BIG DATA FUNDAMENTALS ASSIGNMENT

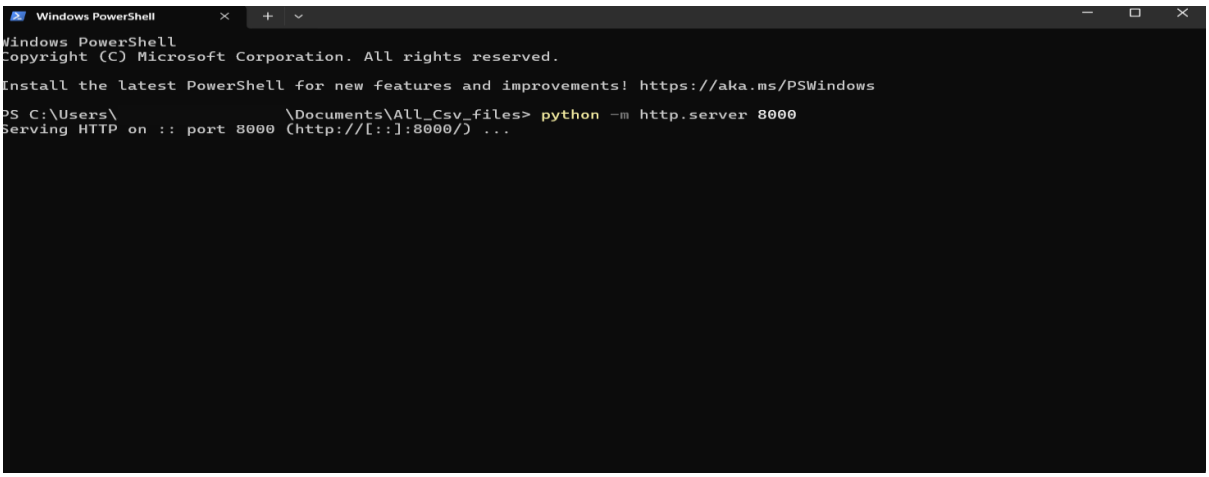
Name : Ganesh Agrahari

Section:Bca ds&ai 23

Roll No. :24

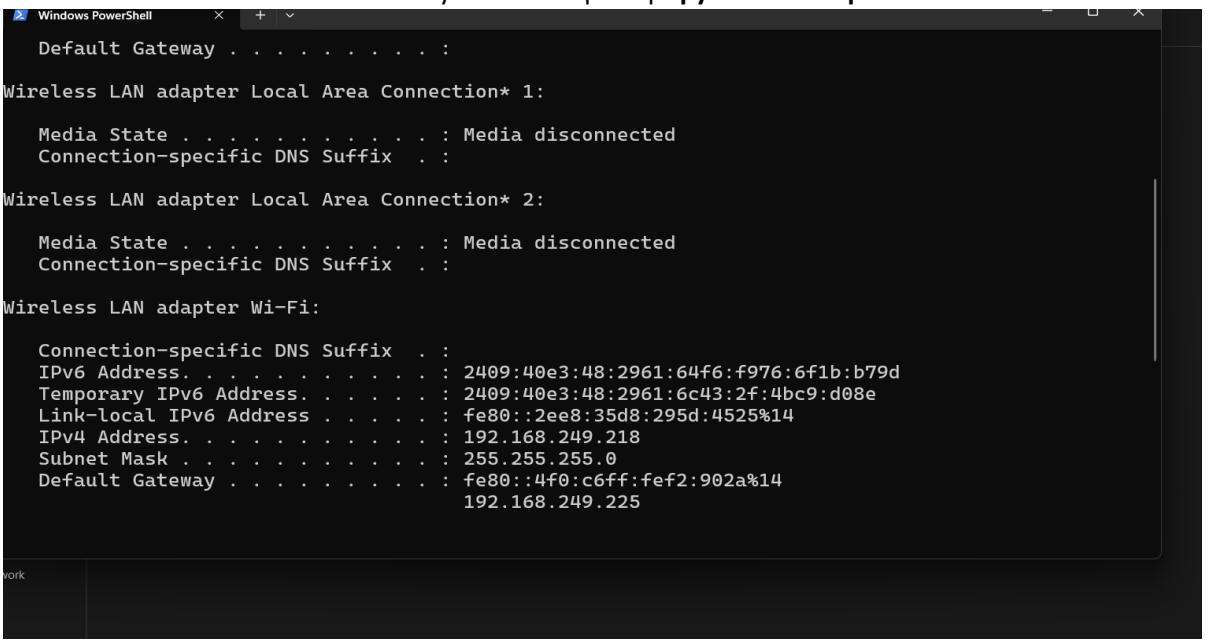
Hive

1. First of all I downloaded the dataset of**merged_dataset.csv** from Kaggle dataset
2. Created a folder in my local machine (windows) **All_Csv_File** and in that folder I move the**merged_dataset.csv**file



