CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2019 Syllabus

San Diego State University -- This page last updated 1/24/19

CS 696 Intro to Big Data: Tools and Methods Fall 2019

Instructor	Roger Whitney
Office	GMCS-561
Phone	619-594-3535
Email	whitney at sdsu.edu
Office Hours	3:15-5:15 pm Tuesday, Thursday

Course WWW Site: http://www.eli.sdsu.edu/courses/spring19/cs696/index.html
The course lecture notes, assignments, course videos and course wiki available at above course web site.

Texts:

Python for Data Analysis, Wes McKinney, O'Reilly Media, Inc, October 10, 2017, 978-1-4919-5766-0

Spark: The Definitive Guide, Matei Zaharia, Bill Chambers, February 2018, 9781491912218

The books are available on-line via SDSU library. See course website for links. In general I use on-line documentation and multiple texts to understand the different technologies we will use in this course. As the course progresses I will add to this list, but will restrict the books to those that are available for free on-line.

Prerequisites: CS310 (Data Structures), Math 254, Math 245

Spark & Amazon's Cloud:
The class will use Amazon's cloud to run Spark on a cluster. Students will create their own accounts on Amazon cloud. This will incur a cost to the student which should be less than the cost of a text book. You can apply for a AWS Education account which gives you $100 of compute time for free. See course web site for link to Amazon's AWS Education program.

Short Description:
Tools, methods and practices to analysis, curate, search, query, visualize big data, ie data characterized by volume, variety, velocity, variability and veracity. These tools, methods include statistics, machine learning, visualization, no-sql databases, MapReduce, Hadoop**.**

Detailed Description:
This course assumes no prior experience with Data Science or Big Data. The goal of Big Data is to gain insight by analyzing data. The course will look at some statistical tech-

niques, machine learning and visualization to help analyze data. However most of the course will focus on software tools (Spark, Cassandra, Kafka, Jupyter Notebooks) to process data on clusters. We will use Jupyter notebooks in assignments to combine graphs, code, and text. Before one can analyze data it needs to be stored. While CSV and DataFrames are fine for small datasets we will look at Hierarchical Data Format (HDF5) and the Hadoop Distributed File System (HDFS). The latter needed when datasets are so large we need multiple machines to process the data. Also for storing data we will use a NoSQL database. NoSQL databases are common in Big Data. Multiple machines are required to analysis larger datasets. We will use Spark to scale to multiple machines. We will be using a cloud provider (Amazon) to run Spark on multiple machines. You will have to create an account on Amazon AWS. Using multiple machines to solve one problem introduces a number of issues. The problem has to be divided into separate pieces that can be solved independently. Data has to be distributed and results need to be communicated between machines. If not done correctly we will not see the benefit of using multiple machines.

While this is the second time the course has been offered it is being changed a fair amount. The course uses a lot of different technologies is a rapidly changing area. As a result there will be some rough edges to the course. You might have problems installing some software on your machines, or have connection issues. This is why the course is offered at the graduate level as graduate students should have the ability to deal with these issues.

Topics Covered

Scala
Hadoop HDFS
Spark
    RDD, Spark SQL, Spark ML
Visualization - Matplotlib
Basic Statistics
    Correlation, conditional probability
No-SQL database
    Cassandra
Machine Learning
    Supervised, Unsupervised
    Linear Regression, K-nearest neighbor
Distributed Streaming
    Kafka

Learning Outcomes:

- Students will be able to implement programs to process and analyze big data
- Students will be able to use utilize Spark and no-sql databases to process and analyze big data
- Students will be able to set up a data pipeline to ingest data from databases (SQL or NO-SQL) and process the data on a cluster
- Use Notebooks to organize & run code, display and explain results

Information & Downloads:

Jupyter - http://jupyter.org/
Spark - http://spark.apache.org/
Kafka https://kafka.apache.org/
Cassandra http://cassandra.apache.org

Grading: Your grade in this course will be determined given below. There will be about 5 assignments.

Homework, Programs	55%
Exams (1)	25%
Project	20%

Crash Policy: After registration closes adding students to the class will be handled via the waitlist system. If you are not able to register for this class you need to add your self to the class waitlist. The instructor has no ability to add any student to the course nor can the instructor see who is on the wait list.

Late Policy: Late homework will be accepted, but with a penalty. An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 80%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.

Email & Assignments: Unless indicated by the assignment all assignments are to be submitted to the course repository. No assignments will be accepted via email.

No Extra Credit: There will not be any extra credit assignments. There will not be any extra credit problems in the assignments.

Cheating: Any one caught cheating will receive an F in the course.

Disabled Students: If you are a student with a disability and believe you will need accommodations for this class, it is your responsibility to contact Student Disability Services at (619) 594-6473. To avoid any delay in the receipt of your accommodations, you should contact Student Disability Services as soon as possible. Please note that accommodations are not retroactive, and that accommodations based upon disability cannot be provided until you have presented your instructor with an accommodation letter from Student Disability Services. Your cooperation is appreciated.