COURSE OVERVIEW

Organizations use their data for decision support and to build data-intensive products and services, such as recommendation, prediction, and diagnostic systems. The collection of skills required by organizations to support these functions has been grouped under the term Data Science. This course will attempt to articulate the expected output of Data Scientists and then teach students how to use PySpark (part of Apache Spark) to deliver against these expectations. The course assignments include Log Mining, Textual Entity Recognition, and Collaborative Filtering exercises that teach students how to manipulate datasets using parallel processing with PySpark.

PREREQUISITES

This course covers advanced undergraduate-level material. It requires a programming background and experience with Python (or the ability to learn it quickly). All exercises will use PySpark (part of Apache Spark), but previous experience with Spark or distributed computing is NOT required. You should take this [Python mini-quiz](http://www.mypythonquiz.com/) before the course and take this [Python mini-course](http://ai.berkeley.edu/tutorial.html#PythonBasics) if you need to learn Python or refresh your Python knowledge.

PIAZZA DISCUSSION GROUP

We are using Piazza for all course questions and discussions. Please sign up at the discussion forum here: <https://piazza.com/edx_berkeley/summer2015/cs1001x> (access key: **cs1001x**)

COURSE CONTENT

**Week 1:** Big Data and Data Science - *Launches June 1 at 16:00 UTC*

* + **Introduction to Big Data and Data Science** - learn about big data and see examples of how data science can leverage big data
  + **Performing Data Science and Preparing Data** - explore data science definitions and topics, and the process of preparing data
  + **Setting up the Course Software Environment** **(Due June 6, 2015 at 00:00 UTC)** - download and install the course software, run your first Apache Spark notebook, and submit your first assignment

**Week 2:** Introduction to Apache Spark  - *Launches June 6 at 16:00 UTC*

* + **Big Data, Hardware Trends, and the History of  Apache Spark** - discuss big data and hardware trends, and learn about the history of Apache Spark
  + **Spark Essentials** - learn about Spark's Resilient Distributed Datasets, transformations, and actions
  + **Lab 1: Learning Apache Spark (Due June 12, 2015 at 00:00 UTC)** - perform your first course lab where you will learn about the Spark data model, transformations, and actions, and write a word counting program to count the words in all of Shakespeare's plays

**Week 3:** Data Management  - *Launches June 13 at 16:00 UTC*

* + **Semi-Structured Data** - explore the concept of semi-structured data and how tabular data is handled in Spark
  + **Structured Data** - learn about structured data, the relational data model, SQL, and joins in SQL and Spark
  + **Lab 2: Web Server Log Analysis with Apache Spark (Due June 19, 2015 at 00:00 UTC)** - use Spark to explore a NASA Apache web server log in the second course lab

**Week 4:** Data Quality, Exploratory Data Analysis, and Machine Learning  - *Launches June 20 at 16:00 UTC*

* + **Data Quality** - learn about the challenges of data quality and cleaning
  + **Exploratory Data Analysis** - understand the statistics of Exploratory Data Analysis and data distributions
  + **Machine Learning** - learn about Spark's machine learning library, mllib
  + **Lab 3: Text Analysis and Entity Resolution (Due June 26, 2015 at 00:00 UTC)** - perform text analysis and entity resolution on Google and Amazon product listings using Spark in the third course lab

**Week 5:** Data Management  - *Launches June 27 at 16:00 UTC*

* + **Lab 4: Introduction to Machine Learning with Apache Spark (Due Jul 3, 2015 at 00:00 UTC)** - use Spark's mllib Machine Learning library to perform collaborative filtering on a movie dataset in the fourth course lab

COURSE GRADING POLICY

The components of the course grade are:

* + 8% Course software setup
  + 12% Quizzes in the lectures (scroll down below some of the videos for multiple choice and check box quizzes)
  + 80% Four lab exercises

The course is graded on the following scale:

* + 75-100: A grade
  + 55-74: B grade
  + 45-54: C grade
  + Below 45: non-passing grade

The course software setup assignment is due June 6, 2015 at 00:00 UTC. Each week's lab exercise will be due the following Friday at 00:00 UTC. We encourage you to start software setup and each exercise as early as possible. *There is an automatic 3-day grace period for submission deadlines. After the grace period, there is a 20% penalty for late submissions.*

CREDITS

This course is sponsored in part by [Databricks](http://databricks.com/) and [UC Berkeley's AMPLab](https://amplab.cs.berkeley.edu/).

The content in this course includes notes and content created by Dan Bruckner, John Canny, Sameer Farooqui, Michael Franklin, Paco Nathan, Kay Ousterhout,  Evan Sparks, Shivaram Venkataraman, Patrick Wendell, and Matei Zaharia.