Lab 3

Student information

- Full name: Qicheng Hu
- E-mail: qhu027@cs.ucr.edu
- UCR NetID: qhu027
- Student ID: X675102

Answers

- (Q1) What do you think the line `job.setJarByClass(Filter.class);` does?

  This method finds Jar of the job by the class of input.

- (Q2) What is the effect of the line `job.setNumReduceTasks(0);`?

  This line sets the number of reduce jobs to 0, which means there is no reduce phase in the whole job, the final result will be the same as the output of map phase.

- (Q3) Where does the `main` function run? (Driver node, Master node, or a slave node).

  The main function sets input and output file stream, the map and reduce task so that it runs on a driver node, which gives information for master and slave to run the tasks.

- (Q4) How many lines do you see in the output?

  There are 27972 lines in the output file.

- (Q5) How many files are produced in the output?

  Run this code on the input data `nasa_19950801.tsv`,

  There is 1 file produced (not counting "_SUCCESS").

  Run this code on the input data `nasa_19950630.22-19950728.12.tsv`,

  There are 5 files produced (not counting "_SUCCESS").

- (Q6) Explain this number based on the input file size and default block size.

  The default block size of my file system is 32MB. This input data `nasa_19950801.tsv` does not exceed the block size so that there is only 1 file; result of input data `nasa_19950630.22-19950728.12.tsv` is split into files of 32MB size.

- (Q7) How many files are produced in the output?

  Run this code on the input data `nasa_1995080.tsv`,

  There is 1 file produced (not counting "_SUCCESS").

  Run this code on the input data `nasa_19950630.22-19950728.12.tsv`,

  There are 2 files produced (not counting "_SUCCESS").

- (Q8) Explain this number based on the input file size and default block size.

  The default block size of HDFS is 128MB. This input data `nasa_19950801.tsv` does not exceed the block size so that there is only 1 file; result of input data `nasa_19950630.22-19950728.12.tsv` is split into files of 128MB size.

- (Q9) How many files are produced in the output directory and how many lines are there in each file?

  2 files are produced (not counting "_SUCCESS").

  4 lines in part-r-00000 and 0 line in part-r-00001.

- (Q10) Explain these numbers based on the number of reducers and number of response codes in the input file.

  There are 2 reducers as defined by `job.setNumReduceTasks(2);` so that there are 2 files in the output folder.

- (Q11) How many files are produced in the output directory and how many lines are there in each file?

  2 files are produced (not counting "_SUCCESS").

  5 lines in part-r-00000 and 2 line in part-r-00001.

- (Q12) Explain these numbers based on the number of reducers and number of response codes in the input file.

  There are 2 reducers as defined by `job.setNumReduceTasks(2);` so that there are 2 files in the output folder.

- (Q13) How many files are produced in the output directory and how many lines are there in each file?

2 file are produced (not counting "_SUCCESS").

1 line in part-r-00000, 0 line in part-r-00001.

- (Q14) Explain these numbers based on the number of reducers and number of response codes in the input file.

  The filter only keeps the code=200 records so that the aggregate adds them to 1 line. 2 files are created because there are two reducers.