In [0]:
```
# Widgets ------------------>>
```

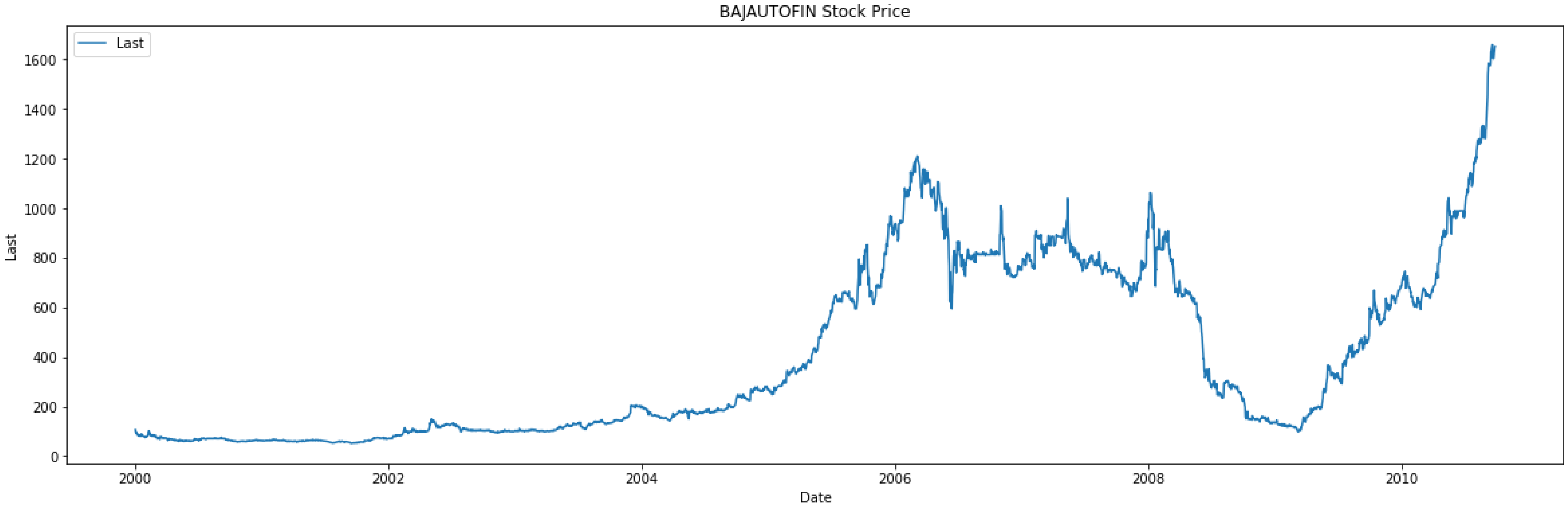
In [0]:
```
dbutils.widgets.dropdown('Symbol','HDFC',['HDFC','TCS','TITAN','BAJAUTOFIN','ASIANPAINT','INFOSYSTCH','HINDLEVER','DRR
from pyspark.sql.window import Window
from pyspark.sql import functions as f
df2 = spark.read.csv('dbfs:/FileStore/PortfolioProject/Stock_price.csv',header = True,inferSchema = True)
T =dbutils.widgets.get('Symbol')
df_p = df2.filter(df2['Symbol']==T).select('Date','Last').toPandas()
df_p
```

	Date	Last
0	2000-01-03	50.75
1	2000-01-04	48.00
2	2000-01-05	44.60
3	2000-01-06	46.00
4	2000-01-07	42.90
...	...	...
2606	2010-09-22	770.00
2607	2010-09-23	771.00
2608	2010-09-24	770.15
2609	2010-09-27	791.05
2610	2010-09-28	793.00

