DATA3404: Scalable Data Management

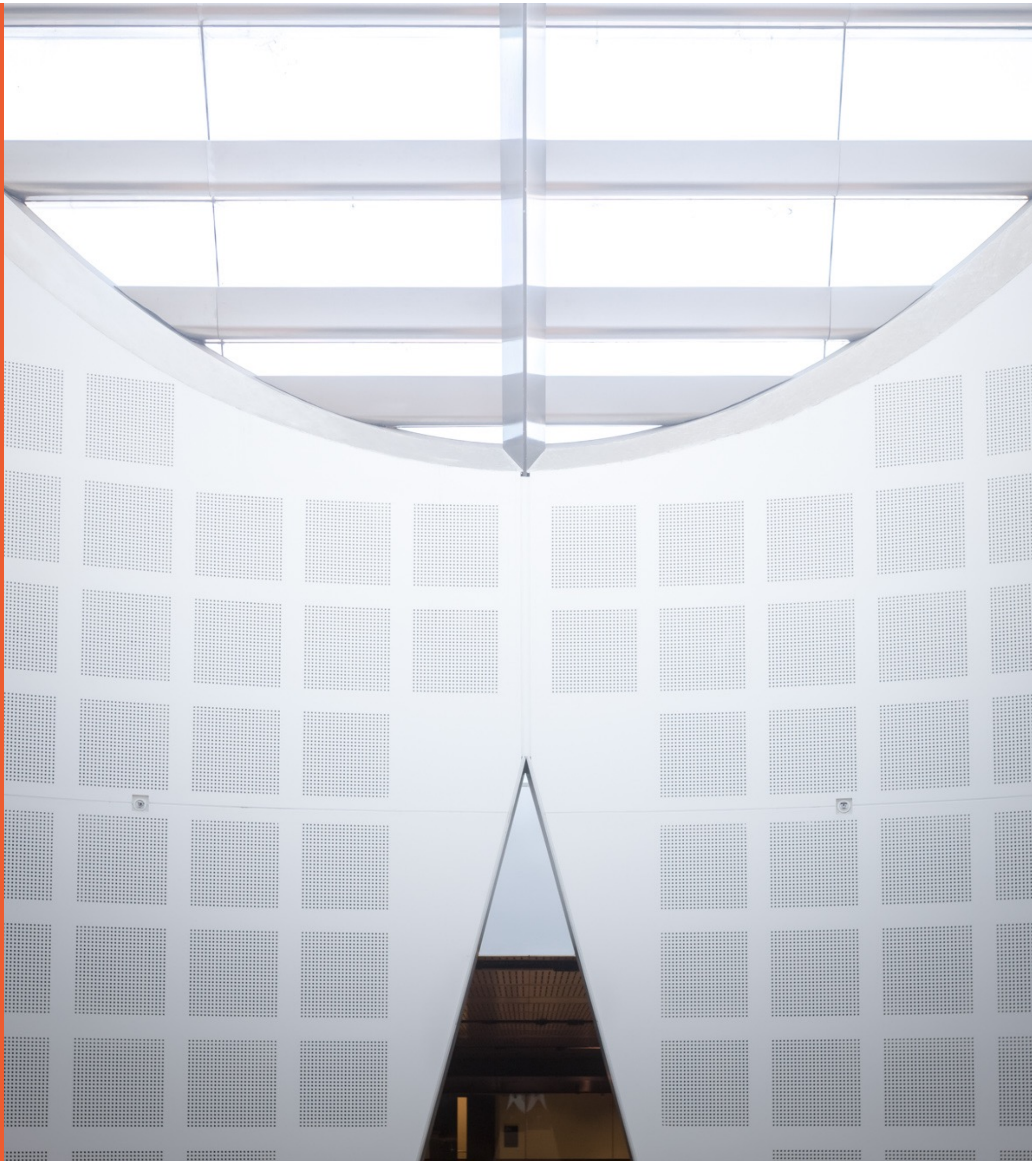
W13: Unit of Study Review

Presented by

A/Prof Uwe Roehm

School of Computer Science

THE UNIVERSITY OF SYDNEY



Outline

- General Summary
- Progressive Work Feedback
- UoS Evaluation
- General UoS Review
- Examination Tips

DATA3404 Grant Theme

- How to efficiently provide SCALABLE data processing?
  - **Large collections of data** (nowadays: terabytes to petabytes)
    - both structured (tuples)
    - and unstructured (text or (key,value) pairs)
    - we are interested on cases where data does not fit into memory….
  - Shared access by large numbers of concurrent users (thousands)
  - Availability – always ON
- Questions:
  - How to efficiently manage large amounts of data?
  - How to efficiently find data in those collections?
  - How to efficiently scale computation on those datasets in a cluster?

Course Objectives

- Main objective: Learn techniques for large-scale data management

- Understand the internals of both database engines and distributed data science platforms
  - storage system and disk-based indexing
  - query execution and optimisation
  - distributing data and computation

- Ability to effective use and tune data processing platform
  - Optimise given database / dataset

- Understanding of distributed data management and parallel processing

Outline of the Semester

	Week	Topic
Storage Layer	Week 1	Introduction & Organisation
	Week 2	Storage Engines / Physical Data Organisation
	Week 3	Disk-based Index Structures: Tree-based Indices
	Week 4	Disk-based Index Structures: Hash and Bitmap Indices
	Week 5	Distributed Data Management
Querying	Week 6	Introduction to Query Processing and External Sorting
	Week 7	Query Execution and Join Algorithms
	Week 8	Query Optimization
	Easter Break	
Big Data Scaling	Week 9	Distributed Querying and Data Processing
	Week 10	Dataflow Platform Optimisation
	Week 11	Data Stream Management Systems
	Week 12	NoSQL
	Week 13	Unit of Study Review

Assessment Package

Component	DATA3404
Weekly Tutorial Participation	5%
Weekly Homework Quizzes	10%
DB Concept Presentation	5%
Assignments (weeks 8 + 13)	20%
Final Exam	60%

- All progressive marks will be consolidated within https://canvas.sydney.edu.au
  - Marks for weekly quizzes are now published (cf. 'Marks' in the sidebar)
  - Report any errors/omissions within 10 days
- A pass requires at least:

  a) >= 40% in exam marks; and

  b) >= 50% overall mark.

Weekly Tutorial Participation

- participation mark for either
  - participation in weekly tutorial (Week 2 until Week 13), and/or
  - submission of finished weekly worksheet from lecture (unmarked)

- Participation Week 12:
  Participation in Group Demo (either in Week 12 or 13)
- Participation Week 13:
  Participation in Peer Reviewing of DB Concept Videos

Weekly Review Quizzes and DB Concept Presentations

- Last Weekly Quiz (Wk11+Wk12) ended yesterday
  - All marks will be available in Canvas this week

- Video Presentations => your input is needed
  - Watch the videos of your peers in your tutorial in Canvas
    - People – *search your own name* - Enrolments - "Tutorial TUT *X*"
      the go to the "Video Presentations Homepage" of your tutorial class
  - Decide which 3 you liked most
  - Enter top-3 (ranked) in Canvas:
    Modules – Week 13 - **Video Presentation Peer Vote**
    - Due: Monday next week (2 June)

Big Data Scalability Assignment 1 and 2 (10 + 10%)

- Practical assignments using PostgreSQL and Databricks/Apache Spark
  - several analysis tasks over Inside AirBnB dataset
  - private install vs. Databricks Community Edition

- Assignment 2 is due this Friday (30 May)
  - Submission page and marking rubric in Canvas
  - Thanks for some of the teams demoing already last week!
  - Only one member per team needs to submit for the whole group;
    she should submit both a ZIP archive under "Assignment 2 Code" and also the PDF of your report in the separate "TurnItIn Dropbox – Assignment 2 Report"
  - Late submissions: -5% of achieved mark per day late

Final Examination

Objective

Assess understanding of unit material; understanding of core data management principles, algorithms, and scalability issues; (e.g., indexing, execution and optimization, distributed "Big Data" processing).

