Assignment 2

Setting up 3-node cluster:

Step1: Setting up hadoop in all 3 nodes.

```
hadoop@aminpri-1-of-3:~/hadoop$ ls
LICENSE-binary  LICENSE.txt  NOTICE-binary  NOTICE.txt  README.txt  bin  etc  include  lib  libexec  licenses-binary  logs  sbin  share
hadoop@aminpri-1-of-3:~/hadoop$
```

```
hadoop@aminpri-2-of-3:~/hadoop$ ls
LICENSE-binary  LICENSE.txt  NOTICE-binary  NOTICE.txt  README.txt  bin  etc  include  lib  libexec  licenses-binary  logs  sbin  share
hadoop@aminpri-2-of-3:~/hadoop$
```

```
hadoop@aminpri-3-of-3:~/hadoop$ ls
LICENSE-binary  LICENSE.txt  NOTICE-binary  NOTICE.txt  README.txt  bin  etc  include  lib  libexec  licenses-binary  logs  sbin  share
hadoop@aminpri-3-of-3:~/hadoop$
```

Step2: Configuring ssh and adding the keys to all 3 nodes to access a node from any of the other 3 nodes.

```
hadoop@aminpri-1-of-3:~/.ssh$ cat authorized_keys
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQDAKXHirf4kktia9wOvJCoDceYnXcdKkgszouzqZKtxtg6bjyk5ykGG8oz6/CeDIV0tKk9NqQWTh8c2k0u20PAUtHbHt2zcXV27KpxkIxNE5Y
JesIqcQt1sBgxdhTI8A4QdRaByrsXia4+5aZMjoAjnhZyEZBLNjX0U+iq0TZa2dwEZEOYUOfpQFjmT6FSlKpixXk/aY9NmVO3gU7KLdaoo5In26V/fKd0nBwEGTl3t/XW3OkfZn3hr6sqP/oBE
HYwAeOD3pq7dJKoTS6uvxpuQFKsJRUOQRAkzxD+k452HKx/VEjkNTVIE0bgHYFO3PZ1uG6UABJVO9DQLZsb4pOWcghJvxRrvS5PrTv9ZKco75D02wVB5oiNV3CW+m0tYjlGE7OCxi2Ahslkxtq
M+6X6Z8I1zumw/RI1yI9w4DQq5VQ3pXi/Bzu2fU2Y3cAYq/ZBeYYWnldZ4G8AoA2UTQzaaDGsffsIldbRYs5RFnaaI2bsyFmR26ocHFIGCm+aJrqk= hadoop@aminpri-1-of-3
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQCkPmkVjohO1O3B7feHnVbXEITSJz6Rs20EWnPoADI3p41iZ9OtFx8nlDtOgGDXEuD4n7vNnfqlYKQewEP1jmXgyISNe8j2Gtk+YHrVBeidfJ
Pnfsxtc4vvclrlFNN29cPL9CP8yRfmkhkPaGsIFBkaOUbvl5VuWduhY/VvkEaDTHHt61wB9TAQmdonqeVARTLrSdRAw3ORXxdUqM7IYgg1BkZxJ2nF4o7d7+LuCF4TCgCizPxLvTZKBTIsWyOK
qZwvVrwUvvgj9HaUWzFekpZaAgpRnAlgg9uh2f9c7eK/VDQnwPuYf2s2ils22+N2wThQ8ZL8qc+g+XLb6sJ22GQtZ56ECekaFMLFEQJOas2z+AS7UF5V8JAmwzfB8U4kYp4qYdVobt65+xefIL
H6o/U0EYV4Yu9/Jq0wCIchNo2zBZNMRyEjIIbavXoMBaYv4XH81R2XcJ9hRnogxtOfatXQ6ExeVIKZBD9OW1FM52vRnR4ld4gSeM7poYbipR94Fo0= hadoop@aminpri-2-of-3
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQCktZpJeO2vw1uIxpQykEOmOrSraDjiUp1FmjYphkWNxeU0RLQYGrbHhG519j8sCp/V+Pu+oLEYqEaEuHGkoQTVn9sehppEmQRruIS1Y4Ep3t
zINB6Id4sCf6lPxgvabf4l2U0yyittRcK5fTwBe3RoQXDGUhg7N5wJ15lfKXKlCtLBAV1T9X1LScI7eMwEzIBJPHDpauMDVT37/3Tj69e/hldOOG1CwPtgm5LevF4jkct/LxCvhkNm/lC07Hrw
gUrb757rrUEWSIZidW2DMS+xx4iFb1y/Smyore8P+Ve5DUjI+cC+64spwHkt4MDzwtdlvfSkcoQ79nUMerX9Hx6d9vuEXbgJRC/h9UuIDtjdSVVyb+rpP2kXB6sXlRqYqcgpw4pD+58UUYYmRT
tcPLXc2Fxzir7Fgv6vlVnPIqgFZyQpKmrfbiz04EsZTTDudT4dXJ4RJyJy/gQRQV9M+fkPGXowXrWxiW3OIruWMMEHrAGy/PJJ7To8uTcQQ9JwD8k= hadoop@aminpri-3-of-3
hadoop@aminpri-1-of-3:~/.ssh$
```

