Course Outline

COMP9313 - Big Data Management

Course Summary

This course introduces the core concepts and technologies involved in managing Big Data. It will first introduce the characteristics of big data and big data analysis. Then, we will present key management aspects in the context of big data projects. Next, we will learn the open-source big data management framework Hadoop. The course will introduce the overall framework and then focus on Hadoop MapReduce and its programming model. In this course, we will also introduce Spark, an open-source and memory-based distributed computing framework. In addition, we will also introduce examples of major NoSQL technologies currently widely used in the big data management ecosystem.

Course Aims and Learning Outcomes

This course aims to introduce students to the concepts behind Big Data, the core technologies used in managing large-scale data sets, and a range of technologies for developing solutions to large-scale data analytics problems.
This course is intended for students who want to understand modern large-scale data analytics systems. It covers a range of topics and technologies, and will prepare students to be able to build such systems as well as use them effectively to address challenges in big data management.
On successfully completing this course, students should be able to:

- Describe the important characteristics of Big Data
- Understand key concerns in the management of Big Data
- Develop an appropriate storage structure for a Big Data repository
- Utilise the Map/Reduce paradigm and the Spark platform to manipulate Big Data
- Use a high-level query language to manipulate Big Data

Pre-requisites and Assumed Knowledge

Prerequisite of this course include COMP9024 and COMP9311. Before commencing this course, students should:

- Have experiences and good knowledge of algorithm design (equivalent to COMP9024)
- Have a solid background in database systems (equivalent to COMP9311)
- Have solid programming skills in Java
- Be familiar with working on a Unix-style operating system
- Be familiar with working with web services (e.g., RESTful services)

Teaching Rationale and Strategies

The course involves lectures/tutorials and practical work. Lectures/Tutorials aim to summarize the concepts and present applications of Big Data. Labs and assignments aim to reinforce the topics covered in lectures.
The teaching strategies include:

- Lectures: Introduce concepts and show examples
- Lab Work: Consultation and activities in the context of the provided assignments
- Tutorials: Discussions on lecture materials and practical examples of Big Data technologies
- Assignments: An important part of the course. They allow students to apply the techniques introduced in the course to real problems.
- Online Quizzes: Revision of the concepts introduced in lectures/tutorials.
- Consultation: Weekly consultation to provide personalized advice to students on their progress in the course.

Course Schedule

The precise schedule is subject to change as the term progresses.

Week	Lecture	Labs
1	**Course Information and Introduction to Big Data**	
2	**Big Data Processes and Management**	
3	**Hadoop Part1**	
4	**Hadoop Part2**	**Lab 1:** *Hadoop*
5	**Hadoop Part3**	**Lab 2:** *Hadoop*
6	**Spark Part1**	**Lab 3:** *Spark*
7	**Spark Part2**	**Lab 4:** *Spark*
8	**NoSQL Technologies for Big data**	**Lab 5:** *Elasticsearch*
9	**NoSQL Technologies for Big data part 2**	**Lab 6:** *Elasticsearch*
10	**Course wrap-up & Review**	

Assessment

The assessment will have the following components:

- Quizzes (15%): This component will help review the concepts introduced in lectures/tutorials.
- Assignments (45%): This component assesses the student's ability to apply big data technologies to solve problems.
- Written Final Exam (40%): This component is going to assess the various facts-and-knowledge level learning outcomes

* Weights for assessment acitivties are subject to change and will be communicated accordingly
** the time of releasing assessment activity specifications depends on progress and feedback

Resources for Students

There is no prescribed textbook for this course. Yet, the following resources are relevant for the topics we will cover in this course:

- Hadoop: The Definitive Guide . Tom White. 4th Edition - O'Reilly Media
- Learning Spark. Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. O'Reilly Media
- Apache MapReduce Tutorial
- Apache Spark Quick Start
- Elasticsearch Reference
- Datasets: Assignments will make use of a number of datasets from different domains (e.g., cybersecurity)

Course Evaluation and Development

This course is evaluated each session using the standard UNSW course evaluation system.

Special Consideration

You can apply for special consideration when illness or other circumstances beyond your control, interfere with your assessment performance. For further details on special consideration, see the UNSW Student website

Student Conduct

The **Student Code of Conduct** (Information , Policy) sets out what the University expects from students as members of the UNSW community. As well as the learning, teaching and research environment, the University aims to provide an environment that enables students to achieve their full potential and to provide an experience consistent with the University's values and guiding principles. A condition of enrolment is that students *inform themselves* of the University's rules and policies affecting them, and conduct themselves accordingly.

In particular, students have the responsibility to observe standards of equity and respect in dealing with every member of the University community. This applies to all activities on UNSW premises and all external activities related to study and research. This includes behaviour in person as well as behaviour on social media, for example Facebook groups set up for the purpose of discussing UNSW courses or course work. Behaviour that is considered in breach of the Student Code Policy as discriminatory, sexually inappropriate, bullying, harassing, invading another's privacy or causing any person to fear for their personal safety, is serious misconduct and can lead to severe penalties, including suspension or exclusion from UNSW.
If you have any concerns, you may raise them with your lecturer, or approach the School Ethics Officer , Grievance Officer , or one of the student representatives.

Plagiarism is defined as using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several on-line sources to help you understand what plagiarism is and how it is dealt with at UNSW:

- Plagiarism and Academic Integrity
- UNSW Plagiarism Procedure

Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW. Plagiarism at UNSW is defined as using the words or ideas of others and passing them off as your own.

If you haven't done so yet, please take the time to read the full text of

- UNSW's policy regarding academic honesty and plagiarism

The pages below describe the policies and procedures in more detail:

- Student Code Policy
- Student Misconduct Procedure
- Plagiarism Policy Statement
- Plagiarism Procedure

You should also read the following page which describes your rights and responsibilities in the CSE context:

- Essential Advice for CSE Students