CIS 3200-02 Term Project Tutorial

Authors: Samuel Mendoza, Israel Alegria, Hector Pedro, Julio Montiel, Marco Rodriguez
Instructor: Jongwook Woo
Date: 05/17/2020

ElasticSearch with CSV Data Import and Visualization

Team 2

Data Processing and Analytics COVID-19 Dashboard

Objectives

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

- Download and extract dataset files from Github
- Upload the files into Elastic
- Map the dataset
- Create 3 types of data visualizations which include a geospatial map, a histogram bar chart and a pie chart to create an insightful dashboard

Log into Elastic Cloud

1. Go to https://www.elastic.co/cloud/as-a-service
2. Log into your ES (Elastic Cloud) account
3. Click on Kibana, following page once logging in
4. New tab should open and login with Kibana account

Step 1: Terminal Command to download COVID-19 dataset from GitHub

---

1. Launch and open ElasticSearch, then go to your Kibana Account.
2. We are going to download 2 datasets of CSV format from a github repository (https://github.com/juliom12/covid_19_cases), one containing confirmed cases of Covid-19, and the other with the confirmed death cases. In your terminal, make sure you navigate into the directory you want the files to be downloaded in before using the curl command.

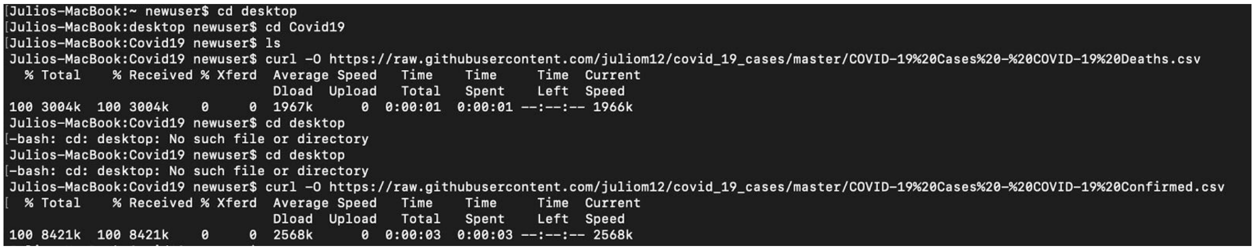
Note: For this download I created a folder on my desktop named Covid19
I navigated into the folder using the terminal by using: cd desktop cd Covid19, your commands may be different but overall make sure you are in the folder you want the files downloaded in.

3. At your Git Bash, download the data files as follows.

curl -O
https://raw.githubusercontent.com/juliom12/covid_19_cases/master/COVID-19%20Cases%20-%20COVID-19%20Confirmed.csv

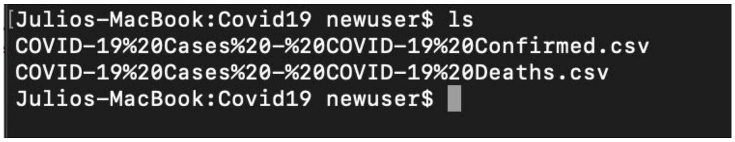
curl -O
https://raw.githubusercontent.com/juliom12/covid_19_cases/master/COVID-19%20Cases%20-%20COVID-19%20Deaths.csv

The result should look like this



```
[Julios-MacBook:~ newuser$ cd desktop
[Julios-MacBook:desktop newuser$ cd Covid19
[Julios-MacBook:Covid19 newuser$ ls
 Julios-MacBook:Covid19 newuser$ curl -O https://raw.githubusercontent.com/juliom12/covid_19_cases/master/COVID-19%20Cases%20-%20COVID-19%20Deaths.csv
   % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                  Dload  Upload   Total   Spent    Left  Speed
 100 3004k  100 3004k    0     0  1967k      0  0:00:01  0:00:01 --:--:-- 1966k
 Julios-MacBook:Covid19 newuser$ cd desktop
[-bash: cd: desktop: No such file or directory
 Julios-MacBook:Covid19 newuser$ cd desktop
[-bash: cd: desktop: No such file or directory
 Julios-MacBook:Covid19 newuser$ curl -O https://raw.githubusercontent.com/juliom12/covid_19_cases/master/COVID-19%20Cases%20-%20COVID-19%20Confirmed.csv
[  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                  Dload  Upload   Total   Spent    Left  Speed
 100 8421k  100 8421k    0     0  2568k      0  0:00:03  0:00:03 --:--:-- 2568k
```

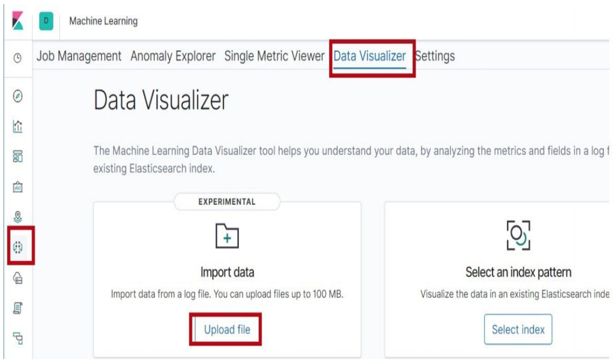
Use the ls command to make sure the files are in your directory, it should look like this.



```
[Julios-MacBook:Covid19 newuser$ ls
 COVID-19%20Cases%20-%20COVID-19%20Confirmed.csv
 COVID-19%20Cases%20-%20COVID-19%20Deaths.csv
 Julios-MacBook:Covid19 newuser$ 
```

Step 2: Upload CSV files into Elastic (Machine Learning-> Data Visualizer)

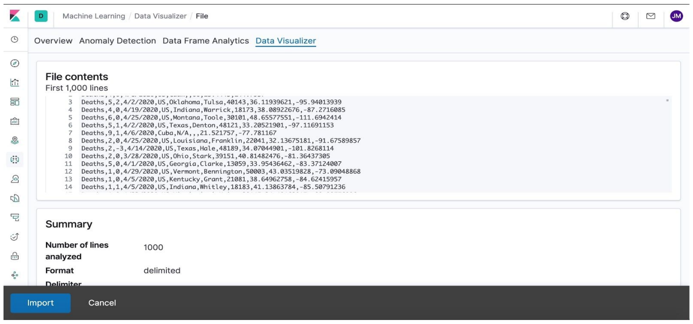
1. Within Kibana, click on Machine Learning
2. In the subnav, click on Data Visualizer



The Machine Learning Data Visualizer tool helps you understand your data, by analyzing the metrics and fields in a log f
existing Elasticsearch index.

EXPERIMENTAL
Import data
Import data from a log file. You can upload files up to 100 MB.
Upload file
Select an index pattern
Visualize the data in an existing Elasticsearch inde
Select index

3. Under Import Data, click Upload File
4. Drag and drop the csv file into the importer, which will show the data



File contents
First 1,000 lines

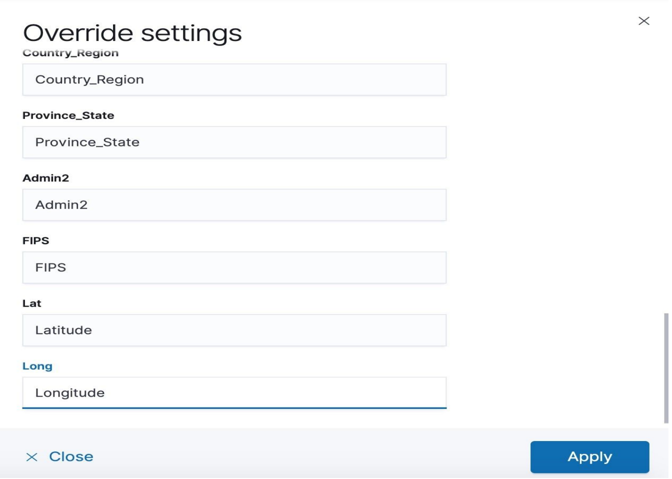
```
Deaths,5,2,4/2/2020,US,Oklahoma,Tulsa,40143,36.11939621,-95.94013939
Deaths,4,0,4/19/2020,US,Indiana,Warrick,18173,38.08922676,-87.2716085
Deaths,6,0,4/25/2020,US,Montana,Toole,30101,48.65577551,-111.6942414
Deaths,5,1,4/2/2020,US,Texas,Denton,48121,33.20521901,-97.11691153
Deaths,9,1,4/6/2020,Cuba,N/A,,,21.521757,-77.781167
Deaths,2,0,4/25/2020,US,Louisiana,Franklin,22041,32.13675181,-91.67589857
Deaths,2,-3,4/14/2020,US,Texas,Hale,48189,34.07044901,-101.8268114
Deaths,2,0,3/28/2020,US,Ohio,Stark,39151,40.81482476,-81.36437305
Deaths,5,0,4/1/2020,US,Georgia,Clarke,13059,33.95436462,-83.37124007
Deaths,1,0,4/29/2020,US,Vermont,Bennington,50003,43.03519828,-73.09048868
Deaths,1,0,4/5/2020,US,Kentucky,Grant,21081,38.64962758,-84.62415957
Deaths,1,1,4/5/2020,US,Indiana,Whitley,18183,41.13863784,-85.50791236
```

Summary

Number of lines analyzed: 1000
Format: delimited
Delimiter

Note: Sometimes the request will timeout. Make sure you have good connection to internet to avoid this from happening

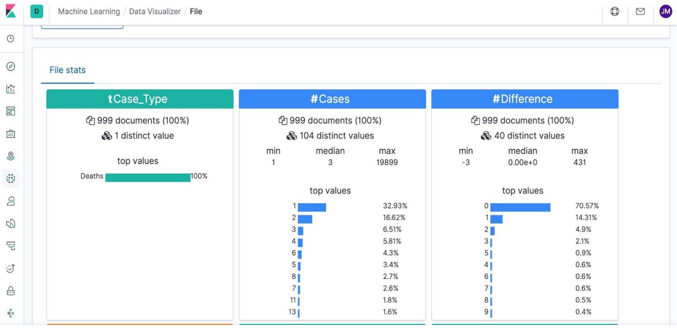
5. Scroll through the list of content. Select Override Settings and <u>change the columns names as</u>: Lat-> Latitude, Long_-> Longitude. Then, select Apply.



Override settings

Country_Region: Country_Region
Province_State: Province_State
Admin2: Admin2
FIPS: FIPS
Lat: Latitude
Long: Longitude

6. Scroll through the list of contents. The fields have types automatically assigned by Kibana



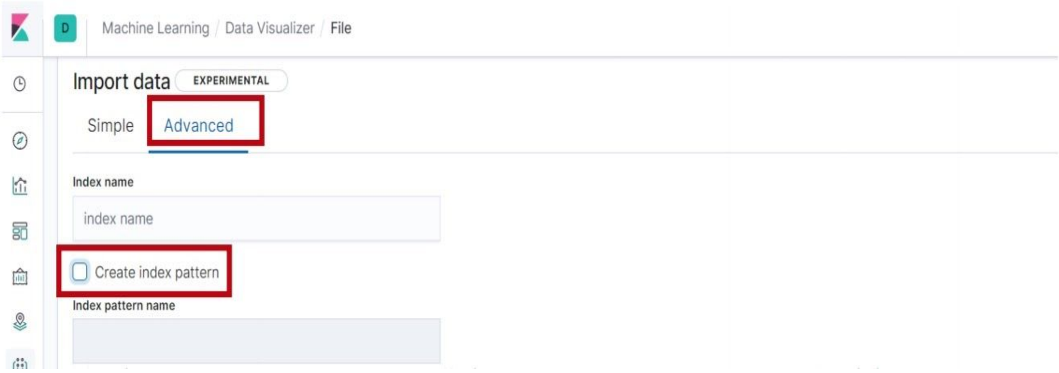
File stats

tCase_Type
999 documents (100%)
1 distinct value
top values
Deaths 100%

#Cases
999 documents (100%)
104 distinct values
min 1, median 3, max 19899
top values
1 32.93%
2 16.62%
3 6.51%
4 5.81%
6 4.3%
5 3.4%
8 2.7%
7 2.6%
11 1.8%
13 1.6%

#Difference
999 documents (100%)
40 distinct values
min -3, median 0.00e+0, max 431
top values
0 70.57%
1 14.31%
2 4.9%
3 2.1%
5 0.9%
4 0.6%
6 0.6%
7 0.6%
8 0.5%
9 0.4%

7. Click Import to upload and index the data. Select Advanced and uncheck “Create Index Pattern”



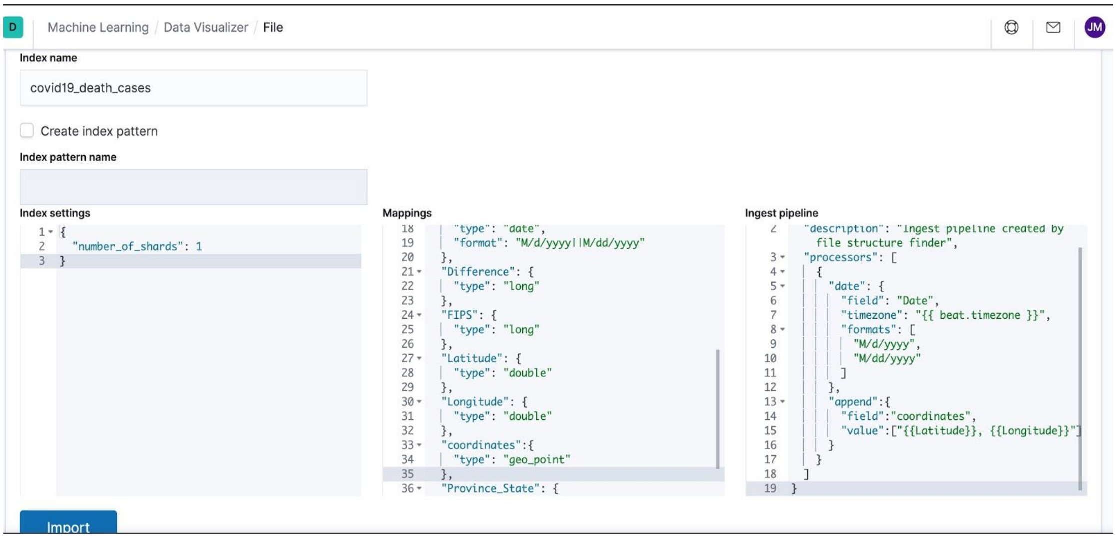
8. In the Mappings frame, add the coordinates elements for geo_point below Longitude:

```
"longitude": {
"type": "double"
 },
   "coordinates": {
   "type": "geo_point"
},
```

9. In the Ingest Pipeline, add coordinates below date elements as follows:

```
        "formats": [
          "M/d/yyyy",
          "M/dd/yyyy"
        ]
      },
     "append":{
        "field":"coordinates",
        "value":["{{Latitude}}, {{Longitude}}"]
      }
```

10. Add the index name: “covid19_death_cases”, the result should look like the following image.



Index name

covid19_death_cases

□ Create index pattern

Index pattern name

Index settings

```
{
  "number_of_shards": 1
}
```

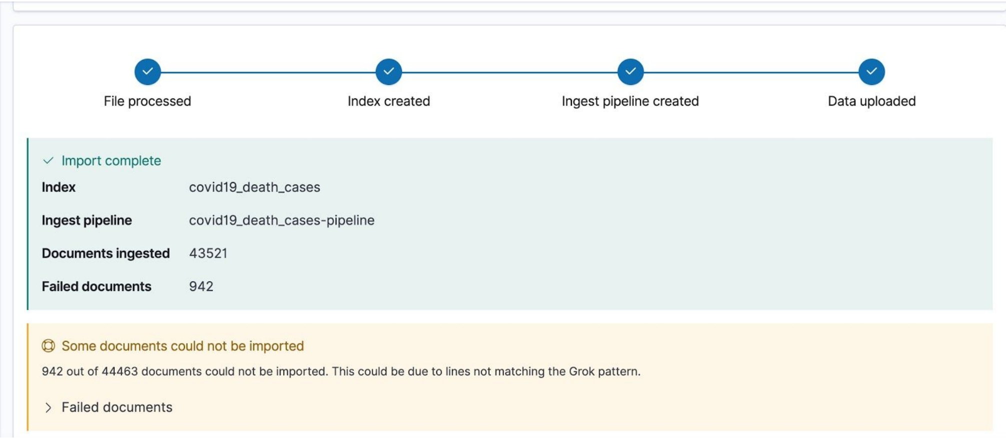
Mappings

```
    "type": "date",
    "format": "M/d/yyyy||M/dd/yyyy"
  },
  "Difference": {
    "type": "long"
  },
  "FIPS": {
    "type": "long"
  },
  "Latitude": {
    "type": "double"
  },
  "Longitude": {
    "type": "double"
  },
  "coordinates":{
    "type": "geo_point"
  },
  "Province_State": {
```

Ingest pipeline

```
  "description": "Ingest pipeline created by
    file structure finder",
  "processors": [
    {
      "date": {
        "field": "Date",
        "timezone": "{{ beat.timezone }}",
        "formats": [
          "M/d/yyyy",
          "M/dd/yyyy"
        ]
      },
      "append":{
        "field":"coordinates",
        "value":["{{Latitude}}, {{Longitude}}"]
      }
    }
  ]
}
```

11. Select import. If no error is found, the following pipeline to import will show up



✓ Import complete

Index covid19_death_cases

Ingest pipeline covid19_death_cases-pipeline

Documents ingested 43521

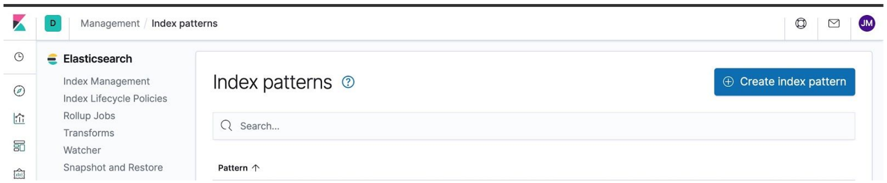
Failed documents 942

Some documents could not be imported

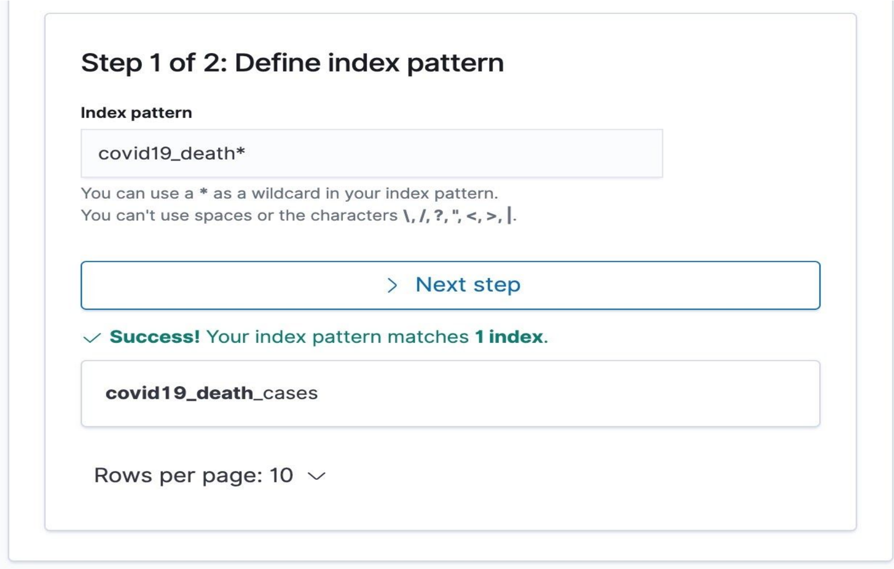
942 out of 44463 documents could not be imported. This could be due to lines not matching the Grok pattern.

Note: it says that some documents could not be imported because they do not match the Grok pattern. This is because some data in the dataset is missing Latitude and Longitude values. However, we will still be able to create a good mapping that gives us a good idea of Covid-19 hotspots.

12. Select Index Patterns Management. Then select "Create Index Pattern"



13. Type in index name: covid19_death, Select Next Step



Step 1 of 2: Define index pattern

Index pattern

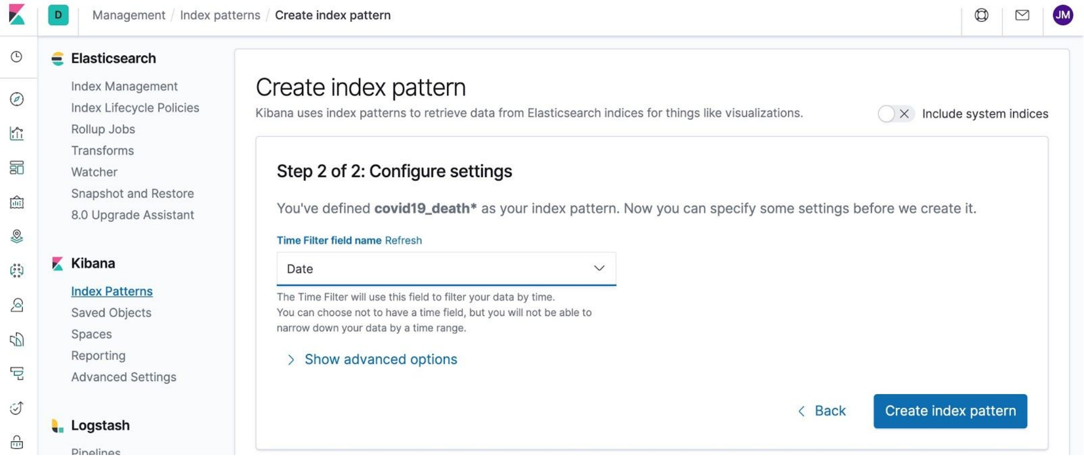
covid19_death*

You can use a * as a wildcard in your index pattern.
You can't use spaces or the characters \, /, ?, ", <, >, |.

✓ **Success!** Your index pattern matches **1 index**.

covid19_death_cases

14. Select Time Filter drop down menu and select Date. Then Select Create Index Pattern



Create index pattern

Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations.

