CIS5200 Term Project Tutorial

San Francisoco Fire Calls & Covid Analysis with Hadoop

Authors: Walter R Giron, Leslile Velasquez, Xavier Colin, Eric Tinoco

Instructor: Jongwook Woo

Date: 11/14/2023

Lab Tutorial By

Walter R Giron (wgiron2@calstatela.edu)

Xavier Colin(xcolin@calstatela.edu)

Erick Tinoco (etinoco4@calstatela.edu)

Leslie Velasquez (ivelasq@calstatela.edu)

Fire Department Calls/Covid Data Analysis

Objectives

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

- Download dataset from https://datasf.org/opendata/
- Upload Dataset to Google Drive.
- Make data shareable via URL.
- Transfer dataset to Oracle Linux server.
- Upload Dataset to Hadoop filesystem.
- Establish a database and create tables using Beeline
- Utilize MapReduce to process and cleanse the data, then download the cleaned Dataset onto a Linux system.
- Retrieve the cleaned Dataset and download it to your local PC
- Visualize the Data in Tableau Desktop/Power BI

Platform Spec

- Oracle Linux (RedHat) Big Data server
- CPU Speed: 1995.312 MHz
- # of CPU cores: 3
- # of nodes: 5 Nodes, 2 Main nodes, and three working nodes
- Total Memory Size: 58GB

Step 1: Downloading Dataset from https://data.sfgov.org

In this step of our process, we'll obtain valuable data by downloading information from data.sfgov, a resource that provides insights into fire department calls for service. Accessing this dataset is essential for gathering the information needed for our analysis. To initiate the download and acquire the data, follow the link (https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3).

Visit the page with the link.

Open the Web browser and navigate to https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3 .

You may click on the provided link to reach the Dataset.

DownLoad the dataset.

Once on the webpage hosting the database, navigate to the Export and select CSV.

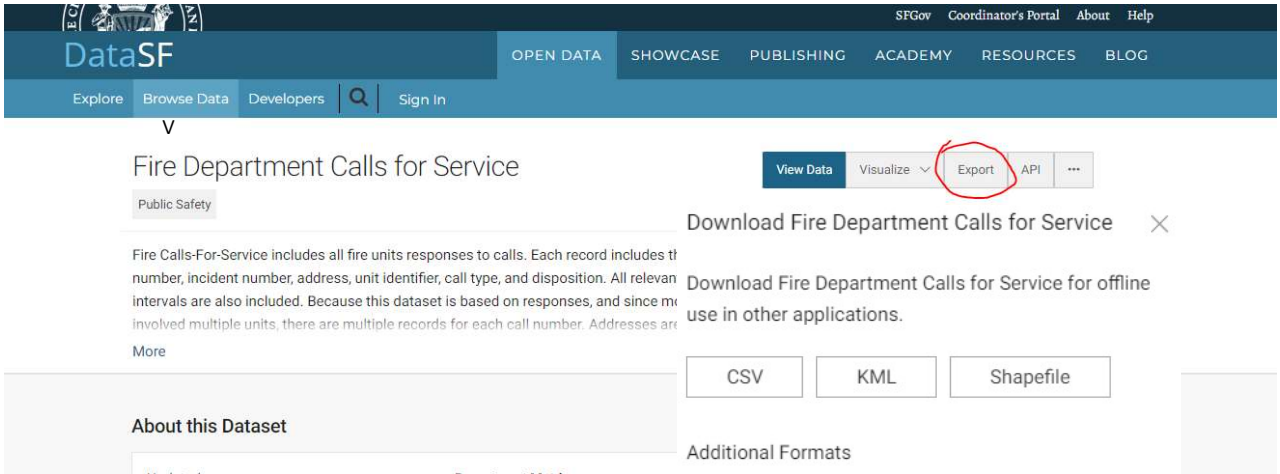
Select a location on your computer to save the dataset file and wait for the download to complete.

Rename Dataset

Locate the Dataset on your computer

Right-click on the file and select rename from the context menu

Enter a new name. name it something descriptive.

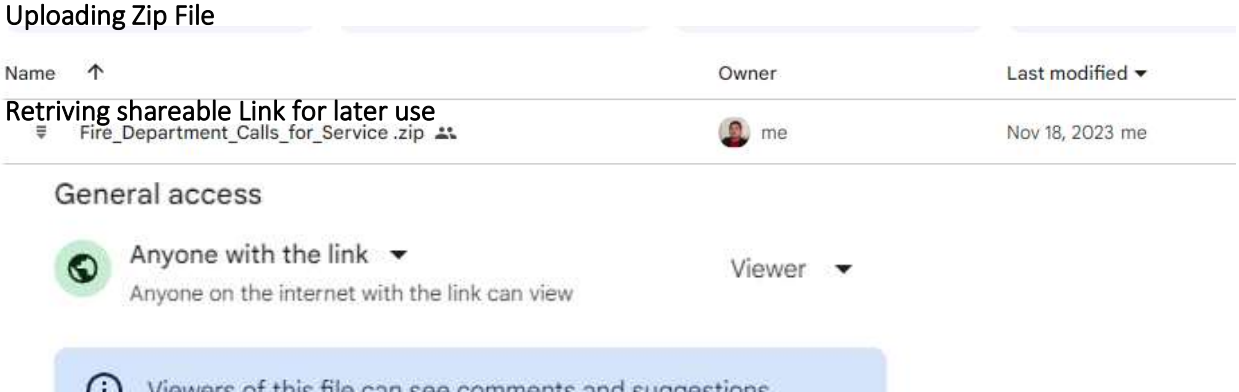


Step 2: Uploading Dataset to Google Drive

In this step, we'll seamlessly transfer the recently acquired Fire_Department_Calls_for_Service.zip folder to Google Drive, ensuring a secure and accessible storage solution. Following this, we'll generate a Shareable URL, providing convenient access for future use and collaboration.