[uploading the file on server 8000]

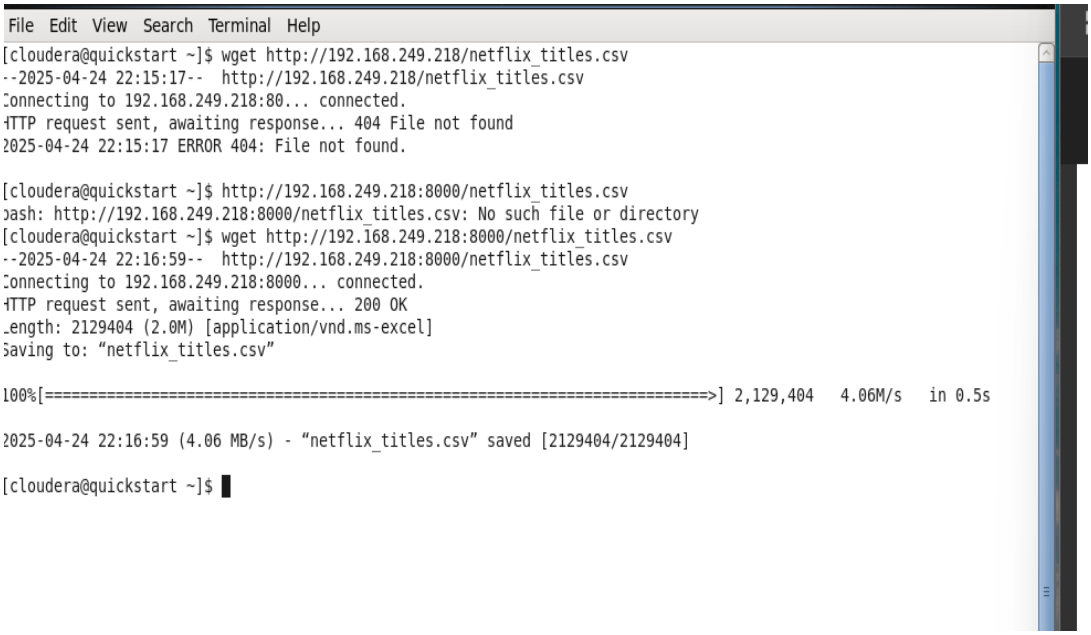
After that I I need to upload this file on **Hadoop Cloudera Manager**so to upload or paste this file on that cloudera I use my command prompt **python -m http.server 8000**



1. After successfully we need to get that files on cloudera so on cloudera terminal we run the command
2. **wget http:// 192.168.249.218/merged_dataset.csv**

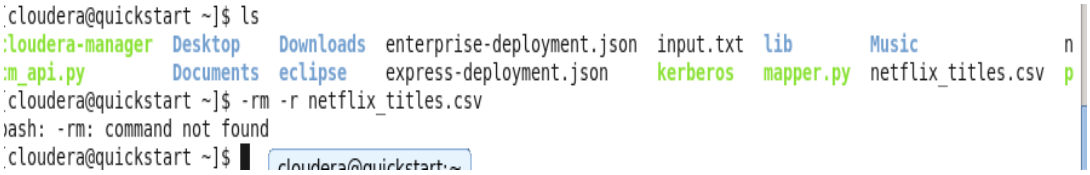
192.168.249.218:- is a IPV4 address of my computer

Note:- I have run this command two times that's why there is a two file of csv are present in my cloudera second name file is **movies**

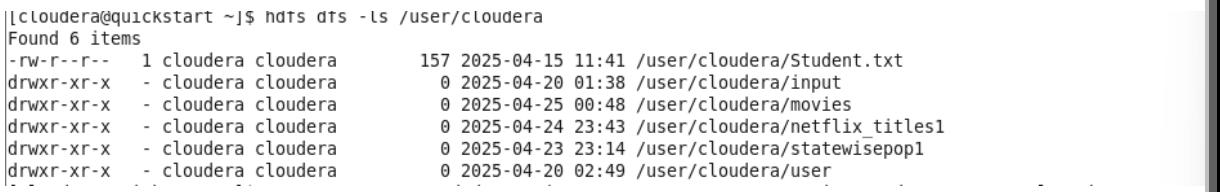


[verifying the folder has import or not]

