In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

How does the code perform?

In [2]:
```
pf = pd.read_csv('perf.csv')
pf = pf[['nlines','processed_time']]
```

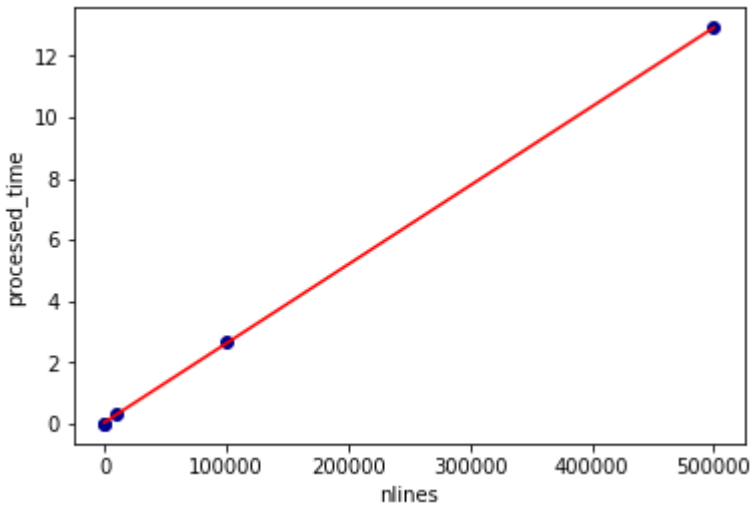
In [3]:
```
lm = np.polyfit(pf.nlines, pf.processed_time, 1)
lm
```

Out[3]:
```
array([  2.57848333e-05,   3.17044523e-02])
```

In [4]:
```
x = np.arange(0,510000,100000)
y = x*lm[0] + lm[1]
```

In [5]:
```
pf.plot.scatter(x='nlines', y='processed_time',  c='DarkBlue', s=35)
plt.plot(x, y, '-', c='Red')
```

Out[5]:
```
[<matplotlib.lines.Line2D at 0xa612da0>]
```



Because the performance scales linearly, one can split a large file into many smaller files, and process them in parallel and distributed way.

What is Harvoni for?

https://www.drugs.com/harvoni.html (https://www.drugs.com/harvoni.html)

In [6]:
```
df = pd.read_csv('top_cost_drug-500k.txt')
```

In [7]:
```
x=df.head(15)[['drug_name','total_cost']].sort_index(ascending=False)
```

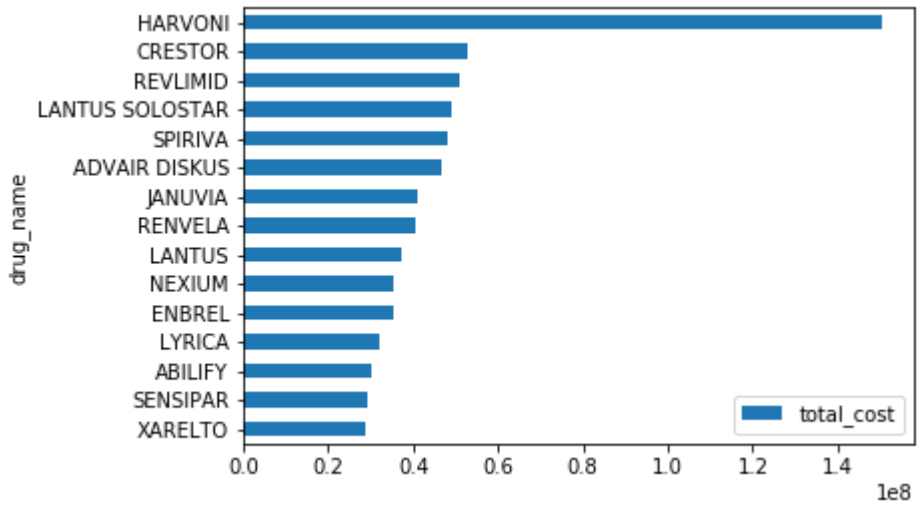
In [8]:
```
x
```

Out[8]:

	drug_name	total_cost
14	XARELTO	2.877651e+07
13	SENSIPAR	2.924247e+07
12	ABILIFY	3.029243e+07
11	LYRICA	3.199746e+07
10	ENBREL	3.524224e+07
9	NEXIUM	3.543758e+07
8	LANTUS	3.731856e+07
7	RENVELA	4.059754e+07
6	JANUVIA	4.090811e+07
5	ADVAIR DISKUS	4.659865e+07
4	SPIRIVA	4.790845e+07
3	LANTUS SOLOSTAR	4.904834e+07
2	REVLIMID	5.111972e+07
1	CRESTOR	5.272927e+07
0	HARVONI	1.503435e+08

In [9]:
```
x.plot.barh(x='drug_name',y='total_cost')
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xaa77780>
```



In [ ]: