Spark

عادل قاضی خانی
هیئت علمی گروه کامپیوتر دانشگاه امام رضا (ع)

تعریف

- بستری برای برای پردازش توزیع شده روی **خوشه ای** [1] از کامپیوترها که سریع و همه منظوره است

1-cluster

RDD

- مجموعه ای از اشیاء توزیع شده
- اشیاء می توانند در ***Python,Scala,Java*** تعریف شوند

Resilient distributed dataset

RDD

- به دو روش ایجاد می شود:

۱- خواندن یک مجموعه داده خارجی

- مثال:

```
scala> val distFile = sc.textFile("data.txt")
distFile: org.apache.spark.rdd.RDD[String] = data.txt MapPartitionsRDD[10] at textFile at
    <console>:26
```

Resilient distributed dataset

RDD

- به دو روش ایجاد می شود:

۲- توزیع کردن مجموعه ای از اشیاء در برنامه راه انداز

- مثال:

```
val data = Array(1, 2, 3, 4, 5)
val distData = sc.parallelize(data)
```

Resilient distributed dataset

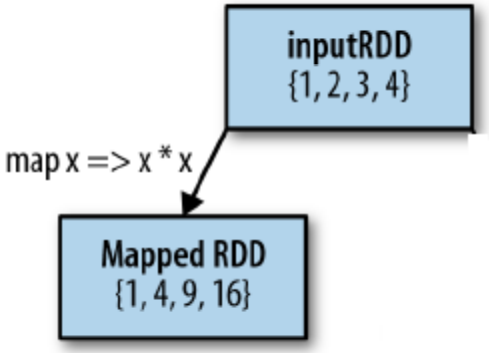
عملیات در RDD

- **دو نوع عملیات:**

۱- تبدیل[1] : نتیجه این عملیات یک RDD جدید است

- **مثال: تبدیل map()**

```
nums = sc.parallelize([1, 2, 3, 4])
squared = nums.map(lambda x: x * x).collect()
for num in squared:
    print "%i " % (num)
```



Resilient distributed dataset

1-Transformation

عملیات در RDD

- **دو نوع عملیات:**

۲- **فعالیت**[1] : نتیجه این عملیات به برنامه راه انداز برگردانده می شود یا روی دیسک نوشته می شود

- **مثال: فعالیت reduce()**

```
sum = rdd.reduce(lambda x, y: x + y)
```

Resilient distributed dataset

1-action

ایجاد کلید مقدار RDD

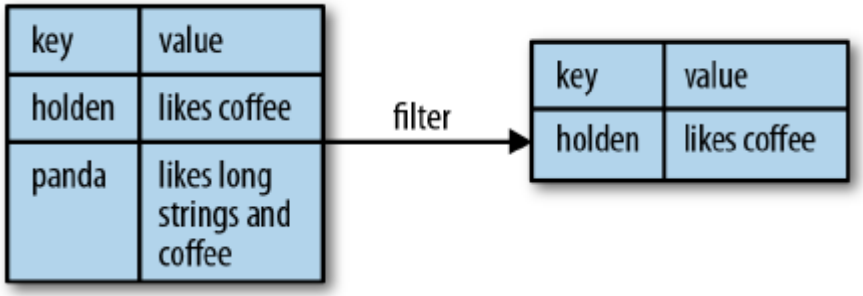
- RDD جفت
- **مثال:**

```
pairs = lines.map(lambda x: (x.split(" ")[0], x))
```

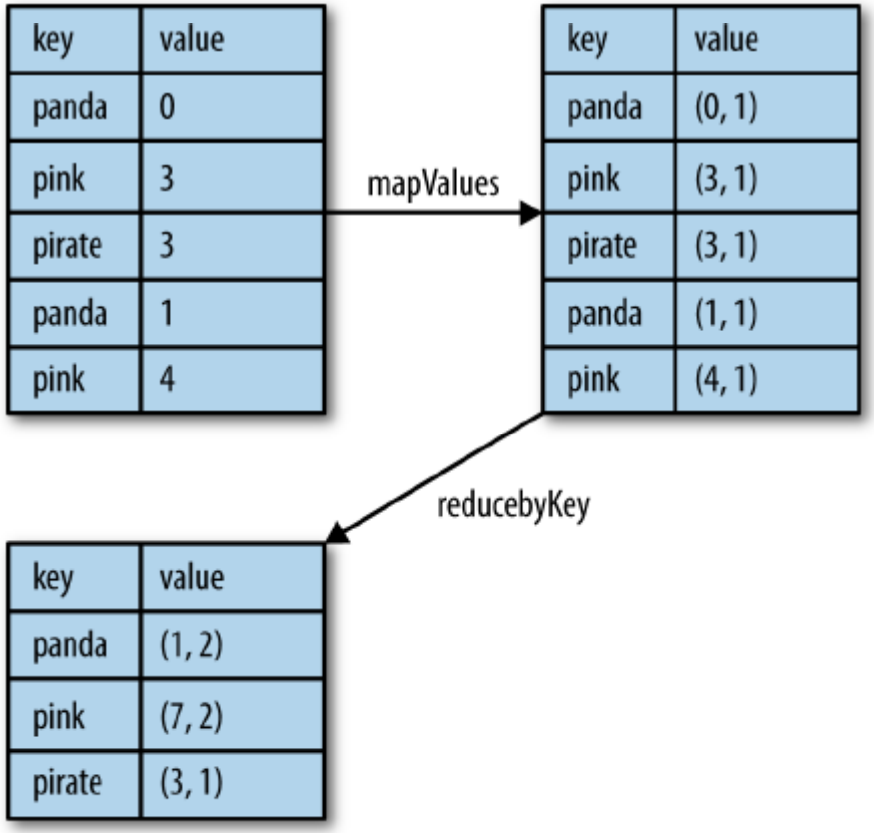
تبدیل ها روی RDD جفت

Function name	Purpose	Example	Result
`reduceByKey(func)`	Combine values with the same key.	`rdd.reduceByKey(x, y) => x + y)`	`{(1, 2), (3, 10)}`
`groupByKey()`	Group values with the same key.	`rdd.groupByKey()`	`{(1, [2]), (3, [4, 6])}`
`combineByKey(createCombiner, mergeValue, mergeCombiners, partitioner)`	Combine values with the same key using a different result type.	See Examples 4-12 through 4-14.	
`mapValues(func)`	Apply a function to each value of a pair RDD without changing the key.	`rdd.mapValues(x => x+1)`	`{(1, 3), (3, 5), (3, 7)}`

تبدیل ها روی RDD جفت – مثال



تبدیل ها روی RDD جفت – مثال

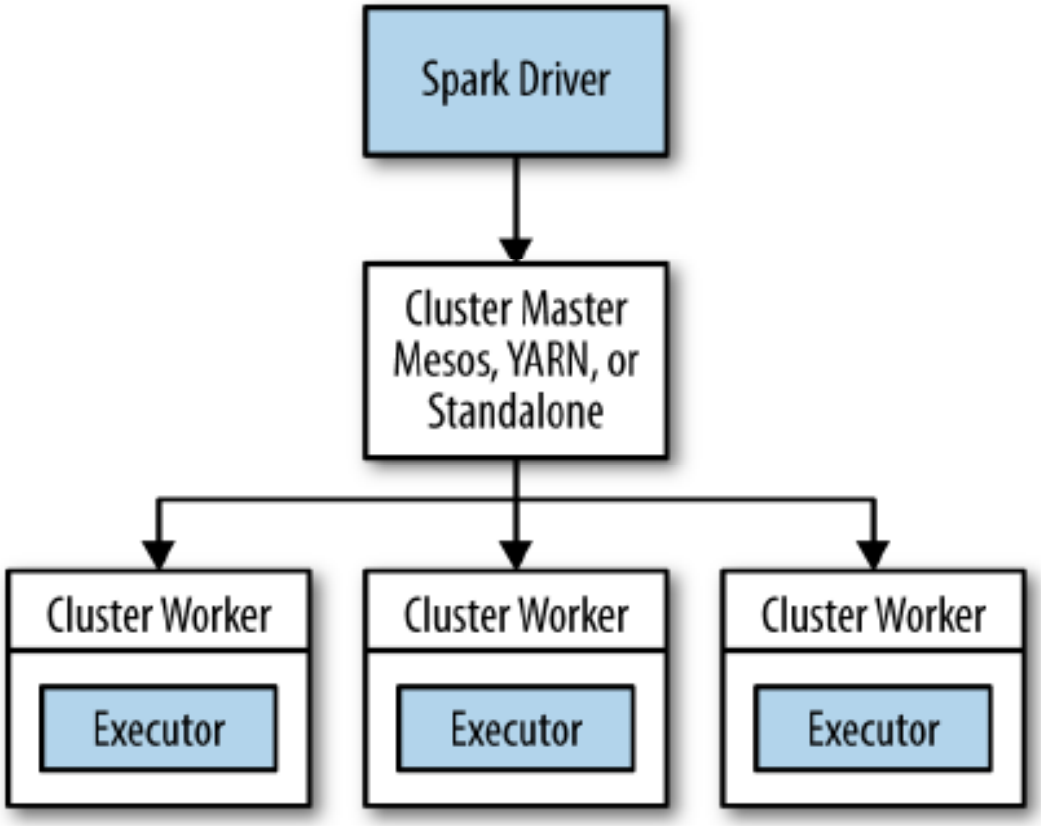


فعالیتها روی RDD جفت

Function	Description	Example	Result
`countByKey()`	Count the number of elements for each key.	`rdd.countByKey()`	`{(1, 1), (3, 2)}`
`collectAsMap()`	Collect the result as a map to provide easy lookup.	`rdd.collectAsMap()`	`Map{(1, 2), (3, 4), (3, 6)}`
`lookup(key)`	Return all values associated with the provided key.	`rdd.lookup(3)`	`[4, 6]`

اجرا کردن برنامه در خوشه

- ساختار



اجرا کردن برنامه در خوشه

- **راه انداز برنامه:** نقطه شروع برنامه، دو کار اصلی

❖تبدیل برنامه به تعدادی task (واحدهای فیزیکی اجرا)

❖برنامه ریزی اجرای task ها روی اجرا کننده ها

اجرا کردن برنامه در خوشه

- **اجرا کننده:** مجری برنامه ها

❖اجرای برنامه و برگرداندن نتایج به راه انداز

❖امکان ذخیره سازی RDD ها در RAM