ON cloudera terminal by doing **ls cmd** I verify the **merged_dataset.csv** has import or not



hdfs dfs –ls /user/cloudera



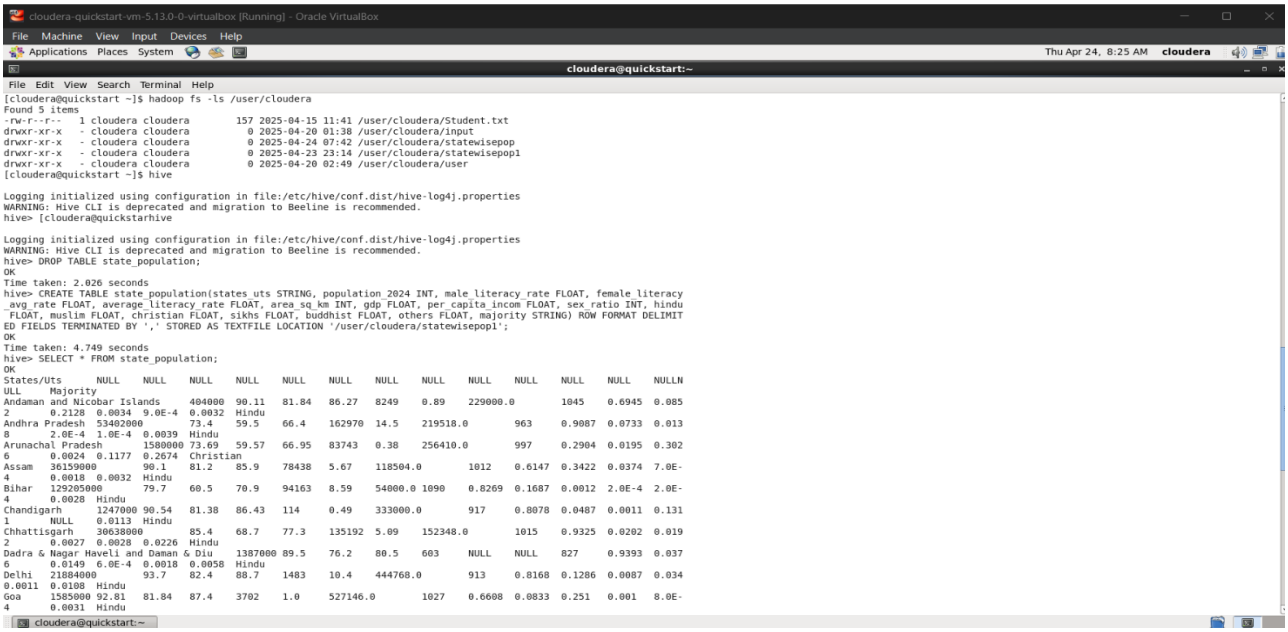
Upload the**merged_dataset.csv** to the Hadoop so, first I've made a folder Hadoop fs -mkdir /user/cloudera/moviesAnd put the **merged_dataset.csv** file in that folder by Hadoop fs -put **merged_dataset.csv** /user/cloudera/movies

[now go to hive]

By just clicking hive we can go to hive (sometimes it take time)

```
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ wget http://192.168.249.218/netflix_titles.csv
--2025-04-24 22:15:17--  http://192.168.249.218/netflix_titles.csv
Connecting to 192.168.249.218:80... connected.
HTTP request sent, awaiting response... 404 File not found
2025-04-24 22:15:17 ERROR 404: File not found.

[cloudera@quickstart ~]$ http://192.168.249.218:8000/netflix_titles.csv
bash: http://192.168.249.218:8000/netflix_titles.csv: No such file or directory
[cloudera@quickstart ~]$ wget http://192.168.249.218:8000/netflix_titles.csv
--2025-04-24 22:16:59--  http://192.168.249.218:8000/netflix_titles.csv
Connecting to 192.168.249.218:8000... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2129404 (2.0M) [application/vnd.ms-excel]
Saving to: "netflix_titles.csv"

100%[======================================================================>] 2,129,404   4.06M/s   in 0.5s

2025-04-24 22:16:59 (4.06 MB/s) - "netflix_titles.csv" saved [2129404/2129404]

[cloudera@quickstart ~]$ ls
cloudera-manager  enterprise-deployment.json  Music              Public           Videos
cm_api.py         express-deployment.json     netflix_titles.csv reducer.py       WordCount.jar
Desktop           input.txt                   parcels            sample.csv       WordCount.java
Documents         kerberos                    Pictures           statewisepop.csv workspace
Downloads         lib                         Placement_Data.csv statewisepop.csv.1
eclipse           mapper.py                   Projectsfolder     Templates
[cloudera@quickstart ~]$ hdfs dfs -mkdir/user/cloudera/netflix_titles.csv
-mkdir/user/cloudera/netflix_titles.csv: Unknown command
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/cloudera/netflix_titles
[cloudera@quickstart ~]$ hdfs dfs -mkdir ls
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 6 items
-rw-r--r--   1 cloudera cloudera        157 2025-04-15 11:41 Student.txt
drwxr-xr-x   - cloudera cloudera          0 2025-04-20 01:38 input
drwxr-xr-x   - cloudera cloudera          0 2025-04-24 22:21 ls
drwxr-xr-x   - cloudera cloudera          0 2025-04-24 22:20 netflix_titles
drwxr-xr-x   - cloudera cloudera          0 2025-04-23 23:14 statewisepop1
drwxr-xr-x   - cloudera cloudera          0 2025-04-20 02:49 user
[cloudera@quickstart ~]$ hdfs dfs -rm -r ls
Deleted ls
[cloudera@quickstart ~]$ hdfs dfs -put netflix_titles.csv /user/cloudera/netflix_titles
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive>
```

[query commands]



SELECT * FROM merged_dataset LIMIT 5;

```
hive> SELECT *  FROM  merged_dataset LIMIT 5;
OK
name    NULL    movie_rated     run_length      genres  release_date    NULL    NULL    NULL
Inception       2010    PG-13   2h 28min        Action; Adventure; Sci-Fi;      16 July 2010 (USA)      8.8     19816
75      3820
The Lord of the Rings: The Fellowship of the Ring       2001    PG-13   2h 58min        Action; Adventure; Drama;   1
9 December 2001 (USA)   8.8     1609165 5365
The Lord of the Rings: The Return of the King   2003    PG-13   3h 21min        Adventure; Drama; Fantasy;      17 De
cember 2003 (USA)       8.9     1593859 3681
The Dark Knight Rises   2012    PG-13   2h 44min        Action; Adventure;      20 July 2012 (USA)      8.4     14703
29      2979
Time taken: 0.133 seconds, Fetched: 5 row(s)
hive>
```

SELECT COUNT(*) FROM merged_dataset;

```
Time taken: 0.133 seconds, Fetched: 5 row(s)
hive> SELECT COUNT (*) FROM merged_dataset;
Query ID = cloudera_20250425013232_e15fcd02-c66a-40c8-ac21-127fb701ba9b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1745503643740_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1745503643740
_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1745503643740_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-04-25 01:32:42,427 Stage-1 map = 0%,  reduce = 0%
2025-04-25 01:32:50,996 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.97 sec
2025-04-25 01:33:09,405 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.07 sec
MapReduce Total cumulative CPU time: 3 seconds 70 msec
Ended Job = job_1745503643740_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.07 sec   HDFS Read: 149321 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 70 msec
OK
1501
Time taken: 42.861 seconds, Fetched: 1 row(s)
hive>
```

SELECT name, COUNT(*) AS total

FROM merged_dataset

GROUP BY year;

```
hive> SELECT type , COUNT(*) AS total FROM merged_dataset GROUP BY type ;
FAILED: SemanticException [Error 10004]: Line 1:61 Invalid table alias or column reference 'type': (possible column n
ames are: name, year, movie_rated, run_length, genres, release_date, rating, num_raters, num_reviews)
hive>
```

merged_dataset to create the following output .