Content

- Questions about all lecture and tutorial material
- Examples and more details see next few slides

Format

- **In-person, written exam** (on campus)
  - scheduled: **Thu 12 June, 1 pm (AEST)**
  - 2 hours 10 min duration
  - Various exam rooms -> check timetable most: Barneys Broadway Main Auditorium
- restricted open-book
  1. allowed: 1 page own notes
  2. allowed: calculator - non-programmable
- 60% of final mark

SIT policy: You must get 40% on the exam and 50% overall to pass DATA3404

More exam details after the break…

Exam Overview (cont'd)

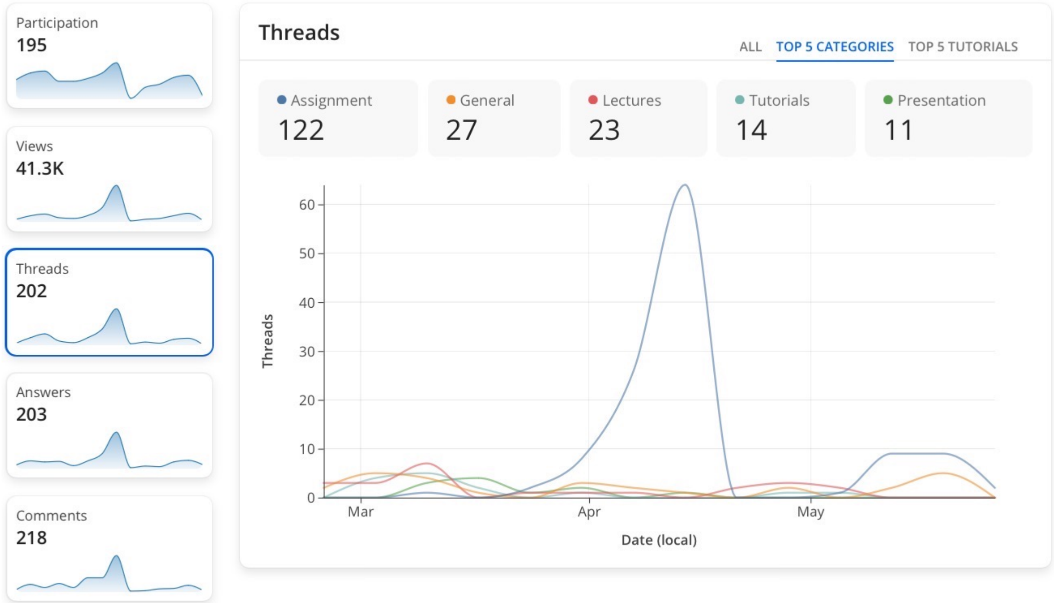
- There will be five to six main topics covered.

- Combination of different types of questions.
  - **Short & Long answer questions, some calculations too**
  - Some Multiple Choice; full marks only for correct selection(s), deducing marks (per question) for wrong choices (min: 0)
  - Typically with increasing level of difficulty.
  - Follows the style of the lecture worksheets, tutorial homework exercises, and to some extend the weekly review quizzes – but note: more short-answer questions than MC questions…

- The exam will have a total of 100 marks
- You need to get a **at least 40/100 in the final exam** to pass

Discussion Forum Participation (2025)

- We used Ed as discussion forum this year; seemed to have worked fine
- 202 questions, over 420 answers and comments



Participation 195

Views 41.3K

Threads 202

Answers 203

Comments 218

Threads

● Assignment 122

● General 27

● Lectures 23

● Tutorials 14

● Presentation 11

and still 1 week to go until + exam preparation

Top askers / commenters:

- actually most were anonymous – why?

Huge shout-out to the tutors (Cabiria, Cody, Danny, Linh and Yan) for promptly answering your requests.

Without their support, this lecture would not have been possible.

Shout-out to the Teaching Team

- Lecturer: Uwe Roehm

- Tutors:
  - Cabiria Liang
  - Cody Hu
  - Danny Chacko
  - Lan Linh Nguyen
  - Yan Rong

Feedback?

- Feedback & suggestions needed
  - USS survey or upcoming “feedback” thread in Ed
- Particularly interested in:
  - Tutorials and in-person lectures
  - In-semester assessment tasks incl. video presentations
  - Coverage of distributed data science platform in assignment
- We care, so please be gentle
- USS survey
  - https://student-surveys.sydney.edu.au/students/

How to make your USS Fedback count

Your Unit of Study Survey (USS) feedback is **confidential.**

It's a way to share what you enjoyed and found most useful in your learning, and to provide constructive feedback. It's also a way to 'pay it forward' for the students coming behind you, so that their **learning experience** in this class is s good, or even better, than your own.

