**CIS3200 Term Project Tutorial**

**Authors: Saul Arguelles, Raul Mancilla, Maria Patricia Santos**

**Instructor:** [**Jongwook Woo**](https://www.linkedin.com/in/jongwook-woo-7081a85)

**Date: 12/5/2020**

**Lab Tutorial**

Saul Arguelles

Raul Mancilla

Maria Patricia Santos

12/5/2020

**COVID-19 Data Analysis between South Korea and United States**

**Objectives**

**List what your objectives are.** In this hands-on lab, you will learn how to:

* Use Azure ML to filter and analyze data
* Use Elastic Search to use GeoMapping areas and graph data sets
* SQL commands to perform the analysis.
* Visualization

**Platform Specs**

·       IBM Bluemix BigInsights

·       CPU Speed: 2.0ghz

·       # of CPU cores: 4

·       # of nodes: 5

·       Total Memory Size: 5.37MB

**AZURE ML**

**Step 1:** In the Studio, at the bottom left, click NEW. Then in the collection of Microsoft samples, select Blank Experiment. This creates a blank experiment, which looks similar to the following image.

Change the title of your experiment from “Experiment created on today’s date” to “COVID-19 Deaths”

**Step 2:** In the experiment items pane on the left, expand Saved Datasets, expand Samples, and drag COVID-19 Data dataset to the experiment canvas in the middle of the page.

Select the COVID-19 Data dataset on the canvas and note that it has a single output port (indicated as a circle containing the value 1 at the bottom of the dataset icon). Right-click this output port and click Visualize to see the data that the dataset contains

**Step 3:** In the dataset, scroll the table pane on the left if necessary to see the deaths column, and then click the death column header so that a column summary and histogram for that column is displayed as shown in the following image. This column represents the count of COVID-19 deaths on a given day and hour.

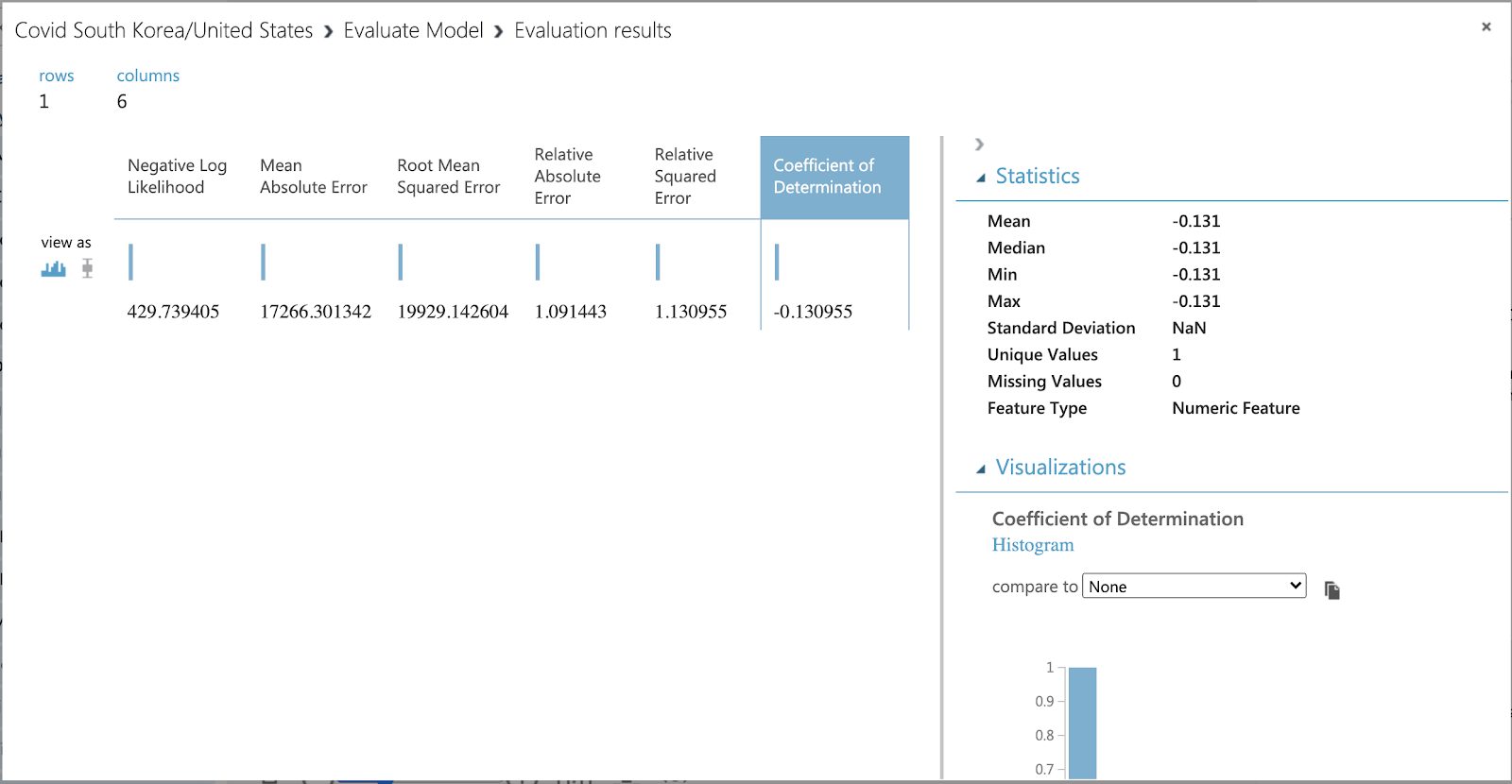
Select any of the other columns (for example date, gender, or age), and note the statistics and histogram that is displayed. By visualizing statistics about the distribution of values in your data, and the relationships between values in columns, you can learn a lot about your data and refine it to build a more effective and accurate predictive model.

