Logo

Description automatically generated

Department of Computer and Information Sciences

COSC 760 –Big Data Analytics

**Fall 2024**

**Instructor:** Dr. Weixian Liao **Email:** wliao@towson.edu

**Office:** 7800 York Road, Room 470 **Phone:** 410.704.2774

**Office Hours:** Thursday 3:00 PM – 4:00 PM and by appointment

**Teaching Assistant**

Name: Papa Pene

Email: [ppene1@students.towson.edu](mailto:ppene1@students.towson.edu)

Office: YR 224

Office Hour: Thursday 4:00 PM – 5:00 PM and by appointment

**Credits:** 3 credits

**Course Time:** Thursday 7:30 PM - 10:10 PM

**Course Location:** 7800 York Rd Room 0302

**Course Website:** Blackboard

**Prerequisite:** COSC 578 or special permission from the instructor

**Course Description:** This course provides students with an understanding of big data. Students will gain knowledge and hands-on experience on various big data framework and NoSQL database tools and real-time stream processing tools such as Spark. The course also includes machine learning based big data analytics. Students are expected to obtain insights on producing and consuming big data collection, ingestion, storage, analytics, and application. In addition, this course explores cutting-edge research in big data analytics published in academic journals and conferences.

**Course Objectives:** Upon completion of the course, students will understand and be familiar with the general concepts associated with Big Data processing and analytics using Big Data framework and NoSQL databases. Specifically, students will be able to:

* Gain in-depth understanding about data-driven decision-making environment with big data;
* Figure out and classify big data datasets, storages, data platforms, analytics mechanisms;
* Formulate big data analytics model and choose a proper framework and tool(s) to discover hidden/implicit values;
* Perform big data analytics using big data cluster framework and real-time stream processing tools;
* Understand machine learning based big data analytical methods with machine learning libraries available on Big Data frameworks;
* Develop big data analytics applications.

**Textbook:**

Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale,

Ofer Mendelevitch, Casey Stella, Douglas Eaoline,

Addison-Wesley, 2017

**Recommended Readings:**

Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQl, and Graph.

David Loshin.

Elsevier, 2013.

Hadoop: The Definitive Guide (Fourth Edition), Storage and Analysis at Internet Scale.

Tom White

O’Reilly Publishers, 2015.

MangoDB: The Definitive Guide (Second or Third Edition), Powerful and Scalable Data Storage.

Shannon Bradshaw, Kristina Chodorow

O’Reilly Publishers, 2013/June 2017 (scheduled).

Elasticsearch in Action.

Radu Gheorghe, Matthew Lee Hinman, Roy Russo.

Manning Publications Co., 2016.

Learning Spark, Lightning-fast Data Analysis.

Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia.

O’Reilly Publishers, 2015.

Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking

Foster Provost and Tom Fawcett

O’Reilly Publishers, 2015.

**References:**Apache Hadoop, http://hadoop.apache.org/

Apache HBase, https://hbase.apache.org/

Apache Spark, http://spark.apache.org/

Apache Mahout, http://mahout.apache.org/

Apache Cassandra, http://cassandra.apache.org/

Apache CouchDB, http://couchdb.apache.org/

MongoDB, http://www.mongodb.com

Elasticsearch, https://www.elastic.co/

Hortonworks Data Platform, http://hortonworks.com/

Cloudera Data Platform, https://www.cloudera.com

MapR Data Platform, https://www.mapr.com/

TensorFlow, https://www.tensorflow.org/

**Course Context/ Background:** Over the last decade, technological advancements have enabled the collection of a vast amount of data (data-at-rest and data-in-motion) from various data sources including Cyber-Physical Systems (CPS), Internet of Things (IoTs), social media, and public/private sectors. The discovery of values in Big Data can enable a better understanding of business, society, the environment, health care, and all aspects of life and a data-driven decision making based on machine learning algorithm. This course provides students with an opportunity to build their Big Data analytical capabilities across all Big Data processing stages including its accumulation the cloud framework, data processing for analysis, and value discovery.

