

# 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](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](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](https://raw.githubusercontent.com/juliom12/covid_19_cases/master/COVID-19%20Cases%20-%20COVID-19%20Deaths.csv)

The result should look like this

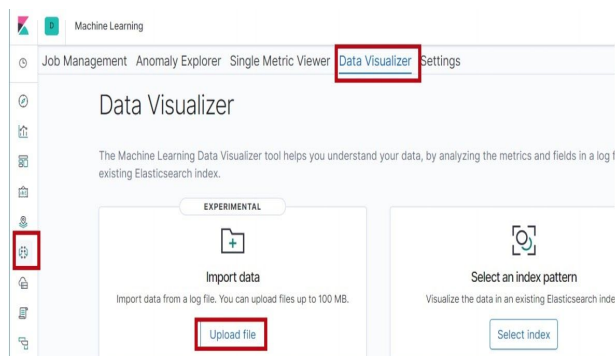
```
Julios-MacBook:~ newuser$ cd desktop
Julios-MacBook:desktop newuser$ cd Covid19
Julios-MacBook:Covid19 newuser$ ls
Julios-MacBook:Covid19 newuser$ curl -O https://raw.githubusercontent.com/juliom12/covid_19_cases/master/COVID-19%20Cases%20-%20COVID-19%20Deaths.csv
% Total    % Received % Xferd  Average 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.

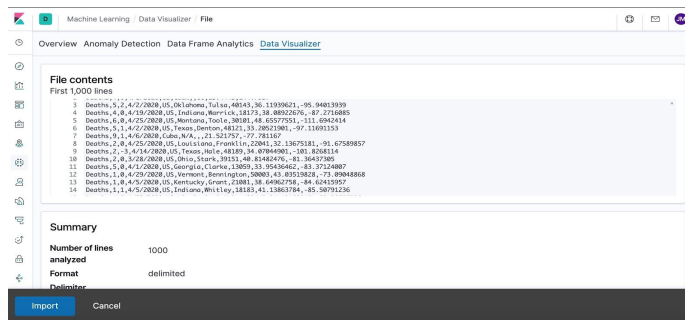
```
Julios-MacBook:Covid19 newuser$ ls
COVID-19%20Cases%20-%20COVID-19%20Confirmed.csv
COVID-19%20Cases%20-%20COVID-19%20Deaths.csv
Julios-MacBook:Covid19 newuser$
```

## 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



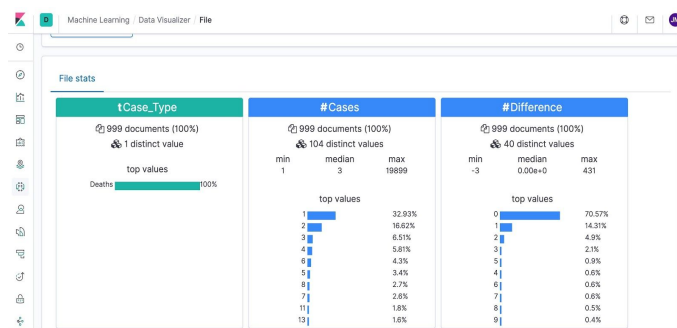
- Under Import Data, click Upload File
- Drag and drop the csv file into the importer, which will show the data



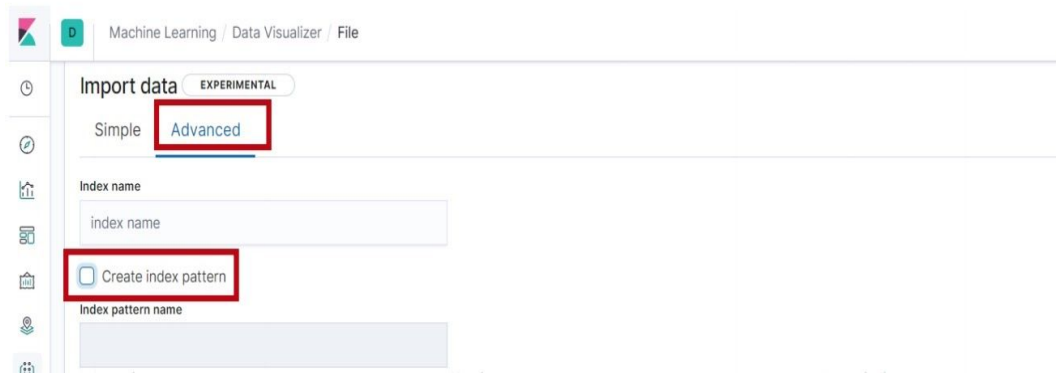
Note: Sometimes the request will timeout. Make sure you have good connection to internet to avoid this from happening

- Scroll through the list of content. Select **Override Settings** and change the columns names as: Lat-> Latitude, Long\_-> Longitude. Then, select Apply.

