**ZikaVirus Analysis and Out-Break**

Insight on Zika Virus and its outbreak

Data set is having 107620 records

This project is done using Pig, HCatalog and Hive

**ZikaVirus Analysis and Out-Break**

This dataset shares publicly available data related to the ongoing Zika epidemic. It is being provided as a resource to the scientific community engaged in the public health response. The data provided here is not official and should be considered provisional and non-exhaustive. The data in reports may change over time, reflecting delays in reporting or changes in classifications. And while accurate representation of the reported data is the objective in the machine readable files shared here, that accuracy is not guaranteed.

Exploratory Questions:

1. Loading and Cleaning of Data

2. Count of Data

3. Most affected country in terms of Zika confirmed and Zika suspected cases, as well as the count of total number of Zika confirmed and Zika suspected cases in all the available countries in the dataset.

4. As the analysis in the dataset is weekly or twice a week, reported date wise analysis of most number of Zika confirmed and Zika discarded cases, which will be helpful in understanding in which month or season a country is most affected to the disease, which in turn helps in understanding the temperature or the climate condition that leading the source of this disease.

5. Individual country’s reported date data to analyze the number of Zika confirmed and suspected cases increasing or decreasing day by day, which helps in identifying the countries where it’s growing adversely as well as about the countries taking precaution against the disease where the numbers are controlled or decreasing.

6. Prepare a report to find out the time (week) which has most Zika effected cases of different countries in a given year.

7. Find out the second top most month which has Zika effected cases.

8. Visualization of the results observed.

**Hive Queries**

/\*Creating a database and a table in database in Hive\*/

create database bigdata\_foundation\_project\_stephy;

use bigdata\_foundation\_project\_stephy;

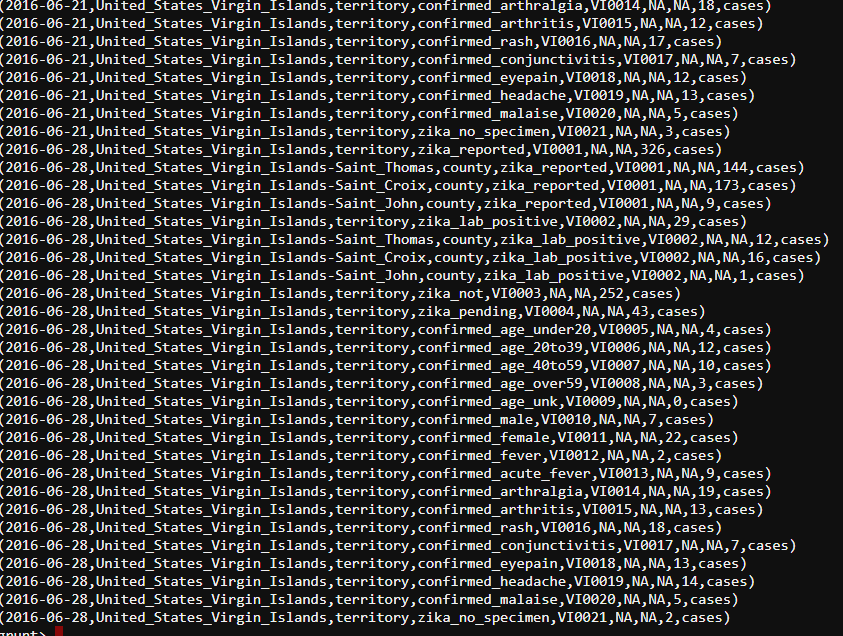
create table zikadata(report\_date string,location string,location\_type string,data\_field string,data\_field\_code string,value int)row format delimited fields terminated by ',';

**Pig Queries**

/\*Loading the data in Pig and cleaning the data in Pig and storing it in Hive using HCatstorer\*/

zika\_data = LOAD '/user/mrstephyjacob4610/Hive/cdc\_zika.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','NOCHANGE','SKIP\_INPUT\_HEADER') AS (report\_date:chararray, location:chararray, location\_type:chararray, data\_field:chararray, data\_field\_code:chararray, time\_period:chararray, time\_period\_type:chararray, value:int, unit:chararray);

