

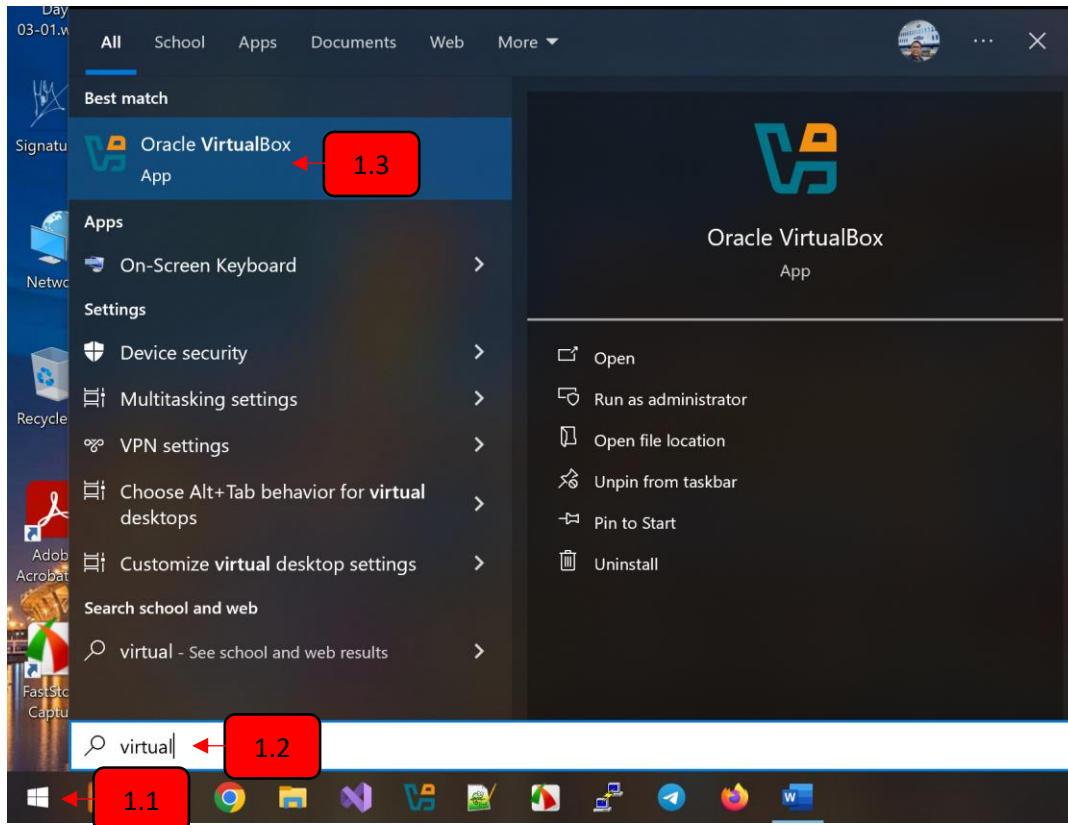
Instruction 1: Big Data Analytics Via Terminal in Ubuntu Linux

Part 1/5: Start VirtualBox and Virtual Machine

Step 1.1: Click the **Windows** Icon.

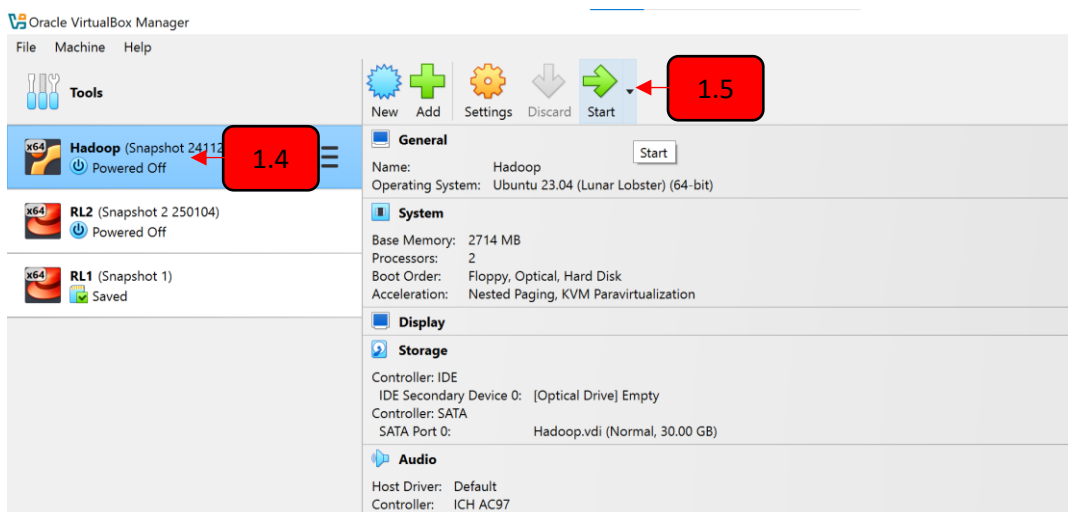
Step 1.2: Type **VirtualBox**.

Step 1.3: Click on the **Oracle VirtualBox** app.



Step 1.4: Select the **Hadoop** Virtual Machine.

Step 1.5: Click on the **Start** icon.



Part 2/5: Start Hadoop Services

Step 2.1: Click on the **Terminal** Icon on the Dash, type **sudo su -** and press the **Enter** key.

Step 2.2: Type the password **Admin1111** and press the **Enter** key.

```
root@UBUNTU24: ~
mony@UBUNTU24:~$ sudo su -
[sudo] password for mony:
root@UBUNTU24:~#
```

Step 2.3: Type `/usr/local/hadoop/sbin/start-all.sh` and press **Enter**.

```
root@UBUNTU24: ~
mony@UBUNTU24:~$ sudo su -
[sudo] password for mony:
root@UBUNTU24:~# /usr/local/hadoop/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
25/06/11 20:21:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Step 2.4: Type `jps` to check if Hadoop services are running.

```
root@UBUNTU24: /home/mony
root@UBUNTU24: /home/mony# jps
2913 NameNode
4582 Jps
3559 SecondaryNameNode
3803 NodeManager
3691 ResourceManager
3118 DataNode
root@UBUNTU24: /home/mony#
```

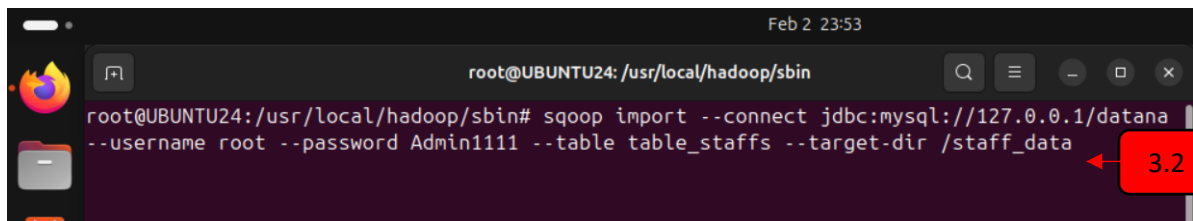
Part 3/5: Check Sqoop Version and Import Data from Database

Step 3.1: Type `sqoop version` and press **Enter**

```
root@UBUNTU24: /home/mony
root@UBUNTU24: /home/mony# sqoop version
Warning: /home/mony/Downloads/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/mony/Downloads/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/mony/Downloads/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/mony/Downloads/sqoop/../zookeeper does not exist! Accumulo imports will fail.
```

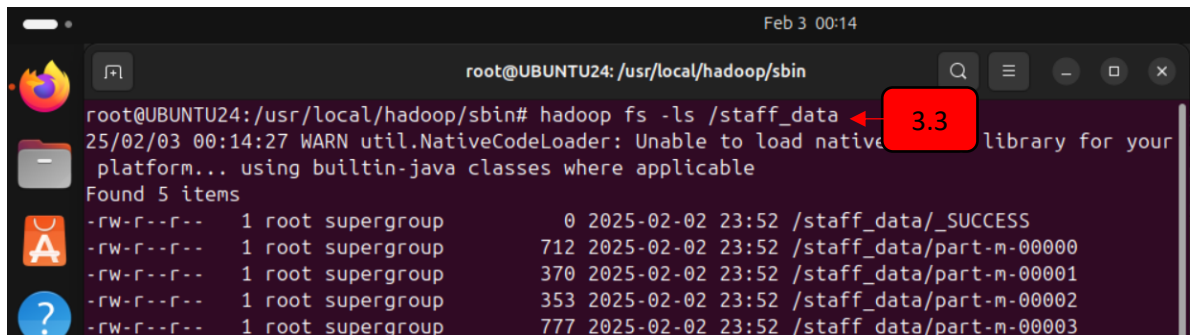
Step 3.2: type the command below and press **Enter**.

```
sqoop import --connect jdbc:mysql://127.0.0.1/dbtest --username usertest --password Admin1111
--table staff_data --target-dir /staff_data --delete-target-dir
```



```
root@UBUNTU24: /usr/local/hadoop/sbin
root@UBUNTU24: /usr/local/hadoop/sbin# sqoop import --connect jdbc:mysql://127.0.0.1/datana
--username root --password Admin1111 --table table_staffs --target-dir /staff_data
```

Step 3.3: Type **hadoop fs -ls /staff_data** and press **Enter** key to view imported data.

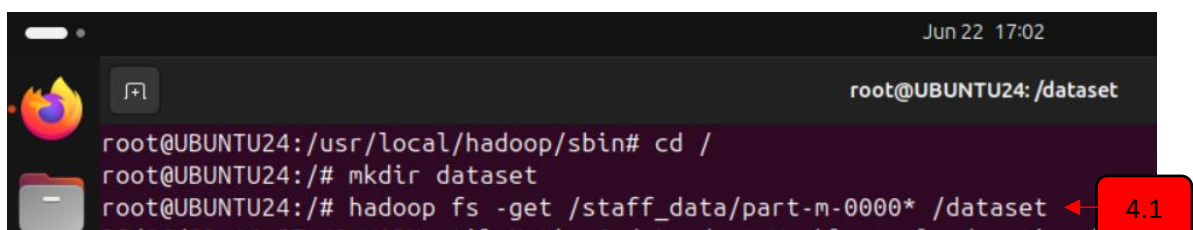


```
root@UBUNTU24: /usr/local/hadoop/sbin
root@UBUNTU24: /usr/local/hadoop/sbin# hadoop fs -ls /staff_data
25/02/03 00:14:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Found 5 items
-rw-r--r-- 1 root supergroup 0 2025-02-02 23:52 /staff_data/_SUCCESS
-rw-r--r-- 1 root supergroup 712 2025-02-02 23:52 /staff_data/part-m-00000
-rw-r--r-- 1 root supergroup 370 2025-02-02 23:52 /staff_data/part-m-00001
-rw-r--r-- 1 root supergroup 353 2025-02-02 23:52 /staff_data/part-m-00002
-rw-r--r-- 1 root supergroup 777 2025-02-02 23:52 /staff_data/part-m-00003
```

Part 4/5: Use Python to Clean Data in HDFS

Step 4.1: Type the following to import data from Hadoop to a local directory:

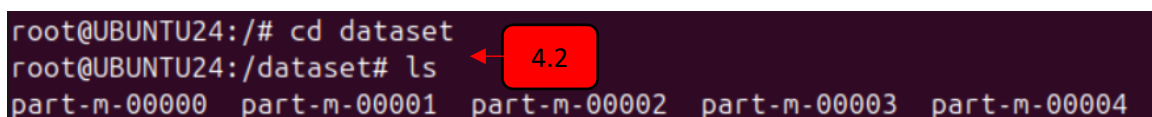
```
cd /
mkdir dataset
hadoop fs -get /staff_data/part-m-0000* /dataset
```



```
root@UBUNTU24: /usr/local/hadoop/sbin
root@UBUNTU24: /usr/local/hadoop/sbin# cd /
root@UBUNTU24: /# mkdir dataset
root@UBUNTU24: /# hadoop fs -get /staff_data/part-m-0000* /dataset
```

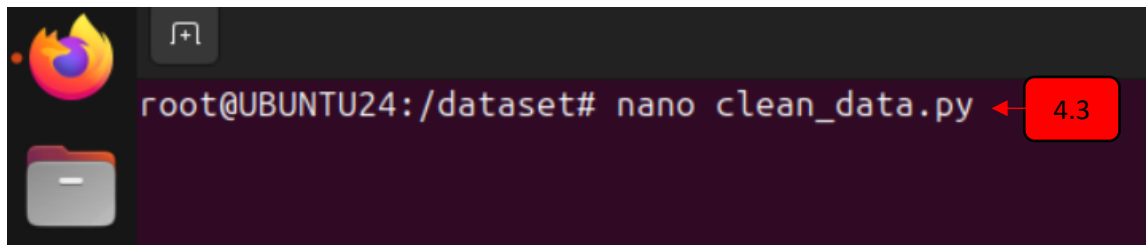
Step 4.2: Verify the imported data:

```
cd dataset
ls
```



```
root@UBUNTU24: /# cd dataset
root@UBUNTU24: /dataset# ls
part-m-00000 part-m-00001 part-m-00002 part-m-00003 part-m-00004
```

Step 4.3: Type **nano clean_data.py** to open a new Python script.



Step 4.4: Type the following Python code, then press **Ctrl + X**, then **Y**, then **Enter**:

```
import glob

# Match all part files downloaded from Hadoop

file_list = glob.glob("part-m-0000*")

cleaned_data = []

for file_name in file_list:

    with open(file_name, "r", encoding="utf-8") as file:

        for line in file:

            parts = line.strip().split(",")

            if len(parts) >= 13:

                # Extract selected fields: staff_id, staff_card_number, staff_full_name, sex, salary

                selected = parts[0:3] + [parts[5], parts[10]]

                cleaned_data.append(selected)

# Write cleaned data to a single output file

with open("staff_filtered.csv", "w", encoding="utf-8") as out:

    for row in cleaned_data:

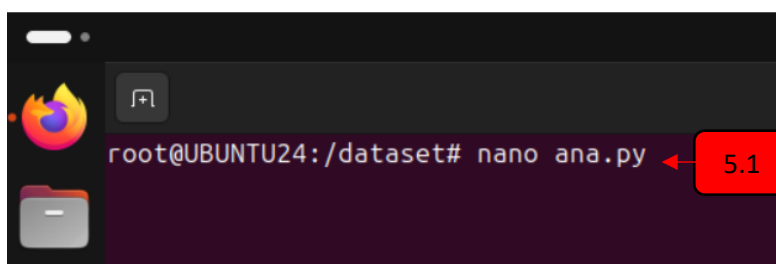
        out.write(",".join(row) + "\n")
```

Step 4.5: Run the script to generate **staff_filtered.csv** with the selected cleaned fields.

```
python3 clean_data.py
```

Part 5/5: Analyze Data Using Python

Step 5.1: Type **nano ana.py** to create a new analysis script.



Step 5.2: Type the following code:

```
with open("staff_filtered.csv", "r", encoding="utf-8") as file:

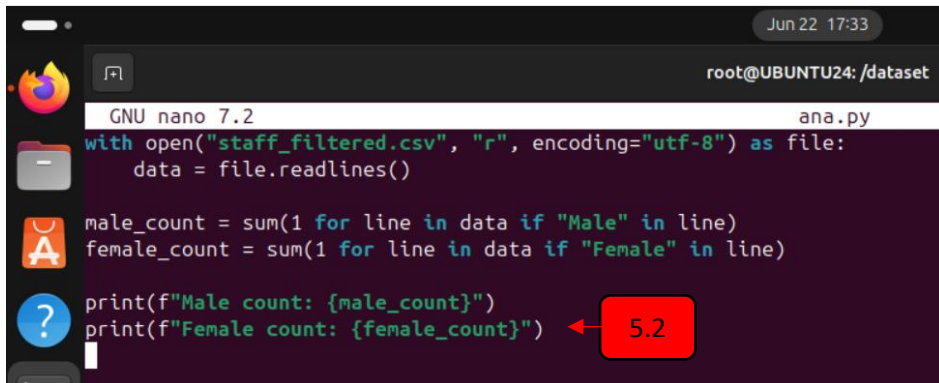
    data = file.readlines()


male_count = sum(1 for line in data if "Male" in line)

female_count = sum(1 for line in data if "Female" in line)


print(f"Male count: {male_count}")

print(f"Female count: {female_count}")
```



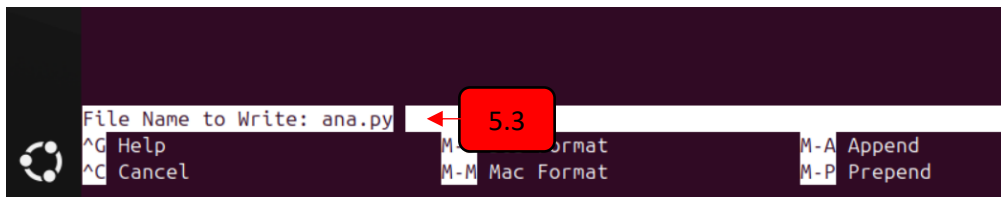
GNU nano 7.2 ana.py

```
with open("staff_filtered.csv", "r", encoding="utf-8") as file:
    data = file.readlines()

male_count = sum(1 for line in data if "Male" in line)
female_count = sum(1 for line in data if "Female" in line)

print(f"Male count: {male_count}")
print(f"Female count: {female_count}")
```

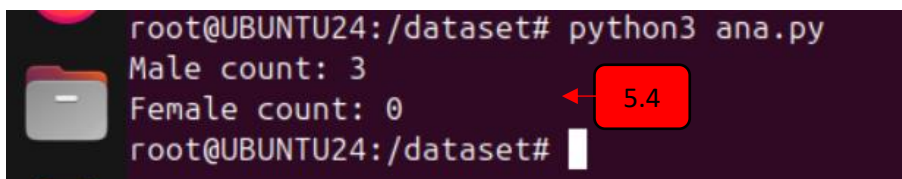
Step 5.3: Save the file by pressing **Ctrl + X**, then **Y**, then **Enter**.



File Name to Write: ana.py

^G Help ^C Cancel M-_ Format M-M Mac Format M-A Append M-P Prepend

Step 5.4: Type **python3 ana.py** and press **Enter** to run the script.



```
root@UBUNTU24:/dataset# python3 ana.py
Male count: 3
Female count: 0
root@UBUNTU24:/dataset#
```