# prova2BES

November 20, 2020

## 1 Prova 2 - Big Data

Considerando o dataset detalhado a seguir, extraia o conjunto de informações solicitadas.

### 1.0.1 Dataset dados de ataques de rede

- Arquivo disponível em /home/dados/ddos/prova.csv
- Dados relativos a ataques em nível de rede em uma rede de computadores
- ~1GB
- ~2M de instâncias

| # | Nome | Descrição |
|---|---|---|
| 0 | Number | Numero |
| 1 | Flow ID | Identificador |
| 2 | Src IP | IP de origem |
| 3 | Src Port | Porta de origem |
| 4 | Dst IP | Ip de destino |
| 5 | Dst Port | Porta de destino |
| 6 | Protocol | Protocolo |
| 7 | Timestamp | Timestamp |
| 8 | Flow duration | Duração do fluxo em microssegundos |
| 9 | total Fwd Packet | Pacotes totais na direção para servidor |
| 10 | total Bwd packets | Pacotes totais na direção para cliente |
| 11 | total Length of Fwd Packet | Tamanho total do pacote na direção para servidor |
| 12 | total Length of Bwd Packet | Tamanho total do pacote na direção para cliente |
| 13 | Fwd Packet Length Min | Tamanho mínimo do pacote na direção para servidor |
| 14 | Fwd Packet Length Max | Tamanho máximo do pacote na direção para servidor |
| 15 | Fwd Packet Length Mean | Tamanho médio do pacote na direção para servidor |
| 16 | Fwd Packet Length Std | Tamanho do desvio padrão do pacote na direção para servidor |

| # | Nome | Descrição |
|---|---|---|
| 17 | Bwd Packet Length Min | Tamanho mínimo do pacote na direção para cliente |
| 18 | Bwd Packet Length Max | Tamanho máximo do pacote na direção para cliente |
| 19 | Bwd Packet Length Mean | Tamanho médio do pacote na direção para cliente |
| 20 | Bwd Packet Length Std | Tamanho do desvio padrão do pacote na direção para cliente |
| 21 | Flow Bytes/s | Número de bytes de fluxo por segundo |
| 22 | Flow Packets/s | Número de pacotes de fluxo por segundo |
| 23 | Flow IAT Mean | Tempo médio entre dois pacotes enviados no fluxo |
| 24 | Flow IAT Std | Tempo de desvio padrão entre dois pacotes enviados no fluxo |
| 25 | Flow IAT Max | Tempo máximo entre dois pacotes enviados no fluxo |
| 26 | Flow IAT Min | Tempo mínimo entre dois pacotes enviados no fluxo |
| 27 | Fwd IAT Min | Tempo mínimo entre dois pacotes enviados na direção para servidor |
| 28 | Fwd IAT Max | Tempo máximo entre dois pacotes enviados na direção para servidor |
| 29 | Fwd IAT Mean | Tempo médio entre dois pacotes enviados na direção para servidor |
| 30 | Fwd IAT Std | Tempo de desvio padrão entre dois pacotes enviados na direção para servidor |
| 31 | Fwd IAT Total | Tempo total entre dois pacotes enviados na direção para servidor |
| 32 | Bwd IAT Min | Tempo mínimo entre dois pacotes enviados no sentido inverso |
| 33 | Bwd IAT Max | Tempo máximo entre dois pacotes enviados no sentido inverso |
| 34 | Bwd IAT Mean | Tempo médio entre dois pacotes enviados na direção para cliente |
| 35 | Bwd IAT Std | Tempo de desvio padrão entre dois pacotes enviados na direção para cliente |
| 36 | Bwd IAT Total | Tempo total entre dois pacotes enviados na direção para cliente |
| 37 | Fwd PSH flags | Número de vezes que a flag PSH foi definida em pacotes que viajam na direção para servidor (0 para UDP) |
| 38 | Bwd PSH Flags | Número de vezes que a flag PSH foi definida em pacotes que viajam na direção para cliente (0 para UDP) |
| 39 | Fwd URG Flags | Número de vezes que a flag URG foi definida em pacotes que viajam na direção para servidor (0 para UDP) |
| 40 | Bwd URG Flags | Número de vezes que a flag URG foi definida em pacotes que viajam na direção para cliente (0 para UDP) |
| 41 | Fwd Header Length | Total de bytes usados para cabeçalhos na direção para servidor |
| 42 | Bwd Header Length | Total de bytes usados para cabeçalhos no sentido inverso |
| 43 | FWD Packets/s | Número de pacotes encaminhados por segundo |
| 44 | Bwd Packets/s | Número de pacotes para trás por segundo |
| 45 | Packet Length Min | Comprimento mínimo de um pacote |
| 46 | Packet Length Max | Comprimento máximo de um pacote |
| 47 | Packet Length Mean | Comprimento médio de um pacote |

| # | Nome | Descrição |
|---|---|---|
| 48 | Packet Length Std | Comprimento do desvio padrão de um pacote |
| 49 | Packet Length Variance | Comprimento de variância de um pacote |
| 50 | FIN Flag Count | Número de pacotes com FIN |
| 51 | SYN Flag Count | Número de pacotes com SYN |
| 52 | RST Flag Count | Número de pacotes com RST |
| 53 | PSH Flag Count | Número de pacotes com PUSH |
| 54 | ACK Flag Count | Número de pacotes com ACK |
| 55 | URG Flag Count | Número de pacotes com URG |
| 56 | CWR Flag Count | Número de pacotes com CWR |
| 57 | ECE Flag Count | Número de pacotes com ECE |
| 58 | down/Up Ratio | Taxa de download e upload |
| 59 | Average Packet Size | Tamanho médio do pacote |
| 60 | Fwd Segment Size Avg | Tamanho médio observado na direção para servidor |
| 61 | Bwd Segment Size Avg | Taxa média do número de bytes em massa no sentido para cliente |
| 62 | Fwd Bytes/Bulk Avg | Taxa média do número de bytes em massa na direção para servidor |
| 63 | Fwd Packet/Bulk Avg | Taxa média do número de pacotes em massa na direção para servidor |
| 64 | Fwd Bulk Rate Avg | Número médio de taxa em massa na direção para servidor |
| 65 | Bwd Bytes/Bulk Avg | Taxa média do número de bytes em massa no sentido para cliente |
| 66 | Bwd Packet/Bulk Avg | Taxa média do número de pacotes em massa na direção para cliente |
| 67 | Bwd Bulk Rate Avg | Número médio de taxa em massa na direção para cliente |
| 68 | Subflow Fwd Packets | O número médio de pacotes em um subfluxo na direção para servidor |
| 69 | Subflow Fwd Bytes | O número médio de bytes em um subfluxo na direção para servidor |
| 70 | Subflow Bwd Packets | O número médio de pacotes em um subfluxo na direção para cliente |
| 71 | Subflow Bwd Bytes | O número médio de bytes em um subfluxo na direção para cliente |
| 72 | Fwd Init Win bytes | O número total de bytes enviados na janela inicial na direção para servidor |
| 73 | Bwd Init Win bytes | O número total de bytes enviados na janela inicial na direção cliente |
| 74 | Fwd Act Data Pkts | Contagem de pacotes com pelo menos 1 byte de carga útil de dados TCP na direção para servidor |
| 75 | Fwd Seg Size Min | Tamanho mínimo do segmento observado na direção para servidor |

| # | Nome | Descrição |
|---|---|---|
| 76 | Active Min | Tempo mínimo em que um fluxo esteve ativo antes de se tornar ocioso |
| 77 | Active Mean | Tempo médio em que um fluxo estava ativo antes de ficar ocioso |
| 78 | Active Max | Tempo máximo em que um fluxo ficou ativo antes de se tornar ocioso |
| 79 | Active Std | Tempo de desvio padrão em que um fluxo estava ativo antes de ficar ocioso |
| 80 | Idle Min | Tempo mínimo em que um fluxo ficou ocioso antes de se tornar ativo |
| 81 | Idle Mean | Tempo médio em que um fluxo ficou ocioso antes de se tornar ativo |
| 82 | Idle Max | Tempo máximo em que um fluxo ficou ocioso antes de se tornar ativo |
| 83 | Idle Std | Tempo de desvio padrão em que um fluxo estava ocioso antes de se tornar ativo |
| 84 | Label | Classe do fluxo |

## 1.1 Extraia as informações solicitadas no PDF entregue, lembre-se de mencionar a informação extraida na célula

[1]:
```
# abre sessão no spark
import os
os.environ['PYSPARK_PYTHON'] = '/usr/bin/python3'

import pyspark
conf = pyspark.SparkConf()

conf.setMaster('spark://spark-master:7077')

sc = pyspark.SparkContext.getOrCreate()
sc.stop()
sc = pyspark.SparkContext(conf = conf)
```

[2]:
```
#carrega arquivo do HDFS em um RDD
arquivoRDD = sc.textFile('hdfs://namenode:9000/prova.csv')
```

### 1.1.1 Informação PYSPARK 1

[3]:
```
# Quais são os 5 dias com mais conexões na base
# TIMESTAMP
arquivoRDD.map(lambda l: l.split(',')[7])\
          .map(lambda l: [(l.split(' ')[0]), 1])\
          .filter(lambda l: l[0] != 'Timestamp')\
          .reduceByKey(lambda x,y: x+y)\
          .sortBy(lambda c: c[1], False)\
          .take(5)
```

[3]:
```
[('20/02/2018', 773676),
 ('16/02/2018', 723196),
 ('22/02/2018', 253894),
 ('21/02/2018', 180242),
 ('03/07/2017', 39505)]
```

[ ]:

[ ]:

### 1.1.2 Informação PYSPARK 2

[4]:
```
# Qual a quantidade de conexões com duração maior que 100 segundos e
# menor que 200 segundos de acordo com seu label
```

