Question 1

a) Write a map-reduce program to determine the most popular video game in North America.

Mapper code (4pt)

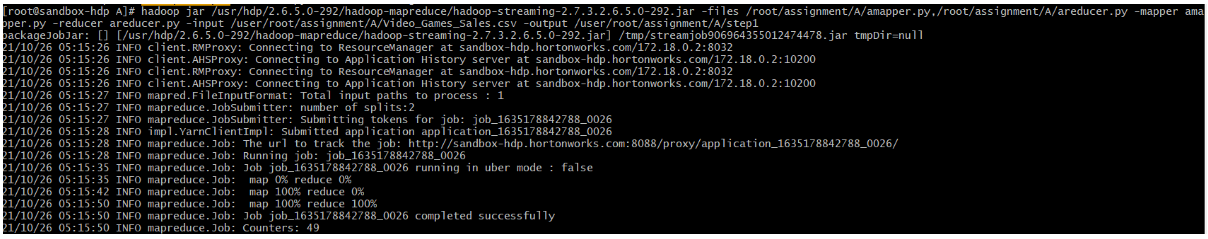
```
#!/usr/bin/env python
import sys
for line in sys.stdin:
  line = line.strip()
  attributes= line.split(',')
  index = 0
  for a in attributes:
    try:
      if len(a) == 4 and a.isdigit():
        name = ",".join(attributes[:index - 1])
        # index is 'year', index - 1 is platform, index [:platform index] is video
game name
        sale = attributes[index + 3] # index(year) + 3 = NA_Sales
        print '%s,%s' % (name, sale)
        break
      else:
        index += 1
    except:
      index += 1
```

Reducer code (4pt)

```
#!/usr/bin/env python
import sys
import operator
topvideo = {}
for line in sys.stdin:
    line = line.strip()
    inputs = line.split(',')
    vid = ",".join(inputs[:-1])
    sale = inputs[-1]
    try:
        sale = float(sale)
        topvideo[vid] = sale
    except:
        continue
sort = sorted(topvideo, key=topvideo.get, reverse=True)
n = 0
for v in sort:
    if n < 10:
        print '%s,%s' % (v, topvideo[v])
        n = n + 1
    else:
     break
```

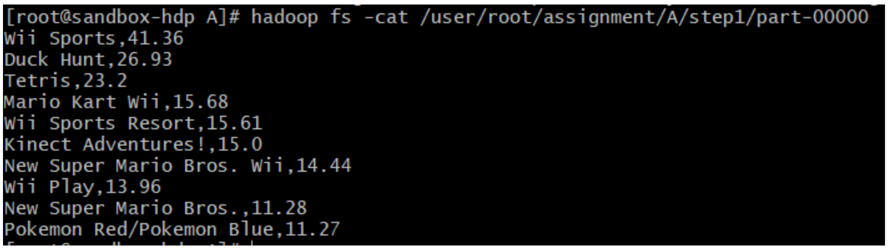
Running MapReduce job (3pt)

hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/A/amapper.py,/root/assignment/A/areducer.py -mapper amapper.py -reducer areducer.py -input /user/root/assignment/A/Video_Games_Sales.csv -output /user/root/assignment/A/step1



```
[root@sandbox-hdp A]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/A/amapper.py,/root/assignment/A/areducer.py -mapper ama
pper.py -reducer areducer.py -input /user/root/assignment/A/Video_Games_Sales.csv -output /user/root/assignment/A/step1
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob906964355012474478.jar tmpDir=null
21/10/26 05:15:26 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/26 05:15:26 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/26 05:15:26 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/26 05:15:26 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/26 05:15:27 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/26 05:15:27 INFO mapreduce.JobSubmitter: number of splits:2
21/10/26 05:15:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0026
21/10/26 05:15:28 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0026
21/10/26 05:15:28 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0026/
21/10/26 05:15:28 INFO mapreduce.Job: Running job: job_1635178842788_0026
21/10/26 05:15:35 INFO mapreduce.Job: Job job_1635178842788_0026 running in uber mode : false
21/10/26 05:15:35 INFO mapreduce.Job:  map 0% reduce 0%
21/10/26 05:15:42 INFO mapreduce.Job:  map 100% reduce 0%
21/10/26 05:15:50 INFO mapreduce.Job:  map 100% reduce 100%
21/10/26 05:15:50 INFO mapreduce.Job: Job job_1635178842788_0026 completed successfully
21/10/26 05:15:50 INFO mapreduce.Job: Counters: 49
```

Producing correct output(3pt)



```
[root@sandbox-hdp A]# hadoop fs -cat /user/root/assignment/A/step1/part-00000
Wii Sports,41.36
Duck Hunt,26.93
Tetris,23.2
Mario Kart Wii,15.68
Wii Sports Resort,15.61
Kinect Adventures!,15.0
New Super Mario Bros. Wii,14.44
Wii Play,13.96
New Super Mario Bros.,11.28
Pokemon Red/Pokemon Blue,11.27
```

Topmost popular video game in North America printed.

Wii Sports is the most popular video in North America

Instructions (4pt) & Screenshots (2pt)

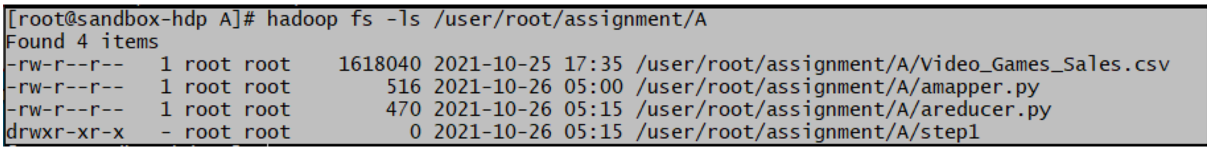
1. Make sure docker is running. Access to Ambari via 127.0.0.1:1080 portal and run services including HDFS, YARN, MAPREDUCE, ZOOKEEPER.
2. Upload files to local system via FileZilla

   host: sftp://localhost username: root password: which you setup for hortonworks port:2222
3. On Git Bash, access to hortonworks sandbox by entering the following command.

   ssh root@sandbox-hdp.hortonworks.com -p 2222
4. Make directory in the HDFS using the following command.

   Hadoop fs -mkdir /user/root/assignment/A
5. Upload files from local system to HDFS using Hadoop fs -put {filename} {HDFS path}

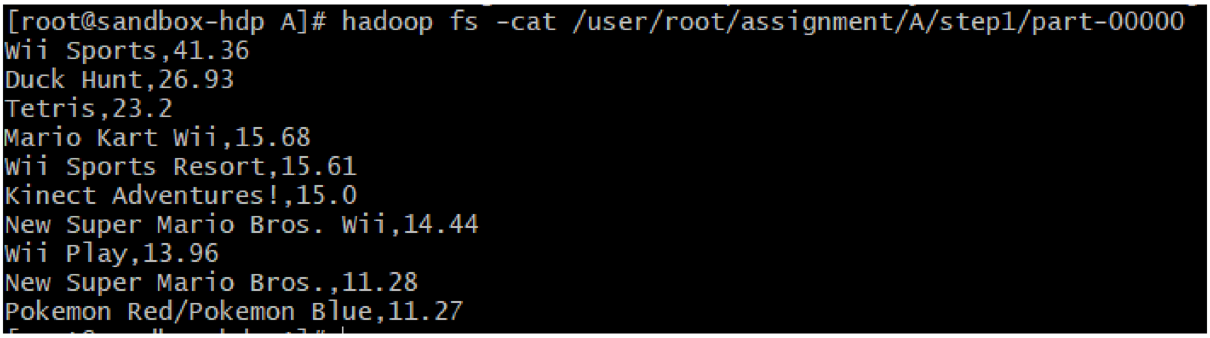
```
[root@sandbox-hdp A]# hadoop fs -put amapper.py /user/root/assignment/A
[root@sandbox-hdp A]# hadoop fs -put areducer.py /user/root/assignment/A
[root@sandbox-hdp A]# hadoop fs -put Video_Games_Sale.csv /user/root/assignment/A
```



```
[root@sandbox-hdp A]# hadoop fs -ls /user/root/assignment/A
Found 4 items
-rw-r--r--   1 root root    1618040 2021-10-25 17:35 /user/root/assignment/A/Video_Games_Sales.csv
-rw-r--r--   1 root root        516 2021-10-26 05:00 /user/root/assignment/A/amapper.py
-rw-r--r--   1 root root        470 2021-10-26 05:15 /user/root/assignment/A/areducer.py
drwxr-xr-x   - root root          0 2021-10-26 05:15 /user/root/assignment/A/step1
```

6. Map-reduce
   - Refer to **Running MapReduce job** section above
7. Checking the output using command

   "hadoop fs -cat {path}/filename"



```
[root@sandbox-hdp A]# hadoop fs -cat /user/root/assignment/A/step1/part-00000
Wii Sports,41.36
Duck Hunt,26.93
Tetris,23.2
Mario Kart Wii,15.68
Wii Sports Resort,15.61
Kinect Adventures!,15.0
New Super Mario Bros. Wii,14.44
Wii Play,13.96
New Super Mario Bros.,11.28
Pokemon Red/Pokemon Blue,11.27
[root@sandbox-hdp A]#
```

b) Write a map-reduce program to determine the most popular video game per genres.

