

Big data: architectures and data analytics

MapReduce - Exercises

2

Exercise #1

- Word count problem
 - Input: (unstructured) textual file
 - Output: number of occurrences of each word appearing in the input file

3

Exercise #1 - Example

- Input file

Toy example file for Hadoop.
 Hadoop running example.
- Output pairs
 - (toy, 1)
 - (example, 2)
 - (file, 1)
 - (for, 1)
 - (hadoop, 2)
 - (running, 1)

4

Exercise #2

- Word count problem
 - Input: a HDFS folder containing textual files
 - Output: number of occurrences of each word appearing in at least one file of the collection (i.e., files of the input directory)
- The only difference with respect to exercise #1 is given by the input
 - Now the input is a collection of textual files

5

Exercise #2 - Example

- Input files

Toy example file for Hadoop.
 Hadoop running example.

Another file for Hadoop.
- Output pairs
 - (another, 1)
 - (example, 2)
 - (file, 2)
 - (for, 2)
 - (hadoop, 3)
 - (running, 1)
 - (toy, 1)

6

Exercise #3

- PM₁₀ pollution analysis
 - Input: a (structured) textual file containing the daily value of PM₁₀ for a set of sensors
 - Each line of the file has the following format
sensorId,date\tPM₁₀ value ($\mu\text{g}/\text{m}^3$)\n
 - Output: report for each sensor the number of days with PM₁₀ above a specific threshold
 - Suppose to set threshold = $50 \mu\text{g}/\text{m}^3$

7

Exercise #3 - Example

- Input file

s1,2016-01-01	20.5
s2,2016-01-01	30.1
s1,2016-01-02	60.2
s2,2016-01-02	20.4
s1,2016-01-03	55.5
s2,2016-01-03	52.5

- Output pairs (s1, 2)
(s2, 1)

8

Exercise #4

- PM₁₀ pollution analysis per city zone
- Input: a (structured) textual file containing the daily value of PM₁₀ for a set of city zones
 - Each line of the file has the following format
zoneId,date\tPM₁₀ value ($\mu\text{g}/\text{m}^3$)\n
- Output: report for each zone the list of dates associated with a PM₁₀ value above a specific threshold
 - Suppose to set threshold = $50 \mu\text{g}/\text{m}^3$

9

Exercise #4 - Example

- Input file

zone1,2016-01-01	20.5
zone2,2016-01-01	30.1
zone1,2016-01-02	60.2
zone2,2016-01-02	20.4
zone1,2016-01-03	55.5
zone2,2016-01-03	52.5

- Output pairs (zone1, [2016-01-03, 2016-01-02])
(zone2, [2016-01-01])

10