Data Ingestion and Schema Validation

Name: Preeti Verma

Batch Code: LISUM19

Summary Text:

The data file used for this task is: data file

• File size: 5.5 GB

• Obtained from Kaggle

Reading file using Pandas:

```
[7] #read file with panadas
    import pandas as pd
    from google.colab import files
    data_to_load = files.upload()
```

```
Choose Files canada_weather.csv
• canada_weather.csv(text/csv) - 9031 bytes, last modified: 3/29/2023 - 100% done
Saving canada_weather.csv to canada_weather.csv
```

```
import time as time
import io
s = time.time()
%time df = pd.read_csv(io.BytesIO(data_to_load["canada_weather.csv"]))
e = time.time()
print("Panadas Loading Time = {}".format(e-s))
```

```
CPU times: user 7.45 ms, sys: 0 ns, total: 7.45 ms
Wall time: 18.5 ms
Panadas Loading Time = 0.020107507705688477
```

The pandas managed to read the file but were slow and took a while to load the data.

Reading File with Modin (Ray):

```
import pandas as pd
import time as time
import io
import modin.pandas as md
from google.colab import files
data_to_load = files.upload()
s = time.time()
%time df = pd.read_csv(io.BytesIO(data_to_load["canada_weather.csv"]))
e = time.time()
print("Modin ray Loading Time = {}".format(e-s))
```

```
Choose Files canada_weather.csv
• canada_weather.csv(text/csv) - 9031 bytes, last modified: 3/29/2023 - 100% done
Saving canada_weather.csv to canada_weather (1).csv
CPU times: user 5.43 ms, sys: 62 µs, total: 5.5 ms
Wall time: 5.77 ms
Modin ray Loading Time = 0.007829666137695312
```

+ Code + Text

```
[ ]
```

When modin was used to read the big file, it was best to read the file and took less time to read the file.

Reading a file with Dask:

```
import dask.dataframe as dd
s = time.time()
%time df = dd.read_csv('canada_weather (4).csv')
e = time.time()
print("Dask Loading Time = {}".format(e-s))
```

```
CPU times: user 15 ms, sys: 1.85 ms, total: 16.9 ms
Wall time: 24.3 ms
Dask Loading Time = 0.02857351303100586
```

Reading the big file with dask was simple and fast. It took more than compared to other two to read CSV file

Conclusion:

In terms of computer efficiency, I believe modin(ray) will be best for this data and most large data sets in general because of its speed and efficiency. It is also mature and has many resources online for those who are new. Unlike dask. Additionally, all the other methods make use of pandas in some way or another. Pandas are good for small datasets, however, when it comes to large datasets, they become slow and less efficient