

Name: Phạm Đức Thế²

ID: 19522253

Class: DS200.M21

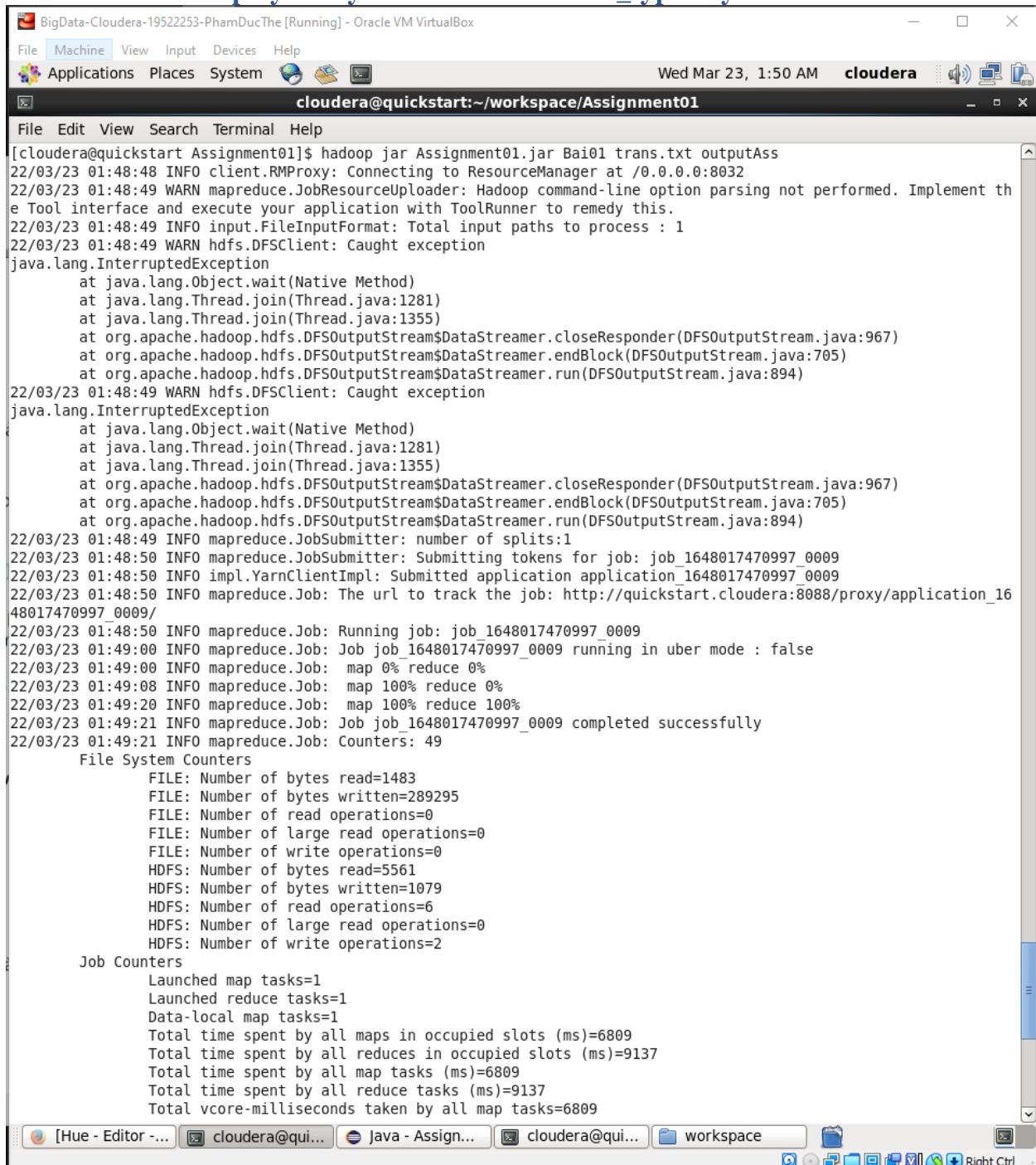
BIG DATA

SUMMARY

Task		Status	Page
Lab 01	Task 1	Hoàn thành	2
	Task 2	Hoàn thành	5

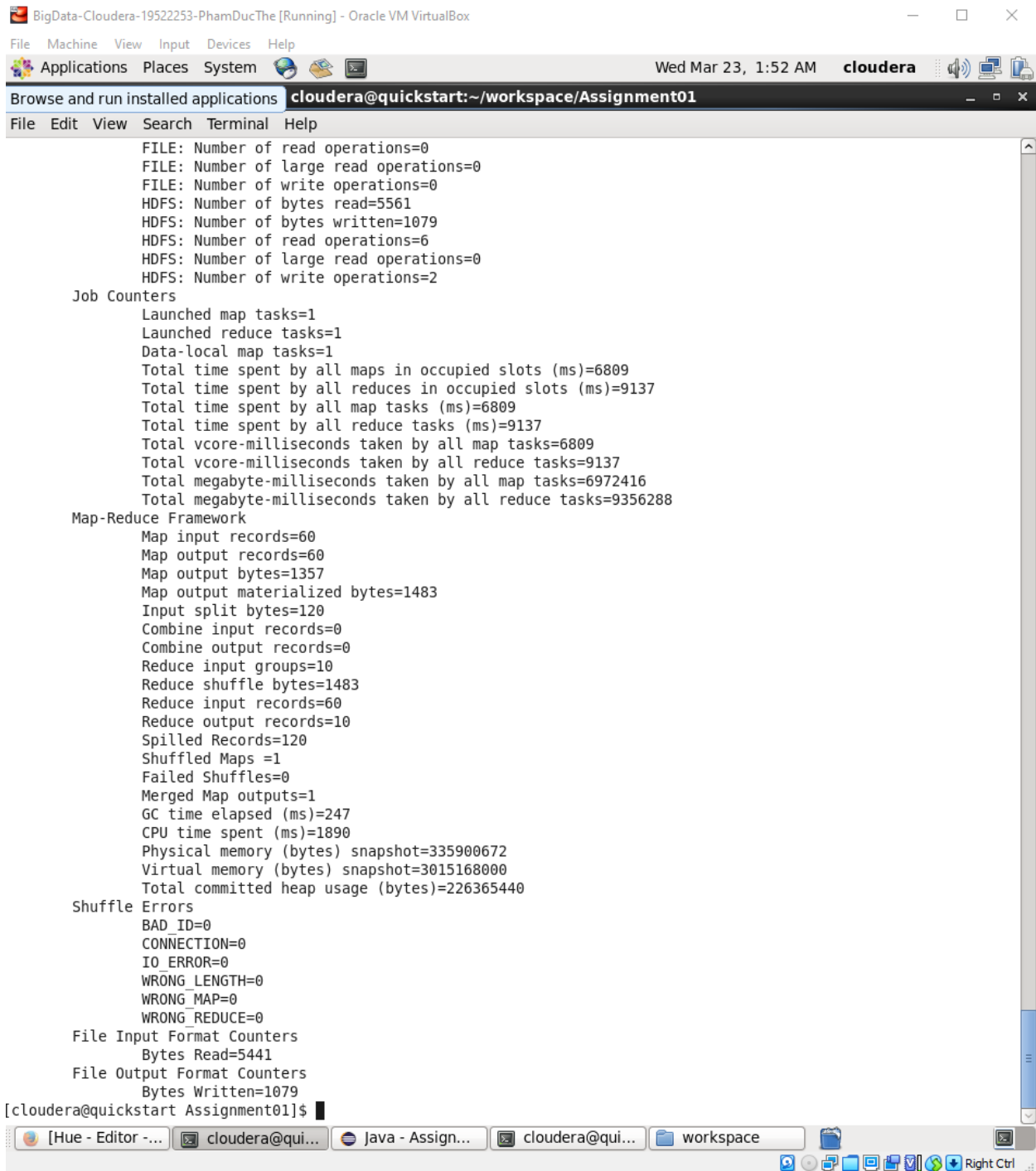
Lab 01

1. Task name 1: Với mỗi Player_ID, xuất danh sách các Game_type được chơi bởi player này và số lần chơi Game_type này



```
[cloudera@quickstart Assignment01]$ hadoop jar Assignment01.jar Bai01 trans.txt outputAss
22/03/23 01:48:48 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/03/23 01:48:49 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
e Tool interface and execute your application with ToolRunner to remedy this.
22/03/23 01:48:49 INFO input.FileInputFormat: Total input paths to process : 1