2611 rows × 2 columns

In [0]:

```
import matplotlib.pyplot as plt
if len(df_p) > 1:
    df_p.set_index('Date', inplace=True)  # Set 'Date' as the index
    df_p['Last'] = (df_p['Last'] / df_p['Last'].iloc[1] * 100)
    df_p.plot(y='Last', figsize=(20, 6))
    plt.xlabel('Date')  # Set x-axis label
    plt.ylabel('Last')  # Set y-axis label
    plt.title(f'{T} Stock Price')  # Set plot title
    plt.show()
else:
    print("DataFrame has insufficient data points for plotting.")
```



```
In [0]: returns_p = (df_p['Last']/df_p['Last'].shift(1)) -1
        returns_p
```

```
Out[200]: Date
2000-01-03         NaN
2000-01-04   -0.054187
2000-01-05   -0.070833
2000-01-06    0.031390
2000-01-07   -0.067391
                ...
2010-09-22   -0.003881
2010-09-23    0.001299
2010-09-24   -0.001102
2010-09-27    0.027138
2010-09-28    0.002465
Name: Last, Length: 2611, dtype: float64
```

```
In [0]: annual_return = returns_p.mean()*250*100
        annual_return
```

```
Out[201]: 40.458734847441505
```

```
In [0]: # Partitioning & Bucketing ---------------->>
In [0]: # Load Datasets
        df = spark.read.csv('dbfs:/FileStore/PortfolioProject/Stock_price.csv',header = True,inferSchema = True)
        df.show()
```

```
+----------+----------+------+----------+------+------+-------+-------+-------+-------+--------+--------+------+-----
-------------+-----------+--------------+------------------+
|      Date|    Symbol|Series|Prev Close|  Open|  High|    Low|   Last|  Close|   VWAP|  Volume|Turnover|Trades|Deliv
erable Volume|%Deliverble|        Sector|          Industry|
+----------+----------+------+----------+------+------+-------+-------+-------+-------+--------+--------+------+-----
-------------+-----------+--------------+------------------+
|2007-11-27|MUNDRAPORT|    EQ|     440.0| 770.0|1050.0|  770.0|  959.0|  962.9| 984.72|27294366| 2.69E15|  null|
9859619|     0.3612|Infrastructure|Ports and Shipping|
|2007-11-28|MUNDRAPORT|    EQ|     962.9| 984.0| 990.0|  874.0|  885.0|  893.9| 941.38| 4581338| 4.31E14|  null|
1453278|     0.3172|Infrastructure|Ports and Shipping|
|2007-11-29|MUNDRAPORT|    EQ|     893.9| 909.0|914.75|  841.0|  887.0|  884.2| 888.09| 5124121| 4.55E14|  null|
1069678|     0.2088|Infrastructure|Ports and Shipping|
|2007-11-30|MUNDRAPORT|    EQ|     884.2| 890.0| 958.0|  890.0|  929.0| 921.55| 929.17| 4609762| 4.28E14|  null|
1260913|     0.2735|Infrastructure|Ports and Shipping|
|2007-12-03|MUNDRAPORT|    EQ|    921.55|939.75| 995.0|  922.0|  980.0|  969.3| 965.65| 2977470| 2.88E14|  null|
816123|     0.2741|Infrastructure|Ports and Shipping|
|2007-12-04|MUNDRAPORT|    EQ|     969.3| 985.0|1056.0|  976.0| 1049.0|1041.45|1015.39| 4849250| 4.92E14|  null|
1537667|     0.3171|Infrastructure|Ports and Shipping|
|2007-12-05|MUNDRAPORT|    EQ|   1041.45|1061.0|1099.5| 1050.0| 1084.0|1082.45|1082.79| 2848209| 3.08E14|  null|
904260|     0.3175|Infrastructure|Ports and Shipping|
|2007-12-06|MUNDRAPORT|    EQ|   1082.45|1089.0|1109.7| 1051.0| 1090.1| 1081.3|1087.03| 1749516|  1.9E14|  null|
825691|      0.472|Infrastructure|Ports and Shipping|
|2007-12-07|MUNDRAPORT|    EQ|    1081.3|1100.0|1134.0| 1078.0| 1100.0| 1102.4|1106.57| 2247904| 2.49E14|  null|
697763|     0.3104|Infrastructure|Ports and Shipping|
|2007-12-10|MUNDRAPORT|    EQ|    1102.4|1110.0|1110.0| 1061.1|1073.55| 1075.4|1080.38| 1012350| 1.09E14|  null|
417514|     0.4124|Infrastructure|Ports and Shipping|
|2007-12-11|MUNDRAPORT|    EQ|    1075.4|1081.0|1089.0| 1041.0| 1046.0|1047.65| 1067.8|  810464| 8.65E13|  null|
415191|     0.5123|Infrastructure|Ports and Shipping|
|2007-12-12|MUNDRAPORT|    EQ|   1047.65|1032.0|1065.0| 1016.0| 1036.9| 1036.8|1043.92|  744799| 7.78E13|  null|
363848|     0.4885|Infrastructure|Ports and Shipping|
|2007-12-13|MUNDRAPORT|    EQ|    1036.8|1040.0|1150.0|1030.25|1131.15|1129.95|1109.09| 3067687|  3.4E14|  null|
1040076|      0.339|Infrastructure|Ports and Shipping|
|2007-12-14|MUNDRAPORT|    EQ|   1129.95|1139.9|1140.0| 1101.1| 1107.0| 1110.5|1119.55| 1070737|  1.2E14|  null|
525239|     0.4905|Infrastructure|Ports and Shipping|
|2007-12-17|MUNDRAPORT|    EQ|    1110.5|1140.0|1168.0| 1021.5| 1052.0|1044.25|1102.42| 1404955| 1.55E14|  null|
670298|     0.4771|Infrastructure|Ports and Shipping|
|2007-12-18|MUNDRAPORT|    EQ|   1044.25|1045.0|1109.9|1031.55| 1085.0|1074.95|1077.84| 1226984| 1.32E14|  null|
449420|     0.3663|Infrastructure|Ports and Shipping|
|2007-12-19|MUNDRAPORT|    EQ|   1074.95|1091.0|1116.0| 1046.3| 1078.0| 1066.9|1082.93|  845666| 9.16E13|  null|
344171|      0.407|Infrastructure|Ports and Shipping|
|2007-12-20|MUNDRAPORT|    EQ|    1066.9|1083.5|1083.5| 1051.0| 1067.0| 1060.2|1065.52|  623288| 6.64E13|  null|
276356|     0.4434|Infrastructure|Ports and Shipping|
|2007-12-24|MUNDRAPORT|    EQ|    1060.2|1095.0|1192.0|1085.25| 1160.0| 1156.8|1160.77| 2060892| 2.39E14|  null|
807879|      0.392|Infrastructure|Ports and Shipping|
|2007-12-26|MUNDRAPORT|    EQ|    1156.8|1175.0|1214.0| 1148.0| 1212.0| 1199.9| 1183.3| 1467031| 1.74E14|  null|
469389|       0.32|Infrastructure|Ports and Shipping|
+----------+----------+------+----------+------+------+-------+-------+-------+-------+--------+--------+------+-----
-------------+-----------+--------------+------------------+
only showing top 20 rows
```

In [0]:

```
%fs rm -r/dbfs:/user/hive/warehouse/partbucketstock
```

res5: Boolean = false

In [0]:

```
# Save the DataFrame as a table

df.write.option('header', True).partitionBy('Sector').bucketBy(5, 'Industry').mode('overwrite').saveAsTable('PartBucke

# Read the table back into a DataFrame
df = spark.table('PartBucketStocks2')

# Show the data from the DataFrame
df.show()
```

```
+----------+------+------+----------+------+------+------+------+------+------+------+--------+------+---------------
---+-----------+--------------------+----------+
|      Date|Symbol|Series|Prev Close|  Open|  High|   Low|  Last| Close|  VWAP|Volume|Turnover|Trades|Deliverable Vol
ume|%Deliverble|            Industry|    Sector|
+----------+------+------+----------+------+------+------+------+------+------+------+--------+------+---------------
---+-----------+--------------------+----------+
|2000-01-03|   M&M|    EQ|    419.75| 453.3|453.35| 448.9|453.35|453.35|453.18| 67195| 3.05E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-04|   M&M|    EQ|    453.35| 489.6|489.65| 489.6|489.65|489.65|489.65| 37470| 1.83E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-05|   M&M|    EQ|    489.65|528.85|528.85|451.15| 519.0|514.85|521.37|227621| 1.19E13|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-06|   M&M|    EQ|    514.85| 528.0| 550.0| 518.0| 521.0|524.55|538.27|198870| 1.07E13|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-07|   M&M|    EQ|    524.55| 515.0| 522.0| 490.0| 498.9| 496.4|508.09| 91052| 4.63E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-10|   M&M|    EQ|     496.4| 509.9| 535.0| 491.7| 495.0| 497.2|509.86| 83454| 4.26E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-11|   M&M|    EQ|     497.2| 514.0| 537.0| 510.1| 537.0| 532.8|527.14|250382| 1.32E13|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-12|   M&M|    EQ|     532.8| 540.0| 550.0| 490.2| 490.2|490.25|515.33|136009| 7.01E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-13|   M&M|    EQ|    490.25| 516.0| 520.0| 465.5| 502.5| 499.0|500.01| 85954|  4.3E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-14|   M&M|    EQ|     499.0| 500.0| 525.0| 490.1| 519.0| 519.0|510.06| 79448| 4.05E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-17|   M&M|    EQ|     519.0| 538.0|560.55| 538.0|560.55|560.55|554.58|163004| 9.04E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-18|   M&M|    EQ|    560.55| 570.0|595.95|550.25| 590.0|573.95|567.55|293023| 1.66E13|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-19|   M&M|    EQ|    573.95| 601.0| 619.9| 572.0| 590.0| 599.2|604.94|327481| 1.98E13|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-20|   M&M|    EQ|     599.2|597.95|597.95| 566.0|575.75|572.35|579.85|107183| 6.22E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-21|   M&M|    EQ|    572.35| 585.0| 585.0| 544.0| 550.0| 550.0|555.65| 91545| 5.09E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-24|   M&M|    EQ|     550.0| 539.1| 549.9|511.25| 516.0| 514.7|529.17| 45765| 2.42E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-25|   M&M|    EQ|     514.7| 516.0|519.75| 491.2| 510.0|507.25|505.87| 91169| 4.61E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-27|   M&M|    EQ|    507.25| 540.0| 540.0| 475.0| 502.0| 499.3|505.17| 92630| 4.68E12|  null|              n
```

```
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-28|   M&M|    EQ|     499.3| 508.0| 519.9| 500.0| 511.0| 511.1|507.13| 56869| 2.88E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
|2000-01-31|   M&M|    EQ|     511.1| 507.0| 517.5| 500.0| 506.0|507.75|509.54| 39681| 2.02E12|  null|              n
ull|       null|Automobiles - Pas...|Automotive|
+----------+------+------+----------+------+------+------+------+------+------+------+--------+------+---------------
---+-----------+--------------------+----------+
only showing top 20 rows
```

In [0]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [0]:

```python
df.printSchema()
```

```
root
 |-- Date: date (nullable = true)
 |-- Symbol: string (nullable = true)
 |-- Series: string (nullable = true)
 |-- Prev Close: double (nullable = true)
 |-- Open: double (nullable = true)
 |-- High: double (nullable = true)
 |-- Low: double (nullable = true)
 |-- Last: double (nullable = true)
 |-- Close: double (nullable = true)
 |-- VWAP: double (nullable = true)
 |-- Volume: integer (nullable = true)
 |-- Turnover: double (nullable = true)
 |-- Trades: integer (nullable = true)
 |-- Deliverable Volume: integer (nullable = true)
 |-- %Deliverble: string (nullable = true)
 |-- Industry: string (nullable = true)
 |-- Sector: string (nullable = true)
```

In [0]:
```
df.describe().toPandas()
```

	summary	Symbol	Series	Prev Close	Open	High	Low	Last	
0	count	235192	235192	235192	235192	235192	235192	235192	
1	mean	None	None	1266.196348727844	1267.7597082383663	1286.5814404826638	1247.4884653814634	1266.388301898019	126
2	stddev	None	None	2581.3703203038617	2585.259609461142	2619.6492164496135	2546.621395806379	2581.3925428034377	2582.
3	min	ADANIPORTS	EQ	0.0	8.5	9.75	8.5	9.1	
4	max	ZEETELE	EQ	32861.95	33399.95	33480.0	32468.1	32849.0	

In [0]:
```
df.count()
```

```
Out[172]: 235192
```

In [0]:
```
df.select(['Open','High','Low','Close']).describe().show()
```

```
+-------+------------------+------------------+------------------+-----------------+
|summary|              Open|              High|               Low|            Close|
+-------+------------------+------------------+------------------+-----------------+
|  count|            235192|            235192|            235192|           235192|
|   mean|1267.7597082383663|1286.5814404826638|1247.4884653814634|  1266.5543507007|
| stddev| 2585.259609461142|2619.6492164496135| 2546.621395806379|2582.140941701451|
|    min|               8.5|              9.75|               8.5|             9.15|
|    max|          33399.95|           33480.0|           32468.1|         32861.95|
+-------+------------------+------------------+------------------+-----------------+
```

In [0]:

```
df.groupBy('Sector').count().show()
```

```
+--------------------+-----+
|              Sector|count|
+--------------------+-----+
|          Healthcare|15918|
|             Finance|36500|
|           Utilities| 7447|
|          Technology|23686|
|     Basic Materials|28018|
|  Financial Services| 8436|
|      Consumer Goods|44369|
|       Manufacturing| 8508|
|              Energy|23501|
|          Automotive| 5306|
|                null| 5305|
|Consumer Goods,Ag...| 5306|
|Energy,Technology...| 5306|
|  Telecommunications| 4774|
|Industrial,Techno...| 4184|
|      Infrastructure| 3322|
|   Consumer Services| 5306|
+--------------------+-----+
In [0]: by_sector = df.select(['Sector','Open','Close']).groupBy('Sector').mean().collect()
        by_sector
```

```
Out[175]: [Row(Sector='Healthcare', avg(Open)=1039.8831542907383, avg(Close)=1038.703615403944),
 Row(Sector='Finance', avg(Open)=802.0072328767147, avg(Close)=801.4350164383554),
 Row(Sector='Utilities', avg(Open)=146.8205451859812, avg(Close)=146.65783536994792),
 Row(Sector='Technology', avg(Open)=1324.4393945790794, avg(Close)=1322.7301486109961),
 Row(Sector='Basic Materials', avg(Open)=1797.9449675208787, avg(Close)=1796.8254015275904),
 Row(Sector='Financial Services', avg(Open)=1925.701446183027, avg(Close)=1925.016945234706),
 Row(Sector='Consumer Goods', avg(Open)=2339.9223252721517, avg(Close)=2337.4382936284487),
 Row(Sector='Manufacturing', avg(Open)=1918.4596203573112, avg(Close)=1917.7150740479542),
 Row(Sector='Energy', avg(Open)=377.5298668141771, avg(Close)=376.9443023701121),
 Row(Sector='Automotive', avg(Open)=687.5602525442888, avg(Close)=686.8726159065222),
 Row(Sector=None, avg(Open)=404.3018850141368, avg(Close)=403.6010367577756),
 Row(Sector='Consumer Goods,Agri-Business,Hotels,Paperboards and Packaging,Information Technology', avg(Open)=420.631
51149641936, avg(Close)=420.2736901620801),
 Row(Sector='Energy,Technology,Retail,Petrochemicals,Textiles', avg(Open)=1012.602374670186, avg(Close)=1011.31683942
70622),
 Row(Sector='Telecommunications', avg(Open)=380.47845622119655, avg(Close)=379.8007645580231),
 Row(Sector='Industrial,Technology,Financial', avg(Open)=1536.5592853728526, avg(Close)=1534.2743546845147),
 Row(Sector='Infrastructure', avg(Open)=344.7630192655023, avg(Close)=344.20162552679227),
 Row(Sector='Consumer Services', avg(Open)=273.9747455710511, avg(Close)=273.2335657745941)]
```

```
In [0]: for row in by_sector:
            print(list(row),end ='\n')
```

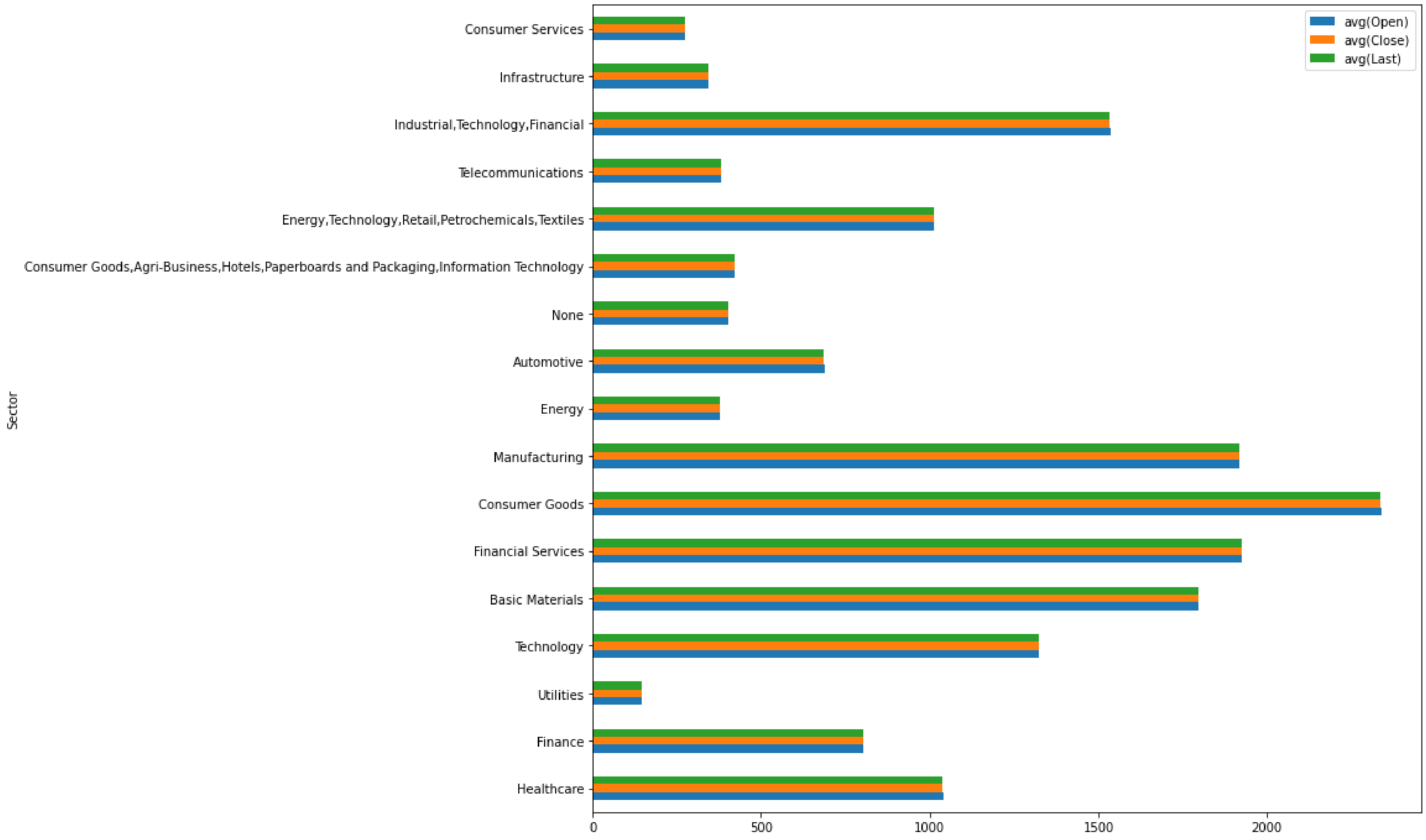
```
['Healthcare', 1039.8831542907383, 1038.703615403944]
['Finance', 802.0072328767147, 801.4350164383554]
['Utilities', 146.8205451859812, 146.65783536994792]
['Technology', 1324.4393945790794, 1322.7301486109961]
['Basic Materials', 1797.9449675208787, 1796.8254015275904]
['Financial Services', 1925.701446183027, 1925.016945234706]
['Consumer Goods', 2339.9223252721517, 2337.4382936284487]
['Manufacturing', 1918.4596203573112, 1917.7150740479542]
['Energy', 377.5298668141771, 376.9443023701121]
['Automotive', 687.5602525442888, 686.8726159065222]
[None, 404.3018850141368, 403.6010367577756]
['Consumer Goods,Agri-Business,Hotels,Paperboards and Packaging,Information Technology', 420.63151149641936, 420.2736
901620801]
['Energy,Technology,Retail,Petrochemicals,Textiles', 1012.602374670186, 1011.3168394270622]
['Telecommunications', 380.47845622119655, 379.8007645580231]
['Industrial,Technology,Financial', 1536.5592853728526, 1534.2743546845147]
['Infrastructure', 344.7630192655023, 344.20162552679227]
['Consumer Services', 273.9747455710511, 273.2335657745941]
```

```
In [0]: sector_df = df.select(['Sector','Open','Close','Last']).groupBy('Sector').mean().toPandas()
        sector_df
```

