HIVE Assingnment

Query1:

```
hive> SELECT split(`issue_date`,'/')[2] as year,count(*) as total_no_of_tckts FROM nyc_ticket_system where `issue_date`
            > IS NOT NULL and split(`issue_date`,'/')[2] is not null GROUP BY split(`issue_date`,'/')[2] order by year;
Query ID = hdfs_20180624111818_cc289e69-130b-4ca7-a05c-89ce24d6ed1f
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 32
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>
Starting Job = job_1529811797645_0446, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0446/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0446
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 32
2018-06-24 11:18:43,629 Stage-1 map = 0%,  reduce = 0%
2018-06-24 11:19:10,060 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 20.49 sec
2018-06-24 11:19:15,515 Stage-1 map = 7%,  reduce = 0%, Cumulative CPU 23.8 sec
2018-06-24 11:19:16,626 Stage-1 map = 13%,  reduce = 0%, Cumulative CPU 29.99 sec
2018-06-24 11:19:22,112 Stage-1 map = 21%,  reduce = 0%, Cumulative CPU 34.23 sec
2018-06-24 11:19:23,192 Stage-1 map = 30%,  reduce = 0%, Cumulative CPU 40.31 sec
2018-06-24 11:19:26,411 Stage-1 map = 38%,  reduce = 0%, Cumulative CPU 42.08 sec
2018-06-24 11:19:45,950 Stage-1 map = 39%,  reduce = 0%, Cumulative CPU 48.71 sec
2018-06-24 11:19:51,432 Stage-1 map = 44%,  reduce = 0%, Cumulative CPU 66.67 sec
2018-06-24 11:19:54,651 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 70.4 sec
2018-06-24 11:19:57,927 Stage-1 map = 62%,  reduce = 0%, Cumulative CPU 78.91 sec
2018-06-24 11:20:02,348 Stage-1 map = 68%,  reduce = 0%, Cumulative CPU 80.84 sec
2018-06-24 11:20:04,569 Stage-1 map = 69%,  reduce = 0%, Cumulative CPU 84.04 sec
2018-06-24 11:20:06,807 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 85.01 sec
2018-06-24 11:20:22,082 Stage-1 map = 79%,  reduce = 0%, Cumulative CPU 94.35 sec
2018-06-24 11:20:27,495 Stage-1 map = 87%,  reduce = 0%, Cumulative CPU 101.65 sec
2018-06-24 11:20:28,563 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 105.58 sec
2018-06-24 11:20:45,905 Stage-1 map = 100%,  reduce = 3%, Cumulative CPU 107.67 sec
2018-06-24 11:20:48,102 Stage-1 map = 100%,  reduce = 6%, Cumulative CPU 109.75 sec
2018-06-24 11:20:49,157 Stage-1 map = 100%,  reduce = 9%, Cumulative CPU 112.01 sec
2018-06-24 11:21:01,193 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 114.39 sec
2018-06-24 11:21:06,632 Stage-1 map = 100%,  reduce = 16%, Cumulative CPU 116.46 sec
2018-06-24 11:21:08,848 Stage-1 map = 100%,  reduce = 19%, Cumulative CPU 118.6 sec
2018-06-24 11:21:18,591 Stage-1 map = 100%,  reduce = 22%, Cumulative CPU 120.6 sec
2018-06-24 11:21:23,963 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 122.66 sec
2018-06-24 11:21:26,164 Stage-1 map = 100%,  reduce = 28%, Cumulative CPU 124.84 sec
2018-06-24 11:21:40,378 Stage-1 map = 100%,  reduce = 31%, Cumulative CPU 126.97 sec
2018-06-24 11:21:42,515 Stage-1 map = 100%,  reduce = 34%, Cumulative CPU 128.96 sec
2018-06-24 11:21:44,681 Stage-1 map = 100%,  reduce = 38%, Cumulative CPU 130.97 sec
2018-06-24 11:21:59,655 Stage-1 map = 100%,  reduce = 41%, Cumulative CPU 133.08 sec
2018-06-24 11:22:02,841 Stage-1 map = 100%,  reduce = 44%, Cumulative CPU 135.03 sec
2018-06-24 11:22:08,161 Stage-1 map = 100%,  reduce = 47%, Cumulative CPU 137.19 sec
2018-06-24 11:22:18,982 Stage-1 map = 100%,  reduce = 53%, Cumulative CPU 141.17 sec
2018-06-24 11:22:26,421 Stage-1 map = 100%,  reduce = 56%, Cumulative CPU 143.09 sec
2018-06-24 11:22:36,262 Stage-1 map = 100%,  reduce = 59%, Cumulative CPU 145.26 sec
2018-06-24 11:22:38,427 Stage-1 map = 100%,  reduce = 63%, Cumulative CPU 147.27 sec
2018-06-24 11:22:41,670 Stage-1 map = 100%,  reduce = 66%, Cumulative CPU 149.36 sec
2018-06-24 11:22:53,626 Stage-1 map = 100%,  reduce = 69%, Cumulative CPU 151.54 sec
2018-06-24 11:22:56,841 Stage-1 map = 100%,  reduce = 72%, Cumulative CPU 153.61 sec
```

```
2018-06-24 11:22:58,989 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 155.63 sec
2018-06-24 11:23:10,866 Stage-1 map = 100%,  reduce = 78%, Cumulative CPU 155.63 sec
2018-06-24 11:23:16,257 Stage-1 map = 100%,  reduce = 81%, Cumulative CPU 159.86 sec
2018-06-24 11:23:18,454 Stage-1 map = 100%,  reduce = 84%, Cumulative CPU 161.95 sec
2018-06-24 11:23:29,469 Stage-1 map = 100%,  reduce = 88%, Cumulative CPU 164.17 sec
2018-06-24 11:23:33,772 Stage-1 map = 100%,  reduce = 91%, Cumulative CPU 166.36 sec
2018-06-24 11:23:34,835 Stage-1 map = 100%,  reduce = 94%, Cumulative CPU 168.81 sec
2018-06-24 11:23:46,718 Stage-1 map = 100%,  reduce = 97%, Cumulative CPU 171.07 sec
2018-06-24 11:23:48,880 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 173.4 sec
MapReduce Total cumulative CPU time: 2 minutes 53 seconds 400 msec
Ended Job = job_1529811797645_0446
Launching Job 2 out of 2
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>
Starting Job = job_1529811797645_0455, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0455/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0455
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 11:24:21,584 Stage-2 map = 0%,  reduce = 0%
2018-06-24 11:24:39,915 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.85 sec
2018-06-24 11:24:55,166 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.93 sec
MapReduce Total cumulative CPU time: 3 seconds 930 msec
Ended Job = job_1529811797645_0455
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8  Reduce: 32   Cumulative CPU: 173.4 sec   HDFS Read: 2146020282 HDFS Write: 4350 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.93 sec   HDFS Read: 18325 HDFS Write: 418 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 57 seconds 330 msec
OK
1972    2
1973    2
1974    1
1976    1
1977    1
1984    1
1985    1
1990    2
1991    3
1994    1
1996    1
1997    1
2000    185
2001    2
2002    1
2003    1
2004    2
2005    1
2006    8
2007    18
2008    4
2009    3
2010    48
2011    22
2012    87
2013    70
2014    120
2015    419
```

