

Exercise 3. Intro to Spark

This is the simple K-means demo that introduces the main concepts in spark. 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.

OR

Run the tutorial container

```
>docker run -it -p 8888:8888 dbgannon/tutorial
```

Both of these have spark built in. Unfortunately, the notebooks.azure.com site does not have spark installed. There is another azure solution for running spark on big data and we will discuss that later.

Load the notebook spark.ipynb (it is in the tutorial container.) or upload it to the data science vm from the tutorial package you installed earlier.

Open it up and follow the instructions in the notebook.