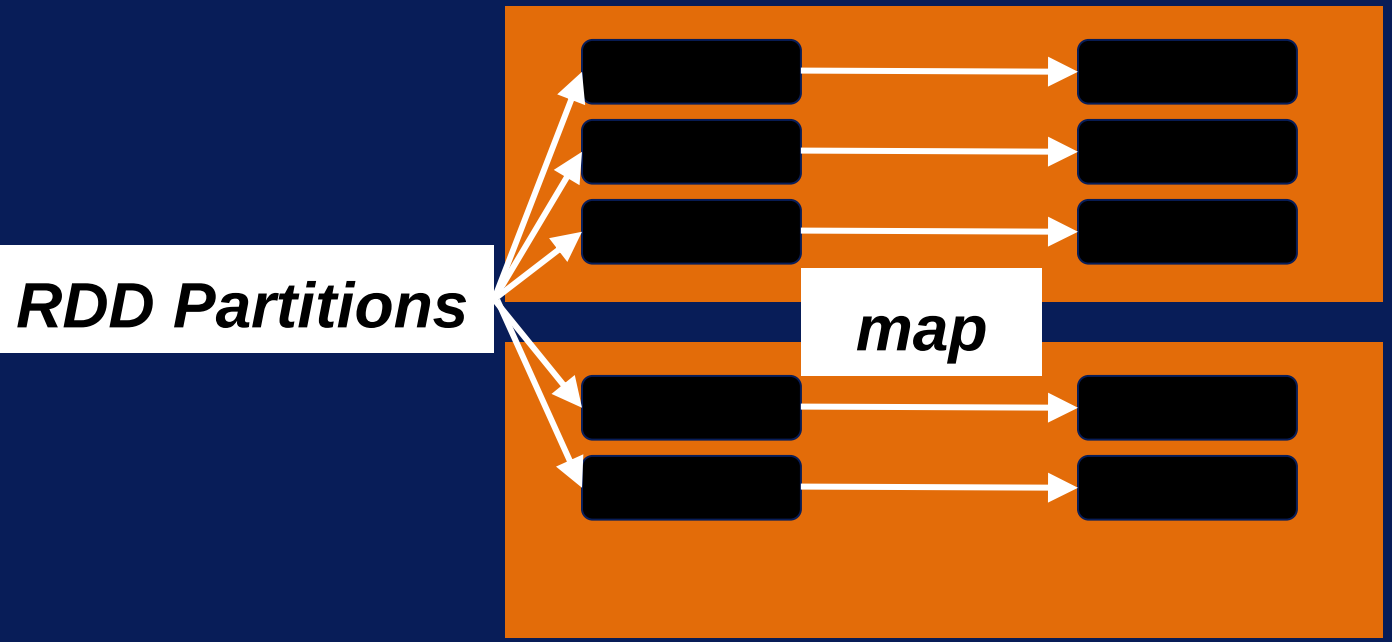
Spark Core: Transformations

After this video you will be able to..

- Explain the difference between a narrow transformation and wide transformation
- Describe map, flatmap, filter and coalesce as narrow transformations
- List two wide transformations

