/\*\*

\* hive exercises

\*

\* part of a longer hadoop video course at http://bit.ly/learn-hadoop/

\*

\* @author rICh <rich@quicloud.com>

\*/

// our sample "page\_view" data looks like this:

1391102617,1234,http://quicloud.com/,http://google.com/,198.211.110.9

1391101111,1234,http://somewhere.com/,http://google.com/,197.211.110.9

1391002637,1234,http://somwhere-else.com/,http://yahoo.com/,196.211.110.9

1391002617,1234,http://quicloud.com/,http://bing.com/,198.211.110.9

// start up hive

hive

// this is an external table

CREATE DATABASE page\_view;

// note: if using the AWS-backed 'cloudera manager' stack, change '/user/cloudera/' to '/user/ubuntu/'

CREATE EXTERNAL TABLE page\_view\_external(viewTime INT, userid BIGINT,page\_url STRING, referrer\_url STRING,ip STRING) COMMENT 'This is the EXTERNAL page view table' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/cloudera/input-data/page\_view/';

DESC page\_view\_external;

DESC EXTENDED page\_view\_external;

//notice tableType (EXTERNAL\_TABLE)

// this command returns quickly (probably seconds). why?

// ...if we are not doing any filtering (like where), and returning all rows, no MR is done

SELECT \* FROM page\_view\_external;

// this command, however, takes several seconds

SELECT \* FROM page\_view\_external WHERE page\_url LIKE '%quicloud%';

SELECT DISTINCT(page\_url) FROM page\_view\_external WHERE page\_url LIKE '%quicloud%';

// this is an internal table

CREATE TABLE page\_view\_internal(viewTime INT, userid BIGINT,page\_url STRING, referrer\_url STRING,ip STRING) COMMENT 'This is the INTERNAL page view table' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA LOCAL INPATH '/home/cloudera/code-and-data/data/page\_view' INTO TABLE page\_view\_internal;

// What happened to our "local" data? In another terminal do:

ls -lat ~/code-and-data/data/page\_view/

// still there

// validate that the data made it into hive

SELECT \* FROM page\_view\_internal;

// in another terminal, let's look for where the data lives

hadoop fs -ls /user/hive/warehouse/page\_view\_internal

// where does our "external" (page\_view\_external) tables data live? this should work, right?

hadoop fs -ls /user/hive/warehouse/

// WRONG! This data is EXTERNAL and in the location that we connected up with the CREATE TABLE command

hadoop fs -ls /user/cloudera/input-data/page\_view/

DESC EXTENDED page\_view\_internal;

// notice tableType (MANAGED\_TABLE)

// you may want to query the hive data more, but when you're done, let's drop each table & notice what happens

SHOW TABLES;

DROP TABLE page\_view\_internal;

DROP TABLE page\_view\_external;

SHOW TABLES;

// in a 2nd terminal, run these commands. Which data do we expect to persist?

hadoop fs -ls /user/hive/warehouse/page\_view\_internal/

hadoop fs -ls /user/cloudera/input-data/page\_view