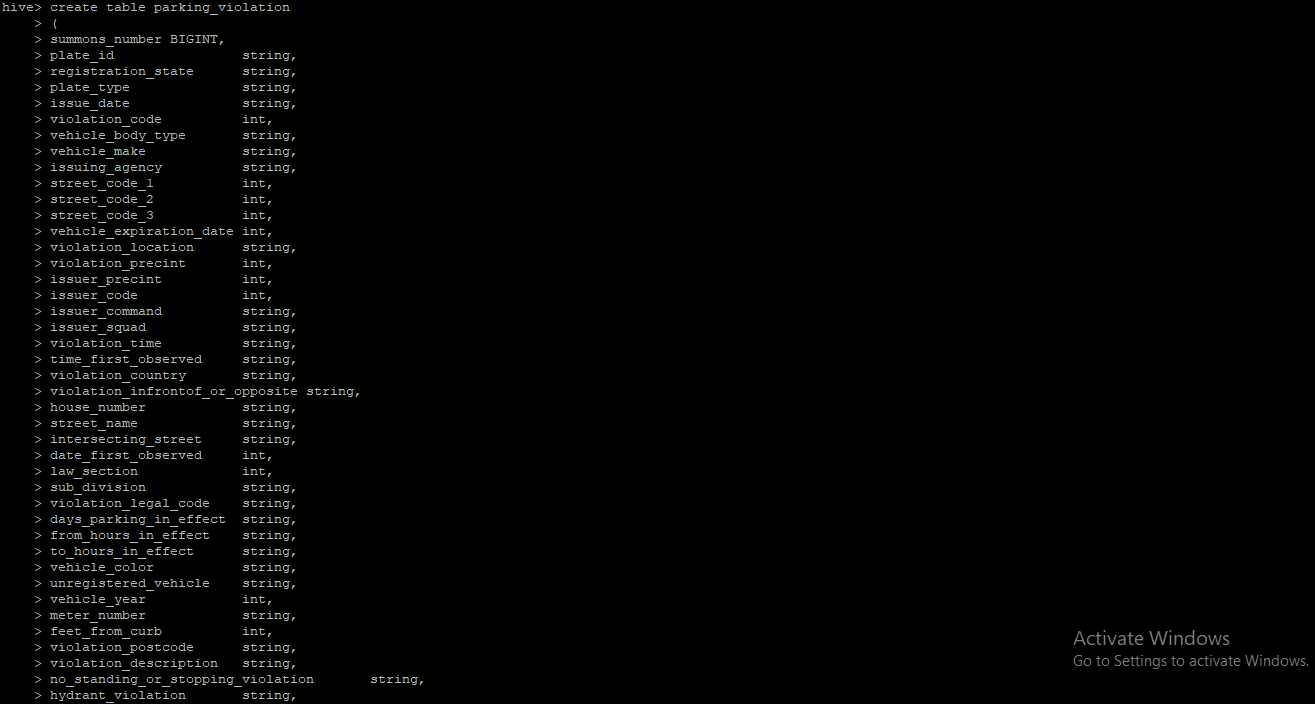
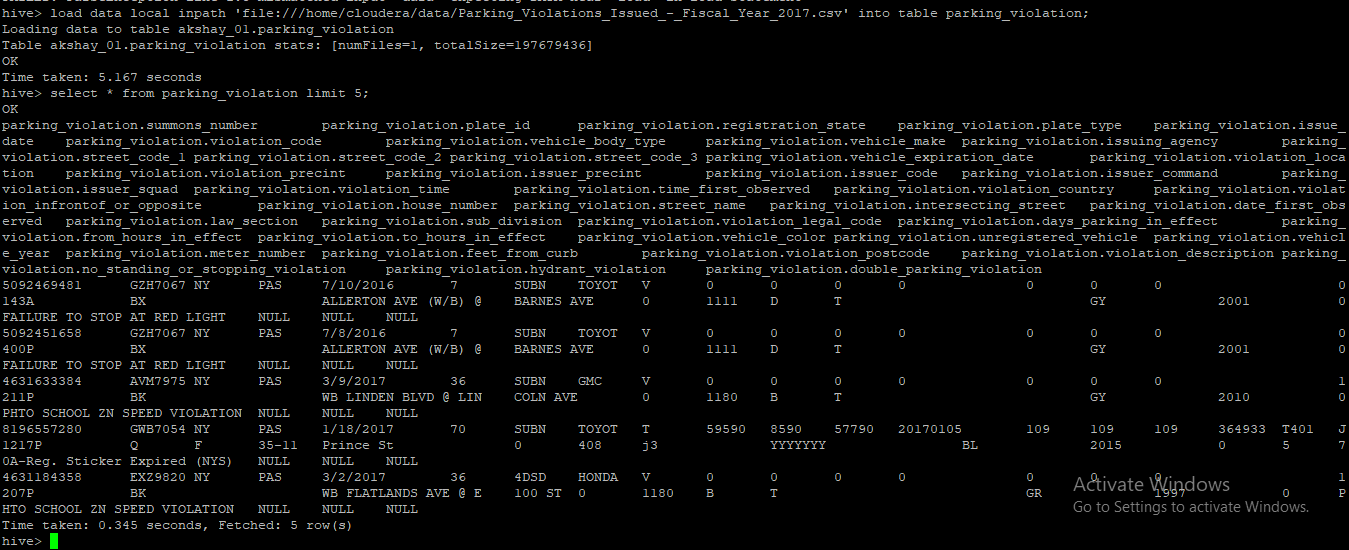
Note: Consider only the year 2017 for analysis and not the Fiscal year.



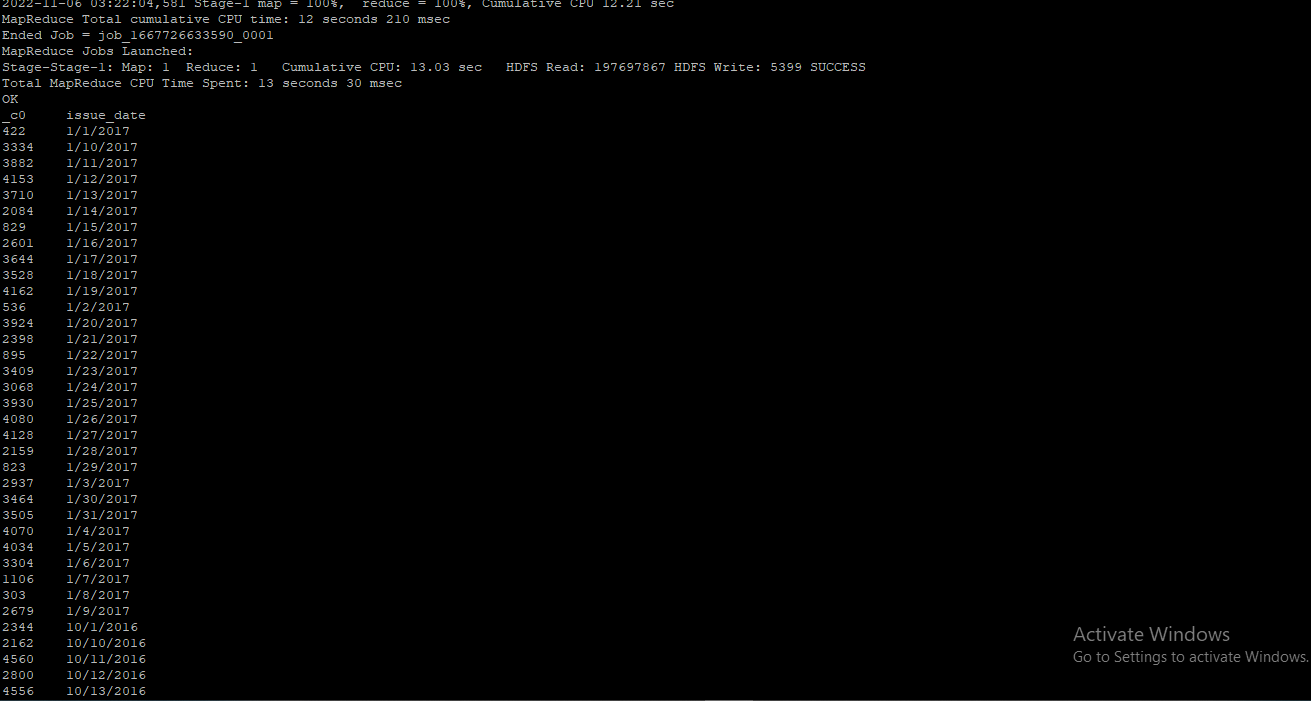


The analysis can be divided into two parts:

Part-I: Examine the data

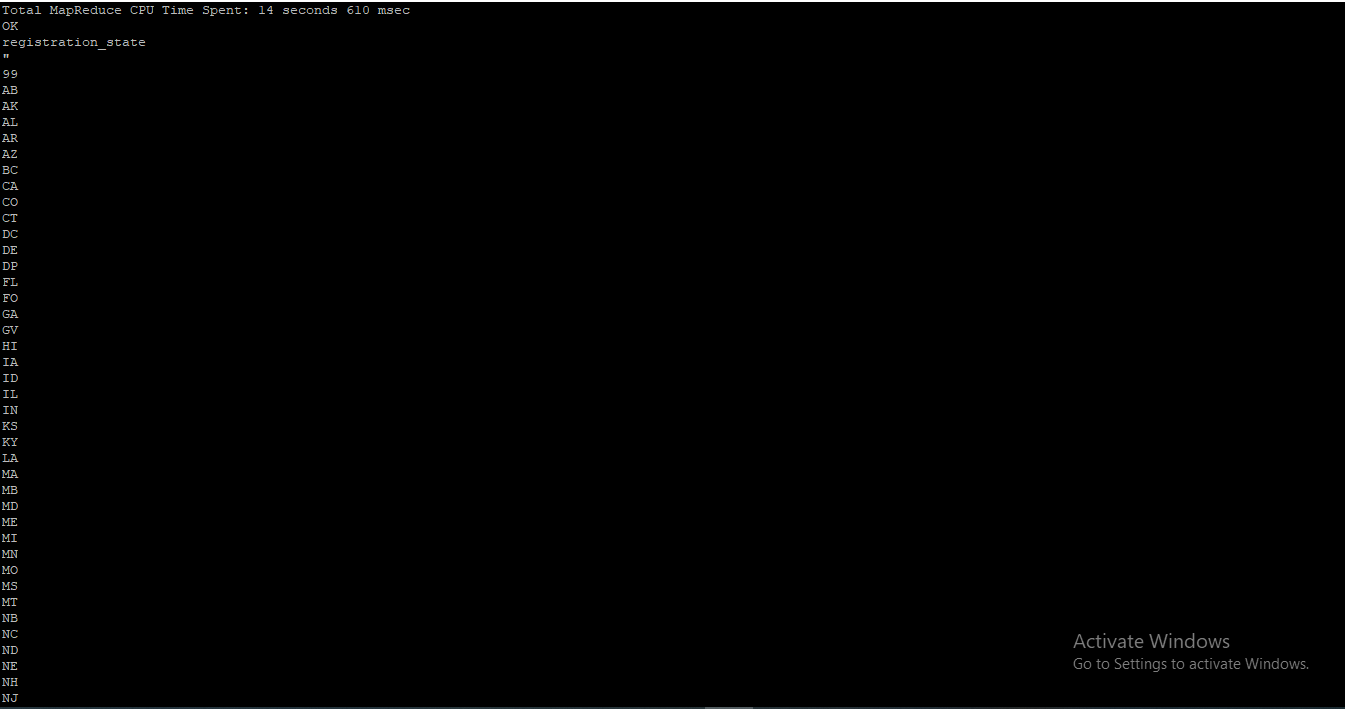
1. Find the total number of tickets for the year.

select count(summons\_number), issue\_date from parking\_violation group by issue\_date;



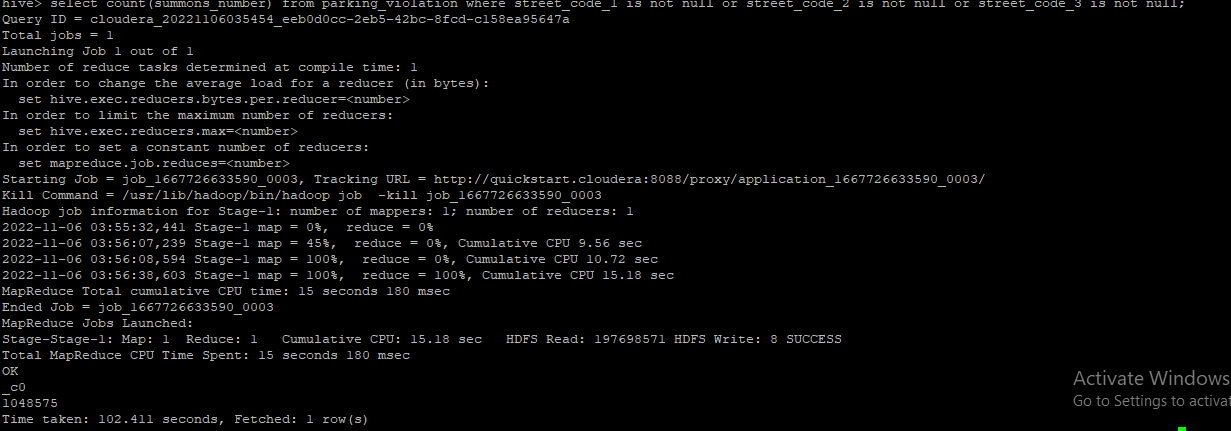
1. Find out how many unique states the cars which got parking tickets came from.

select distinct(registration\_state) from parking\_violation where summons\_number is not null;



1. 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 )

