جامعة الأخوين

AL AKHAWAYN

UNIVERSITY

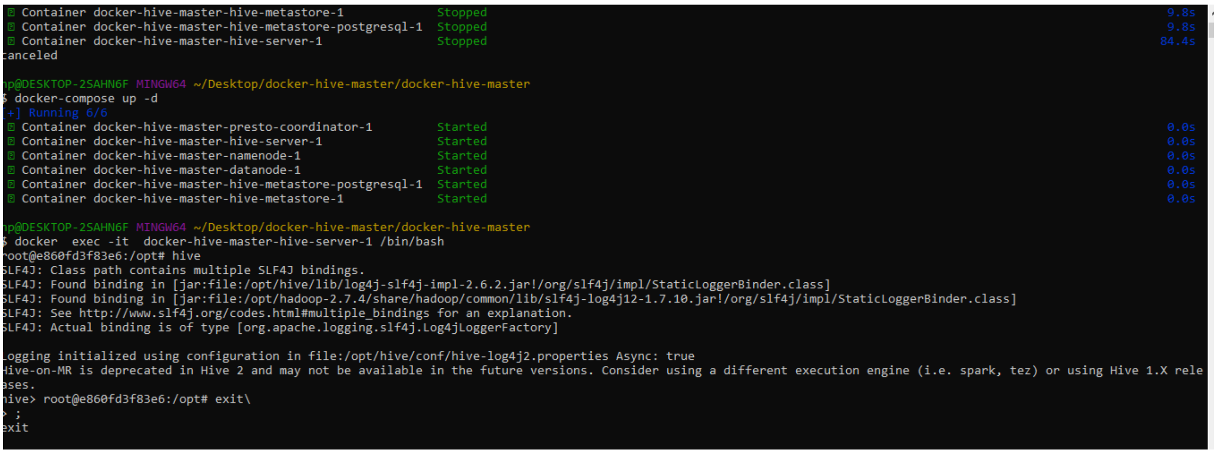
Project N°2

Intro to Big Data Environment

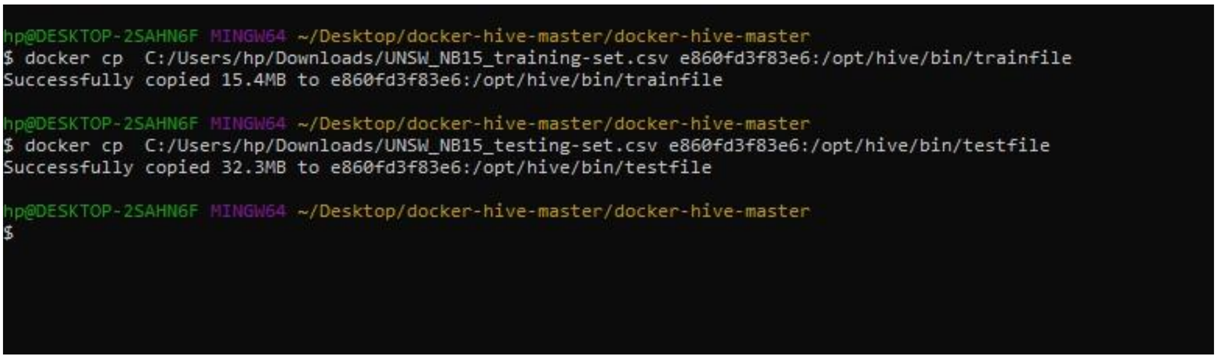
Fall 2023

1- **Download the Network traces in this link and load the into HDFS. Carefully explore it and understand each file of the dataset. Use directories if necessary.**

by executing the command hive:



```
 □ Container docker-hive-master-hive-metastore-1              Stopped                                                                   9.8s
 □ Container docker-hive-master-hive-metastore-postgresql-1   Stopped                                                                   9.8s
 □ Container docker-hive-master-hive-server-1                 Stopped                                                                  84.4s
anceled

np@DESKTOP-2SAHN6F MINGW64 ~/Desktop/docker-hive-master/docker-hive-master
$ docker-compose up -d
[+] Running 6/6
 □ Container docker-hive-master-presto-coordinator-1          Started                                                                   0.0s
 □ Container docker-hive-master-hive-server-1                 Started                                                                   0.0s
 □ Container docker-hive-master-namenode-1                    Started                                                                   0.0s
 □ Container docker-hive-master-datanode-1                    Started                                                                   0.0s
 □ Container docker-hive-master-hive-metastore-postgresql-1   Started                                                                   0.0s
 □ Container docker-hive-master-hive-metastore-1              Started                                                                   0.0s

np@DESKTOP-2SAHN6F MINGW64 ~/Desktop/docker-hive-master/docker-hive-master
$ docker  exec -it  docker-hive-master-hive-server-1 /bin/bash
root@e860fd3f83e6:/opt# hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rele
ases.
hive> root@e860fd3f83e6:/opt# exit\
> ;
exit
```

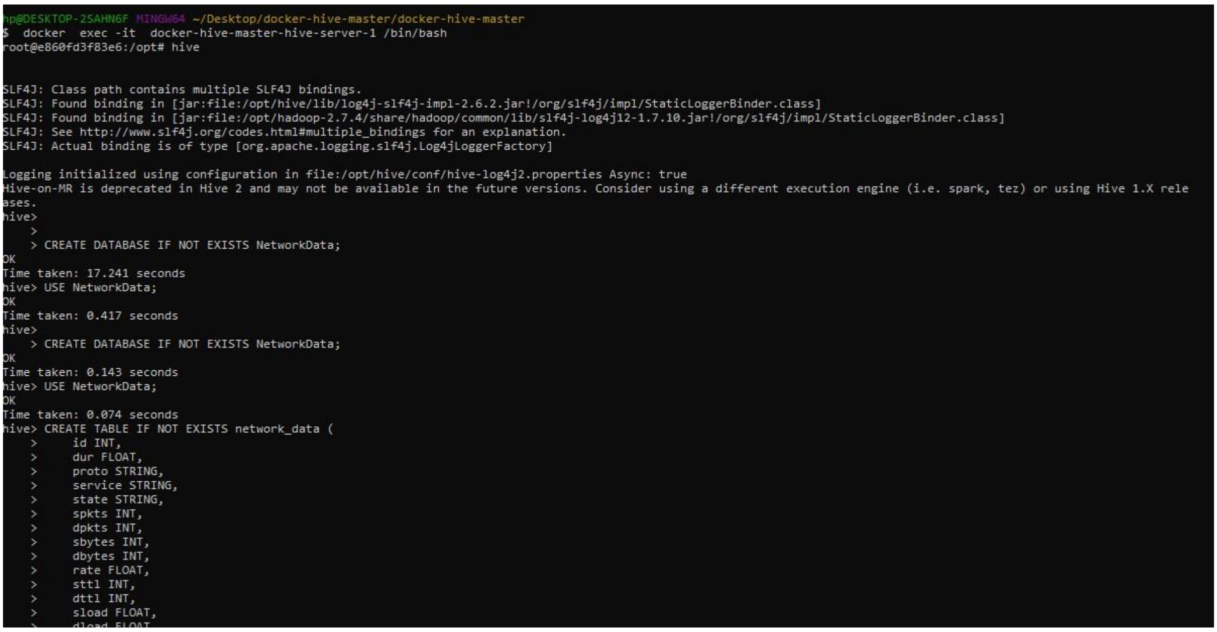


```
hp@DESKTOP-2SAHN6F MINGW64 ~/Desktop/docker-hive-master/docker-hive-master
$ docker cp  C:/Users/hp/Downloads/UNSW_NB15_training-set.csv e860fd3f83e6:/opt/hive/bin/trainfile
Successfully copied 15.4MB to e860fd3f83e6:/opt/hive/bin/trainfile

hp@DESKTOP-2SAHN6F MINGW64 ~/Desktop/docker-hive-master/docker-hive-master
$ docker cp  C:/Users/hp/Downloads/UNSW_NB15_testing-set.csv e860fd3f83e6:/opt/hive/bin/testfile
Successfully copied 32.3MB to e860fd3f83e6:/opt/hive/bin/testfile

hp@DESKTOP-2SAHN6F MINGW64 ~/Desktop/docker-hive-master/docker-hive-master
$
```

2- **Build an appropriate schema for ingesting the Network traces (separate training from test data), -include partitioning of the data**