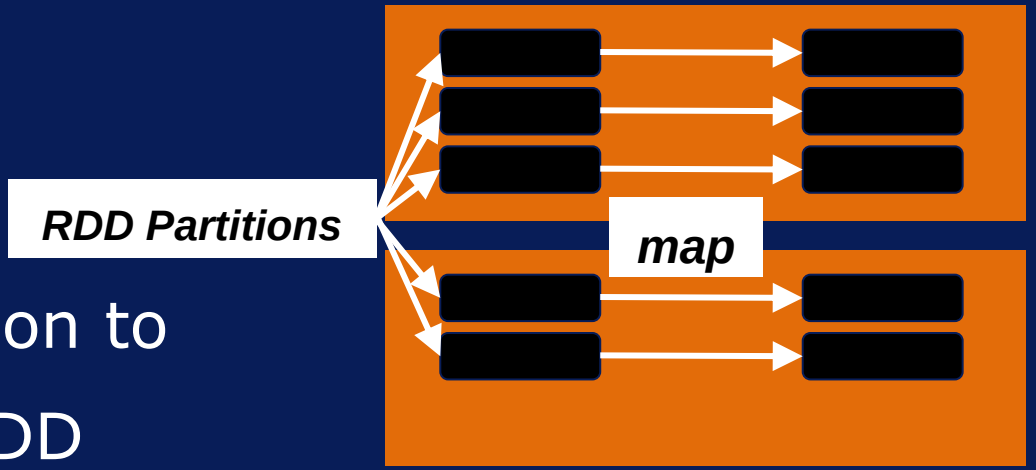
map

map : apply function to each element of RDD



map

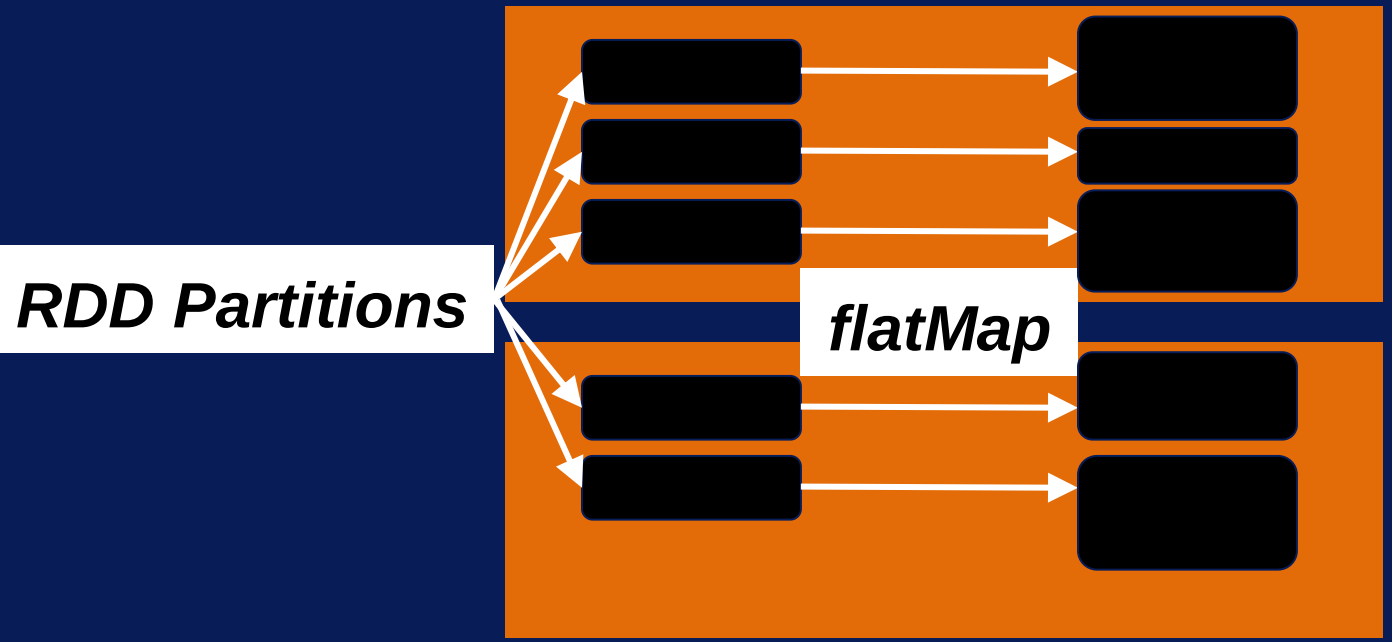
map : apply function to each element of RDD




```
def lower(line):
    return line.lower()
lower_text_RDD = text_RDD.map(lower)
```

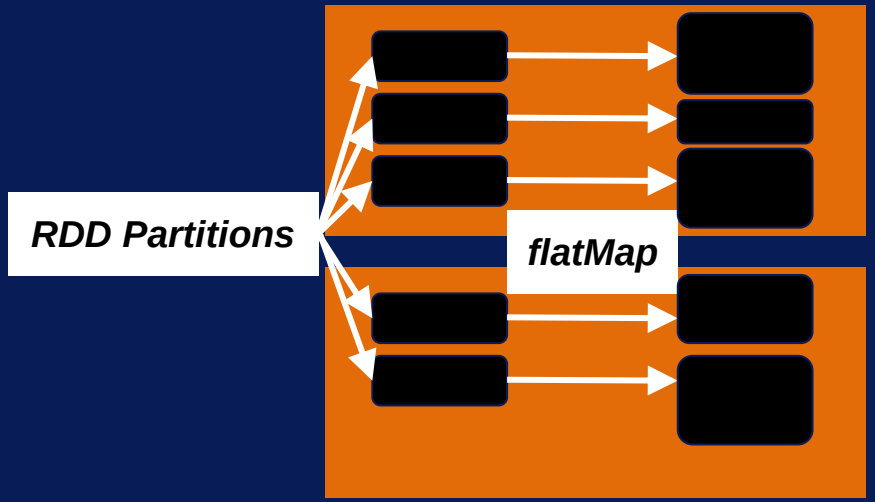
flatMap

flatMap : map then flatten output



flatMap

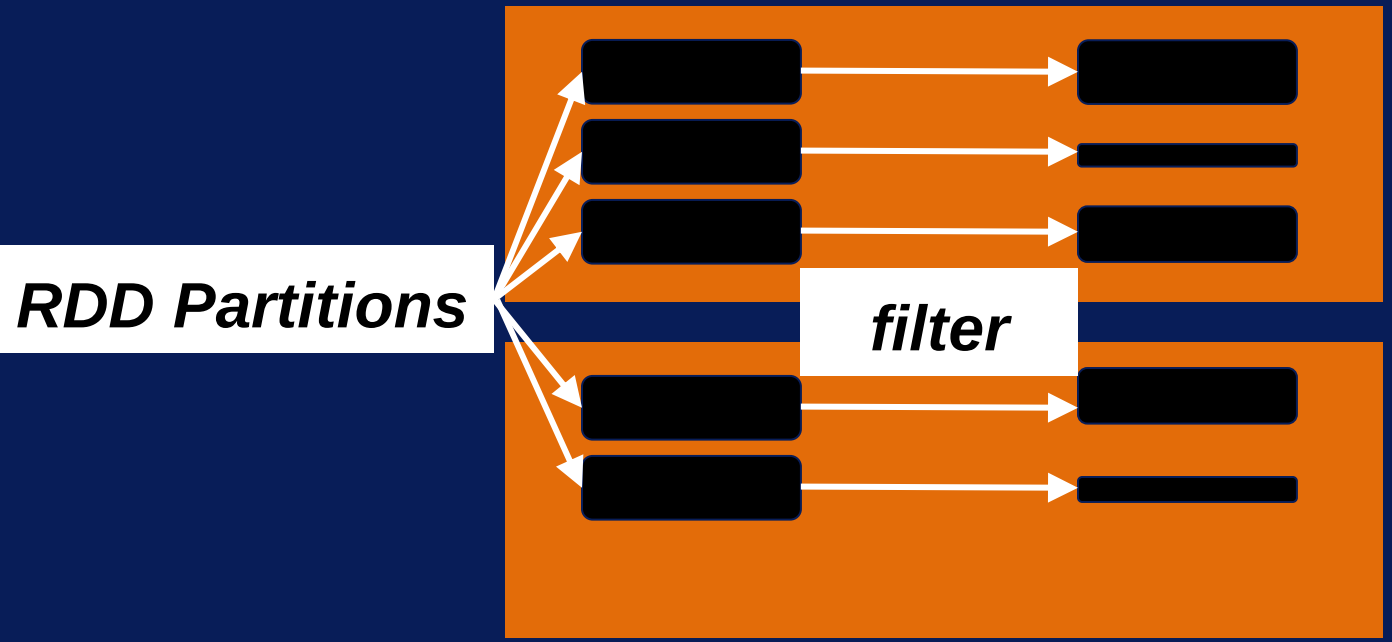
flatMap : map then flatten output



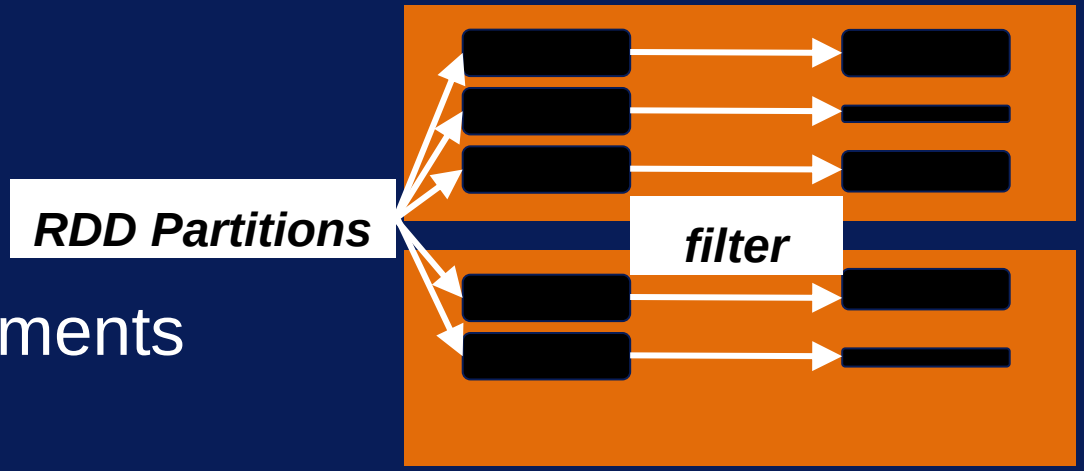

```
def split_words(line):
    return line.split()

words_RDD = text_RDD.flatMap(split_words)
words_RDD.collect()
```

