BUSINESS SCHOOL
Te Kura Pakihi

INFO 408

Management of Large-Scale Data

Second Semester, 2021

COURSE OUTLINE

Contents

1 TL;DR **1**

2 Paper description and aims **2**

3 Learning outcomes **2**

4 Teaching staff **2**

5 Delivery **3**
5.1 On campus . . . 3
5.2 Online . . . 3
5.3 Learning resources . . . 3
5.4 COVID-19 contingency plans . . . 4

6 Expectations and workload **4**

7 Assessment **5**
7.1 Assessment submission requirements . . . 5
7.2 Essay . . . 6
7.3 Online participation . . . 6
7.4 Assessed labs . . . 6
7.5 Project . . . 6
7.6 Final examination . . . 6
7.7 Final grade and passing the paper . . . 6
7.8 Special consideration for internal assessment . . . 7

8 Further information **7**

9 Disclaimer **7**

1 TL;DR

(aka all the most critical information)

Dr. Nigel Stanger (coordinator & lecturer)
nigel.stanger@otago.ac.nz
OBS 704

Prof. Stephen Cranefield (lecturer)
stephen.cranefield@otago.ac.nz
OBS 717

What	When	Where
Lecture (on campus)	Wednesday 1pm–3pm*	OBS 227
Lab (on campus)	Friday 2pm–4pm	OBS 327
Lecture/lab discussion (online)	Monday 6pm–8pm*	Microsoft Teams
Participation (5%)	throughout semester	
Assessed labs (20%)	weeks 3–11 (4% each)	
Essay (10%)	Monday 16 August at 5pm	
Project (20%)	Monday 4 October at 5pm	
Final Exam (45%)	TBA	TBA

* We may not always need to use the full time allocated. All times are in New Zealand time.

Labs start in the *first* week.

TO PASS INFO 408 YOU MUST

achieve a total overall mark of at least 50%.

Continue reading for the full, gory details! ☺

2 Paper description and aims

The modern world is awash in a sea of data. Improvements in storage, compute, and networking infrastructure have made it possible to store, process, and transport ever-increasing amounts of data, while the rise of data-driven companies like Google and Facebook has driven rapid developments in data management software. Today, the world's largest databases are highly-scalable distributed systems that hold hundreds of petabytes of data or more. In INFO 408 we will study concepts and issues associated with the management and use of such databases.

There are no formal prerequisites for INFO 408, but prior knowledge is assumed, particularly regarding database systems and programming. Ideally, you will have passed an introductory database paper that covered relational databases and SQL, and have some general programming skills (language is not important). If you do not have this background, please contact the teaching staff as soon as possible to discuss how to bring your knowledge up to a sufficient level.

3 Learning outcomes

You will become proficient in both theoretical and practical aspects relating to the management of large-scale databases (loosely referred to as "big data"). In particular, those who successfully complete INFO 408 should be able to:

1. discuss ethical implications and issues raised by big data, and how these may be addressed;
2. compare and contrast different software and infrastructural architectures that can be applied to big data management, and choose an appropriate architecture for a given problem;
3. compare and contrast different implementation approaches for big data systems;
4. design and build a large, scalable database, and use it to answer complex queries;
5. identify and deal with architectural and implementation performance bottlenecks in big data environments.

4 Teaching staff

Paper coordinator & lecturer: Dr. Nigel Stanger
nigel.stanger@otago.ac.nz
OBS 704
Dr. Stanger will coordinate the paper and present about two thirds of the classes.

Lecturer: Prof. Stephen Cranefield
stephen.cranefield@otago.ac.nz
OBS 717
Prof. Cranefield will present about a third of the classes.

You are welcome to contact teaching staff about anything relating to INFO 408. We generally have an "open door" policy: if the office door is open, we are probably available; if closed, probably not. Scheduled office hours will also be posted on Blackboard and our office doors. We will make our best effort to be available during these times, but this may not always be possible due to other commitments. You are welcome to email us to ask questions or to arrange a meeting.

5 Delivery

All teaching material for INFO 408 will be made available through Blackboard, Otago Capture, or Microsoft Teams. Blackboard is the primary entry point for the paper, so make sure you check Blackboard regularly for announcements, updates, and corrections.

The planned teaching schedule is at the end of this document. The schedule is not fixed and may be subject to change during the semester. The latest version can always be found on Blackboard. Any major changes will be made in full consultation with the class. All material is examinable unless otherwise stated.

5.1 On campus

There will be one lecture each week: Wednesday 1pm–3pm in OBS 227, starting on **Wednesday 8th July**. Lectures will normally involve an interactive discussion of readings and/or videos made available in advance on Blackboard. **Make sure that you prepare for the lecture by reviewing any assigned material.**

There is also one laboratory class each week: Friday 2pm–4pm in OBS 327. Laboratory classes give you the opportunity to carry out practical work related to the lecture material, and start on **Friday 10th July**. The exercises will be released on Blackboard early each week. Nine of the thirteen laboratory exercises will be assessed (see Section 7.4).