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



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



- In the Mappings frame, add the coordinates elements for geo\_point below Longitude:

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

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

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

- Add the index name: “covid19\_death\_cases”, the result should look like the following image.

Machine Learning / Data Visualizer / File

Index name  
covid19\_death\_cases

☐ Create index pattern

Index pattern name

Index settings

```

1 - {
2   "number_of_shards": 1
3 }

```

Mappings

```

18 | "type": "date",
19 | "format": "M/d/yyyy||M/dd/yyyy"
20 |
21 - "Difference": {
22 | "type": "long"
23 |
24 - "FIPS": {
25 | "type": "long"
26 |
27 - "Latitude": {
28 | "type": "double"
29 |
30 - "Longitude": {
31 | "type": "double"
32 |
33 - "coordinates": {
34 | "type": "geo_point"
35 |
36 - "Province_State": {

```

Ingest pipeline

```

2 | "description": "Ingest pipeline created by
3 | file structure finder",
4 | "processors": [
5 | {
6 |   "date": {
7 |     "field": "Date",
8 |     "timezone": "{{ beat.timezone }}",
9 |     "formats": [
10 |       "M/d/yyyy",
11 |       "M/dd/yyyy"
12 |     ]
13 |   },
14 |   "append": {
15 |     "field": "coordinates",
16 |     "value": "[{{Latitude}}, {{Longitude}}]"
17 |   }
18 | }
19 | ]

```

Import

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

File processed ✓ Index created ✓ Ingest pipeline created ✓ Data uploaded ✓

✓ 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.

> Failed documents

**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”

Management / Index patterns

Elasticsearch

Index Management  
Index Lifecycle Policies  
Rollup Jobs  
Transforms  
Watcher  
Snapshot and Restore

Index patterns ⓘ

Create index pattern

Search...

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 \, /, ?, ", <, >, |.

> Next step

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

covid19\_death\_cases

Rows per page: 10 ▾

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

Management / Index patterns / Create index pattern

Elasticsearch

Index Management

Index Lifecycle Policies

Rollup Jobs

Transforms

Watcher

Snapshot and Restore

8.0 Upgrade Assistant

Kibana

Index Patterns

Saved Objects

Spaces

Reporting

Advanced Settings

Logstash

Dinalline

## Create index pattern

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

☐ Include system indices

### 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.

> Show advanced options

< Back Create index pattern

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)

Q 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.

Management / Index patterns / Create index pattern

Elasticsearch

- Index Management
- Index Lifecycle Policies
- Rollup Jobs
- Transforms
- Watcher
- Snapshot and Restore
- 8.0 Upgrade Assistant

Kibana

- [Index Patterns](#)
- Saved Objects
- Spaces
- Reporting
- Advanced Settings

### 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

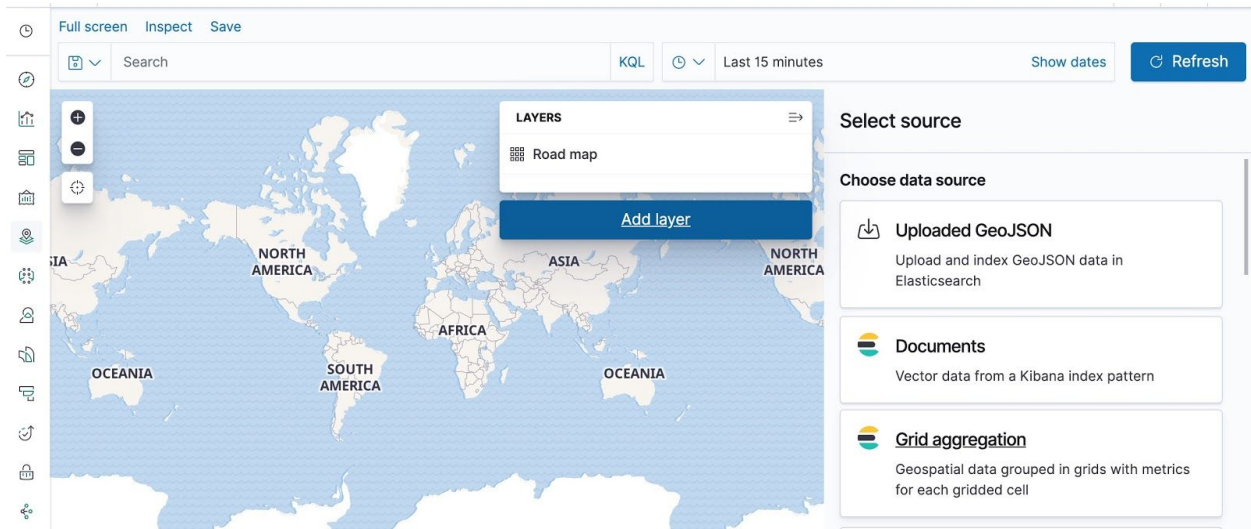
[Create map](#)

Search...

