CS218

Lab #8 – Spark

This lab introduces Apache Spark, Resilient Distributed Datasets, and Running a simple Spark application

Goals of Lab

- Know the basic knowledge of Apache Spark, its architecture, and write a simple Spark Word Count Program.

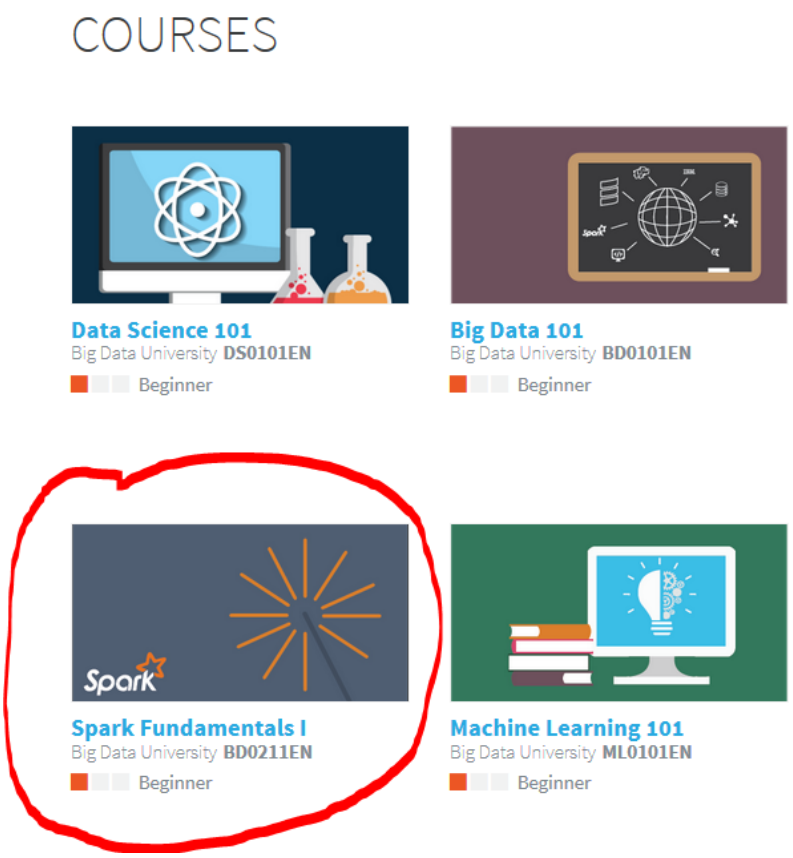
Pre-requisites

- VirtualBox with Linux

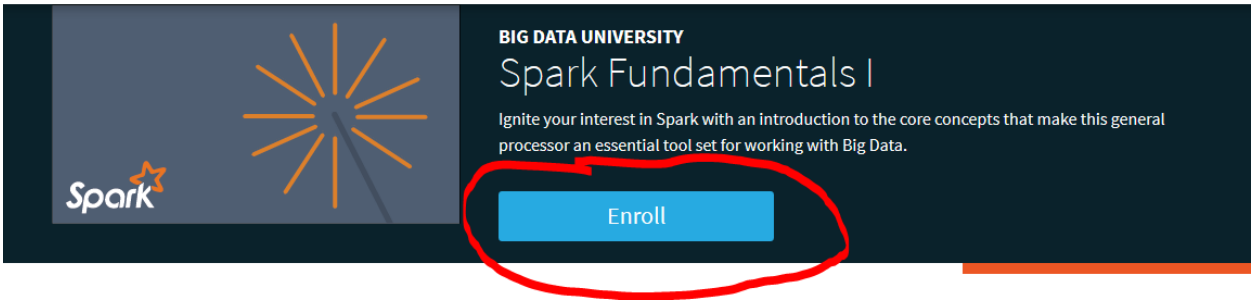
Section 1 – Introduction to Spark ecosystem, RDDs, and Applications

1. Register for the website:
   https://bigdatauniversity.com/

2. Go to Courses and Select Spark Fundamentals



3. Select Enroll:



4. Watch the videos for: **Module 1(Introduction to Spark), Module 2 (Resilient Distributed Dataset), and Module 3 (Application programming)**
    - Make sure to understand the benefits of Spark versus Hadoop's MapReduce
    - Make sure to understand RDDs (Resilient Distributed Datasets)

Homework

Note: Apache Spark release tested for this lab was 1.6.2 .

1. Answer the Review Questions for Module 1,2 and 3. Take a screenshot of both answered reviews.
2. In Module 3, Complete the Optional hands-on lab exercise (Download the Guide for Exercise 3).
    a. Setup the VM to run Spark with Scala. Download from: http://www.scala-sbt.org/download.html This is a tool already setup so that you do not have to setup Spark manually.
    b. Run the Sample Scala application to calculate the value of Pi. Take a screenshot of the output of the run of spark-submit.
3. Spark can be run with Java or Python as well. Continuing the same hands-on lab exercise, setup/write the WordCount program in Java based on the sample given. Run the wordcount program and take a screenshot of the output. Try the program with your own input textfile afterwards(Copy the contents of the textfile with wordcount along with the Java WordCount output).

Extra Credit

4. If you want to learn to setup Spark manually, you can install spark shell using the guide: https://www.tutorialspoint.com/apache_spark/apache_spark_installation.htm
5. Go to: https://www.tutorialspoint.com/apache_spark/apache_spark_core_programming.htm
    a. This time we will run the wordcount program from the command line (spark shell). Follow the tutorial in creating a WordCount program.
    b. Create your own input text file. Run the wordcount program. Paste the contents of your input textfile along with the contents of parts-00000 and part-00001.

Deliverables

- Create a {Microsoft Word| PDF } document containing the output from the homework section.
- Name the file <Last Name>_<First Name>_Lab08.{docx|pdf}
- Send the file to me via slack
  with the subject line "LAB## LastName FirstName" (ex.: LAB08 Smith John)

Lab Written By: Dennis Hsu