Department of Computer Science and Cybersecurity
College of Health, Science, and Technology
University of Central Missouri
Course Syllabus
DSA 5620 – Big Data Analytics

Instructor: Dr. Mohammad Rawashdeh
Email: rawashdeh@ucmo.edu
Office: MIC D158
Office Hours:
TBD
Other times by appointment

Accessibility Services:

Students with disabilities who are seeking accommodation should contact the Office of Accessibility Services at Union 222, 660-543-4421. If you want to share information about your needs that I should be aware of, such as emergency medical information or special arrangements for field trips or internships, please see me privately after class or during office hours.

Sexual Discrimination and Sexual Misconduct Statement:

The University of Central Missouri seeks to foster a safe and healthy environment built on mutual respect and trust. Sex discrimination, including sexual harassment, sexual violence, and other forms of sexual misconduct will not be tolerated. All faculty and most staff are considered mandated reports by the University and must disclose all information they receive about sexual misconduct to the Title IX Coordinator. As a faculty or staff member of the University, I am a mandated reporter. This means I am required to report information shared with me regarding sex discrimination and sexual misconduct.

If you, or someone you know, has experienced sex discrimination or sexual misconduct, please know assistance and options are available. UCM strongly encourages all members of the community to seek support and report incidents of this nature to the Title IX Coordinator. Anyone who wishes to report sexual misconduct, to learn more about the University process and options available, or to utilize a confidential resource, please visit http://ucmo.edu/titleix

Diversity, Equity, and Inclusion:

The University of Central Missouri strives to develop a campus environment that welcomes and recognizes all dimensions of diversity and inclusiveness. What this means is that all students are welcomed in the classroom, and differences are to be recognized rather than erased or denied. Dimension of diversity can include sex, race, age, national origin, ethnicity, gender identity and expression, intellectual and physical ability, sexuality, income, faith, and non-faith perspectives, socio-economic class, primary language, family status, military experience, and more. Inclusive learning is facilitated by creative and innovative thought and mutual respect; being in this classroom means that you, your faculty member, and your peers pledge to foster a welcoming and equitable environment for all.

Purpose of the Course:
The course covers Spark and the key components of the Hadoop ecosystem that are used in Big Data applications.

Objectives:
During this course, students will learn:

- An overview of the various tools and framework available under Hadoop umbrella: the purpose of each tool and where it fits.
- About Hadoop core components, namely, HDFS, MapReduce, and Yarn: architecture, daemons, how to start-stop, basic commands, modes of operation, etc.
- How to develop Hadoop applications in Java MapReduce, submit, monitor, and access their output.
- About Hive services, modes of operation, basic configurations, and limitations.
- About Hive CLI commands, and HiveQL statements
- About Spark deploy modes and architecture
- How to start and use Spark python shell
- How to write applications in Spark using its RDD, DataFrame, and Dataset APIs
- How to write queries in Spark SQL
- How to perform simple data streaming using Kafka or Spark streaming.
- How to perform common machine learning tasks in Spark.

References:
Required

- Hadoop: The Definitive Guide, by Tom White
- Programming Hive, by Capriolo, Wampler, and Rutherglen
- Learning Spark, by Karau, Konwinski, Wendell, and Zaharia

Optional

- Advanced Analytics with Spark, by Ryza, Laserson, Owen, and Wills

Note: UCM students have free access to all of the listed books above and more on O'Reilly Learning platform. If the online platform is not working, it is the student responsibility to obtain a copy of the required text.

Course Outline:

A. Introduction to Hadoop Ecosystem
    a. History
    b. Why Hadoop?
    c. Where it Fits and Doesn't Fit
    d. An Overview of Hadoop Various Services and Tools
B. HDFS
    a. HDFS Blocks, and Replication
    b. Daemons
    c. HDFS Federation
    d. High Availability
    e. Rack Awareness
    f. Basic HDFS commands
    g. Moving Files Between HDFS and Other File Systems
C. MapReduce
    a. An Abstract Model of Computation
    b. Data Flow and Phases
    c. Tasks, and Task Attempts
    d. Examples of MapReduce Applications
    e. Optimization Techniques: Data locality and Speculations
    f. Hadoop Streaming
D. YARN
    a. YARN Daemons
    b. Anatomy of a YARN Application Run
    c. YARN Compared to MapReduce 1
    d. Scheduling in YARN: FIFO, Capacity, and Fair
E. Hive
    a. Why Hive? WordCount Example in Java Versus Hive
    b. Architecture, Configuration, and Services
    c. Hive CLI, and Beeline
    d. Data Types and File Formats
    e. HiveQL DDL
    f. Hive Tables: Partitioned, Managed, and External
    g. HiveQL DML: Loading Data, Writing, and Submitting Different Queries
F. Spark
    a. Introduction to Unified Analytics with Spark
    b. Getting Started with Spark
    c. Spark Structured APIs: RDDs, DataFrames, and Datasets
    d. Spark SQL
G. Data-Streaming with Kafka and Spark: Introduction, and Basics (Producer-Consumer, Connectors, Zookeeper, etc.)
H. Machine Learning and Data Analysis:

Statistics, Feature Extraction, Measuring Similarity, Classification and Regression, Clustering, Association-Rule Mining, Collaborative Filtering and Recommendation, Dimensionality Reduction, Model Evaluation, etc.

Procedures and Assessments:

- **Midterm and Final Exams (200 pts.)**
- **Quizzes, Homework Assignments, and Attendance (100 pts.):**
- **Grading Scale:** your final grade will be determined by the percentages

Percent	Grade
90% and above	A
80% – 89.99%	B
70% – 79.99%	C
60% – 69.99%	D
Below 60%	F

Attendance Policy:

Attendance will be taken every class. Students are expected to attend every class on time either face-to-face or online. You don't have to ask for a permission if you must skip a class but I will appreciate it if you let me know about it. If you are an international student, please be aware that **USCIS** might request your attendance records for the time you spent at UCM.

Academic Dishonesty:

Academic dishonesty will not be tolerated in this class. Any form of cheating will be dealt with according to the Academic Dishonesty Policy on-line at https://www.ucmo.edu/offices/general-counsel/university-policy-library/academic-policies/academic-honesty-policy

Other Policies:

1. Your UCM email account will be used, frequently, by the instructor, to communicate messages. It is your responsibility to check this account regularly.
2. Class notes and assignments will be posted by email and on Blackboard. It is the responsibility of the student to frequently check their email and Blackboard for course changes and updates.
3. The assigned textbook(s) is required, either physical or digital.
4. Advanced arrangement for unavoidable absences should be made whenever possible. **Neither absence nor notification of absence relieves you** of the responsibility of meeting all course requirements.
5. You are expected to **attend every class**. Your attendance for every class will be marked as either: face-to-face, online, or absent. You are

allowed **one** unexcused absences. Beyond that, you will start losing points.

6. Make-up tests will be given only for absences if determined by the instructor to be acceptable. Scheduling of the make-up test is the student's responsibility. .Advance arrangements for unavoidable absence(s) should be made whenever possible. Contact me by email prior to the class you would miss. Make sure to bring a doctor's note for the absence caused by health issues to the office below. For official process of documented absence, contact:
   The office of student experience and engagement,
   Admin Bld. 214,
   Mrs. Keri Busker (busker@ucmo.edu).
7. In case if any exam, or a quiz, is scheduled online, Honorlock might be used to proctor the exam. This doesn't apply to take-home exams. Online exams, if any, most probably will be scheduled on Mondays.
8. There will be no make-up for quizzes.
9. When you contact me via email **please make sure that you add your student and class information to your inquiry** otherwise you might not get any response.
10. Please silence all cellphones and do not text, receive calls, or make calls during class. If you must do so, you can leave the class and come back later without distributing the class.