22/03/23 01:48:49 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/03/23 01:48:49 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/03/23 01:48:49 INFO mapreduce.JobSubmitter: number of splits:1
22/03/23 01:48:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1648017470997_0009
22/03/23 01:48:50 INFO impl.YarnClientImpl: Submitted application application_1648017470997_0009
22/03/23 01:48:50 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_16
48017470997_0009/
22/03/23 01:48:50 INFO mapreduce.Job: Running job: job_1648017470997_0009
22/03/23 01:49:00 INFO mapreduce.Job: Job job_1648017470997_0009 running in uber mode : false
22/03/23 01:49:00 INFO mapreduce.Job:  map 0% reduce 0%
22/03/23 01:49:08 INFO mapreduce.Job:  map 100% reduce 0%
22/03/23 01:49:20 INFO mapreduce.Job:  map 100% reduce 100%
22/03/23 01:49:21 INFO mapreduce.Job: Job job_1648017470997_0009 completed successfully
22/03/23 01:49:21 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=1483
        FILE: Number of bytes written=289295
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=5561
        HDFS: Number of bytes written=1079
        HDFS: Number of read operations=6
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=6809
        Total time spent by all reduces in occupied slots (ms)=9137
        Total time spent by all map tasks (ms)=6809
        Total time spent by all reduce tasks (ms)=9137
        Total vcore-milliseconds taken by all map tasks=6809
```