When you complete your USS survey (https://student-surveys.sydney.edu.au), please:

Be specific.
Which class tasks, assessments or other activities helped you to learn? Why were they helpful? Which one(s) *didn't* help you to learn? Why didn't they work for you?

Be constructive.
What practical changes can you suggest to class tasks, assessments or other activities, to help the next class learn better?

Be relevant.
Imagine you are the teacher. What sort of feedback would you find most useful to help make your teaching more effective?

Exercise: Questions and Suggestions

- Please take 10 min now to complete the survey
  - Browse to https://student-surveys.sydney.edu.au/students/
  - Log in if you aren't already
  - Complete survey for DATA3404

Content Review

THE UNIVERSITY OF SYDNEY

The Key Principles of Data Platforms

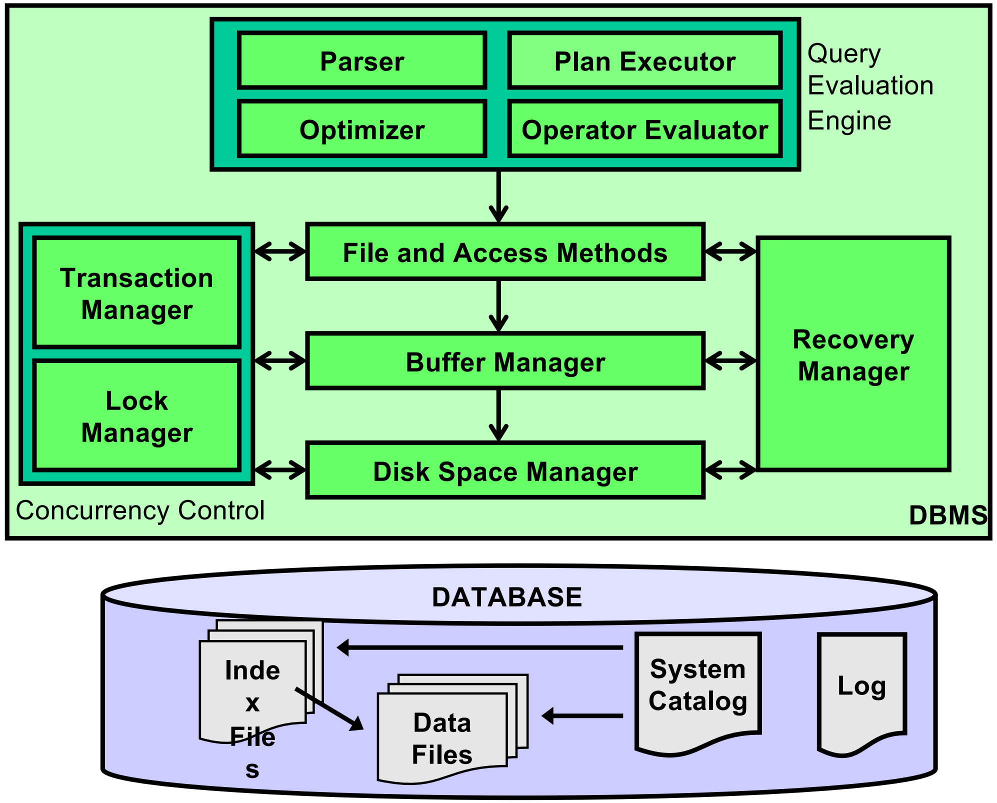
- Data Independence
  - Application programs decoupled from the physical data layout
  - You can optimise the physical schema without changing the application

- High-Level, declarative interface
  - SQL (or data transformation operators) allows you to specify "*what* rather than *how*."
  - Unfortunately, this means 'fighting the optimiser' in some cases

- Space can be reused but not time
  - Indexes can massively speed-up lookups and joins,
  - Just be careful to not hamper updates too much…

- But sometimes adding indexes isn't enough

DBMS Architecture



Review: Storage Layer

- Important Aspects:
  - **DBMS Storage Hierarchy**
  - Buffer Management
    - E.g. **buffer replacement**/page eviction algorithms, pinning of pages
  - Disk Storage Organisation
    - E.g. page layout, record structures, file structures
    - **Row Stores vs. Column Stores**

- Typical exam questions:
  - What is the role of a buffer manager in a DBMS?
  - Estimate/calculate the minimum and maximum storage costs for a given schema.
  - Differences between row and column stores? When use which?

