Part A

Q1 What is an Intrusion Detection System? Is it possible to implement an Intrusion Detection System on this dataset? Explain the workflow described in the paper for implementing the Intrusion Detection System.

Intrusion Detection System is a system used to detect if there is any intrusion exist in the system by analyzing the big data. It is designed for high volume, variety and high speed of data, it can monitor the system in an accurate and efficient process.

It is possible to apply IDS o this dataset. KDD99 is a high volume dataset. It has 41 attributes and the 'class' attributes which indicates whether the given instance is normal or attack.

Data preprocessing:

Remove any noisy from the data, and convert the categorical data to numerical data.

Standardization:

Apply the standardization to make sure all numerical data range within the scale.

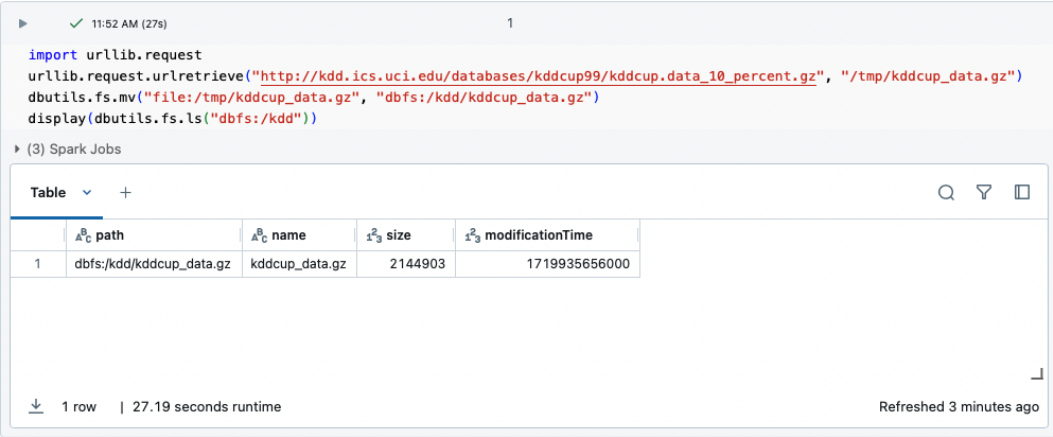
Feature selection:

Remove the features that are redundant or irrelevant from the data to ensure an efficient computation network

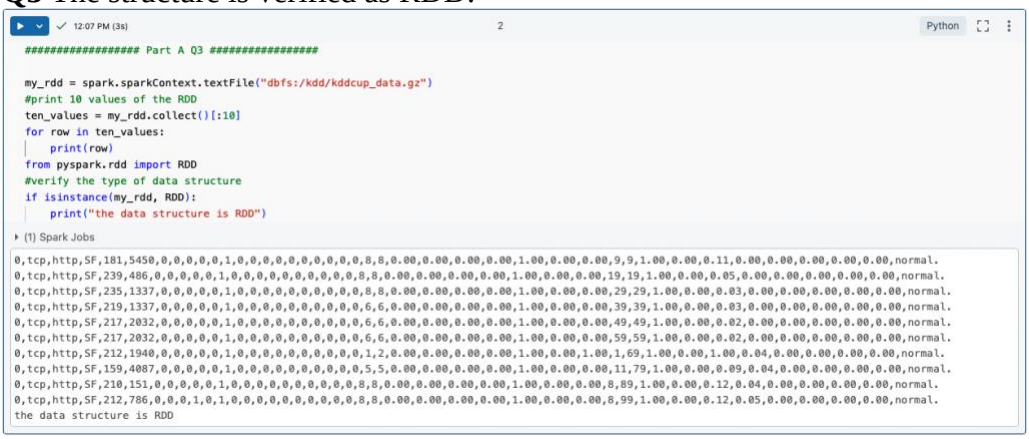
Model classifier:

Apply Support Vector Machine(SVM) to the data as a supervised learning method for classification.

Q2



Q3 The structure is verified as RDD.



Q4 There are in total 42 columns.