Ensure you have a Gmail account for **Google Drive** access. Upload the 2 to 3GB.zip file to your Google Drive, taking into consideration the potential impact of your upload speed on the duration. After the upload, copy and retrieve the link. This URL is crucial for subsequent steps, so be sure to save it for future use.

Uploading Zip File



Retriving shareable Link for later use

Step 3: Extract File to Oracle Linux Server

Open your GitBash Terminal.

Access the Oracle Linux Server using the following command:

ssh yourUsername@***.***.**.***

Enter Server Password

Beeline to enter into bash

Paste your newly edited wget command from step into the terminal.

wget "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=1inFQ6rwWOyJ54KjGDWxEfNe9cyTnqcNF' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1inFQ6rwWOyJ54KjGDWxEfNe9cyTnqcNF" -O Fire_Department_Calls_for_Service .zip && rm -rf /tmp/cookies.txt

Wait for the download to complete.

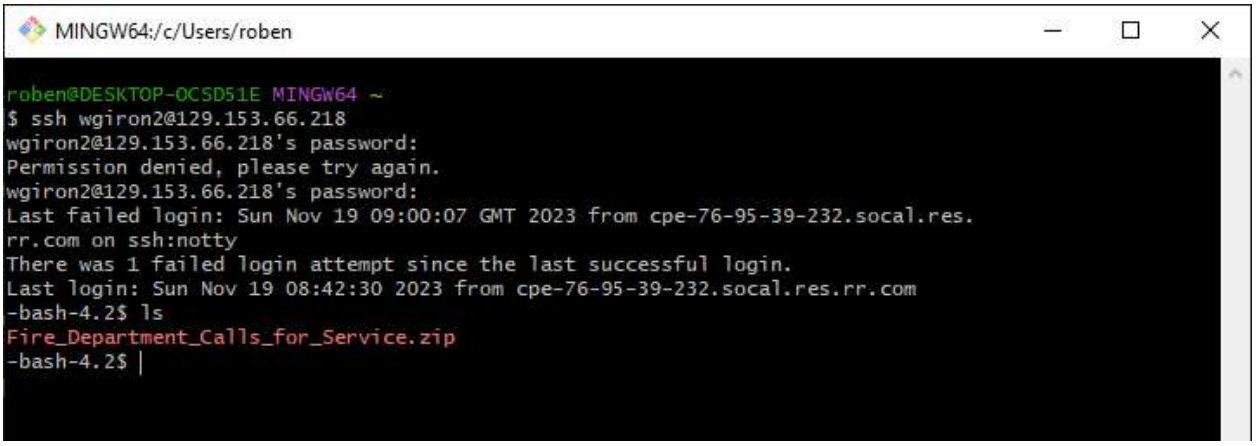
ls to verifie if Zip file was downloaded

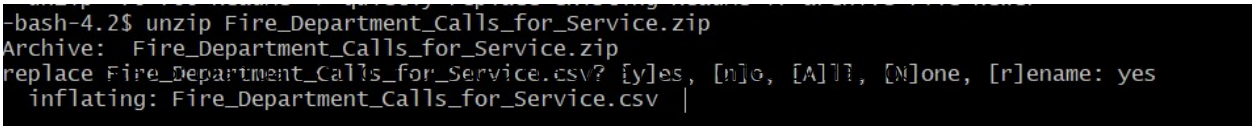
Step 4: Unziping Zip file to Linux Server

In the upcoming steps, our first action involves unpacking the file we downloaded earlier. Subsequently, we'll go beyond a mere extraction process and delve into a thorough verification to ensure the successful completion of this operation. This additional step adds an extra layer of assurance that our file is not only unzipped but also validated for its integrity and completeness. extraction, ensuring that the data is now accessible on the Linux server.

- ls to make sure the data has been succuesfully downloaded

- unzip Fire_Department_Calls_for_Service.zip





In this stage, we'll transfer the **Fire_Department_Calls_for_Service.csv** file to HDFS. Initially, we'll create a directory within our HDFS and then proceed to upload the file into that specific folder.

- hdfs dfs -mkdir /user/yourUserName/Fire_Department_Calls_for_Service
- hdfs dfs -mkdir /user/ yourUserName/tmp
- hdfs dfs -ls
- hdfs dfs -put Fire_Department_Calls_for_Service.csv /user/ yourUserName/Fire_Department_Calls_for_Service
- hdfs dfs -ls Fire_Department_Calls_for_Service.csv/

```
-bash-4.2$ hdfs dfs -mkdir /user/wgiron2/tmp
-bash-4.2$ hdfs dfs -ls
Found 3 items
drwx------   - wgiron2 hdfs          0 2023-11-19 06:55 .Trash
drwxr-xr-x   - wgiron2 hdfs          0 2023-11-19 07:39 Fire_Department_Calls_for_Service
drwxr-xr-x   - wgiron2 hdfs          0 2023-11-19 07:39 tmp
-bash-4.2$ hdfs dfs -put Fire_Department_Calls_for_Service.csv /user/wgiron2/Fire_Department_Calls_for_Service
-bash-4.2$ hdfs dfs -ls Fire_Department_Calls_for_Service/
Found 1 items
-rw-r--r--   3 wgiron2 hdfs 2403931416 2023-11-19 07:39 Fire_Department_Calls_for_Service/Fire_Department_Calls_for_Service.csv
-bash-4.2$ |
```

Step 5: Establish Database and Define Tables through Beeline

In this phase, we kickstart the process of crafting tables using Beeline. The primary aim is to strategically construct these tables, paving the way for a seamless visualization of the expansive dataset we are managing.

First Initiate **beeline**

In this instance, we have decided to create a new database and it will be called Fire_Department_Calls_for_Service

create database Fire_Department_Calls_for_Service;

In order to use the the table

Use Fire_Department_Calls_for_Service;

Create External Table in order to visuallize the tables.

- DROP TABLE IF EXISTS Fire_Department_Calls_for_Service;

  CREATE EXTERNAL TABLE Fire_Department_Calls_for_Service (

Call_Number STRING,

Unit_ID STRING,

Incident_Number STRING,

Call_Type STRING,

Call_Date STRING,

Watch_Date STRING,

Received_DtTm STRING,

Entry_DtTm STRING,

Dispatch_DtTm STRING,

Response_DtTm STRING,

On_Scene_DtTm STRING,

Transport_DtTm STRING,

Hospital_DtTm STRING,

Call_Final_Disposition STRING,

Available_DtTm STRING,

Address STRING,

City STRING,

Zipcode_of_Incident STRING,

Battalion STRING,

Station_Area STRING,

Box STRING,

Original_Priority STRING,

Priority STRING,

Final_Priority STRING,

```
ALS_Unit BOOLEAN,

Call_Type_Group STRING,

Number_of_Alarms INT,

Unit_Type STRING,

Unit_sequence_in_call_dispatch INT,

Fire_Prevention_District STRING,

Supervisor_District STRING,

Neighborhoods_Analysis_Boundaries STRING,

RowID STRING,

case_location STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ","

LOCATION "/user/username/Fire_Department_Calls_for_Service"

TBLPROPERTIES ('skip.header.line.count' = '1');
```

Executing the provided command reassured us that the data was successfully implemented into our previously created database.

SELECT * FROM Fire_Department_Calls_for_Service LIMIT 3;

```
MINGW64:/c/Users/roben
bles
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231119090616_09dc7d4c-f287-41d4-b6ea-78aeb9f06556
); Time taken: 0.208 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------+
|              tab_name             |
+-----------------------------------+
| fire_department_calls_for_service |
+-----------------------------------+
1 row selected (0.251 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> |
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> ;
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> SELECT * FROM Fire_Department_Calls_for_Service LIMIT 3;
INFO  : Compiling command(queryId=hive_20231119090804_bd407903-648b-4d4a-a7a4-72b575c215d8): SELECT * FROM Fire_Department_Calls_for_Service LIMIT 3
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:fire_department_calls_for_service.call_number, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service.uni
INFO  : Completed compiling command(queryId=hive_20231119090804_bd407903-648b-4d4a-a7a4-72b575c215d8); Time taken: 0.329 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231119090804_bd407903-648b-4d4a-a7a4-72b575c215d8): SELECT * FROM Fire_Department_Calls_for_Service LIMIT 3
INFO  : Completed executing command(queryId=hive_20231119090804_bd407903-648b-4d4a-a7a4-72b575c215d8); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------------------+-----------------------------------------+-----------------------------------------------+-----------------------------------------+---
| fire_department_calls_for_service.call_number | fire_department_calls_for_service.unit_id | fire_department_calls_for_service.incident_number | fire_department_calls_for_service.call_type | fi
+-----------------------------------------------+-----------------------------------------+-----------------------------------------------+-----------------------------------------+---
| 221210313                                     | E36                                     | 22054955                                      | Outside Fire                            | 05
| 220190150                                     | E29                                     | 22008871                                      | Alarms                                  | 01
| 211233271                                     | T07                                     | 21053032                                      | Alarms                                  | 05
+-----------------------------------------------+-----------------------------------------+-----------------------------------------------+-----------------------------------------+---
3 rows selected (0.409 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> |
```

Step 6: Clean data using MapReduce

Our strategy involves leveraging MapReduce to optimize and minimize the data, capitalizing on its efficiency in data reduction. This method aims to streamline the dataset, thereby improving processing efficiency. The overarching objective is to effectively condense information through the utilization of MapReduce.The following Steps will guide the user to implement this technique.

In beeline use the database you have previously made for instance our databases was named Fire_Department_Calls_for_Service.

