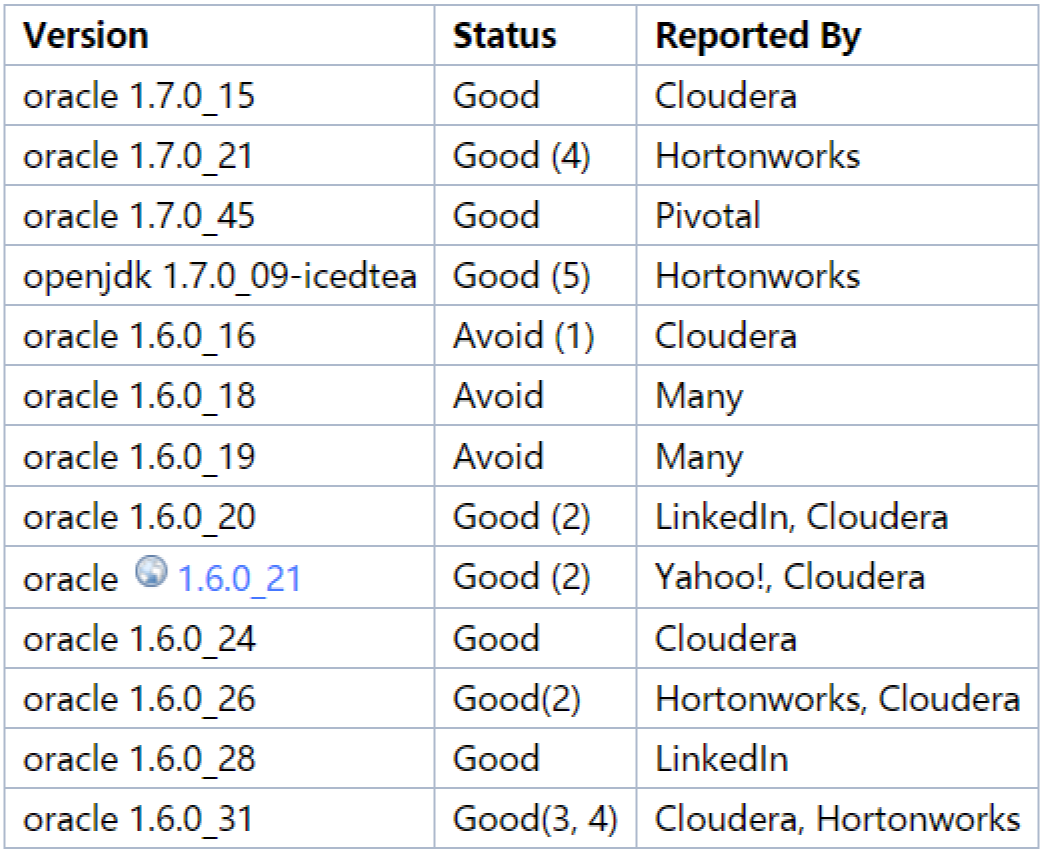
1. How to choose the java version?

   You can check the table which is from Hadoop website:



Version	Status	Reported By
oracle 1.7.0_15	Good	Cloudera
oracle 1.7.0_21	Good (4)	Hortonworks
oracle 1.7.0_45	Good	Pivotal
openjdk 1.7.0_09-icedtea	Good (5)	Hortonworks
oracle 1.6.0_16	Avoid (1)	Cloudera
oracle 1.6.0_18	Avoid	Many
oracle 1.6.0_19	Avoid	Many
oracle 1.6.0_20	Good (2)	LinkedIn, Cloudera
oracle 1.6.0_21	Good (2)	Yahoo!, Cloudera
oracle 1.6.0_24	Good	Cloudera
oracle 1.6.0_26	Good(2)	Hortonworks, Cloudera
oracle 1.6.0_28	Good	LinkedIn
oracle 1.6.0_31	Good(3, 4)	Cloudera, Hortonworks

   You can check the website for details:

   https://wiki.apache.org/hadoop/HadoopJavaVersions

2. Why I cannot connect to my VM?

   If you have checked the IP address and your ssh client, the most possible condition is that your VM is attacked and occupied by some adversaries to be used to do some DDoS attack because the VMs are directly connected to the Internet which makes them vulnerable. So the cloud provider banned your VM network in the datacenter.

   You need to contact the TA to solve this problem.

   After that, you can follow the suggestions which are provided by the DigitOcean:

   1. Always use SSH keys, this will disable the chance of getting your root account bruteforced:
   Linux or Mac: https://www.digitalocean.com/community/tutorials/how-to-use-ssh-keys-with-digitalocean-droplets

Windows: https://www.digitalocean.com/community/tutorials/how-to-use-ssh-keys-with-putty-on-digitalocean-droplets-windows-users

2. Create a sudo user, then copy your "authorization_keys" file from your /root/.ssh to your user's /home/user/.ssh directory, and set a password for this user. (This can be found in any of our initial setup tutorials for your distribution of choice)

3. Lock down your /etc/ssh/sshd_config, to disable RootLogin and PasswordAuthentication. Make sure you can login from your sudo user with SSH keys, and "sudo -i" or "sudo su -" to become root. Then restart SSH.

4. Enable your firewall, deny all incoming ports, and open up the ports that you need. You may need to use commands like "netstat -plnt" or "ss -plnt" to see what ports you may need to be open. (There is no need to open every port that it lists)
UFW: https://www.digitalocean.com/community/tutorials/how-to-set-up-a-firewall-with-ufw-on-ubuntu-14-04
IPTables: https://www.digitalocean.com/community/tutorials/how-to-set-up-a-firewall-using-iptables-on-ubuntu-14-04

5. Finally, ensure that any services that do not need to be customer facing (such as web services) are only listening to the localhost or private network interfaces.

Attention: If you plan to change the default user, root, to another user from the above suggestions, the steps of installing the Hadoop on the cluster should also be changed to use that user.

3. Which Hadoop version should I choose?

   Any version you satisfy. But the latest stable version can be a good choice.

4. What should I do if the VMs meet a memory issues when running some MapReduce examples?

   You can either reduce the number of slaves or why an example with smaller input.

   If you cannot process the whole data on the two VMs cluster, you can do it on your own computer with a configured Hadoop environment.

5. In the Docker setup, what should I submit?

   You need to submit the Dockerfile and the configuration files(e.g. core-site.xml, hdfs-site.xml, ...) for a pseudo-distributed (single node) Hadoop, which is much simpler than what you did in the real Hadoop cluster setup. You can check https://hadoop.apache.org/docs/r2.7.3/hadoop-project-dist/hadoop-common/SingleCluster.html for some configuration examples. The test of it can be done on your own computer.

6. Should I do the performance test for the n-gram Part 3 with a randomly generated dataset?

   No, you have not to do that. This part is recommended but not mandatory.

7. If I have messed up, what should I do if I want to start from the beginning?

   The steps you need to do are:

   i. Stop all the Hadoop services

   ii. Delete the Hadoop directory

   iii. Delete the HDFS directory

   iv. Delete the template files in /tmp directory for the SeconderyNameNode

   v. Re-check any configuration files you may edited before, they can be the files as below but not restricted to:

      a) /etc/hosts

      b) /etc/hostname

      c) ~/.bachrc

      d) /etc/environment

      e) /etc/ssh/sshd_config

      f) /etc/profile

      g) …

   vi. Re-extract the Hadoop package to the directory

   vii. Re-do the Hadoop cluster setup from the beginning