New York University Tandon School of Engineering
Computer Science and Engineering
Course Outline CS6513 Big Data

Professors: Raman Kannan rk1750@nyu.edu

Office Hours: email and weekly virtual meetings

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 in 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

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.

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

Course Pre-requisites

This offering of the course is for students who wish to prepare for a career in processing very large amounts of data. As prerequisite, 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
5. 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 R and/or Spark over java 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 and get hands-on experience analyzing large data sets using a combination of R,MySQL and mongo or any other non-relational database.
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

Materials posted on classes 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 two projects.

Readings

The required text for the course is: **Mining of Massive Datasets**. Rajaraman and Ullman, Cambridge University Press, 2011. Available online at http://infolab.stanford.edu/~ullman/mmds/book.pdf

Additional reading: **Data-Intensive Text Processing with MapReduce**. J. Lin and Chris Dyer, Morgan Claypool , 2010. Available online at http://lintool.github.io/MapReduceAlgorithms/

A list of journal and conference papers, available on the internet or via the Dibner electronic library, challenges from real-world, additional notes and presentations will be provided.

Software Requirements

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

1. The R Project for Statistical Computing, http://www.r-project.org/
   Optional R Studio, http://www.rstudio.com/
2. MySQL for relational, mongo for document oriented data will be provided.
3. Spark over java or scala will also be provided.

All class related work must be done IBM cloud so the work can be centrally evaluated. A login will be provided free of charge. There is no installation required. Students are encouraged to use Xming (on

Windows) and Quartz (on Mac) if there is aversion to command line interactivity.

.

Other Technical Requirements

We will be performing all our work on IBM Cloud. All the (functional) projects and tests required for this course have to be delivered on the IBM Cloud.

Access to IBM Cloud will be provided by the instructor free of cost.

Course requirements

Students are expected to do, and will be graded on: (a) 2 significant homework projects giving them hands-on experience in high volume data processing and applications (60%), (b) two tests (20%), (c) class|discussions (20%).

Course Topics by Week: Subject to adjustment/revision

Week 1: Course Overview. This is course is project driven and processing large number of files or

Records held in a persistent store is of particular interest. Some of the practical applications I am interested in are:

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. Building and enhancing R with a framework for very large scale distributed computing in R using RServ and other distributed computing primitives available in R. Both bigmemory, Rdsm are installed and available on IBM Cloud.
   7. Students are expected to learn and engineer a non-trivial application.

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, RServ, etc

Week 7: Text Processing in R and Spark – basics TF/IDF, Word2Vec, LDA, Entity Extraction.

Week 8: Learning for Scalable Text Analysis

Week 9: Algorithms for Big Data: Finding Similar Items

Week 10: parallelizing Cross Validation, LOOCV

Week 11: parallelizing Stochastic Gradient Descent, Boosting, Locally Sensitive Hashing

Week 12: SparkML – certain fundamental algorithms used in Machine Learning

Week 13: SparkKnife, Occams Razor, Classifiers as instructions and MISD

Week 14: Matching Resumes with Project Specifications, Comparing Handwritten Signature -- Introduction, Issues, Techniques

Grade distribution will be as follows:

Top 30% of students will get A and A-,

Next 25% B+,

Next 25% B,

Next 20% other grades

Hadoop is not included in the topics and nor will any work done in python is acceptable for this course.

This is not a course on analytics but many programming challenges are drawn from many disciplines including machine learning, and data mining.

Schedule:

For calendar visit here https://www.nyu.edu/registrar/calendars/university-academic-calendar.html#1186, first week starts May 21,2018 (Mon) and the last week ends on Aug 10th (FRI).

First project proposal due June 8th Wednesday 5PM.
First project completed and submitted on July 6th 5PM. (25%)
Second project proposal due July 13th 5PM
Second project completed and submitted on August 8th 5PM (35%)

First test to be turned in by June 29th, 5 PM (10%)
Second test to turned in by August 10th, 5PM (10%)

Our virtual class will be held over the zoom on classes from 8 to 9 PM on TUESDAY, each week. Attendance and active participation is required and submitting a weekly progress report is also required to receive class participation grade of 1% point toward your grade except for the first and the last week (12%). First week students are asked to submit an expectation report (4%) and a reflection report on the final week (4%). Submitting a weekly progress report is also required to receive class participation grade. Based on student feedback, the deadlines are rigid and I will not grant any extension.

This course requires a lot of work, allowing students to define their own projects.

Central learning objective are

1. distributed problem solving leveraging large volume and variety of data. This course does not address velocity, the third V of Big Data.

2. Thinking and devising distributed problem solving architecture is an essential learning objective.
3. Implementing/engineering solutions using R,java,scala and Spark is the third learning objective.

To achieve these objectives we will use data from financial services, text analytics. Our focus will not be analytics or statistics or the mathematics. But given some analytical function how to compute solve problems using that function in a distributed environment. Given some statistics or mathematics how to setup parallel solution so that problems that cannot be solved in a single computer is solved using a cluster of computers.

Students are encouraged to study Ken Birman (Cornell, ISIS), Yale (Linda Gelertner) and Condor (Processor hunter) and PVM …from the 90s. Think about strategies to incorporate missing features into R and Spark.