clean_and_convert

June 1, 2024

1 data info

source: https://www.unb.ca/cic/datasets/ids-2018.html

Selected days: - Wed-14-02-2018 - Thurs-15-02-2018 - Fri-16-02-2018

Available attacks: - FTP-BruteForce - SSH-Bruteforce - DoS-GoldenEye - DoS-Slowloris - DoS-SlowHTTPTest - DoS-Hulk

[ ]:
```
!pip install -r requirements.txt
```

[3]:
```
from pandas import StringDtype

types = {
    'Dst Port': 'int64',
    'Protocol': 'int64',
    'Timestamp': 'int64',
    'Flow Duration': 'int64',
    'Tot Fwd Pkts': 'int64',
    'Tot Bwd Pkts': 'int64',
    'TotLen Fwd Pkts': 'int64',
    'TotLen Bwd Pkts': 'int64',
    'Fwd Pkt Len Max': 'int64',
    'Fwd Pkt Len Min': 'int64',
    'Fwd Pkt Len Mean': 'float64',
    'Fwd Pkt Len Std': 'float64',
    'Bwd Pkt Len Max': 'int64',
    'Bwd Pkt Len Min': 'int64',
    'Bwd Pkt Len Mean': 'float64',
    'Bwd Pkt Len Std': 'float64',
    'Flow Byts/s': 'float64',
    'Flow Pkts/s': 'float64',
    'Flow IAT Mean': 'float64',
    'Flow IAT Std': 'float64',
    'Flow IAT Max': 'int64',
    'Flow IAT Min': 'int64',
    'Fwd IAT Tot': 'int64',
    'Fwd IAT Mean': 'float64',
    'Fwd IAT Std': 'float64',
```

```
    'Fwd IAT Max': 'int64',
    'Fwd IAT Min': 'int64',
    'Bwd IAT Tot': 'int64',
    'Bwd IAT Mean': 'float64',
    'Bwd IAT Std': 'float64',
    'Bwd IAT Max': 'int64',
    'Bwd IAT Min': 'int64',
    'Fwd PSH Flags': 'int64',
    'Bwd PSH Flags': 'int64',
    'Fwd URG Flags': 'int64',
    'Bwd URG Flags': 'int64',
    'Fwd Header Len': 'int64',
    'Bwd Header Len': 'int64',
    'Fwd Pkts/s': 'float64',
    'Bwd Pkts/s': 'float64',
    'Pkt Len Min': 'int64',
    'Pkt Len Max': 'int64',
    'Pkt Len Mean': 'float64',
    'Pkt Len Std': 'float64',
    'Pkt Len Var': 'float64',
    'FIN Flag Cnt': 'int64',
    'SYN Flag Cnt': 'int64',
    'RST Flag Cnt': 'int64',
    'PSH Flag Cnt': 'int64',
    'ACK Flag Cnt': 'int64',
    'URG Flag Cnt': 'int64',
    'CWE Flag Count': 'int64',
    'ECE Flag Cnt': 'int64',
    'Down/Up Ratio': 'int64',
    'Pkt Size Avg': 'float64',
    'Fwd Seg Size Avg': 'float64',
    'Bwd Seg Size Avg': 'float64',
    'Fwd Byts/b Avg': 'int64',
    'Fwd Pkts/b Avg': 'int64',
    'Fwd Blk Rate Avg': 'int64',
    'Bwd Byts/b Avg': 'int64',
    'Bwd Pkts/b Avg': 'int64',
    'Bwd Blk Rate Avg': 'int64',
    'Subflow Fwd Pkts': 'int64',
    'Subflow Fwd Byts': 'int64',
    'Subflow Bwd Pkts': 'int64',
    'Subflow Bwd Byts': 'int64',
    'Init Fwd Win Byts': 'int64',
    'Init Bwd Win Byts': 'int64',
    'Fwd Act Data Pkts': 'int64',
    'Fwd Seg Size Min': 'int64',
    'Active Mean': 'float64',
```

```
    'Active Std': 'float64',
    'Active Max': 'int64',
    'Active Min': 'int64',
    'Idle Mean': 'float64',
    'Idle Std': 'float64',
    'Idle Max': 'int64',
    'Idle Min': 'int64',
    'Label': StringDtype()
}
```

- merging files into single file
- loading data from csv
- removing broken rows (duplicated header row on far index)
- converting Timestamp from string to seconds (int)
- saving in different formats
- comparing file sizes and load times

[1]:
```
import pandas as pd

# Load the first CSV file
csv1 = pd.read_csv('Thursday-15-02-2018_TrafficForML_CICFlowMeter.csv')

# Load the second CSV file
csv2 = pd.read_csv('Wednesday-14-02-2018_TrafficForML_CICFlowMeter.csv')

# Merge the two DataFrames
merged_csv = pd.concat([csv1, csv2], ignore_index=True)

# Save the merged DataFrame to a new CSV file
merged_csv.to_csv('data.csv', index=False)
```

[9]:
```
from pandas import read_csv, read_parquet, read_orc, read_pickle, to_datetime
from time import time

file="data.csv"

print(f"Converting file {file}")
file_prefix = file.removesuffix('.csv')
start = time()
df = read_csv(file)
print(f"Time for csv: {time() - start}")
for index, port in enumerate(df['Dst Port']):
    try:
        test = int(port)
    except ValueError as exc:
        print(f"{exc}, index: {index}, value: '{port}'")
        df = df[df['Dst Port'] != port]
```

```
# converting time string to seconds
df['Timestamp'] = to_datetime(df['Timestamp'], format='%d/%m/%Y %H:%M:%S').
 ↪apply(lambda x: x.to_pydatetime().timestamp())

print(f"Attack labels: {set(df['Label'])}")
print("shape: ", df.shape)

df = df.astype(types).reset_index(drop=True)
df.to_pickle(f"{file_prefix}.pickle")
df.to_parquet(f"{file_prefix}.parquet")
df.to_orc(f"{file_prefix}.orc")


start = time()
read_pickle(f"{file_prefix}.pickle")
print(f"Time for pickled: {time() - start}")
start = time()
read_parquet(f"{file_prefix}.parquet")
print(f"Time for parquet: {time() - start}")
start = time()
read_orc(f"{file_prefix}.orc")
print(f"Time for orc: {time() - start}")
```

```
Converting file data.csv
Time for csv: 5.93277382850647
Attack labels: {'Benign', 'FTP-BruteForce', 'DoS attacks-GoldenEye', 'SSH-
Bruteforce', 'DoS attacks-Slowloris'}
shape:  (2097150, 80)
Time for pickled: 0.18784666061401367
Time for parquet: 0.37416887283325195
Time for orc: 1.134911298751831
```

[10]:
```
from pandas import read_parquet, set_option

df = read_parquet('data.parquet')
set_option('display.max_columns', None)
print(df.head(1))
```

```
   Dst Port  Protocol   Timestamp  Flow Duration  Tot Fwd Pkts  Tot Bwd Pkts  \
0         0         0  1518679518      112641158             3             0

   TotLen Fwd Pkts  TotLen Bwd Pkts  Fwd Pkt Len Max  Fwd Pkt Len Min  \
0                0                0                0                0

   Fwd Pkt Len Mean  Fwd Pkt Len Std  Bwd Pkt Len Max  Bwd Pkt Len Min  \
0               0.0              0.0                0                0

   Bwd Pkt Len Mean  Bwd Pkt Len Std  Flow Byts/s  Flow Pkts/s  Flow IAT Mean  \
```

```
0              0.0              0.0          0.0     0.026633     56320579.0

   Flow IAT Std  Flow IAT Max  Flow IAT Min  Fwd IAT Tot  Fwd IAT Mean  \
0    704.278354      56321077      56320081    112641158    56320579.0

   Fwd IAT Std  Fwd IAT Max  Fwd IAT Min  Bwd IAT Tot  Bwd IAT Mean  \
0   704.278354     56321077     56320081            0           0.0

   Bwd IAT Std  Bwd IAT Max  Bwd IAT Min  Fwd PSH Flags  Bwd PSH Flags  \
0          0.0            0            0              0              0

   Fwd URG Flags  Bwd URG Flags  Fwd Header Len  Bwd Header Len  Fwd Pkts/s  \
0              0              0               0               0    0.026633

   Bwd Pkts/s  Pkt Len Min  Pkt Len Max  Pkt Len Mean  Pkt Len Std  \
0         0.0            0            0           0.0          0.0

   Pkt Len Var  FIN Flag Cnt  SYN Flag Cnt  RST Flag Cnt  PSH Flag Cnt  \
0          0.0             0             0             0             0

   ACK Flag Cnt  URG Flag Cnt  CWE Flag Count  ECE Flag Cnt  Down/Up Ratio  \
0             0             0               0             0              0

   Pkt Size Avg  Fwd Seg Size Avg  Bwd Seg Size Avg  Fwd Byts/b Avg  \
0           0.0               0.0               0.0               0

   Fwd Pkts/b Avg  Fwd Blk Rate Avg  Bwd Byts/b Avg  Bwd Pkts/b Avg  \
0               0                 0               0               0

   Bwd Blk Rate Avg  Subflow Fwd Pkts  Subflow Fwd Byts  Subflow Bwd Pkts  \
0                 0                 3                 0                 0

   Subflow Bwd Byts  Init Fwd Win Byts  Init Bwd Win Byts  Fwd Act Data Pkts  \
0                 0                -1                 -1                  0

   Fwd Seg Size Min  Active Mean  Active Std  Active Max  Active Min  \
0                 0          0.0         0.0           0           0

    Idle Mean    Idle Std  Idle Max  Idle Min   Label
0  56320579.0  704.278354  56321077  56320081  Benign
```