

Transformations and Actions: Takeaways



by Dataquest Labs, Inc. - All rights reserved © 2020

Syntax

- Generate a sequence of values from an RDD:

```
def hamlet_speaks(line):  
    id = line[0]  
    speaketh = False  
    if "HAMLET" in line:  
        speaketh = True  
    if speaketh:  
        yield id, "hamlet speaketh!"  
  
hamlet_spoken = split_hamlet.flatMap(lambda x: hamlet_speaks(x))
```

- Return the number of elements in an RDD:

```
hamlet_spoken_lines.count()
```

- Return a list representation of an RDD:

```
hamlet_spoken_lines.collect()
```

Concepts

- `yield` is a Python technique that allows the interpreter to generate data on the fly and pull it when necessary as opposed to storing to the memory immediately.
- Spark takes advantage of 'yield' to improve the speed of computations.
- `flatMap()` is useful when you want to generate a sequence of values from an RDD.

Resources

- [Python yield](#)

- [Difference between map and flatMap in Apache Spark](#)



Takeaways by Dataquest Labs, Inc. - All rights reserved © 2020