Hình 1: Chạy lệnh submit map reduce bài 1



The screenshot shows a terminal window titled "cloudera@quickstart:~/workspace/Assignment01". The terminal output displays the results of a Hadoop MapReduce job. The output is organized into several sections: File operations, HDFS statistics, Job Counters, Map-Reduce Framework, Shuffle Errors, and File Input/Output Format Counters. The terminal window is part of an Oracle VM VirtualBox environment, as indicated by the title bar. The bottom of the screen shows a taskbar with several open applications, including Hue, cloudera@qui..., Java - Assign..., cloudera@qui..., and workspace.

```
BigData-Cloudera-19522253-PhamDucThe [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Browse and run installed applications cloudera@quickstart:~/workspace/Assignment01
File Edit View Search Terminal Help

FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5561
HDFS: Number of bytes written=1079
HDFS: Number of read operations=6
HDFS: Number of large read operations=0
HDFS: Number of write operations=2

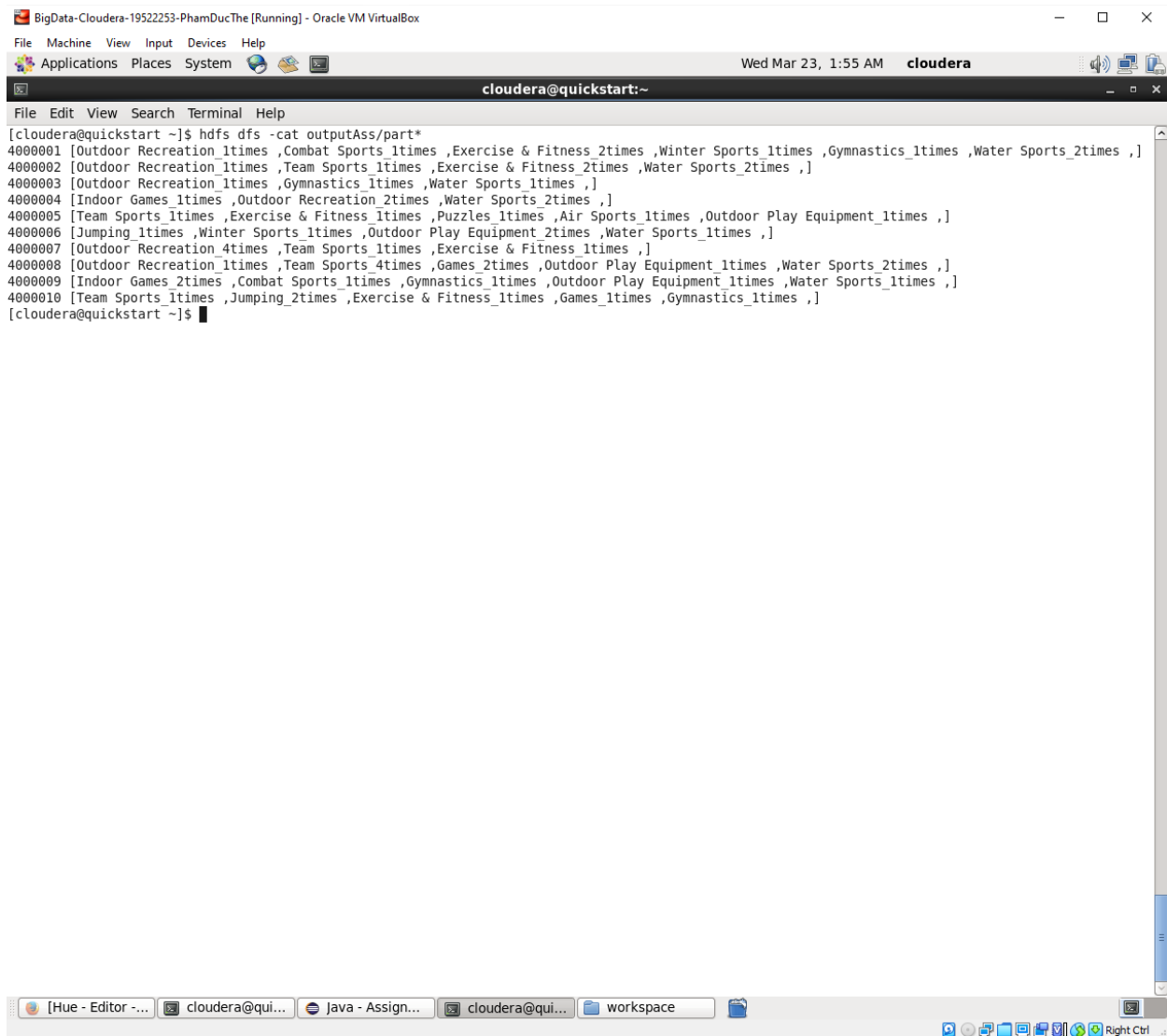
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=6809
  Total time spent by all reduces in occupied slots (ms)=9137
  Total time spent by all map tasks (ms)=6809
  Total time spent by all reduce tasks (ms)=9137
  Total vcore-milliseconds taken by all map tasks=6809
  Total vcore-milliseconds taken by all reduce tasks=9137
  Total megabyte-milliseconds taken by all map tasks=6972416
  Total megabyte-milliseconds taken by all reduce tasks=9356288

Map-Reduce Framework
  Map input records=60
  Map output records=60
  Map output bytes=1357
  Map output materialized bytes=1483
  Input split bytes=120
  Combine input records=0
  Combine output records=0
  Reduce input groups=10
  Reduce shuffle bytes=1483
  Reduce input records=60
  Reduce output records=10
  Spilled Records=120
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=247
  CPU time spent (ms)=1890
  Physical memory (bytes) snapshot=335900672
  Virtual memory (bytes) snapshot=3015168000
  Total committed heap usage (bytes)=226365440

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=5441
File Output Format Counters
  Bytes Written=1079
[cloudera@quickstart Assignment01]$
```

