Course Syllabus for SIADS 516: Big Data: Scalable Data Processing

Course Overview and Prerequisites

This course will introduce students to the use of the Spark data analysis framework for the analysis of Big Data. We will cover the theory and application of MapReduce strategies, the use of Resilient Distributed Datasets in cluster computing, and higher-level abstractions such as DataFrames and SparkSQL, which facilitate efficient analysis of large amounts of data.

The prerequisites for SIADS 516 include:

SIADS 505: Data Manipulation

SIADS 511: SQL & Databases

Instructor and Course Assistants

Instructor: Chris Teplovs, Lecturer IV & Research Investigator, School of Information

Course Assistants:

Kris Steinhoff, Intermittent Lecturer in Information, School of Information and Staff Software Engineer at Woven Planet

Oleg Nikolsky, Graduate Student Instructor and Intermittent Lecturer in Information, School of Information

Naga Santa, Intermittent Lecturer in Information, School of Information

Toby Kemp, Intermittent Lecturer in Information, School of Information

Jake Huang, Graduate Student Instructor and Intermittent Lecturer in Information, School of Information

Chris Teplovs designed and developed this course.

Course Communication Expectations

If you have questions about course content (e.g. lecture videos, quizzes, or assignments), please use the class Slack channel to discuss with classmates and the instructional team. Instructor and course assistant response time to Slack messages will be within 24 hours.

Personal communication that may involve sensitive information can be emailed directly to *Kris Steinhoff <steinhof@umich.edu>*. Instructor and course assistant response time to email messages will be within 48 hours.

Help Desk(s): How to get Help

Need help? You may reach out to UMSI or Coursera depending on the type of question you have.

- Degree program questions or general help - umsimadshelp@umich.edu
- Coursera's Technical Support (24/7) - https://learner.coursera.help/

Weekly Readings

- See the Library's access instructions for the readings hosted by O'Reilly Online.

Week 1 Readings

Data Science from Scratch by Joel Grus (1st Edition: Chapter 24, 2nd. Edition: Chapter 25)—Required

Hadoop with Python by Zachary Radtka (Chapter 2)—Required

Advanced Analytics with Spark by Josh Wills, Uri Laserson, Sandy Ryza, and Sean Owen (Chapter 1)—Optional

High-Performance Spark by Rachel Warren and Holden Karau (Chapter 2)—Optional

Week 2 Reading

Spark, the Definitive Guide by Matei Zaharia and Bill Chambers (Chapter 12)—Required

Debugging Apache Spark by Holden Karau (Video) - Optional

Week 3 Reading

High-Performance Spark by Rachel Warren and Holden Karau (Chapter 3)—Required

Week 4 Readings

High-Performance Spark by Rachel Warren and Holden Karau (Chapter 4)—Required

Learning Spark (2nd ed.) by Jules Damji, Denny Lee, Brooke Wenig, and Tathagata Das—Optional

Learning Outcomes

Explain the relationships between MapReduce, Spark and Hadoop

Use MapReduce to sort, filter, and summarize a dataset

Use Spark RDDs to manipulate, summarize, and analyze a text-based dataset

Use Spark DataFrames to manipulate, summarize and analyze heterogeneous data

Use Spark SQL to perform complex queries on heterogeneous Big Data

Course Schedule

- **This course begins on Tuesday, November 22, 2022 and ends on Monday, December 19, 2022.**
- Weekly assignments will be **due on Mondays at 11:59 pm** (Ann Arbor, Michigan time-Eastern Standard/Daylight Time - EDT, UTC -4).

In **Week 1**, after some discussion of how we can define Big Data, you will be introduced to the split-apply-combine workflow common in Big-Data work. After overviewing scaling of distributed systems, week one progresses further into discussing how one can use libraries such as MapReduce, Hadoop, and MrJob for distributed solutions to big-data problems.

In **Week 2**, the course shifts focus towards introducing the Apache Spark functionality, structure, and stack. After further exploring the relationship between Spark and Hadoop, students will learn about the creation and manipulation of resilient distributed datasets (RDDs) using Spark and Python.

In **Week 3**, you will focus on working with Spark dataframes, including initialization from various file-types, column manipulation, and operations such as sorting, grouping, and user-defined functionality.

In **Week 4**, the course shifts focus to Spark's support for Structured Query Language (SQL) and explores methods available for data selection, filtration, and cross-table joining with SQL and Spark.

Weekly Office Hours via Zoom (Ann Arbor, Michigan time):

Your instructor will hold weekly, synchronous office hours using the video-conferencing tool, Zoom. The schedule of office hours can be found by clicking on the **Live Events** link in the left-hand navigation menu. Additionally, all office hours will be recorded and archived so that you can retrieve at a later date. Archived office hours can be found in the last lesson for each week with the heading **SIADS 516 Office Hour Recordings**.

Grading

Course Item	**Percentage of Final Grade**	**Due**
Week 1 Homework	25%	Monday, November 28, 2022
Week 2 Homework	25%	Monday, December 5, 2022
Week 3 Homework	25%	Monday, December 12, 2022
Week 4 Homework	25%	Monday, December 19, 2022
Total	**100%**	

Note: All assignments are required to earn credit for this course.

Letter Grades, Course Grades, and Late Submission Policy

We realize that the occasional crisis might mess up your schedule enough to require a bit of extra time in completing a course assignment. Thus, we have instituted the following late policy that gives you a limited number of flexible “late day” credits.

You have three (3) free late days to use during SIADS 516. One late day equals exactly one 24-hour period after the due date of the assignment (including weekends). No fractional late days: they are all or nothing. Once you have used up your late days, 25% penalty for each subsequent 24-hour period after the deadline that an assignment is late. For example, if the due date is 11:59pm Monday, with no late days left, penalties would be:

Before 11:59pm Tuesday: 25% deduction

Before 11:59pm Wednesday: 50% deduction

Before 11:59pm Thursday: 75% deduction

After 11:59pm Thursday: 100% deduction

You don't need to explain or get permission to use late days, and we will track them for you. Note that resubmissions after the deadline will be counted as late submissions.

The grading scale for this course is as follows:

A	95%
A-	90%
B+	87%
B	83%
B-	80%

C+	77%
C	73%
C-	70%
D+	67%
D	63%
D-	60%
F	0%

Academic Integrity/Code of Conduct

Refer to the Academic and Professional Integrity section of the UMSI Student Handbook. (access to Student Orientation course required).

Accommodations

Refer to the Accommodations for Students with Disabilities section of the UMSI Student Handbook.

Use the Student Application Form in Accommodate to begin the process of working with the University's Office of Services for Students with Disabilities

Accessibility

Refer to the Screen reader configuration for Jupyter Notebook Content document to learn accessibility tips for Jupyter Notebooks.

Library Access

Refer to the U-M Library's information sheet on accessing library resources from off-campus. For more information regarding library support services, please refer to the U-M Library Resources section of the UMSI Student Handbook (access to the Student Orientation course required).

Student Mental Health

Refer to the University's Resources for Stress and Mental Health website for a listing of resources for students.

Student Services

Refer to the Introduction to UMSI Student Life section of the UMSI Student Handbook (access to the Student Orientation course required).

Technology Tips

Recommended Technology

- This program requires Jupyter Notebook for completion of problem sets and Adobe or other PDF viewer for reading articles.

Working Offline

- While the Coursera platform has an integrated Jupyter Notebook system, you can work offline on your own computer by installing Python 3.5+ and the Jupyter software packages, including pyspark. For more details, consult the Jupyter Notebook FAQ.