IT4043E
Tích hợp và xử lý dữ liệu lớn

IT4043E

12/2022
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

```
import sys
from pyspark import SparkContext
sc = SparkContext(appName="WordCountExample")
lines = sc.textFile(sys.argv[1])
counts = lines.flatMap(lambda x: x.split(' ')) \
                  .map(lambda x: (x, 1)) \
                  .reduceByKey(lambda x,y:x+y)
output = counts.collect()
for (word, count) in output:
    print "%s: %i" % (word, count)
sc.stop()
```

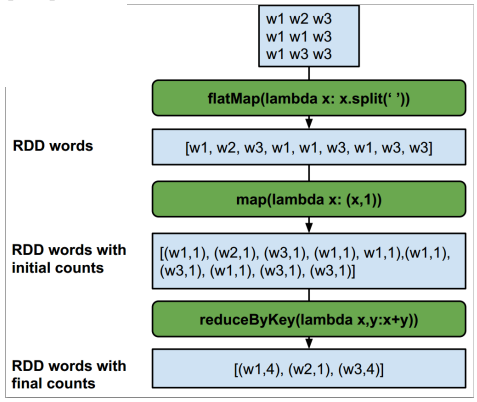

Figure from [1]

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

7