Output of the command SELECT * FROM **merged_dataset** ;

```
Jon Jon 2013    R       1h 30min        Comedy; Romance;        27 September 2013 (USA) 6.5     224896  338
The Fountain    2006    PG-13   1h 37min        Romance;        22 November 2006 (USA)  7.2     224185  985
Marriage Story  2019    R       2h 17min        Comedy; Romance;        6 December 2019 (USA)   8.0     219894  1178
Bridget Jones's Diary   2001    R       1h 37min        Comedy; Romance;        13 April 2001 (USA)     6.7     21871
5       588
The Beach       2000    R       1h 59min        Romance;        11 February 2000 (USA)  6.7     218197  587
Warm Bodies     2013    PG-13   1h 38min        Comedy; Romance;        1 February 2013 (USA)   6.9     215679  378
Just Go with It 2011    PG-13   1h 57min        Comedy; Romance;        11 February 2011 (USA)  6.4     215475  271
Singin' in the Rain     1952    G       1h 43min        Comedy; Romance;        11 April 1952 (USA)     8.3     21137
0       664
Mr. Nobody      2009    R       2h 21min        Romance;        26 September 2013 (USA) 7.8     210352  393
Meet Joe Black  1998    PG-13   2h 58min        Romance;        13 November 1998 (USA)  7.2     208791  799
Mamma Mia!      2008    PG-13   1h 48min        Comedy; Romance;        18 July 2008 (USA)      6.4     208596  752
No Strings Attached     2011    R       1h 48min        Comedy; Romance;        21 January 2011 (USA)   6.2     20828
9       214
The Lobster     2015    R       1h 59min        Comedy; Romance;        16 October 2015 (Ireland)       7.2     20745
4       646
Romeo + Juliet  1996    PG-13   2h      Romance;        1 November 1996 (USA)   6.7     205296  576
Garden State    2004    R       1h 42min        Comedy; Romance;        20 August 2004 (USA)    7.4     204812  973
How to Lose a Guy in 10 Days    2003    PG-13   1h 56min        Comedy; Romance;        7 February 2003 (USA)   6.4 2
03365   383
Match Point     2005    R       2h 4min Romance;        20 January 2006 (USA)   7.6     201777  786
True Romance    1993    R       1h 59min        Romance;        10 September 1993 (USA) 7.9     200842  537
The Ugly Truth  2009    R       1h 36min        Comedy; Romance;        24 July 2009 (USA)      6.4     199488  195
Closer  2004    R       1h 44min        Romance;        3 December 2004 (USA)   7.2     198838  952
The Girl Next Door      2004    R       1h 49min        Comedy; Romance;        9 April 2004 (USA)      6.7     19748
9       540
Me Before You   2016    PG-13   1h 46min        Romance;        3 June 2016 (USA)       7.4     197474  457
P.S. I Love You 2007    PG-13   2h 6min Comedy; Romance;        21 December 2007 (USA)  7.0     197346  276
Call Me by Your Name    2017    R       2h 12min        Romance;        19 January 2018 (USA)   7.9     195583  951
A Walk to Remember      2002    PG      1h 41min        Romance;        25 January 2002 (USA)   7.4     195515  1010
Dirty Dancing   1987    PG-13   1h 40min        Romance;        21 August 1987 (USA)    7.0     194838  462
"I Love You     NULL    2009    R       1h 45min        Comedy; Romance;        NULL    7       191014
Ghost   1990    PG-13   2h 7min Romance;        13 July 1990 (USA)      7.0     190746  301
Revolutionary Road      2008    R       1h 59min        Romance;        23 January 2009 (USA)   7.3     190069  486
When Harry Met Sally... 1989    R       1h 35min        Comedy; Romance;        21 July 1989 (USA)      7.6     18907
9       326
Bad Teacher     2011    R       1h 32min        Comedy; Romance;        24 June 2011 (USA)      5.6     189027  325
Inception       2010    PG-13   2h 28min        Action; Adventure; Sci-Fi;      16 July 2010 (USA)      8.8     19816
88      3820
The Matrix      1999    R       2h 16min        Action; Sci-Fi;         31 March 1999 (USA)     8.7     1619772 4281
Interstellar    2014    PG-13   2h 49min        Adventure; Drama; Sci-Fi;       7 November 2014 (USA)   8.6     14308
10      3977
The Avengers    2012    PG-13   2h 23min        Action; Adventure; Sci-Fi;      4 May 2012 (USA)        8.0     12360
53      1929
The Prestige    2006    PG-13   2h 10min        Drama; Mystery; Sci-Fi;         20 October 2006 (USA)   8.5     11445
21      1394
Back to the Future      1985    PG      1h 56min        Adventure; Comedy; Sci-Fi;      3 July 1985 (USA)       8.5 1
```

