Big data: architectures and data analytics

Exercise #31

- Log analysis
  - Input: log of a web server (i.e., a textual file)
    - Each line of the file is associated with a URL request
  - Output: the list of distinct IP addresses associated with the connections to a google page (i.e., connections to URLs containing the term "www.google.com")
    - Store the output in an HDFS folder

2

Exercise #31 - Example

- Input file

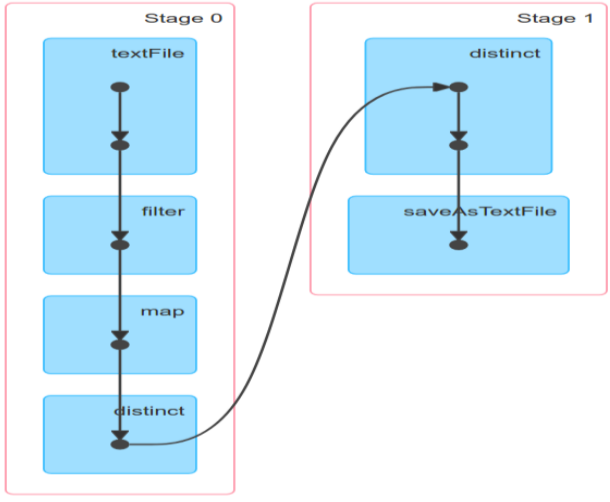
```
66.249.69.97 - - [24/Sep/2014:22:25:44 +0000] "GET http://www.google.com/bot.html"
66.249.69.97 - - [24/Sep/2014:22:26:44 +0000] "GET http://www.google.com/how.html"
66.249.69.97 - - [24/Sep/2014:22:28:44 +0000] "GET http://dbdmg.polito.it/course.html"
71.19.157.179 - - [24/Sep/2014:22:30:12 +0000] "GET http://www.google.com/faq.html"
66.249.69.95 - - [24/Sep/2014:31:28:44 +0000] "GET http://dbdmg.polito.it/thesis.html"
66.249.69.97 - - [24/Sep/2014:56:26:44 +0000] "GET http://www.google.com/how.html"
56.249.69.97 - - [24/Sep/2014:56:26:44 +0000] "GET http://www.google.com/how.html"
```

- Output

```
66.249.69.97
71.19.157.179
56.249.69.97
```

3

Exercise #31 - DAG



4

Exercise #31 - Simulation

- Suppose that Sparks splits the RDD associated with the input file in two partitions
- Part. #1

```
66.249.69.97 - - [24/Sep/2014:22:25:44 +0000] "GET http://www.google.com/bot.html"
66.249.69.97 - - [24/Sep/2014:22:26:44 +0000] "GET http://www.google.com/how.html"
66.249.69.97 - - [24/Sep/2014:22:28:44 +0000] "GET http://dbdmg.polito.it/course.html"
```

- Part. #2

```
71.19.157.179 - - [24/Sep/2014:22:30:12 +0000] "GET http://www.google.com/faq.html"
66.249.69.95 - - [24/Sep/2014:31:28:44 +0000] "GET http://dbdmg.polito.it/thesis.html"
66.249.69.97 - - [24/Sep/2014:56:26:44 +0000] "GET http://www.google.com/how.html"
56.249.69.97 - - [24/Sep/2014:56:26:44 +0000] "GET http://www.google.com/how.html"
```

