Objectives:
Create an EC2 instance - t2.medium & 15 GB storage
Setup Hadoop 3 environment 3.3.1
Execute following jobs -
- Pi
- Wordcount
- Sudoku
- Teragen if possible

Setup jupyterlab on the same instance
Access Jupyter on webUI
Install spark
Download any csv file on your machine using jupyter terminal
Fetch the file in python notebook using spark
Create a dataframe
Perform any three operations on this dataframe. Once done upto this point, inform me.
We'll have two more objectives included in the setup! ;)
Task:
On this system, you already have some csv file at this point. Copy this csv into your
hdfs storage.
Fetch the csv file from hdfs using spark. [Hint path will update to hdfs://localhost:9000/foldername/filename.csv]
Create dataframe from this file and perform any 2-3 spark operations.
Note - Once you're done with the objective, take screenshots, download python
notebook and terminate instance

1: ssh-keygen

cd .ssh/

ls

cat id_rsa.pub >> authorized_keys

cd

ssh localhost

```
ubuntu@ip-172-31-34-18:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ubuntu/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/ubuntu/.ssh/id_rsa
Your public key has been saved in /home/ubuntu/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:XdCuuNKdG6MHm2AzM4RDCMrFHhkE6qCznZUcwT6gY3E ubuntu@ip-172-31-34-18
The key's randomart image is:
+---[RSA 3072]----+
|.o+*+     ..     |
|+.+Eo.     ..    |
|+o++o.     ..    |
|*. o=o.  . ..    |
|+o  ++  S...     |
| + o  B o .      |
|. o  . B *o.     |
|      . =.+o     |
|       .....     |
+----[SHA256]-----+
ubuntu@ip-172-31-34-18:~$ cd .ssh/
ubuntu@ip-172-31-34-18:~/.ssh$ ls
authorized_keys  id_rsa  id_rsa.pub
ubuntu@ip-172-31-34-18:~/.ssh$ cat id_rsa.pub >> authorized_keys
ubuntu@ip-172-31-34-18:~/.ssh$ cd
ubuntu@ip-172-31-34-18:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:smTCvm6xSeOxzyZ3Ct5UOFoI3p2R4o2Z796r+rqLWJM.
This key is not known by any other names
```

2: sudo apt update

sudo apt install openjdk-17-jdk –y

```
Last login: Thu Oct 12 05:06:58 2023 from 49.37.25.64
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ip-172-31-34-18:~$ sudo apt update
Hit:1 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Get:3 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy-backports InRelease [109 kB]
Get:4 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy/universe amd64 Packages [14.1 MB]
Get:5 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy/universe Translation-en [5652 kB]
Get:6 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy/universe amd64 c-n-f Metadata [286 kB]
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
134 packages can be upgraded. Run 'apt list --upgradable' to see them.
ubuntu@ip-172-31-34-18:~$ sudo apt install openjdk-17-jdk -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  adwaita-icon-theme alsa-topology-conf alsa-ucm-conf at-spi2-core ca-certificates-java dconf-gsettings-backend dconf-servi
```

3:

Nano .bashrc

export HADOOP_PREFIX=/usr/local/hadoop/
export PATH=$PATH:$HADOOP_PREFIX/bin
export HADOOP_HOME=/usr/local/hadoop/
export PATH=$PATH:$HADOOP_HOME/sbin
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export PATH=$PATH:$JAVA_HOME

Wget https://archive.apache.org/dist/hadoop/core/hadoop-3.3.1/hadoop-3.3.1.tar.gz

```
No containers need to be restarted.

No user sessions are running outdated binaries.

No VM guests are running outdated hypervisor (qemu) binaries on this host.
ubuntu@ip-172-31-34-18:~$ nano .bashrc
ubuntu@ip-172-31-34-18:~$ wget https://archive.apache.org/dist/hadoop/core/hadoop-3.3.1/hadoop-3.3.1.tar.gz
--2023-10-12 05:35:22--  https://archive.apache.org/dist/hadoop/core/hadoop-3.3.1/hadoop-3.3.1.tar.gz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 605187279 (577M) [application/x-gzip]
Saving to: 'hadoop-3.3.1.tar.gz'

hadoop-3.3.1.tar.gz                 61%[=================================================>
```

tar -xvzf hadoopfile

```
2023-10-12 05:35:54 (18.1 MB/s) - 'hadoop-3.3.1.tar.gz' saved [605187279/605187279]

ubuntu@ip-172-31-34-18:~$ tar -xvzf hadoop-3.3.1.tar.gz
```

sudo mv hadoop-2.5.0 /usr/local/hadoop/
ls
whereis java
readlink -f /usr/bin/javac | sed "s:/bin/javac::"
cd /usr/local/hadoop/
Ls
Cd etc/
Ls
Cd hadoop/
Ls

```
hadoop-3.3.1/include/Pipes.hh
ubuntu@ip-172-31-34-18:~$ sudo mv hadoop-3.3.1 /usr/local/hadoop/
ubuntu@ip-172-31-34-18:~$ ls
hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-34-18:~$ cd /usr/local/hadoop/etc/
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc$ ls
hadoop
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc$ cd hadoop/
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ ls
capacity-scheduler.xml      hadoop-policy.xml                kms-acls.xml          mapred-queues.xml.template     yarn-env.cmd
configuration.xsl           hadoop-user-functions.sh.example kms-env.sh            mapred-site.xml                yarn-env.sh
container-executor.cfg      hdfs-rbf-site.xml                kms-log4j.properties   shellprofile.d                 yarn-site.xml
core-site.xml               hdfs-site.xml                    kms-site.xml          ssl-client.xml.example         yarnservice-log4j.properties
hadoop-env.cmd              httpfs-env.sh                    log4j.properties      ssl-server.xml.example
hadoop-env.sh               httpfs-log4j.properties          mapred-env.cmd        user_ec_policies.xml.template
hadoop-metrics2.properties  httpfs-site.xml                  mapred-env.sh         workers
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ ^C
```

