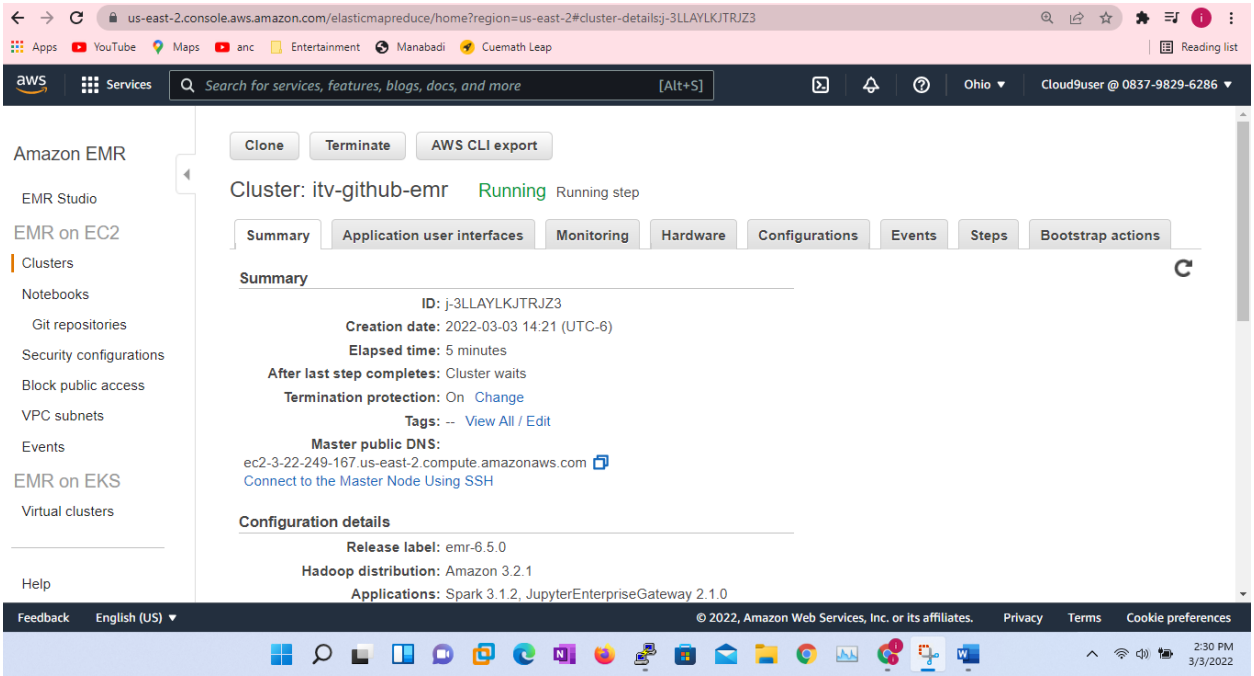
Deploying Spark Application using AWS EMR

Setup the EMR cluster to deploy application using IAM user



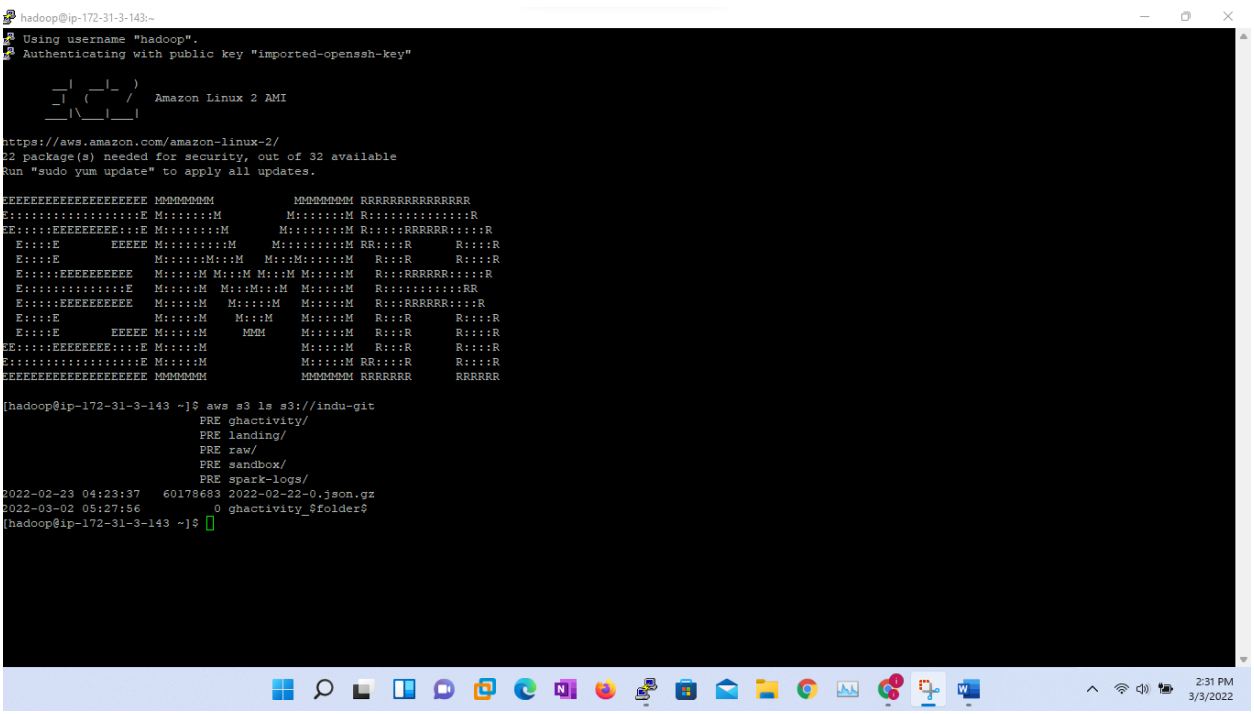
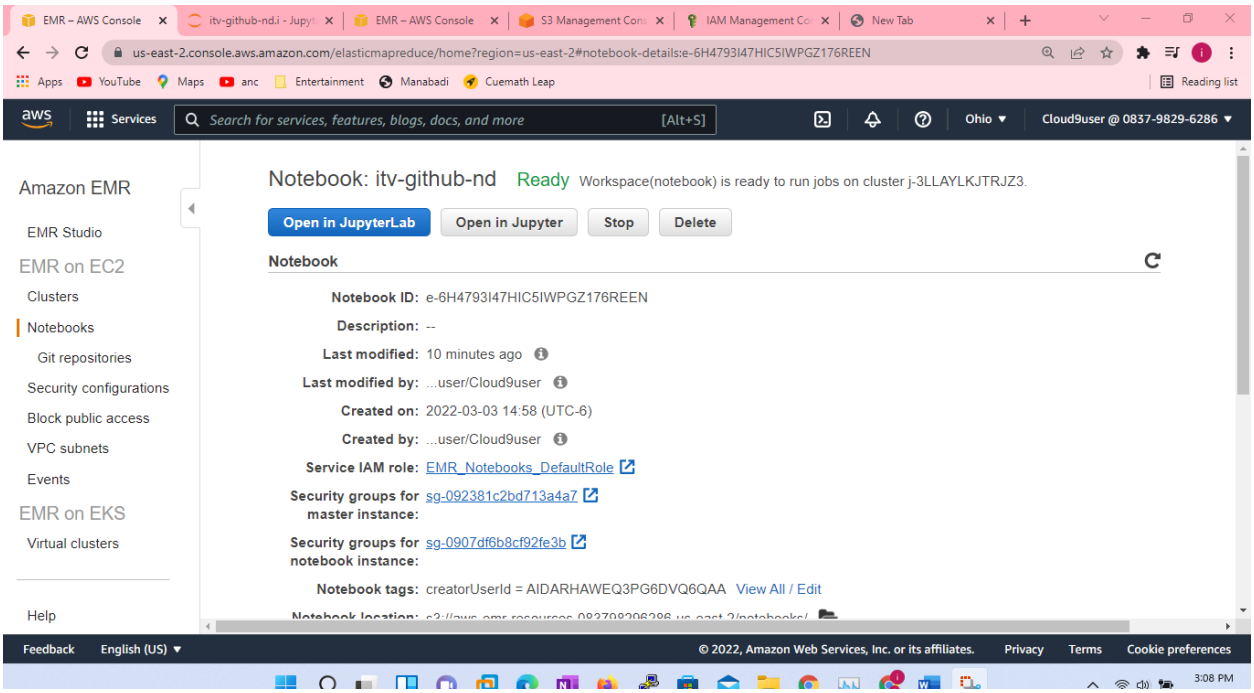


Figure 1validated SSH connectivity to master node of AWS EMR Cluster

I have created the note book on top of the EMR Cluster With that notebook I opened the Jupyter lab environment



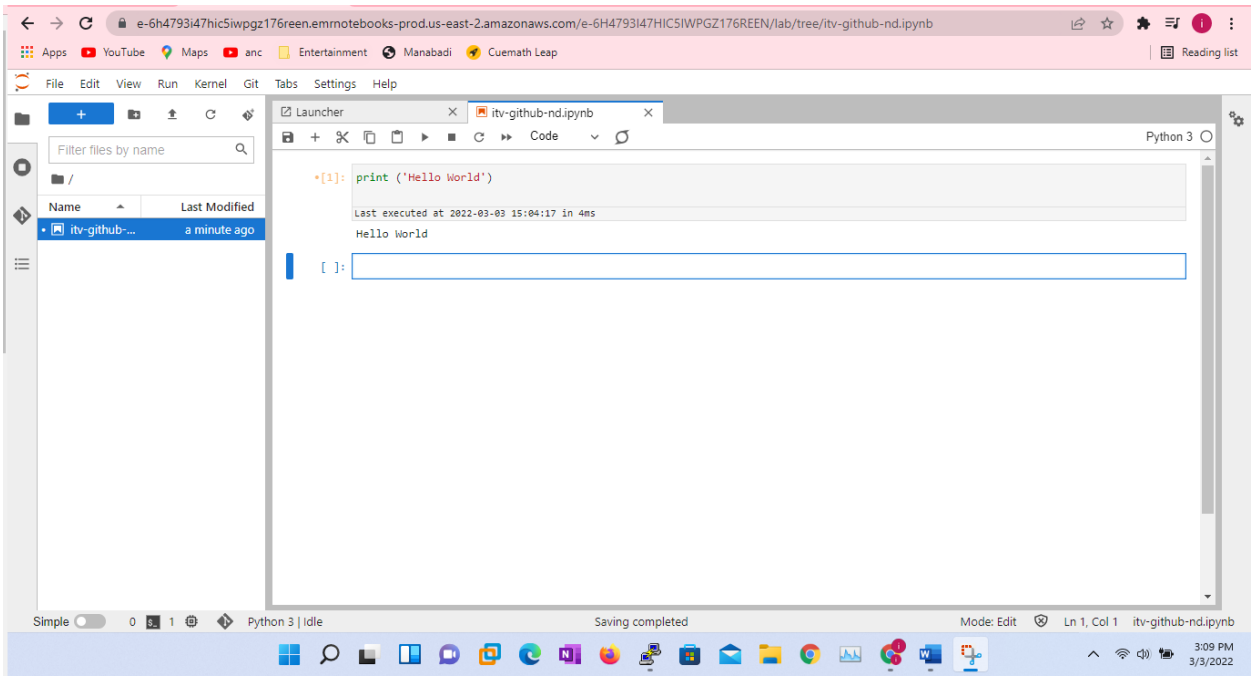


Figure 2Jupyter lab environment

Created the s3 Bucket on aws cli using the command

aws s3 mb s3://indu-github-emr --region us-east-2

Created a directory ghactivity in that folder downloaded the file from ghacive.org

wget https://data.gharchive.org/2021-01-13-{0..23}.json.gz

wget https://data.gharchive.org/2021-01-14-{0..23}.json.gz

wget https://data.gharchive.org/2021-01-15-{0..23}.json.gz

and I had uploaded the files into s3 bucket indu-github-emr using the command

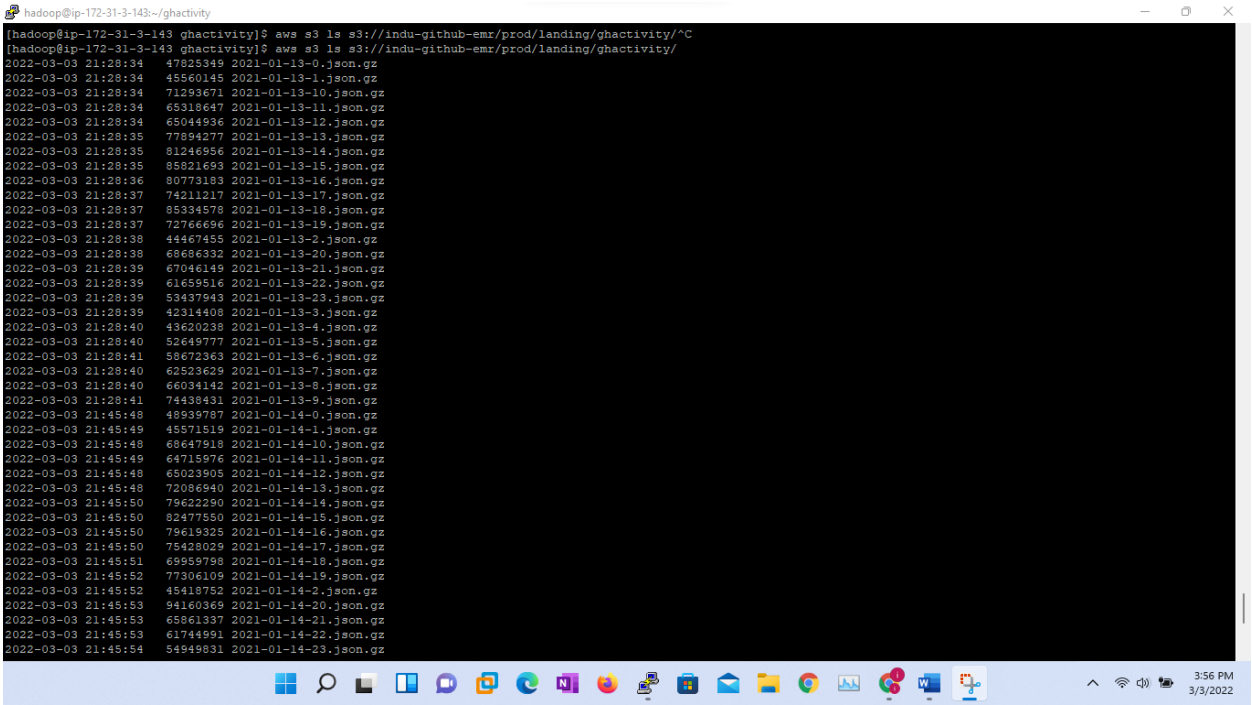
aws s3 cp . s3://indu-github-emr/prod/landing/ghactivity/ --exclude "*" --include "2021-01-13*" --recursive

