Text

Description automatically generated with medium confidence

ISGB | BYGB 7990 - Big Data Analytics

Course Syllabus, Spring 2022

# Contact Information

Instructor: Dawit [David] Demissie, Ph.D.

Office Hours: Friday, 2-5 pm & by appointment

Office Location: LC, 45 Columbus Circle, Room 617 | RH, Faber Hall 446.

E-mail: ddemissie@fordham.edu

Voice **|** Text: 518-421-5506 (*In most cases, I will return your emails, texts & calls within 8 hours)*

Graduate Assistants:

Tanya Agarwal. Email: tagarwal6@fordham.edu

Manas Pachauri,Email: [msureshkumarpachauri@fordham.edu](mailto:msureshkumarpachauri@fordham.edu)

# Course Information

Course **|** Section: ISGB | BYGB 7990 - Big Data Analytics | Section 001 & 002.

Course Delivery Format: Face-to-Face (In-Person) Teaching

Date **|** Time **|** Place: Tuesday, 3:30 – 5:30pm | 5:45 – 7:45pm, 140W, 334

# Course Description

Many of the world's biggest discoveries and decisions in science, technology, business, medicine, politics, and society as a whole, are now being made on the basis of analyzing massive data sets. Those typical datasets include millions of online customer reviews, social comments from Facebook, Twitter and other popular social platforms, shopping transaction records, mobile messages, financial news, climate data, IoT data and others. Today’s organizations face the key challenge of analyzing massive data to gain key insight to make informed decisions.

This course provides a broad and practical introduction to Big Data from three perspectives: (1) introduction to the Big Data problem, current challenges, trends, and applications; (2) technologies for Big Data management, Big Data technology and tools, special consideration made to the MapReduce paradigm and the Hadoop ecosystem; (3) algorithms for Big Data analysis, mining and learning algorithms that have been developed specifically to deal with large datasets.

The technologies are still evolving very rapidly. Therefore, there is a level of experimentation with new material that will take place during the semester.

# Prerequisites

Data Mining is a required prerequisite course.

Basic Python programming skills are required.

# Course Objectives

Upon completing this course, the students are expected to achieve the following objectives:

* To understand the organizational, strategic, and managerial concepts surrounding big data and big data analytic.
* To have working knowledge and exposure to key elements of a big data technology framework and platform including Hadoop and MapReduce, Cloudera and AWS.
* To have hands-on experience with distributed data management and NoSQL techniques such as Hive, Pig, Sqoop, and Kafka.
* To gain understanding of popular models and algorithms of big data analytics, including recommender systems, clustering, classification, topic models, and network analysis and their implementation with Spark on Hadoop.

# Course Materials and Textbook

* Lecture slides and handouts distributed by the instructor.
* Reference books

**For practical Big Data Analytics: Mining of Massive Datasets**

by Jure Leskovec, Anand Rajaraman, and Jeff Ullman

ISBN-10: 1107015359 | ISBN-13: 978-1107015357

Copy can be downloaded from <http://infolab.stanford.edu/~ullman/mmds/book.pdf>hardcopy is available from Amazon.

**Learning Spark: Lightning-Fast Big Data Analysis**

by Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia

ISBN-13: 978-1449358624 | ISBN-10: 1449358624

**Programming Hive: Data Warehouse and Query Language for Hadoop**

by Edward Caprio, Dean Wampler, and Jason Rutherglen

ISBN-13: 978-1449319335 | ISBN-10: 1449319335

**For advanced mastering of Big Data Framework Hadoop: The Definitive Guide**

by Tom White

ISBN-10: 9352130672 | ISBN-13: 978-9352130672

**Data-Intensive Text Processing with MapReduce**

by Jimmy Lin and Chris Dyer

ISBN-10: 9781608453429 | ISBN-13: 978-1608453429

Copy can be downloaded from

<http://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf>

# Classroom Equipment

Students will be required to bring their laptops during some classes with wireless Internet access to complete the class demo and exercises.

# Office Hours

Instructor’s office hours: Friday 1-5pm or by appointment (*45 Columbus Circle, Room 617*). You may also send e-mails to ask questions or set up appointments outside of office hours. You may also send e-mails to ask questions or set up appointments outside of office hours. **Please type “BYGB 7990 Section 1” or “BYGB 7990 Section 2” in the subject line of every email that you send**. If no reply is received within 12 hours, your email may have been overlooked and please feel free to send another email.

