COMP5349 – Cloud Computing

Week 13: Course Review

A/Prof Uwe Roehm
School of Information Technologies

COMP5349 Schedule in 2017

Week	Topic
Week 1	Cloud Computing Overview and Service Models
Week 2	Data Centres and Virtualization Technology
Week 3	Distributed File Systems: GFS and HDFS
Week 4	The Map/Reduce Paradigm and Hadoop
Week 5	Map/Reduce Design Patterns and YARN Resource Management
Week 6	Data Analytical Extensions for MapReduce
14 April – 23 April Easter Break	
Week 7	Dataflow Engines for Big Data Analysis: Spark and Flink
Week 8	Programming with Apache Spark and Flink
Week 9	Programming with Apache Spark and Flink, Part II
Week 10	Cloud-based NoSQL Databases
Week 11	Data Consistency in Cloud Computing
Week 12	Benchmarking Cloud Services
Week 13	Unit of Study Review

The Big Picture

- Cloud Computing
  - Shared IT services
  - On different levels (IaaS, PaaS, SaaS)
  - Made possible through data center technology
- **Big Data Analysis**
  - **Computation Model**
  - **Data Analytics**
  - **Data Storage**
- Benchmarking

Reprise: Cloud Service Models

- Three main models
  - IaaS, PaaS and SaaS
- Virtualization and Multitenancy
  - Virtualization is the basic technology for IaaS
  - Multitenancy is the basic technology for PaaS and SaaS
    - It refers to the architecture/principle that allows a single instance of a software/application to serve multiple client organizations (tenants).
    - It is different to multi-user system as the **performance isolation** should be guaranteed at the application level
    - There is no standard multitenancy solution equivalent to “virtual machine” in virtualization area
    - Industries and academic researchers are still experimenting various resource reservation and performance isolation techniques

Reprise: Big Data Computational Model

- Basic Computational Model
  - Storage: distributed file systems (GFS, HDFS)
  - Programming Paradigm: Map/Reduce
  - Hadoop MapReduce as specific (open source) example
- New computation models
  - **Spark**
    - Based on Resilient Distributed Datasets (RDDs)
  - **Flink**
    - Data flow engine with generic (parallelizable) operator model
  - Both try to be more main-memory based than disk/batch based

Reprise: Big Data Analytics and Storage

- Big Data Analytics
  - Declarative Data Querying on top of M/R
  - Several examples: Pig Latin, **HIVE**, Tensing, …
- Big Data Storage
  - NoSQL storage models
    - (Dynamo/Cassandra and HBase/BigTable as explicit examples)
  - Underlying data management principles
    - (partitioning & replication)
  - Consistency Models and Algorithms
    - eventual consistency vs. strong consistency
    - CAP theorem
    - 2PC and Paxos

Reprise: Assessment Package

Assessment Tasks

- Practical Assignments with Hadoop, Spark & Flink (Wks7-13) 45%
  - Late submissions: 10% penalty for each day (EXCEPT A3)
- Written Exam on Cloud Concepts 55%

Marks will be published on eLearning

- Report any errors or omissions within 10 days!

School Policy

You must obtain at least **40% in the final exam**, as well as an **overall mark of at least 50%**, to pass the unit!

Research Activity in 2017

This semester, we also conduct a study of the usability and 'learnability' of cloud computing frameworks.

You worked with three different platforms – Hadoop, Spark and Flink; and some **self-reflective surveys** are part of the assignment submissions.

Note: No submission of A3 without having done surveys for A1 and A2 before!

You can opt-in to use your assignment submissions and survey data for our study.

The study outcome will **not** be used for marking and everyone will need to do the same tasks; but if you opt-in, the **anonymised meta-data** of your assignment submissions and the answers from the self-reflective survey will be used for the study.

Scheduling of the Final Exam

- **Thursday, 22nd of June, 1:50pm – 4:00pm**
  - Venues: typically several different rooms around the campus
    for your room and time, please check *Sydney Student*
- Standard two-hour exam
- What you can bring:
  - You are allowed to bring **one** A4 sheet of own notes (double-sided), but no books or papers
- **Don't forget to bring your University student card!**
- cf. http://sydney.edu.au/current_students/student_administration/examinations/students.shtml
- http://sydney.edu.au/students/preparing-for-exams.html

Exam Paper Overview

- There will be five to six questions.
- Questions are typically structured into multiple sub-questions.
- Write your answers in the space provided in the exam paper
  - Use a pen! **Pencils are not allowed!**
- The exam will have a total of 60 marks
- The exam has a 40% barrier
  - **You need to get at least 24 of 60 points in the final exam to pass this subject**

Exam Techniques

- Read each question carefully and make sure you understand it thoroughly before answering it
  - You will have 10 minutes to just read the exam paper
  - Plan your time, choose the easiest questions to do first
- When answered, cross-check whether the question is answered
  - Example: if asked for two advantages/disadvantages each, your answer should give a total of four points
- Check for "Justify your choice", "Briefly explain", "Discuss" or "Give an example" parts.
  - Such questions test your *understanding* of an area
  - Please answer BRIEFLY
    - (One or two sentences are typically OK)
- Write the answer as neatly as possible into the answer space; make use of rough paper work to develop your answer if needed.
  - The better you structure your answer the more you help to mark it.
  - **Pencils ARE NOT allowed. Use a black or blue pen**
- General preparation tips: http://sydney.edu.au/students/preparing-for-exams.html

Exam Content Tips

- Assessable:
  - All lecture content including slides and readings
  - Tutorial material; all assignments
- The focus will be on **concepts and algorithms**
  - Less on programming; if algorithms, it is about the algorithm itself (such as complexity, request sequence, data flow, etc)
  - BUT: there is the possibility that small coding sections are included too
    - You are expected to be able to READ and WRITE code
    - And don't forget everything about your assignment submissions ;)
- Do not purely learn definitions and acronyms
  - Understanding each concept and their relation & limitation is important too
- We also did several tasks on evaluating cloud computing approaches and even had a whole lecture on benchmarking
  - How do you judge and assess a given approach based on some measurement results?
  - What scalability/performance behavior do you expect?

Last Tutorials This Week 13

- Tutors will help again with assignment 3
- Please also finish the Self-reflection Surveys for A2 if you have not done so yet
- **In the waiting time till the tutorials, please fill in the UoS evaluation in Blackboard!**

Arrangement of access to the local cluster

- The cluster will stay there
  - Some upgrade may happen
- Your home directory on HDFS will be removed before next semester starts unless you inform us you want that to be kept.

Reminder: Unit of Study Evaluation

- ONLINE survey at
  - http://www.itl.usyd.edu.au/surveys/complete/default.cfm