Hình 2: Chạy lệnh submit map reduce bài 1

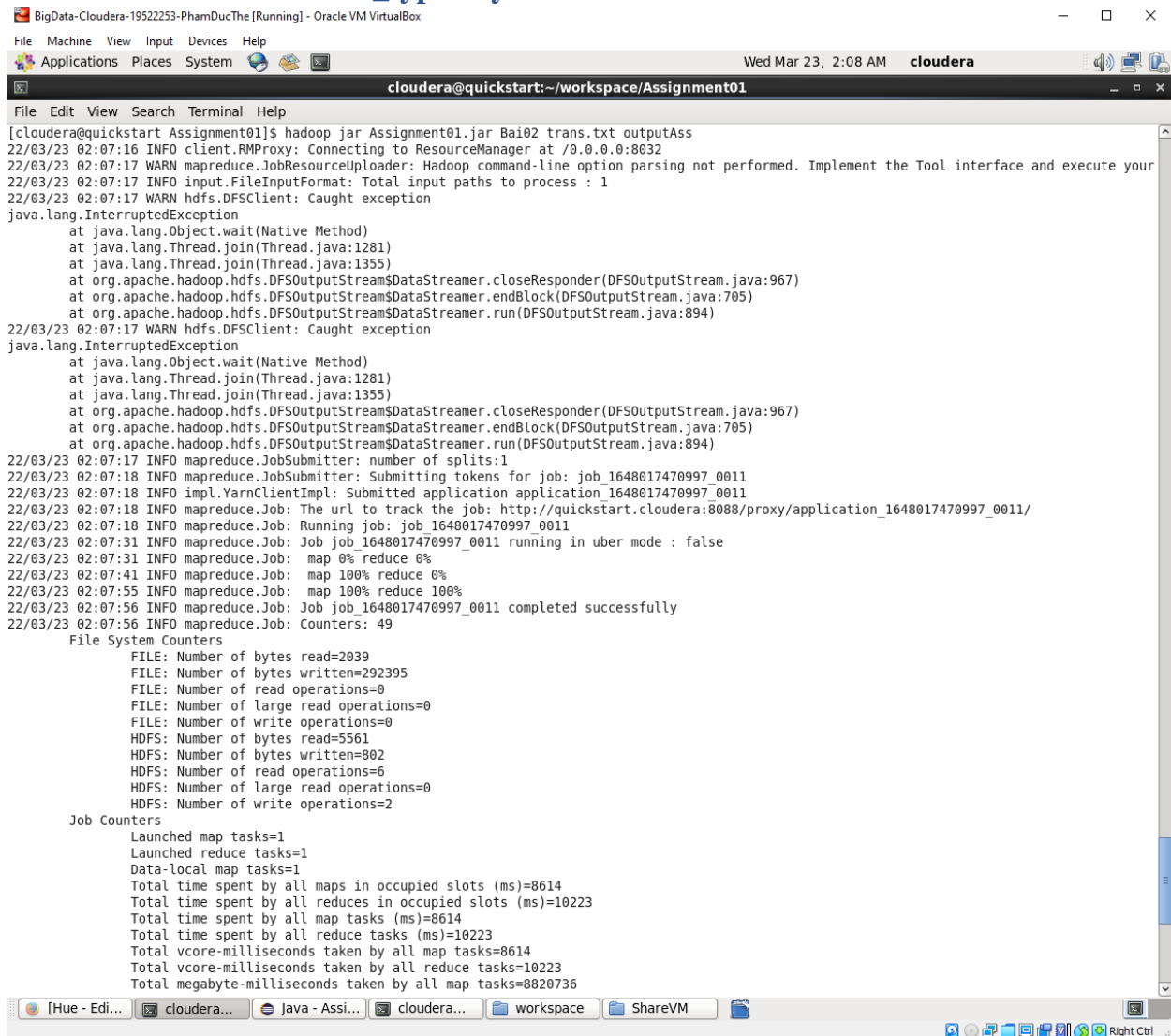


The screenshot shows a terminal window titled "cloudera@quickstart:~". The terminal displays the output of the command `hdfs dfs -cat outputAss/part*`. The output consists of ten lines of text, each representing a file's contents. The files are named with IDs from 4000001 to 4000010. The contents of these files are lists of sports-related terms, such as "Outdoor Recreation", "Combat Sports", "Exercise & Fitness", "Winter Sports", "Gymnastics", and "Water Sports", followed by a count of occurrences (e.g., "1times", "2times", "4times").

```
[cloudera@quickstart ~]$ hdfs dfs -cat outputAss/part*
4000001 [Outdoor Recreation 1times ,Combat Sports 1times ,Exercise & Fitness 2times ,Winter Sports 1times ,Gymnastics 1times ,Water Sports 2times ,]
4000002 [Outdoor Recreation 1times ,Team Sports 1times ,Exercise & Fitness 2times ,Water Sports 2times ,]
4000003 [Outdoor Recreation 1times ,Gymnastics 1times ,Water Sports 1times ,]
4000004 [Indoor Games 1times ,Outdoor Recreation 2times ,Water Sports 2times ,]
4000005 [Team Sports 1times ,Exercise & Fitness 1times ,Puzzles 1times ,Air Sports 1times ,Outdoor Play Equipment 1times ,]
4000006 [Jumping 1times ,Winter Sports 1times ,Outdoor Play Equipment 2times ,Water Sports 1times ,]
4000007 [Outdoor Recreation 4times ,Team Sports 1times ,Exercise & Fitness 1times ,]
4000008 [Outdoor Recreation 1times ,Team Sports 4times ,Games 2times ,Outdoor Play Equipment 1times ,Water Sports 2times ,]
4000009 [Indoor Games 2times ,Combat Sports 1times ,Gymnastics 1times ,Outdoor Play Equipment 1times ,Water Sports 1times ,]
4000010 [Team Sports 1times ,Jumping 2times ,Exercise & Fitness 1times ,Games 1times ,Gymnastics 1times ,]
```

Hình 3: Kết quả bài 1

2. Task name 2: Với mỗi Game_type, xuất danh sách tên và tuổi của những player (không trùng ID) đã chơi game_type này và tổng số tiền các player đã trả cho Game_type này.

