**Assignment-5: Running Page Rank code on a multi node cluster**

1. **Setting up a multi node cluster:**

I created three virtual machines, One master and two slaves.

I edited /etc/hosts and /etc/hostname files to customize my hostnames as “master, slave1, slave2” respectively.

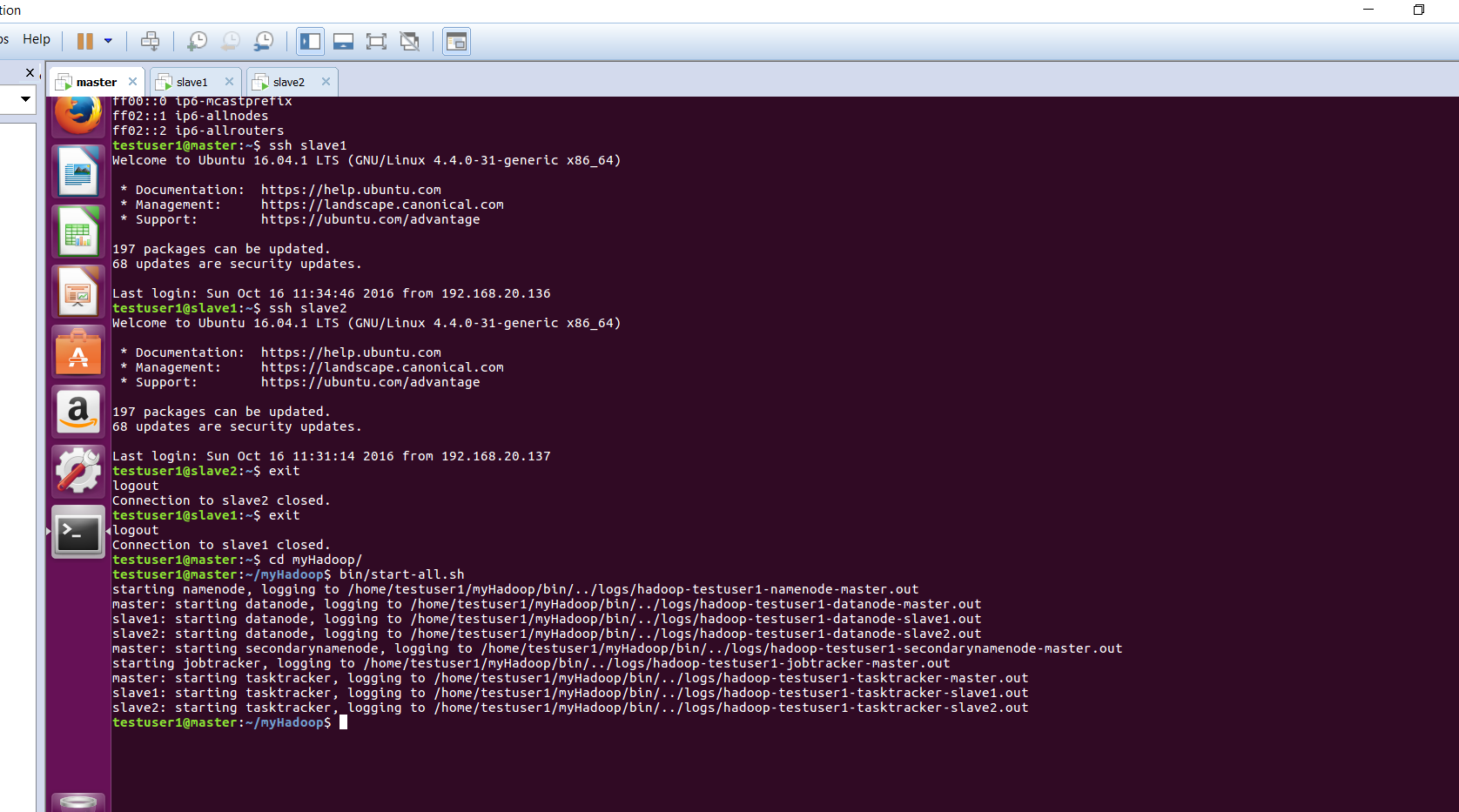
Then, I installed Java 8 on all the VMs’ followed by Hadoop installations.

