Question 1

In [36]:
```
import findspark
findspark.init()
findspark.find()

import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import monotonically_increasing_id
from pyspark.sql.functions import lit

spark = SparkSession.builder.master("local[*]").appName("GuestLoginOverride").getOrCreat
```

In [37]:
```
col = ["duration","protocol_type","service","flag","src_bytes",
    "dst_bytes","land","wrong_fragment","urgent","hot","num_failed_logins",
    "logged_in","num_compromised","root_shell","su_attempted","num_root",
    "num_file_creations","num_shells","num_access_files","num_outbound_cmds",
    "is_host_login","is_guest_login","count","srv_count","serror_rate",
    "srv_serror_rate","rerror_rate","srv_rerror_rate","same_srv_rate",
    "diff_srv_rate","srv_diff_host_rate","dst_host_count","dst_host_srv_count",
    "dst_host_same_srv_rate","dst_host_diff_srv_rate","dst_host_same_src_port_rate",
    "dst_host_srv_diff_host_rate","dst_host_serror_rate","dst_host_srv_serror_rate",
    "dst_host_rerror_rate","dst_host_srv_rerror_rate","classes","difficulty_level"]

df = spark.read.csv("/Users/kiranprasadjp/Desktop/KDDTrain+.txt",header=False, inferSche
df.show(5,vertical=True)
```

```
-RECORD 0-------------------------------
 duration                     | 0
 protocol_type                | tcp
 service                      | ftp_data
 flag                         | SF
 src_bytes                    | 491
 dst_bytes                    | 0
 land                         | 0
 wrong_fragment               | 0
 urgent                       | 0
 hot                          | 0
 num_failed_logins            | 0
 logged_in                    | 0
 num_compromised              | 0
 root_shell                   | 0
 su_attempted                 | 0
 num_root                     | 0
 num_file_creations           | 0
 num_shells                   | 0
 num_access_files             | 0
 num_outbound_cmds            | 0
 is_host_login                | 0
 is_guest_login               | 0
 count                        | 2
 srv_count                    | 2
 serror_rate                  | 0.0
 srv_serror_rate              | 0.0
 rerror_rate                  | 0.0
 srv_rerror_rate              | 0.0
 same_srv_rate                | 1.0
 diff_srv_rate                | 0.0
 srv_diff_host_rate           | 0.0
 dst_host_count               | 150
 dst_host_srv_count           | 25
 dst_host_same_srv_rate       | 0.17
```

```
dst_host_diff_srv_rate      | 0.03
dst_host_same_src_port_rate | 0.17
dst_host_srv_diff_host_rate | 0.0
dst_host_serror_rate        | 0.0
dst_host_srv_serror_rate    | 0.0
dst_host_rerror_rate        | 0.05
dst_host_srv_rerror_rate    | 0.0
classes                     | normal
difficulty_level            | 20
-RECORD 1------------------------------
duration                    | 0
protocol_type               | udp
service                     | other
flag                        | SF
src_bytes                   | 146
dst_bytes                   | 0
land                        | 0
wrong_fragment              | 0
urgent                      | 0
hot                         | 0
num_failed_logins           | 0
logged_in                   | 0
num_compromised             | 0
root_shell                  | 0
su_attempted                | 0
num_root                    | 0
num_file_creations          | 0
num_shells                  | 0
num_access_files            | 0
num_outbound_cmds           | 0
is_host_login               | 0
is_guest_login              | 0
count                       | 13
srv_count                   | 1
serror_rate                 | 0.0
srv_serror_rate             | 0.0
rerror_rate                 | 0.0
srv_rerror_rate             | 0.0
same_srv_rate               | 0.08
diff_srv_rate               | 0.15
srv_diff_host_rate          | 0.0
dst_host_count              | 255
dst_host_srv_count          | 1
dst_host_same_srv_rate      | 0.0
dst_host_diff_srv_rate      | 0.6
dst_host_same_src_port_rate | 0.88
dst_host_srv_diff_host_rate | 0.0
dst_host_serror_rate        | 0.0
dst_host_srv_serror_rate    | 0.0
dst_host_rerror_rate        | 0.0
dst_host_srv_rerror_rate    | 0.0
classes                     | normal
difficulty_level            | 15
-RECORD 2------------------------------
duration                    | 0
protocol_type               | tcp
service                     | private
flag                        | S0
src_bytes                   | 0
dst_bytes                   | 0
land                        | 0
wrong_fragment              | 0
urgent                      | 0
hot                         | 0
num_failed_logins           | 0
logged_in                   | 0
```

```
num_compromised             | 0
root_shell                  | 0
su_attempted                | 0
num_root                    | 0
num_file_creations          | 0
num_shells                  | 0
num_access_files            | 0
num_outbound_cmds           | 0
is_host_login               | 0
is_guest_login              | 0
count                       | 123
srv_count                   | 6
serror_rate                 | 1.0
srv_serror_rate             | 1.0
rerror_rate                 | 0.0
srv_rerror_rate             | 0.0
same_srv_rate               | 0.05
diff_srv_rate               | 0.07
srv_diff_host_rate          | 0.0
dst_host_count              | 255
dst_host_srv_count          | 26
dst_host_same_srv_rate      | 0.1
dst_host_diff_srv_rate      | 0.05
dst_host_same_src_port_rate | 0.0
dst_host_srv_diff_host_rate | 0.0
dst_host_serror_rate        | 1.0
dst_host_srv_serror_rate    | 1.0
dst_host_rerror_rate        | 0.0
dst_host_srv_rerror_rate    | 0.0
classes                     | neptune
difficulty_level            | 19
-RECORD 3-------------------------------
duration                    | 0
protocol_type               | tcp
service                     | http
flag                        | SF
src_bytes                   | 232
dst_bytes                   | 8153
land                        | 0
wrong_fragment              | 0
urgent                      | 0
hot                         | 0
num_failed_logins           | 0
logged_in                   | 1
num_compromised             | 0
root_shell                  | 0
su_attempted                | 0
num_root                    | 0
num_file_creations          | 0
num_shells                  | 0
num_access_files            | 0
num_outbound_cmds           | 0
is_host_login               | 0
is_guest_login              | 0
count                       | 5
srv_count                   | 5
serror_rate                 | 0.2
srv_serror_rate             | 0.2
rerror_rate                 | 0.0
srv_rerror_rate             | 0.0
same_srv_rate               | 1.0
diff_srv_rate               | 0.0
srv_diff_host_rate          | 0.0
dst_host_count              | 30
dst_host_srv_count          | 255
dst_host_same_srv_rate      | 1.0
```

```
 dst_host_diff_srv_rate      | 0.0
 dst_host_same_src_port_rate | 0.03
 dst_host_srv_diff_host_rate | 0.04
 dst_host_serror_rate        | 0.03
 dst_host_srv_serror_rate    | 0.01
 dst_host_rerror_rate        | 0.0
 dst_host_srv_rerror_rate    | 0.01
 classes                     | normal
 difficulty_level            | 21
-RECORD 4-----------------------------
 duration                    | 0
 protocol_type               | tcp
 service                     | http
 flag                        | SF
 src_bytes                   | 199
 dst_bytes                   | 420
 land                        | 0
 wrong_fragment              | 0
 urgent                      | 0
 hot                         | 0
 num_failed_logins           | 0
 logged_in                   | 1
 num_compromised             | 0
 root_shell                  | 0
 su_attempted                | 0
 num_root                    | 0
 num_file_creations          | 0
 num_shells                  | 0
 num_access_files            | 0
 num_outbound_cmds           | 0
 is_host_login               | 0
 is_guest_login              | 0
 count                       | 30
 srv_count                   | 32
 serror_rate                 | 0.0
 srv_serror_rate             | 0.0
 rerror_rate                 | 0.0
 srv_rerror_rate             | 0.0
 same_srv_rate               | 1.0
 diff_srv_rate               | 0.0
 srv_diff_host_rate          | 0.09
 dst_host_count              | 255
 dst_host_srv_count          | 255
 dst_host_same_srv_rate      | 1.0
 dst_host_diff_srv_rate      | 0.0
 dst_host_same_src_port_rate | 0.0
 dst_host_srv_diff_host_rate | 0.0
 dst_host_serror_rate        | 0.0
 dst_host_srv_serror_rate    | 0.0
 dst_host_rerror_rate        | 0.0
 dst_host_srv_rerror_rate    | 0.0
 classes                     | normal
 difficulty_level            | 21
only showing top 5 rows
```

In [38]:

```
#initial
guest_df = df.select("protocol_type","is_guest_login")

guest_df.show()
```

```
+-------------+--------------+
|protocol_type|is_guest_login|
+-------------+--------------+
|          tcp|             0|
|          udp|             0|
```

```
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|           tcp|              0|
|          icmp|              0|
|           tcp|              0|
|           tcp|              0|
+--------------+---------------+
only showing top 20 rows
```

In [39]:
```
guest_traffic_df = df.filter(df.is_guest_login == 1).withColumn("protocol_type", lit("tc
guest_traffic_df.show(4,vertical=True)
```

```
-RECORD 0----------------------------------
 duration                     | 26
 protocol_type                | tcp
 service                      | ftp
 flag                         | SF
 src_bytes                    | 273
 dst_bytes                    | 903
 land                         | 0
 wrong_fragment               | 0
 urgent                       | 0
 hot                          | 5
 num_failed_logins            | 0
 logged_in                    | 1
 num_compromised              | 0
 root_shell                   | 0
 su_attempted                 | 0
 num_root                     | 0
 num_file_creations           | 0
 num_shells                   | 0
 num_access_files             | 0
 num_outbound_cmds            | 0
 is_host_login                | 0
 is_guest_login               | 1
 count                        | 1
 srv_count                    | 1
 serror_rate                  | 0.0
 srv_serror_rate              | 0.0
 rerror_rate                  | 0.0
 srv_rerror_rate              | 0.0
 same_srv_rate                | 1.0
 diff_srv_rate                | 0.0
 srv_diff_host_rate           | 0.0
 dst_host_count               | 176
 dst_host_srv_count           | 49
 dst_host_same_srv_rate       | 0.28
 dst_host_diff_srv_rate       | 0.02
 dst_host_same_src_port_rate  | 0.01
 dst_host_srv_diff_host_rate  | 0.0
 dst_host_serror_rate         | 0.01
 dst_host_srv_serror_rate     | 0.02
 dst_host_rerror_rate         | 0.0
```

```
 dst_host_srv_rerror_rate    | 0.0
 classes                     | normal
 difficulty_level            | 21
-RECORD 1----------------------------------
 duration                    | 15159
 protocol_type               | tcp
 service                     | ftp
 flag                        | SF
 src_bytes                   | 350
 dst_bytes                   | 1185
 land                        | 0
 wrong_fragment              | 0
 urgent                      | 0
 hot                         | 6
 num_failed_logins           | 0
 logged_in                   | 1
 num_compromised             | 0
 root_shell                  | 0
 su_attempted                | 0
 num_root                    | 0
 num_file_creations          | 0
 num_shells                  | 0
 num_access_files            | 0
 num_outbound_cmds           | 0
 is_host_login               | 0
 is_guest_login              | 1
 count                       | 1
 srv_count                   | 2
 serror_rate                 | 0.0
 srv_serror_rate             | 0.0
 rerror_rate                 | 0.0
 srv_rerror_rate             | 0.0
 same_srv_rate               | 1.0
 diff_srv_rate               | 0.0
 srv_diff_host_rate          | 1.0
 dst_host_count              | 255
 dst_host_srv_count          | 142
 dst_host_same_srv_rate      | 0.56
 dst_host_diff_srv_rate      | 0.02
 dst_host_same_src_port_rate | 0.0
 dst_host_srv_diff_host_rate | 0.0
 dst_host_serror_rate        | 0.0
 dst_host_srv_serror_rate    | 0.0
 dst_host_rerror_rate        | 0.0
 dst_host_srv_rerror_rate    | 0.0
 classes                     | warezclient
 difficulty_level            | 2
-RECORD 2----------------------------------
 duration                    | 1
 protocol_type               | tcp
 service                     | ftp
 flag                        | SF
 src_bytes                   | 1238
 dst_bytes                   | 2451
 land                        | 0
 wrong_fragment              | 0
 urgent                      | 0
 hot                         | 28
 num_failed_logins           | 0
 logged_in                   | 1
 num_compromised             | 0
 root_shell                  | 0
 su_attempted                | 0
 num_root                    | 0
 num_file_creations          | 0
 num_shells                  | 0
```

```
num_access_files            | 0
num_outbound_cmds           | 0
is_host_login               | 0
is_guest_login              | 1
count                       | 1
srv_count                   | 1
serror_rate                 | 0.0
srv_serror_rate             | 0.0
rerror_rate                 | 0.0
srv_rerror_rate             | 0.0
same_srv_rate               | 1.0
diff_srv_rate               | 0.0
srv_diff_host_rate          | 0.0
dst_host_count              | 255
dst_host_srv_count          | 123
dst_host_same_srv_rate      | 0.48
dst_host_diff_srv_rate      | 0.02
dst_host_same_src_port_rate | 0.0
dst_host_srv_diff_host_rate | 0.0
dst_host_serror_rate        | 0.0
dst_host_srv_serror_rate    | 0.0
dst_host_rerror_rate        | 0.01
dst_host_srv_rerror_rate    | 0.0
classes                     | warezclient
difficulty_level            | 15
-RECORD 3----------------------------------
duration                    | 30
protocol_type               | tcp
service                     | ftp
flag                        | SF
src_bytes                   | 1458
dst_bytes                   | 4152
land                        | 0
wrong_fragment              | 0
urgent                      | 0
hot                         | 30
num_failed_logins           | 0
logged_in                   | 1
num_compromised             | 0
root_shell                  | 0
su_attempted                | 0
num_root                    | 0
num_file_creations          | 0
num_shells                  | 0
num_access_files            | 0
num_outbound_cmds           | 0
is_host_login               | 0
is_guest_login              | 1
count                       | 1
srv_count                   | 1
serror_rate                 | 0.0
srv_serror_rate             | 0.0
rerror_rate                 | 0.0
srv_rerror_rate             | 0.0
same_srv_rate               | 1.0
diff_srv_rate               | 0.0
srv_diff_host_rate          | 0.0
dst_host_count              | 158
dst_host_srv_count          | 40
dst_host_same_srv_rate      | 0.25
dst_host_diff_srv_rate      | 0.03
dst_host_same_src_port_rate | 0.01
dst_host_srv_diff_host_rate | 0.0
dst_host_serror_rate        | 0.0
dst_host_srv_serror_rate    | 0.0
dst_host_rerror_rate        | 0.0
```

```
 dst_host_srv_rerror_rate    | 0.0
 classes                     | normal
 difficulty_level            | 20
only showing top 4 rows
```