filter

filter : keep only elements where function is true



filter



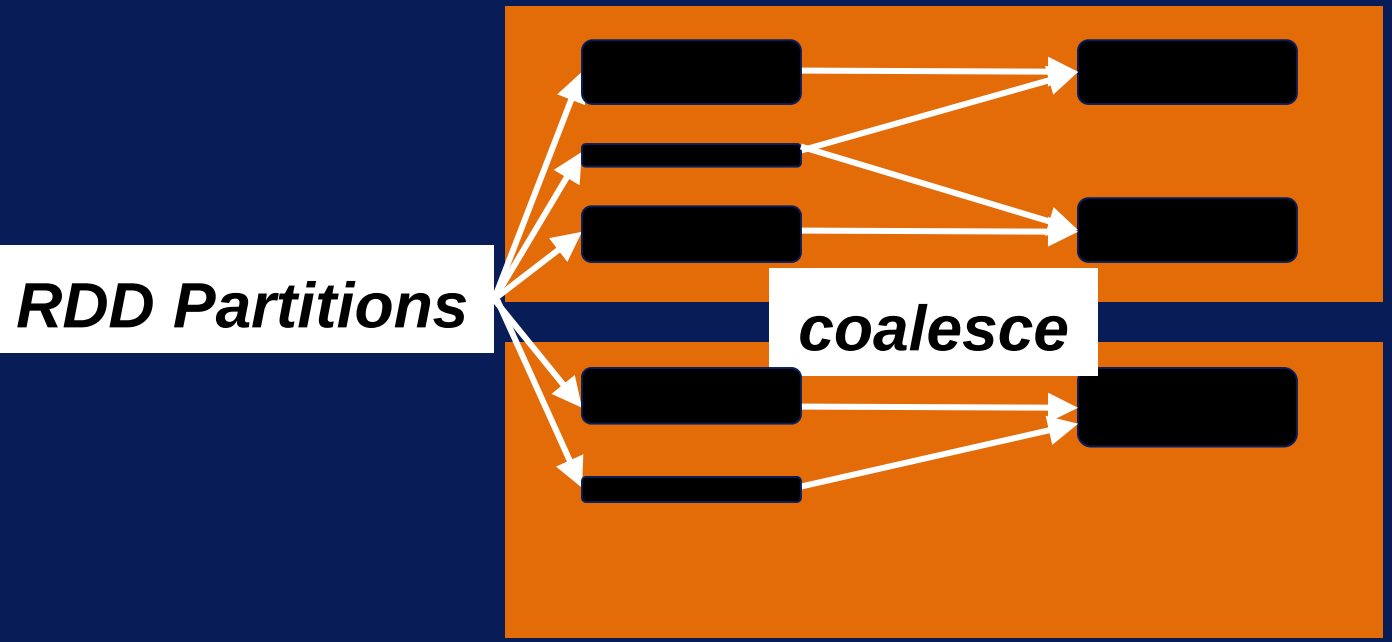
filter : keep only elements

where function is true

```
def starts_with_a(word):
    return word.lower().startswith("a")
words_RDD.filter(starts_with_a).collect()
```

