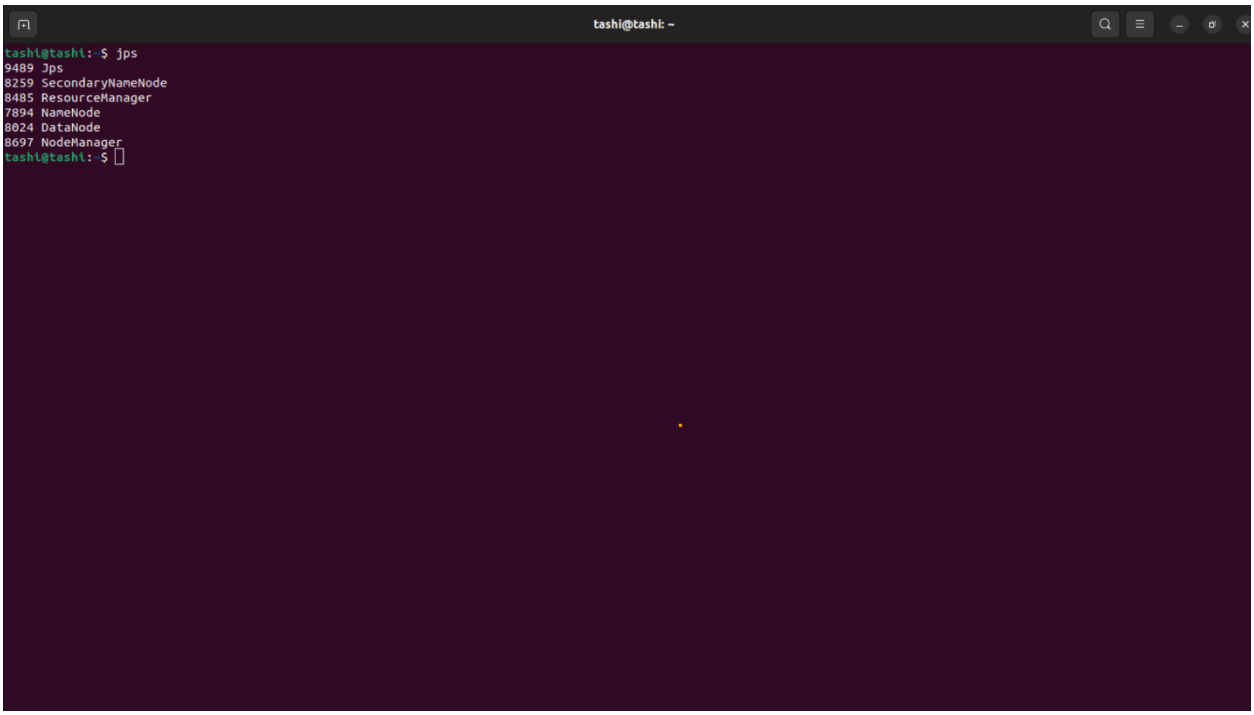
BIG DATA ANALYTICS

Final Project

FEBRUARY 2, 2024

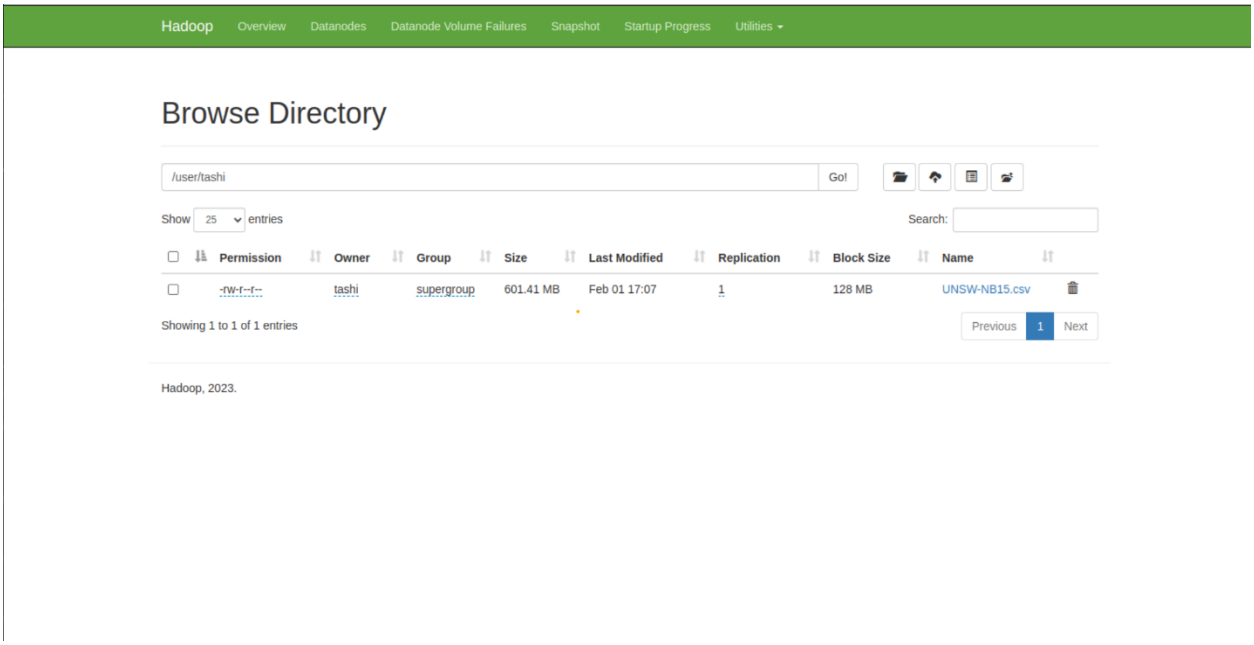
1.1 Hadoop is Running Successfully

- The jps command output confirms that necessary Hadoop services are running, including:
    - SecondaryNameNode
    - ResourceManager
    - NameNode
    - DataNode
    - NodeManager
- This ensures that the Hadoop Distributed File System (HDFS) and YARN are active.



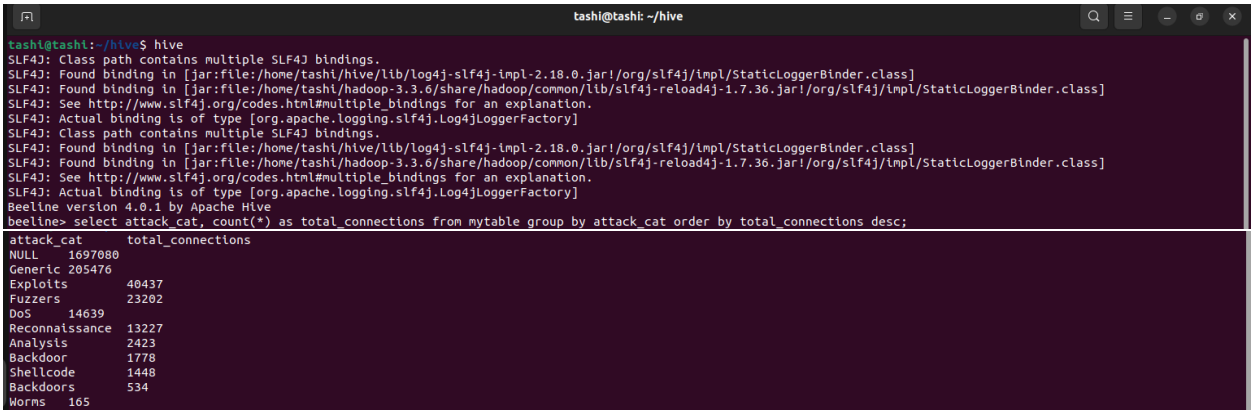
1.2 Dataset Uploaded to HDFS

- Hadoop Web UI confirms that the UNSW-NB15.csv dataset is successfully uploaded to HDFS.
- The file size is 601.41 MB, stored under /user/tashi/ with appropriate read/write permissions.



2. Apache Hive Queries Execution

Query 1:



Query 2:

```
tashi@tashi:~/hive$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
Beeline version 4.0.1 by Apache Hive
beeline> select proto, count(*) as protocol_count from mytable group by proto order by protocol_count desc;
```

```
proto   protocol_count
tcp     1128336
udp     827834
unas    13682
arp     6914
ospf    6234
sctp    1465
any     348
gre     303
sun-nd  241
pim     241
swipe   241
mobile  241
sep     239
rsvp    232
ipv6    230
icmp    213
mfe-nsp 116
rvd     116
stp     116
sdrp    116
micp    116
mux     116
aris    116
ipcv    116
a/n     116
il      116
ipx-n-ip        116
ddx     116
ptp     116
rdp     116
emcon   116
smp     116
wb-expak        116
secure-vmt      116
ipcomp  116
trunk-2 116
hmp     116
iso-tp4 116
3pc     116
xnet    116
ipip    116
pnni    116
```

```
pri-enc 116
vmtp    116
crtp    116
i-nlsp  116
dgp     116
pipe    116
trunk-1 116
st2     116
zero    116
pup     116
chaos   116
iatp    116
ggp     116
idpr    116
etherip 116
wsn     116
sccopmce        116
br-sat-mon      116
igp     116
xns-idp 116
sps     116
bna     116
cpnx    116
srp     116
visa    116
xtp     116
ipv6-opts       116
ipv6-frag       116
compaq-pee      116
uti     116
ib      116
idpr-cmtp       116
ippc    116
tlsp    116
nvp     116
idrp    116
ttp     116
narp    116
larp    116
igmp    62
rtp     7
```

Query 3:

```
tashi@tashi:~/hive$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
Beeline version 4.0.1 by Apache Hive
beeline> select srcip, dstip, dsport, proto, state from mytable where dsport > 50000 order by dsport asc;
```

srcip	dstip	dsport	proto	state
59.166.0.9	149.171.126.7	50001	udp	CON
59.166.0.1	149.171.126.8	50001	tcp	FIN
59.166.0.9	149.171.126.5	50001	tcp	FIN
59.166.0.7	149.171.126.6	50001	tcp	FIN
59.166.0.1	149.171.126.3	50001	tcp	FIN
59.166.0.6	149.171.126.5	50001	tcp	FIN
59.166.0.7	149.171.126.1	50001	tcp	FIN
59.166.0.2	149.171.126.8	50002	tcp	FIN
59.166.0.7	149.171.126.7	50002	tcp	FIN
175.45.176.1	149.171.126.10	50002	tcp	FIN
175.45.176.0	149.171.126.17	50002	tcp	FIN
175.45.176.1	149.171.126.16	50002	tcp	FIN
175.45.176.2	149.171.126.13	50002	tcp	FIN
175.45.176.3	149.171.126.14	50002	tcp	FIN
59.166.0.5	149.171.126.8	50002	tcp	FIN
59.166.0.1	149.171.126.6	50002	udp	CON
59.166.0.2	149.171.126.3	50002	tcp	FIN
59.166.0.0	149.171.126.1	50002	tcp	FIN
175.45.176.2	149.171.126.12	50002	tcp	FIN
59.166.0.4	149.171.126.2	50002	tcp	FIN
59.166.0.2	149.171.126.8	50003	tcp	FIN
59.166.0.1	149.171.126.9	50003	tcp	FIN
59.166.0.5	149.171.126.9	50003	tcp	FIN
59.166.0.3	149.171.126.3	50003	tcp	FIN
59.166.0.1	149.171.126.4	50003	tcp	FIN
59.166.0.0	149.171.126.7	50003	tcp	FIN
59.166.0.8	149.171.126.5	50003	tcp	FIN
59.166.0.3	149.171.126.2	50003	tcp	FIN
59.166.0.2	149.171.126.1	50004	tcp	FIN
59.166.0.9	149.171.126.5	50004	tcp	FIN
59.166.0.6	149.171.126.1	50004	tcp	FIN
59.166.0.1	149.171.126.0	50004	tcp	FIN
59.166.0.3	149.171.126.8	50004	tcp	FIN
175.45.176.0	149.171.126.11	50004	tcp	FIN
175.45.176.0	149.171.126.11	50004	tcp	FIN
59.166.0.1	149.171.126.8	50005	tcp	FIN
59.166.0.8	149.171.126.0	50005	tcp	FIN
59.166.0.2	149.171.126.3	50005	tcp	FIN
59.166.0.3	149.171.126.9	50005	tcp	FIN
59.166.0.6	149.171.126.3	50005	udp	CON
59.166.0.4	149.171.126.1	50005	tcp	FIN
59.166.0.4	149.171.126.3	50005	tcp	FIN
59.166.0.5	149.171.126.3	50101	tcp	FIN
59.166.0.9	149.171.126.2	50101	tcp	FIN
59.166.0.1	149.171.126.6	50102	tcp	FIN
59.166.0.0	149.171.126.6	50102	tcp	FIN
59.166.0.4	149.171.126.6	50102	tcp	FIN
59.166.0.5	149.171.126.9	50102	udp	CON
59.166.0.3	149.171.126.9	50102	tcp	FIN
59.166.0.9	149.171.126.1	50102	tcp	FIN
59.166.0.2	149.171.126.6	50102	udp	CON
59.166.0.7	149.171.126.7	50103	tcp	FIN
59.166.0.9	149.171.126.0	50103	tcp	FIN
59.166.0.2	149.171.126.3	50103	tcp	FIN
59.166.0.2	149.171.126.7	50103	tcp	FIN
59.166.0.0	149.171.126.0	50103	tcp	FIN
59.166.0.5	149.171.126.5	50103	tcp	FIN
59.166.0.7	149.171.126.8	50103	tcp	FIN
59.166.0.3	149.171.126.6	50103	tcp	FIN
59.166.0.0	149.171.126.0	50103	tcp	FIN
59.166.0.6	149.171.126.5	50103	tcp	FIN
59.166.0.5	149.171.126.0	50103	tcp	FIN
59.166.0.9	149.171.126.8	50104	tcp	FIN
175.45.176.2	149.171.126.15	50104	tcp	FIN
59.166.0.1	149.171.126.2	50104	tcp	FIN
59.166.0.8	149.171.126.7	50104	tcp	FIN
59.166.0.5	149.171.126.2	50104	tcp	CON
59.166.0.1	149.171.126.1	50104	tcp	FIN
59.166.0.9	149.171.126.3	50104	tcp	FIN
59.166.0.0	149.171.126.8	50104	tcp	FIN
59.166.0.4	149.171.126.3	50105	udp	CON
59.166.0.2	149.171.126.7	50105	udp	CON
59.166.0.0	149.171.126.0	50105	tcp	FIN
59.166.0.9	149.171.126.2	50105	tcp	FIN
59.166.0.2	149.171.126.0	50105	tcp	FIN
59.166.0.1	149.171.126.0	50105	tcp	FIN
59.166.0.1	149.171.126.1	50105	tcp	FIN
59.166.0.2	149.171.126.9	50105	tcp	FIN
59.166.0.4	149.171.126.4	50106	udp	CON
59.166.0.1	149.171.126.1	50106	tcp	FIN
59.166.0.3	149.171.126.7	50106	tcp	FIN
59.166.0.8	149.171.126.8	50106	tcp	FIN
59.166.0.5	149.171.126.0	50106	tcp	FIN

Query 4:

```
tashi@tashi:~/hive$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
Beeline version 4.0.1 by Apache Hive
beeline> select state, avg(dur) as avg_duration from mytable group by state order by avg_duration desc;
```

state	avg_duration
REQ	23.6520025320951
RST	4.71968852654839
PAR	2.65564513043483
FIN	0.81862227860841
ECO	0.776047062315775
CON	0.379094692684259
ACC	0.306560067668
TXD	0.2316186
INT	0.162004784681629
CLO	0.119018461137742
URN	0.00551671428571429
TST	7.75000075E-06
ECR	6.625001375E-06
MAS	5.33333346666667E-06
no	0
URH	0

Query 5:

```
tashi@tashi:~/hive$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/tashi/hive/lib/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/tashi/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.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]
Beeline version 4.0.1 by Apache Hive
beeline> select top 10 srcip, dstip, sbytes, dbytes, proto from mytable order by sbytes desc;
```

srcip	dstip	sbytes	dbytes	proto
175.45.176.1	149.171.126.15	14355774	68026	tcp
175.45.176.0	149.171.126.13	13677393	65274	tcp
175.45.176.2	149.171.126.13	12965233	62436	tcp
175.45.176.2	149.171.126.17	12594395	60802	tcp
175.45.176.0	149.171.126.13	12536928	61446	tcp
175.45.176.1	149.171.126.15	12500367	61132	tcp
175.45.176.2	149.171.126.17	12029826	59288	tcp
175.45.176.2	149.171.126.13	11854358	58530	tcp
175.45.176.3	149.171.126.13	11631181	58598	tcp
175.45.176.0	149.171.126.16	11085528	55276	tcp

3.1 Analyze and Interpret Big Data

Methods and Findings:

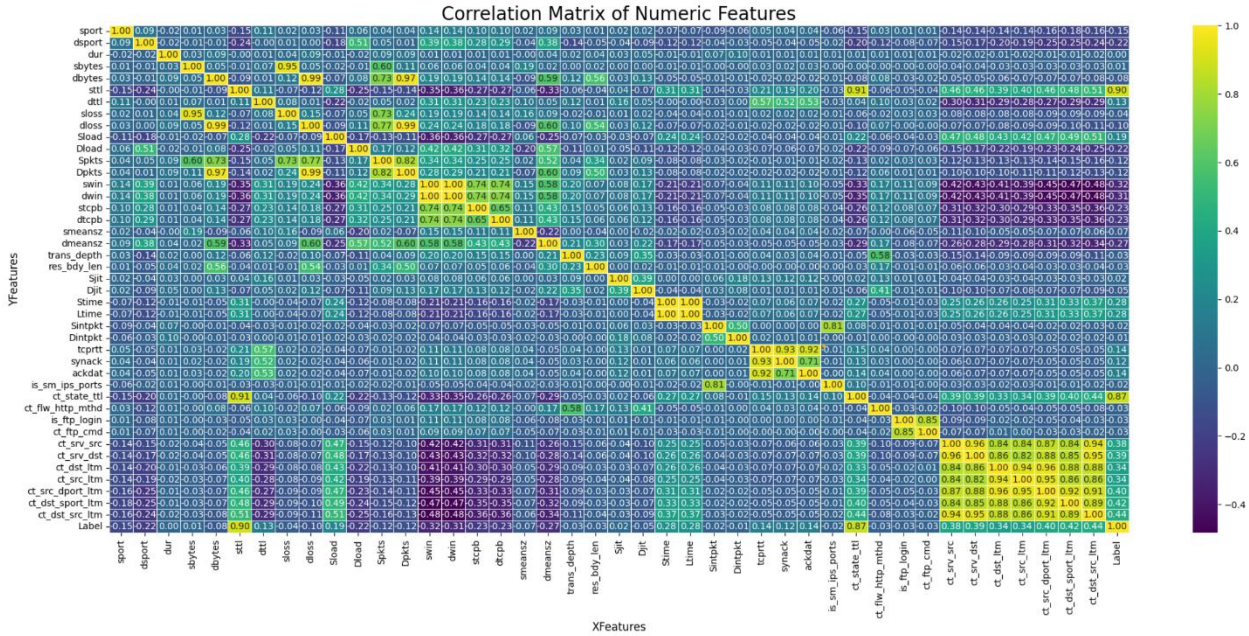
1. **Descriptive Statistics**
   - Summary statistics were generated for all numerical features.
   - This knowledge allowed teams to check for distribution patterns and detect data irregularities.

```
| Descriptive Statistics |
```

	sport	dsport	dur	sbytes	dbytes	sttl	dttl	sloss	dloss	Sload	...	is ftp login	ct ftp cmd	ct srv src	ct srv dst	ct dst ltm	ct src ltm	ct src dport ltm	ct dst sp
count	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	...	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.539739e+06	2.5397
mean	3.053693e+04	1.123510e+04	6.588634e-01	4.340072e+03	3.643201e+04	6.278150e+01	3.077044e+01	5.164547e+00	1.633142e+01	3.694928e+07	...	1.735336e-02	2.056038e-02	9.207912e+00	8.989883e+00	6.439727e+00	6.901640e+00	4.642572e+00	3.5930
std	2.044122e+04	1.843820e+04	1.392577e+01	5.640940e+04	1.611053e+05	7.462670e+01	4.285192e+01	2.251837e+01	5.659789e+01	1.186041e+08	...	1.334651e-01	1.843730e-01	1.083708e+01	1.082281e+01	8.162330e+00	8.205340e+00	8.478001e+00	6.1747
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.0000
25%	1.123100e+04	5.300000e+01	1.037000e-03	2.000000e+02	1.780000e+02	3.100000e+01	2.900000e+01	0.000000e+00	0.000000e+00	1.353769e+05	...	0.000000e+00	0.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00	1.000000e+00	1.0000
50%	3.169000e+04	8.000000e+01	1.586400e-02	1.470000e+03	1.820000e+03	3.100000e+01	2.900000e+01	3.000000e+00	4.000000e+00	5.893038e+05	...	0.000000e+00	0.000000e+00	5.000000e+00	5.000000e+00	3.000000e+00	4.000000e+00	1.000000e+00	1.0000
75%	4.743900e+04	1.497000e+04	2.147540e-01	3.182000e+03	1.490800e+04	3.100000e+01	2.900000e+01	7.000000e+00	1.400000e+01	2.038365e+06	...	0.000000e+00	0.000000e+00	1.000000e+01	1.000000e+01	6.000000e+00	7.000000e+00	2.000000e+00	1.0000
max	6.553500e+04	6.553500e+04	8.786638e+03	1.435577e+07	1.465753e+07	2.550000e+02	2.540000e+02	5.319000e+03	5.507000e+03	5.988000e+09	...	4.000000e+00	8.000000e+00	6.700000e+01	6.700000e+01	6.700000e+01	6.700000e+01	6.700000e+01	6.0000

2. Correlation Analysis

- A correlation matrix received visual treatment through heatmap representation.
- Several features demonstrated strong correlations which suggested overlap or duplication between them.

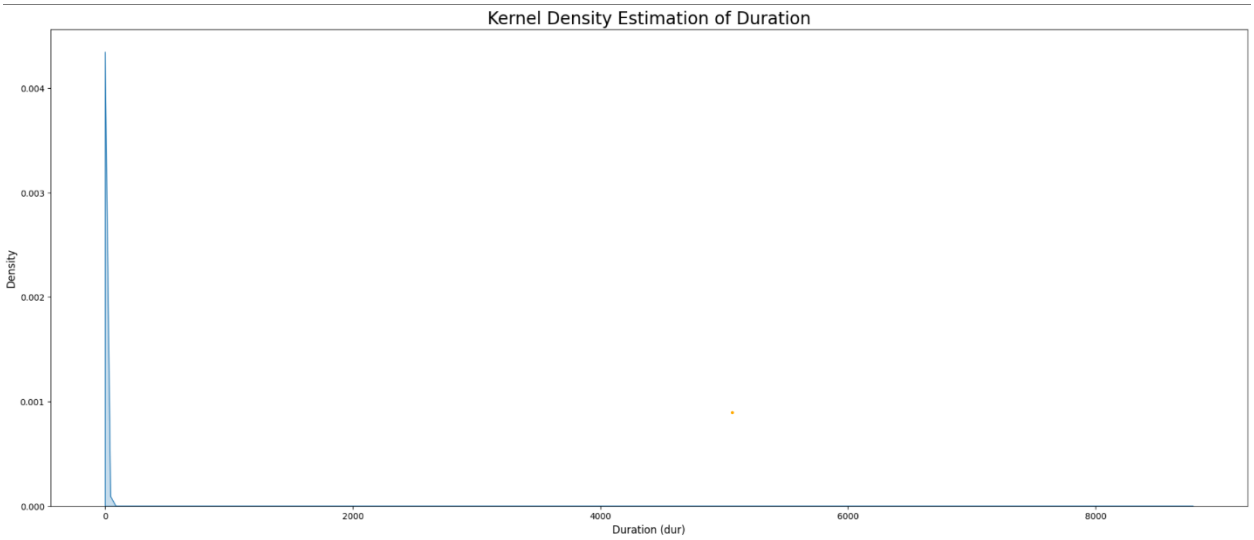


3. Hypothesis Testing (ANOVA - F-statistic)

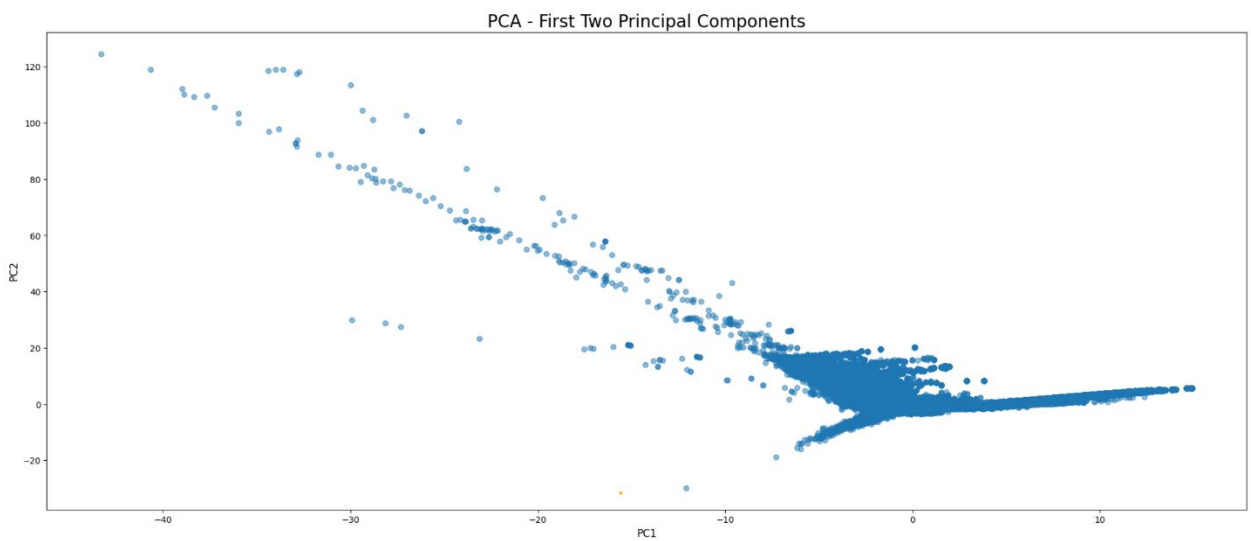
- ANOVA was conducted to analyze the relationship between attack categories and duration.
- **P-value:** NaN (suggesting insufficient data).
- **Conclusion:** Statistics showed no important correlations between variables.

4. Density Estimation (T-statistic for Feature Significance)

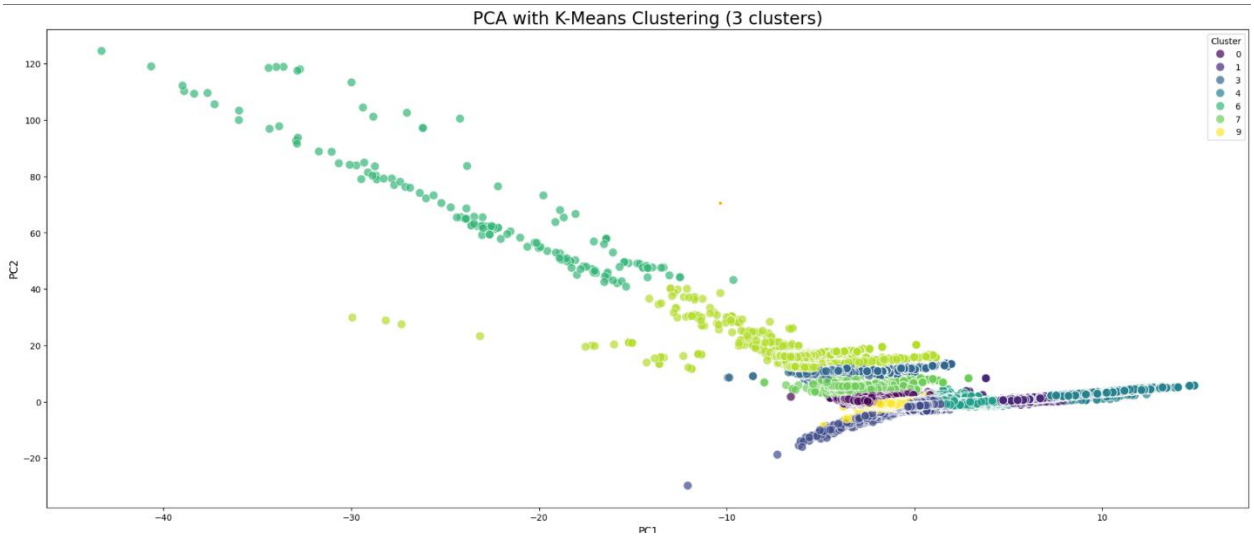
- **P-value:** 3.677e-59
- **Conclusion:** The hypothesis rejection establishes that important differences exist between these tested features.



5. **Principal Component Analysis (PCA)**
    - Variance captured by **PC1:** 22.76%
    - Variance captured by **PC2:** 10.37%
    - **Conclusion:** The first and second principal components successfully explain most of the data variability which suggests benefits for reducing information dimensions.



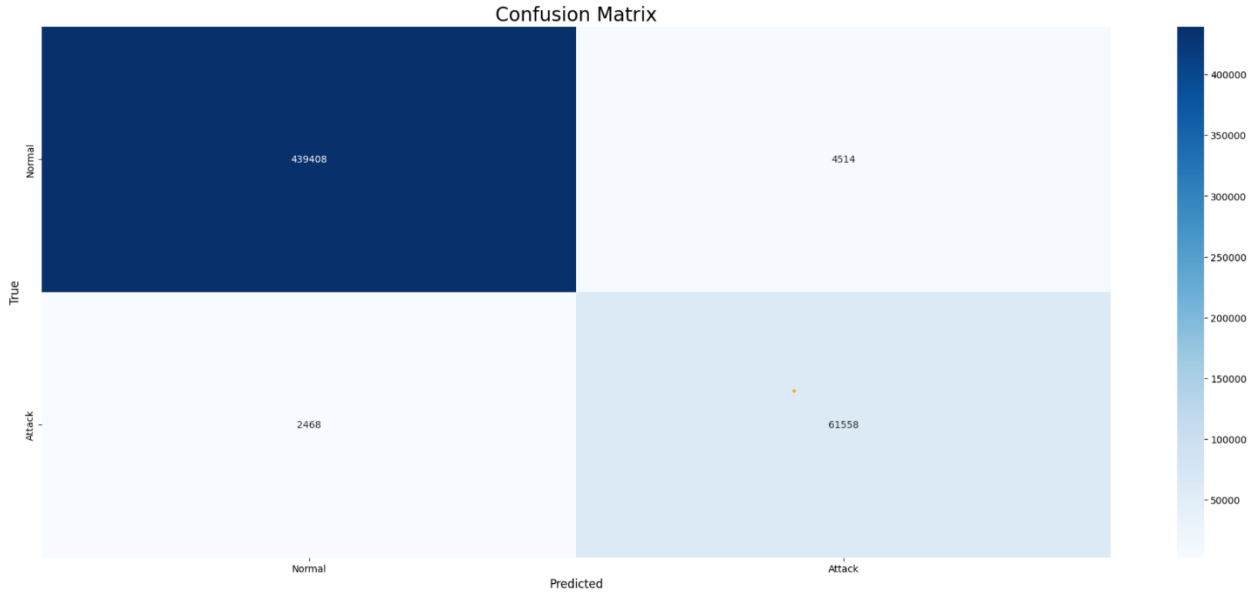
6. **K-Means Clustering**
    - Identified **10 cluster centers** with varied distributions.
    - The extreme cluster values including [-22.72, 67.22] could suggest the presence of outliers and attack clusters.



3.2 Design and Build a Classifier

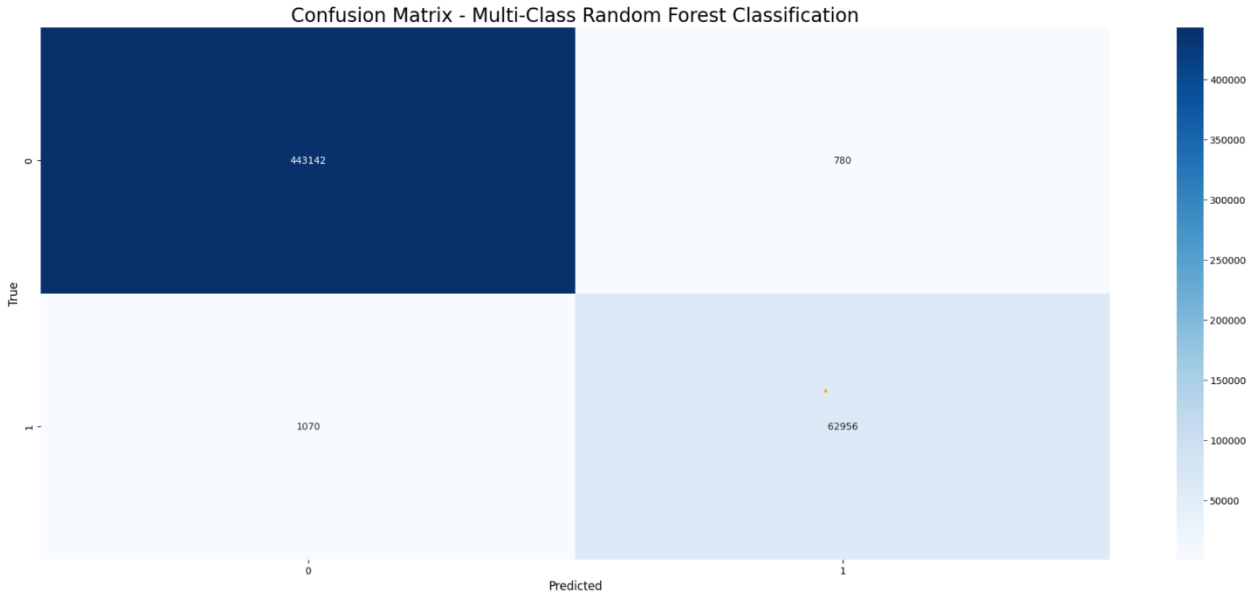
(a) Binary Classification

- **Model:** Logistic Regression
- **Performance Metrics:**
  - **Accuracy:** 98.6%
  - **Precision:** 99% (Normal) | 93% (Attack)
  - **Recall:** 99% (Normal) | 96% (Attack)
  - **F1-Score:** 99% (Normal) | 95% (Attack)
- **Findings:**
  - The model performed **exceptionally well**, with high accuracy.
  - Attack detection was slightly lower in precision but had high recall.



(b) Multi-Class Classification

- **Model:** Random Forest Classifier
- **Classes:** 10 (Normal + 9 Attack Categories)
- **Accuracy:** 99.63%
- **Findings:**
    - The similar patterns between DoS and Fuzzers attack categories produced misclassification in the analysis.
    - System performance was stable for common attack classes although it failed to identify less frequently occurring attacks.



4. Individual Assessment

1. Alternative Technologies for Tasks 2 and 3

- **Apache Impala:**
  - The high-speed parallel processing SQL engine functions as a solution for big data analytical queries which also delivers low latency performance.
  - Impala operates differently from Apache Hive because it performs query execution through a distributed query engine instead of converting queries into MapReduce jobs which leads to interactive workload speedup.
- **Apache Flink:**
  - Flink functions as an optimized distributed framework for real-time data processing of both streaming and batch operations.
  - Flink runs with continuous event processing at low latency because it does not utilize Apache Spark's micro-batch operations which makes it optimal for cybersecurity real-time anomaly detection.

2. New Thinking Evoked & Neglected

New Thinking:

- **Apache Flink for Real-Time Processing:**
  - Running Flink operations through streaming streams rather than Spark batch analytics would create a more effective solution for detecting real-time anomalies.
- **Apache Impala for Faster Queries:**
  - Moving from Hive to Impala platform will reduce query execution times significantly when performing ad-hoc analysis due to MapReduce independence.

Neglected Considerations:

- **Resource Management:**
  - Effective memory optimization becomes essential for Flink to operate fast and efficient event-streaming processes.
- **Compatibility:**
  - Implementation of Impala would involve both schema modifications and performance refinements when migrating from Hive.

References

1. Apache Hive Documentation: https://cwiki.apache.org/confluence/display/Hive

2.UNSW-NB15 Dataset: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/