In [40]:
```
# Select and show only the modified 'protocol_type' column
modified_protocol_df = guest_traffic_df.select("protocol_type","is_guest_login")

# Show a sample output of the modified 'protocol_type' column
modified_protocol_df.show()
```

```
+-------------+--------------+
|protocol_type|is_guest_login|
+-------------+--------------+
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
|          tcp|             1|
+-------------+--------------+
only showing top 20 rows
```

Question 2

In [43]:
```
logged_in_users = df.filter(df.logged_in == 1)  # Assuming 'logged_in' is 1 for logged-i
non_logged_in_users = df.filter(df.logged_in == 0)  # Assuming 'logged_in' is 0 for non-
```

In [44]:
```
logged_in_protocol_counts = logged_in_users.groupBy("protocol_type").count()

non_logged_in_protocol_counts = non_logged_in_users.groupBy("protocol_type").count()
```

In [45]:
```
print("Counts of protocol_type for logged-in users:")
logged_in_protocol_counts.show()

print("\nCounts of protocol_type for non-logged-in users:")
non_logged_in_protocol_counts.show()
```

```
Counts of protocol_type for logged-in users:
+-------------+-----+
|protocol_type|count|
+-------------+-----+
|          tcp|49852|
+-------------+-----+


Counts of protocol_type for non-logged-in users:
+-------------+-----+
```

```
|protocol_type|count|
+-------------+-----+
|          tcp|52837|
|          udp|14993|
|         icmp| 8291|
+-------------+-----+
```

In [ ]:

In [ ]:

In [ ]: