CIS 8795: IT Infrastructure for Big Data

Table of Contents

CIS 8795: IT Infrastructure for Big Data....1
1 Catalog Description....2
1.1 Prerequisites....2
1.2 Sections....2
2 Instructor....2
2.1 Contact the instructor... Please!....2
2.2 Course web site....2
2.3 Course materials....2
3 Overview....2
3.1 Intended audience....2
3.2 Learning objectives....3
4 Schedule Sessions & Readings....3
4.1 References....5
4.2 Lynda.com....5
4.3 Software....5
5 Evaluation....5
6 In class labs & demos....7
7 Big Data Topic....7
8 Homework....9
8.1 Data notebooks....9
8.2 Big Data Analytics proposal....9
8.3 Big Data Analytics presentation....9
8.4 Big Data Analytics project deliverable M1....9
8.5 Big Data Analytics project deliverable M2....9
9 Examinations....9
9.1 Quizzes....9
9.2 Exams....9
10 How to scan Computing literature....9
10.1 Software....9
10.2 Literature review....9
11 Failing and Late Work....10
12 Workload Expectations....10
13 Student Behavior....10
13.1 Discrimination and harassment....11
13.2 Official CIS department class policies....11

As with any document, be aware that this may contain clerical errors. Please tell me if you spot one.

The instructor reserves the right to modify the syllabus as necessary to improve student learning and provide appropriate evaluation. Students will be notified of any such modification in-class and via the web site.

1 Catalog Description

The current university catalog description of this course can be obtained in the University's Catalog:

http://www.gsu.edu/es/catalogs_courses.html

A recent university catalog description follows:

Big Data infrastructure. Prerequisite: Programming and database courses. This course builds upon the student's foundation of programming and database principles for large-scale data analysis. Major areas covered include distributed data infrastructure design and usage, and parallel processing large-scale data over computing grids. Managerial and strategic concerns of big data grids will be introduced. Students will design and implement distributed data storage and processing schemes and demonstrate their effectiveness with a Big Data application.

1.1 Prerequisites

Required: CIS 3270 (pre-requisite) and CIS 8040 (co-requisite); assumes some good programming and database skills

1.2 Sections

Days	Time	Room
Monday	5:30 pm – 9:45	Buckhead 306

2 Instructor

Dr. William N. Robinson; http://wrobinson.cis.gsu.edu; wrobinson@gsu.edu

Office (404) 413-7374; Dept: (404) 413-7360; FAX: (404) 413-7394

Office hours: TBA & by Appointment. Send me **Instant Messaging** *(Skype contact wrobinson@gsu.edu).*

2.1 Contact the instructor… Please!

During the term, it is **highly recommended** that you contact the instructor, in-person or via email. I am available to help you focus your projects, gain access to resources, and answer your questions. Please try to see me, phone me, or e-mail me at least once during the term to discuss your project. Your class members are also a good source of help.

2.2 Course web site

Web sites for our course is on GSU's iCollege and uses a combination of files, and videos to deliver content.

2.3 Course materials

Lecture notes, exams, assignments are all intellectual materials, which have copyright. Student are explicitly not permitted to share these materials. Sharing of materials will be addressed by the university legal services. Punishment may include removal of class credit and prosecution of copyright infringement.

3 Overview

This class covers the principles of Big Data infrastructure emphasizing distributed data processing, which includes distributed algorithm and data design and processing using patterns, techniques, and tools, and demonstrating applications for large-scale data analysis.

3.1 Intended audience

Anyone with a keen interest in Big Data design and analysis using distributed data networks will do well in this course. It's mainly geared towards the needs of Big Data Analysts and their ability to design and use data grids.

