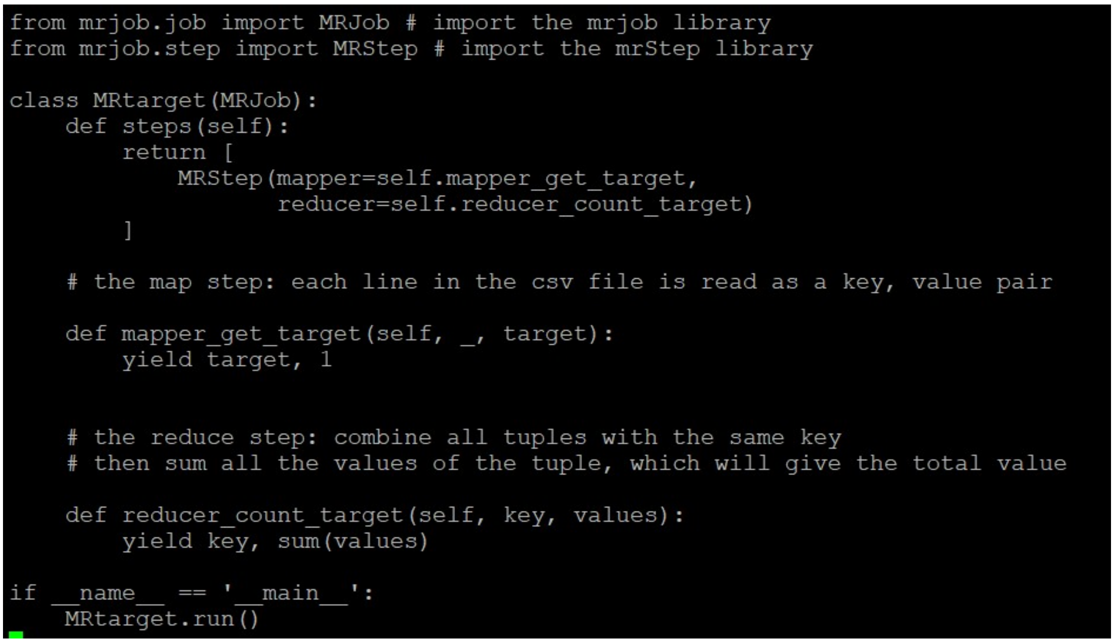
Count the change categories for each year, month and day

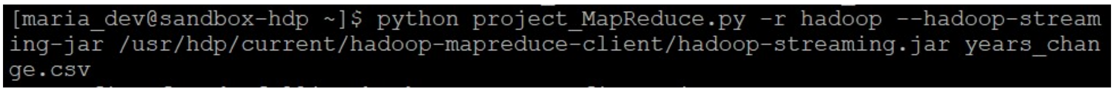
- MapReduce function:



```
from mrjob.job import MRJob # import the mrjob library
from mrjob.step import MRStep # import the mrStep library

class MRtarget(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_target,
                   reducer=self.reducer_count_target)
        ]

    # the map step: each line in the csv file is read as a key, value pair

    def mapper_get_target(self, _, target):
        yield target, 1


    # the reduce step: combine all tuples with the same key
    # then sum all the values of the tuple, which will give the total value

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

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

1- Year:

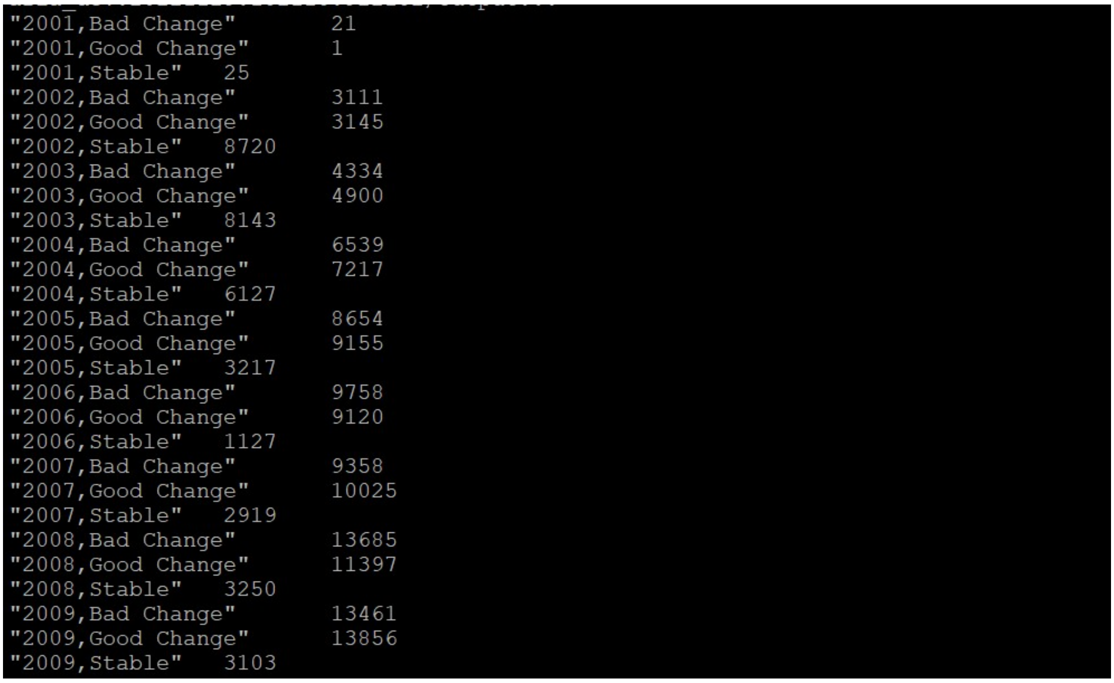
- Running Hadoop Cluster on a year_change data:



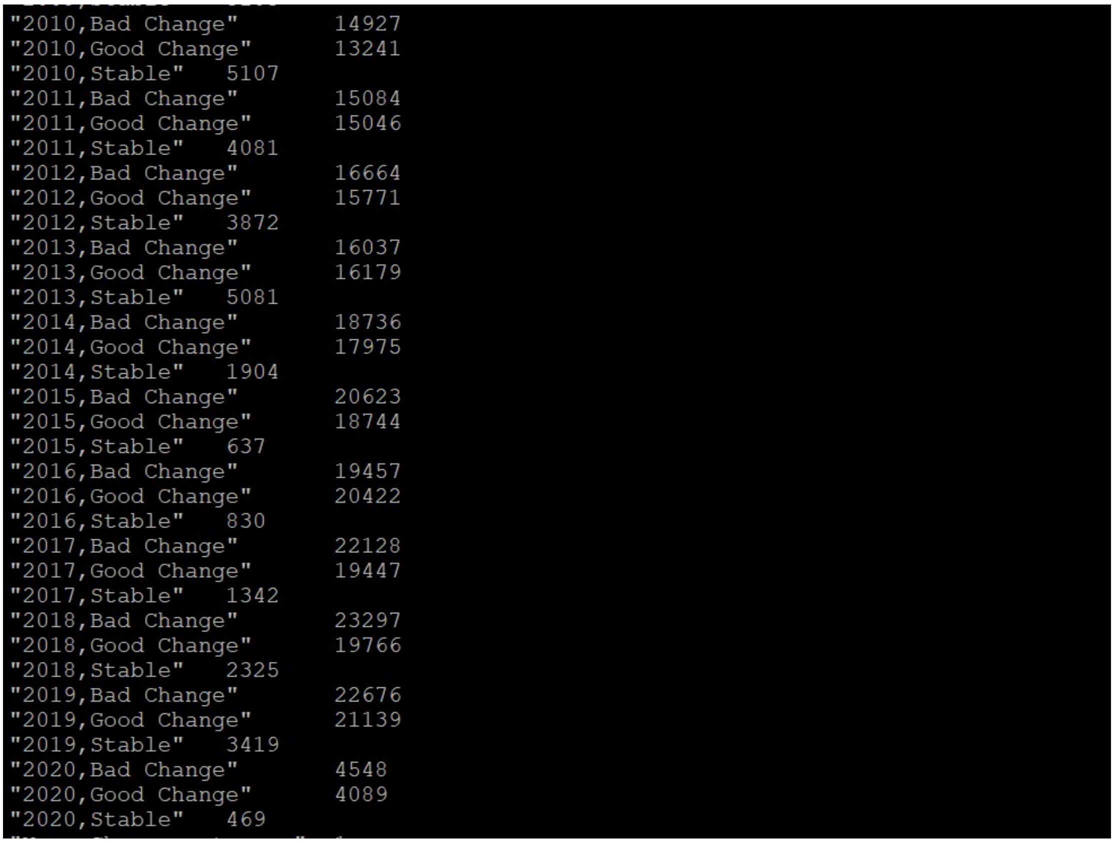
```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar years_chan
ge.csv
```

- Output:

- 2019 is the year with the highest Good change (21139)
- 2018 is the year with the highest Bad Change (23297)
- 2002 is the most stable year (8720)



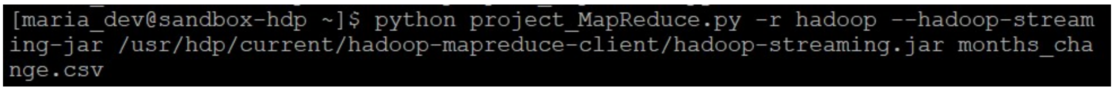
```
"2001,Bad Change"       21
"2001,Good Change"      1