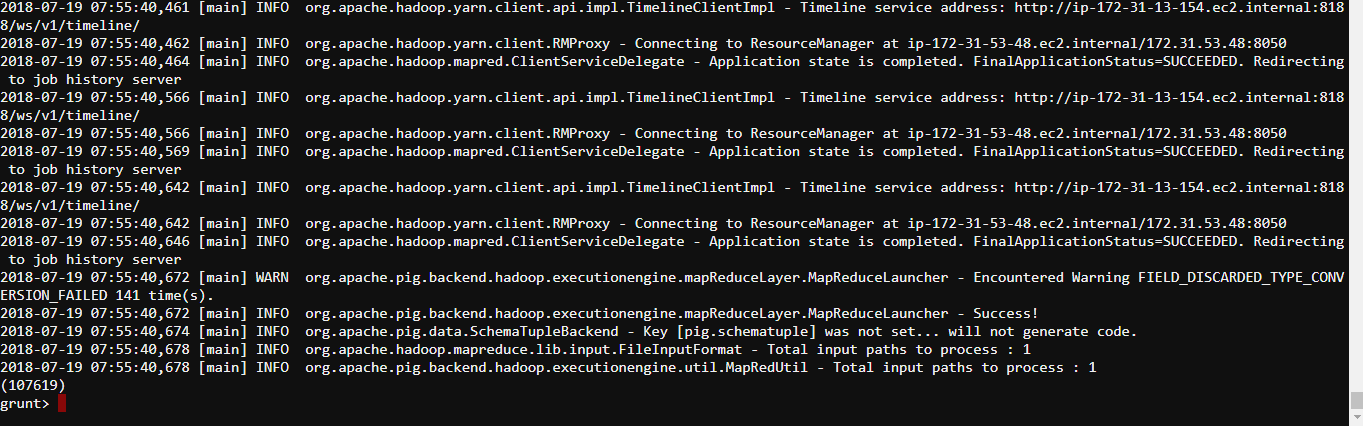
dump zika\_data;



zika\_data\_group = GROUP zika\_data all;

count\_zika\_data = FOREACH zika\_data\_group GENERATE COUNT(zika\_data.report\_date);

dump count\_zika\_data;

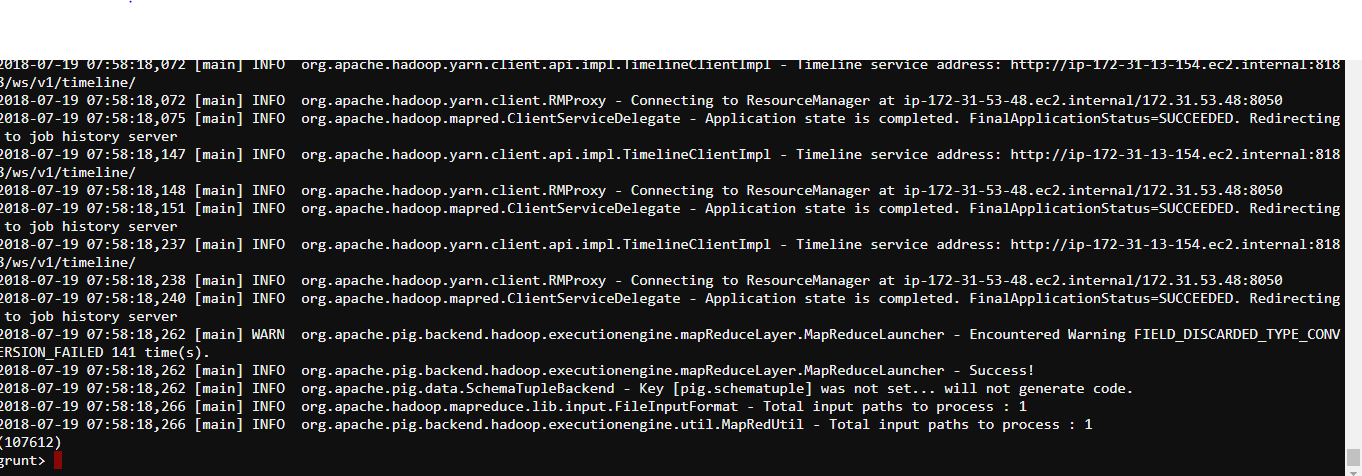


zika\_data\_filter = FILTER zika\_data BY (report\_date != 'NA');

zika\_data\_filter\_group = GROUP zika\_data\_filter all;

count\_zika\_data\_filter\_group = FOREACH zika\_data\_filter\_group GENERATE COUNT(zika\_data\_filter.report\_date);

dump count\_zika\_data\_filter\_group;

