HU Extension Assignment 03 E63 Big Data Analytics

Handed out: 09/15/2017 Due by 11:59 AM EST on Saturday, 09/23/2017

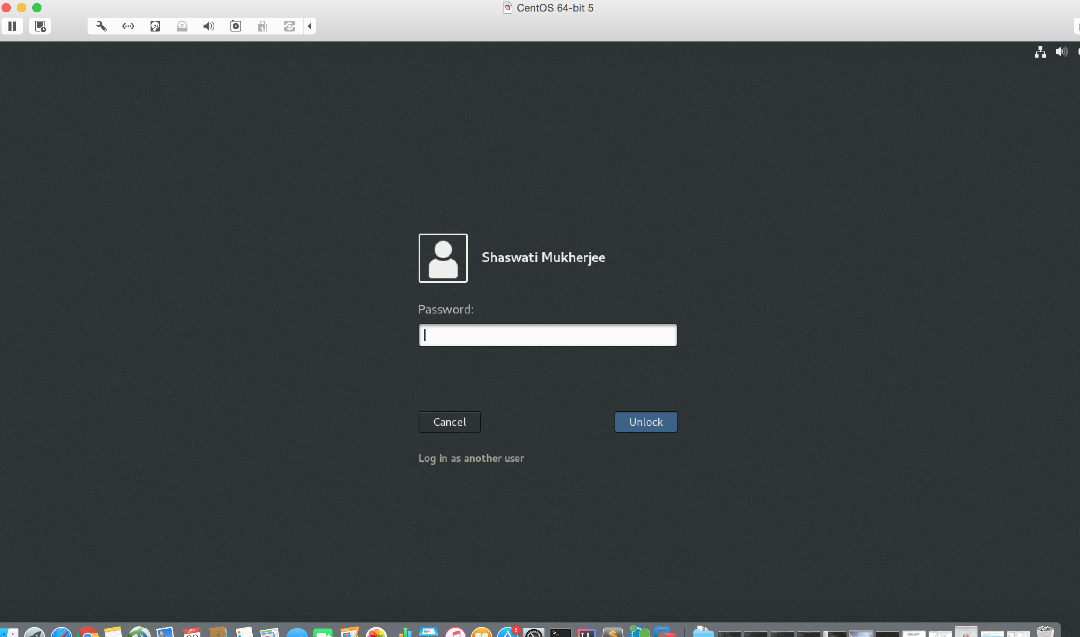
Use either VMWare Workstation or VMWare Fusion

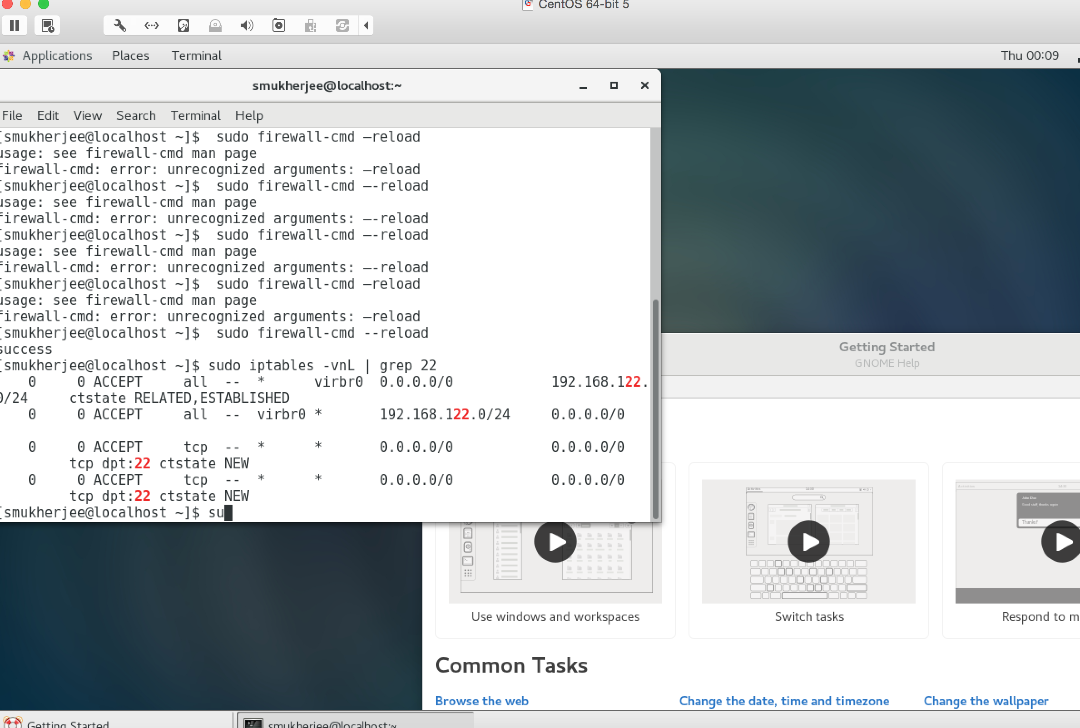
<https://e5.onthehub.com/WebStore/Welcome.aspx?ws=4185a0dc-d0d1-e511-9416-b8ca3a5db7a1&vsro=8>

If you feel comfortable with a virtualization technology different from VMWare, please be free to use it. Rather than creating a VM with CentOS, you can create the VM with a different flavor of Linux you more familiar with or consider better.

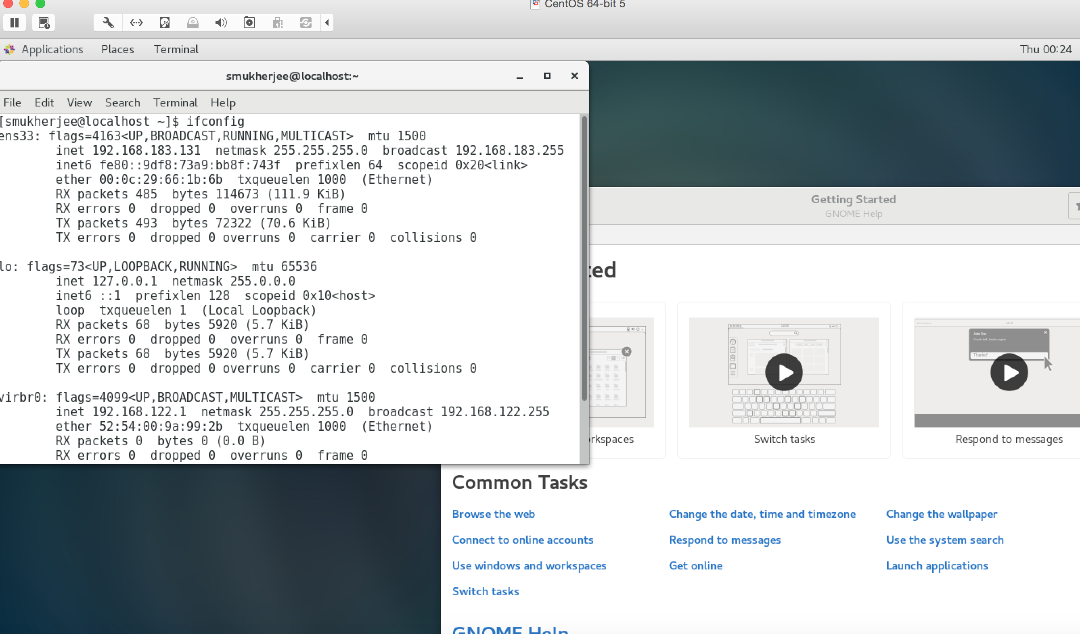
**Problem 1**. Create your own Virtual Machine with a Linux operating system. The lecture notes speak about CentOS. You are welcome to work with another Linux OS. When creating the VM, create an administrative user. Call that user whatever you feel like. Please record the password of the new user. Once the VM is created transfer the attached text file Ulysses10.txt to the home of new user. You can do it using scp (secure copy command) or email. Examine the version of Java, Python and Scala on your VM. If any of those versions is below requirements for Spark 2.2 install proper version. Set JAVA\_HOME environmental variable. Set your PATH environmental variable properly, so that you can invoke: java, sbt and python commands from any directory on your system. [20%]

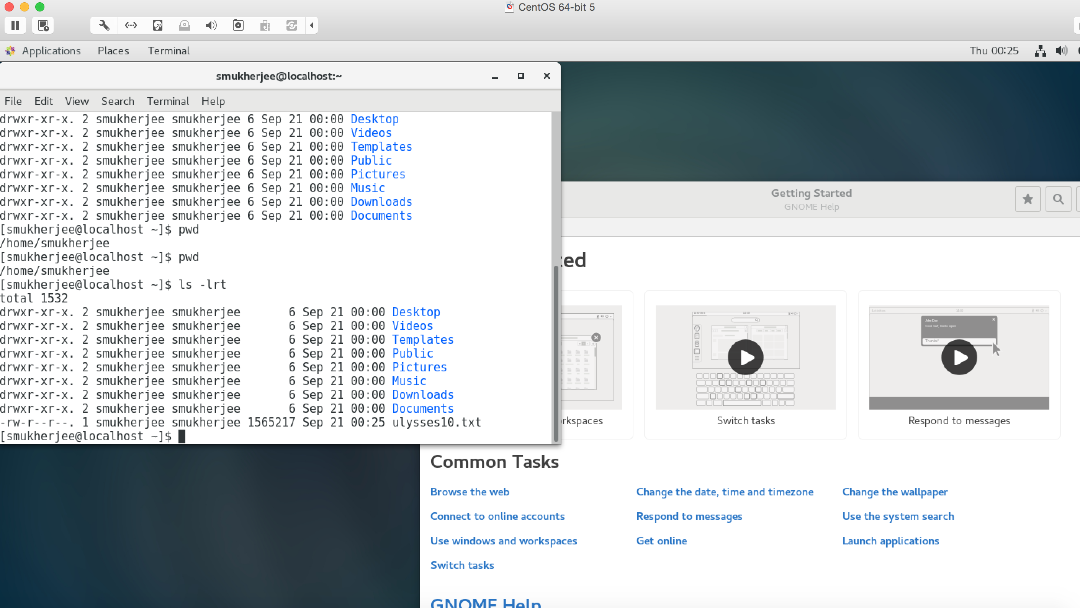
I followed the instructions on the slides to install VMWare and then CentOS on it, which I had downloaded. Installed Scala on it. Attaching some screenshots of it.

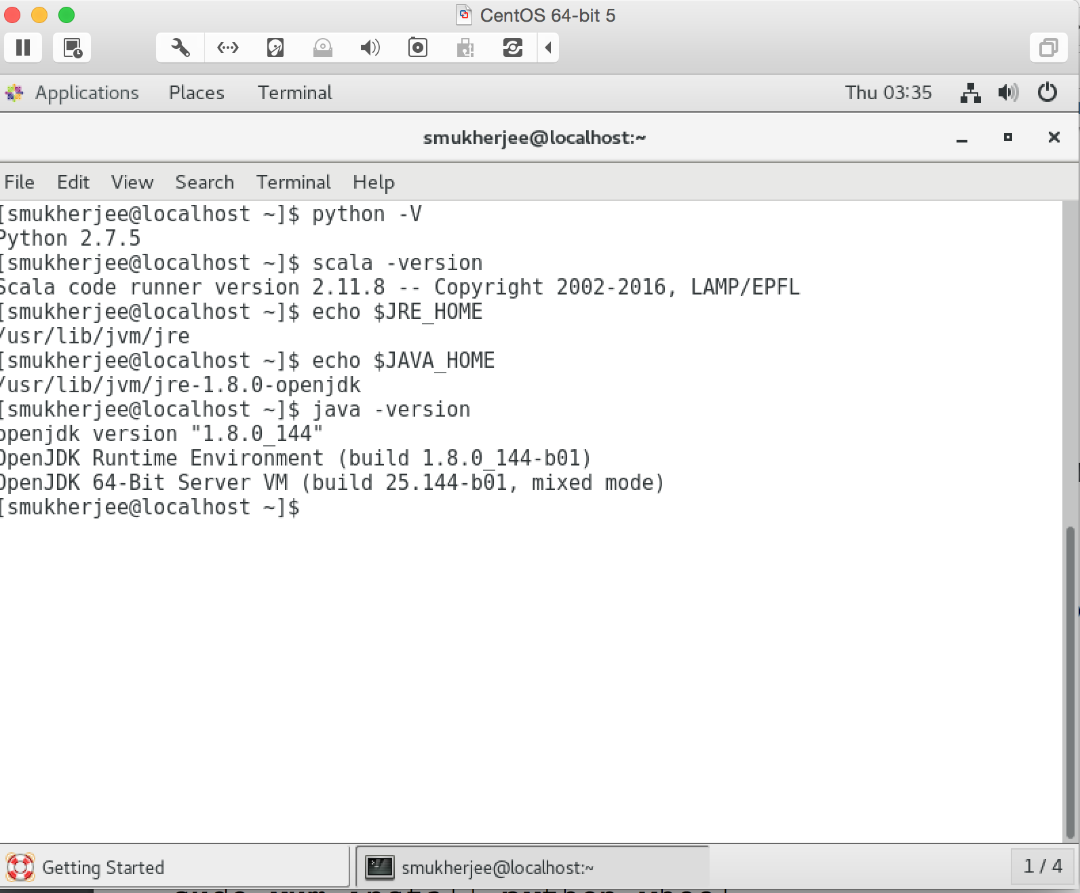






****

****

****

[smukherjee@localhost ~]$ sbt -version

[info] Loading project definition from /home/smukherjee/project

[info] Set current project to smukherjee (in build file:/home/smukherjee/)

[warn] The `-` command is deprecated in favor of `onFailure` and will be removed in 0.14.0

[smukherjee@localhost ~]$

**Problem 2**. Install Spark 2.2 on your VM. Make sure that pyspark is also installed. Demonstrate that you can successfully open spark-shell and that you can eliminate most of WARNing messages. [15%]

I installed spark 2.2 next. I also updated updated log4j.properties after creating it from log4j.properties.template. Then I updated the log4j.rootCategory.

smukherjee@localhost ~]$ spark-shell

Setting default log level to "ERROR".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

Spark context Web UI available at http://192.168.190.128:4040

Spark context available as 'sc' (master = local[\*], app id = local-1506153158500).

Spark session available as 'spark'.

Welcome to

      \_\_\_\_              \_\_

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

    \_\ \/ \_ \/ \_ `/ \_\_/  '\_/

   /\_\_\_/ .\_\_/\\_,\_/\_/ /\_/\\_\   version 2.2.0

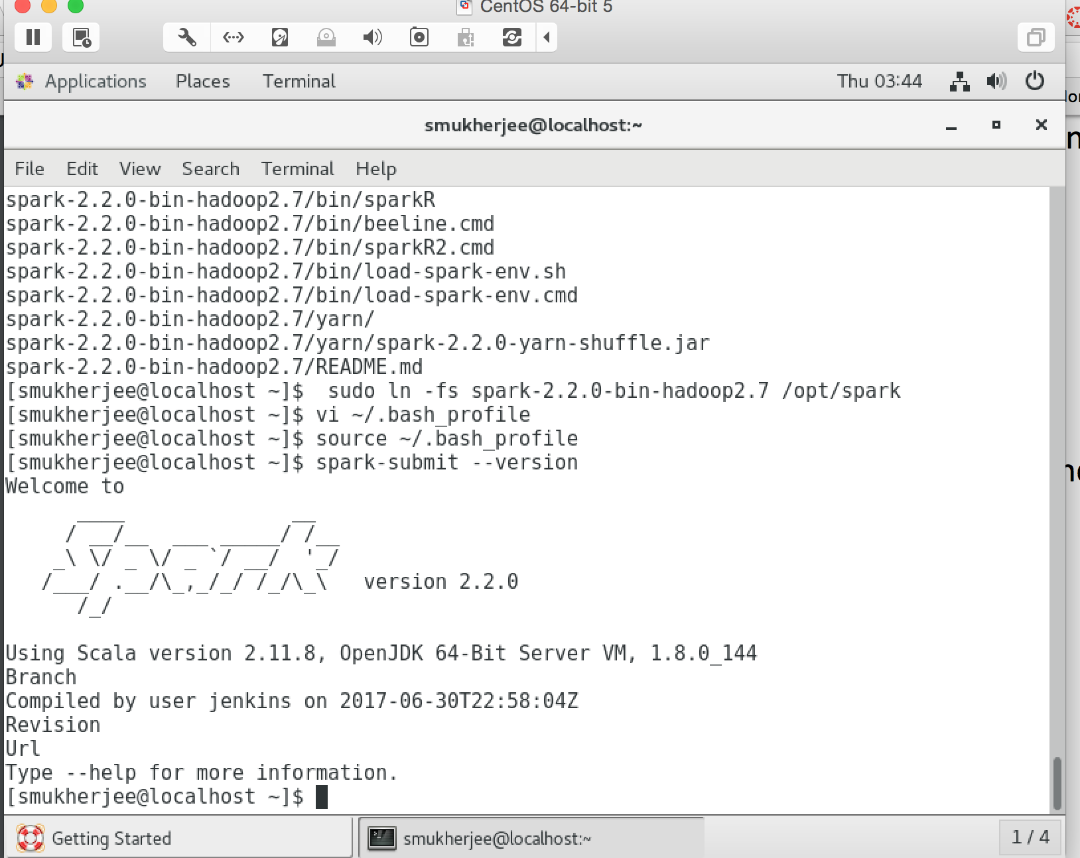
      /\_/

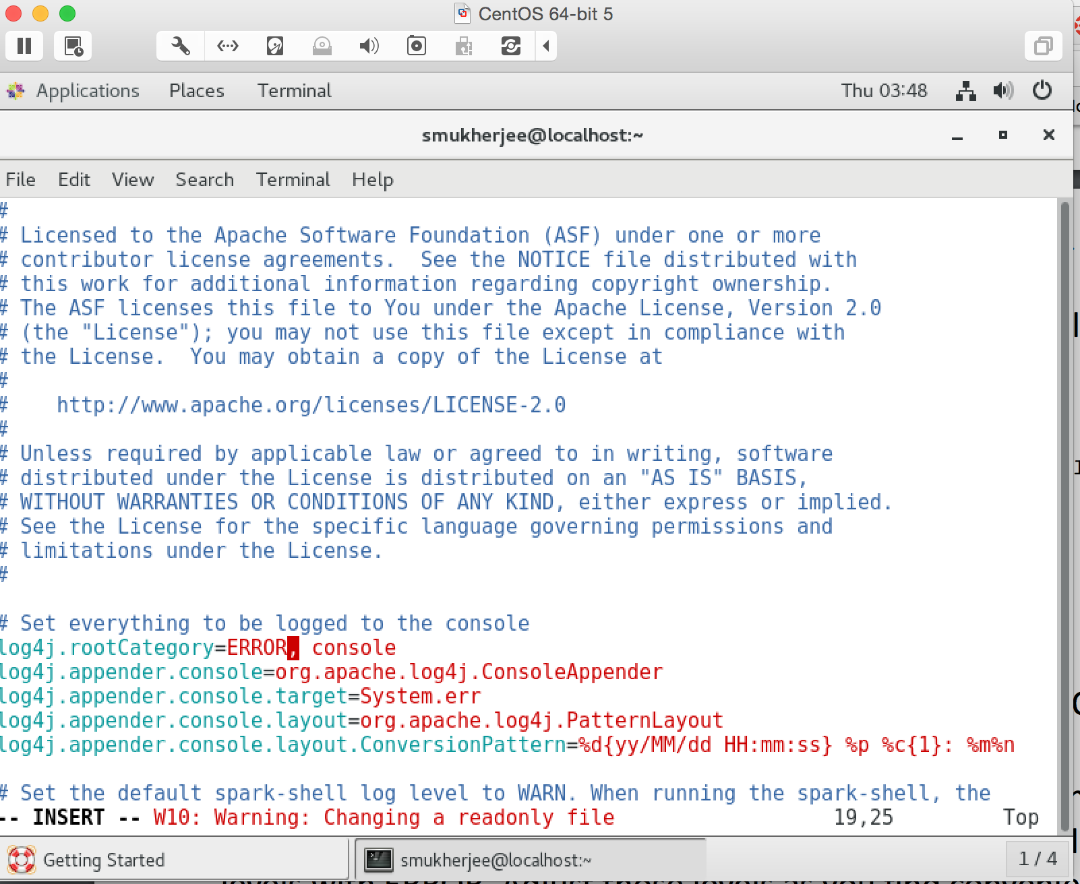
Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0\_144)

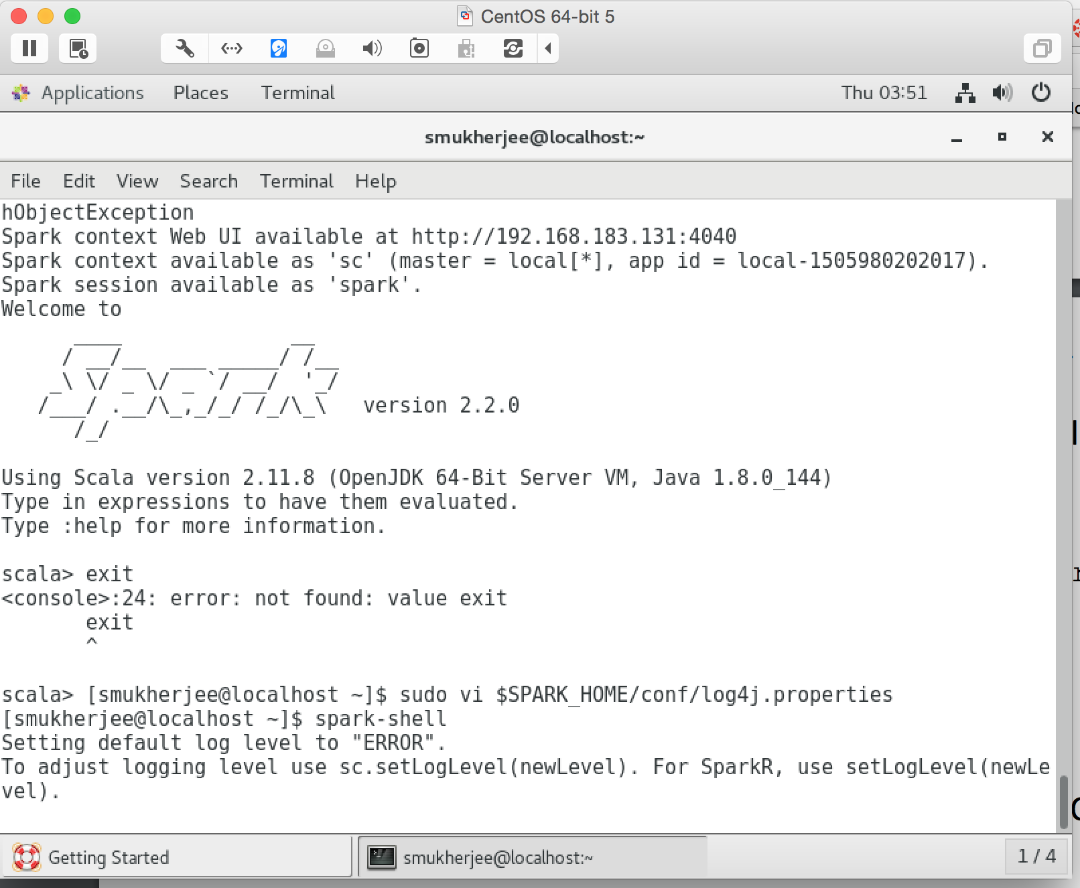
Type in expressions to have them evaluated.

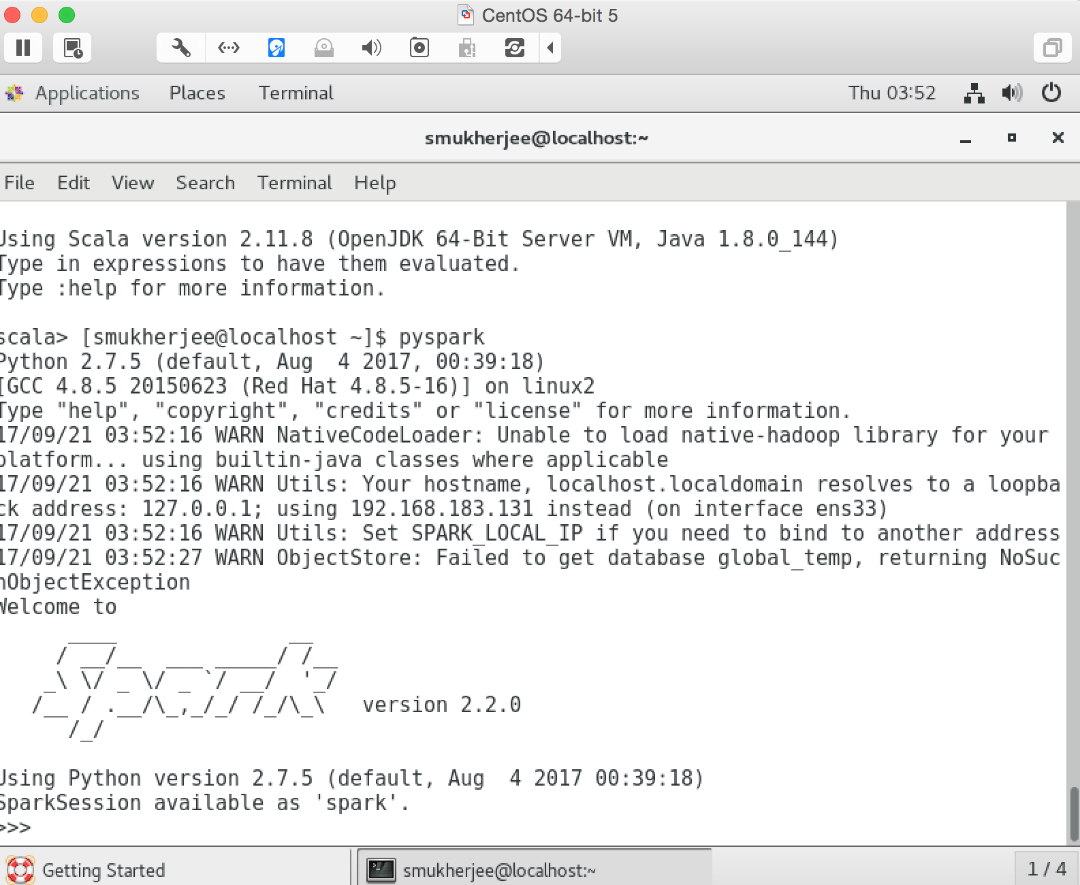
Type :help for more information.

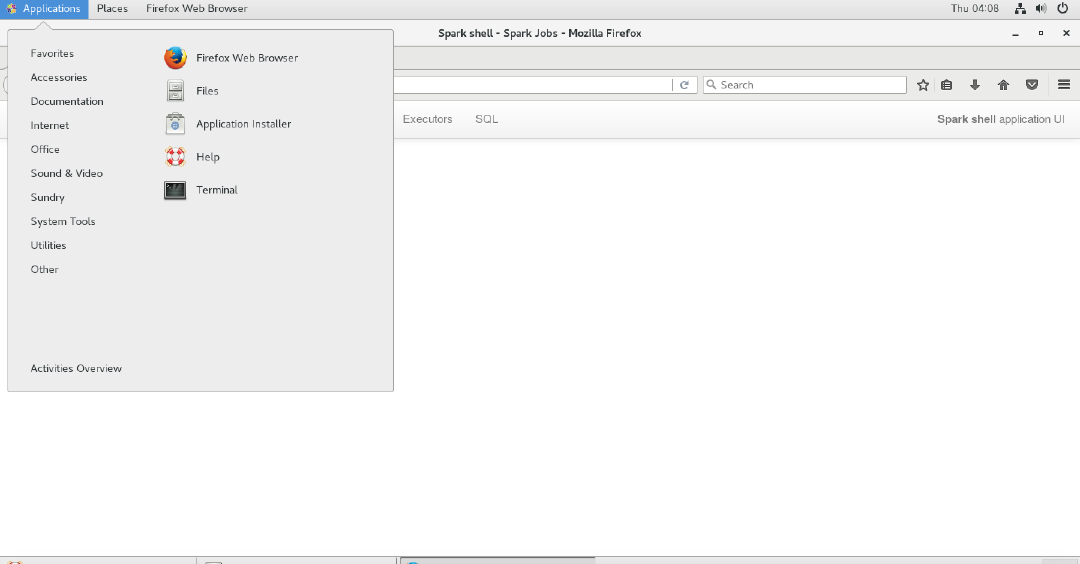
scala>











**Problem 3**. Find the number of lines in the text file ulysses10.txt that contain word “afternoon” or “night” or “morning”. In this problem use RDD API. Do this in two ways, first create a lambda function which will test whether a line contains any one of those 3 words. Second, create a named function in the language of choice that returns TRUE if a line passed to it contains any one of those three words. Demonstrate that the count is the same. Use pyspark and Spark Python API. If convenient you are welcome to implement this problem in any other language: Scala, Java or R. [15%]

I use scala to do this. Using sparkContext sc I use textFile to get the file as input and create the rdd. Next I filter using the lambda function. I use lowercase so I can cover both uppercase and lowercase.

val input = sc.textFile("ulysses10.txt");

scala> input.filter(line => line.toLowerCase.contains("afternoon")||line.toLowerCase.contains("morning")|| line.toLowerCase.contains("night")).count

**res9: Long = 462**

Next I create a named function in scala. Then I use it for filtering the rdd.