	Sector	avg(Open)	avg(Close)	avg(Last)
0	Healthcare	1039.883154	1038.703615	1038.657529
1	Finance	802.007233	801.435016	801.391548
2	Utilities	146.820545	146.657835	146.648852
3	Technology	1324.439395	1322.730149	1322.680803
4	Basic Materials	1797.944968	1796.825402	1796.265902
5	Financial Services	1925.701446	1925.016945	1925.077365
6	Consumer Goods	2339.922325	2337.438294	2337.069374
7	Manufacturing	1918.459620	1917.715074	1917.506653
8	Energy	377.529867	376.944302	376.946030
9	Automotive	687.560253	686.872616	686.834640
10	None	404.301885	403.601037	403.514722
11	Consumer Goods,Agri-Business,Hotels,Paperboard...	420.631511	420.273690	420.250207
12	Energy,Technology,Retail,Petrochemicals,Textiles	1012.602375	1011.316839	1011.157143
13	Telecommunications	380.478456	379.800765	379.798502
14	Industrial,Technology,Financial	1536.559285	1534.274355	1534.166551
15	Infrastructure	344.763019	344.201626	344.239539
16	Consumer Services	273.974746	273.233566	273.184075

```
In [0]: sector_df.plot(kind='barh', x='Sector', y=sector_df.columns.tolist()[1:],
                       figsize=(12,12))
```

```
Out[178]: <AxesSubplot:ylabel='Sector'>
```



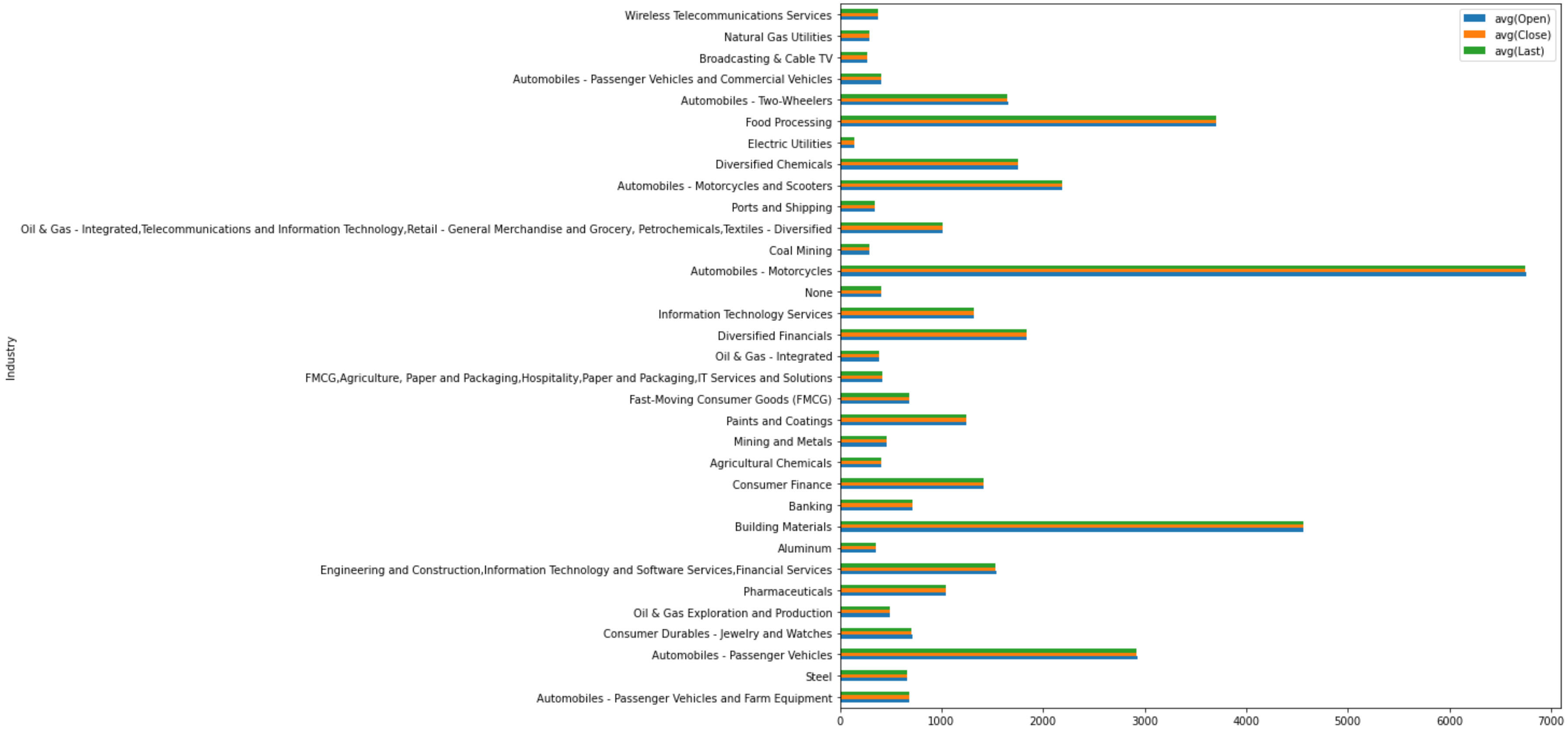
In [0]:

```
industry_df = df.select(['Industry','Open','Close','Last'])\
                .groupBy('Industry').mean().toPandas()
```