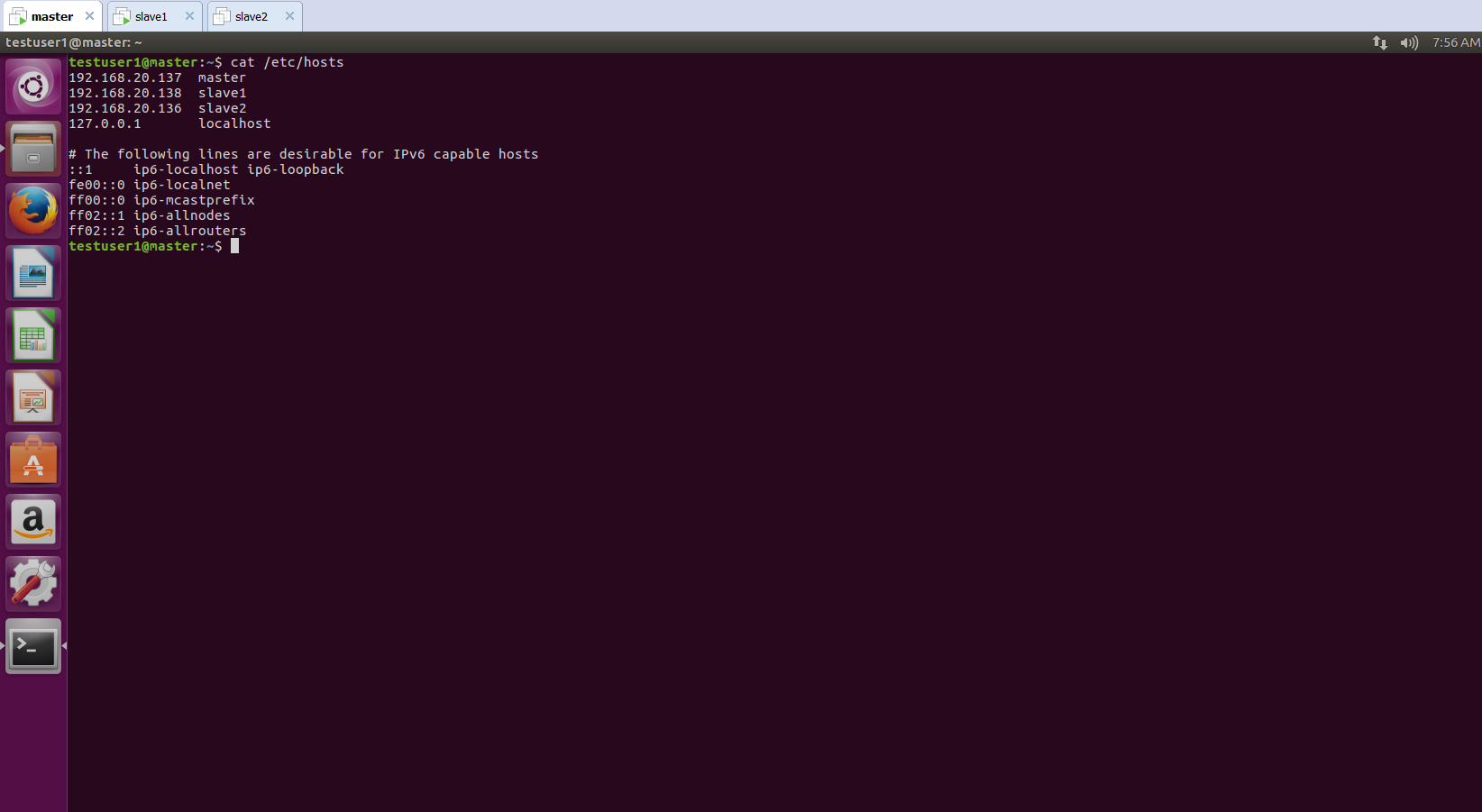
After installing Hadoop, I configured master slave communication, by overwriting the hosts files on both the files mentioned above.

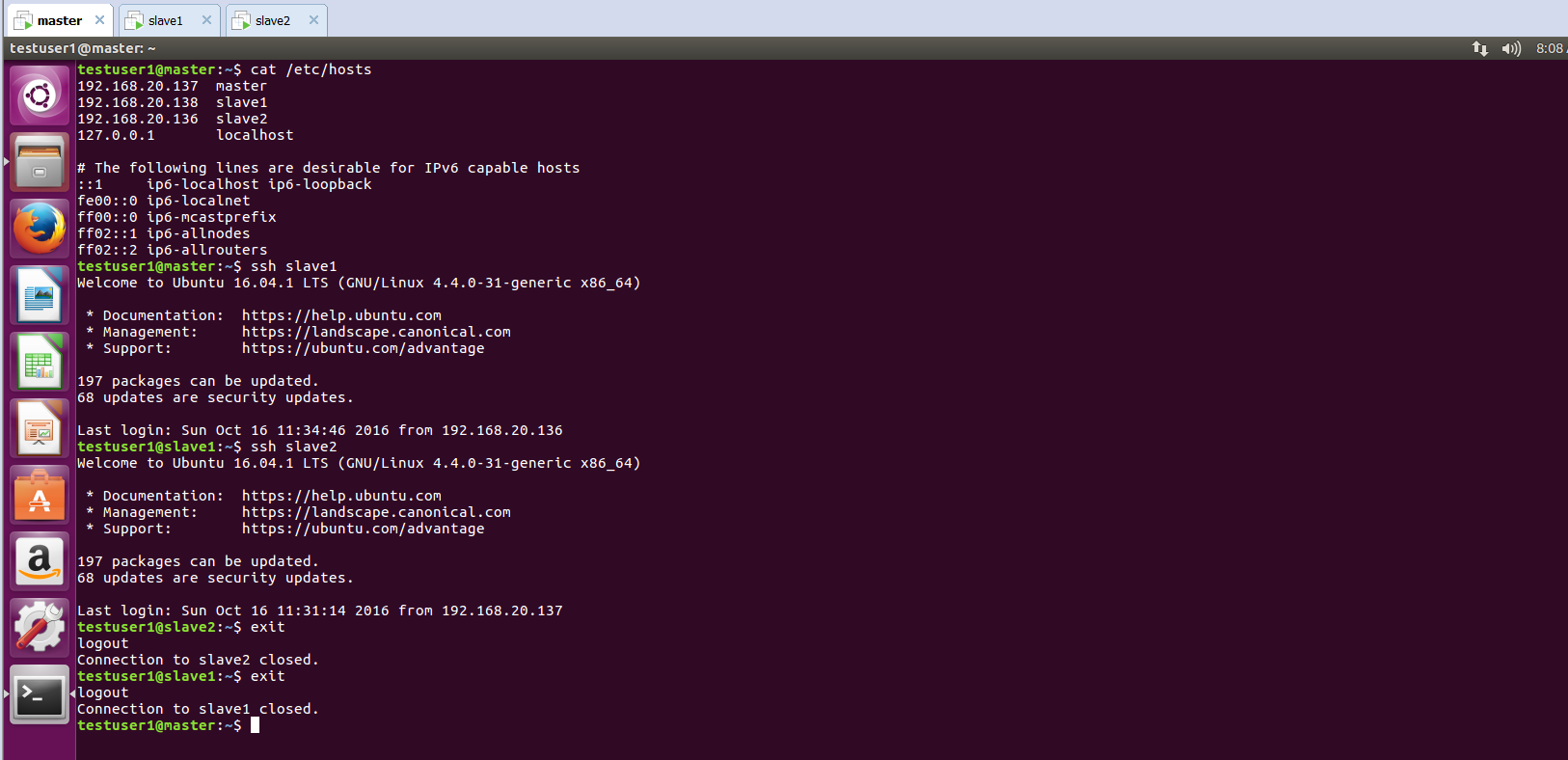
I made configuration changes (replace localhost as master) on mapred-site.xml and core-site.xml.