3.2 Learning objectives

Upon successful completion of this course, you will accomplish the following objectives and outcomes. In particular, students who complete this course will gain "Ready for work" skills (along with theory), including:

1. Design and processing large-scale, distributed data
2. Applying patterns, techniques, and tools to design and use distributed data networks

Specific objectives include the following:

1. Understanding distributed data networks
    a. Hadoop spark ecosystem
    b. Hadoop distributed file system
    c. Heterogenous data
    d. Data distribution patterns
    e. MapReduce
2. Processing distributed data
    a. Data ingest, compression
    b. Processing patterns
    c. Data transformations
    d. Data mining interfaces with distributed data networks
    e. Distributed data applications
3. Big Data in Context
    a. Strategic goals of distributed data infrastructure
    b. Management and applications using distributed data infrastructure
4. Programing and tools
    a. Data cluster
    b. Hadoop
    c. Spark
    d. Analytic Notebooks (PySpark, Scala, R, SQL, etc.)
    e. Code infrastructure: Docker, Kubernetes, GitHub, Docker Hub
5. Demonstrate critical thinking, integrative reasoning, & communication skills

4 Schedule Sessions & Readings

The following defines the schedule. However, the topics and readings may change according to the interests and abilities of the class. See the Academic Calendar.
On the web, the underlined items link to supporting information. Materials may be updated 24 hours prior to class; please check before attending class.

Readings provide content for class discussions. Thus, readings must be read prior the class. For example, week 1 readings must be read prior to class on week 1. The readings are in order of importance. Thus, where there are several readings, you may need to scan the last articles.

Don't get more than one week ahead of the class in the readings. Readings may be changed one week prior to their presentation in class.

Below, some readings are web link, which can be clicked to obtain. Other references are publications, listed below, which may be accessed through library.gsu.edu or purchased.

		Topic	Due
Big Data Programming	11-Oct	**1. Introduction with PySpark Notebooks** 1.1. Introduction to big data (online PDF) 1.2. **su**: Watch LinkedIn Learning: Big Data in the Age of AI 1.3. Python vs PySpark DataFrame: §1-4 (Chambers and Zaharia 2018), **su** DataFrame diffsv, cheat sheet 1.4. Python vs PySpark, regression	Quiz
	18-Oct	**2. Big Data Programming** 2.1. Watch LinkedIn Learning: Apache PySpark by Example 2.2. Spark Data: §5-13 (in §13 stop @ CoGroups)	hw: Notebook 1 Quiz
	25-Oct	**3. Big Data Analytics** 3.1. Specifying features: §24-25 3.2. Classification: §26 3.3. Regression: §27 3.4. Unsupervised learning: §29	hw: Notebook 2 Quiz
Big Data DevOps	1-Nov	**4. Cluster Development** 4.1. ML continued 4.2. Processing on a Cluster §15 – 19 4.3. https://docs.databricks.com/dev-tools/index.html 4.4. Programming guide: production job 4.5. Watch LinkedIn Learning: Cloud Architecture: Advanced Concepts 4.6. Install Miniconda (WSL) 4.7. Install Spark (windows, Mac, Linux) 4.8. Install Docker (on Windows, install WSL 2; see timestamps) 4.9. Install PyCharm (configure for spark, config WSL) 4.10. Install Google Cloud SDK (windows, install in WSL2 via Ubuntu) 4.11. Install gCloud (video)	hw: Notebook 3 **Exam**
	8-Nov	**5. Cluster DevOps** 5.1. Google Cloud Platform (dataproc, GKE) 5.2. Streaming data §20 – 23 5.3. Data pipelines	hw: Cluster 1 Group Proposal Quiz
Big Data Architecture	15-Nov	**6. Cluster Management: YARN, Kubernetes** 6.1. MapReduce §2 *sc* 7 (White 2014) 6.2. Spark job: §15 (Chambers and Zaharia 2018)	hw: Cluster 2 hw: M1 Topic Presentations* Quiz
	29-Nov	**7. Cluster Components: HDFS, S3** 7.1. HDFS §1,3 (White 2014) 7.2. https://www.flackbox.com/object-storage-overview-tutorial	Topic Presentations* Quiz
	6-Dec	**8. Evaluation**	M2 Presentations **Exam**
		9. Grades due Dec 16	**Due Saturday (11th):** hw: M2 hw: Topic article

- Exams and Quizzes are presented at the end of class

hw means homework, due at the beginning of class
sc means *scan the text* (so that you can use it as a reference later, if needed).
su means read **only if you need** supplemental information to clarify misunderstandings
§ means chapter or section number. Note that much of the Chambers book is reference material. So, ensure that you run the examples, but you can skim the reference material.
§ **without a reference, refers to the main textbook**: (Chambers and Zaharia 2018).
* topic presentations scheduled throughout the term; topic paper (based on presentation) due at end of term.

