## **Big Data Technology “IS 6733”**

### Tools

The following cloud-based environment will be provided to each student: Jupyter, Spark, Python, Python libraries for big data analytics. Github will be used for course content, projects, and assignments.

### Recommended Background

This course is for those new to data science. No prior programming experience is needed, although the ability to install applications and utilize a virtual machine is necessary to complete the hands-on assignments.

### About the Course

According to McKinsey’s report, *The Age of Analytics: Competing in a Data Driven World*, December 2016, “Data and analytics capabilities have made a leap forward in recent years. The volume of available data has grown exponentially, more sophisticated algorithms have been developed, and computational power and storage have steadily improved. The convergence of these trends is fueling rapid technology advances and business disruptions”. This course is designed for students new to data science and interested in understanding why the Big Data Era has come to be. In this course, you will gain an understanding of what insights big data can provide through hands-on experience with the tools and frameworks used by big data scientists and engineers. This course provides an introduction to one of the most common frameworks, Spark, that has made big data analysis easier and more accessible. By following along with provided lessons, you will experience with data processing and leverage graph analytics to model problems. In the final course project, you’ll apply the skills you learned to do basic analyses of big data, to develop and implement a big data application.

### Syllabus

* Introduction to Big Data
* Big Data Characteristics
* Hadoop Ecosystems
* Big Data Integration
* Big Data Processing
* Graph Analytics for Big Data
* Machine Learning for Big Data
* Course Big Data Project
* Big Data Visualization (If time permits)

### Grading Criteria

* Quizzes and Exams (30%)
* Assignments (30%)
* Course project (40%)
* Extra grade (5%)

### What is the course project?

The course project is designed to help you practice, apply, and showcase the skills you’ve learned. In some Specializations, the Capstone Project is spread out across the courses. In others, it is the last course in the Specialization.

### Books

No required text book. Slides will be uploaded in GitHub.

### Recommended Books:

[1] Martella, Claudio, et al. Practical graph analytics with apache giraph. Apress, 2015.

[2] Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron

[3] Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale by Tom White

### Prerequisite: None

This course is for those new to data science. No prior programming experience is needed, although the ability to install applications and utilize a virtual machine is necessary to complete the hands-on assignments.

### Class Schedule

### Class Time: Monday 6:00 pm – 8:45 pm

### Office Hours:

* Monday 4:00 pm – 5:45 pm
* Wednesday 3:00 pm – 5:00 pm

### Office Location:

* NPB 3.138E08 @ Open Cloud Institute (OCI)

|  |  |  |
| --- | --- | --- |
| **No.** | **Date** | **Topic** |
| 1 | August 28th | **Introduction to Big Data** At the end of this module you will be able to:   * Describe the Big Data landscape * The three key sources of Big Data * Characteristics of Big Data * Data at Scale -- Working with Big Data   **Wednesday, August 30 - Deadline to add a class or late register** |
| 2 | September 4th  **Labor Day holiday** | Thursday, September 7: Census Date, 5 p.m. |
| 3 | September 11th | **Big Data Characteristics** At the end of these modules you will be able to:   * Recognize different data types * Data models (vector Space vs. Graph Data Model) * Streaming data vs. static data   September 8 – October 24 Drop Time Frame (Automatic “W”) |
| 4 | September 18th | **Hadoop Ecosystem (Part I)**  * HDFS, YARN and MapReduce |
| 5 | September 25th | **Hadoop Ecosystem (Part II)** At the end of these modules you will be able to:   * MapReduce Programming |
| 6 | October 2nd  **Submit Project Definition**  ***Quiz #1*** |  |
| 7 | October 9th | **Big Data Integration**  At the end of these modules you will be able to:   * SQL vs. NoSQL * Retrieve data from databases and big data management systems * Execute simple big data integration and processing on Spark platforms   Tuesday, October 10 Midterm Grades Due; 2 p.m. Deadline for faculty to enter midterm grades. |
| 8 | October 16th | **Big Data Processing (Part I)** At the end of these modules you will be able to:   * Introduction to Apache Spark * SparkSQL and Spark DataFrames |
| 9 | October 23rd | **Big Data Processing (Part II)**  * Spark Streaming * Sensor Data and Spark Streaming * Deep Learning |
| 10 | October 30th | **Graph Analytics for Big Data (part I)** At the end of these modules you will be able to:   * Introduction to Graphs * Examples (Smart Cities, Social Network, Etc.) * Path Analytics |
| 11 | November 6th  ***Quiz #2*** | **Graph Analytics for Big Data (Part II)**  * Connectivity Analytics * Connectedness: Indegree and Outdegree * Global Property: Modularity * Key Player, Eigenvector Centrality * Large Scale Graph Processing Giraph and GraphX |
| 12 | November 13th | **Machine Learning with Big Data (Part I)**  * Graph Computaiton and Tensorflow * Deep Learning and Convolution Neural Network |
| 13 | November 20th | November 23-24 Thanksgiving holiday. University closed. |
| 14 | November 27th | **Machine Learning with Big Data (Part II)** At the end of these modules you will be able to:   * Deep Learning and Convolution Neural Network |
| 15 | December 4th | **Big Data Project Review** Wednesday, December 6 Last Day of Classes |
| 16 | December 11th | **Final** December 9-15 Final Exams. No Final Exams on Sunday. |
|  |  | 1. December 19th: Final Grades Due; 2 p.m., Deadline for faculty to enter final grades. 2. December 20th: Students may view final grades on ASAP |