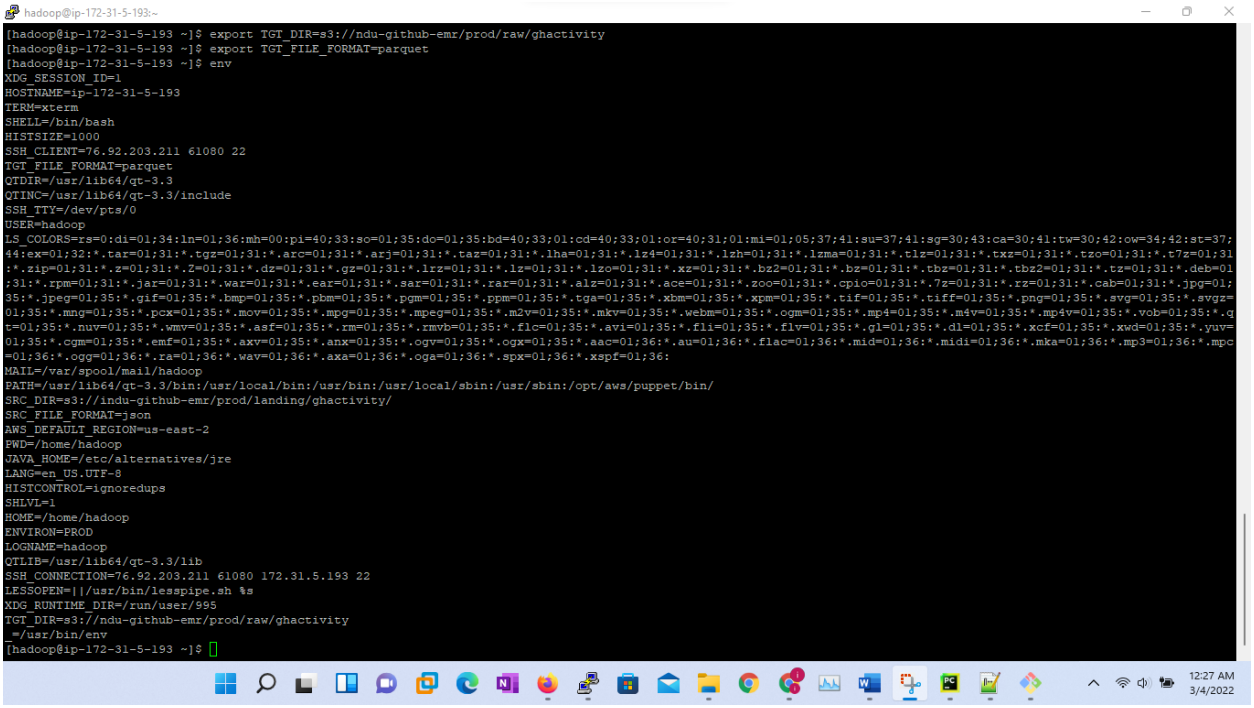
aws s3 cp . s3://indu-github-emr/prod/landing/ghactivity/ --exclude "*" --include "2021-01-14*" --recursive

aws s3 cp . s3://indu-github-emr/prod/landing/ghactivity/ --exclude "*" --include "2021-01-15*" --recursive

if you want to see the list of the files use the command:

aws s3 ls s3://indu-github-emr/prod/landing/ghactivity/





```
hadoop@ip-172-31-5-193:~
[hadoop@ip-172-31-5-193 ~]$ export TGT_DIR=s3://ndu-github-emr/prod/raw/ghactivity
[hadoop@ip-172-31-5-193 ~]$ export TGT_FILE_FORMAT=parquet
[hadoop@ip-172-31-5-193 ~]$ env
XDG_SESSION_ID=1
HOSTNAME=ip-172-31-5-193
TERM=xterm
SHELL=/bin/bash
HISTSIZE=1000
SSH_CLIENT=76.92.203.211 61080 22
TGT_FILE_FORMAT=parquet
QTDIR=/usr/lib64/qt-3.3
QTINC=/usr/lib64/qt-3.3/include
SSH_TTY=/dev/pts/0
USER=hadoop
LS_COLORS=rs=0:di=01;34:ln=01;36:mh=00:pi=40;33:so=01;35:do=01;35:bd=40;33;01:cd=40;33;01:or=40;31;01:mi=01;05;37;41:su=37;41:sg=30;43:ca=30;41:tw=30;42:ow=34;42:st=37;
44:ex=01;32:*.tar=01;31:*.tgz=01;31:*.arc=01;31:*.arj=01;31:*.taz=01;31:*.lha=01;31:*.lz4=01;31:*.lzh=01;31:*.lzma=01;31:*.tlz=01;31:*.txz=01;31:*.tzo=01;31:*.t7z=01;31
:*.zip=01;31:*.z=01;31:*.Z=01;31:*.dz=01;31:*.gz=01;31:*.lrz=01;31:*.lz=01;31:*.lzo=01;31:*.xz=01;31:*.bz2=01;31:*.bz=01;31:*.tbz=01;31:*.tbz2=01;31:*.tz=01;31:*.deb=01
;31:*.rpm=01;31:*.jar=01;31:*.war=01;31:*.ear=01;31:*.sar=01;31:*.rar=01;31:*.alz=01;31:*.ace=01;31:*.zoo=01;31:*.cpio=01;31:*.7z=01;31:*.rz=01;31:*.cab=01;31:*.jpg=01;
35:*.jpeg=01;35:*.gif=01;35:*.bmp=01;35:*.pbm=01;35:*.pgm=01;35:*.ppm=01;35:*.tga=01;35:*.xbm=01;35:*.xpm=01;35:*.tif=01;35:*.tiff=01;35:*.png=01;35:*.svg=01;35:*.svgz=
01;35:*.mng=01;35:*.pcx=01;35:*.mov=01;35:*.mpg=01;35:*.mpeg=01;35:*.m2v=01;35:*.mkv=01;35:*.webm=01;35:*.ogm=01;35:*.mp4=01;35:*.m4v=01;35:*.mp4v=01;35:*.vob=01;35:*.q
t=01;35:*.nuv=01;35:*.wmv=01;35:*.asf=01;35:*.rm=01;35:*.rmvb=01;35:*.flc=01;35:*.avi=01;35:*.fli=01;35:*.flv=01;35:*.gl=01;35:*.dl=01;35:*.xcf=01;35:*.xwd=01;35:*.yuv=
01;35:*.cgm=01;35:*.emf=01;35:*.axv=01;35:*.anx=01;35:*.ogv=01;35:*.ogx=01;35:*.aac=01;36:*.au=01;36:*.flac=01;36:*.mid=01;36:*.midi=01;36:*.mka=01;36:*.mp3=01;36:*.mpc
=01;36:*.ogg=01;36:*.ra=01;36:*.wav=01;36:*.axa=01;36:*.oga=01;36:*.spx=01;36:*.xspf=01;36:
MAIL=/var/spool/mail/hadoop
PATH=/usr/lib64/qt-3.3/bin:/usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/opt/aws/puppet/bin/
SRC_DIR=s3://indu-github-emr/prod/landing/ghactivity/
SRC_FILE_FORMAT=json
AWS_DEFAULT_REGION=us-east-2
PWD=/home/hadoop
JAVA_HOME=/etc/alternatives/jre
LANG=en_US.UTF-8
HISTCONTROL=ignoredups
SHLVL=1
HOME=/home/hadoop
ENVIRON=PROD
LOGNAME=hadoop
QTLIB=/usr/lib64/qt-3.3/lib
SSH_CONNECTION=76.92.203.211 61080 172.31.5.193 22
LESSOPEN=||/usr/bin/lesspipe.sh %s
XDG_RUNTIME_DIR=/run/user/995
TGT_DIR=s3://ndu-github-emr/prod/raw/ghactivity
_=/usr/bin/env
[hadoop@ip-172-31-5-193 ~]$
12:27 AM
3/4/2022
```

Figure 3Environment variable is set