Intitiating Spark Session & Loading Data

In [1]:
```
import findspark
findspark.init('/home/ubuntu/spark-2.4.5-bin-hadoop2.7')
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('kmean').getOrCreate()
```

In [2]:
```
df = spark.read.csv('hack_data.csv', header=True, inferSchema=True)
```

In [3]:
```
df.printSchema()
```

```
root
 |-- Session_Connection_Time: double (nullable = true)
 |-- Bytes Transferred: double (nullable = true)
 |-- Kali_Trace_Used: integer (nullable = true)
 |-- Servers_Corrupted: double (nullable = true)
 |-- Pages_Corrupted: double (nullable = true)
 |-- Location: string (nullable = true)
 |-- WPM_Typing_Speed: double (nullable = true)
```

In [4]:
```
for i in df.head(3):
    print(i,'\n')
```

```
Row(Session_Connection_Time=8.0, Bytes Transferred=391.09, Kali_Trace_Used=
1, Servers_Corrupted=2.96, Pages_Corrupted=7.0, Location='Slovenia', WPM_Ty
ping_Speed=72.37)

Row(Session_Connection_Time=20.0, Bytes Transferred=720.99, Kali_Trace_Used
=0, Servers_Corrupted=3.04, Pages_Corrupted=9.0, Location='British Virgin I
slands', WPM_Typing_Speed=69.08)

Row(Session_Connection_Time=31.0, Bytes Transferred=356.32, Kali_Trace_Used
=1, Servers_Corrupted=3.71, Pages_Corrupted=8.0, Location='Tokelau', WPM_Ty
ping_Speed=70.58)
```

In [5]:
```
df.show()
```

```
+----------------------+----------------+--------------+---------------
-+--------------+-------------------+---------------+
|Session_Connection_Time|Bytes Transferred|Kali_Trace_Used|Servers_Corrupte
d|Pages_Corrupted|            Location|WPM_Typing_Speed|
+----------------------+----------------+--------------+---------------
-+--------------+-------------------+---------------+
|                    8.0|          391.09|             1|             2.9
6|           7.0|           Slovenia|          72.37|
|                   20.0|          720.99|             0|             3.0
4|           9.0|British Virgin Is...|          69.08|
|                   31.0|          356.32|             1|             3.7
1|           8.0|             Tokelau|          70.58|
|                    2.0|          228.08|             1|             2.4
8|           8.0|             Bolivia|           70.8|
|                   20.0|           408.5|             0|             3.5
7|           8.0|                Iraq|          71.28|
|                    1.0|          390.69|             1|             2.7
9|           9.0|    Marshall Islands|          71.57|
|                   18.0|          342.97|             1|              5.
1|           7.0|             Georgia|          72.32|
|                   22.0|          101.61|             1|             3.0
3|           7.0|         Timor-Leste|          72.03|
|                   15.0|          275.53|             1|             3.5
3|           8.0|Palestinian Terri...|          70.17|
|                   12.0|          424.83|             1|             2.5
3|           8.0|          Bangladesh|          69.99|
|                   15.0|          249.09|             1|             3.3
9|           9.0|Northern Mariana ...|          70.77|
|                   32.0|          242.48|             0|             4.2
4|           8.0|            Zimbabwe|          67.93|
|                   23.0|          514.54|             0|             3.1
8|           8.0|         Isle of Man|          68.56|
|                    9.0|          284.77|             0|             3.1
2|           9.0|Sao Tome and Prin...|          70.82|
|                   27.0|          779.25|             1|             2.3
7|           8.0|              Greece|          72.73|
|                   12.0|          307.31|             1|             3.2
2|           7.0|     Solomon Islands|          67.95|
|                   21.0|          355.94|             1|              2.
0|           7.0|       Guinea-Bissau|           72.0|
|                   10.0|          372.65|             0|             3.3
3|           7.0|        Burkina Faso|          69.19|
|                   20.0|          347.23|             1|             2.3
3|           7.0|            Mongolia|          70.41|
|                   22.0|          456.57|             0|             1.5
2|           8.0|             Nigeria|          69.35|
+----------------------+----------------+--------------+---------------
-+--------------+-------------------+---------------+
only showing top 20 rows
```

Feature Engineering

In [6]:
```
df.columns
```

Out[6]:
```
['Session_Connection_Time',
 'Bytes Transferred',
 'Kali_Trace_Used',
 'Servers_Corrupted',
 'Pages_Corrupted',
 'Location',
 'WPM_Typing_Speed']
```

In [7]:
```
ession_Connection_Time', 'Bytes Transferred', 'Kali_Trace_Used', 'Servers_Cor
```

In [8]:
```
df.describe('Session_Connection_Time', 'Bytes Transferred', 'Kali_Trace_Used'
```

```
+-------+----------------------+------------------+------------------+
|summary|Session_Connection_Time| Bytes Transferred|   Kali_Trace_Used|
+-------+----------------------+------------------+------------------+
|  count|                   334|               334|               334|
|   mean|    30.008982035928145| 607.2452694610777|0.5119760479041916|
| stddev|    14.088200614636158|286.33593163576757|0.5006065264451406|
|    min|                   1.0|              10.0|                 0|
|    max|                  60.0|            1330.5|                 1|
+-------+----------------------+------------------+------------------+
```

In [9]:
```
df.describe('Servers_Corrupted', 'Pages_Corrupted', 'WPM_Typing_Speed').show(
```

```
+-------+-----------------+------------------+------------------+
|summary|Servers_Corrupted|   Pages_Corrupted|  WPM_Typing_Speed|
+-------+-----------------+------------------+------------------+
|  count|              334|               334|               334|
|   mean|5.258502994011977|10.838323353293413|57.342395209580864|
| stddev| 2.30190693339697|  3.06352633036022| 13.41106336843464|
|    min|              1.0|               6.0|              40.0|
|    max|             10.0|              15.0|              75.0|
+-------+-----------------+------------------+------------------+
```

In [10]:
```
from pyspark.ml.feature import VectorAssembler, StandardScaler
```

In [11]:
```
assembler = VectorAssembler(inputCols=['Session_Connection_Time', 'Bytes Tran
                            outputCol='features')
```

In [12]:
```
df = assembler.transform(df)
```

In [13]:
```
scale = StandardScaler(inputCol='features', outputCol='scaled_features', with
```

In [14]:
```
scaler = scale.fit(df)
```

In [15]:
```
df = scaler.transform(df)
```

In [16]:
```
df.show()
```

```
+----------------------+----------------+--------------+----------------
-+--------------+---------------+-------------------+------------------
-+
|Session_Connection_Time|Bytes Transferred|Kali_Trace_Used|Servers_Corrupte
d|Pages_Corrupted|WPM_Typing_Speed|            features|     scaled_feature
s|
+----------------------+----------------+--------------+----------------
-+--------------+---------------+-------------------+------------------
-+
|                    8.0|          391.09|             1|             2.9
6|           7.0|          72.37|[8.0,391.09,1.0,2...|[0.5678510846650
5...|
|                   20.0|          720.99|             0|             3.0
4|           9.0|          69.08|[20.0,720.99,0.0,...|[1.4196277116626
3...|
|                   31.0|          356.32|             1|             3.7
1|           8.0|          70.58|[31.0,356.32,1.0,...|[2.2004229530770
7...|
|                    2.0|          228.08|             1|             2.4
8|           8.0|           70.8|[2.0,228.08,1.0,2...|[0.1419627711662
6...|
|                   20.0|           408.5|             0|             3.5
7|           8.0|          71.28|[20.0,408.5,0.0,3...|[1.4196277116626
3...|
|                    1.0|          390.69|             1|             2.7
9|           9.0|          71.57|[1.0,390.69,1.0,2...|[0.0709813855831
3...|
|                   18.0|          342.97|             1|              5.
1|           7.0|          72.32|[18.0,342.97,1.0,...|[1.2776649404963
6...|
|                   22.0|          101.61|             1|             3.0
3|           7.0|          72.03|[22.0,101.61,1.0,...|[1.5615904828288
9...|
|                   15.0|          275.53|             1|             3.5
3|           8.0|          70.17|[15.0,275.53,1.0,...|[1.0647207837469
7...|
|                   12.0|          424.83|             1|             2.5
3|           8.0|          69.99|[12.0,424.83,1.0,...|[0.8517766269975
7...|
|                   15.0|          249.09|             1|             3.3
9|           9.0|          70.77|[15.0,249.09,1.0,...|[1.0647207837469
7...|
|                   32.0|          242.48|             0|             4.2
4|           8.0|          67.93|[32.0,242.48,0.0,...|[2.2714043386602
0...|
|                   23.0|          514.54|             0|             3.1
8|           8.0|          68.56|[23.0,514.54,0.0,...|[1.6325718684120
2...|
|                    9.0|          284.77|             0|             3.1
2|           9.0|          70.82|[9.0,284.77,0.0,3...|[0.6388324702481
8...|
|                   27.0|          779.25|             1|             2.3
7|           8.0|          72.73|[27.0,779.25,1.0,...|[1.9164974107445
5...|
```

```
|                  12.0|           307.31|               1|             3.2
2|            7.0|           67.95|[12.0,307.31,1.0,...|[0.8517766269975
7...|
|                  21.0|           355.94|               1|              2.
0|            7.0|            72.0|[21.0,355.94,1.0,...|[1.4906090972457
6...|
|                  10.0|           372.65|               0|             3.3
3|            7.0|           69.19|[10.0,372.65,0.0,...|[0.7098138558313
1...|
|                  20.0|           347.23|               1|             2.3
3|            7.0|           70.41|[20.0,347.23,1.0,...|[1.4196277116626
3...|
|                  22.0|           456.57|               0|             1.5
2|            8.0|           69.35|[22.0,456.57,0.0,...|[1.5615904828288
9...|
+----------------------+-----------------+---------------+----------------
-+---------------+----------------+--------------------+------------------
-+
only showing top 20 rows
```

Model Building

In [17]:
```
from pyspark.ml.clustering import KMeans
```

In [18]:
```
kmean2 = KMeans(featuresCol='scaled_features', k=2)
kmean3 = KMeans(featuresCol='scaled_features', k=3)
kmean4 = KMeans(featuresCol='scaled_features', k=4)
kmean5 = KMeans(featuresCol='scaled_features', k=5)
kmean6 = KMeans(featuresCol='scaled_features', k=6)
```

In [19]:
```
model_k2 = kmean2.fit(df)
model_k3 = kmean3.fit(df)
model_k4 = kmean4.fit(df)
model_k5 = kmean5.fit(df)
model_k6 = kmean6.fit(df)
```

In [20]:
```
result_k2 = model_k2.transform(df)
result_k3 = model_k3.transform(df)
result_k4 = model_k4.transform(df)
result_k5 = model_k5.transform(df)
result_k6 = model_k6.transform(df)
```

In [21]:
```
result_k2.groupBy('prediction').count().show()
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|  167|
|         0|  167|
+----------+-----+
```

In [22]:
```
result_k3.groupBy('prediction').count().show()
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   88|
|         2|   79|
|         0|  167|
+----------+-----+
```

In [23]:
```
result_k4.groupBy('prediction').count().show()
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   88|
|         3|   83|
|         2|   79|
|         0|   84|
+----------+-----+
```

In [24]:
```
result_k5.groupBy('prediction').count().show()
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   88|
|         3|   38|
|         4|   45|
|         2|   79|
|         0|   84|
+----------+-----+
```

In [25]:
```
result_k6.groupBy('prediction').count().show()
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   47|
|         3|   79|
|         5|   83|
|         4|   41|
|         2|   63|
|         0|   21|
+----------+-----+
```

The number shows that there were 2 or 4 hackers in the cyber attack