In [0]:

```
industry_df.plot(kind='barh', x='Industry', y=sector_df.columns.tolist()[1:],
               figsize=(12,12))
```

```
Out[180]: <AxesSubplot:ylabel='Industry'>
```



In [0]:

```
import pyspark.sql.functions as f
```

In [0]:

```
health = df.filter(f.col('Sector')=='Healthcare')
health.show()
```

```
+----------+------+------+----------+------+-------+-------+-------+-------+-------+------+--------+------+----------
--------+-----------+---------------+----------+
|      Date|Symbol|Series|Prev Close|  Open|   High|    Low|   Last|  Close|   VWAP|Volume|Turnover|Trades|Deliverabl
e Volume|%Deliverble|       Industry|    Sector|
+----------+------+------+----------+------+-------+-------+-------+-------+-------+------+--------+------+----------
--------+-----------+---------------+----------+
|2000-01-03| CIPLA|    EQ|    1349.4|1410.0|1457.35|1380.05|1457.35|1457.35|1441.36| 21060| 3.04E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-04| CIPLA|    EQ|   1457.35|1537.0| 1537.0| 1430.0|1466.05|1465.25|1460.43| 30215| 4.41E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-05| CIPLA|    EQ|   1465.25|1474.0| 1474.0| 1365.0| 1441.0|1435.05|1428.11| 33799| 4.83E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-06| CIPLA|    EQ|   1435.05|1434.0| 1435.0| 1349.0| 1365.0|1355.85|1390.55| 33083|  4.6E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-07| CIPLA|    EQ|   1355.85|1370.0| 1389.9| 1247.4| 1247.4|1247.55|1267.49| 66536| 8.43E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-10| CIPLA|    EQ|   1247.55|1288.0| 1299.0| 1191.0|1197.15| 1205.9|1222.23|105912| 1.29E13|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-11| CIPLA|    EQ|    1205.9|1225.0| 1225.0|1109.45| 1125.0|1114.25|1156.31|186975| 2.16E13|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-12| CIPLA|    EQ|   1114.25|1185.0| 1203.4| 1185.0| 1203.4| 1203.4|1202.76|  7416| 8.92E11|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-13| CIPLA|    EQ|    1203.4|1299.7| 1299.7| 1281.2| 1299.7|1297.05|1298.53| 90379| 1.17E13|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-14| CIPLA|    EQ|   1297.05|1299.0|1304.55| 1220.0| 1275.0| 1280.7|1275.38| 70729| 9.02E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-17| CIPLA|    EQ|    1280.7|1335.0| 1340.0|1250.15| 1265.0|1270.05|1292.22| 54938|  7.1E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-18| CIPLA|    EQ|   1270.05|1294.0| 1294.0| 1200.0| 1235.0|1220.15|1227.43| 51691| 6.34E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-19| CIPLA|    EQ|   1220.15|1175.0| 1219.9| 1132.0| 1200.0|1203.85|1189.27|132669| 1.58E13|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-20| CIPLA|    EQ|   1203.85|1205.0| 1223.0| 1201.0| 1208.0| 1208.8|1212.22| 44602| 5.41E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-21| CIPLA|    EQ|    1208.8|1210.0| 1210.0| 1160.0| 1202.0| 1201.1|1198.65| 43168| 5.17E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-24| CIPLA|    EQ|    1201.1|1218.0| 1223.9| 1185.0| 1212.0| 1212.0|1210.61| 67930| 8.22E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-25| CIPLA|    EQ|    1212.0|1195.0| 1208.0|1176.05| 1197.9| 1194.3|1195.43| 65851| 7.87E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-27| CIPLA|    EQ|    1194.3|1225.0| 1225.0| 1185.0| 1195.0| 1190.3|1197.07| 33549| 4.02E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-28| CIPLA|    EQ|    1190.3|1210.0| 1215.0| 1171.0| 1183.0|1183.75|1188.57| 25834| 3.07E12|  null|
null|       null|Pharmaceuticals|Healthcare|
|2000-01-31| CIPLA|    EQ|   1183.75|1162.0| 1184.9| 1155.0|1169.35| 1173.1|1172.11| 30473| 3.57E12|  null|
null|       null|Pharmaceuticals|Healthcare|
+----------+------+------+----------+------+-------+-------+-------+-------+-------+------+--------+------+----------
--------+-----------+---------------+----------+
only showing top 20 rows
```

In [0]:

```
# Start/End Date, min, max, avg
from pyspark.sql.functions import col, min, max, avg
```

In [0]:

```
df.groupBy('Sector')\
    .agg(
        min('Date').alias('Start'),
        max('Date').alias('End'),

        min('open').alias('Minimum Opening'),
        max('open').alias('Maximum Opening'),
        avg('open').alias('Average Opening'),

        min('Close').alias('Minimum Closing'),
        max('Close').alias('Maximum Closing'),
        avg('Close').alias('Average Closing'),
    ).show(truncate=True)
```

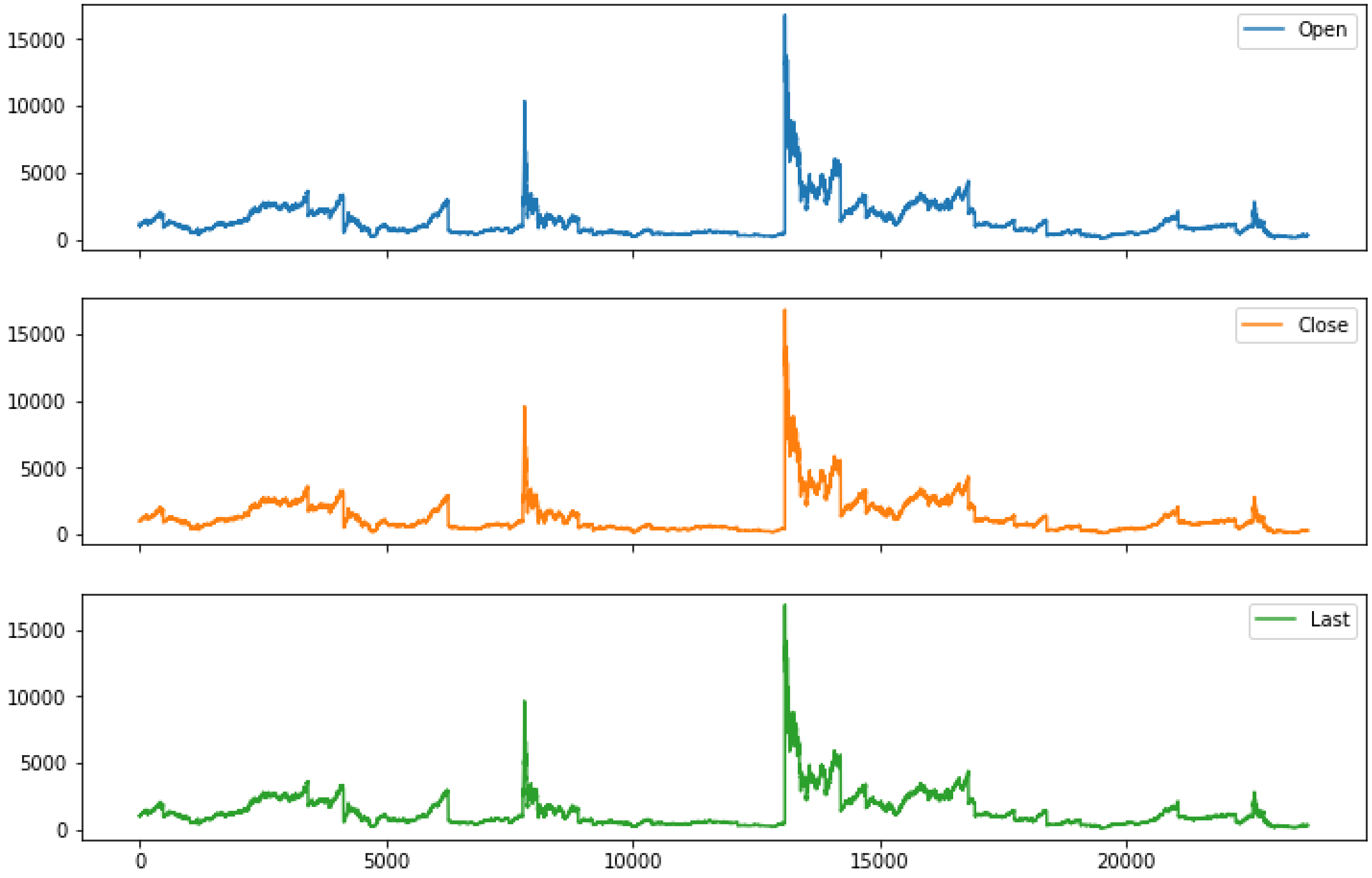
```
+--------------------+----------+----------+---------------+---------------+------------------+---------------+------
---------+------------------+
|              Sector|     Start|       End|Minimum Opening|Maximum Opening|   Average Opening|Minimum Closing|Maximu
m Closing|   Average Closing|
+--------------------+----------+----------+---------------+---------------+------------------+---------------+------
---------+------------------+
|          Healthcare|2000-01-03|2021-04-30|         150.55|        5440.95|1039.8831542907383|          160.1|
5416.8| 1038.703615403944|
|             Finance|2000-01-03|2021-04-30|            8.5|         3505.0| 802.0072328767147|           9.15|
3489.95| 801.4350164383554|
|           Utilities|2004-11-05|2021-04-30|           61.7|          289.0| 146.8205451859812|           58.0|
284.65|146.65783536994792|
|          Technology|2000-01-03|2021-04-30|           87.1|        16800.0|1324.4393945790794|           89.7|
16855.9|1322.7301486109961|
|     Basic Materials|2000-01-03|2021-04-30|          24.75|        31682.4|1797.9449675208787|           24.1|
31748.75|1796.8254015275904|
|  Financial Services|2000-01-03|2021-04-30|           25.2|        11300.0| 1925.701446183027|           24.5|
11393.3| 1925.016945234706|
|      Consumer Goods|2000-01-03|2021-04-30|          17.05|       33399.95|2339.9223252721517|           17.7|
32861.95|2337.4382936284487|
|       Manufacturing|2000-01-03|2021-04-30|         170.25|         5286.0|1918.4596203573112|          172.5|
5286.1|1917.7150740479542|
|              Energy|2000-01-03|2021-04-30|           44.8|        1480.55| 377.5298668141771|           43.5|
1484.2| 376.9443023701121|
|          Automotive|2000-01-03|2021-04-30|           52.0|         1560.0| 687.5602525442888|           51.8|
1556.3| 686.8726159065222|
|                null|2000-01-04|2021-04-30|           66.0|         1024.0| 404.3018850141368|          67.25|
1034.0| 403.6010367577756|
|Consumer Goods,Ag...|2000-01-03|2021-04-30|          115.0|         1946.0|420.63151149641936|         115.45|
1940.1| 420.2736901620801|
|Energy,Technology...|2000-01-03|2021-04-30|          205.5|         3298.0| 1012.602374670186|          203.2|
3220.85|1011.3168394270622|
|  Telecommunications|2002-02-18|2021-04-30|           21.1|         1133.9|380.47845622119655|          20.75|
1125.65| 379.8007645580231|
|Industrial,Techno...|2004-06-23|2021-04-30|          500.0|         4510.0|1536.5592853728526|         562.05|
4506.7|1534.2743546845147|
|      Infrastructure|2007-11-27|2021-04-30|          108.0|        1310.25| 344.7630192655023|          108.0|
1307.45|344.20162552679227|
|   Consumer Services|2000-01-03|2021-04-30|           62.0|         1640.0| 273.9747455710511|           62.3|
1541.7| 273.2335657745941|
+--------------------+----------+----------+---------------+---------------+------------------+---------------+------
```

```
---------+------------------+
```