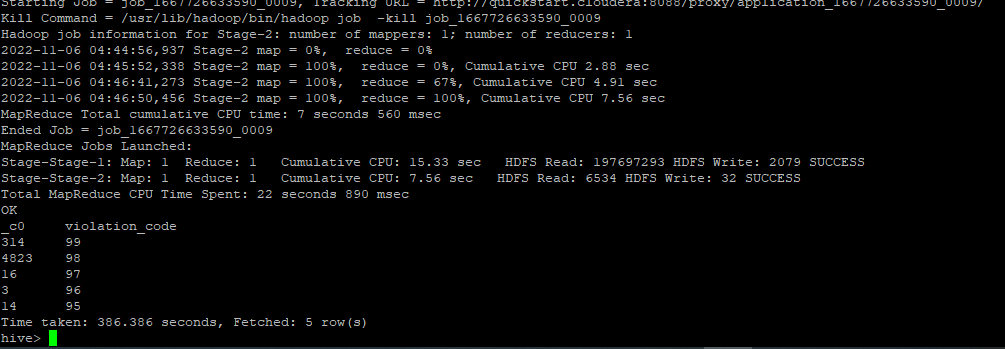
select count(summons\_number) from parking\_violation where street\_code\_1 is not null or street\_code\_2 is not null or street\_code\_3 is not null



Part-II: Aggregation tasks

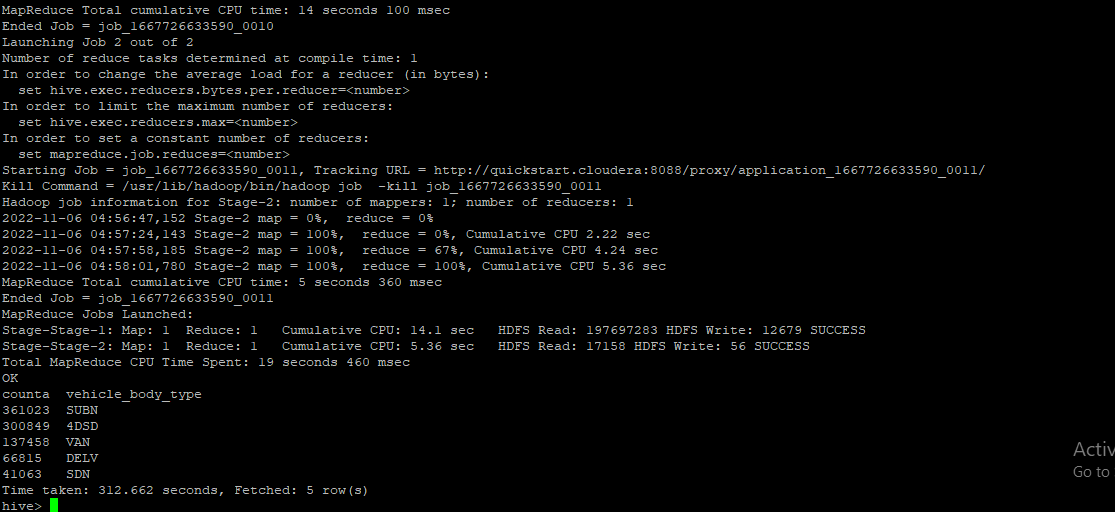
1. How often does each violation code occur? (frequency of violation codes - find the top 5)

select count(violation\_code),violation\_code from parking\_violation group by violation\_code order by violation code desc limit 5;

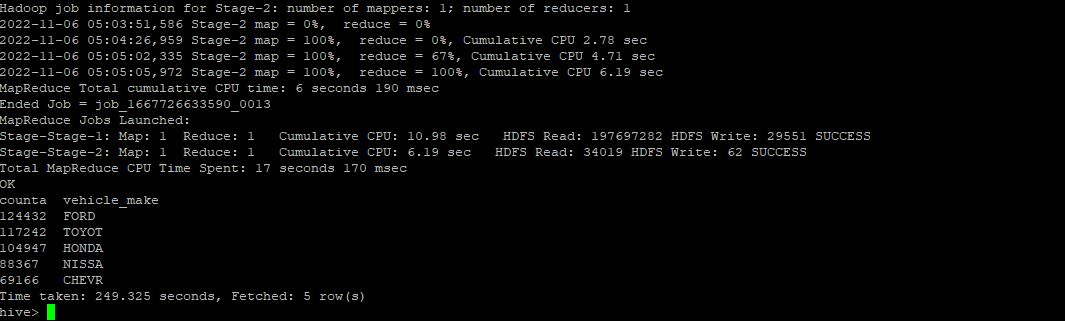


1. How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)

select count(summons\_number)AS countA,vehicle\_body\_type from parking\_violation group by vehicle\_body\_type order by countA desc limit 5;



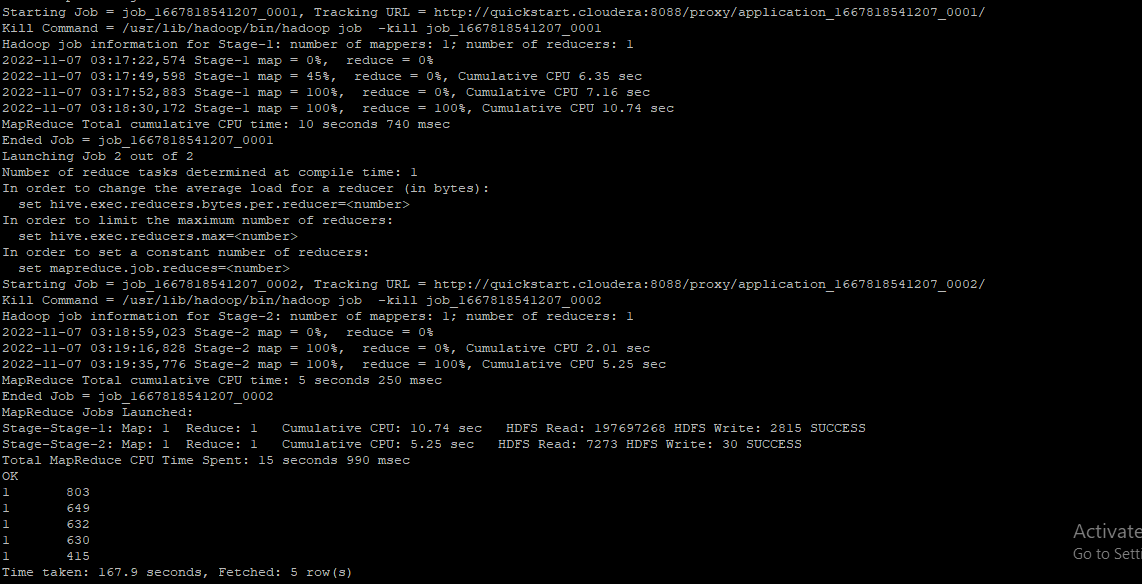
select count(summons\_number)AS countA,vehicle\_make from parking\_violation group by vehicle\_make order by countA desc limit 5;



3.) A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:

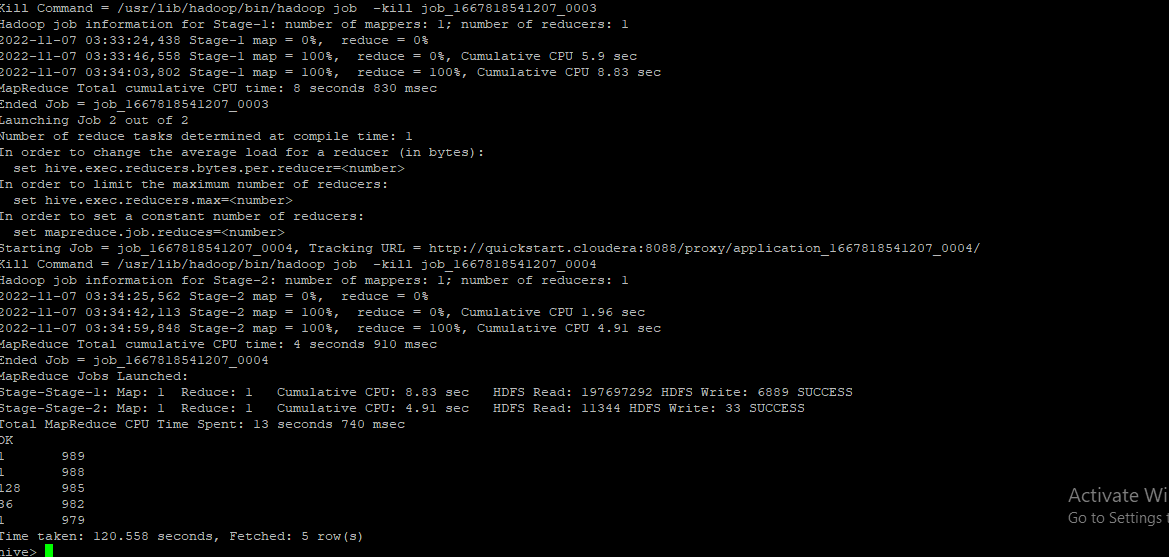
a.) Violating Precincts (this is the precinct of the zone where the violation occurred)

select count(violation\_precint), violation\_precint from parking\_violation group by violation\_precint order by violation\_precint desc limit 5;



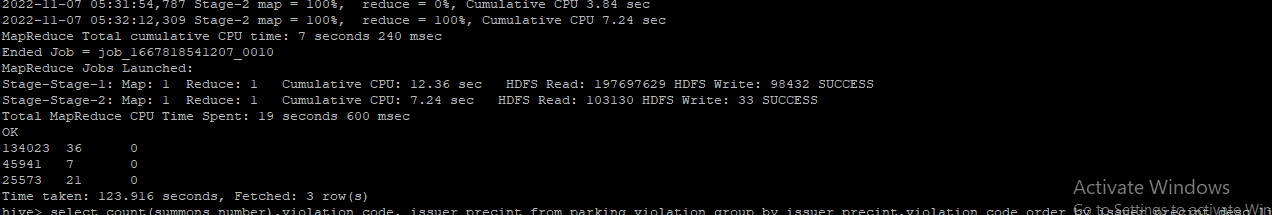
b.) Issuer Precincts (this is the precinct that issued the ticket)

select count(issuer\_precint), issuer\_precint from parking\_violation group by issuer\_precint order by issuer\_precint desc limit 5;



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?

select count(summons\_number),violation\_code, issuer\_precint from parking\_violation group by issuer\_precint,violation\_code order by issuer\_precint desc limit 3;



Issuer\_precint with value 0 has issued maximum number of tickets. They have violation codes of 36,7,21 with a frequency of 1 each.

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

To divide the 24 hrs into 6 equal discrete bins of time , First created view on partitioned table:

hive> create view vw\_parking\_violations\_partitoned\_2017

partitioned on (Violation\_Code)

as

SELECT

Summons\_Number,

Violation\_Time,

Issuer\_Precinct,

case

when substring(Violation\_Time,1,2) in ('00','01','02','03','12') and upper(substring(Violation\_Time,-1))='A' then 1

when substring(Violation\_Time,1,2) in ('04','05','06','07') and upper(substring(Violation\_Time,-1))='A' then 2

when substring(Violation\_Time,1,2) in ('08','09','10','11') and upper(substring(Violation\_Time,-1))='A' then 3

when substring(Violation\_Time,1,2) in ('12','00','01','02','03') and upper(substring(Violation\_Time,-1))='P' then 4

when substring(Violation\_Time,1,2) in ('04','05','06','07') and upper(substring(Violation\_Time,-1))='P' then 5

when substring(Violation\_Time,1,2) in ('08','09','10','11') and upper(substring(Violation\_Time,-1))='P'then 6

