Parallel and Distributed Computing
Lab2
Basic RDDs

In this lab you will work the basic RDDs, maps, filters and reductions.

Getting the Cluster up and running

Log into DataBricks with the account you created. Go to the Cluster button and Create a New Cluster. Name the cluster whatever you want – although lab2cluster is a good name. Create it with the 6.4 databrick runtime. Once it has started, go to Workspace and create a new notebook. Your new notebook should be a Python notebook and be tied to the cluster you just started. You can call your notebook lab2.

Note that your Cluster will be good until you leave it idle for 2 hours. When you log into Databricks the next time you will find the Cluster is terminated, but still listed under Clusters. You can go to your cluster and Clone it. You can even give it the same name if you want. And then you can delete the old terminated cluster. Make sure you attach your notebook to the new cluster is just made.

Prime exercise

In one databricks notebook command, create Spark code that counts the number of primes in a given range. Start by first creating a Python list of all the numbers in the range 100..10,000. Then use Spark commands to create a parallel RDD from this list. Using only Spark map, filter, reduce and/or count, count the number of primes in this range in parallel. You may use lambdas or standard Python functions in your maps/filters/reductions.

Celsius exercise

In one databricks notebook command, create Spark code that works with temperature in the following way. Start with creating 1000 random Fahrenheit temperatures between 0..100 degrees F. This should be done in a standard Python list. Normally, we would load this data from 1000 different observations, but for this lab we will simply generate random test data. Next use Spark RDDs (only single ones – no pairRDDs) and only the Spark commands map, filter, reduce and/or count to first convert the full list to Celsius. Then find the average of all the Celsius values above freezing. You should print that average. You are only to use lambdas in your maps/filters/reductions. And you should persist RDDs if helps reduce computations.

Submission

To submit your work, you should first export your notebook. Please export just a Source File through the Notebook's Export option. This will produce a .py file. Please upload this file as your submission. Make sure you put your name in a comment on the first line of your source code.