Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS	**Scalable Data Analytics** **MSA 8050**

Course Syllabus

Instructor:

Dr. Kai Zhao
Email: kzhao4@gsu.edu
Office: Buckhead, 536
Office Hours: Mon 3:30pm-4:30pm

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 design and implementation of large scale analytics pipelines for structured and unstructured data.

Course schedule: Mon 4:30pm-7pm
Classroom: 501

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:

Three mini-projects (Homeworks), each 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 analyse 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

Participation (in-class & 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 01/22	Introduction to Big Data Analytics		
Class 2 01/29	Dealing with Data Volume: Streaming Computation		
Class 3 02/05	Higher Order Function and Parallel Computing		
Class 4 02/12	MapReduce		MP 1 Quiz 1 (20 mins)
Class 5 02/19	Hadoop		
Class 6 02/26	Intro to Data Analytics with Spark	Spark Chapter 1 Spark Chapter 2	
Class 7 03/05	Programming with RDDs	Spark Chapter 3	MP1 due
Class 8 03/19	Working with Key/Value Pairs	Spark Chapter 4	
Class 9 03/26	Loading and Saving your Data and Advance Spark Programming	Spark Chapter 5 Spark Chapter 6	MP2 assigned Quiz 2 (20 mins)
Class 10 04/02	Running on a Cluster	Spark Chapter 7 Spark Chapter 8	Final Project released
Class 11 04/09	Spark Streaming	Spark Chapter 10	MP2 due MP3 assigned
Class 12 04/16	Machine Learning with MLlib	Spark Chapter 11	

Class 13 04/23	Machine Learning with MLlib Invited Speaker Ashwin Jagarapu: How Companies use Spark for Analysing Big Data	Spark Chapter 11	Quiz 3 (20 mins) MP3 Due
Class 14 04/30	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.