**Clustering Consulting Project**

**Summary**

***PROJECT DESCRIPTION:***

A large technology firm needs your help, they've been hacked! Luckily their forensic engineers have grabbed valuable data about the hacks, including information like session time, locations, wpm typing speed, etc. The forensic engineer relates to you what she has been able to figure out so far, she has been able to grab meta data of each session that the hackers used to connect to their servers. These are the features of the data:

* 'Session\_Connection\_Time': How long the session lasted in minutes
* 'Bytes Transferred': Number of MB transferred during session
* 'Kali\_Trace\_Used': Indicates if the hacker was using Kali Linux
* 'Servers\_Corrupted': Number of servers corrupted during the attack
* 'Pages\_Corrupted': Number of pages illegally accessed
* 'Location': Location attack came from (Probably useless because the hackers used VPNs)
* 'WPM\_Typing\_Speed': Their estimated typing speed based on session logs.

The technology firm has 3 potential hackers that perpetrated the attack. Their certain of the first two hackers but they aren't very sure if the third hacker was involved or not. They have requested your help! Can you help figure out whether or not the third suspect had anything to do with the attacks, or was it just two hackers? It's probably not possible to know for sure, but maybe what you've just learned about Clustering can help!

**One last key fact, the forensic engineer knows that the hackers trade off attacks. Meaning they should each have roughly the same amount of attacks. For example, if there were 100 total attacks, then in a 2-hacker situation each should have about 50 hacks, in a three-hacker situation each would have about 33 hacks. The engineer believes this is the key element to solving this, but doesn't know how to distinguish this unlabelled data into groups of hackers.**

***STEPS FOLLOWED IN THE PROJECT***

* installing pyspark
* importing necessary packages like SparkSession, PipeLine,Vectors,VectorAssembler,MultiClassClassificationEvaluator,mean etc.
* Created the spark session: A SparkSession can be used create DataFrame , register DataFrame as tables, execute SQL over tables, cache tables, and read parquet files.
* Assigning the path of the csv file to the variable named as 'data\_path'
* Read the csv file
* checking the total number of rows present in the csv file using count() function.
* checking the attributes present in the dataset using df.columns
* checking the datatypes of all the columns using df.dtypes.
* displaying the dataframe using df.show()
* describing the dataset using df.describe().show() method : It will give statistical values of the numerical datatypes
* displaying the schema of the dataset df.printSchema()
* Creating a user defined function named null\_value\_ count ,function use to print feature with null values and null count in each column of the dataset

**Conclusion:** *If there are two hackers(if k=2) then both of them were hacked equally that is 167 and 167.*

*If there are Three hackers(if k=3) then then probability of hacking is 1st hacker 167, 2nd hacker 84 and 3rd hacker is 83.*