**Step 4:** Close the dataset, and in the experiment items pane, in the search box, type “Project Columns”. Then, in the filtered experiment items pane, under Data Transformation and Manipulation, drag the Project Columns – Select Columns in Data Set- module to the canvas and place it under the COVID-19 Data dataset.

Click the output port of the COVID-19 Data dataset and drag it to the input port at the top of the Project Columns (Select Columns in Data Set) module to connect the items.

**Step 5:** Select the Project Columns module, and in the Properties pane on the right, click Launch column selector. The column selector is a common user interface element in Azure ML modules and enables you to select the columns you want to use in the module. In this case, the Project Columns module is used to filter out columns you don’t need, so that only the columns you want to use are passed (or projected) into the data flow for the next module.

In the Select columns dialog box, select option With Rules to begin with all columns, and exclude the registered and casual column names as shown in the image below. Then click the OK icon to close the column selector.



**Step 6:** On the toolbar at the bottom of the page, click SAVE to save the experiment. Then click RUN to run the experiment.

When the experiment has finished running, note the status displayed at the top-right of the experiment canvas and the green check mark that indicates that the Project Columns module completed successfully.

**Step 7:** Visualize the Result Dataset output port of the Project Columns module, and verify that the registered and casual columns have been removed, as shown here. Then close the results dataset.

**Step 8:** Drag the COVID-19 Data dataset to the canvas for the COVID-19 Data experiment.

Right-click the output port for the COVID-19 data dataset on the canvas and click Visualize to view the data in the dataset.

Verify that the dataset contains the data you viewed in the source file, and then close the dataset.

