Cleveland State University  
Department of Electrical Engineering and Computer Science  

CIS 490/590 Big Data (3-0-3)  

**Prerequisites:** CIS430/530 and CIS408 or equivalents are required.  
**Instructor:** Dr. Sunnie (Sun) Chung  
**Office Location:** FH222  
**Phone:** 216 687 4661  
**Email:** sschung.cis@gmail.com  
**Webpage:** [http://eecs.csuohio.edu/~sschung](http://eecs.csuohio.edu/~sschung)  

**Catalog Description:** The course introduces fundamental concepts to design and implement an intelligent system (AI) as big data processing systems and/or big data analytic applications. It explores a number of related subjects such as common characteristics of machine generated big data format from social network sites or any web servers, semi structured and unstructured data models and processing techniques, major big data processing systems with semi structured database systems, modeling and design of semi-structured databases, and manipulation and retrieval methods as backend systems of an intelligent system as a big data application. The course also introduces topics related to optimization strategies for the performance of big data processing with parallel data processing systems and parallel programming paradigms Map Reduce in Hadoop Distributed File System (HDFS) that is used as physical data storage and index structures for big data. The course also introduces topics on basic methods and algorithms for big data analytics from Machine Learning, Information Retrieval, Natural Language Processing (NLP) disciplines for unstructured text analysis that are commonly employed in an industry level of big data analytic applications. Furthermore, the course introduces new platforms and architectures of big data processing systems of the existing Cloud and hand on experience on the Cloud based systems. Finally, this course provides design strategies and methodologies to build big data applications as intelligent web applications that process machine generated big data from social network sites or web servers/application servers for big data analytics with machine learning algorithms.

**Key Concepts:** Big data, Big data Processing, Intelligent System, Big Data Analytics, Machine Learning, Semi-structured and Unstructured Data Processing, XHTML, XML (eXtensible Markup Language), Document Object Model (DOM), JSON (Java Script Object Notation), Social Media Site Logging Structures, Semi-structured Database Systems, Mongo DB, Hadoop File Distributed System (HDFS), Map Reduce, Parallel Data Processing System, Information Retrieval, Parsing Techniques, Inverted Index, Term Frequency and Inverted Document Frequency (TF-IDF), Text Analysis, Natural Language Processing Techniques, POS (Part-Of-Speech) Tagging, Information Extraction (IE), Intelligent Web Application, Cloud Systems

**List of Required Systems:**  
(Instructions for each system will be given in class)

1. Any RDBMS: One of the following SQL Servers  
   SQL Server 2014 or higher with Visual Studio 2014 or higher available at  
   [https://portal.azure.com/#home](https://portal.azure.com/#home)  
   PostgreSQL available at [https://www.postgresql.org/](https://www.postgresql.org/)  
   Oracle Database 11g or higher with Eclipse IDE  
3. Hadoop File Distributed System (HDFS) with Map Reduce  

One of the following Data Analytic Tools:
5. R Programming

MS Software are also available at the Microsoft Academic Alliance programs: [https://azureforeducation.microsoft.com/devtools](https://azureforeducation.microsoft.com/devtools)

**Text:**
1. Selective Lecture Notes from Related Disciplines from Advanced Database, Information Retrieval, Machine Learning, and Artificial Intelligence
2. Selective Lecture Notes from Related Leading Industry Online Manuals

Class Web page: [http://cis.csuohio.edu/~sschung/CIS593/CIS593.html](http://cis.csuohio.edu/~sschung/CIS593/CIS593.html)
Go to my web page at [http://eecs.csuohio.edu/~sschung](http://eecs.csuohio.edu/~sschung) then choose Teaching->CIS 493/593 Link

**Note:** The Class Webpage Link might be changed per each class. The changed link will be announced on the blackboard of each class.

**Official Academic Calendar**

Please consult the university academic calendar for the semester and the final exam schedules at: [https://www.csuohio.edu/registrar/academic-calendar](https://www.csuohio.edu/registrar/academic-calendar)

**Grading:** The course grade is based on a student's overall performance through the entire Semester. The final grade is distributed among the following components:

Required:
1. Exams 40% (20% Midterm, 20% Final)
2. Computer Labs 35% (about 4 Lab Assignments)
3. Project and Presentation on Big Data Processing and Applications: 2 person group project (20%)
4. Class Attendance: 5%

Optional: Extra Project and/or Extra Assignments may be given for the students who want to outperform, or honor students in a contract course or students in CIS 593.

**NOTE:** The instructor preserves the right to change grading scheme depending on the performance of each class.

A 94% + A: Outstanding (student's performance is genuinely excellent)
A- 90% - 93.9% 
B+ 87% - 89.9% 
B 80% - 86.9% B: Student's performance is satisfied for every course requirement and acceptable but not necessarily distinguishable
B- 77% - 79.9% 
C 70% - 76.9% C: For CIS493, student's performance is satisfied for every course requirement and acceptable but not necessarily good; For CIS593, student's performance is not satisfied for every course requirement and is not acceptable to pass
D 65%-69.9% D: Student's performance fails to meet course objectives and standards
F <65% F: Failure (student's performance is unacceptable)
**Examination Policy:** Students are allowed to bring to the final exam a summary page (standard letter size) with their own notes. The note page is not allowed to midterm. During the exams: (1) the use of books, cell phones, calculators, or any electronic devices is prohibited, and (2) students must not share any materials.

**Make-Up Exam Policy:** No makeup exams will be given unless notified and agreed to in advance. Requests will be considered only in case of exceptional demonstrated need.

