New York University Tandon School of Engineering
Computer Science and Engineering
Course Outline CS9223 Big Data Analytics

Professors: Raman Kannan rk1750@nyu.edu
Office Hours will be posted in the course website

**Statement of Academic Integrity**

Students are expected to follow standards of excellence set forth by New York University. Such standards include respect, honesty, and responsibility. This class does not tolerate violations to academic integrity including:

- Plagiarism
- Cheating on an exam
- Submitting your own work toward requirements in more than one course without prior approval from the instructor
- Collaborating with other students for work expected to be completed individually
- Giving your work to another student to submit as his/her own
- Purchasing or using papers or work online or from a commercial firm and presenting it as your own work

Please refer students to the Tandon code-of-conduct for addition information at: [http://engineering.nyu.edu/life/student-affairs/code-of-conduct](http://engineering.nyu.edu/life/student-affairs/code-of-conduct)

**Instructor allows students to source knowledge from any source including friends, colleagues, internet, library, papers and books.**

**All evaluations are open book and open notes and your problem solving abilities and your ability to work with other students are assessed.**

**However, some students in the past, have either failed to do any work or even violate the above policy and norms of academic integrity.**

**Instructor does not have time for handling such incidents. Students will be turned over to Deans Office and in the past Instructor has given F.**

What follows is an elaborate description and consistent with the code-of-conduct.
Course Pre-requisites

This offering of the course is for students admitted to a special program. As prerequisite, students must have been admitted to the program, and to be admitted to the program, students must have significant experience in programming, mathematical background, and some knowledge of algorithms. Of benefit for this course, but not required, is some basic knowledge in databases.

Course Description

Big Data requires the storage, organization, and processing of data at a scale and efficiency that go well beyond the capabilities of conventional information technologies. The course reviews the state of the art in Big Data analytics and in addition to covering the specifics of different platforms, models, and languages, students will look at real applications that perform massive data analysis and how they can be implemented on Big Data platforms.

Topics discussed include:

1. DataStores: SQL and NoSQL stores,
2. Map reduce over Mongo,
3. Apache Spark,
4. large-scale data mining using R and visualization.

The curriculum will primarily consist of technical readings and discussions and will also include programming projects where participants will prototype data-intensive applications using existing Big Data tools and platforms, namely R, Relational, non-Relational, and Spark. Students may choose to use python and/or Scala.

Course Objectives

1. To learn about basic concepts, technical challenges, and opportunities in big data management and big data analysis technologies.
2. To learn about common algorithmic and statistical techniques used to perform big data analysis.
3. To learn and get hands-on experience analyzing large data sets using Apache Spark.
4. To learn about different types of scenarios and applications in big data analysis, including for structured, semi structured, and unstructured data.

Course Structure

Lectures (recorded), plus intensive interaction via the e-learning platform. There will also be a reading list of research papers, and students are expected to perform hands-on homeworks and a possibly a project.

Readings


Lecture notes are prepared and provided by the faculty, and a list of journal and conference papers, available on the internet or via the Dibner electronic library, will be provided.

Software Requirements

The course requires the following software packages, all freely available:

Other recommended software for this course:
Eclipse IDE, https://eclipse.org/ (for Java)

Tutorials will be provided in video, slides and other suitable form on how to install/use these packages.

**Other Technical Requirements**

We will be performing all our work on IBM Cloud. Students may choose to use other cloud environments including Google AppEngine clouds. Access to IBM Cloud will be provided by the instructor free of cost and is not responsible for other cloud environment.

**Course requirements**

Students are expected to do, and will be graded on: (a) 3 significant homework projects giving them hands-on experience in data analysis tools and applications (60%), (b) a series of small tests (20%), (c) class discussions (20%).

**Course Topics by Week: Subject to adjustment/revision**

**Week 1: Course Overview. This is course is project driven**

1. Comparing Handwritten Signature -- Introduction, Issues, Techniques;  
2. Matching Resumes with Project Specifications;  
3. Applied Image Analysis  
4. Month-End Problem :  
   1. Portfolio Valuation for millions of customers every month (think of Fidelity, Visa, MC, TD),  
   2. Billing Period Invoice Generation (think of ATT or VZ),  
   3. Option Tree Premium Update (examine Option Tree for ticker symbol IBM and the number of updates to Option premium)  
   4. Hedging portfolios with Options mitigating market meltdown and bubbles  
   5. Predicting and forecasting 25% market corrections (systemic crashes)  
   6. RDD -- is not the Spark Resilient Distributed Data that comes with Spark. Building and enhancing R with a framework for very large scale distributed computing in R by integrating ISIS (Cornell Ken Birman et. Al.), Linda (Yale Gelertner et. Al) using R Serv and other distributed computing primitives available in R.

**Week 2: Databases and Big Data: Persistence, Transactions, Querying, Indexing and SQL**

**Week 3: Introduction to R Programming Language from Data Analytics Perspective I**

**Week 4: Introduction to R Programming Language from Data Analytics Perspective II**

**Week 5: Basic Data Mining and Statistics in SQL and R**

**Week 6: Distributed Problem Solving in R:Shiny, R Serv, etc**

**Week 7: Text Processing in R and Spark**
Week 8: Learning for Scalable Text Analysis
Week 9: Algorithms for Big Data: Finding Similar Items
Week 10: Comparing Handwritten Signature -- Introduction, Issues, Techniques
Week 11: Matching Resumes with Project Specifications
Week 12: SparkML – certain fundamental algorithms used in Machine Learning
Week 13: Clustering Algorithms -- EM
Week 14: Class Project Presentations

Grade distribution will be as follows:
- Top 20% of students will get A,
- Next 10% A-,
- Next 25% B+,
- Next 25% B,
- Next 20% other grades

I am leaving material relating to Hadoop etc from earlier versions of this course for the benefit of students, and I will introduce material consistent with revised curriculum throughout the semester.

Schedule:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>09/05/17 Introduction sample projects, student projects</td>
<td>Grading, Teams, Project, Systems, Lecture</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>09/13/17 Databases, Big Data, My SQL Introduction – Introduction to SQL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>09/20/17 Introduction to R Part I</td>
<td>SCMP with W from 09/19/2017</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>09/27/17 Introduction to R Part II</td>
<td>Students must have IBM Cloud access or their own cloud</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10/04/17 Basic Data Mining, Statistics, SQL and R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>10/11/17 Distributed Problem Solving in R, Shiny, Rserv etc</td>
<td>Lab 60 minutes</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10/18/17 Text Processing in R and Spark</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10/25/17 Scalable Text Analysis</td>
<td>Lab 60 minutes</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>11/01/17 MIDTERM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11/08/17 Algorithms for Big Data: Finding Similar Items</td>
<td>Lab 60 minutes</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11/15/17 Comparing Handwritten Signature Introduction, Issue, Techniques</td>
<td>Lab 60 minutes</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11/22/17 SparkML – NL using Spark, certain basic algorithms in ML</td>
<td>Lab 60 minutes</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>11/29/17 Clustering Algorithms – EM, Local/Global Maxima</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>12/06/17 Project Presentations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>12/13/17 Review of Course</td>
<td>J&amp;A</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>12/20/17 Class Finals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www.nyu.edu/registrar/calendars/university-academic-calendar.html