Mapper code (4pt)

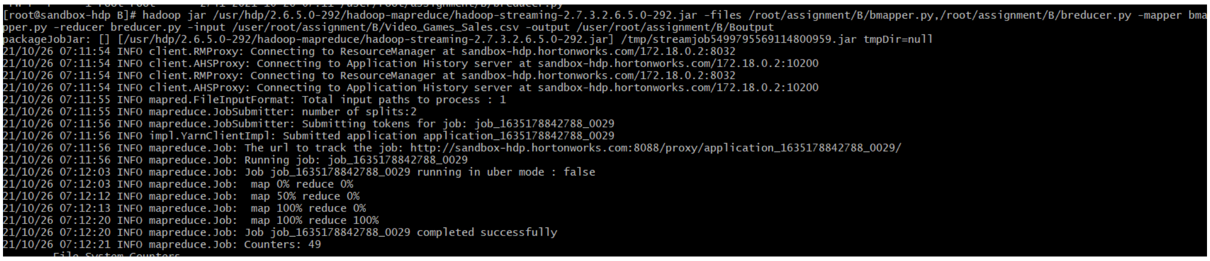
```
#!/usr/bin/env python
import sys
for line in sys.stdin:
  line = line.strip()
  attributes= line.split(',')
  index = 0
  for a in attributes:
    try:
      if len(a) == 4 and a.isdigit():
        name = ",".join(attributes[:index - 1])
        # index is 'year', index - 1 is platform, index [:platform index] is video game name
        genre = attributes[index + 1] # index is 'year', index + 1 is Genre
        gsale = attributes[index + 7] # index(year) + 7 = global_Sales
        print '%s₩t%s₩t%s' % (name, genre, gsale)
        break
      else:
        index += 1
    except:
      index += 1
```

Reducer code (4pt)

```
#!/usr/bin/env python
import sys
import operator
Sports = {} #create dictionary for each genre
Platform = {}
Racing = {}
Puzzle = {}
Misc = {}
Shooter = {}
Simulation = {}
Action = {}
Fighting = {}
Adventure = {}
Strategy = {}

for line in sys.stdin:
    inputs = line.split('₩t')
    name = inputs[0]
    genre = inputs[1]
    sale = inputs[2]
    try:
        if genre == "Sports":
            Sports[name] = sale
        elif genre == "Platform":
            Platform[name] = sale
        elif genre == "Racing":
            Racing[name] = sale
        elif genre == "Puzzle":
            Puzzle[name] = sale
        elif genre == "Misc":
            Misc[name] = sale
        elif genre == "Shooter":
            Shooter[name] = sale
        elif genre == "Simulation":
            Simulation[name] = sale
        elif genre == "Action":
            Action[name] = sale
        elif genre == "Fighting":
            Fighting[name] = sale
        elif genre == "Adventure":
            Adventure[name] = sale
        elif genre == "Strategy":
            Strategy[name] = sale
    except:
        continue
print '%s₩t%s₩t%s' % ("Sports", max(Sports, key =Sports.get), Sports[max(Sports, key =Sports.get)])
print '%s₩t%s₩t%s' % ("Platform", max(Platform, key =Platform.get), Platform[max(Platform, key =Platform.get)])
print '%s₩t%s₩t%s' % ("Racing", max(Racing, key =Racing.get), Racing[max(Racing, key =Racing.get)])
print '%s₩t%s₩t%s' % ("Puzzle", max(Puzzle, key =Puzzle.get), Puzzle[max(Puzzle, key =Puzzle.get)])
print '%s₩t%s₩t%s' % ("Misc", max(Misc, key =Misc.get), Misc[max(Misc, key =Misc.get)])
print '%s₩t%s₩t%s' % ("Shooter", max(Shooter, key =Shooter.get), Shooter[max(Shooter, key =Shooter.get)])
print '%s₩t%s₩t%s' % ("Simulation", max(Simulation, key =Simulation.get), Simulation[max(Simulation, key =Simulation.get)])
print '%s₩t%s₩t%s' % ("Action", max(Action, key =Action.get), Action[max(Action, key =Action.get)])
print '%s₩t%s₩t%s' % ("Fighting", max(Fighting, key =Fighting.get), Fighting[max(Fighting, key =Fighting.get)])
print '%s₩t%s₩t%s' % ("Adventure", max(Adventure, key =Adventure.get), Adventure[max(Adventure, key =Adventure.get)])
print '%s₩t%s₩t%s' % ("Strategy", max(Strategy, key =Strategy.get), Strategy[max(Strategy, key =Strategy.get)])
```

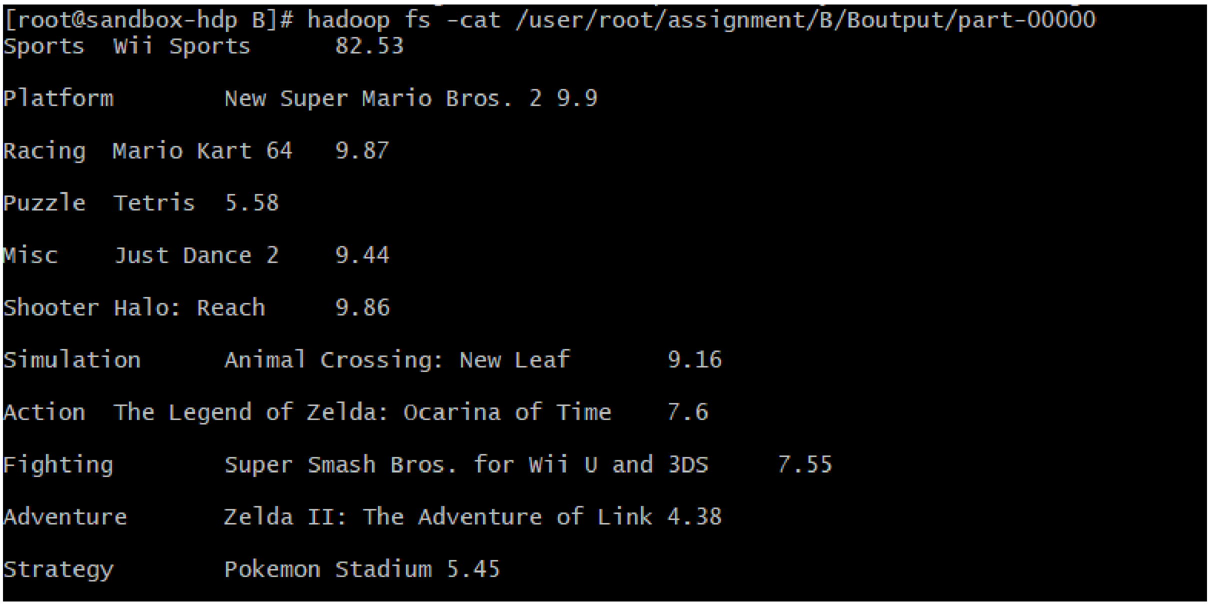
Running MapReduce job (3pt)

hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/B/bmapper.py,/root/assignment/B/breducer.py -mapper bmapper.py -reducer breducer.py -input /user/root/assignment/B/Video_Games_Sales.csv -output /user/root/assignment/B/Boutput



```
[root@sandbox-hdp B]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/B/bmapper.py,/root/assignment/B/breducer.py -mapper bma
pper.py -reducer breducer.py -input /user/root/assignment/B/Video_Games_Sales.csv -output /user/root/assignment/B/Boutput
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob5499795569114800959.jar tmpDir=null
21/10/26 07:11:54 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/26 07:11:54 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/26 07:11:54 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/26 07:11:54 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/26 07:11:55 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/26 07:11:55 INFO mapreduce.JobSubmitter: number of splits:2
21/10/26 07:11:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0029
21/10/26 07:11:56 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0029
21/10/26 07:11:56 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0029/
21/10/26 07:11:56 INFO mapreduce.Job: Running job: job_1635178842788_0029
21/10/26 07:12:03 INFO mapreduce.Job: Job job_1635178842788_0029 running in uber mode : false
21/10/26 07:12:03 INFO mapreduce.Job:  map 0% reduce 0%
21/10/26 07:12:12 INFO mapreduce.Job:  map 50% reduce 0%
21/10/26 07:12:13 INFO mapreduce.Job:  map 100% reduce 0%
21/10/26 07:12:20 INFO mapreduce.Job:  map 100% reduce 100%
21/10/26 07:12:20 INFO mapreduce.Job: Job job_1635178842788_0029 completed successfully
21/10/26 07:12:21 INFO mapreduce.Job: Counters: 49
        File System Counters
```

Producing correct output(3pt)



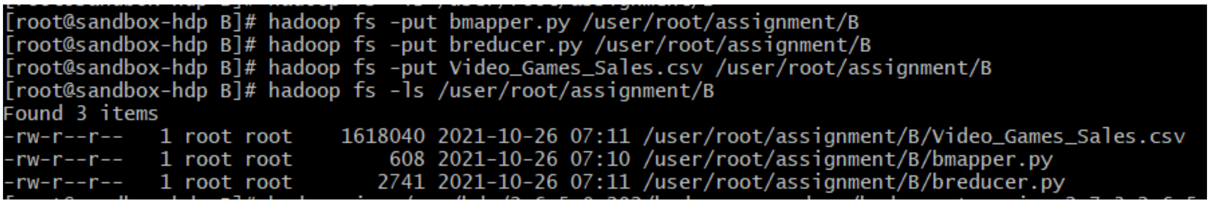
```
[root@sandbox-hdp B]# hadoop fs -cat /user/root/assignment/B/Boutput/part-00000
Sports  Wii Sports      82.53

Platform        New Super Mario Bros. 2 9.9

Racing  Mario Kart 64   9.87

Puzzle  Tetris  5.58

Misc    Just Dance 2    9.44

Shooter Halo: Reach     9.86

Simulation      Animal Crossing: New Leaf       9.16

Action  The Legend of Zelda: Ocarina of Time    7.6

Fighting        Super Smash Bros. for Wii U and 3DS     7.55

Adventure       Zelda II: The Adventure of Link 4.38

Strategy        Pokemon Stadium 5.45
```