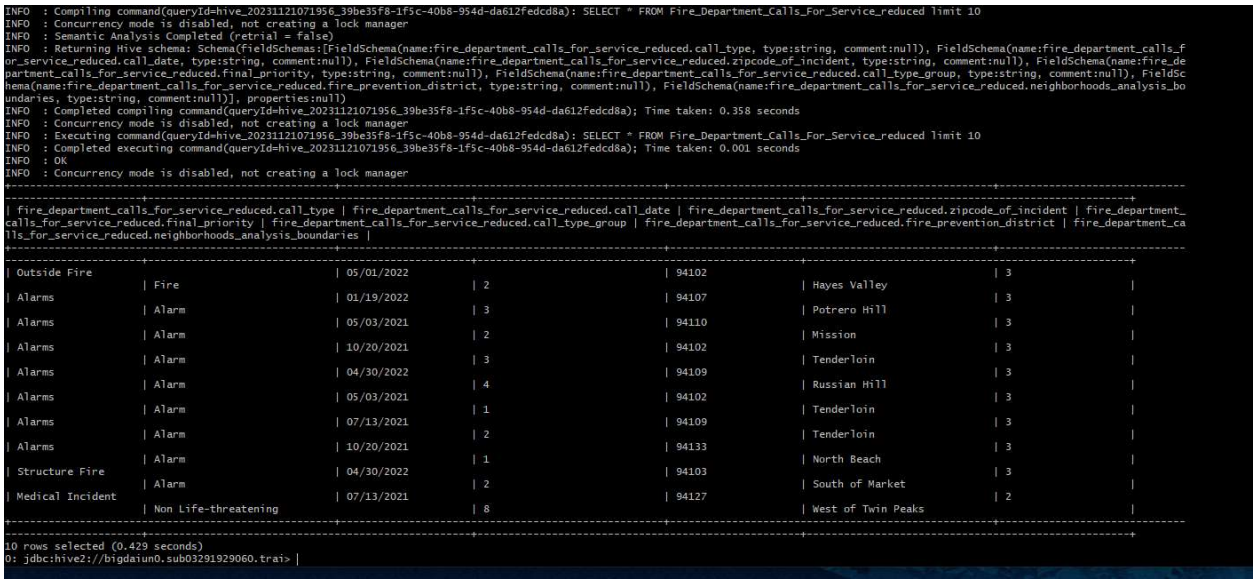
```
use Fire_Department_Calls_for_Service;
```

Secondly, we establish the view to define the specific columns for analysis

```
CREATE VIEW Fire_Department_Calls_For_Service_reduced  AS SELECT Call_Type, Call_Date,
Zipcode_of_Incident, Final_Priority, Call_Type_Group, Fire_Prevention_District,
Neighborhoods_Analysis_Boundaries FROM Fire_Department_Calls_for_Service;
```

Selecting Fire_Department_Calls_For_Service_reduced to view are new data table made

```
SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10;
```



```
INFO  : Compiling command(queryId=hive_20231121071956_39be35f8-1f5c-40b8-954d-da612fedcd8a): SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:fire_department_calls_for_service_reduced.call_type, type:string, comment:null), FieldSchema(name:fire_department_calls_f
or_service_reduced.call_date, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.zipcode_of_incident, type:string, comment:null), FieldSchema(name:fire_de
partment_calls_for_service_reduced.final_priority, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.call_type_group, type:string, comment:null), FieldSc
hema(name:fire_department_calls_for_service_reduced.fire_prevention_district, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.neighborhoods_analysis_bo
undaries, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231121071956_39be35f8-1f5c-40b8-954d-da612fedcd8a); Time taken: 0.358 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231121071956_39be35f8-1f5c-40b8-954d-da612fedcd8a): SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10
INFO  : Completed executing command(queryId=hive_20231121071956_39be35f8-1f5c-40b8-954d-da612fedcd8a); Time taken: 0.001 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+
| fire_department_calls_for_service_reduced.call_type | fire_department_calls_for_service_reduced.call_date | fire_department_calls_for_service_reduced.zipcode_of_incident | fire_department_calls_for_service_reduced.final_priority | fire_department_calls_for_service_reduced.call_type_group | fire_department_calls_for_service_reduced.fire_prevention_district | fire_department_calls_for_service_reduced.neighborhoods_analysis_boundaries |
+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+
| Outside Fire                                        | 05/01/2022                                          | 94102                                               | 3                                                   | Fire                                                | 2                                                   | Hayes Valley                                        |
| Alarms                                              | 01/19/2022                                          | 94107                                               | 3                                                   | Alarm                                               | 3                                                   | Potrero Hill                                        |
| Alarms                                              | 05/03/2021                                          | 94110                                               | 3                                                   | Alarm                                               | 2                                                   | Mission                                             |
| Alarms                                              | 10/20/2021                                          | 94102                                               | 3                                                   | Alarm                                               | 3                                                   | Tenderloin                                          |
| Alarms                                              | 04/30/2022                                          | 94109                                               | 3                                                   | Alarm                                               | 4                                                   | Russian Hill                                        |
| Alarms                                              | 05/03/2021                                          | 94102                                               | 3                                                   | Alarm                                               | 1                                                   | Tenderloin                                          |
| Alarms                                              | 07/13/2021                                          | 94109                                               | 3                                                   | Alarm                                               | 2                                                   | Tenderloin                                          |
| Alarms                                              | 10/20/2021                                          | 94133                                               | 3                                                   | Alarm                                               | 1                                                   | North Beach                                         |
| Structure Fire                                      | 04/30/2022                                          | 94103                                               | 3                                                   | Alarm                                               | 2                                                   | South of Market                                     |
| Medical Incident                                    | 07/13/2021                                          | 94127                                               | 2                                                   | Non Life-threatening                                | 8                                                   | West of Twin Peaks                                  |
+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+
10 rows selected (0.429 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> |
```

Re-runing the command in order to get better visuals by expanding the terminal

```
SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10;
```

```
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10;
INFO  : Compiling command(queryId=hive_20231121072307_9cc2efdd-3c1e-47b8-8ac8-cca2cef3ca2f): SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:fire_department_calls_for_service_reduced.call_type, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.call_date, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.zipcode_of_incident, type:string, co
INFO  : Completed compiling command(queryId=hive_20231121072307_9cc2efdd-3c1e-47b8-8ac8-cca2cef3ca2f); Time taken: 0.324 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231121072307_9cc2efdd-3c1e-47b8-8ac8-cca2cef3ca2f): SELECT * FROM Fire_Department_Calls_For_Service_reduced limit 10
INFO  : Completed executing command(queryId=hive_20231121072307_9cc2efdd-3c1e-47b8-8ac8-cca2cef3ca2f); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+
| fire_department_calls_for_service_reduced.call_type | fire_department_calls_for_service_reduced.call_date | fire_department_calls_for_service_reduced.zipcode_of_incident | fire_department_calls_for_service_reduced.final_priority | fire_department_calls_for_service_reduced.call_type_group | fire_department_calls_for_service_reduced.fire_prev
+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+
| Outside Fire                                        | 05/01/2022                                          | 94102                                               | 3                                                   | Fire                                                | 2                                                   | Hayes Valley
| Alarms                                              | 01/19/2022                                          | 94107                                               | 3                                                   | Alarm                                               | 3                                                   | Potrero Hill
| Alarms                                              | 05/03/2021                                          | 94110                                               | 3                                                   | Alarm                                               | 2                                                   | Mission
| Alarms                                              | 10/20/2021                                          | 94102                                               | 3                                                   | Alarm                                               | 3                                                   | Tenderloin
| Alarms                                              | 04/30/2022                                          | 94109                                               | 3                                                   | Alarm                                               | 4                                                   | Russian Hill
| Alarms                                              | 05/03/2021                                          | 94102                                               | 3                                                   | Alarm                                               | 1                                                   | Tenderloin
| Alarms                                              | 07/13/2021                                          | 94109                                               | 3                                                   | Alarm                                               | 2                                                   | Tenderloin
| Alarms                                              | 10/20/2021                                          | 94133                                               | 3                                                   | Alarm                                               | 1                                                   | North Beach
| Structure Fire                                      | 04/30/2022                                          | 94103                                               | 3                                                   | Alarm                                               | 2                                                   | South of Market
| Medical Incident                                    | 07/13/2021                                          | 94127                                               | 2                                                   | Non Life-threatening                                | 8                                                   | West of Twin Peaks
+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+-----------------------------------------------------+
10 rows selected (0.388 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> |
```

This will create the new table that will be used to create visuals.