# Course Homepage

Class-related materials will be posted on Blackboard. Please check the page frequently for updates. You will find the following materials: syllabus, lecture schedule, lecture notes, group project, frequently asked questions, and course resources. Please print out the necessary materials for yourself.

# Grading

Student grades will approximately consist of the following elements as related to learning objectives:

|  |  |
| --- | --- |
| **Grading Item** | **Weight** |
| Homework and Labs | 30% |
| Group Project | 25% |
| Midterm Exam | 20% |
| Final Exam | 15% |
| Class Participation | 10% |
| Programming Contest | 5% (extra-credit) |

Letter grades will be assigned approximately as follows. The grade distribution may be curved.

93-100% A

90-93% A-

87-90% B+

83-87% B

80-83% B-

77-80% C+

73-77% C

70-73% C-

67-70% D+

63-67% D

60% to 63%

D- Below 60% F

# Attendance

Class activities are intended to give an overview of the material, provide examples, and conduct discussions to help you think critically. You are encouraged to ask questions during the class. You are responsible for all lecture material regardless of whether you attend each class, and you must get your own notes from your classmates if you miss a class. Please note that office hours should not be used as a substitute for class attendance.

Although it should not need to be said, I expect you to maintain a reasonable level of decorum in class. This means that there is usually no eating or drinking in class. **Cell phones are suggested to be turned off**. You'd better not walk in late or walk in and out of the room during lecture.

# Assignments

Several assignments will be given throughout the semester. They are INDIVIDUAL assignments and must be completed INDEPENDENTLY to receive full credits. Discussions among students about the assignments are allowed. However, **copied work is absolutely NOT allowed and will be turned into the honor court**. If you have any doubt, ask the instructor. All assignments, unless otherwise stated, should be submitted electronically on Blackboard.

Assignment grade can only be appealed within one week from the day when that assignment is returned. **No appeal is accepted after that period**.

Late submission will have points deducted (10% for each day being late).

# Group Project

Students will have the opportunity to further sharpen their skills and acquire hands-on experience with practical business applications through a team project. Students will form groups consisting of **4 or 5** people (exceptions have to be approved by instructor).

For your project you can use any publicly availabledatasets or crawl your own datasets from online resources, like Facebook, Twitter, Yelp, Amazon.com, Yahoo financial news, etc. To manage your group effort, if possible, follow the following 6 steps:

1. Define the topic of interest.
2. Obtain the necessary data.
3. Explore the data.
4. Create analytical models.
5. Interpret data / gather insights.
6. Communicate the results.

The project deliverables are:

* 1. **Proposal presentation.** Maximum 6 slides of PowerPoint slides, what business problem are you solving, which dataset you are working on, how you plan to preprocess and analyze the dataset. what are the expected outcomes are?
  2. **Final Presentation.** There will be 1 presentation during the last class in the semester, major results are expected during the presentation.
  3. **Technical Report.** Maximum 15 pages report (double-space, 12-point font) highlighting consisting of introduction, method, design, models, results, and conclusion.
  4. **Confidential peer-evaluation form**. You will evaluate the contribution of each of your group members.

# Exams

You must take exams at the scheduled time slots. No early or make-up exams will be permitted except for extremely exceptional cases that are approved by the instructor 3 weeks BEFORE the exam date. If you miss an exam for any reason except serious and/or debilitating illness/injury, you will receive a zero grade for that exam.

A student with a conflicting exam schedule or with more than two officially scheduled examinations in twenty-four hours may reschedule an examination with permission sought by the student from the student's college dean no later than 3 weeks prior to the start of Final Examinations and by arrangement with the appropriate instructor.

# Programming Contest

Students will attend a programming contest. The contest consists of 3 problems and the students are only required to successfully accomplish one of them to receive 5 extra credit points.

# Integrity

We assume that you have complete integrity in all your class efforts. Violations of the University's Honor Code will be taken extremely seriously, and they will be addressed promptly according to the established procedures. Students are to adhere to the Code of Student Conduct, and other policies and regulations as adopted and promulgated by appropriate University authorities. Copies of these documents may be obtained from the Office of the Dean of Students or from the offices of the academic deans. No infractions will be tolerated. Students violating the Code of Student Conduct will be dismissed from class and will receive an “F” for the course.

# Disability Policy

Under the Americans with Disabilities Act and Section 504 of the Vocational Rehabilitation Act of 1973, all students, with or without disabilities, are entitled to equal access to the programs and activities of Fordham University. If you believe that you have a disability that may interfere with your ability to participate in the activities, coursework, or assessment of the object of this course, you may be entitled to accommodations. Please schedule a meeting to speak with someone at the Office of Disability Services (Rose Hill - O’Hare Hall, Lower Level, x0655 or at Lincoln Center – Room 207, x6282).

# Religious Holiday Policy

The university policy regarding religious holidays is that it is the student’s responsibility to notify faculty during the first week of the semester of their intention to be absent from class on their days(s) of religious observance. The faculty continues to extend to these students the courtesy of absence without penalty on such occasions, including permission to make up examinations.

# Class Schedule (Tentative)

*Here is a tentative schedule of lectures, readings, and labs for this course. We will try to keep approximately to this schedule.*

|  |  |  |
| --- | --- | --- |
| **Session** | **Reading Checklist** | **Assignments** |
| Session 1:  1/18 | **Lecture #1 – Intro to Big Data**    **Course Overview - Syllabus** | Form Project Groups |
|  | **Module I: Data Management and Platforms** |  |
| Session 2:  1/25 | **Lecture #2- Intro to Distributed System and Hadoop**  **Lecture #2.1 -Intro to Linux OS**    Lab #1: UNIX / Linux commands | Project Milestone 1 (Group Members) |
| Session 3:  2/1 | **Lecture #3 - Intro to MapReduce**  **Lecture #3.1 - Git-GitHub**    Lab #2: Git-GitHub | HW #1 - Posted |
| Session 4:   2/8 | **Lecture #4 - Google Cloud Platform (GCP)**  **Lecture #4.1 - Apache Hive**    Lab #3: Configuration & Installation of Hadoop  Lab #4: Hive and HQL |  |
| Session 5:  2/15 | **Lecture #5 - Apache Hbase**    Lab #5: Cloudera on Google Cloud | HW #1- Due  HW #2- Posted |
| Session 6:  2/22 | **Lecture #6 - Amazon AWS**    Lab #6: Jupyter Notebook & Google Cloud Storage |  |
| Session 7:  3/8 | **Lecture #7 - Intro to Spark**    Lab #7: Jupyter Notebook & S3 storage on AWS | HW #2 - Due  HW #3 - Posted |
| 3/15 | Spring Recess - No Classes |  |
| 3/22 | Midterm Exam (*Time – TBA)* |  |
|  | **Module II: Data Analytics** |  |
| Session 8:  3/29 | **Project Milestone 3 (Midterm Design Presentation)** |  |
| Session 9:  3/30 | **Lecture #8 - Association Rules & Recommendation System – Part 1**  Lab #8: Spark Configuration Guide & Rec Sys. | HW #3 - Due  HW #4 - Posted |
| Session 10:  4/5 | **Lecture #8 - Association Rules & Recommendation System – Part 2**  Lab #9.0: Apache Spark SQL on GCP using Dataproc  Lab #9.1: Using Databricks to Run a RS with MLLib & Spark |  |
| Session 11:  4/12 | **Lecture #9 - Google BigQuery**  Lab #10: Google BigQuery | HW #4 - Due |
| Session 12:  4/20 | **Lecture #10 – Dataiku DSS**  Lab #11: Data Exploration & Visualization Recipe with Dataiku  Lab #11.1: Prepare & Visual Recipes with Dataiku DSS |  |
| Session 13:  4/19 | **Lecture # 11 & 12 - ML, Networks & Facial Recognition** **(Dr. Yilu’s Research Project for References)**  Lab #12: |  |
| Session 14:  4/26 | Group Project Presentation I |  |
| Session 15:  5/3 | Group Project Presentation II |  |
| Final Exam 5/10 | * Final Exam: * Project Report Deadline |  |