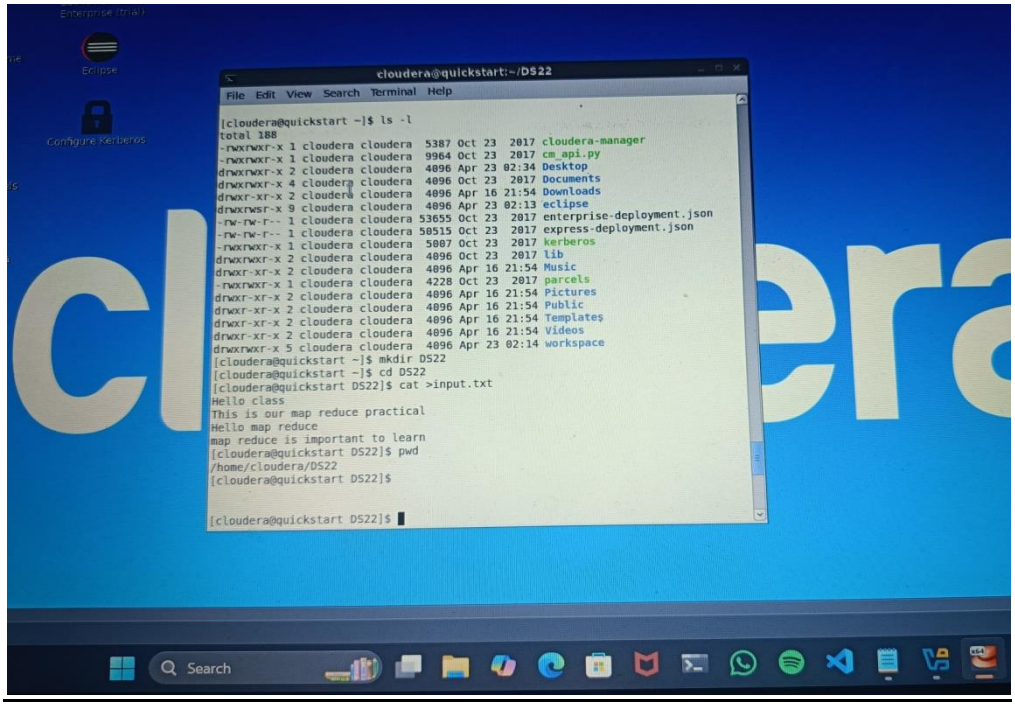
MapReduce

Use the **l s–l** command to show the list of the files

Create the directory DS22 by **mkdir**.

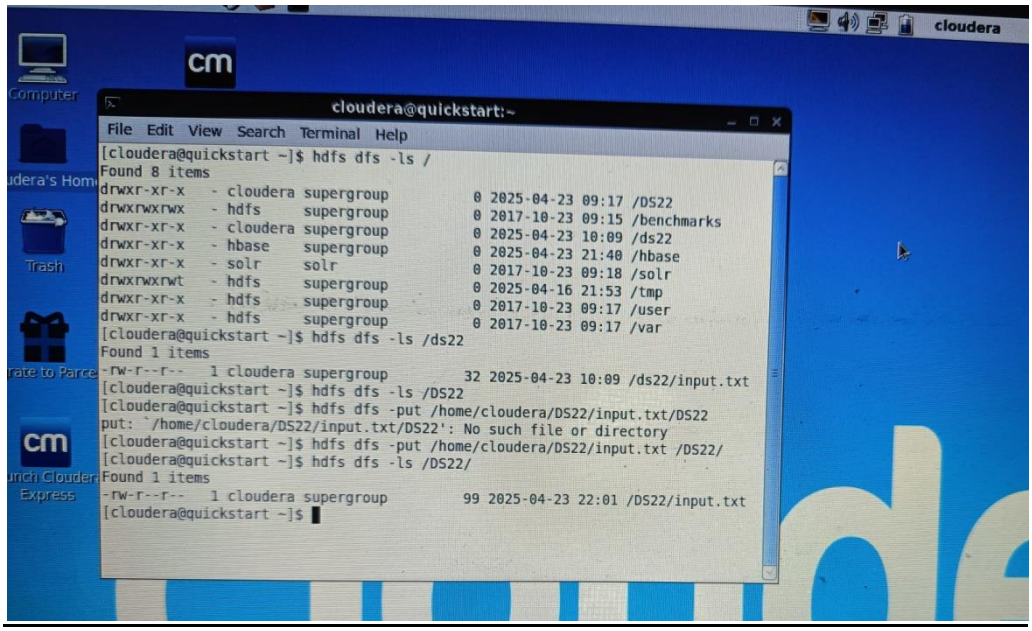
Change the directory to DS22 by **cd** command

Use cat >input.txt to create the file named **input.txt** by the **cat** command and enter the data.