In [0]:

```
# Time Series
tech = df.where(col('Sector')=='Technology').select('Date','Open','Close','Last')
tech.toPandas().plot(subplots=True,figsize=(12,8))
```

```
Out[206]: array([<AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



```
In [0]: df.filter(df['Last'].between(100,500)).show(5)
```

```
+----------+------+------+----------+------+-----+------+------+------+------+-------+--------+------+---------------
---+-----------+--------------------+----------+
|      Date|Symbol|Series|Prev Close|  Open| High|   Low|  Last| Close|  VWAP| Volume|Turnover|Trades|Deliverable Vol
ume|%Deliverble|            Industry|    Sector|
+----------+------+------+----------+------+-----+------+------+------+------+-------+--------+------+---------------
---+-----------+--------------------+----------+
|2008-10-16|   TCS|    EQ|     543.1| 525.0|528.7| 467.9| 493.0| 495.0|494.57|2007950| 9.93E13|  null|           1017
019|     0.5065|Information Techn...|Technology|
|2008-10-17|   TCS|    EQ|     495.0| 500.0|529.0| 445.0| 445.8|453.85|489.66|2435885| 1.19E14|  null|           1395
432|     0.5729|Information Techn...|Technology|
|2008-10-20|   TCS|    EQ|    453.85|496.65|505.0|455.25| 500.0|491.35|482.28|3103265|  1.5E14|  null|           1962
918|     0.6325|Information Techn...|Technology|
|2008-10-24|   TCS|    EQ|     547.3| 503.7|536.4| 440.0| 495.0|498.85|515.83|2119984| 1.09E14|  null|            851
277|     0.4015|Information Techn...|Technology|
|2008-11-06|   TCS|    EQ|     506.4| 528.5|528.5| 482.1|499.85| 500.2|505.05|1965479| 9.93E13|  null|            584
162|     0.2972|Information Techn...|Technology|
+----------+------+------+----------+------+-----+------+------+------+------+-------+--------+------+---------------
---+-----------+--------------------+----------+
only showing top 5 rows
```

In [0]:

```
# Using lit is useful when you want to perform operations that involve constant values or when you need to compare Dat
from pyspark.sql.functions import lit
df.filter((col('Date')>=lit('2020-01-01')) & (col('Date') <= lit('2020-01-31'))).show(5)
```

```
+----------+------+------+----------+-------+-------+------+-------+-------+-------+-------+--------+------+---------
---------+-----------+--------------------+----------+
|      Date|Symbol|Series|Prev Close|   Open|   High|   Low|   Last|  Close|   VWAP| Volume|Turnover|Trades|Deliverab
le Volume|%Deliverble|            Industry|    Sector|
+----------+------+------+----------+-------+-------+------+-------+-------+-------+-------+--------+------+---------
---------+-----------+--------------------+----------+
|2020-01-01|   TCS|    EQ|    2161.7| 2168.0| 2183.9|2154.0| 2170.0| 2167.6|2170.54|1354908| 2.94E14| 44438|
164490|     0.1214|Information Techn...|Technology|
|2020-01-02|   TCS|    EQ|    2167.6|2179.95|2179.95|2149.2| 2157.0|2157.65|2158.63|2380752| 5.14E14| 99242|
1204079|     0.5058|Information Techn...|Technology|
|2020-01-03|   TCS|    EQ|   2157.65| 2164.0| 2223.0|2164.0| 2201.0|2200.65|2199.26|4655761| 1.02E15|123516|
1833823|     0.3939|Information Techn...|Technology|
|2020-01-06|   TCS|    EQ|   2200.65| 2205.0|2225.95|2187.9|2201.35|2200.45|2204.89|3023209| 6.67E14|135360|
1000021|     0.3308|Information Techn...|Technology|
|2020-01-07|   TCS|    EQ|   2200.45| 2200.5|2214.65|2183.8| 2205.0|2205.85|2203.53|2429317| 5.35E14| 95018|
966753|      0.398|Information Techn...|Technology|
+----------+------+------+----------+-------+-------+------+-------+-------+-------+-------+--------+------+---------
---------+-----------+--------------------+----------+
only showing top 5 rows
```

In [0]:

```
df.select('Open','Close',f.when(df['Last'] >= 200, 1).otherwise(0).alias('Strategy')).show()
```

```
+------+------+--------+
|  Open| Close|Strategy|
+------+------+--------+
| 453.3|453.35|       1|
| 489.6|489.65|       1|
|528.85|514.85|       1|
| 528.0|524.55|       1|
| 515.0| 496.4|       1|
| 509.9| 497.2|       1|
| 514.0| 532.8|       1|
| 540.0|490.25|       1|
| 516.0| 499.0|       1|
| 500.0| 519.0|       1|
| 538.0|560.55|       1|
| 570.0|573.95|       1|
| 601.0| 599.2|       1|
|597.95|572.35|       1|
| 585.0| 550.0|       1|
| 539.1| 514.7|       1|
| 516.0|507.25|       1|
| 540.0| 499.3|       1|
| 508.0| 511.1|       1|
| 507.0|507.75|       1|
+------+------+--------+
only showing top 20 rows
```

In [0]:

```
df.select('Sector',df['Sector'].rlike('^[B,C]').alias('Sector Starts with B or C')).distinct().show()
```

```
+--------------------+-------------------------+
|              Sector|Sector Starts with B or C|
+--------------------+-------------------------+
|           Utilities|                    false|
|          Healthcare|                    false|
|             Finance|                    false|
|     Basic Materials|                     true|
|          Technology|                    false|
|  Financial Services|                    false|
|       Manufacturing|                    false|
|      Consumer Goods|                     true|
|          Automotive|                    false|
|              Energy|                    false|
|                null|                     null|
|Consumer Goods,Ag...|                     true|
|  Telecommunications|                    false|
|Energy,Technology...|                    false|
|Industrial,Techno...|                    false|
|      Infrastructure|                    false|
|   Consumer Services|                     true|
+--------------------+-------------------------+
```