Lecture notes are available online. Each file is prefixed with the session in which it is to be presented.

4.1 References

The readings are two main textbooks, supplemented with published articles and blog posts.

The two textbooks are available on Kindle at Amazon.com:

For PySpark coding (first part of class):

- Chambers, Bill, and Matei Zaharia. 2018. *Spark: the definitive guide: big data processing made simple* (" O'Reilly Media, Inc.").

For cluster infrastructure (latter part of class):

- White, Tom. 2014. *Hadoop: The definitive guide* ("O'Reilly Media, Inc.").

Texbooks

Chambers, Bill, and Matei Zaharia. 2018. *Spark: the definitive guide: big data processing made simple* (" O'Reilly Media, Inc.").

White, Tom. 2014. *Hadoop: The definitive guide* ("O'Reilly Media, Inc.").

4.2 Linked In Learning

GSU students have access to Linked In Learning, which has many online courses. Go the http://www.gsu.edu and search for Linked In Learning. The current link is:

- https://technology.gsu.edu/technology-services/it-services/training-and-learning-resources/linkedin-learning/

4.3 Software

Additionally, much of the software is available for download, either from the instructor, or from the CIS agreements with MSDNAA and the IBM Academic Initiative.

- Spark notebooks
  - https://community.cloud.databricks.com/login.html
- Docker
  - https://training.play-with-docker.com/
  - https://labs.play-with-docker.com/
- Kubernetes
  - https://labs.play-with-k8s.com/
  - https://www.katacoda.com/courses/kubernetes/playground
- GSU virtual machines (with Cloudera)
  - https://cis-vctr6app.gsu.edu/

5 Evaluation

Students are evaluated by the deliverables summarized in Table 1. The course credits are earned according to the following Table 1.

Table 1 Relative weights assigned to course deliverables.

	Assignment	Percentage
Individual	Quizzes	20
	Exam 1	20
	Exam 2	20
	Data notebooks (individual with team aid)	5
Team	Topic article (team)	5
	Topic presentation (team)	5
	Big Data Analytics proposal (team)	0
	Big Data Analytics presentation (team)	5
	Big Data Analytics project deliverable M1 (team)	10
	Big Data Analytics project deliverable M2 (team)	10
	Total	100

The following table overviews how credit will be assigned. Note that all group work includes a peer review, which can distinguish an individual's assigned points from the group's assigned points. (See Self-Managed Teams in the Failing and Late Work section.)

Table 2 Grading standards.

Work quality	Percent
Absolutely fantastic, walk on water, overflow grade	110
Excellent answer on all counts	100
Excellent answer on most counts	90
Very good answer, but not excellent	80
Professionally done and adequate	70
Inadequate, needs work	60
Varying degrees of inadequacy	0 - 50

The following breakout depicts how grades will be assigned under this system.

Grade	Percentage
A+	≥ 97
A	≥ 90
A-	≥ 87
B+	≥ 83
B	≥ 80
B-	≥ 77
C+	≥ 73
C	≥ 70
C-	≥ 67
D	≥ 60
F	< 60

6 In class labs & demos

Each exercise is intended as a group effort, which illustrates important concepts introduced in the associated readings. **More detailed description and associated materials is on the course web site**.

- **Deliver** your results to the course web site.
  - Authors shall receive credit for each in-class exercise.
  - Prominently (at the top) of the delivered document, place the names of authors.
  - Do not include the name of anyone who is absent or did not contribute. Doing so will result in zero credit for all 'authors'.
  - Late deliverables (after class) shall receive a 10% deduction per day; after the third day the deduction is 100% (zero credit).

7 Big Data Topic

Your student team will describe concept, method, or tool. See our web site for some ideas. You can find your own topic as well. You will deliver:

- a paper
- a PowerPoint presentation summary (10 - 25 slides)

The goal is to have students know some important topics in detail, through research and presentation, while the entire class becomes familiar with a wide range of topics. Please ensure that your research goes beyond that presented in the class discussion and readings.

