Lab 4

Student information

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

Answers

- (Q1) Do you think it will use your local cluster? Why or why not?

  It does not use the cluster. In the web interface there is no running or completed application in the list after I run the command.

- (Q2) Does the application use the cluster that you started? How did you find out?

  It uses the cluster. The web interface shows the application in the completed application list.

- (Q3) What is the Spark master printed on the standard output on IntelliJ IDEA?

  In stdout, there are only 2 lines:

  Using Spark master 'local[*]' Number of lines in the log file 30970

- (Q4) What is the Spark master printed on the standard output on the terminal?

  In stdout, there are only 2 lines:

  Using Spark master 'local[*]'

  Number of lines in the log file 30970

- (Q5) For the previous command that prints the number of matching lines, list all the processed input splits.

INFO HadoopRDD: Input split: file:/Users/qhu/Documents/167/Projects/qhu027_lab4/nasa_19950801.tsv:1169610+1169610

INFO HadoopRDD: Input split: file:/Users/qhu/Documents/167/Projects/qhu027_lab4/nasa_19950801.tsv:0+1169610

- (Q6) For the previous command that counts the lines and prints the output, how many splits were generated?

  4 splits were generated.

- (Q7) Compare this number to the one you got earlier.

  The output that replated to 'split' are:

  INFO HadoopRDD: Input split: file:/Users/qhu/Documents/167/Projects/qhu027_lab4/nasa_19950801.tsv:0+1169610 INFO HadoopRDD: Input split: file:/Users/qhu/Documents/167/Projects/qhu027_lab4/nasa_19950801.tsv:1169610+1169610

  INFO HadoopRDD: Input split: file:/Users/qhu/Documents/167/Projects/qhu027_lab4/nasa_19950801.tsv:1169610+1169610 INFO HadoopRDD: Input split: file:/Users/qhu/Documents/167/Projects/qhu027_lab4/nasa_19950801.tsv:0+1169610

- (Q8) Explain why we get these numbers.

  It seems that it reads the file twice this time (from start to end and from end to start).

- (Q9) What can you do to the current code to ensure that the file is read only once?