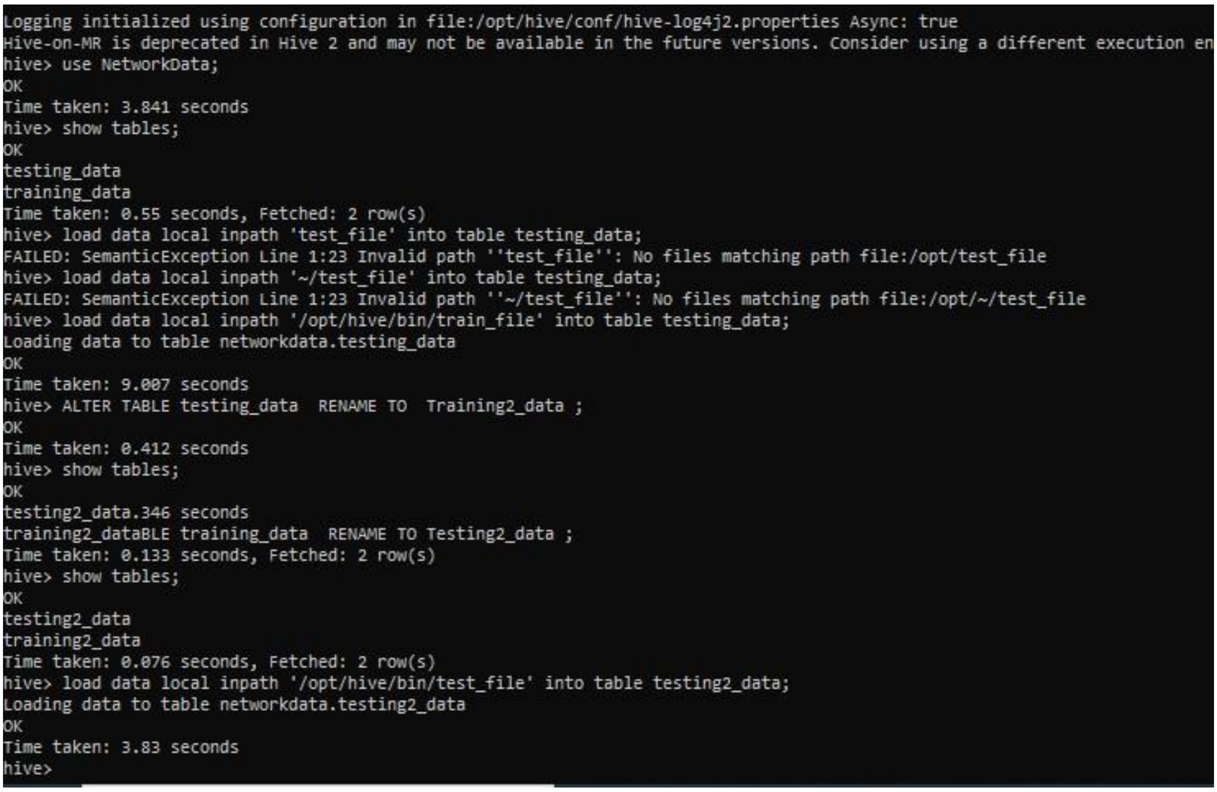
Create Database :

networkData



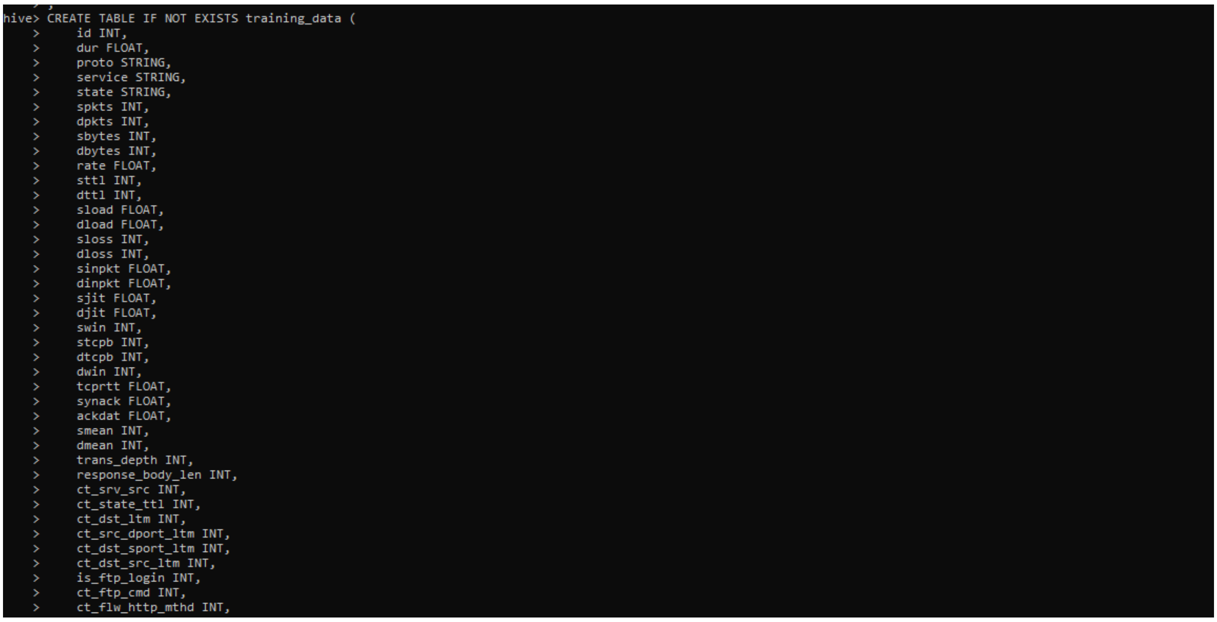
```
hp@DESKTOP-2SAHN6F MINGW64 ~/Desktop/docker-hive-master/docker-hive-master
$  docker  exec -it  docker-hive-master-hive-server-1 /bin/bash
root@e860fd3f83e6:/opt# hive


SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rele
ases.
hive>
    >
    > CREATE DATABASE IF NOT EXISTS NetworkData;
OK
Time taken: 17.241 seconds
hive> USE NetworkData;
OK
Time taken: 0.417 seconds
hive>
    > CREATE DATABASE IF NOT EXISTS NetworkData;
OK
Time taken: 0.143 seconds
hive> USE NetworkData;
OK
Time taken: 0.074 seconds
hive> CREATE TABLE IF NOT EXISTS network_data (
    >     id INT,
    >     dur FLOAT,
    >     proto STRING,
    >     service STRING,
    >     state STRING,
    >     spkts INT,
    >     dpkts INT,
    >     sbytes INT,
    >     dbytes INT,
    >     rate FLOAT,
    >     sttl INT,
    >     dttl INT,
    >     sload FLOAT,
    >     dload FLOAT
```

Create the table testingdata



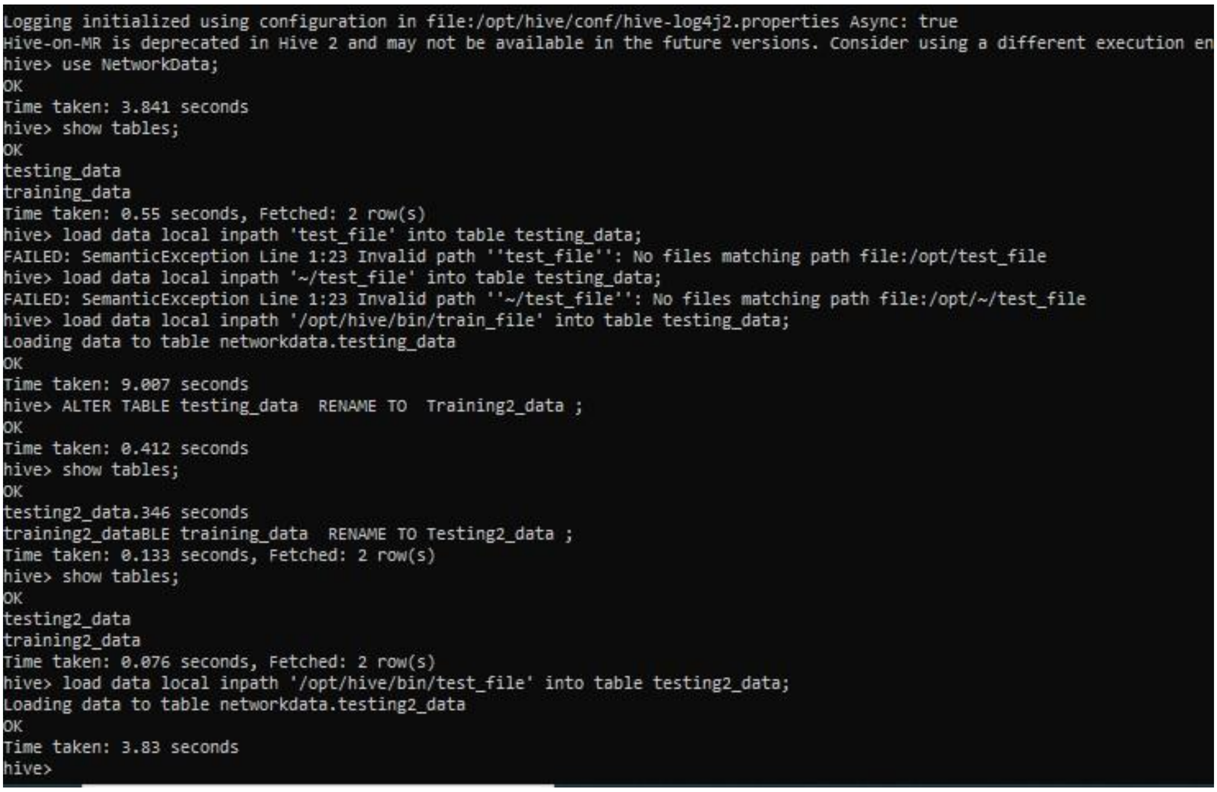
```
Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution en
hive> use NetworkData;
OK
Time taken: 3.841 seconds
hive> show tables;
OK
testing_data
training_data
Time taken: 0.55 seconds, Fetched: 2 row(s)
hive> load data local inpath 'test_file' into table testing_data;
FAILED: SemanticException Line 1:23 Invalid path ''test_file'': No files matching path file:/opt/test_file
hive> load data local inpath '~/test_file' into table testing_data;
FAILED: SemanticException Line 1:23 Invalid path ''~/test_file'': No files matching path file:/opt/~/test_file
hive> load data local inpath '/opt/hive/bin/train_file' into table testing_data;
Loading data to table networkdata.testing_data
OK
Time taken: 9.007 seconds
hive> ALTER TABLE testing_data  RENAME TO  Training2_data ;
OK
Time taken: 0.412 seconds
hive> show tables;
OK
testing2_data.346 seconds
training2_dataBLE training_data  RENAME TO Testing2_data ;
Time taken: 0.133 seconds, Fetched: 2 row(s)
hive> show tables;
OK
testing2_data
training2_data
Time taken: 0.076 seconds, Fetched: 2 row(s)
hive> load data local inpath '/opt/hive/bin/test_file' into table testing2_data;
Loading data to table networkdata.testing2_data
OK
Time taken: 3.83 seconds
hive>
```

Create the table trainingdata



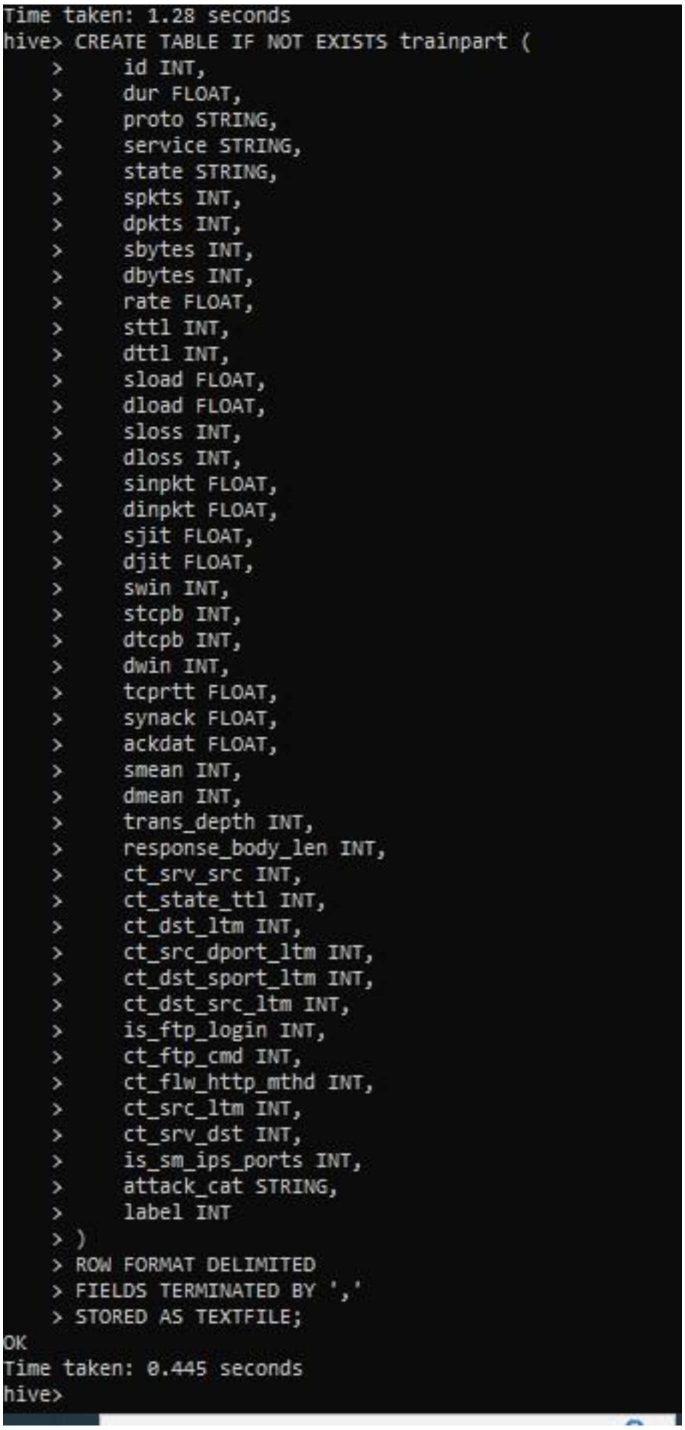
```
    > ;
hive> CREATE TABLE IF NOT EXISTS training_data (
    >     id INT,
    >     dur FLOAT,
    >     proto STRING,
    >     service STRING,
    >     state STRING,
    >     spkts INT,
    >     dpkts INT,
    >     sbytes INT,
    >     dbytes INT,
    >     rate FLOAT,
    >     sttl INT,
    >     dttl INT,
    >     sload FLOAT,
    >     dload FLOAT,
    >     sloss INT,
    >     dloss INT,
    >     sinpkt FLOAT,
    >     dinpkt FLOAT,
    >     sjit FLOAT,
    >     djit FLOAT,
    >     swin INT,
    >     stcpb INT,
    >     dtcpb INT,
    >     dwin INT,
    >     tcprtt FLOAT,
    >     synack FLOAT,
    >     ackdat FLOAT,
    >     smean INT,
    >     dmean INT,
    >     trans_depth INT,
    >     response_body_len INT,
    >     ct_srv_src INT,
    >     ct_state_ttl INT,
    >     ct_dst_ltm INT,
    >     ct_src_dport_ltm INT,
    >     ct_dst_sport_ltm INT,
    >     ct_dst_src_ltm INT,
    >     is_ftp_login INT,
    >     ct_ftp_cmd INT,
    >     ct_flw_http_mthd INT,
```

Load the data:

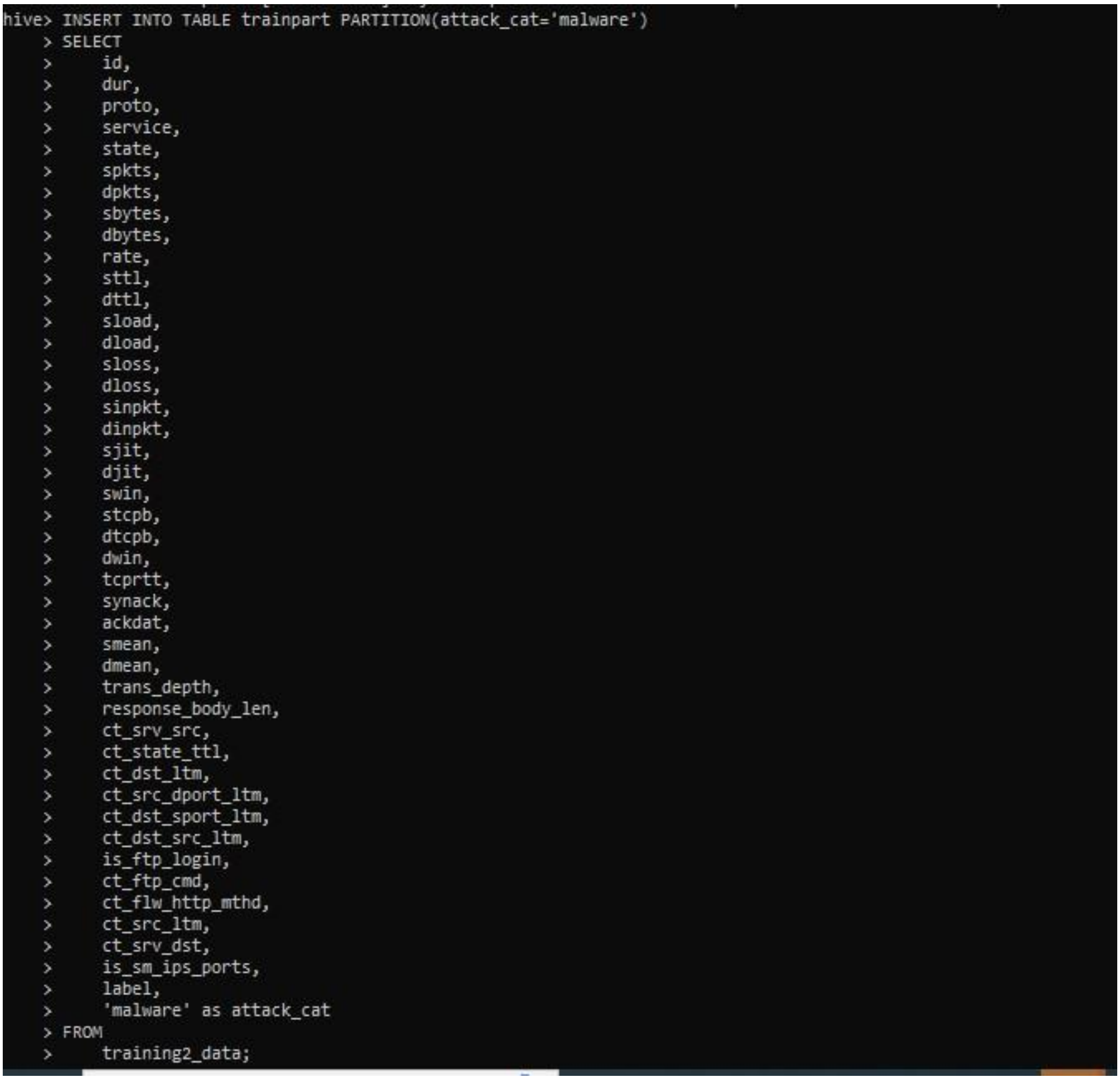


```
Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution en
hive> use NetworkData;
OK
Time taken: 3.841 seconds
hive> show tables;
OK
testing_data
training_data
Time taken: 0.55 seconds, Fetched: 2 row(s)
hive> load data local inpath 'test_file' into table testing_data;
FAILED: SemanticException Line 1:23 Invalid path ''test_file'': No files matching path file:/opt/test_file
hive> load data local inpath '~/test_file' into table testing_data;
FAILED: SemanticException Line 1:23 Invalid path ''~/test_file'': No files matching path file:/opt/~/test_file
hive> load data local inpath '/opt/hive/bin/train_file' into table testing_data;
Loading data to table networkdata.testing_data
OK
Time taken: 9.007 seconds
hive> ALTER TABLE testing_data  RENAME TO  Training2_data ;
OK
Time taken: 0.412 seconds
hive> show tables;
OK
testing2_data.346 seconds
training2_dataBLE training_data  RENAME TO Testing2_data ;
Time taken: 0.133 seconds, Fetched: 2 row(s)
hive> show tables;
OK
testing2_data
training2_data
Time taken: 0.076 seconds, Fetched: 2 row(s)
hive> load data local inpath '/opt/hive/bin/test_file' into table testing2_data;
Loading data to table networkdata.testing2_data
OK
Time taken: 3.83 seconds
hive>
```