#Function

scala> def f(line:String) = {line.toLowerCase.contains("afternoon")|| line.toLowerCase.contains("night")|| line.toLowerCase.contains("morning") }

input.filter(line => f(line)).count

**res3: Long = 462**

**Problem 4.** Implement the above task, finding the number of lines with one of those three words in file ulysses10.txt using Dataset/DataFrame API. Again, use the language of your choice. [20%]

I now create a dataset using SparkSession (spark). Again I read the input in. Next I register it as temp table. And run a sql query on it to filter. The count matches to the one in Prob #3

val df = spark.sqlContext.read.text("ulysses10.txt").as[String]

scala> df.show

+--------------------+

| value|

+--------------------+

|The Project Guten...|

| by James Joyce|

|(#4 in our series...|

| |

|Copyright laws ar...|

|copyright laws fo...|

|Project Gutenberg...|

scala> df.registerTempTable("df")

scala> sql("select \* from df where lower(value) like '%afternoon%' or lower(value) like '%night%' or lower(value) like '%morning%'").count

**res1: Long = 462**

**Problem 5**. Create a standalone Python script that will count all words in file ulysses10.txt. You are expected to produce a single number. Do it using RDD API. If convenient, you are welcome to implement this problem in other languages: Scala, Java or R. [%15]

I use scala for this.Again I create the rdd from the file. I do a flatmap to do a flattened list by splitting based on space (“ “) as delimiter. I then get a count of elements this flattened Array to get the word count.

**val input = sc.textFile(“ulysses10.txt")**

scala> **input.flatMap(line => line.split(" ")).collect**

res12: Array[String] = Array(The, Project, Gutenberg, Etext, of, Ulysses, by, James, Joyce, (#4, in, our, series, by, James, Joyce), "", Copyright, laws, are, changing, all, over, the, world., Be, sure, to, check, the, copyright, laws, for, your, country, before, distributing, this, or, any, other, Project, Gutenberg, file., "", We, encourage, you, to, keep, this, file,, exactly, as, it, is,, on, your, own, disk,, thereby, keeping, an, electronic, path, open, for, future, readers., "", Please, do, not, remove, this., "", This, header, should, be, the, first, thing, seen, when, anyone, starts, to, view, the, etext., Do, not, change, or, edit, it, without, written, permission., The, words, are, carefully, chosen, to, provide, users, with, the, information, they, need, to, understand, what...

scala> input.flatMap(line => line.split(" ")).count

**res13: Long = 278555**

**Problem 6.** Create a standalone Python script that will count all words in file ulysses10.txt. You are expected to produce a single number. Do it using Dataset/DataFrame API. If convenient, you are welcome to implement this problem in other languages: Scala, Java or R. [%15]

Below I have create a spark jar. I have also shown the contents of my build.sbt. I package the jar, scp it to VM and then do a spark-submit. I create a dataset from the input which I do a flatmap, splitting with space (“ “) as delimiter. I then get a count. This matches #5

**build.sbt**

*name* := "sparkHW"

*version* := "1.0"

*scalaVersion* := "2.11.8"

*libraryDependencies* += "org.apache.spark" %% "spark-sql" % "2.2.0"

\*\*\*\*

**WordCount.scala**

**package** com.spark.hw.practice

**import** org.apache.spark.sql

**import** org.apache.spark.sql.SparkSession

*/\*\**

*\* Created by smukherjee on 9/22/17.*

*\*/*

**object** WordCount {

**def** main(args: Array[String]): Unit = {

**val** spark = SparkSession.*builder*().getOrCreate()

**val** ds = spark.read.text("ulysses10.txt").as[String].flatMap(x => x.split(" "))

*println*(ds.count)

spark.close()

}

}

**Packaging:**

**sbt package**

**info] Packaging /Users/smukherjee/vm/SparkHW/target/scala-2.11/sparkhw\_2.11-1.0.jar ...**

**scp /Users/smukherjee/vm/SparkHW/target/scala-2.11/sparkhw\_2.11-1.0.jar smukherjee@192.168.190.128:/home/smukherjee/**

**[smukherjee@localhost ~]$ spark-submit --class "com.spark.hw.practice.WordCount" sparkhw\_2.11-1.0.jar**

**17/09/23 02:18:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable**

**17/09/23 02:18:45 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.190.128 instead (on interface ens33)**

**17/09/23 02:18:45 WARN Utils: Set SPARK\_LOCAL\_IP if you need to bind to another address**

**278555**

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document with explanations of the purpose of those commands. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE COPY Snippets of your Code but DO NOT EMBED files into your MS Word document. For issues and comments visit the class Discussion Board on Piazza.