CIS 731 ZA: Programming Techniques for Data Science and Analytics

Spring 2020

Hours: 3 hours; extended course project option (CIS 690, 798, 890) 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: Online - https://k-state.instructure.com/courses/91181 (Reference #18137)
Instructor: William H. Hsu, Department of Computer Science (http://www.cs.ksu.edu/~bhsu, bhsu@ksu.edu)
Office: 2178 Engineering Hall (DUE) Google Voice (office/home/cell): +1 785 236 8247
TA: CS graduate TA, 2221 DUE Instructional alias: bigdataclassta@listserv.ksu.edu
Schedule: http://www.kddresearch.org/page/53 Research & teaching wiki: http://www.kddresearch.org
Office hours: 13:00 – 14:00 Mon; 14:00 – 15:00 Tue; 09:30 – 10:30 Wed; 08:30 – 09:30 Fri; by appointment
Web page: http://bit.ly/kstate-datascience-class (public)
MediaSite lectures: Linked from both K-State Canvas (official) and mirror (public)
GTA: Luis Bobadilla (happystep@ksu.edu), lab/TA office phone +1 785 380 7422
GTA office hours: 16:00-17:00 Mon, 09:00-11:00 Wed in DUE 1119

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 PySpark, MongoDB, scikitl-earn, PyTorch**. 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, Spark SQL, Solr, Flink, and Kafka. 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%	

Calendar/Syllabus/Readings: Miner & Shook (MS), Lin & Dyer (LD), Drabas & Lee (DL), etc.

Lecture	Date	Topic	Reading (Default: MS)
0	Wed 22 Jan 2020	MapReduce / PySpark 1 of 3; Overview	Missive; (Ch.) 1; DL 1
1	Fri 24 Jan 2020	MR / PySpark 2 of 3: **RDDs**; **BYOD**	2; LD 1-2; BigTable Hand.
2	Mon 27 Jan 2020	MR / PySpark 3 of 3: Five Basic Patterns	LD 3.1-3.4
3	Wed 29 Jan 2020	Python, **Amazon Web Services**	LD 3.5-3.6, Python Hand.
4	Fri 31 Jan 2020	PySpark in Jupyter; **Google Colab**	DL 2
5	Mon 03 Feb 2020	Introduction to Queries & SQL	5; SQL Handout
6	Wed 05 Feb 2020	Pig & Hive 1 of 3: Queries; **Google Cloud**	1; Pig & Hive Handouts
7	Fri 07 Feb 2020	Pig & Hive 2 of 3: Exercises	5; MapR, Pig, Hive Hand.
8	Mon 10 Feb 2020	Pig & Hive 3 of 3; **More PySpark**	PySpark Pipeline Handout
9	Wed 12 Feb 2020	PySpark 1 of 3: Transformations, Actions	3; FP, Spark Handouts
10	**Fri 14 Feb 2020**	**PySpark 2 of 3; Exam 1 Review**	**DL 2; PySpark Hand.**
11	Mon 17 Feb 2020	PySpark 3 of 3; **DataFrames**; cogroup	5; PySpark Handouts (3)
12	Wed 19 Feb 2020	Pipelines 1; **AWS Deployment**	DL 3
13	**Fri 21 Feb 2020**	Pipelines 2; **Azure/GCP/Atlas**; **Exam 1**	**Avro/ORC Handout**
14	Mon 24 Feb 2020	MR Design Patterns 1 of 3: Basics	Project Handout
15	Wed 28 Feb 2020	MR Design Patterns 2 of 3: Exercise	UMAP, OSEMN Handouts
16	**Fri 28 Feb 2020**	**MR Design Patterns 3 of 3: Algorithms**	2-3; Spark MR Handout
17	Mon 02 Mar 2020	MR Apps 1 / 5: Analytics, Docker, Projects	4-5
18	Wed 04 Mar 2020	MR Apps 2 / 5: SaaS, BI, CRM, Watson	Analytics Handout
19	Fri 06 Mar 2020	MR Apps 3 / 5: Cloud, NoSQL; **GCP/Atlas**	Cloud Computing Handout
20	Mon 16 Mar 2020	MR Apps 4 / 5: **AWS/Azure**, IR, **Projects**	NoSQL, Scalability Hand.
21	Wed 18 Mar 2020	MR Apps 5 / 5: IR **– ELK/Elasticsearch**	LD 4; NoSQL Handout
22	**Fri 20 Mar 2020**	**MR Synopsis; Search; Exam 2 Review**	**Elasticsearch Handout**
23	Mon 23 Mar 2020	Data Science 1 of 5: Data Mining	ETL, sklearn Handouts
24	Wed 25 Mar 2020	Data Mining 2 of 5: Machine Learning	Data Mining Handout
25	**Fri 27 Mar 2020**	**Data Mining 3 of 5: Evaluation; Exam 2**	Data Mining Handout
26	Mon 30 Mar 2020	Data Mining 4 of 5: Association, Recsys	Recsys Handout
27	Wed 01 Apr 2020	Data Mining 5 of 5: **sklearn**, other APIs	sklearn Handout
28	**Fri 03 Apr 2020**	**NoSQL 1 of 3: MongoDB in depth**	**MongoDB Handout**
29	Mon 06 Apr 2020	NoSQL 2 of 3: MongoDB uses	MongoDB Handout
30	Wed 08 Apr 2020	NoSQL 3 of 3: Cassandra & Neo4j	Cassandra, Neo4j Hand.
31	Fri 10 Apr 2020	Search/Analytics – **Solr**	Search Handout
32	Mon 13 Apr 2020	Infovis/Analytics 1 of 2: **Logstash/Kibana**	Infovis Handout
33	**Wed 15 Apr 2020**	**Infovis/Analytics 2 of 2: Tableau, Plotly**	**Infovis Handout**
34	Fri 17 Apr 2020	Stats & Analytics 1 of 3: descriptive	Statistics Handout
35	Mon 20 Apr 2020	Stats & Analytics 2 of 3: streaming; **Flink**	Flink Handout
36	**Wed 22 Apr 2020**	**Stats & Analytics 2 of 3: networks/SNA**	**Graph/Neo4j Handouts**
37	Fri 24 Apr 2020	Real-time, IoT, **Lambda Architecture**	Lambda Handout
38	Mon 27 Apr 2020	Project Requirements, Writeup	Writing Handout
39	Wed 29 Apr 2020	Project Presentation	Talk Handout
	Fri 01 May 2020	Final Projects & Peer Review 1	N/A (Student Handouts)
	Mon 04 May 2020	Final Projects & Peer Review 2	N/A (Student Handouts)
40	Wed 06 May 2020	Performance & Scalability Issues	HPC Handout
41	Fri 08 May 2020	Advanced Big Data & Analytics Topics	Advanced Topics Handout

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.