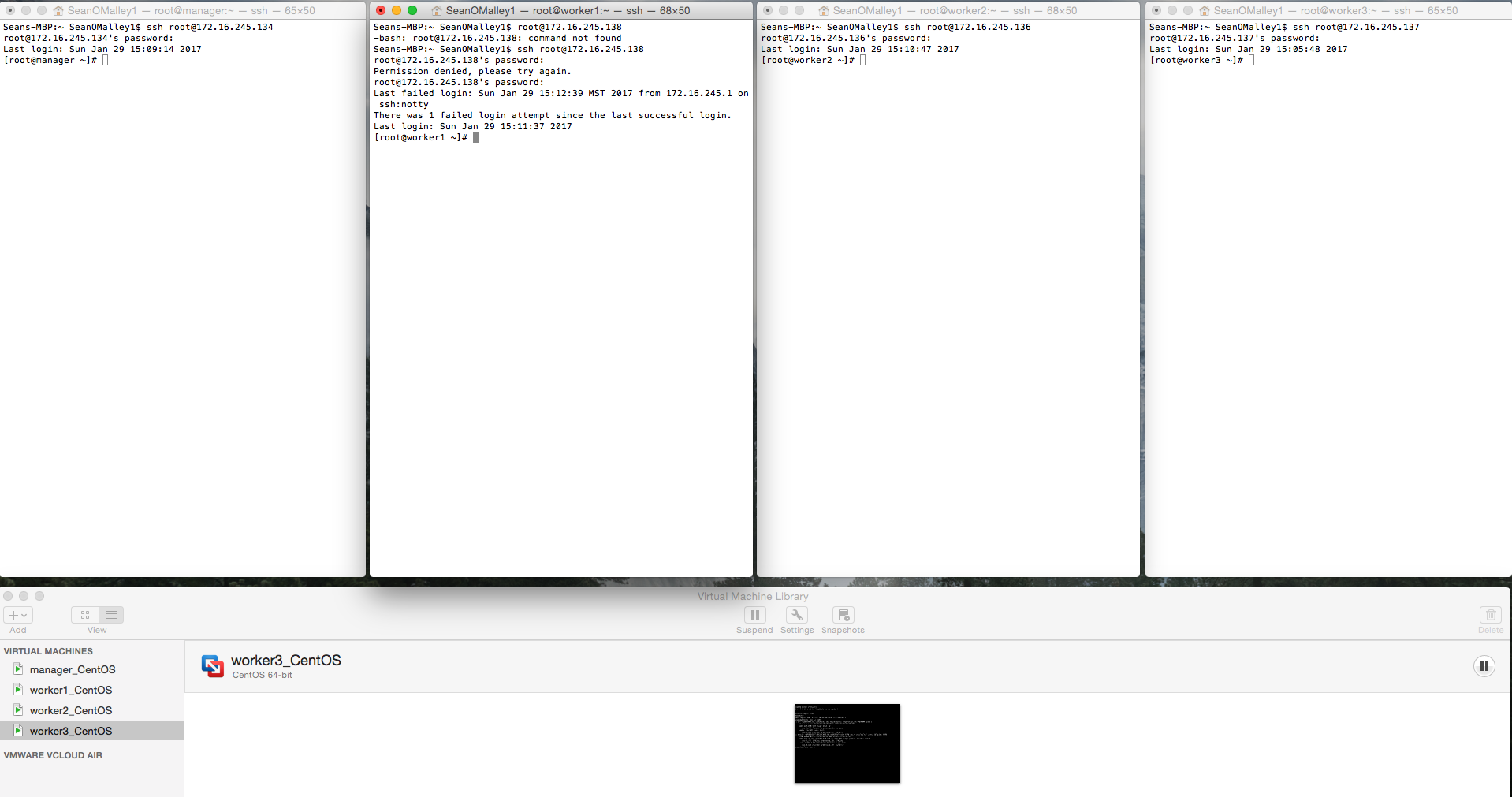
Data Engineering

Week 2 Assn

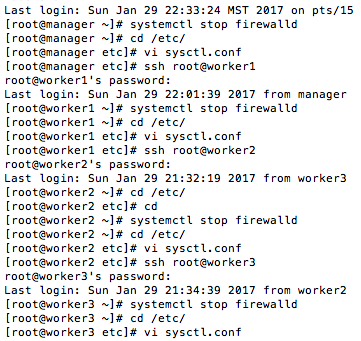
Sean O’Malley

Install Hadoop on a cluster.

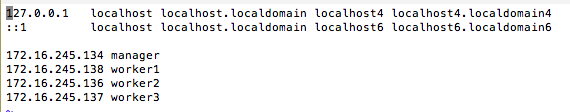
I first created 4 virtual machines, I created a manager and 3 worker nodes, named them accordingly and also added a passwordless Hadoop user on each cluster.



