CPSC 436C: CLOUD COMPUTING FOR DATA SCIENCE

The University of British Columbia (Vancouver)
Winter Session 2024 Term 1
CEME-Floor 1-Room 1202 | Days: Tue/Thu – 11AM -12:30PM

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
(a) CPSC 203 or CPSC 210, and (b) CPSC 304 or 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 1PM - 3PM (Feel free to reach out anytime by appointment.)
TA Team	Mahmoud Alkhatib Email: makhatib@students.cs.ubc.ca Office Hours: Monday at 1pm-3pm Arman Moztarzadeh Email: arman88@student.ubc.ca Office Hours: Monday at 10am-12pm Sopuruchi Chisom Email: schisom@student.ubc.ca Office Hours: Friday at 12pm-1:30pm Chen ZHao Email: laozc@student.ubc.ca Office Hours: Friday at 1:30pm-3pm

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 11AM- 12:30PM.
The course will be in person in CEME- 1202**.**
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 the cloud;
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 mandatory text book for this course. All learning materials 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 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 11:00AM – 12:30PM on Tuesdays and Thursdays each week. All deadlines, dates and times are given in Pacific Standard Time (PST).

Week	Topic	Reading	Labs	Quizzes	Assignments
Week 1 Sep 2-7	**Imagine UBC 2024** Introduction to Data Centres and Cloud				
Week 2 Sep 9-13	Function as a service & Containerization				

Week 3 Sep 16-20	Virtualization		Lab 1: **Setting up a cloud Account**		Assignment 0 - Release : Sep 19th - Due: Sep 26st at 11:59pm
Week 4 Sep 23-27	Big Data		Lab 2: **Setting up Cloud Computing Services**	Quiz 1 (week 1-3)	Assignment 1 - Release : Sep 26th - Due: Oct 3rd at 11:59pm
Week 5 Sep 30 - Oct 4	Data Stores				
Week 6 Oct 7-11	Data Management Systems		Lab 3: **Setting up Cloud Storage Services**		Assignment 2 - Release : Oct 8st Due: Oct 15th at 11:59pm
Week 7 Oct 14-18	Data Processing			Quiz 2 (week 4-6)	Assignment 3 - Release : Oct 15th - Due: Oct 29th at 11:59pm
Week 8 Oct 21-25	Structured Data Processing Machine Learning		Lab 4 : **Setting up Spark Streaming**		
Week 9 Oct 28 - Nov 1	Stream Processing				**Proposal Submission Deadline :** Nov 1st at 11:59pm
Week 10 Nov 4-8	Graph Processing Cloud Security-Part1		Lab 5 : **Setting up a Spark Cluster**		Assignment 4 - Release : Nov 5th - Due: Nov 17th at 11:59pm
Week 11 Nov 11-13 Nov 14-15	Reading Break Cloud Security-Part2				
Week 12 Nov 18-22	Guest speaker for Cloud Security (from AWS) Guest speaker for Cloud Security (Anthony Green)			Quiz 3 (Week 8-11)	

Week 13 Nov 25-29	Guest speaker for Advanced topics (Tom Jin) Project Presentation				
Week 14 Dec 2-6	Project Presentation Project presentation				
Week 15 Dec 9-13	Course wrap up				**Project Final Report Submission Deadline :** Dec 10th at 11:59pm **Final Exam :** **Dec 15th at 8:30 am**

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	25%
Assignment 1	5	
Assignment 2	5	
Assignment 3	5	
Assignment 4	5	
Quiz 1	5	15%
Quiz 2	5	
Quiz 3	5	
Project	25	25%
Final Exam	30	30%
In-class participation	5	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 the cloud and find the best trade-off between dataaccess, performance, and cost based on the application requirements.

Assignment 4: Building a Machine Learning Pipeline through Jupyter Notebook (5%)
Addressing course learning outcome #3

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

Assignments must be submitted on Canvas by the specified due date. Extensions beyond this period are granted only until the release of the subsequent assignments. Late submissions incur a penalty of 30% if submitted after the main due date and 50% if submitted after the extended due date, unless a valid reason such as illness or another extreme condition is provided by the student. If you cannot finish any assignment even with the extra hours, write to teaching team along with the reasons. We will exempt you from the assignment and shift the weight to the other submitted assignments.

Quizzes (15%)
Addresses course learning outcomes #1, #2, #3 and 4?
In the quizzes, students will be evaluated on modules 1, 2 and 3 of the course. There will be 3 quizzes roughly at three weeks intervals. If you are going to miss a quiz, write to TA team as soon as you are

aware of the situation. If you miss one quiz, the remaining quizzes will be considered for your final grade. If you miss more than one quiz, we will move the weights to other quizzes. All quizzes will be available starting Monday of a week and will span three days.

Course Project (25%): In teams of up to 5, students will design and implement a cloud-based data processing pipeline using one storage system, one computing engine, and one database to store or visualize the results. The project topic is student-chosen but must cover the entire pipeline, giving students the opportunity to apply the practical skills learned in the course to solve a real-world problem. Evaluation will focus on the quality of the architecture, cost-efficiency, performance, and security. The project will culminate in a final presentation at the end of the semester.

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

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. We will use iClicker questions for marking the presence. In each lecture class, there will be **iClicker questions** worth a small number of points. We will drop roughly 10-20% of these components for the final grading. However, absence more than five times without any reason is subject to losing the entire attendance mark.

Computer Based Test Facility (CBTF)

CBTF portal: https://ca.prairietest.com/

Quizzes

The quizzes will be conducted in the Computer Based Test Facility, which allows you the flexibility of taking the quizzes at your convenient date and time. Each of the three quizzes will be run for three days in the week they are scheduled. You will be able to register for a 1-hour slot on the CBTF portal one week before the week of a quiz. Students must ensure timely completion within this window, as the deadline cannot be extended except for valid reasons such as illness or university-related activities, which require prompt notification to the instructor or course coordinator with supporting documentation. In cases where a student misses a quiz due to a valid reason and provides timely notification with supporting documentation, the weight of the missed quiz will be redistributed to other quizzes or the final exam, as determined by the course grading policy.

Inside the CBTF, you will be giving the test on a computer via PrairieLearn. The CBTF only supports exams that are closed book and closed notes; our quizzes will comply with this requirement. No electronic devices, such as laptops and tablets will be allowed.

Final exam

Our plan is to use the CBTF for the final exam as well, although this might change. The exam format will be announced well ahead of time.

Online Learning Tools

- **PrairieLearn:** for Quizzes and Final Exam
  - CPSC 436C PrairieLearn - 2024W1
- **iClicker Cloud** (for in-class participation questions):
  - CPSC 436C https://join.iclicker.com/
- **Piazza** (for class discussions): Sign up link

- **Canvas:** for lecture slides, assignments, project and the other resources

University Policies Weather Contingency Plan for Class, Quizzes and Exam

If in-person activities in our course are canceled due to weather conditions (e.g., snow), please monitor https://www.cs.ubc.ca for information, the pinned posts section on Piazza for announcements on our course, and your UBC-registered e-mail. Specifically, in the event of such cancellations, we will:

- Make up missed lectures and tutorials via posted recordings or alternate exercises; follow Piazza for details.
- Adjust the deadline of quizzes, assignments, and in-class exercises during the affected period.
- Add scheduling dates to exams during the affected period and post instructions on how to re-register. (For the final exam, in the unlikely event that we run out of days available in the exam period, we will work with scheduling services and announce an alternate solution similar to normal exam rescheduling. Follow Piazza and your UBC-registered e-mail for more information.)

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 AnalyticsLearning 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

Academic Integrity

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/