MapReduce Tasks:

Task 4. Write MapReduce codes to perform the tasks using the files you've downloaded on your EMR Instance:

We ran the MR Jobs on the files in Hadoop Cluster using the following command:

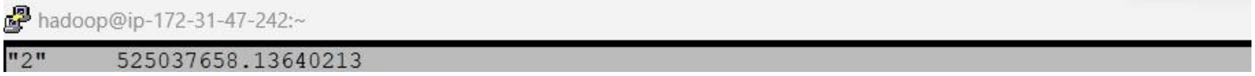
python *mrjobfile.py* -r Hadoop hdfs://user/Hadoop/*csv > *outputfile.txt*

- *mrjobfile.py* – is the python file which has the code for MR Job
- All the 6 data files are in Hadoop cluster in path - */user/Hadoop/*
- The outputs will be saved in file – ***outputfile.txt***

<u>Answers</u>

a) Which vendors have the most trips, and what is the total revenue generated by that vendor?

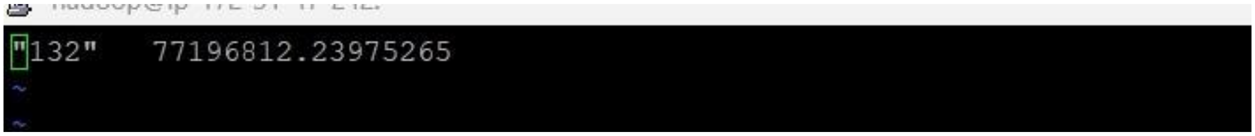
Ans. Vendor "2" has the most trips with total revenue of 525037658.13640213.



```
"2"      525037658.13640213
```

b) Which pickup location generates the most revenue?

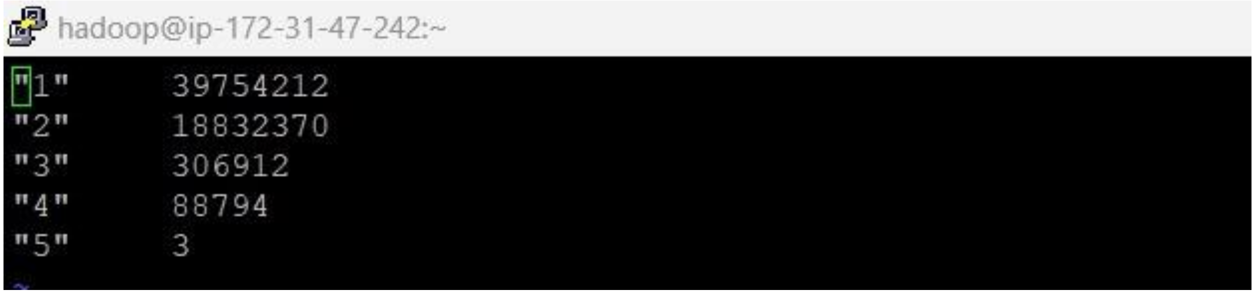
Ans. Pick Up Location 132 generates the highest revenue of 77196812.23975265



```
"132"    77196812.23975265
~
~
```

c) What are the different payment types used by customers and their count? The final results should be in a sorted format.

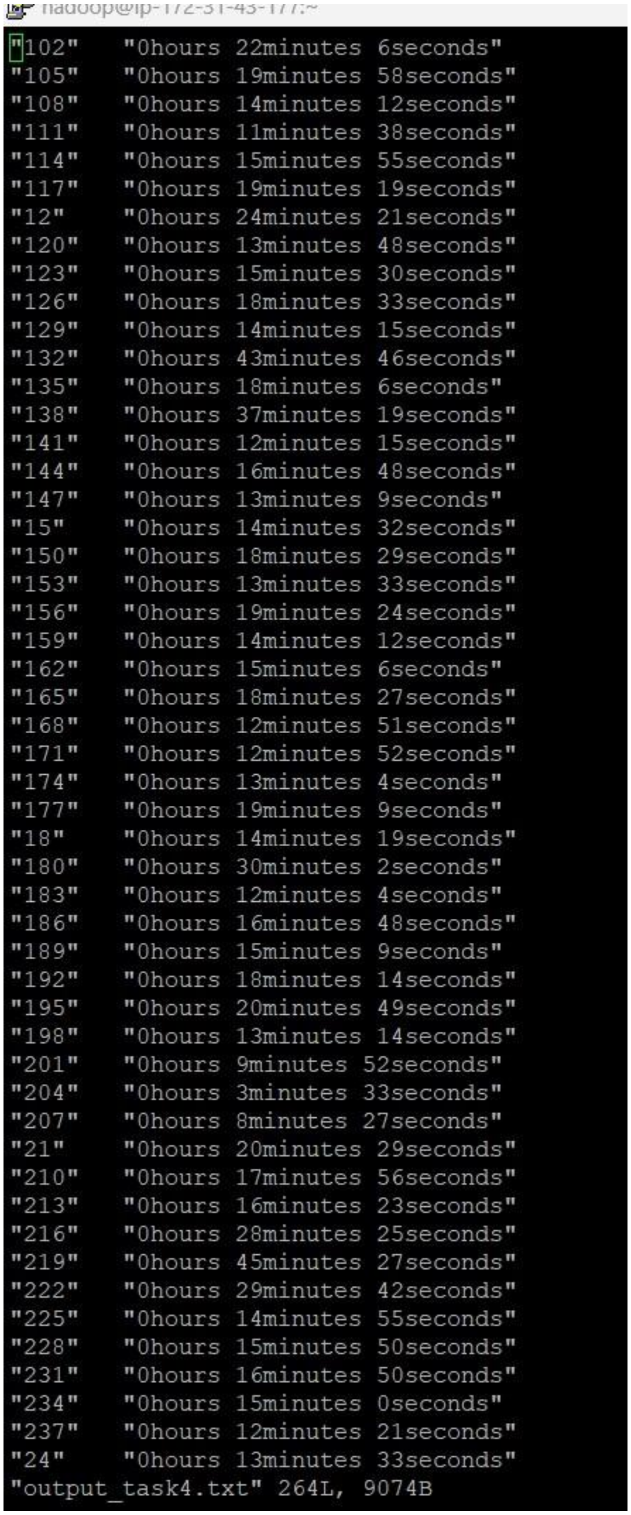
Ans. The payment types used by the customers in descending order of their counts are – 1,2,3,4 and 5.



```
"1"      39754212
"2"      18832370
"3"      306912
"4"      88794
"5"      3
~
```

d) What is the average trip time for different pickup locations?

Ans. The average trip time for different pickup locations are as follows. There are a total of 264 Pick Up Locations. Also attaching the final output file – **output_task4.txt**



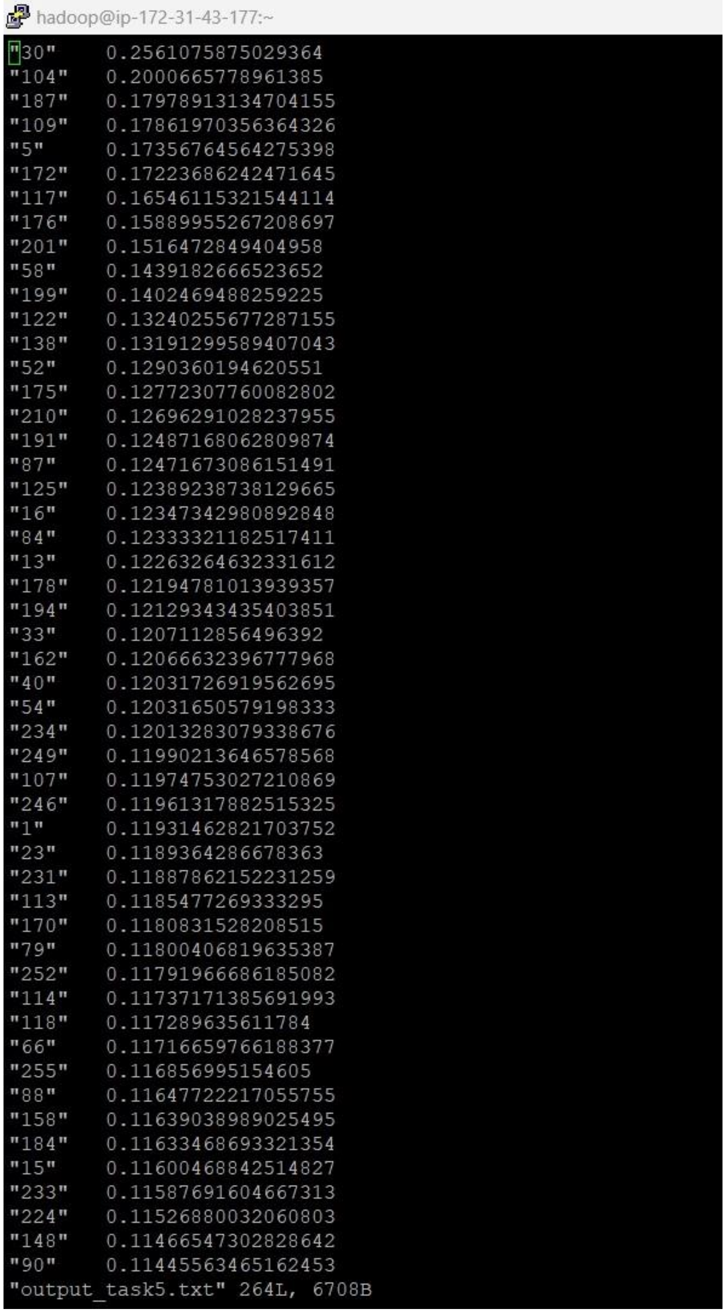
hadoop@ip-172-31-43-177:~

```
"102"    "0hours 22minutes 6seconds"
"105"    "0hours 19minutes 58seconds"
"108"    "0hours 14minutes 12seconds"
"111"    "0hours 11minutes 38seconds"
"114"    "0hours 15minutes 55seconds"
"117"    "0hours 19minutes 19seconds"
"12"     "0hours 24minutes 21seconds"
"120"    "0hours 13minutes 48seconds"
"123"    "0hours 15minutes 30seconds"
"126"    "0hours 18minutes 33seconds"
"129"    "0hours 14minutes 15seconds"
"132"    "0hours 43minutes 46seconds"
"135"    "0hours 18minutes 6seconds"
"138"    "0hours 37minutes 19seconds"
"141"    "0hours 12minutes 15seconds"
"144"    "0hours 16minutes 48seconds"
"147"    "0hours 13minutes 9seconds"
"15"     "0hours 14minutes 32seconds"
"150"    "0hours 18minutes 29seconds"
"153"    "0hours 13minutes 33seconds"
"156"    "0hours 19minutes 24seconds"
"159"    "0hours 14minutes 12seconds"
"162"    "0hours 15minutes 6seconds"
"165"    "0hours 18minutes 27seconds"
"168"    "0hours 12minutes 51seconds"
"171"    "0hours 12minutes 52seconds"
"174"    "0hours 13minutes 4seconds"
"177"    "0hours 19minutes 9seconds"
"18"     "0hours 14minutes 19seconds"
"180"    "0hours 30minutes 2seconds"
"183"    "0hours 12minutes 4seconds"
"186"    "0hours 16minutes 48seconds"
"189"    "0hours 15minutes 9seconds"
"192"    "0hours 18minutes 14seconds"
"195"    "0hours 20minutes 49seconds"
"198"    "0hours 13minutes 14seconds"
"201"    "0hours 9minutes 52seconds"
"204"    "0hours 3minutes 33seconds"
"207"    "0hours 8minutes 27seconds"
"21"     "0hours 20minutes 29seconds"
"210"    "0hours 17minutes 56seconds"
"213"    "0hours 16minutes 23seconds"
"216"    "0hours 28minutes 25seconds"
"219"    "0hours 45minutes 27seconds"
"222"    "0hours 29minutes 42seconds"
"225"    "0hours 14minutes 55seconds"
"228"    "0hours 15minutes 50seconds"
"231"    "0hours 16minutes 50seconds"
"234"    "0hours 15minutes 0seconds"
"237"    "0hours 12minutes 21seconds"
"24"     "0hours 13minutes 33seconds"
"output_task4.txt" 264L, 9074B
```

e) Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.