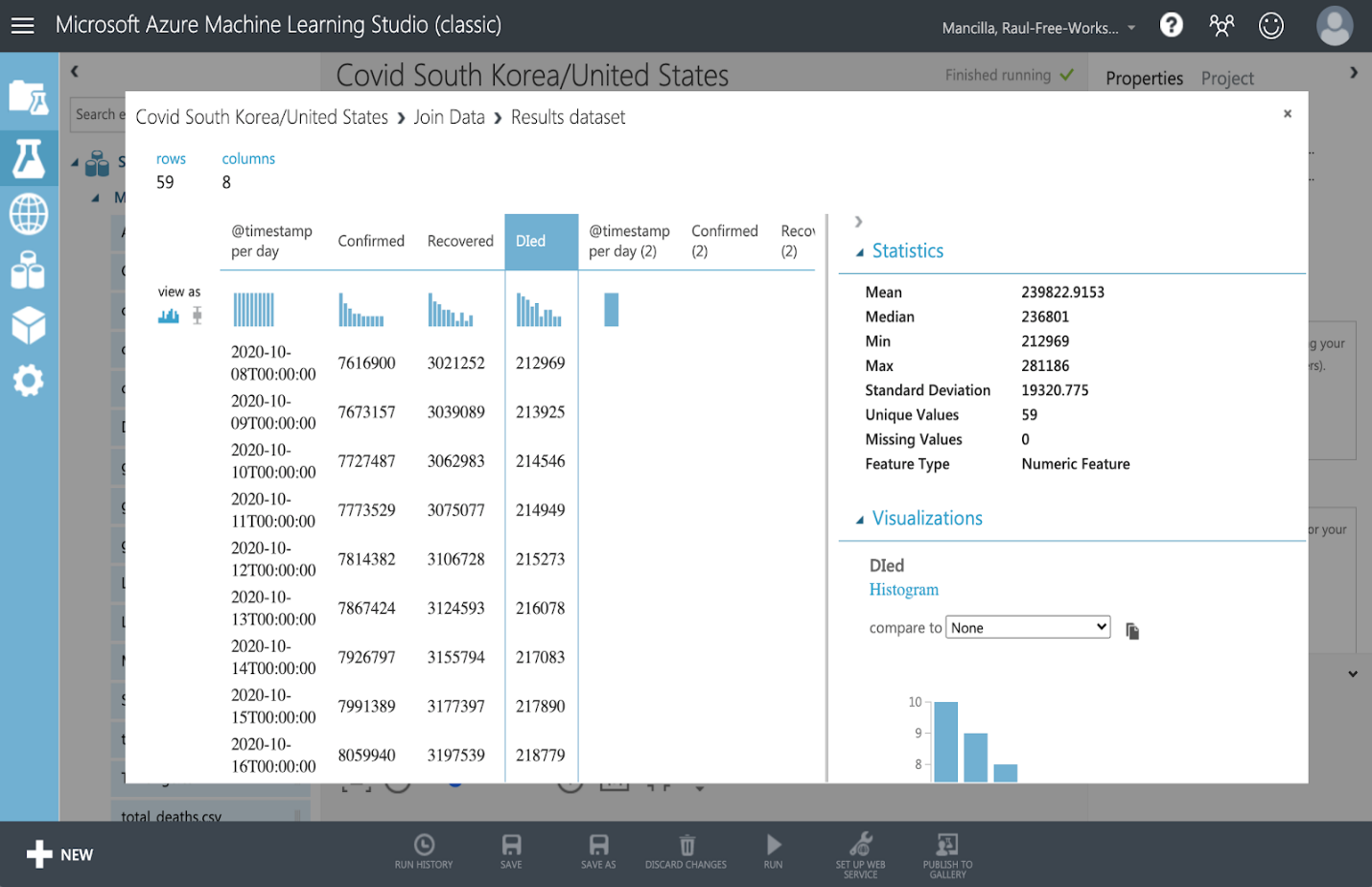
In Azure ML Studio, ensure that the COVID-19 Data experiment you created in the previous exercise is open.

Add a new dataset by uploading the covid\_data.csv file in the folder where you extracted the lab files. Name the new dataset Admissions Mapping.

In the COVID-19 Data Azure ML experiment, search for the US data dataset and drag it to the canvas, next to the South Korea COVID-19 data dataset. Do not connect it to any other modules at this point but visualize the dataset output port and note that the dataset contains rows that map death\_data codes to death\_data text values.

Close the dataset.

**Step 9:** Search for the Join module, and drag it to the canvas below both the COVID-19 Data dataset and the Covid\_US\_data dataset. Then connect the output ports from the Diabetic Data dataset and the Admission Mapping dataset to the Dataset1 and Dataset2 input ports of the Join module, respectively.



Select the Join module, and in the Properties pane, set the following properties:

Join key columns for L: Launch the column selector and select Death\_rate. b. Join key columns for R: Launch the column selector and select Death-rate. c. Join type: Left Outer Join

Keep right key column: Unselected

When the experiment has finished running, visualize the Results dataset output port of the Join module, and verify that it contains the cleansed patient admission data from the original Diabetic Data dataset and the death column from the Covid\_data dataset.

Close the results dataset

**Step 10:** In the experiment items pane, search for “Execute R Script”. Then drag the Execute R Script module to the experiment canvas, under the Project Columns module; and connect the output port from the Project Columns module to the first (left-most) input port of the Execute R Script module.

Select the Execute R Script module, and in the Properties pane, replace the default R script with the following code. You can copy and paste this code from the Covid\_data.txt file in the folder where you extracted the lab files for this lab:

frame1 <- maml.mapInputPort(1)

## Delete age and gender columns

frame1$age <- NULL

frame1$gender<- NULL

maml.mapOutputPort('frame1')

This code creates an R data frame from the data that is passed to the first input port of the script module, and then removes the age and gender columns, before passing the modified data frame to the output port of the script module.

Right-click the Execute R Script module and click Edit Comment. Then type “Remove columns” in the comment box and click a blank area of the canvas to finish editing the comment, and click the v icon in the Execute R Script module to expand the comment.

In the experiment items pane, search for “Execute Python Script”. Then drag the Execute Python Script module to the experiment canvas, under the Execute R Script module, and connect the Result Dataset output port (the left-most output) from the Execute R Script module to the first (left-most) input port of the Execute Python Script module.

Select the Execute Python Script module, and in the Properties pane, replace the default Python script with the following code. You can copy and paste this code from the Covid\_data.txt file in the folder where you extracted the lab files for this.

def azureml\_main(dataframe1 = None, dataframe2 = None):

## delete medium and cases columns

dataframe1.drop(['medium', 'cases'], 1, inplace = True)

return dataframe1

**Step 11:** In the experiment items pane, search for “Apply SQL Transformation”. Then drag the Apply SQL Transformation module to the experiment canvas, under the Execute Python Script module, and connect the Result Dataset output port (the left-most output) from the Execute Python Script module to the first (left-most) input port of the Apply SQL Transformation module.

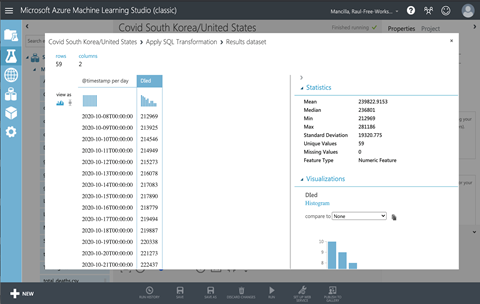
Select the Apply SQL Transformation module, and in the Properties pane, replace the default SQL script with the following code. You can copy and paste this code from the Bike Rentals Code.txt file in the folder where you extracted the lab files for this lab:

       SELECT \*, death\*death AS deathcombined FROM t1

       WHERE age BETWEEN 1 AND 99

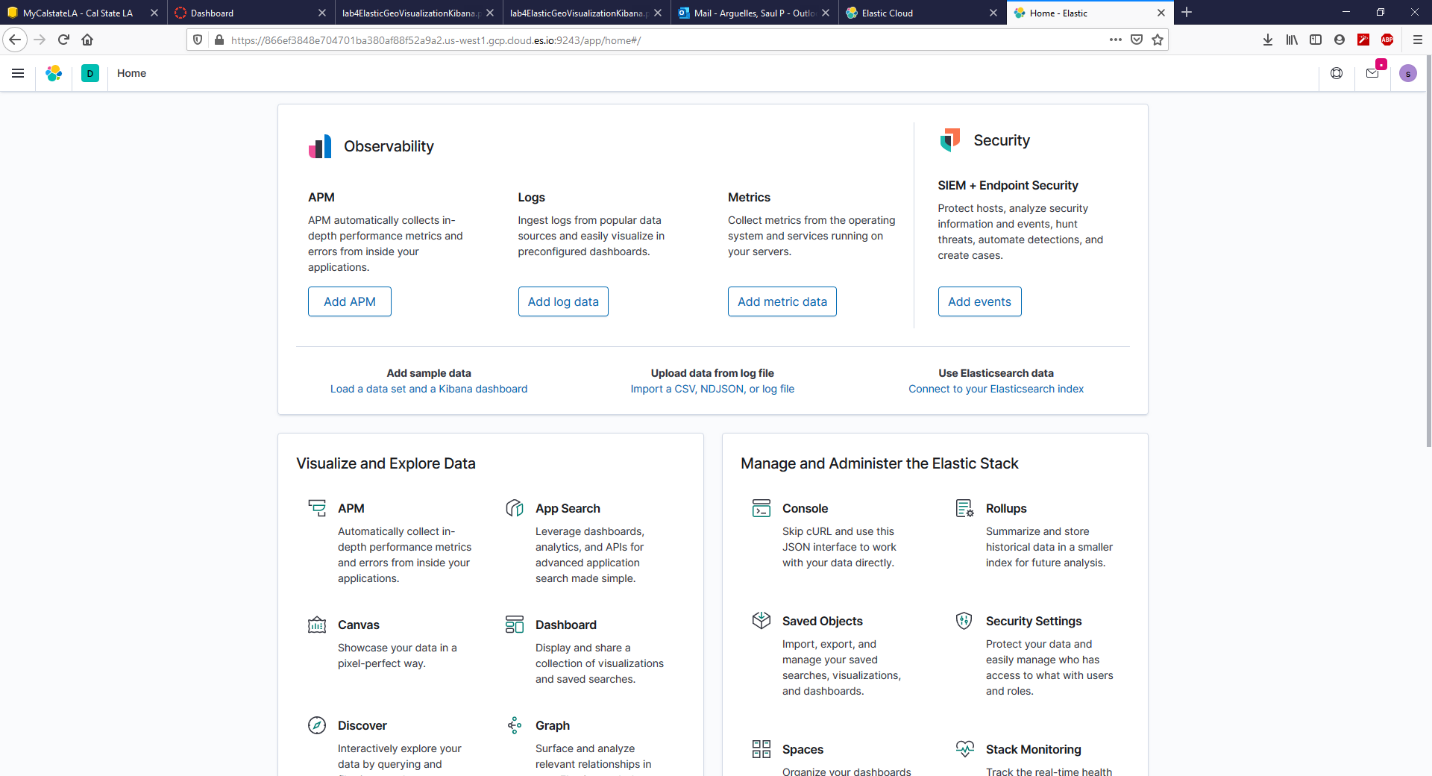
       AND day BETWEEN 0 AND 31;

Save and run the experiment. At each node, the green circling should be shown while it is running. Then, when the experiment has finished with “Finished running” at the top of the pane, visualize the Results dataset output of the Apply SQL Transformation module, and view the filtered data, as shown here. Then close the results dataset



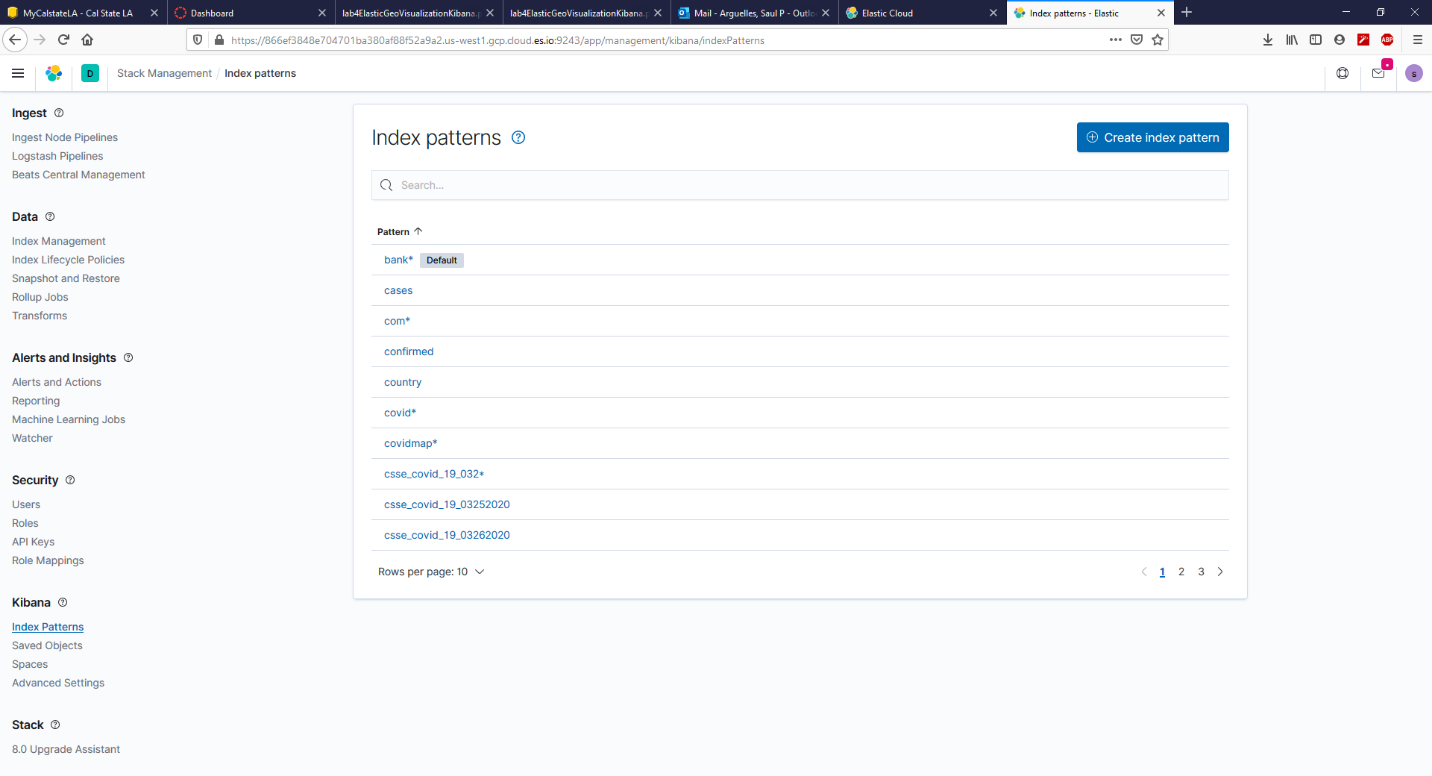
**Elastic Search (Geo Mapping/ Graphing)**

Before all this you must upload data from log file in the begging.



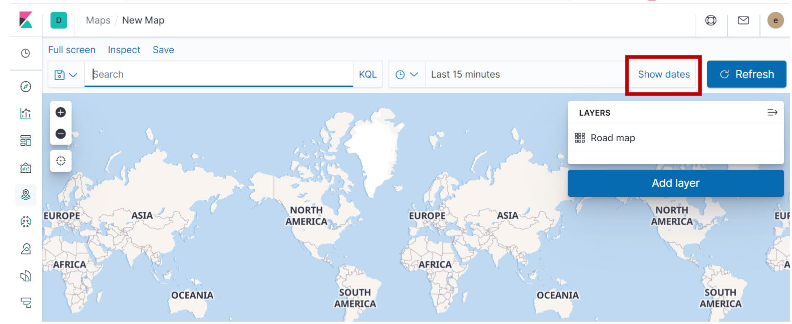
V. Indexing your data

You’ll use a coordinate chart to look at the logs.jsonl data set, which are ranged in 3 days. 1.Click Management > Index Patterns to index logs data as you have imported.



Select Create New Index Pattern and type in something\*. Then, select Next Step > Create Index Pattern. Make sure you see the name you put as the index. Set the index as default and it should be good to go

Go to visualize and create new visualization and click map



In the time filter, clickShow dates. See the figure in the next page:

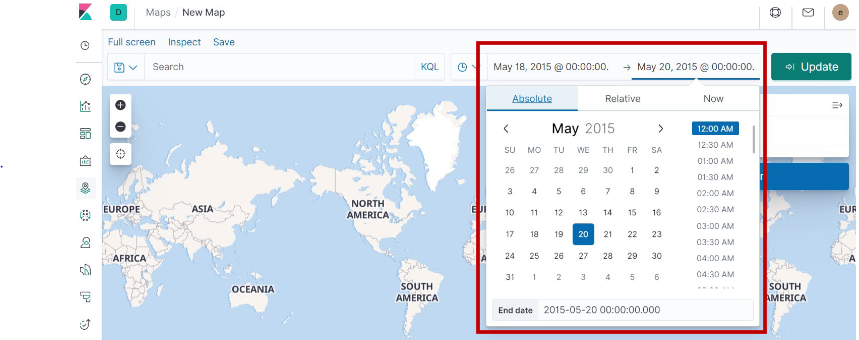
a)Click the start date, thenAbsolute.

b)Set theStart date to May 18, 2015, 12:00AM.

c)In the time filter, clicknow, thenAbsolute.

d)Set theEnd date to May 20, 2015, 12:00AM.

e)ClickUpdate



Click Map the geo coordinates from the log files.

a)Click Add layer.

b) Click theGrid aggregation data source.

c)Set Index pattern to something\* or whatever you named your index.

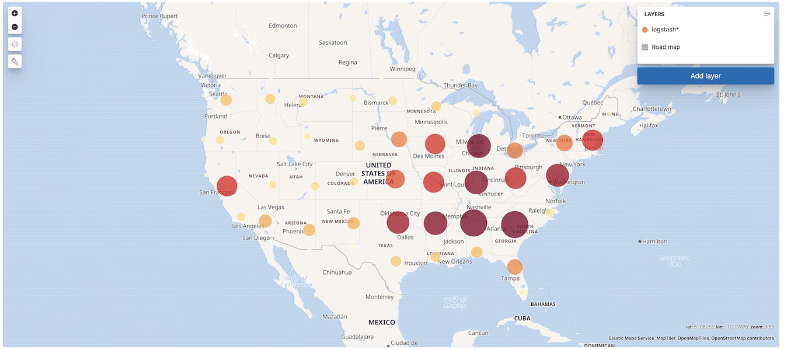
d)Click the Add layer button.

6.Set the layer style.

a) For Fill color, select the yellow to red color ramp.

b) For Border color, select white.

c) Click Save & close.



Import the data from CSV

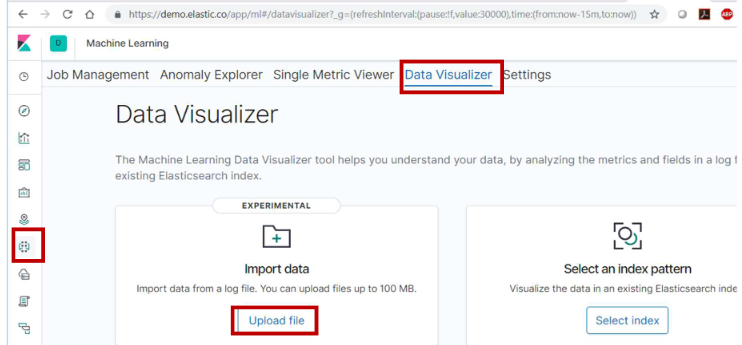
Ok, now that we have data to analyze and an Elasticsearch cluster that can do the analyzing, we need to import and set up index.

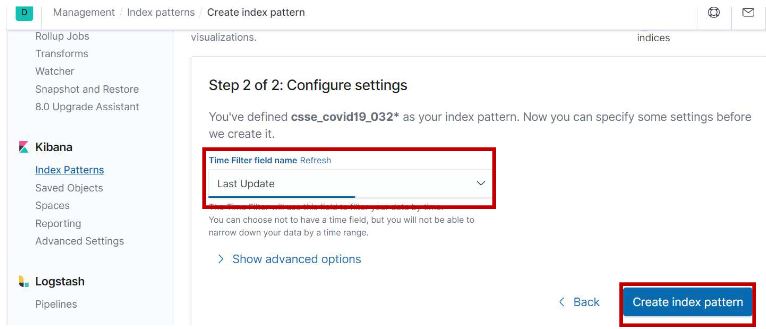
1.Within Kibana, click on Machine Learning.

2.In the sub nav, click on Data Visualizer.