else null

end as Violation\_Time\_bin,

Violation\_Code

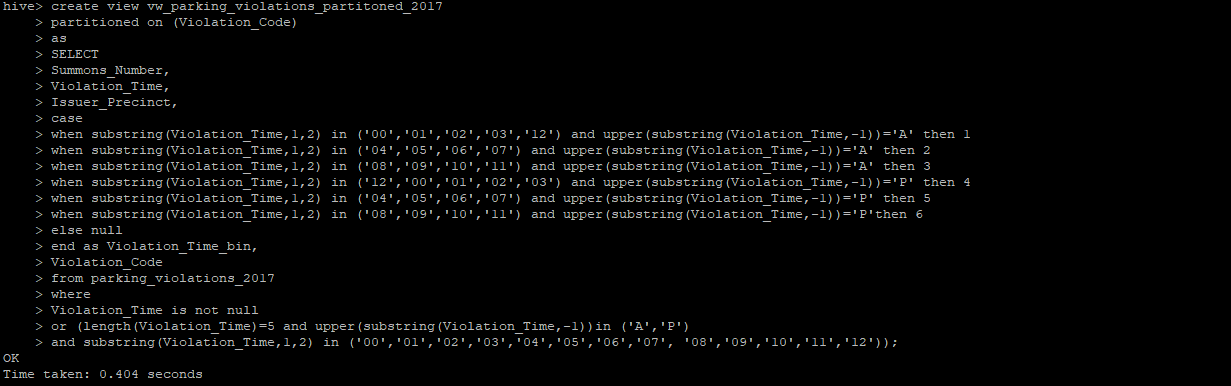
from parking\_violations\_2017

where

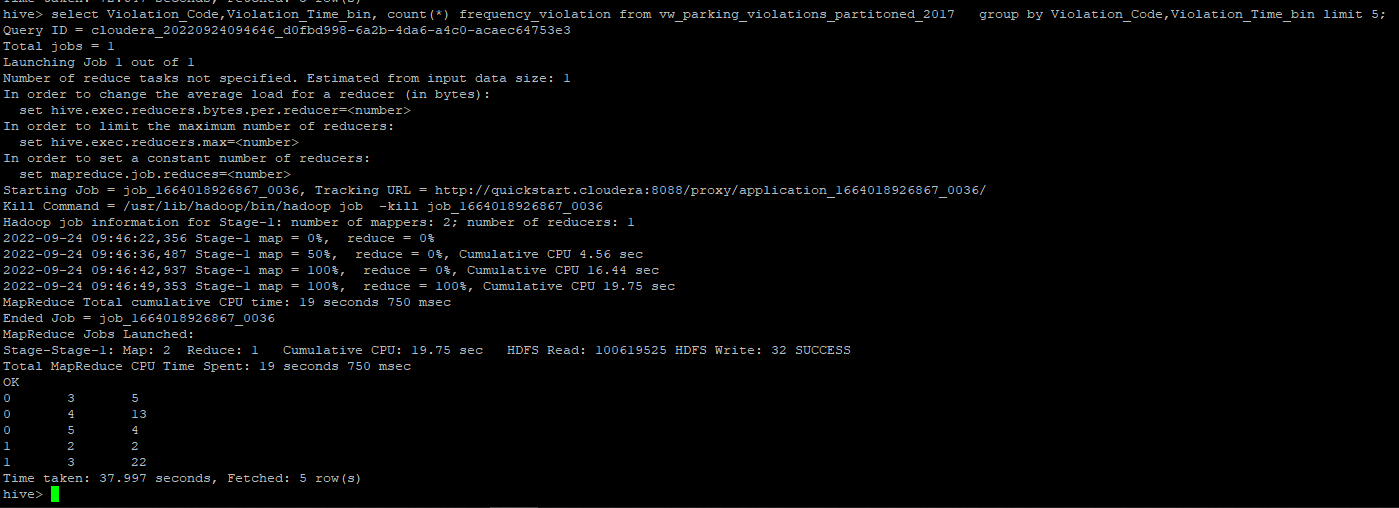
Violation\_Time is not null

or (length(Violation\_Time)=5 and upper(substring(Violation\_Time,-1))in ('A','P')

and substring(Violation\_Time,1,2) in ('00','01','02','03','04','05','06','07', '08','09','10','11','12'));

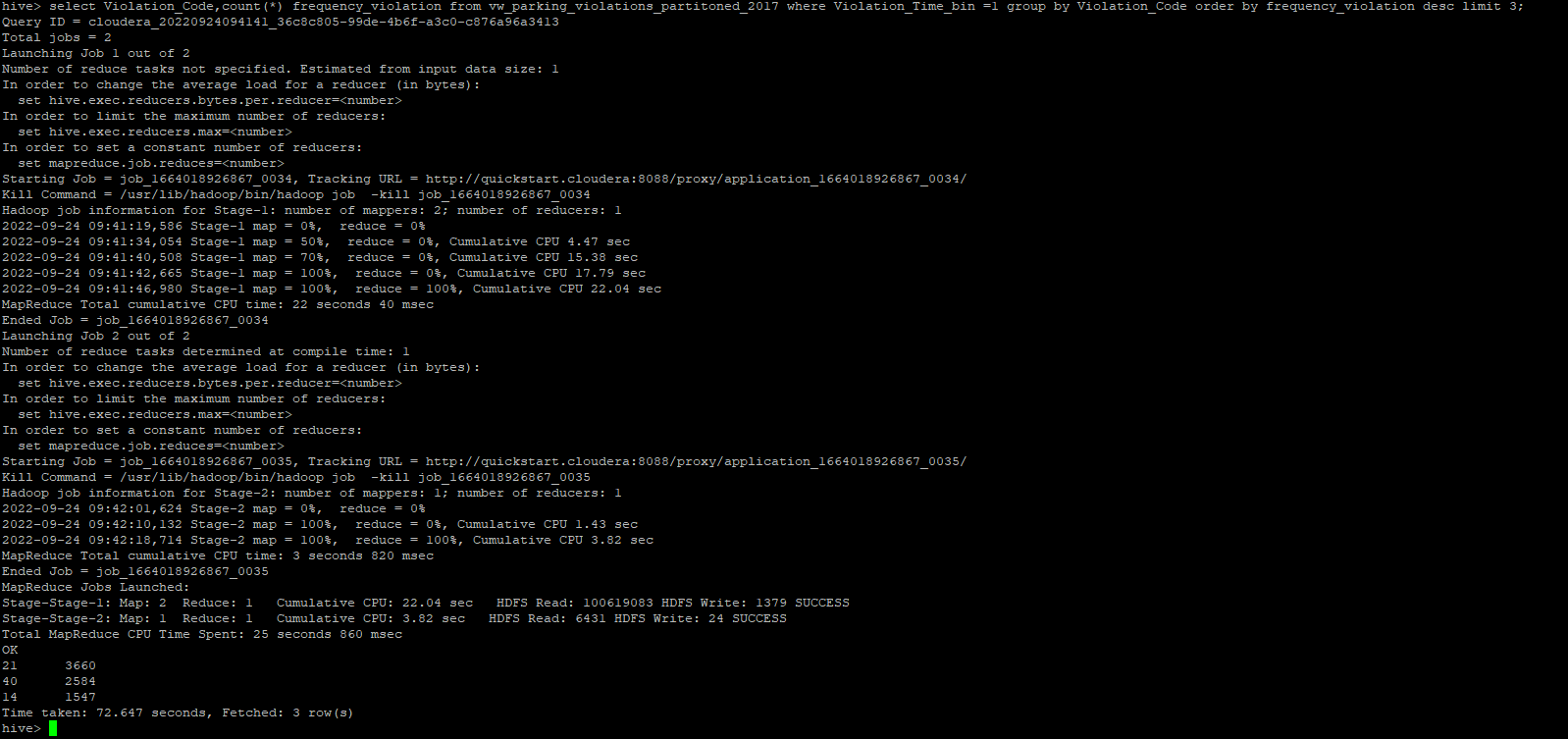


select Violation\_Code,Violation\_Time\_bin, count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 group by Violation\_Code,Violation\_Time\_bin limit 5;



🡺bin1

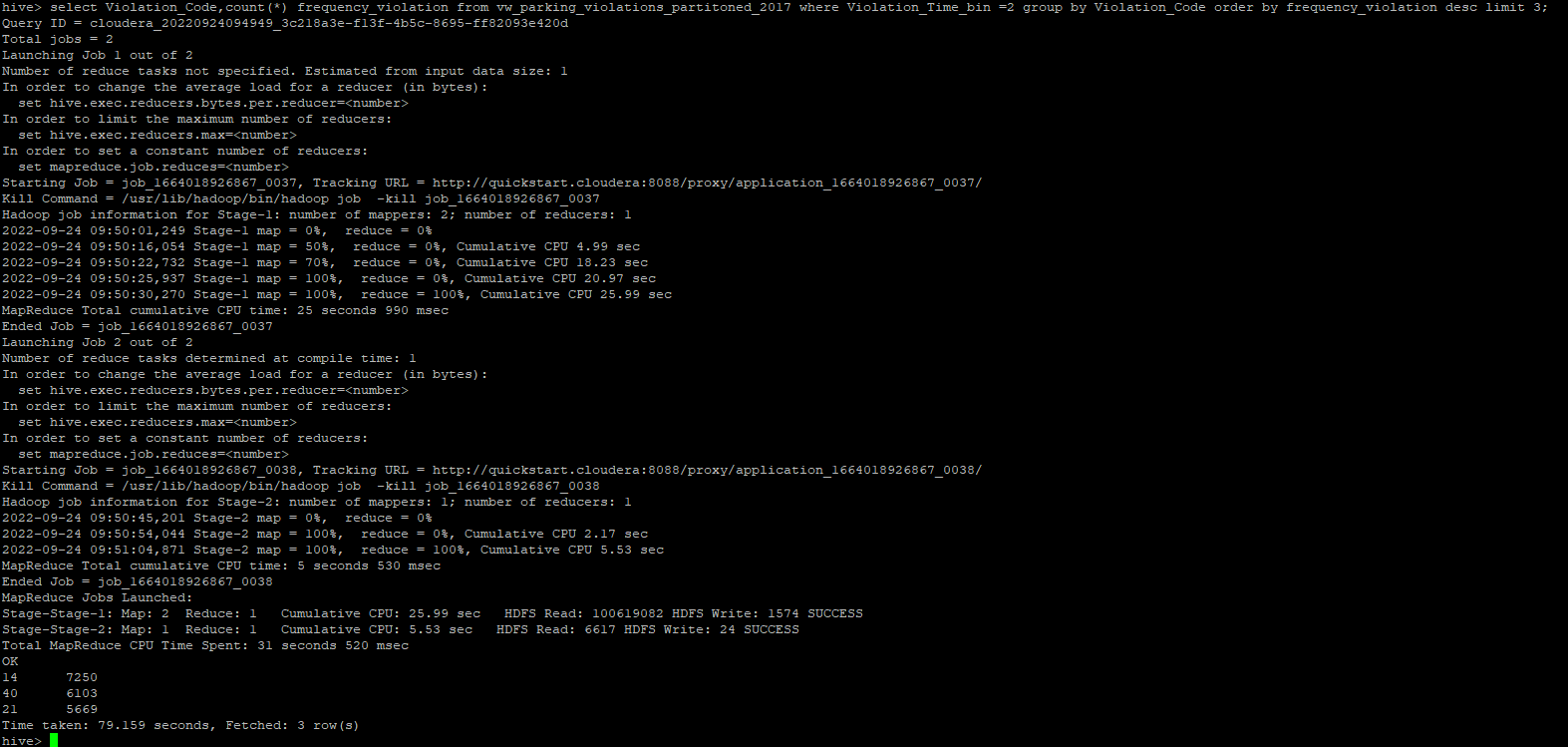
select Violation\_Code,count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 where Violation\_Time\_bin =1 group by Violation\_Code order by frequency\_violation desc limit 3;b



|  |  |
| --- | --- |
| **Violation\_code** | **frequency\_violation** |
| **21** | **3660** |
| **40** | **2584** |
| **14** | **1574** |

🡺bin 2

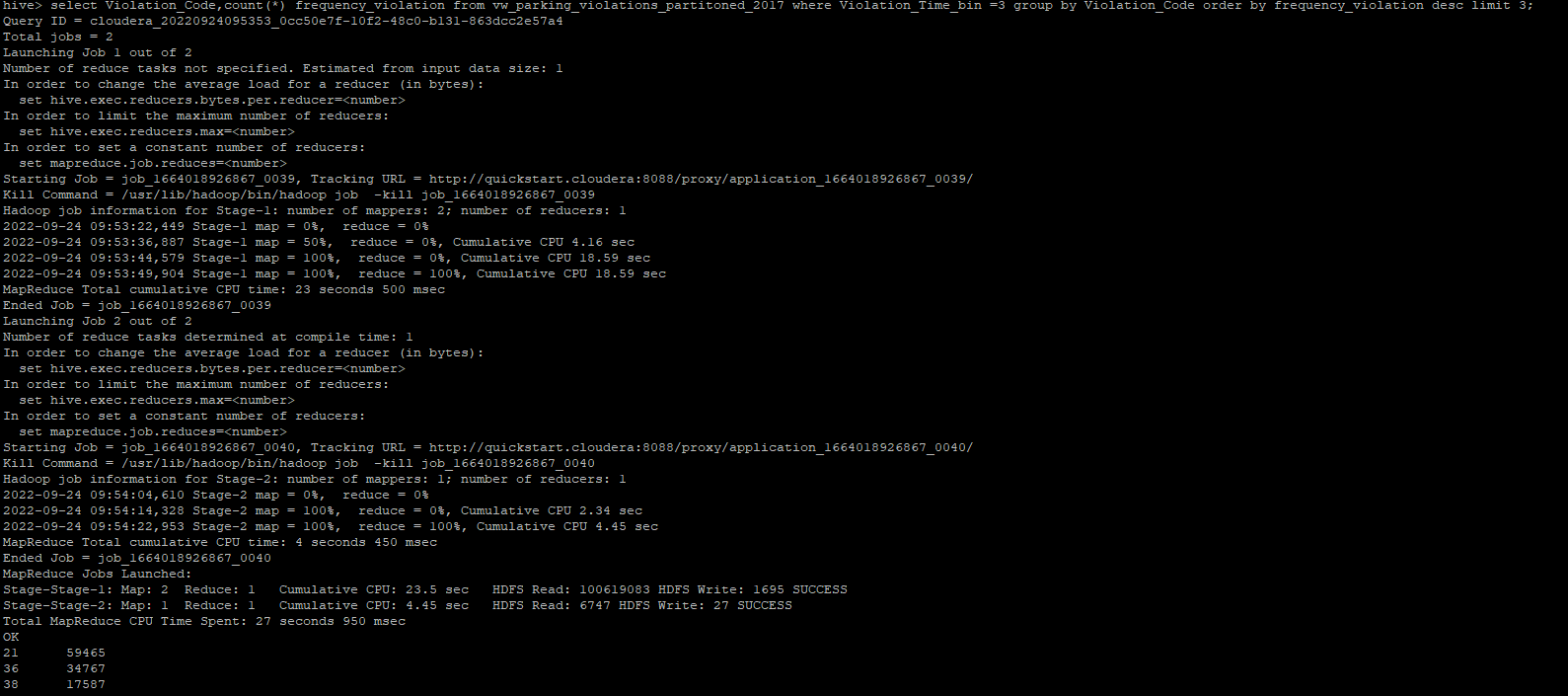
select Violation\_Code,count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 where Violation\_Time\_bin =2 group by Violation\_Code order by frequency\_violation desc limit 3;



|  |  |
| --- | --- |
| Violation\_code | frequency\_violation |
| 14 | 7250 |
| 40 | 6403 |
| 21 | 5669 |

🡺bin 3

select Violation\_Code,count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 where Violation\_Time\_bin =3 group by Violation\_Code order by frequency\_violation desc limit 3;



|  |  |
| --- | --- |
| Violation\_code | frequency\_violation |
| 21 | 59465 |
| 36 | 37767 |
| 38 | 17587 |

🡺bin 4

select Violation\_Code,count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 where Violation\_Time\_bin =4 group by Violation\_Code order by frequency\_violation desc limit 3;

|  |  |
| --- | --- |
| Violation\_code | frequency\_violation |
| 36 | 28600 |
| 38 | 23877 |
| 37 | 16777 |

