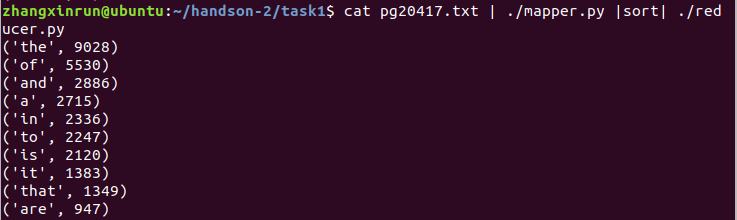
Name: Xinrun Zhang

Date: 15/10/2018

Task1

1. Test mapper.py and reducer.py:



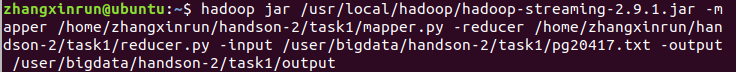
1. Make a new directory in Hadoop cluster:

/Users/zhangxinrun/Desktop/屏幕快照 2018-10-15 22.25.00.png

1. Copy the pg20417.txt file into Hadoop cluster:

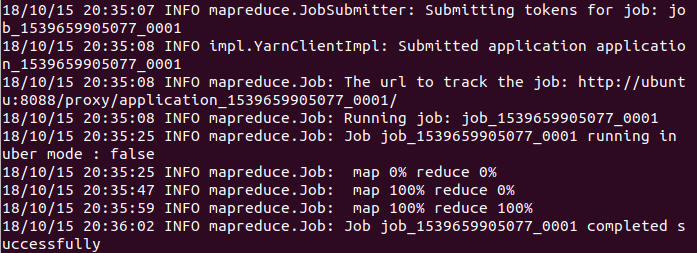
/Users/zhangxinrun/Desktop/屏幕快照 2018-10-15 22.29.07.png

1. Run the Hadoop job and save the output into a new folder:



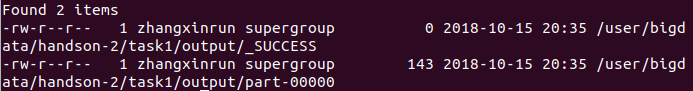
1. Result:

The Hadoop job is running

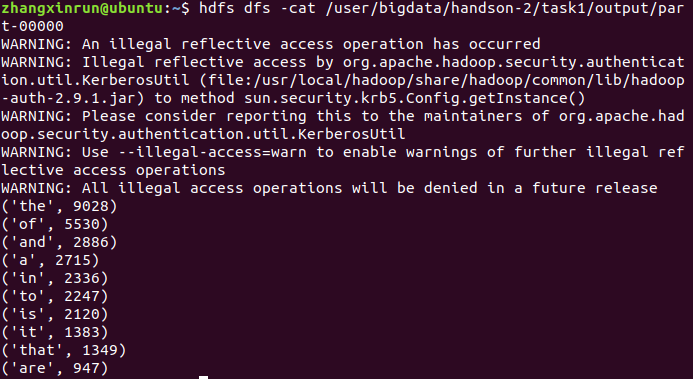


Take a look at Hadoop output fold and the outputs are showed:

/Users/zhangxinrun/Desktop/屏幕快照 2018-10-15 22.39.17.png



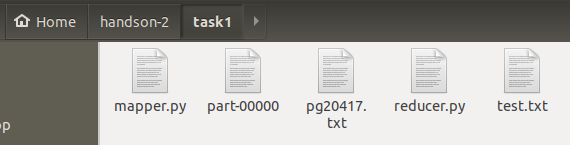
Use –cat command to execute output. The outputs are exactly same as test results:



Download outputs into local file system.