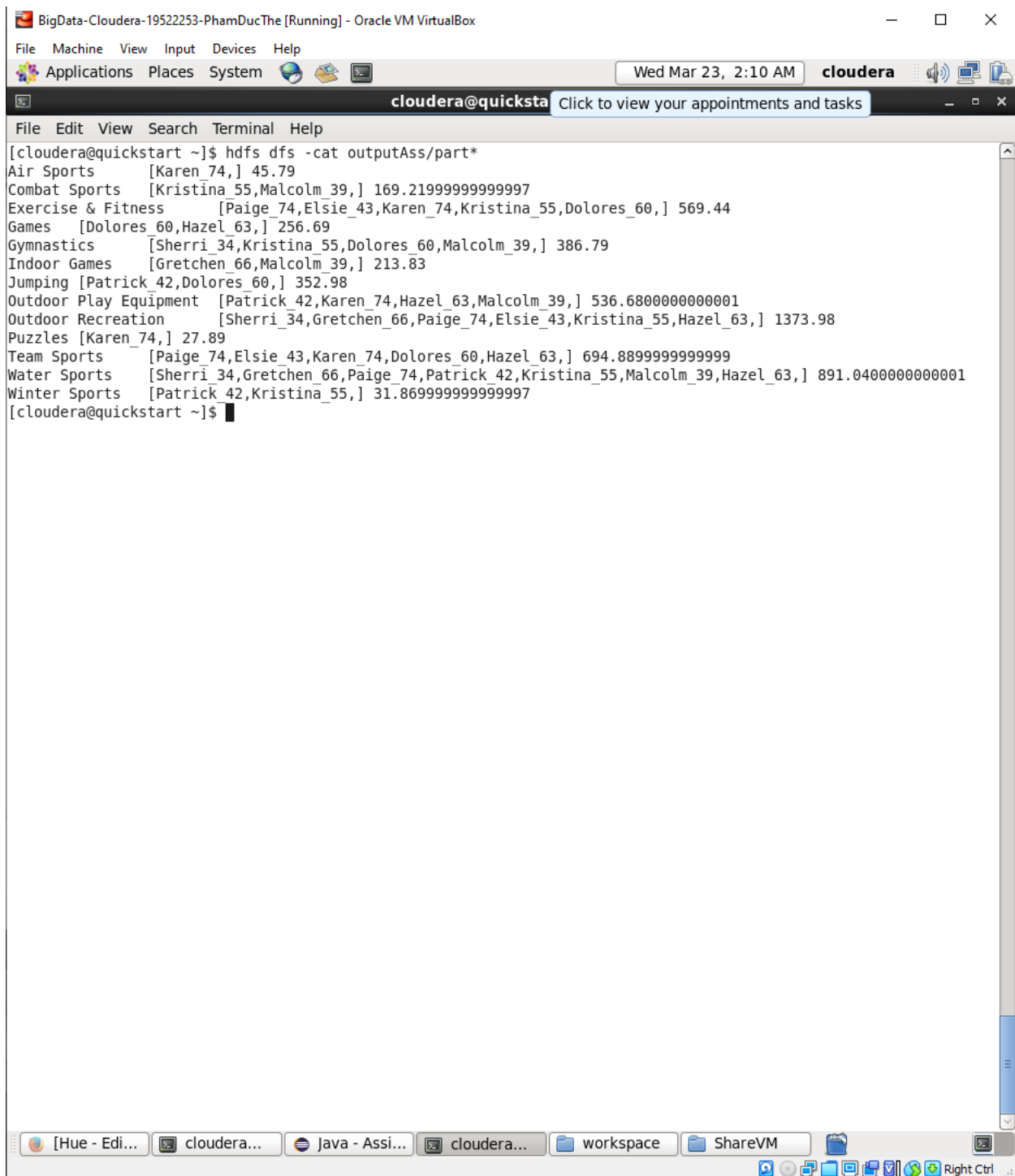


```
BigData-Cloudera-19522253-PhamDucThe [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Wed Mar 23, 2:08 AM cloudera
cloudera@quickstart:~/workspace/Assignment01
File Edit View Search Terminal Help
[cloudera@quickstart Assignment01]$ hadoop jar Assignment01.jar Bai02 trans.txt outputAss
22/03/23 02:07:16 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/03/23 02:07:17 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your
22/03/23 02:07:17 INFO input.FileInputFormat: Total input paths to process : 1
22/03/23 02:07:17 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/03/23 02:07:17 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
22/03/23 02:07:17 INFO mapreduce.JobSubmitter: number of splits:1
22/03/23 02:07:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1648017470997_0011
22/03/23 02:07:18 INFO impl.YarnClientImpl: Submitted application application_1648017470997_0011
22/03/23 02:07:18 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1648017470997_0011/
22/03/23 02:07:18 INFO mapreduce.Job: Running job: job_1648017470997_0011
22/03/23 02:07:31 INFO mapreduce.Job: Job job_1648017470997_0011 running in uber mode : false
22/03/23 02:07:31 INFO mapreduce.Job: map 0% reduce 0%
22/03/23 02:07:41 INFO mapreduce.Job: map 100% reduce 0%
22/03/23 02:07:55 INFO mapreduce.Job: map 100% reduce 100%
22/03/23 02:07:56 INFO mapreduce.Job: Job job_1648017470997_0011 completed successfully
22/03/23 02:07:56 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=2039
  FILE: Number of bytes written=292395
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=5561
  HDFS: Number of bytes written=802
  HDFS: Number of read operations=6
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=8614
  Total time spent by all reduces in occupied slots (ms)=10223
  Total time spent by all map tasks (ms)=8614
  Total time spent by all reduce tasks (ms)=10223
  Total vcore-milliseconds taken by all map tasks=8614
  Total vcore-milliseconds taken by all reduce tasks=10223
  Total megabyte-milliseconds taken by all map tasks=8820736
```