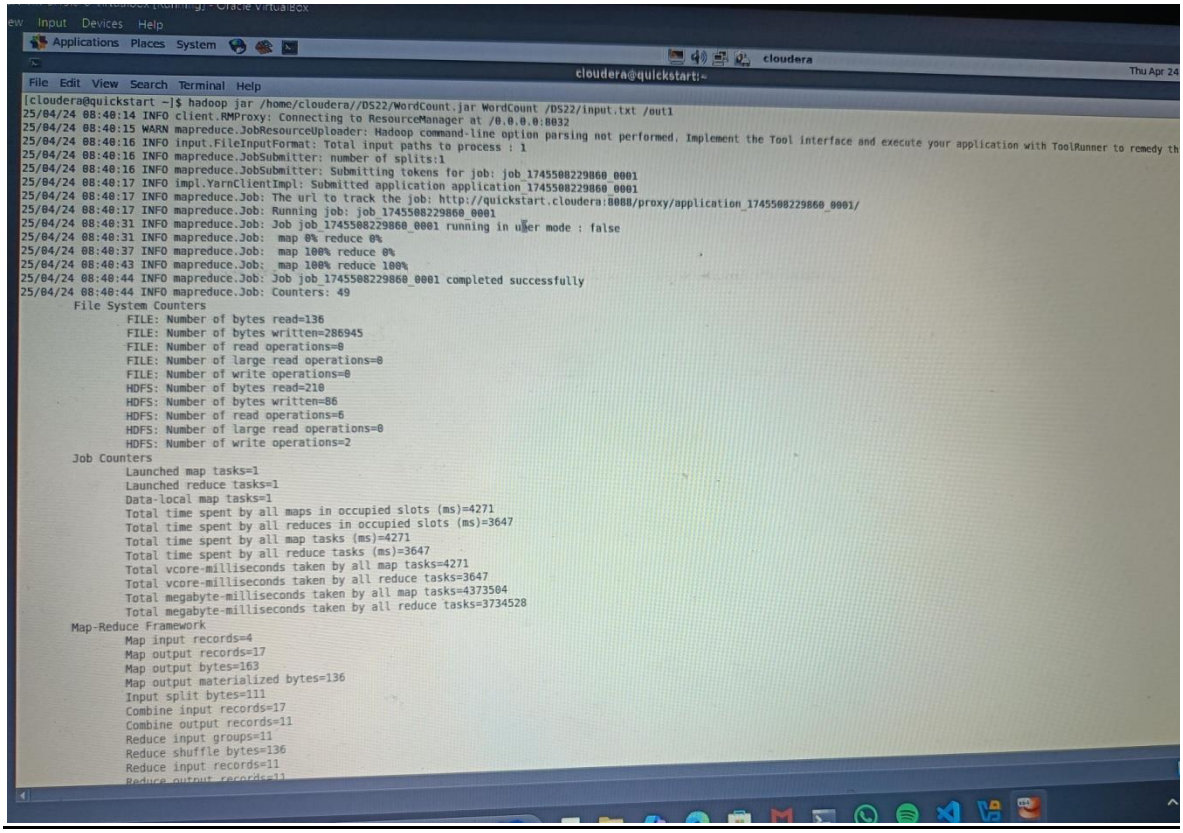


use **put** cmd to put data of input.txt to DS22.