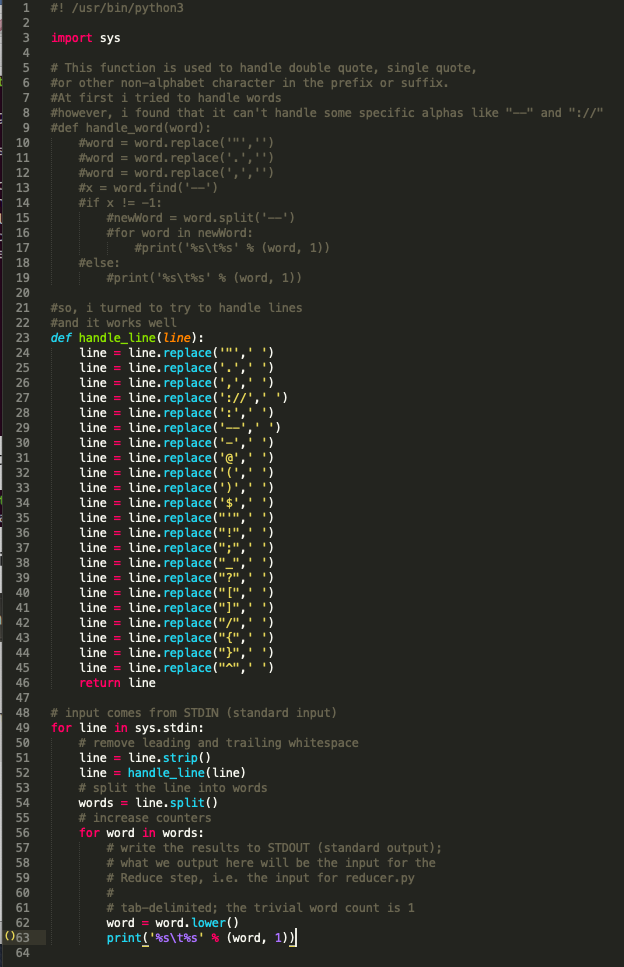
/Users/zhangxinrun/Desktop/屏幕快照 2018-10-15 22.41.13.png

the file in local file system, it’s a .txt file:



1. Code:

Mapper.py:



I defined a handle\_line() function to handle the alphas in each line. It replaced alphas with a blank. Then the words in each line were separated naturally and clearly.

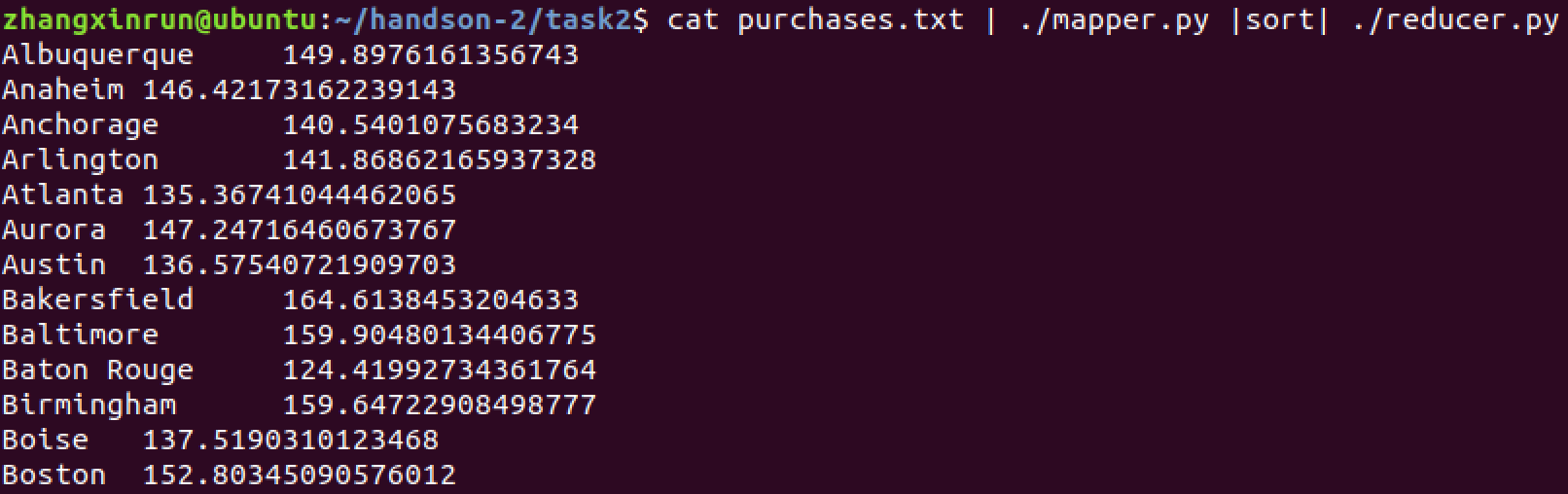
Reducer.py:



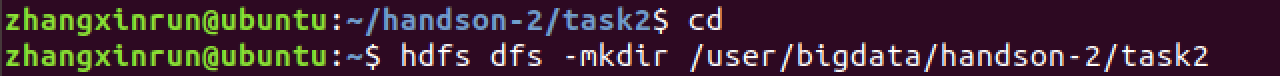
To display top 10 frequency words, I saved the outputs of reducer in a dictionary newDict{} and sorted it. After that, I displayed the first ten keys and values of dictionary.

