MSA 6500 – Big Data Analytics

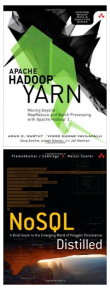
BGSU Computer Science **Spring 2015 Syllabus**
Department Office: HAYES 221
Department Phone: (419) 372–2337

Instructor: Dr. Robert Dyer
E-mail: **rdyer@bgsu.edu**
My office: HAYES 238
My phone: (419) 372–3469

Office Hours: MTWR 4:00pm – 5:30pm
OR by appointment (email me)

Class Meeting Time: Tuesdays and Thursdays, 2:30pm – 3:45pm, HAYES 114

Textbooks:



- "Apache Hadoop YARN" by Murthy and Vavilapalli, Addison-Wesley Professional, 2014, ISBN: 978-0321934505
- "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence" by Sadalage and Fowler, Addison-Wesley Professional, 2012, ISBN: 978-0321826626

Learning Outcomes for the Course: After successfully completing MSA 6500, students will be able to:

- explain Big Data concerns in the field;
- define the role of a data scientist;
- describe the life cycle of data analytics;
- save and retrieve ultra large scale data;
- use parallel processing to scale-up applications and speed up their execution;
- understand Hadoop-like distributed computing frameworks and the tools available;
- use Hadoop and/or other distributed file systems to store vast quantity of data;
- write and execute MapReduce programs to analyze unstructured data;
- utilize tools to monitor the health and performance of Hadoop-like clusters; and
- describe the advantages and disadvantages of using NoSQL databases for Big Data.

Withdrawal Deadline: Friday, April 10, 2015. University policy states that after this date, anybody withdrawing from the course will have the grade automatically turn into an F.

Grading: The final grade will be composed of the following weights.
(The instructor reserves the right to make changes at any time.)

Assessments

Item	*Each*	*Total*
Exams (2)	100	200
Project		100
Assignments (4)	25	100
Survey Paper and Presentation		100
	Overall Total	**500**

Grading Scale

Grade	*Percent Range*	*Point Range*
A	90 – 100%	450 – 500
B	80 – 89%	400 – 449
C	70 – 79%	350 – 399
D	60 – 69%	300 – 349

Homework and Projects: Homework is essential to learning the material, and you are expected to submit all homework assignments and projects. You are responsible for planning ahead to allow yourself enough time to complete all homework and projects by the deadlines. **Start your work early**. Homework and projects are due at the BEGINNING of the class on the due date.

Homework and projects will be done in small groups/pairs. Groups are randomly assigned and will change after each assignment. You are allowed to discuss the assignment only with your group members. **Discussing with other groups is considered academic dishonesty.** Group members are expected to contribute equally and all group members must write and understand the source code. Each group member will anonymously rate the other group members. **The ratings will affect individual grades.** More details will come later.

Survey Paper: Students must develop a publishable survey paper by the end of the course. The paper will examine a large number of works in a related area of big data. The paper must be in a standard ACM/IEEE conference format and include a sufficient overview of related works. The paper will be judged as if being reviewed for acceptance to a conference in the field. Students will also give a short presentation on their survey.

Attendance: Students are expected to attend each class and be on time. I do not believe in any specific grade incentives for class attendance, or penalties for absence.

Academic honesty: All coursework for this class is expected to be YOUR OWN work. The MINIMUM penalty for copying someone's work (including current classmates, students from a previous offering of the course, or postings found on the web) or knowingly allowing someone to copy your work is a zero for the homework/project/exam/presentation. The offense is also reported to the dean of your college. *Turnitin*, a plagiarism detection tool, will be used in this course. I follow the University's general codes of conduct defined in the *BGSU Student Handbook*. For details refer to:

1. *BGSU Student Handbook,* page 33 (http://www.bgsu.edu/content/dam/BGSU/student-affairs/Student-Conduct/documents/2014-15-Student-Handbook.pdf)
2. *Department of Computer Science Academic Honesty Policy* (http://www.bgsu.edu/arts-and-sciences/computer-science/policies-for-current-students.html)
3. *The Academic Charter*, section B-I.G (http://www.bgsu.edu/content/dam/BGSU/faculty-senate/documents/academic-charter/B-I-G-Academic-Honesty-Policy.pdf)

Canvas: The syllabus, schedule, and course policies are available on Canvas. Your grades will also be available on Canvas throughout the semester.

Class Notes/Handouts: Every student is responsible for taking notes in the class, collecting class handouts, and generally keeping up with the class, even if they must miss a class meeting for any reason. I do not always post class notes, so if you miss a class please borrow notes from another student.

Office Hours and Help: Please check your Canvas course site, Canvas messages, and your BGSU email **regularly**. *[You may have your Canvas messages forwarded to your email, but do check it (multiple times) daily.]* I do forward my own Canvas messages to my BGSU email and check my email frequently (with rare exceptions). I check email more often than I access Canvas so if you need to contact me urgently, use email. I will do my best to accommodate you ASAP, even if outside my posted office hours and without appointment. In general, if you need to see me in my office outside of my regular office hours, please make an appointment.

Make-up policy: If you cannot take an exam as scheduled, you (or an authorized person, only in case you are unable to do so) must contact me ahead of time with the reason. Make-ups are considered typically for health emergencies only. Taking the FINAL exam at a time other than the university-scheduled time requires approval by the MSA adviser. If you feel that you have a valid reason to request a change in FINAL exam time, inform me and obtain the request form in the Computer Science department office, Hayes 221. You must sign an academic honesty statement specifically in connection with the exam.

Disability Policy: In accordance with the University policy, students with disabilities must verify their eligibility through the Office of Disability Services for Students, 38 College Park Office Building, 419–372–8495 (http://www.bgsu.edu/disability-services.html). Contact me as soon as possible this semester to arrange any accommodations needed to assist with your success in this course.

Religious Holidays: It is the policy of the University to make every reasonable effort allowing students to observe their religious holidays without academic penalty. In such cases, it is the obligation of the student to provide the instructor with reasonable notice of the dates of religious holidays on which he or she will be absent. Absence from classes or examinations for religious reasons does not relieve the student of responsibility for completing required work missed. Following the necessary notification, the student should consult with the instructor to determine what appropriate alternative opportunity will be provided, allowing the student to fully complete his or her academic responsibilities (*The Academic Charter*, section B–I.F–4.b at: http://www.bgsu.edu/content/dam/BGSU/faculty-senate/documents/academic-charter/B-I-F-Classroom-Related-Responsibilities.pdf).

Tentative Course Schedule

Week	Day	Date	Topics	Assigned	Due
1	T	Jan 13	Introduction		
	R	Jan 15	Big Data Analytics overview		
2	T	Jan 20	GFS / HDFS		
	R	Jan 22	Using HDFS		
3	T	Jan 27	MapReduce		
	R	Jan 29	Hadoop 1.x	**HW1**	
4	T	Feb 3	Hadoop YARN		
	R	Feb 5	Column-oriented Storage		
5	T	Feb 10	R + Hadoop	**HW2**	**HW1**
	R	Feb 12	Work Day / Lab		
6	T	Feb 17	Dremel		
	R	Feb 19	Impala		**HW2**
7	T	Feb 24	Amazon Web Services (AWS)		
	R	Feb 26	Hadoop on AWS		**Survey Topic**
8	T	Mar 3	Exam Review	**HW3**	
	R	Mar 5	**Exam 1**		
9	T	Mar 10	No Class – Spring Break		
	R	Mar 12			
10	T	Mar 17	Apache Pig		
	R	Mar 19	Work Day / Lab	**Project**	
11	T	Mar 24	Boa		
	R	Mar 26	FlumeJava	**HW4**	**HW3**
12	T	Mar 31	CloudDataflow		**Survey Draft 1**
	R	Apr 2	NoSQL		
13	T	Apr 7	Bigtable		**HW4**
	R	Apr 9	HBase		
14	T	Apr 14	Cassandra		**Survey Draft 2**
	R	Apr 16	Project Voldemort		
15	T	Apr 21	Spark		
	R	Apr 23	Pregel		**Project**
16	T	Apr 28	GraphX = Spark+Pregel		
	R	Apr 30	Exam Review	**Exam 2**	**Final Survey**
17	R	May 7	**Project Presentations – 10:45am–12:45pm – Hayes 114**		

NOTE: If there is a discrepancy between the due dates here and on actual assignments, the one on the assignment applies.