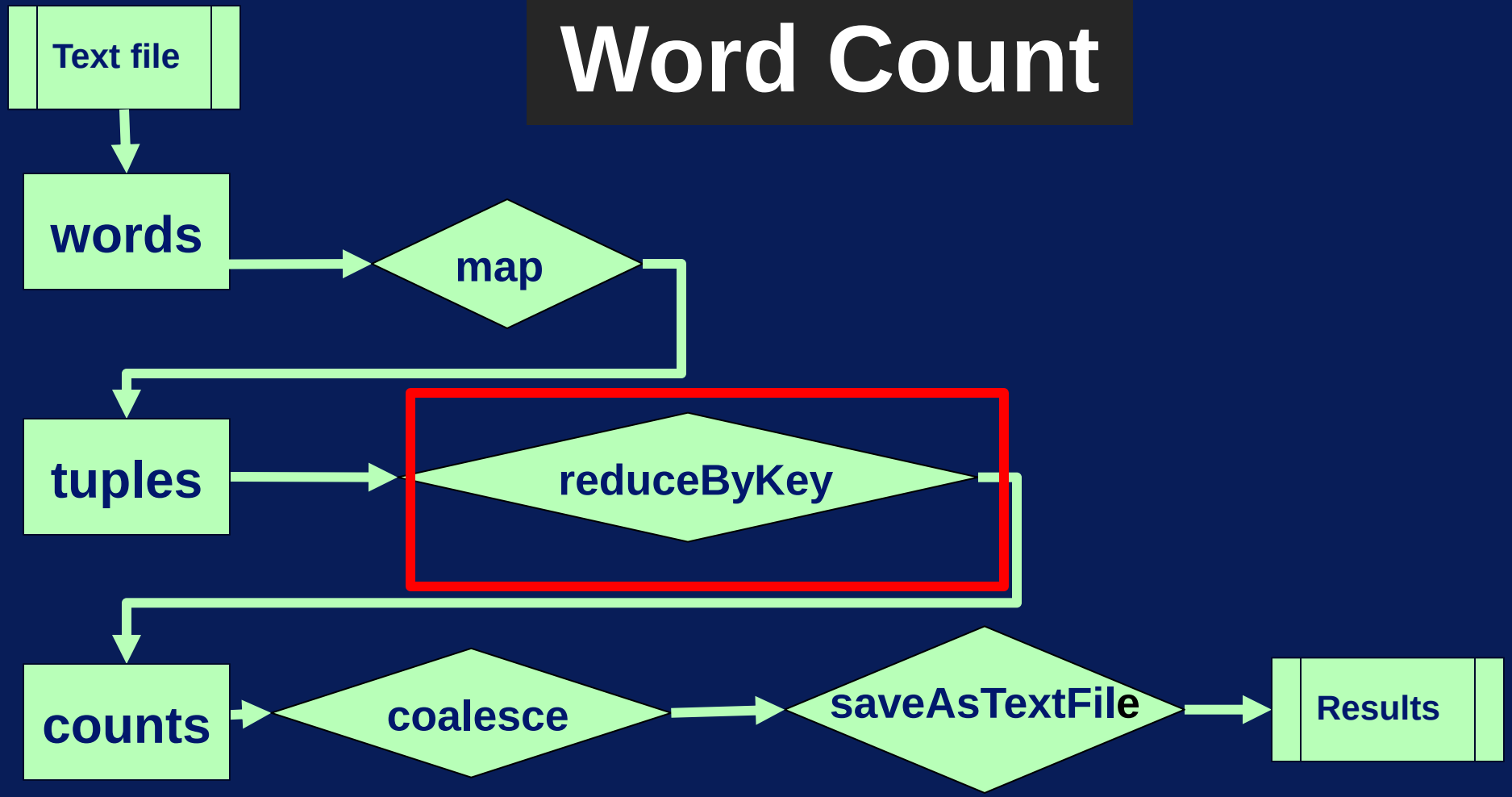
coalesce

coalesce : reduce the number of partitions

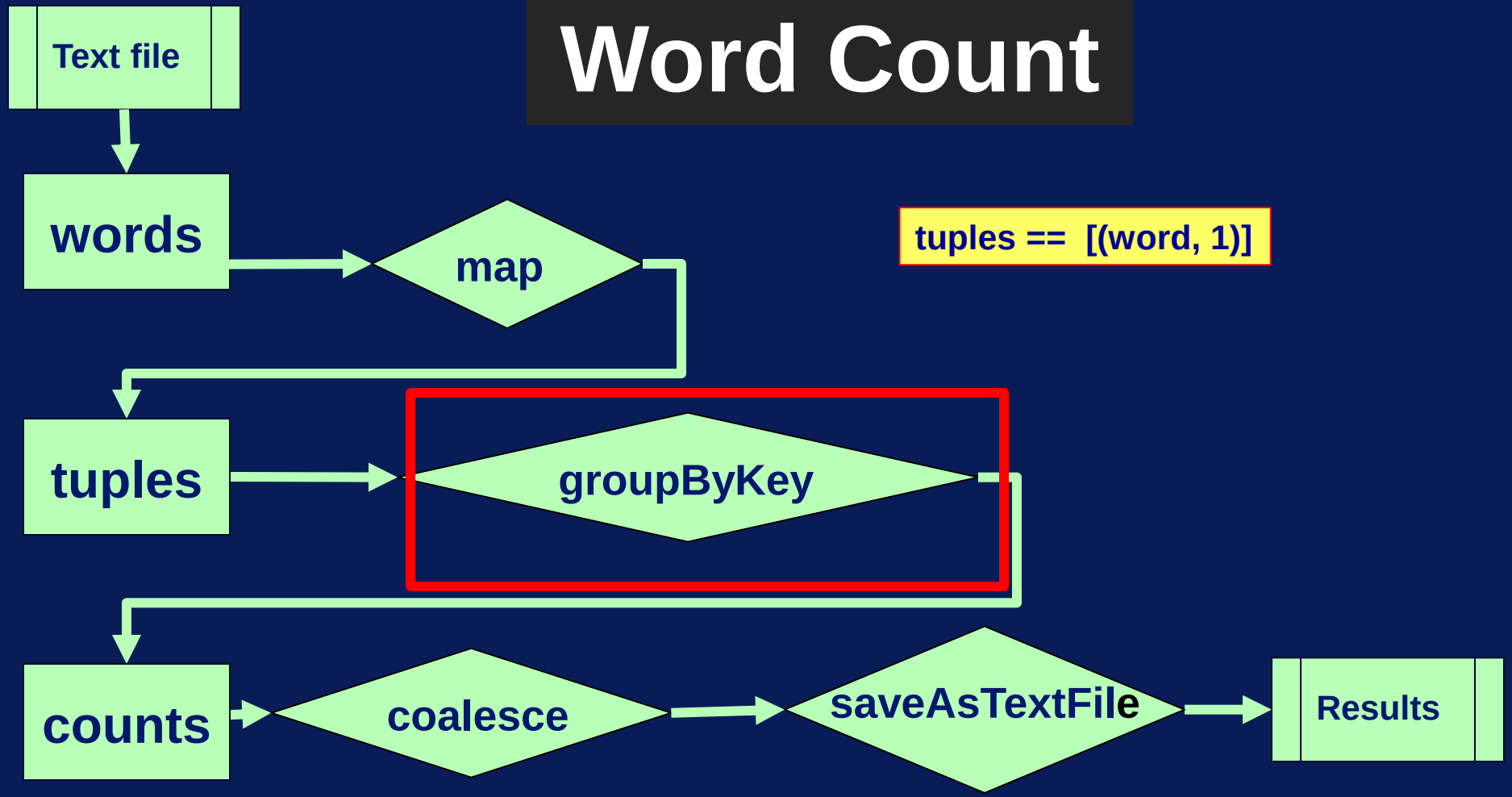


Wide Transformations

Word Count