Most popular video game per genres

Output format: Genre, Video_game_name, global_sale

Instructions (4pt) & Screenshots (2pt)

1. Make sure docker is running. Access to Ambari via 127.0.0.1:1080 portal and run services including HDFS, YARN, MAPREDUCE, ZOOKEEPER.
2. Upload files to local system via FileZilla
   host: sftp://localhost username: root password: which you setup for hortonworks port:2222
3. On Git Bash, access to hortonworks sandbox by entering the following command.
   ssh root@sandbox-hdp.hortonworks.com -p 2222
4. Make directory in the HDFS using the following command.
   Hadoop fs -mkdir /user/root/assignment/A
5. Upload files to HDFS using Hadoop fs -put {filename} {HDFS path}



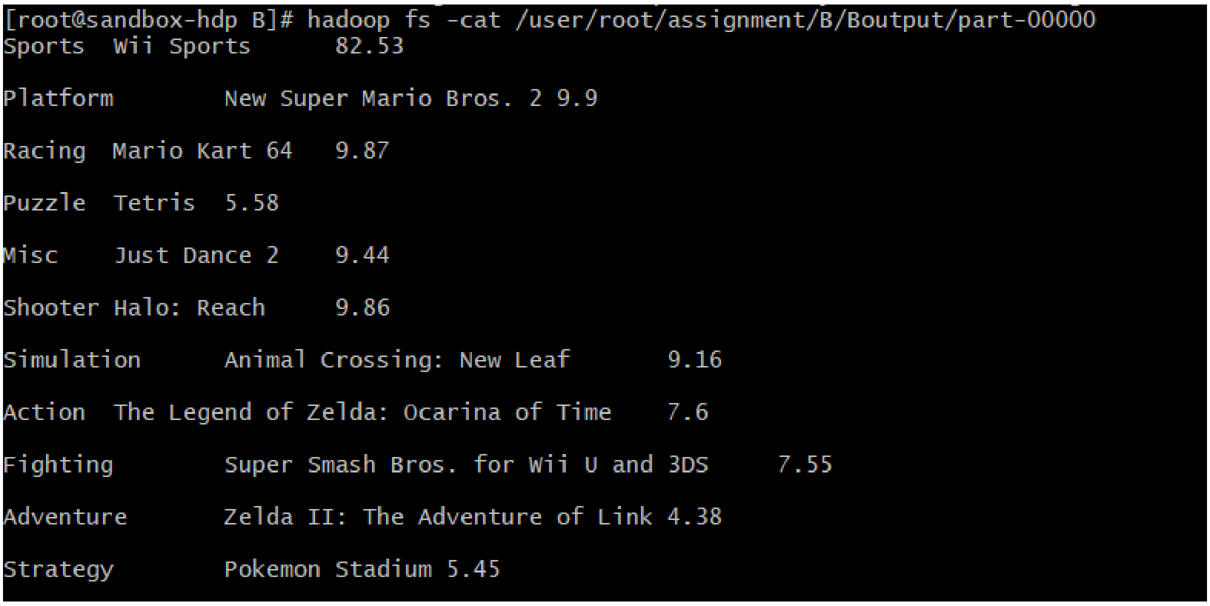
```
[root@sandbox-hdp B]# hadoop fs -put bmapper.py /user/root/assignment/B
[root@sandbox-hdp B]# hadoop fs -put breducer.py /user/root/assignment/B
[root@sandbox-hdp B]# hadoop fs -put Video_Games_Sales.csv /user/root/assignment/B
[root@sandbox-hdp B]# hadoop fs -ls /user/root/assignment/B
Found 3 items
-rw-r--r--   1 root root    1618040 2021-10-26 07:11 /user/root/assignment/B/Video_Games_Sales.csv
-rw-r--r--   1 root root        608 2021-10-26 07:10 /user/root/assignment/B/bmapper.py
-rw-r--r--   1 root root       2741 2021-10-26 07:11 /user/root/assignment/B/breducer.py
```

6. Map-reduce

   - Refer to **Running MapReduce job** section above

7. Checking the output using command

   "hadoop fs -cat {path}/filename



```
[root@sandbox-hdp B]# hadoop fs -cat /user/root/assignment/B/Boutput/part-00000
Sports  Wii Sports      82.53

Platform        New Super Mario Bros. 2 9.9

Racing  Mario Kart 64   9.87

Puzzle  Tetris  5.58

Misc    Just Dance 2    9.44

Shooter Halo: Reach     9.86

Simulation      Animal Crossing: New Leaf       9.16

Action  The Legend of Zelda: Ocarina of Time    7.6

Fighting        Super Smash Bros. for Wii U and 3DS     7.55

Adventure       Zelda II: The Adventure of Link 4.38

Strategy        Pokemon Stadium 5.45
```

c) Write a map-reduce program to determine the year in which North America had highest video games sales.

Mapper code (4pt)

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.strip()
    attributes= line.split(',')
    index = 0
    for a in attributes:
        try:
            if len(a) == 4 and a.isdigit():
                year = a
                sale = attributes[index + 3]
                print '%s,%s' % (year, sale)
                break
            else:
                index += 1
        except:
            index += 1
```

Reducer code (4pt)

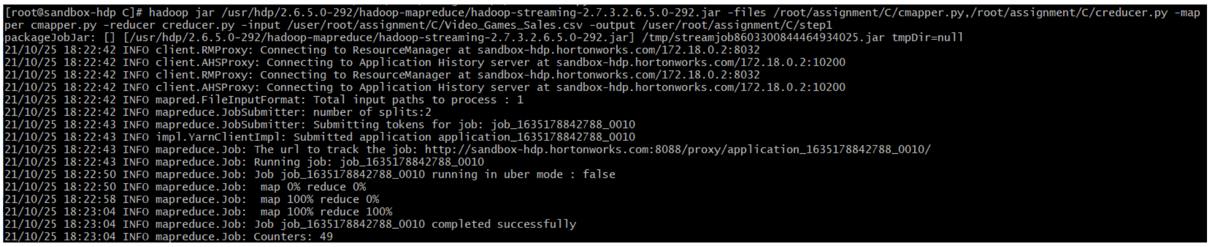
```
#!/usr/bin/env python
import sys
current_year = None
current_sale = 0
year = None
# input comes from STDIN
for line in sys.stdin:
    line = line.strip()
    year, sale = line.split(',')
    try:
        sale = float(sale)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue
    if current_year == year:
        current_sale += sale
    else:
        if current_year:
            print '%s,%s' % (current_year, current_sale)
        current_sale = sale
        current_year = year
if current_year == year:
    print '%s,%s' % (current_year, current_sale)
```

Mapper and reducer codes for sorting : print top 10 popular video game in North America

```
#!/usr/bin/env python
import sys
import operator
topsaleyear = {}
for line in sys.stdin:
    line = line.strip()
    vid, sale = line.split(",")
    try:
        sale = float(sale)
        topsaleyear[vid] = sale
    except:
        continue
sort = sorted(topsaleyear, key=topsaleyear.get, reverse=True)
n = 0
for v in sort:
    if n < 10:
        print '%s,%s' % (v, topsaleyear[v])
        n = n + 1
    else:
     break
```

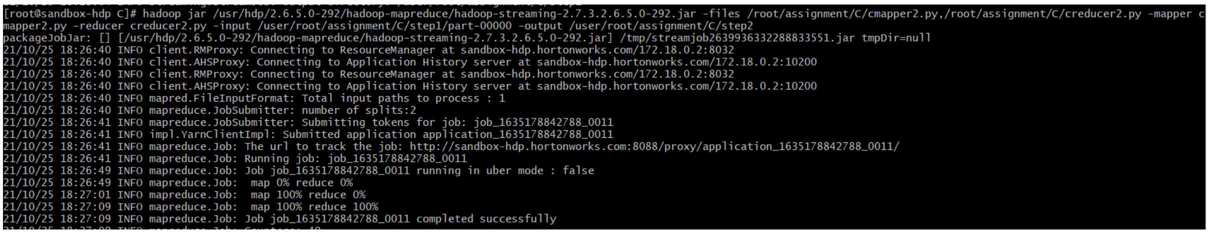
Running MapReduce job (3pt)

hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/C/cmapper.py,/root/assignment/C/creducer.py -mapper cmapper.py -reducer creducer.py -input /user/root/assignment/C/Video_Games_Sales.csv -output /user/root/assignment/C/step1



```
[root@sandbox-hdp C]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/C/cmapper.py,/root/assignment/C/creducer.py -map
per cmapper.py -reducer creducer.py -input /user/root/assignment/C/Video_Games_Sales.csv -output /user/root/assignment/C/step1
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob8603300844464934025.jar tmpDir=null
21/10/25 18:22:42 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 18:22:42 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 18:22:42 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 18:22:42 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 18:22:42 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/25 18:22:42 INFO mapreduce.JobSubmitter: number of splits:2
21/10/25 18:22:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0010
21/10/25 18:22:43 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0010
21/10/25 18:22:43 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0010/
21/10/25 18:22:43 INFO mapreduce.Job: Running job: job_1635178842788_0010
21/10/25 18:22:50 INFO mapreduce.Job: Job job_1635178842788_0010 running in uber mode : false
21/10/25 18:22:50 INFO mapreduce.Job:  map 0% reduce 0%
21/10/25 18:22:58 INFO mapreduce.Job:  map 100% reduce 0%
21/10/25 18:23:04 INFO mapreduce.Job:  map 100% reduce 100%
21/10/25 18:23:04 INFO mapreduce.Job: Job job_1635178842788_0010 completed successfully
21/10/25 18:23:04 INFO mapreduce.Job: Counters: 49
```

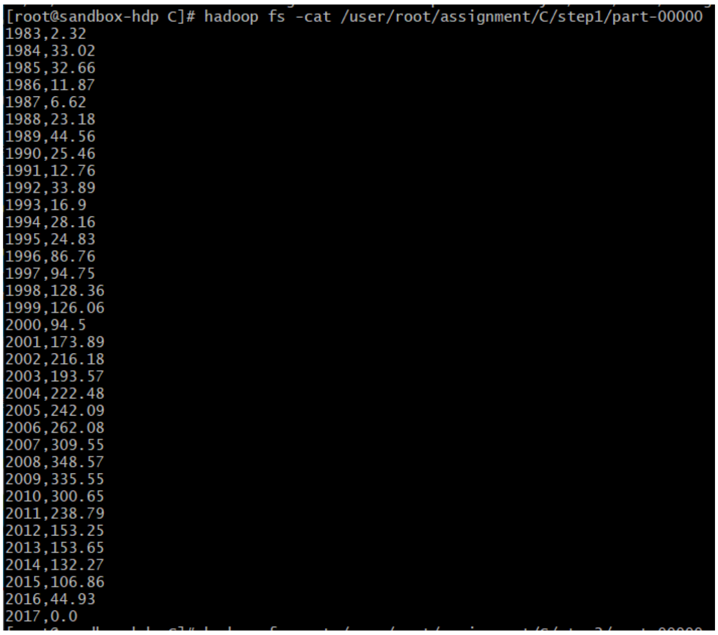
hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/C/cmapper2.py,/root/assignment/C/creducer2.py -mapper cmapper2.py -reducer creducer2.py -input /user/root/assignment/C/step1/part-00000 -output /user/root/assignment/C/step2



```
[root@sandbox-hdp C]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/C/cmapper2.py,/root/assignment/C/creducer2.py -mapper c
mapper2.py -reducer creducer2.py -input /user/root/assignment/C/step1/part-00000 -output /user/root/assignment/C/step2
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob2639936332288833551.jar tmpDir=null
21/10/25 18:26:40 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 18:26:40 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 18:26:40 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 18:26:40 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 18:26:40 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/25 18:26:40 INFO mapreduce.JobSubmitter: number of splits:2
21/10/25 18:26:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0011
21/10/25 18:26:41 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0011
21/10/25 18:26:41 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0011/
21/10/25 18:26:41 INFO mapreduce.Job: Running job: job_1635178842788_0011
21/10/25 18:26:49 INFO mapreduce.Job: Job job_1635178842788_0011 running in uber mode : false
21/10/25 18:26:49 INFO mapreduce.Job:  map 0% reduce 0%
21/10/25 18:27:01 INFO mapreduce.Job:  map 100% reduce 0%
21/10/25 18:27:09 INFO mapreduce.Job:  map 100% reduce 100%
21/10/25 18:27:09 INFO mapreduce.Job: Job job_1635178842788_0011 completed successfully
```

Producing correct output(3pt)

1. **First map-reduce job: map-reduce without sorting**

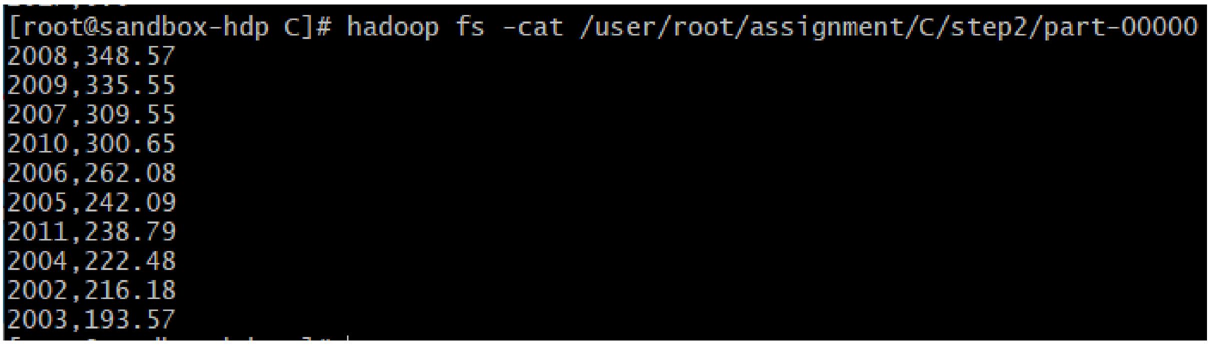


```
[root@sandbox-hdp C]# hadoop fs -cat /user/root/assignment/C/step1/part-00000
1983,2.32
1984,33.02
1985,32.66
1986,11.87
1987,6.62
1988,23.18
1989,44.56
1990,25.46
1991,12.76
1992,33.89
1993,16.9
1994,28.16
1995,24.83
1996,86.76
1997,94.75
1998,128.36
1999,126.06
2000,94.5
2001,173.89
2002,216.18
2003,193.57
2004,222.48
2005,242.09
2006,262.08
2007,309.55
2008,348.57
2009,335.55
2010,300.65
2011,238.79
2012,153.25
2013,153.65
2014,132.27
2015,106.86
2016,44.93
2017,0.0
```

2. **Second map-reduce job: map-reduce after sorting,**

   printing top year in which north America had highest video games sales.

   2008 had the highest video games sales in North America



```
[root@sandbox-hdp C]# hadoop fs -cat /user/root/assignment/C/step2/part-00000
2008,348.57
2009,335.55
2007,309.55
2010,300.65
2006,262.08
2005,242.09
2011,238.79
2004,222.48
2002,216.18
2003,193.57
```

Instructions (4pt) & Screenshots (2pt)

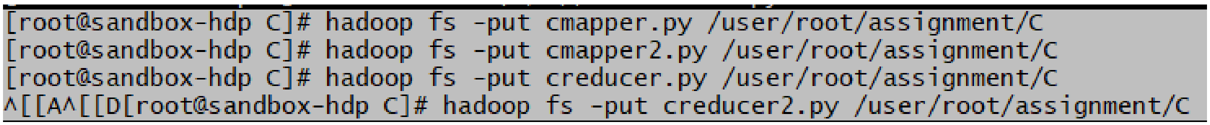
1. Make sure docker is running. Access to Ambari via 127.0.0.1:1080 portal and run services including HDFS, YARN, MAPREDUCE, ZOOKEEPER.
2. Upload files to local system via FileZilla

   host: sftp://localhost username: root password: which you setup for hortonworks port:2222
3. On Git Bash, access to hortonworks sandbox by entering the following command.

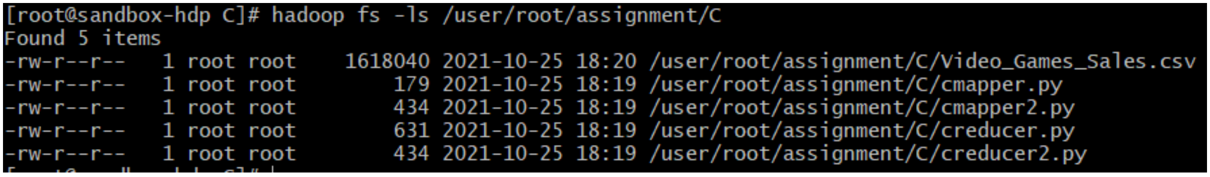
   ssh root@sandbox-hdp.hortonworks.com -p 2222
4. Make directory in the HDFS using the following command.

   Hadoop fs -mkdir /user/root/assignment/A

5. Upload files to HDFS using Hadoop fs -put {filename} {HDFS path}



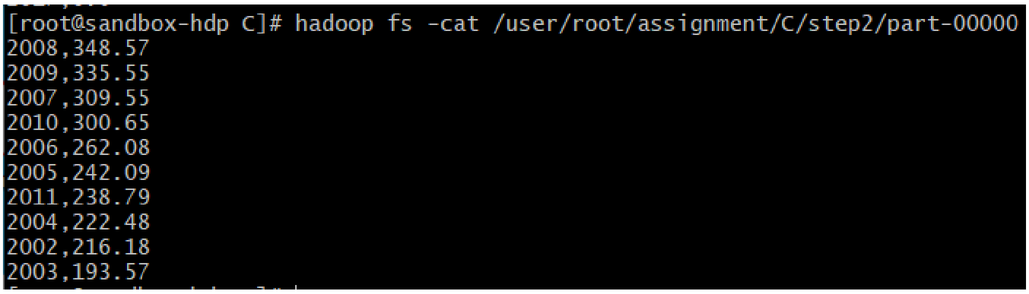
```
[root@sandbox-hdp C]# hadoop fs -put cmapper.py /user/root/assignment/C
[root@sandbox-hdp C]# hadoop fs -put cmapper2.py /user/root/assignment/C
[root@sandbox-hdp C]# hadoop fs -put creducer.py /user/root/assignment/C
^[[A^[[D[root@sandbox-hdp C]# hadoop fs -put creducer2.py /user/root/assignment/C
```



```
[root@sandbox-hdp C]# hadoop fs -ls /user/root/assignment/C
Found 5 items
-rw-r--r--   1 root root    1618040 2021-10-25 18:20 /user/root/assignment/C/Video_Games_Sales.csv
-rw-r--r--   1 root root        179 2021-10-25 18:19 /user/root/assignment/C/cmapper.py
-rw-r--r--   1 root root        434 2021-10-25 18:19 /user/root/assignment/C/cmapper2.py
-rw-r--r--   1 root root        631 2021-10-25 18:19 /user/root/assignment/C/creducer.py
-rw-r--r--   1 root root        434 2021-10-25 18:19 /user/root/assignment/C/creducer2.py
```

6. Map-reduce in two steps: 1) map-reduce for counting 2) map-reduce for sorting
   - Refer to **Running MapReduce job** section above
7. Checking the output using command

   "hadoop fs -cat {path}/filename"



```
[root@sandbox-hdp C]# hadoop fs -cat /user/root/assignment/C/step2/part-00000
2008,348.57
2009,335.55
2007,309.55
2010,300.65
2006,262.08
2005,242.09
2011,238.79
2004,222.48
2002,216.18
2003,193.57
```

d) Write a map-reduce program to determine the genre had highest video games sales globally.

Mapper code (4pt)

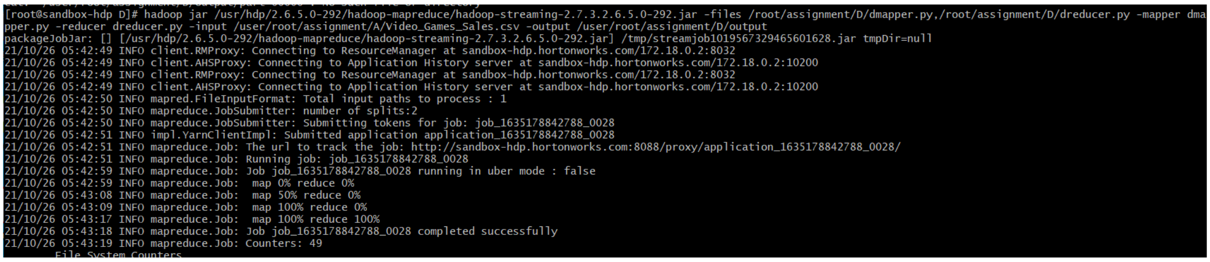
```
#!/usr/bin/env python
import sys
for line in sys.stdin:
  line = line.strip()
  attributes= line.split(',')
  index = 0
  for a in attributes:
    try:
      if len(a) == 4 and a.isdigit():
        genre = attributes[index + 1] # index is 'year', index + 1 is Genre
        sale = attributes[index + 7] # index(year) + 7 = global_Sales
        print '%s,%s' % (genre, sale)
        break
      else:
        index += 1
    except:
      index += 1
```

Reducer code (4pt)

```
#!/usr/bin/env python
import sys
import operator
topgenre = {}
for line in sys.stdin:
    line = line.strip()
    inputs = line.split(',')
    genre = ",".join(inputs[:-1])
    sale = inputs[-1]
    try:
        sale = float(sale)
        topgenre[genre] = sale
    except:
        continue
sort = sorted(topgenre, key=topgenre.get, reverse=True)
n = 0
for g in sort:
    if n < 10:
        print '%s,%s' % (g, topgenre[g])
        n = n + 1
    else:
     break
```

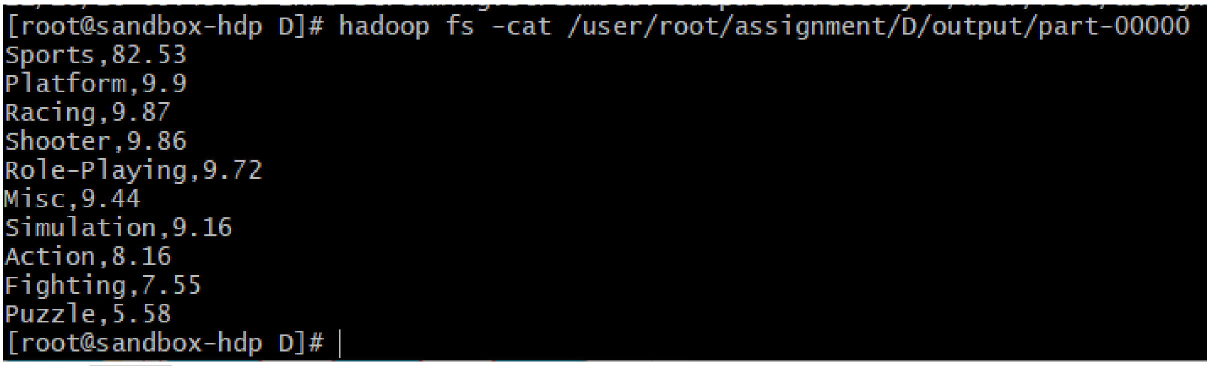
Running MapReduce job (3pt)

[root@sandbox-hdp D]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/D/dmapper.py,/root/assignment/D/dreducer.py -mapper dmapper.py -reducer dreducer.py -input /user/root/assignment/A/Video_Games_Sales.csv -output /user/root/assignment/D/output



```
[root@sandbox-hdp D]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/D/dmapper.py,/root/assignment/D/dreducer.py -mapper dma
pper.py -reducer dreducer.py -input /user/root/assignment/A/Video_Games_Sales.csv -output /user/root/assignment/D/output
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob1019567329465601628.jar tmpDir=null
21/10/26 05:42:49 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/26 05:42:49 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/26 05:42:49 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/26 05:42:49 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/26 05:42:50 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/26 05:42:50 INFO mapreduce.JobSubmitter: number of splits:2
21/10/26 05:42:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0028
21/10/26 05:42:51 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0028
21/10/26 05:42:51 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0028/
21/10/26 05:42:51 INFO mapreduce.Job: Running job: job_1635178842788_0028
21/10/26 05:42:59 INFO mapreduce.Job: Job job_1635178842788_0028 running in uber mode : false
21/10/26 05:42:59 INFO mapreduce.Job:  map 0% reduce 0%
21/10/26 05:43:08 INFO mapreduce.Job:  map 50% reduce 0%
21/10/26 05:43:09 INFO mapreduce.Job:  map 100% reduce 0%
21/10/26 05:43:17 INFO mapreduce.Job:  map 100% reduce 100%
21/10/26 05:43:18 INFO mapreduce.Job: Job job_1635178842788_0028 completed successfully
21/10/26 05:43:19 INFO mapreduce.Job: Counters: 49
        File System Counters
```

Producing correct output(3pt)



```
[root@sandbox-hdp D]# hadoop fs -cat /user/root/assignment/D/output/part-00000
Sports,82.53
Platform,9.9
Racing,9.87
Shooter,9.86
Role-Playing,9.72
Misc,9.44
Simulation,9.16
Action,8.16
Fighting,7.55
Puzzle,5.58
[root@sandbox-hdp D]# |
```

Genre: **Sports** had highest video games sales globally

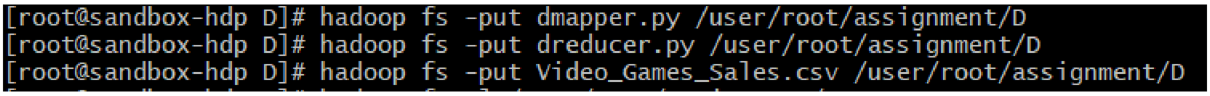
Instructions (4pt) & Screenshots (2pt)

1. Make sure docker is running. Access to Ambari via 127.0.0.1:1080 portal and run services including HDFS, YARN, MAPREDUCE, ZOOKEEPER.
2. Upload files to local system via FileZilla

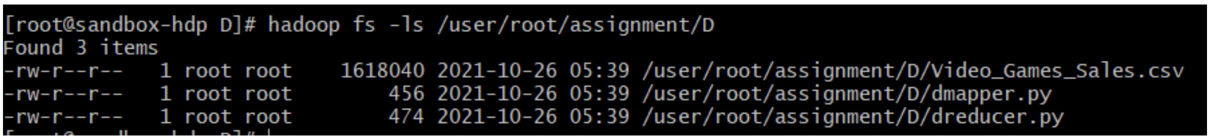
   host: sftp://localhost username: root password: which you setup for hortonworks port:2222
3. On Git Bash, access to hortonworks sandbox by entering the following command.

   ssh root@sandbox-hdp.hortonworks.com -p 2222
4. Make directory in the HDFS using the following command.

   Hadoop fs -mkdir /user/root/assignment/A
5. Upload files to HDFS using Hadoop fs -put {filename} {HDFS path}



```
[root@sandbox-hdp D]# hadoop fs -put dmapper.py /user/root/assignment/D
[root@sandbox-hdp D]# hadoop fs -put dreducer.py /user/root/assignment/D
[root@sandbox-hdp D]# hadoop fs -put Video_Games_Sales.csv /user/root/assignment/D
```



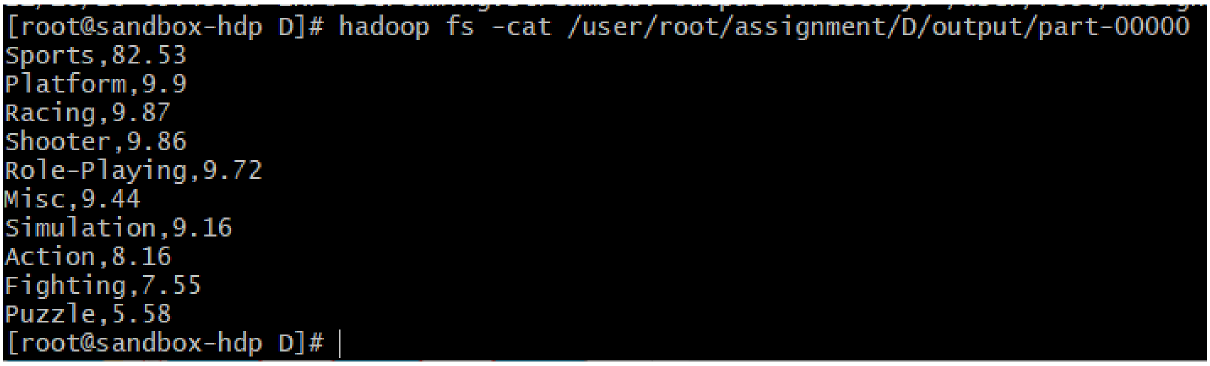
```
[root@sandbox-hdp D]# hadoop fs -ls /user/root/assignment/D
Found 3 items
-rw-r--r--   1 root root    1618040 2021-10-26 05:39 /user/root/assignment/D/Video_Games_Sales.csv
-rw-r--r--   1 root root        456 2021-10-26 05:39 /user/root/assignment/D/dmapper.py
-rw-r--r--   1 root root        474 2021-10-26 05:39 /user/root/assignment/D/dreducer.py
```

6. Map-reduce

   - Refer to **Running MapReduce job** section above

7. Checking the output using command

   “hadoop fs -cat {path}/filename



```
[root@sandbox-hdp D]# hadoop fs -cat /user/root/assignment/D/output/part-00000
Sports,82.53
Platform,9.9
Racing,9.87
Shooter,9.86
Role-Playing,9.72
Misc,9.44
Simulation,9.16
Action,8.16
Fighting,7.55
Puzzle,5.58
[root@sandbox-hdp D]#
```

Question 2

Mapper code (4pt) & Reducer code (4pt)

First step for counting) Mapper code

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print '%s₩t%s' % (word, 1)
```

First step for counting) Reducer code

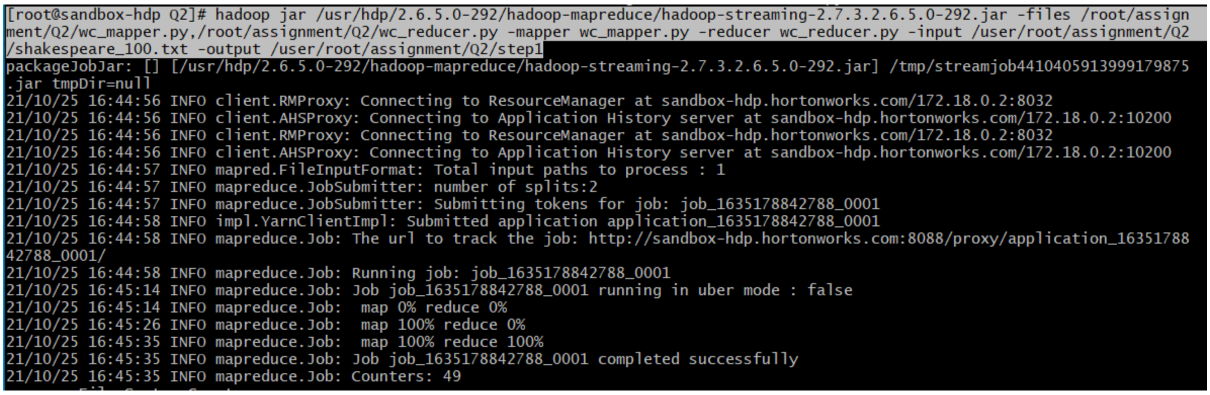
```
#!/usr/bin/env python
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
# input comes from STDIN
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('₩t', 1)
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue
    if current_word == word:
        current_count += count
    else:
        if current_word:
            print '%s₩t%s' % (current_word, current_count)
        current_count = count
        current_word = word
if current_word == word:
    print '%s₩t%s' % (current_word, current_count)
```

Second step for sorting) Mapper and reducer code are the same

```
#!/usr/bin/env python
import sys
import operator
topnwords = {}   #dictionary to sort the words
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    if len(words) == 2:
        try:
            count = int(words[1])    # word count
            topnwords[words[0]] = count # add word and count to dictionary
        except:
            continue
# list of words sorted by count
sort = sorted(topnwords, key=topnwords.get, reverse=True)

n = 0    #counter to only print up to top 10 words
for w in sort:     #iterate through all words sorted by count
    if n < 10:        # print up to top 10 words
        print '%s₩t%s' % (w, topnwords[w])
        n = n + 1
    else:
        continue
```

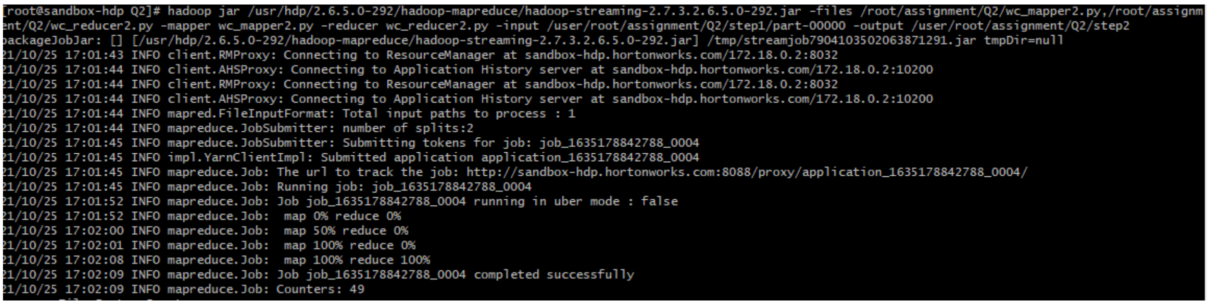
Running MapReduce job (3pt)

1) First step: Map-reduce job

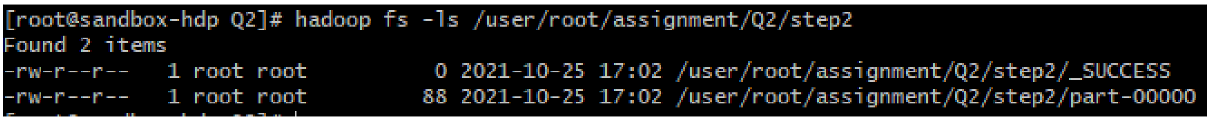


```
[root@sandbox-hdp Q2]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assign
ment/Q2/wc_mapper.py,/root/assignment/Q2/wc_reducer.py -mapper wc_mapper.py -reducer wc_reducer.py -input /user/root/assignment/Q2
/shakespeare_100.txt -output /user/root/assignment/Q2/step1
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob4410405913999179875
.jar tmpDir=null
21/10/25 16:44:56 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 16:44:56 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 16:44:56 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 16:44:56 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 16:44:57 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/25 16:44:57 INFO mapreduce.JobSubmitter: number of splits:2
21/10/25 16:44:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0001
21/10/25 16:44:58 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0001
21/10/25 16:44:58 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_16351788
42788_0001/
21/10/25 16:44:58 INFO mapreduce.Job: Running job: job_1635178842788_0001
21/10/25 16:45:14 INFO mapreduce.Job: Job job_1635178842788_0001 running in uber mode : false
21/10/25 16:45:14 INFO mapreduce.Job:  map 0% reduce 0%
21/10/25 16:45:26 INFO mapreduce.Job:  map 100% reduce 0%
21/10/25 16:45:35 INFO mapreduce.Job:  map 100% reduce 100%
21/10/25 16:45:35 INFO mapreduce.Job: Job job_1635178842788_0001 completed successfully
21/10/25 16:45:35 INFO mapreduce.Job: Counters: 49
```

2) **Second step: Map-reduce job**



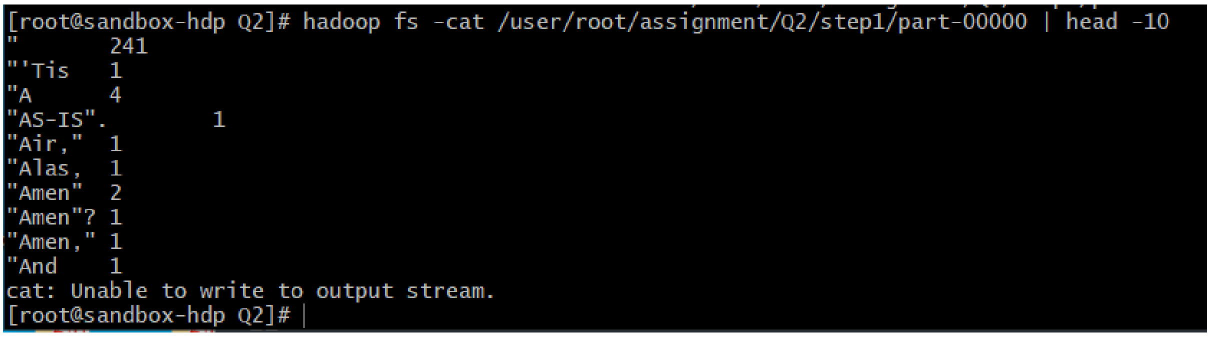
```
[root@sandbox-hdp Q2]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/Q2/wc_mapper2.py,/root/assignm
ent/Q2/wc_reducer2.py -mapper wc_mapper2.py -reducer wc_reducer2.py -input /user/root/assignment/Q2/step1/part-00000 -output /user/root/assignment/Q2/step2
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob7904103502063871291.jar tmpDir=null
21/10/25 17:01:43 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 17:01:44 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 17:01:44 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 17:01:44 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 17:01:44 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/25 17:01:44 INFO mapreduce.JobSubmitter: number of splits:2
21/10/25 17:01:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0004
21/10/25 17:01:45 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0004
21/10/25 17:01:45 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0004/
21/10/25 17:01:45 INFO mapreduce.Job: Running job: job_1635178842788_0004
21/10/25 17:01:52 INFO mapreduce.Job: Job job_1635178842788_0004 running in uber mode : false
21/10/25 17:01:52 INFO mapreduce.Job:  map 0% reduce 0%
21/10/25 17:02:00 INFO mapreduce.Job:  map 50% reduce 0%
21/10/25 17:02:01 INFO mapreduce.Job:  map 100% reduce 0%
21/10/25 17:02:08 INFO mapreduce.Job:  map 100% reduce 100%
21/10/25 17:02:09 INFO mapreduce.Job: Job job_1635178842788_0004 completed successfully
21/10/25 17:02:09 INFO mapreduce.Job: Counters: 49
```



```
[root@sandbox-hdp Q2]# hadoop fs -ls /user/root/assignment/Q2/step2
Found 2 items
-rw-r--r--   1 root root          0 2021-10-25 17:02 /user/root/assignment/Q2/step2/_SUCCESS
-rw-r--r--   1 root root         88 2021-10-25 17:02 /user/root/assignment/Q2/step2/part-00000
```

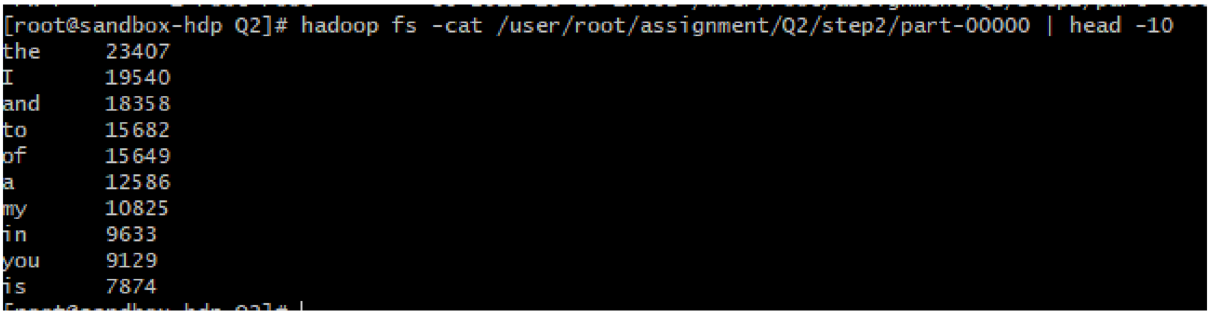
Producing correct output (3pt)

Output after step1) : unsorted



```
[root@sandbox-hdp Q2]# hadoop fs -cat /user/root/assignment/Q2/step1/part-00000 | head -10
"        241
"'Tis    1
"A       4
"AS-IS".        1
"Air,"  1
"Alas,  1
"Amen"  2
"Amen"? 1
"Amen," 1
"And    1
cat: Unable to write to output stream.
[root@sandbox-hdp Q2]# |
```

Output after step2) : sorted



```
[root@sandbox-hdp Q2]# hadoop fs -cat /user/root/assignment/Q2/step2/part-00000 | head -10
the     23407
I       19540
and     18358
to      15682
of      15649
a       12586
my      10825
in      9633
you     9129
is      7874
```

Instructions (4pt) & screenshots (2pt)

1) Make sure docker is running. Access to Ambari via 127.0.0.1:1080 portal and run services including HDFS, YARN, MAPREDUCE, ZOOKEEPER.
2) Upload files to local system via FileZilla
host: sftp://localhost username: root password: which you setup for hortonworks port:2222

3) On Git Bash, access to hortonworks sandbox by entering the following command.

   ssh root@sandbox-hdp.hortonworks.com -p 2222

4) Make directory in the HDFS using the following command.

   Hadoop fs -mkdir /user/root/assignment/A

5) Upload files from HDP Sandbox to HDFS using command

   (Hadoop fs -put filename file_location):

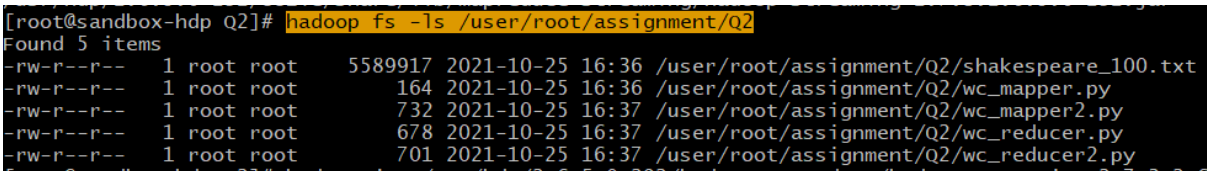
   hadoop fs -put shakespeare_100.txt ./lab/A1

   hadoop fs -put wc_mapper.py ./lab/A1

   hadoop fs -put wc_reducer.py ./lab/A1

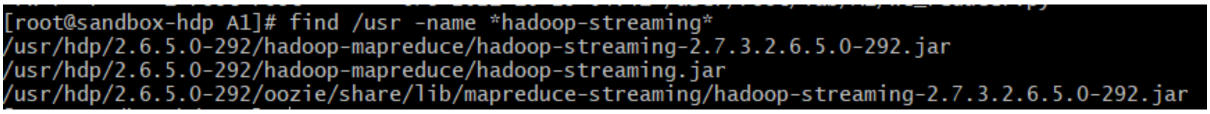
   hadoop fs -put wc_mapper2.py ./lab/A1

   hadoop fs -put wc_reducer2.py ./lab/A1



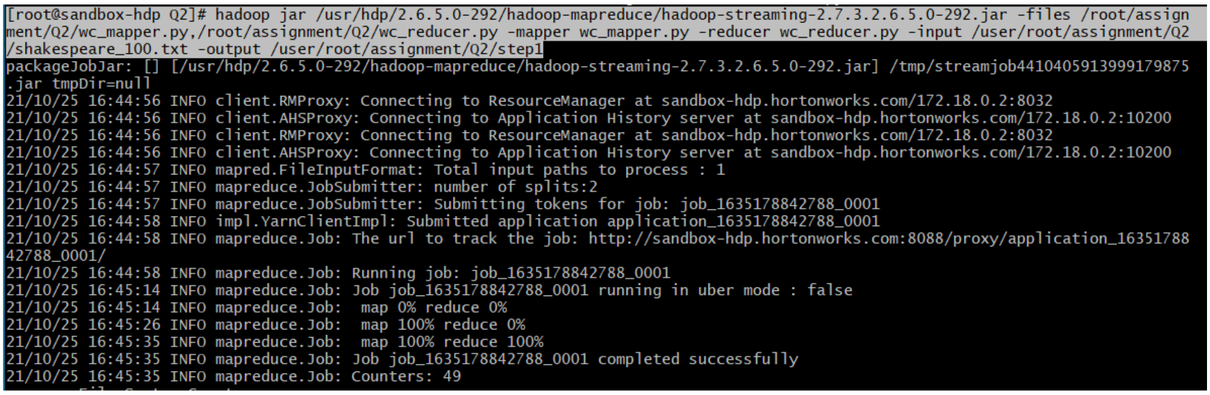
```
[root@sandbox-hdp Q2]# hadoop fs -ls /user/root/assignment/Q2
Found 5 items
-rw-r--r--   1 root root    5589917 2021-10-25 16:36 /user/root/assignment/Q2/shakespeare_100.txt
-rw-r--r--   1 root root        164 2021-10-25 16:36 /user/root/assignment/Q2/wc_mapper.py
-rw-r--r--   1 root root        732 2021-10-25 16:37 /user/root/assignment/Q2/wc_mapper2.py
-rw-r--r--   1 root root        678 2021-10-25 16:37 /user/root/assignment/Q2/wc_reducer.py
-rw-r--r--   1 root root        701 2021-10-25 16:37 /user/root/assignment/Q2/wc_reducer2.py
```

