Schema Design:

1) Created external table called “violation_table_ext” and loaded with Parking_Violations_Issued_-_Fiscal_Year_2017.csv data.
2) Created the second external partitioned table “violation_cleaned_partition_ext” based on “Season” value which is derived from “Issue_Date” column (divided into 4 seasons “Spring”,”Winter”,”Fall”,”summer”)
3) “violation_cleaned_partition_ext” table is inserted with violation_table_ext table data with dynamic partition using “Season” column
4) Also “Violation_Time” column is divided into 6 slots (like 'EarlyMorning',”Morning”,”late Morning”, “AfterNoon”, “Evening”,”Night”) and inserted as new “DaySlot” column.
5) Only the required columns are inserted into the second table

LOGS for Table creation:

External Table created in AWS S3 bucket:

```
hive> create external table violation_table_ext(
    > Summons_Number bigint, Plate_ID  string, Registration_State  string, Plate_Type
string, Issue_Date  string, Violation_Code  int, Vehicle_Body_Type  string, Vehicle_Make
string, Issuing_Agency  string, Street_Code1  int, Street_Code2  int, Street_Code3  int,
Vehicle_Expiration_Date  string, Violation_Location  string, Violation_Precinct  int,
Issuer_Precinct  int, Issuer_Code  int, Issuer_Command  string, Issuer_Squad string,
Violation_Time  string, Time_First_Observed  string, Violation_County  string,
Violation_In_Front_Of_Or_Opposite  string, House_Number string, Street_Name  string,
Intersecting_Street  string, Date_First_Observed  string, Law_Section  string, Sub_Division
string, Violation_Legal_Code  string, Days_Parking_In_Effect  string,
From_Hours_In_Effect  string, To_Hours_In_Effect  string, Vehicle_Color  string,
Unregistered_Vehicle  string, Vehicle_Year  int, Meter_Number  string, Feet_From_Curb
string, Violation_Post_Code string, Violation_Description  string,
No_Standing_or_Stopping_Violation  string, Hydrant_Violation  string,
Double_Parking_Violation  string)
    > row format delimited fields terminated by ','
    > location 's3a://hiveviolationdemo/violation_table_ext'
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 5.342 seconds
```

LOADING data to table:

```
hive> load data inpath 's3a://hiveviolationdemo/Parking_Violations_Issued_-
_Fiscal_Year_2017.csv' into table violation_table_ext;
Loading data to table default.violation_table_ext
Table default.violation_table_ext stats: [numFiles=1, totalSize=2086913576]
OK
```

Time taken: 8.776 seconds

Setting the Partition related configuration:

```
hive> set hive.exec.dynamic.partition =true;
hive> set hive.exec.dynamic.partition.mode=nonstrict;
```

Creating secondary external partition table:

```
hive> create external table violation_cleaned_partition_ext (
    > Summons_Number bigint,Registration_State  string, Issue_Date  string, Violation_Code int,
    > Vehicle_Body_Type  string, Vehicle_Make  string, Street_Code1  int, Street_Code2  int, Street_Code3  int,Violation_Precinct  int, Issuer_Precinct  int,
    > Violation_Time  string, Violation_Description string, DaySlot string)
    > PARTITIONED BY (Season String)
    > location 's3a://hiveviolationdemo/violation_cleaned_partition_table_ext';
OK
Time taken: 0.513 seconds
hive> desc violation_cleaned_partition_ext;
OK
summons_number      bigint
registration_state      string
issue_date              string
violation_code          int
vehicle_body_type       string
vehicle_make            string
street_code1            int
street_code2            int
street_code3            int
violation_precinct      int
issuer_precinct         int
violation_time          string
violation_description   string
dayslot                 string
season                  string

# Partition Information
# col_name              data_type               comment

season                  string
Time taken: 0.131 seconds, Fetched: 20 row(s)
```