Similarly, we will add it to hadoop@aminpri-2-of-3 and hadoop@aminpri-3-of-3

Step3: Configuring hadoop@aminpri-1-of-3 as pour master node and the other two nodes as slave nodes. Obtain the ip address of all 3 nodes using ip addr.

```
hadoop@aminpri-1-of-3:~$ ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp1s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 8900 qdisc fq_codel state UP group default qlen 1000
    link/ether fa:16:3e:c9:f3:0a brd ff:ff:ff:ff:ff:ff
    inet 10.0.195.85/24 metric 100 brd 10.0.195.255 scope global dynamic enp1s0
       valid_lft 85409sec preferred_lft 85409sec
    inet6 fe80::f816:3eff:fec9:f30a/64 scope link
       valid_lft forever preferred_lft forever
```

Master: 10.0.195.85
Slave nodes: 10.0.195.92 and 10.0.195.197

Step4: Edit the core-site.xml file and add ip address of master for all 3 nodes

```
<configuration>
<property>
        <name>fs.defaultFS</name>
        <value>hdfs://10.0.195.85:9000</value>
    </property>
</configuration>
hadoop@aminpri-1-of-3:~$
```

```
<configuration>
<property>
        <name>fs.defaultFS</name>
        <value>hdfs://10.0.195.85:9000</value>
    </property>
</configuration>
hadoop@aminpri-2-of-3:~/hadoop$
```

```
<configuration>
<property>
        <name>fs.defaultFS</name>
        <value>hdfs://10.0.195.85:9000</value>
    </property>
</configuration>
hadoop@aminpri-3-of-3:~/hadoop$
```

Step5: Edit the hdfs-site.xml file. Change the NameNode and DataNode directory paths.
Namenode - master
Datanode - slave nodes

```
<configuration>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
    </property>

    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
    </property>
</configuration>
hadoop@aminpri-1-of-3:~$
```

```
<configuration>
<property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
    </property>
</configuration>
hadoop@aminpri-2-of-3:~/hadoop$
```

```
<configuration>
<property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
    </property>
</configuration>
hadoop@aminpri-3-of-3:~/hadoop$
```

Step6: Edit the mapred-site.xml file. This would be the same for all nodes.

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
    <property>
        <name>mapreduce.admin.user.env</name>
        <value>HADOOP_MAPRED_HOME=$HADOOP_COMMON_HOME</value>
    </property>
    <property>
        <name>yarn.app.mapreduce.am.env</name>
        <value>HADOOP_MAPRED_HOME=$HADOOP_COMMON_HOME</value>
    </property>
</configuration>
hadoop@aminpri-1-of-3:~$
```

Step7: Edit the yarn-site.xml file. Add the ip of the master node as resource manager for all nodes.

```
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
    <description>The hostname of the RM.</description>
    <name>yarn.resourcemanager.hostname</name>
    <value>10.0.195.85</value>
</property>
<property>
    <description>The address of the applications manager interface in the RM.</description>
    <name>yarn.resourcemanager.address</name>
    <value>10.0.195.85:8032</value>
</property>
</configuration>
hadoop@aminpri-1-of-3:~$
```

```
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
    <description>The hostname of the RM.</description>
    <name>yarn.resourcemanager.hostname</name>
    <value>10.0.195.85</value>
</property>
<property>
    <description>The address of the applications manager interface in the RM.</description>
    <name>yarn.resourcemanager.address</name>
    <value>10.0.195.85:8032</value>
</property>
</configuration>
hadoop@aminpri-2-of-3:~/hadoop$
```

```
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
    <description>The hostname of the RM.</description>
    <name>yarn.resourcemanager.hostname</name>
    <value>10.0.195.85</value>
</property>
<property>
    <description>The address of the applications manager interface in the RM.</description>
    <name>yarn.resourcemanager.address</name>
    <value>10.0.195.85:8032</value>
</property>
</configuration>
hadoop@aminpri-3-of-3:~/hadoop$
```

Step8: Add the ip address of the slave nodes in the workers file of the master node. No change for slave nodes

