

Hadoop project

Ninjaz

Lina Almansour
Zainab Alibrahim
Esraa Alyoubi
Alaa bin Othman
Amjad Saleem



Agenda



- 1 THE DATASET

- 2 RATING FILES

- 3 MOVIES FILES

- 4 USERS FILES

- 5 INSTALLING DATASET INTO HDFS

- 6 RUNNING HADOOP CLUSTER ON A LARGER DATASET

- 7 MAPREDUCE PROGRAMMING CHALLENGE

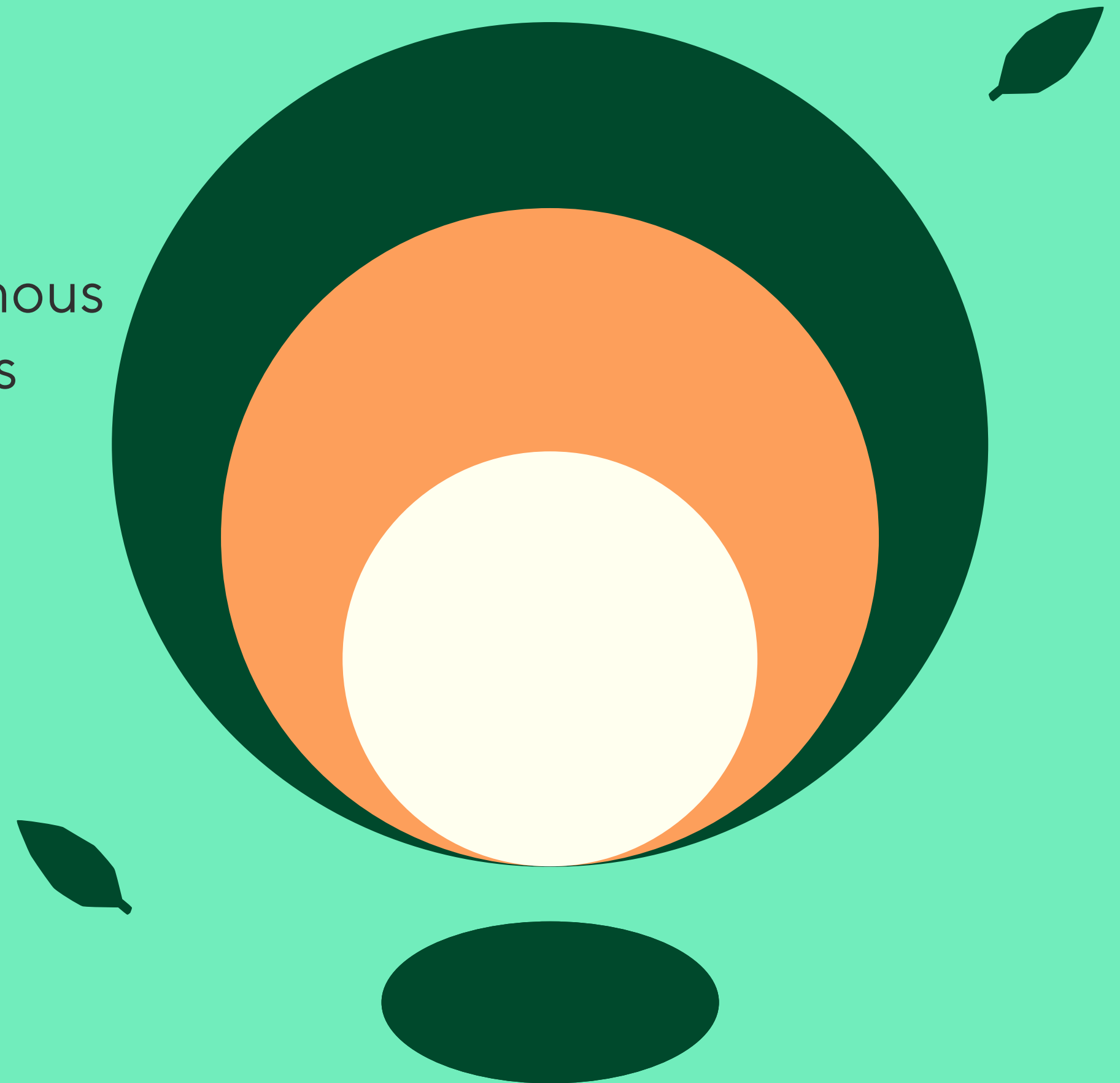
The Dataset

These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000.

