Data Glacier

File Ingestion and Schema Validation

Name: Abishek James

Batch Code: LISUM19

Date: 10 April 2023

Submitted to: Data Glacier

Submitted link: https://github.com/abishekjames/Data-Glacier-Week6

Dataset

Dataset used for this project is from Kaggle which is 2Gb of size.

https://www.kaggle.com/datasets/ekibee/car-sales-information?select=region41.csv

region25.csv is renamed as carsales.csv.

Reading file in pandas

```
import pandas as pd
```

```
%%time
df = pd.read_csv('carsales.csv')
```

```
CPU times: total: 47.3 s
Wall time: 49 s
```

file in modin and ray

```
import modin.pandas as pd
import ray
ray.shutdown()
ray.init()
start = time.time()
df = pd.read_csv('carsales.csv')
end = time.time()
print("Read csv with modin and ray: ",(end-start),"sec")
```

```
2023-04-04 15:28:05,145 INFO worker.py:1553 -- Started a local Ray instance.
Read csv with modin and ray:  73.02560472488403 sec
```

Reading file in Dask

```
from dask import dataframe as dd
```

```
%%time
dask_df = dd.read_csv('carsales.csv')
```

```
CPU times: total: 15.6 ms
Wall time: 87.1 ms
```

```
start = time.time()
dask_df = dd.read_csv('carsales.csv')
end = time.time()
print("Read csv with dask: ",(end-start),"sec")
```

```
Read csv with dask:  0.020333051681518555 sec
```

Conclusion

Each of the libraries has different strengths, weaknesses, and scaling strategies. From the above reading with data, Dask in faster. It reads CSV files (Size of 2GB) faster than Pandas,Modin and ray.

So as a conclusion, Dask is better than pandas and Modin with computational time of 0.025 s, whereas the other two took 73 s and 50 s respectively.

File in pip separated the text file in gz format

```
# Write csv in gz format in pipe separated text file (|)
df.to_csv("car_sales.gz", sep = '|', index = False)
```

```
#size of the gz format folder
import os
os.path.getsize('C:\\Users\\Admin\\Desktop\\Data Glacier\\Week6\\car_sales.gz')
```

```
457589697
```

```
entries = os.listdir('C:\\Users\\Admin\\Desktop\\Data Glacier\\Week6\\car_sales.gz')
len(entries)
```

```
38
```

Total size of gz format folder 436MB