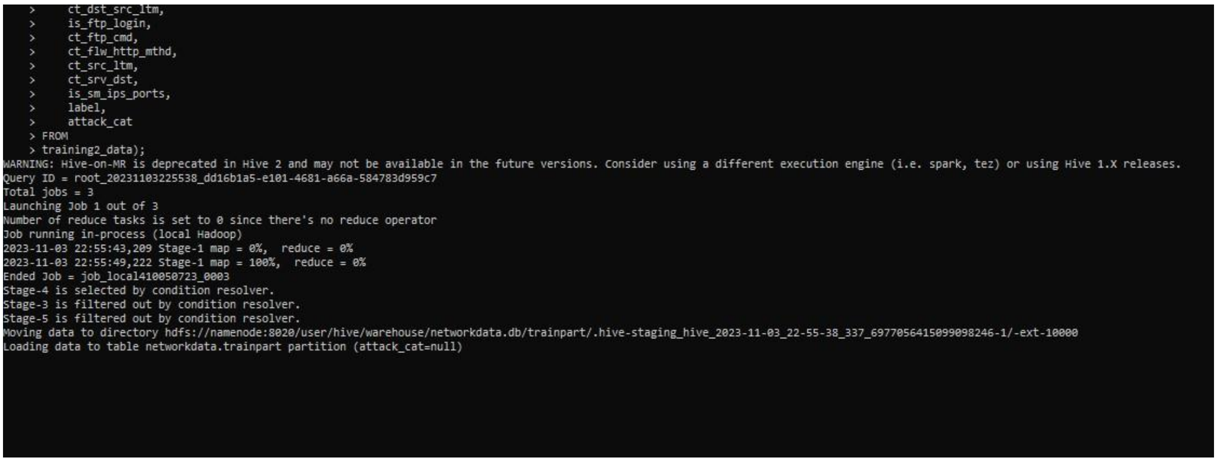
Partitioning and checking:



```
Time taken: 1.28 seconds
hive> CREATE TABLE IF NOT EXISTS trainpart (
    >     id INT,
    >     dur FLOAT,
    >     proto STRING,
    >     service STRING,
    >     state STRING,
    >     spkts INT,
    >     dpkts INT,
    >     sbytes INT,
    >     dbytes INT,
    >     rate FLOAT,
    >     sttl INT,
    >     dttl INT,
    >     sload FLOAT,
    >     dload FLOAT,
    >     sloss INT,
    >     dloss INT,
    >     sinpkt FLOAT,
    >     dinpkt FLOAT,
    >     sjit FLOAT,
    >     djit FLOAT,
    >     swin INT,
    >     stcpb INT,
    >     dtcpb INT,
    >     dwin INT,
    >     tcprtt FLOAT,
    >     synack FLOAT,
    >     ackdat FLOAT,
    >     smean INT,
    >     dmean INT,
    >     trans_depth INT,
    >     response_body_len INT,
    >     ct_srv_src INT,
    >     ct_state_ttl INT,
    >     ct_dst_ltm INT,
    >     ct_src_dport_ltm INT,
    >     ct_dst_sport_ltm INT,
    >     ct_dst_src_ltm INT,
    >     is_ftp_login INT,
    >     ct_ftp_cmd INT,
    >     ct_flw_http_mthd INT,
    >     ct_src_ltm INT,
    >     ct_srv_dst INT,
    >     is_sm_ips_ports INT,
    >     attack_cat STRING,
    >     label INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.445 seconds
hive>
```

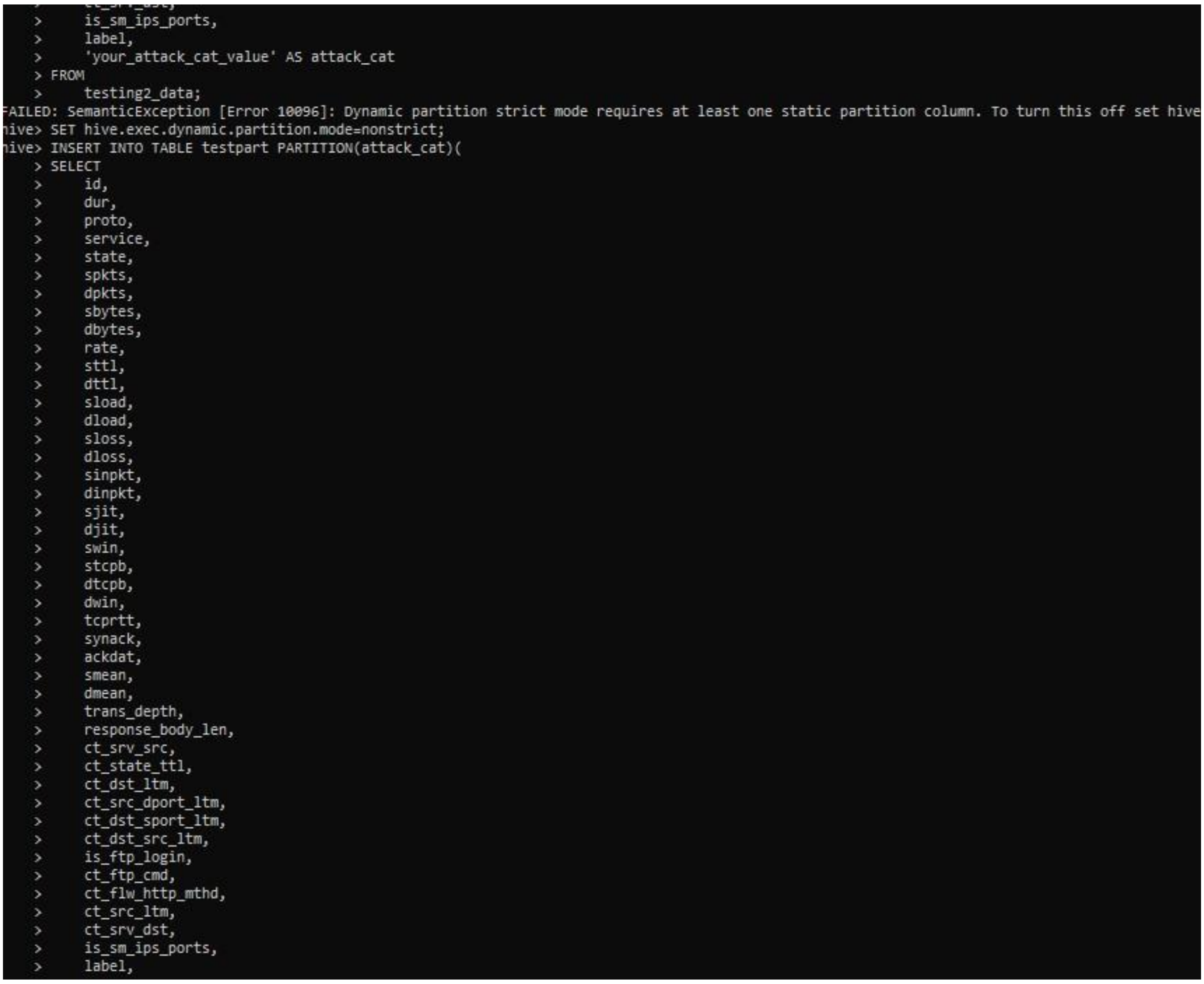


```
hive> INSERT INTO TABLE trainpart PARTITION(attack_cat='malware')
    > SELECT
    >     id,
    >     dur,
    >     proto,
    >     service,
    >     state,
    >     spkts,
    >     dpkts,
    >     sbytes,
    >     dbytes,
    >     rate,
    >     sttl,
    >     dttl,
    >     sload,
    >     dload,
    >     sloss,
    >     dloss,
    >     sinpkt,
    >     dinpkt,
    >     sjit,
    >     djit,
    >     swin,
    >     stcpb,
    >     dtcpb,
    >     dwin,
    >     tcprtt,
    >     synack,
    >     ackdat,
    >     smean,
    >     dmean,
    >     trans_depth,
    >     response_body_len,
    >     ct_srv_src,
    >     ct_state_ttl,
    >     ct_dst_ltm,
    >     ct_src_dport_ltm,
    >     ct_dst_sport_ltm,
    >     ct_dst_src_ltm,
    >     is_ftp_login,
    >     ct_ftp_cmd,
    >     ct_flw_http_mthd,
    >     ct_src_ltm,
    >     ct_srv_dst,
    >     is_sm_ips_ports,
    >     label,
    >     'malware' as attack_cat
    > FROM
    >     training2_data;
```

