<https://github.com/jc7mendoza/vaccinedata>

**1. Place file in hadoop.**

Hdfs dfs -mkdir tmp/vaccine

Hdfs dfs -put CIS\_CSV\_Selected\_States\_Dataset\_new1.csv tmp/vaccine

**2. Connect to hive to use the database.**

-bash-4.1$ beeline

!connect jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181,bigdai-nov-bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin

**3. Create a table schema by using values from the csv file.**

CREATE EXTERNAL TABLE IF NOT EXISTS vaccine (

State STRING,

Estimated\_hesitant STRING,

Percent\_adults\_fully\_vaccinated\_against\_COVID\_19 STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ","

LOCATION "/user/jmend161/tmp/vaccine"

TBLPROPERTIES ('skip.header.line.count' = '1');

**4. Run pig file to organize values from the csv file.**

hesitancy\_dirty = LOAD '/user/jdiaz231/project1/CIS\_CSV\_Selected\_States\_Dataset.csv' using PigStorage(',');

hesitancy\_clean = FOREACH hesitancy\_dirty generate (chararray) $1 as County, (chararray) $2 as State, (double) $3 as EstimatedHesitancy, (double) $4 as StrongHesitancy;

StateHesitancy\_Descending = ORDER hesitancy\_clean BY EstimatedHesitancy DESC;

DUMP StateHesitancy\_Descending LIMIT 5;

StateHesitancy\_Ascending = ORDER hesitancy\_clean BY EstimatedHesitancy ASC;

DUMP StateHesitancy\_Ascending LIMIT 5;