Nano hadoop-env.sh

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_OPTS=-Djava.net.preferIPV4Stack=true

Nano core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
```

```
  <value>hdfs://localhost:9000</value>
 </property>
</configuration>
```

Nano hdfs-site.xml

```
<configuration>
 <property>
  <name>dfs.replication</name>
  <value>1</value>
 </property>
</configuration>
```

Nano mapred-site.xml

```
<property>
<name>mapred.job.tracker</name>
<value>hdfs://localhost:9001</value>
</property>
```

Nano .bashrc

```
export HADOOP_PREFIX=/usr/local/hadoop/
export PATH=$PATH:$HADOOP_PREFIX/bin
export HADOOP_HOME=/usr/local/hadoop/
export PATH=$PATH:$HADOOP_HOME/sbin
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export PATH=$PATH:$JAVA_HOME
```

Exec bash
Hadoop namenode -format

```
hadoop-env.cmd              httpfs-env.sh                   log4j.properties        ssl-server.xml.example
hadoop-env.sh               httpfs-log4j.properties         mapred-env.cmd          user_ec_policies.xml.template
hadoop-metrics2.properties  httpfs-site.xml                 mapred-env.sh           workers
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ ^C
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ nano hadoop-env.sh
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ nano core-site.xml
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ nano hdfs-site.xml
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ nano mapred-site.xml
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ nano .bashrc
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ exec bash
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ hadoop namenode -format
```

Cd

Exec bash
Hadoop namenode -format

```
5503 SecondaryNameNode
ubuntu@ip-172-31-34-18:/usr/local/hadoop/etc/hadoop$ cd
ubuntu@ip-172-31-34-18:~$ nano .bashrc
ubuntu@ip-172-31-34-18:~$ exec bash
ubuntu@ip-172-31-34-18:~$ hadoop namenode -format
```

Start-dfs.sh

Jps

```
namenode is running as process 5131.  Stop it first and ensure /tmp/hadoop-ubuntu-namenode.pid file is empty before retry.
ubuntu@ip-172-31-34-18:~$ start-dfs.sh
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Starting namenodes on [localhost]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
localhost: namenode is running as process 5131.  Stop it first and ensure /tmp/hadoop-ubuntu-namenode.pid file is empty before retry.
Starting datanodes
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
localhost: datanode is running as process 5268.  Stop it first and ensure /tmp/hadoop-ubuntu-datanode.pid file is empty before retry.
Starting secondary namenodes [ip-172-31-34-18]
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ip-172-31-34-18: secondarynamenode is running as process 5503.  Stop it first and ensure /tmp/hadoop-ubuntu-secondarynamenode.pid file is
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ubuntu@ip-172-31-34-18:~$ jps
5268 DataNode
5131 NameNode
6191 Jps
5503 SecondaryNameNode
ubuntu@ip-172-31-34-18:~$
```

cd /usr/local/hadoop/share/hadoop/mapreduce/
hadoop jar hadoop-mapreduce-examples-3.3.1.jar pi 10 10000

```
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
namenode is running as process 5131.  Stop it first and ensure /tmp/hadoop-ubuntu-namenode.pid file is empty before retry.
ubuntu@ip-172-31-34-18:~$ cd /usr/local/hadoop/share/hadoop/mapreduce/
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.5.0.jar pi 10 10000
```

```
JAR does not exist or is not a normal file: /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.0.jar
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.3.1.jar pi 10 10000
```

```
                Failed Shuffles=0
                Merged Map outputs=10
                GC time elapsed (ms)=40
                Total committed heap usage (bytes)=1880096768
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1180
        File Output Format Counters
                Bytes Written=97
Job Finished in 2.82 seconds
Estimated value of Pi is 3.14120000000000000000
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$
```

hadoop jar hadoop-mapreduce-examples-3.3.1.jar teragen 5207890 input

```
                Bytes Written=97
Job Finished in 2.82 seconds
Estimated value of Pi is 3.14120000000000000000
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.3.1.jar teragen 5207890 input
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2023-10-12 06:20:36,292 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-10-12 06:20:36,465 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-10-12 06:20:36,465 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-10-12 06:20:36,669 INFO terasort.TeraGen: Generating 5207890 using 1
```

hadoop jar hadoop-mapreduce-examples-3.3.1.jar terasort input output

```
        File Output Format Counters
                Bytes Written=520789000
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.3.1.jar terasort input output
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2023-10-12 06:22:25,370 INFO terasort.TeraSort: starting
2023-10-12 06:22:26,296 INFO input.FileInputFormat: Total input files to process : 1
Spent 130ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
Computing input splits took 132ms
Sampling 4 splits of 4
```

hadoop jar hadoop-mapreduce-examples-3.3.1.jar teravalidate input validate

```
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=520789000
        File Output Format Counters
                Bytes Written=520789000
2023-10-12 06:23:07,769 INFO terasort.TeraSort: done
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.3.1.jar teravalidate input validate
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2023-10-12 06:26:01,371 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-10-12 06:26:01,508 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-10-12 06:26:01,508 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-10-12 06:26:01,893 INFO input.FileInputFormat: Total input files to process : 1
Spent 110ms computing base-splits.
Spent 19ms computing TeraScheduler splits.
```

hadoop fs –ls

```
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=520789000
        File Output Format Counters
                Bytes Written=268196368
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop fs -ls
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Found 3 items
drwxr-xr-x   - ubuntu supergroup          0 2023-10-12 06:20 input
drwxr-xr-x   - ubuntu supergroup          0 2023-10-12 06:23 output
drwxr-xr-x   - ubuntu supergroup          0 2023-10-12 06:26 validate
```

nano puzzle.txt
hadoop jar hadoop-mapreduce-examples-3.3.1.jar sudoku puzzle.txt

```
ubuntu@ip-172-31-34-18: /usr/local/hadoop/share/hado
  GNU nano 6.2
? 9 7 ? ? ? ? ? 5
? 6 3 ? 4 ? 2 ? ?
? ? ? 9 ? ? ? 8 ?
? ? 9 ? ? ? ? 7 ?
? ? ? 1 ? 6 ? ? ?
2 5 4 8 3 ? ? ? 1
? 7 ? ? ? 1 8 ? ?
? 8 ? ? 7 ? 6 ? 4
5 ? ? ? ? 2 ? 9 ?
```

```
Found 3 items
drwxr-xr-x   - ubuntu supergroup          0 2023-10-12 06:20 input
drwxr-xr-x   - ubuntu supergroup          0 2023-10-12 06:23 output
drwxr-xr-x   - ubuntu supergroup          0 2023-10-12 06:26 validate
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ nano puzzle.txt
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.3.1.jar sudoku puzzle.txt
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Solving puzzle.txt
1 9 7 6 2 8 4 3 5
8 6 3 7 4 5 2 1 9
4 2 5 9 1 3 7 8 6
6 1 9 2 5 4 3 7 8
7 3 8 1 9 6 5 4 2
2 5 4 8 3 7 9 6 1
9 7 2 4 6 1 8 5 3
3 8 1 5 7 9 6 2 4
5 4 6 3 8 2 1 9 7

Found 1 solutions
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$
```

download sample file - wget
https://raw.githubusercontent.com/ErikSchierboom/sentencegenerator/master/samples/the-king-james-bible.txt > sample.txt
Cat theking > sample.txt
Ls

```
Found 1 solutions
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ wget https://raw.githubusercontent.com/ErikSchierboom/sentencegenerator/master/samples/the-king-james-b
ible.txt > sample.txt
--2023-10-12 06:34:56--  https://raw.githubusercontent.com/ErikSchierboom/sentencegenerator/master/samples/the-king-james-bible.txt
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.109.133, 185.199.110.133, 185.199.111.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4332499 (4.1M) [text/plain]
Saving to: 'the-king-james-bible.txt'

the-king-james-bible.txt            100%[===================================================================>]   4.13M  --.-KB/s    in 0.08s
```

cat the-king-james-bible.txt > sample.txt

hdfs dfs -mkdir -p “/user/ubuntu”

hdfs dfs -mkdir input1

hdfs dfs -put sample.txt input1

hadoop jar hadoop-mapreduce-examples-3.3.1.jar wordcount input1 output1

```
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ cat the-king-james-bible.txt > sample.txt
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -mkdir -p "/user/ubuntu"
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -mkdir input1
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -put sample.txt input1
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.3.1.jar wordcount input1 output1
```

hdfs dfs -ls output1

hdfs dfs -tail output1/part-r-00000 | tail

```
        Bytes Read=4332499
        File Output Format Counters
                Bytes Written=341453
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -ls output1
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
Found 2 items
-rw-r--r--   1 ubuntu supergroup          0 2023-10-12 06:41 output1/_SUCCESS
-rw-r--r--   1 ubuntu supergroup     341453 2023-10-12 06:41 output1/part-r-00000
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -tail output1/part-r-00000 | tail
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
youth:  7
youth;  8
youth?  2
youthful        1
youths  1
youths, 1
zeal    13
zeal,   3
zealous 8
zealously       2
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$
```

hdfs dfs -get output1 output1

Tail output1/part-r-00000

```
zealously       2
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ hdfs dfs -get output1 output1
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ tail output1/part-r-00000
youth:  7
youth;  8
youth?  2
youthful        1
youths  1
youths, 1
zeal    13
zeal,   3
zealous 8
zealously       2
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$
```

Ls

Cd

sudo apt update

```
zealously       2
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ ls
hadoop-mapreduce-client-app-3.3.1.jar   hadoop-mapreduce-client-jobclient-3.3.1-tests.jar   hadoop-mapreduce-examples-3.3.1.jar   sample.txt
hadoop-mapreduce-client-common-3.3.1.jar   hadoop-mapreduce-client-jobclient-3.3.1.jar   input1   source
hadoop-mapreduce-client-core-3.3.1.jar   hadoop-mapreduce-client-nativetask-3.3.1.jar   lib-examples   the-king-james-bible.tx
hadoop-mapreduce-client-hs-3.3.1.jar   hadoop-mapreduce-client-shuffle-3.3.1.jar   output1
hadoop-mapreduce-client-hs-plugins-3.3.1.jar   hadoop-mapreduce-client-uploader-3.3.1.jar   puzzle.txt
ubuntu@ip-172-31-34-18:/usr/local/hadoop/share/hadoop/mapreduce$ cd
ubuntu@ip-172-31-34-18:~$ sudo apt update
Hit:1 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Hit:3 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Fetched 229 kB in 1s (430 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
133 packages can be upgraded. Run 'apt list --upgradable' to see them.
ubuntu@ip-172-31-34-18:~$
```

sudo apt install python3-pip

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
133 packages can be upgraded. Run 'apt list --upgradable' to see them.
ubuntu@ip-172-31-34-18:~$ sudo apt install python3-pip
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  build-essential bzip2 cpp cpp-11 dpkg-dev fakeroot g++ g++-11 gcc gcc-11 gcc-11-base gcc-12-base javascript-common libalgorithm-diff-perl libalgorith
  libalgorithm-merge-perl libasan6 libatomic1 libc-dev-bin libc-devtools libc6 libc6-dev libcc1-0 libcrypt-dev libdpkg-perl libexpat1-dev libfakeroot
  libfile-fcntllock-perl libgcc-11-dev libgcc-s1 libgd3 libgomp1 libisl23 libitm1 libjs-jquery libjs-sphinxdoc libjs-underscore liblsan0 libmpc3 libnsl
```

pip3 install jupyterlab

```
No VM guests are running outdated hypervisor (qemu) binaries on this host.
ubuntu@ip-172-31-34-18:~$ pip3 install jupyterlab
Defaulting to user installation because normal site-packages is not writeable
Collecting jupyterlab
  Downloading jupyterlab-4.0.7-py3-none-any.whl (9.2 MB)
     ---------------------------------------- 9.2/9.2 MB 46.9 MB/s eta 0:00:00
Collecting traitlets
```

sudo apt update
exec bash
exit

```
2.8.19.14 typing-extensions-4.8.0 uri-template-1.3.0 wcwidth-0.2.8 webcolors-1.13 webencodin
ubuntu@ip-172-31-34-18:~$ sudo apt update
Hit:1 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:4 http://security.ubuntu.com/ubuntu jammy-security InRelease
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
124 packages can be upgraded. Run 'apt list --upgradable' to see them.
ubuntu@ip-172-31-34-18:~$ exec bash
ubuntu@ip-172-31-34-18:~$ exit
```

Open new putty
Jupyter server –generate-config
Nano /home/ubuntu/.jupyter/jupyter_server_config.py
C.ServerApp.ip = ‘*’
C.ServerApp.port = 8500

```
ubuntu@ip-172-31-34-18:~$ jupyter server --generate-config
Writing default config to: /home/ubuntu/.jupyter/jupyter_server_config.py
ubuntu@ip-172-31-34-18:~$ nano /home/ubuntu/.jupyter/jupyter_server_config.py
ubuntu@ip-172-31-34-18:~$
```

```
# Configuration file for jupyter-server.

c = get_config()  #noqa
C = ServerApp.ip = '*'
C = ServerApp.port = 8500
#------------------------------------------------------------------------------
# Application(SingletonConfigurable) configuration
#------------------------------------------------------------------------------
## This is an application.
```

Screen
(two time spacebar)
jupyter-lab --no-browser

```
ubuntu@ip-172-31-34-18: ~
ubuntu@ip-172-31-34-18:~$ jupyter server --generate-config
Writing default config to: /home/ubuntu/.jupyter/jupyter_server_config.py
ubuntu@ip-172-31-34-18:~$ nano /home/ubuntu/.jupyter/jupyter_server_config.py
ubuntu@ip-172-31-34-18:~$ screen
        self._read_file_as_dict()
      File "/home/ubuntu/.local/lib/python3.10/site-packages/traitlets/config/loader.py", lin
        exec(compile(f.read(), conf_filename, "exec"), namespace, namespace)  # noqa
      File "/home/ubuntu/.jupyter/jupyter_server_config.py", line 4
        C = ServerApp.ip = '*'
```

```
[I 2023-10-12 07:36:53.730 ServerApp] http://localhost:8500/lab?token=1d96bbdc1859ce84e98e734b98ebca26d58cdcae340995a6
[I 2023-10-12 07:36:53.730 ServerApp]     http://127.0.0.1:8500/lab?token=1d96bbdc1859ce84e98e734b98ebca26d58cdcae340995a6
[I 2023-10-12 07:36:53.730 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 2023-10-12 07:36:53.733 ServerApp]

    To access the server, open this file in a browser:
        file:///home/ubuntu/.local/share/jupyter/runtime/jpserver-10233-open.html
    Or copy and paste one of these URLs:
        http://localhost:8500/lab?token=1d96bbdc1859ce84e98e734b98ebca26d58cdcae340995a6
        http://127.0.0.1:8500/lab?token=1d96bbdc1859ce84e98e734b98ebca26d58cdcae340995a6
[I 2023-10-12 07:36:53.756 ServerApp] Skipped non-installed server(s): bash-language-server, dockerfile-language-server-nodejs, javascript-typescript-l
anguage-server, julia-language-server, pyright, python-language-server, python-lsp-server, r-languageserver, sql-language-server, texlab, typescript-la
ified-language-server, vscode-css-languageserver-bin, vscode-html-languageserver-bin, vscode-json-languageserver-bin, yaml-language-server
[I 2023-10-12 07:37:05.212 ServerApp] Malformed HTTP message from 49.37.25.64: no colon in header line
[I 2023-10-12 07:38:07.642 ServerApp] Malformed HTTP message from 49.37.25.64: Malformed HTTP request line
```

```
jupyter server list
```

will show you the URLs of running servers with their tokens, which you can copy and paste into your browser. For example:

```
Currently running servers:
http://localhost:8888/?token=c8de56fa... :: /Users/you/notebooks
```

or you can paste just the token value into the password field on this page.

See **the documentation on how to enable a password** in place of token authentication, if you would like to avoid dealing with random tokens.

Cookies are required for authenticated access to the Jupyter server.

## Setup a Password

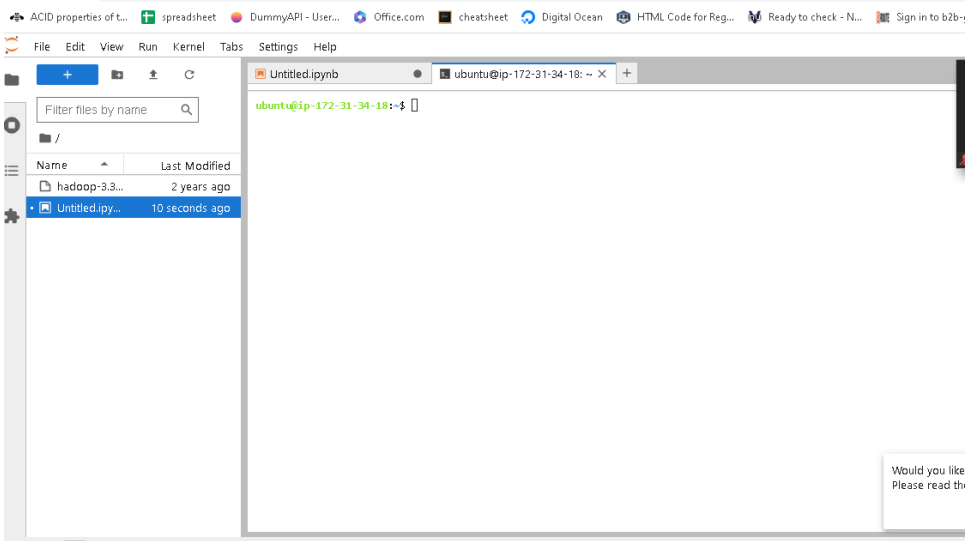
You can also setup a password by entering your token and a new password on the fields below:

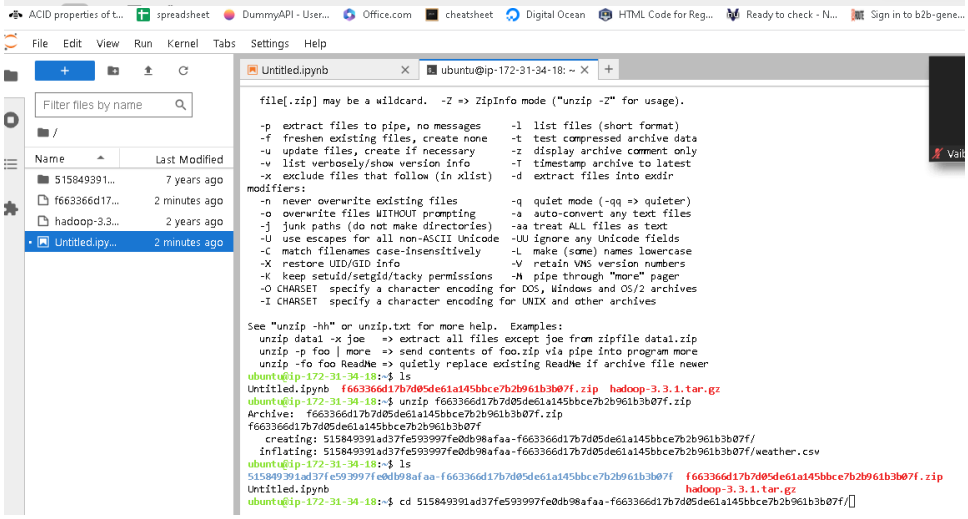
### Token

### New Password

Log in and set new password

1d96bbdc1859ce84e98e734b98ebca26d58cdcae340995a6



```
  -j  junk paths (do not make directories)   -aa treat ALL files as text
  -U  use escapes for all non-ASCII Unicode  -UU ignore any Unicode fields
  -C  match filenames case-insensitively     -L  make (some) names lowercase
  -X  restore UID/GID info                   -V  retain VMS version numbers
  -K  keep setuid/setgid/tacky permissions   -N  pipe through "more" pager
  -O CHARSET  specify a character encoding for DOS, Windows and OS/2 archives
  -I CHARSET  specify a character encoding for UNIX and other archives

See "unzip -hh" or unzip.txt for more help.  Examples:
  unzip data1 -x joe   => extract all files except joe from zipfile data1.zip
  unzip -p foo | more  => send contents of foo.zip via pipe into program more
  unzip -fo foo ReadMe => quietly replace existing ReadMe if archive file newer
ubuntu@ip-172-31-34-18:~$ ls
Untitled.ipynb  f663366d17b7d05de61a145bbce7b2b961b3b07f.zip  hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-34-18:~$ unzip f663366d17b7d05de61a145bbce7b2b961b3b07f.zip
Archive:  f663366d17b7d05de61a145bbce7b2b961b3b07f.zip
f663366d17b7d05de61a145bbce7b2b961b3b07f
   creating: 515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f/
  inflating: 515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f/weather.csv
ubuntu@ip-172-31-34-18:~$ ls
515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f  f663366d17b7d05de61a145bbce7b2b961b3b07f.zip
Untitled.ipynb                                                             hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-34-18:~$ cd 515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f/
ubuntu@ip-172-31-34-18:~/515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f$ ls
weather.csv
ubuntu@ip-172-31-34-18:~/515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f$ pwd
/home/ubuntu/515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f
ubuntu@ip-172-31-34-18:~/515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f$ cp weather.csv /home/ubuntu/
ubuntu@ip-172-31-34-18:~/515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f$ cd
ubuntu@ip-172-31-34-18:~$ ls
515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f  f663366d17b7d05de61a145bbce7b2b961b3b07f.zip  weather.csv
Untitled.ipynb                                                             hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-34-18:~$
```

```
[*]: pip install pyspark

     Defaulting to user installation because normal site-packages is not writeable
     Collecting pyspark
       Downloading pyspark-3.5.0.tar.gz (316.9 MB)
       ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 316.9/316.9 MB 104.9 MB/s eta 0:00:01
[ ]:
```

```
[3]: from pyspark.sql import*

[5]: spark=SparkSession.builder.appName('zen').getOrCreate()

     Setting default log level to "WARN".
     To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
     23/10/12 08:18:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wher
     e applicable

[6]: spark

[6]: SparkSession - in-memory

     SparkContext

     Spark UI

     Version    v3.5.0
     Master     local[*]
     AppName    zen

[7]: f1=spark.read.option('header','true').csv('/home/ubuntu/weather.csv')

[8]: f1.show()
```

File Edit View Run Kernel Tabs Settings Help

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code

Filter files by name

/

| Name | Last Modified |
|---|---|
| 515849391... | 7 years ago |
| f663366d17... | 39 minutes ago |
| hadoop-3.3... | 2 years ago |
| Untitled.ipy... | 44 seconds ago |
| weather.csv | 35 minutes ago |

```
[7]: f1=spark.read.option('header','true').csv('/home/ubuntu/weather.csv')

[8]: f1.show()

+--------+-----------+--------+-----+----+
| outlook|temperature|humidity|windy|play|
+--------+-----------+--------+-----+----+
|overcast|        hot|    high|FALSE| yes|
|overcast|       cool|  normal| TRUE| yes|
|overcast|       mild|    high| TRUE| yes|
|overcast|        hot|  normal|FALSE| yes|
|   rainy|       mild|    high|FALSE| yes|
|   rainy|       cool|  normal|FALSE| yes|
|   rainy|       cool|  normal| TRUE|  no|
|   rainy|       mild|  normal|FALSE| yes|
|   rainy|       mild|    high| TRUE|  no|
|   sunny|        hot|    high|FALSE|  no|
|   sunny|        hot|    high| TRUE|  no|
|   sunny|       mild|    high|FALSE|  no|
|   sunny|       cool|  normal|FALSE| yes|
|   sunny|       mild|  normal| TRUE| yes|
+--------+-----------+--------+-----+----+

[9]: print(type(f1))

<class 'pyspark.sql.dataframe.DataFrame'>
```

