Hive Lab : Temperature data Handling and Other operations

Objective : To get familiarity with Substring function

[training@localhost ~]$ gedit tmpr.txt  
[training@localhost ~]$ cat tmpr.txt  
xxxxx2006xxx34xxx  
xxxxx2006xxx35xxx  
xxxxx2006xxx32xxx  
xxxxx2007xxx24xxx  
xxxxx2007xxx21xxx  
xxxxx2008xxx37xxx  
xxxxx2008xxx39xxx  
[training@localhost ~]$   
  
hive> create database mytmpr;  
OK  
Time taken: 1.756 seconds  
hive> use mytmpr;                     
OK  
Time taken: 0.032 seconds  
hive> create table raw(line string);  
OK  
Time taken: 0.345 seconds  
hive>   
  
[training@localhost ~]$ hadoop fs -ls /user/hive/warehouse/mytmpr.db  
Found 1 items  
drwxr-xr-x   - training supergroup          0 2016-06-23 20:31 /user/hive/warehouse/mytmpr.db/raw  
[training@localhost ~]$   
  
-- when database is created, with name database and .db extension one directory will be created in warehouse location.  
  
 /user/hive/warehouse/mytmpr.db  
  
hive> load data local inpath  'tmpr.txt'   
    >  into table raw;   
  
[training@localhost ~]$ hadoop fs -ls /user/hive/warehouse/mytmpr.db/raw  
Found 1 items  
-rw-r--r--   1 training supergroup        126 2016-06-23 20:35 /user/hive/warehouse/mytmpr.db/raw/tmpr.txt  
[training@localhost ~]$   
hive> select \* from raw;  
OK  
xxxxx2006xxx34xxx  
xxxxx2006xxx35xxx  
xxxxx2006xxx32xxx  
xxxxx2007xxx24xxx  
xxxxx2007xxx21xxx  
xxxxx2008xxx37xxx  
xxxxx2008xxx39xxx  
Time taken: 0.186 seconds  
hive>   
  
hive> create table tmpr(y int, t int);  
OK  
Time taken: 0.058 seconds  
hive> insert overwrite table tmpr  
    >   select substr(line,6,4),  
    >     substr(line,13,2) from raw;  
  
  
hive> select \* from tmpr;  
OK  
2006    34  
2006    35  
2006    32  
2007    24  
2007    21  
2008    37  
2008    39  
Time taken: 0.057 seconds  
hive> describe tmpr;  
OK  
y       int  
t       int  
Time taken: 0.065 seconds  
hive>   
hive> create table results(y int,   
 max int, min int);  
hive> insert overwrite table results  
   select y, max(t), min(t)  
   from tmpr group by y;  
  
hive> select \* from results;  
OK  
2006    35      32  
2007    24      21  
2008    39      37  
Time taken: 0.058 seconds  
hive>   
  
[training@localhost ~]$ hadoop fs -cat /user/hive/warehouse/mytmpr.db/tmpr/000000\_0  
2006 34  
2006 35  
2006 32  
2007 24  
2007 21  
2008 37  
2008 39  
[training@localhost ~]$   
  
   default delimiter for hive table is  \001.  
  
[training@localhost ~]$ hadoop fs -ls /user/hive/warehouse/mytmpr.db/results  
Found 1 items  
-rw-r--r--   1 training supergroup         33 2016-06-23 20:48 /user/hive/warehouse/mytmpr.db/results/000000\_0  
[training@localhost ~]$ hadoop fs -cat /user/hive/warehouse/mytmpr.db/results/000000\_0  
2006 35 32  
2007 24 21  
2008 39 37  
[training@localhost ~]$   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
if temperature contains both positives and negatives..  
  
[training@localhost ~]$ cat tmpr2.txt  
xxxxx2006xxx34xxx  
xxxxx2006xxx-35xxx  
xxxxx2006xxx-14xxx  
xxxxx2006xxx32xxx  
xxxxx2007xxx-24xxx  
xxxxx2007xxx-21xxx  
xxxxx2007xxx35xxx  
xxxxx2008xxx-37xxx  
xxxxx2008xxx39xxx  
[training@localhost ~]$   
hive> create database urtmpr;  
OK  
Time taken: 0.022 seconds  
hive> use urtmpr;  
OK  
Time taken: 0.011 seconds  
hive> create table raw(line string);  
OK  
Time taken: 0.032 seconds  
hive> load data local inpath 'tmpr2.txt'  
    >  into table raw;  
  
hive> create table tmpr(y int, t int);  
OK  
Time taken: 0.035 seconds  
hive> insert overwrite table tmpr  
    >  select \* from (  
    >    select substr(line,6,4),  
    >      substr(line,13,2) from raw  
    >   where substr(line,13,1)!='-'  
    >      union all  
    >    select substr(line,6,4),      
    >      substr(line,13,3) from raw  
    >   where substr(line,13,1)='-') x;  
  
  
-- in hive only "union all" available., which allows duplicates.  
-- hive union should be placed as subquery.  
-- subquery should have aliase.  
  
hive> select \* from tmpr;  
OK  
2006    34  
2006    -35  
2006    -14  
2006    32  
2007    -24  
2007    -21  
2007    35  
2008    -37  
2008    39  
Time taken: 0.056 seconds  
hive>   
  
hive> create table results(y int,   
    >   max int, min int);  
OK  
Time taken: 0.047 seconds  
hive> insert overwrite table results  
    >   select y, max(t) , min(t)  
    >   from tmpr group by y;  
  
  
hive> select \* from results;  
OK  
2006    34      -35  
2007    35      -24  
2008    39      -37  
Time taken: 0.054 seconds  
hive>   
  
  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
  
Both inner and external tables can use custom locations.  
  
  
hive> create table hvtab1(line string)  
  location '/user/ursloc';  
hive> load data local inpath 'file1'  
   into table hvtab1;  
  
[training@localhost ~]$ hadoop fs -ls /user/ursloc  
Found 1 items  
-rw-r--r--   1 training supergroup         61 2016-06-23 21:15 /user/ursloc/file1  
[training@localhost ~]$   
  
A Single Location can be applied for mutliple tables.  
  
hive> create table hvtab2(line string)  
   location '/user/ursloc';  
  
now tables use same data and same location.  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
how to reuse:  
  
hive> create external table   
    >   xtab(line string);  
OK  
Time taken: 0.055 seconds  
hive> load data local inpath 'file1'  
    >  into table xtab;  
  
hive> select \* from xtab;  
OK  
aaaaaaaaaaaaaaaaa  
aaaaaaaaaaaaaaaaaaa  
aaaaaaaaaaaaaaaaaaaaaa  
Time taken: 0.059 seconds  
hive> drop table xtab;  
OK  
Time taken: 0.062 seconds  
hive> show tables;  
OK  
raw  
results  
tmpr  
Time taken: 0.042 seconds  
hive>   
  
hive> create table xtab(line string);  
OK  
Time taken: 0.031 seconds  
hive> select \* from xtab;  
OK  
aaaaaaaaaaaaaaaaa  
aaaaaaaaaaaaaaaaaaa  
aaaaaaaaaaaaaaaaaaaaaa  
Time taken: 0.06 seconds  
hive>   
  
-- create behavior:  
  if directory is already existed, it will use it, if not existed, it will create a new directory.  
  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
what if, backend directory is deleted.  
 what will happen to table?  
  
 table will not be deleted.  
  
[training@localhost ~]$ hadoop fs -ls /user/hive/warehouse/urtmpr.db/raw  
Found 1 items  
-rw-r--r--   1 training supergroup        167 2016-06-23 21:00 /user/hive/warehouse/urtmpr.db/raw/tmpr2.txt  
[training@localhost ~]$ hadoop fs -rmr /user/hive/warehouse/urtmpr.db/raw  
Deleted hdfs://localhost/user/hive/warehouse/urtmpr.db/raw  
  
  
  
hive> show tables;  
OK  
raw  
results  
tmpr  
xtab  
Time taken: 0.045 seconds  
hive> select \* from raw;  
OK  
Time taken: 0.036 seconds  
hive>