MSA 6500 Spring 2018 – Big Data Analytics

Contact Information

Instructor	Dr. Robert Dyer
Office Hours	MWF 2:20-3:30pm, TR 3:45-5p OR by appointment
E-mail	rdyer@bgsu.edu
Office	HAYES 244
Phone	(419) 372–3469

Class Meeting Time

Tuesdays/Thursdays, 2:30-3:45pm, OLSCAMP 208

Textbook

None! We will rely on instructor provided PDFs.

Outcomes for the course

After successfully completing MSA/CS 6500, students will be able to say:

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

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	15%
Problem Set Reflections	25%
Team-Based Assessments (TBAs)	30%
Projects	30%

Grading Scale

Percentage	Grade
90 - 100%	A
80 - 89%	B
70 - 79%	C
60 - 69%	D
below 60%	F

Teams

This course relies heavily on team work. First day of class we will form TWO teams for the entirety of the semester. You will work with one team on all course projects. You will work with the other team for in-class activities such as the TBAs and problem set discussions. Team members are expected to all contribute equally and team members will be rating each other's effort to help ensure fairness.

Assessments

Readings

Readings are absolutely essential to learning in this course. Almost every lecture there are readings due, meaning there are approximately 2 readings due each week. Readings will be done on the website Perusall (note: **only access Perusall via the links in Canvas!**) 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 an insightful question, answering other student's questions, finding problems with the text, etc. Your annotations should be spread out through the whole assigned reading (and not all in one 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 grade everyone's solution based on completeness. We will then spend the first portion of the class time in teams, discussing the solutions. Each team will write up a team solution. I will then provide official solutions and allow further discussion. Your solution and the team's solution are not graded. It is quite ok to have a wrong solution, 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 own knowledge, and seek additional help. The reflection is graded (using a multiplier based on your completeness score).

Team-Based Assessments (TBAs)

To help me gauge where the class is at, we will have in-class assessments at regular intervals. These will contain a small number (5-10) of relatively difficult questions. Students will take the assessment in the first half of the class by themselves. You will then collaborate with your team members and re-take the assessment as a team. The score will be the average of your individual score and the team score.

Projects

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

Technology

Canvas

The syllabus, all assignments, and due dates are posted on Canvas. Your grades will also be available on Canvas throughout the semester. Canvas is the main entry point for this course - everything you need to do is linked and organized from the Canvas course. Always start there!

Perusall

The reading annotation website we use is Perusall. To access Perusall, please click the link for a particular reading assignment in Canvas. This should open the reading (in Perusall) in a new tab/window. THERE IS NO NEED TO CREATE A USER ON PERUSALL (and doing so will break your grading!).

Perusally provides some documentation to explain what kind of comments we are looking for and also some details on the grading.

Plickers

Each student will be assigned their own Plickers card. Plickers cards are 3D barcodes, and depending on the orientation of the card (4 possible sides can face up) you are able to respond to questions with answers A, B, C, or D. This allows quick, interactive feedback from the class. I also use these to quickly record attendance near the start of each class.

Docker

All course related technologies will be access via a Docker image. Docker is a lightweight virtual machine, and will provide consistent access to the course technologies.

Course Policies

Withdrawal Deadline

Friday, April 6, 2018. University policy states that after this date, anybody withdrawing from the course will have the grade automatically turn into an F.

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 BGSU/other email, and have your BGSU email forwarded to another favorite email address, if necessary, but do check it (multiple times) daily.] I do forward my own Canvas messages to my BGSU email and check my BGSU email multiple times everyday (with rare exceptions). I check BGSU email more often than I access Canvas, so if you need to contact me urgently, use both Canvas and BGSU email, if necessary multiple times. 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.

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.

Make-up policy

If you cannot take an RAA 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. Note however that any make-up RAA will count 100% toward your score (there will be no averaging with the team RAA score). Make-ups are considered typically for health emergencies only.

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 details refer to:

1. Department of Computer Science Academic Honesty Policy
2. BGSU Code of Academic Conduct
3. The Academic Charter, section B-I.G

Disability Policy

In accordance with the University policy, students with disabilities must verify their eligibility through the Office of Disability Services, 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).

Tentative Course Schedule

Week	Day	Date	Topics
1	T	Jan 9	Introduction
	R	Jan 11	Big Data Analytics overview / Python intro
2	T	Jan 16	GFS
	R	Jan 17	HDFS Architecture/API
3	T	Jan 23	**PS 1 discussion**
	R	Jan 25	**RAA 1**
4	T	Jan 30	MapReduce
	R	Feb 1	Hadoop 1.x / YARN
5	T	Feb 6	Hadoop Streaming w/ Python
	R	Feb 8	Sawzall / Boa
6	T	Feb 13	**PS 2 discussion**
	R	Feb 15	**RAA 2**
7	T	Feb 20	FlumeJava / CloudDataflow
	R	Feb 22	Spark
8	T	Feb 27	Spark Streaming
	R	Mar 1	Pregel / GraphX
9	~~T~~	~~Mar 6~~	~~No class - Spring Break~~
	~~R~~	~~Mar 8~~	~~No class - Spring Break~~
10	T	Mar 13	**PS 3 discussion**
	R	Mar 15	**RAA 3**
11	T	Mar 20	Column-oriented Storage
	R	Mar 22	Dremel / Impala
12	T	Mar 27	Apache Pig
	R	Mar 29	Hive
13	T	Apr 3	**PS 4 discussion**
	R	Apr 5	**RAA 4**
14	T	Apr 10	NoSQL
	R	Apr 12	Amazon Web Services (AWS)
15	T	Apr 17	Bigtable/ HBase
	R	Apr 19	Cassandra
16	T	Apr 24	**PS 5 discussion**
	R	Apr 26	**RAA 5**
17	W	May 2	3:30-5:30pm **Project Presentations**