Data Intake Report

Name: File ingestion and schema validation
Report date: July 11, 2021
Internship Batch: LISUM01
Version:<1.0>
Data intake by: Nicolette Peterkin
Data intake reviewer: N/A
Data storage location: https://github.com/NicolettePeterkin/Data-Ingestion-Pipeline.week6.git

Tabular data details:

Total number of observations	10626899
Total number of files	1
Total number of features	51
Base format of the file	.csv
Size of the data	2 GB

Time taken to read files comparison

1. Pandas took 69.10192894935608 seconds
2. Pandas took with chunk size 1.3400604724884033 seconds
3. Modin [Ray] 193.85107588768005 seconds
4. Dask took 0.1345052719116211 seconds

```
DtypeWarning: Columns (17,18,20,21,22,23,29,30,31,32,34,36,38,39) have mixed types.Specify dtype option on import or set low_memory=False.
Pandas took 69.10192894935608 seconds
Pandas took with chunksize 1.3400604724884033 seconds
UserWarning: Ray execution environment not yet initialized. Initializing...
To remove this warning, run the following python code before doing dataframe operations:

    import ray
    ray.init()

2021-07-12 01:10:30,902 WARNING services.py:1740 -- WARNING: The object store is using /tmp instead of /dev/shm because /dev/shm has only 630823
(pid=4811) tcmalloc: large alloc 1075970048 bytes == 0x55ce2ac20000 @  0x7f3f4756a1e7 0x55ce26f94e68 0x55ce26f5f637 0x55ce27040a6e 0x55ce26f62b5
(pid=4810) tcmalloc: large alloc 1075970048 bytes == 0x55e36d614000 @  0x7f026756f1e7 0x55e36a9d2e68 0x55e36a99d637 0x55e36aa7ea6e 0x55e36a9a0b5
2021-07-12 01:12:11,145 WARNING worker.py:1123 -- A worker died or was killed while executing a task by an unexpected system error. To troublesho
(pid=4894) tcmalloc: large alloc 1075970048 bytes == 0x55dfc7152000 @  0x7f75b7b581e7 0x55dfc3ff0e68 0x55dfc3fbb637 0x55dfc409ca6e 0x55dfc3fbeb5
Modin[Ray] 193.85107588768005 seconds
Dask took 0.1345052719116211 seconds
```

Conclusion

After running the analysis dask took the shortest amount of time to run while modin ray took the longest. Pandas also took a long time. However, when pandas was used with chunk size it took way less time than the original almost as fast as dask.

Summary of the Pipeline:

```
import pandas as pd
import utility as util

# Read config file
config_data = util.read_config_file("configuration.yaml")

# read the file using config file
file_type = config_data['file_type']
source_file = "./" + config_data['file_name'] + f'.{file_type}'
#print("",source_file)
df = pd.read_csv(source_file,config_data['inbound_delimiter'],)


if util.col_header_val(df,config_data)==0:
    print("validation failed")
    # write code to reject the file
else:
    print("col validation passed")
    # write the code to perform further action
    # in the pipleine
    util.file_summary(df,config_data)
    util.saveFile(df,config_data)
```

```
/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarning: Columns (17,18,20,21,22,23,29,30,31,32,34,36,38,39) have mixed types.Specify dtype option on import or set
  interactivity=interactivity, compiler=compiler, result=result)
col validation passed
FILE SUMMARY FOR:  parking_violations_issued_fiscal_year_2016.csv
Total number of rows:  10626899
Total number of columns:  51
File size:  2 GB
```