Both sessions are timetabled for two hours, but we may not always need to make full use of the allocated time. You are also welcome to join the online follow-up session (see Section 5.2).

5.2 Online

The on campus lecture (see Section 5.1) will be recorded, and the recording and any associated notes will be made available within 24 hours. **Make sure that you prepare first by by reviewing any assigned material** that is published in advance on Blackboard.

Lab exercises will be released on Blackboard early each week and should be worked on in your own time. Nine of the thirteen laboratory exercises will be assessed (see Section 7.4).

A weekly follow-up discussion session has been scheduled: Monday 6pm–8pm (New Zealand time) in Microsoft Teams. This is timetabled for two hours, but we may not always need to make full use of the allocated time. We are always available via email.

5.3 Learning resources

INFO 408 uses free, open source software that runs on all the major platforms (Linux, macOS, Windows). For those on campus, these tools are available through the Linux desktop running

in labs OBS 327 and North CAL. Most are also available via the Student Desktop. All software can also be installed on your own personal computer.

There is no prescribed textbook for INFO 408, but we will make extensive use of various online resources as necessary. You are expected to make appropriate use of these and any other resources you might find to broaden your knowledge. You can document any relevant resources that you find in the INFO 408 wikis on Microsoft Teams. You may also be provided with copies of additional relevant material where necessary.

5.4 COVID-19 contingency plans

Hopefully the COVID-19 alert level will not change, but we must be prepared. As a precaution, ensure that you have installed all relevant software on your personal computer so that you can continue to work online if necessary. We have put in place the following contingency plans:

- All lectures will be recorded, and we will switch to fully online lectures (via Zoom or Teams) at Alert Level 3 or 4. On campus lectures may move to a different room under Alert Level 2 in order to provide physical distancing.
- On campus labs will move fully online under Alert Level 3 or 4 (see Section 5.2). On campus labs may move to North CAL at Alert Level 2 in order to provide physical distancing.
- Ad hoc Zoom or Teams sessions can be organised as necessary.

We will provide further details and instructions should any of these plans need to be activated.

6 Expectations and workload

What you can expect of us:

- We will keep you informed of important developments by means of announcements both in class and on Blackboard. Critical notices will also be sent to your student email address.
- We will answer queries within a reasonable timeframe. Typically this will be the same day for urgent queries, and within 48 hours for non-urgent queries. **Closeness of an assessment deadline does not automatically make queries urgent.** Please include your name and student ID in all correspondence in order to facilitate any necessary follow-up.
- We will maintain regular contact with class reps, and will consult the class before changing the content or structure of the assessment schedule.
- We will mark internal assessments within a reasonable timeframe, typically within two to three weeks of the submission deadline, depending on other commitments.
- We will make our best effort to be available in our offices during scheduled office hours, depending on other commitments. We will make alternative arrangements where necessary, and operate an "open door" policy at other times (see Section 4).

What we expect of you:

- You will read the Course Outline. ☺
- You will keep up to date with announcements and regularly check your student email.

- You are familiar with the relevant prerequisite material. While we will provide links to revision materials, it is your responsibility to make effective use of these.
- You will prepare for, attend (on time), and actively participate in (or complete) as many of the lecture and lab classes as you can.
- You will seek out relevant supplementary material. Independent learning is normal in industry. Simply reading or watching videos about a topic is not enough to truly understand it, however; you will be most successful when you can *apply* your knowledge.
- You will manage your time effectively. This includes completing assessments on time, and working on lab and assessment tasks in your own time outside scheduled classes.
- You will submit all assessments on time (see also Section 7.8), and in the correct format (see Section 7.1). Assessment deadlines have already been set (see Section 7); work submitted late without prior arrangement will not be accepted.

The expected workload for INFO 408 is about 200 hours of work per student for the whole semester, or about fifteen hours per week on average. A rough model of how this could be broken down is as follows (see also Section 5):

Contact hours*	Lectures	13×1.5	$\approx$ 20 hours
	Labs/follow-up	13×1.5	$\approx$ 20 hours
Assessment	Essay		$\approx$ 15 hours
	Project		$\approx$ 40 hours
	Online participation	13×1	$=$ 13 hours
	Final exam	1×2	$=$ 2 hours
Personal study	Lectures	13×1.5	$\approx$ 20 hours
	Lab exercises	13×1.5	$\approx$ 20 hours
	Final exam	$\approx$ 1 week	$\approx$ 15 hours
	Additional study		$\approx$ 35 hours
TOTAL:			$\approx$ 200 hours

* Either face-to-face or online, as appropriate.

7 Assessment

7.1 Assessment submission requirements

All internal assessment is to be submitted online through Blackboard unless otherwise specified. **Late submissions will not be accepted without prior arrangement** (see Section 7.8).

A high standard of presentation is expected in all your submitted work. All submitted documents must include both your name and student ID, and must be either plain text (`.txt`), Markdown, or PDF—**no Word or OpenOffice documents!** Use only commonly available fonts. Multiple files may be combined in a Zip or 7-Zip archive file. These requirements have been put in place to ensure that submitted work can be opened and marked successfully across multiple platforms. Work that fails to meet these requirements runs the risk of being unidentifiable or unreadable, and will not be marked until it is re-submitted in an acceptable form.