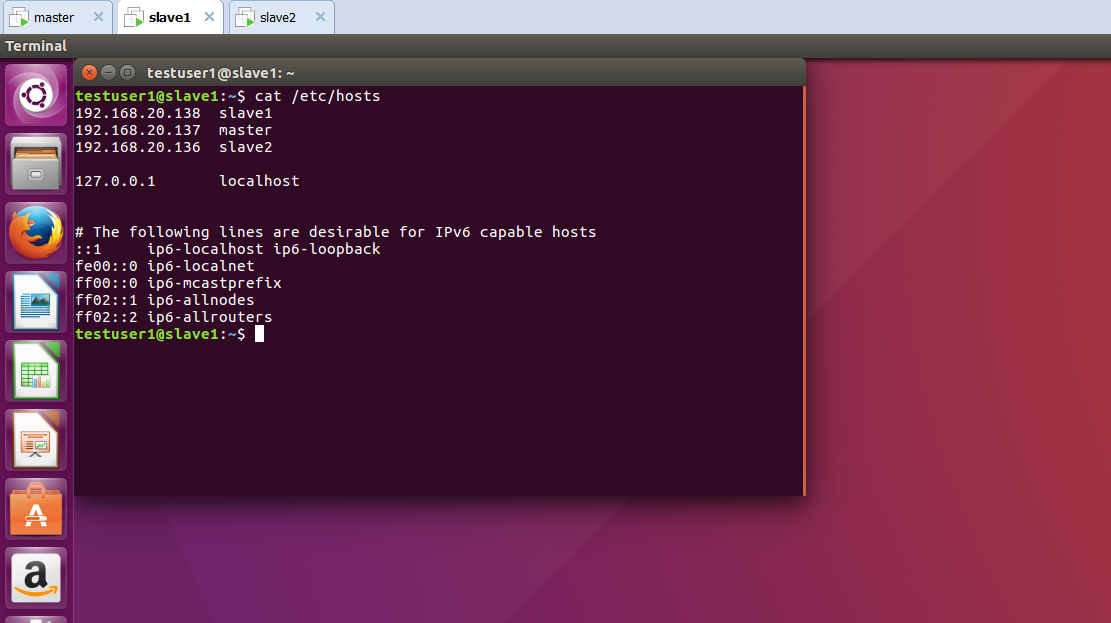
I confirmed that my Hadoop cluster was up and running by checking the browser results at master:50070.

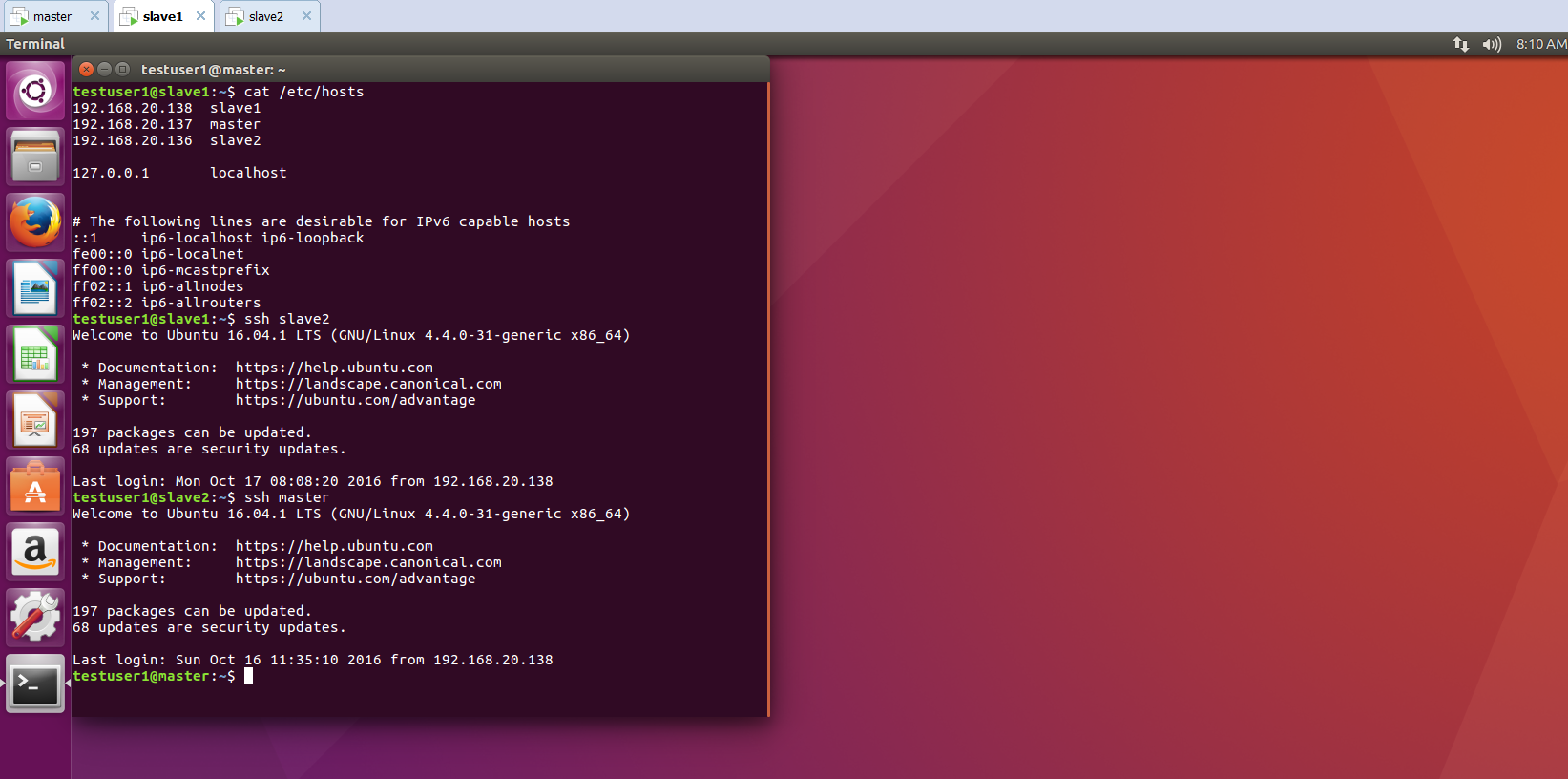
Screenshots below:

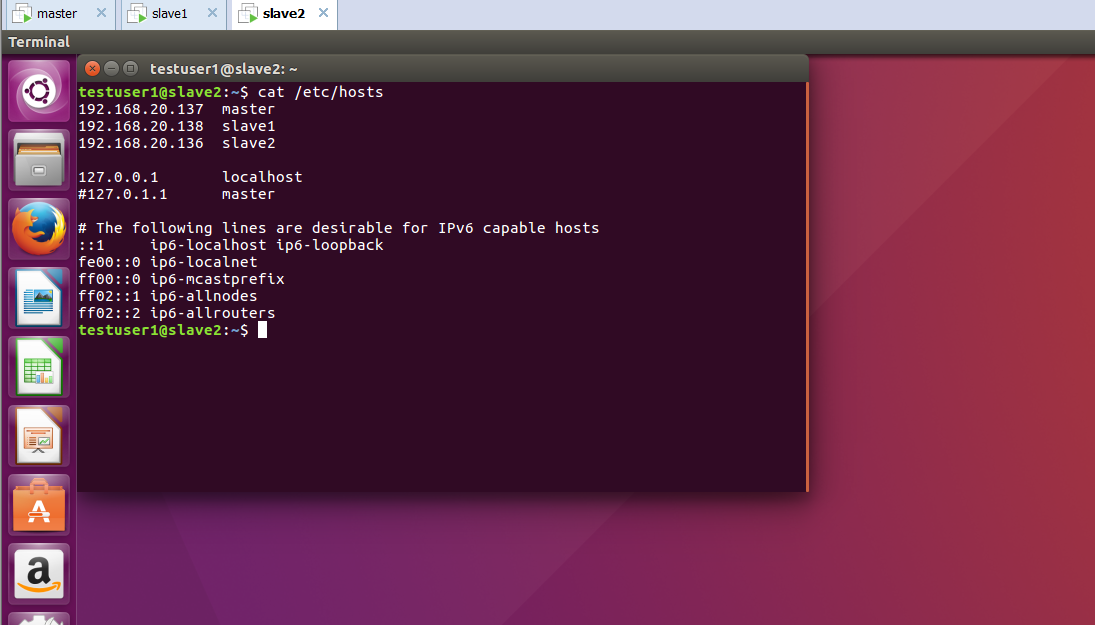
Hadoop started at master:

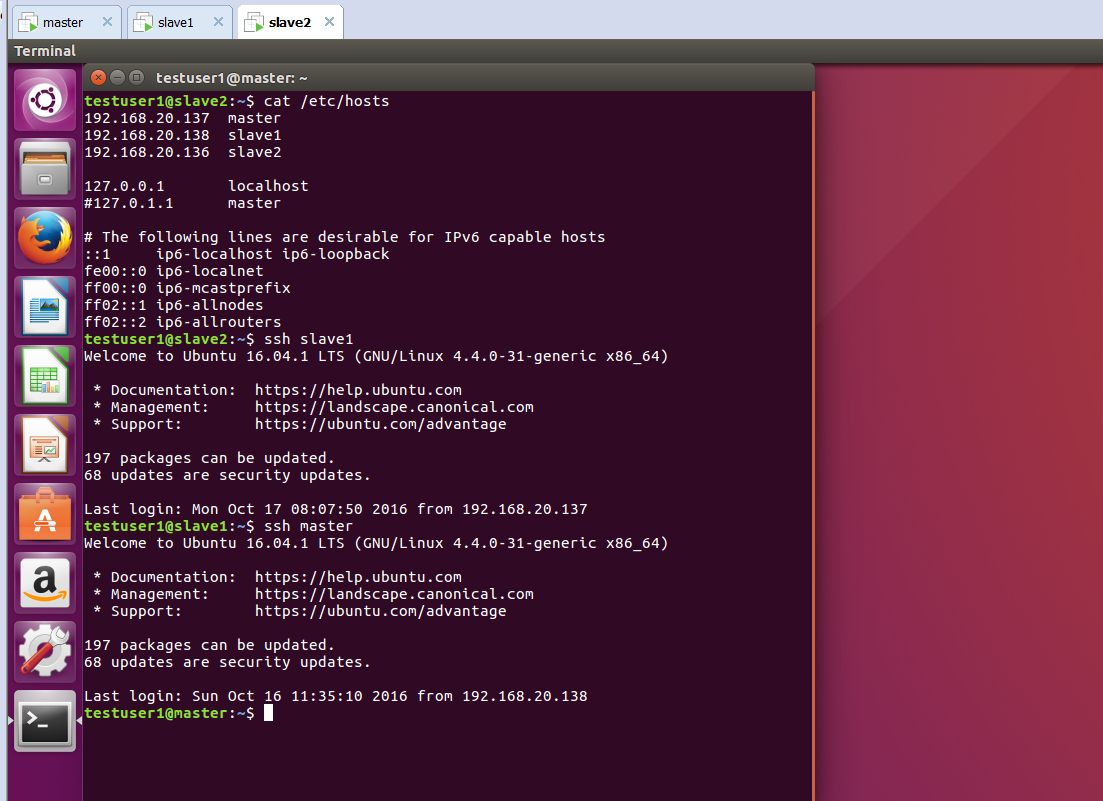
Master hosts file :

Ssh master outputs :

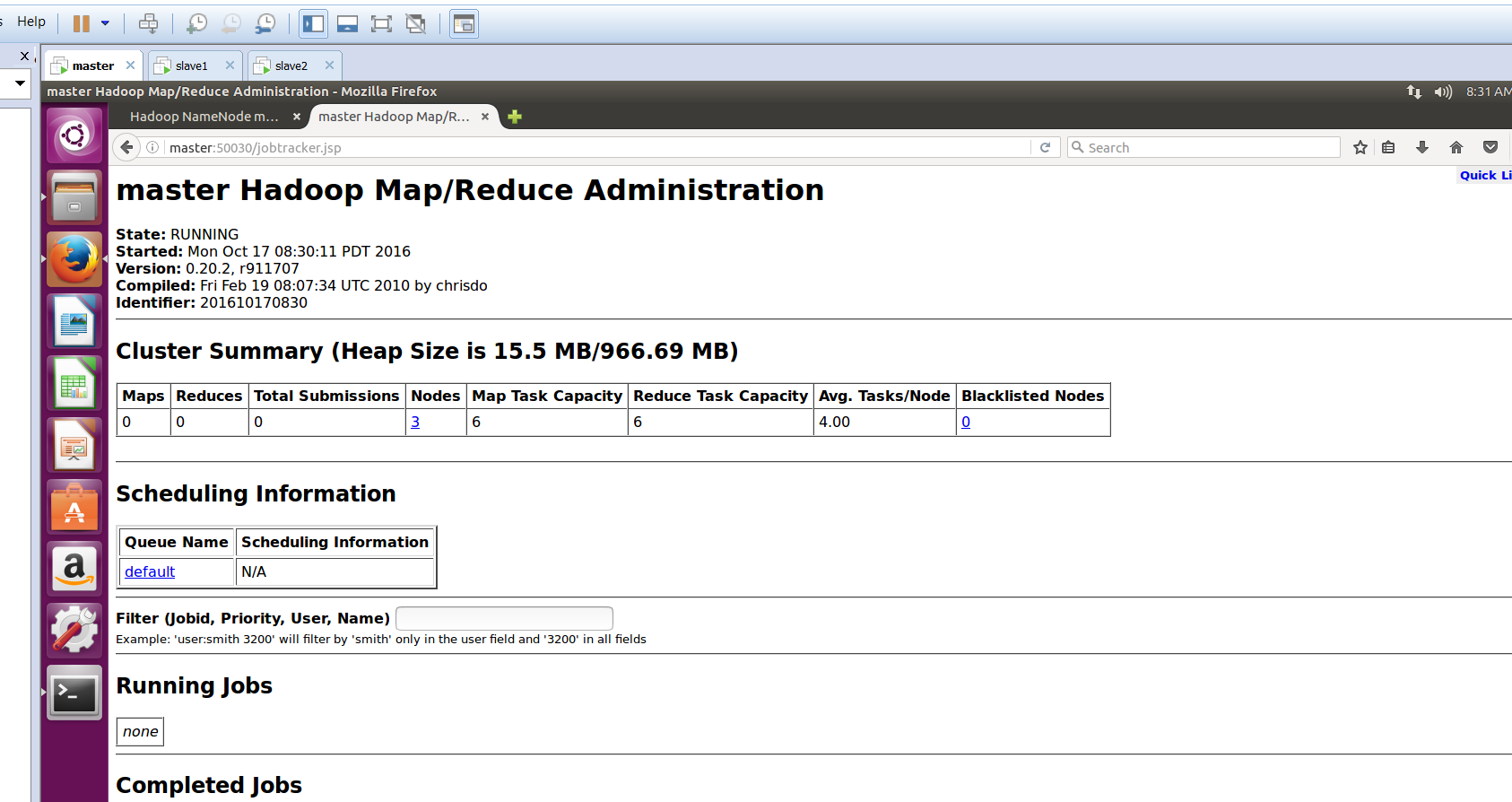
slave-1 hosts :

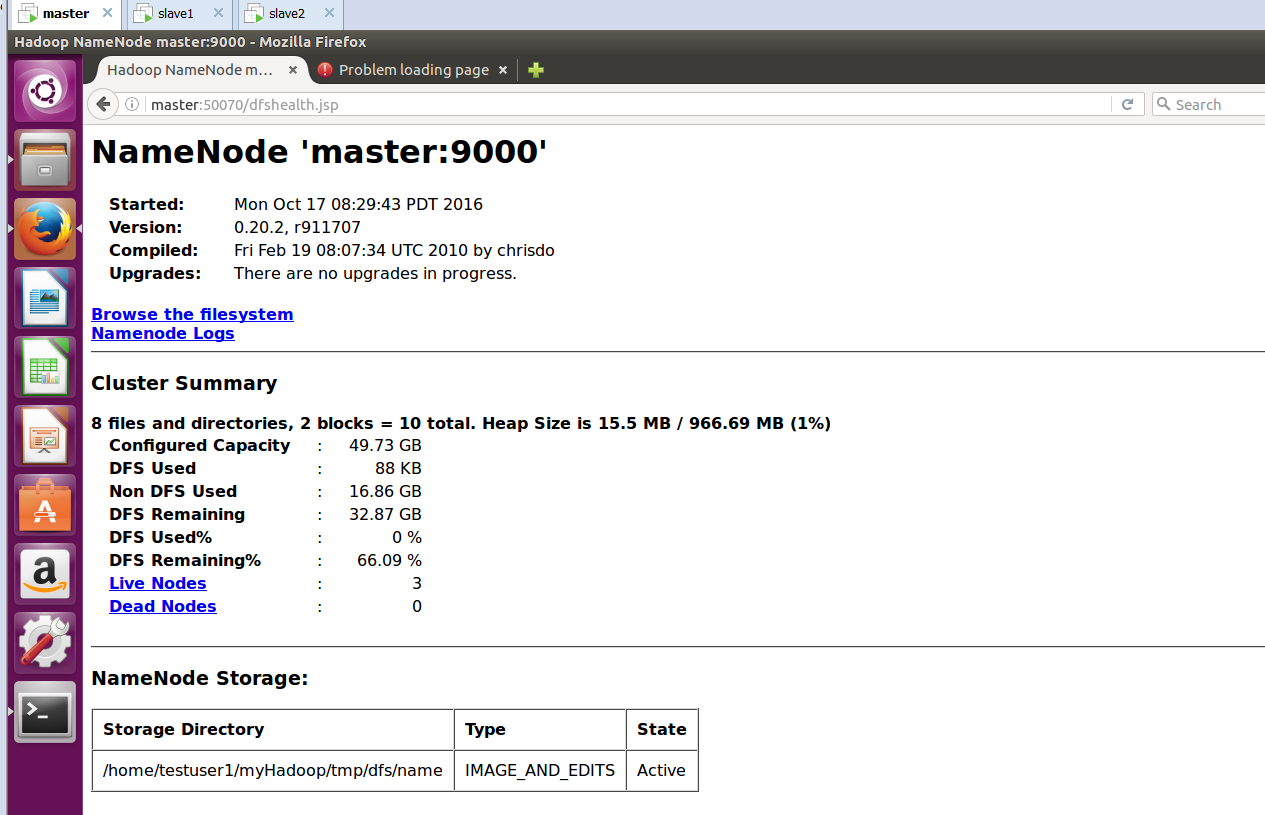
ssh at slave-1:

slave -2 - hosts:

ssh at slave 2:

Job tracker at master:



Browse master:

**2.PageRank Code:**

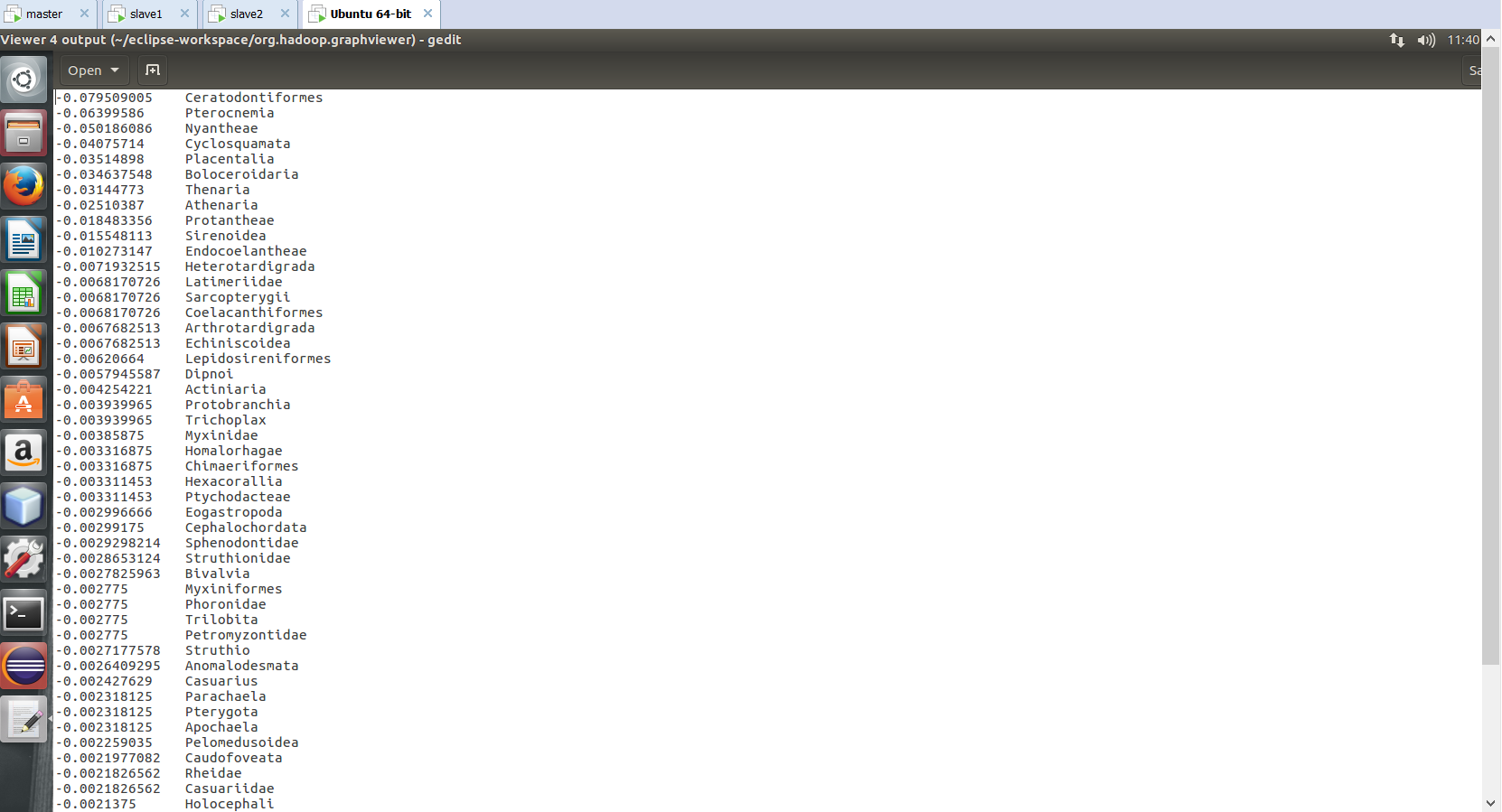
Following are the code changes I have performed to arrive at the results.