🡺bin 5

select Violation\_Code,count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 where Violation\_Time\_bin =5 group by Violation\_Code order by frequency\_violation desc limit 3;

|  |  |
| --- | --- |
| Violation\_code | frequency\_violation |
| 38 | 10148 |
| 14 | 7609 |
| 37 | 6944 |

🡺bin 6

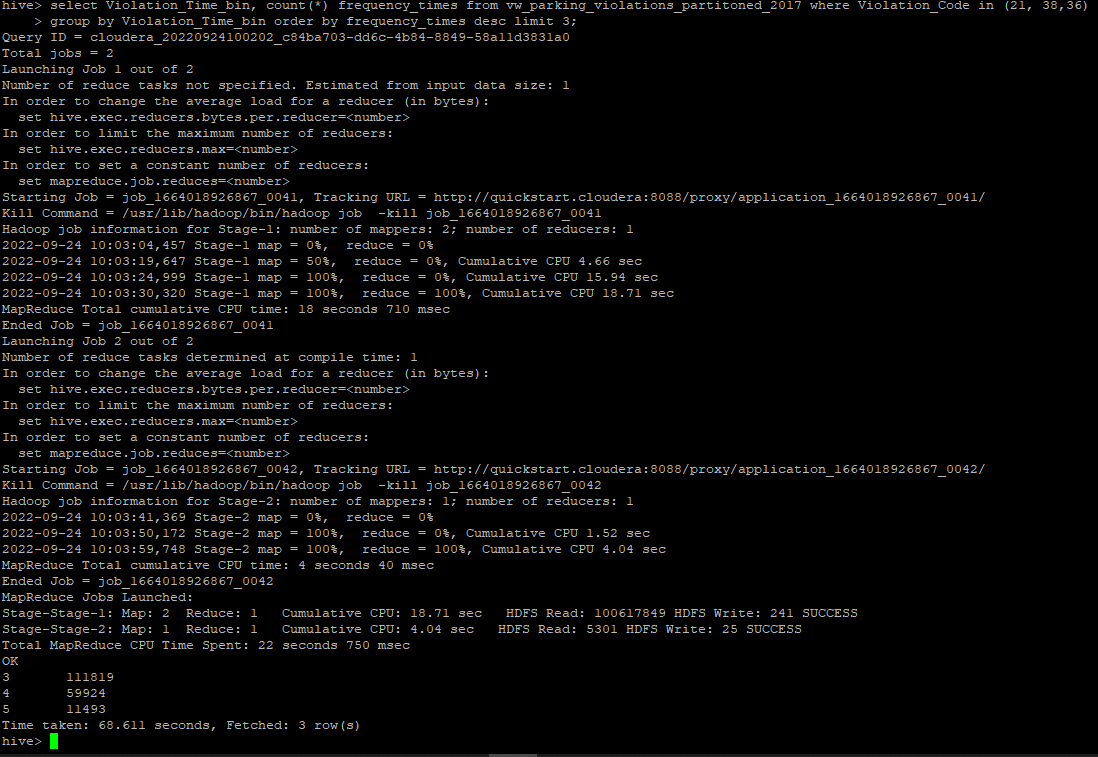
select Violation\_Code,count(\*) frequency\_violation from vw\_parking\_violations\_partitoned\_2017 where Violation\_Time\_bin =6 group by Violation\_Code order by frequency\_violation desc limit 3;

|  |  |
| --- | --- |
| Violation\_code | frequency\_violation |
| 7 | 2602 |
| 40 | 2159 |
| 14 | 2091 |

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)

Hive> select Violation\_Time\_bin, count(\*) frequency\_times from vw\_parking\_violations\_partitoned\_2017 where Violation\_Code in (21, 38,36)

group by Violation\_Time\_bin order by frequency\_times desc limit 3;



|  |  |
| --- | --- |
| Violation\_Time\_bin | frequency\_times |
| 3 | 116785 |
| 4 | 76701 |
| 5 | 18437 |

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

a.) 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))

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

🡺Created View to divide the year into Seasons:

Hive> create view vw\_tickets\_issued\_2017

partitioned on (Violation\_Code)

as

select Issuer\_Precinct,

case

when MONTH(Issue\_Date) between 03 and 05 then 'spring'

when MONTH(Issue\_Date) between 06 and 08 then 'summer'

when MONTH(Issue\_Date) between 09 and 11 then 'autumn'

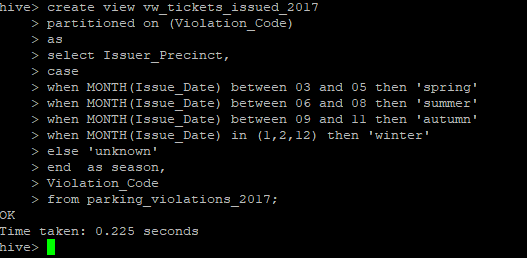
when MONTH(Issue\_Date) in (1,2,12) then 'winter'

else 'unknown'

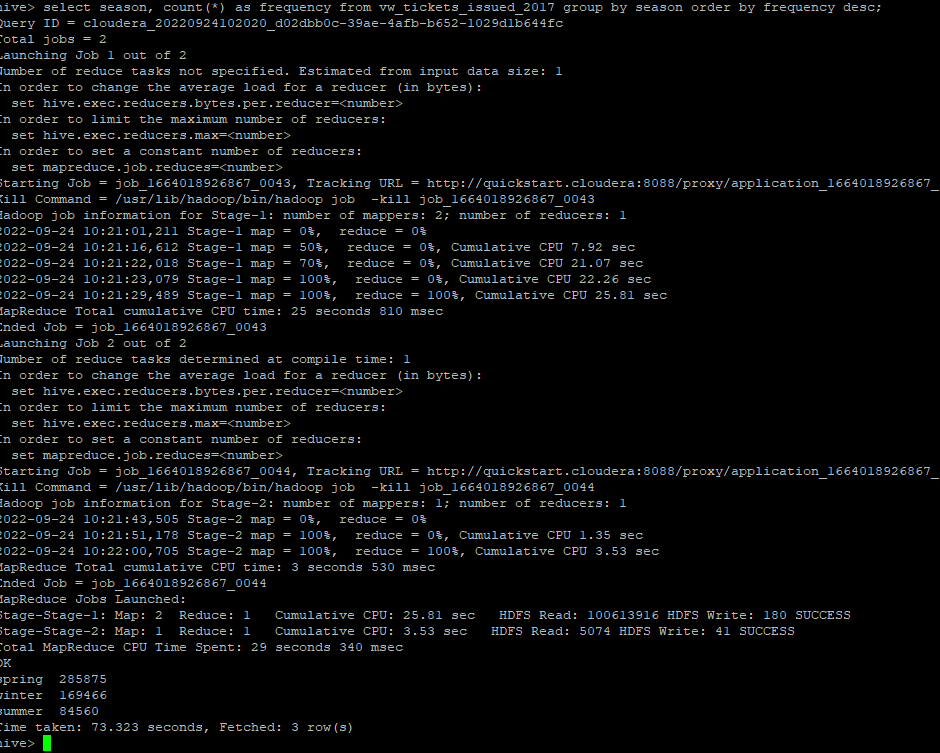
end as season,

Violation\_Code

from parking\_violations\_2017;



Hive> select season, count(\*) as frequency from vw\_tickets\_issued\_2017 group by season order by frequency desc;



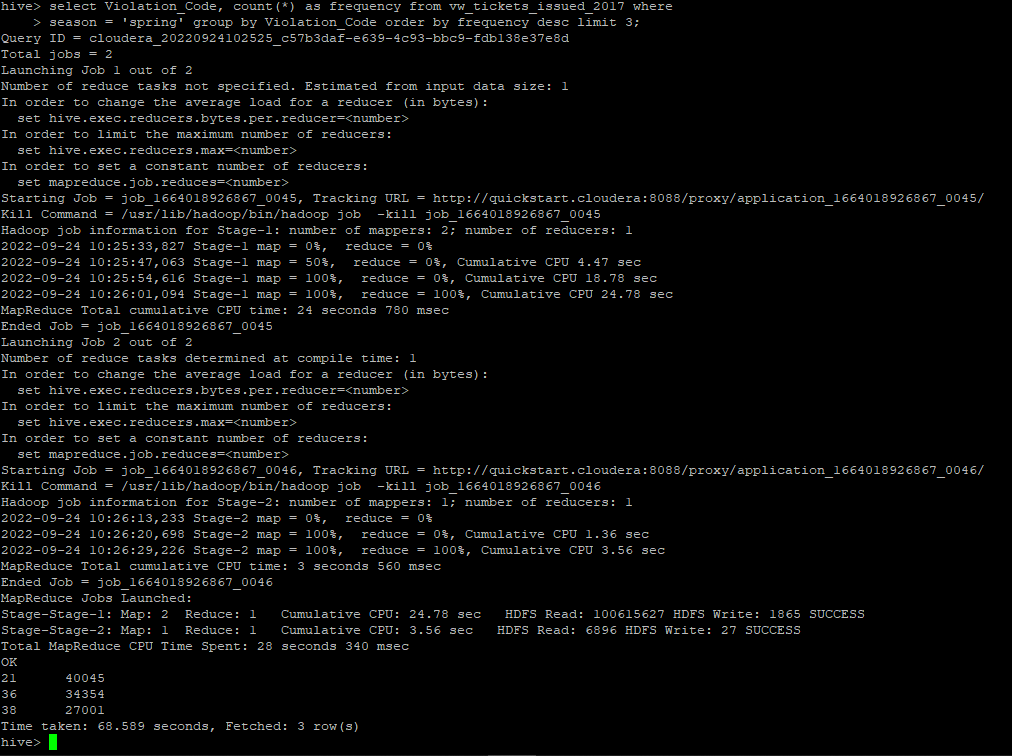
|  |  |
| --- | --- |
| Season | frequency |
| Spring | 285875 |
| Winter | 169466 |
| Summer | 84560 |
| autumn | 0 |

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

🡺spring season

select Violation\_Code, count(\*) as frequency from vw\_tickets\_issued\_2017 where

season = 'spring' group by Violation\_Code order by frequency desc limit 3;



|  |  |
| --- | --- |
| Violation\_Code | frequency |
| 21 | 40045 |
| 36 | 34354 |
| 38 | 27001 |

🡺winter season

select Violation\_Code, count(\*) as frequency from vw\_tickets\_issued\_2017 where

season = 'winter' group by Violation\_Code order by frequency desc limit 3;

|  |  |
| --- | --- |
| Violation\_Code | frequency |
| 21 | 23684 |
| 36 | 22084 |
| 38 | 18450 |

🡺 summer season

select Violation\_Code, count(\*) as frequency from vw\_tickets\_issued\_2017 where

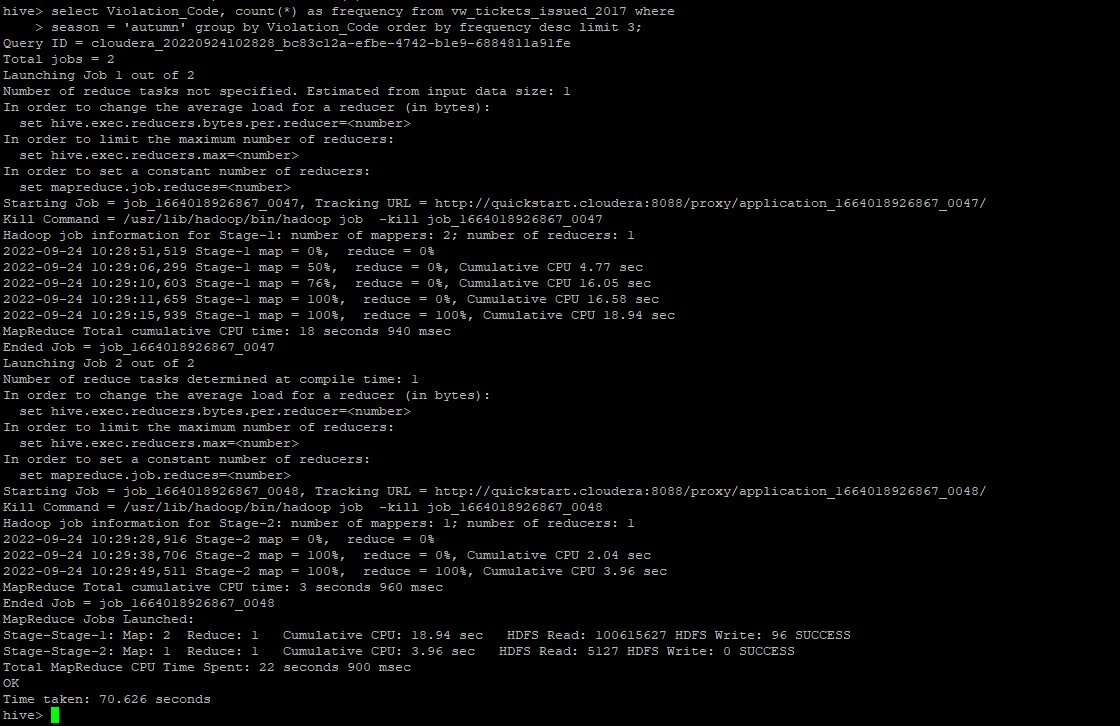
season = 'summer' group by Violation\_Code order by frequency desc limit 3;

|  |  |
| --- | --- |
| Violation\_Code | frequency |
| 21 | 12565 |
| 36 | 9655 |
| 38 | 8331 |

🡺 autumn season

select Violation\_Code, count(\*) as frequency from vw\_tickets\_issued\_2017 where

season = 'autumn' group by Violation\_Code order by frequency desc limit 3;



|  |  |
| --- | --- |
| Violation\_Code | frequency |
|  |  |

Note: Please ensure you make necessary optimizations to your queries like selecting the appropriate table format, using partitioned/bucketed tables. Marks will be awarded for keeping the performance also in mind.

