Desafio Semantix !!!

Desenvolvimento da Segunda Parte do Desafio:

Responda as seguintes questões devem ser desenvolvidas em Spark utilizando a sua linguagem de preferência.

1) Número de hosts únicos.
r) 161884

In [2]:
```
#testando a variavel Spark
sc
```

Out[2]:

SparkContext

Spark UI (http://LAPTOP-OODQM8OG:4040)

Version

v2.3.0

Master

local[*]

AppName

PySparkShell

In [3]:
```
#Acessando o arquivo de Julho
base = sc.textFile("C:/Users/thiag/Desktop/Semantix/access_log_Jul
95.txt")
base
```

Out[3]:
```
C:/Users/thiag/Desktop/Semantix/access_log_Jul95.txt MapPartition
sRDD[1] at textFile at NativeMethodAccessorImpl.java:0
```

In [4]:
```
#Checando assinatura da row
base.first()
```

Out[4]:
```
'199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apol
lo/ HTTP/1.0" 200 6245'
```

In [5]:
```
#Splitando por ' - - ', e retiramos as partes da string diferente
dos holts
temp1 = base.flatMap(lambda k: k.split(" - - "))
temp1
```

Out[5]:
```
'199.72.81.55'
```