**Exercise n**

*Get started with Apache Spark and Python*

**Prior Knowledge**

Unix Command Line Shell

Simple Python

**Learning Objectives**

Understand the Spark system

Use the Spark Python shell to interactively work with data

Submit Spark jobs locally and using YARN

Write SparkSQL code in Python

**Software Requirements**

(see separate document for installation of these)

* Apache Spark 1.5.1
* Python 2.7.x
* Nano text editor or other text editor

**Part A. Spark Python Shell (pySpark)**

1. Apache Spark has a useful Python shell, which we can use to interactively test and run code. Since we have our data in HDFS, *we need to ensure HDFS is running.* (Follow the instructions from the Hadoop lab).
2. Let’s load some books into HDFS:
3. In a terminal window, change to the Spark directory:  
   cd ~/spark-1.5.1
4. Now start the Spark Python command line tool – pyspark  
   bin/pyspark  
   1. You should see a lot of log come up, ending in something like:

15/10/25 23:39:52 INFO BlockManagerMaster: Registered BlockManager

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_ / .\_\_/\\_,\_/\_/ /\_/\\_\ version 1.5.1

/\_/

Using Python version 2.7.6 (default, Jun 22 2015 17:58:13)

SparkContext available as sc, SQLContext available as sqlContext.

1. Once you have loaded the
2. Now let’s load some data. We already have a SparkContext object defined in the shell (in a program you need to define one, which we will see later)  
   1. Type   
      winds = sc.textFile("/usr/hduser/datafiles/books/\*.txt")