INSERT OVERWRITE DIRECTORY '/user/wgiron2/tmp/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM Fire_Department_Calls_For_Service_reduced;

```
INFO  : Compiling command(queryId=hive_20231121074809_432a046f-c0a1-426a-a7da-86b8cff47a7c): INSERT OVERWRITE DIRECTORY '/user/wgiron2/tmp/' ROW FORMAT DELIMITED FI
ELDS TERMINATED BY ',' SELECT * FROM Fire_Department_Calls_For_Service_reduced
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:fire_department_calls_for_service_reduced.call_type, type:string, comment:null), FieldSchema(na
me:fire_department_calls_for_service_reduced.call_date, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.zipcode_of_incident,
type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.final_priority, type:string, comment:null), FieldSchema(name:fire_department_
calls_for_service_reduced.call_type_group, type:string, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.fire_prevention_district, type:str
ing, comment:null), FieldSchema(name:fire_department_calls_for_service_reduced.neighborhoods_analysis_boundaries, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231121074809_432a046f-c0a1-426a-a7da-86b8cff47a7c); Time taken: 0.495 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231121074809_432a046f-c0a1-426a-a7da-86b8cff47a7c): INSERT OVERWRITE DIRECTORY '/user/wgiron2/tmp/' ROW FORMAT DELIMITED FI
ELDS TERMINATED BY ',' SELECT * FROM Fire_Department_Calls_For_Service_reduced
INFO  : Query ID = hive_20231121074809_432a046f-c0a1-426a-a7da-86b8cff47a7c
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20231121074809_432a046f-c0a1-426a-a7da-86b8cff47a7c
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: INSERT OVERWRITE DIRE..._For_Service_reduced (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1698804105104_0371)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 22.88 s
--------------------------------------------------------------------------------
INFO  : Status: DAG finished successfully in 22.85 seconds
INFO  :
INFO  : Query Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  : OPERATION                            DURATION
INFO  : ----------------------------------------------------------------------------------------------
INFO  : Compile Query                          0.49s
INFO  : Prepare Plan                           5.50s
INFO  : Get Query Coordinator (AM)             0.00s
INFO  : Submit Plan                            0.45s
INFO  : Start DAG                              0.52s
INFO  : Run DAG                               22.85s
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  :   VERTICES      DURATION(ms)   CPU_TIME(ms)    GC_TIME(ms)   INPUT_RECORDS   OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
INFO  :     Map 1          20312.00         27,300           201       6,362,049                0
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : org.apache.tez.common.counters.DAGCounter:
INFO  :    NUM_SUCCEEDED_TASKS: 1
INFO  :    TOTAL_LAUNCHED_TASKS: 1
INFO  :    AM_CPU_MILLISECONDS: 2910
INFO  :    WALL_CLOCK_MILLIS: 19915
INFO  :    AM_GC_TIME_MILLIS: 0
INFO  : File System Counters:
INFO  :    HDFS_BYTES_READ: 2403931416
INFO  :    HDFS_BYTES_WRITTEN: 405232758
INFO  :    HDFS_READ_OPS: 4
INFO  :    HDFS_LARGE_READ_OPS: 0
INFO  :    HDFS_WRITE_OPS: 2
INFO  : org.apache.tez.common.counters.TaskCounter:
INFO  :    GC_TIME_MILLIS: 201
INFO  :    CPU_MILLISECONDS: 27300
INFO  :    WALL_CLOCK_MILLISECONDS: 19781
INFO  :    PHYSICAL_MEMORY_BYTES: 983564288
INFO  :    VIRTUAL_MEMORY_BYTES: 5442547712
INFO  :    COMMITTED_HEAP_BYTES: 983564288
INFO  :    INPUT_RECORDS_PROCESSED: 6362049
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 2403931416
INFO  :    OUTPUT_RECORDS: 0
INFO  : HIVE:
INFO  :    CREATED_FILES: 1
INFO  :    DESERIALIZE_ERRORS: 0
INFO  :    RECORDS_IN_Map_1: 6362049
INFO  :    RECORDS_OUT_1: 6362049
INFO  :    RECORDS_OUT_INTERMEDIATE_Map_1: 0
INFO  :    RECORDS_OUT_OPERATOR_FS_4: 6362049
INFO  :    RECORDS_OUT_OPERATOR_MAP_0: 0
INFO  :    RECORDS_OUT_OPERATOR_SEL_3: 6362049
INFO  :    RECORDS_OUT_OPERATOR_TS_0: 6362049
INFO  : TaskCounter_Map_1_INPUT_fire_department_calls_for_service:
INFO  :    INPUT_RECORDS_PROCESSED: 6362049
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 2403931416
INFO  : TaskCounter_Map_1_OUTPUT_out_Map_1:
INFO  :    OUTPUT_RECORDS: 0
INFO  : org.apache.hadoop.hive.ql.exec.tez.HiveInputCounters:
INFO  :    GROUPED_INPUT_SPLITS_Map_1: 1
INFO  :    INPUT_DIRECTORIES_Map_1: 1
INFO  :    INPUT_FILES_Map_1: 1
INFO  :    RAW_INPUT_SPLITS_Map_1: 1
INFO  : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Moving data to directory /user/wgiron2/tmp from hdfs://bigdaimn0.sub03291929060.trainingvcn.oraclevcn.com:8020/user/wgiron2/tmp/.hive-staging_hive_2023-11-2
1_07-48-09_023_2220264222452147685-949/-ext-10000
INFO  : Completed executing command(queryId=hive_20231121074809_432a046f-c0a1-426a-a7da-86b8cff47a7c); Time taken: 29.334 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (29.843 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>
```

Checking if the file was successfully made.

hdfs dfs -ls tmp/

```
-bash-4.2$ hdfs dfs -ls tmp/
Found 1 items
-rw-r--r--   3 wgiron2 hdfs  405232758 2023-11-21 07:48 tmp/000000_0
-bash-4.2$
```

To retrive the new file that was cleanded with mapreduce

hdfs dfs -get /user/wgiron2/tmp/000000_0

Verifing the file in bash

ls

du – h 000000_0

```
-bash-4.2$ hdfs dfs -get /user/wgiron2/tmp/000000_0
-bash-4.2$ ls
000000_0  Fire_Department_Calls_for_Service.zip
-bash-4.2$ du -h 000000_0
387M    000000_0
-bash-4.2$
```

Step 7: Downloading new cleaned data to PC

The subsequent step involves downloading the recently cleaned dataset, facilitating the generation of more comprehensible and visually appealing data representations. This action is pivotal in enhancing the accessibility and clarity of the visualizations we aim to create.

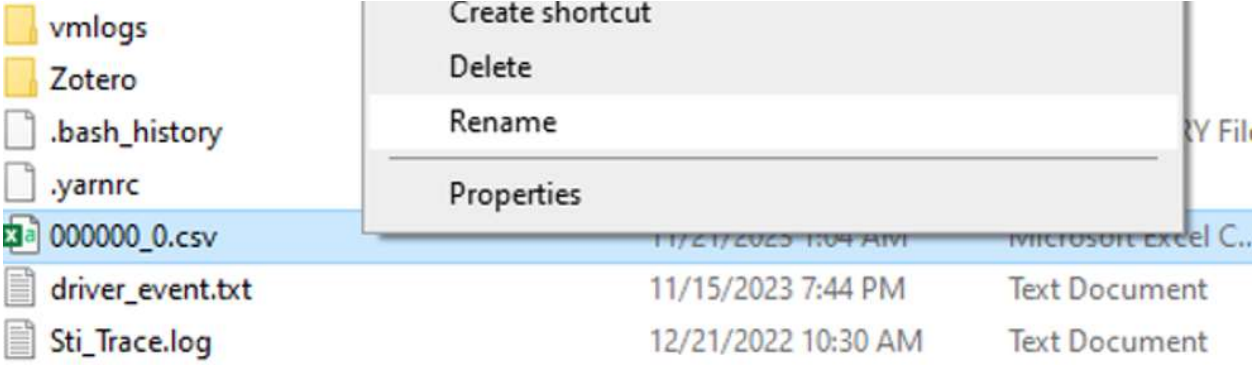
Select git bash terminal and write the following

scp wgiron2@129.153.66.218:/home/wgiron2/000000_0 .