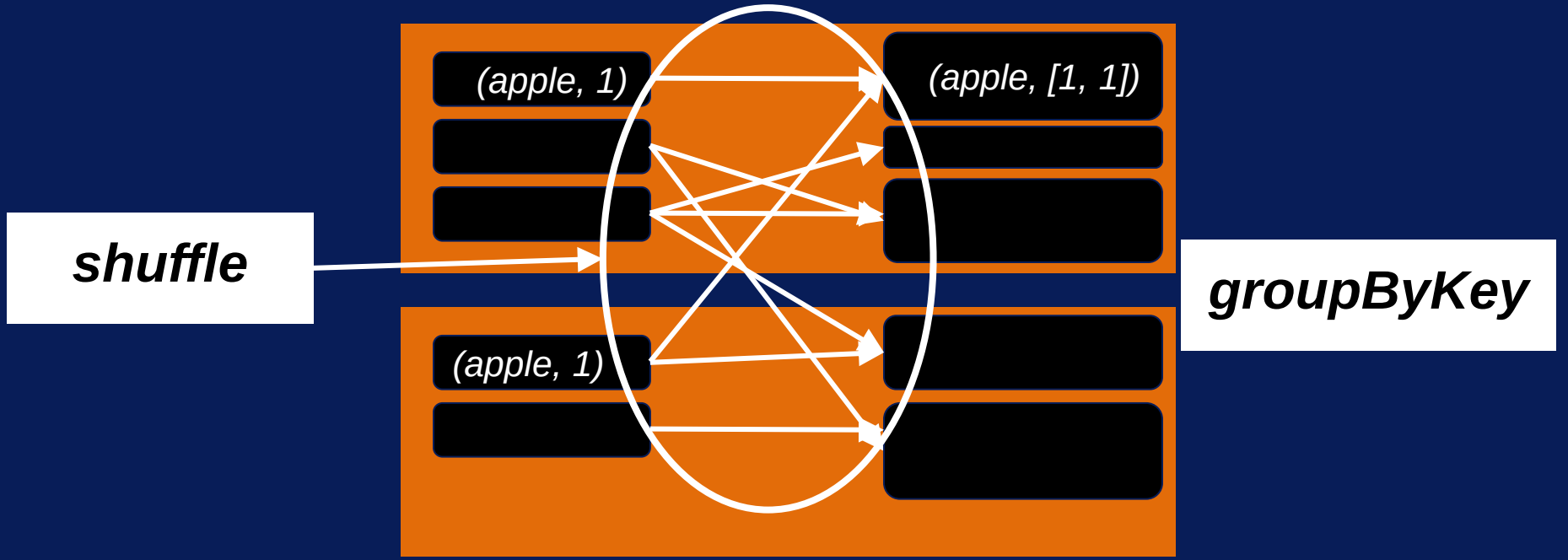

Text file
words
map
tuples
reduceByKey
counts
coalesce
saveAsTextFile
Results

Word Count

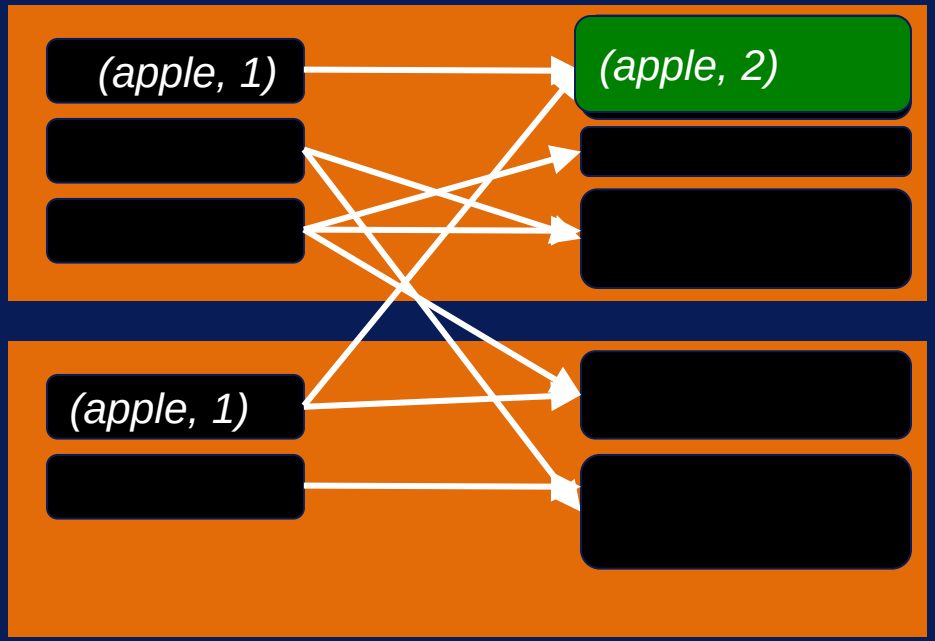

Text file
words
map
