| | |
|---|---|
| Georgia State University \| J. MACK ROBINSON COLLEGE OF BUSINESS | **Scalable Data Analytics**<br>**MSA 8050** |

# Sample Syllabus

**Instructor:**

Dr. Kai Zhao
Email: kaizhaofrank@gmail.com
In-person class: Buckhead, 1203
Online class: Zoom
Class Hours: Mon 1:00pm-3:30pm  Mon 6:00pm-8:30pm

**Prerequisites**:

MSA 8010.

**Textbook:**

- Learning Spark: Lightning-Fast Big Data Analysis, by Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, 2015.

**Course Description:**

This course covers essential concepts and tools for large scale data analytics. Topics include 1) functional and parallel programming paradigms and languages, 2) core components of large scale platforms, and 3) scalable machine learning algorithms. Programming projects demonstrate the design and implementation of large-scale analytics pipelines for structured and unstructured data.

**Course Objectives:**
By the end of the semester students will be able to:

- Broad understanding of big data and its ecosystem
- Ability to identify big data challenges and how to tackle them
- Understanding of big data programming paradigm
- Knowledge in how to implement analytical tools using Apache and Hadoop

**Homeworks:**

Four mini-projects (Homeworks), every three weeks, students complete a hands-on project that further explores the topic/technique covered in class. This is an individual activity. With these

mini-projects, students gain proficiency in the various tools assigned for this class.

**Final Project:**

The project consists of a research report and presentation on a student-selected topic that is relevant to the course. It is group-based. On the last day of class, each group will present their findings to the class (a 15-min presentation and a written report). The grade will be based on the evaluation from the instructor. In the final project, the students will work on a big data set that cannot be processed by a laptop. The students need to work in groups and analyze the data set using the cluster environment.

**Typical class session:**

Class sessions will comprise (1) lectures/discussions of relevant techniques, concepts, and features, (2) instructor demonstrations, and student lab sessions with hands-on work. The purpose of this pedagogical approach is to introduce and reinforce ideas and skill sets so that you can master these on your own after class hours.
To bring this knowledge to a highly proficient, professional level, you will have to spend time and effort outside of class reviewing and practicing the class material.
To ensure that you have the basic knowledge that will allow you to function on your own after class, be sure to ask the instructor questions during class, either during the lecture/discussion, demo, or lab.

**Classroom guidelines:**

Coming to class fully prepared and contributing to the discussion help deepening the learning.
Individual deliverables are to be submitted individually and group work is collaborative.
Refer to http://www2.gsu.edu/~wwwfhb/sec400.html for additional information on instructional information.

**Grading:**

Deliverables

| | |
|---|---|
| Quiz | 30% |
| Homeworks | 45% |
| Final Project | 25% |
| **Total** | 100% |

*Late submission policy: deliverables submitted after their due date will be penalized 50% the first week after. No submission is accepted after one week.

Letter Grade Scale

| A+ | A | A- | B+ | B | B- | C+ | C | C- | D | F |
|---|---|---|---|---|---|---|---|---|---|---|
| 97.0% -100 % | 91.0 – 96.9 % | 89.5 – 90.9% | 87.0 – 89.4% | 83.0 – 86.9 % | 79.5 – 82.9% | 77.0 – 79.4% | 72.0 – 76.9 % | 69.5 – 71.9 % | 60.0– 69.4 % | Below 59.9% |

**Class Schedule (adjustments may be necessary)**

| Date | Topic | Reading Spark* | Mini-projects (MP) |
|---|---|---|---|
| Class 1 01/11 | Introduction to Big Data Analytics<br>Dealing with Data Volume: Streaming Computation<br><br>*Understand Methodology to handle the data stream<br>*Write Python code with generator | | |
| Class 1 01/25 | Higher-Order Function and Parallel Computing<br><br>*Write map, reduce and filter functions | | |
| Class 3 02/01 | MapReduce<br><br>*Understand the map and the reduce function in Hadoop | | MP 1 assigned<br>Quiz 1<br>(20 mins) |
| Class 4 02/08 | Hadoop<br><br>*Understand the Hadoop system<br>*Write Linux command lines | | |
| Class 5 02/15 | Intro to Data Analytics with Spark<br><br>*Understand the basics of the Spark programme<br>*Install the Spark in each laptop. | Spark Chapter 1<br>Spark Chapter 2 | MP1 due<br>MP2 assigned |
| Class 6 02/22 | Programming with RDDs<br><br>*Learn how to programme with RDD | Spark Chapter 3 | MP2 due |
| Class 7 03/01 | Working with Key/Value Pairs<br><br>*Use Python to do transformation with key/value pairs | Spark Chapter 4 | Quiz 2<br>(20 mins)<br>MP3 assigned |
| Class 8 03/08 | Running on a Cluster<br><br>*Learn Linux command line and how to make a programme run on a cluster | Spark Chapter 7<br>Spark Chapter 8 | Final Project released |

| | | | |
|---|---|---|---|
| Class 9<br>03/22 | Machine Learning with MLlib<br><br>*Write MLlib code<br>*Run Mlib to analyse data sets and extract insights from data set | Spark Chapter 10 | MP3 due |
| Class 10<br>03/29 | Machine Learning with MLlib<br>*Loading and Saving your Data and Advance Spark Programming | Spark Chapter 5<br>Chapter 6 | MP4 assigned |
| Class 11<br>04/05 | Spark DataFrames and GraphX<br><br>*Write Spark DataFramework to advance preprocessing data | Spark Chapter 10 | Quiz 3<br>(20 mins) |
| Class 12<br>04/12 | Spark Streaming<br>Invited Speaker Ashwin Jagarapu: How Companies use Spark for Analysing Big Data<br><br>*Write real-time big data processing code<br>*How companies use cloud computing and big data tools | Spark Chapter 11 | MP4 Due |
| Class 13<br>04/19 | Deep learning with Spark<br><br>*Write deep learning spark codes | | |
| Class 14<br>04/26 | Project presentation & report submission | | |

* Textbook – Learning Spark: Lightning-Fast Big Data Analysis, by Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, 2015.