**Loading data from HDFS to Hive**

**Steps 1**

Create devbox environment

* AWS ec2 instance with open internet access and size and ARM64
* Preferably t4g.xlarge,30 gb
* Connect to the instance
  + sudo yum update -y or sudo apt-get update -y
  + sudo yum install -y docker or sudo apt-get install docker.io -y
  + docker --version (check installation) or sudo apt-get install docker.io -y
  + sudo service docker start (Start the Docker service)
  + sudo usermod -aG docker $USER(Add the user to the docker group)
  + Exit and re connect for it to take action
  + sudo service docker restart
  + sudo curl -L "https://github.com/docker/compose/releases/download/v2.6.1/docker-compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
  + sudo chmod +x /usr/local/bin/docker-compose **(apply executable permissions to the binary)**
  + docker-compose --version
  + docker pull mentornirmal/devbox-coder:latest
  + docker pull mentornirmal/mysql:8.0
  + Mkdir project (for folder)
  + Pwd and replace the path in volume (/home/ubuntu/project:/home/coder/project)
  + **In docker-compose.yml (CMS: nano docker-compose.yml)**

version: "3.8"

services:

   coder:

     image: mentornirmal/devbox-coder:latest

ports:

       - "8080:8080"    # Port mapping for code-server

       - "9870:50070"    # HDFS NameNode

       - "9864:50010"    # HDFS DataNode

       - "8042:8042"    # YARN NodeManager

       - "8088:8032"    # YARN ResourceManager

     volumes:

       - /home/ec2-user/nirmal:/home/coder/project

      depends\_on:

        - mysql

    mysql:

      image: mentornirmal/mysql:8.0

      environment:

        MYSQL\_ROOT\_PASSWORD: rootpassword

        MYSQL\_USER: coder

        MYSQL\_PASSWORD: coderpassword

      ports:

       - "3306:3306"

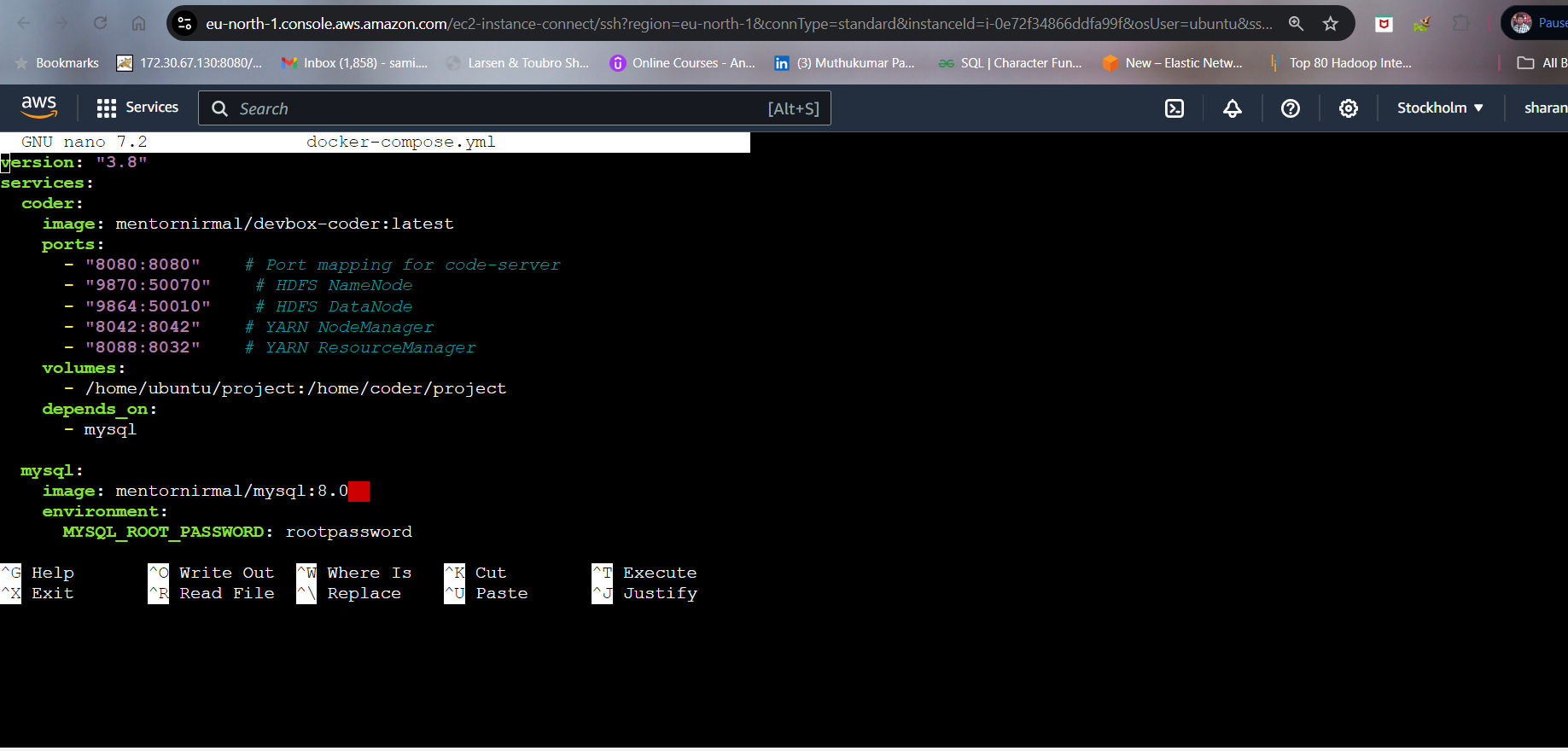
      volumes:

       - mysql\_data:/var/lib/mysql

     restart: always

volumes:

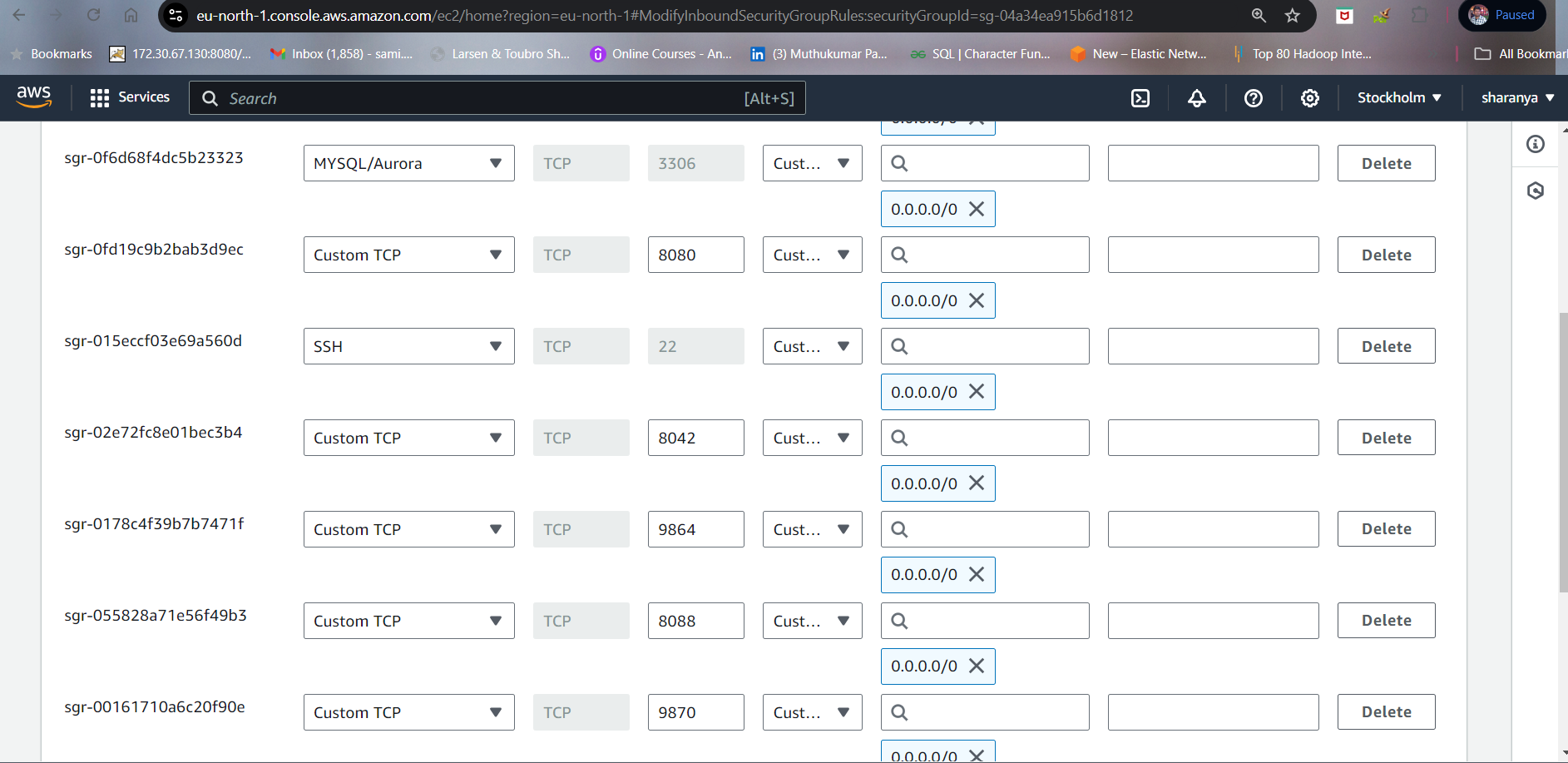
    mysql\_data:



* docker-compose up -d
* Open the instance ,into the security group ,edit inbound rules ,add rules,give the below port numbers and save .
* Expose all the ports in inbound as follows in security group

**Ports to Expose**

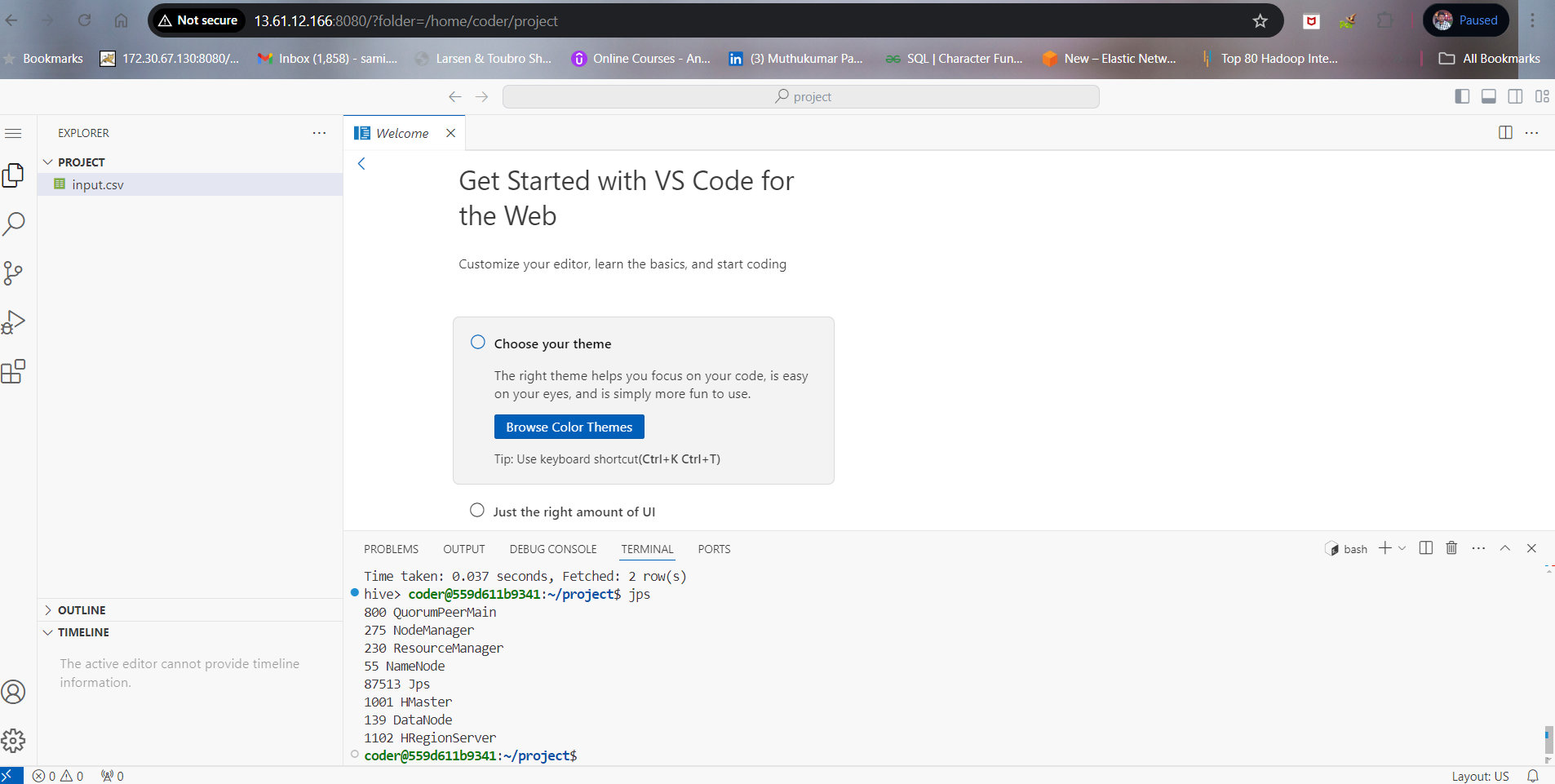
1. **Port 8080** for code-server.
2. **Port 9870** for HDFS NameNode.
3. **Port 9864** for HDFS DataNode.
4. **Port 8042** for YARN NodeManager.
5. **Port 8088** for YARN ResourceManager.
6. **Port 3306** for MySQL.



When we clcik the ip address of the instance a window will open, delete ‘s’ from https and at the end give :8080. The Dev environment will open

**Step2**

Into the dev environment



* sudo chown -R coder:coder /home/coder/project (to create an play with files)
  + Password ;codeserver
* After giving permission
* Create an input.csv file
* Copy the content in csv file and paste into this input.csv file

**Step3**

1. hdfs dfs -mkdir -p /user/hadoop/input
2. hdfs dfs -put input.csv /user/hadoop/input/
3. hdfs dfs -ls /user/hadoop/input/
4. hdfs dfs -cat /user/hadoop/input/input.csv

**step4**

Create a new database

CREATE DATABASE test\_db;

SHOW DATABASES;

USE test\_db;

CREATE TABLE census\_data (

year\_of\_estimate INT,

fips\_state\_and\_county\_codes STRING,

race\_sex\_indicator STRING,

under\_5\_years INT,

from\_5\_to\_9\_years INT,

from\_10\_to\_14\_years INT,

from\_15\_to\_19\_years INT,

from\_20\_to\_24\_years INT,

from\_25\_to\_29\_years INT,

from\_30\_to\_34\_years INT,

from\_35\_to\_39\_years INT,

from\_40\_to\_44\_years INT,

from\_45\_to\_49\_years INT,

from\_50\_to\_54\_years INT,

from\_55\_to\_59\_years INT,

from\_60\_to\_64\_years INT,

from\_65\_to\_69\_years INT,

from\_70\_to\_74\_years INT,

from\_75\_to\_79\_years INT,

from\_80\_to\_84\_years INT,

above\_85\_years INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

LINES TERMINATED BY '\n'

STORED AS TEXTFILE;

LOAD DATA INPATH '/user/hadoop/input/input.csv' INTO TABLE census\_data;

show tables in test\_db;

step5

**Query the Hive Table**

**Select all data from the table**

**SELECT \* FROM census\_data;**