Exercise 3. Intro to Spark and Data analytics

The first part of this is the simple K-means demo that introduces the main concepts in spark. The second part is an illustration of using Spark to analyze data. To run this you must either bring up Jupyter in your linux data science vm by going to https://youvm-IP-address:8000 and signing in with the vm userid and password you set in the last exercise.

Next you need to upload two notebooks. Login to your VM. In the shell type

```
$ cd notebooks
$ wget https://SciEngCloud.github.io/spark.ipynb
$ wget https://SciEngCloud.github.io/interactivelyExploreDataWithSpark.ipynb
```

Now refresh your jupyter page and you should see these two new notebooks. Go through the first one, spark.ipynb to get a feel for spark doing a map reduce style computation. This is actually a very poor k-means algorithm, but it does nicely illustrate how spark works. You can try changing the number of partitions to see how that changes the performance. You just need to change the variable in step 8 and then run steps 8 through to the end again.

Now do the DataExplore example. The data in this case is real data (but only about 140Mbytes) from the City of Chicago. It is from food service inspections. After you are done running the basics, see if you can find the names of some of the worst restaurants.

Each notebook has its own limited but essential documentation.

For the data exploration with spark example, we have another version of the exercise in which you deploy an HDInsight cluster. This takes more time, but if you want to try it, go to \Content\HDInsight\HDInsight Spark in the azure training course you downloaded from github.