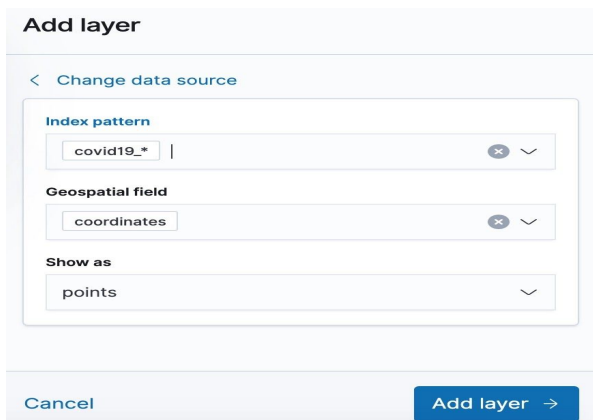
☐ Title      Description

☐ [Map Example](#)

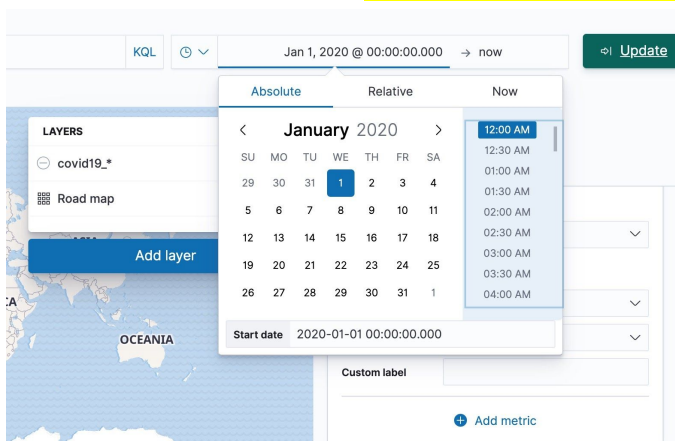
## 2. Then select Add Layer > Grid Aggregation



## 3. At Index Pattern on Grid Aggregation, then select drop-box menu to find your index name: covid19\_\*

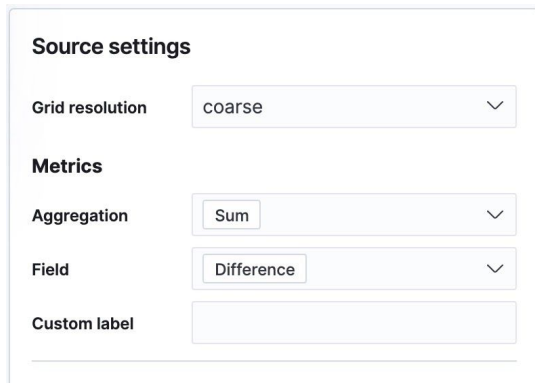


## 4. Change the start time: Jan 1, 2020 @ 00:00:00:000 & end time: now. Select Update





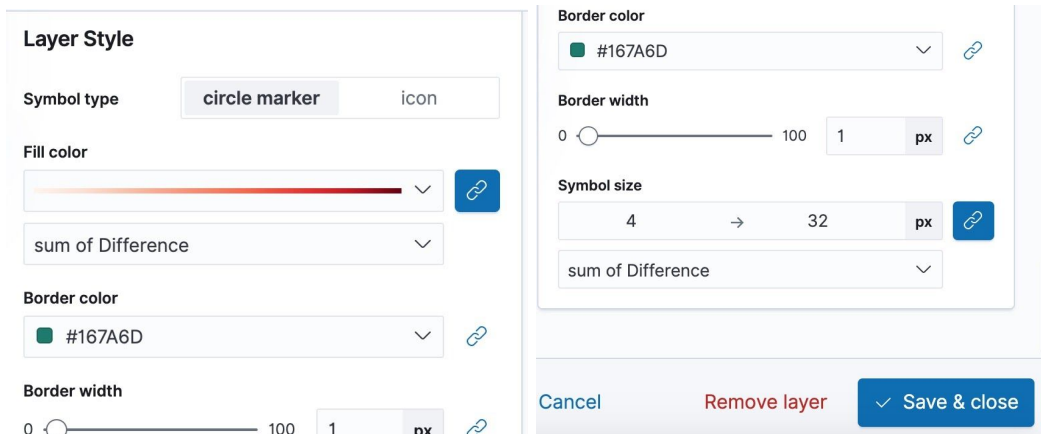
5. Fill in the source setting as follows. For **Aggregation** select: **Sum** and for **Field** select: **Difference**



The 'Source settings' form contains the following fields:

- Grid resolution:** A dropdown menu with 'coarse' selected.
- Metrics:**
  - Aggregation:** A dropdown menu with 'Sum' selected.
  - Field:** A dropdown menu with 'Difference' selected.
  - Custom label:** An empty text input field.

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



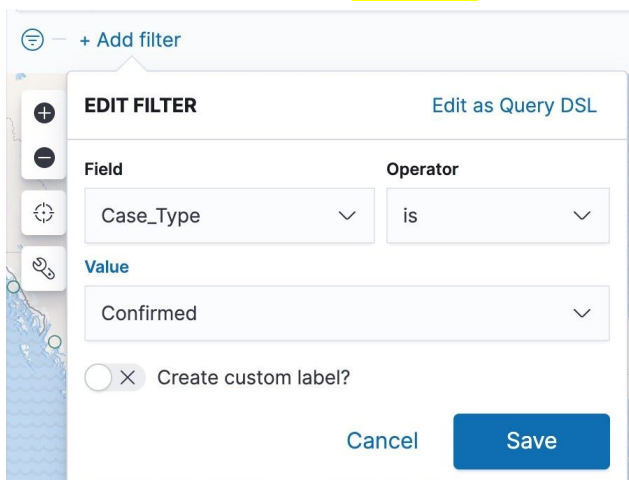
The 'Layer Style' form for 'circle marker' includes the following settings:

- Symbol type:** 'circle marker' (selected) and 'icon'.
- Fill color:** A color bar with a gradient from orange to red, and a dropdown menu showing 'sum of Difference'.
- Border color:** A dropdown menu with '#167A6D' selected.
- Border width:** A slider from 0 to 100, with '1' selected and 'px' as the unit.
- Symbol size:** A range from 4 to 32, with '32' selected and 'px' as the unit.
- sum of Difference:** A dropdown menu.

At the bottom, there are three buttons: 'Cancel', 'Remove layer', and 'Save & close'.

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**

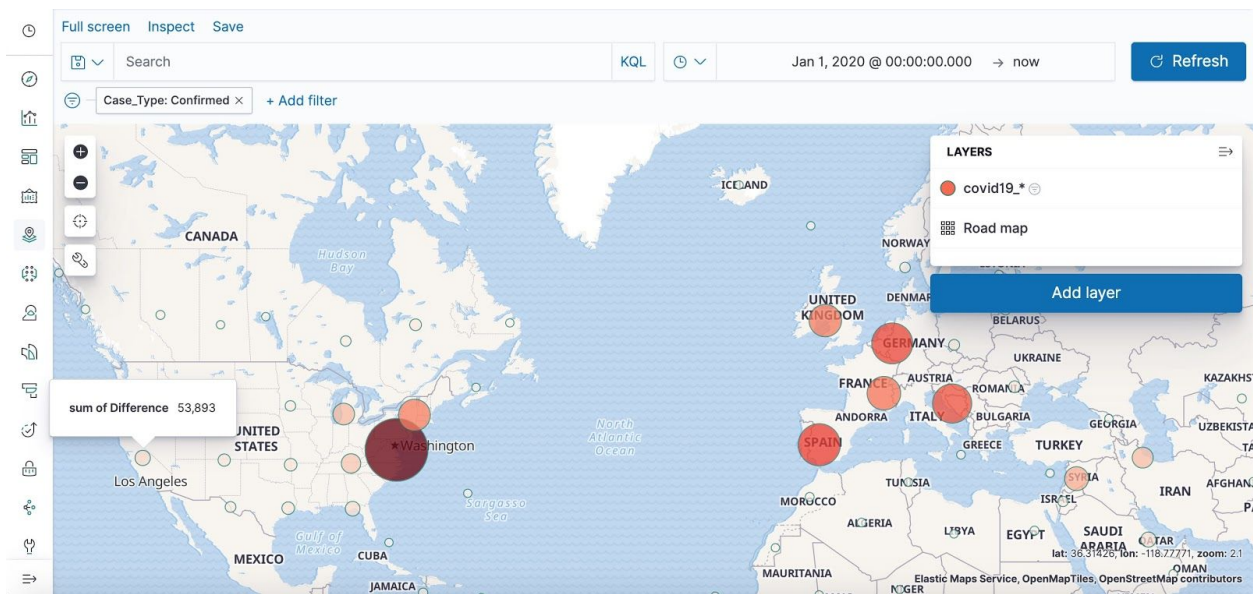


The 'Add Filter' dialog shows the following configuration:

- Field:** 'Case\_Type' (selected from a dropdown).
- Operator:** 'is' (selected from a dropdown).
- Value:** 'Confirmed' (selected from a dropdown).
- Create custom label?:** An unchecked checkbox.

Buttons at the bottom: 'Cancel' and 'Save'.

You will then get the following result:



**Note** that when we hover over California, we can see the number of confirmed cases is over 50,000. 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 New York have a larger amount of confirmed cases than some entire European countries!

8. **Save** the visualization as: Covid19\_map

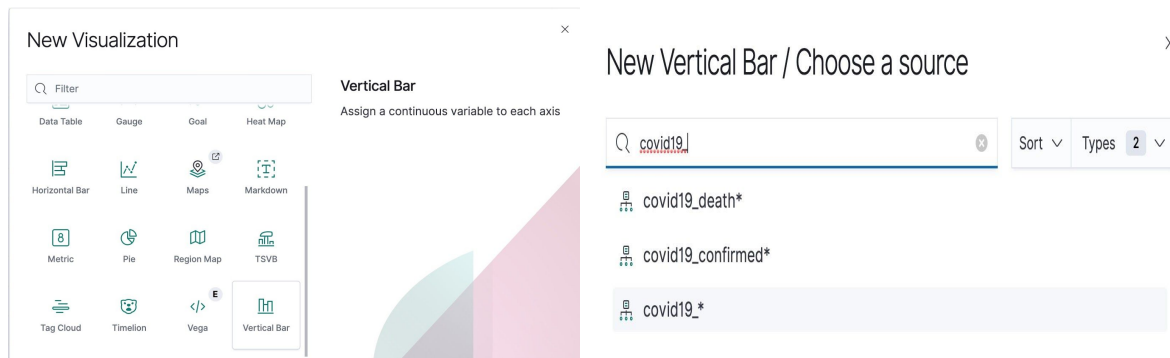
Now we will create a time series visualization which will help us visualize the increasing sum of cases over time.

## Step 4: Create Histogram Bar Chart Visualization

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

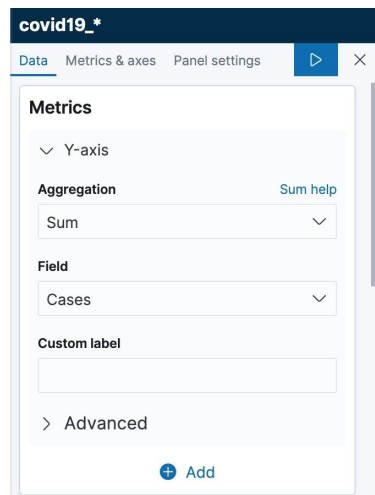


2. Select and Create a **Vertical Bar** chart & search for covid19\_\* index. Select create chart.



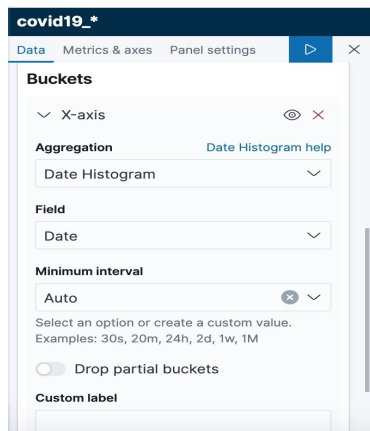
### Metric Aggregation: Y-Axis configuration

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



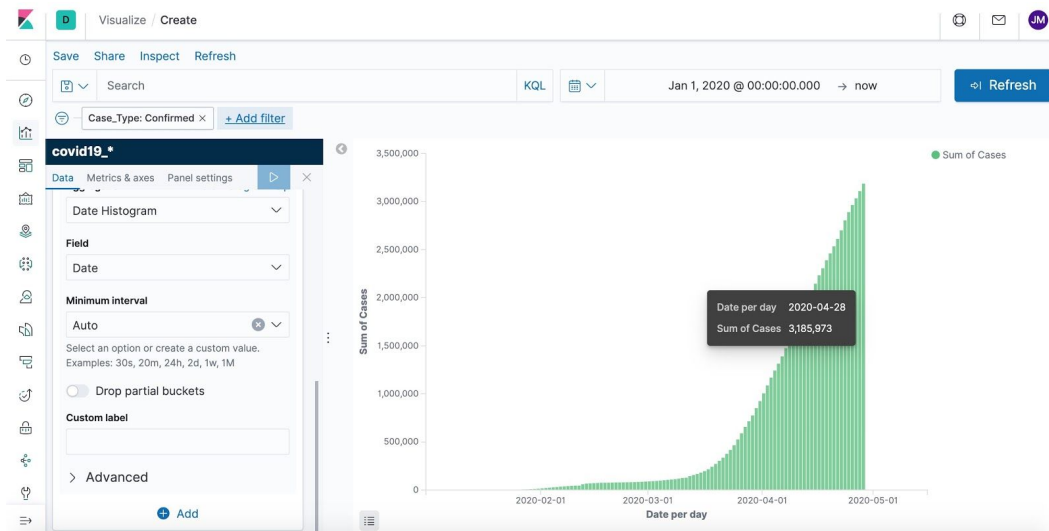
### Buckets Aggregation: X-Axis configuration

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



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**

- 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**
- Click the blue “Run” button, which should give you the following result.

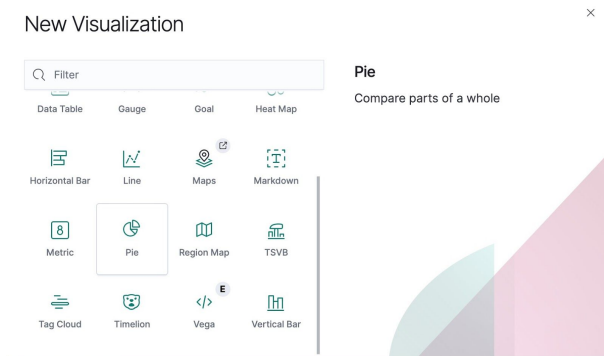


**Note:** We can see that the number of confirmed cases was consistent through the beginning of March, then skyrocketed through the end of March to now. (04/28/2020 was the last case in the dataset). Thus, we can see the total cases have exceeded over 3 million !

- Remove the Filter as we will be adding one later in the dashboard.**  
Save the visualization as covid19\_histogram\_bar\_chart

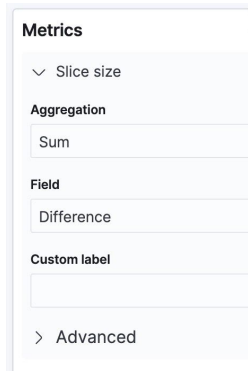
## Step 5: Create a Pie Chart Visualization

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



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

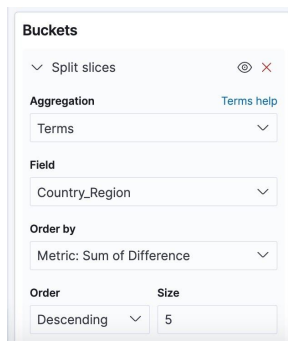
Custom label

> Advanced

### 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

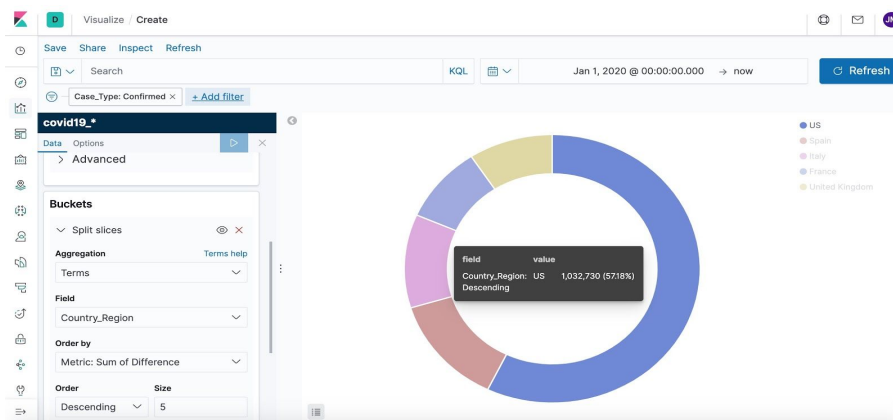
Descending

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.



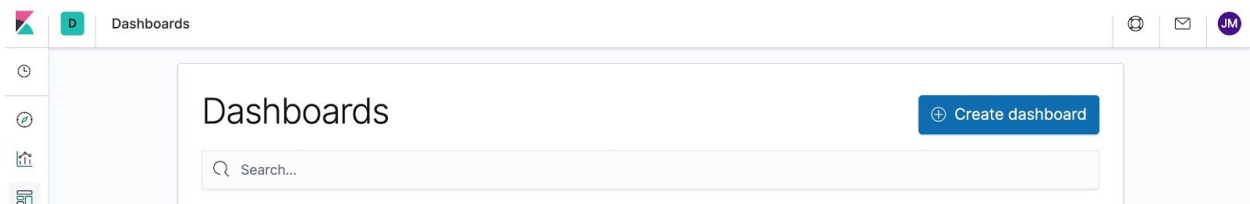
Note: As we hover over the “US” field, we can observe that they account for over 50% of cases worldwide with over 1 million confirmed, followed by Spain, Italy, France and the United Kingdom.

**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**

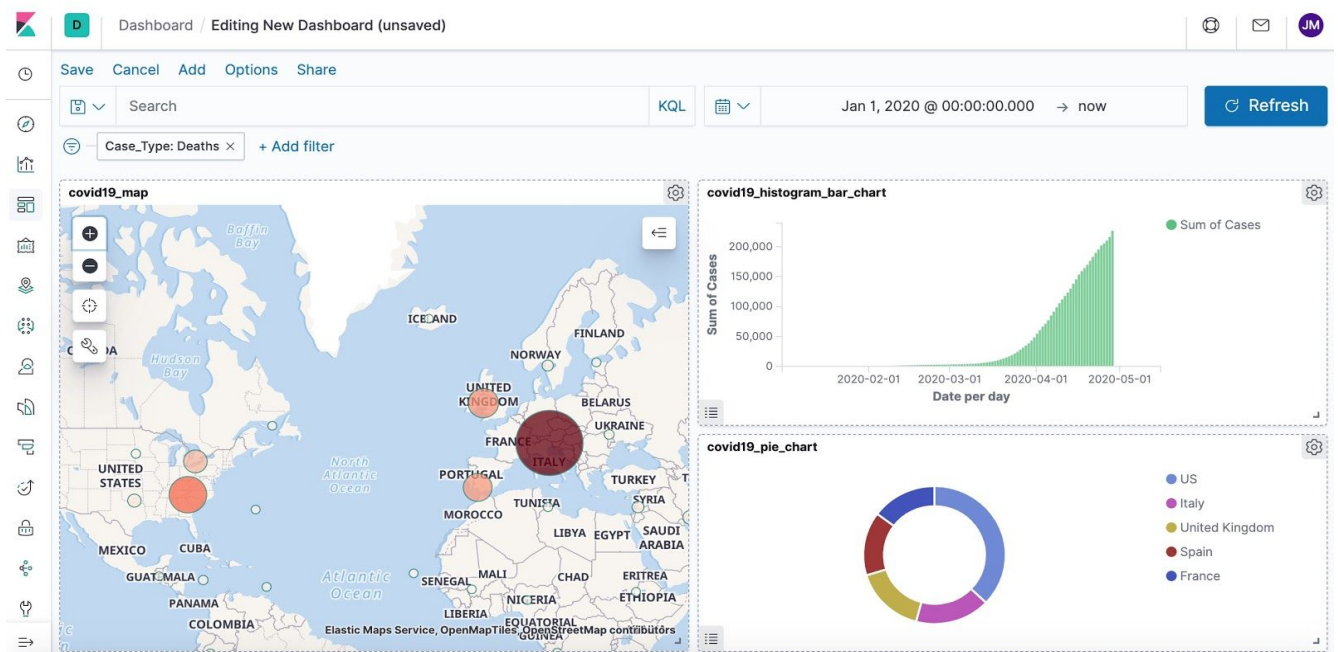
**EDIT FILTER** [Edit as Query DSL](#)

Field	Operator	Value
Case_Type	is	Deaths

☐ Create custom label?

[Cancel](#) [Save](#)

You should now get the following result:

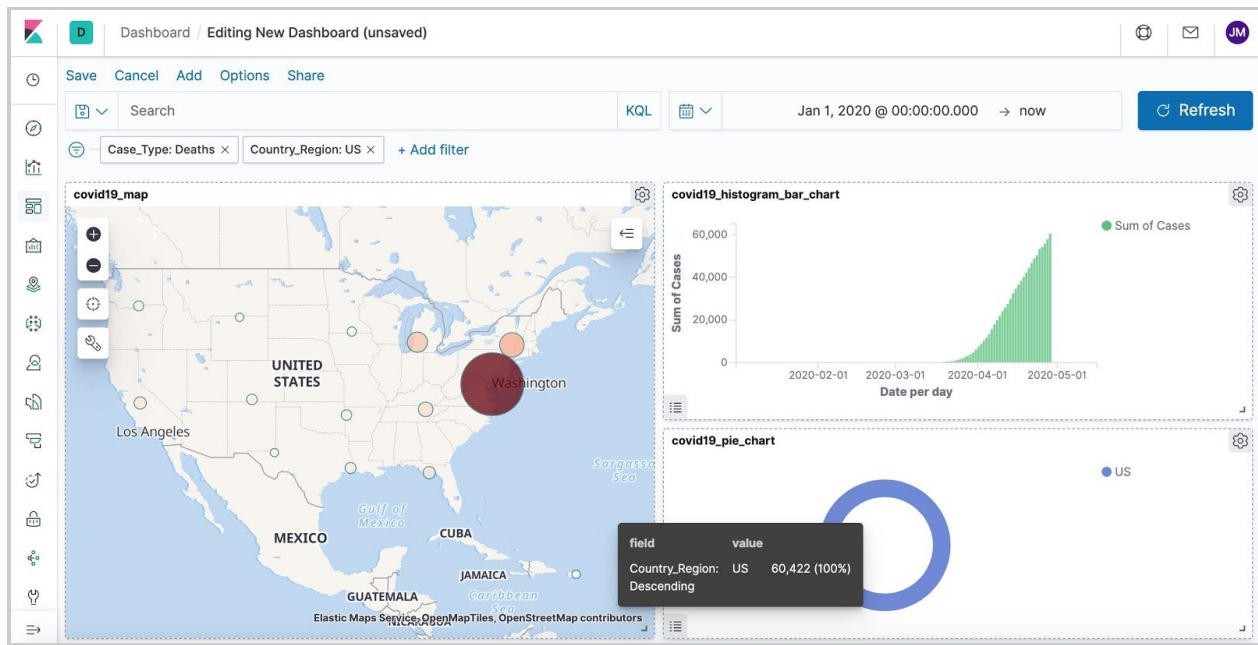


Note: Now, we are looking at the Death cases around the world. Overall, we can see that there are just over 200,000 confirmed death cases worldwide. Also, Italy now climbs up in the ranks to 2nd within the pie chart and Spain moves down to 4th in the Top 5 of confirmed death cases worldwide. We can also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



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. 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>
2. GitHub, [https://github.com/juliom12/covid\\_19\\_cases](https://github.com/juliom12/covid_19_cases)
3. Dropbox,  
<https://www.dropbox.com/s/smw5aein7i8zr1d/covid-19-data-resource-hub-covid-19-case-counts%20%281%29.zip?dl=0>
4. Elasticsearch <https://www.elastic.co/>