Revised: October 11th, 2018

**Course Title**: **Big Data Technology**

**Program**: **Business Intelligence and Analytics (BI&A)**

**Course #**: **BIA 678**

**Instructor: Dr. David Belanger**

**Catalog Description:**

This course covers the several areas necessary to function in the world of Big Data. These include: why (what applications are possible with Big Data that were difficult or impossible with classical techniques), what (the various technologies that make Big Data possible), and how (use of technologies, governance, and management of volume, velocity, and variety). It is intended for students who will eventually execute and/or manage Big Data Projects. As such, it includes both hands on use of technology (e.g. Hadoop, Spark), and coverage of management techniques such as privacy, security, and metadata management.

This course is required for students in Business Intelligence and Analytics, and open to all other graduate students including Information Systems, Computer Science, Financial Engineering, and Masters of Business Administration, or with the permission of the instructor.

**Course Objectives:**

This course purpose is to prepare students to function in a world of big data and associated technologies such as machine learning and data management.

It presents the underlying technologies such as; distributed (and cloud) computing, data base/stream management at scale (e.g. NoSQL), analytic techniques at scale including recommender systems, ensembles, data stream analysis, visualization at scale, and governance.

The course assumes basic knowledge of multivariate statistics and data management, and is oriented to the skills needed to operate at large volume and velocity, and both structured and non-structured data. It includes coverage of some of the basic tools in the area. Active use of both Hadoop and Spark, and coverage of tools like MongoDB, Cassandra, and Storm. It is also oriented to give students a framework for the inevitable changes in available tools and techniques in the area.

**List of Course Outcomes:**

|  |
| --- |
| After taking this course, students will be able to:   1. Understand and discuss what big data is, and how it differs from traditional approaches and when and why one would apply it. 2. Plan and use the primary tools associated with big data. 3. Plan and execute a team project that both uses big data technology, and analyzes the impact of scale on those technologies. 4. Conduct research and write a term paper on some aspect of big data, in preparation for understanding where it is going.. |

**Grading Percentages**: Homework/Class work **25%** Final Project **40% Term Paper 35%**

|  |
| --- |
| **Final Project:**  Student teams have an opportunity to build a big data application as their final project. In addition to building the application, students must do an oral presentation and a written report on the project. The reports can be on a variety of data analytic and/or data management problems, but must include an analysis of the impact of scale on the various techniques used, both in terms of performance and of accuracy.  **Assignments:**  Each week there are reading assignments in which the students must read one or two papers, write a short (1/2 page) overview of what they learned, and be prepared (by random selection) to orally report on the paper in class. In addition, there are several, 4 – 5 programming assignments to be executed in Hadoop and Spark. In addition, each student is required to write a 5 – 8 page Term Paper on a topic of their choice related to Big Data, and to be a member of a 1 – 5 person team which executes a semester long big data project. |

**Textbook(s) or References**

There are no texts used, but there are on the order of 20 – 25 papers, ranging from research to technology documentation to governance (e.g. privacy, security). These change from semester to semester as new papers become available.

# **Ethical Conduct**

|  |
| --- |
| The following statement is printed in the Stevens Graduate Catalog and applies to all students taking Stevens courses, on and off campus.  “Cheating during in-class tests or take-home examinations or homework is, of course, illegal and immoral. A Graduate Academic Evaluation Board exists to investigate academic improprieties, conduct hearings, and determine any necessary actions. The term ‘academic impropriety’ is meant to include, but is not limited to, cheating on homework, during in-class or take home examinations and plagiarism.“  Consequences of academic impropriety are severe, ranging from receiving an “F” in a course, to a warning from the Dean of the Graduate School, which becomes a part of the permanent student record, to expulsion.  *Reference: The Graduate Student Handbook, Academic Year 2003-2004 Stevens*  *Institute of Technology, page 10.*  Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.  Signature \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_  Please note that assignments in this class may be submitted to [www.turnitin.com](http://www.turnitin.com/), a web-based anti-plagiarism system, for an evaluation of their originality. |

**Course/Teacher Evaluation**

Continuous improvement can only occur with feedback based on comprehensive and appropriate surveys. Your feedback is an important contributor to decisions to modify course content/pedagogy which is why we strive for 100% class participation in the survey.

All course teacher evaluations are conducted on-line.  You will receive an e-mail one week prior to the end of the course informing you that the survey site (<https://www.stevens.edu/assess>) is open along with instructions for accessing the site.  Login using your Campus (email) username and password. This is the same username and password you use for access to Moodle. Simply click on the course that you wish to evaluate and enter the information. All responses are strictly anonymous.  We especially encourage you to clarify your position on any of the questions and give explicit feedbacks on your overall evaluations in the section at the end of the formal survey which allows for written comments.  We ask that you submit your survey prior to end of the examination period.

**Syllabus**:

The course is divided into modules, some of which will extend across more than a single class meeting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Topic(s)** | **Reading(s)** | **Class Exercises** | **Assignments** |
| Week 1 | Introduction to Big Data |  |  |  |
| Week 2 | Introduction to Big Data 2 - Applications | Usually papers from Twitter and McKinsey on Big Data | Report on papers assigned in previous week. | Each week has reading assignments. |
| Week 3 | Distributed Computing 1 – HDFS, Hadoop | Usually seminal papers on MapReduce and GFS |  |  |
| Week 4 | Distributed Computing 2 - Spark | Usually Seminal paper on Spark from UCB |  |  |
| Week 5 | Introduction to scale in data base management |  |  | Abstract of proposed term paper due. |
| Week 6 | Data base management at scale – Column Store, NoSQL, Hbase, Cassandra, MongoDB | Usually paper on Column vs Row Store, and documentation of Cassandra and MongoDB |  | First programming assignment due. |
| Week 7 | Data Stream Management | Usually a paper on survey of data stream management |  |  |
| Week 8 | Data Stream Analytics |  |  | Hadoop Prog. Assign due. |
| Week 9 | Analytics at scale – Recommender systems | Usually papers such as Koran’s on Matrix Factorization, and a case study on recommender systems |  | Term paper due. Abstract of proposed team project due. |
| Week 10 | Recommender systems continued (collaborative filetering) and start of Ensembles (bagging, boosting) |  |  | Spark prog. Proj. due. |
| Week 11 | Ensemble (boosting, adaboost, xgboost), and other selected analytic techniques. | Usually includes paper by Freund on Adaboost, with one other paper. |  |  |
| Week 12 | Big Data Governance - Overview |  |  | Spark prog. Proj. 2 due. |
| Week 13 | Big Data Governance Technology – e.g. privacy, security, metadata management. |  |  |  |
| Week 14 | Team Presentations |  |  | Team Presentation. Written report due at EOW. |