Step 2 of 2: Configure settings

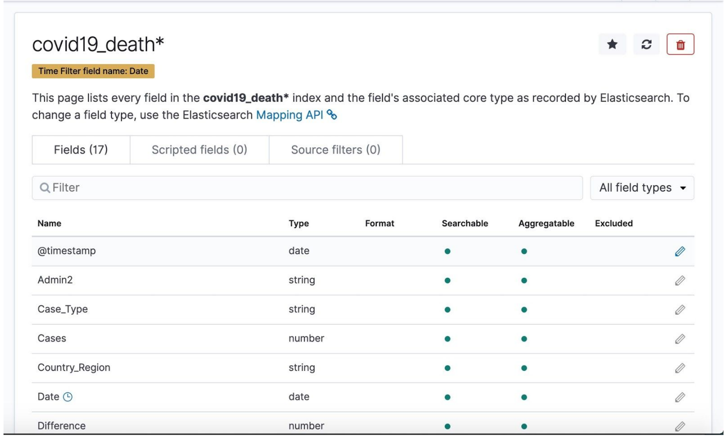
You've defined **covid19_death*** as your index pattern. Now you can specify some settings before we create it.

Time Filter field name Refresh

Date

The Time Filter will use this field to filter your data by time.
You can choose not to have a time field, but you will not be able to narrow down your data by a time range.

15. The following page should appear, showing the process was successful



covid19_death*

Time Filter field name: Date

This page lists every field in the **covid19_death*** index and the field's associated core type as recorded by Elasticsearch. To change a field type, use the Elasticsearch Mapping API

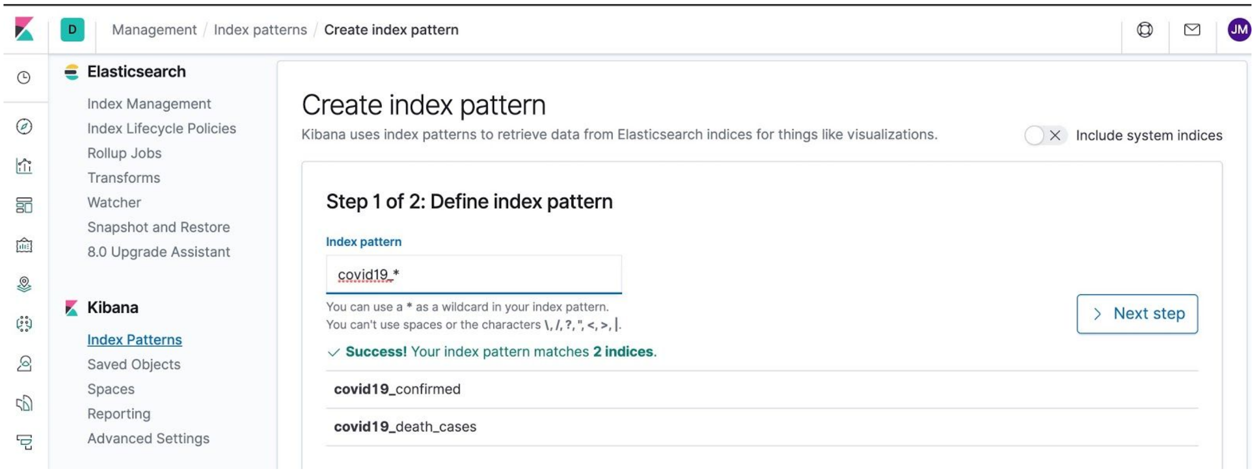
Fields (17) | Scripted fields (0) | Source filters (0)

Filter | All field types

Name	Type	Format	Searchable	Aggregatable	Excluded
@timestamp	date		●	●	
Admin2	string		●	●	
Case_Type	string		●	●	
Cases	number		●	●	
Country_Region	string		●	●	
Date	date		●	●	
Difference	number		●	●	

Now upload the **Confirmed cases data set** and repeat steps 2-14.

You need to create an index pattern that contains these 2 data files: COVID-19….Confirmed.csv, and COVID-19….Deaths.csv. For example, the general index pattern name can be covid19_* to include both files. Select: Management > Index Pattern. Then, fill in the name as covid19_*. Then, select Next step.



Create index pattern

Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations.

Include system indices

Step 1 of 2: Define index pattern

Index pattern

covid19_*

You can use a * as a wildcard in your index pattern.
You can't use spaces or the characters \, /, ?, ", <, >, |.

✓ **Success!** Your index pattern matches **2 indices**.

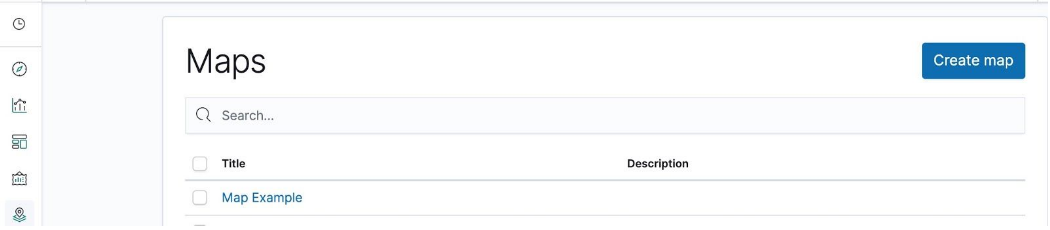
covid19_confirmed

covid19_death_cases

Next step

Step 3: Create Kibana Visualizations using Maps

1. On the left side menu, click on Maps. Then on create map.

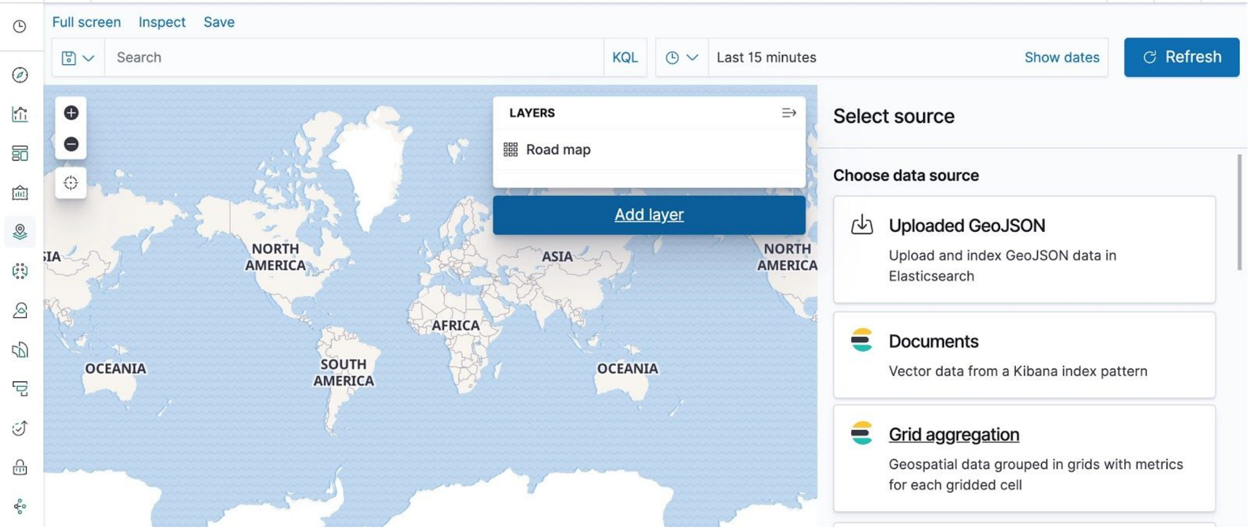


Maps

Search...

□ Title	Description
□ Map Example	

2. Then select Add Layer > Grid Aggregation



Select source

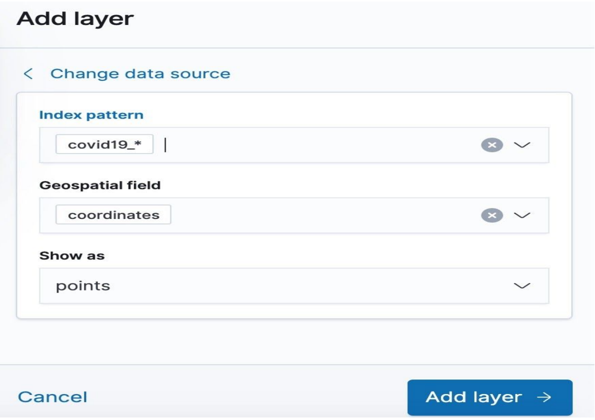
Choose data source

Uploaded GeoJSON
Upload and index GeoJSON data in Elasticsearch

Documents
Vector data from a Kibana index pattern

Grid aggregation
Geospatial data grouped in grids with metrics for each gridded cell

3. At Index Pattern on Grid Aggregation, then select drop-box menu to find your index name: covid19_*



Add layer

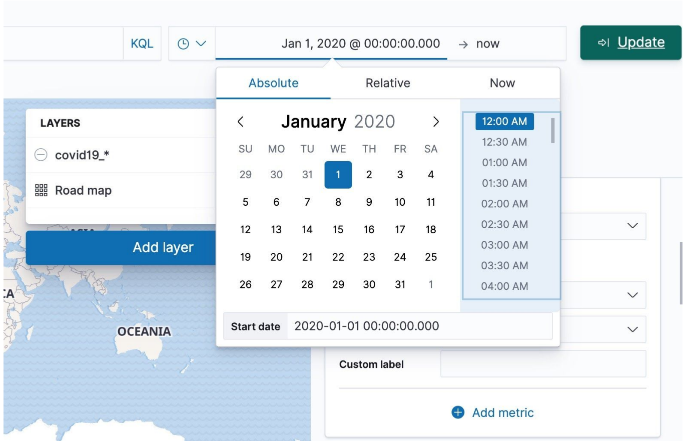
< Change data source

Index pattern: covid19_*

Geospatial field: coordinates

Show as: points

4. Change the start time: Jan 1, 2020 @ 00:00:00:000 & end time: now. Select Update



Jan 1, 2020 @ 00:00:00.000 → now

January 2020

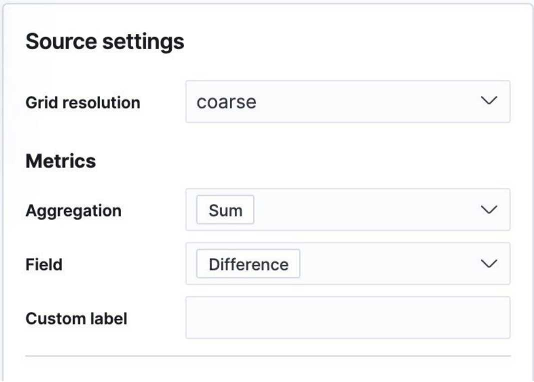
SU	MO	TU	WE	TH	FR	SA
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1

12:00 AM
12:30 AM
01:00 AM
01:30 AM
02:00 AM
02:30 AM
03:00 AM
03:30 AM
04:00 AM

Start date 2020-01-01 00:00:00.000

Custom label

5. Fill in the source setting as follows. For **Aggregation** select: Sum and for **Field** select: Difference



Source settings

Grid resolution: coarse

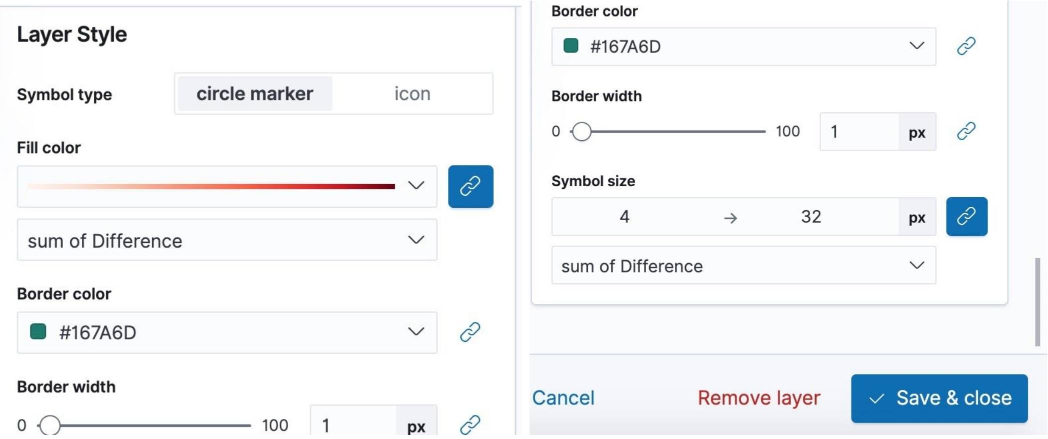
Metrics

Aggregation: Sum

Field: Difference

Custom label:

6. Fill out the form for Circle Markers as follows, then Select Save & Close.



Layer Style

Symbol type: circle marker | icon

Fill color

sum of Difference

Border color

#167A6D

Border width

0 100 1 px

Border color

#167A6D

Border width

0 100 1 px

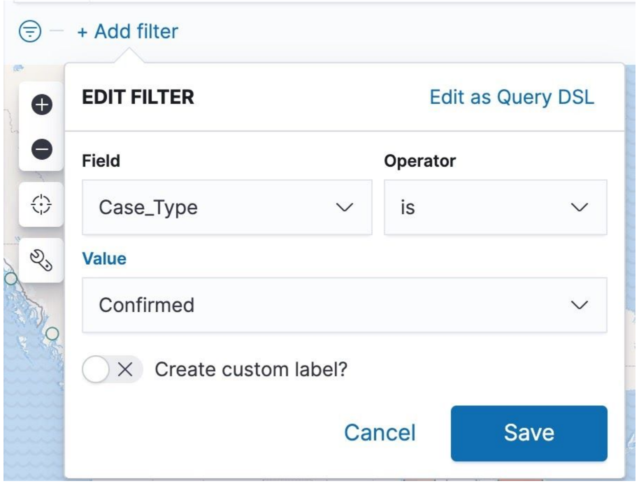
Symbol size

4 → 32 px

sum of Difference

Now we will add a filter by case type to make sure we are looking at the data intended. First we will be looking at Confirmed case type.

7. At the top of the map select **Add Filter,** then for **Field** select: Case_Type, for **Operator** select: is, and for **Value** select: Confirmed



+ Add filter

EDIT FILTER

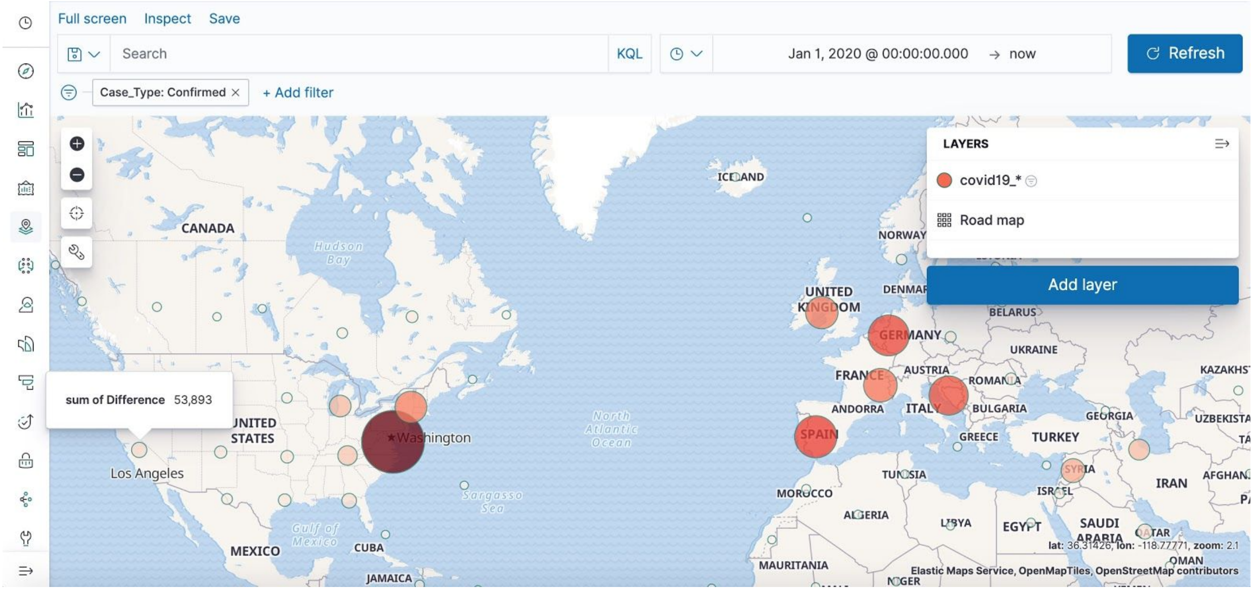
Field: Case_Type

Operator: is

Value: Confirmed

Create custom label?

You will then get the following result:



Note that when we hover over California, we can see the number of confirmed cases is over 50,0000. We can also see that the east coast of the United States has some areas with the greatest amount of confirmed cases, reflected by the bigger and darker geo-point. In fact, states like NewYork have a larger amount of confirmed cases than some entire European countries!

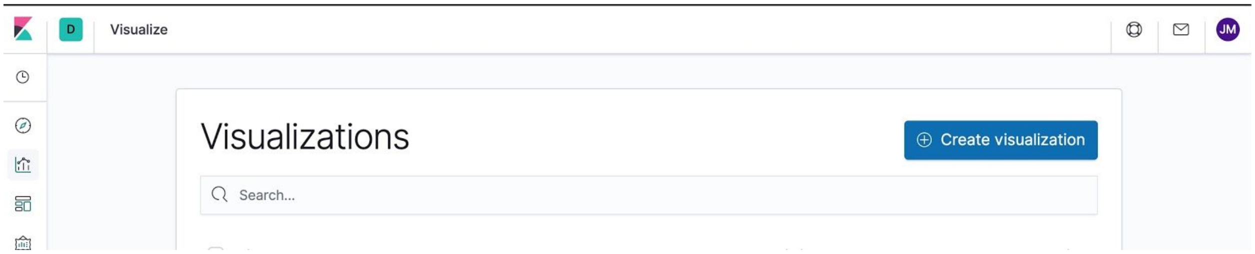
8. **Save** the visualization as: Covid19_map

Now we will create a time series visualization which will help us visualize the increasing sum of cases over time.

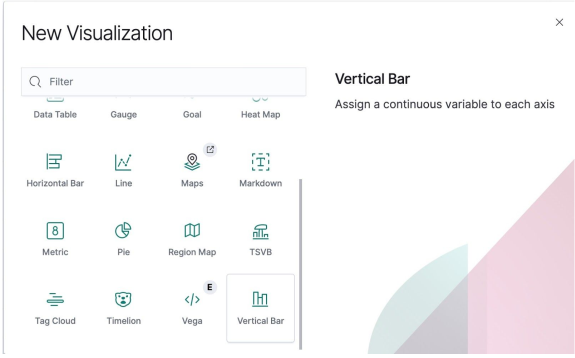
Step 4: Create Histogram Bar Chart Visualization

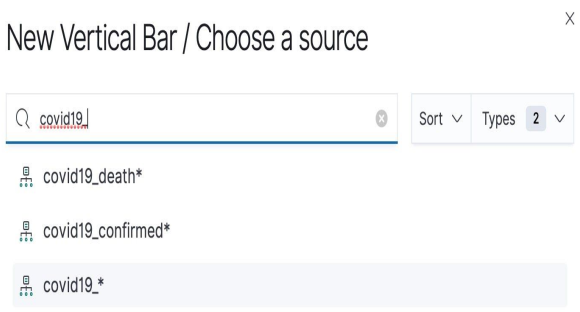
---

1. Select Visualize menu in the left frame, then Select Create Visualization



2. Select and Create a **Vertical Bar** chart & search for covid19_* index. Select create chart.

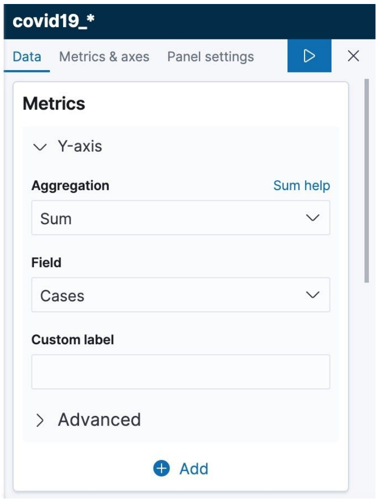



New Vertical Bar / Choose a source

covid19_

Metric Aggregation: Y-Axis configuration

3. In the Metrics pane, expand **Y-Axis** . For **Aggregation** select: Sum and for **Field** select: Cases


covid19_*

Metrics

Y-axis

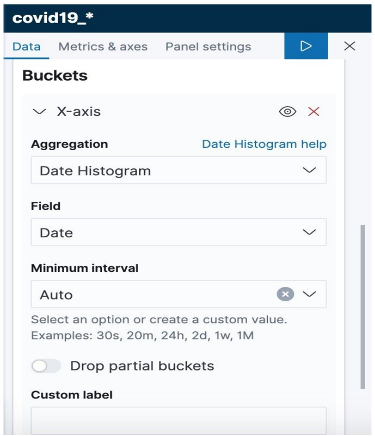
Aggregation: Sum

Field: Cases

Custom label

Buckets Aggregation: X-Axis configuration

4. In the Buckets pane, click Add > Add Buckets > X-Axis. For **Aggregation** select: Date Histogram and for **Field** select: Date


covid19_*

Buckets

X-axis

Aggregation: Date Histogram

Field: Date

Minimum interval: Auto

Select an option or create a custom value. Examples: 30s, 20m, 24h, 2d, 1w, 1M

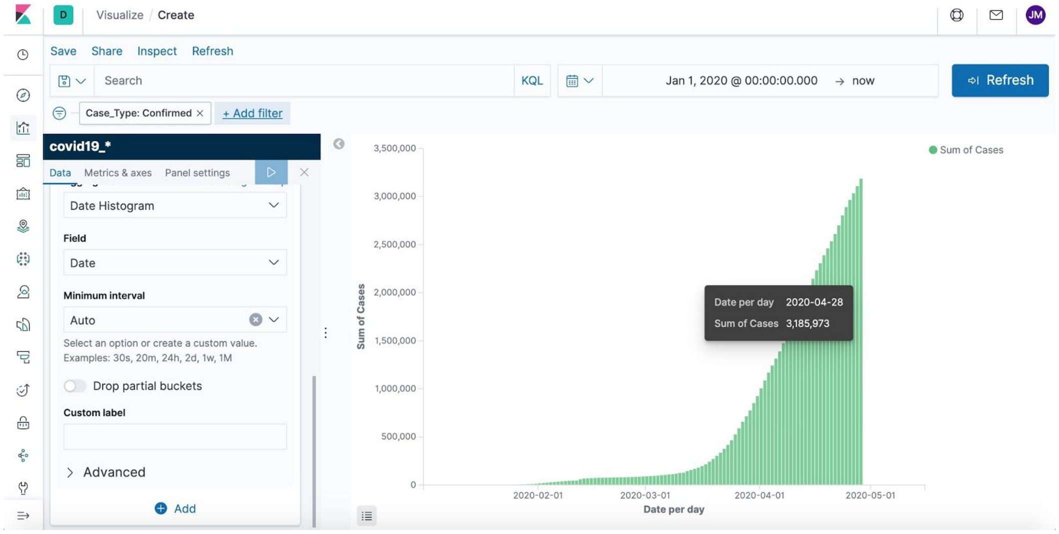
○ Drop partial buckets

Custom label

Just like in the map visualization, make sure you select the **start time** to be: Jan 1, 2020 @ 00:00:00:000 & end time: now

5. Now **filter** the data by clicking +Add Filter, then click the **Field dropdown** and select: Case_Type. For **Operator** select: is. For **Value** select: Confirmed

6. Click the blue "Run" button, which should give you the following result.



Jan 1, 2020 @ 00:00:00.000 → now

Case_Type: Confirmed

covid19_*

Date Histogram

Field: Date

Minimum interval: Auto

Select an option or create a custom value. Examples: 30s, 20m, 24h, 2d, 1w, 1M

Drop partial buckets

Custom label

Date per day: 2020-04-28
Sum of Cases: 3,185,973

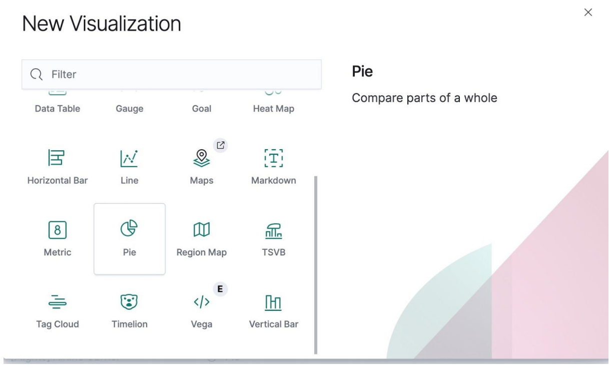
Note: We can see that the number of confirmed cases was consistent through the beginning of March, then skyrocketed through the end of march to now. (04/28/2020 was the last case in the dataset). Thus, we can see the total cases have exceeded over 3 million !

7. **Remove the Filter as we will be adding one later in the dashboard.**

Save the visualization as covid19_histogram_bar_chart

Step 5: Create a Pie Chart Visualization

1. Select Visualize Menu. Click Create a Visualization. Click on Pie

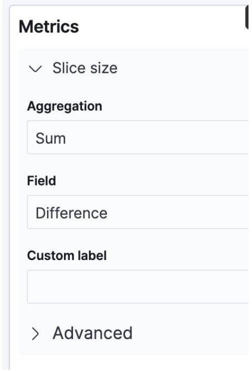


Pie

Compare parts of a whole

2. Select covid19_* index for the visualization. Select create visualization.

In the Metrics pane, dropdown the **Slice size** option, then In **Aggregation** select: Sum, and in **Field** select: Difference



Metrics

Slice size

Aggregation

Sum

Field

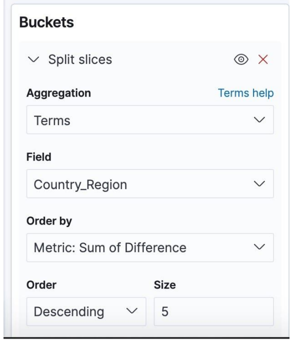
Difference

Custom label

Bucket Aggregation

1. In the Buckets pane, click +Add -> Split Slices.
2. In the **Aggregation** field, select: Terms. In **Field** select: Country_Region, and in **Order by** select Metric: Sum of Difference. Select **Order**: Descending, and **Size**: 5

Note: You can increase the size to greater than five but for clarity, we will only be looking at the Top 5 Countries with confirmed cases.



Buckets

Split slices

Aggregation

Terms

Field

Country_Region

Order by

Metric: Sum of Difference

Order

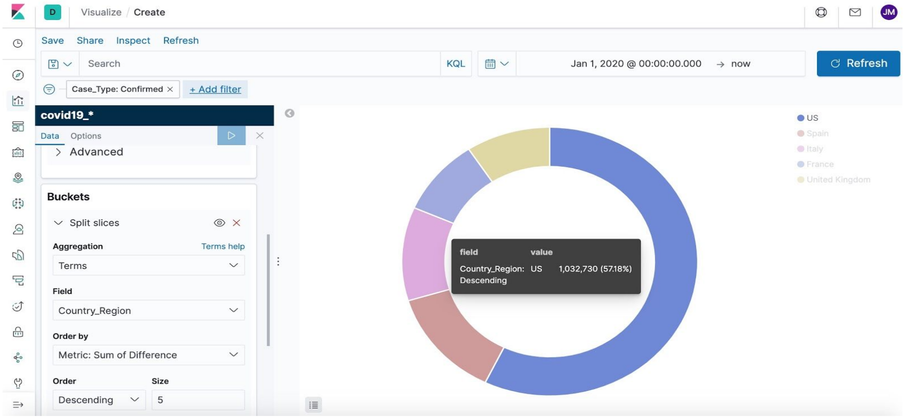
Descending

Size

5

3. Click Apply Changes and run the visualization.
4. Add the same Filter as in the previous visualization: **Field**: Case_Type, **Operator**: is, **Value**: Confirmed.

Once the changes have been applied you should get the following result.



Jan 1, 2020 @ 00:00:00.000 → now

Case_Type: Confirmed

covid19_*

Aggregation: Terms

Field: Country_Region

Order by: Metric: Sum of Difference

Order: Descending

Size: 5

field	value
Country_Region: US	1,032,730 (57.18%)
Descending	

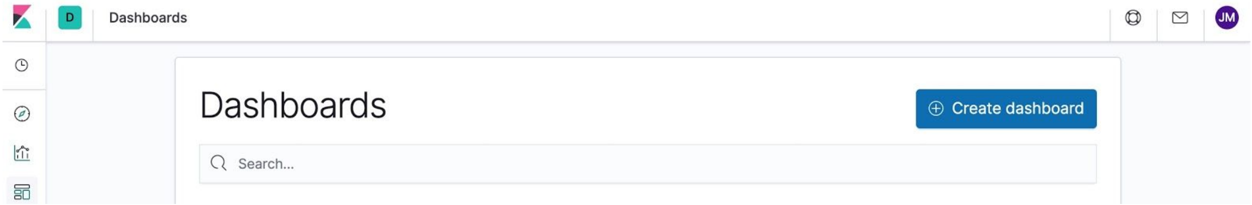
Note: As we hover over the "US" field, we can observe that they account for over 50% of cases worldwide with over 1 million confirmed, followed by Spain, Italy, France and the United Kingdom.

Remove the filter as we will be adding one in the dashboard.

Save visualization as : covid19_pie_chart

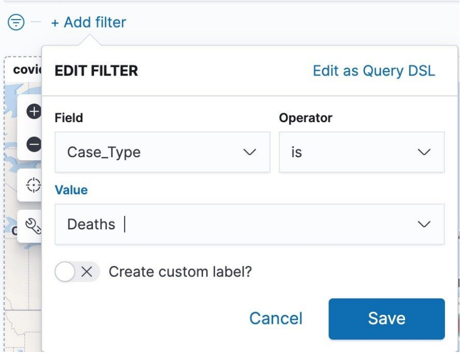
Now with all of our visualizations created we will be integrating them into a dashboard.

1. Navigate to the Dashboard button on the menu to the left-hand side. Select **Create dashboard**



2. In the dashboard pane, select **Add** then when prompted select: **covid19_map**, **covid19_histogram_bar_chart**, and **covid19_pie_chart** visualizations.

3. Now select **Add Filter**. For **Field** select: Case_Type, **Operator** select: is, and **Value** select: Deaths



EDIT FILTER

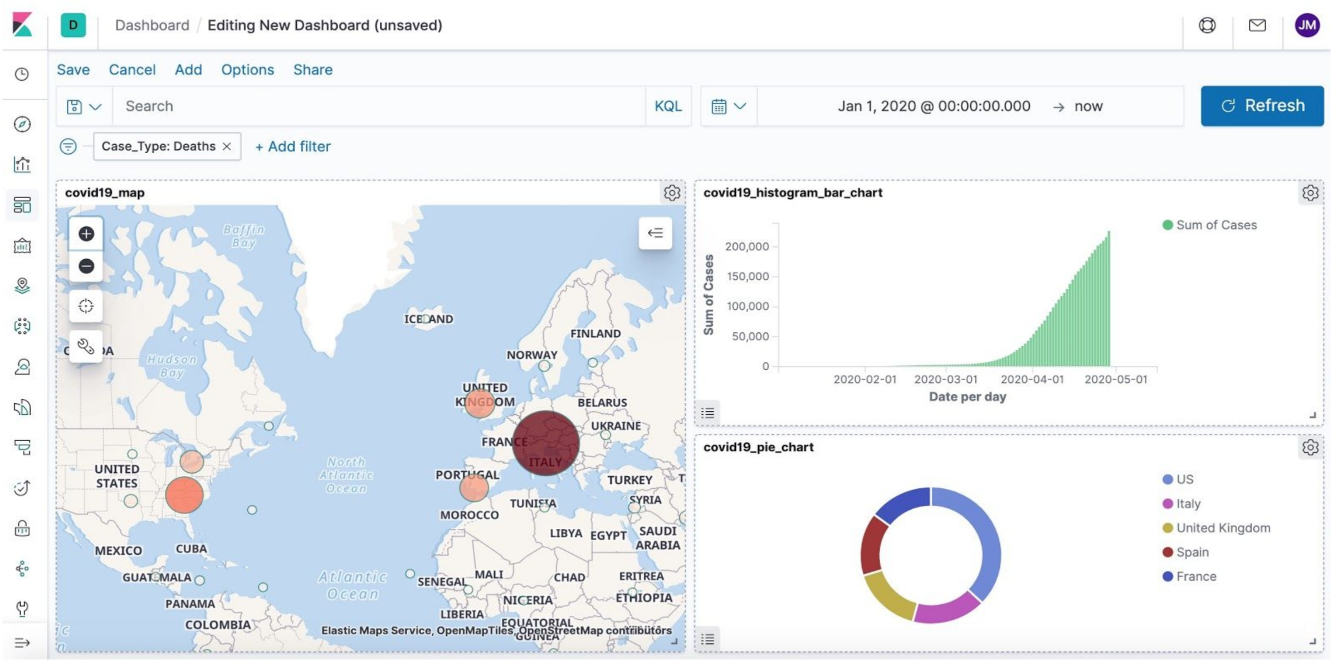
Field: Case_Type

Operator: is

Value: Deaths

○ Create custom label?

You should now get the following result:



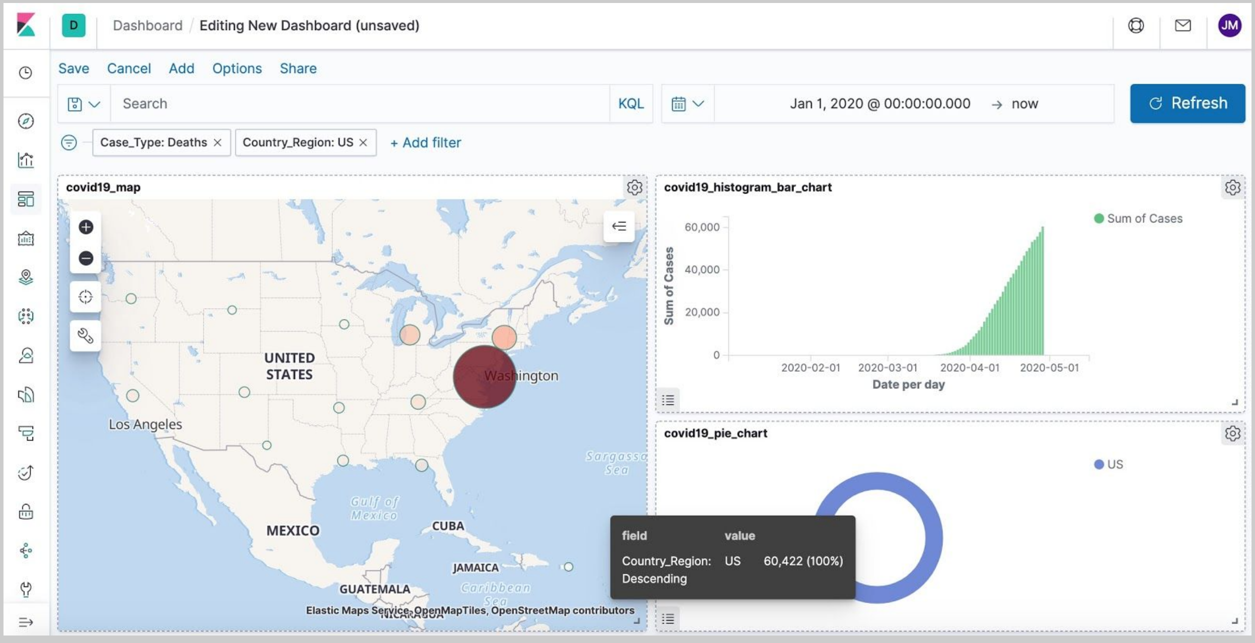
Case_Type: Deaths

Jan 1, 2020 @ 00:00:00.000 → now

Note: Now, we are looking at the Death cases around the world. Overall, we can see that there are just over 200,000 confirmed death cases worldwide. Also, Italy now climbs up in the ranks to 2nd within the pie chart and Spain moves down to 4th in the Top 5 of confirmed death cases worldwide. We can also navigate within the map to look for hotspots of Covid-19 death cases.

We can go one step further by adding another filter to get more insight.
Select **+Add Filter**, then For **Field** select: Country_Region. For **Operator** select: is. For **Value** select: US

You will then get the following result:



Jan 1, 2020 @ 00:00:00.000 → now

Case_Type: Deaths ×   Country_Region: US ×

field	value	
Country_Region: Descending	US	60,422 (100%)

Note: We can now see how rapidly the death cases grew within the US. Around the beginning of April, the amount of cases skyrocketed, where as of 4/28/2020 (The last case in the dataset) there were a little over 60,000 death cases in the US, most of them coming from the east coast from around the New York area.

References

1. ElasticSearch with CSV Data Import and Visualization [1] Jongwook Woo, PhD
2. Data World, John Hopkins University Dataset, https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv
3. GitHub, https://github.com/juliom12/covid_19_cases
4. Dropbox, https://www.dropbox.com/s/smw5aein7i8zr1d/covid-19-data-resource-hub-covid-19-case-counts%20%281%29.zip?dl=0
5. Elasticsearch https://www.elastic.co/