Next, I disabled the firewalls on all nodes and disabled the IPv6 in the sysctl on all nodes



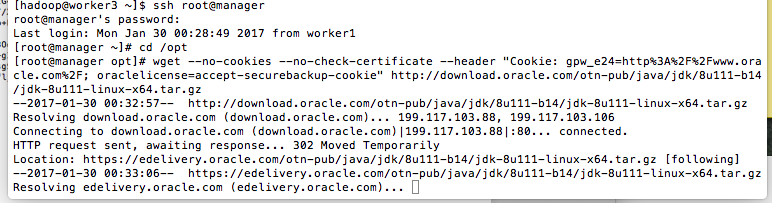
I then added the IP addresses and node names to the /etc/hosts file



I created the password-less Hadoop user in each node correctly, but ran into trouble with the authorized keys and id\_rsa progression. After some searching on the internet, I did my best to mimic the functionality of the ssh-copy-id function that is roughly outlined in the instructions. The explicitly entered authorization keys using vi function on each node are listed below.



After creating each authorization key for each node, then giving 0600 permissions to each Hadoop node interaction, I then downloaded the java tar.gz onto my root manager user.

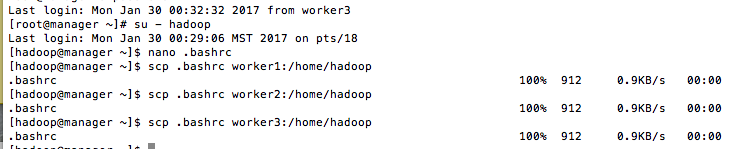


I then unpacked the tar file, changed file owner to root then copied the java to all the worker root nodes. The output of which is a little too verbose for a screenshot.

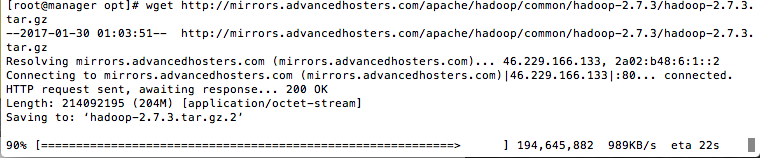
I then installed the java alternatives on each root node, as instructed.

../Desktop/Screen%20Shot%202017-01-30%20at%2011.05.06%20PM.png

Next, I switched to the Hadoop user, inserted the jdk into the .bashrc file then copied that Hadoop user bashrc file to the worker nodes



I then installed Hadoop on the root@manager account then used the scp –r function to put a copy of Hadoop 2.7.3 in the /opt folder of each root worker node



Next, I made the directories for Hadoop and gave access to Hadoop users for name and data nodes.

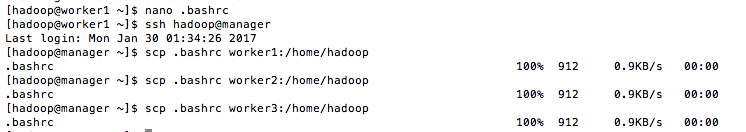
../Desktop/Screen%20Shot%202017-01-30%20at%2011.36.26%20PM.png

../Desktop/Screen%20Shot%202017-01-30%20at%2011.36.42%20PM.png

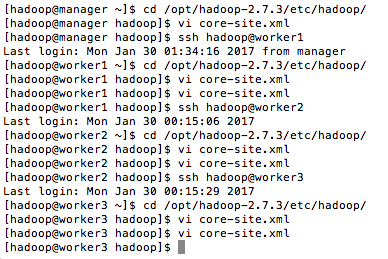
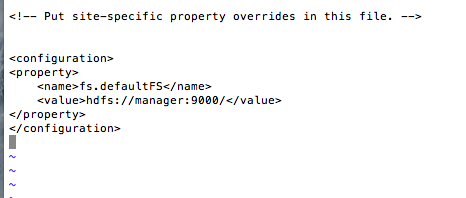
I then configured the following manager hadoop user .bashrc file using the nano function, giving the below output in the bashrc file.



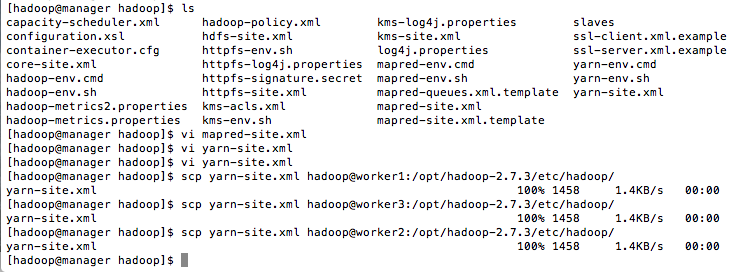
Then passed that information to the Hadoop worker nodes using the scp code:



After this, I copied the name node code into the core-site.xml of each node, also highlighting my seem-less transition between Hadoop clusters, of which took me over two days to finally figure out my bug, yikes.



I then added the necessary changes to the mapred and yarn xml’s then pushed the yarn xml changes to the individual data nodes from the manager data node. Afterwards adding the vm names to the slaves file.



I then successfully started up both dfs.sh and yarn.sh.