"2001,Stable"   25
"2002,Bad Change"       3111
"2002,Good Change"      3145
"2002,Stable"   8720
"2003,Bad Change"       4334
"2003,Good Change"      4900
"2003,Stable"   8143
"2004,Bad Change"       6539
"2004,Good Change"      7217
"2004,Stable"   6127
"2005,Bad Change"       8654
"2005,Good Change"      9155
"2005,Stable"   3217
"2006,Bad Change"       9758
"2006,Good Change"      9120
"2006,Stable"   1127
"2007,Bad Change"       9358
"2007,Good Change"      10025
"2007,Stable"   2919
"2008,Bad Change"       13685
"2008,Good Change"      11397
"2008,Stable"   3250
"2009,Bad Change"       13461
"2009,Good Change"      13856
"2009,Stable"   3103
```



```
"2010,Bad Change"       14927
"2010,Good Change"      13241
"2010,Stable"   5107
"2011,Bad Change"       15084
"2011,Good Change"      15046
"2011,Stable"   4081
"2012,Bad Change"       16664
"2012,Good Change"      15771
"2012,Stable"   3872
"2013,Bad Change"       16037
"2013,Good Change"      16179
"2013,Stable"   5081
"2014,Bad Change"       18736
"2014,Good Change"      17975
"2014,Stable"   1904
"2015,Bad Change"       20623
"2015,Good Change"      18744
"2015,Stable"   637
"2016,Bad Change"       19457
"2016,Good Change"      20422
"2016,Stable"   830
"2017,Bad Change"       22128
"2017,Good Change"      19447
"2017,Stable"   1342
