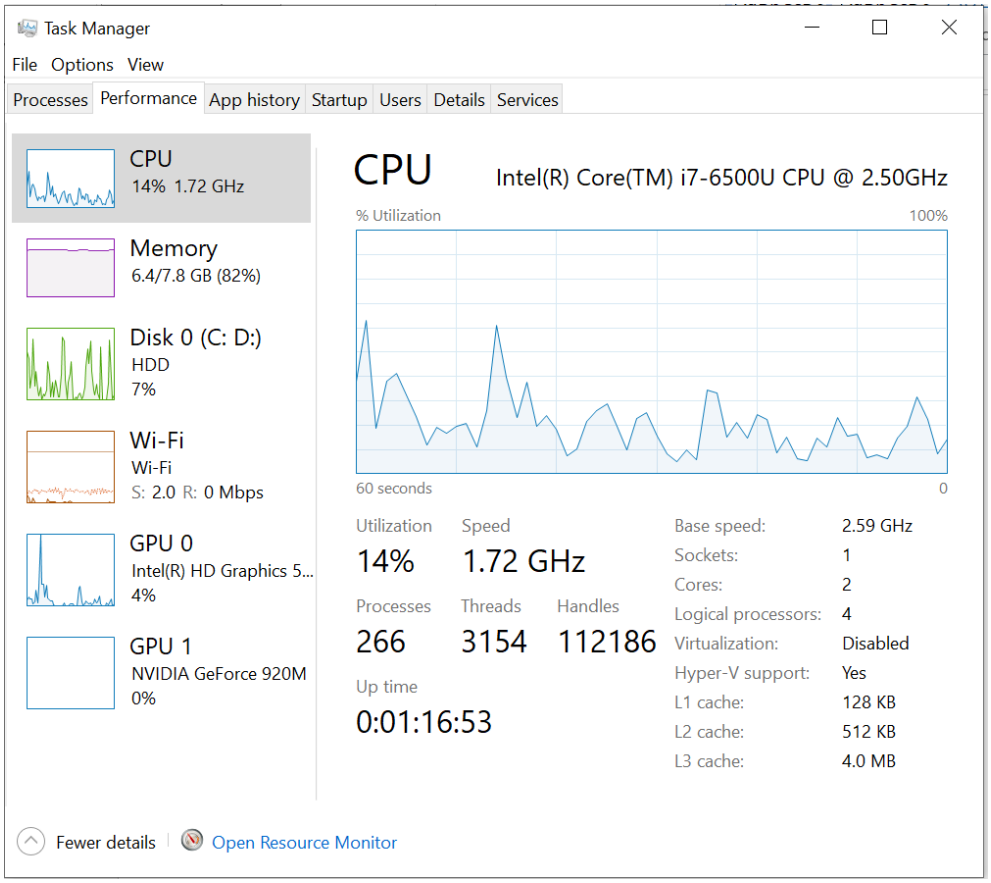
SDE Assignment 3

A.Jyothi

M22CS007

1. Youtube link: https://youtu.be/pAaqGDj9G5g

2. Technical Specifications of your system/laptop, Including L1 and L2 cache size.



RAM: 8 GB

ROM: 1 TB

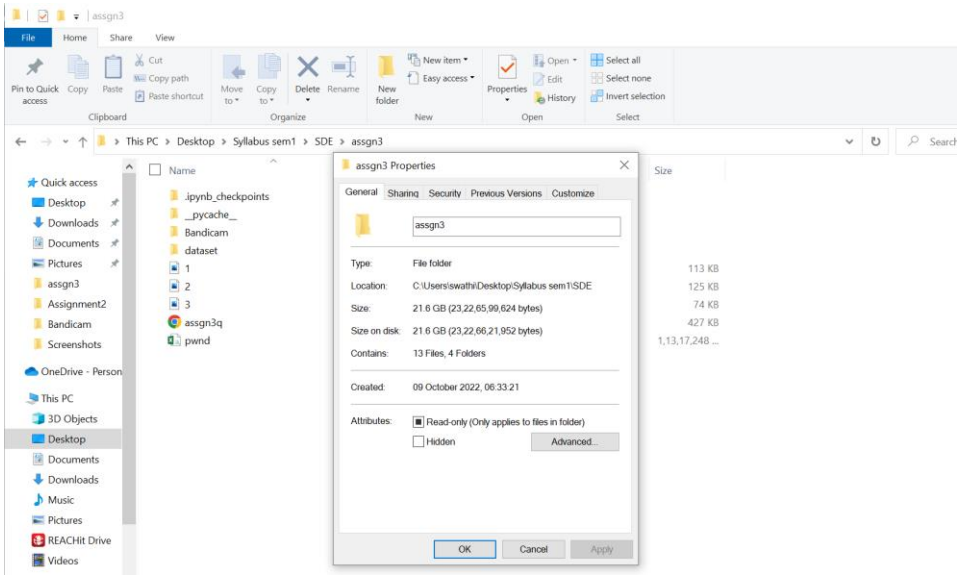
L1 Cache: 128KB

L2 Cache: 512KB

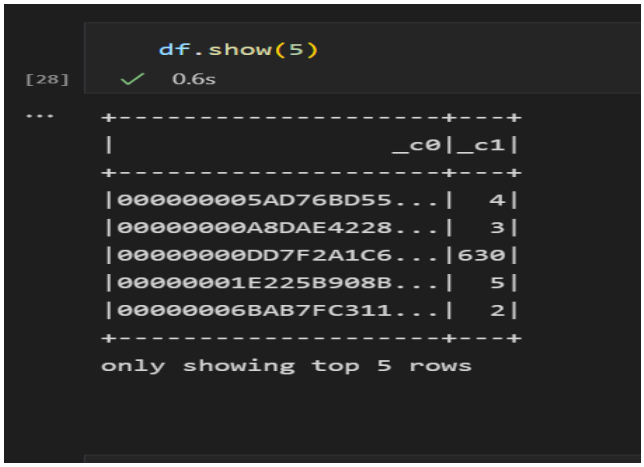
L3 Cache: 4.0 MB

3. Mention metadata of the benchmarking data in tabular and graphical format.

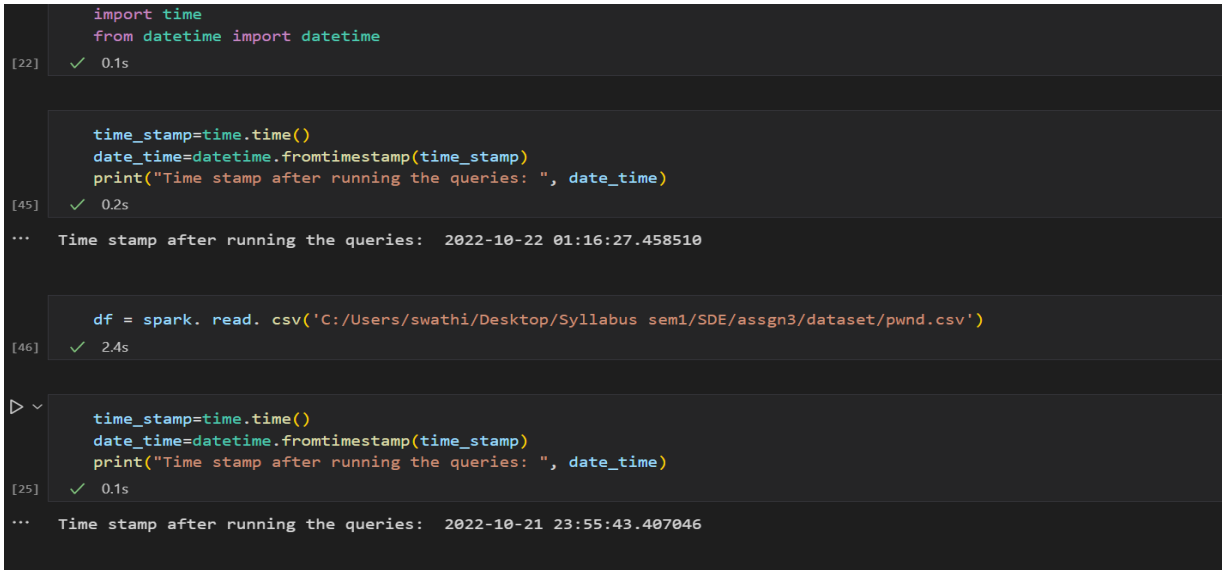
Dataset Size:

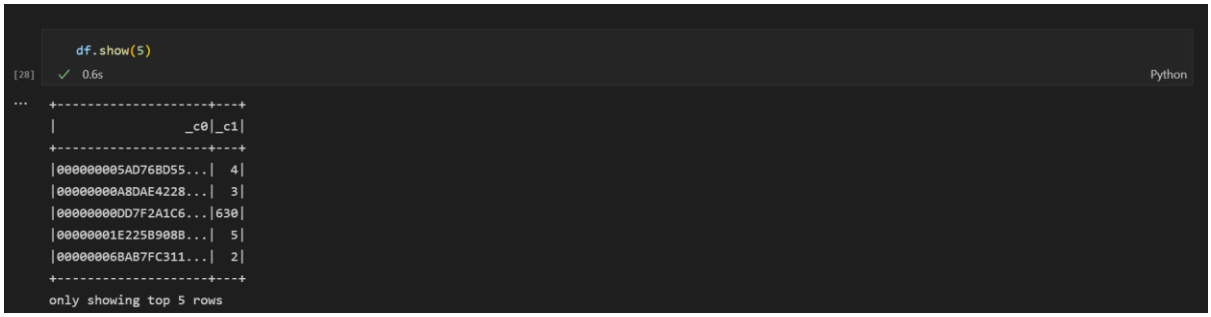


10.7GB



Operations	Pandas	Pyspark
Execution time for uploading the data and reading it	Ran for 30 minutes and the process got killed, as shown in the video.	2.4 seconds
Execution time for count operation	Could not read the dataset. Got error.	3m 57 seconds
Execution time for filter operations	Could not read the dataset. Got error.	1.4 seconds
Execution time for Groupby operation	Could not read the data. Got error.	7minutes





```
df.count()
```

✓ 3m 57.2s — Python

```
262974241
```

```
df.filter(df._c1 == "4").show(truncate=False)
```

[34] ✓ 1.4s — Python

```
+----------------------------------------+---+
|_c0                                     |_c1|
+----------------------------------------+---+
|000000005AD76BD555C1D6D771DE417A4B87E4B4|4  |
|00000008CD1806EB7B9B46A8F87690B2AC16F617|4  |
|00000016C6C075173C163757BCEA8139D4CC69CF|4  |
|00000060F035EF7BA6F001322EDEB8D4ACA2E4EA|4  |
|0000006C4904972F8806EE91A3F999F3AAF54833|4  |
|0000006FCED58039CEF7D334324DC9850CBF36C6|4  |
|00000080C2C27E1E062F010E67AC2A9AF05E722B|4  |
|0000009528888365C2871F32BEB84FC83B220301|4  |
|000000A9EE46773928B13458F482294FABE7DA51|4  |
|000000C37922FE4AC8F30FE4B1C990B9793796E9|4  |
|000000E1B0AE2470DEF64ECCCF85BA3F839F06F3|4  |
```

```
time_stamp=time.time()
date_time=datetime.fromtimestamp(time_stamp)
print("Time stamp after running the queries: ", date_time)
```

[41] ✓ 0.9s — Python

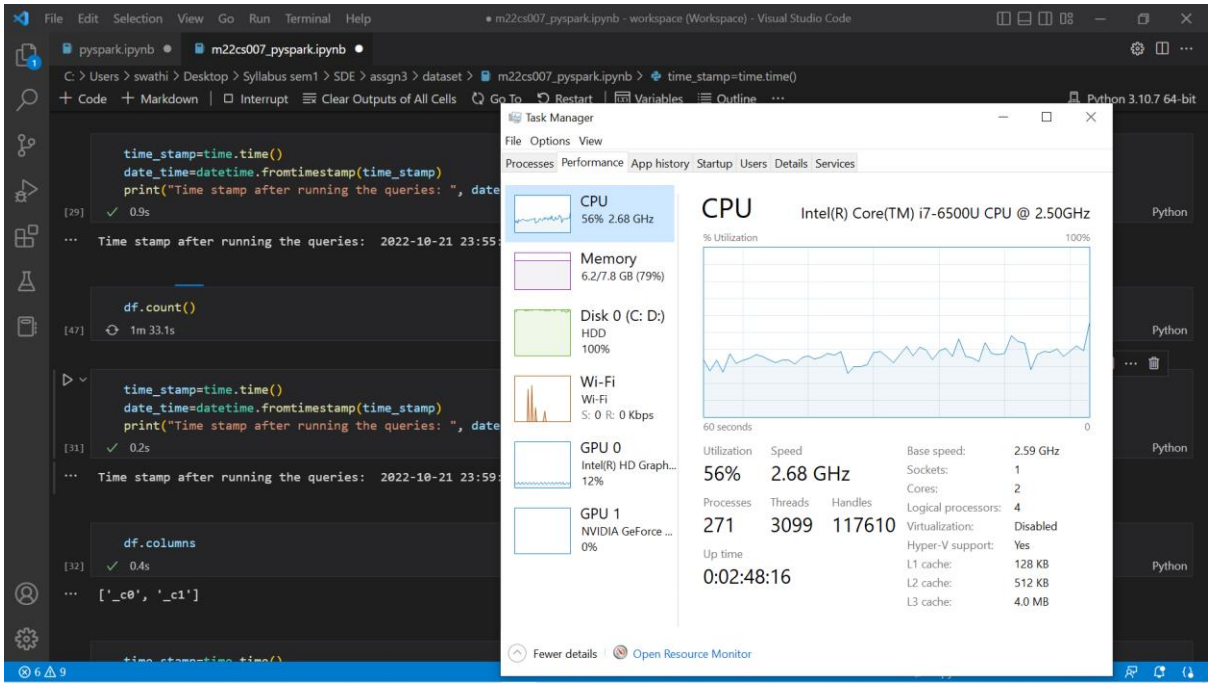
```
Time stamp after running the queries:  2022-10-21 23:59:47.750527
```

```
df.groupby("_c1").count().show()
```

[42] ✓ 7m 28.7s — Python

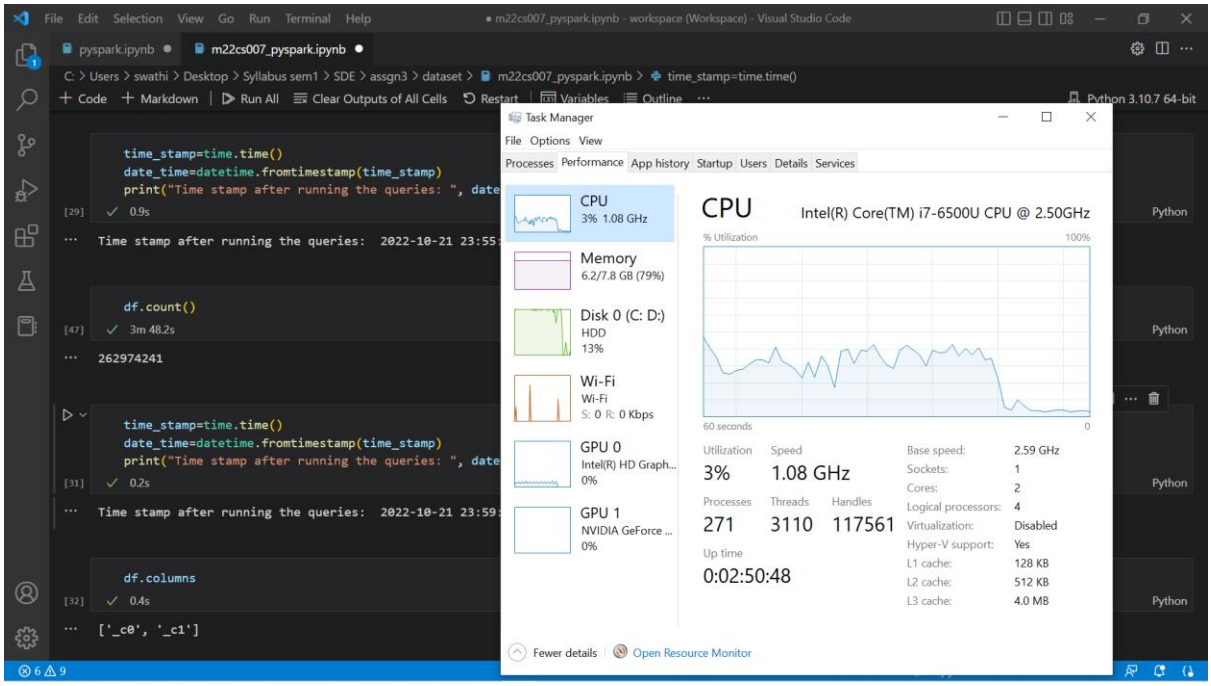
```
+----+-----+
| _c1|count|
+----+-----+
| 296| 1081|
| 467|  462|
| 675|  234|
| 691|  214|
| 829|  150|
|1436|   42|
|1090|   64|
|1159|   60|
|3414|   11|
|2088|   25|
|2162|   17|
```

Operations	Pandas	PySpark
CPU utilization	Out of memory	56% while running a function (count) 3% after finishing the function.
Memory	Out of memory	79% while running
Disk Utilization	Out of memory	100% while running 13% after finishing
Speed	Out of memory	2.68 GHz while running 1.08 GHz after finishing
Threads	Out of memory	3099 while running 3110 after finishing a function
Handles	Out of memory	117610



While running the function,

After finishing the task.



Task(C): Here, For pandas it might have taken 100% memory utilization, and so it has shown out of memory alert. For pyspark the memory utilization is 79% while running the count(). And the disk is 100% while running, but 13% after finishing the task. So, pySpark is efficient in handling larger datasets.

Task (b): After analysing and performing the operations on the dataset, Clearly pyspark has the upper hand. In my case, the dataset is not uploading in pandas environment, it is showing out of memory. So, Pyspark is far better than pandas. And its execution time to upload the dataset of 10GB is few seconds. So, Pyspark is better than Pandas in handling larger datasets.