Your group will present a Big Data topic that has been in the wider-press headlines or technical articles within the last 5 years. The goals of the assignment are to:

- Show the relevance of Big Data for everyone
- Present Big Data course materials in the context of real, ongoing, problems
- Generate discussion about Big Data—in particular, tradeoffs, decision-making, and consequences of Big Data for organizations and people
- May includes technical demonstrations

In your presentation:

- Show a news article(s), blogs, scholarly article, demonstration program, etc.
- Present a about 15 PowerPoint slides summarizing the source, big data and/or societal issues, and provide for questions for subsequent discussion
- Moderate a brief discussion

Deliver to our web site:

- Your PowerPoint slides
- Any notes that might be relevant to aid further study

7.1.1 Topic article

Your class peers are the intended audience for this article. They are among the best IT managers, engineers, and scientists. The will want to know the details of the method you describe, as well as "why do I care?" ☺

- See the web site for a list of potential topics. However, you may define our own topic. Note: each topic must be unique; thus, no two groups will present the same topic.
- Topics will be scheduled throughout the term
- The paper is due at the end of the term

In describing your topic, please consider the following aspects:

- Answer questions of what, why, when, where, how, and who.
- Describe both theory and computer tool support.
- Describe implications for practitioners, i.e., does the method really help?
- Include at least **two** academic references (peer-reviewed articles) in your research.
- Do not summarize (substantially) the course materials; assume them as background and add new materials.
- Write the article using the wikipedia.org *style* (meaning, a short, to the point, encyclopedia description).
- Relate the topic material to the course materials

Ensure that your article:

- Style, grammar, and spell checked (in Word)
- Minimum of 1,500 words; about 3 pages (single spaced)
    - figures, tables, and references do not included in page count
    - Liberal use of quotations are allowed; however, the **sources must be referenced**. No more than 10 percent of the paper word count can be quotations
- References to articles using a standard format (e.g., Chicago); see EndNote.

Include appropriate web links and article references using EndNote; See How to scan Computing literature.

- **Deliver** your article to the instructor as an email attachment
    - The article must be a Word document. Ensure that it is spell and grammar checked.

7.1.2 Topic presentation

- **Deliver** a PowerPoint presentation of your topic on the scheduled date
    - You will have about 15 minutes, depending on class size
- **Deliver** your PowerPoint slides to our web site

7.1.2.1 Presentation Guidelines

The project presentation should reflect your article. Be prepared to setup, present, and leave the presentation areas within the allotted time, depending on the number and type of presentations. Poorly prepared groups will find they must end their presentation before getting to their most important points.

Consider the following structure for your presentation:

1. Introduction
2. 1 Summary
3. < topic 1 details> …
4. Take Away Points (1 – 3)
5. 2 Summary

6. < topic 2 details> …
7. Take Away Points (1 – 3)
8. …
9. Bibliography

Presentations shall be professional, of course—minimally, PowerPoint sides.

I recommend that your presentation include:

- Introductory, illustrative problem
- Theoretical approach to the problem
- Practical or example solutions to the problem
- Tools, techniques, or specific suggestions that can be applied in practice
- 2 Multiple Choice questions summarizing the most importance ideas

8 Homework

See the web site for the most recent and detailed information on these assignments. The following is provided as an introduction to the assignments.

8.1 Data notebooks

8.2 Big Data Analytics proposal

8.3 Big Data Analytics presentation

8.4 Big Data Analytics project deliverable M1

8.5 Big Data Analytics project deliverable M2

9 Examinations

Online review guides to be updated one-half week prior to the exam.

9.1 Quizzes

9.2 Exams

See the online exam review for a description.

10 How to scan Computing literature

10.1 Software

Install EndNote:

1. **Free** EndNote @ GSU

10.2 Literature review

Search for peer reviewed articles using keywords:

2. Scan the web
    a. www.google.com
3. Scan the web using scholar search engines
    a. http://scholar.google.com/
        i. Set the Google Scholar Preferences to
            1. Show library access links for Georgia State University

2. Show links to import citations into EndNote
   b. http://academic.live.com/
   c. http://citeseer.ist.psu.edu/
4. Scan using library databases (@GSU)
   a. http://www.galileo.usg.edu
   b. In particular, the following databases
      i. **ABI/INFORM Complete**
      ii. **ACM Digital Library**
      iii. **IEEE Xplore**

11 Failing and Late Work

Work that is delivered late will receive a 10 percent deduction per date until 5 days, at which time it will receive a zero.

Students who deliver a project deliverable that receives a failing grade will be given one-half week to resubmit the work (after it is graded). A second submission will receive a late penalty deduction.

12 Workload Expectations

Students should plan for 2 - 3 hours of work outside of class each week for each course credit hour. Thus, a 3-credit course averages between 6 and 9 hours of student work outside of the classroom, *each week*. See GSU site for Academic Success. Students must take responsibility for their learning. In contrast to high school, college has fewer opportunities for student teacher interactions. Consequently, students must prepare to gain the most from each interaction.

Self-Managed Teams: Teams will be allowed for some activities during the term. Please note that unless the activity is explicitly identified as a "team activity", I expect everyone to perform their own work (your hands on the keyboard). For team activities, you will be allowed to work with partners (of your choosing).

- Initial teams must be established by the second week of classes. Established teams may continue working together on subsequent team activities. Team membership may change during the term, if problems arise. However, team members must be designated within one week of the due date for the team activity. Exception: you may withdraw from a team at any time and submit an assignment individually.
- Teams will submit one assignment for all team members. In most cases, each member of the team will get the same score. However, an individual's score may be reduced at the discretion of the instructor.
- Each team assignment **must include the following**:
  - Tasks completed by each member.
  - Percentage of the total work completed by each member.
- Any individual with a low team contribution will be removed from their team.

Arbitration: There will be a one-week arbitration period after graded activities are returned. Within that one-week period, you are encouraged to discuss any assumptions and/or misinterpretations that you made on the activity that may have influenced your grade.

Attendance: If you are unable to attend a class session, it is your responsibility to acquire the class notes, assignments, announcements, etc. from a classmate. The instructor will not give private lectures for those that miss class.

Submission of Deliverables: Unless specific, prior approval is obtained, no deliverable will be accepted after the specified due date.

If you have a legitimate personal emergency (e.g., health problem) that may impair your ability to submit a deliverable on time, you must take the initiative to contact the instructor before the due date/time (or as soon after your emergency as possible) to communicate the situation.

Make-up exams will not be given: However, if a student has a planned absence, he or she may take the exam **earlier** with the permission of the instructor.

13 Student Behavior

Behavior in class should be professional at all times. People must treat each other with dignity and respect in order for scholarship to thrive. Behaviors that are disruptive to learning will not be tolerated and may be referred to the Office of the Dean of Students for disciplinary action.

13.1 Discrimination and harassment

Discrimination and/or harassment will not be tolerated in the classroom. In most cases, discrimination and/or harassment violates Federal and State laws and/or University Policies and Regulations. Intentional discrimination and/or harassment will be referred to the Affirmative Action Office and dealt with in accordance with the appropriate rules and regulations.

Unintentional discrimination and/or harassment is just as damaging to the offended party. But, it usually results from people not understanding the impact of their remarks or actions on others, or insensitivity to the feelings of others. We must all strive to work together to create a positive learning environment. This means that each individual should be sensitive to the feelings of others, and tolerant of the remarks and actions of others. If you find the remarks and actions of another individual to be offensive, please bring it to their attention. If you believe those remarks and actions constitute intentional discrimination and/or harassment, please bring it to my attention.