Simple 1 1 Python 3 (ipykernel) | Idle Mode: Edit

File Edit View Run Kernel Tabs Settings Help Nimanshi Rustagi

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code Python 3 (ipykernel)

Filter files by name

/

| Name | Last Modified |
|---|---|
| 515849391... | 7 years ago |
| f663366d17... | 40 minutes ago |
| hadoop-3.3... | 2 years ago |
| Untitled.ipy... | 1 minute ago |
| weather.csv | 36 minutes ago |

```
[9]: print(type(f1))

<class 'pyspark.sql.dataframe.DataFrame'>

[10]: f1.head(3)

[10]: [Row(outlook='overcast', temperature='hot', humidity='high', windy='FALSE', play='yes'),
       Row(outlook='overcast', temperature='cool', humidity='normal', windy='TRUE', play='yes'),
       Row(outlook='overcast', temperature='mild', humidity='high', windy='TRUE', play='yes')]

[11]: print(type(f1))

<class 'pyspark.sql.dataframe.DataFrame'>

[ ]:
```

Simple 1 1 Python 3 (ipykernel) | Idle Mode: Edit Ln 1, Col 1 Untitled.ipynb

# IMDB file added

| | |
|---|---|
| f663366d17... | 44 minutes ago |
| hadoop-3.3... | 2 years ago |
| imdb_1000.... | 1 minute ago |
| Untitled.ipy... | 5 minutes ago |
| weather.csv | 40 minutes ago |

```
ubuntu@ip-172-31-34-18:~/515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f$ cd
ubuntu@ip-172-31-34-18:~$ ls
515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f  f663366d17b7d05de61a145bbce7b2b961b3b07f.zip  weather.csv
Untitled.ipynb                                                            hadoop-3.3.1.tar.gz
ubuntu@ip-172-31-34-18:~$ wget "C:\Users\Admin\Downloads\Sample - Superstore.csv"
--2023-10-12 08:06:29--  ftp://c/%5CUsers%5CAdmin%5CDownloads%5CSample%20-%20Superstore.csv
           => '\\Users\\Admin\\Downloads\\Sample - Superstore.csv'
Resolving c (c)... failed: Temporary failure in name resolution.
wget: unable to resolve host address 'c'
ubuntu@ip-172-31-34-18:~$ wget https://a-nw-nishant.s3.eu-west-1.amazonaws.com/imdb_1000.csv
--2023-10-12 08:27:39--  https://a-nw-nishant.s3.eu-west-1.amazonaws.com/imdb_1000.csv
Resolving a-nw-nishant.s3.eu-west-1.amazonaws.com (a-nw-nishant.s3.eu-west-1.amazonaws.com)... 52.218.60.176, 52.218.85.72, 52.218.88.128, ...
Connecting to a-nw-nishant.s3.eu-west-1.amazonaws.com (a-nw-nishant.s3.eu-west-1.amazonaws.com)|52.218.60.176|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 91499 (89K) [text/csv]
Saving to: 'imdb_1000.csv'

imdb_1000.csv                       100%[===================================================================>]  89.35K  --.-KB/s    in 0.001s

2023-10-12 08:27:39 (76.9 MB/s) - 'imdb_1000.csv' saved [91499/91499]

ubuntu@ip-172-31-34-18:~$ ls
515849391ad37fe593997fe0db98afaa-f663366d17b7d05de61a145bbce7b2b961b3b07f  f663366d17b7d05de61a145bbce7b2b961b3b07f.zip  imdb_1000.csv
Untitled.ipynb                                                            hadoop-3.3.1.tar.gz                           weather.csv
ubuntu@ip-172-31-34-18:~$
```

File Edit View Run Kernel Tabs Settings Help Nimanshi Rust

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code Python

Filter files by name

/

| Name | Last Modified |
|---|---|
| 515849391... | 7 years ago |
| f663366d17... | 49 minutes ago |
| hadoop-3.3... | 2 years ago |
| imdb_1000.... | 6 minutes ago |
| Untitled.ipy... | 22 seconds ago |
| weather.csv | 45 minutes ago |

```
[12]: f2=spark.read.option('header','true').csv('/home/ubuntu/imdb_1000.csv')

[13]: f2.show()

+-----------+--------------------+--------------+--------+--------+--------------------+
|star_rating|               title|content_rating|   genre|duration|          actors_list|
+-----------+--------------------+--------------+--------+--------+--------------------+
|        9.3|The Shawshank Red...|             R|   Crime|     142|[u'Tim Robbins', ...|
|        9.2|       The Godfather|             R|   Crime|     175|[u'Marlon Brando'...|
|        9.1|The Godfather: Pa...|             R|   Crime|     200|[u'Al Pacino', u'...|
|        9.0|     The Dark Knight|         PG-13|  Action|     152|[u'Christian Bale...|
|        8.9|         Pulp Fiction|             R|   Crime|     154|[u'John Travolta'...|
|        8.9|            12 Angry Men|     NOT RATED|   Drama|      96|[u'Henry Fonda', ...|
|        8.9|The Good, the Bad...|     NOT RATED|  Western|     161|[u'Clint Eastwood...|
|        8.9|The Lord of the R...|         PG-13|Adventure|     201|[u'Elijah Wood', ...|
|        8.9|    Schindler's List|             R|Biography|     195|[u'Liam Neeson', ...|
|        8.9|          Fight Club|             R|   Drama|     139|[u'Brad Pitt', u'...|
|        8.8|The Lord of the R...|         PG-13|Adventure|     178|[u'Elijah Wood', ...|
|        8.8|           Inception|         PG-13|  Action|     148|[u'Leonardo DiCap...|
|        8.8|Star Wars: Episod...|            PG|  Action|     124|[u'Mark Hamill', ...|
|        8.8|         Forrest Gump|         PG-13|   Drama|     142|[u'Tom Hanks', u'...|
|        8.8|The Lord of the R...|         PG-13|Adventure|     179|[u'Elijah Wood', ...|
|        8.7|        Interstellar|         PG-13|Adventure|     169|[u'Matthew McCona...|
|        8.7|One Flew Over the...|             R|   Drama|     133|[u'Jack Nicholson...|
|        8.7|       Seven Samurai|      UNRATED|   Drama|     207|[u'Toshir\xf4 Mif...|
|        8.7|          Goodfellas|             R|Biography|     146|[u'Robert De Niro...|
|        8.7|           Star Wars|            PG|  Action|     121|[u'Mark Hamill', ...|
+-----------+--------------------+--------------+--------+--------+--------------------+
```