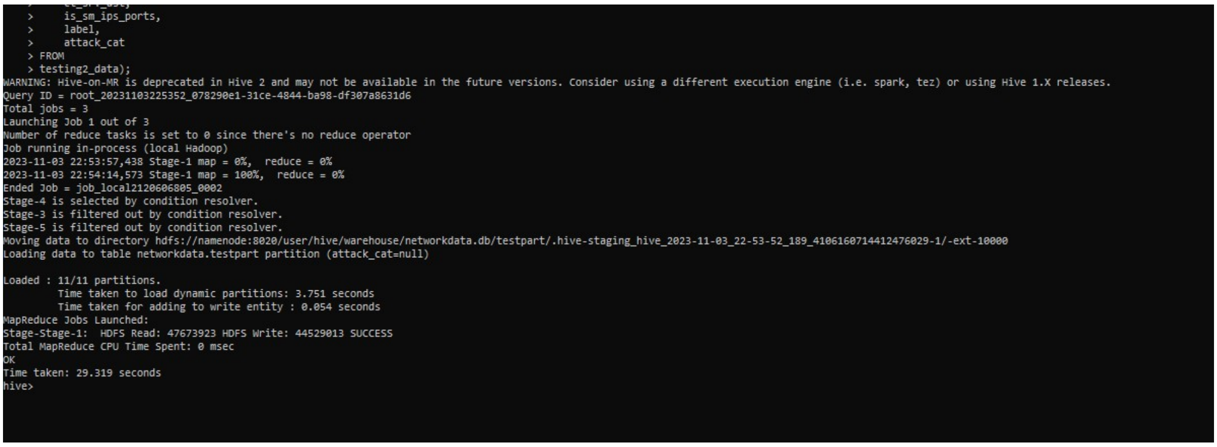


```
    >     ct_dst_src_ltm,
    >     is_ftp_login,
    >     ct_ftp_cmd,
    >     ct_flw_http_mthd,
    >     ct_src_ltm,
    >     ct_srv_dst,
    >     is_sm_ips_ports,
    >     label,
    >     attack_cat
    > FROM
    > training2_data);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103225538_dd16b1a5-e101-4681-a66a-584783d959c7
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2023-11-03 22:55:43,209 Stage-1 map = 0%,  reduce = 0%
2023-11-03 22:55:49,222 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local410050723_0003
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://namenode:8020/user/hive/warehouse/networkdata.db/trainpart/.hive-staging_hive_2023-11-03_22-55-38_337_6977056415099098246-1/-ext-10000
Loading data to table networkdata.trainpart partition (attack_cat=null)
```

Testing partitioning:



```
    >     is_sm_ips_ports,
    >     label,
    >     'your_attack_cat_value' AS attack_cat
    > FROM
    >     testing2_data;
FAILED: SemanticException [Error 10096]: Dynamic partition strict mode requires at least one static partition column. To turn this off set hive
hive> SET hive.exec.dynamic.partition.mode=nonstrict;
hive> INSERT INTO TABLE testpart PARTITION(attack_cat)(
    > SELECT
    >     id,
    >     dur,
    >     proto,
    >     service,
    >     state,
    >     spkts,
    >     dpkts,
    >     sbytes,
    >     dbytes,
    >     rate,
    >     sttl,
    >     dttl,
    >     sload,
    >     dload,
    >     sloss,
    >     dloss,
    >     sinpkt,
    >     dinpkt,
    >     sjit,
    >     djit,
    >     swin,
    >     stcpb,
    >     dtcpb,
    >     dwin,
    >     tcprtt,
    >     synack,
    >     ackdat,
    >     smean,
    >     dmean,
    >     trans_depth,
    >     response_body_len,
    >     ct_srv_src,
    >     ct_state_ttl,
    >     ct_dst_ltm,
    >     ct_src_dport_ltm,
    >     ct_dst_sport_ltm,
    >     ct_dst_src_ltm,
    >     is_ftp_login,
    >     ct_ftp_cmd,
    >     ct_flw_http_mthd,
    >     ct_src_ltm,
    >     ct_srv_dst,
    >     is_sm_ips_ports,
    >     label,
```

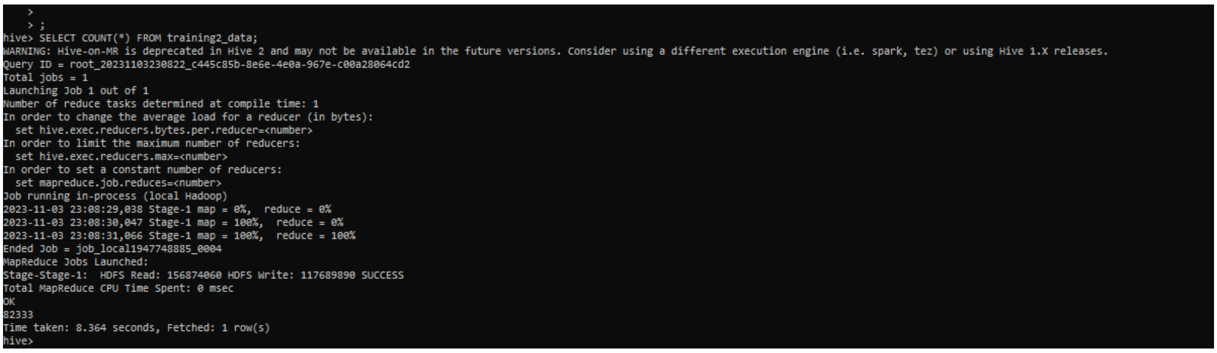


```
    >     is_sm_ips_ports,
    >     label,
    >     attack_cat
    > FROM
    > testing2_data);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103225352_078290e1-31ce-4844-ba98-df307a8631d6
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2023-11-03 22:53:57,438 Stage-1 map = 0%,  reduce = 0%
2023-11-03 22:54:14,573 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local2120606805_0002
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://namenode:8020/user/hive/warehouse/networkdata.db/testpart/.hive-staging_hive_2023-11-03_22-53-52_189_4106160714412476029-1/-ext-10000
Loading data to table networkdata.testpart partition (attack_cat=null)

Loaded : 11/11 partitions.
         Time taken to load dynamic partitions: 3.751 seconds
         Time taken for adding to write entity : 0.054 seconds
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 47673923 HDFS Write: 44529013 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 29.319 seconds
hive>
```

3. Write HQL queries to confirm the various statistics of the dataset

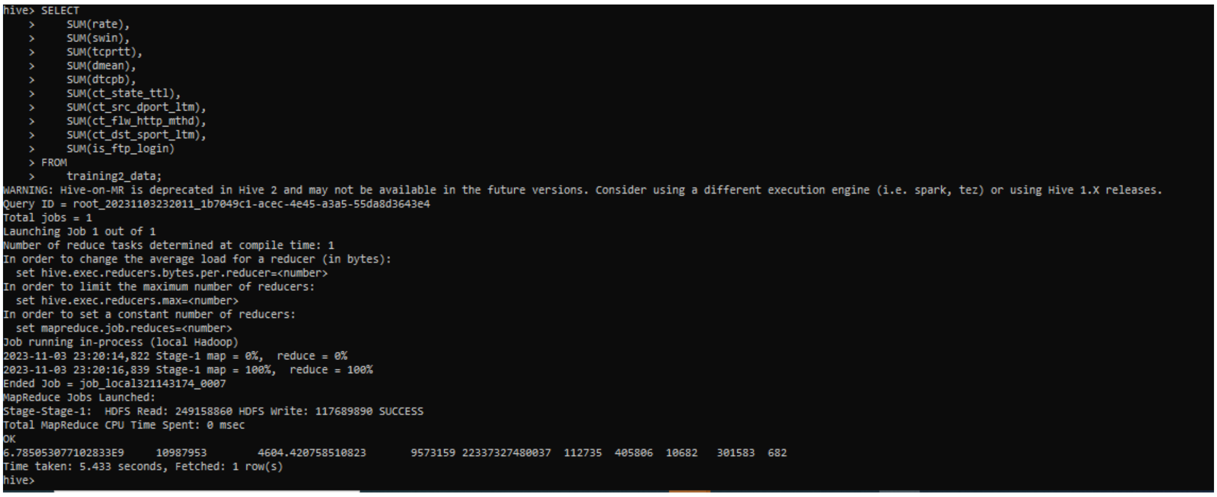
Statistics:

Count:



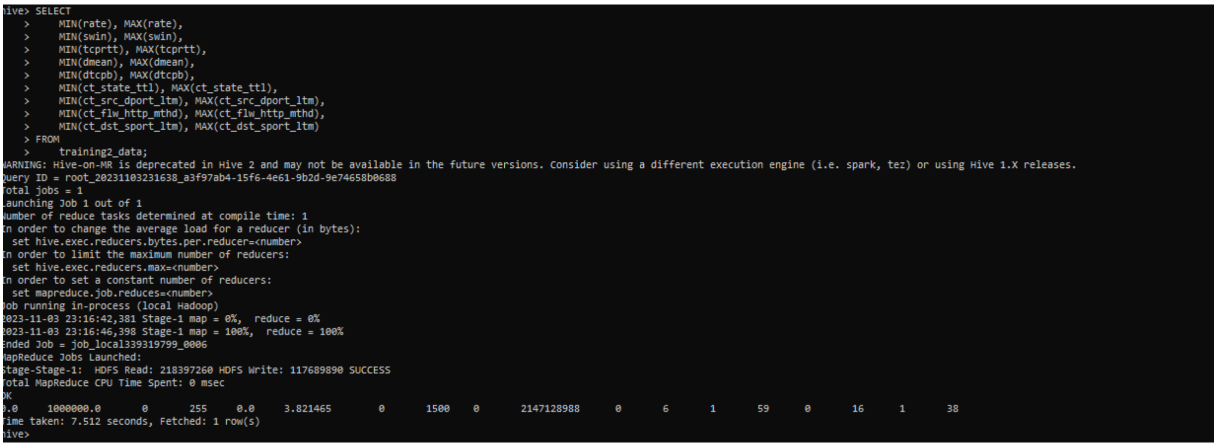
```
    >
    > ;
hive> SELECT COUNT(*) FROM training2_data;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103230822_c445c85b-8e6e-4e0a-967e-c00a28064cd2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:08:29,038 Stage-1 map = 0%,  reduce = 0%
2023-11-03 23:08:30,047 Stage-1 map = 100%,  reduce = 0%
2023-11-03 23:08:31,066 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1947748885_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 156874060 HDFS Write: 117689890 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
82333
Time taken: 8.364 seconds, Fetched: 1 row(s)
hive>
```

Sum:



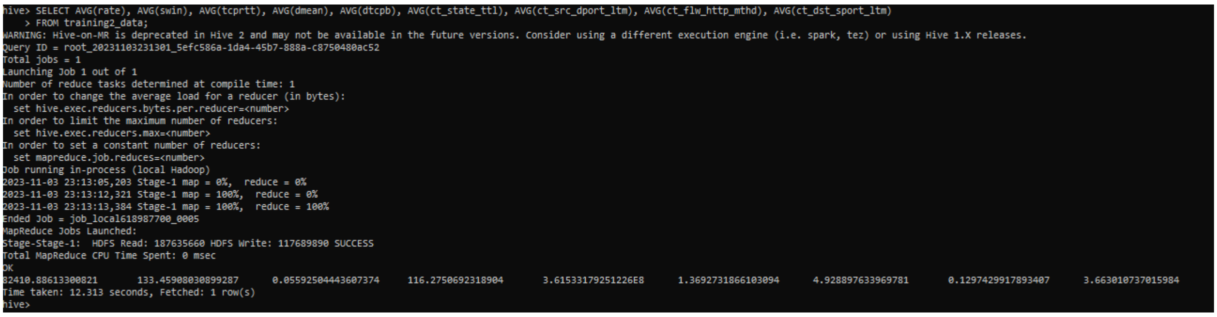
```
hive> SELECT
    >     SUM(rate),
    >     SUM(swin),
    >     SUM(tcprtt),
    >     SUM(dmean),
    >     SUM(dtcpb),
    >     SUM(ct_state_ttl),
    >     SUM(ct_src_dport_ltm),
    >     SUM(ct_flw_http_mthd),
    >     SUM(ct_dst_sport_ltm),
    >     SUM(is_ftp_login)
    > FROM
    >     training2_data;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103232011_1b7049c1-acec-4e45-a3a5-55da8d3643e4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:20:14,822 Stage-1 map = 0%,  reduce = 0%
2023-11-03 23:20:16,839 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local321143174_0007
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 249158860 HDFS Write: 117689890 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
6.785053077102833E9     10987953        4604.420758510823       9573159 22337327480037  112735  405806  10682   301583  682
Time taken: 5.433 seconds, Fetched: 1 row(s)
hive>
```

Min/max



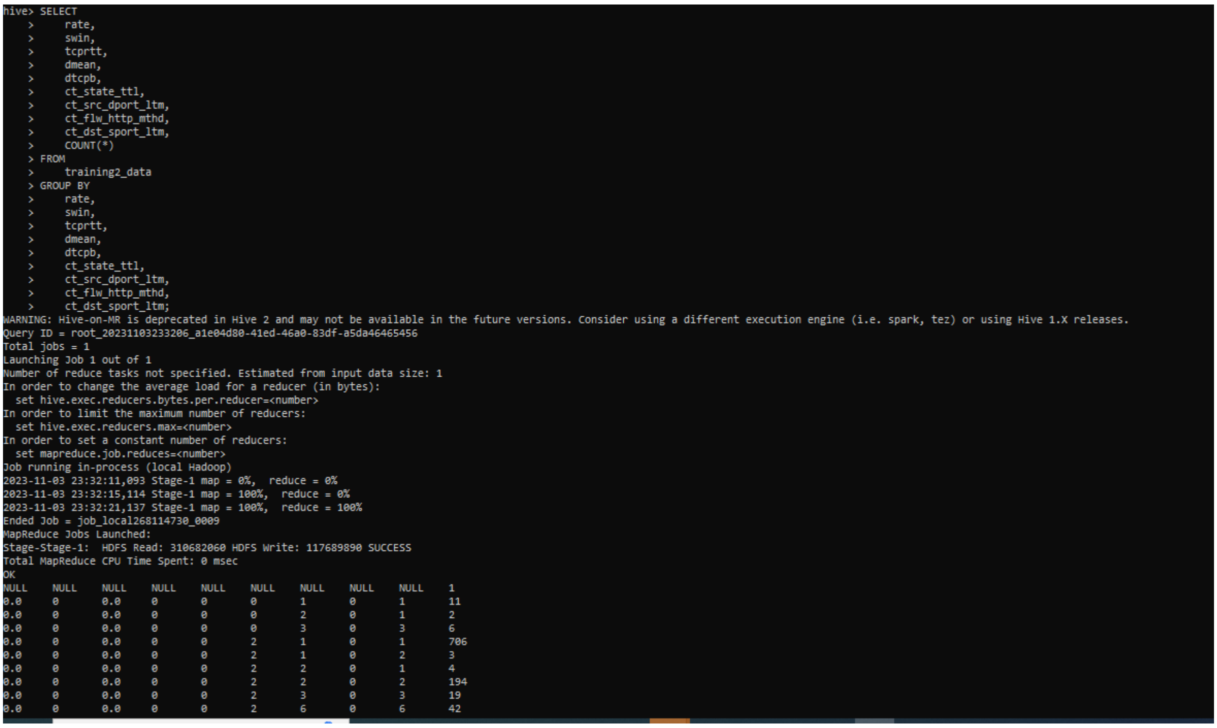
```
ive> SELECT
   >     MIN(rate), MAX(rate),
   >     MIN(swin), MAX(swin),
   >     MIN(tcprtt), MAX(tcprtt),
   >     MIN(dmean), MAX(dmean),
   >     MIN(dtcpb), MAX(dtcpb),
   >     MIN(ct_state_ttl), MAX(ct_state_ttl),
   >     MIN(ct_src_dport_ltm), MAX(ct_src_dport_ltm),
   >     MIN(ct_flw_http_mthd), MAX(ct_flw_http_mthd),
   >     MIN(ct_dst_sport_ltm), MAX(ct_dst_sport_ltm)
   > FROM
   >     training2_data;
ARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
uery ID = root_20231103231638_a3f97ab4-15f6-4e61-9b2d-9e74658b0688
otal jobs = 1
aunching Job 1 out of 1
umber of reduce tasks determined at compile time: 1
n order to change the average load for a reducer (in bytes):
 set hive.exec.reducers.bytes.per.reducer=<number>
n order to limit the maximum number of reducers:
 set hive.exec.reducers.max=<number>
n order to set a constant number of reducers:
 set mapreduce.job.reduces=<number>
ob running in-process (local Hadoop)
023-11-03 23:16:42,381 Stage-1 map = 0%,  reduce = 0%
023-11-03 23:16:46,398 Stage-1 map = 100%,  reduce = 100%
nded Job = job_local339319799_0006
apReduce Jobs Launched:
tage-Stage-1:  HDFS Read: 218397260 HDFS Write: 117689890 SUCCESS
otal MapReduce CPU Time Spent: 0 msec
K
.0     1000000.0       0       255     0.0     3.821465        0       1500    0       2147128988      0       6       1       59      0       16      1       38
ime taken: 7.512 seconds, Fetched: 1 row(s)
ive>
```

Average:



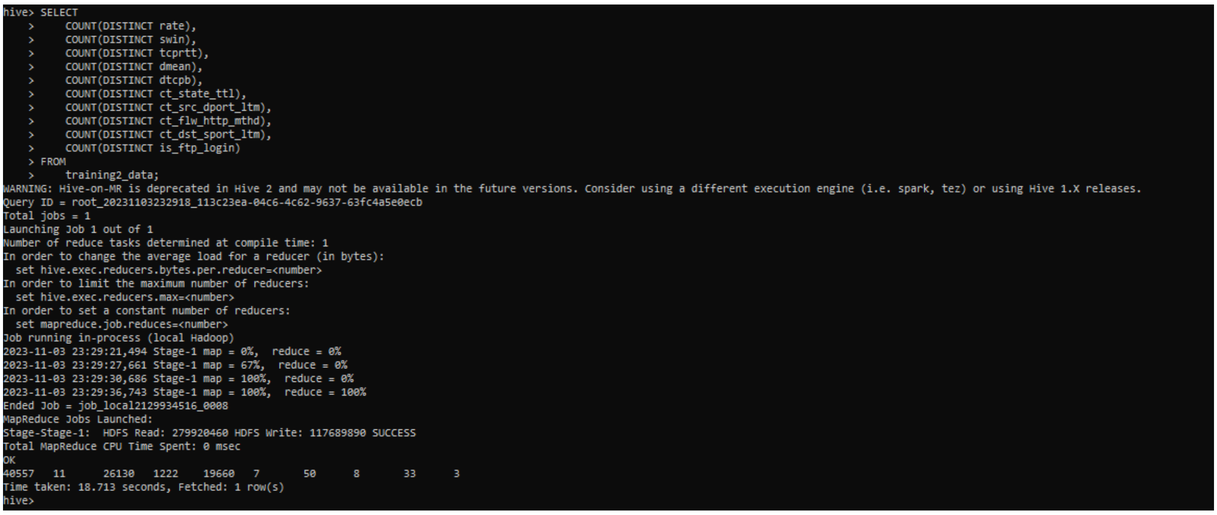
```
hive> SELECT AVG(rate), AVG(swin), AVG(tcprtt), AVG(dmean), AVG(dtcpb), AVG(ct_state_ttl), AVG(ct_src_dport_ltm), AVG(ct_flw_http_mthd), AVG(ct_dst_sport_ltm)
    > FROM training2_data;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103231301_5efc586a-1da4-45b7-888a-c8750480ac52
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:13:05,203 Stage-1 map = 0%,  reduce = 0%
2023-11-03 23:13:12,321 Stage-1 map = 100%,  reduce = 0%
2023-11-03 23:13:13,384 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local618987700_0005
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 187635660 HDFS Write: 117689890 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
82410.88613300821       133.45908030899287      0.05592504443607374     116.2750692318904       3.61533179251226E8      1.3692731866103094      4.928897633969781       0.1297429917893407      3.663010737015984
Time taken: 12.313 seconds, Fetched: 1 row(s)
hive>
```