13.2 Official CIS department class policies

1. Prerequisites are strictly enforced. Students failing to complete any of the prerequisites with a grade of "C" or higher will be administratively withdrawn from this course with *loss of tuition fees*. **There are no exceptions, except as granted by the instructor with the approval of the department.**
2. Students are expected to attend all classes and group meetings, except when precluded by emergencies, religious holidays, or bona fide extenuating circumstances.
3. Students who, for non-academic reasons beyond their control, are unable to meet the full requirements of the course should notify the instructor, by email, as soon as this is known and prior to the class meeting. Incompletes may be given if a student has ONE AND ONLY ONE outstanding assignment.
4. A "W" grade will be assigned if a student withdraws before mid-semester if (and only if) he/she has maintained a passing grade up to the point of withdrawal. Withdrawals after the mid-semester date will result in a grade of "WF". See the GSU catalog or registrar's office for details.
5. Spirited class participation is encouraged and informed discussion in class is expected. This requires completing readings and assignments **before** class.
6. All exams and individual assignments are to be completed by the student alone with **no** help from any other person.
7. Collaboration within groups is encouraged for project work. However, collaboration between project groups will be considered cheating.
8. Copying work from the Internet without a proper reference is considered plagiarism and subject to disciplinary action as delineated in the GSU Student Handbook.
9. Any non-authorized collaboration will be considered cheating and the student(s) involved will have an Academic Dishonesty charge completed by the instructor and placed on file in the Dean's office and the CIS Department. All instructors regardless of the type of assignment will apply this Academic Dishonesty policy equally to all students. Abstracted from GSU's Student Handbook Student Code of Conduct "Policy on Academic Honesty and Procedures for Resolving Matters of Academic Honesty"
    a. http://www2.gsu.edu/%7Ewwwdos/codeofconduct_conpol.html
    b. http://www2.gsu.edu/~wwwcam/

    As members of the academic community, students are expected to recognize and uphold standards of intellectual and academic integrity. The University assumes as a basic and minimum standard of conduct in academic matters that students be honest and that they submit for credit only the products of their own efforts. Both the ideals of scholarship and the

need for fairness require that all dishonest work be rejected as a basis for academic credit. They also require that students refrain from any and all forms of dishonorable or unethical conduct related to their academic work.

Students are expected to discuss with faculty the expectations regarding course assignments and standards of conduct. Here are some examples and definitions that clarify the standards by which academic honesty and academically honorable conduct are judged at GSU.

Plagiarism. Plagiarism is presenting another person's work as one's own. Plagiarism includes any paraphrasing or summarizing of the works of another person without acknowledgment, including the submitting of another student's work as one's own. Plagiarism frequently involves a failure to acknowledge in the text, notes, or footnotes the quotation of the paragraphs, sentences, or even a few phrases written or spoken by someone else. The submission of research or completed papers or projects by someone else is plagiarism, as is the unacknowledged use of research sources gathered by someone else when that use is specifically forbidden by the faculty member. Failure to indicate the extent and nature of one's reliance on other sources is also a form of plagiarism. Any work, in whole or part, taken from the Internet or other computer based resource without properly referencing the source (for example, the URL) is considered plagiarism. A complete reference is required in order that all parties may locate and view the original source. Finally, there may be forms of plagiarism that are unique to an individual discipline or course, examples of which should be provided in advance by the faculty member. The student is responsible for understanding the legitimate use of sources, the appropriate ways of acknowledging academic, scholarly or creative indebtedness, and the consequences of violating this responsibility.

Cheating on Examinations. Cheating on examinations involves giving or receiving unauthorized help before, during, or after an examination. Examples of unauthorized help include the use of notes, texts, or "crib sheets" during an examination (unless specifically approved by the faculty member), or sharing information with another student during an examination (unless specifically approved by the faculty member). Other examples include intentionally allowing another student to view one's own examination and collaboration before or after an examination if such collaboration is specifically forbidden by the faculty member.

Unauthorized Collaboration. Submission for academic credit of a work product, or a part thereof, represented as its being one's own effort, which has been developed in substantial collaboration with another person or source or with a computer-based resource is a violation of academic honesty. It is also a violation of academic honesty knowingly to provide such assistance. Collaborative work specifically authorized by a faculty member is allowed.

Falsification. It is a violation of academic honesty to misrepresent material or fabricate information in an academic exercise, assignment or proceeding (e.g., false or misleading citation of sources, the falsification of the results of experiments or of computer data, false or misleading information in an academic context in order to gain an unfair advantage).

Multiple Submissions. It is a violation of academic honesty to submit substantial portions of the same work for credit more than once without the explicit consent of the faculty member(s) to whom the material is submitted for additional credit. In cases in which there is a natural development of research or knowledge in a sequence of courses, use of prior work may be desirable, even required; however the student is responsible for indicating in writing, as a part of such use, that the current work submitted for credit is cumulative in nature.