```
hadoop@aminpri-1-of-3:~/hadoop/etc/hadoop$ cat workers
10.0.195.85
10.0.195.92
10.0.195.197
hadoop@aminpri-1-of-3:~/hadoop/etc/hadoop$
```

```
hadoop@aminpri-2-of-3:~/hadoop/etc/hadoop$ cat workers
localhost
hadoop@aminpri-2-of-3:~/hadoop/etc/hadoop$
```

```
hadoop@aminpri-3-of-3:~/hadoop/etc/hadoop$ cat workers
localhost
hadoop@aminpri-3-of-3:~/hadoop/etc/hadoop$
```

Step9: Format the Namenode as a hadoop user (only on Master) and run the command start-all.sh (only on Master).
Check the status of all nodes using jps.
You should see the NameNode, SecondaryNameNode, and ResourceManager daemons running on your master, and DataNode and NodeManager daemons on your slaves.

```
hadoop@aminpri-1-of-3:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [aminpri-1-of-3.js2local]
Starting datanodes
Starting secondary namenodes [aminpri-1-of-3]
Starting resourcemanager
Starting nodemanagers
hadoop@aminpri-1-of-3:~$ jps
11329 Jps
10531 DataNode
10886 ResourceManager
10392 NameNode
11000 NodeManager
10698 SecondaryNameNode
hadoop@aminpri-1-of-3:~$
```

```
hadoop@aminpri-2-of-3:~$ jps
132131 DataNode
132228 NodeManager
132349 Jps
hadoop@aminpri-2-of-3:~$
```

```
hadoop@aminpri-3-of-3:~$ jps
132484 Jps
132266 DataNode
132363 NodeManager
hadoop@aminpri-3-of-3:~$
```

A 3-node cluster has been setup successfully.

References:
1. *Installation of Hadoop on Ubuntu 2023-*
   https://sanjayshankar1428.medium.com/in-the-world-of-big-data-managing-and-processing-vast-amounts-of-information-efficiently-is-2ffdc0204357
2. *Hadoop Multi Node Cluster Setup - https://www.youtube.com/watch?v=_iP2Em-5Abw*

NY Parking Violations

1. **When are tickets most likely to be issued?**

```
+-------------+-----+
|violation_time|count|
+-------------+-----+
|        0836A|20112|
|        0839A|19531|
|        0840A|19436|
|        0838A|19433|
|        0906A|19412|
|        1139A|19255|
|        1140A|19154|
|        1141A|19135|
|        0841A|18987|
|        1142A|18936|
|        1145A|18917|
|        0837A|18893|
|        0842A|18806|
|        1143A|18738|
|        0910A|18698|
|        1138A|18679|
|        0845A|18648|
|        0908A|18648|
|        0909A|18604|
|        1136A|18541|
|        1144A|18524|
|        0907A|18464|
|        0843A|18363|
|        1146A|18324|
|        1147A|18313|
|        0844A|18185|
|        1137A|17977|
|        0846A|17964|
|        1148A|17963|
|        0912A|17929|
+-------------+-----+
only showing top 30 rows
```

2. **What are the most common years and types of cars to be ticketed?**

```
+-----------+-----------+--------------+
|vehicle_type|vehicle_year|violation_count|
+-----------+-----------+--------------+
|       SUBN|       2021|        468828|
|       SUBN|       2022|        452377|
|       SUBN|       2023|        447136|
|       SUBN|       2019|        345021|
|       SUBN|       2020|        343283|
|       SUBN|       2018|        275702|
|       SUBN|       2017|        226828|
|       SUBN|       2016|        186233|
|       SUBN|       2015|        180925|
|       4DSD|       2017|        155318|
|       4DSD|       2019|        152339|
|       4DSD|       2018|        146010|
|       SUBN|       2014|        142950|
|       4DSD|       2020|        138040|
|       4DSD|       2023|        134263|
|       4DSD|       2021|        134002|
|       4DSD|       2022|        131561|
|       SUBN|       2013|        130515|
|       4DSD|       2015|        129270|
|       4DSD|       2016|        126511|
|       SUBN|       2011|        117729|
|       SUBN|       2012|        117674|
|       4DSD|       2013|        108203|
|       4DSD|       2014|        102357|
|       SUBN|       2008|         93048|
|       SUBN|       2010|         88975|
|       4DSD|       2012|         87130|
|       SUBN|       2007|         77786|
|       4DSD|       2010|         74395|
|        VAN|       2019|         73346|
+-----------+-----------+--------------+
only showing top 30 rows
```

3. Where are tickets most commonly issued?

