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
qZwvVrwUvvgj9HaUWrFekpZaAgpRnAlgg9uh2f9c7eK/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.

```
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
6736 ResourceManager
6544 SecondaryNameNode
6850 NodeManager
6251 NameNode
7180 Jps
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:~$
```

3-node cluster has been setup successfully.

```
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ chmod +x download_nyc_data.sh
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ ./download_nyc_data.sh
Downloading rows 0 to 50000...
Downloading rows 50000 to 100000...
Downloading rows 100000 to 150000...
Downloading rows 150000 to 200000...
Downloading rows 200000 to 250000...
Downloading rows 250000 to 300000...
Downloading rows 300000 to 350000...
```

```
Downloading rows 10600000 to 10650000...
Downloading rows 10650000 to 10700000...
Downloading rows 10700000 to 10750000...
Download completed.
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ ls -lrt
total 2800052
-rw-rw-r-- 1 hadoop hadoop     263059 Mar 16 18:52 nyc_data_old.csv
-rw-rw-r-- 1 hadoop hadoop       3736 Apr 11 03:01 parking_violation.py
-rwxrwxr-x 1 hadoop hadoop       1148 Apr 11 03:17 download_nyc_data.sh
-rw-rw-r-- 1 hadoop hadoop 2866971972 Apr 11 03:28 nyc_data_full.csv
```

```
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ head -n 5 nyc_data_full.csv
"summons_number","plate_id","registration_state","plate_type","issue_date","violation_code","vehicle_body_type","vehicle
_make","issuing_agency","street_code1","street_code2","street_code3","vehicle_expiration_date","violation_location","vio
lation_precinct","issuer_precinct","issuer_code","issuer_command","issuer_squad","violation_time","time_first_observed",
"violation_county","violation_in_front_of_or_opposite","house_number","street_name","intersecting_street","date_first_ob
served","law_section","sub_division","violation_legal_code","days_parking_in_effect","from_hours_in_effect","to_hours_in
_effect","vehicle_color","unregistered_vehicle","vehicle_year","meter_number","feet_from_curb","violation_post_code","vi
olation_description","no_standing_or_stopping_violation","hydrant_violation","double_parking_violation"
"1159637337","KZH2758","NY","PAS","2023-06-09T00:00:00.000","67","VAN","HONDA","P","0","0","0","20250201","0043","43","4
3","972773","0043","0000","0911A",,"BX",,,"I/O TAYLOR AVE","GUERLAIN","0","408","E5",,"BBBBBBB","ALL","ALL","BLUE","0","
2006","-","0",,,,,
"1252960645","JPD8746","NY","PAS","2023-06-30T00:00:00.000","87","SUBN","LINCO","M","17870","25390","32670","20240210","
0014","14","968","271057","0968","0000","0717A",,"NY","O","51","E 44TH ST",,"0","408","D",,"BBBBBBB","ALL","ALL","GRAY",
"0","2020","-","0",,,,,
"1252960669","JPD8746","NY","PAS","2023-06-30T00:00:00.000","31","SUBN","LINCO","M","17870","25390","32670","20240210","
0014","14","968","271057","0968","0000","0823A",,,"O","51","E 44TH STREET",,"0","408","O3",,"BBBBBBB","ALL","ALL","GRAY"
,"0","2020","-","0",,,,,
"1252994126","MBH9245","99","PAS","2023-07-06T00:00:00.000","20","SDN","KIA","M","12690","41700","61090","20231231","108
","108","968","272834","0968","0000","1150P",,,"F","39-41","60TH ST",,"0","408","D",,"BBBBBBB","ALL","ALL","WHITE","0","
0","-","0",,,,,
```

```
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ wc -l nyc_data_full.csv
10717483 nyc_data_full.csv
```

```
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ ls -lrt
total 2800052
-rw-rw-r-- 1 hadoop hadoop     263059 Mar 16 18:52 nyc_data_old.csv
-rw-rw-r-- 1 hadoop hadoop       3736 Apr 11 03:01 parking_violation.py
-rwxrwxr-x 1 hadoop hadoop       1148 Apr 11 03:17 download_nyc_data.sh
-rw-rw-r-- 1 hadoop hadoop 2866971972 Apr 11 03:28 nyc_data.csv
```

```
hadoop@aminpri-1-of-3:~/assignment2/nyc_data$ python3 parking_violation.py
/usr/lib/python3/dist-packages/scipy/__init__.py:146: UserWarning: A NumPy version >=1.17.3 and <1.25.0 is required for this version of SciPy (det
ected version 1.26.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/04/11 05:22:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Most Issued Times:
+--------------+-----+
|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|
+--------------+-----+
only showing top 20 rows
```

```
Most common years and types of cars to be ticketed:
+------------+-----------------+------+
|vehicle_year|vehicle_body_type| count|
+------------+-----------------+------+
|        2021|             SUBN|468828|
|        2022|             SUBN|452377|
|        2023|             SUBN|447136|
|        2019|             SUBN|345021|
|        2020|             SUBN|343283|
|        2018|             SUBN|275702|
|        2017|             SUBN|226828|
|        2016|             SUBN|186233|
|        2015|             SUBN|180925|
|        2017|             4DSD|155318|
|        2019|             4DSD|152339|
|        2018|             4DSD|146010|
|        2014|             SUBN|142950|
|        2020|             4DSD|138040|
|        2023|             4DSD|134263|
|        2021|             4DSD|134002|
|        2022|             4DSD|131561|
|        2013|             SUBN|130515|
|        2015|             4DSD|129270|
|        2016|             4DSD|126511|
+------------+-----------------+------+
only showing top 20 rows
```

```
Most Common Locations:
+------------------+------+
|violation_location| count|
+------------------+------+
|              0019|276203|
|               114|213205|
|              0006|207636|
|              0013|189589|
|              0014|178348|
|               109|153765|
|              0001|148286|
|              0018|147809|
|              0009|142074|
|               115|135832|
|              0061|116439|
|              0066|115903|
|              0020|115747|
|               112|109812|
|              0070|107721|
|              0084|104404|
|               103|104246|
|              0052|103097|
|               108|102733|
|              0046| 98620|
+------------------+------+
only showing top 20 rows
```

```
Most Issued Colors:
+-------------+-------+
|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|
+-------------+-------+
only showing top 20 rows
```

```
Showing the clustered data:
+------------+------------+------------+-------------+--------------------+----------+
|street_code1|street_code2|street_code3|vehicle_color|            features|prediction|
+------------+------------+------------+-------------+--------------------+----------+
|           0|           0|           0|         BLUE|           (3,[],[])|         1|
|       17870|       25390|       32670|         GRAY|[17870.0,25390.0,...|         0|
|       17870|       25390|       32670|         GRAY|[17870.0,25390.0,...|         0|
|       12690|       41700|       61090|        WHITE|[12690.0,41700.0,...|         2|
|       12690|       41700|       61090|        WHITE|[12690.0,41700.0,...|         2|
|           0|           0|           0|           BK|           (3,[],[])|         1|
|        8690|       21690|       21740|         NULL|[8690.0,21690.0,2...|         0|
|           0|       61090|           0|          GRY|   [0.0,61090.0,0.0]|         0|
|           0|           0|           0|          WHT|           (3,[],[])|         1|
|           0|       40404|       40404|         BLUE|[0.0,40404.0,4040...|         2|
|       59590|        9440|           0|         GREY|[59590.0,9440.0,0.0]|         0|
|       37890|       51290|       51895|        WHITE|[37890.0,51290.0,...|         2|
|       24090|       10290|       52490|          BLK|[24090.0,10290.0,...|         0|
|       23840|       57790|       42120|          RED|[23840.0,57790.0,...|         2|
|       28940|        9690|       57790|         GOLD|[28940.0,9690.0,5...|         2|
|       21790|        5940|           0|         BLUE|[21790.0,5940.0,0.0]|         1|
|       36400|       24240|       46090|         NULL|[36400.0,24240.0,...|         2|
|        8440|       65490|       58590|          BLK|[8440.0,65490.0,5...|         2|
|        6190|       23090|       23190|           WH|[6190.0,23090.0,2...|         0|
|        8590|       53390|       65490|           RD|[8590.0,53390.0,6...|         2|
+------------+------------+------------+-------------+--------------------+----------+
only showing top 20 rows
```

```
Probability of black vehicle getting ticket in different clusters:
+----------+------+----------+------------------+
|prediction|  Count|Total_Cars|       probability|
+----------+------+----------+------------------+
|         0| 411818|   2890149|  0.1424902314724 95|
|         1|1134880|   4759961|0.23842212152578562|
|         2| 365911|   2052254|0.17829713086196933|
+----------+------+----------+------------------+

Probability of Black vehicle parking illegally at 34510, 10030, 34050 getting ticket:
+----------+------+----------+-----------------+
|prediction| Count|Total_Cars|      probability|
+----------+------+----------+-----------------+
|         0|411818|   2890149|0.1424902314724 95|
+----------+------+----------+-----------------+
```