Capstone queries :

WITH TEM AS (

select count(\*) CNT,pickup\_latitude,pickup\_longitude from lpep2015\_triprecord\_vts

group by pickup\_latitude,pickup\_longitude)

SELECT MAX(CNT) FROM TEM WHERE CNT IS NOT NULL

select count(\*) CNT,pickup\_longitude,pickup\_latitude from lpep2015\_triprecord\_vts

where pickup\_longitude IS NOT NULL

group by pickup\_longitude,pickup\_latitude

order by 1 desc

--- all the tables in the tables

select \* from lpep2015\_breadcrumb\_cmt

select \* from lpep2015\_breadcrumb\_vts

select \* from lpep2015\_triprecord\_cmt

select \* from lpep2015\_triprecord\_vts

--- view creation

create view temp\_vts as

select trip\_type,substr(cast(pickup\_latitude as string),1,8) as lat,

substr(cast(pickup\_longitude as string),1,9) as longitude,taximeter\_pickup\_datetime ,

YEAR(taximeter\_pickup\_datetime) AS YR,HOUR(taximeter\_pickup\_datetime) as HR,

MINUTE(taximeter\_pickup\_datetime) as MI,shl\_number,hack\_number

from lpep2015\_triprecord\_vts

where pickup\_latitude!=0 or pickup\_longitude!=0

create view temp\_vts\_byhr as

select count(\*) as cnt,lat,longitude,hr,mi from temp\_vts

group by lat,longitude,hr,mi

order by cnt desc

--- creating view of breadcrumb vts table

create view breadcrumb\_vts as

select shl\_number,hack\_number,substr(cast(latitude as string),1,8) as lat,

substr(cast(longitude as string),1,9) as longitude,

YEAR(lpep\_datetime) AS YR,MONTH(lpep\_datetime) as MON,HOUR(lpep\_datetime) as HR,

MINUTE(lpep\_datetime) as MI

from lpep2015\_breadcrumb\_vts

where latitude!=0 or longitude!=0

select substr(cast(latitude as string),1,8) as lat,

substr(cast(longitude as string),1,9) as longitude,

DAY(lpep\_datetime) DAY\_1,

TO\_DATE(lpep\_datetime) Da,MONTH(lpep\_datetime) MO,

HOUR(lpep\_datetime) as HR,

MINUTE(lpep\_datetime) as MI,shl\_number,hack\_number

from lpep2015\_breadcrumb\_vts

where (DAY(lpep\_datetime) between 7 and 13)

--- table columns :

|  |  |
| --- | --- |
| [lpep2015\_breadcrumb\_vts.shl\_number](javascript:void(0)) | STRING\_TYPE |
| [lpep2015\_breadcrumb\_vts.hack\_number](javascript:void(0)) | STRING\_TYPE |
| [lpep2015\_breadcrumb\_vts.lpep\_datetime](javascript:void(0)) | STRING\_TYPE |
| [lpep2015\_breadcrumb\_vts.latitude](javascript:void(0)) | FLOAT\_TYPE |
| [lpep2015\_breadcrumb\_vts.longitude](javascript:void(0)) | FLOAT\_TYPE |
| [lpep2015\_breadcrumb\_vts.type](javascript:void(0)) | TINYINT\_TYPE |

|  |
| --- |
| select bbl\_index, count(distinct shl\_number) as cnt from lpep2015\_breadcrumb\_total |
| where count!=0 |
| group by bbl\_index; |
|  |
| to subset wav lpep cmt breadcrumb data |
|  |
| select shl\_number,lpep\_datetime,latitude,longitude from lpep2015\_breadcrumb\_cmt |
| where lpep2015\_breadcrumb\_cmt.shl\_number in (select shl\_number from lpep\_wav); |
|  |
| **/\* create wheel acess vehicle data from tpep and lpep for a select medallion number \*/** |
| **LPEP breadcrumb wav : cab 1930 :count : 126 million** |
| create table lpep\_wav\_tot |
| select shl\_number,lpep\_datetime,latitude,longitude from lpep2015\_breadcrumb\_cmt |
| where lpep2015\_breadcrumb\_cmt.shl\_number in (select distinct(shl\_number) from lpep\_wav) |
| union all |
| select shl\_number,lpep\_datetime,latitude,longitude from lpep2015\_breadcrumb\_vts |
| where lpep2015\_breadcrumb\_vts.shl\_number in (select distinct(shl\_number) from lpep\_wav); |
|  |
| **To save the table to csv** |
| create table csv\_dump\_bhagwat ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' as |
| select \* from tlc.lpep\_wav\_tot; |
|  |
| hadoop fs -cat /user/hive/warehouse/csv\_dump\_bhagwat/\* > ~/csv\_dump\_latest.csv |
| drop table csv\_dump\_bhagwat; |
|  |
| **TPEP breadcrumb wav : cab : 892 : count : 187 million** |
| create table tpep\_wav\_tot as |
| select medallion,crumb\_tpep\_datetime as tpep\_datetime,latitude,longitude from tpep2015\_breadcrumb\_vts |
| where tpep2015\_breadcrumb\_vts.medallion in (select distinct(medallion) from tpep\_wav) |
| union all |
| select medallion,tpep\_datetime,latitude,longitude from tpep2015\_breadcrumb\_cmt |
| where tpep2015\_breadcrumb\_cmt.medallion in (select distinct(medallion) from tpep\_wav); |
|  |
|  |
| **LPEP triprecord wav : cab 1932 :count : 2841977** |
|  |
| create table lpep\_wav\_triprecord\_total as |
| select shl\_number,lpep\_pickup\_datetime as pickup\_datetime,lpep\_dropoff\_datetime as dropoff\_datetime ,pickup\_latitude as latitude,pickup\_longitude as longitude |
| from lpep2015\_triprecord\_cmt |
| where lpep2015\_triprecord\_cmt.shl\_number in (select distinct(shl\_number) from lpep\_wav) |
| union all |
| select shl\_number,taximeter\_pickup\_datetime as pickup\_datetime,taximeter\_dropoff\_datetime as dropoff\_datetime ,pickup\_latitude as latitude,pickup\_longitude as longitude |
| from lpep2015\_triprecord\_vts |
| where lpep2015\_triprecord\_vts.shl\_number in (select distinct(shl\_number) from lpep\_wav); |
|  |
| **TPEP triprecord wav : cab : 892 :count : 7812304** |
|  |
| create table tpep\_wav\_triprecord\_total as |
| select medallion,tpep\_pickup\_datetime as pickup\_datetime,tpep\_dropoff\_datetime as dropoff\_datetime ,pickup\_latitude as latitude,pickup\_longitude as longitude |
| from tpep2015\_triprecord\_cmt |
| where tpep2015\_triprecord\_cmt.medallion in (select distinct(medallion) from tpep\_wav) |
| union all |
| select medallion,tpep\_pickup\_datetime as pickup\_datetime,tpep\_dropoff\_datetime as dropoff\_datetime, pickup\_latitude as latitude,pickup\_longitude as longitude |
| from tpep2015\_triprecord\_vts |
| where tpep2015\_triprecord\_vts.medallion in (select distinct(medallion) from tpep\_wav); |

Analysis :

For CMT LPEP time is the more reliable timestamp while for VTS taximeter time is more reliable.

For vts taxitrip cnt is taken for values >20

And for breadcrumb vts cnt is taken for values >10