Review: Indexing

- Important aspects:
  - **B+ Tree**
  - Static and Dynamic **Hashing**
  - **Bitmap Indexes**
  - **LSM Tree** (Wk 12)
  - Index Classification
  - **Query Tuning using Indexing**
- Typical exam questions:
  - When are indexes good, when are they bad? Role for querying?
  - Explain the differences / access costs of B+-Tree and Hash indices.
  - Given a database schema and a workload specification, suggest a set of suitable indexes to improve the performance of the system.

Review: Query Execution

- Important Aspects:
  - Query processing steps
    - **Pipelining** vs. **materialization**
  - Relational algebra expression and query execution plans
  - Physical operator algorithms:
    - **Join algorithms, External sorting, …**

- Typical exam questions:
  - When can we use pipelining and when materialization in query processing?
  - How do logical and physical query operations relate?
  - What role does sorting play for different query execution algorithms?
  - Compare the costs of block-NLJ, merge-join and hash join for a certain scenario.
  - Determine the number of runs for a n-way sort-merge over *X* tuples.

Review: Query Optimization

- Important Aspects:
  - Basic query optimisation steps
  - Heuristic query optimizations
    (algebraic query transformation, equivalent RA expressions)
  - **Cost-based query optimization**
    - though some systems do it differently.; cf. MongoDB’s unique FPTP approach
  - Role of database statistics

- Typical exam questions:
  - Briefly explain cost-based query optimization. What is the goal?
  - What role do database statistics play in query optimization?
    How do they affect, e.g., join orders?
  - Draw the best (left-deep) query plan for the following SQL query. Why left-deep?
  - Give an SQL query, find a good query execution plan for it. Explain your choices.

Review: Distributed Data Management

- Important Aspects:
  - Distributed system architectures, **CAP Theorem**
  - Data **replication**, data **partitioning** / **sharding**
  - Distributed query processing; **distributed join algorithms**

- Typical exam questions:
  - What is the meaning of the CAP theorem?
  - Suggest a partitioning strategy for a given scenario
  - Difference between primary copy and update anywhere replication?
  - How is a distributed join algorithm executed for a given data set?

Review: Big Data Processing

- Important Aspects:
  - Scale-Agnostic Computation: **MapReduce Principle**
  - Distributed Data Processing Frameworks; example: **Apache Spark**
  - lazy evaluation in Apache Spark
  - **Data Stream Processing**; notions of time; window processing; examples: **Kafka** and **Apache Flink**

- Typical exam questions:
  - When would you use a DBMS, when MapReduce, when Spark or Flink?
  - Given a Spark program, in which tasks or stages will it be executed?
  - Role of lazy evaluation in Spark.
  - What is the difference between batch and stream processing?

Review: NoSQL

- Important Aspects:
  - **NoSQL** background and classification
  - **Key-Value Stores**; data model; querying; example: **RocksDB & Dynamo**
  - **(Distributed) Column Stores**; data model; querying; example: **HBASE**
  - **Document Stores**; data model; querying; example: **MongoDB**

- Yes, this was late in the semester – but note it is still examinable!
- Typical exam questions:
  - What is a key-value store?
  - What is the difference between Amazon Dynamo, HBASE and MongoDB? What is the difference between a relational DBMS like PostgreSQL and a document store like MongoDB?
  - How does the NoSQL system *X* implement the CAP theorem?

