**Steps for Analytics:**

IMPORTANT: In addition to my dataset, my analytics will also require a secondary dataset that is cleaned by Yuejia (Stephen) Tong. Please follow his data ingest and ETL code instructions to obtain the national covid-19 dataset. Finally, my dataset, which you can obtain by following the steps below, will also be used in Stephen’s analytics.

1. Create Hive tables for analytics with the two datasets. First login to Hive, then run the following commands
   1. use NetID;
   2. create external table happiness (country string, happiness\_score double, gdp\_per\_capita double, social\_support double, life\_expectancy double, freedom double, generosity double, government\_trust double)

row format delimited fields terminated by ‘,’

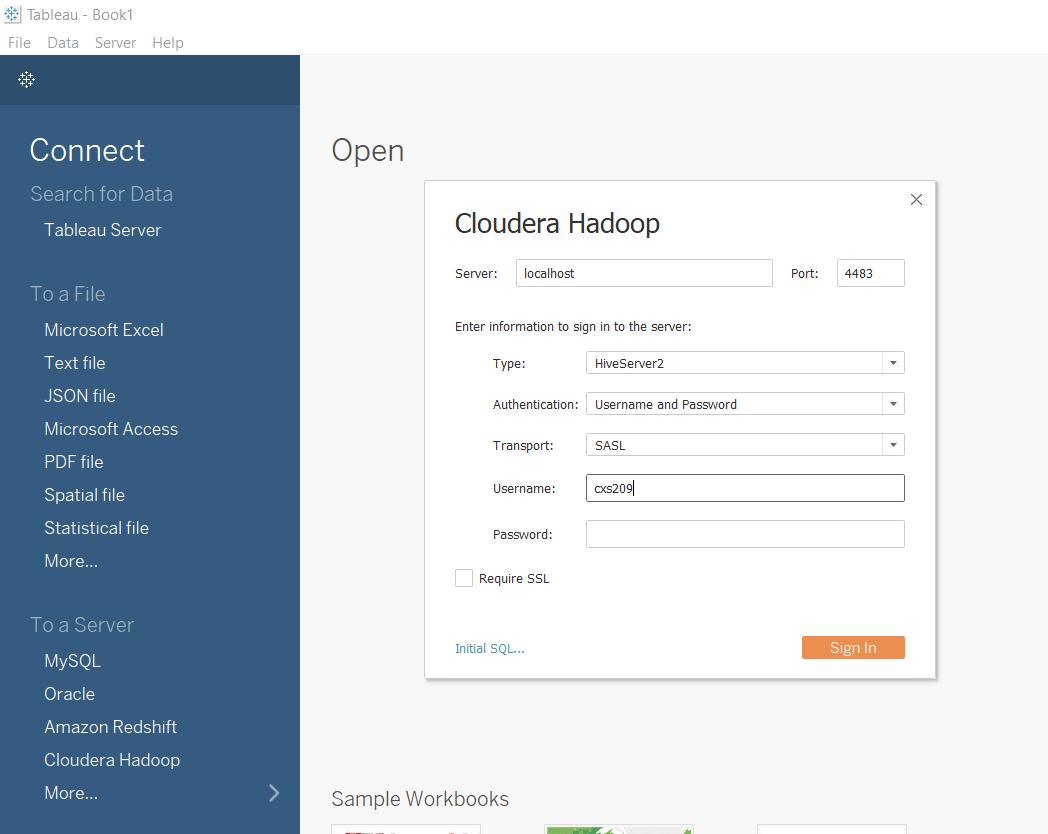
location ‘/user/NetID/FinalProject/hiveInput1/’;

* 1. create external table covid (country string, confirmed int, death int, recovered int, active int, death\_rate double, recovered\_rate double, one\_week\_increase int, one\_week\_increase\_rate double)

row format delimited fields terminated by ‘,’

location ‘/user/NetID/FinalProject/hiveInput2/’;

1. Log out from Dumbo. Re-login with port forwarding using the command “ssh -L 4483:babar.es.its.nyu.edu:10000 NetID@dumbo.hpc.nyu.edu”
2. Open Tableau and connect to Cloudera Hadoop.



1. Select NetID in “Schema”
2. We will attempt to look at the correlation between every numeric field in the happiness table vs the confirmed Covid cases and recovered rate. Specifically, I am responsible for analyzing 1) overall happiness score 2) GDP per capita 3) social support. I will list the steps for overall happiness score, and the steps for the other two are the same.
   1. Click “New Custom SQL”
   2. The query will be “select t1.country, t1.happiness\_score, t2.confirmed, t2.death, t2.recovered, t2.active, t2.recovered\_rate from NetID.happiness t1 inner join NetID.covid t2 on t1.country = t2.country”
   3. Plot a symbol map where “country” is the country on the map, “sum(happiness\_score)” shows the color, and “sum(recovered\_rate)” reflects the size of the circle. We are essentially analyzing happiness score vs recovered rate on a national scale by plotting and visualization. Note that you may need to change the color scheme and also adjust the median point for the coloring.
   4. Similarly, plot a symbol map where “country” is the country on the map, “sum(happiness\_score)” shows the color, and “sum(confirmed)” reflects the size of the circle. We are essentially analyzing happiness score vs confirmed cases on a national scale.
   5. Repeat a-d for GDP and Social Support. We have 6 graphs in the end.