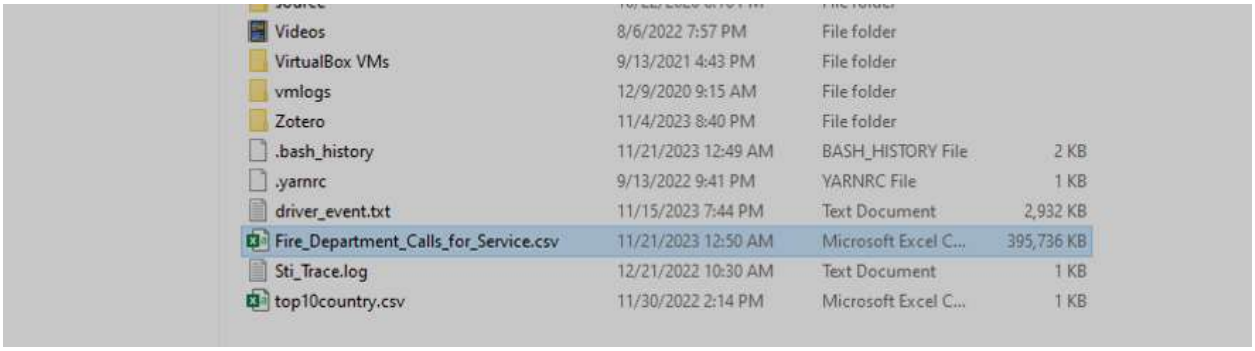
```
MINGW64:/c/Users/roben
roben@DESKTOP-OCSD51E MINGW64 ~
$ scp wgiron2@129.153.66.218:/home/wgiron2/000000_0 .
wgiron2@129.153.66.218's password:
000000_0                                                    100%  386MB  32.1MB/s   00:12

roben@DESKTOP-OCSD51E MINGW64 ~
$
```

Renaming the file to from **00000_0** to **Fire_Department_Calls_For_Service**



Fire_Department_Calls_for_Service.csv



Ending of Fire_Department_calls_For_Service.cs

Step 2.1: Replicating Steps 2-9 Steps

This analysis is centered in San Francisco, exploring the interplay between COVID-19 testing data and fire department responses. Focusing on this specific location enables us to draw meaningful connections between public health measures and emergency services in the urban context. By revisiting steps 2-9 of the process, our goal is to provide a concise yet comprehensive understanding of the relationship between COVID-19 testing activities and fire department calls for service in San Francisco.

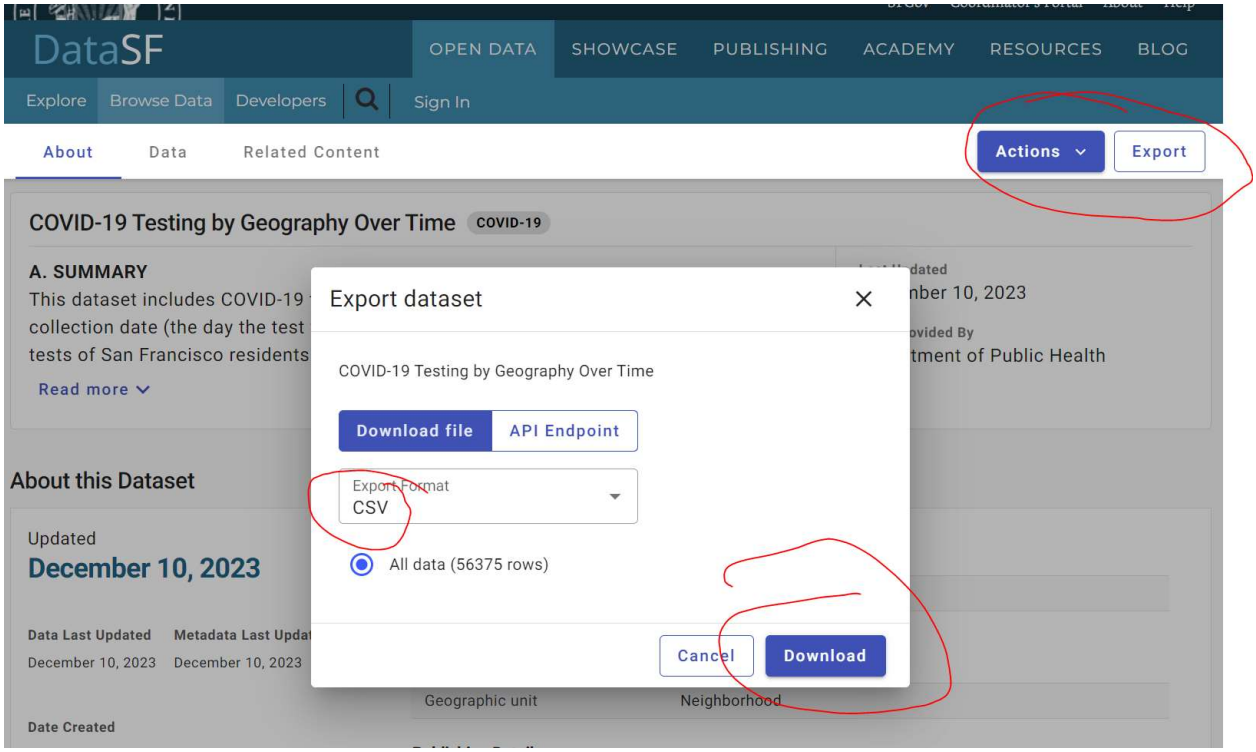
1. Visit the webpage with COVID-19 testing data. Open your web browser and navigate to the relevant data source, such as a local health department website