```
+---------------+------------------+
|Number_of_tickets|Violation_Location|
+---------------+------------------+
|         276203|               19|
|         213205|              114|
|         207636|                6|
|         189589|               13|
|         178348|               14|
|         153765|              109|
|         148286|                1|
|         147809|               18|
|         142074|                9|
|         135832|              115|
|         116439|               61|
|         115903|               66|
|         115747|               20|
|         109812|              112|
|         107721|               70|
|         104404|               84|
|         104246|              103|
|         103097|               52|
|         102733|              108|
|          98620|               46|
|          94933|               10|
|          88065|              110|
|          84947|               90|
|          84910|              102|
|          83649|               43|
|          83088|              104|
|          81657|               24|
|          80699|               62|
|          76885|               44|
|          76378|               17|
+---------------+------------------+
only showing top 30 rows
```

4. Which color of the vehicle is most likely to get a ticket?

```
+------------+-------+
|Vehicle_Color|  count|
+------------+-------+
|          GY|2086349|
|          WH|1924604|
|          BK|1821703|
|        NULL|1015118|
|          BL| 688918|
|       WHITE| 610935|
|       BLACK| 401993|
|          RD| 393388|
|        GREY| 303176|
|        BLUE| 140721|
|          GR| 134699|
|       SILVE| 134375|
|       BROWN| 129885|
|         RED| 116436|
|         BLK|  89406|
|          TN|  70852|
|          BR|  68615|
|          YW|  65868|
|         GRY|  64505|
|         WHI|  54907|
|          GL|  39134|
|       OTHER|  33184|
|          MR|  30204|
|       GREEN|  26247|
|          OR|  24657|
|         SIL|  22166|
|        GRAY|  21722|
|         BLU|  20578|
|       YELLO|  15621|
|        LTGY|  12880|
+------------+-------+
only showing top 30 rows
```

5. **Given a Black vehicle parking illegally at 34510, 10030, 34050 (street codes). What is the probability that it will get a ticket?**

```
Probability of Black vehicle parking illegally at 34510, 10030, 34050 getting ticket:
[Stage 74:=================================================> (21 + 1) / 22]
+----------+------+----------+------------------+
|prediction| Count|Total_Cars|       probability|
+----------+------+----------+------------------+
|         2|533813|   3575206|0.1493097180973628|
+----------+------+----------+------------------+
```

NBA Shot Logs

1. **For each pair of the players (A, B), we define the fear score of A when facing B as the hit rate, such that B is the closest defender when A is shooting. Based on the fear sore, for each player, please find out who is his "most unwanted defender"**

```
+---------------+-----------------+--------+-----------+
|    player_name| CLOSEST_DEFENDER|hit_rate|total_shots|
+---------------+-----------------+--------+-----------+
|    aaron brooks|    Nurkic, Jusuf|     0.0|          5|
|    aaron gordon|   Rivers, Austin|     0.0|          2|
|  al farouq aminu|   Johnson, James|     0.0|          3|
|      al horford|      Diaw, Boris|     0.0|          4|
|    al jefferson|Hardaway Jr., Tim|     0.0|          3|
|   alan anderson|       Leuer, Jon|     0.0|          5|
|     alan crabbe| Sefolosha, Thabo|     0.0|          3|
|         alex len|  Knight, Brandon|     0.0|          4|
|    alexis ajinca|     Meeks, Jodie|     0.0|          3|
|       alonzo gee|     Korver, Kyle|     0.0|          3|
|amare stoudemire|      Deng, Luol|     0.0|          3|
|    amir johnson|       Tucker, PJ|     0.0|          3|
|   andre drummond|    James, LeBron|     0.0|          5|
|   andre iguodala|     Lowry, Kyle|     0.0|          3|
|     andre miller|     Turner, Evan|     0.0|          2|
|   andre roberson|   Ginobili, Manu|     0.0|          2|
|     andrew bogut|     Ibaka, Serge|     0.0|          6|
|   andrew wiggins|   Roberts, Brian|     0.0|          9|
|  anthony bennett|   Ajinca, Alexis|     0.0|          5|
|    anthony davis|Mbah a Moute, Luc|     0.0|          5|
+---------------+-----------------+--------+-----------+
only showing top 20 rows
```

2. **For each player, we define the comfortable zone of shooting is a matrix of, {SHOT DIST, CLOSE DEF DIST, SHOT CLOCK}.Considering the hit rate, which zone is the best for James Harden, Chris Paul, Stephen Curry, and Lebron James.**

```
+-------------+----------------+------------------+
| player_name|prediction (zone)|      Best_zone_res|
+-------------+----------------+------------------+
|    chris paul|               2|0.5531914893617021|
|  james harden|               0|0.5604395604395604|
|  lebron james|               0|0.6613545816733067|
|  stephen curry|              0|0.6350710900473934|
+-------------+----------------+------------------+
```