Ans. The average tips to revenue ratio for different pickup locations are as follows. Also attaching the final output file **output_task5.txt**

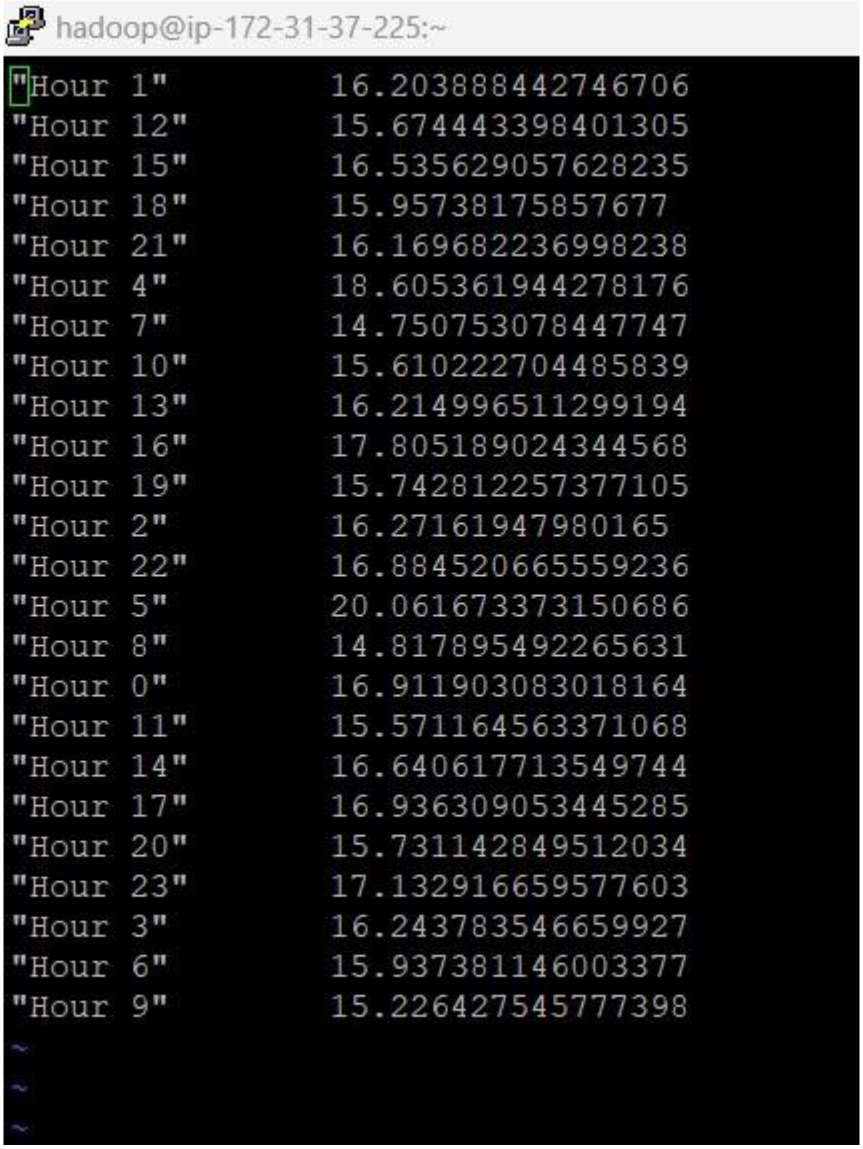
Pick Up Location "30" has the highest tips to revenue ratio of 0.2561. There are a total of 264 Pick Up Locations.



```
"30"    0.2561075875029364
"104"   0.2000665778961385
"187"   0.17978913134704155
"109"   0.17861970356364326
"5"     0.17356764564275398
"172"   0.17223686242471645
"117"   0.16546115321544114
"176"   0.15889955267208697
"201"   0.1516472849404958
"58"    0.1439182666523652
"199"   0.1402469488259225
"122"   0.13240255677287155
"138"   0.13191299589407043
"52"    0.1290360194620551
"175"   0.12772307760082802
"210"   0.12696291028237955
"191"   0.12487168062809874
"87"    0.12471673086151491
"125"   0.12389238738129665
"16"    0.12347342980892848
"84"    0.12333321182517411
"13"    0.12263264632331612
"178"   0.12194781013939357
"194"   0.12129343435403851
"33"    0.1207112856496392
"162"   0.12066632396777968
"40"    0.12031726919562695
"54"    0.12031650579198333
"234"   0.12013283079338676
"249"   0.11990213646578568
"107"   0.11974753027210869
"246"   0.11961317882515325
"1"     0.11931462821703752
"23"    0.1189364286678363
"231"   0.11887862152231259
"113"   0.1185477269333295
"170"   0.1180831528208515
"79"    0.11800406819635387
"252"   0.11791966686185082
"114"   0.11737171385691993
"118"   0.117289635611784
"66"    0.11716659766188377
"255"   0.116856995154605
"88"    0.11647722217055755
"158"   0.11639038989025495
"184"   0.11633468693321354
"15"    0.11600468842514827
"233"   0.11587691604667313
"224"   0.11526880032060803
"148"   0.11466547302828642
"90"    0.11445563465162453
"output_task5.txt" 264L, 6708B
```