Groupby:



```
hive> SELECT
    >     rate,
    >     swin,
    >     tcprtt,
    >     dmean,
    >     dtcpb,
    >     ct_state_ttl,
    >     ct_src_dport_ltm,
    >     ct_flw_http_mthd,
    >     ct_dst_sport_ltm,
    >     COUNT(*)
    > FROM
    >     training2_data
    > GROUP BY
    >     rate,
    >     swin,
    >     tcprtt,
    >     dmean,
    >     dtcpb,
    >     ct_state_ttl,
    >     ct_src_dport_ltm,
    >     ct_flw_http_mthd,
    >     ct_dst_sport_ltm;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103233206_a1e04d80-41ed-46a0-83df-a5da46465456
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:32:11,093 Stage-1 map = 0%,  reduce = 0%
2023-11-03 23:32:15,114 Stage-1 map = 100%,  reduce = 0%
2023-11-03 23:32:21,137 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local268114730_0009
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 310682060 HDFS Write: 117689890 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    1
0.0     0       0.0     0       0       0       1       0       1       11
0.0     0       0.0     0       0       0       2       0       1       2
0.0     0       0.0     0       0       0       3       0       3       6
0.0     0       0.0     0       0       2       1       0       1       706
0.0     0       0.0     0       0       2       1       0       2       3
0.0     0       0.0     0       0       2       2       0       1       4
0.0     0       0.0     0       0       2       2       0       2       194
0.0     0       0.0     0       0       2       3       0       3       19
0.0     0       0.0     0       0       2       6       0       6       42
```

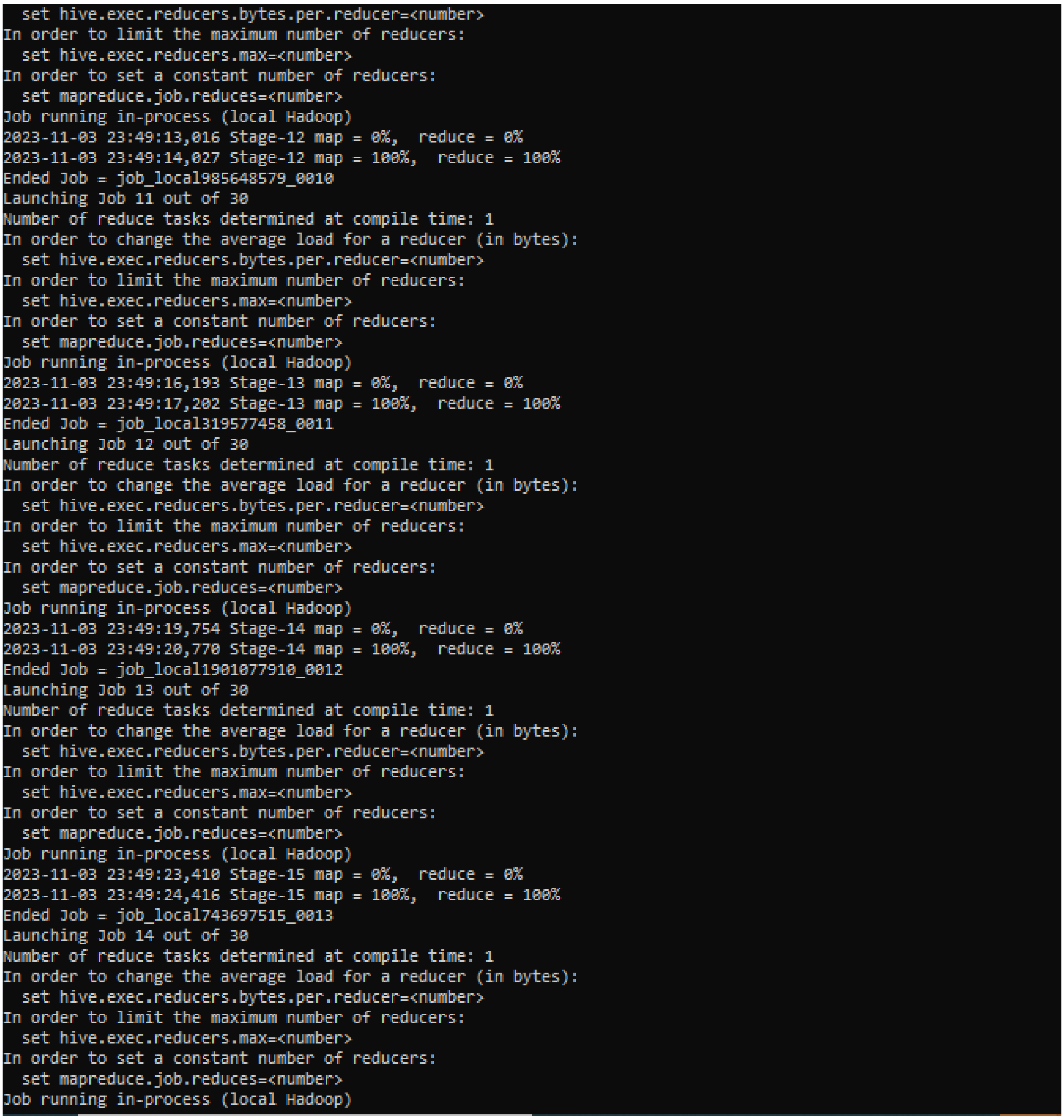
Count distinct:



```
hive> SELECT
    >     COUNT(DISTINCT rate),
    >     COUNT(DISTINCT swin),
    >     COUNT(DISTINCT tcprtt),
    >     COUNT(DISTINCT dmean),
    >     COUNT(DISTINCT dtcpb),
    >     COUNT(DISTINCT ct_state_ttl),
    >     COUNT(DISTINCT ct_src_dport_ltm),
    >     COUNT(DISTINCT ct_flw_http_mthd),
    >     COUNT(DISTINCT ct_dst_sport_ltm),
    >     COUNT(DISTINCT is_ftp_login)
    > FROM
    >     training2_data;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231103232918_113c23ea-04c6-4c62-9637-63fc4a5e0ecb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:29:21,494 Stage-1 map = 0%,  reduce = 0%
2023-11-03 23:29:27,661 Stage-1 map = 67%,  reduce = 0%
2023-11-03 23:29:30,686 Stage-1 map = 100%,  reduce = 0%
2023-11-03 23:29:36,743 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local2129934516_0008
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 279920460 HDFS Write: 117689890 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
40557   11      26130   1222    19660   7       50      8       33      3
Time taken: 18.713 seconds, Fetched: 1 row(s)
hive>
```

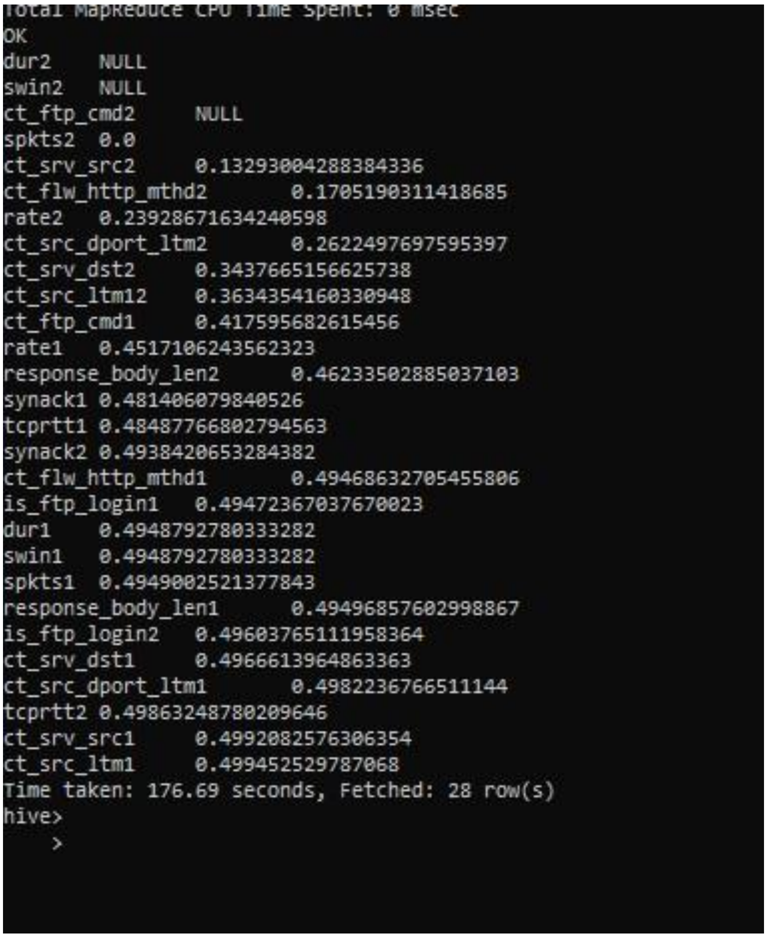
4. Write the HQL queries for computing the Gini impurity for each feature. Split the data according to the best (least impurity) feature.

