Troubleshooting Hadoop

There are many aspects where Hadoop will run into problems. The following aspects are the common areas where Hadoop faces configuration problems. The only way to know for sure and fix your specific problem is by looking at the logs, which can be found in `/hadoop/logs/`.

1. **Network**
    - Make sure all nodes are connected on a local network, with distinct names.
    - Each machine should have a name configured on /etc/hostname
    - Make sure the networks can be accessed by each other (given all permissions)
        - `ALL:ALL` is present in `/etc/hosts.allow`. (I'm not sure if this step is required.)
    - Ensure each machine has a static IP address and is configured in etc/hosts
2. **RSA KEY (SSH configuration)**
    - pdsh is installed
    - ssh is installed
    - rcmd_type is set as ssh
    - Master's key is shared among slaves
    - permissions of key are set to readable and executable only (not writable)
3. Ensure `hdfs-site.xml` is configured seperately for each node.
    - Datanode should only have an entry for datanode in hdfs-site.xml (and corresponding directory)
    - Namenode was run with an entry for datanode (and correspomding directory) as well, but I think it can be deleted. > **NOTE:** The entries of the HDFS data locations on the local filesystem in the .xml file should not begin their paths with "file://".
4. Check the other configuration files
    - core-site.xml
    - mapred-site.xml
    - yarn-site.xml > **Information:** Default configurations for block size and stoarge size were used!!!
5. Ensure the directory owner of the HDFS file locations is the 'user' and not 'root'
    - chmod 755 for all underlying directories
6. Remove all tmp files. Sometimes they cause problems
    - /tmp/hadoop-*

7. Ensure `hadoop/etc/hadoop/workers` is correctly configured > Should have the list of workers (the computer's names that are configured as datanodes)
8. Finally format namenode > `hdfs namenode -format`
9. Finally, run HDFS (and start-yarn.sh, if configured as the resource manager) >
   `start-dfs.sh  start-yarn.sh`
10. Check running JVM processes using `jps` > Ensure that NameNode instance and resourceManager instance are running on the machine configured as the nameNode. > > Ensure the DataNode and nodeManager instances are running on each machine configured as the dataNode.

Sources

1. The idea this error had to do something with permissions

Further Reading

1. Adding nodes to a HDFS system