**Seattle Air Pollution Dataset for Pollution Control using Apache Hive**

hive> show databases;

hive> show databases;

OK

default

Time taken: 0.271 seconds, Fetched: 1 row(s)

hive>

hive> create database Seattle\_AirPollution;

hive> create database Seattle\_AirPollution;

OK

Time taken: 0.215 seconds

hive>

hive> show databases;

hive> show databases;

OK

default

Seattle\_AirPollution

Time taken: 0.112 seconds, Fetched: 2 row(s)

hive>

hive>create database if not exists Seattle\_AirPollution;

hive>create database if not exists Seattle\_AirPollution;

OK

Time talen: 0.107 seconds

hive>create database Seattle\_AirPollution

>WITH DBPROPERTIES ( ‘Owner’=’Krishna Sindhu’ , ‘date’=’2021-08-23’); --DBPROPERTIES takes multiple arguments in form of Key-Value pair.

hive>describe database extended Seattle\_AirPollution;

hive>describe database extended Seattle\_AirPollution;

OK

Seattle\_AirPollution USER {creator=Krishna Sindhu, date=2021-08-23 }

Time taken: 0.228 seconds, Fetched: 1 row(s)

hive>

hive>create table Seattle.AirPollution (date int, month int, year int, dow int, pm1 Double, temp int, stagno int, admyng I int, gasyng int)

> row format delimited

> fields terminated by ‘,’ ;

hive>create table Seattle.AirPollution.Pollution\_data (date int, month int, year int, dow int, pm1 Double, temp int, stagno int, admyng I int, gasyng int)

>roe format delimited

>fields terminated by ‘,’

OK

Time taken: 0.461 seconds

hive>

hive>describe Seattle\_AirPollution.Pollution\_data

hive>describe Seattle\_AirPollution.Pollution\_data

OK

date int

month int

year int

dow int

pm1 Double

temp int

stagno int

admyng int

gasyng int

Time taken: 0.37 seconds, Fetched: 9 row(s)

hive>

hive>local data local inpath ‘C:\Users\dksin\Dropbox\My PC (LAPTOP-E383CH2K)\Desktop’ into table Seattle\_AirPollution.Pollution\_data;

hive>local data local inpath ‘C:\Users\dksin\Dropbox\My PC (LAPTOP-E383CH2K)\Desktop\hive\Seattle\_data’ into table Seattle\_AirPollution.Pollution\_data;

Loading data to table Seattle\_AirPollution.Pollution\_data

Table Seattle\_AirPollution.Pollution\_data stats: [numFiles=1, numRows=0, totalSize=50, rawDataSize=0]

OK

Time taken: 7.505 seconds

hive>

hive> select \* from Seattle\_AirPollution.Pollution\_data;

hive> select \* from Seattle\_AirPollution.Pollution\_data;

OK

10000 5 87 3 5.549286979 52 7 2 0

10001 5 87 4 10.27945433 55 7.5 1 1

10002 5 87 5 7.933674644 59 10.5 1 1

10003 5 87 6 11.1673955 55 8 6 1

10004 5 87 7 8.795092998 58 7.5 4 2

10005 5 87 1 8.308595578 58 11 3 1

10006 5 87 2 8.632475678 53 2.5 6 1

Time taken: 3.32 seconds, Fetched: 7 row(s)

hive>

hive>create table if not exista Seattle\_AirPollution

>like Seattle\_AirPollution.copy\_of\_Polluted\_data;

hive>create table if not exista Seattle\_AirPollution

>like Seattle\_AirPollution.copy\_of\_polluted\_data;

OK

Time taken: 0.606 seconds

hive>

hive>describe copy\_of\_polluted\_data;

hive>describe copy\_of\_polluted\_data;

OK

date int

month int

year int

dow int

pm1 Double

temp int

stagno int

admyng int

gasyng int

Time taken: 0.97 seconds, Fetched: 9 row(s)

hive>

hive>show tables;

hive>show tables;

OK

Pollution\_data

Copy\_of\_polluted\_data

Time taken: 17.3 seconds

hive>

hive>drop table copy\_of\_polluted\_data;

hive>show tables;

hive>show tables;

OK

Pollution\_data

Time taken: 0.37 seconds

>hive

hive>use Seattle\_AirPollution;

hive>create table partitioned\_table (date int, month int, dow int, pm1 Double, temp int,stagno int,admyng int, gasyng

int)

>partitioned by(year int)

>row format delimited

>fields terminated by ‘,’;

hive>create table partitioned\_table (date int, month int, dow int, pm1 Double, temp int,stagno int,admyng int, gasyng

int)

>partitioned by(year int)

>row format delimited

>fields terminated by ‘,’;

OK

Time taken: 3.402 seconds

hive>

hive>describe partitioned\_table;

hive>describe partitioned\_table;

OK

Date int

Month int

Dow int

Pm1 double

Temp int

Stagno int

Admyng int

Gasyng int

#Partition information

#col\_name data\_type comment

Year int

Time taken: 12.56 seconds, Fetched: 14 row(s)

hive>

hive>load data local inpath ‘’ C:\Users\dksin\Dropbox\My PC (LAPTOP-E383CH2K)\Desktop \hive\Seattle\_data“ into table partitioned\_table partition (year=87);

hive>load data local inpath ‘’ C:\Users\dksin\Dropbox\My PC (LAPTOP-E383CH2K)\Desktop \hive\Seattle\_data“ into table partitioned\_table

>partition(year=87);

Loading data to table Seattle\_AirPollution.partitioned\_table partition (year=87)

OK

Time taken: 9.765 seconds

hive>

hive>load data local inpath “ C:\Users\dksin\Dropbox\My PC (LAPTOP-E383CH2K)\Desktop \hive\seattle\_Data” into table partitioned\_table partition (year=93);

hive>load data local inpath “ C:\Users\dksin\Dropbox\My PC (LAPTOP-E383CH2K)\Desktop \hive\seattle\_Data” into

table partitioned\_table

> partition (year=93);

Loading data to table Seattle\_AirPollution.partitioned\_table partition (year=87)

OK

Time taken: 8.085 seconds

hive>

hive>select \* from partitioned\_table;

hive>select \* from partitioned\_table;

OK

12222 6 93 6 4.35597661521393 67 5 9 1

12223 6 93 7 4.45525676827392 61 3.3 5 0

12224 6 93 1 4.02667236515958 63 3.3 4 1

12225 6 93 2 3.65804746665376 56 1.7 5 1

12226 6 93 3 2.35249047449037 55 1.7 6 1

12227 6 93 4 4.20924189884728 58 2.7 8 2

12228 6 93 5 4.52170929312068 62 4.7 2 0

12328 10 93 7 15.8073524313881 65 11 6 0

12329 10 93 1 21.0572013736568 65 10.3 8 1

12341 10 93 6 10.0758106784738 54 10.7 6 0

12342 10 93 7 11.3527795580989 55 12 10 1

12344 10 93 2 17.3584722594861 52 3 6 1

10000 5 87 3 5.549286979 52 7 2 0

10001 5 87 4 10.27945433 55 7.5 1 1

10002 5 87 5 7.933674644 59 10.5 1 1

10003 5 87 6 11.1673955 55 8 6 1

10004 5 87 7 8.795092998 58 7.5 4 2

10197 12 87 4 9.882118978 48 3.5 4 1

Time taken: 9.123 seconds, Fetched: 19 row(s)

hive>

hive>use Seattle\_AirPollution

hive>select year, avg(pm1) from partitioned\_table group by year having year=87 and year=93;

hive>select year, avg(pm1) from partitioned\_table group by year having year=87 and year=93;

OK

87 1237.987

93 627.097

Time taken: 987.654 sdeconds, Fetched: 4 row(s)

hive>

hive>use Seattle\_AirPollution;

hive>select \* from partitioned\_table order by temp desc;

hive>Drop table partitioned\_table;

hive>show tables;

hive>show tables;

OK

Pollution\_data

Copy\_of\_polluted\_data

Time take: 0.112 seconds , Fetched: 2 row(s)

hive>