STSCI 5065 Quiz 4
(3/27/2019)

Name:
NetID:

1. A Hive table called **employees** is in the **default** database and the data of the table is first partitioned by **country** and then by **state**. Right now, you only have data for country US and its three states (NY, CA and PA) in the table. Where is the data of the employees table actually stored in HDFS (give the detailed location(s))? What is the relationship of two partitions and the relationship of the state level partitions? (50 points)

   The location of the **employees** table: /apps/hive/warehouse/employees.

   The **state** partition is a subdirectory under the **country=us** directory.

   The **state level partitions** are **state=NY**, **state=CA** and **state=PA** and they are same-level subdirectories of the **country=US** directory, which is a subdirectory of the **employees** directory. The following are the locations of the three state level subdirectories (it is not required to list the following):

   ```
   /apps/hive/warehouse/employees/country=us/state=NY
   /apps/hive/warehouse/employees/country=us/state=CA
   /apps/hive/warehouse/employees/country=us/state=PA
   ```

2. What is a hint in Hive join operation? State its purpose and give an example of using a hint (in the example, you should specify the Hive tables and their column name). (50 points)

   A hint, expressed as **/*+ STREAMTABLE(s) */**, in Hive is a mechanism to tell the Hive query optimizer which table should be streamed in join operations. The purpose is that you don't have to put the largest table last in a join operation and that the query optimizer will automatically stream the largest table even you do not put the largest table last in your query. An example of using a Hive hint is as follows (the following tables and their columns are the same as those discussed in the class).

   ```
   SELECT /*+ STREAMTABLE(s) */ s.ymd, s. symbol, s.price_close, d.dividend
   FROM stocks s JOIN dividens d ON s.ymd = d.ymd AND s.symbol = d.symbol;
   ```