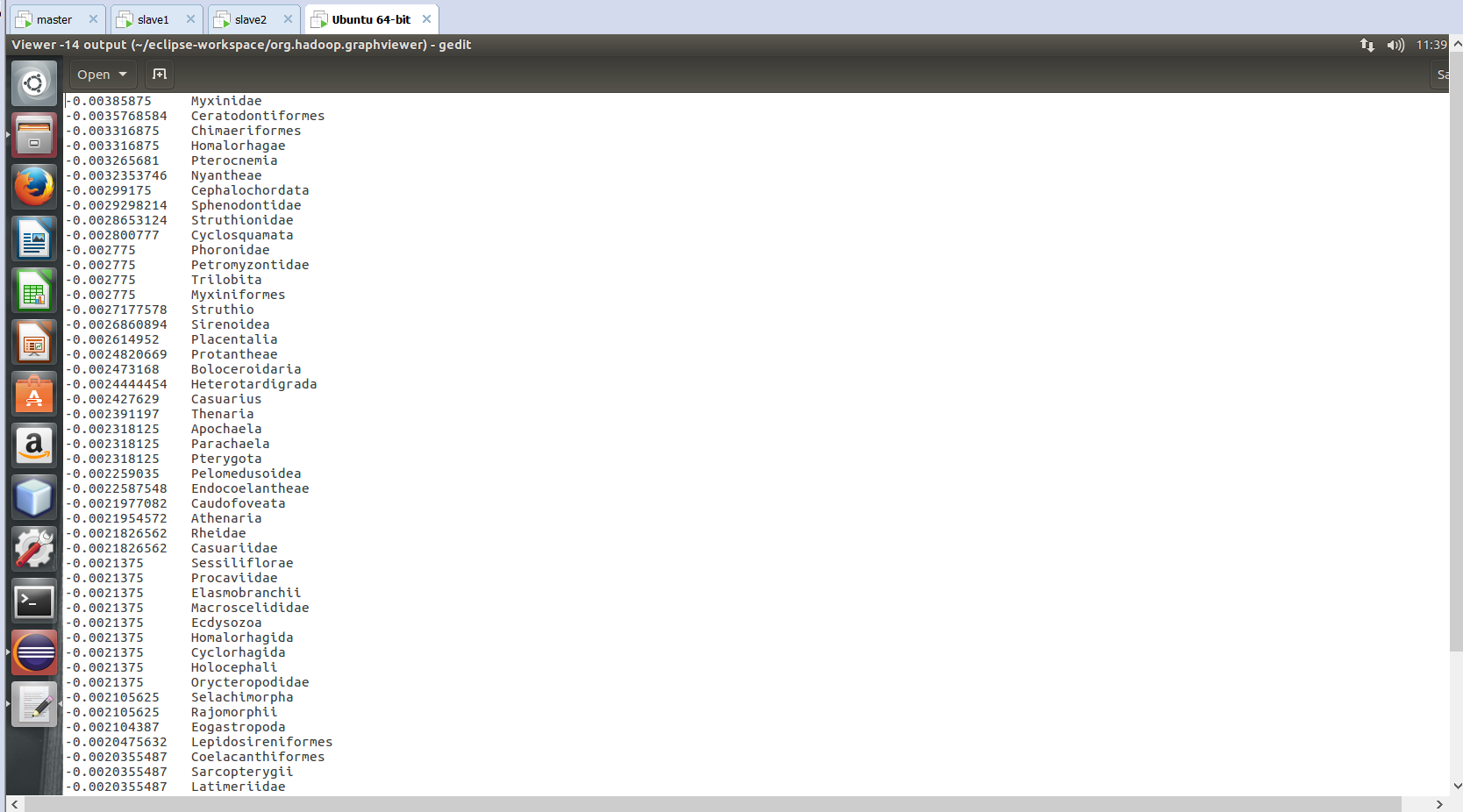
Steps taken:

1. Cleaned a lot of junk value on builder reducer.
2. Changed the damping factor difference in the emit value of iterator reducer to 0.0015 for a factor of 0.85.

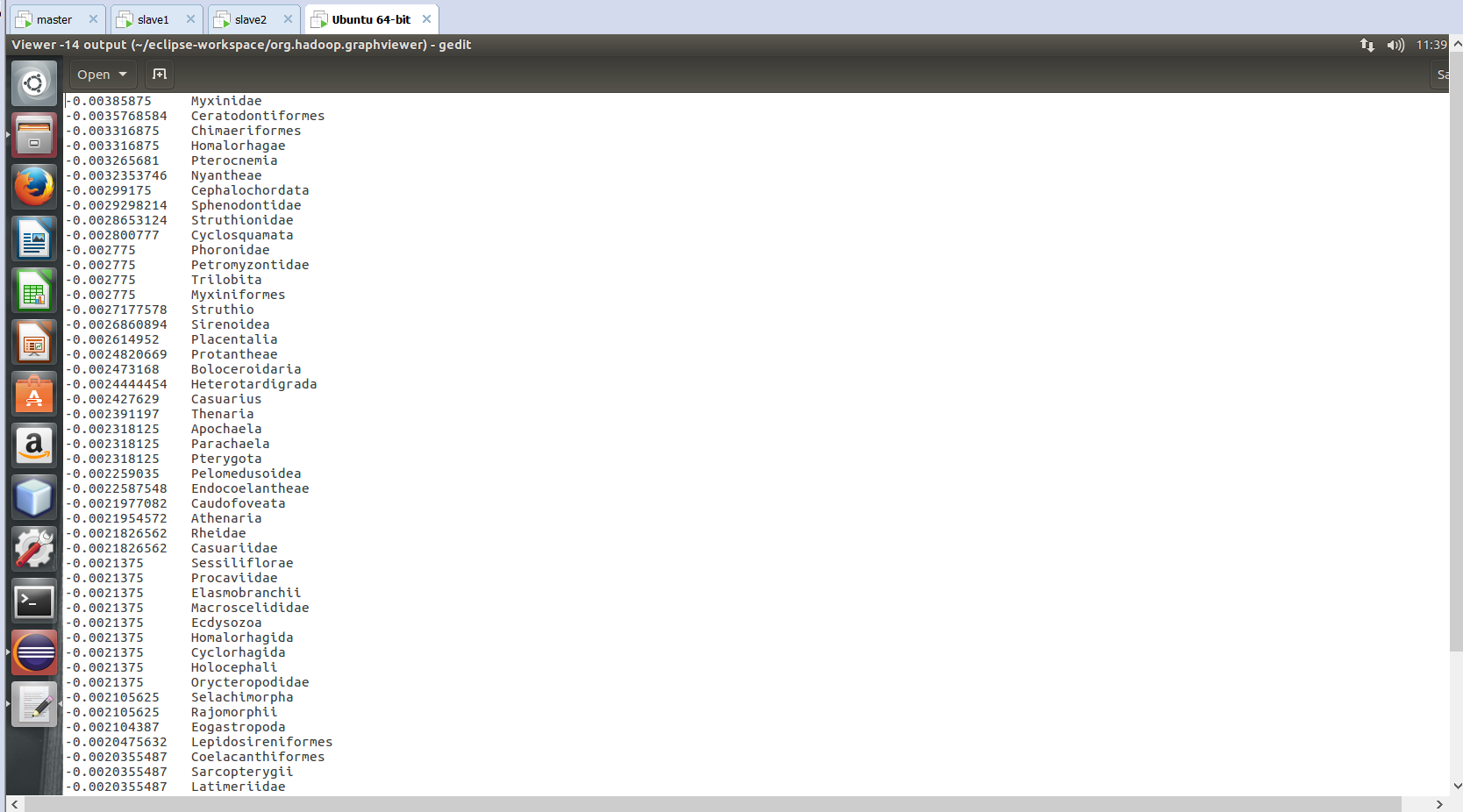
From the code results,

1. The range is converging.
2. Totally 254 unique species in the result.
3. The convergence is seen after 10 iterations approximately, for a factor of 0.85.

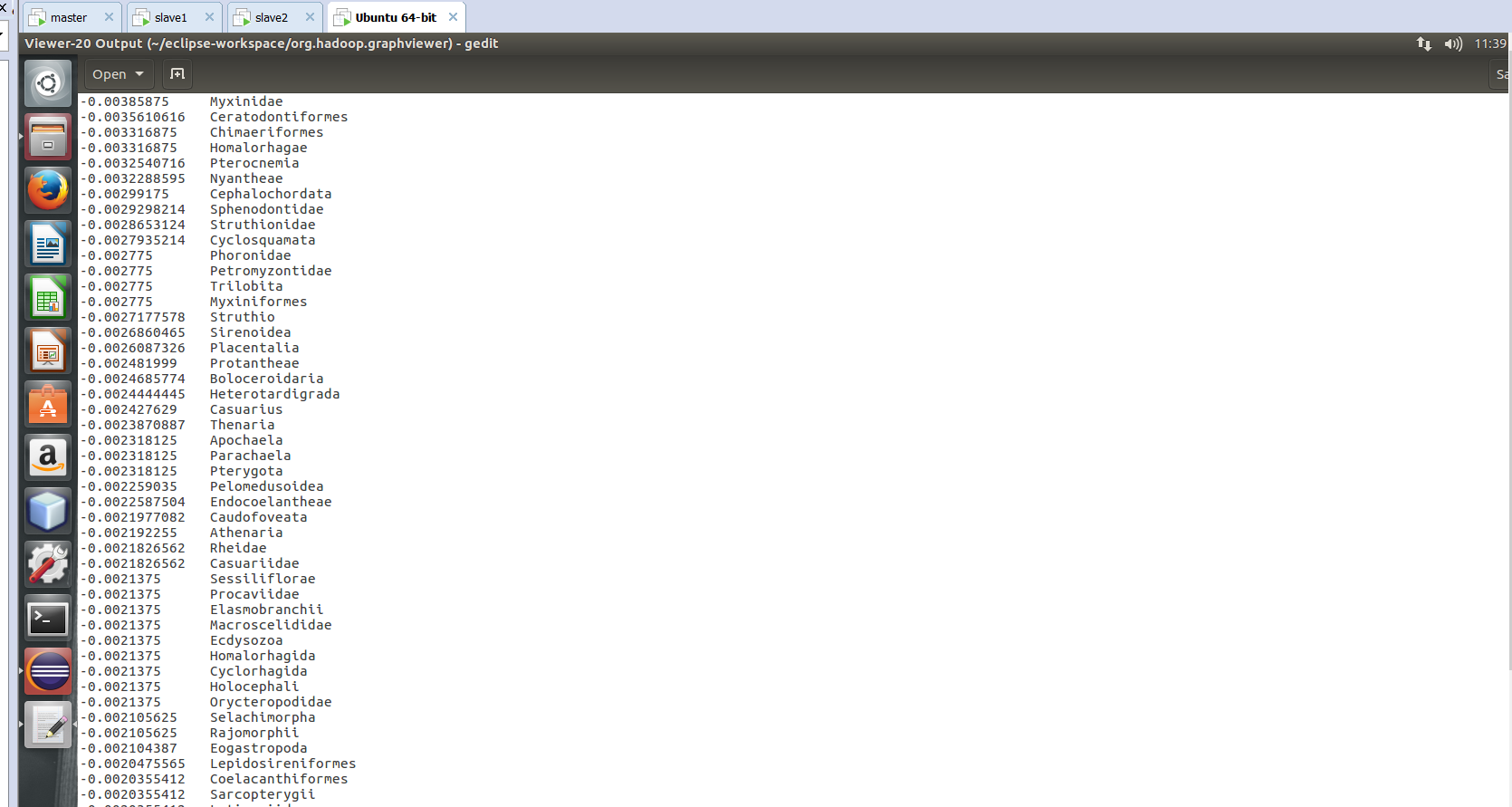
Output-4:

Output -14:

Output-20:



Output 14:



Data cleansing:

