PAIR RDDS

There are 2 types of RDDs

Basic RDDs Each element is a single object

Pair RDDs Each element is a **Key/Value pair**

Basic RDDs

Till now, we have only worked with Basic RDDs

ie. we treat each record in the RDD as a single object

Basic RDDs

All our transformations, actions act on each record as a whole

There are 2 types of RDDs

Basic RDDs Each element is a single object

Pair RDDs Each element is a **Key/Value pair**

Pair RDDs

Each element is a Key/Value pair

Many data processing tasks can be easily expressed using Key, Value pairs

Ex: Delays by Airline, Sales by City, Word Counts etc

Pair RDDs

Pair RDDs are special RDDs where each record is treated as tuple

Pair RDDs

All the basics RDD transformations and actions work for Pair RDDs too

Special Transformations and Actions exist for Pair RDDs

Pair RDDs

Transformations

keys
values
mapValues
groupByKey
reduceByKey
combineByKey

A few transformations for pair RDDs

Pair RDDs

Transformations

keys
values
mapValues
groupByKey
reduceByKey
combineByKey

Return RDDs with only the keys or the values

Pair RDDs

Transformations

mapValues

keys

values

groupByKey

reduceByKey

combineByKey

Takes a function and applies it on the values of the key, value pairs

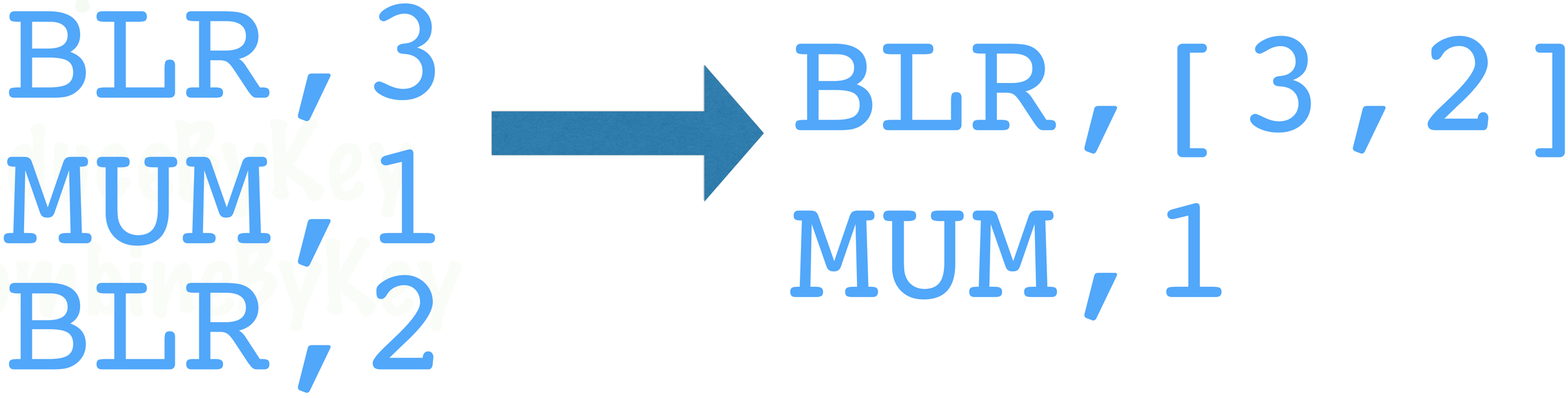
Pair RDDs

Transformations

groupByKey

Groups the values which have the same key into a list/collection

keys
values
mapValues
reduceByKey
combineByKey



cogroup is also like groupByKey

But it can group values across RDDS

Pair RDDs

Transformations

reduceByKey

This is like reduce on Basic RDDS

It takes a function to combine 2 values

It combines values with the same key

Pair RDDs

Transformations

keys

values

mapValues

groupByKey

reduceByKey

combineByKey

Using a Pair RDD of City, Sales

you can find the sum of sales for each city

Pair RDDs

Transformations

combineByKey

keys
values
mapValues
groupByKey
reduceByKey

Just as for basic RDDs, we have reduce and aggregate

For Pair RDDs, we have reduceByKey and combineByKey

Pair RDDs

Transformations

combineByKey

keys

values

reduce and aggregate

mapValues

groupByKey

reduceByKey and combineByKey

reduceByKey

Note one important difference!

Pair RDDs

Transformations

combineByKey

keys

values

reduce and aggregate

Actions on basic RDDs

mapValues

groupByKey

reduceByKey and combineByKey

Transformations on Pair RDDs

Pair RDDs

Transformations

One of the most common operations is

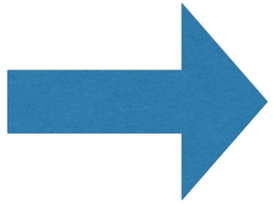
to merge 2 Pair RDDs

based on the keys

Pair RDDs

Transformations

Pair RDD1	Pair RDD2	Merge 2 Pair RDDs
BLR,3	BLR,"B"	BLR,[3,"B"]
MUM,1	MUM,"M"	MUM,[1,"M"]
DEL,2	DEL,"D"	DEL,[2,"D"]



Such operations are called joins

Pair RDDs

Transformations

joins

join

left outer join

right outer join

These are similar to their counter parts in SQL

Pair RDDs

Transformations

join

left outer join

right outer join

A join will return a new Pair RDD

Values from the input RDDs whose keys match are grouped together

Pair RDDs

Transformations

join

Only keys which exist in both RDDs are returned

Like an inner join in SQL

left outer join

right outer join

Pair RDDs

Transformations

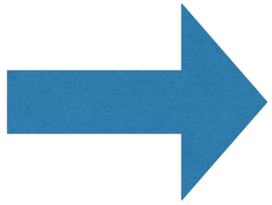
join

Pair RDD1

```
BLR,3
MUM,1
DEL,2
```

Pair RDD2

```
BLR,"B"
MUM,"M"
```



```
BLR,[3,"B"]
MUM,[1,"M"]
```

Pair RDDs

Transformations

join

left outer join

right outer join

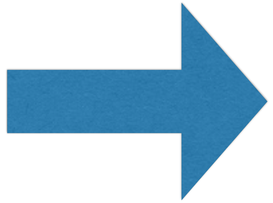
All keys from the left RDD are returned

Pair RDDs

Transformations

left outer join

Pair RDD1	Pair RDD2
BLR,3	BLR,"B"
MUM,1	MUM,"M"
DEL,2	KOL,"D"



BLR,[3,"B"]

MUM,[1,"M"]

DEL,[2,None]

Pair RDDs

Transformations

join

left outer join

right outer join

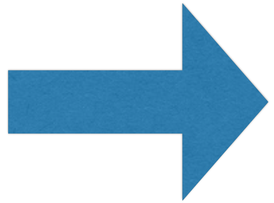
All keys from the right RDD are returned

Pair RDDs

Transformations

right outer join

Pair RDD1	Pair RDD2	
BLR,3	BLR,"B"	BLR,[3,"B"]
MUM,1	MUM,"M"	MUM,[1,"M"]
DEL,2	KOL,"D"	KOL,[None,"D"]



Pair RDDs

countByKey

lookup

collectAsMap

Actions

A few special actions are available for pair RDDs

Pair RDDs

Actions

countByKey count the number of values per key

lookup returns all values for a specific key

collectAsMap returns a dict with all the key value pairs