Big data: architectures and data analytics

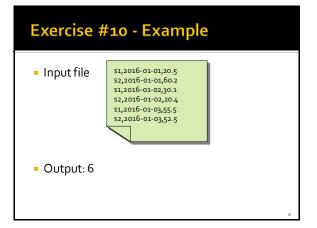
MapReduce - Exercises

Exercise #9

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

Exercise #10

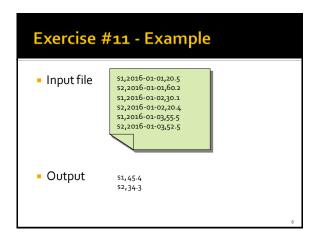
- Total count
 - Input: a collection of (structured) textual csv files containing the daily value of PM10 for a set of sensors
 - Each line of the files has the following format sensorId, date, PM10 value (μg/m³)\n
 - Output: total number of records



Exercise #11

Average

- Input: a collection of (structured) textual csv files containing the daily value of PM10 for a set of sensors
- Each line of the files has the following format sensorld,date,PM10 value (μg/m³)\n
- Output: report for each sensor the average value of PM10
- Suppose the number of sensors is equal to 2 and their ids are s1 and s2



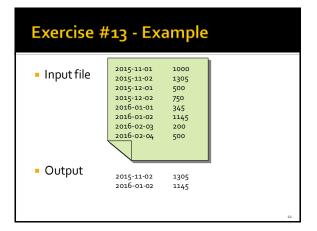
Exercise #12

- Select outliers
 - Input: a collection of (structured) textual files containing the daily value of PM10 for a set of sensors
 - Each line of the files has the following format sensorld,date\tPM10 value (μg/m³)\n
 - Output: the records with a PM10 value below a user provided threshold (the threshold is an argument of the program)

Exercise #12 - Example 51,2016-01-01 Input file 20.5 \$2,2016-01-01 60.2 30.1 \$1,2016-01-02 52,2016-01-02 20.4 s1,2016-01-03 55.5 \$2,2016-01-03 52.5 Threshold: 21 Output s1,2016-01-01 20.5 \$2,2016-01-02 20.4

Exercise #13

- Top 2 most profitable dates
 - Input: a (structured) textual csv files containing the daily income of a company
 - Each line of the files has the following format date\tdaily income\n
 - . .
 - Output:
 - Select the date and income of the top 2 most profitable dates



Exercise #14

- Dictionary
 - Input: a collection of news (textual files)
 - Output:
 - List of distinct words occurring in the collection

Exercise #14 - Example Input file Toyexample Hadoop, Hadoop,

Exercise #15

- Dictionary Mapping word integer
 Input: a collection of news (textual files)
 - Input: a collection of news (textual in
 - Output:
 - List of distinct words occurring in the collection associated with a set of unique integers
 Each word is associated with a unique integer (and viceversa)

###