Exam Overview

THE UNIVERSITY OF
SYDNEY

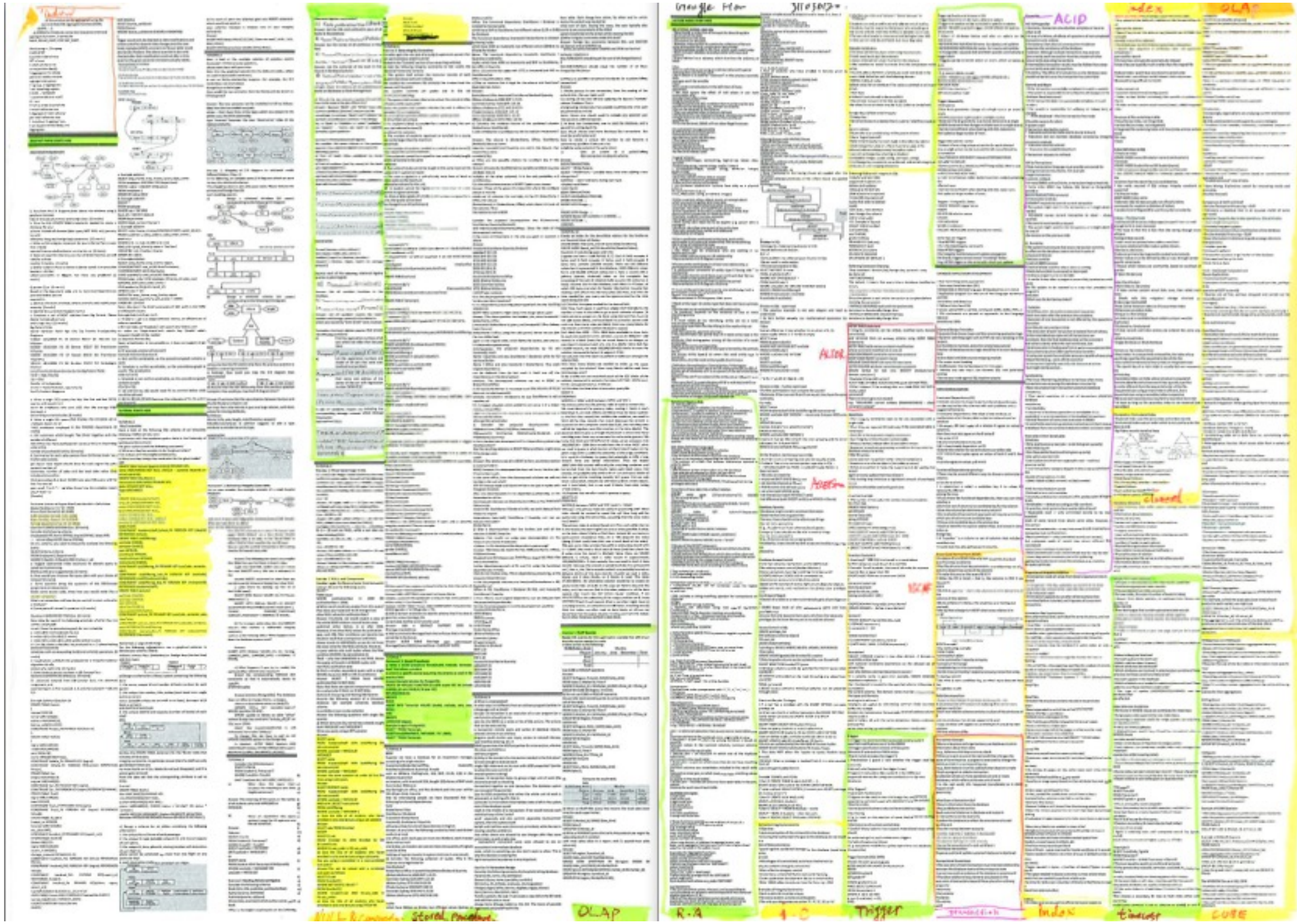
Exam Overview

Thursday, 12 June, 13:00 – 15:10 (early afternoon)

Venue: Barneys Broadway – Main Auditorium (check your timetable)
(St Barnabas Broadway church at corner of Broadway and Mountain Street
https://www.barneys.org.au/venue-hire/)

- Restricted open book exam: handwritten notes, printed notes
  - One A4 sheet of paper of own notes (hand-written or typed; double-sided)
  - Approved, **non-programmable calculators** are allowed too
- Two-hour, supervised in-person examination
  - Read through the in-person exam site for advice and preparation tips.
    - cf. https://www.sydney.edu.au/students/exams/in-person.html
  - You will need to bring one form of valid photo identification to your exam
    - **Don't forget to bring your University student card!**

Examples of Exam Notes


stored procedure
OLAP
R-A
Trigger
index
cube
ACID
OLAP

Exam Overview (cont'd)

- There will be five to six main topics covered.