Rating files

Users files

Movies files

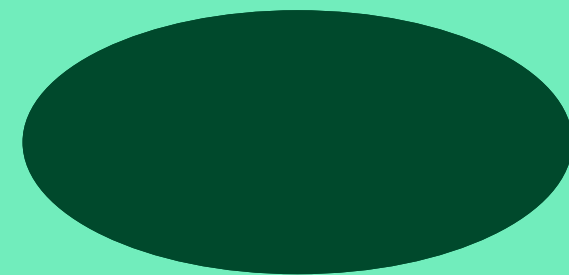


Rating files:



UserID::MovieID::Rating::Timestamp

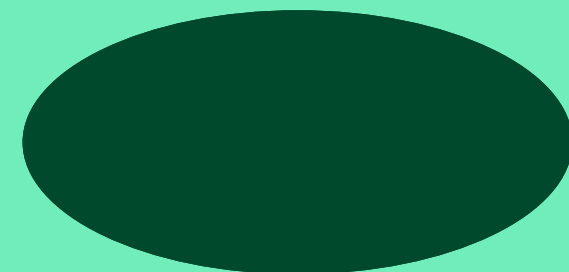
- UserIDs range between 1 and 6040
- MovieIDs range between 1 and 3952
- Ratings are made on a 5-star scale (whole-star ratings only)
- Timestamp is represented in seconds since the epoch as returned by time(2)
- Each user has at least 20 ratings



Users files:



- Gender is denoted by a "M" for male and "F" for female
- Age is chosen from the following ranges:
 - * 1: "Under 18"
 - * 18: "18-24"
 - * 25: "25-34"
 - * 35: "35-44"
 - * 45: "45-49"
 - * 50: "50-55"
 - * 56: "56+"
- Occupation

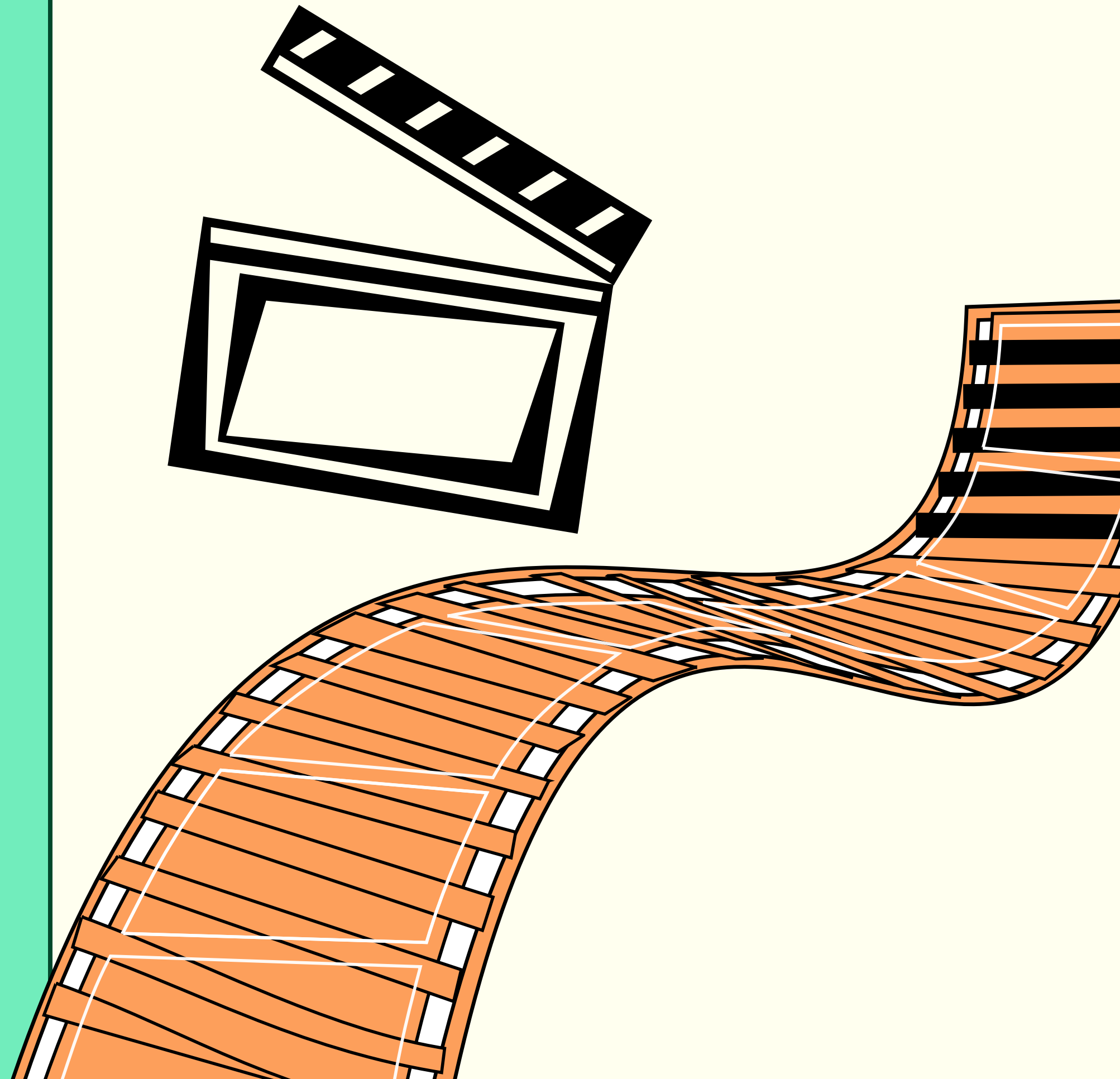
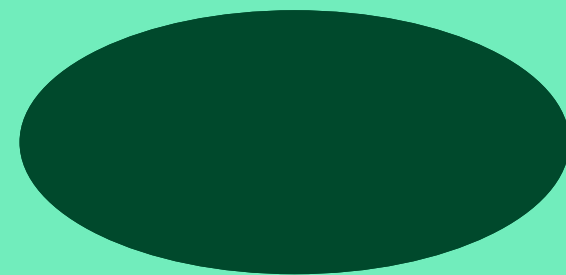




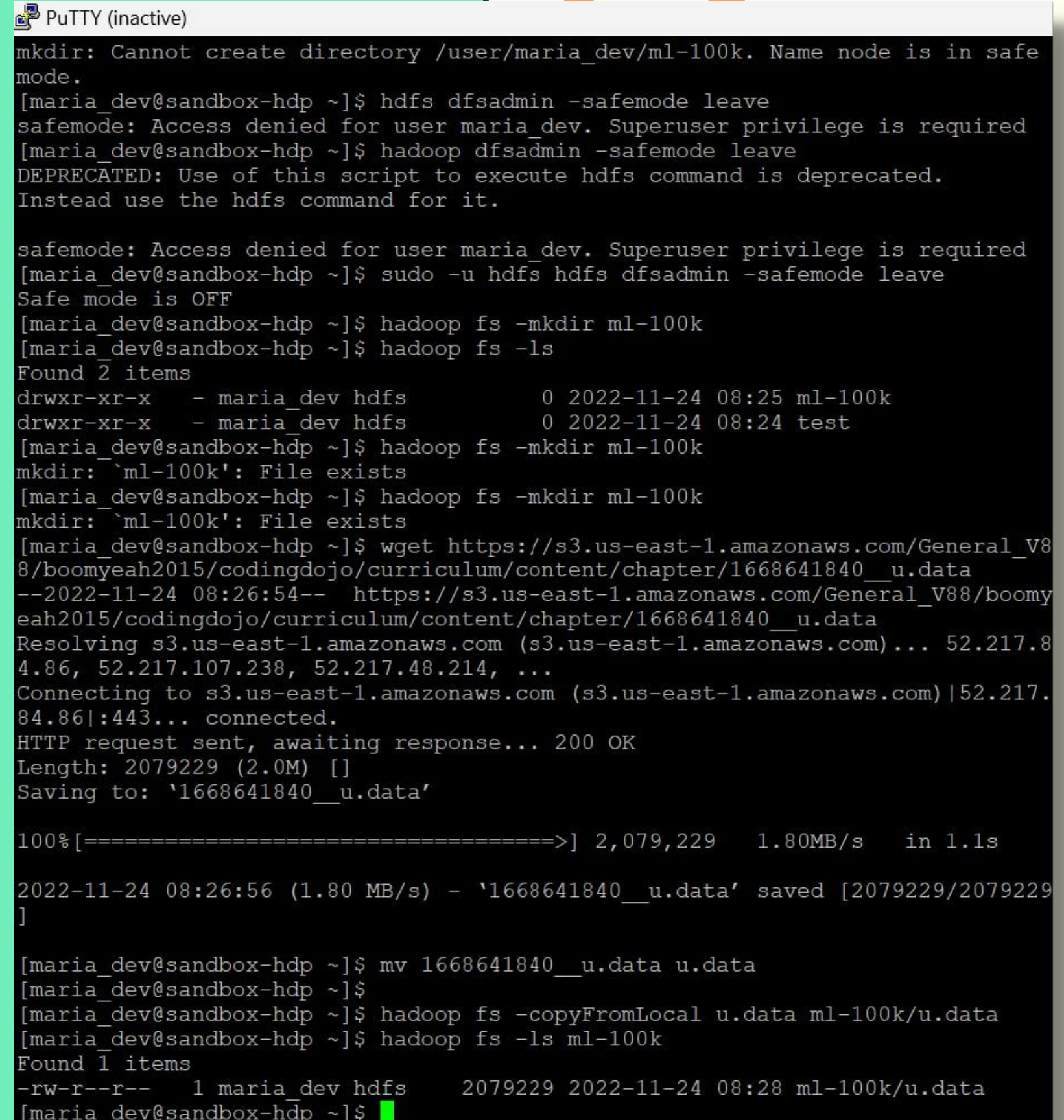
Movies files:

MovieID::Title::Genres

- Titles are identical to titles provided by the IMDB (including year of release)
- Genres are pipe-separated



Installing Dataset into HDFS



```
PuTTY (inactive)
mkdir: Cannot create directory /user/maria_dev/ml-100k. Name node is in safe
mode.
[maria_dev@sandbox-hdp ~]$ hdfs dfsadmin -safemode leave
safemode: Access denied for user maria_dev. Superuser privilege is required
[maria_dev@sandbox-hdp ~]$ hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

safemode: Access denied for user maria_dev. Superuser privilege is required
[maria_dev@sandbox-hdp ~]$ sudo -u hdfs hdfs dfsadmin -safemode leave
Safe mode is OFF
[maria_dev@sandbox-hdp ~]$ hadoop fs -mkdir ml-100k
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x  - maria_dev hdfs          0 2022-11-24 08:25 ml-100k
drwxr-xr-x  - maria_dev hdfs          0 2022-11-24 08:24 test
[maria_dev@sandbox-hdp ~]$ hadoop fs -mkdir ml-100k
mkdir: `ml-100k': File exists
[maria_dev@sandbox-hdp ~]$ hadoop fs -mkdir ml-100k
mkdir: `ml-100k': File exists
[maria_dev@sandbox-hdp ~]$ wget https://s3.us-east-1.amazonaws.com/General_V8
8/boomyeah2015/codingdojo/curriculum/content/chapter/1668641840__u.data
--2022-11-24 08:26:54-- https://s3.us-east-1.amazonaws.com/General_V88/boomye
eah2015/codingdojo/curriculum/content/chapter/1668641840__u.data
Resolving s3.us-east-1.amazonaws.com (s3.us-east-1.amazonaws.com)... 52.217.8
4.86, 52.217.107.238, 52.217.48.214, ...
Connecting to s3.us-east-1.amazonaws.com (s3.us-east-1.amazonaws.com)|52.217.
84.86|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2079229 (2.0M) []
Saving to: `1668641840__u.data'

100%[=====>] 2,079,229  1.80MB/s  in 1.1s

2022-11-24 08:26:56 (1.80 MB/s) - `1668641840__u.data' saved [2079229/2079229
]

[maria_dev@sandbox-hdp ~]$ mv 1668641840__u.data u.data
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$ hadoop fs -copyFromLocal u.data ml-100k/u.data
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls ml-100k
Found 1 items
-rw-r--r--  1 maria_dev hdfs      2079229 2022-11-24 08:28 ml-100k/u.data
[maria_dev@sandbox-hdp ~]$
```


