Hive Lab: Url Data Processing

$ cat  urls  
http://yoursite.com/bigdata/mongodb?  
  
name=Raj&age=25&city=hyd  
http://yoursite.com/bigdata/spark?  
  
name=Swapna&city=del&desig=se  
http://yoursite.com/bigdata/spark?  
  
name=Satwik&age=26&city=hyd  
http://yoursite.com/oops/java?  
  
name=Manoj&sex=m&age=35  
  
Url contains , 3 parts of information.  
  
 i) Host name:    
  
    yoursite.com  
  
ii) Path:  
  
   ex:    bigdata/mongodb  
  
iii) Query string.  
  
   name=Manoj&sex=m&age=35  
  
Hive hase predefined parsers ,   
  to extract basic entities of url.  
  
 i) parse\_url() ---> udf  
 ii) parse\_url\_tuple() ---> udtf.  
  
  
hive> create database urls;  
OK  
Time taken: 1.727 seconds  
hive> use urls;  
OK  
Time taken: 0.019 seconds  
hive> create table raw(line string);  
OK  
Time taken: 0.349 seconds  
hive> load data local inpath 'urls'  
    >  into table raw;  
Copying data from   
  
file:/home/training/urls  
Copying file: file:/home/training/urls  
Loading data to table urls.raw  
OK  
Time taken: 0.2 seconds  
hive>   
  
hive> select \* from raw;  
OK  
http://yoursite.com/bigdata/mongodb?  
  
name=Raj&age=25&city=hyd  
http://yoursite.com/bigdata/spark?  
  
name=Swapna&city=del&desig=se  
http://yoursite.com/bigdata/spark?  
  
name=Satwik&age=26&city=hyd  
http://yoursite.com/oops/java?  
  
name=Manoj&sex=m&age=35  
Time taken: 0.064 seconds  
hive>   
  
hive> select   
    >     parse\_url(line,'HOST'),  
    >     parse\_url(line,'PATH'),  
    >     parse\_url(line,'QUERY')  
    > from raw;  
  
hive> create table raw2(host string,   
    >  path string,   
    >  query string);  
OK  
Time taken: 0.039 seconds  
hive> insert overwrite table raw2  
    >   select x.\* from raw  
    >  lateral view  
    > parse\_url\_tuple  
  
(line,'HOST','PATH','QUERY')  
    >  x as h,p,q;  
  
hive> select \* from raw2;  
OK  
yoursite.com    /bigdata/mongodb         
  
 name=Raj&age=25&city=hyd  
yoursite.com    /bigdata/spark    
  
name=Swapna&city=del&desig=se  
yoursite.com    /bigdata/spark    
  
name=Satwik&age=26&city=hyd  
yoursite.com    /oops/java        
  
name=Manoj&sex=m&age=35  
Time taken: 0.063 seconds  
hive>   
  
  
  
  
hive> create table raw3(host string,   
    >  path array<string>,  
    >  qmap map<string,string>);  
  
  
  
hive> insert overwrite table raw3  
    >  select host,   
    >   split(path,'/'),  
    >  str\_to\_map(query,'&','=')  
    > from raw2;  
  
hive> select \* from raw3;  
  
hive> select \* from raw3;  
OK  
yoursite.com      
  
["","bigdata","mongodb"]          
  
{"age":"25","name":"Raj","city":"hyd"}  
yoursite.com    ["","bigdata","spark"]   
  
  
  
{"desig":"se","name":"Swapna","city":"  
  
del"}  
yoursite.com    ["","bigdata","spark"]   
  
  
  
{"age":"26","name":"Satwik","city":"hy  
  
d"}  
yoursite.com    ["","oops","java"]       
  
 {"sex":"m","age":"35","name":"Manoj"}  
Time taken: 0.064 seconds  
hive>   
  
  
  
hive> create table info(host string,    
    >  catogiry string,  course   
  
string,  
    >  name string, age int,   
    >  sex string, city string, desig   
  
string)  
    > row format delimited          
    >   fields terminated by ',';  
OK  
Time taken: 0.038 seconds  
hive> insert overwrite table info  
    >   select host, path[1], path[2],  
    >  qmap['name'], qmap['age'],  
    >  qmap['sex'], qmap['city'],  
    >  qmap['desig'] from  raw3;  
  
hive> select \* from info;  
OK  
yoursite.com    bigdata mongodb Raj      
  
 25      NULL      hyd     NULL  
yoursite.com    bigdata spark   Swapna   
  
 NULL    NULL      del     se  
yoursite.com    bigdata spark   Satwik   
  
 26      NULL      hyd     NULL  
yoursite.com    oops    java    Manoj    
  
 35      mNULL     NULL  
Time taken: 0.055 seconds  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
to extract all keys of a map   
  
collection/  
  
hive> select  map\_keys(qmap)      from   
  
raw3;  
  
["age","name","city"]  
["desig","name","city"]  
["age","name","city"]  
["sex","age","name"]  
  
hive> select  map\_values(qmap)        
  
from raw3;  
  
  
["25","Raj","hyd"]  
["se","Swapna","del"]  
["26","Satwik","hyd"]  
["m","35","Manoj"]  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
  
hive> create table raw4(host string,   
    >   k   array<string>,  
    >  v array<string>);  
OK  
Time taken: 0.037 seconds  
hive> insert overwrite table raw4  
    >   select host,   
    >     map\_keys(qmap), map\_values  
  
(qmap)   
    >   from raw3;  
  
  
hive> select \* from raw4;  
OK  
yoursite.com    ["age","name","city"]    
  
 ["25","Raj","hyd"]  
yoursite.com      
  
["desig","name","city"]   
  
["se","Swapna","del"]  
yoursite.com    ["age","name","city"]    
  
 ["26","Satwik","hyd"]  
yoursite.com    ["sex","age","name"]     
  
 ["m","35","Manoj"]  
Time taken: 0.061 seconds  
hive>