2016 5368366
2017 5431903
2018 1057
2019 472
2020 22
2021 22
2022 4
2023 5
2024 3
2025 6
2026 24
2027 50
2028 8
2029 2
2030 12
2031 5
2033 2
2036 1
2041 1
2047 2
2053 1
2060 2
2061 1
2062 2
2063 2
2068 1
2069 4
Time taken: 412.245 seconds, Fetched: 55 row(s)
hive>

Query for part1 2:

hive> SELECT count(DISTINCT `violation_county`) as Unique_States from nyc_ticket_system where `violation_county` IS NOT NULL;
Query ID = hdfs_20180624113333_3edffefd-f34f-41ab-a3c9-5565e0f9f07e
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>
Starting Job = job_1529811797645_0464, Tracking URL = http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0464/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529811797645_0464
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2018-06-24 11:34:24,921 Stage-1 map = 0%, reduce = 0%
2018-06-24 11:34:49,260 Stage-1 map = 5%, reduce = 0%, Cumulative CPU 7.09 sec
2018-06-24 11:34:50,340 Stage-1 map = 9%, reduce = 0%, Cumulative CPU 13.4 sec
2018-06-24 11:34:51,433 Stage-1 map = 21%, reduce = 0%, Cumulative CPU 21.43 sec
2018-06-24 11:34:55,788 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 25.87 sec
2018-06-24 11:35:14,802 Stage-1 map = 42%, reduce = 0%, Cumulative CPU 32.68 sec
2018-06-24 11:35:18,144 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 34.53 sec
2018-06-24 11:35:21,461 Stage-1 map = 59%, reduce = 0%, Cumulative CPU 46.74 sec
2018-06-24 11:35:25,816 Stage-1 map = 66%, reduce = 0%, Cumulative CPU 48.62 sec
2018-06-24 11:35:27,975 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 50.87 sec
2018-06-24 11:35:43,511 Stage-1 map = 82%, reduce = 0%, Cumulative CPU 58.07 sec
2018-06-24 11:35:44,779 Stage-1 map = 88%, reduce = 0%, Cumulative CPU 58.8 sec
2018-06-24 11:35:48,063 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 64.26 sec

```
2018-06-24 11:36:01,125 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 66.31 sec
MapReduce Total cumulative CPU time: 1 minutes 6 seconds 310 msec
Ended Job = job_1529811797645_0464
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8  Reduce: 1   Cumulative CPU: 66.31 sec   HDFS Read: 2145852079 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 6 seconds 310 msec
OK
19
Time taken: 141.235 seconds, Fetched: 1 row(s)

Query part1 3:
hive> SELECT count(*) as Number FROM nyc_ticket_system WHERE
            > `street_code1` is null Or `street_code2` is null Or `street_code3` is null OR
            > `street_code1` = 0 Or `street_code2` = 0 Or `street_code3`  = 0 ;
Query ID = hdfs_20180624114040_4bd3abc9-3d59-4ce7-b4ef-e00e0d59a3bf
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>
Starting Job = job_1529811797645_0481, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0481/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0481
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 1
2018-06-24 11:40:55,509 Stage-1 map = 0%,  reduce = 0%
2018-06-24 11:41:21,957 Stage-1 map = 8%,  reduce = 0%, Cumulative CPU 13.12 sec
2018-06-24 11:41:23,083 Stage-1 map = 11%,  reduce = 0%, Cumulative CPU 19.69 sec
2018-06-24 11:41:26,320 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 23.13 sec
2018-06-24 11:41:27,381 Stage-1 map = 38%,  reduce = 0%, Cumulative CPU 28.62 sec
2018-06-24 11:41:50,544 Stage-1 map = 43%,  reduce = 0%, Cumulative CPU 36.12 sec
2018-06-24 11:41:51,637 Stage-1 map = 46%,  reduce = 0%, Cumulative CPU 42.88 sec
2018-06-24 11:41:52,725 Stage-1 map = 52%,  reduce = 0%, Cumulative CPU 50.19 sec
2018-06-24 11:41:54,849 Stage-1 map = 59%,  reduce = 0%, Cumulative CPU 51.83 sec
2018-06-24 11:41:55,930 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 55.87 sec
2018-06-24 11:42:17,770 Stage-1 map = 94%,  reduce = 0%, Cumulative CPU 69.34 sec
2018-06-24 11:42:18,836 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 70.28 sec
2018-06-24 11:42:31,689 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 72.52 sec
MapReduce Total cumulative CPU time: 1 minutes 12 seconds 520 msec
Ended Job = job_1529811797645_0481
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8  Reduce: 1   Cumulative CPU: 72.52 sec   HDFS Read: 2145862708 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 12 seconds 520 msec
OK
3667559
Time taken: 136.398 seconds, Fetched: 1 row(s)
hive>

Query for Part 2:
Question 1:
CREATE TABLE if not exists intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`violation_code`) ,`violation_code`
from nyc_ticket_system where `violation_code` is not null group by `violation_code`;
select `value` as Count ,`key` as ViolationCode from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
========
count    violationcode
    count    violationcode
```