5

Exercise #31 - Simulation

Part. #1

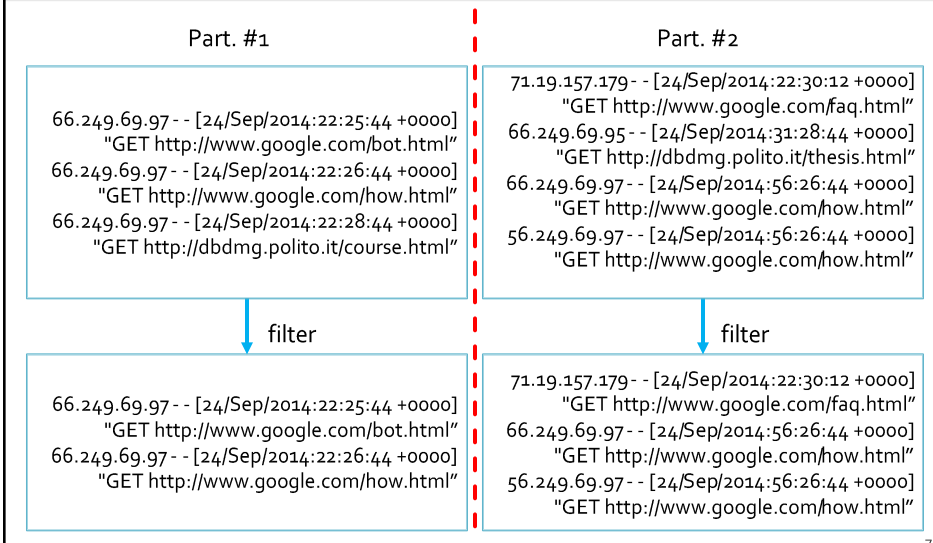
```
66.249.69.97 - - [24/Sep/2014:22:25:44 +0000]
"GET http://www.google.com/bot.html"
66.249.69.97 - - [24/Sep/2014:22:26:44 +0000]
"GET http://www.google.com/how.html"
66.249.69.97 - - [24/Sep/2014:22:28:44 +0000]
"GET http://dbdmg.polito.it/course.html"
```

Part. #2

```
71.19.157.179 - - [24/Sep/2014:22:30:12 +0000]
"GET http://www.google.com/faq.html"
66.249.69.95 - - [24/Sep/2014:31:28:44 +0000]
"GET http://dbdmg.polito.it/thesis.html"
66.249.69.97 - - [24/Sep/2014:56:26:44 +0000]
"GET http://www.google.com/how.html"
56.249.69.97 - - [24/Sep/2014:56:26:44 +0000]
"GET http://www.google.com/how.html"
```

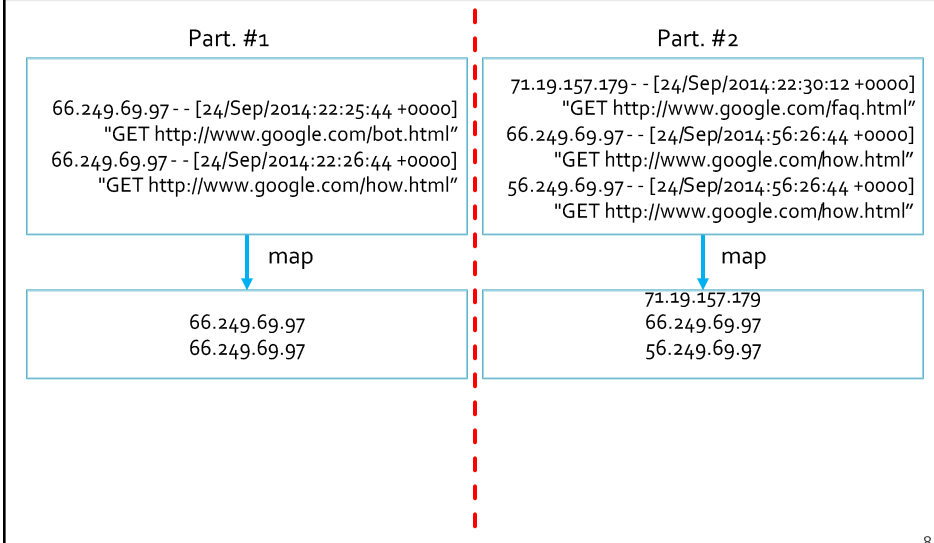
6

Exercise #31 - Simulation



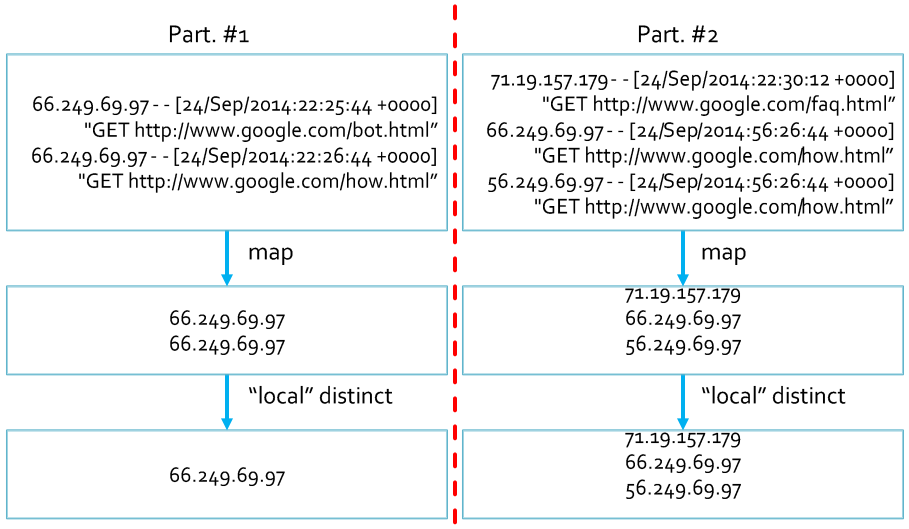
7

Exercise #31 - Simulation



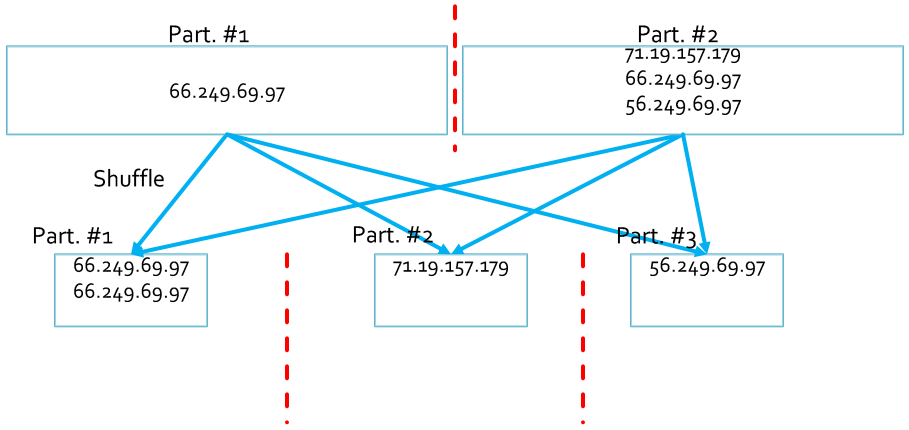
8

Exercise #31 - Simulation



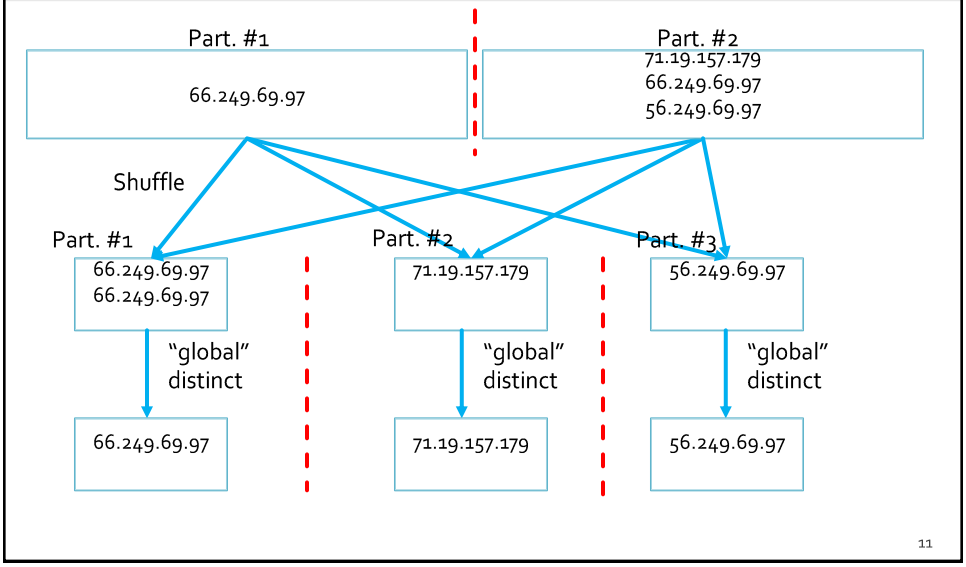
9

Exercise #31 - Simulation



10

Exercise #31 - Simulation



Exercise #31 - Simulation