"2018,Bad Change"       23297
"2018,Good Change"      19766
"2018,Stable"   2325
"2019,Bad Change"       22676
"2019,Good Change"      21139
"2019,Stable"   3419
"2020,Bad Change"       4548
"2020,Good Change"      4089
"2020,Stable"   469
```

2- months:

- Running Hadoop Cluster on a month_change data:



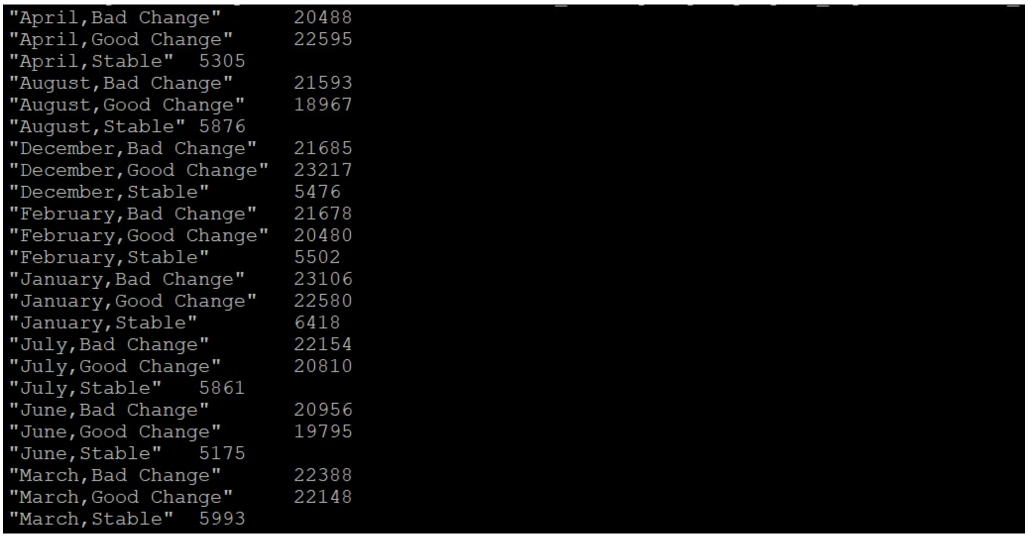
```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar months_cha
nge.csv
```

Output:

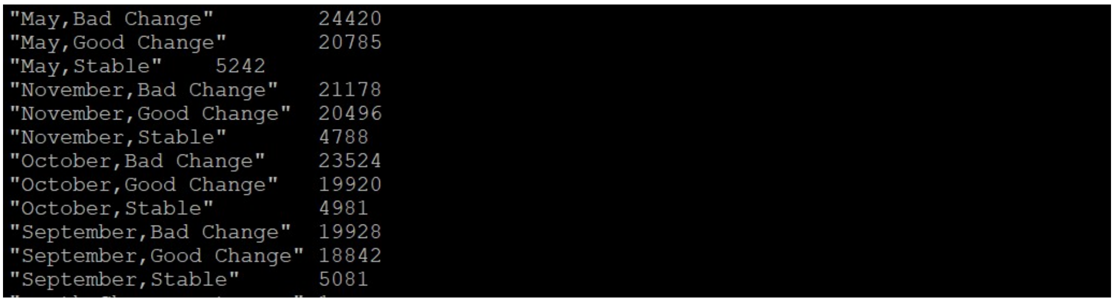
- December is the month with the highest Good change (23217)

- May is the month with the highest Bad Change (24420)

- January is the most stable month (6418)



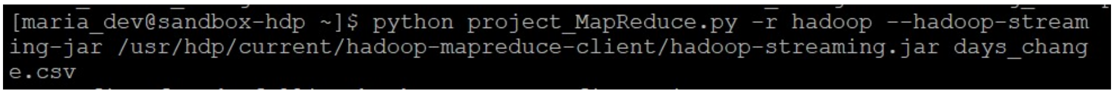
```
"April,Bad Change"      20488
"April,Good Change"     22595
"April,Stable"  5305
"August,Bad Change"     21593
"August,Good Change"    18967
"August,Stable" 5876
"December,Bad Change"   21685
"December,Good Change"  23217
"December,Stable"       5476
"February,Bad Change"   21678
"February,Good Change"  20480
"February,Stable"       5502
"January,Bad Change"    23106
"January,Good Change"   22580
"January,Stable"        6418
"July,Bad Change"       22154
"July,Good Change"      20810
"July,Stable"   5861
"June,Bad Change"       20956
"June,Good Change"      19795
"June,Stable"   5175
"March,Bad Change"      22388
