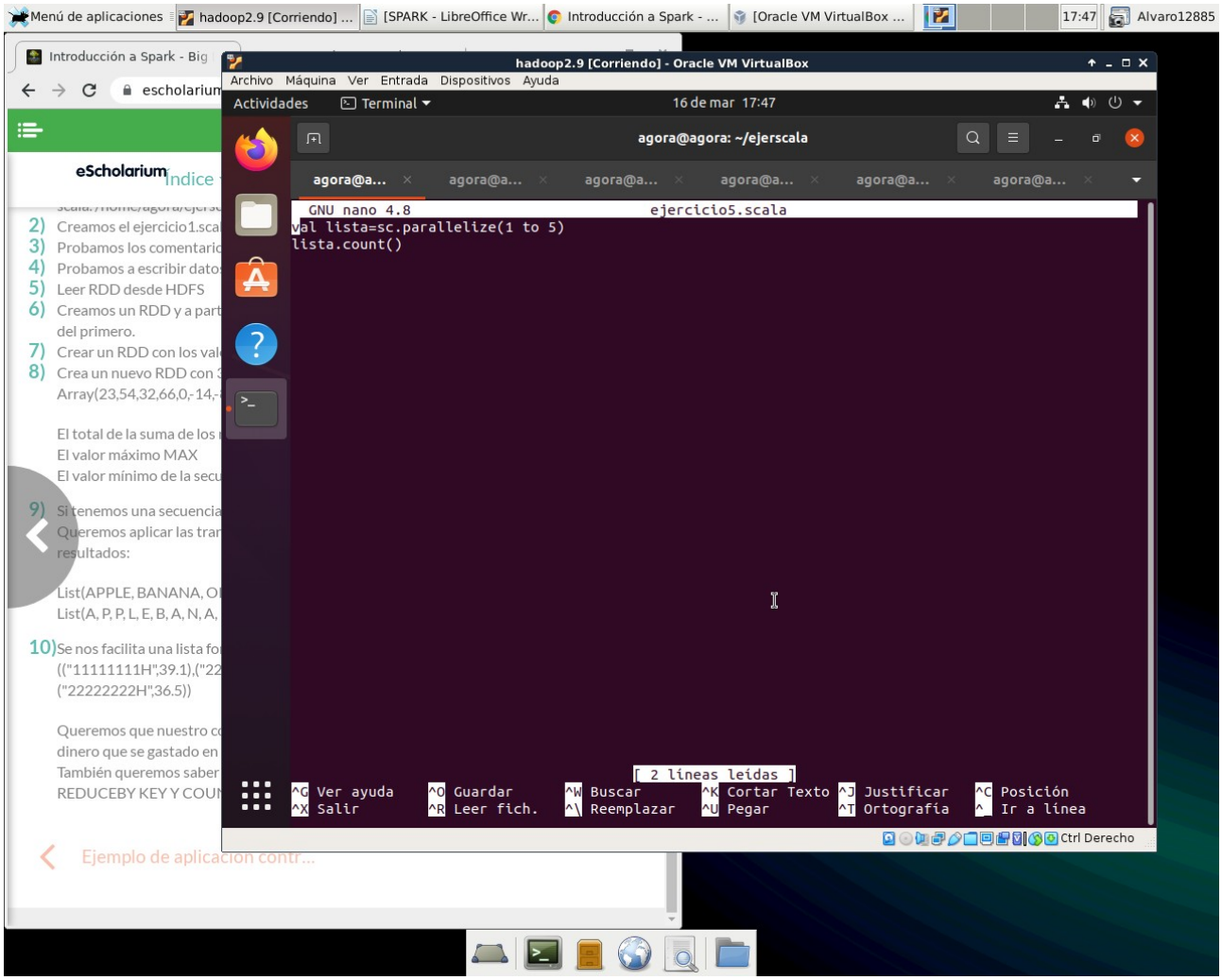
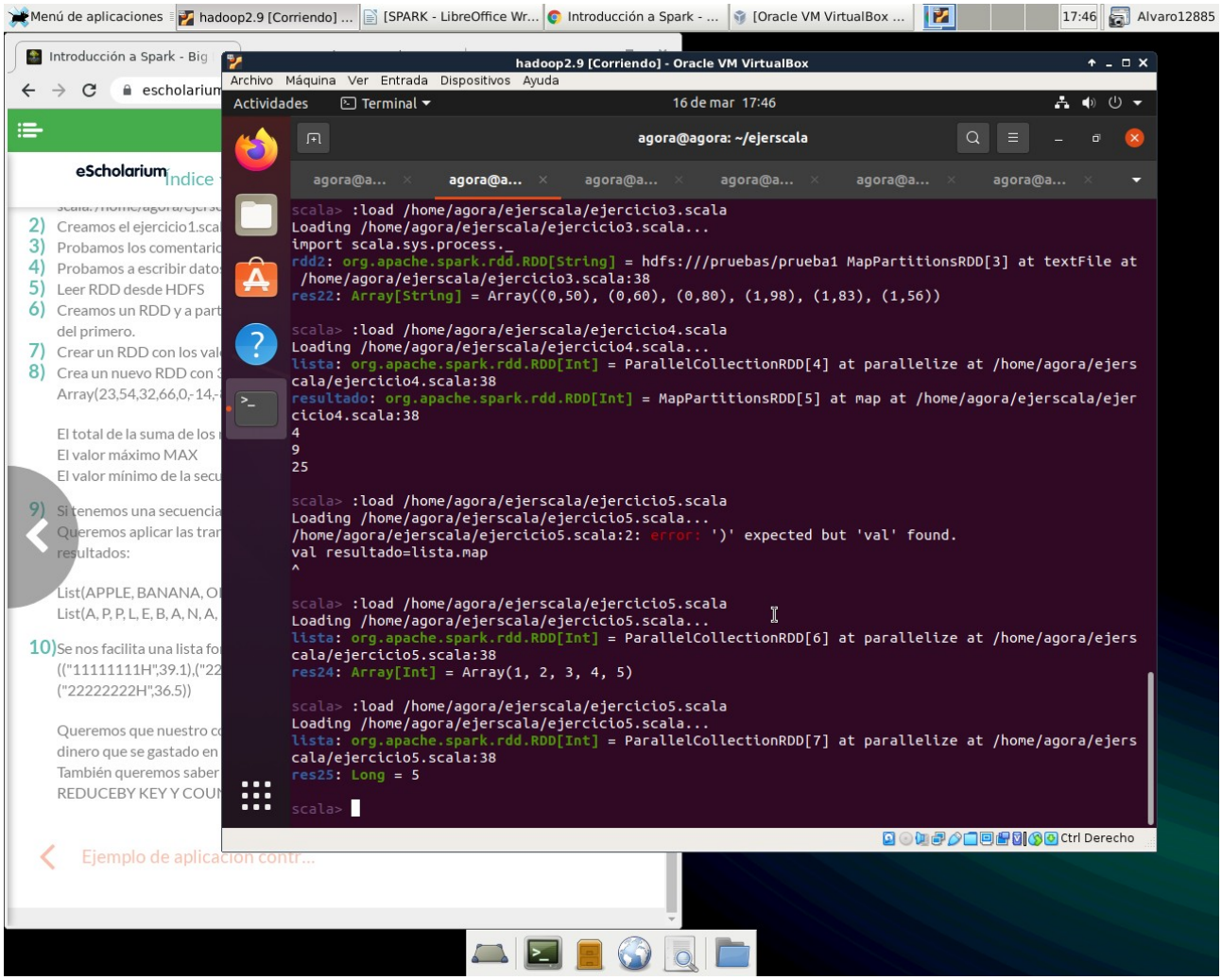
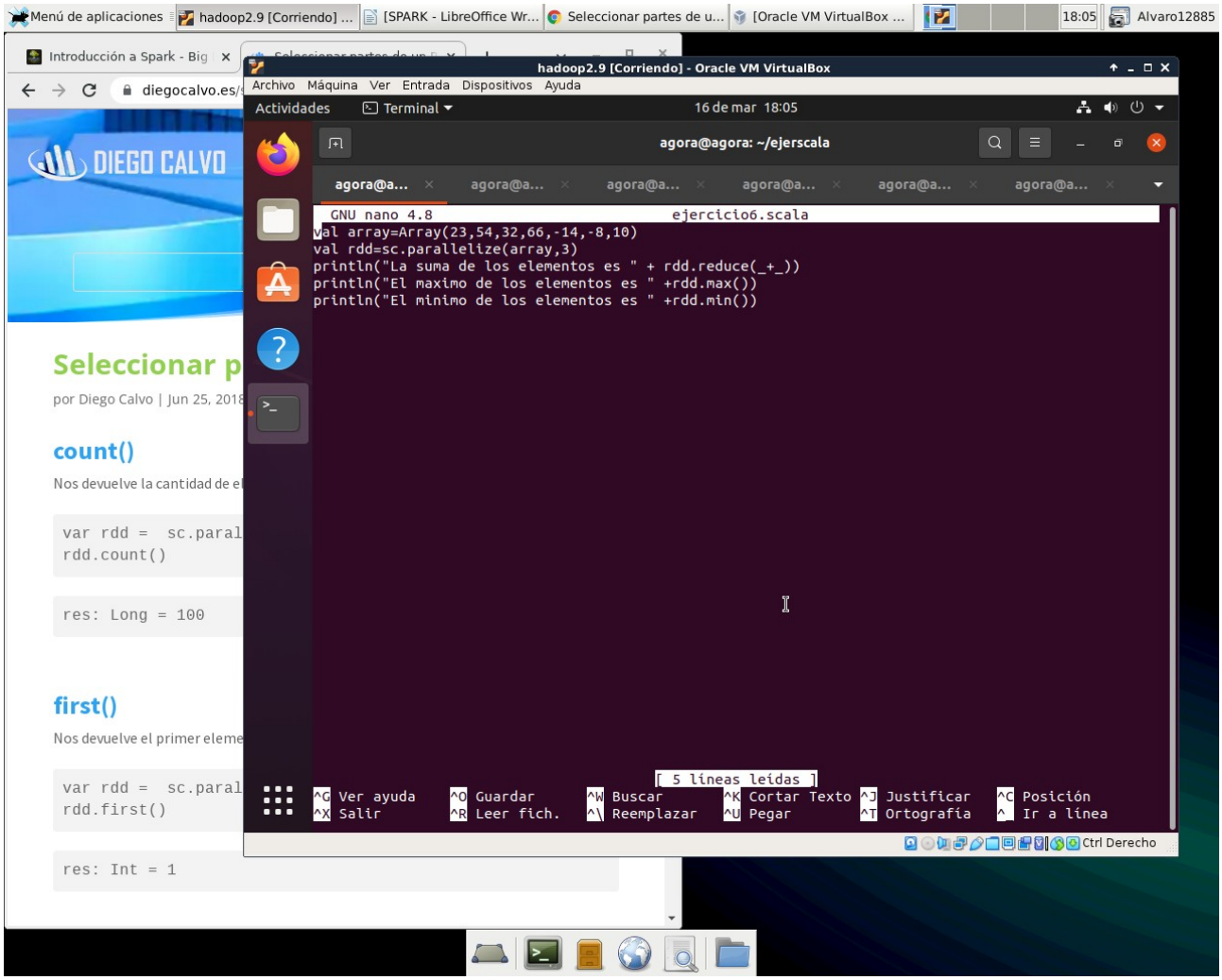
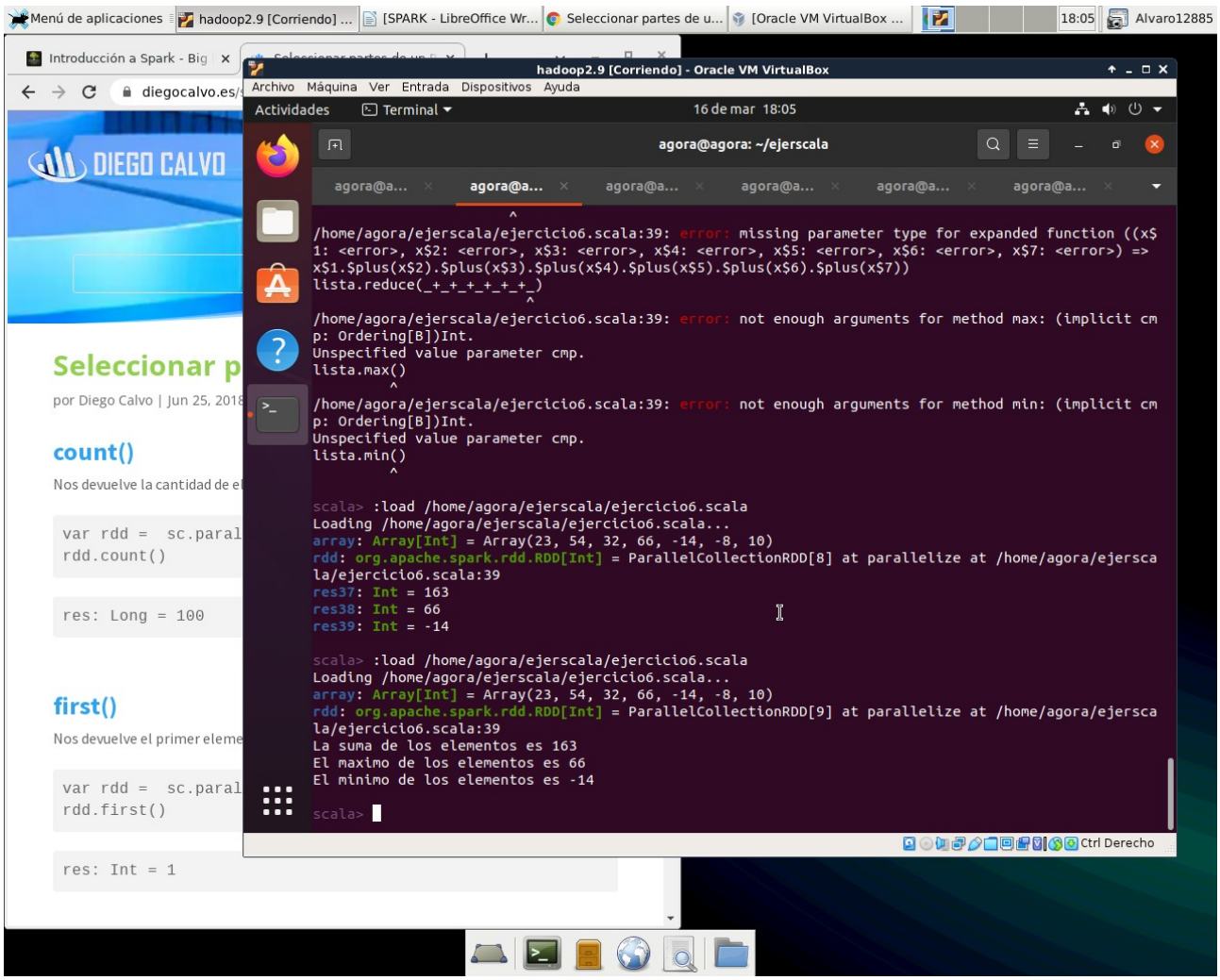
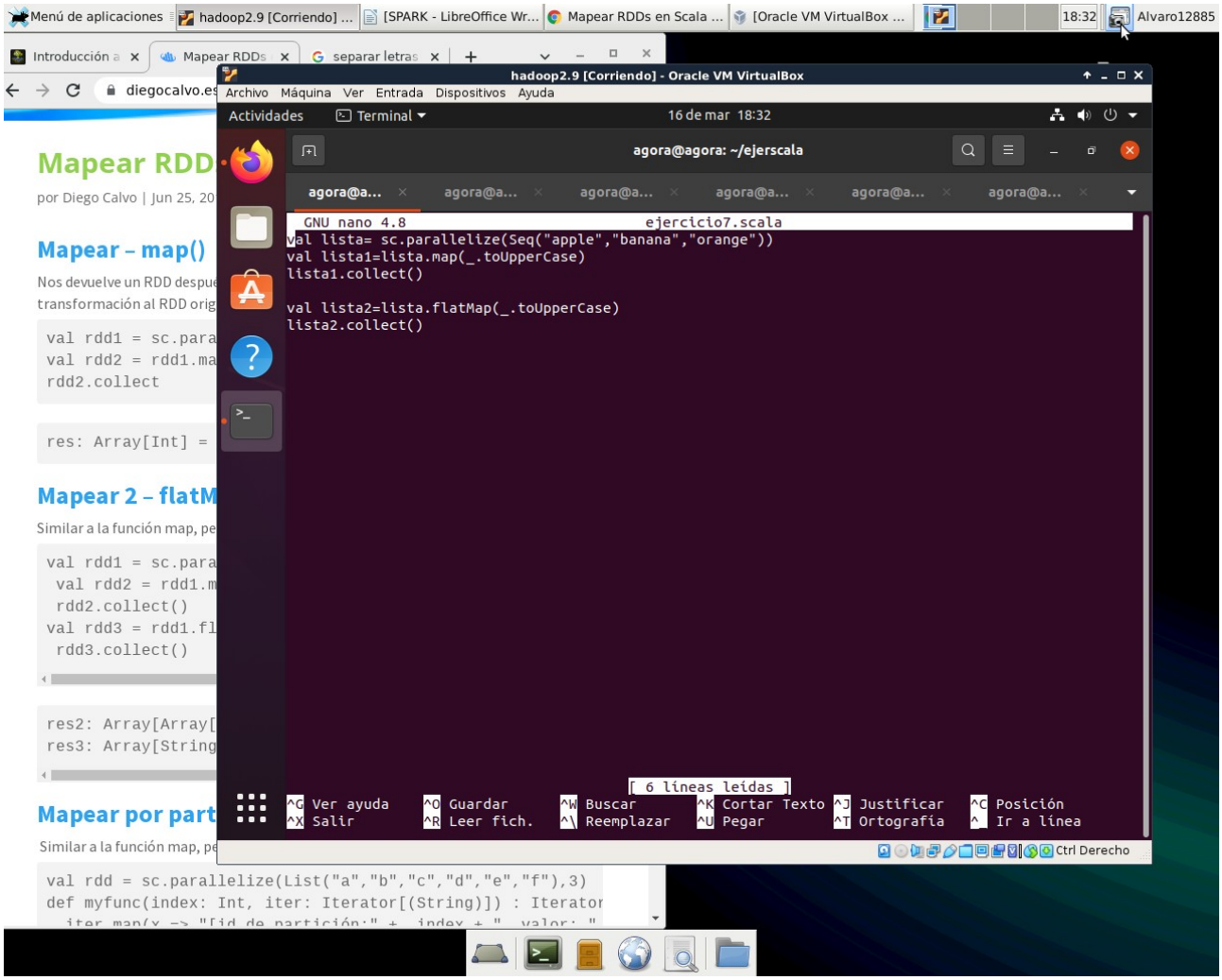
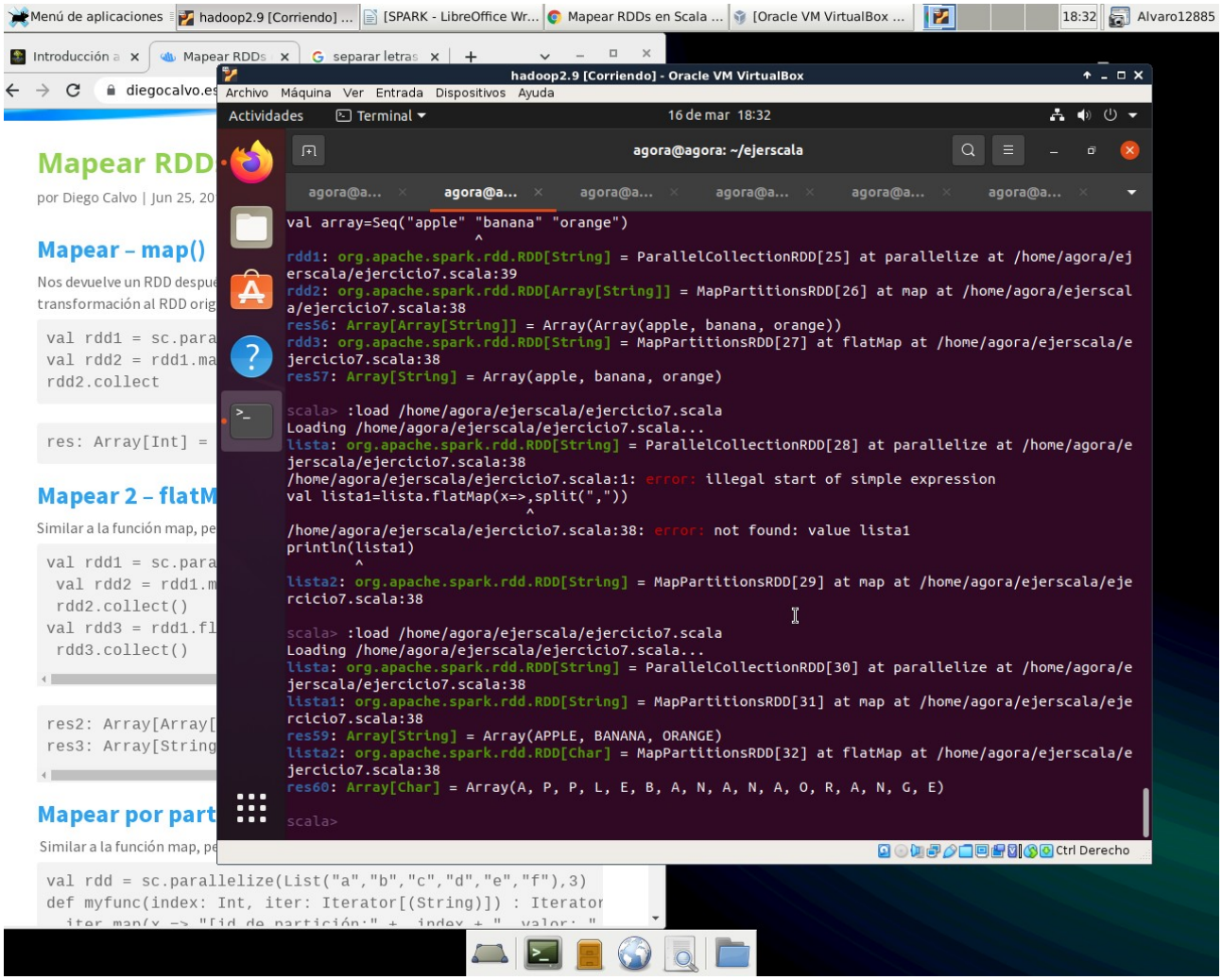

Menú de aplicaciones
hadoop2.9 [Corriendo] ...
[SPARK - LibreOffice Wr...
Introducción a Spark - ...
[Oracle VM VirtualBox ...
17:47
Alvaro12885
Introducción a Spark - Big
escholarium
eScholarium
Indice
2) Creamos el ejercicio1.scal
3) Probamos los comentario
4) Probamos a escribir dato
5) Leer RDD desde HDFS
6) Creamos un RDD y a part
del primero.
7) Crear un RDD con los val
8) Crea un nuevo RDD con
Array(23,54,32,66,0,-14,-
El total de la suma de los
El valor máximo MAX
