

Lecture 4: Build simple Spark applications 1

IT5427  
Tích hợp và xử lý dữ liệu lớn

IT5427 Spring 2020  
Thanh-Chung Dao Ph.D.

1

**Spark running mode**

- Local
- Clustered
  - Spark Standalone
  - Spark on Apache Mesos
  - Spark on Hadoop YARN

2

2

## Hello World: Word-Count

```

1 import sys
2 from pyspark import SparkContext
3 sc = SparkContext(appName="WordCountExample")
4 lines = sc.textFile(sys.argv[1])
5 counts = lines.flatMap(lambda x: x.split(' ')) \
6                 .map(lambda x: (x, 1)) \
7                 .reduceByKey(lambda x,y:x+y)
8 output = counts.collect()
9 for (word, count) in output:
10     print "%s: %i" % (word, count)
11 sc.stop()

```

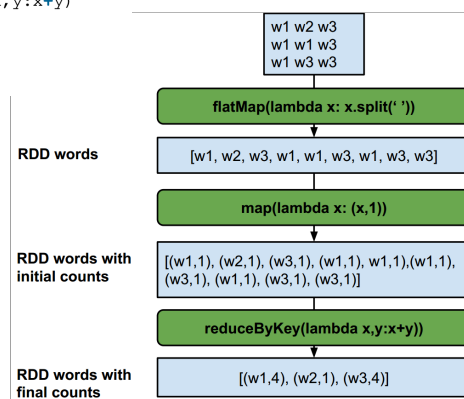


Figure from [1]

3

3

## Run using command line

- Turn on docker bash
- `spark-submit wordcount.py README.md`
- Result will be shown as follows

```

19/05/19 07:56:51 INFO scheduler.TaskSchedulerImpl: Removed
pool
19/05/19 07:56:51 INFO scheduler.DAGScheduler: ResultStage 1
) finished in 0.150 s
19/05/19 07:56:51 INFO scheduler.DAGScheduler: Job 0 finishe
0, took 2.050066 s
Turks: 1
States,294: 1
Algeria,United: 1
States,2025: 1
States,955: 1
States,Czech: 1
Colombia,United: 1
States,588: 1
States,Dominican: 1

```

4

4

### **Lab: Word-Count**

- Lab on the Zeppelin notebook
- Github source code
  - <https://github.com/bk-blockchain/big-data-class>

5

5

### **Flight data:**

- Analyzing flight data from the United States Bureau of Transportation statistics
- Lab on the Zeppelin notebook
- Github source code
  - <https://github.com/bk-blockchain/big-data-class>

6

6

## References

- [1]  
<https://datamize.wordpress.com/2015/02/08/visualizing-basic-rdd-operations-through-wordcount-in-pyspark/>

7