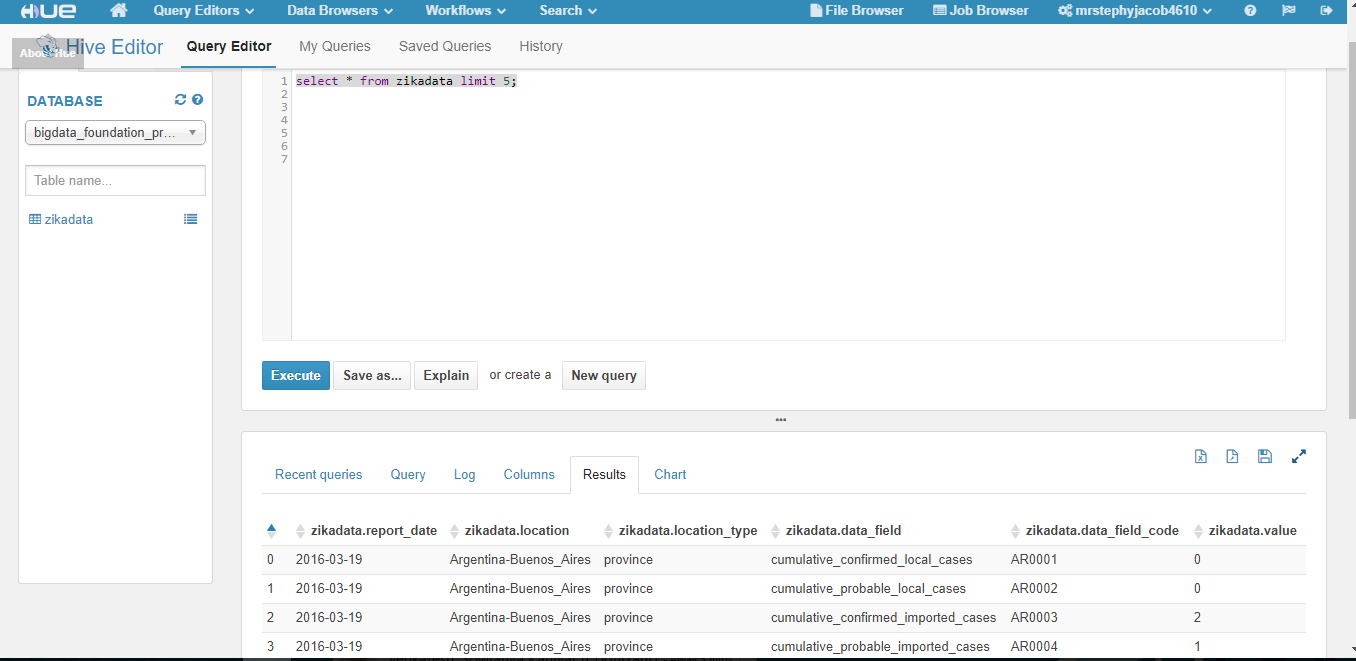


zika\_data\_cleaned = FOREACH zika\_data\_filter GENERATE report\_date,location,location\_type,data\_field,data\_field\_code,value;

STORE zika\_data\_cleaned INTO 'bigdata\_foundation\_project\_stephy.zikadata' USING org.apache.hive.hcatalog.pig.HCatStorer();

**Hive Queries**

select \* from zikadata limit 5;



/\*Count of total number of Zika confirmed across all countries in the dataset\*/

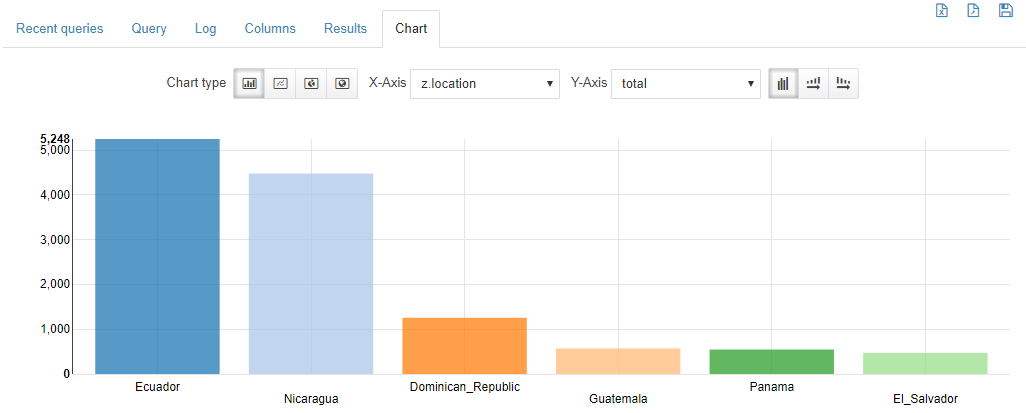
select z.location,sum(z.value) as total from zikadata z where z.location\_type='country'

and z.data\_field like ('%confirmed%')

and z.value is not null

group by z.location order by total desc;





/\*Count of total number of Zika suspected across all countries in the dataset\*/

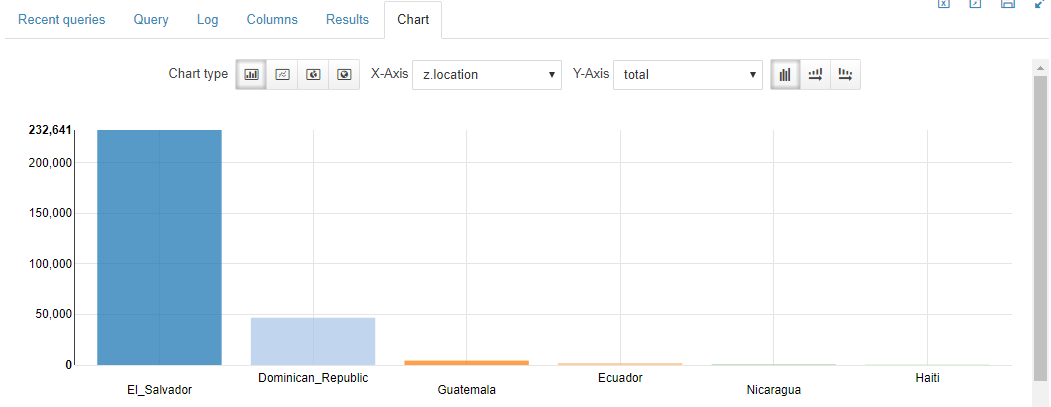
select z.location,sum(z.value) as total from zikadata z where z.location\_type='country'

and z.data\_field like ('%suspect%')

and z.value is not null

group by z.location order by total desc;





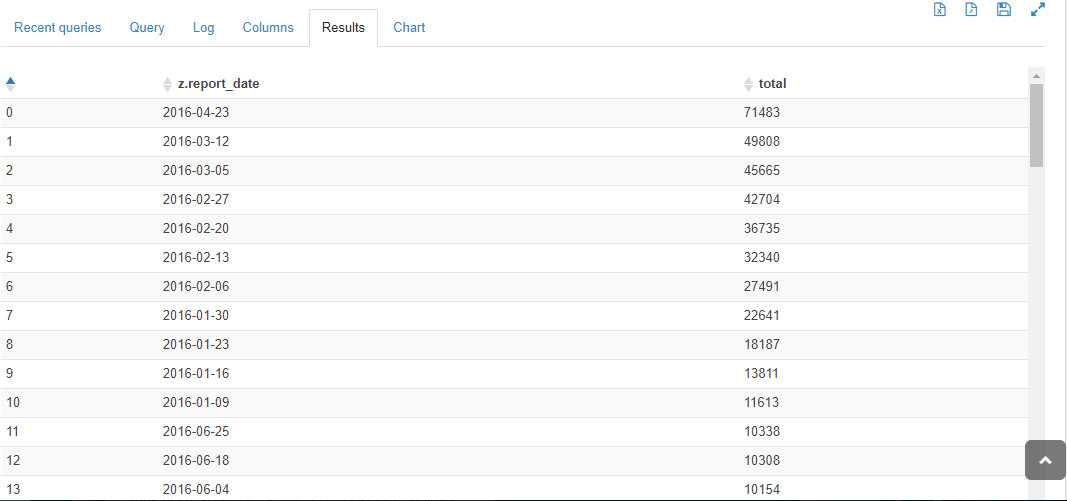
/\* Reported date wise analysis for Discarded and Confirmed cases\*/

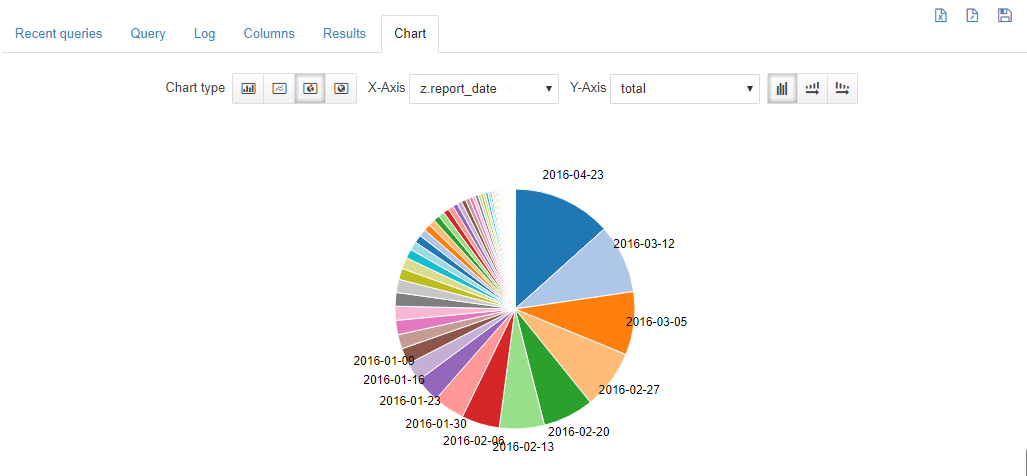
select z.report\_date,sum(z.value) as total from zikadata z

where (z.data\_field like ('%discard%') or z.data\_field like ('%confirm%'))

and z.value is not null

group by z.report\_date order by total desc;





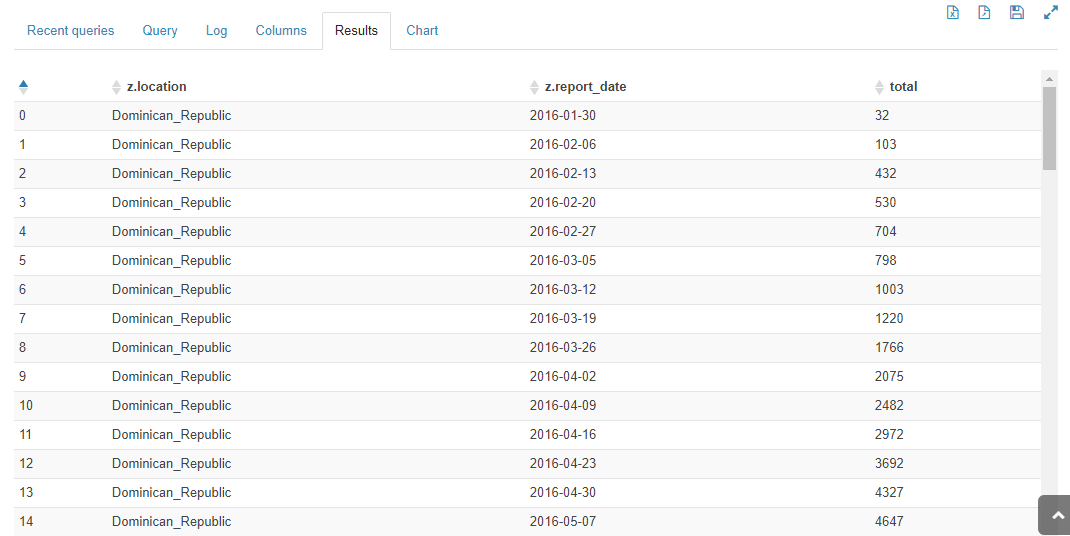
/\* Individual country’s Reported date wise analysis for Suspected and Confirmed cases\*/

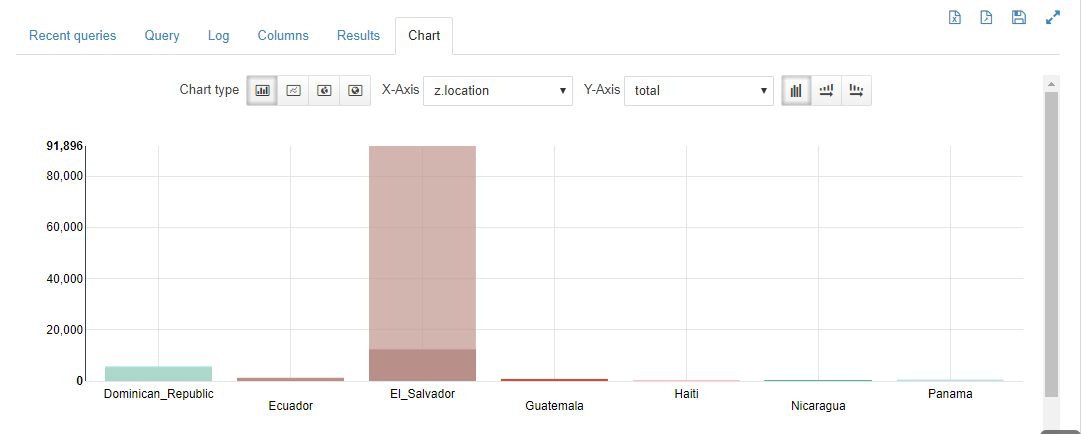
select z.location,z.report\_date,sum(z.value) as total from zikadata z

where (z.data\_field like ('%suspect%') or z.data\_field like ('%confirm%'))

and z.location\_type='country' and z.value>0

group by z.location,z.report\_date order by z.location;





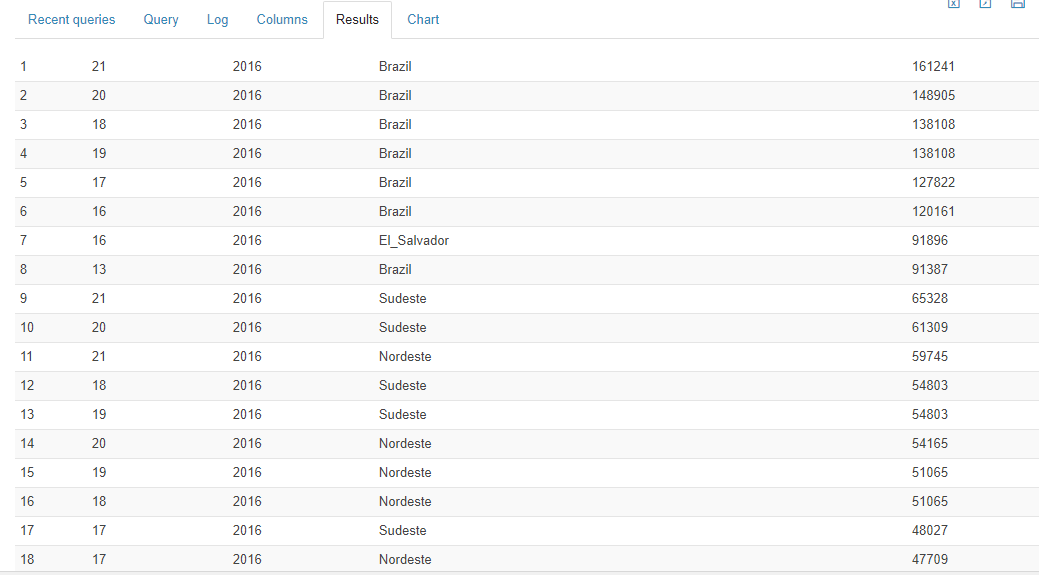
/\* report to find out the time (week) which has most zika effected cases of different countries in a given year \*/

SET hive.groupby.orderby.position.alias=true;

select weekofyear(cast(report\_date as date)) as week,year(cast(report\_date as date)) year1,location,sum(value) as total

from zikadata

group by 1,2,3 order by total desc;



/\* the second top most month which has zika effected cases

select month(cast(report\_date as date)) as month,year(cast(report\_date as date)) year1,location,sum(value) as total

from zikadata

group by 1,2,3 order by total desc limit 2;