6) Find streaming file location



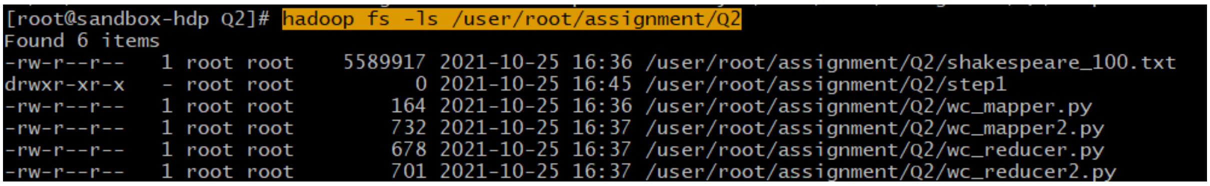
```
[root@sandbox-hdp A1]# find /usr -name *hadoop-streaming*
/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar
/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming.jar
/usr/hdp/2.6.5.0-292/oozie/share/lib/mapreduce-streaming/hadoop-streaming-2.7.3.2.6.5.0-292.jar
```

7) First step Map-reduce job

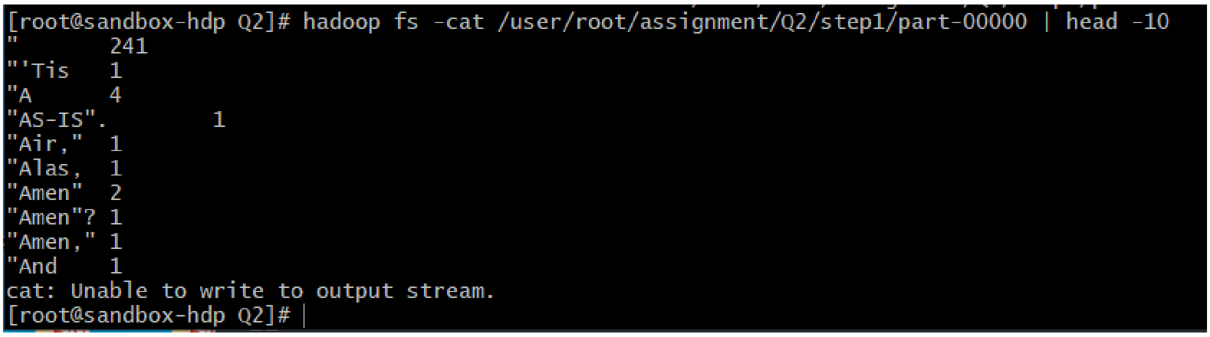


```
[root@sandbox-hdp Q2]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assign
ment/Q2/wc_mapper.py,/root/assignment/Q2/wc_reducer.py -mapper wc_mapper.py -reducer wc_reducer.py -input /user/root/assignment/Q2
/shakespeare_100.txt -output /user/root/assignment/Q2/step1
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob4410405913999179875
.jar tmpDir=null
21/10/25 16:44:56 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 16:44:56 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 16:44:56 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 16:44:56 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 16:44:57 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/25 16:44:57 INFO mapreduce.JobSubmitter: number of splits:2
21/10/25 16:44:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0001
21/10/25 16:44:58 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0001
21/10/25 16:44:58 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_16351788
42788_0001/
21/10/25 16:44:58 INFO mapreduce.Job: Running job: job_1635178842788_0001
21/10/25 16:45:14 INFO mapreduce.Job: Job job_1635178842788_0001 running in uber mode : false
21/10/25 16:45:14 INFO mapreduce.Job:  map 0% reduce 0%
21/10/25 16:45:26 INFO mapreduce.Job:  map 100% reduce 0%
21/10/25 16:45:35 INFO mapreduce.Job:  map 100% reduce 100%
21/10/25 16:45:35 INFO mapreduce.Job: Job job_1635178842788_0001 completed successfully
21/10/25 16:45:35 INFO mapreduce.Job: Counters: 49
```

8) Map-reduce output (part-00000) saved in the step1 folder

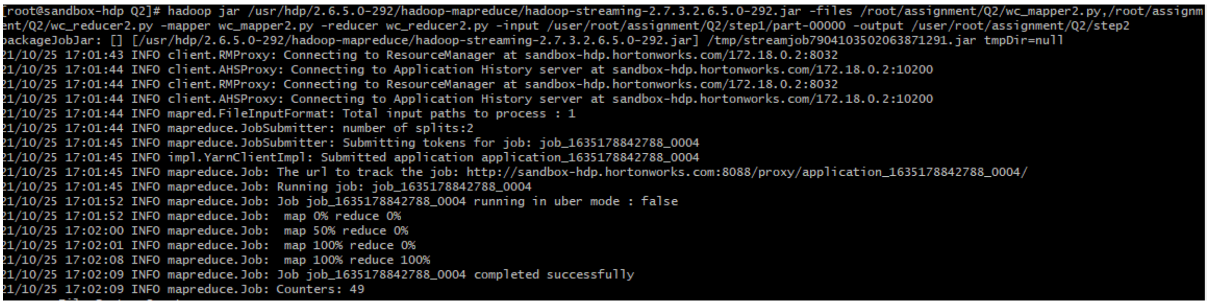


```
[root@sandbox-hdp Q2]# hadoop fs -ls /user/root/assignment/Q2
Found 6 items
-rw-r--r--   1 root root    5589917 2021-10-25 16:36 /user/root/assignment/Q2/shakespeare_100.txt
drwxr-xr-x   - root root          0 2021-10-25 16:45 /user/root/assignment/Q2/step1
-rw-r--r--   1 root root        164 2021-10-25 16:36 /user/root/assignment/Q2/wc_mapper.py
-rw-r--r--   1 root root        732 2021-10-25 16:37 /user/root/assignment/Q2/wc_mapper2.py
-rw-r--r--   1 root root        678 2021-10-25 16:37 /user/root/assignment/Q2/wc_reducer.py
-rw-r--r--   1 root root        701 2021-10-25 16:37 /user/root/assignment/Q2/wc_reducer2.py
```

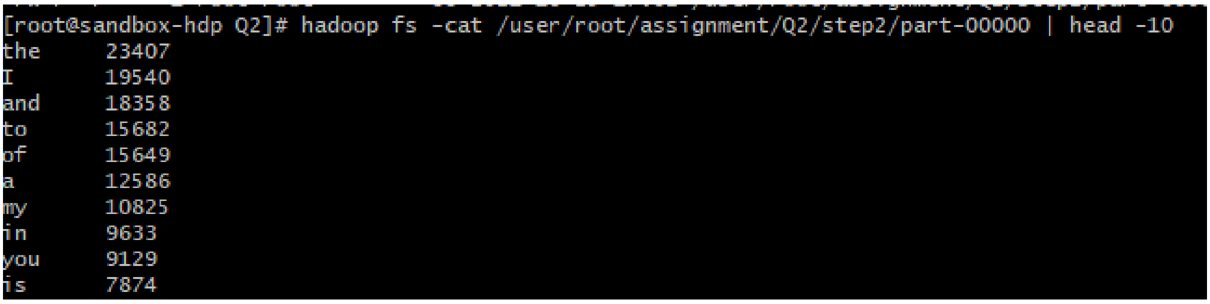


```
[root@sandbox-hdp Q2]# hadoop fs -cat /user/root/assignment/Q2/step1/part-00000 | head -10
"       241
"'Tis   1
"A      4
"AS-IS".        1
"Air,"  1
"Alas,  1
"Amen"  2
"Amen"? 1
"Amen," 1
"And    1
cat: Unable to write to output stream.
[root@sandbox-hdp Q2]# |
```

9) 2nd map-reduce job (input: step1/part-00000 > output: step2)



```
[root@sandbox-hdp Q2]# hadoop jar /usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar -files /root/assignment/Q2/wc_mapper2.py,/root/assignm
ent/Q2/wc_reducer2.py -mapper wc_mapper2.py -reducer wc_reducer2.py -input /user/root/assignment/Q2/step1/part-00000 -output /user/root/assignment/Q2/step2
packageJobJar: [] [/usr/hdp/2.6.5.0-292/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.0-292.jar] /tmp/streamjob7904103502063871291.jar tmpDir=null
21/10/25 17:01:43 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 17:01:44 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 17:01:44 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
21/10/25 17:01:44 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
21/10/25 17:01:44 INFO mapred.FileInputFormat: Total input paths to process : 1
21/10/25 17:01:44 INFO mapreduce.JobSubmitter: number of splits:2
21/10/25 17:01:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635178842788_0004
21/10/25 17:01:45 INFO impl.YarnClientImpl: Submitted application application_1635178842788_0004
21/10/25 17:01:45 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1635178842788_0004/
21/10/25 17:01:45 INFO mapreduce.Job: Running job: job_1635178842788_0004
21/10/25 17:01:52 INFO mapreduce.Job: Job job_1635178842788_0004 running in uber mode : false
21/10/25 17:01:52 INFO mapreduce.Job:  map 0% reduce 0%
21/10/25 17:02:00 INFO mapreduce.Job:  map 50% reduce 0%
21/10/25 17:02:01 INFO mapreduce.Job:  map 100% reduce 0%
21/10/25 17:02:08 INFO mapreduce.Job:  map 100% reduce 100%
21/10/25 17:02:09 INFO mapreduce.Job: Job job_1635178842788_0004 completed successfully
21/10/25 17:02:09 INFO mapreduce.Job: Counters: 49
```



```
[root@sandbox-hdp Q2]# hadoop fs -cat /user/root/assignment/Q2/step2/part-00000 | head -10
the     23407
I       19540
and     18358
to      15682
of      15649
a       12586
my      10825
in      9633
you     9129
is      7874
```