Header ID**	**Contribution: 70% of course**
Course Leader: Hai Huang	**Coursework_part2**		**Deadline Date: 29 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 29th 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

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. References should be in Harvard format. You may include material discussed in the lectures or labs, but additional credit will be given for independent research.

PART A: MapReduce Programming [300 words ±10% excluding java codes] (30 marks)

There is a text file ("papers.txt" is uploaded in Moodle) about computer science bibliography. Each line of this file describes the details of one paper in the following format: *Authors|Title|Conference|Year*. The different fields are separated by the "|"character, and the list of authors are separated by commas (","). An example line is given below:

D Zhang, J Wang, D Cai, J Lu|Self-Taught Hashing for Fast Similarity Search|SIGIR|2010

You can assume that there are no duplicate records, and each distinct author or conference as a different name.

TASK: Please write a MapReduce program to calculate for each conference the number of papers.

Requirement: Please include your MapReduce pseudo code (or real Java code) and the detailed description of your algorithm including main functions (and their functionalities), execution steps, performance analysis in the report.

PART B: Big Data Project Analysis [2000 words ±10% excluding references]
(70 marks)

Precision agriculture (PA) is the science of improving crop yields and assisting management decisions using high technology sensor and analysis tools. The AgrBIG company is a leading provider of agronomy services, technology and strategic advice. They plan to develop a big data system. The users can be farmers, research laboratories, policy makers, public administration, consulting or logistic companies, etc. The system aims to help worldwide users better understanding the implications of the weather and making contingency plans; buying supplies, such as fertilizer and seeds; as well as maintaining and monitoring the quality of yield, whether livestock or crops; knowing the variety of cultivated plants, conditions of its growth and its needs of seeds; choosing the type of fertilizer and pesticides, understanding their employment conditions and their impact on the climate- soil-plant; recognizing daily water needs for each kind of plant; calculating the median and mean values of yield; studying the conditions of natural environment; estimating the financial revenue and manage the potential risks.

The sources of data will be from various Sensors, Satellites, Drones, Social media, Market data, Online news feed, Logistic Corporate data, etc..

Example Data	
Streaming	Soil Sensor data, Streaming Agricultural product prices (from Market data vendors)
Batch	Historical Weather Data
On-Line:	Online news feed
Unstructured:	Satellite Image Data, Drone Video Data, Text

Note that some data such as price and customer data are confidential. The data volume is expected to be 300 Petabytes. The applications need to be highly available, scalable and accessible from worldwide.

You need to design a big data project by solving the following tasks for the AgrBIG company:

Task1 (25 marks): Produce a Big Data Architecture for the AgrBIG company with the following components in detail:

- Data sources,
- Data extraction and cleaning,
- Data storage,
- Batch processing,
- Real time message ingestion,
- Analytical Data store

For each of the above, discuss various options and produce your recommendation which best meets the business requirement.

Task2 (10 marks): The AgrBIG company needs to store a large collection of plants, corps, diseases, symptoms, pests and their relationships. They also want to facilitate queries such as: "find all corn diseases which are directly or indirectly caused by Zinc deficiency". Please recommend a data store for that purpose and justify your choice.

Task3 (10 marks): MapReduce has become the standard for performing batch processing on big data analysis tasks. However, data analysists and researchers in the AgrBIG company found that MapReduce coding can be quite challenging to them for data analysis tasks. Please recommend an alternative way for those people who are more familiar with SQL language to do the data analysis tasks or business intelligence tasks on big data and justify your recommendation.

Task 4 (15 marks): The AgrBIG company needs near real time performance for some services such as soil moisture prediction service. It has been suggested the parallel distributed processing on a cluster should use MapReduce to process this requirement. Provide a detailed assessment of whether MapReduce is optimal to meet this requirement and If not, what would be the best approach.

Task 5 (10 marks): Design a detailed hosting strategy for this Big Data project and how this will meet the scalability, high availability requirements for this global business.

Assessment criteria

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

1. An excellent/very good implementation of the coding task, all components are working and provide a very good result.
2. An excellent/very good research demonstrating a very good/ excellent understanding of big data concepts and techniques.

Note: In order to be eligible for a very high mark (**80 and over**) you will need to have:
1. An exceptional implementation of the coding task, showing all requirements implemented to a higher standard.
2. An exceptional research, demonstrating a thorough understanding of the big data concepts and techniques.

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

1. A good implementation of the coding task, with components are working and providing a good result.
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. A satisfactory implementation of the coding task with majority of the components are working.

2. A satisfactory research showing some understanding of big data concepts and techniques.

For a mark below 50:

1. Very few requirements of the tasks are met.
2. A poor or little understanding of big data concepts and techniques.