Inserting data into the dynamic partition table:

```
hive> insert into table violation_cleaned_partition_ext partition(season)
    > select
    > Summons_Number, Registration_State, Issue_Date, Violation_Code,
Vehicle_Body_Type, Vehicle_Make, Street_Code1, Street_Code2, Street_Code3,
Violation_Precinct, Issuer_Precinct, Violation_Time,violation_description,
    > case
    > when CONCAT(SUBSTR(Violation_Time, 1, 2), SUBSTR(Violation_Time, 5, 1)) in (
'00A', '01A', '02A', '03A') then 'EarlyMorning'
    > when CONCAT(SUBSTR(Violation_Time, 1, 2), SUBSTR(Violation_Time, 5, 1)) in (
'04A', '05A', '06A', '07A') then 'Morning'
    > when CONCAT(SUBSTR(Violation_Time, 1, 2), SUBSTR(Violation_Time, 5, 1)) in (
'08A', '09A', '10A', '11A') then 'LateMorning'
    > when CONCAT(SUBSTR(Violation_Time, 1, 2), SUBSTR(Violation_Time, 5, 1)) in (
'12P', '01P', '02P', '03P') then 'AfterNoon'
    > when CONCAT(SUBSTR(Violation_Time, 1, 2), SUBSTR(Violation_Time, 5, 1)) in (
'04P', '05P', '06P', '07P') then 'Evening'
    > when CONCAT(SUBSTR(Violation_Time, 1, 2), SUBSTR(Violation_Time, 5, 1)) in (
'08P', '09P', '10P', '11P') then 'Night'
    > else null
    > end,
    > case
    > when SUBSTR(Issue_Date,1,2) in ( '12','01','02') then 'winter'
    > when SUBSTR(Issue_Date,1,2) in ( '03','04','05') then 'spring'
    > when SUBSTR(Issue_Date,1,2) in ( '06','07','08') then 'summer'
    > when SUBSTR(Issue_Date,1,2) in ( '09','10','11') then 'fall'
    > else null
    > end
    > from violation_table_ext where Issue_Date rlike '2017';

Query ID = root_20180624135151_8641a687-da3c-4cee-b428-5fa841b3b64f
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1529846747726_0001, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0001/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0001
Hadoop job information for Stage-1: number of mappers: 8; number of reducers: 0
2018-06-24 13:52:05,707 Stage-1 map = 0%,  reduce = 0%
2018-06-24 13:52:32,435 Stage-1 map = 6%,  reduce = 0%, Cumulative CPU 52.76 sec
2018-06-24 13:52:38,709 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 69.6 sec
2018-06-24 13:52:39,740 Stage-1 map = 38%,  reduce = 0%, Cumulative CPU 74.7 sec
2018-06-24 13:53:12,700 Stage-1 map = 63%,  reduce = 0%, Cumulative CPU 147.87 sec
2018-06-24 13:53:13,731 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 153.32 sec
2018-06-24 13:53:40,088 Stage-1 map = 88%,  reduce = 0%, Cumulative CPU 199.1 sec
2018-06-24 13:53:42,149 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 202.27 sec
MapReduce Total cumulative CPU time: 3 minutes 23 seconds 820 msec
```

```
Ended Job = job_1529846747726_0001
Stage-4 is filtered out by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is selected by condition resolver.
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1529846747726_0002, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0002/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0002
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 0
2018-06-24 13:54:03,555 Stage-5 map = 0%,  reduce = 0%
2018-06-24 13:54:21,129 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 13.48 sec
MapReduce Total cumulative CPU time: 13 seconds 550 msec
Ended Job = job_1529846747726_0002
Moving data to: s3a://hiveviolationdemo/violation_cleaned_partition_table_ext/.hive-
staging_hive_2018-06-24_13-51-54_841_294417287626062777-1/-ext-10000/season=fall
Moving data to: s3a://hiveviolationdemo/violation_cleaned_partition_table_ext/.hive-
staging_hive_2018-06-24_13-51-54_841_294417287626062777-1/-ext-10000/season=spring
Moving data to: s3a://hiveviolationdemo/violation_cleaned_partition_table_ext/.hive-
staging_hive_2018-06-24_13-51-54_841_294417287626062777-1/-ext-10000/season=winter
Loading data to table default.violation_cleaned_partition_ext partition (season=null)
        Time taken for load dynamic partitions : 4170
        Loading partition {season=fall}
        Loading partition {season=winter}
        Loading partition {season=spring}
        Loading partition {season=summer}
        Time taken for adding to write entity : 2
Partition default.violation_cleaned_partition_ext{season=fall} stats: [numFiles=1,
numRows=979, totalSize=75349, rawDataSize=74370]
Partition default.violation_cleaned_partition_ext{season=spring} stats: [numFiles=8,
numRows=2873380, totalSize=282296426, rawDataSize=279423046]
Partition default.violation_cleaned_partition_ext{season=summer} stats: [numFiles=1,
numRows=852864, totalSize=85580434, rawDataSize=84727570]
Partition default.violation_cleaned_partition_ext{season=winter} stats: [numFiles=8,
numRows=1704680, totalSize=168416224, rawDataSize=166711544]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8   Cumulative CPU: 203.82 sec   HDFS Read: 94829 HDFS Write:
2338 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 13.55 sec   HDFS Read: 4780 HDFS Write: 0
SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 37 seconds 370 msec
OK
Time taken: 161.23 seconds
```

Tables created Successfully:

```
hive> show tables;
OK
violation_cleaned_partition_ext
violation_table_ext
Time taken: 0.015 seconds, Fetched: 2 row(s)

 [ec2-user@ip-10-0-0-71 ~]$ aws s3 ls hiveviolationdemo
                           PRE violation_cleaned_partition_table_ext/
                           PRE violation_table_ext/
[ec2-user@ip-10-0-0-71 ~]$
```

Part 1: Examine the data

1.Find the total number of tickets for the year.

```
hive> select count(*) from violation_cleaned_partition_ext;
Query ID = root_20180624142121_153ddc87-1129-42b4-88e3-7789aaececba
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_1529846747726_0003, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0003/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0003
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2018-06-24 14:21:38,452 Stage-1 map = 0%,  reduce = 0%
2018-06-24 14:21:54,288 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 21.58 sec
2018-06-24 14:22:00,581 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 23.07 sec
MapReduce Total cumulative CPU time: 23 seconds 70 msec
Ended Job = job_1529846747726_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 23.07 sec   HDFS Read: 18169 HDFS
Write: 8 SUCCESS

Total MapReduce CPU Time Spent: 23 seconds 70 msec
OK
5431903
Time taken: 35.476 seconds, Fetched: 1 row(s)
```

2. Find out how many unique states the cars which got parking tickets came from

hive> **select count(distinct Registration_State) from violation_cleaned_partition_ext where Registration_State != '99';**
Query ID = root_20180624142424_ffeaecba-5a6c-46e7-ab63-61db1155b61c
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_1529846747726_0005, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0005/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0005
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2018-06-24 14:24:20,503 Stage-1 map = 0%, reduce = 0%
2018-06-24 14:24:37,182 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 12.2 sec
2018-06-24 14:24:38,213 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 24.87 sec
2018-06-24 14:24:44,436 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 26.34 sec
MapReduce Total cumulative CPU time: 26 seconds 340 msec
Ended Job = job_1529846747726_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 1   Cumulative CPU: 26.34 sec   HDFS Read: 18513 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 340 msec
OK
64
Time taken: 35.414 seconds, Fetched: 1 row(s)

3. Some parking tickets don't have addresses on them, which is cause for concern. Find out how many such tickets there are(i.e. tickets where either "Street Code 1" or "Street Code 2" or "Street Code 3" is empty)

hive> **select count(*) from violation_cleaned_partition_ext where Street_Code1=0 or Street_Code2=0 or Street_Code3=0;**
Query ID = root_20180624142626_1997a7f2-db8c-4dd0-9417-b2d6775f851b
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_1529846747726_0006, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0006/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0006
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2018-06-24 14:26:57,239 Stage-1 map = 0%,  reduce = 0%
2018-06-24 14:27:14,051 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.85 sec
2018-06-24 14:27:21,292 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 26.44 sec
MapReduce Total cumulative CPU time: 26 seconds 440 msec
Ended Job = job_1529846747726_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 26.44 sec   HDFS Read: 19920 HDFS
Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 440 msec
OK
1816814
Time taken: 34.053 seconds, Fetched: 1 row(s)
hive>
```

Part-II: Aggregation tasks

1. How often does each violation code occur? (frequency of violation codes - find the top 5)

```
hive> select violation_code, count(violation_code) as count from
violation_cleaned_partition_ext group by violation_code order by count desc limit 5;
Query ID = root_20180624142828_fd234c6b-c066-460a-a59a-0ab2919486e3
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0007, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0007/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0007
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 14:28:59,400 Stage-1 map = 0%,  reduce = 0%
2018-06-24 14:29:15,293 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 11.86 sec
2018-06-24 14:29:16,325 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.32 sec
2018-06-24 14:29:24,908 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 26.17 sec
2018-06-24 14:29:25,944 Stage-1 map = 100%,  reduce = 38%, Cumulative CPU 30.01 sec
2018-06-24 14:29:32,476 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 31.77 sec
2018-06-24 14:29:34,578 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 35.0 sec
```

2018-06-24 14:29:39,856 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 38.49 sec
MapReduce Total cumulative CPU time: 38 seconds 490 msec
Ended Job = job_1529846747726_0007
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_1529846747726_0008, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0008/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0008
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 14:29:49,365 Stage-2 map = 0%, reduce = 0%
2018-06-24 14:29:54,560 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.13 sec
2018-06-24 14:30:00,799 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.64 sec
MapReduce Total cumulative CPU time: 2 seconds 640 msec
Ended Job = job_1529846747726_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 8   Cumulative CPU: 38.49 sec   HDFS Read: 43954 HDFS Write: 2823 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 2.64 sec   HDFS Read: 9820 HDFS Write: 50 SUCCESS
Total MapReduce CPU Time Spent: 41 seconds 130 msec
OK
**21	768082**
**36	662765**
**38	542079**
**14	476660**
**20	319646**
Time taken: 72.516 seconds, Fetched: 5 row(s)
hive>

2. How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)

hive> **select vehicle_make, count(vehicle_make) as count from violation_cleaned_partition_ext group by vehicle_make order by count desc limit 5;**
Query ID = root_20180624144343_d9b18e0b-f96c-4284-a50f-76e0e7ac6b88
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0009, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0009/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0009
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 14:43:51,812 Stage-1 map = 0%,  reduce = 0%
2018-06-24 14:44:08,498 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 13.16 sec
2018-06-24 14:44:09,527 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 26.5 sec
2018-06-24 14:44:21,139 Stage-1 map = 100%,  reduce = 38%, Cumulative CPU 32.39 sec
2018-06-24 14:44:29,824 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 38.61 sec
2018-06-24 14:44:35,085 Stage-1 map = 100%,  reduce = 88%, Cumulative CPU 40.45 sec
2018-06-24 14:44:36,115 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 41.95 sec
MapReduce Total cumulative CPU time: 41 seconds 950 msec
Ended Job = job_1529846747726_0009
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_1529846747726_0010, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0010/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0010
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 14:44:44,591 Stage-2 map = 0%,  reduce = 0%
2018-06-24 14:44:50,816 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.76 sec
2018-06-24 14:44:58,073 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.19 sec
MapReduce Total cumulative CPU time: 3 seconds 190 msec
Ended Job = job_1529846747726_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 8   Cumulative CPU: 41.95 sec   HDFS Read: 43974 HDFS
Write: 75657 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.19 sec   HDFS Read: 82660 HDFS
Write: 64 SUCCESS
Total MapReduce CPU Time Spent: 45 seconds 140 msec
OK
FORD 636842
TOYOT        605290
HONDA        538884
NISSA462017
CHEVR        356032
Time taken: 76.944 seconds, Fetched: 5 row(s)
hive>
```

hive> **select vehicle_body_type, count(vehicle_body_type) as count from violation_cleaned_partition_ext group by vehicle_body_type order by count desc limit 5;**
Query ID = root_20180624144646_9f816383-6b04-4c4f-aecb-2234ce4067eb
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0011, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0011/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0011
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 14:46:18,454 Stage-1 map = 0%, reduce = 0%
2018-06-24 14:46:34,240 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 11.98 sec
2018-06-24 14:46:35,279 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 24.8 sec
2018-06-24 14:46:44,815 Stage-1 map = 100%, reduce = 25%, Cumulative CPU 28.52 sec
2018-06-24 14:46:46,956 Stage-1 map = 100%, reduce = 38%, Cumulative CPU 30.29 sec
2018-06-24 14:46:53,332 Stage-1 map = 100%, reduce = 63%, Cumulative CPU 34.62 sec
2018-06-24 14:46:54,388 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 36.27 sec
2018-06-24 14:46:59,661 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 39.61 sec
MapReduce Total cumulative CPU time: 39 seconds 610 msec
Ended Job = job_1529846747726_0011
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_1529846747726_0012, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0012/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 14:47:11,007 Stage-2 map = 0%, reduce = 0%
2018-06-24 14:47:17,204 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.63 sec
2018-06-24 14:47:23,415 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.11 sec
MapReduce Total cumulative CPU time: 3 seconds 110 msec
Ended Job = job_1529846747726_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 8   Cumulative CPU: 39.61 sec   HDFS Read: 44034 HDFS Write: 27188 SUCCESS

Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.11 sec HDFS Read: 34221 HDFS Write: 60 SUCCESS
Total MapReduce CPU Time Spent: 42 seconds 720 msec
OK
SUBN 1883953
4DSD 1547307
VAN 724025
DELV 358982
SDN 194197
Time taken: 76.653 seconds, Fetched: 5 row(s)
hive>

3. A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:
Violating Precincts (this is the precinct of the zone where the violation occurred)
Issuer Precincts (this is the precinct that issued the ticket)

hive> **select violation_precinct, count(violation_precinct) as count from violation_cleaned_partition_ext group by violation_precinct order by count desc limit 5;**

Query ID = root_20180624144848_a83d6739-d174-4125-b5ca-f3a9aa0b26e6
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0013, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0013/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0013
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 14:48:38,768 Stage-1 map = 0%, reduce = 0%
2018-06-24 14:48:54,577 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 11.65 sec
2018-06-24 14:48:55,606 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 23.91 sec
2018-06-24 14:49:06,188 Stage-1 map = 100%, reduce = 13%, Cumulative CPU 25.92 sec
2018-06-24 14:49:07,231 Stage-1 map = 100%, reduce = 38%, Cumulative CPU 29.73 sec
2018-06-24 14:49:13,715 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 31.92 sec
2018-06-24 14:49:14,768 Stage-1 map = 100%, reduce = 63%, Cumulative CPU 33.72 sec
2018-06-24 14:49:15,833 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 35.41 sec
2018-06-24 14:49:20,040 Stage-1 map = 100%, reduce = 88%, Cumulative CPU 37.14 sec
2018-06-24 14:49:21,073 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 38.71 sec
MapReduce Total cumulative CPU time: 38 seconds 710 msec
Ended Job = job_1529846747726_0013
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_1529846747726_0014, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0014/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 14:49:30,337 Stage-2 map = 0%, reduce = 0%
2018-06-24 14:49:36,538 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.12 sec
2018-06-24 14:49:42,771 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.68 sec
MapReduce Total cumulative CPU time: 2 seconds 680 msec
Ended Job = job_1529846747726_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 8   Cumulative CPU: 38.71 sec   HDFS Read: 44034 HDFS Write: 4269 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 2.68 sec   HDFS Read: 11290 HDFS Write: 48 SUCCESS
Total MapReduce CPU Time Spent: 41 seconds 390 msec
OK
**0	925596**
**19	274443**
**14	203552**
**1	174702**
**18	169131**
Time taken: 73.242 seconds, Fetched: 5 row(s)

hive> **select issuer_precinct, count(issuer_precinct) as count from violation_cleaned_partition_ext group by issuer_precinct order by count desc limit 5;**

Query ID = root_20180624145050_24a7b4be-1149-4aec-a27c-5897011faa56
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0022, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0022/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0022

```
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 14:51:05,411 Stage-1 map = 0%,  reduce = 0%
2018-06-24 14:51:30,830 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 7.72 sec
2018-06-24 14:51:37,168 Stage-1 map = 35%,  reduce = 0%, Cumulative CPU 22.35 sec
2018-06-24 14:51:42,495 Stage-1 map = 62%,  reduce = 0%, Cumulative CPU 24.96 sec
2018-06-24 14:51:43,563 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 27.73 sec
2018-06-24 14:51:46,797 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 29.57 sec
2018-06-24 14:52:04,101 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 31.49 sec
2018-06-24 14:52:05,165 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 33.79 sec
2018-06-24 14:52:07,338 Stage-1 map = 100%,  reduce = 38%, Cumulative CPU 35.73 sec
2018-06-24 14:52:21,421 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 37.6 sec
2018-06-24 14:52:22,533 Stage-1 map = 100%,  reduce = 63%, Cumulative CPU 39.57 sec
2018-06-24 14:52:24,692 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 41.67 sec
2018-06-24 14:52:37,724 Stage-1 map = 100%,  reduce = 88%, Cumulative CPU 43.62 sec
2018-06-24 14:52:38,834 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 45.48 sec
MapReduce Total cumulative CPU time: 45 seconds 480 msec
Ended Job = job_1529846747726_0022
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_1529846747726_0024, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0024/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0024
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 14:53:00,440 Stage-2 map = 0%,  reduce = 0%
2018-06-24 14:53:06,793 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.38 sec
2018-06-24 14:53:20,568 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.28 sec
MapReduce Total cumulative CPU time: 3 seconds 280 msec
Ended Job = job_1529846747726_0024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 8   Cumulative CPU: 45.48 sec   HDFS Read: 43974 HDFS
Write: 11405 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.28 sec   HDFS Read: 18408 HDFS
Write: 49 SUCCESS
Total MapReduce CPU Time Spent: 48 seconds 760 msec
OK
0       1078403
19      266959
14      200494
1       168740
18      162994
Time taken: 169.093 seconds, Fetched: 5 row(s)
hive>
```

4. Find the violation code frequency across 3 precincts which have issued the most number of tickets - do these precinct zones have an exceptionally high frequency of certain violation codes?

5. **Find out the properties of parking violations across different times of the day: The Violation Time field is specified in a strange format.**
Find a way to make this into a time attribute that you can use to divide into groups.

6. Divide 24 hours into 6 equal discrete bins of time. The intervals you choose are at your discretion. For each of these groups, find the 3 most commonly occurring violations

```
hive> Select dayslot, violation_code, viol_count
    > FROM (select dayslot, violation_code,
    > count(*) as viol_count,
    > rank() over (partition by dayslot order by count(*) desc) as row_num
    > FROM violation_cleaned_partition_ext
    > GROUP BY dayslot, violation_code
    > ) T Where row_num <= 3;
Query ID = root_20180624170303_685a239a-f436-4b8c-a363-0637edec7591
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0172, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0172/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0172
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 17:04:08,590 Stage-1 map = 0%,  reduce = 0%
2018-06-24 17:04:34,363 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 14.78 sec
2018-06-24 17:04:39,736 Stage-1 map = 29%,  reduce = 0%, Cumulative CPU 22.19 sec
2018-06-24 17:04:46,173 Stage-1 map = 56%,  reduce = 0%, Cumulative CPU 28.3 sec
2018-06-24 17:04:47,303 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 29.01 sec
2018-06-24 17:04:50,539 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 31.37 sec
2018-06-24 17:05:11,316 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 35.14 sec
2018-06-24 17:05:12,375 Stage-1 map = 100%,  reduce = 38%, Cumulative CPU 37.0 sec
2018-06-24 17:05:26,504 Stage-1 map = 100%,  reduce = 63%, Cumulative CPU 41.0 sec
2018-06-24 17:05:28,625 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 42.97 sec
2018-06-24 17:05:42,733 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 46.8 sec
MapReduce Total cumulative CPU time: 46 seconds 800 msec
Ended Job = job_1529846747726_0172
Launching Job 2 out of 2
```

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>
Starting Job = job_1529846747726_0175, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0175/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0175
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 17:06:14,609 Stage-2 map = 0%, reduce = 0%
2018-06-24 17:06:29,055 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.57 sec
2018-06-24 17:06:45,009 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.44 sec
MapReduce Total cumulative CPU time: 4 seconds 440 msec
Ended Job = job_1529846747726_0175
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 8   Cumulative CPU: 46.8 sec   HDFS Read: 45314 HDFS Write: 18457 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 4.44 sec   HDFS Read: 28346 HDFS Write: 371 SUCCESS
Total MapReduce CPU Time Spent: 51 seconds 240 msec
OK
NULL 7 3949
NULL 21 2288
NULL 40 2260
AfterNoon 36 286284
AfterNoon 38 240721
AfterNoon 37 167025
EarlyMorning 21 34703
EarlyMorning 40 23628
EarlyMorning 14 14168
Evening 38 102855
Evening 14 75902
Evening 37 70345
LateMorning 21 598060
LateMorning 36 348165
LateMorning 38 176570
Morning 14 74114
Morning 40 60652
Morning 21 57896
Night 7 26293
Night 40 22337
Night 14 21045
Time taken: 188.476 seconds, Fetched: 21 row(s)

7. Now, try another direction. For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part)

8. Let's try and find some seasonality in this data

1) First, divide the year into some number of seasons, and find frequencies of tickets for each season. (Hint: A quick Google search reveals the following seasons in NYC: Spring(March, April, March); Summer(June, July, August); Fall(September, October, November); Winter(December, January, February))

hive> **select season, count (*) as season_cnt from violation_cleaned_partition_ext group by season order by season_cnt desc;**

```
Query ID = root_20180624152626_bff355b3-0111-433a-8295-fd70f5d16810
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0065, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0065/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop
job  -kill job_1529846747726_0065
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 15:27:02,966 Stage-1 map = 0%,  reduce = 0%
2018-06-24 15:27:27,666 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 14.64 sec
2018-06-24 15:27:32,972 Stage-1 map = 25%,  reduce = 0%, Cumulative CPU 18.44 sec
2018-06-24 15:27:34,049 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 21.84 sec
2018-06-24 15:27:36,183 Stage-1 map = 75%,  reduce = 0%, Cumulative CPU 23.52 sec
2018-06-24 15:27:37,253 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.88 sec
2018-06-24 15:27:53,567 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 26.67 sec
2018-06-24 15:27:54,664 Stage-1 map = 100%,  reduce = 38%, Cumulative CPU 30.47 sec
2018-06-24 15:28:12,128 Stage-1 map = 100%,  reduce = 63%, Cumulative CPU 34.69 sec
2018-06-24 15:28:14,250 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 36.78 sec
2018-06-24 15:28:27,153 Stage-1 map = 100%,  reduce = 88%, Cumulative CPU 38.69 sec
2018-06-24 15:28:28,250 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 40.56 sec
MapReduce Total cumulative CPU time: 40 seconds 560 msec
Ended Job = job_1529846747726_0065
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_1529846747726_0066, Tracking URL = http://ip-10-0-0-
71.ec2.internal:8088/proxy/application_1529846747726_0066/
```

Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0066
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 15:29:00,053 Stage-2 map = 0%, reduce = 0%
2018-06-24 15:29:13,302 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.54 sec
2018-06-24 15:29:28,203 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.66 sec
MapReduce Total cumulative CPU time: 3 seconds 660 msec
Ended Job = job_1529846747726_0066
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 8 Cumulative CPU: 40.56 sec HDFS Read: 43444 HDFS Write: 877 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.66 sec HDFS Read: 7689 HDFS Write: 53 SUCCESS
Total MapReduce CPU Time Spent: 44 seconds 220 msec
OK
spring 2873380
winter 1704680
summer 852864
fall 979
Time taken: 180.995 seconds, Fetched: 4 row(s)

2) **Then, find the 3 most common violations for each of these seasons.**

hive> **Select season, violation_code, viol_count**
> FROM (select season, violation_code,
> count(*) as viol_count,
> rank() over (partition by season order by count(*) desc) as row_num
> FROM violation_cleaned_partition_ext
> GROUP BY season, violation_code)T Where row_num <= 3;
Query ID = root_20180624164242_3e89ab82-b612-4e44-9219-88f75405ad63
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 8
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_1529846747726_0150, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0150/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0150
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 8
2018-06-24 16:43:22,671 Stage-1 map = 0%, reduce = 0%
2018-06-24 16:43:52,979 Stage-1 map = 11%, reduce = 0%, Cumulative CPU 15.47 sec
2018-06-24 16:43:54,100 Stage-1 map = 19%, reduce = 0%, Cumulative CPU 18.61 sec
2018-06-24 16:43:58,470 Stage-1 map = 30%, reduce = 0%, Cumulative CPU 21.17 sec
2018-06-24 16:44:00,631 Stage-1 map = 39%, reduce = 0%, Cumulative CPU 24.09 sec

2018-06-24 16:44:03,849 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 26.29 sec
2018-06-24 16:44:06,058 Stage-1 map = 79%, reduce = 0%, Cumulative CPU 29.61 sec
2018-06-24 16:44:07,160 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 30.0 sec
2018-06-24 16:44:23,685 Stage-1 map = 100%, reduce = 13%, Cumulative CPU 31.85 sec
2018-06-24 16:44:25,933 Stage-1 map = 100%, reduce = 38%, Cumulative CPU 35.58 sec
2018-06-24 16:44:40,502 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 37.5 sec
2018-06-24 16:44:42,664 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 41.41 sec
2018-06-24 16:44:54,950 Stage-1 map = 100%, reduce = 88%, Cumulative CPU 43.32 sec
2018-06-24 16:44:56,052 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 45.18 sec
MapReduce Total cumulative CPU time: 45 seconds 180 msec
Ended Job = job_1529846747726_0150
Launching Job 2 out of 2
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>
Starting Job = job_1529846747726_0151, Tracking URL = http://ip-10-0-0-71.ec2.internal:8088/proxy/application_1529846747726_0151/
Kill Command = /opt/cloudera/parcels/CDH-5.14.0-1.cdh5.14.0.p0.24/lib/hadoop/bin/hadoop job -kill job_1529846747726_0151
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-24 16:45:27,671 Stage-2 map = 0%, reduce = 0%
2018-06-24 16:45:41,606 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.52 sec
2018-06-24 16:45:56,567 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 4.42 sec
MapReduce Total cumulative CPU time: 4 seconds 420 msec
Ended Job = job_1529846747726_0151
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 8   Cumulative CPU: 45.18 sec   HDFS Read: 45270 HDFS Write: 9782 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1   Cumulative CPU: 4.42 sec   HDFS Read: 19665 HDFS Write: 187 SUCCESS
Total MapReduce CPU Time Spent: 49 seconds 600 msec
OK
fall 46 231
fall 21 128
fall 40 116
spring 21 402424
spring 36 344834
spring 38 271167
summer 21 127350
summer 36 96663
summer 38 83518
winter 21 238180
winter 36 221268
winter 38 187386
Time taken: 186.447 seconds, Fetched: 12 row(s)
hive>