Just now (6s) 3 Python

```python
################## Part A Q4 #################
from pyspark.sql import Row

#define the column from http://kdd.ics.uci.edu/databases/kddcup99/kddcup.names
column_titles = ["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_hot_login", "is_guest_login", "count", "serror_rate", "rerror_rate", "same_srv_rate", "diff_srv_rate", "srv_count", "srv_serror_rate",
"srv_rerror_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","label"]

# Split the rows by the comma
split_data = my_rdd.map(lambda row: row.split(","))

split_df = split_data.toDF(column_titles)
print("The total number of columns are: " + str(len(split_df.columns)))
#print result
split_df.show()
```

(2) Spark Jobs

split_df: pyspark.sql.dataframe.DataFrame = [duration: string, protocol_type: string ... 40 more fields]

```
The total number of columns are: 42

|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|n
um_root|num_file_creations|num_shells|num_access_files|num_outbound_cmds|is_hot_login|is_guest_login|count|serror_rate|rerror_rate|same_srv_rate|diff_srv_rat
e|srv_count|srv_serror_rate|srv_rerror_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|   label|

|       0|          tcp|   http|  SF|      181|     5450|   0|             0|     0|  0|                0|        1|              0|         0|           0|
```

Q5 Extracted data and schema are shown below.

```python
################## Part A Q5 #################
df_6 = split_df[["duration", "protocol_type", "service", "src_bytes","dst_bytes", "flag" , "label"]]
df_6.printSchema()
df_6.show(10)
```

(1) Spark Jobs

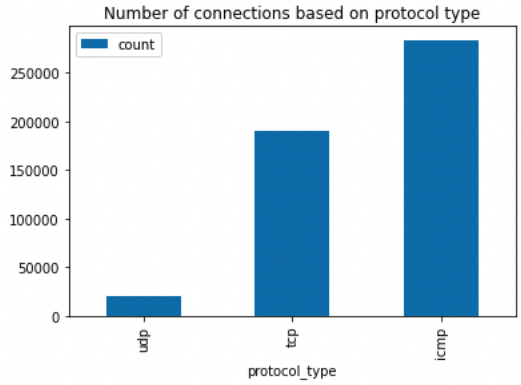
df_6: pyspark.sql.dataframe.DataFrame = [duration: string, protocol_type: string ... 5 more fields]

```
 |-- protocol_type: string (nullable = true)
 |-- service: string (nullable = true)
 |-- src_bytes: string (nullable = true)
 |-- dst_bytes: string (nullable = true)
 |-- flag: string (nullable = true)
 |-- label: string (nullable = true)
```

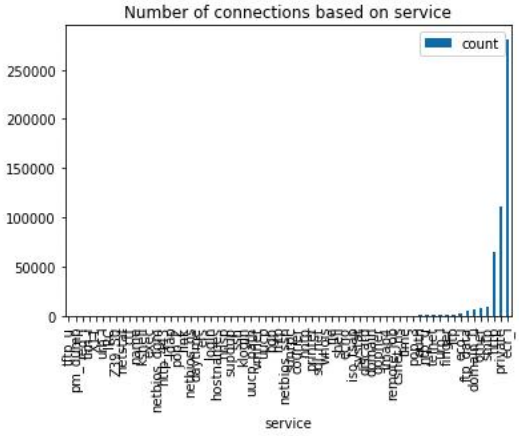
duration	protocol_type	service	src_bytes	dst_bytes	flag	label
0	tcp	http	181	5450	SF	normal.
0	tcp	http	239	486	SF	normal.
0	tcp	http	235	1337	SF	normal.
0	tcp	http	219	1337	SF	normal.
0	tcp	http	217	2032	SF	normal.
0	tcp	http	217	2032	SF	normal.
0	tcp	http	212	1940	SF	normal.
0	tcp	http	159	4087	SF	normal.
0	tcp	http	210	151	SF	normal.
0	tcp	http	212	786	SF	normal.