El valor mínimo de la secu
9) Si tenemos una secuencia
Queremos aplicar las tra
resultados:
List(APPLE, BANANA, O
List(A, P, P, L, E, B, A, N, A,
10) Se nos facilita una lista fo
(("11111111H",39.1),("22
("22222222H",36.5))
Queremos que nuestro c
dinero que se gastado en
También queremos saber
REDUCEBY KEY Y COU
Ejemplo de aplicación contr...
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 17:47
agora@agora: ~/ejerscala
agora@a...
GNU nano 4.8
ejercicio5.scala
val lista=sc.parallelize(1 to 5)
lista.count()
[2 líneas leídas]
^G Ver ayuda
^O Guardar
^W Buscar
^K Cortar Texto
^J Justificar
^C Posición
^X Salir
^R Leer fich.
^\ Reemplazar
^U Pegar
^T Ortografía
^_ Ir a línea
Ctrl Derecho

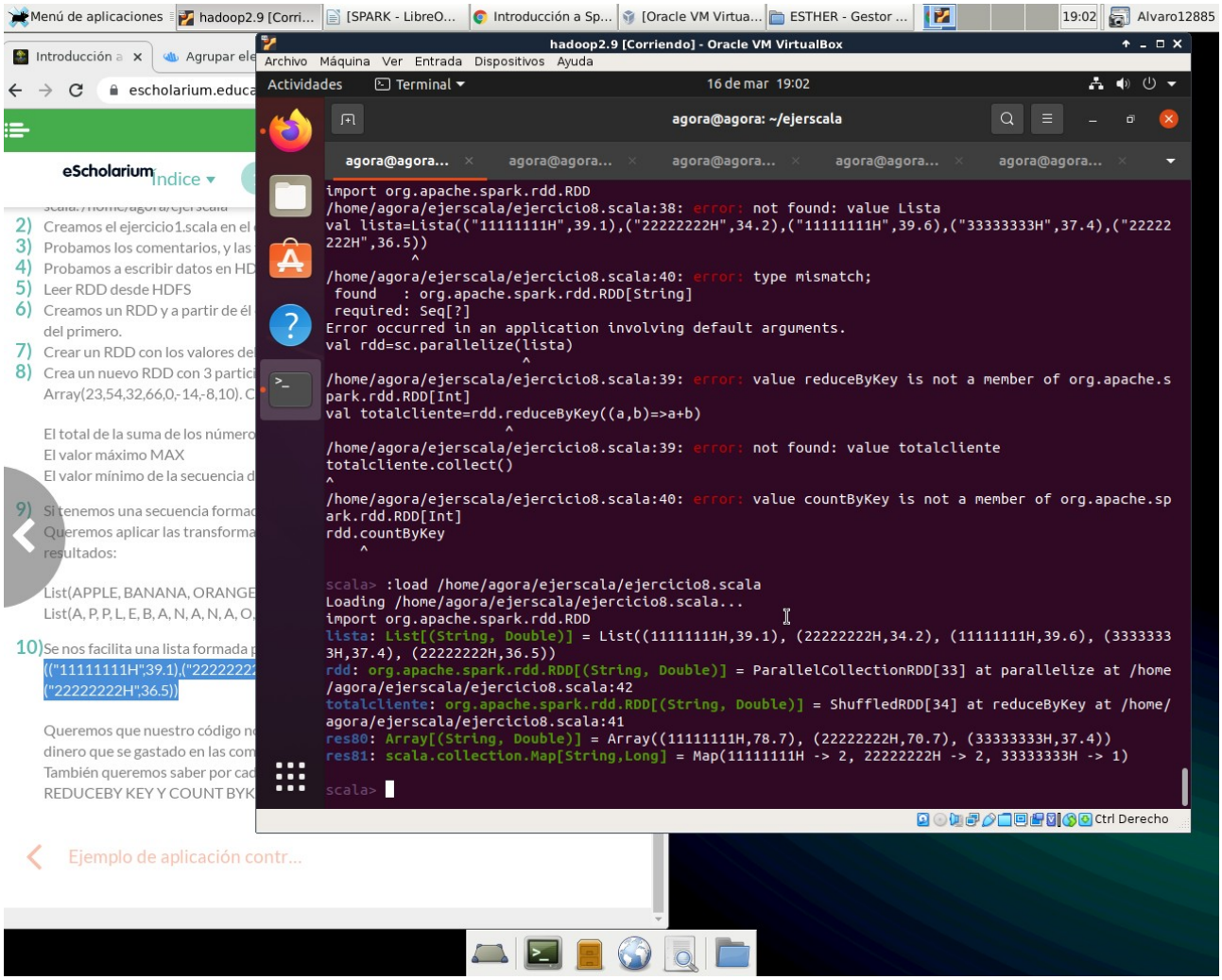

Menú de aplicaciones
hadoop2.9 [Corriendo] ...
[SPARK - LibreOffice Wr...
Introducción a Spark - ...
[Oracle VM VirtualBox ...
17:46
Alvaro12885
Introducción a Spark - Big
escholarium
eScholarium
Indice
2) Creamos el ejercicio1.scal
3) Probamos los comentario
4) Probamos a escribir dato
5) Leer RDD desde HDFS
6) Creamos un RDD y a part
del primero.
7) Crear un RDD con los val
8) Crea un nuevo RDD con
Array(23,54,32,66,0,-14,-
El total de la suma de los
El valor máximo MAX
El valor mínimo de la secu
9) Si tenemos una secuencia
Queremos aplicar las tra
resultados:
List(APPLE, BANANA, O
List(A, P, P, L, E, B, A, N, A,
10) Se nos facilita una lista fo
(("11111111H",39.1),("22
("22222222H",36.5))
Queremos que nuestro c
dinero que se gastado en
También queremos saber
REDUCEBY KEY Y COU
Ejemplo de aplicación contr...
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 17:46
agora@agora: ~/ejerscala
agora@a...
scala> :load /home/agora/ejerscala/ejercicio3.scala
Loading /home/agora/ejerscala/ejercicio3.scala...
import scala.sys.process._
rdd2: org.apache.spark.rdd.RDD[String] = hdfs:///pruebas/prueba1 MapPartitionsRDD[3] at textFile at
/home/agora/ejerscala/ejercicio3.scala:38
res22: Array[String] = Array((0,50), (0,60), (0,80), (1,98), (1,83), (1,56))
scala> :load /home/agora/ejerscala/ejercicio4.scala
Loading /home/agora/ejerscala/ejercicio4.scala...
lista: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[4] at parallelize at /home/agora/ejers
cala/ejercicio4.scala:38
resultado: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[5] at map at /home/agora/ejerscala/ejer
cicio4.scala:38
4
9
25
scala> :load /home/agora/ejerscala/ejercicio5.scala
Loading /home/agora/ejerscala/ejercicio5.scala...
/home/agora/ejerscala/ejercicio5.scala:2: error: ')' expected but 'val' found.
val resultado=lista.map
^
scala> :load /home/agora/ejerscala/ejercicio5.scala
Loading /home/agora/ejerscala/ejercicio5.scala...
lista: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[6] at parallelize at /home/agora/ejers
cala/ejercicio5.scala:38
res24: Array[Int] = Array(1, 2, 3, 4, 5)
scala> :load /home/agora/ejerscala/ejercicio5.scala
Loading /home/agora/ejerscala/ejercicio5.scala...
lista: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[7] at parallelize at /home/agora/ejers
cala/ejercicio5.scala:38
res25: Long = 5
scala>
Ctrl Derecho


Menú de aplicaciones
hadoop2.9 [Corriendo] ...
[SPARK - LibreOffice Wr...
Seleccionar partes de u...
[Oracle VM VirtualBox ...
18:05
Alvaro12885
Introducción a Spark - Big
diegocalvo.es/
DIEGO CALVO
Seleccionar p
por Diego Calvo | Jun 25, 2018
count()
Nos devuelve la cantidad de el
var rdd = sc.paral
rdd.count()
res: Long = 100
first()
Nos devuelve el primer eleme
var rdd = sc.paral
rdd.first()
res: Int = 1
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 18:05
agora@agora: ~/ejerscala
agora@a...
GNU nano 4.8
ejercicio6.scala
val array=Array(23,54,32,66,-14,-8,10)
val rdd=sc.parallelize(array,3)
println("La suma de los elementos es " + rdd.reduce(_+_))
println("El maximo de los elementos es " +rdd.max())
println("El minimo de los elementos es " +rdd.min())
[5 líneas leídas]
^G Ver ayuda
^O Guardar
^W Buscar
^K Cortar Texto
^J Justificar
^C Posición
^X Salir
^R Leer fich.
^\ Reemplazar
^U Pegar
^T Ortografía
^_ Ir a línea
Ctrl Derecho


Menú de aplicaciones
hadoop2.9 [Corriendo] ...
[SPARK - LibreOffice Wr...
Seleccionar partes de u...
[Oracle VM VirtualBox ...
18:05
Alvaro12885
Introducción a Spark - Big
diegocalvo.es/
DIEGO CALVO
Seleccionar p
por Diego Calvo | Jun 25, 2018
count()
Nos devuelve la cantidad de el
var rdd = sc.paral
rdd.count()
res: Long = 100
first()
Nos devuelve el primer eleme
var rdd = sc.paral
rdd.first()
res: Int = 1
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 18:05
agora@agora: ~/ejerscala
agora@a...
/home/agora/ejerscala/ejercicio6.scala:39: error: missing parameter type for expanded function ((x$1: <error>, x$2: <error>, x$3: <error>, x$4: <error>, x$5: <error>, x$6: <error>, x$7: <error>) => x$1.$plus(x$2).$plus(x$3).$plus(x$4).$plus(x$5).$plus(x$6).$plus(x$7))
lista.reduce(_+_+_+_+_+_+_)
^
/home/agora/ejerscala/ejercicio6.scala:39: error: not enough arguments for method max: (implicit cmp: Ordering[B])Int.
Unspecified value parameter cmp.
lista.max()
^
/home/agora/ejerscala/ejercicio6.scala:39: error: not enough arguments for method min: (implicit cmp: Ordering[B])Int.
Unspecified value parameter cmp.
lista.min()
^
scala> :load /home/agora/ejerscala/ejercicio6.scala
Loading /home/agora/ejerscala/ejercicio6.scala...
array: Array[Int] = Array(23, 54, 32, 66, -14, -8, 10)
rdd: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[8] at parallelize at /home/agora/ejerscala/ejercicio6.scala:39
res37: Int = 163
res38: Int = 66
res39: Int = -14
scala> :load /home/agora/ejerscala/ejercicio6.scala
Loading /home/agora/ejerscala/ejercicio6.scala...
array: Array[Int] = Array(23, 54, 32, 66, -14, -8, 10)
rdd: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[9] at parallelize at /home/agora/ejerscala/ejercicio6.scala:39
La suma de los elementos es 163
El maximo de los elementos es 66
El minimo de los elementos es -14
scala>
Ctrl Derecho


Menú de aplicaciones
hadoop2.9 [Corriendo] ...
[SPARK - LibreOffice Wr...
Mapear RDDs en Scala ...
[Oracle VM VirtualBox ...
18:32
Alvaro12885
Introducción a
Mapear RDDs
separar letras
diegocalvo.es
Mapear RDD
por Diego Calvo | Jun 25, 20
Mapear – map()
Nos devuelve un RDD despue
transformación al RDD orig
val rdd1 = sc.para
val rdd2 = rdd1.ma
rdd2.collect
res: Array[Int] =
Mapear 2 – flatM
Similar a la función map, pe
val rdd1 = sc.para
val rdd2 = rdd1.m
rdd2.collect()
val rdd3 = rdd1.fl
rdd3.collect()
res2: Array[Array[
res3: Array[String
Mapear por part
Similar a la función map, pe
val rdd = sc.parallelize(List("a","b","c","d","e","f"),3)
def myfunc(index: Int, iter: Iterator[(String)]) : Iterator
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 18:32
agora@agora: ~/ejerscala
agora@a...
GNU nano 4.8
ejercicio7.scala
val lista= sc.parallelize(Seq("apple","banana","orange"))
val lista1=lista.map(_.toUpperCase)
lista1.collect()
val lista2=lista.flatMap(_.toUpperCase)
lista2.collect()
[6 líneas leídas]
^G Ver ayuda
^X Salir
^O Guardar
^R Leer fich.
^W Buscar
^\ Reemplazar
^K Cortar Texto
^U Pegar
^J Justificar
^T Ortografía
^C Posición
^_ Ir a línea
Ctrl Derecho


Menú de aplicaciones
hadoop2.9 [Corriendo] ...
[SPARK - LibreOffice Wr...
Mapear RDDs en Scala ...
[Oracle VM VirtualBox ...
