Course Outline

Contents

- Course Details
- Course Summary
- Course Timetable
- Course Aims
- Student Learning Outcomes
- Assumed Knowledge
- Teaching Rationale
- Teaching Strategies
- Assessment
- Course Schedule
- Resources for Students
- Student Conduct
- Course Evaluation and Development

Course Details

Course Code	COMP9313
Course Title	Big Data Management
Units of Credit	6
Course Website	https://webcms3.cse.unsw.edu.au/COMP9313/18s2/ (https://webcms3.cse.unsw.edu.au/COMP9313/18s2/)
Handbook Entry	http://www.handbook.unsw.edu.au/postgraduate/courses/current/COMP9313.html (http://www.handbook.unsw.edu.au/postgraduate/courses/current/COMP9313.html)

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 learn the open-source big data management framework Hadoop. We will mainly focus on the learning of Hadoop MapReduce, and HDFS, HBase, and Hive will be briefly introduced. We will also learn an open-source memory-based distributed computing framework Spark. Another major focus of this course is algorithm design on large-scale data sets based on the big data management frameworks, in a variety of domains such as link analysis, data stream mining, graph data processing, and recommender systems.

Course Timetable

Lecture Time (3 hours):

- **Wednesday 6pm to 9pm** , in Ritchie Theatre (K-G19-LG02)

Consultation Time (1 hour):

- **Wednesday 3pm to 4pm** , in K17-201D

The complete course timetable is available here (/COMP9313/18s2/timetable) .

Course Aims

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 wide range of topics and technologies, and will prepare students to be able to build such systems as well as use them efficiently and effectively address challenges in big data management.

Student Learning Outcomes

After completing this course, students will be able to:

- describe the important characteristics 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
- develop efficient solutions for analytical problems involving Big Data

Assumed Knowledge

Official prerequisite of this course is 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 (VERY IMPORTANT!!!)
- have basic knowledge of linear algebra (e.g., vector spaces, matrix multiplication), probability theory and statistics , and graph theory

Teaching Rationale

The course involves lectures and practical work. Lectures aim to summarize the concepts and present case studies. The lab exercises aim to reinforce the topics covered in lectures (without assessment), while the assignments and projects aim to do the same (but are assessed). The course will have an emphasis on problem solving for large-scale data sets.

Teaching Strategies

- Lectures: the main way to introduce concepts, and will explain with detailed examples and include exercises, demonstrations and live coding examples.
- Lab Work: lab exercises to practice the knowledge on big data management learned in lectures.
- Projects and Assignments: a very important part of the course. They allow students to apply the techniques introduced in the course to real problems.
- Consultation: weekly consultation to provide personalized advice to students on their progress in the course.

Assessment

Number	Name	Full Mark
1**	Coding Project 1	25
2**	Coding Project 2	25
3**	Coding Project 3	40
4*	Assignment	10
5	Final Exam	100

Later Submission Penalties:
* : zero marks
** : 10% reduction of your marks for the 1st day, 30% reduction/day for the following days

The final mark is calculated by the harmonic mean:
Final Mark= 2 * (proj1 + proj2 + proj3 + ass1) * FinalExam / (proj1 + proj2 + proj3 + ass1 + FinalExam)

You also need to achieve at least 40 marks in the final exam to pass the course.

Course Schedule

The order that topics are covered in this course is probably not the best order for presenting them. The material is presented in lectures in an order that ensures that you are best prepared for the assignments.

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

Week	Lecturer
1	Course information + introduction to big data

2	Hadoop MapReduce 1
3	Hadoop MapReduce 2
4	Hadoop MapReduce 3 (proj1)
5	Spark 1
6	Spark 2
7	Data stream mining (proj2)
8	Spark 3
9	Finding Similar Items
10	NoSQL, HBase, and Hive (proj3)
11	Recommender Systems
12	Revision and exam preparation (ass1)

Resources for Students

The textbooks include:

- Hadoop: The Definitive Guide (http://hadoopbook.com/) . Tom White. 4th Edition - O'Reilly Media
- Data-Intensive Text Processing with MapReduce (https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf) . Jimmy Lin and Chris Dyer. University of Maryland, College Park.
- Mining of Massive Datasets (http://www.mmds.org/) (http://www.mmds.org/) . Jure Leskovec, Anand Rajaraman, Jeff Ullman . 2nd edition - Cambridge University Press
- Learning Spark (http://shop.oreilly.com/product/0636920028512.do) . Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. O'Reilly Media

Other references include:

- Apache MapReduce Tutorial (https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html)
- Apache Spark Quick Start (http://spark.apache.org/docs/latest/quick-start.html)

Student Conduct

The **Student Code of Conduct** (Information (https://student.unsw.edu.au/conduct) , Policy (https://www.gs.unsw.edu.au/policy/documents/studentcodepolicy.pdf)) 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 (mailto:ethics-officer@cse.unsw.edu.au) , Grievance Officer (mailto:grievance-officer@cse.unsw.edu.au) , or one of the student representatives.

Plagiarism is defined as (https://student.unsw.edu.au/plagiarism) 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 (https://student.unsw.edu.au/plagiarism)
- UNSW Plagiarism Procedure (https://www.gs.unsw.edu.au/policy/documents/plagiarismprocedure.pdf)

Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible 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. 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 (https://student.unsw.edu.au/plagiarism)

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

- Student Code Policy (https://www.gs.unsw.edu.au/policy/documents/studentcodepolicy.pdf)
- Student Misconduct Procedure (https://www.gs.unsw.edu.au/policy/documents/studentmisconductprocedures.pdf)
- Plagiarism Policy Statement (https://www.gs.unsw.edu.au/policy/documents/plagiarismpolicy.pdf)
- Plagiarism Procedure (https://www.gs.unsw.edu.au/policy/documents/plagiarismprocedure.pdf)

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

- Essential Advice for CSE Students (https://www.engineering.unsw.edu.au/computer-science-engineering/about-us/organisational-structure/student-services/policies/essential-advice-for-cse-students)

Course Evaluation and Development

This course is evaluated each session using the myExperience system. Students are also encouraged to provide informal feedback during the session and to let the lecturer in charge know of any problems as soon as they arise. Suggestions will be listened to very openly, positively, constructively and thankfully, and every reasonable effort will be made to address them. Your feedback is important and will be considered seriously. Student feedback via the myExperience system will enable improvements to future offerings of this subject.

Student feedback from last offerings indicated that students were satisfied with the course. The students suggest that more examples could be provided to help understand the theoretical concepts. We will endeavor to achieve this in this offering.

Resource created 3 years ago (Wednesday 18 July 2018, 02:24:06 AM), last modified 3 years ago (Wednesday 17 October 2018, 04:36:40 PM).

Comments

(/COMP9313/18s2/forums/search?forum_choice=resource/17640) (/COMP9313/18s2/forums/resource/17640)

Arthur Pegios (/users/z5194172) 3 years ago (Mon Aug 13 2018 12:57:05 GMT+1000 (AEST))

Hi, I can't seem to access the projects. When will they be made available?

Xiaohan Peng (/users/z5106189) 3 years ago (Sat Aug 04 2018 08:31:37 GMT+1000 (AEST))

Hi,

May I ask about the approximate due dates of these projects?

Tianle Qiu (/users/z5174807) 3 years ago (Fri Jul 27 2018 01:11:43 GMT+1000 (AEST))

I was wondering whether it is possible to add more stuff about nitty-gritty details of how hadoop/spark handling those typical challenges like fail-safe, coordination between different nodes, parallelism, etc. For how to write applications on top of hadoop and spark, one could find plenty of examples and tutorials from the Internet. These techniques, however, require much more effort for one to gain some ideas and it would be really helpful to have someone guide us through.

Xin Cao (/users/z3515164) 3 years ago (Fri Jul 27 2018 03:15:51 GMT+1000 (AEST))

Some system level mechanism of mapreduce and spark will be briefly explained. However, due to the time limit, we will still focus on the high level algorithm design.