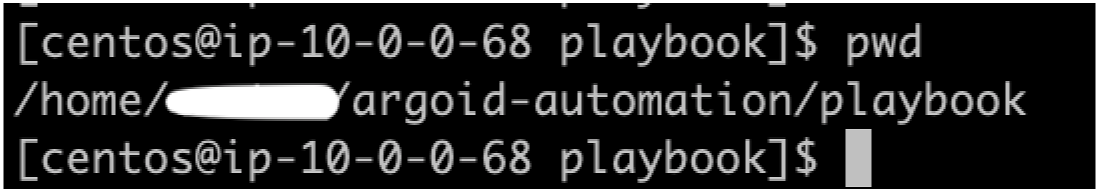
HDFS Cluster setup

Prerequisites:

- Java-8 needs to be installed
- Zookeeper Setup - Zookeeper setup
- Check with leads, whether to host the HDFS data on boot(root, os) disk or on secondary(data) disks
  If HDFS data to be kept on a secondary disk,
  * then create new data disks through cloud console UI,**(note down the new disk , it could be /dev/sdb or /dev/sdc)**
  * Create ext4 filesystem on the new disk
  `mkfs.ext4  /dev/sdc` **(here as example the disk device is /dev/sdc , please check with new disk device accordingly to the environment/cloud)**
  * Create a new directory where new data disks will get mounted
  `mkdir /data/1/`
  `mount /dev/sdc /data/1/`

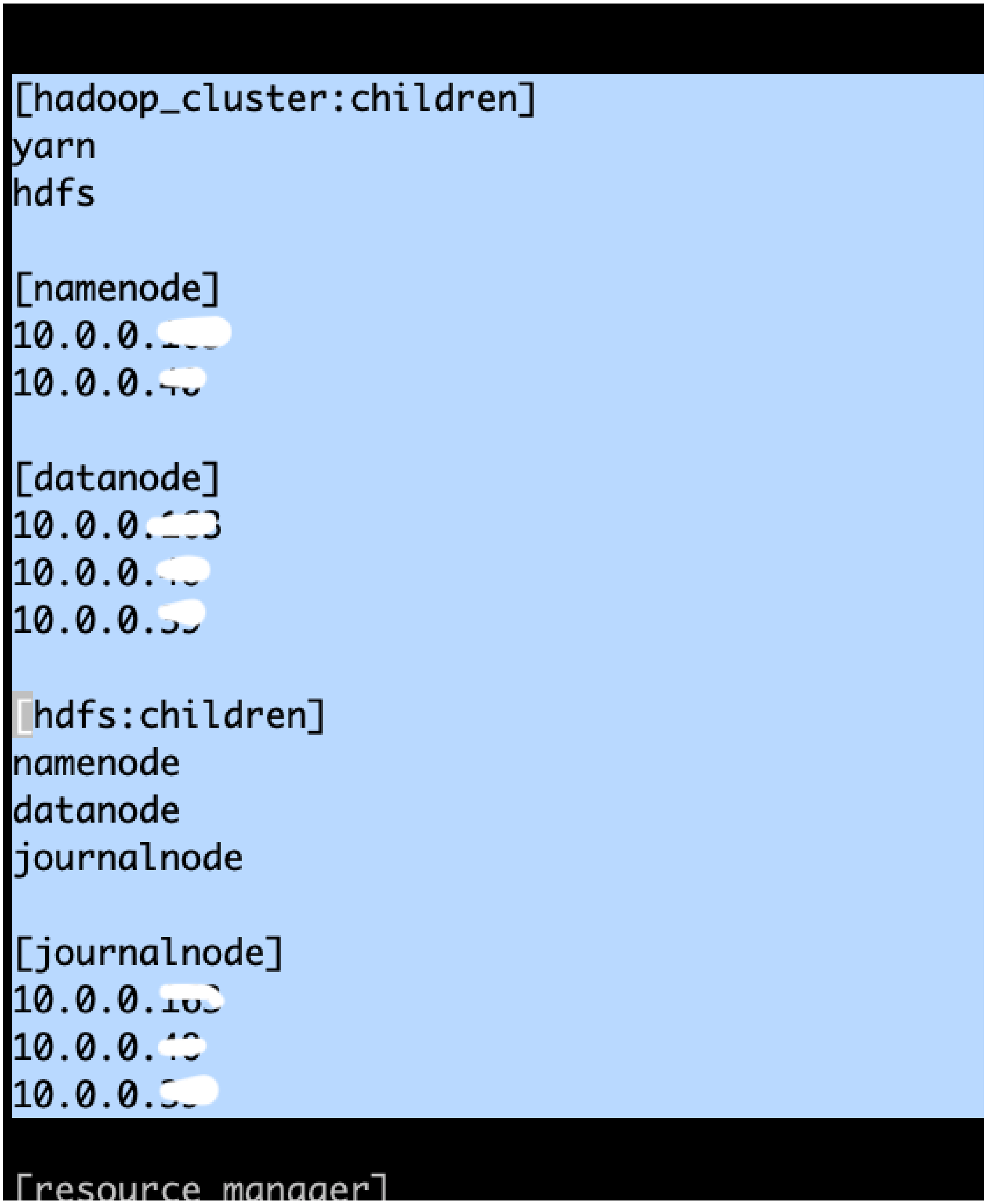
Ansible Run:

- Login to VM where ansible-playbooks are placed
- Change the current working directory to ansible-playbook



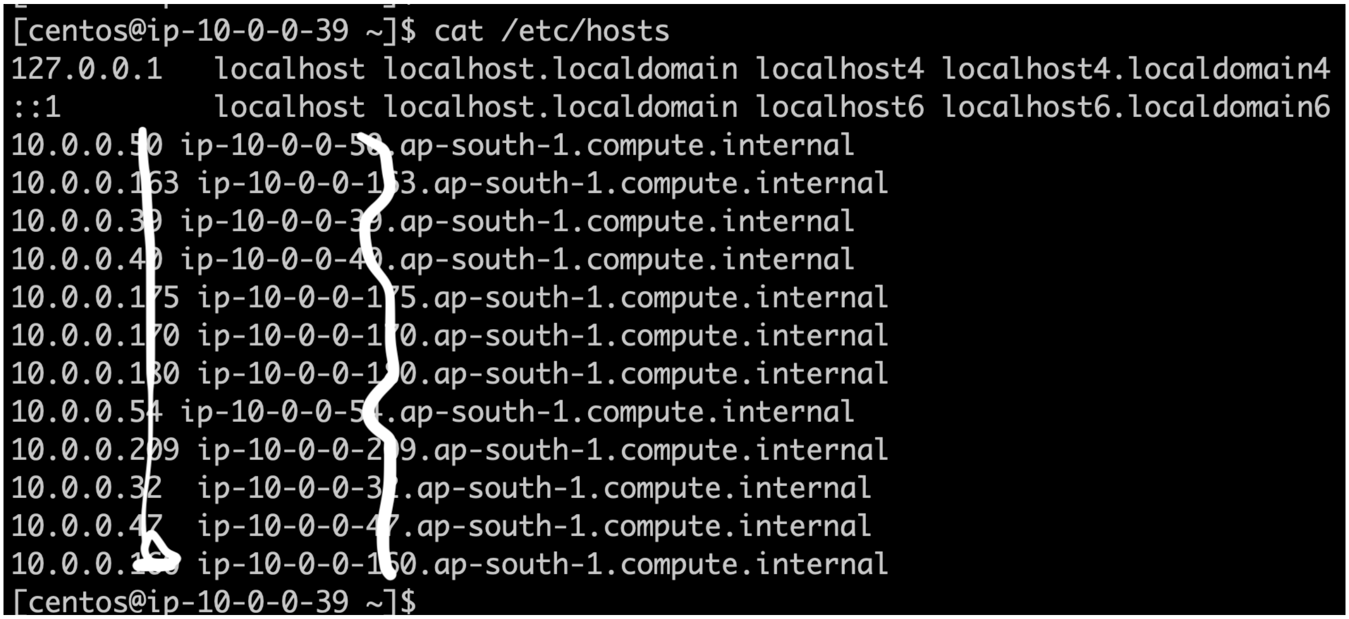
```
[centos@ip-10-0-0-68 playbook]$ pwd
/home/[redacted]/argoid-automation/playbook
[centos@ip-10-0-0-68 playbook]$
```

- Modify `namenode,datanode,journalnode` inventory IP addresses (accordingly to the environment IP addresses) in `inventory/env_name.ini` file
  Note: Do not use `env_name.ini` as an inventory name in your case, here it is shown just for example purpose, in your case name of the inventory file will be different



```
[hadoop_cluster:children]
yarn
hdfs

[namenode]
10.0.0.[redacted]
10.0.0.[redacted]

[datanode]
10.0.0.[redacted]
10.0.0.[redacted]
10.0.0.[redacted]

[hdfs:children]
namenode
datanode
journalnode

[journalnode]
10.0.0.[redacted]
10.0.0.[redacted]
10.0.0.[redacted]

[resource manager]
```

- If you are in need to keep hdfs data in `/data/2` `/data/3` directory , then modify the values for the keys `hdfs_datanode_data_dir` `hdfs_namenode_name_dir` `hdfs_journalnode_edits_dir` (by default the values are set to `/data/1/`) in inventory file
- Set replication factor required through this key `hdfs_replication_factor` in inventory file
- Set HDFS nameservice name through this key `hdfs_cluster_id` in inventory file
- Run ansible playbook
  `ansible-playbook -i inventory/env_name.ini hadoophdfsha.yml --private-key=files/common/id_rsa --limit=10.0.0.x,10.0.0.y,10.0.0.z --tags=cluster_setup`
- Contact infra team , if there are any errors
- Add host mappings in each datanode /etc/hosts files manually



```
[centos@ip-10-0-0-39 ~]$ cat /etc/hosts
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
10.0.0.50 ip-10-0-0-50.ap-south-1.compute.internal
10.0.0.163 ip-10-0-0-163.ap-south-1.compute.internal
10.0.0.39 ip-10-0-0-39.ap-south-1.compute.internal
10.0.0.40 ip-10-0-0-40.ap-south-1.compute.internal
10.0.0.175 ip-10-0-0-175.ap-south-1.compute.internal
10.0.0.170 ip-10-0-0-170.ap-south-1.compute.internal
10.0.0.180 ip-10-0-0-180.ap-south-1.compute.internal
10.0.0.54 ip-10-0-0-54.ap-south-1.compute.internal
10.0.0.209 ip-10-0-0-209.ap-south-1.compute.internal
10.0.0.32  ip-10-0-0-32.ap-south-1.compute.internal
10.0.0.47  ip-10-0-0-47.ap-south-1.compute.internal
10.0.0.160 ip-10-0-0-160.ap-south-1.compute.internal
[centos@ip-10-0-0-39 ~]$
```

- Validation
  execute this command, if there are any errors contact the infra team.
  `hdfs dfs -ls /`
- Also, validate by executing the `copyFromLocal` command operations