18:32
Alvaro12885
Introducción a
Mapear RDDs
separar letras
diegocalvo.es
Mapear RDD
por Diego Calvo | Jun 25, 20
Mapear – map()
Nos devuelve un RDD despue
transformación al RDD orig
val rdd1 = sc.para
val rdd2 = rdd1.ma
rdd2.collect
res: Array[Int] =
Mapear 2 – flatM
Similar a la función map, pe
val rdd1 = sc.para
val rdd2 = rdd1.m
rdd2.collect()
val rdd3 = rdd1.fl
rdd3.collect()
res2: Array[Array[
res3: Array[String
Mapear por part
Similar a la función map, pe
val rdd = sc.parallelize(List("a","b","c","d","e","f"),3)
def myfunc(index: Int, iter: Iterator[(String)]) : Iterator
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 18:32
agora@agora: ~/ejerscala
agora@a...
agora@a...
agora@a...
agora@a...
agora@a...
agora@a...
val array=Seq("apple" "banana" "orange")
^
rdd1: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[25] at parallelize at /home/agora/ejerscala/ejercicio7.scala:39
rdd2: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[26] at map at /home/agora/ejerscala/ejercicio7.scala:38
res56: Array[Array[String]] = Array(Array(apple, banana, orange))
rdd3: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[27] at flatMap at /home/agora/ejerscala/ejercicio7.scala:38
res57: Array[String] = Array(apple, banana, orange)
scala> :load /home/agora/ejerscala/ejercicio7.scala
Loading /home/agora/ejerscala/ejercicio7.scala...
lista: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[28] at parallelize at /home/agora/ejerscala/ejercicio7.scala:38
/home/agora/ejerscala/ejercicio7.scala:1: error: illegal start of simple expression
val lista1=lista.flatMap(x=>,split(","))
^
/home/agora/ejerscala/ejercicio7.scala:38: error: not found: value lista1
println(lista1)
^
lista2: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[29] at map at /home/agora/ejerscala/ejercicio7.scala:38
scala> :load /home/agora/ejerscala/ejercicio7.scala
Loading /home/agora/ejerscala/ejercicio7.scala...
lista: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[30] at parallelize at /home/agora/ejerscala/ejercicio7.scala:38
lista1: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[31] at map at /home/agora/ejerscala/ejercicio7.scala:38
res59: Array[String] = Array(APPLE, BANANA, ORANGE)
lista2: org.apache.spark.rdd.RDD[Char] = MapPartitionsRDD[32] at flatMap at /home/agora/ejerscala/ejercicio7.scala:38
res60: Array[Char] = Array(A, P, P, L, E, B, A, N, A, N, A, O, R, A, N, G, E)
scala>
Ctrl Derecho


Menú de aplicaciones
hadoop2.9 [Corri...
[SPARK - LibreO...
Introducción a Sp...
[Oracle VM Virtua...
ESTHER - Gestor ...
19:02
Alvaro12885
hadoop2.9 [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Actividades
Terminal
16 de mar 19:02
agora@agora: ~/ejerscala
import org.apache.spark.rdd.RDD
/home/agora/ejerscala/ejercicio8.scala:38: error: not found: value Lista
val lista=Lista(("11111111H",39.1),("22222222H",34.2),("11111111H",39.6),("33333333H",37.4),("22222222H",36.5))
^
/home/agora/ejerscala/ejercicio8.scala:40: error: type mismatch;
found : org.apache.spark.rdd.RDD[String]
required: Seq[?]
Error occurred in an application involving default arguments.
val rdd=sc.parallelize(lista)
^
/home/agora/ejerscala/ejercicio8.scala:39: error: value reduceByKey is not a member of org.apache.spark.rdd.RDD[Int]
val totalcliente=rdd.reduceByKey((a,b)=>a+b)
^
/home/agora/ejerscala/ejercicio8.scala:39: error: not found: value totalcliente
totalcliente.collect()
^
/home/agora/ejerscala/ejercicio8.scala:40: error: value countByKey is not a member of org.apache.spark.rdd.RDD[Int]
rdd.countByKey
^
scala> :load /home/agora/ejerscala/ejercicio8.scala
Loading /home/agora/ejerscala/ejercicio8.scala...
import org.apache.spark.rdd.RDD
lista: List[(String, Double)] = List((11111111H,39.1), (22222222H,34.2), (11111111H,39.6), (33333333H,37.4), (22222222H,36.5))
rdd: org.apache.spark.rdd.RDD[(String, Double)] = ParallelCollectionRDD[33] at parallelize at /home/agora/ejerscala/ejercicio8.scala:42
totalcliente: org.apache.spark.rdd.RDD[(String, Double)] = ShuffledRDD[34] at reduceByKey at /home/agora/ejerscala/ejercicio8.scala:41
res80: Array[(String, Double)] = Array((11111111H,78.7), (22222222H,70.7), (33333333H,37.4))
res81: scala.collection.Map[String,Long] = Map(11111111H -> 2, 22222222H -> 2, 33333333H -> 1)
scala>
Ctrl Derecho
Introducción a
Agrupar ele
escholarium.educa
eScholarium
Indice
2) Creamos el ejercicio1.scala en el
3) Probamos los comentarios, y las
4) Probamos a escribir datos en HD
5) Leer RDD desde HDFS
6) Creamos un RDD y a partir de él
del primero.
7) Crear un RDD con los valores de
8) Crea un nuevo RDD con 3 partici
Array(23,54,32,66,0,-14,-8,10). C
El total de la suma de los número
El valor máximo MAX
El valor mínimo de la secuencia d
9) Si tenemos una secuencia formad
Queremos aplicar las transforma
resultados:
List(APPLE, BANANA, ORANGE
List(A, P, P, L, E, B, A, N, A, N, A, O,
10) Se nos facilita una lista formada p
(("11111111H",39.1),("2222222
("22222222H",36.5))
Queremos que nuestro código n
dinero que se gastado en las com
También queremos saber por cad
REDUCEBY KEY Y COUNT BYK
Ejemplo de aplicación contr...