Lecture 0: Course Logistics

CSCI4180 (Fall 2022)

Introduction to Cloud Computing and Storage

About Me

- Patrick P. C. Lee (http://www.cse.cuhk.edu.hk/~pclee)
  - B. Eng. in IE, CUHK, 2001
  - M. Phil. in CSE, CUHK, 2003
  - Ph. D. in Computer Science, Columbia, 2008
- Research interests:
  - Applied/systems topics on improving the dependability of large-scale software systems, including storage systems, distributed systems and networks, and cloud computing.
  - Focus on system prototyping and implementation

Course Information

- Course website:
  - http://www.cse.cuhk.edu.hk/~pclee/csci4180

- TAs (all in SHB 118)
  - Ren, Yanjing, yjren22@cse (Assignment 1)
  - Puyang, Huancheng, hcpuyang22@cse (Assignment 2)
  - Zhao, Jia, jzhao@cse (Assignment 3)
  - Tang, Kaicheng, kctang@cse (Group and VM Management)

- Tutorials start on Sep 14

Course Prerequisites

- CSCI3150, or equivalent
  - If you are now taking CSCI3150 this semester, we also let you enroll, but keep in mind that we assume you know the key concepts in operating systems (e.g., parallel computing and file systems)
    - Send me email, with a proof that you're taking CSCI3150 this term
- Comfortable with Java programming
  - We will provide crash courses in tutorials
- Comfortable with Linux

Course Newsgroup

- Piazza:
  - https://piazza.com/configure-classes/fall2022/csci4180
    - Access code: fa224180
  - I will make announcements in class and Piazza

- Please post your assignment questions to Piazza
  - Avoid sending email to the instructor or TAs about your assignment questions
  - Your questions are also the questions from your classmates

Course Materials

- Reference books:
  - Jimmy Lin and Chris Dyer, **"Data-Intensive Text Processing with MapReduce"**, Morgan and Claypool, 2010.
  - Tom White, **"Hadoop: The Definitive Guide"**, Fourth Edition, O'Reilly Media, 2015
- Free online copies are on CUHK library website
- Notes and reference materials are posted on Blackboard (https://blackboard.cuhk.edu.hk)
- ***It's important to read!!***

Course Assessment

- 3 programming assignments (50%)
  - Group assignments of **1-2** people
  - Assignment 1: Counting by MapReduce
  - Assignment 2: Iterative MapReduce
  - Assignment 3: Deduplicated Storage
- Final exam (50%)
  - Face-to-face (default); online for special cases

Assignment Policies

- On late submissions:
  - Finish your assignments and submit them before the deadline
  - You can submit multiple times; the last one will be graded
  - We allow late submissions until the start of the demo time (which is likely on the next day), with a penalty of 20 marks
  - No late submissions are allowed after the demo starts
- On demo
  - No code change is allowed during the demo
  - TAs may allow *slight* modifications of the code (e.g., changing 1-2 lines of code), with a penalty of 10 marks for *each* change
  - TAs will have the final call to decide if code changes are allowed
- Please refer to the course website for details

Exam Policies

- 2-hour written exam
- Cover all lecture notes, tutorials, assignments
  - The scope may be narrowed down in the last lecture
- **Two A4 cheat-sheets (both sides) are allowed**
  - You can write anything that you think is important
- Approved calculators allowed
  - No notebooks, smartphones
- Make assumptions if needed, and provide justifications

Make-up Exam Policy

- You must inform the instructor within 12 hours after the original exam time **and** file a formal request to RES
- If you take a make-up exam, **your final exam score will be automatically decremented by 30%.** I hope you can understand this policy, since we need to be fair to the students who work hard to prepare for the regular exam.

Amazon EC2

- Trial on real production cloud: **Amazon EC2**

- CUHK has an agreement with Amazon. Students receive US$100 credit for the use of AWS services (renewable annually)
  - https://www.itsc.cuhk.edu.hk/all-it/teaching-learning-and-research/aws-educate/

- Credit card tie-in not required

Windows Azure

- https://azure.microsoft.com/en-us/free/students/
  - USD 100 free credit for 1 year upon activation with university-associated email
  - NOT bounded to personal credit card
  - Free tier services available after credit consumed

Video Recordings

- Lectures will be video-recorded.
- Videos posted on Blackboard
  - https://blackboard.cuhk.edu.hk/
- It's meant to help you review the lectures. It's NOT meant to let you skip lectures.

Academic Honesty

- In short, don't cheat!
- Don't copy code or solutions from your classmates or third-party sources, and don't let others copy yours.
- Cases will be reported to the school
- Details:
  - CUHK: http://www.cuhk.edu.hk/policy/academichonesty/
  - Faculty of engineering: http://www.cse.cuhk.edu.hk/v5/other/A5_BookletN3.pdf
- Ask me if you are unsure

- ***Don't*** *post your assignments online, even after the end of semester*

Course Objectives

- **Goals**:
  - Understand the essentials of cloud computing and cloud storage
  - Learn the applied methodologies of using cloud computing and cloud storage for solving practical engineering problems

10 Questions

- What is (and is not) cloud computing?
- How does Google analyze and store BIG data in a scalable, reliable way?
- How do we write elegant programs for BIG data processing (like Google)?
- How do we solve BIG graph problems with cloud computing?
- How does Google make data searchable?
- How does Dropbox make profit?
- How does Amazon achieve highly available storage at scale?
- How does Facebook manage your photos?
- How does Yahoo! coordinate thousands of machines?
- How do Google and NetApp deal with tails?

