How to spark

1. Prerequisites
    a. Setup with SSH
    b. Setup hadoop hdfs
    c. bashrc setup
        i. export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/
        ii. export PATH=$JAVA_HOME/bin:$PATH
        iii. source .bashrc

2. Download SPARK (on master)
    a. wget https://dlcdn.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz

3. Unpack (on master)
    a. tar -xzf spark-3.3.2-bin-hadoop3.tgz

4. edit spark-env.sh (on master)
    a. create new config file from template
        i. cp /home/ubuntu/spark-3.3.2-bin-hadoop3/conf/spark-env.sh.template /home/ubuntu/spark-3.3.2-bin-hadoop3/conf/spark-env.sh
    b. edit config file
        i. sudo nano /home/ubuntu/spark-3.3.2-bin-hadoop3/conf/spark-env.sh
            1. export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/
5. add workers (on master)
    a. create new workers file from template
        i. cp /home/ubuntu/spark-3.3.2-bin-hadoop3/conf/workers.template /home/ubuntu/spark-3.3.2-bin-hadoop3/conf/workers
    b. edit workers file
        i. sudo nano /home/ubuntu/spark-3.3.2-bin-hadoop3/conf/workers
            1. 192.168.2.121
            2. 192.168.2.33

6. copy files from master to workers
    a. tarball spark folder
        i. tar czf spark.tar.gz spark-3.3.2-bin-hadoop3
    b. copy to workers
        i. scp ./spark.tar.gz 192.168.2.121:./
        ii. scp ./spark.tar.gz 192.168.2.33:./
    c. untar it on workers
        i. tar -xzf spark.tar.gz

7. DNS settings (on all nodes)
    a. for i in {1..4};
       do
            for j in {1..255};
            do
                echo "192.168.$i.$j host-192-168-$i-$j" | sudo tee -a /etc/hosts
            done
       done

8. Set hostname of your vm (on all nodes)
    a. sudo hostname host-$(hostname -I | awk '{$1=$1};1' | sed 's/\./-/'g)

    b. echo "sudo hostname host-$(hostname -I | awk '{$1=$1};1' | sed 's/\./-/'g)" | sudo tee -a /home/ubuntu/.profile

9. Set environment (on all nodes)
    a. export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
       echo JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64 | sudo tee -a /etc/environment

10. Manage cluster (from master node)
    a. start
        i. /home/ubuntu/spark-3.3.2-bin-hadoop3/sbin/start-all.sh
        ii.
    b. close
        i. /home/ubuntu/spark-3.3.2-bin-hadoop3/sbin/stop-all.sh