[5]:
```
def limite(x):
    try:
        x = int(x)
        if x < 200 and x > 100:
            return 1
        else:
            return 0
    except:
        return 0

arquivoRDD.map(lambda l: [l.split(',')[84], l.split(',')[8]])\
        .map(lambda l: [l[0],limite(l[1])])\
        .reduceByKey(lambda x,y: x+y)\
        .filter(lambda l: l[0]!='Label')\
        .collect()
```

[5]:
```
[('ddos', 3986), ('Benign', 22217)]
```

[ ]:

### 1.1.3 Informação PYSPARK 3

[6]:
```
# Qual a ocorrência de cada label
arquivoRDD.map(lambda l: [l.split(',')[84], 1])\
    .filter(lambda l: l[0] != 'Label')\
    .reduceByKey(lambda x,y: x+y)\
    .collect()
```

[6]:
```
[('ddos', 1294529), ('Benign', 705470)]
```

[ ]:

[ ]:

### 1.1.4 Informação PYSPARK 4

[7]:
```
# Quais os 5 IPs com maior ocorrência como origem de fluxo da conexão
```

[8]:
```
arquivoRDD.map(lambda l: [l.split(',')[2], 1])\
        .filter(lambda l: l[0] != 'Src IP')\
        .reduceByKey(lambda x,y: x+y)\
        .sortBy(lambda l: l[1], False)\
        .take(5)
```

[8]:
```
[('172.31.69.25', 353151),
 ('18.219.193.20', 348970),
 ('172.31.69.28', 185081),
 ('18.218.229.235', 37035),
 ('18.216.200.189', 36992)]
```

[ ]:

### 1.1.5 Informação PYSPARK 5

[9]:
```
# Qual o fluxo com maior duração para cada label de acordo com seu protocolo
```

[10]:
```
# Para comparar valores string muito grandes e difíceis de representar como int.
def compare(x, y):
    arr = [x,y]
    arr.sort()
    return arr[1]

arquivoRDD.map(lambda l: [str(l.split(',')[84])+"-"+str(l.split(',')[6]), l.
 ↪split(',')[8]])\
       .reduceByKey(lambda x,y: compare(x,y))\
       .collect()
```

[10]:
```
[('Label-Protocol', 'Flow Duration'),
 ('ddos-6', '99999947'),
 ('Benign-0', '99999990'),
 ('Benign-6', '999999'),
 ('Benign-17', '99992'),
 ('ddos-17', '99999830')]
```

[ ]:

### 1.1.6 Informação PYSPARK 6

[11]:
```
# Sabendo que um IP é caracterizado por 4 bytes no formato BYTE1.BYTE.BYTE3.
 ↪BYTE4
# Determine quais os 5 valores que ocorrem com maior frequência no BYTE1 do ip␣
 ↪de destino.
```

[12]:
```
arquivoRDD.map(lambda l: [l.split(',')[4].split('.')[0], 1])\
        .reduceByKey(lambda x,y: x+y)\
        .sortBy(lambda l: l[1], False)\
        .take(5)
```

[12]:
```
[('172', 1182960),
 ('18', 518799),
 ('169', 49418),
 ('52', 32581),
 ('192', 31510)]
```

[ ]:

### 1.1.7 Informação PYSPARK 7

[13]:
```
# Sabendo que um serviço é caracterizado pelo campo IP de destino e porta de␣
 ↪destino, quais são os 5 serviços
# mais acessados.
```

[14]:
```
arquivoRDD.map(lambda l: [l.split(',')[4]+"::"+l.split(',')[5], 1])\
        .reduceByKey(lambda x,y: x+y)\
        .sortBy(lambda l: l[1], False)\
        .take(5)
```

[14]:
```
[('172.31.69.25::80', 475265),
 ('172.31.69.28::80', 250274),
 ('172.31.0.2::53', 218991),
 ('169.254.169.254::80', 48885),
 ('172.31.69.25::21', 21245)]
```

[15]:
```
# " ... :: ... " = " IP :: Port"
```

### 1.1.8 Informação PYSPARK 8

[16]:
```
# Qual a quantidade de conexões para cada tipo considerado(Flow Duration)
```

[17]:
```
def tipo(x):
    try:
        x = int(x)

        if x < 101 and x > 0:
            return 'pequena'
        elif x < 1001:
            return 'media'
        elif x < 10001:
            return 'grande'
        else:
            return 'jumbo'
    except:
        return 'erro'

arquivoRDD.map(lambda l: l.split(',')[8])\
        .map(lambda l: [tipo(l), 1])\
        .reduceByKey(lambda x,y: x+y)\
        .filter(lambda l: l[0]!='erro')\
        .take(5)
```

[17]:
```
[('pequena', 112337),
 ('media', 226615),
 ('jumbo', 1232358),
 ('grande', 428689)]
```

[ ]:
```

```