Commands to start hadoop from terminal

1. Start HDFS (NameNode and DataNode):

```
start-dfs.sh
```

NameNode Web UI: http://localhost:9870

```
jps
```

if not started do following

1. Install OpenSSH server (if not installed)

```
sudo apt update
sudo apt install openssh-server
```

2. Enable passwordless SSH login

Run the following commands:

```
ssh-keygen -t rsa -P ""
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
```

Now test it:

```
ssh localhost
```

If it logs in without asking for a password, you're good. If it asks, let me know and we can fix it.

Start HDFS Components Manually

1. NameNode

```
hdfs --daemon start namenode
```

2. DataNode

```
hdfs --daemon start datanode
```

3. SecondaryNameNode

```
hdfs --daemon start secondarynamenode
```

Start YARN Components Manually

4. ResourceManager

```
yarn --daemon start resourcemanager
```

5. NodeManager

```
yarn --daemon start nodemanager
```

Verify Everything

Run:

```
jps
```

You should now see:

```
nginx
CopyEdit
NameNode
DataNode
SecondaryNameNode
ResourceManager
NodeManager
Jps
```

Practical 12

1. Get a Sample Dataset

You can use a simplified version of the NOAA weather dataset, or use a dummy text file like this:

sample_weather.txt

txt
CopyEdit
Date,Temperature,DewPoint,WindSpeed
2025-04-01,22.5,18.3,12.0
2025-04-01,24.0,19.0,15.5
2025-04-02,23.2,18.1,14.2
2025-04-02,25.1,20.0,13.0
2025-04-03,21.8,17.6,16.1

Each line has:

- **Date**
- **Temperature (°C)**
- **Dew Point (°C)**
- **Wind Speed (km/h)**

Mapper: mapper.py

python
CopyEdit

```
#!/usr/bin/env python3
import sys

for line in sys.stdin:
    if line.startswith("Date") or not line.strip():
        continue
    parts = line.strip().split(",")
    if len(parts) != 4:
        continue
    date, temp, dew, wind = parts
    print(f"{date}\t{temp},{dew},{wind}")
```

Reducer: reducer.py

python
CopyEdit

```
#!/usr/bin/env python3
import sys
from collections import defaultdict
```

```
data = defaultdict(lambda: [0.0, 0.0, 0.0, 0])  # sum_temp, sum_dew, sum_wind, count

for line in sys.stdin:
    date, values = line.strip().split("\t")
    temp, dew, wind = map(float, values.split(","))
    sums = data[date]
    sums[0] += temp
    sums[1] += dew
    sums[2] += wind
    sums[3] += 1

for date, (temp_sum, dew_sum, wind_sum, count) in data.items():
    print(f"{date}\tAvgTemp: {temp_sum/count:.2f}, AvgDew: {dew_sum/count:.2f}, AvgWind: {wind_sum/count:.2f}")
```

Upload the File to HDFS

You need to copy your input file from local to HDFS like this:

```
hadoop fs -mkdir -p /user/te
hadoop fs -put sample_weather.txt /user/te/
```

Then run the job again with the **HDFS path**:

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
  -input /user/te/sample_weather.txt \
  -output /user/te/weather_output \
  -mapper ./mapper.py \
  -reducer ./reducer.py
```

If Output Directory Already Exists

Hadoop doesn't overwrite output folders, so if you get an error like "Output directory already exists", remove it first:

```
hadoop fs -rm -r /user/te/weather_output
```

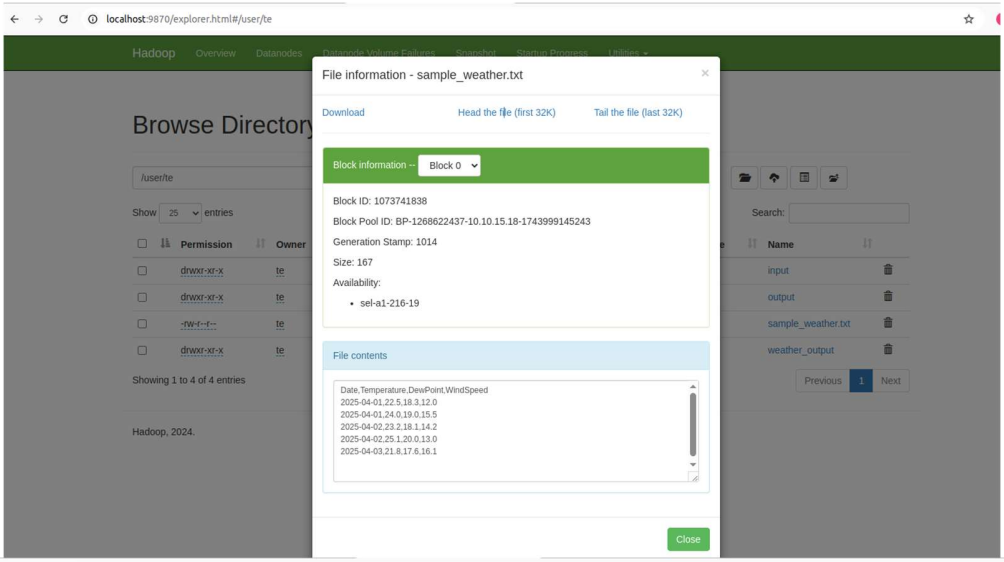
To View the Result

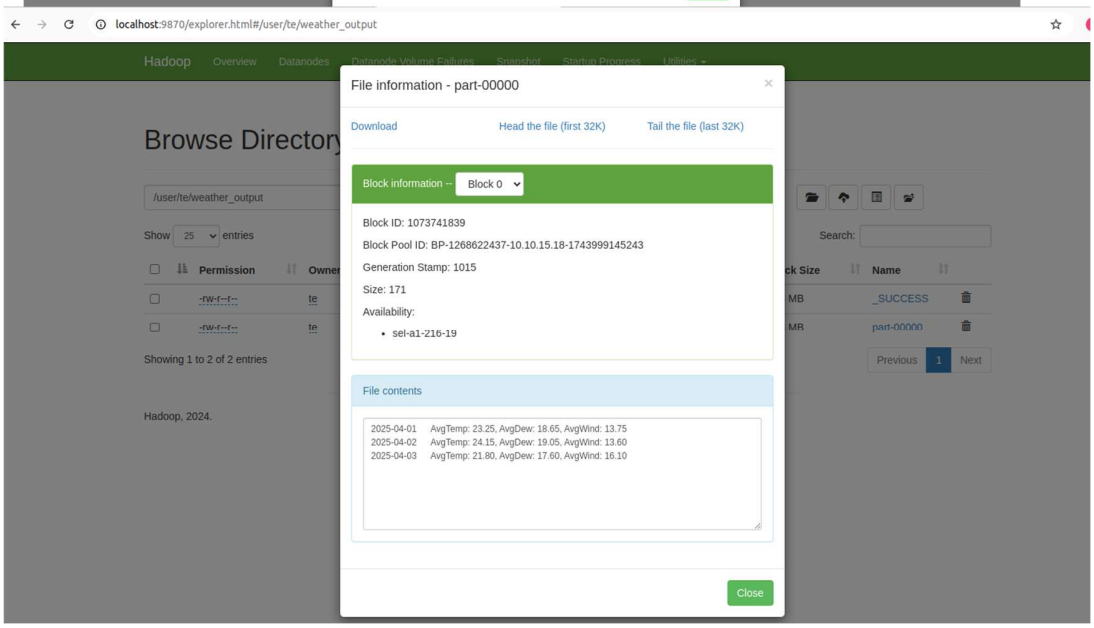
Once the job completes:

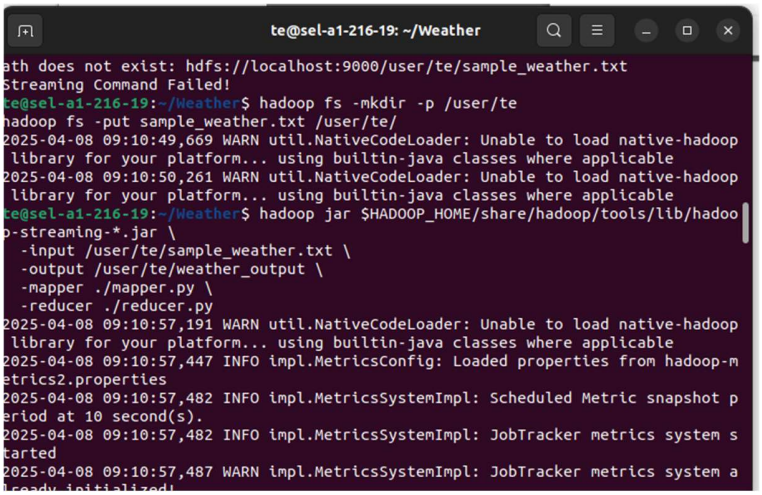
```
hadoop fs -cat /user/te/weather_output/part-00000
```

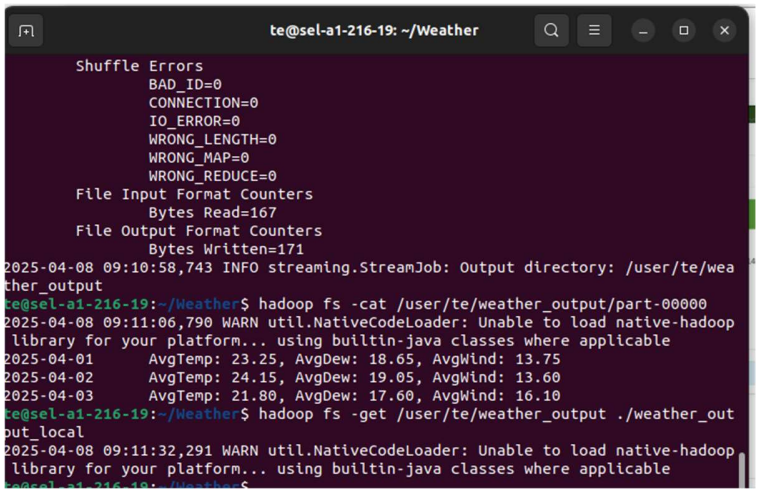
Download Output from HDFS to Local

```
hadoop fs -get /user/te/weather_output ./weather_output_local
```


localhost:9870/explorer.html#/user/te
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/user/te
Show 25 entries
Search:
Permission
Owner
Name
drwxr-xr-x
te
input
output
sample_weather.txt
weather_output
-rw-r--r--
Showing 1 to 4 of 4 entries
Previous
1
Next
Hadoop, 2024.
File information - sample_weather.txt
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information --
Block 0
Block ID: 1073741838
Block Pool ID: BP-1268622437-10.10.15.18-1743999145243
Generation Stamp: 1014
Size: 167
Availability:
sel-a1-216-19
File contents
Date,Temperature,DewPoint,WindSpeed
2025-04-01,22.5,18.3,12.0
2025-04-01,24.0,19.0,15.5
2025-04-02,23.2,18.1,14.2
2025-04-02,25.1,20.0,13.0
2025-04-03,21.8,17.6,16.1
Close


localhost:9870/explorer.html#/user/te/weather_output
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/user/te/weather_output
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
te
_SUCCESS
part-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2024.
File information - part-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information --
Block 0
Block ID: 1073741839
Block Pool ID: BP-1268622437-10.10.15.18-1743999145243
Generation Stamp: 1015
Size: 171
Availability:
sel-a1-216-19
File contents
2025-04-01 AvgTemp: 23.25, AvgDew: 18.65, AvgWind: 13.75
2025-04-02 AvgTemp: 24.15, AvgDew: 19.05, AvgWind: 13.60
2025-04-03 AvgTemp: 21.80, AvgDew: 17.60, AvgWind: 16.10
Close


te@sel-a1-216-19: ~/Weather
ath does not exist: hdfs://localhost:9000/user/te/sample_weather.txt
Streaming Command Failed!
te@sel-a1-216-19:~/Weather$ hadoop fs -mkdir -p /user/te
hadoop fs -put sample_weather.txt /user/te/
2025-04-08 09:10:49,669 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
2025-04-08 09:10:50,261 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
te@sel-a1-216-19:~/Weather$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoo
p-streaming-*.jar \
-input /user/te/sample_weather.txt \
-output /user/te/weather_output \
-mapper ./mapper.py \
-reducer ./reducer.py
2025-04-08 09:10:57,191 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
2025-04-08 09:10:57,447 INFO impl.MetricsConfig: Loaded properties from hadoop-m
etrics2.properties
2025-04-08 09:10:57,482 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot p
eriod at 10 second(s).
2025-04-08 09:10:57,482 INFO impl.MetricsSystemImpl: JobTracker metrics system s
tarted
2025-04-08 09:10:57,487 WARN impl.MetricsSystemImpl: JobTracker metrics system a


te@sel-a1-216-19: ~/Weather
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=167
File Output Format Counters
Bytes Written=171
2025-04-08 09:10:58,743 INFO streaming.StreamJob: Output directory: /user/te/wea
ther_output
te@sel-a1-216-19:~/Weather$ hadoop fs -cat /user/te/weather_output/part-00000
2025-04-08 09:11:06,790 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
2025-04-01 AvgTemp: 23.25, AvgDew: 18.65, AvgWind: 13.75
2025-04-02 AvgTemp: 24.15, AvgDew: 19.05, AvgWind: 13.60
2025-04-03 AvgTemp: 21.80, AvgDew: 17.60, AvgWind: 16.10
te@sel-a1-216-19:~/Weather$ hadoop fs -get /user/te/weather_output ./weather_out
put_local
2025-04-08 09:11:32,291 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable