CPSC 436C: CLOUD COMPUTING FOR DATA SCIENCE

The University of British Columbia (Vancouver)
Winter Session 2023Term 2

Hugh Dempster Pavilion (DMP) - Room: 110 | Days: Tue/Thu – 8AM -9:30PM

ACKNOWLEDGEMENT

UBC's Point Grey Campus is located on the traditional, ancestral, and unceded territory of the xwməθkwəy̓əm (Musqueam) people. The land it is situated on has always been a place of learning for the Musqueam people, who for millennia have passed on in their culture, history, and traditions from one generation to the next on this site.

COURSE INFORMATION

Course Title	Course Code Number	Credit Value
Cloud Computing for Data Science	CPSC 436C	3 credits

Prerequisites

CPSC 203, CPSC 330, CPSC 368, and third year standing in a computing specialization.

Corequisites

N/A

CONTACTS

Instructor	Dr. Maryam R.Aliabadi Email: mraiyata@cs.ubc.ca Office: X465, ICICS *You have to expect a response within 24 hours after sending an email.* Instructor Office Hours: Thu at 10AM - 12 PM (Feel free to reach out anytime by appointment.)
TA Team	Aryan Bhairaw Email: baryan01@student.ubc.ca Office Hours: TBD Arman Moztarzadeh Email: arman88@student.ubc.ca Office Hours: TBD Ryan Dick Email: rdick01@student.ubc.ca Office Hours: TBD

COURSE DESCRIPTION

This course is an introduction to cloud computing designed for the students who wish to use the cloud for data science applications. It covers the topics of how cloud computing can be used to support data science workflows, including data storage, processing, analysis, and visualization. It also includes security considerations for the entire pipeline. Overall, the course provides students with the skills and knowledge necessary to effectively use cloud computing for design, implementation, test, and deployment of data science applications.

TARGET AUDIENCE

This course is specifically designed for undergraduate students in their third year of a computing specialization, with a focus on those pursuing a data science minor.

COURSE STRUCTURE

This course lectures are offered on Tue/Thu at 8AM- 9:30 PM.
The course will be in person in **DMP 110.**
The TA office hours and lab tutorials will be announced soon.
The discussions will happen in Piazza.
The course material will be available on Canvas and course website.

The course will be a standard 13-week in-person course, structured as a sequence of five topic areas:

- Cloud service delivery models
- Cloud storage systems
- Batch processing
- Stream processing
- Cloud security.

There will be 3 hours of lecture and 1 hour lab per week, and students are expected to spend up to 8 hours per week on readings and homework assignments. Lab sections (2 hours per week) will provide dedicated time for students to work on homework, as well as access to TA support.

The course materials including lecture slides, assignments and projects are available through UBC Canvas learning management system. Every student will be provided with an account on two cloud platforms (AWS and Azure) to do the course assignments and projects.

LEARNING OUTCOMES

After completing this course, students are able to:

1. **Deploy** different service delivery models such as virtualization, containers and serverless in thecloud;
2. **Identify** trade-offs in different data storage solutions;
3. **Identify** the appropriate tools and architectures to implement a cloud-based design,
4. **Create** a distributed computing pipeline using cloud-based execution engines, including scheduling the jobs that comprise the pipeline;
5. **Analyze** large datasets using a distributed computing pipeline for different applications;
6. **Analyze** the trade-offs between performance and cost using different designs to meet real-world constraints;
7. **Identify** basic cloud security services such as Identity and access management, network security, compliance and incident response.

LEARNING MATERIALS

Required Textbooks

There is no manatory text book for this course. All learning material are made available on UBC Canvas. However, there is an optional textbook:
1- Learning Spark: Lightning-fast Data Analytics, by: Jules Damji , Brooke Wenig , Tathagata Das

LEARNING APPROACH AND ACTIVITIES

In this course, students will be provided with 7 assignments. Each assignment exercises a concept or group of concepts from the course. Students will be asked to submit code for each assignment, a demo file showing their deployed application in running in the cloud, and a report. We have designed 5 lab tutorials that will be run by TAs within office hours to streamline the course assignments' completion.

There will be one midterm exam and a final, which will test students' ability to reason about Cloud concepts and their understanding of the course material.

Group work is central to the course. In forming groups (of 2-3), we will balance student preferences with other criteria such as ensuring groups can bring at least one of their own computers suitable for lab and hands-on assignments and taking into account diversity of backgrounds and gender balance.

SCHEDULE OF TOPICS

Note that class runs at 8:00AM – 9:30PM on Tusdays and Thursdays each week. All deadlines, dates and times are given in Pacific Standard Time (PST).

Week	Topic	Class Sessions	Labs	Assignments
Week 1 Jan 8-12	**Topic 1.1:** Introduction to Datacentres and Cloud	Tue Jan 9 • Lecture Thu Jan 11 • Lecture		

Week 2 Jan 15-19	**Topic 1.2:** Function as a service & Containerization	Tue Jan 16 • Lecture Thu Jan 18 • Lecture		
Week 3 Jan 22-26	**Topic 1.3:** Virtualization	Tue Jan 23 • Lecture Thu Jan 25 • Lecture • Group activity	Tue/Thu • Lab1-AWS Mon/Fri • Lab1-Azure	Assignment 0 - Release : Jan 25th - Due: Feb 1st at 11:59pm
Week 4 Jan 29-Feb 2	**Topic 2.1:** Big Data	Tue Jan 30 • Lecture Thu Feb 1 • Lecture	Tue/Thu • Lab2-AWS Mon/Fri • Lab2-Azure	Assignment 1 - Release : Feb 1st - Due: Feb 8th at 11:59pm
Week 5 Feb 5-9	**Topic 2.2:** Data Stores	Tue Feb 4 • Lecture Thu Feb 6 • Lecture	Tue/Thu • Lab3-AWS Mon/Fri • Lab3-Azure	Assignment 2 - Release : Feb 8th Due: Feb 15th at 11:59pm
Week 6 Feb 12-16	**Topic 2.3:** Data Management Systems	Tue Feb 13 • Lecture Thu Feb 15 • Lecture		Assignment 3 - Release : Feb 15th Due: Feb 29th at 11:59pm
Week 7 Feb 19-23	Readig break			
Week 8 Feb 26-Mar 1	**Topic 3.1:** Data Processing	Tue Feb 27 • Lecture Thu Feb 29 • Lecture		
Week 9 Mar 4-8	**Topic 3.2:** Structured Data Processing Machine Learning	Tue Mar 5 • Lecture Thu Mar 7 • Exam		**Midterm Exam : Mar 7th**
Week 10 Mar 11-15	**Topic 3.3:** Distributed Machine Learning	Tue Mar 12 • Lecture Thu Mar 14 • Lecture		Assignment 4 - Release : Mar 14th Due: Mar 21st at 11:59pm
Week 11 Mar 18-22	**Topic 3.4:** Stream Processing	Tue Mar 19 • Lecture Thu Mar 21 • Lecture		Assignment 5 - Release : Mar 21th Due: Mar 28th at 11:59pm
Week 11 Mar 25-29	**Topic 3.5:** Graph Processing Resource Management	Tue Mar 26 • Lecture Thu Mar 28 Guest speaker		
Week 12	**Topic 3.6:**	Tue Apr 2		Assignment 6

Apr 1-5	Cloud Security	• Lecture Thu Apr 4 Guest speaker		- Release : Apr 4th - Due: Apr 11th at 11:59pm
Week 13 Apr 8-12	**Topic 3.7:** Advanced topics	Tue Apr 9 • Lecture Thu Apr 11 • Guest speaker		
Week 14 Apr 15-19	**Final Exam**			**Final Exam : Apr 18th**

Note: Certain topics and assignments outlined in the syllabus are considered *tentative* and will be covered if we have sufficient time. This approach is due to the first offering of the course, and our aim is to ensure a comprehensive understanding of the core concepts before potentially exploring these additional areas.

ASSESSMENTS OF LEARNING

Assessments to student learning include the following components in this course. Each component must be passed to successfully complete the course and receive credits. The passing grade is 50%.

Components	Points/Marks	Weight
Assignment 0	5	5%
Assignment 1	5	5%
Assignment 2	5	5%
Assignment 3	5	5%
Assignment 4	5	5%
Assignment 5	5	5%
Assignment 6	10	10%
Midterm Exam	25	25%
Final Exam	30	30%
In-class participation	5	5%

Note: **Group-based activities (55%)**

- Assignment 0-5: 5 points each (30%)
- Assignment 6: 10 points (10%)

Individual activities (45%)

- Midterm exam (25%)
- Final Exam (30%)
- In-class participation (5%)

Student final letter grade will be given based on the following:

Letter Grade	Percentage
A+	90% - 100%

A	85% - 89%
A-	80% - 84%
B+	76% - 79%
B	72% - 75%
B-	68% - 71%
C+	64% - 67%
C	60% - 63%
C-	55% - 59%
D	50% - 54%
F (Fail)	0% - 49%

Assignment 0: Go Serverless (5%)
Addressing course learning outcome #1.

Goal: To set up the cloud environment using one of the main cloud providers: AWS, Azure or GCP, and create, deploy and test a function as a service on the target cloud platform.

Assignment 1: Containerization Vs. Serverless (5%)
Addressing course learning outcomes #1 and 7

Goal: to give the students hands-on experience with building and deploying containers for data science applications and compare containers with serverless.

Assignment 2: Running Image recognition on a Virtual Machine (5%)
Addressing course learning outcomes #1 and 7

Goal: to learn how to launch and use a virtual machine for running and deployment of a data science application in the cloud.

Assignment 3: Running image recognition in a VM using Object Store (5%)
Addressing course learning outcomes #2, 3, and 7

Goal: to understand the flexibility and cost-effective options for storing data in cloud and find the best trade-off between dataaccess, performance, and cost based on the applic ation requirements.

Assignment 4: Clustering and Batch Processing (5%)
Addressing course learning outcomes #3 and 4

Goal: to learn how clustering speeds up big data analysis and to give students hands-on experience in setting up and configuring a cluster and find tradeoff between cost and performance of different cluster sizes.

Assignment 5: Streaming Text Analysis using Spark (5%)
Addressing course learning outcomes #3 and 4

Goal: to learn how to perform real-time processing, for which they will set up a local Spark Streaming environment, and implement a real-time application.

Assignment 6: Building a Machine Learning Pipeline through Jupyter Notebook (10%)
Addressing course learning outcome #3

Goal: to learn how to build a machine learning pipeline including training, saving, loading, deployment, and predicting through a jupyetr notebook in the cloud.

Midterm Exam (25%)
Addresses course learning outcomes #1, #2, #3 and 4?
In midterm exam, students will be evaluated on module 1, and 2 of the course.

Final Exam (30%)
Addresses course learning outcomes #5, #6 and #7
In final exam, students will be evaluated on module 3 of the course.

In-class Participation (5%)
Active in-class participation constitutes 5% of the final grade. Engaging in discussions, collaborative activities, and asking questions demonstrates the students' commitment to learning. Critical thinking, respectful interactions, and contributions to group work are key aspects.

Policies on Late Submissions and Re-grading

Late submissions for the assignments are always accepted. However, the late submission of each assignment will be subject to 15% and 50% penalty in garde, if happen by one week and one month after deadline, respectively.

Participation Expectations

In-class active participation comprises 5% of final grade. However, abcences more than three times without any reason is subject to failure in the course.

UNIVERSITY POLICIES

UBC provides resources to support student learning and to maintain healthy lifestyles but recognizes that sometimes crises arise and so there are additional resources to access including those for survivors of sexual violence. UBC values respect for the person and ideas of all members of the academic community. Harassment and discrimination are not tolerated nor is suppression of academic freedom. UBC provides appropriate accommodation for students with disabilities and for religious observances. UBC values academic honesty and students are expected to acknowledge the ideas generated by others and to uphold the highest academic standards in all of their actions.

Details of the policies and how to access support are available on the UBC Senate website**.**

OTHER COURSE POLICIES

Learning Analytics

Learning analytics includes the collection and analysis of data about learners to improve teaching and learning. This course will be using the following learning technologies: Canvas. Many of these tools capture data about your activity and provide information that can be used to improve the quality of teaching and learning. In this course, I plan to use analytics data to:

- View overall class progress

- Track your progress in order to provide you with personalized feedback
- Review statistics on course content being accessed to support improvements in the course
- Track participation in discussion forums
- Assess your participation in the course

Copyright

All materials of this course (course handouts, lecture slides, assessments, course readings, etc.) are the intellectual property of the Course Instructor or licensed to be used in this course by the copyright owner. Redistribution of these materials by any means without permission of the copyright holder(s) constitutes a breach of copyright and may lead to academic discipline.

You are not allowed to record and share the course without permission.

Academic Integrity Policy

All of the work you submit must be done by you and your work must not be submitted by someone else. Plagiarism is academic fraud and is taken very seriously. The department uses software that compares programs for evidence of similar code/report. Please read the Rules and Regulations from the UBC Academic Integrity website: https://academicintegrity.ubc.ca/student-start/