tuples == [(word, 1)]
tuples
groupByKey
counts
coalesce
saveAsTextFile
Results

groupByKey

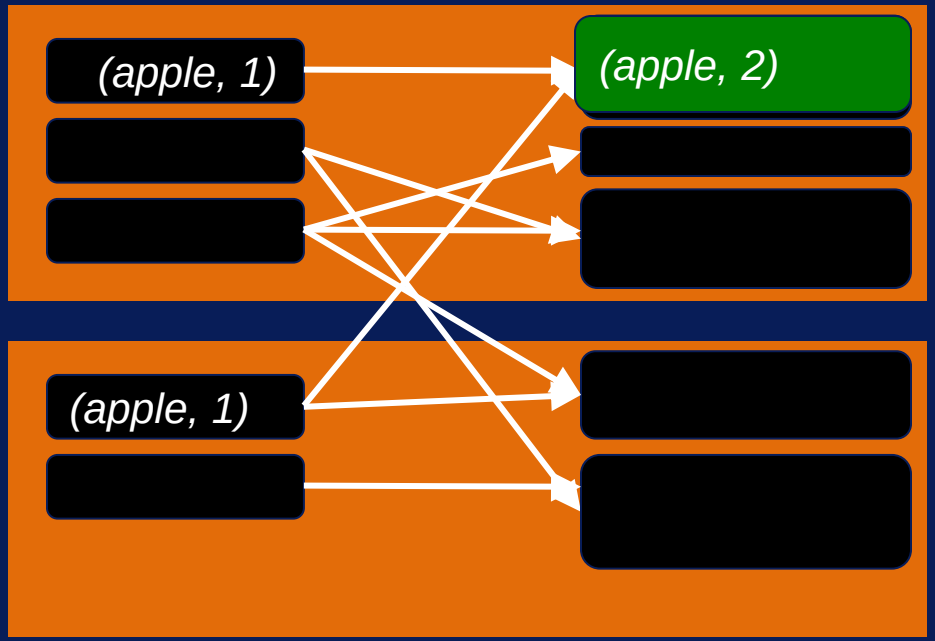
groupByKey : (K, V) pairs => (K, list of all V)



