**Big Data Mining and Applications**

Homework3

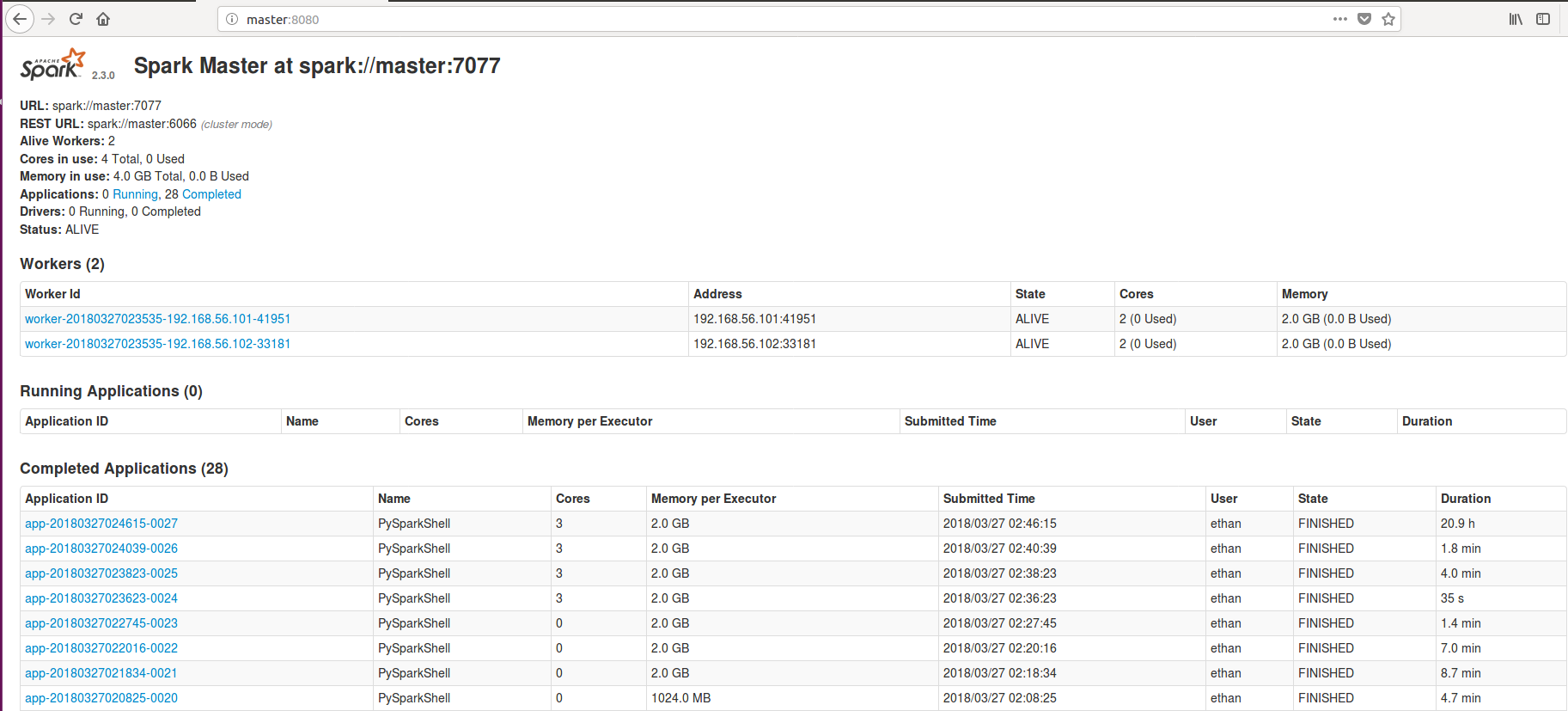
106598005 盧家馨

2018/05/04

* My cluster environment setup:

|  |  |  |
| --- | --- | --- |
| PC | Memory | Ip address |
| Master | 2G | 192.168.56.100 |
| Node1 | 2G | 192.168.56.101 |
| Node2 | 2G | 192.168.56.102 |

* Connection Status



* Source code link

<https://github.com/Jessieluu/Spark2018/tree/master/hw3>

* Output link

<https://drive.google.com/drive/folders/1rS-mA7p2B_-eQrIPIip6cya7yKUvJ-WW?usp=sharing>

Because the file is too large to compute, so I choose 5 document to compute and generate the matrix with (shingle \* document(5) )