CIS 798: Programming Techniques for Data Science and Analytics
Fall 2018

Hours: 3 hours; extended course project option (CIS 597/598, 690) available
Prerequisite: CIS 200 (Fundamentals of Programming) or CIS 209 (C Programming for Engineers) or instructor permission; **programming background (Python, R, or Matlab for analytics) recommended**
Textbook: Drabas, T., & Lee, D. (2017). *Learning PySpark*. Birmingham, UK: Packt Publishing. URL: http://bit.ly/learning-pyspark-mapt
Venue: MWF 12:30 – 13:20 U.S. Central Time, 1116 Engineering Hall (Reference #15132, CIS 490: #15130)
Instructor: William H. Hsu, Department of Computer Science (http://www.cs.ksu.edu/~bhsu, bhsu@ksu.edu)
Office: 2164 Engineering Hall (DUE) Google Voice (office/home/cell): +1 785 236 8247
TA: CS graduate TA (TBD), 1119 DUE Instructional alias: bigdataclassta@listserv.ksu.edu
Schedule: http://www.kddresearch.org/page/53 Research & teaching wiki: http://www.kddresearch.org
Office hours: 10:30 – 11:30, 15:30 – 16:30 Mon, Wed; 08:00 – 09:00 Fri; 09:00 – 10:00 Tue; 18:30 Thu for distance students; by appointment
Web page: http://bit.ly/kstate-datascience-class (public)
MediaSite lectures: Linked from both K-State Canvas (official) and mirror (public)

Course Description

This is an intermediate (upper undergraduate / beginning graduate-level) course on programming techniques for data science, including large (terascale) data sets and complex, heterogeneous data from various domains. It is intended for students who have had at least one introductory programming course and **know or are prepared to learn the Python programming language along with requisite software libraries. These include NumPy, SciPy, PySpark, and one or more of: scikitlearn, PyTorch, and TensorFlow.** Prior experience with technical computing (numerical and statistical computations) using Matlab/Octave or R is a plus, but no additional background is assumed. The course will survey programming concepts that underlie MapReduce, Apache Hadoop and Spark, and specific tools such as Apache Hive, Storm, and Flink. It will give a practical hands-on introduction to programming tools for statistical computing in big data environments and methods for evaluation and validation of results. NoSQL, search tools, graph databases, and visualization will be presented in the context of real-world problems involving data integration and transformation. Students will be given full-scale data sets and problems to work with and may bring data and problems to work on from other disciplines.

Recommended text (on reserve in K-State Library, excerpts on Canvas)

Hwang, K. (2017). *Cloud Computing for Machine Learning and Cognitive Applications*. Cambridge, MA, USA: MIT Press. URL: http://bit.ly/cloud-computing-mit
Hastie, T., Tibshirani, R. & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition.* URL: http://www.springer.com/gp/book/9780387848570
Scott, J. A. (2015). *Getting Started with Apache Spark: From Inception to Production.* San Jose, CA, USA: MapR Technologies, Inc. URL: https://mapr.com/ebooks/spark/
(Additional bibliography on text analytics, deep learning, other machine learning in online materials)

Course Requirements

Component	Components	Grade Value	Total Value
Exams and quizzes	2 online hour exams	20% (10% each)	45%
	1 online final exam	20%	
	5 of 6 online quizzes	5% (1% each)	
Homework	5 of 6 programs	10% (2% each)	20%
	5 of 6 exercise sets (written)	10% (2% each)	
Term project	Planning/design, interview	8%	25%
	Intermediate milestone	8%	
	Implementation, report	8%	
	Peer review	1%	
Class participation	Forum participation	6%	10%
	Answering questions/discussion	2%	
	10 of 15 labs	2%	

Course Calendar, Syllabus, and Readings (textbook chapters unless otherwise indicated)

Lecture	Date	Topic	Reading (Before Class)
0	Mon 20 Aug 2018	MapReduce/Hadoop 1 of 3; Overview	Missive; Preface, (Ch.) 1
1	Wed 22 Aug 2018	MapReduce/Hadoop 2 of 3: **BYOD**	2; 1-2 Lin; BigTable Article
2	Fri 24 Aug 2018	MapReduce/Hadoop 3 of 3: Tools	3 Lin; Stats Handout
3	Mon 27 Aug 2018	Python, **Amazon Web Services**	Python Handout
4	Wed 29 Aug 2018	Python Streaming; **Full Scala/sbt Intro**	Streaming Handout
5	Fri 31 Aug 2018	Introduction to Queries & SQL; **Scalding**	SQL Handout; 5
6	Wed 05 Sep 2018	Pig & Hive 1 of 3: Queries; **Google Cloud**	Pig & Hive Handouts; 1
7	Fri 07 Sep 2018	Pig & Hive 2 of 3: Exercises; **Azure**	Pig & Hive Handouts
8	Mon 10 Sep 2018	Pig & Hive 3 of 3; **More Scala/sbt**	FP Basics Handout
9	Wed 12 Sep 2018	PySpark 1 of 3: Overview; **IBM Bluemix**	PySpark Handout
10	**Fri 14 Sep 2018**	**PySpark 2 of 3; Exam 1 Review**	**PySpark Handouts**
11	Mon 17 Sep 2018	PySpark 3 of 3; AWS Deployment	PySpark Handouts
12	Wed 19 Sep 2018	Project Topics Review	Projects Handout
13	**Fri 21 Sep 2018**	**Exam 1; Flink**	**Avro Handout**
14	Mon 24 Sep 2018	MR Design Patterns 1 of 3: Basics	UMAP Papers; 2
15	Wed 26 Sep 2018	MR Design Patterns 2 of 3: Exercise, **Drill**	RecSys Survey, Papers; 3
16	**Fri 28 Sep 2018**	**MR Design Patterns 3 of 3: Apps**	**Cloudera Handout; 4**
17	Mon 01 Oct 2018	NoSQL 1 of 5: Intro, CAP, Sharding	NoSQL Handout; 2-3
18	Wed 03 Oct 2018	NoSQL 2 of 5: Key-Value Stores	Key-Value Handout; 4
19	Fri 05 Oct 2018	NoSQL 3 of 5: **Cassandra (Columnar)**	Cassandra Handout; 5
20	Mon 08 Oct 2018	NoSQL 4 of 5: **MongoDB (Document)**	6; Lin 4; Mongo Handout
21	Wed 10 Oct 2018	NoSQL 5 of 5: **Neo4j (Graph)**	7; Lin 4; Graph Handout
22	**Fri 12 Oct 2018**	**MR Synopsis; Exam 2 Review**	**8; Lin 5, ETL Handout**
23	Mon 15 Oct 2018	Data Mining 1 of 3: Machine Learning	Mahout Handout
24	Wed 17 Oct 2018	Data Mining 2 of 3: sklearn/MLLib, Mahout	Learning Handout
25	**Fri 19 Oct 2018**	**Data Mining 3 of 3: KDD; Exam 2**	**KDD, CRM/BI Handouts**
26	Mon 22 Oct 2018	Search 1 of 5: Indexing	Search Handout; Lin 4
27	Wed 24 Oct 2018	Search 2 of 5: PageRank	3; Lin 6; Search Handout
28	**Fri 26 Oct 2018**	**Search 3 of 5: TFIDF, Text Analytics**	**Text Analytics Handout**
29	Mon 29 Oct 2018	Search 4 of 5: Solr	Solr Handouts
30	Wed 31 Oct 2018	Search 5 of 5: ElasticSearch	ElasticSearch Handout
31	Fri 02 Nov 2018	Mesos, **Lambda Architecture**	Shard Basics Handout
32	**Mon 05 Nov 2018**	**Synopsis, Final Review**	**NoSQL Handout**
33	Wed 07 Nov 2018	KVP vs. Columnar DBs	NoSQL Handout
34	Fri 09 Nov 2018	Graph Databases 1 of 3 :More Neo4j	Graph & Neo4j Handouts
35	**Mon 12 Nov 2018**	**Graph Databases 2 of 3 :Examples**	**Neo4j Handout**
36	Wed 14 Nov 2017	Graph Databases 3 of 3 :Applications	Graph & Neo4j Handouts
37	Fri 16 Nov 2018	Statistics & Analytics 1 of 3: Descriptive	Lin 5; Stats Handout
38	Mon 26 Nov 2018	Statistics & Analytics 2 of 3: Big Data	Stats Handout
39	Wed 28 Nov 2018	Statistics & Analytics 2 of 3: Info Vis	Visualization Handout
	Fri 30 Nov 2018	Final Projects & Peer Review 1	N/A (Student Handouts)
	Mon 03 Dec 2018	Final Projects & Peer Review 2	N/A (Student Handouts)
40	Wed 05 Dec 2018	Performance & Scalability Issues	Lin 7; HPC Handout
41	Fri 07 Dec 2018	Advanced Big Data Analytics Topics	Advanced Topics Handout
		PROJECT SUBMISSIONS DUE	

Lightly-shaded entries (with bold text) denote the due date of a written problem set.
Heavily-shaded entries denote the due date of a machine problem (programming assignment).
Green-shaded entries denote the date of a lab (distance students may participate online that week).
Green bold-faced text denotes the date of an exam preparation review.
Red bold-faced text denotes the date of an exam (distance students may use online proctoring that week),
Blue bold-faced text denotes the date of a post-exam (model solution) review.
Project proposals are due on the day of Lecture 23. Interim project interviews will be held right after the second hour **exam**.
The **highlighted date** is the due date of the draft project report and demo, with interviews and presentations to be held the last two weeks of class.