HIVE

1.1 Objective:

Identify the data of trees in the New York city

1.2 Dataset:

Source: NYC Open data

'https://github.com/sushilsani/HIVE/blob/master/input/tree_data.txt'

Data Format: csv, Number of records: 9259, File size 577 KB

1.3 Approach:

a. Copy the file to Ubuntu file system using WinScp
b. In Hive, create a table to match the dataset
c. Load the file into the table
d. Run a HIVE Query to select the required data
e. Write the output to a file

1.4 Hive Commands:

Hive command

```
create database tree_info;
create table tree(RecordId INT, Address STRING, House_Number STRING, Street STRING, Zip_Original INT, Site STRING, Diameter INT, Condition STRING, Sidewalk_Condition STRING, Spc_Common STRING, Spc_Latin STRING, Spc_Combined STRING)
row format delimited
fields terminated by ','
stored as textfile;
LOAD DATA LOCAL INPATH '/home/sushil/datasets/tree_data.txt' OVERWRITE INTO TABLE tree;
```

Hive Execution

```
sushil@singlegen1:~$ hive
WARNING: org.apache.hadoop.metrics.jvm.EventCounter is deprecated. Please use org.apache.hadoop.log.metrics.EventCounter in all the log4j.properties files.
Logging initialized using configuration in jar:file:/home/sushil/hive-0.9.0-bin/lib/hive-common-0.9.0.jar!/hive-log4j.properties
Hive history file=/tmp/sushil/hive_job_log_sushil_201606021330_2013680688.txt
hive> create database tree_info;
OK
Time taken: 0.09 seconds
hive> create table tree(RecordId INT, Address STRING, House_Number STRING, Street STRING, Zip_Original INT, Site STRING, Diameter INT, Condition STRING, Sidewalk_Condition STRING, Spc_Common STRING, Spc_Latin STRING, Spc_Combined STRING)
    > row format delimited
    > fields terminated by ','
```

```
   > stored as textfile;
OK
Time taken: 0.262 seconds
hive> LOAD DATA LOCAL INPATH '/home/sushil/datasets/tree_data.txt' OVERWRITE INTO TABLE tree;
Copying data from file:/home/sushil/datasets/tree_data.txt
Copying file: file:/home/sushil/datasets/tree_data.txt
Loading data to table default.tree
Deleted hdfs://localhost:8020/user/hive/warehouse/tree
OK
Time taken: 0.69 seconds
hive> select Spc_combined,count(*) from tree group by spc_combined;
Total MapReduce jobs = 1
Launching Job 1 out of 1
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 mapred.reduce.tasks=<number>
Starting Job = job_201606021330_0001, Tracking URL = http://localhost:50030/jobd
etails.jsp?jobid=job_201606021330_0001
Kill Command = /home/sushil/hadoop-1.2.1/libexec/../bin/hadoop job  -Dmapred.job
.tracker=hdfs://localhost:8021 -kill job_201606021330_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2016-06-02 13:36:41,685 Stage-1 map = 0%,  reduce = 0%
2016-06-02 13:36:44,722 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:45,734 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:46,752 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:47,771 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:48,793 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:49,829 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:50,857 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:51,878 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2016-06-02 13:36:52,890 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 1.4 sec
2016-06-02 13:36:53,904 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.35 sec
2016-06-02 13:36:54,912 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.35 sec
2016-06-02 13:36:55,931 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.35 sec
MapReduce Total cumulative CPU time: 3 seconds 350 msec
Ended Job = job_201606021330_0001
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 3.35 sec   HDFS Read: 1123257 HDFS Write: 1686 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 350 msec
OK
AMUR CORKTREE (PHELLODENDRON AMURENSE)  1
APPLE (MALUS PUMILA)    32
ASH GREEN (FRAXINUS PENNSYLVANICA)      222
BEECH AMERICAN (FAGUS GRANDIFOLIA)      6
BIRCH OTHER (BETULA SPECIES)    11
BLACK LOCUST (ROBINIA PSEUDOACACIA)     8
BLACKGUM (NYSSA SYLVATICA)      10
```

```
CHERRY BLACK (PRUNUS SEROTINA)  21
CHERRY CORNELIAN (CORNUS MAS)   176
CHERRY OTHER (PRUNUS SPECIES)   57
EASTERN HOP HORNBEAM (OSTRYA VIRGINIANA)        96
ELM AMERICAN (ULMUS AMERICANA)  307
ELM OTHER (ULMUS SPECIES)       34
FIR OTHER (ABIES SPECIES)       35
GINKGO (GINKGO BILOBA)  1107
HACKBERRY (CELTIS OCCIDENTALIS) 72
HAWTHORN OTHER (CRATAEGUS SPECIES)      109
HONEYLOCUST (GLEDITSIA TRIACANTHOS)     1931
ILEX (ILEX SPECIES)     3
JAPANESE PAGODA TREE (SOPHORA JAPONICA) 392
LINDEN LITTLE LEAF (TILIA CORDATA)      699
LINDEN OTHER (TILIA SPECIES)    49
LONDON PLANETREE (PLATANUS ACERIFOLIA)  933
MAGNOLIA OTHER (MAGNOLIA SPECIES)       1
MAPLE NORWAY (ACER PLATANOIDES) 100
MAPLE OTHER (ACER SPECIES)      12
MAPLE RED (ACER RUBRUM) 19
MAPLE SILVER (ACER SACCHARINUM) 10
MAPLE SUGAR (ACER SACCHARUM)    26
MAPLE SYCAMORE (ACER PSEUDOPLATANUS)    36
MULBERRY WHITE (MORUS ALBA)     1
OAK NORTHERN RED (QUERCUS RUBRA)        78
OAK OTHER (QUERCUS SPECIES)     8
OAK PIN (QUERCUS PALUSTRIS)     209
OAK WILLOW (QUERCUS PHELLOS)    127
PEAR CALLERY (PYRUS CALLERYANA) 1496
PINE OTHER (PINUS SPECIES)      4
PLANTING SITE (PLANTING SITE)   184
REDWOOD COAST (SEQUOIA SEMPERVIRENS)    4
SERVICEBERRY OTHER (AMELANCHIER SPECIES)        1
SWEETGUM (LIQUIDAMBAR STYRACIFLUA)      5
TREE OF HEAVEN (AILANTHUS ALTISSIMA)    79
UNKNOWN DEAD TREES (UNKNOWN DEAD TREES )        60
UNKNOWN LIVE TREES (UNKNOWN LIVE TREES )        230
UNKNOWN SHAFT (UNKNOWN SHAFT )  1
UNKNOWN STUMP (UNKNOWN STUMP )  68
WILLOW SPECIES (SALIX SPECIES)  8
ZELKOVA JAPANESE (ZELKOVA SERRATA)      181
Time taken: 20.276 seconds
```

1.5 Save the Output:

insert overwrite local directory '/home/sushil/hive_output/tree_info' select spc_combined,count(*) from tree group by spc_combined;

Output file:

'https://github.com/sushilsani/HIVE/blob/master/output/000000_0'