Running Hadoop Cluster on a Larger Dataset

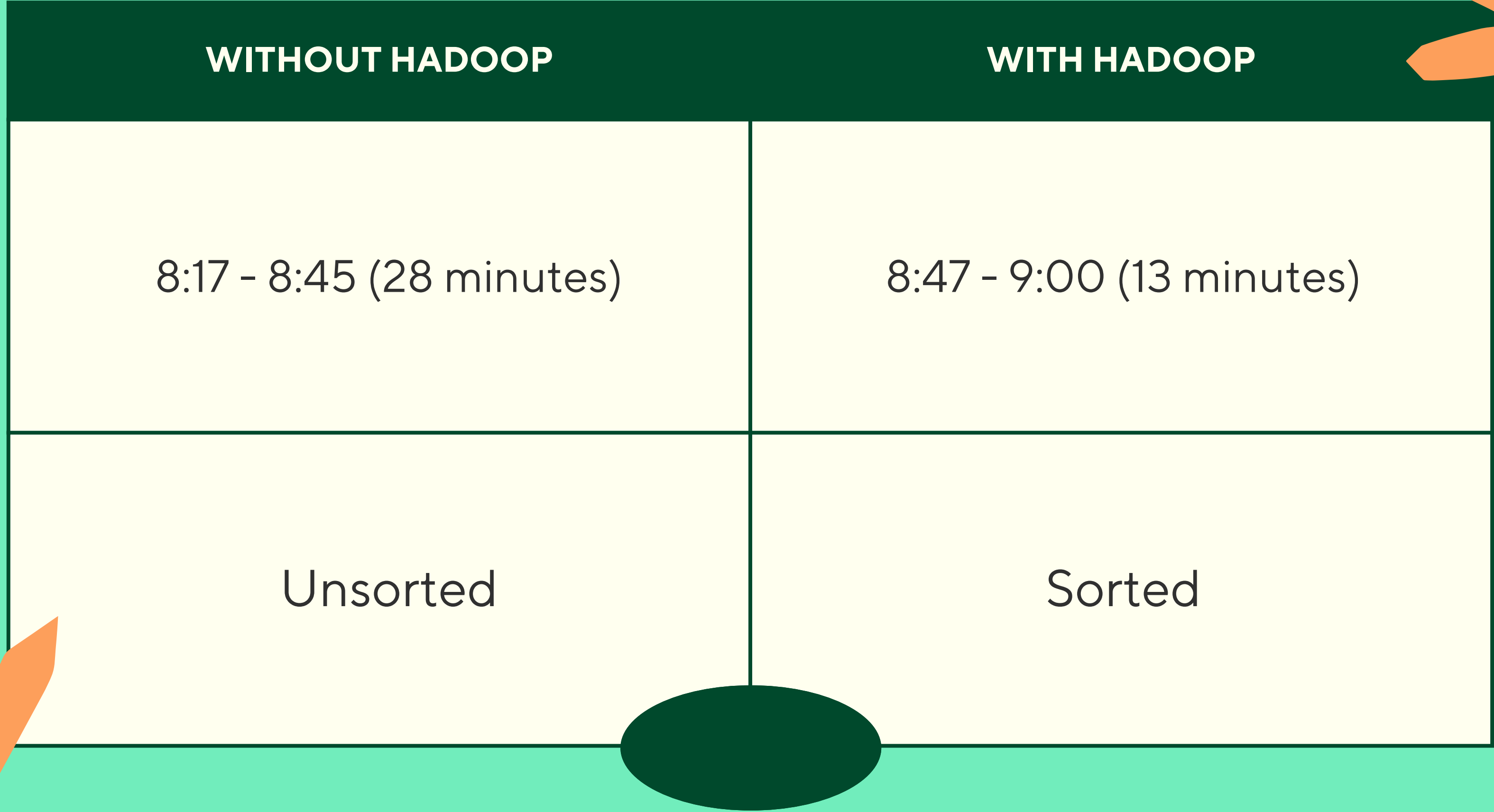
Without Hadoop:

```
sandbox-hdp maria_dev
sandbox-hdp login: maria_dev
maria_dev@sandbox-hdp.hortonworks.com's password:
Last login: Tue Nov 22 15:36:09 2022 from 172.18.0.2
[maria_dev@sandbox-hdp ~]$ nano ml-20m/1661211802__RatingsBreakdown.py
[maria_dev@sandbox-hdp ~]$ nano 1661211802__RatingsBreakdown.py
[maria_dev@sandbox-hdp ~]$ chmod 777 ml-20m/ratings.csv
[maria_dev@sandbox-hdp ~]$ chmod 777 1661211802__RatingsBreakdown.py
chmod: changing permissions of '1661211802__RatingsBreakdown.py': Operation not permitted
[maria_dev@sandbox-hdp ~]$ sudo su
[root@sandbox-hdp maria_dev]# ls
1661211748__u.data    1661211802__RatingsBreakdown.py    1661211802__RatingsBreakdown.py.save  ml-1m.zip    ml-20m    u.data
1661211748__u.data.1 1661211802__RatingsBreakdown.py.1  ml-1m                                ml-1m.zip.1  ml-20m.zip wget-log
[root@sandbox-hdp maria_dev]# chmod 777 1661211802__RatingsBreakdown.py
[root@sandbox-hdp maria_dev]# nano 1661211802__RatingsBreakdown.py
[root@sandbox-hdp maria_dev]# python 1661211802__RatingsBreakdown.py ml-20m/ratings.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/1661211802__RatingsBreakdown.root.20221122.155332.317684
Running step 1 of 1...
job output is in /tmp/1661211802__RatingsBreakdown.root.20221122.155332.317684/output
Streaming final output from /tmp/1661211802__RatingsBreakdown.root.20221122.155332.317684/output...
"3.5"    2200156
"4.0"    5561926
"4.5"    1534824
"5.0"    2898660
"0.5"    239125
"1.0"    680732
"1.5"    279252
"2.0"    1430997
"2.5"    883398
"3.0"    4291193
"rating" 1
Removing temp directory /tmp/1661211802__RatingsBreakdown.root.20221122.155332.317684...
[root@sandbox-hdp maria_dev]#
```


Running Hadoop Cluster on a Larger Dataset

With Hadoop:

```
WRONG_REDUCE=0
job output is in hdfs:///user/root/tmp/mrjob/1661211802__RatingsBreakdown.root.20221122.162323.833046/output
Streaming final output from hdfs:///user/root/tmp/mrjob/1661211802__RatingsBreakdown.root.20221122.162323.833046/output...
"0.5"    239125
"1.0"    680732
"1.5"    279252
"2.0"    1430997
"2.5"    883398
"3.0"    4291193
"3.5"    2200156
"4.0"    5561926
"4.5"    1534824
"5.0"    2898660
"rating"      1
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/1661211802__RatingsBreakdown.root.20221122.162323.833046...
Removing temp directory /tmp/1661211802__RatingsBreakdown.root.20221122.162323.833046...
[root@sandbox-hdp maria_dev]# {.{bash{_{history,logout,profile},rc},cache},1661211{748__u.data{,.1},802__RatingsBreakdown.py{,.{1,save}}},ml-{1m{,.zip{,.1}},20m{,.zip}},u.data,wget-log}
```



WITHOUT HADOOP	WITH HADOOP
8:17 - 8:45 (28 minutes)	8:47 - 9:00 (13 minutes)
Unsorted	Sorted

MapReduce Programming Challenge

```
GNU nano 2.3.1 File: ...211802_RatingsBreakdown.py

from mrjob.job import MRJob
from mrjob.step import MRStep

class RatingsBreakdown(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_ratings,
                  reducer=self.reducer_count_ratings)
        ]

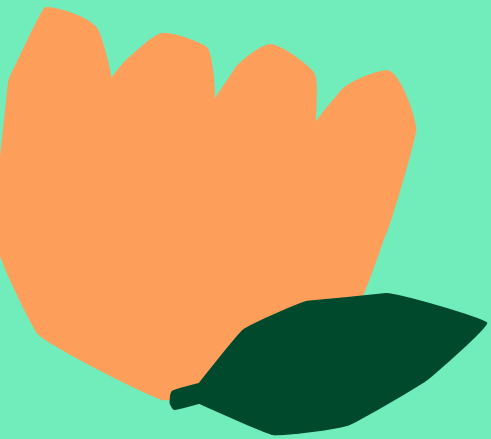
    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield movieID, 1

    def reducer_count_ratings(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    RatingsBreakdown.run()
```

```
^G Get Help ^O Write Out ^R Read File ^Y Prev Page ^K Cut Text ^C Cur Pos
^X Exit    ^J Justify  ^W Where I  ^V Next Pa ^U UnCut T  ^T To Spell
```

"971"	34
"972"	28
"973"	4
"974"	32
"975"	44
"976"	12
"977"	49
"978"	27
"979"	35
"98"	390
"980"	22
"981"	8
"982"	20
"983"	15
"984"	44
"985"	22
"986"	23
"987"	4
"988"	86
"989"	32
"99"	172
"990"	33
"991"	25
"992"	4
"993"	66
"994"	7
"995"	31
"996"	14
"997"	16
"998"	16



**Thank you for
listening!**