"March,Good Change"     22148
"March,Stable"  5993
```



```
"May,Bad Change"        24420
"May,Good Change"       20785
"May,Stable"    5242
"November,Bad Change"   21178
"November,Good Change"  20496
"November,Stable"       4788
"October,Bad Change"    23524
"October,Good Change"   19920
"October,Stable"        4981
"September,Bad Change"  19928
"September,Good Change" 18842
"September,Stable"      5081
```

3- days:

- ## Running Hadoop Cluster on a days_change data:



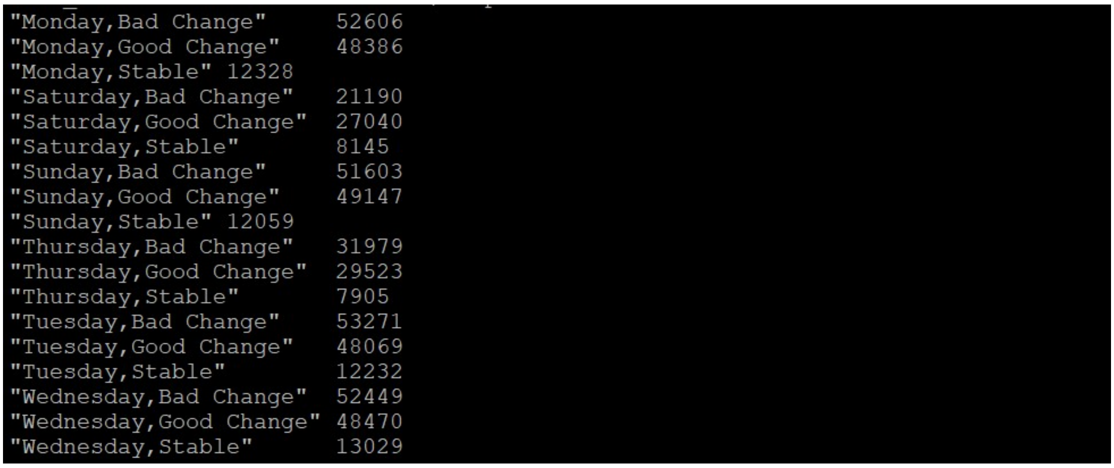
```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar days_chang
e.csv
```

Output:

- Sunday is the day with the highest good change (49147)

- Tuesday is the day with the highest Bad Change (53271)

- Wednesday is the most stable day (13029)



```
"Monday,Bad Change"     52606
"Monday,Good Change"    48386
"Monday,Stable" 12328
"Saturday,Bad Change"   21190
"Saturday,Good Change"  27040
"Saturday,Stable"       8145
"Sunday,Bad Change"     51603
"Sunday,Good Change"    49147
"Sunday,Stable" 12059
"Thursday,Bad Change"   31979
"Thursday,Good Change"  29523
"Thursday,Stable"       7905
"Tuesday,Bad Change"    53271
"Tuesday,Good Change"   48069
"Tuesday,Stable"        12232
"Wednesday,Bad Change"  52449
"Wednesday,Good Change" 48470
"Wednesday,Stable"      13029
```