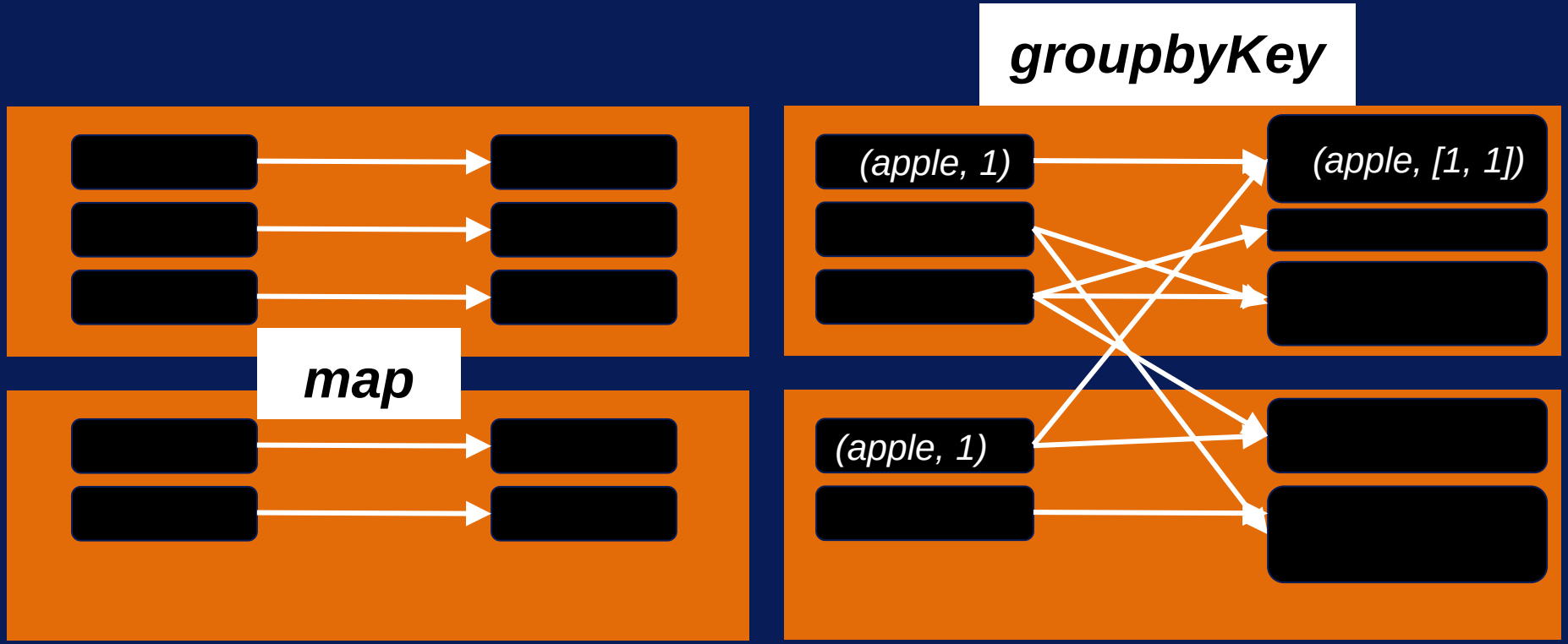
groupByKey + reduce



reduceByKey



Narrow vs Wide



Many more transformations…

Full list of transformations at:

https://spark.apache.org/docs/1.2.0/programming-guide.html#transformations