Centered around two main areas: ***Computing*** *and* ***Storage***

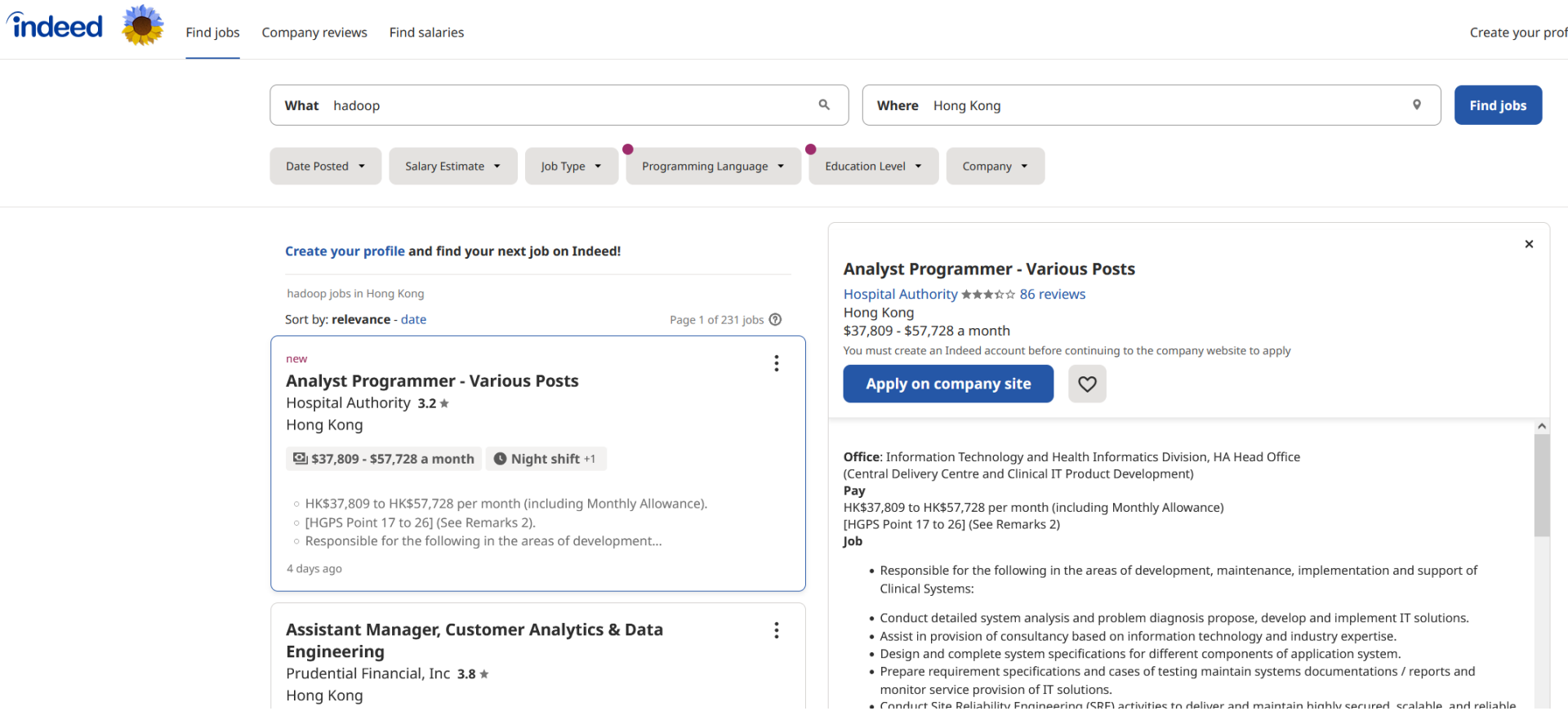
Topics to Cover

- Fundamentals of Cloud Computing
- Big data computation
  - Overview of Hadoop: MapReduce and HDFS
  - MapReduce Programming
  - MapReduce Algorithm Design
  - MapReduce Applications (e.g., PageRank)
  - Spark
- Hadoop data management
  - BigTable, HBase
  - Zookeeper

Topics to Cover

- Cloud storage:
  - Deduplication
  - Dropbox
  - Amazon's Dynamo
  - Facebook's Haystack and f4
- Tail latency issues in clouds:
  - Google and NetApp's case studies
- Containerization and serverless computing

Hadoop Jobs in HK?



Extracted on July 28, 2022

Student/Faculty Expectations

- Goal: to enhance teaching and learning qualities

- http://www.erg.cuhk.edu.hk/Student-Faculty-Expectations