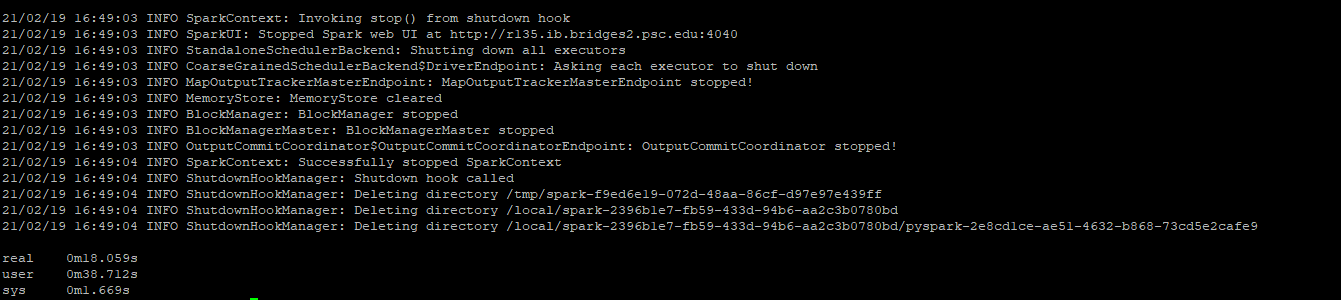
Lab 5 submission

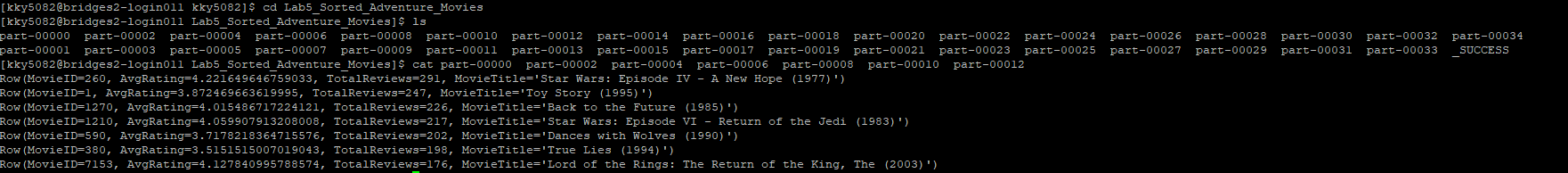
Screen shots of computing time for each of the 4 cases below:

a). The small movie reviews and movies dataset, original PySpark code

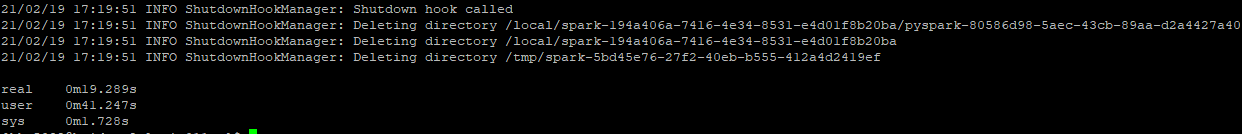
screenshot of computing time



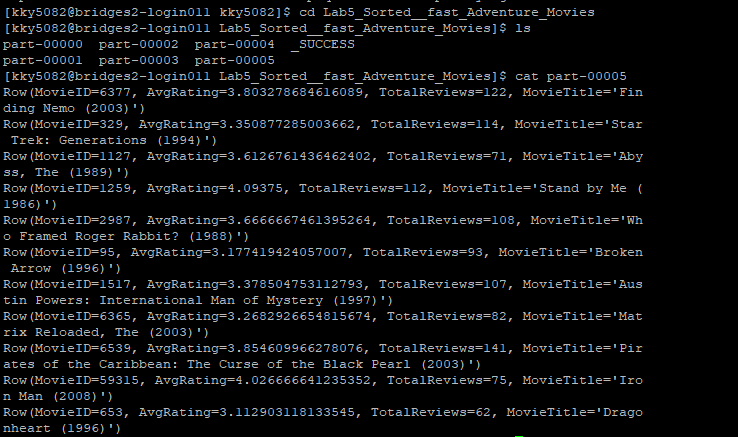
Screenshot of output



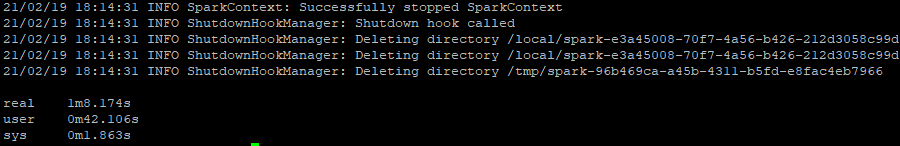
b). The small movie reviews and movies dataset, improved PySpark code



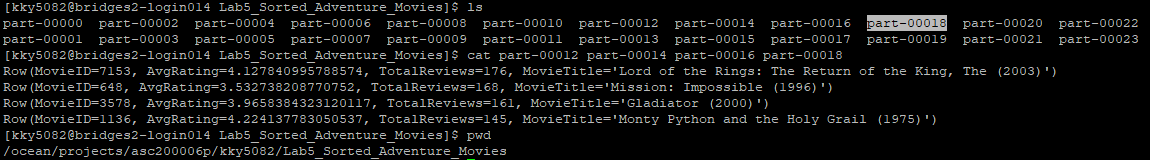
Screenshot of output



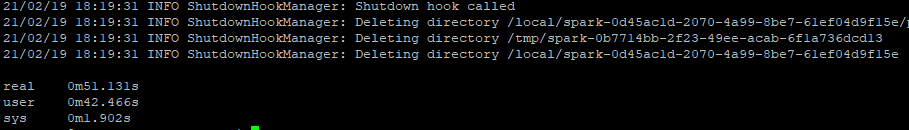
c). The large movie reviews and movies dataset, original PySpark code



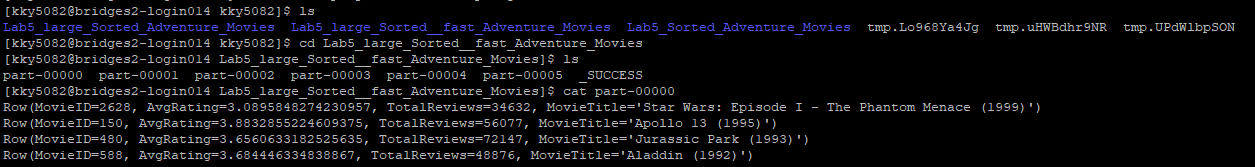
Screenshot of output



d). The large movie reviews and movies dataset, improved PySpark code



Screenshot of output



Discussion of my observation

Running time

A: 18.059 sec (real)

B: 19.289 sec (real)

C: 68.147 sec (real)

D: 51.131 sec (real)

The running time of A and B have very small difference, running time of C and D have comparable difference. The difference between A and B make me surprise, because the improved spark session run slower than original spark session, although there is just small difference. But the C and D have much lager difference and the improve spark session run much faster.

To my surprised, the running time difference between large and small dataset are not too large, because the large data size has very large size. So, it may because the feature of spark which running on cluster could save time.