Hình 4: Chạy lệnh submit map reduce bài 2

```
BigData-Cloudera-1952253-PhamDucThe [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/workspace/Assignment01
File Edit View Search Terminal Help
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=5561
HDFS: Number of bytes written=802
HDFS: Number of read operations=6
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
  Launched map tasks=1
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=8614
  Total time spent by all reduces in occupied slots (ms)=10223
  Total time spent by all map tasks (ms)=8614
  Total time spent by all reduce tasks (ms)=10223
  Total vcore-milliseconds taken by all map tasks=8614
  Total vcore-milliseconds taken by all reduce tasks=10223
  Total megabyte-milliseconds taken by all map tasks=8820736
  Total megabyte-milliseconds taken by all reduce tasks=10468352
Map-Reduce Framework
  Map input records=60
  Map output records=60
  Map output bytes=1913
  Map output materialized bytes=2039
  Input split bytes=120
  Combine input records=0
  Combine output records=0
  Reduce input groups=13
  Reduce shuffle bytes=2039
  Reduce input records=60
  Reduce output records=13
  Spilled Records=120
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=213
  CPU time spent (ms)=2030
  Physical memory (bytes) snapshot=342413312
  Virtual memory (bytes) snapshot=3015159808
  Total committed heap usage (bytes)=226365440
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=5441
File Output Format Counters
  Bytes Written=802
[cloudera@quickstart Assignment01]$
```

Hình 5: Chạy lệnh submit map reduce bài 2



The screenshot shows a terminal window titled "BigData-Cloudera-19522253-PhamDucThe [Running] - Oracle VM VirtualBox". The terminal is running the command `hdfs dfs -cat outputAss/part*`. The output lists various sports and games with associated names and numerical values. The terminal window has a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar. The bottom of the screen shows a taskbar with several open applications: Hue - Edi..., cloudera..., Java - Assi..., cloudera..., workspace, and ShareVM.

```
[cloudera@quickstart ~]$ hdfs dfs -cat outputAss/part*
Air Sports      [Karen 74,] 45.79
Combat Sports   [Kristina 55,Malcolm 39,] 169.21999999999997
Exercise & Fitness [Paige 74,Elsie 43,Karen 74,Kristina 55,Dolores 60,] 569.44
Games [Dolores 60,Hazel 63,] 256.69
Gymnastics      [Sherri 34,Kristina 55,Dolores 60,Malcolm 39,] 386.79
Indoor Games    [Gretchen 66,Malcolm 39,] 213.83
Jumping [Patrick 42,Dolores 60,] 352.98
Outdoor Play Equipment [Patrick 42,Karen 74,Hazel 63,Malcolm 39,] 536.6800000000001
Outdoor Recreation [Sherri 34,Gretchen 66,Paige 74,Elsie 43,Kristina 55,Hazel 63,] 1373.98
Puzzles [Karen 74,] 27.89
Team Sports     [Paige 74,Elsie 43,Karen 74,Dolores 60,Hazel 63,] 694.8899999999999
Water Sports    [Sherri 34,Gretchen 66,Paige 74,Patrick 42,Kristina 55,Malcolm 39,Hazel 63,] 891.0400000000001
Winter Sports   [Patrick 42,Kristina 55,] 31.869999999999997
[cloudera@quickstart ~]$
```

Hình 6: Kết quả bài 2

3. Task name 3

...