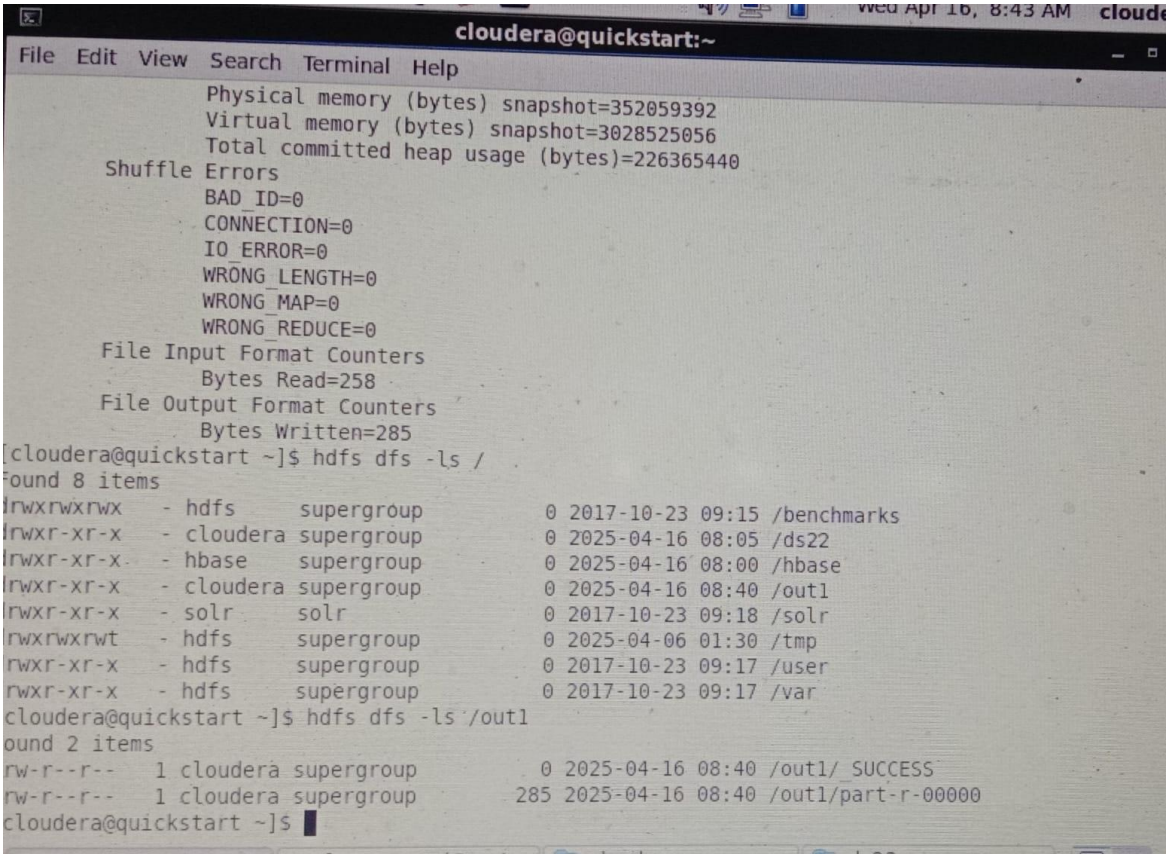
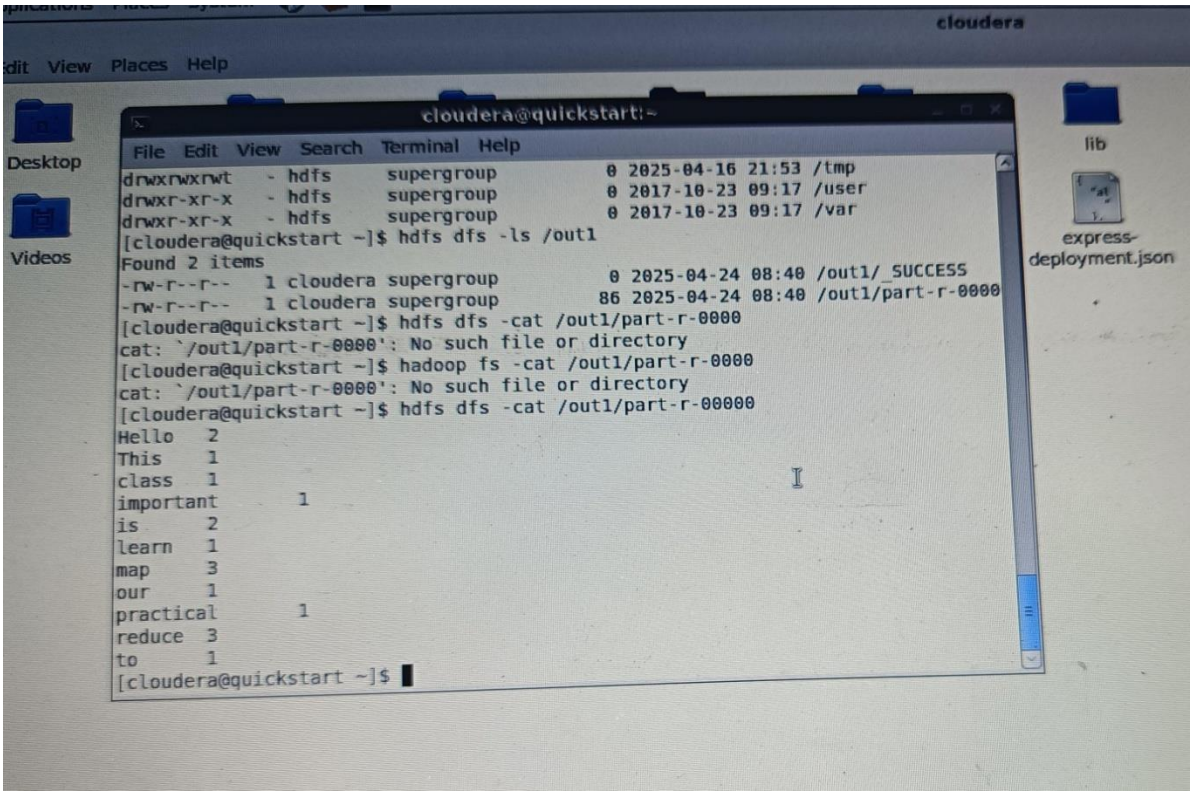
Then check the the data by ls.



Imported the **mapReduce** framework in to perform the the the task of mapReduce work.



check the directory by cmd **hdfs dfs –ls /out1** (created a out1 to store the output).



Use the cmd **hdfs dfs -cat /out1/part-r-00000** to get the output of map_reduce .