BerkeleyX: CS105x Introduction to Apache Spark

Bookmarks

Week 1 - Apache Spark Programming Model

Lecture 1: Apache Spark Architecture and Programming Model
Quizzes

Setting up the Course Software Environment (Due September 10, 2016 at 23:59 UTC)
Setup

(Optional) Survey about your machine and setup experience

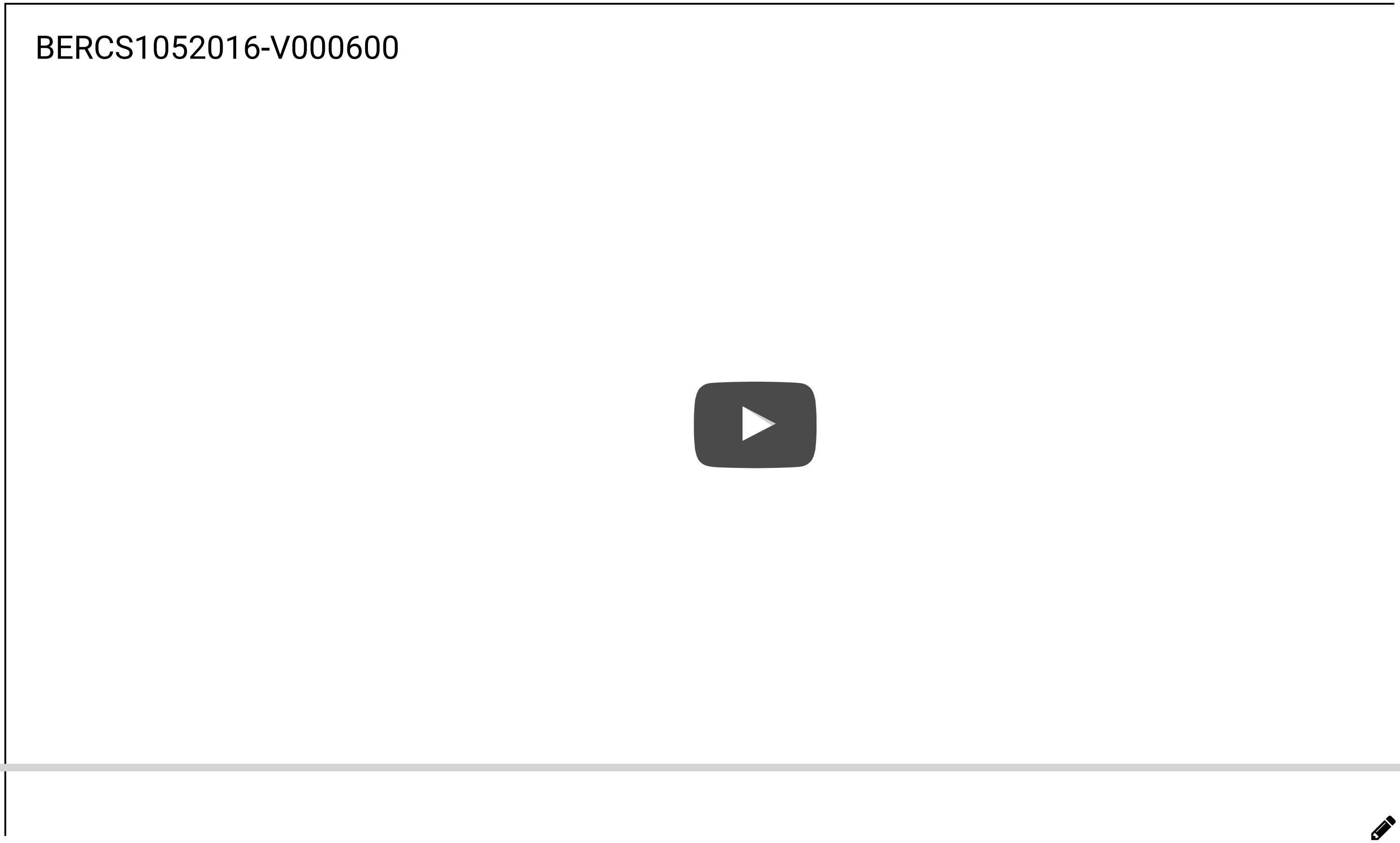
Week 2 - The Structured Query Language and Spark SQL

Week 3 - Analyzing Semi-Structured Data with Apache Spark

Week 1 - Apache Spark Programming Model > Lecture 1: Apache Spark Architecture and Programming Model > Sources of Big Data

Bookmark

Sources of Big Data



▶ 0:00 / 3:39 1.0x

Download video Download transcript .srt

Here is a huge list on GitHub of Awesome Public Datasets, most of which are free. After taking this course, you can download one of the datasets, import it into your Databricks Community Edition account, and explore it with Apache Spark!

(Optional Reading) This paper, Structured Open Urban Data: Understanding the Landscape, examines over 9,000 open data sets from 20 cities in North America, and presents general statistics about the content, size, nature, and popularity of the different data sets, and also examines the data quality issues and time-related aspects of the various datasets.

The City of San Francisco has an extensive collection of online city records. These data cover public safety, health, transportation, housing and many other topics. The data, together with public social media, can provide an unprecedented window into the City's operations. The data is freely available for anyone to explore. Many other cities are also putting their records online. Let's consider a couple of the types of questions one can ask using this data, but there are many others:

1. In the health section of sf.data.gov there is an extensive set of records about restaurant inspections. San Francisco has the largest number of restaurants per capita of any major city in the United States. Tracking and maintaining the quality of those restaurants is an ongoing challenge for inspectors. An interesting question would be could you create an "early warning system" based on social media (e.g., is it possible to predict restaurants in need of inspection from Yelp reviews?)? This question

could be partially answered by building a machine learning classifier using historical social media reviews (e.g., from Yelp or Trip Advisor) and city records of inspections.

2. The City receives many 3-1-1 reports (non-emergency incident reports from citizens). But some of these reports predict serious incidents, which may be recorded later in police reports. Consider the challenge of mining the CABLE reported incidents, and looking for text markers that predict future police reports. This challenge would require tying the two tabular datasets together, a process that is complicated by noisy data - would it be better to tie the datasets together by the name of the protagonist or the location (address) of the incident? This is an entity resolution problem, as the protagonists' names might be listed differently in the two datasets (e.g., Anthony Joseph versus A. Joseph), and the same applies to address information in the two datasets (e.g., SF City Hall versus 1 Dr Carlton B Goodlett Pl, San Francisco, CA 94102). You will explore the entity resolution problem in Lab #3. After combining the two datasets, the next step would be to look for keywords in the police report marking the type of incident, and attempt to predict incidents from the full text of the CABLE report.

These two are just two challenges, but there are many more are possible ones from this dataset.

Sources of Big Data

(1/1 point)

Which of the following items are potential sources of Big Data:

☑ Credit card transactions ✓

☑ City-block level weather (temperature, humidity, barometric pressure, light) measurements in a city ✓

☑ Tweets about an national vote ✓

☑ Traffic reports from the Waze crowdsource traffic application ✓

☑ Video viewing actions by edX students ✓

✓

Note: Make sure you select all of the correct options—there may be more than one!

EXPLANATION

These are all potential sources of Big Data! Big Data is all around us and can be generated by us, our actions, and our surroundings.

You have used 1 of 4 submissions