To calculate the Gini impurity to define the root of our decision tree, we need to query the element from each column and calculate its probabilities given the label is equal to “0” (no attack) or “1” (attack).



```
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:49:13,016 Stage-12 map = 0%,  reduce = 0%
2023-11-03 23:49:14,027 Stage-12 map = 100%,  reduce = 100%
Ended Job = job_local985648579_0010
Launching Job 11 out of 30
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:49:16,193 Stage-13 map = 0%,  reduce = 0%
2023-11-03 23:49:17,202 Stage-13 map = 100%,  reduce = 100%
Ended Job = job_local319577458_0011
Launching Job 12 out of 30
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:49:19,754 Stage-14 map = 0%,  reduce = 0%
2023-11-03 23:49:20,770 Stage-14 map = 100%,  reduce = 100%
Ended Job = job_local1901077910_0012
Launching Job 13 out of 30
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-03 23:49:23,410 Stage-15 map = 0%,  reduce = 0%
2023-11-03 23:49:24,416 Stage-15 map = 100%,  reduce = 100%
Ended Job = job_local743697515_0013
Launching Job 14 out of 30
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
```

As a results the minimum values are: is



```
Total MapReduce CPU Time Spent: 0 msec
OK
dur2    NULL
swin2   NULL
ct_ftp_cmd2     NULL
spkts2  0.0
ct_srv_src2     0.13293004288384336
ct_flw_http_mthd2       0.1705190311418685
rate2   0.23928671634240598
ct_src_dport_ltm2       0.2622497697595397
ct_srv_dst2     0.3437665156625738
ct_src_ltm12    0.3634354160330948
ct_ftp_cmd1     0.417595682615456
rate1   0.4517106243562323
response_body_len2      0.46233502885037103
synack1 0.481406079840526
tcprtt1 0.48487766802794563
synack2 0.4938420653284382
ct_flw_http_mthd1       0.49468632705455806
is_ftp_login1   0.49472367037670023
dur1    0.4948792780333282
swin1   0.4948792780333282
spkts1  0.4949002521377843
response_body_len1      0.49496857602998867
is_ftp_login2   0.49603765111958364
ct_srv_dst1     0.4966613964863363
ct_src_dport_ltm1       0.4982236766511144
tcprtt2 0.49863248780209646
ct_srv_src1     0.4992082576306354
ct_src_ltm1     0.499452529787068
Time taken: 176.69 seconds, Fetched: 28 row(s)
hive>
    >
```

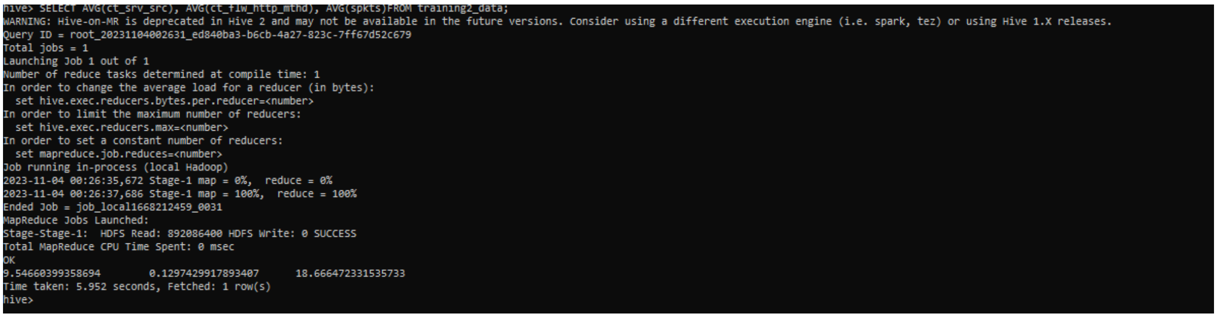
Based on these result we notice that the minimized values and the e minimum impurities are :

0 for spkts

0.17 for Ct-flw-http-mthd

0.13 of ct-src-src

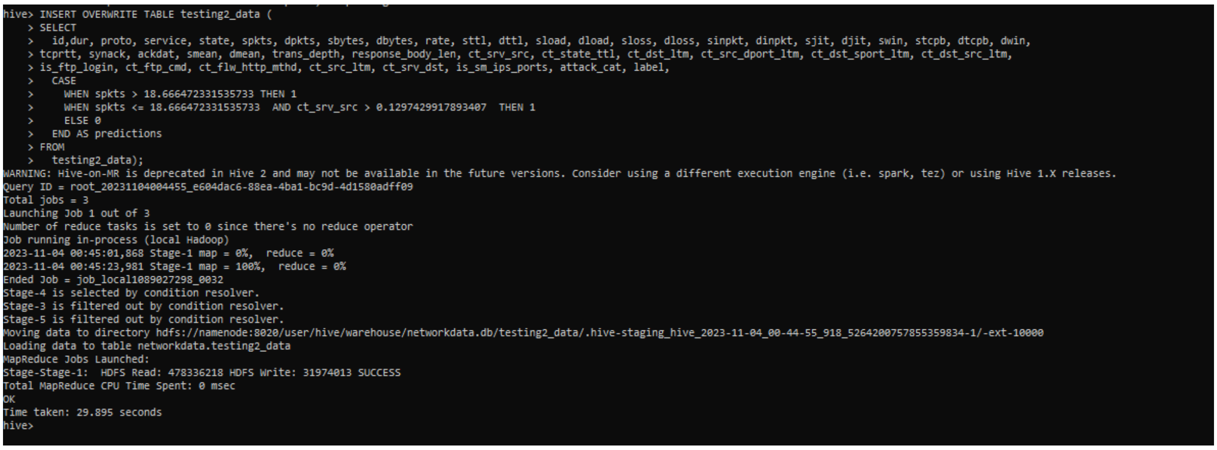
We use those attribute in the decision tree:



```
hive> SELECT AVG(ct_srv_src), AVG(ct_flw_http_mthd), AVG(spkts)FROM training2_data;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231104002631_ed840ba3-b6cb-4a27-823c-7ff67d52c679
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-04 00:26:35,672 Stage-1 map = 0%,  reduce = 0%
2023-11-04 00:26:37,686 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1668212459_0031
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 892086400 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
9.54660399358694        0.1297429917893407      18.666472331535733
Time taken: 5.952 seconds, Fetched: 1 row(s)
hive>
```

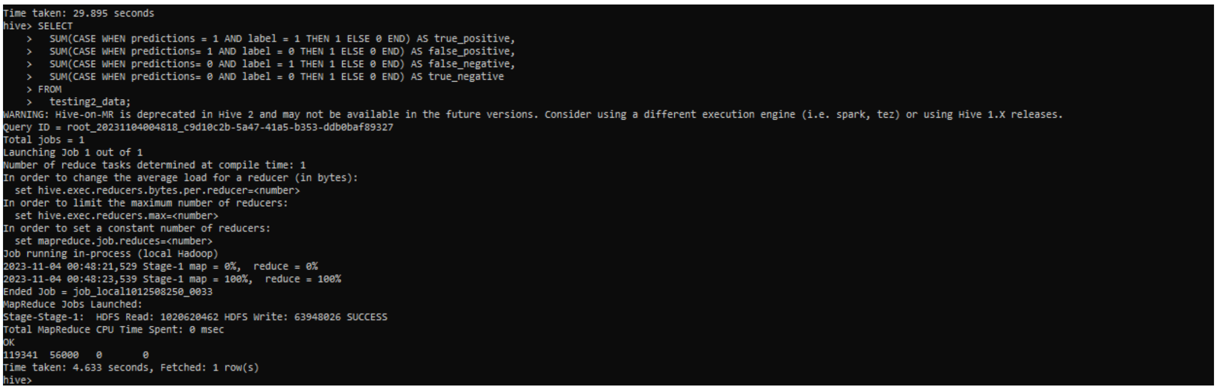
6. test traces apply your detection, mark the prediction and produce the confusion matrix:

Confusion matrix



```
hive> INSERT OVERWRITE TABLE testing2_data (
    > SELECT
    >   id,dur, proto, service, state, spkts, dpkts, sbytes, dbytes, rate, sttl, dttl, sload, dload, sloss, dloss, sinpkt, dinpkt, sjit, djit, swin, stcpb, dtcpb, dwin,
    > tcprtt, synack, ackdat, smean, dmean, trans_depth, response_body_len, ct_srv_src, ct_state_ttl, ct_dst_ltm, ct_src_dport_ltm, ct_dst_sport_ltm, ct_dst_src_ltm,
    > is_ftp_login, ct_ftp_cmd, ct_flw_http_mthd, ct_src_ltm, ct_srv_dst, is_sm_ips_ports, attack_cat, label,
    >   CASE
    >     WHEN spkts > 18.666472331535733 THEN 1
    >     WHEN spkts <= 18.666472331535733  AND ct_srv_src > 0.1297429917893407  THEN 1
    >     ELSE 0
    >   END AS predictions
    > FROM
    >   testing2_data);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231104004455_e604dac6-88ea-4ba1-bc9d-4d1580adff09
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2023-11-04 00:45:01,868 Stage-1 map = 0%,  reduce = 0%
2023-11-04 00:45:23,981 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local1089027298_0032
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://namenode:8020/user/hive/warehouse/networkdata.db/testing2_data/.hive-staging_hive_2023-11-04_00-44-55_918_5264200757855359834-1/-ext-10000
Loading data to table networkdata.testing2_data
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 478336218 HDFS Write: 31974013 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 29.895 seconds
hive>
```

Confusion matrix



```
Time taken: 29.895 seconds
hive> SELECT
    >   SUM(CASE WHEN predictions = 1 AND label = 1 THEN 1 ELSE 0 END) AS true_positive,
    >   SUM(CASE WHEN predictions= 1 AND label = 0 THEN 1 ELSE 0 END) AS false_positive,
    >   SUM(CASE WHEN predictions= 0 AND label = 1 THEN 1 ELSE 0 END) AS false_negative,
    >   SUM(CASE WHEN predictions= 0 AND label = 0 THEN 1 ELSE 0 END) AS true_negative
    > FROM
    >   testing2_data;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20231104004818_c9d10c2b-5a47-41a5-b353-ddb0baf89327
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-04 00:48:21,529 Stage-1 map = 0%,  reduce = 0%
2023-11-04 00:48:23,539 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1012508250_0033
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 1020620462 HDFS Write: 63948026 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
119341  56000   0       0
Time taken: 4.633 seconds, Fetched: 1 row(s)
hive>
```