Syllabus

IST 410 Data Analytics at Scale

(Programming Models for Big Data)

Tu Th 3:05-4:20 pm; Room: 210 IST Building

Instructor: John Yen (Office Hours: Tu Th 4:30-5:30 pm; Room: 313J)

TA: Yu Luo (Office Hours: 2-3 pm, Friday; Room: 325D IST)

Learning Objectives: The goal of this course is to introduce the programming models for processing massive datasets. Specifically, we will cover two programming models: (1) MapReduce, and (2) Spark. A dataset of scholarly publications gathered from PubMed will be made available for labs and projects. A cluster computing platform, using Hadoop Distributed File Systems, will be made available through VLab. A high-level MapReduce programming language (Pig), Spark, Scala (the language in which Spark is implemented), and a machine learning library on Spark (MLlib) are all available from the cluster. The learning objectives of the course are the following:

- Gain hands-on experience about programming models that are designed to be scalable for handling massive datasets.
- Be able to design and implement MapReduce applications using Pig.
- Be able to design and implement an iterative machine learning algorithm (i.e., clustering) using Spark in a cluster environment
- Be able to choose Spark configurations in a cluster to compare and tune the performance.

Textbook:

1. Learning Spark, by Holden Karua, Andy Konwinski, Patrick Wendell, and Martei Zaharia, 2015 (L)

References:

1. Programming in Scala, Martin Odersky, Lex Spoon, and Bill Venners, second edition, Artima Press, 2012. (P)
2. Hadoop: The Definitive Guide, Tom White, 2012 (H)
3. Big Data Analytics Beyond Hadoop, Vijay Agneeswaran, 2014 (B)

Topics and Schedule:

We will focus on one topic each week. Most of the Tuesday classes will also include a hands-on "lab" (often related to the Thursday lecture in the previous week). The list of topics and labs are listed below. Papers and reading assignment for each topic are made available through Canvas.

The hands-on labs will give you hands-on experience on programming MapReduce and Spark on a Hadoop Distributed File System.

The term project is an experiment regarding scalability of clustering a massive Pub med dataset using Spark. Many of the labs will help you to achieve your goals for the project. You also need to design what you want to configure for comparing performance. The project abstract is due on 1/26. I will provide feedback to assist you in refining the project ideas.

Date	Topic (Th)	Readings	Lab (Tu)	Project
1/10, 12	Overview of trends that introduce the opportunities for data analytics at scale	B: Chap 1; H: Chap 1	Challenges of these opportunities not met by previous approaches	
1/17, 19	The MapReduce Programming Model for Big Data	H: Chap 2,	Vlab and UNIX commands	
1/24, 1/26	Pig: A High-level Programming Language for MapReduce	H: Chapt 11	Map and Reduce in Pig	Project Abstract (1/26)
1/31, 2/2	HDFS: Hadoop Distributed File System	H: Chap 3	HDFS	
2/7, 2/9	Software Stacks for Data Analytics at Scale	B: Chap 2; L: Chap 1, 2	Spark Shell	
2/14, 16	Functional and scalable programming for Big Data (Scala)	P: Chap 2, 3	Hashtag count using Scala	
2/21, 2/23	Data flow supports for modern cluster computing environment (RDD)	L: Chap 3, 4	Hashtag counting using Spark	
2/28, 3/2	Submit Spark to Hadoop Cluster	L: Chap 7	Spark Submit	Midterm Project Report (3/2)
3/6-3/10	Spring Break			
3/14, 3/16	K-means Clustering	L: Chap 11	Clustering small PubMed abstracts Using Spark	
3/21, 3/23	Scalable Machine Learning and Distributed Statistical Query	L: Chap 11	Clustering massive PubMed abstracts Using Spark-submit and MLlib	
3/28, 3/30	Advanced Spark Programming	L: Chap 6		
4/4, 4/6	Tuning and Debugging Spark	L: Chap 8	Configure Spark for Performance	
4/11, 4/13	Stream Processing Using Spark	L: Chap 10		

4/18, 4/20	Emerging Topic: Network Analysis at Scale			
4/25, 4/27	Project Presentation			Project Presentation
4/25-4/28	Final Project Demonstration			Project Demo
5/1	Final Project Report			Final Project Report

Course Materials: Additional reading materials and papers will be posted in Canvas.

