clean_and_convert

May 17, 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

[1]:

```
!pip install -r requirements.txt
```

```
Requirement already satisfied: pandas in /venv/lib64/python3.12/site-packages
(from -r requirements.txt (line 1)) (2.2.2)
Requirement already satisfied: pyarrow in /venv/lib64/python3.12/site-packages
(from -r requirements.txt (line 2)) (16.1.0)
Requirement already satisfied: numpy>=1.26.0 in /venv/lib64/python3.12/site-
packages (from pandas->-r requirements.txt (line 1)) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
/venv/lib64/python3.12/site-packages (from pandas->-r requirements.txt (line 1))
(2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /venv/lib64/python3.12/site-
packages (from pandas->-r requirements.txt (line 1)) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in /venv/lib64/python3.12/site-
packages (from pandas->-r requirements.txt (line 1)) (2024.1)
Requirement already satisfied: six>=1.5 in /venv/lib64/python3.12/site-packages
(from python-dateutil>=2.8.2->pandas->-r requirements.txt (line 1)) (1.16.0)
```

[2]:

```python
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()
}
```

- 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

[8]:

```
from pandas import read_csv, read_parquet, read_orc, read_pickle, to_datetime
from time import time

files = ['Thursday-15-02-2018_TrafficForML_CICFlowMeter.csv',
        'Wednesday-14-02-2018_TrafficForML_CICFlowMeter.csv',
        'Friday-16-02-2018_TrafficForML_CICFlowMeter.csv']

for file in files:
    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'])}")

    df = df.astype(types).reset_index()
    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 Thursday-15-02-2018_TrafficForML_CICFlowMeter.csv
Time for csv: 2.606449604034424
Attack labels: {'DoS attacks-GoldenEye', 'DoS attacks-Slowloris', 'Benign'}
Time for pickled: 0.08860635757446289
Time for parquet: 0.25847291946411133
Time for orc: 0.5763046741485596
Converting file Wednesday-14-02-2018_TrafficForML_CICFlowMeter.csv
Time for csv: 2.413745164871216
Attack labels: {'SSH-Bruteforce', 'FTP-BruteForce', 'Benign'}
Time for pickled: 0.09118247032165527
Time for parquet: 0.2338240146636963
Time for orc: 0.525947093963623
Converting file Friday-16-02-2018_TrafficForML_CICFlowMeter.csv
```

```
/tmp/ipykernel_4668/4107320491.py:12: DtypeWarning: Columns (0,1,3,4,5,6,7,8,9,1
0,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,
37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63
,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78) have mixed types. Specify dtype
option on import or set low_memory=False.
  df = read_csv(file)
```

```
Time for csv: 3.345402956008911
```

```
invalid literal for int() with base 10: 'Dst Port', index: 999999, value: 'Dst
Port'
Attack labels: {'DoS attacks-SlowHTTPTest', 'DoS attacks-Hulk', 'Benign'}
Time for pickled: 0.10001444816589355
Time for parquet: 0.2265605926513672
Time for orc: 0.5186152458190918
```

[9]:

```python
from pandas import read_parquet, set_option

df = read_parquet('Wednesday-14-02-2018_TrafficForML_CICFlowMeter.parquet')
set_option('display.max_columns', None)
print(df.head(1))
```

```
   index  Dst Port  Protocol   Timestamp  Flow Duration  Tot Fwd Pkts  \
0      0         0         0  1518593461      112641719             3

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

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

   Bwd Pkt Len Min  Bwd Pkt Len Mean  Bwd Pkt Len Std  Flow Byts/s  \
0                0               0.0              0.0          0.0

   Flow Pkts/s  Flow IAT Mean  Flow IAT Std  Flow IAT Max  Flow IAT Min  \
0     0.026633     56320859.5    139.300036      56320958      56320761

   Fwd IAT Tot  Fwd IAT Mean  Fwd IAT Std  Fwd IAT Max  Fwd IAT Min  \
0    112641719    56320859.5   139.300036     56320958     56320761

   Bwd IAT Tot  Bwd IAT Mean  Bwd IAT Std  Bwd IAT Max  Bwd IAT Min  \
0            0           0.0          0.0            0            0

   Fwd PSH Flags  Bwd PSH Flags  Fwd URG Flags  Bwd URG Flags  Fwd Header Len  \
0              0              0              0              0               0

   Bwd Header Len  Fwd Pkts/s  Bwd Pkts/s  Pkt Len Min  Pkt Len Max  \
0               0    0.026633         0.0            0            0

   Pkt Len Mean  Pkt Len Std  Pkt Len Var  FIN Flag Cnt  SYN Flag Cnt  \
0           0.0          0.0          0.0             0             0

   RST Flag Cnt  PSH Flag Cnt  ACK Flag Cnt  URG Flag Cnt  CWE Flag Count  \
0             0             0             0             0               0

   ECE Flag Cnt  Down/Up Ratio  Pkt Size Avg  Fwd Seg Size Avg  \
0             0              0           0.0               0.0
```

```
   Bwd Seg Size Avg  Fwd Byts/b Avg  Fwd Pkts/b Avg  Fwd Blk Rate Avg  \
0               0.0               0               0                 0

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

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

   Init Bwd Win Byts  Fwd Act Data Pkts  Fwd Seg Size Min  Active Mean  \
0                 -1                  0                 0          0.0

   Active Std  Active Max  Active Min   Idle Mean    Idle Std  Idle Max  \
0         0.0           0           0  56320859.5  139.300036  56320958

   Idle Min   Label
0  56320761  Benign
```

```
[ ]:
```