Big Data Lab

Exercise 1 - Basic Linux Commands

List of commands is available in the LMS course page.

Exercise 2 - Install Hadoop

Install and Configure Hadoop in Ubuntu.

Exercise 3 - Implementing MapReduce

Implement a simple map-reduce code for the wordcount problemusing
Java/Python.

(Create the jar files and run the code using HDFS.)

Exercise 4 - Implementing MapReduce 2

1. Implement map reduce for NCDC weather dataset using Hadoop -fine the max and min temperature.
2. Implement Apriori algorithm using map reduce paradigm.

Exercise 5 - Spark and PySpark

- Install spark and pyspark (ubuntu only)
- Run a spark shell and test the installation.
- Run the wordcount program that you did using hadoop usingpyspark.
- Use the movielens dataset available in the LMS theory pageand try to find out for each movie, how are the ratings distributed.

Exercise 6 - PySpark 2

1. Use the “friends_test” dataset. Col1 is ID, Col2 is name, Col 3 is Age, Col 4 is num of friends. Understand **mapvalues function of RDD** in spark and find the average number of friends for each unique age present in the dataset.

2. Use the “temp.csv” dataset. Column headers are present in the dataset. Understand filter operations and filter out only the “TMIN” values from the “desc” column. With the resultant data (RDD) find the following:

    a. Minimum temperature (overall)
    b. Minimum temperature for every ItemID
    c. Minimum temperature for every StationID.

3. Use the same dataset, filter only “TMAX” column and find the maximum temperatures just like the ones mentioned above.

Exercise 7 - Hadoop and Docker

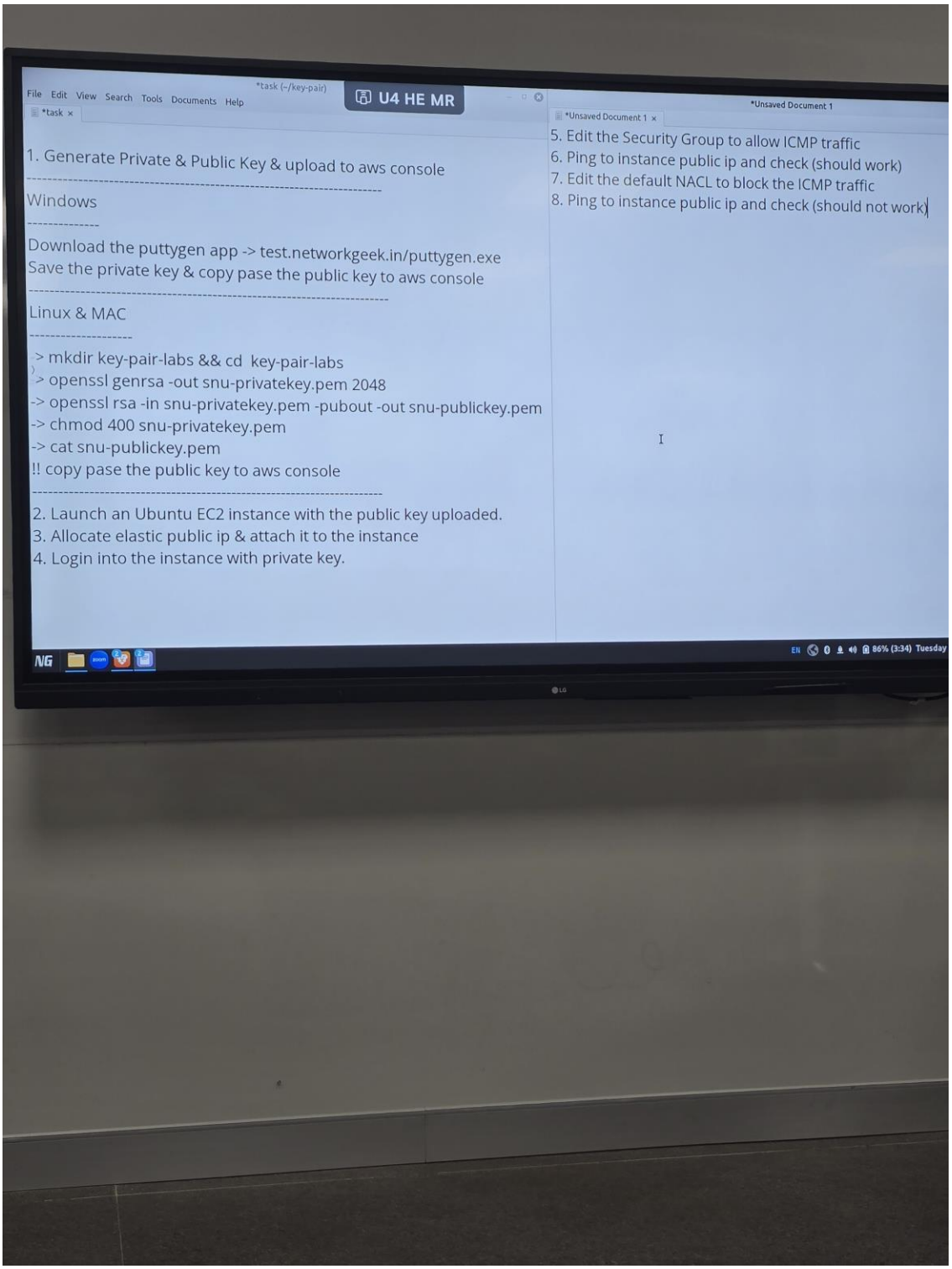
Set up a simple Hadoop environment using Docker containers, including at least one NameNode and one DataNode. Ensure the containers are properly configured to interact with each other.After the setup, verify that the Hadoop cluster is operational by running a simple HDFS file operation (e.g., uploading a fileto HDFS).

Points to note:

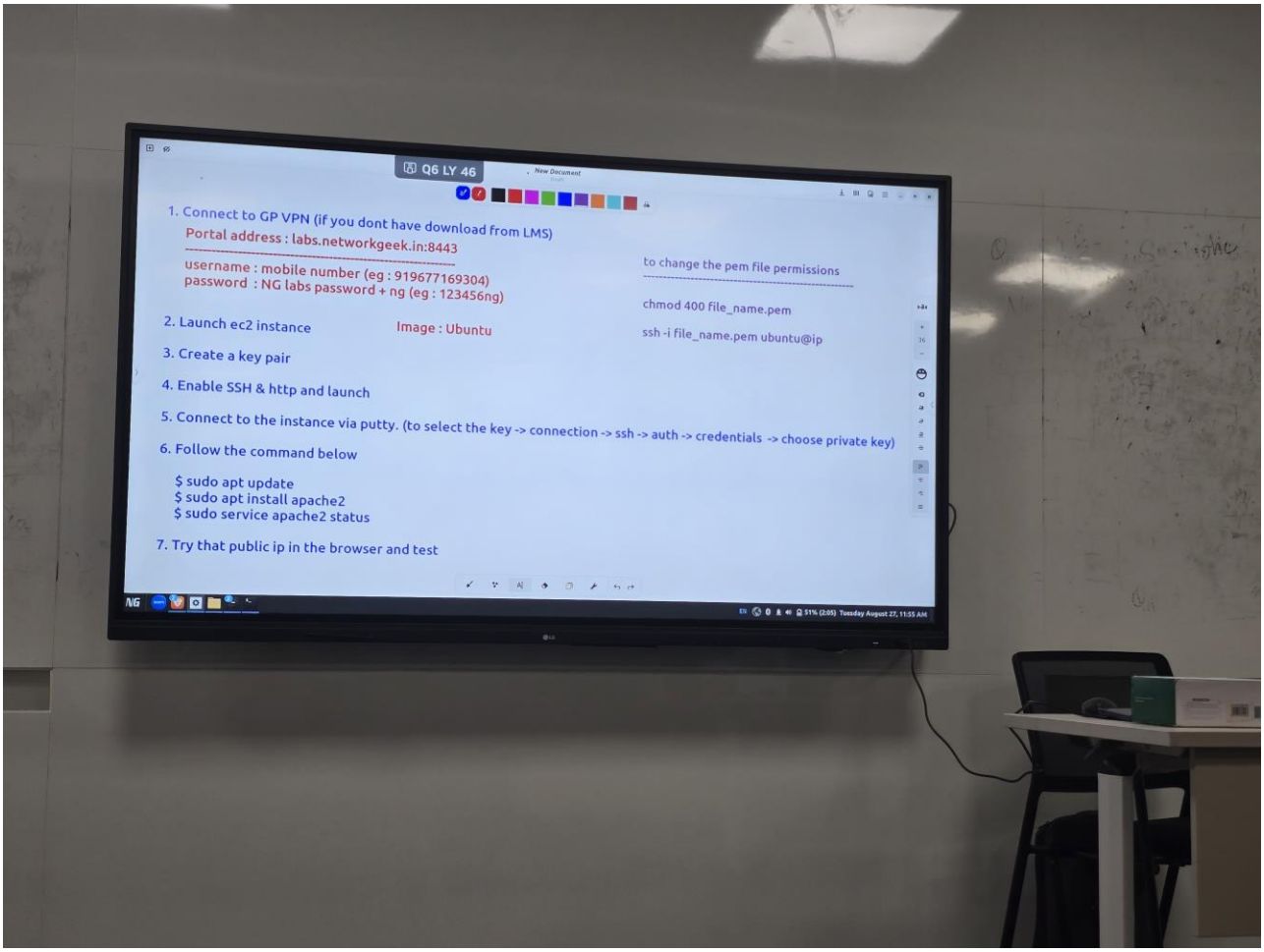
- Use Docker to containerize Hadoop's components (**NameNode** and **DataNode**).
- You can refer to the Hadoop Docker images available on Docker Hub, or **create your own Dockerfile** for the setup.
- Ensure that the **NameNode** and **DataNode** communicate over the network within the Docker environment.
- Validate your setup by running basic commands to save a file from local system to the HDFS, displaying the contents of the file.
- You can use **docker logs <container_id>** to view the output from the containers while running the application.

AWS exercises

Exercise 8 - Public and Private Keys


File Edit View Search Tools Documents Help
*task (~/key-pair)
U4 HE MR
*task ×
1. Generate Private & Public Key & upload to aws console
Windows
Download the puttygen app -> test.networkgeek.in/puttygen.exe
Save the private key & copy pase the public key to aws console
Linux & MAC
> mkdir key-pair-labs && cd key-pair-labs
-> openssl genrsa -out snu-privatekey.pem 2048
-> openssl rsa -in snu-privatekey.pem -pubout -out snu-publickey.pem
-> chmod 400 snu-privatekey.pem
-> cat snu-publickey.pem
!! copy pase the public key to aws console
2. Launch an Ubuntu EC2 instance with the public key uploaded.
3. Allocate elastic public ip & attach it to the instance
4. Login into the instance with private key.
*Unsaved Document 1
*Unsaved Document 1 ×
5. Edit the Security Group to allow ICMP traffic
6. Ping to instance public ip and check (should work)
7. Edit the default NACL to block the ICMP traffic
8. Ping to instance public ip and check (should not work)
86% (3:34) Tuesday

Exercise 9 - EC2




```
cd /var/www/html/
sudo chmod 777 index.html

echo "website" > index.html
```

Exercise 10 - Route53

DNS and GoDaddy Lab
Login into Godaddy Domain :

URL : https://www.godaddy.com/
Username : aws-ng
password : Welcome1!

Create a sub domain with your 12 digit registration number

eg : 21100101001.ngaws.xyz

b) Create a Hosted Zone in Route 53 service for your subdomain

c) Get the name server(any one) information from the Route 53 dashboard and update the NS record in godaddy portal for your subdomain.

d) Create an EC2 instance (Name : Web Server) with elastic public ip

e) Install apache application(web service) and make vm as web server

f) Create a A record in your hosted zone with the following details

subdomain : www
IP : elastic ip allotted for your account
routing policy : simple

g) check the website reachability with the url

url : www.<registration_no>.ngaws.xyz

eg : www.21100101001.ngaws.xyz

Exercise 11 - IAM

IAM Lab
---

1. Create IAM users(alice & bob), put them under group(server admin) and give full access to EC2 services
2. Create IAM users(cathy & david) put them under group(dns admin) and give full access to Route 53 services
3. Create IAM user(eve) give him access to billing
4. Create IAM user(your_name) and give full access to all the services
5. Create an alias name for your account & check the login via alias url instead of account ID.