3.Under Import Data, click Upload File. And override setting with lagitutde and longitude and hot import and create a new index pattern

4. follow steps from the previous map to see if index pattern is correct



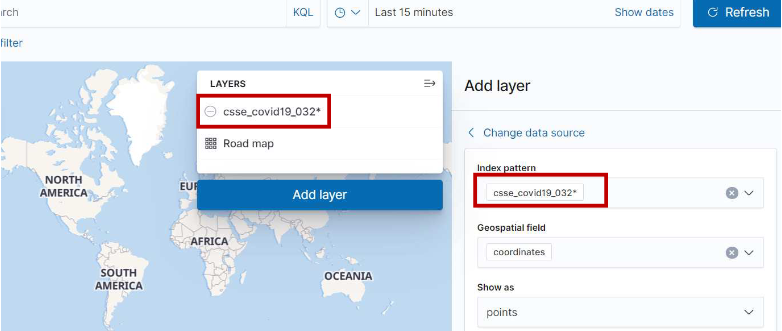


Creating Kibana visualizations using Maps

1.On the left side navigation, click on Maps. Then, select Add Layer > Grid Aggregation

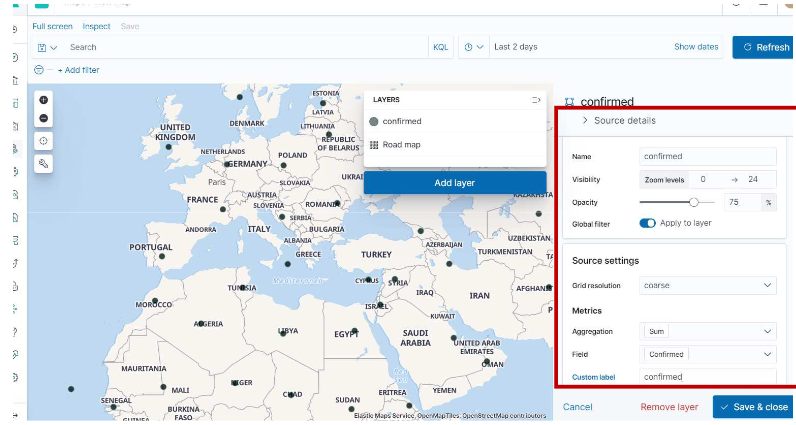
2.Select “Clusters and grids”

3.At Index pattern of Grid aggregation, select drop-box menu to find out the index name: something\* or covidmap\*. It will show the geo coordinates and points automatically.

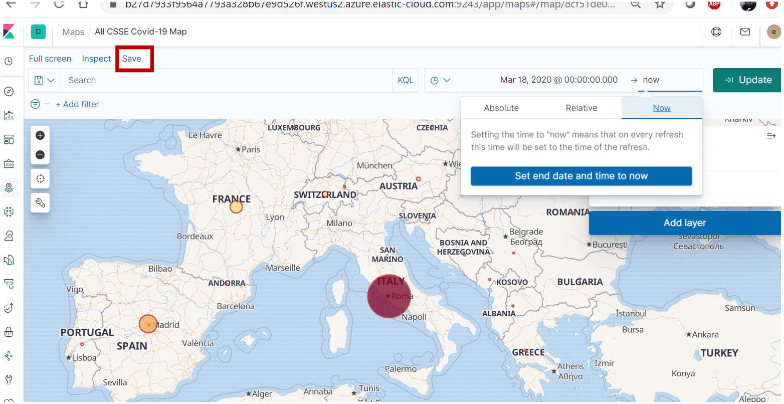


Scroll up back up to the step about time frame and do the same thing again to get it in March

Then, you need to add another layer by selecting Add Layer in the bottom of the right frame. Then, fill in the form as the next step.



Save map as COVID-19 map and should look like this or something related.



Why use elastic? Elastic is an open-source search and analytics engine for all type of data, from maps to geographic visuals to structured data in order to show a visualization in real time. It works by using a variety of sources including logs, system metric and web applications to process data and re create distribution to back up summaries of reports. In my portion of the lab, I tried to show the COVID-19 cases in real time now from confirmed, recovered and death tools to create a result to show the audience. Knowing that things are not good I use this comparison to south Korea which is still stagnant in numbers but has completely cut off the spread of COVID-19. Unlike the United States where cases are growing every day even know. In March is was a small number of cases compared to know with over a million in some states. People are ravaging from looting to holding toilet paper to staying inside due the home orders. I used a source to compare the 14000 cases in March to the millions of cases confirmed with two different geographic maps to create a live comparison so the audience can see how real this and why we need to take this covid-19 on as a serious threat. Lastly the biggest fact or comparison I saw was a province like Puerto Rico has more confirmed cases than South Korea.