Task2

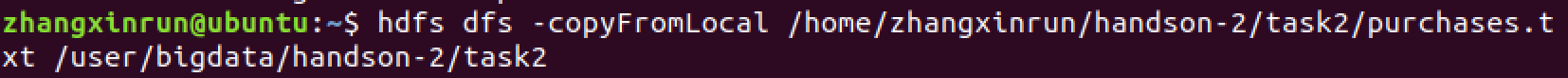
1. Test mapper.py and reducer.py



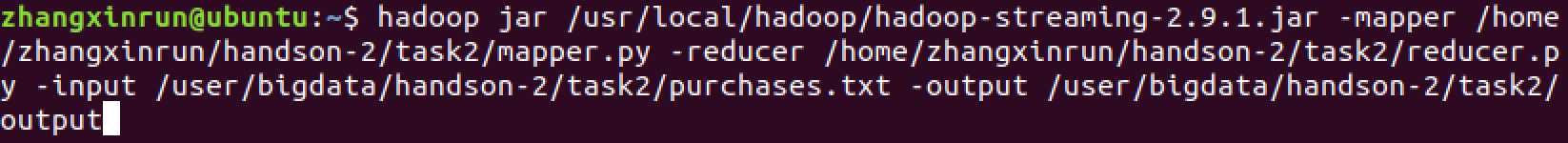
1. Make a new directory in Hadoop cluster:



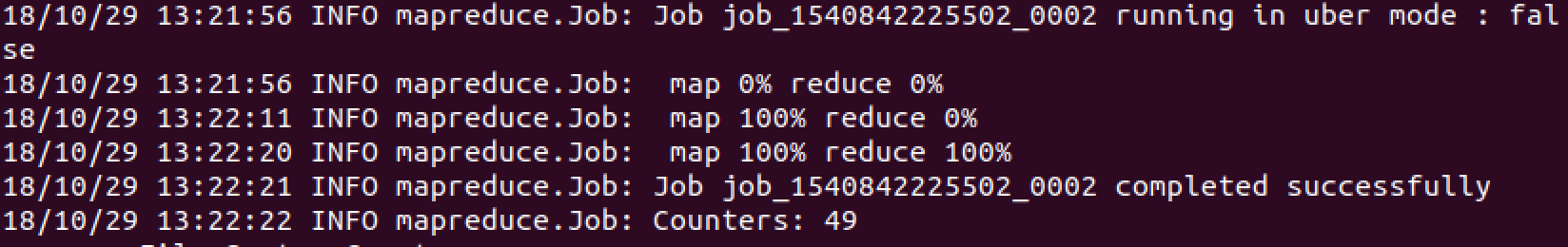
1. Copy the purchases.txt into Hadoop cluster:

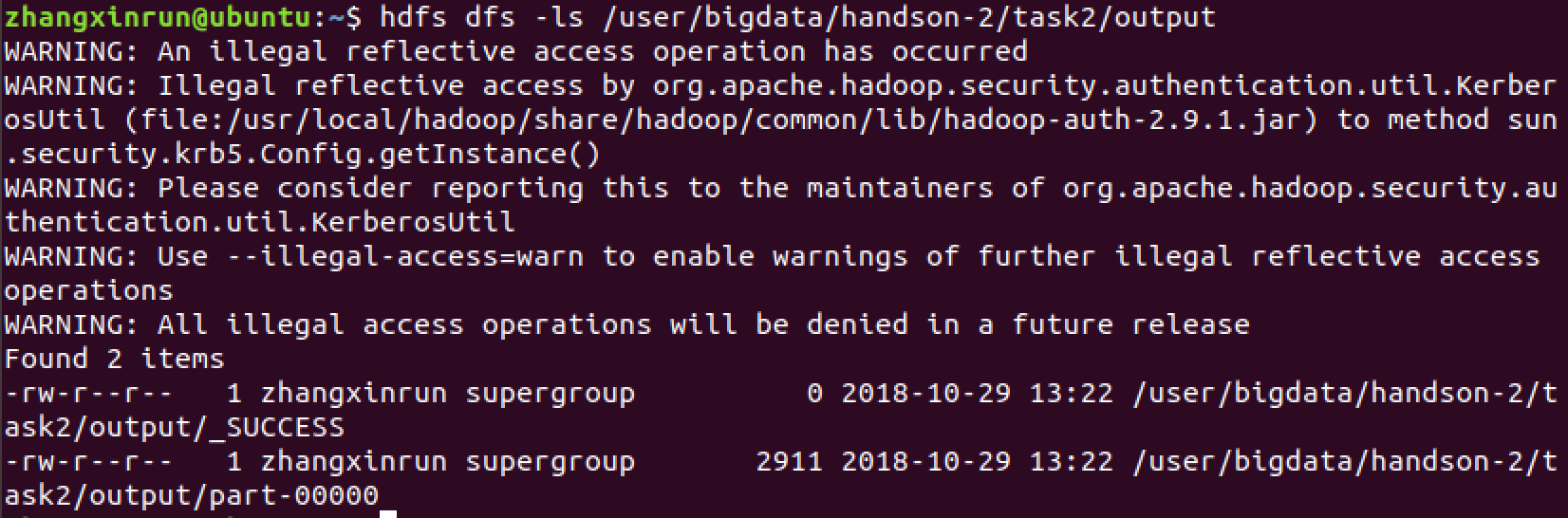


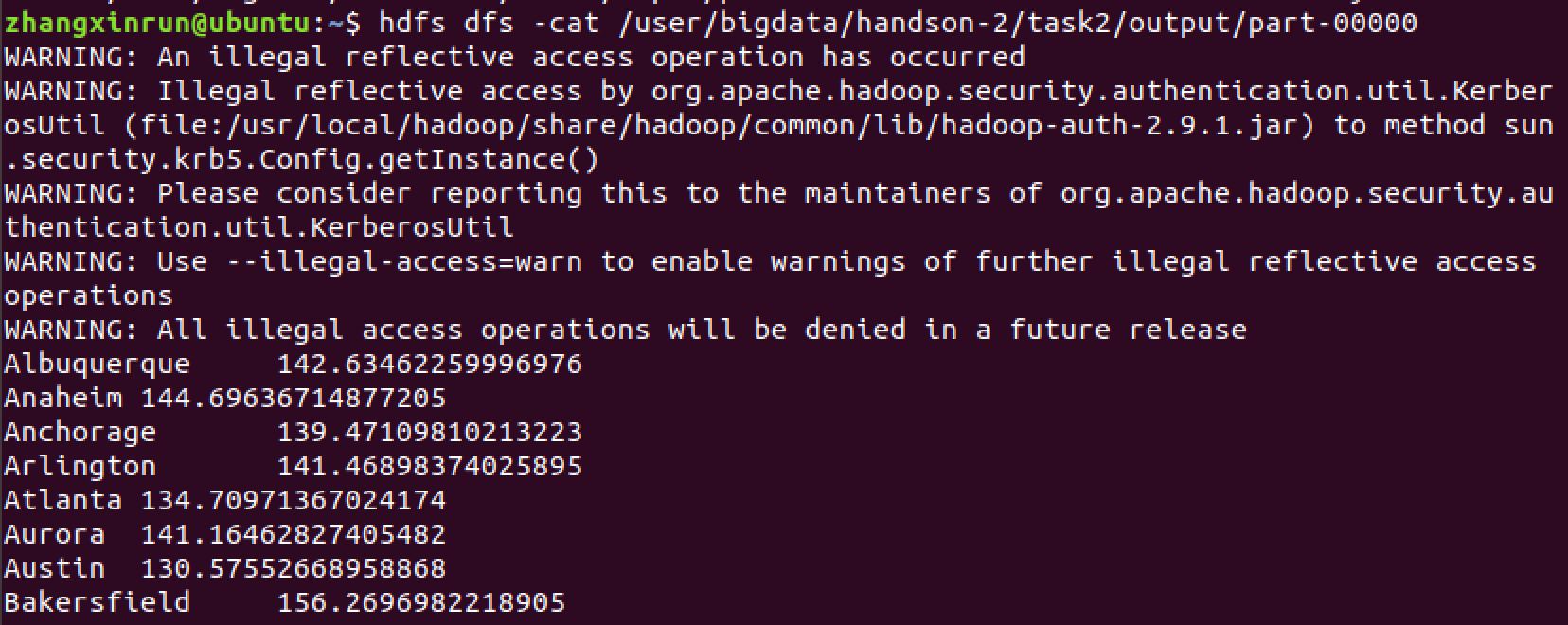
1. Run the Hadoop job and save the output into a new folder:



1. Result:

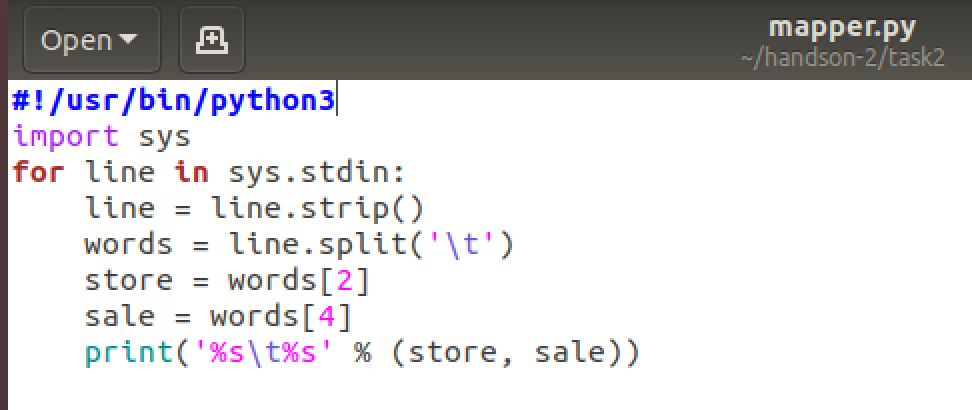




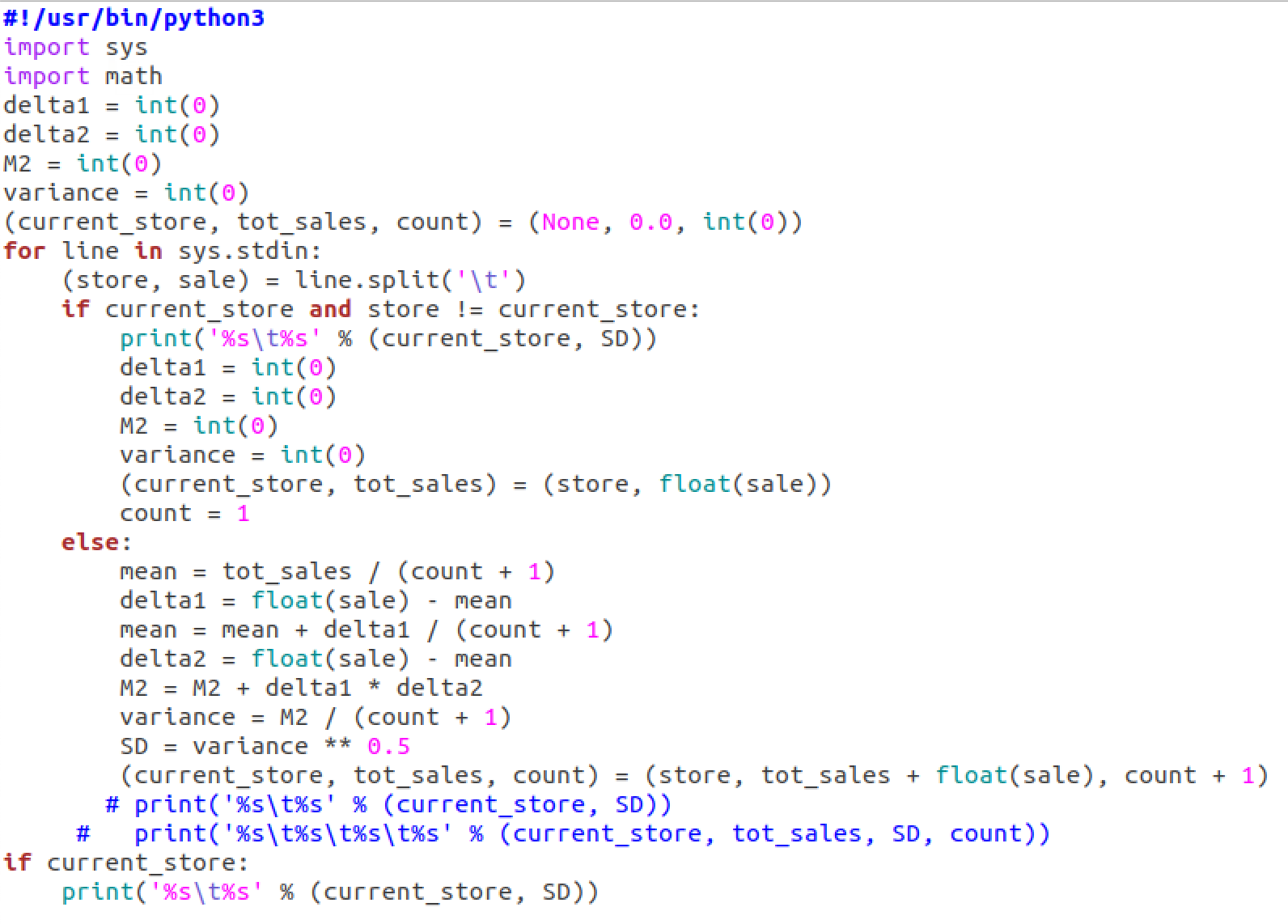


1. Code:

mapper.py:

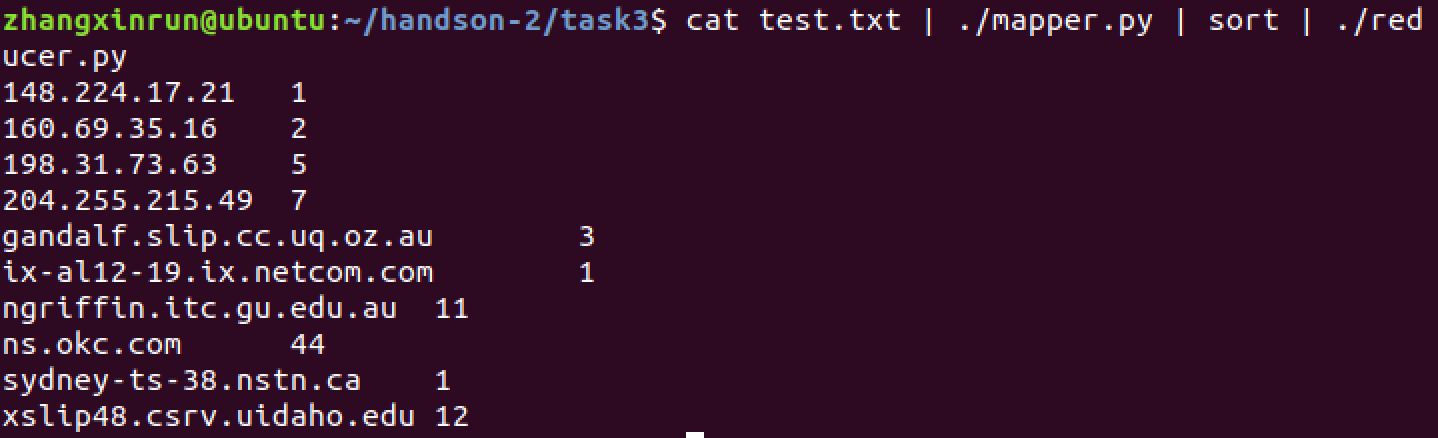


reducer.py:



Task3

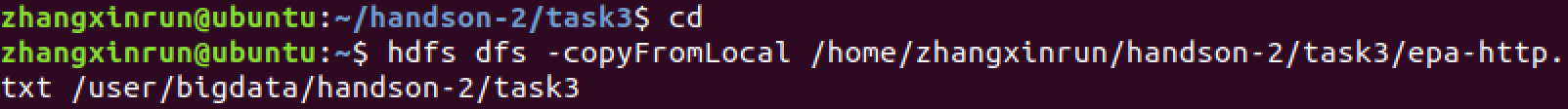
1. Test mapper.py and reducer.py



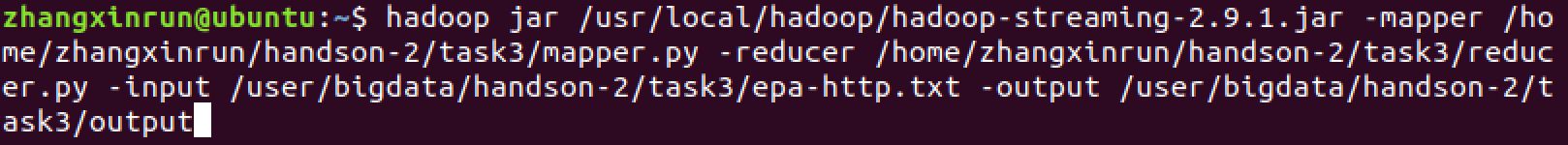
1. Make a new directory in Hadoop cluster



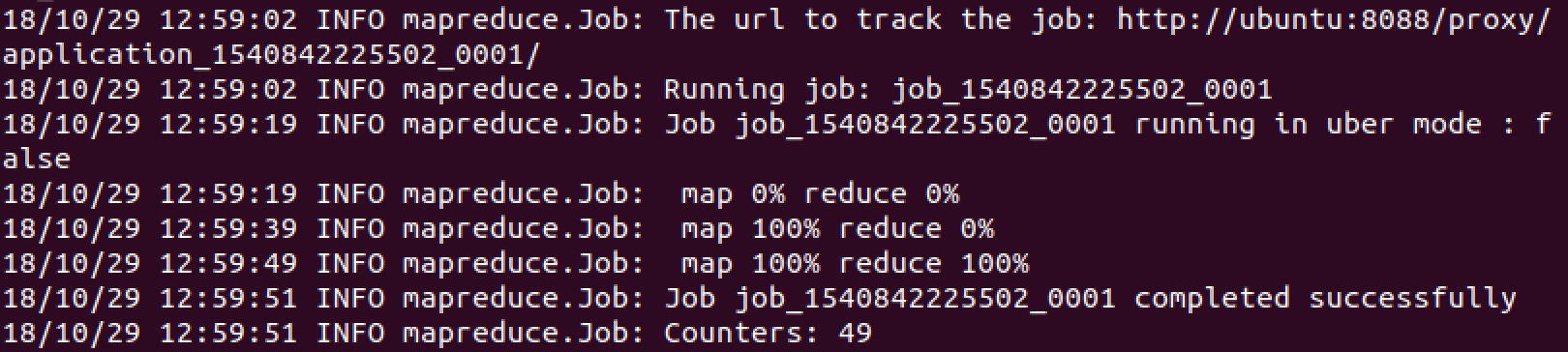
1. Copy the epa-http.txt into the directory:

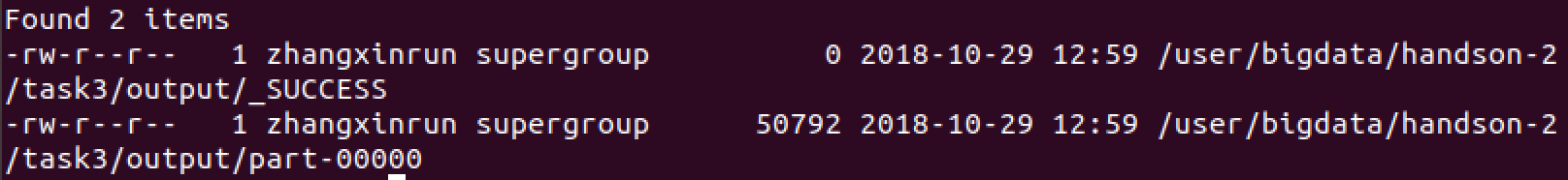


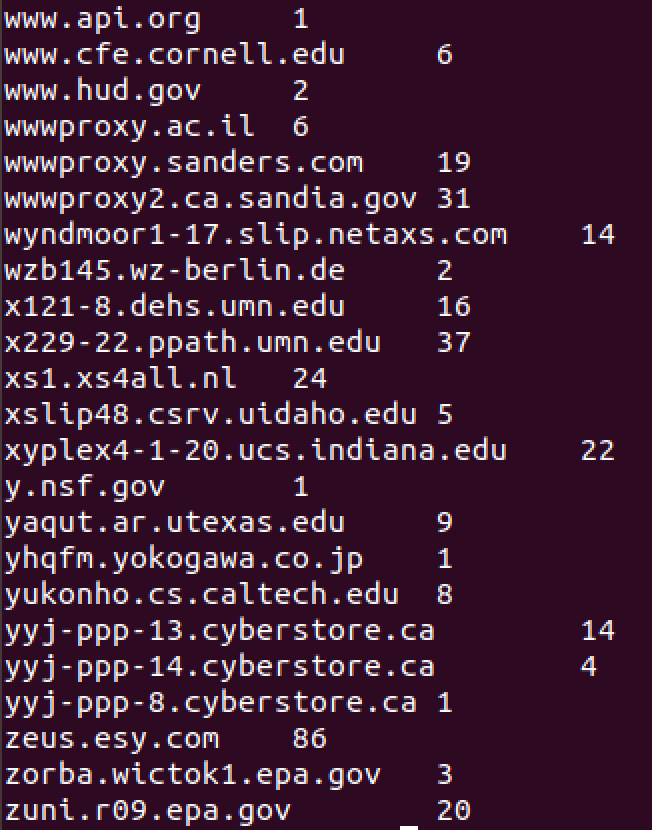
1. Run the Hadoop job and save the output into a new folder:



1. Result:

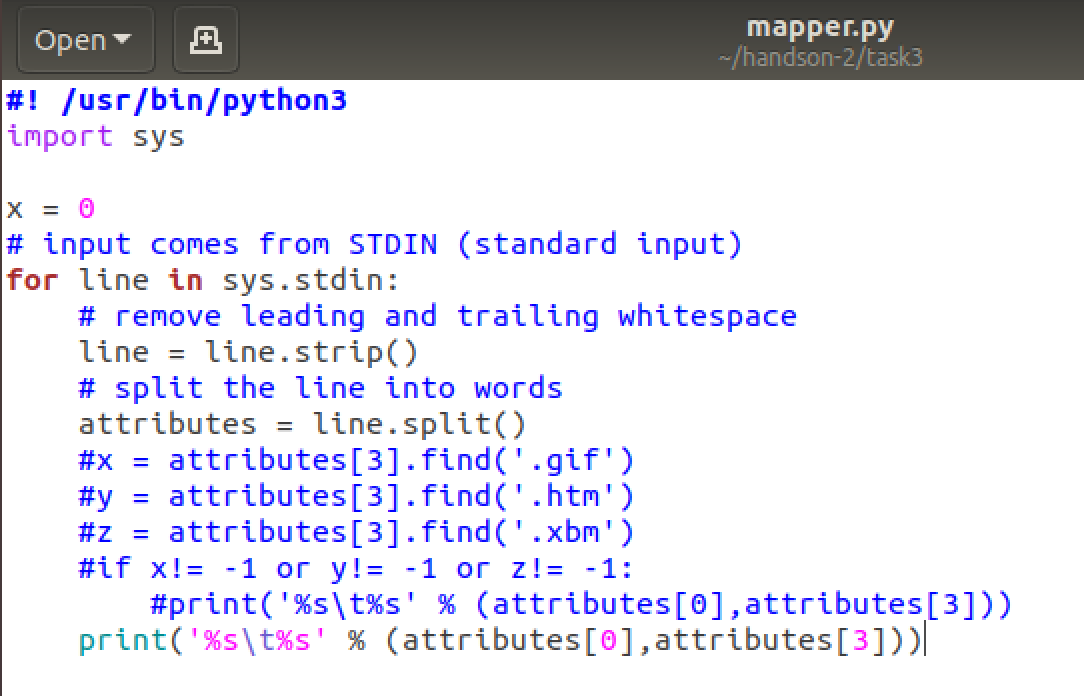






1. Code:

mapper.py:



reducer.py:

