In [1]:
```
from pyspark.sql import SparkSession
from pyspark.sql import SQLContext
from pyspark.sql import Row
```

In [2]:
```
print(sc)
```

```
<SparkContext master=local[*] appName=PySparkShell>
```

In [3]:
```
spSession = SparkSession.builder.master("local").appName("NasaSQL").getOrCre
ate()
```

In [4]:
```
sqlContext = SQLContext(sc)
```

In [21]:
```
RDD = sc.textFile("/home/neto/nasa/nasa_temp")
```

In [22]:
```
RDD.count()
```

Out[22]:
```
106
```

In [23]:
```
RDD1 = RDD.map(lambda x: x.replace(" - - [", " "))
```

In [24]:
```
RDD2 = RDD1.map(lambda x: x.replace("] \"GET ", " GET"))
```

In [25]:
```
RDD3 = RDD2.map(lambda x: x.replace("\" ", " "))
```

In [26]:
```
RDD4 = RDD3.map(lambda x: x.replace(" -", "-"))
```

In [27]:
```
RDD5 = RDD4.map(lambda x: x.replace(" HTTP", "-HTTP"))
```

In [28]:
```
RDD6 = RDD5.map(lambda x: x.replace(" ", ","))
```

In [29]:
```
RDD7 = RDD6.map(lambda line: line.split(","))
```

In [30]:
```
RDD8 = RDD7.map(lambda p: Row(host = p[0], data = p[1], cod_http = p[3], qtd
_bytes = int(p[4])))
```

In [31]:
```
nasaDF = spSession.createDataFrame(RDD8)
```

In [32]:
```
nasaDF.show()
```

```
+--------+--------------------+--------------------+---------+
|cod_http|                data|                host|qtd_bytes|
+--------+--------------------+--------------------+---------+
|     200|01/Aug/1995:00:00...|   in24.inetnebr.com|     1839|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     1713|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     1687|
|     200|01/Aug/1995:00:00...|piweba4y.prodigy.com|    11853|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     9202|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     3635|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     1173|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     3047|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|        133.43.96.45|    10566|
|     200|01/Aug/1995:00:00...|kgtyk4.kj.yamagat...|     7280|
|     200|01/Aug/1995:00:00...|kgtyk4.kj.yamagat...|     5866|
|     200|01/Aug/1995:00:00...|     d0ucr6.fnal.gov|     2743|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     6849|
|     200|01/Aug/1995:00:00...|     d0ucr6.fnal.gov|    14897|
+--------+--------------------+--------------------+---------+
only showing top 20 rows
```

In [33]:
```
nasaDF.select("*").show()
```

```
+--------+--------------------+--------------------+---------+
|cod_http|                data|                host|qtd_bytes|
+--------+--------------------+--------------------+---------+
|     200|01/Aug/1995:00:00...|   in24.inetnebr.com|     1839|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     1713|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     1687|
|     200|01/Aug/1995:00:00...|piweba4y.prodigy.com|    11853|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     9202|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     3635|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     1173|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     3047|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|        133.43.96.45|    10566|
|     200|01/Aug/1995:00:00...|kgtyk4.kj.yamagat...|     7280|
|     200|01/Aug/1995:00:00...|kgtyk4.kj.yamagat...|     5866|
|     200|01/Aug/1995:00:00...|     d0ucr6.fnal.gov|     2743|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     6849|
|     200|01/Aug/1995:00:00...|     d0ucr6.fnal.gov|    14897|
+--------+--------------------+--------------------+---------+
only showing top 20 rows
```

In [34]:
```
nasaDF.createOrReplaceTempView("dadosNasa")
```

In [35]:
```
spSession.sql("select * from dadosNasa").show()
```

```
+--------+--------------------+--------------------+---------+
|cod_http|                data|                host|qtd_bytes|
+--------+--------------------+--------------------+---------+
|     200|01/Aug/1995:00:00...|   in24.inetnebr.com|     1839|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     1713|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     1687|
|     200|01/Aug/1995:00:00...|piweba4y.prodigy.com|    11853|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     9202|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     3635|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     1173|
|     200|01/Aug/1995:00:00...|slppp6.intermind.net|     3047|
|     304|01/Aug/1995:00:00...|     uplherc.upl.com|        0|
|     200|01/Aug/1995:00:00...|        133.43.96.45|    10566|
|     200|01/Aug/1995:00:00...|kgtyk4.kj.yamagat...|     7280|
|     200|01/Aug/1995:00:00...|kgtyk4.kj.yamagat...|     5866|
|     200|01/Aug/1995:00:00...|     d0ucr6.fnal.gov|     2743|
|     200|01/Aug/1995:00:00...|ix-esc-ca2-07.ix....|     6849|
|     200|01/Aug/1995:00:00...|     d0ucr6.fnal.gov|    14897|
+--------+--------------------+--------------------+---------+
only showing top 20 rows
```

In [38]:
```
spSession.sql("select count(1) as qtd_host_unicos from (select host, count
(1) qtd from dadosNasa group by host having count(1) = 1)").show()
```

```
+---------------+
|qtd_host_unicos|
+---------------+
|              4|
+---------------+
```

In [53]:
```
spSession.sql("select * from (select host, qtd_erros_404, rank()over(partiti
on by chv order by qtd_erros_404 desc) as rk from (select host, count(1) as
qtd_erros_404, 1 as chv from dadosNasa where cod_http = '200' group by hos
t)) where rk < 6").show()
```

```
+--------------------+-------------+---+
|                host|qtd_erros_404| rk|
+--------------------+-------------+---+
|     uplherc.upl.com|           16|  1|
|        133.43.96.45|           11|  2|
|www-d3.proxy.aol.com|           10|  3|
|piweba4y.prodigy.com|            8|  4|
|       133.68.18.180|            6|  5|
|ix-esc-ca2-07.ix....|            6|  5|
|slppp6.intermind.net|            6|  5|
+--------------------+-------------+---+
```

In [58]:
```
spSession.sql("select substr(data, 1, 11) data, count(1) from dadosNasa wher
e cod_http = '304' group by substr(data, 1, 11)").show()
```

```
+-----------+--------+
|       data|count(1)|
+-----------+--------+
|01/Aug/1995|      17|
+-----------+--------+
```

In [60]:
```
spSession.sql("select sum(qtd_bytes) total_bytes from dadosNasa").show()
```

```
+-----------+
|total_bytes|
+-----------+
|    1115275|
+-----------+
```

In [ ]:
```

```