- Combination of different types of questions.
  - **Short & Long answer questions, some calculations too**
  - Multiple choice; full marks only for correct selection(s),
    deducing marks (per question) for wrong choices (min: 0)
  - Typically with increasing level of difficulty.
  - Follows the style of the lecture worksheets, tutorial homework exercises, and to some extend the weekly quizzes – but note: more long-answer questions than MC questions…

- The exam will have a total of 100 marks
- You need to get a **at least 40/100 in the final exam** to pass

MC Question Marking Example

What is 7 x 6? You may choose more than one answer. **[2 points]**

□ 40

□ 35

□ 42

□ 49

□ The Answer to the Ultimate Question of Life, the Universe, and Everything.

42 => 1 point

49 => 0 points

42 & 49 => 1-1 = 0 points

42 & The Answer to Everything => 2 points

DON'T
PANIC

Exam Preparation

- Check that you have covered the whole lecture
  - Important topics have been pointed out
  - Also revisit the review quizzes (but exam will have more short answer questions)

- The weekly lecture worksheets, the tutorial exercises, the db concept videos, the practical assignments, as well as the weekly quizzes should give you a good idea what the exam will be like
  - Note: Tutorial homework question and understanding of concepts are more important in an exam than programming code.
  - There was a previous exam (sub-)question in almost every week's lecture or tutorial worksheet.
  - Try as many exercises as possible and check against the example solutions.
  - You should be able to understand the example solutions.

Exam Techniques

- You will get an exam script with exam questions which you can answer in the order of your liking
  - make sure that you answer **all** questions

- Questions will be a mix of multiple choice, matching, calculation and short answer questions
  - The latter have to be answered in the answer boxes provided in the exam script
  - Whenever there is a tip on how to format your answer, please follow this
    - E.g. to label answer parts to correspond to different question parts

Exam Techniques (cont’d)

- In the short-answer questions, check for “Justify your choice”, “Briefly explain”, “Describe”, “Why?”, “Discuss” or “Give an example” parts.
  - Such questions test your *understanding* of an area.
    - A simple yes/no is not enough!
  - Please answer BRIEFLY
    (one or two sentences are typically OK, but NOT a whole page)
    - E.g. if you have to compare two techniques, a good approach is to first define the techniques in one sentence each, before actually comparing them in more detail (and don’t forget that last part ;)

  - Say it in your own words, don’t just copy from the textbook.
  - Please write complete, English sentences!
    - You want the marker to understand what you wanted to say…

Further Database Units

- **4th Year:** COMP5349: Cloud Computing
  - Much more on cloud infrastructure, cloud storage and cloud compute services
  - Microservices and serverless architectures
  - Containerization and automation; Cloud Security

- **4th Year:** COMP5338 "Advanced Data Models"
  - (post-relational databases): ORDBMS
  - Spatial and Temporal Databases and Index Structures
  - Graph Databases

- **Year 4:** COMP5318 "Machine Learning and Data Mining"
  - Machine Learning and Data Mining
  - Classification and Clustering Algorithms; Linear Regression; Decision Trees
  - (Deep) Neural Networks; Reinforcement Learning

- **Further:** COMP5328 – Advanced Machine Learning and COMP5329 – Deep Learning

Research Opportunities

- Research Areas:
  - Distributed data systems, queries and transactions
  - Machine Learning for self-tuning databases ; machine learning support in DBMS
  - Graph Databases
  - Key-Value Stores on modern hardware

- USYD Database Research Group (DBRG)
  - Uwe Roehm
  - Alan Fekete
  - Michael Cahill
  - Lijun Chang
  - Ying Zhou

Final Activities

- Tutorial Week 13: Assignment demos
- Peer-Vote for the DB Concept Video Presentation in your lab by Monday 2nd June
  - Will provide example exam for students doing this peer reviewing
- Housekeeping (before exam):
  - Review your marks and assignment results on Canvas
  - **Read through the Taking in-person exams site**
- Feedback:
  - Submit USS survey (**https://student-surveys.sydney.edu.au/students**/)
- Revision:
  - Review lecture material and cross-check with topics highlighted in Week 13
  - Review/attempt all worksheets/tutorial exercises and check the example solutions
  - Try reading the suggested texts again
- If needed: Arrange consultation sessions
  - alternatively: Use ED STEM to ask questions

- **All the best for your assignment submission and in your exam**