MSA 6500 – Big Data Analytics

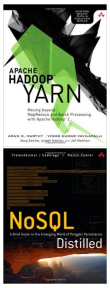
BGSU Computer Science **Spring 2016 Syllabus**
Department Office: HAYES 221
Department Phone: (419) 372–2337

Instructor: Dr. Robert Dyer
E-mail: **rdyer@bgsu.edu**
My office: HAYES 244
My phone: (419) 372–3469

Office Hours: MTWR 3:45pm – 5:15pm
OR by appointment (email me)

Class Meeting Time: Tuesdays and Thursdays, 2:30pm – 3:45pm, OLSCAMP 208

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 8, 2016. 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	***Weight***
Reading Annotations	25%
Problem Set Reflections	25%
Readiness Assurance Assessments (RAAs)	25%
Projects	25%

Grading Scale	
Range	***Grade***
[90–100]%	A
[80–90)%	B
[70–80)%	C
[60–70)%	D

Readings: Readings are absolutely essential to learning in this course. Every lecture, readings are due meaning there are 2 readings due each week. Readings will be done on the website Perusall which let's you annotate the text being read. **These annotations are required and graded.** The system will automatically grade each annotation. It keeps the 5 highest annotations. Thus, to ensure the best possible score you should aim for 7-10 **good** annotations. Annotations could be you summarizing a piece of text for the class, asking a detailed question, finding problems with the text, etc. Your annotations should be spread out through the whole assigned reading (and not all in 1 small area). You are graded on 3 criteria: quantity of annotations, average quality of annotations, and spacing of the annotations in the text.

Problem Sets and Reflections: Problem sets will be assigned for each unit. You are expected to attempt all problems on your own and bring your solutions to the assigned class. I will walk around and evaluate everyone's solution based solely on completeness. We will then spend the first portion of the class time in groups, discussing the solutions. Each group will write up a group solution. I will then provide my solutions and allow further group discussion. It's important to remember that your solution and the group's solution **are not graded**. It is quite ok to have wrong solutions, as long as you gave the problem an honest attempt!

You are then expected to go home and write a reflection on the problem set due the next class period. This is a chance to outline your own efforts, identify gaps in your knowledge, and seek additional help. **The reflection is graded** (using a multiplier based on your completeness score).

Readiness Assurance Assessments (RAAs): To help me gauge where the class is at, we will have in-class assessments at regular intervals. These will contain a small number (3–5) of relatively difficult questions. Students will take the assessment in the first half of the class individually. You will then collaborate with your group members and re-take the assessment as a group. *Your score will be the average of your individual score and the group score.*

Projects: There will be several small projects during the system. The goal of the projects is to give you hands on experience with the technologies we are studying. Projects are done in groups.

Groups: This course relies heavily on group work. You will work with a group on all course projects, the RAAs, and revising the problem sets. **Group members are expected to all contribute equally** and group members will be **rating each other's effort** to help ensure fairness. We will frequently change group membership.

Perusall: The reading annotation website we use is Perusall. To log in, you will go to Canvas and follow the link. Do not directly go to the Perusall website and log in, as that will create a second user. For an idea of what type of comments we are looking for, please view this document: http://perusall.com/downloads/scoring-examples.pdf and this document http://perusall.com/downloads/rubric.docx for details on the grading.

Attendance: Students are expected to attend each class and be on time. I take attendance at the start of each lecture. I typically use good attendance as a factor when considering final grades. I reserve the right to penalize students up to 1% of their final grade, per absence, for more than 3 un-excused absences.

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/paper/presentation. The offense is also reported to the dean of your college. *Turnitin* and *Moss*, plagiarism detection tools, will be used in this course. I will follow the Department's policies and the University's code of academic conduct as defined in the *BGSU Student Handbook*. For specific details refer to:

1. *Department of Computer Science Academic Honesty Policy* (http://www.bgsu.edu/arts-and-sciences/computer-science/policies-for-current-students.html)
2. *BGSU Code of Academic Conduct* (http://www.bgsu.edu/content/dam/BGSU/student-handbook/documents/Academic-Code-of-Conduct-Chapter.pdf)
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 post class notes, so if you miss a class please borrow notes from another student.

Office Hours and Help: Please check Canvas, Canvas messages, and your BGSU email **regularly**. *[You may have your Canvas messages forwarded to your BGSU/other email, if necessary, but do check it (multiple times) daily.]* I forward my own Canvas messages to my BGSU email and check my BGSU email multiple times everyday (with rare exceptions). 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 if you are unable to do so) must contact me **prior to the exam** with the reason. Make-ups are considered for **health emergencies only** and **require a written note from a medical professional**. There are **no** exceptions to this policy.

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 12	Introduction		
	R	Jan 14	Big Data Analytics overview		
2	T	Jan 19	GFS	**PS1**	
	R	Jan 21	HDFS Architecture/API		
3	T	Jan 26	**PS 1** discussion		**PS1**
	R	Jan 28	**RAA 1**		**PS1 reflection**
4	T	Feb 2	MapReduce		
	R	Feb 4	Hadoop 1.x / YARN		
5	T	Feb 9	Hadoop Streaming w/ Python	**PS2**	
	R	Feb 11	Sawzall / Boa		
6	T	Feb 16	**PS 2** discussion		**PS2**
	R	Feb 18	**RAA 2**		
7	T	Feb 23	Spark		**PS2 reflection**
	R	Feb 25	No Class		
8	T	Mar 1	Pregel / GraphX	**PS3**	
	R	Mar 3	FlumeJava / CloudDataflow		
9	T	Mar 8	No Class – Spring Break		
	R	Mar 10			
10	T	Mar 15	**PS 3** discussion		**PS3**
	R	Mar 17	**RAA 3**		**PS3 reflection**
11	T	Mar 22	Column-oriented Storage		
	R	Mar 24	Dremel / Impala		
12	T	Mar 29	Apache Pig	**PS4**	
	R	Mar 31	Hive		
13	T	Apr 5	**PS 4** discussion		**PS4**
	R	Apr 7	**RAA 4**		**PS4 reflection**
14	T	Apr 12	NoSQL		
	R	Apr 14	Amazon Web Services (AWS)		
15	T	Apr 19	Bigtable / HBase	**PS5**	
	R	Apr 21	Cassandra		
16	T	Apr 26	**PS 5** discussion		**PS5**
	R	Apr 28	**RAA 5**		**PS5 reflection**
17	R	May 5	**Project Presentations – 3:30pm–5:30pm – Olscamp 208**		

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