File Edit View Run Kernel Tabs Settings Help

Nimanshi Rustagi

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code | Python 3 (ipykernel)

Filter files by name

/

| Name | Last Modified |
|---|---|
| 515849391... | 7 years ago |
| f663366d17... | 53 minutes ago |
| hadoop-3.3... | 2 years ago |
| imdb_1000.... | 10 minutes ago |
| Untitled.ipy... | 10 seconds ago |
| weather.csv | 49 minutes ago |

```
only showing top 20 rows
```

[14]:
```
print(type(f2))
```

```
<class 'pyspark.sql.dataframe.DataFrame'>
```

[15]:
```
f2.head(3)
```

[15]:
```
[Row(star_rating='9.3', title='The Shawshank Redemption', content_rating='R', genre='Crime', duration='142', actors_list="[u'Tim R
obbins', u'Morgan Freeman', u'Bob Gunton']"),
 Row(star_rating='9.2', title='The Godfather', content_rating='R', genre='Crime', duration='175', actors_list="[u'Marlon Brando',
u'Al Pacino', u'James Caan']"),
 Row(star_rating='9.1', title='The Godfather: Part II', content_rating='R', genre='Crime', duration='200', actors_list="[u'Al Paci
no', u'Robert De Niro', u'Robert Duvall']")]
```

[16]:
```
f2.tail(3)
```

[16]:
```
[Row(star_rating='7.4', title='Master and Commander: The Far Side of the World', content_rating='PG-13', genre='Action', duration
='138', actors_list="[u'Russell Crowe', u'Paul Bettany', u'Billy Boyd']"),
 Row(star_rating='7.4', title='Poltergeist', content_rating='PG', genre='Horror', duration='114', actors_list='"[u\'JoBeth William
s\', u""Heather O\'Rourke""]"'),
 Row(star_rating='7.4', title='Wall Street', content_rating='R', genre='Crime', duration='126', actors_list="[u'Charlie Sheen',
u'Michael Douglas', u'Tamara Tunie']")]
```

[17]:
```
f2.createOrReplaceTempView("f3")
```

[18]:
```
spark.sql("select * from f3").show()
```

Simple | Python 3 (ipykernel) | Idle | Mode: Edit | Ln 1, Col 1 | Untitled.ipynb

File Edit View Run Kernel Tabs Settings Help

Nimanshi Rustagi

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code | Python 3 (ipykernel)

Filter files by name

/

| Name | Last Modified |
|---|---|
| 515849391... | 7 years ago |
| f663366d17... | 53 minutes ago |
| hadoop-3.3... | 2 years ago |
| imdb_1000.... | 10 minutes ago |
| Untitled.ipy... | 33 seconds ago |
| weather.csv | 49 minutes ago |

[17]:
```
f2.createOrReplaceTempView("f3")
```

[18]:
```
spark.sql("select * from f3").show()
```

```
+----------+--------------------+--------------+---------+--------+--------------------+
|star_rating|               title|content_rating|    genre|duration|         actors_list|
+----------+--------------------+--------------+---------+--------+--------------------+
|       9.3|The Shawshank Red...|             R|    Crime|     142|[u'Tim Robbins', ...|
|       9.2|       The Godfather|             R|    Crime|     175|[u'Marlon Brando'...|
|       9.1|The Godfather: Pa...|             R|    Crime|     200|[u'Al Pacino', u'...|
|         9|     The Dark Knight|         PG-13|   Action|     152|[u'Christian Bale...|
|       8.9|        Pulp Fiction|             R|    Crime|     154|[u'John Travolta'...|
|       8.9|        12 Angry Men|     NOT RATED|    Drama|      96|[u'Henry Fonda', ...|
|       8.9|The Good, the Bad...|     NOT RATED|  Western|     161|[u'Clint Eastwood...|
|       8.9|The Lord of the R...|         PG-13|Adventure|     201|[u'Elijah Wood', ...|
|       8.9|    Schindler's List|             R|Biography|     195|[u'Liam Neeson', ...|
|       8.9|          Fight Club|             R|    Drama|     139|[u'Brad Pitt', u'...|
|       8.8|The Lord of the R...|         PG-13|Adventure|     178|[u'Elijah Wood', ...|
|       8.8|           Inception|         PG-13|   Action|     148|[u'Leonardo DiCap...|
|       8.8|Star Wars: Episod...|            PG|   Action|     124|[u'Mark Hamill', ...|
|       8.8|        Forrest Gump|         PG-13|    Drama|     142|[u'Tom Hanks', u'...|
|       8.8|The Lord of the R...|         PG-13|Adventure|     179|[u'Elijah Wood', ...|
|       8.7|       Interstellar|         PG-13|Adventure|     169|[u'Matthew McCona...|
|       8.7|One Flew Over the...|             R|    Drama|     133|[u'Jack Nicholson...|
|       8.7|       Seven Samurai|      UNRATED|    Drama|     207|[u'Toshir\xf4 Mif...|
|       8.7|          Goodfellas|             R|Biography|     146|[u'Robert De Niro...|
|       8.7|           Star Wars|            PG|   Action|     121|[u'Mark Hamill', ...|
+----------+--------------------+--------------+---------+--------+--------------------+
```

