Data Ingestion and Schema Validation

Name: Robin Masawi

Batch Code: LISP01

Submission Date: 28 May 2021

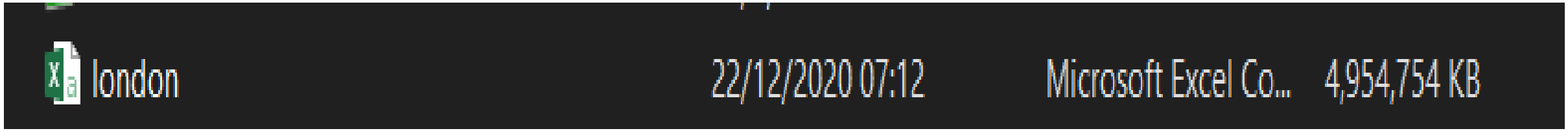
Submitted To: Data Glacier

Data File

- The data file used for this task is:

  File size: 4.95 GB

  Obtained from: kaggle



Reading File with Pandas

Data Glacier
Your Deep Learning Partner

```
#read file with pandas
import pandas as pd
import time as time
s = time.time()
%time df = pd.read_csv("/content/drive/MyDrive/london.csv")
e = time.time()
print("Pandas Loading Time = {}".format(e-s))
```

```
CPU times: user 1min 2s, sys: 21.5 s, total: 1min 23s
Wall time: 1min 59s
Pandas Loading Time = 119.00196409225464
```

Pandas took a while to read the file, but not as long as was expected.

Reading File with Dask

Data Glacier
Your Deep Learning Partner

```
#reading with dask
import time as time
s = time.time()
%time  df = dd.read_csv("/content/drive/MyDrive/london.csv")
e = time.time()
print("Pandas Loading Time = {}".format(e-s))
```

```
CPU times: user 605 ms, sys: 327 ms, total: 932 ms
Wall time: 978 ms
Pandas Loading Time = 0.982053279876709
```

- Reading the file with dask was much more faster.
- It took milliseconds.

Reading File with Modin (Ray)

Data Glacier
Your Deep Learning Partner

```
[14] #read file with modin
     import time as time
     import modin.pandas as md
     s = time.time()
     %time df = md.read_csv("/content/drive/MyDrive/london.csv")
     e = time.time()
     print("Modin ray Loading Time = {}".format(e-s))
```

```
(pid=966) tcmalloc: large alloc 2536841216 bytes == 0x55ee92056000 @  0x7ff5c67211e7 0x55ee8e348e68 0x55ee8e313637 0x55ee8e3f4a6e 0x55ee8e316b59 0x55ee8e407fed 0x55ee8e38a988 0
(pid=965) tcmalloc: large alloc 2536841216 bytes == 0x564f6090c000 @  0x7f4b6a5ab1e7 0x564f5d157e68 0x564f5d122637 0x564f5d203a6e 0x564f5d125b59 0x564f5d216fed 0x564f5d199988 0
2021-05-28 10:36:29,394 WARNING worker.py:1034 -- A worker died or was killed while executing task cb230a572350ff44ffffffff01000000.
(pid=966) tcmalloc: large alloc 2536841216 bytes == 0x55ee92056000 @  0x7ff5c67211e7 0x55ee8e348e68 0x55ee8e313637 0x55ee8e3f4a6e 0x55ee8e316b59 0x55ee8e407fed 0x55ee8e38a988 0
2021-05-28 10:38:06,082 WARNING worker.py:1034 -- A worker died or was killed while executing task cb230a572350ff44ffffffff01000000.
(pid=1049) tcmalloc: large alloc 2536841216 bytes == 0x55952cba0000 @  0x7f1de5bac1e7 0x559529351e68 0x55952931c637 0x5595293fda6e 0x55952931fb59 0x559529410fed 0x559529393988
2021-05-28 10:39:35,054 WARNING worker.py:1034 -- A worker died or was killed while executing task cb230a572350ff44ffffffff01000000.
2021-05-28 10:41:05,263 WARNING worker.py:1034 -- A worker died or was killed while executing task cb230a572350ff44ffffffff01000000.
```

- Modin failed to read the file.
- It used up a lot of memory.
- It took a long time and still failed to read the file.

Conclusion

Data Glacier
Your Deep Learning Partner

- Dask is best for this data and most large data sets in general because of its speed and efficiency.
- It has many resources online for those who are new.
- Pandas is good for smaller datasets, when it comes to large datasets it becomes slow and less efficient.

Thank You