**Course Topics:** This course provides students with a background in the concepts and applications of Big Data analytics. Potential course topics include the following:

* Introduction to data science and Big Data analytics
* Big Data cluster framework
* Introduction to Hadoop Ecosystem/Hadoop Stack
* Big Data classification
* Hadoop Distributed File System (HDFS)
* Map/Reduce Big Data processing approach
* Hadoop Application Data Frameworks
* NoSQL databases for non-structured datasets
* Real-time stream data processing
* Introduction to machine learning algorithms
* Cutting-edge Big Data analytics trends and research directions

**Course Format:** Active learning techniques, lectures, discussion sessions (in class AND online), presentations, lab exercises, hands-on programming, and projects may be used. Students are expected to read and use current contents on the course subjects using the library, internet and provided resources. Some assignments will necessitate collaborative learning while others will require individual research and presentation.

**Attendance Policy:** Students are expected to attend all classes in order to remain current in the coursework. It is the student’s responsibility to remain current on the handouts, assignments and notes if class is missed. The instructor will allow only students with documented excuses (see below) to make up missed work or assignments *when it is feasible*. If the student is absent from an exam during the scheduled time for that exam, the student will automatically receive a grade of zero (0) for the exam unless: the student notifies the instructor of the absence prior to the exam AND the student is ill and supplies a written doctor’s excuse explaining the absence OR there is an extraordinary situation which the instructor allows as an acceptable excuse. Only under above circumstances will arrangements for a makeup exam be made.

**Labs and Assignments:** Big data analysis tasks will be assigned throughout the semester to give students hands-on experience using big data cluster framework and real-time stream data processing software tools such as Hadoop, MongoDB, CouchDB, Spark, Mahout, Elasticsearch, etc. Students are expected to select a tool as their big data clustering framework and NoSQL database, present the summary of the tools, and share the installation procedures during the correspondent lab sessions. We will also complete a number of big data analytics labs using the Python/Scala/Java to experiment with a number of datasets: data-at-rest and data-in-motion. There will be three big data analytics assignments that require students to setup a big data cluster framework and implement big data analytical approaches on real datasets. All assignments are due at the assigned time on Blackboard. All work should be thoroughly backed up before turning it in and all submissions should be well documented with proper citations. Homework may also be assigned in the lecture at the instructor’s discretion and as the need arises.

**Article Summary and Presentation:** Each student will be required to select a research article related to a course topic in the beginning of the semester. The student will write a summary of the article to be submitted to the professor and prepare a 20-minute presentation on the article. The presentation should give details about the approach presented in the article and summarize the results. During the first class, students will select a topic and sign up for a presentation slot.

**Project:** The semester project will require students to apply a big data analytics technique covered during the semester to discover values in a real-world dataset. Project examples include classification of Baltimore City crime/income data, world-wide/nation-wide climate data, clustering of stock market data, social network data, transaction data, CPS/IoT data, and many other public/private datasets. Students will be required to submit a project proposal to the instructor before carrying out any execution of the project. Projects should generally follow the steps of the big data analytics process and should give details on the data pre-processing, clustering and mapping, reducing and transforming, modeling and analyzing, discovering knowledge and value of the data, and learning from the data. This will include selecting an appropriate algorithm and a software tool based on the knowledge and value discovery problem at hand. Projects should also report on the evaluation of the approach and use visualization to present the discovered knowledge and value. Students are encouraged to use available big data cluster framework and real-time stream data processing software tools such as Hadoop, MongoDB, CouchDB, Spark, Elasticsearch, etc.

**Grading Policy:** Students will be evaluated on the following basis:

Labs: 28% (7 lab assignments)

Assignments: 10% (2 assignments)

Project: 30% (1 group project)

Paper Summary: 30% (Written summary and presentation)

Attendance/participation: 2%

Final course grades will abide by the following scale:

|  |  |  |  |
| --- | --- | --- | --- |
| A | 93 – 100 | B | 75 – 82.9 |
| A- | 88 – 92.9 | C | 60 – 74.9 |
| B+ | 83 – 87.9 | F | Below 60 |

All assigned work (i.e., labs, assignments, and projects) is expected to be completed and submitted by the stated deadline. *No late work will be accepted* and a grade of zero (0) will be given. Students are encouraged to discuss homework and projects as a means to share knowledge, experience and lessons learned as part of the learning process, but academic honesty should be strictly observed (see below).

**Professionalism:** All materials submitted for this course should be professional including the use of correct grammar and spelling. Please ensure all cell phones and other devices that could potentially disrupt the class are turned off upon entering the classroom except for laptops during the lab sessions.

**Posting of Grades:** University policy prohibits posting of grades in any form. The instructor will not report grades via email or in response to phone calls. Grades for the semester can be accessed online.

**Cheating and Plagiarism:** Academic honesty is strongly observed. This course may consist of both individual and team assignments. A team project is an assignment in which collaboration is allowed and highly encouraged. However, the work of the team *must* be of the team’s creation and not plagiarized from other sources. Individual assignments *must* reflect the work of the individual student and of his/her creation. While studying together, discussion and collaboration is encouraged, individual assignments *must be individually prepared* – copying or sharing files, diagrams and/or code is considered cheating. The penalty for cheating will, at a minimum, consist of a grade of zero for the dishonest work and may lead to the possibility of **course failure** depending on the severity. Students are responsible for reading and knowing Towson University’s policy regarding academic dishonesty, located in Appendix F in the Undergraduate Catalog and familiarizing themselves with the policies detailed at <http://cooklibrary.towson.edu/avoidingPlagiarism.cfm>.

**Classroom and Lab Policy:** Food and drink are not allowed in the labs and classrooms with the exception of water in the classroom only. All cell phones should be turned off or put on silent mode to avoid disruptions and distractions.

**Repeat Policy:** Students may not repeat a course more than once without prior permission of the Academic Standards Committee.

**Note:**

**This syllabus may change according to class needs. If you need course adaptations of accommodations because of a learning disability and/or for any reasons, please advise the instructor as early as possible in the course.**

**Tentative Schedule:** The following is a tentative schedule. Note that these topics and chapters are subject to change based on time and discretion of the instructor.

|  |  |  |
| --- | --- | --- |
| Week | **Lectures** | **Labs/Assignments:**  **Reading Article, Discussions, and**  **Development with Tools** |
| Week 1 | Course Overview/Introduction to Data Science and Big Data Analytics | Reading/Overview:  Introduction to Python and Scala |
| Week 2 | Big Data Cluster Framework | Reading/Overview:  Hadoop Ecosystem |
| Week 3 | Introduction to  Hadoop Ecosystem | LAB #1:  Hadoop Stack and Hadoop Application Data Framework |
| Week 4 | Big Data Classification  (data-at-rest & data-in-motion) | ASSIGNMENT #1: Supplemental Readings  Article Presentations |
| Week 5 | HDFS, Big Data Store | LAB #2: HDFS as a Big Data Store/MapReduce |
| Week 6 | MapReduce | ASSIGNMENT #2: MapReduce Applications |
| Week 7 | Hadoop Application Data Framework | LAB #2: HDFS as a Big Data Store  Hortonworks, Cloudera, MapR, Amazon EMR, etc. |
| Week 8 | Project Discussion | Project Discussion |
| Week 9 | NoSQL Data Management for unstructured data processing | LAB #3: NoSQL Databases  MongoDB, CouchDB, Elasticsearch, Cassandra, etc.  \*Students choose and install their preferred NoSQL database. Lab 4 |
| Week 10 | Building NoSQL Database | ASSIGNMENT #3: NoSQL Application  Unstructured data (ex, JSON documents) into the selected NoSQL Database. Lab 5 |
| Week 11 | Cyber-Physical Systems (CPS)/Internet of Things (IoTs)/Smart Grid Environment | LAB #4: Real-time Stream Data Processing for CPS/IoTs  Apache Spark, etc. |
| Week 12 | Real-time Stream Data Processing: Spark | ASSIGNMENT #4: Real-time Stream Data Processing for CPS/IoTs Application |
| Week 13 | Introduction to Machine Learning | Optional work: Machine Learning Tools  Apache Mahout, Spark MLlib, TensorFlow, etc.  \*Students choose their preferred machine learning library or tool for their term project |
| Week 14 | Machine Learning Algorithms | Supplemental Readings:  Machine Learning Algorithms |
| Week 15 | Project Presentation/Final exam | Project Presentations |