Simple | Python 3 (ipykernel) | Idle | Mode: Edit | Ln 1, Col 1 | Untitled.ipynb

File Edit View Run Kernel Tabs Settings Help

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code

Filter files by name

/

| Name | Last Modified |
|---|---|
| 515849391... | 7 years ago |
| f663366d17... | 54 minutes ago |
| hadoop-3.3... | 2 years ago |
| imdb_1000.... | 11 minutes ago |
| Untitled.ipy... | 1 minute ago |
| weather.csv | 51 minutes ago |

[18]:
```
spark.sql("select * from f3").show()
```

```
+----------+--------------------+--------------+---------+--------+--------------------+
|star_rating|               title|content_rating|    genre|duration|         actors_list|
+----------+--------------------+--------------+---------+--------+--------------------+
|       9.3|The Shawshank Red...|             R|    Crime|     142|[u'Tim Robbins', ...|
|       9.2|       The Godfather|             R|    Crime|     175|[u'Marlon Brando'...|
|       9.1|The Godfather: Pa...|             R|    Crime|     200|[u'Al Pacino', u'...|
|         9|     The Dark Knight|         PG-13|   Action|     152|[u'Christian Bale...|
|       8.9|        Pulp Fiction|             R|    Crime|     154|[u'John Travolta'...|
|       8.9|        12 Angry Men|     NOT RATED|    Drama|      96|[u'Henry Fonda', ...|
|       8.9|The Good, the Bad...|     NOT RATED|  Western|     161|[u'Clint Eastwood...|
|       8.9|The Lord of the R...|         PG-13|Adventure|     201|[u'Elijah Wood', ...|
|       8.9|    Schindler's List|             R|Biography|     195|[u'Liam Neeson', ...|
|       8.9|          Fight Club|             R|    Drama|     139|[u'Brad Pitt', u'...|
|       8.8|The Lord of the R...|         PG-13|Adventure|     178|[u'Elijah Wood', ...|
|       8.8|           Inception|         PG-13|   Action|     148|[u'Leonardo DiCap...|
|       8.8|Star Wars: Episod...|            PG|   Action|     124|[u'Mark Hamill', ...|
|       8.8|        Forrest Gump|         PG-13|    Drama|     142|[u'Tom Hanks', u'...|
|       8.8|The Lord of the R...|         PG-13|Adventure|     179|[u'Elijah Wood', ...|
|       8.7|       Interstellar|         PG-13|Adventure|     169|[u'Matthew McCona...|
|       8.7|One Flew Over the...|             R|    Drama|     133|[u'Jack Nicholson...|
|       8.7|       Seven Samurai|      UNRATED|    Drama|     207|[u'Toshir\xf4 Mif...|
|       8.7|          Goodfellas|             R|Biography|     146|[u'Robert De Niro...|
|       8.7|           Star Wars|            PG|   Action|     121|[u'Mark Hamill', ...|
+----------+--------------------+--------------+---------+--------+--------------------+
only showing top 20 rows
```

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code

[20]:
```
spark.sql("select title from f3").show()
```

```
+--------------------+
|               title|
+--------------------+
|The Shawshank Red...|
|       The Godfather|
|The Godfather: Pa...|
|     The Dark Knight|
|        Pulp Fiction|
|        12 Angry Men|
|The Good, the Bad...|
|The Lord of the R...|
|    Schindler's List|
|          Fight Club|
|The Lord of the R...|
|           Inception|
|Star Wars: Episod...|
|        Forrest Gump|
|The Lord of the R...|
|        Interstellar|
|One Flew Over the...|
|       Seven Samurai|
|          Goodfellas|
|           Star Wars|
+--------------------+
only showing top 20 rows
```

515849391... 7 years ago
f663366d17... 57 minutes ago
hadoop-3.3... 2 years ago
imdb_1000.... 14 minutes ago
Untitled.ipy... 40 seconds ago
weather.csv 53 minutes ago

[22]:
```
spark.sql("select count(title) from f3").show()
```

```
+------------+
|count(title)|
+------------+
|         979|
+------------+
```

[ ]:

File Edit View Run Kernel Tabs Settings Help

Nimanshi Rustag

Filter files by name

Name | Last Modified

515849391... 7 years ago
f663366d17... 59 minutes ago
hadoop-3.3... 2 years ago
imdb_1000.... 16 minutes ago
Untitled.ipy... 25 seconds ago
weather.csv 55 minutes ago

Untitled.ipynb | ubuntu@ip-172-31-34-18: ~

Code | Python 3

[23]:
```
spark.sql("select star_rating,title,duration from f3 where content_rating='R'").show()
```

```
+-----------+--------------------+--------+
|star_rating|               title|duration|
+-----------+--------------------+--------+
|        9.3|The Shawshank Red...|     142|
|        9.2|       The Godfather|     175|
|        9.1|The Godfather: Pa...|     200|
|        8.9|        Pulp Fiction|     154|
|        8.9|    Schindler's List|     195|
|        8.9|          Fight Club|     139|
|        8.7|One Flew Over the...|     133|
|        8.7|          Goodfellas|     146|
|        8.7|          The Matrix|     136|
|        8.7|         City of God|     130|
|        8.7|  The Usual Suspects|     106|
|        8.7|                Se7en|     127|
|        8.6|The Silence of th...|     118|
|        8.6|Leon: The Profess...|     110|
|        8.6|    The Intouchables|     112|
|        8.6|           Whiplash|     107|
|        8.6|  American History X|     119|
|        8.6| Saving Private Ryan|     169|
|        8.6|              Psycho|     109|
```

Simple 1 1 Python 3 (ipykernel) | Idle Mode: Edit Ln 1, Col 1 Unt