# Big Data Hadoop Training

Session 13 Assignment 2 Solution

**1. Create a simple pairRDD of (1, 2), (3, 4), (3, 6).**

1. Run all Spark Daemons and start “spark scala” shell :

For creating a pairRDD:

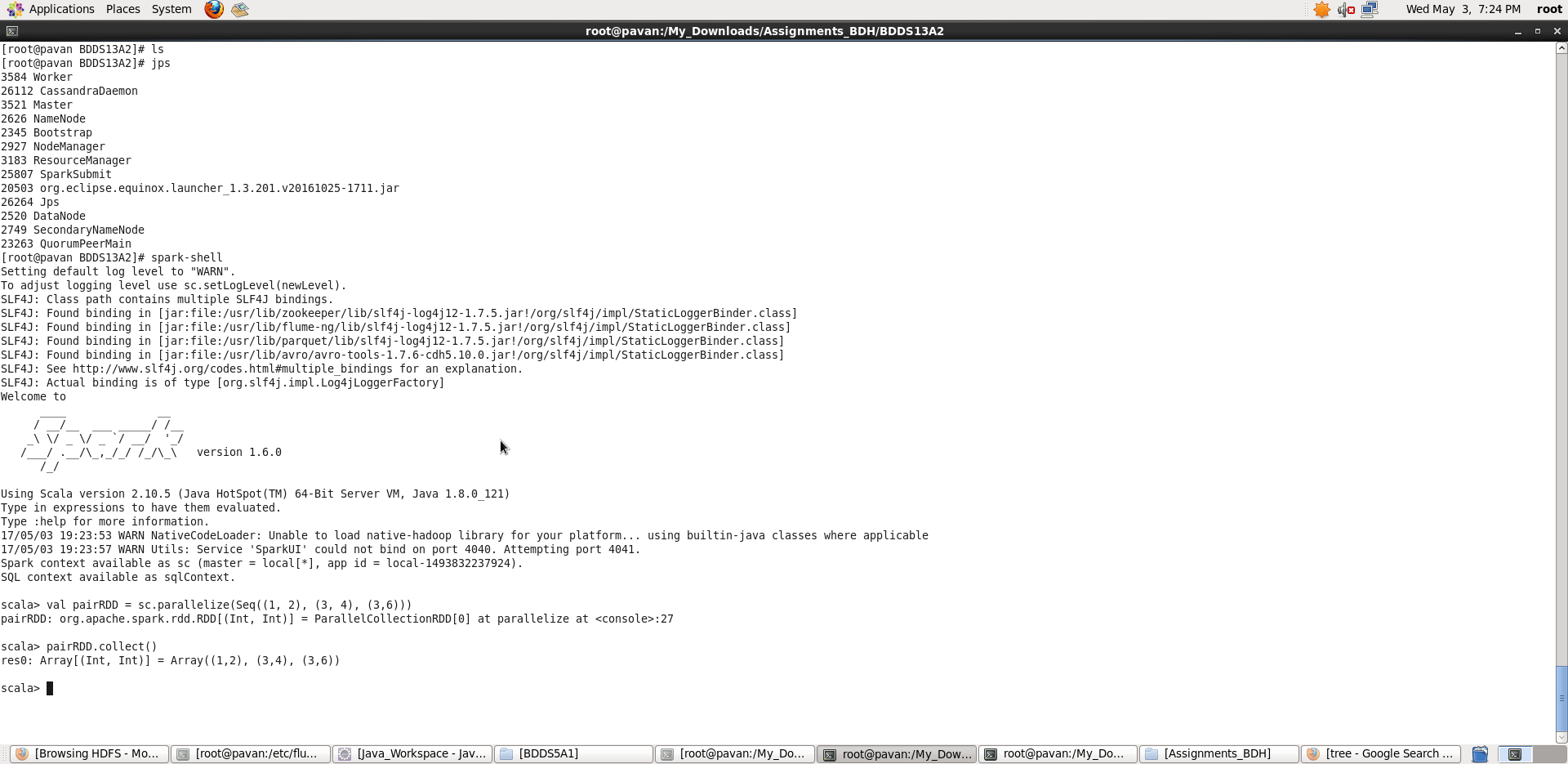
**Scala> val pairRDD = sc.parallelize(Seq((1,2),(3,4),(3,6)))**

Here sc is SparkContext.

We can see in the screenshot, an pairRDD being created of type RDD[(Int, Int)]

**Scala> pairRDD.collect()**

**Collect() –** returns all the elements of the dataset as an array at the driver program.

****

**2. Transform an RDD of ("a","b","c","d","e") to PairRDD (a,0), (b,1), (c,2), (d,3), (e,4)**

1. Run all Spark Daemons and start “scala” terminal :

First, create a simple RDD and then transform it into given pairRDD:

**Scala> val a = sc.parallelize(List(“a”,”b”,”c”,”d”,”e”))**

**Scala> val pairRDD = a.zipWithIndex.map { case (element, index) => (element,index)}**

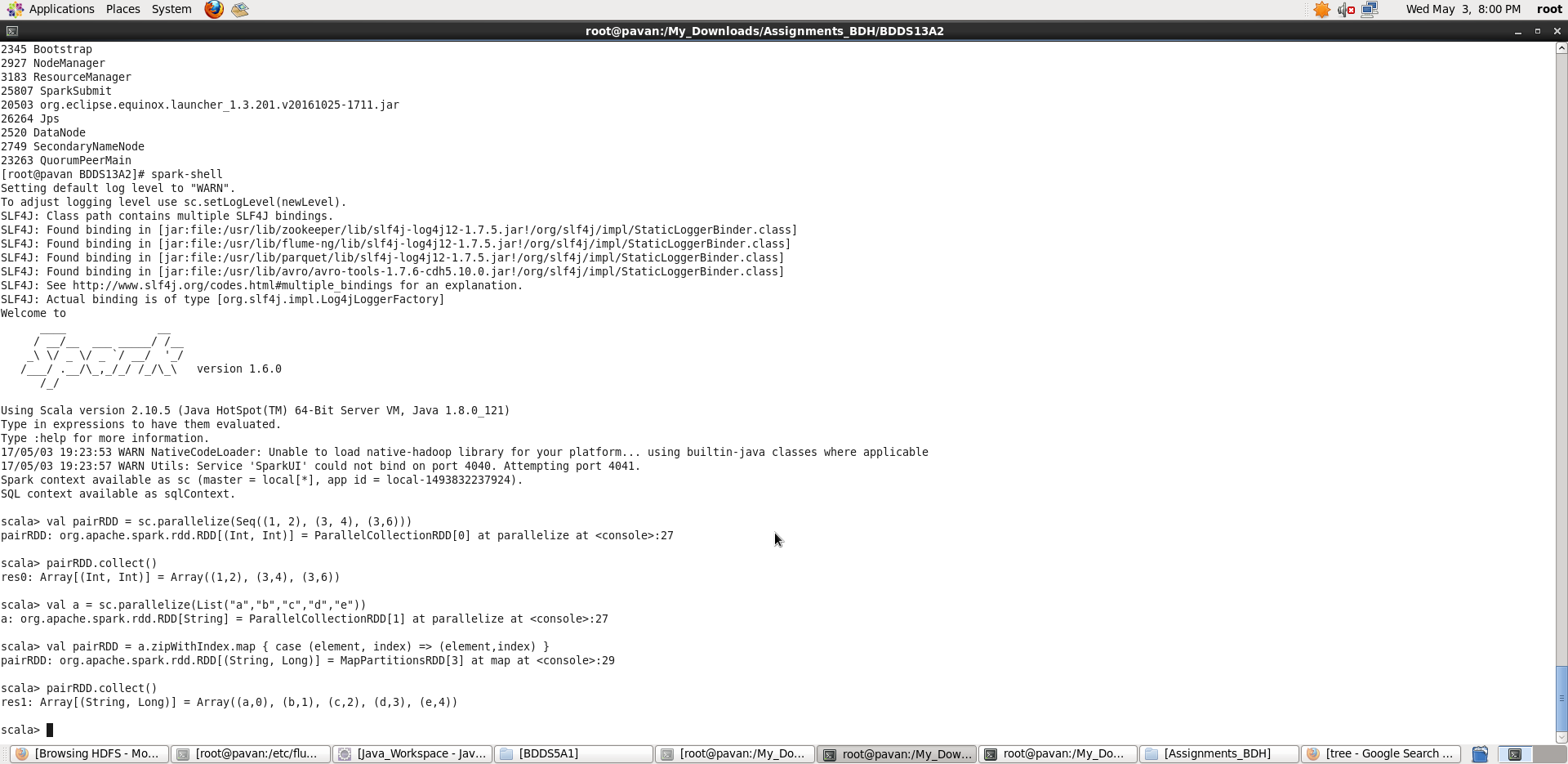
**Scala> pairRDD.collect()**

Here sc is SparkContext.

We can see in the screenshot, a pairRDD of specified value is being created. Here, I have used the index of each element inside the List ‘a’ to be the second element in the pairRDD.

**Map(func) –** returns a new distributed dataset formed by passing each element of the source through a function **func.**

**Collect() –** returns all the elements of the dataset as an array at the driver program

****

Thus, in Scala prompt, we have performed the required tasks.