**Homework Policy:** The students are expected to attend all classes. The students are responsible for collecting the notes, handouts and any other course material distributed during the class period. All assignments must be individually and independently completed and must represent the effort of the student turning in the assignment. Should two or more students turn in *substantially the same solution* or output, in the judgment of the instructor, the solution will be considered group effort. All involved in group effort homework will receive a zero grade for that assignment. A student turning in a group effort assignment more than once will automatically receive an “F” grade for the course.

**Late Assignment:** All lab assignments are due at the beginning of class on the date specified. Laboratory Assignments handed in after the class has begun will be accepted with a 25% grade penalty for up to a week and then not accepted at all. All laboratory assignments must be completed. *Failure to do so will lower your course grade one additional letter grade.*

**Student Conduct:** Students are expected to do their own work. Academic misconduct, student misconduct, cheating and plagiarism will not be tolerated. Violations will be subject to disciplinary action as specified in the CSU Student Conduct Code. A copy can be obtained on the web page at: [http://www.csuohio.edu/studentlife/StudentCodeOfConduct.pdf](http://www.csuohio.edu/studentlife/StudentCodeOfConduct.pdf) or by contacting Valerie Hinton Hannah, Judicial Affairs Officer in the Department of Student Life (MC 106 email v.hintonhannah@csuohio.edu). For more information consult the following web page *CSU Judicial Affairs* available at [http://www.csuohio.edu/studentlife/jaffairs/faq.html](http://www.csuohio.edu/studentlife/jaffairs/faq.html)

**Contract Course Requirement for Honor Students and CIS593 Students:** An Honor student in a Contract Course or a CIS593 student should complete a group project that is approved as a substantial work.

**Course Schedule:** The schedule of topics to be covered is given below. The schedule and topics covered may vary depending upon the progress of each class made. Please see the class website for the detailed schedule of the subjects to be covered.

**Tentative Course Schedule is as follow.**
**Please See the Class Webpage for More Updated Contents and Recent Schedules**

<table>
<thead>
<tr>
<th>Week of</th>
<th>Topic</th>
</tr>
</thead>
</table>
| 1-2     | Introduction to Big Data, Big Data Processing, and Big Data Analytics **Review:** Web Application Basics: XHTML, AJAX, AngularJS, Node JS, REST API **Object Relational Mapping (ORM) for Modern Database Programming with Web Applications:** JDBC/ODBC User Defined Function (UDF), User Defined Type (UDT), User Defined Aggregate (UDA), Table Function, CLR, LINQ: .NET Database Programmability and Extensibility in Microsoft SQL Server by José A. Blakeley, *et al* (*Microsoft Corporation*) *in the proceedings of SIGMOD 2008*
| 3-4     | **Big Data:** Semi-structured and Unstructured data Machine Generated Big Data Formats in Web Applications |
Data Exchange Formats:
HTML/XHTML (Hyper Text Markup Language)
XML (eXtensible Markup Language)
Key Value Stores: JSON (Java Script Object Notation)

Semi Structured Data Processing Methods
Web Data Processing
XHTML, XML with DOM
Enhanced Data Models for Advanced Applications
XML Data Model and Processing:
- XML Schema, Syntax/Semantics, Protocol
- XPath, XQuery
Key Value Stores: JSON Data Model and Processing
Data Transformation techniques between Semi Structured/Unstructured and Relational Models
Social Media Data Format: Twitter, Facebook Logging Structure

Semi-Structured NoSQL Database System: Mongo DB
Mongo DB Scheme Design and Creation
Mongo DB Retrieval Methods
Mongo DB Data Pipelining

Unstructured Data Processing
Introduction to Information Retrieval and Web Data Processing
Parsing Methods for Unstructured Data Processing
Inverted Index Construction

Text Analysis Basics for Intelligent Web Application System
Document Search Methods
- Google Search Engine
Term Frequency (TF) and Inverted Document Frequency (IDF)
Document Similarity Measures

Big Data Analytic Basics:
Machine Learning,
Data Mining with Classification and Predictive Models:
Decision Tree
Neural Network
Support Vector Machine

Big Data Processing and Parallel Distributed Computing:
Introduction of Parallel Distributed Data Processing System for Big Data
Google’s Map Reduce Parallel Programming Paradigm
- MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean (Google) and Sanjay Ghemawat (Google) in the proceedings of OSDI 2004
Apache Hadoop File System (HDFS) for Parallel Processing
- Apathy Hadoop in White Papers by Apache, Yahoo

Cloud Computing for Big Data Processing and Data Analytics
Amazon Elastic Compute Could (EC2)
Cloud Computing
- Microsoft Azure
- Amazon Elastic Cloud (EC2)

Natural Language Processing Techniques for Text Analysis
Text Preprocessing Methods for Text Analysis
Presentation on Group Project in Big Data Processing, Big Data Analytics, and Cloud Computing
List of Suggested Projects will be given in class.

Labs/Project: Big Data Processing and Analytics with Real Life Big Data
- Social Network Data Collection Methods -- How to obtain and process real-time stream data from Twitter
- How to Obtain and Process Data from Websites
- How to Process Big Data in Various Formats
- How to Analyze Big Data

**NOTE:** The instructor reserves the right of the intellectual property to retain, for pedagogical reasons or any, either the original or a copy of your work submitted either individually or as a group project for this class. It is illegal (a violation of an intellectual property right) to submit the labs or group projects of this course to other courses or use a copy of the contents for any other purposes without a permission of the instructor ahead.

**ADA Adherence:**

If you need course adaptations or accommodations because of a disability, if you have emergency medical information to share with me, or if you need special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible. My office location and hours are listed on top of this syllabus. If you need further information, please contact the Office of Disability Services (Main Classroom 147), phone number 216.687.2015, on the web at [http://www.csuohio.edu/offices/disability/](http://www.csuohio.edu/offices/disability/).