Q6

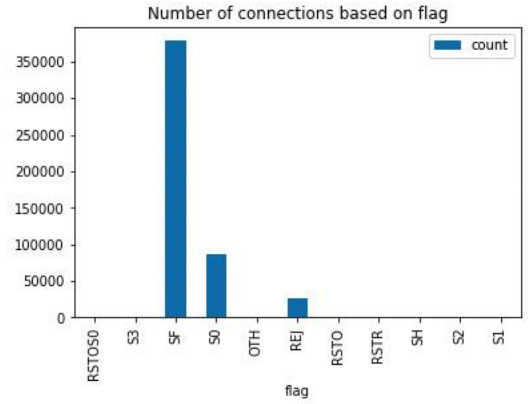
```
+-------------+------+
|protocol_type| count|
+-------------+------+
|          udp| 20354|
|          tcp|190065|
|         icmp|283602|
+-------------+------+
```

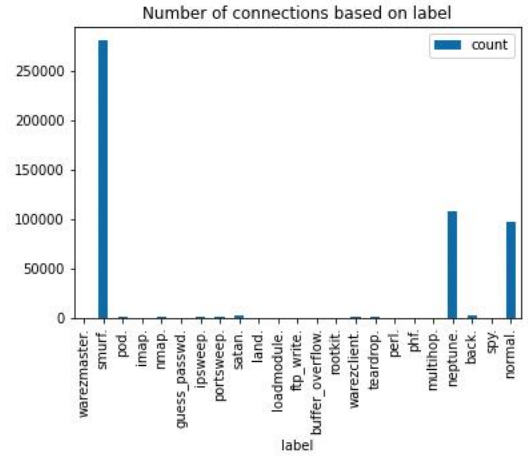


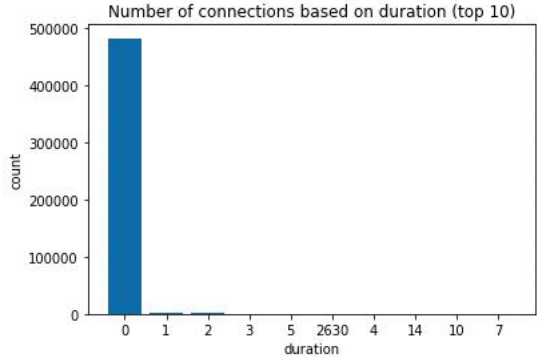

```
+-----------+-----+
|    service|count|
+-----------+-----+
|     tftp_u|    1|
|    pm_dump|    1|
|      red_i|    1|
|      tim_i|    7|
|        X11|   11|
|      urh_i|   14|
|        IRC|   43|
|     Z39_50|   92|
|    netstat|   95|
|        ctf|   97|
|       name|   98|
|     kshell|   98|
|       exec|   99|
|netbios_dgm|   99|
|   http_443|   99|
|       ldap|  101|
|      pop_2|  101|
|       link|  102|
| netbios_ns|  102|
|    daytime|  103|
+-----------+-----+
only showing top 20 rows
```



Q7 I plot the number of connections based on service, flag and duration(top 10)







Q8

I used SVM for the machine learning model. SVM is supervised and designed for classification, which suits our dataset the best(a simple binary classification).

To show the accuracy of my model, I simply calculate the number of correct prediction and divide it by the total number of data in the test set.

```
the model has accuracy:  0.9769717282015244
```

Part B

Q1.

1). True

PaaS provides a framework that developers can build upon to develop or customize cloud-based applications.

2). True

PaaS features scalability, high-availability, and multi-tenant capability.

Q2. D

Relational database is structured, so B and C are not related.

Relational database is static schema, so A is also wrong.

Q3. D

The provider will be responsible for A and B, and user can directly use the ability of configuration but do not need to install it.

Q4.

1). False

Company can start with a public cloud and then combine that with private cloud

2). True

We can use public cloud to extend the capacity.

3). False

It is not limited to guest user, if the user has been given account and authentication, he can access the resources in the cloud.

Q5.

a. Fault tolerance (tolerant to a failure)
b. Disaster recovery (recover from a failure)
c. Dynamic Scalability (deal with different demand)
d. Low latency (quick access)