1 1528577 21
2 1400614 36
3 1062302 38
4 893493 14
5 618592 20
==========
Qustion 2a:

```
CREATE TABLE intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`vehicle_body_type`) ,`vehicle_body_type`
from nyc_ticket_system where `vehicle_body_type` is not null group by `vehicle_body_type`;
select `value` as Count ,`key` as vehicleBodyType from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
```

==========
count vehiclebodytype
count vehiclebodytype
1 3719796 SUBN
2 3082006 4DSD
3 1411964 VAN
4 687324 DELV
5 438191 SDN
==========

Question 2b:

```
CREATE TABLE intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`vehicle_make`) ,`vehicle_make`
from nyc_ticket_system where `vehicle_make` is not null group by `vehicle_make`;
select `value` as Count ,`key` as vehicleMake from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
```

==========
count vehiclemake
count vehiclemake
1 1280956 FORD
2 1211447 TOYOT
3 1079237 HONDA
4 918590 NISSA
5 714654 CHEVR
==========
Question 3a:

```
CREATE TABLE intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`issuer_precinct`) ,`issuer_precinct`
from nyc_ticket_system where `issuer_precinct` is not null and `issuer_precinct` != '0' group by `issuer_precinct`;
select `value` as Count ,`key` as IssuerPrecinct from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
```

==========
count issuerprecinct
count issuerprecinct
1 521513 19
2 344977 14
3 321170 1
4 296554 18
5 289950 114
==========

Question 3b:

```
CREATE TABLE intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`violation_precinct`) ,`violation_precinct`
from nyc_ticket_system where `violation_precinct` is not null and `violation_precinct` != '0' group by
`violation_precinct`;
select `value` as Count ,`key` as ViolationPrecinct from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
```

==========

count violationprecinct

	count	violationprecinct
1	535671	19
2	352450	14
3	331810	1
4	306920	18
5	296514	114

==========

Question 5:
To parse violation time I have used Java class and created temporary function to link that java class functionality to the value.

Ex: If '0143A' is the value I assumed it as 01:43:AM and if '0543P' = 05:43:PM .
The java function will validate the data and convert 0143A to 01:43:00 and 0543P to 17:43:00.
and in the floor(hour(timehandler(`violation_time`))/4) function it will do:
hour() will return the hour value and floor() will divide the hours into 6 different bins[As we have used group by with this value.],
java function:

```
public String evaluate(String value){
   String ret = "";
   if(value.length() ==5 && value.matches("^[AP0-9_]*$")){

            char[] c = value.toCharArray();
            int h1 = Integer.parseInt(c[0]+"");
            int h2 = Integer.parseInt(c[1]+"");
            String HH = h1+""+h2;
            int Hour = Integer.parseInt(HH);
            int m1 = Integer.parseInt(c[2]+"");
            int m2 = Integer.parseInt(c[3]+"");
            String MM = m1+""+m2;
            int Minutes = Integer.parseInt(MM);

            if(c[4] == 'A'){
                     ret= HH+":"+MM+":00";
            }else if (c[4] == 'P'){
                     int hourIn24 = Hour + 12;
                     ret = hourIn24+":"+MM+":00";
            }                        return ret;
   }
   else{
            return "InvalidDate";
   }
```

```
cp /home/ec2-user/bigdata-0.0.1-SNAPSHOT.jar /var/lib/hadoop-hdfs/
add jar bigdata-0.0.1-SNAPSHOT.jar;
list jars;
create temporary function timehandler as 'com.nikhil.bigdata.TimeHandler';
insert overwrite table intermediatory_table_seasons select count(`violation_code`)
,`violation_code`,floor(hour(timehandler(`violation_time`))/4)
from nyc_ticket_system where `violation_time` is not null and length(`violation_time`) = 5
and floor(hour(timehandler(`violation_time`))/4) is not null group by
floor(hour(timehandler(`violation_time`))/4),`violation_code`;
```

By this the data is divided into bins and the description field [0,1,2,3,4,5] identifies it.
The below query returns the top 3 rows in that time division.

Question 6:

```
select * from intermediatory_table_seasons where description = '0' order by value desc limit 3;
```

==========

```
hive> select * from intermediatory_table_seasons where description = '0' order by value desc limit 3;
```

```
Query ID = hdfs_20180624130202_b9eea353-f64c-4ddc-8d1d-fff244e408a2
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>
Starting Job = job_1529811797645_0562, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0562/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0562
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-24 13:03:17,808 Stage-1 map = 0%,  reduce = 0%
2018-06-24 13:03:34,493 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.4 sec
2018-06-24 13:03:50,962 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.61 sec
MapReduce Total cumulative CPU time: 4 seconds 610 msec
Ended Job = job_1529811797645_0562
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.61 sec   HDFS Read: 17424 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 610 msec
OK
216842  21          0
211434  36          0
106868  38          0
Time taken: 71.351 seconds, Fetched: 3 row(s)
============================================================
select * from intermediatory_table_seasons where description = '1' order by value desc limit 3;
============================================================
hive> select * from intermediatory_table_seasons where description = '1' order b
        y value desc limit 3;
Query ID = hdfs_20180624130404_ebe6839b-3869-4a5d-9273-dfab6339060e
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>
Starting Job = job_1529811797645_0563, Tracking URL = http://ip-10-0-0-39.ap-sou
        th-1.compute.internal:8088/proxy/application_1529811797645_0563/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin
        /hadoop job  -kill job_1529811797645_0563
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-24 13:05:10,724 Stage-1 map = 0%,  reduce = 0%
2018-06-24 13:05:31,548 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.47 se
        c
2018-06-24 13:05:49,011 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.93 sec
MapReduce Total cumulative CPU time: 4 seconds 930 msec
Ended Job = job_1529811797645_0563
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.93 sec   HDFS Read: 17420 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 930 msec
OK
141275  14          1
119470  21          1
112187  40          1
Time taken: 78.83 seconds, Fetched: 3 row(s)
```

```
hive>
================================================================
select * from intermediatory_table_seasons where description = '2' order by value desc limit 3;
================================================================
hive> select * from intermediatory_table_seasons where description = '2' order by value desc limit 3;
Query ID = hdfs_20180624130505_c939b252-9e19-46b8-9c13-525103b15fc3
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>
Starting Job = job_1529811797645_0570, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0570/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0570
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-24 13:06:28,942 Stage-1 map = 0%,  reduce = 0%
2018-06-24 13:06:47,603 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2018-06-24 13:07:05,083 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.93 sec
MapReduce Total cumulative CPU time: 4 seconds 930 msec
Ended Job = job_1529811797645_0570
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.93 sec   HDFS Read: 17507 HDFS Write: 37 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 930 msec
OK
1182665 21          2
751422  36          2
346518  38          2
Time taken: 78.404 seconds, Fetched: 3 row(s)
hive>
================================================================
select * from intermediatory_table_seasons where description = '3' order by value desc limit 3;
================================================================
hive> select * from intermediatory_table_seasons where description = '3' order by value desc limit 3;
Query ID = hdfs_20180624151313_dfc4e9a0-b556-417e-9e21-fb08ab82391a
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>
Starting Job = job_1529811797645_0685, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0685/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0685
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-24 15:13:46,907 Stage-1 map = 0%,  reduce = 0%
2018-06-24 15:14:03,189 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.38 sec
2018-06-24 15:14:20,837 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.59 sec
MapReduce Total cumulative CPU time: 4 seconds 590 msec
Ended Job = job_1529811797645_0685
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.59 sec   HDFS Read: 17424 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 590 msec
OK
```

```
376961  36      3
356353  38      3
265866  37      3
Time taken: 74.088 seconds, Fetched: 3 row(s)
============================================================
select * from intermediatory_table_seasons where description = '4' order by value desc limit 3;
============================================================
hive> select * from intermediatory_table_seasons where description = '4' order by value desc limit 3;
Query ID = hdfs_20180624151616_a28a7205-aa5c-4a00-b388-1ff057b1dfc8
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>
Starting Job = job_1529811797645_0693, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0693/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0693
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-24 15:16:44,912 Stage-1 map = 0%,  reduce = 0%
2018-06-24 15:17:02,599 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.47 sec
2018-06-24 15:17:18,992 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.68 sec
MapReduce Total cumulative CPU time: 4 seconds 680 msec
Ended Job = job_1529811797645_0693
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.68 sec   HDFS Read: 17424 HDFS Write: 36 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 680 msec
OK
203233  38      4
145784  37      4
144748  14      4
Time taken: 76.419 seconds, Fetched: 3 row(s)
============================================================
select * from intermediatory_table_seasons where description = '5' order by value desc limit 3;
============================================================
hive> select * from intermediatory_table_seasons where description = '5' order by value desc limit 3;
Query ID = hdfs_20180624151616_6fe5d3ae-be30-4241-b588-16ebeb28cc11
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>
Starting Job = job_1529811797645_0695, Tracking URL =
http://ip-10-0-0-39.ap-south-1.compute.internal:8088/proxy/application_1529811797645_0695/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job  -kill
job_1529811797645_0695
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-24 15:17:02,852 Stage-1 map = 0%,  reduce = 0%
2018-06-24 15:17:19,740 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.45 sec
2018-06-24 15:17:35,178 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.81 sec
MapReduce Total cumulative CPU time: 4 seconds 810 msec
Ended Job = job_1529811797645_0695
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.81 sec   HDFS Read: 17424 HDFS Write: 32 SUCCESS
```

Total MapReduce CPU Time Spent: 4 seconds 810 msec
OK
65593 7 5
47030 38 5
44778 14 5
Time taken: 72.05 seconds, Fetched: 3 row(s)
hive>
==========
Question 7:
The below query will generate the list for the most violation codes and the description field will give the time of the day.[0 for 0-4 hours, 1 for 4-8,2 for 8-12,3 for 12-16,4 for 16-20,5 for 20-14 hours]
intermediatory_table_seasons.value intermediatory_table_seasons.key intermediatory_table_seasons.description

	intermediatory_table_seasons.value	intermediatory_table_seasons.key	intermediatory_table_seasons.description
1	1182665	21	2
2	751422	36	2
3	376961	36	3

select * from intermediatory_table_seasons order by value desc limit 3;
Question 8:
CREATE TABLE intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`issuer_precinct`) ,`issuer_precinct`
from nyc_ticket_system where `issuer_precinct` is not null and `issuer_precinct` != '0' group by `issuer_precinct`;
select `value` as Count ,`key` as IssuerPrecinct from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
==========
count issuerprecinct
count issuerprecinct
1 521513 19
2 344977 14
3 321170 1
4 296554 18
5 289950 114
==========

CREATE TABLE intermediatory_table (`value` bigint,`key` string);
insert overwrite table intermediatory_table SELECT count(`violation_precinct`) ,`violation_precinct`
from nyc_ticket_system where `violation_precinct` is not null and `violation_precinct` != '0' group by
`violation_precinct`;
select `value` as Count ,`key` as ViolationPrecinct from intermediatory_table order by Count desc limit 5;
drop table intermediatory_table;
==========
count violationprecinct
count violationprecinct
1 535671 19
2 352450 14
3 331810 1
4 306920 18
5 296514 114
==========

CREATE TABLE if not exists intermediatory_table_seasons (`value` bigint,`key` string,`description` string);

```
insert overwrite table intermediatory_table_seasons select count(`violation_code`) as
Count_for_Winter,`violation_code`,`violation_description` from nyc_ticket_system where split(`issue_date`,'/')[1] in
('12','01','02') group by `violation_code`,`violation_description`;
select `value` as Count_for_Winter,`key` as ViolationCode,`description` as ViolationDescription from
intermediatory_table_seasons order by Count_for_Winter desc limit 5;
drop table intermediatory_table_seasons;
==================
count    violationcode    violationdescription
    count    violationcode    violationdescription
1   37712   21   21-No Parking (street clean)
2   32877   38   38-Failure to Display Muni Rec
3   28752   36   PHTO SCHOOL ZN SPEED VIOLATION
4   26947   14   14-No Standing
5   18761   37   37-Expired Muni Meter

==================
select sum(value) as Frequency_of_tckts_for_Winter from intermediatory_table_seasons;
==================
frequency_of_tckts_for_winter
    frequency_of_tckts_for_winter
1   1097424
==================


CREATE TABLE intermediatory_table_seasons (`value` bigint,`key` string,`description` string);
insert overwrite table intermediatory_table_seasons select count(`violation code`) as Count_for_Spring,`violation
code`,`violation description` from nyc_ticket_system where split(`issue date`,'/')[1] in ('03','04','05') group by `violation
code`,`violation description`;
select `value` as Count_for_Spring,`key` as ViolationCode,`description` as ViolationDescription from
intermediatory_table_seasons order by Count_for_Spring desc limit 5;
drop table intermediatory_table_seasons;
====================
count_for_spring    violationcode    violationdescription
    count_for_spring    violationcode    violationdescription
1   123436   36   PHTO SCHOOL ZN SPEED VIOLATION
2   105267   21   21-No Parking (street clean)
3   92131   38   38-Failure to Display Muni Rec
4   83460   71   71A-Insp Sticker Expired (NYS)
5   74038   14   14-No Standing

select sum(value) as Frequency_of_tckts_for_Spring from intermediatory_table_seasons;
====================
frequency_of_tckts_for_spring
    frequency_of_tckts_for_spring
1   1021551
====================


CREATE TABLE intermediatory_table_seasons (`value` bigint,`key` string,`description` string);
insert overwrite table intermediatory_table_seasons select count(`violation code`) as Count_for_Fall,`violation
code`,`violation description` from nyc_ticket_system where split(`issue date`,'/')[1] in ('09','10','11') group by `violation
code`,`violation description`;
select `value` as Count_for_Fall,`key` as ViolationCode,`description` as ViolationDescription from
intermediatory_table_seasons order by Count_for_Fall desc limit 5;
drop table intermediatory_table_seasons;
========================
count_for_fall    violationcode    violationdescription
    count_for_fall    violationcode    violationdescription
1   115875   36   PHTO SCHOOL ZN SPEED VIOLATION
2   105554   38   38-Failure to Display Muni Rec
3   95665   21   21-No Parking (street clean)
4   77680   14   14-No Standing
```

5 55660 37 37-Expired Muni Meter
==========
select sum(value) as Frequency_of_tckts_for_Fall from intermediatory_table_seasons;
==========
frequency_of_tckts_for_fall
frequency_of_tckts_for_fall
1 988917
==========

CREATE TABLE intermediatory_table_seasons (`value` bigint,`key` string,`description` string);
insert overwrite table intermediatory_table_seasons select count(`violation code`) as Count_for_Summer,`violation code`,`violation description` from nyc_ticket_system where split(`issue date`,'/')[1] in ('06','07','08') group by `violation code`,`violation description`;
select `value` as Count_for_Summer,`key` as ViolationCode,`description` as ViolationDescription from intermediatory_table_seasons order by Count_for_Summer desc limit 5;
drop table intermediatory_table_seasons;
==========
count_for_summer violationcode violationdescription
count_for_summer violationcode violationdescription
1 149586 36 PHTO SCHOOL ZN SPEED VIOLATION
2 123722 21 21-No Parking (street clean)
3 102709 38 38-Failure to Display Muni Rec
4 82095 14 14-No Standing
5 61889 37 37-Expired Muni Meter
==========

select sum(value) as Frequency_of_tckts_for_Summer from intermediatory_table_seasons;

frequency_of_tckts_for_summer
frequency_of_tckts_for_summer
1 1116026
==========