Authors: Tony Fong, Michael Medina, Victor Verduzco
Pablo Ilabaca, John Martinez
Instructor: Jongwook Woo
Date: 12/7/2022

Lab Tutorial

12/7/2022

Covid 19 Surveillance Data

Objectives:
In this hand-on lab, you will learn how to:

- Split a dataset using HDFS commands
- Upload the dataset
- Query data
- Visualize data

Platform Specifications:

- Oracle Cloud
- CPU Speed: 1995. 309 MHz:
- # of CPU Cores: 32
- # of nodes: 3
- Total Memory Size: 58GB

1. open a shell terminal – git bash, minty, putty etc- and run the ssh command to connect to the Hadoop Cloud.

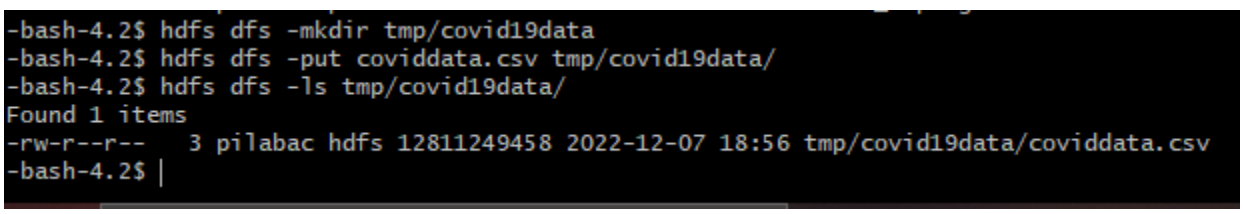
$ssh yourusername@ipaddress

2. Download the file using wget

```
wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies
/tmp/cookies.txt --keep-session-cookies --no-check-certificate
'https://docs.google.com/uc?export=download&id=1NP3feB6JvFAlv4rnatskW5047cedEI8C
' -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1NP3feB6JvFAlv4rnatskW5047cedEI8C" -O
coviddata.csv && rm -rf /tmp/cookies.txt
```

3.You have to upload the files to hdfs folder coviddata. Run the following HDFS commands to create and list coviddata directory in HDFS:

```
$ hdfs dfs -mkdir tmp/covid19data
$ hdfs dfs -put coviddata.csv tmp/covid19data/
```



4.open hive

```
$ beeline
```

5. Create your own database and use that database

```
$ create database Covid19;
$ use database Covid19;
```

6. Create external table “Covid19Data”

```
- - create the covid19 table on comma-seperated covid19data
```

```
create external table if not exists Covid19 (case_months string,
res_state string,
state_fips_code string,
res_country string,
county_fips_county string,
age_group string,
sex string,
race string,
ethnicity string,
case_positive_specimen_interval int,
case_onset_interval int,
process string,
exposure_yn string,
current_status string,
symptom_status string,
hosp_yn string,
icu_yn string,
death_yn string,
underlying_conditions_yn string)
row format delimited fields terminated by ","
stored as textfile location '/user/tfong9/tmp/covid19data'
tblproperties ('skip.header.line.count' = '1');
```

Now run the following HiveQL at the query editor to see how the dataset looks like

```
select * from covid19 limit 10;
```

covid19.case_months	covid19.res_state	covid19.state_fips_code	covid19.res_country	covid19.county_fips_county	covid19.age_group	covid19.sex	covid19.race	covid19.ethnicity	covid19.case_positive_specimen_interval	covid19.case_onset_interval	covid19.process	covid19.exposure_yn	covid19.current_status	covid19.symptom_status	covid19.hosp_yn	covid19.icu_yn	covid19.death_yn	covid19.underlying_conditions_yn
2021-12	CA	06	VENTURA	06111	18 to 49 years	Female	White	Non-Hispanic/Latino	NULL	NULL	Missing	Missing	Missing	Missing	Missing	Missing	Missing	
2021-09	TX	48	TARRANT	48439	18 to 49 years	Male	White	Non-Hispanic/Latino	NULL	NULL	Missing	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	
2022-01	MA	25	MIDDLESEX	25017	18 to 49 years	Female	Unknown	Unknown	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	
2020-12	NY	36	KINGS	36047	65+ years	Female	White	Non-Hispanic/Latino	0	0	Missing	Missing	Laboratory-confirmed case	Symptomatic	Missing	Missing	Missing	
2022-01	NJ	34	ESSEX	34013	0 - 17 years	Male	White	Non-Hispanic/Latino	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	No	Missing	No	
2022-06	CA	06	SACRAMENTO	06067	18 to 49 years	Female	Unknown	Non-Hispanic/Latino	NULL	NULL	Missing	Missing	Laboratory-confirmed case	Unknown	Missing	Missing	Missing	
2021-12	NJ	34	OCEAN	34029	50 to 64 years	Female	White	Non-Hispanic/Latino	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	No	Missing	No	
2021-09	NY	36	MONROE	36055	0 - 17 years	Female	Black	Non-Hispanic/Latino	NULL	0	Missing	Missing	Laboratory-confirmed case	Symptomatic	No	Missing	No	
2021-07	FL	12	PALM BEACH	12099	18 to 49 years	Male	Black	Non-Hispanic/Latino	NULL	0	Missing	Missing	Laboratory-confirmed case	Symptomatic	No	Missing	No	
2022-05	FL	12	PINELLAS	12103	18 to 49 years	Male	White	Non-Hispanic/Latino	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	

7.create external table “patient_profile”

```
- - create the patient_profile table on comma-seperated covid19data
CREATE EXTERNAL TABLE IF NOT EXISTS patient_profile(age_group STRING, sex STRING, race STRING, ethnicity STRING, res_state STRING, underlying_conditions STRING)
row format delimited fields terminated by ","
STORED AS TEXTFILE LOCATION '/user/tfong9/tmp/covid19data';

insert overwrite table patient_profile
select age_group, sex, race, ethnicity, res_state, underlying_conditions_yn
```

```
from covid19;
```

Now run the following HiveQL at the query editor to see how the dataset looks like

```
Select * from patient_profile limit 10;
```

patient_profile.age_group	patient_profile.sex	patient_profile.race	patient_profile.ethnicity	patient_profile.res_state	patient_profile.underlying_conditions
18 to 49 years	Female	White	Non-Hispanic/Latino	CA	
18 to 49 years	Male	White	Non-Hispanic/Latino	TX	
18 to 49 years	Female	Unknown	Unknown	MA	
65+ years	Female	White	Non-Hispanic/Latino	NY	
0 - 17 years	Male	White	Non-Hispanic/Latino	NJ	
18 to 49 years	Female	Unknown	Non-Hispanic/Latino	CA	
50 to 64 years	Female	White	Non-Hispanic/Latino	NJ	
0 - 17 years	Female	Black	Non-Hispanic/Latino	NY	
18 to 49 years	Male	Black	Non-Hispanic/Latino	FL	
18 to 49 years	Male	White	Non-Hispanic/Latino	FL	

8. Now run the following HiveQL at the Query editor to see the number of cases

```
select case_months, count(sex) as number_of_cases
from covid19
group by case_months;
```

case_months	number_of_cases
2022-01	2648280
2020-05	153218
2020-04	210489
2021-12	1412096
2022-03	175036
2020-03	110842
2020-07	345947
2020-12	985109
2022-04	269741
2020-09	183285
2021-07	299067
2022-02	423524
2020-11	598596
2021-01	1024501
2021-03	380664
2021-06	77038
2022-10	109129
2021-05	122949
2022-09	221328
2021-09	426063
2020-01	548
2020-10	259393
2022-07	598233
2022-05	699285
2022-06	519352
2021-11	307190
2022-08	451627
2020-02	1227
2021-02	391593
2021-08	585600
2021-10	295130
2020-06	213922
2021-04	274905
2020-08	225092

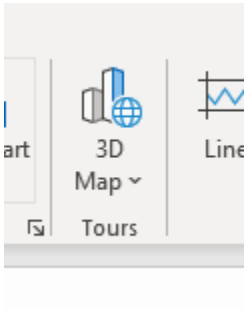
9. Now download data ito your PC

```
- - download to local file
hdfs dfs -get tmp/covid19data/000000_0
- - download file to your PC
scp tfong9@144.24.14.145:/home/tfong9/000000_0 covid19data.csv
```

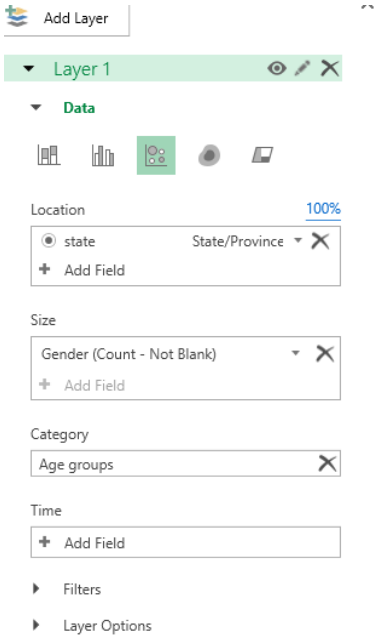
10. Loading Data into and Visualizing using Power Map in Excel

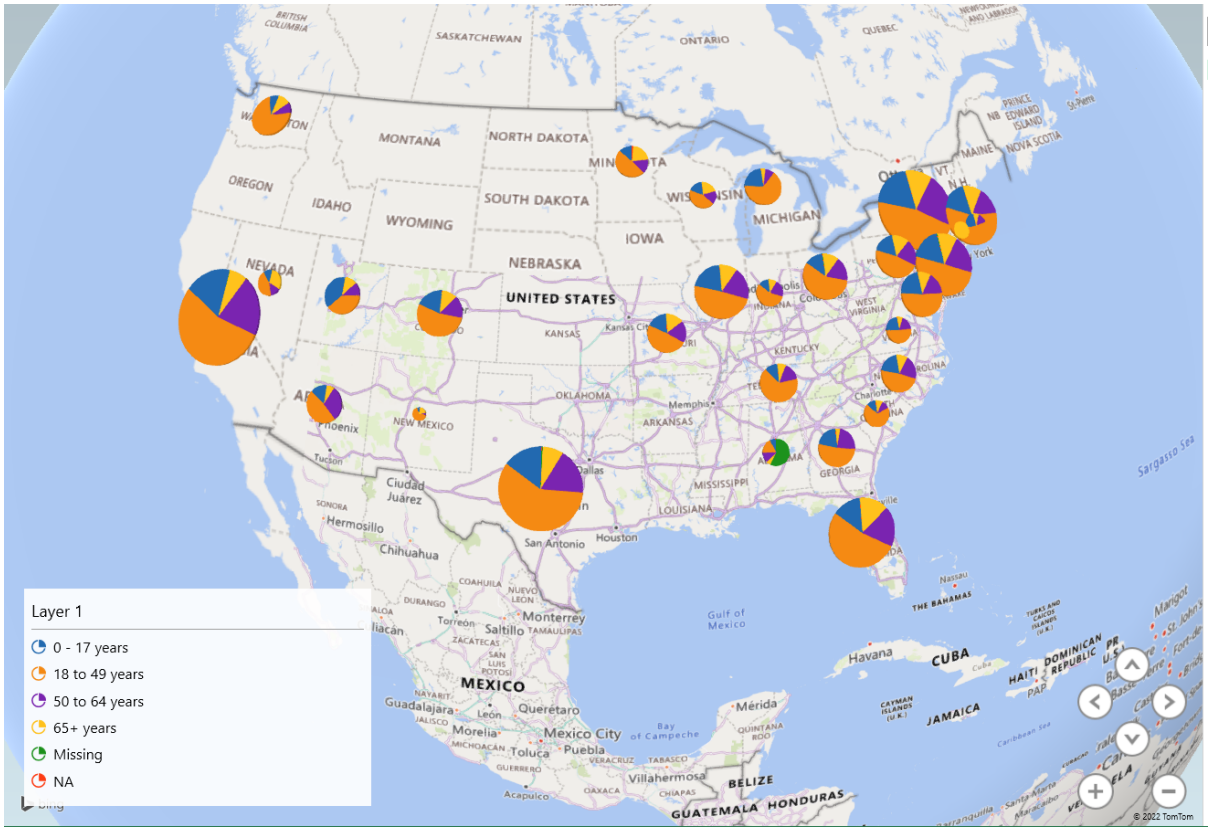
Create column names for each column

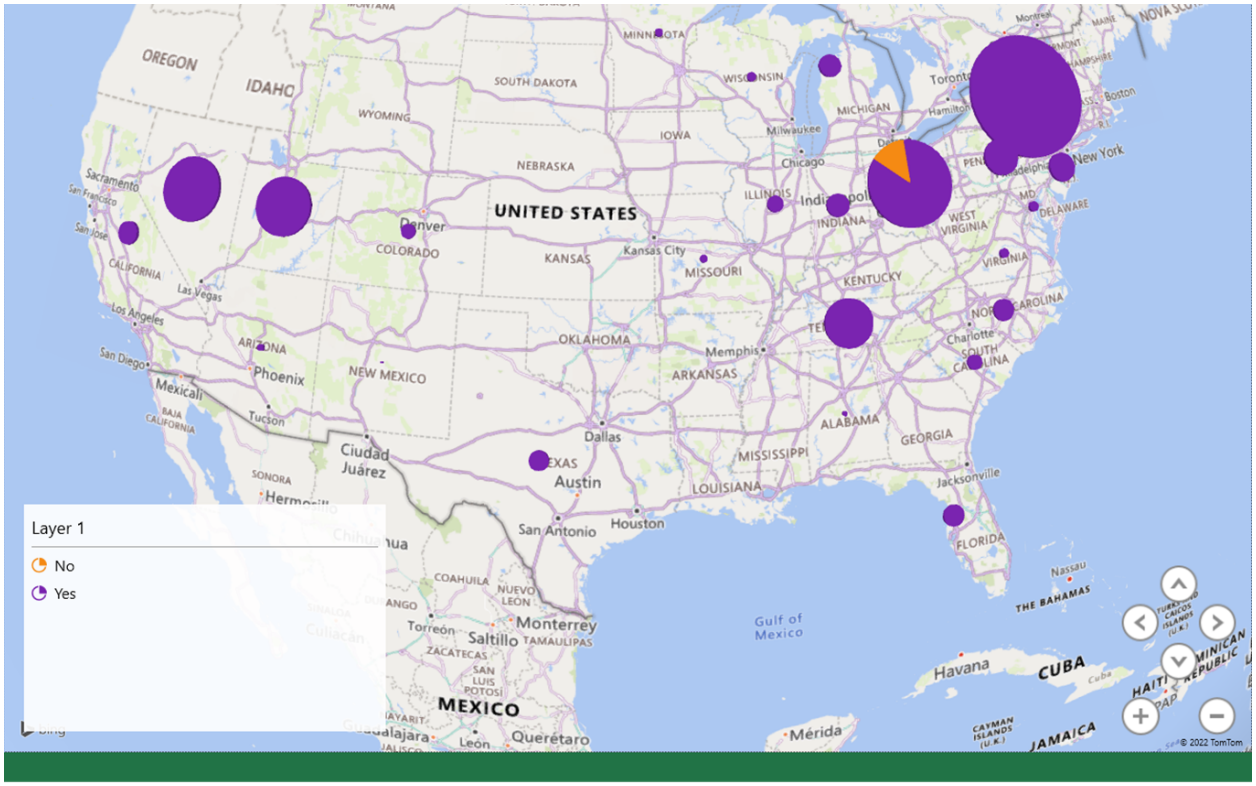
Open up 3d maps

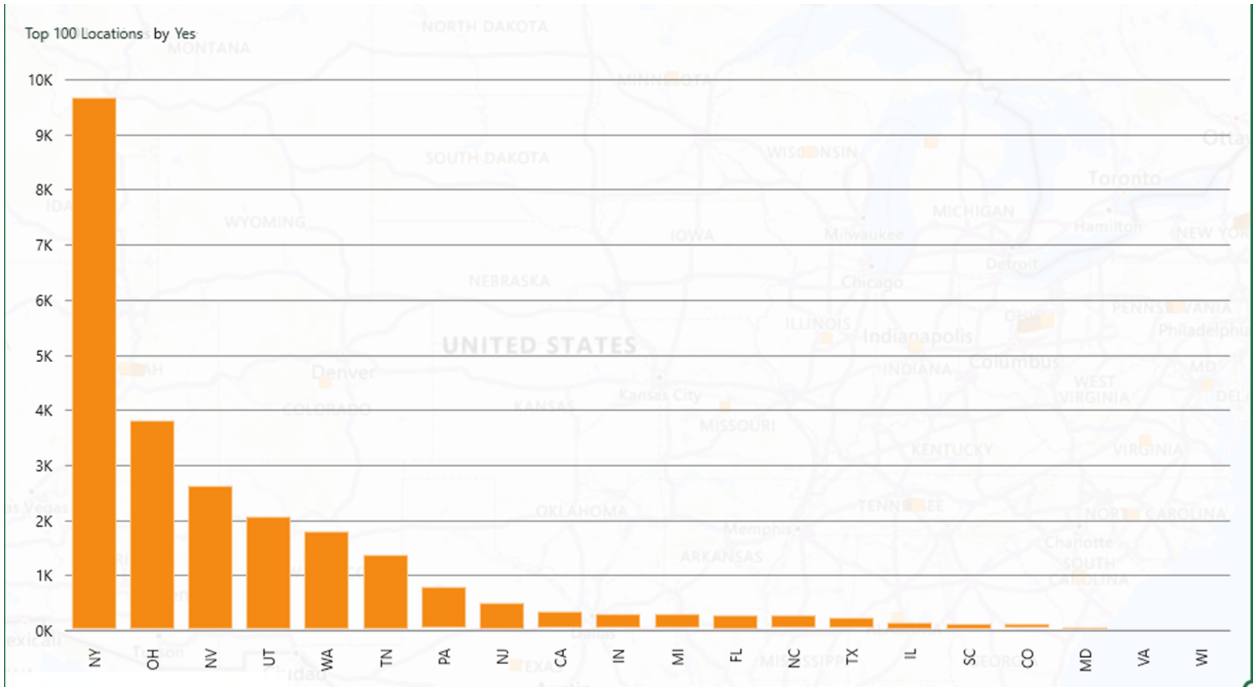


You need to select the properties and values in the layer as follows.


Add Layer
Layer 1
Data
Location
100%
state
State/Province
Add Field
Size
Gender (Count - Not Blank)
Add Field
Category
Age groups
Time
Add Field
Filters
Layer Options


Layer 1
0 - 17 years
18 to 49 years
50 to 64 years
65+ years
Missing
NA
UNITED STATES
MEXICO
CUBA
JAMAICA
HAITI
DOMINICAN REPUBLIC
BELIZE
GUATEMALA
HONDURAS
THE BAHAMAS
Gulf of Mexico
Sargasso Sea
© 2022 TomTom


Layer 1
No
Yes
UNITED STATES
MEXICO
Gulf of Mexico
© 2022 TomTom


Top 100 Locations by Yes
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
NY
OH
NV
UT
WA
TN
PA
NJ
CA
IN
MI
FL
NC
TX
IL
SC
CO
MD
VA
WI