f) Part1 : How does revenue vary over time? Calculate the average trip revenue analyzing it by hour of the day – Hour 0, Hour 1, Hour 2 etc.

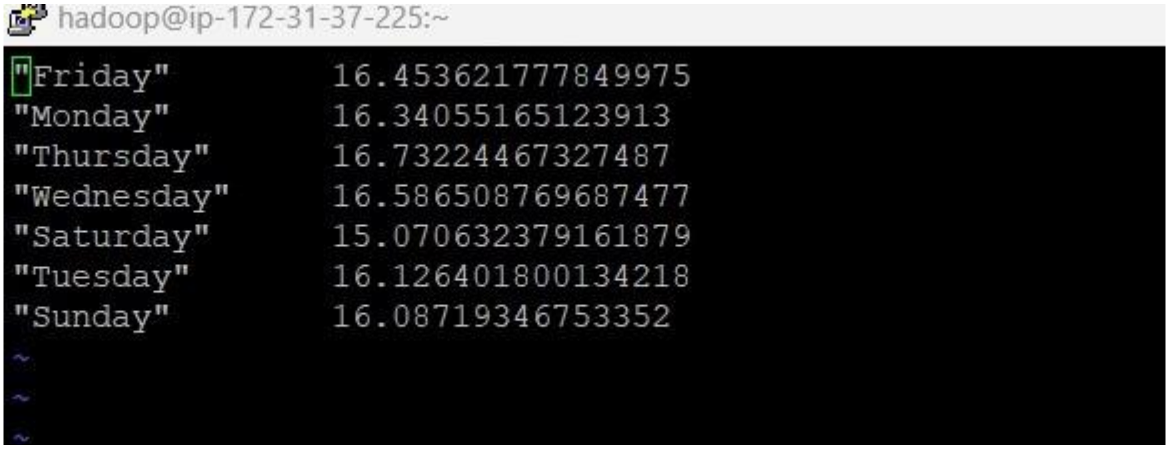
Ans. Following is the result when we executed the job to calculate the average trip revenue per hour of the day. Hour 5 (5AM to 5.59AM) has the highest revenue average of 20.06, Hour 7 (7AM to 7.59AM) has the lowest revenue average of 14.75



```
"Hour 1"        16.203888442746706
"Hour 12"       15.674443398401305
"Hour 15"       16.535629057628235
"Hour 18"       15.95738175857677
"Hour 21"       16.169682236998238
"Hour 4"        18.605361944278176
"Hour 7"        14.750753078447747
"Hour 10"       15.610222704485839
"Hour 13"       16.214996511299194
"Hour 16"       17.805189024344568
"Hour 19"       15.742812257377105
"Hour 2"        16.27161947980165
"Hour 22"       16.884520665559236
"Hour 5"        20.061673373150686
"Hour 8"        14.817895492265631
"Hour 0"        16.911903083018164
"Hour 11"       15.571164563371068
"Hour 14"       16.640617713549744
"Hour 17"       16.936309053445285
"Hour 20"       15.731142849512034
"Hour 23"       17.132916659577603
"Hour 3"        16.243783546659927
"Hour 6"        15.937381146003377
"Hour 9"        15.226427545777398
~
~
~
```

Part2 : How does revenue vary over time? Calculate the average trip revenue analyzing it by day of the week – Monday, Tuesday, Wednesday etc.

Ans. Following is the result when we executed the job to calculate the average trip revenue by day of the week. Thursday has the highest average revenue of 16.73 while Saturday has the lowest average revenue of 15.07.



```
"Friday"        16.453621777849975
"Monday"        16.34055165123913
"Thursday"      16.73224467327487
"Wednesday"     16.586508769687477
"Saturday"      15.070632379161879
"Tuesday"       16.126401800134218
"Sunday"        16.08719346753352
~
~
~
```