Use APA style[1] for referencing where applicable. Wikipedia can be a useful starting point for finding relevant literature on a topic, but should *never* be used as a primary source in submitted work. Inappropriate referencing will be penalised.

7.2 Essay

You will write an essay exploring ethical issues associated with big data. The essay counts towards **10%** of your final grade, and is due **Monday 16 August at 5pm**. Further details will be provided early in the semester.

7.3 Online participation

You are expected to actively participate in topic wikis on Microsoft Teams. Online participation counts towards **5%** of your final grade. Details of the assessment criteria will be provided early in the semester, but timeliness of contributions will be a factor. (You are also expected to actively participate in class, but this will not be assessed.)

7.4 Assessed labs

Lab exercises 3–11 will be assessed (see the teaching schedule at the end of this document). The **best five** of these nine labs count towards **20%** of your final grade (that is, 4% each). You will normally submit your work before midnight on the Thursday of the week after the exercise is released. Details of the deadline and assessment criteria will be provided with each assessed lab.

7.5 Project

You will undertake a project to implement a database to manage a large, real-world dataset. The project counts towards **20%** of your final grade, and is due **Monday 4 October at 5pm**. Further details will be provided later in the semester.

7.6 Final examination

There will be a two-hour final examination that counts for **45%** of your final grade. Details of the date, time, and venue will be provided later in the semester. All material covered in classes is examinable unless otherwise stated. INFO 408 has no terms requirement, so you are automatically eligible to sit the final examination.

7.7 Final grade and passing the paper

TO PASS INFO 408 YOU MUST

achieve a total overall mark of at least 50%.

[1] http://otago.libguides.com/citation_styles/APA

Your final grade for INFO 408 will be calculated as follows:

Assessment	%	Notes	Learning outcome(s)
Essay	10%	(due Monday 16 August at 5pm)	1
Participation	5%		1, 2, 3, 4, 5
Assessed labs	20%	(9 labs, 4% each for best *five*)	1, 2, 3, 4, 5
Project	20%	(due Monday 4 October at 5pm)	2, 3, 4, 5
Final examination	45%		1, 2, 3, 5
TOTAL	**100%**		

You can check your internal assessment marks online through Blackboard. It is *essential* that you verify all your internal assessment marks when they are posted and promptly notify the teaching staff of any errors or omissions.

7.8 Special consideration for internal assessment

If you feel unwell, please stay at home.

Send us an email as soon as you can so that we can organise alternative arrangements for labs and/or internal assessments.

Please submit special consideration requests for internal assessment to Dr. Stanger. Should you be unable to complete an internal assessment component for medical or personal reasons, you should contact Dr. Stanger *as early as possible* so that alternative arrangements can be made. **Contact Ask Otago if you wish to apply for special consideration for the final examination.**

If you have a disability, please let us know how we can help. We are happy to offer whatever assistance we can, but need to know in advance of any potential difficulties that might arise.

8 Further information

Refer to the **Information** section on Blackboard for additional general information that is not specific to INFO 408, such as student support services, academic integrity, and student feedback.

9 Disclaimer

While every effort is made to ensure that the information contained in this document is accurate, it is subject to change. Changes will be notified in class and via Blackboard, and this document will be updated as required. The latest version can always be found on Blackboard, so you should check Blackboard regularly for updates. It is your responsibility to be informed.

INFO 408 Schedule, Second Semester 2021

as at 20th July 2021

Week	Starting	Lecture	Lab	Assessment
1	12 July	Introduction: What is "big data"?	Introduction to lab environment	
2	19 July	Examples of big data (student research)	Database and SQL refresher	
3	26 July	Ethics of big data	The variety problem †	
4	2 August	Big data management landscape	Automating ETL pipelines †	
5	9 Aug	Big data integration: lakes, meshes, … (TBC)	TBC †	
6	16 Aug	NoSQL DBMSs 1: key-value, document	Document DBMSs: MongoDB †	**Essay (10%)** *Mon 16 Aug @ 5pm*
7	23 Aug	NoSQL DBMSs 2: columns, graph, …	Graph DBMSs: OrientDB †	
Mid-Semester Break (20 August–3 September)				
8	6 Sep	Trade-offs in big data management	NoSQL trade-offs (MongoDB) †	
9	13 Sep	Performance bottlenecks	Performance: querying †	
10	20 Sep	Scalability of DBMSs	Performance: data loading †	
11	27 Sep	Processing big data	Apache Spark (1) †	
12	4 Oct	NewSQL DBMSs	Apache Spark (2)	**Project (20%)** *Mon 4 Oct @ 5pm*
13	11 Oct	Wrap-up	Streaming data	

Note: The contents of this schedule may be subject to change during the semester. The first week of semester is **academic week 28**.

Second semester exam period runs from Wednesday 20th October to Saturday 13th November.

† Assessed lab exercise.