Contact Information for Instructor and TA:

Instructor	Prof. John Yen
Office	313J IST Building
Phone	(814) 865-6174
Office Hours	Tu Th 4:30-5:30 pm or by Appointment
E-mail	jyen@ist.psu.edu
Web Site	http://faculty.ist.psu.edu/yen
TA	Yu Luo
Office Hours	F 2-3 pm
E-mail	yzl5709@ist.psu.edu

Course Policies:

- Due to many in-class assignments of the course, attendance of the course is mandatory. Excused absences need to be approved by the instructor before the class to be missed. After three unexcused absences, penalty (10% of class attendance for each absence) will be applied to the final grade. A zero will be assigned (to the absent student) for each unexcused absence from in-class assignments.
- Late homework will receive a penalty of 25% for each day after the due date.
- Questions and class participation are encouraged and will be taken into consideration in the final grade.
- **Academic Integrity: According to the Penn State Principles and University Code of Conduct:** Academic integrity is the pursuit of scholarly activity in an open, honest and responsible manner. Academic integrity is a basic guiding principle for all academic activity at The Pennsylvania State University, and all members of the University community are expected to act in accordance with this principle. Consistent with this expectation, students should act with personal integrity, respect other students' dignity, rights and property, and help create and maintain an environment in which all can succeed through the fruits of their efforts. Academic integrity includes a commitment not to engage in or tolerate acts of falsification, misrepresentation or deception. Such acts of dishonesty violate the fundamental ethical principles of the University community and compromise the worth of work completed by others. Academic dishonesty

includes, but is not limited to, cheating, plagiarism, fabrication of information or citations, facilitation of acts of academic dishonesty by others, unauthorized possession of examinations, submitting work of another person or work previously used without informing the instructor, and tampering with the academic work of other students (also see
Faculty Senate Policy 49-20 and G-9 Procedures).

- **Affirmative Action & Sexual Harassment:** The Pennsylvania State University is committed to a policy that all persons shall have equal access to programs, facilities, admission, and employment without regard to personal characteristics not related to ability, performance, or qualifications as determined by University policy or by Commonwealth or Federal authorities. Penn State does not discriminate against any person because of age, ancestry, color, disability or handicap, national origin, race, religious creed, sex, sexual orientation, or veteran status. Direct all inquiries to the Affirmative Action Office, 328 Boucke, University Park, PA 16802, (814) 863-0471.
- **Americans with Disabilities Act:** The College of Information Sciences and Technology welcomes persons with disabilities to all of its classes, programs, and events. If you need accommodations, or have questions about access to buildings where IST activities are held, please contact us in advance of your participation or visit. If you need assistance during a class, program, or event, please contact the member of our staff or faculty in charge.
- **An Invitation to Students with Learning Disabilities:** It is Penn State's policy to not discriminate against qualified students with documented disabilities in its educational programs. If you have a disability-related need for modifications in your testing or learning situation, your instructor should be notified during the first week of classes so that your needs can be accommodated. You will be asked to present documentation from the Office of Disability Services (located in 116 Boucke Building, 863-1807) that describes the nature of your disability and the recommended remedy. You may refer to the Nondiscrimination Policy in the Student Guide to University Policies and Rules.

Grading:

Evaluation of knowledge and understanding of materials will be based on term project, lab assignments, quizzes, and in-class activities.

Project Proposal	5 %
Mid-term Project Report	5 %
Project Demo and Presentation	10 %
Final Project Report	30 %
Quiz	10%
Attendance/In-class Activities	5%
Lab Assignments	35 %
Total	100%