2. Download the COVID-19 Testing dataset. Navigate to the export option on the webpage and choose CSV as the format.

3. Save the dataset on your computer. Select a destination on your computer, initiate the download, and wait for it to finish.

4. Rename the COVID-19 Testing dataset. Locate the downloaded file, right-click on it, and choose the rename option from the context menu.

5. Enter a new descriptive name for the file to enhance clarity in your analysis, such as "covid-19_Testing."



Step 2.2: Replicating Steps 2-9 Steps

In this step, transfer the COVID-19_Testing.zip folder to Google Drive for secure storage. Use a Gmail account for Google Drive access, and upload the 2 to 3GB.zip file, considering upload speed impact. After uploading, copy and save the Shareable URL for future use. Note that these steps are a reuse of the former process with a different dataset.

Step 2.3: Replicating Steps 2-9 Steps

Open your GitBash Terminal.

Access the Oracle Linux Server using the following command:

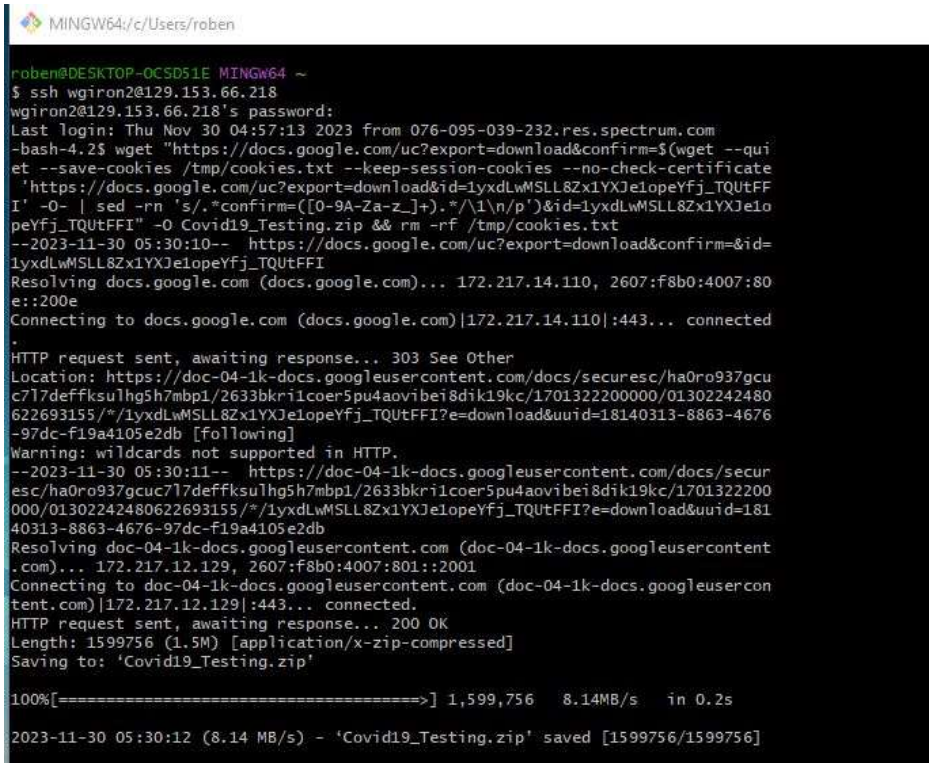
ssh yourUsername@***.***.**.***

Enter Server Password

Beeline to enter into bash

Paste your newly edited wget command from step into the terminal.

[wget "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id=1yxdLwMSLL8Zx1YXJe1opeYfj_TQUtFFI' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1yxdLwMSLL8Zx1YXJe1opeYfj_TQUtFFI" -O Covid19_Testing.zip && rm -rf /tmp/cookies.txt](https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id=1yxdLwMSLL8Zx1YXJe1opeYfj_TQUtFFI' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1yxdLwMSLL8Zx1YXJe1opeYfj_TQUtFFI" -O Covid19_Testing.zip && rm -rf /tmp/cookies.txt)



Step 2.4: Replicating Steps 2-9 Steps

In the next steps, we'll unzip the downloaded file. We'll then verify the extraction to ensure its completeness and integrity, adding an extra layer of assurance. This guarantees the data's accessibility on the Linux server.

- ls to make sure the data has been succuesfully downloaded
- unzip Covid19_Testing.zip

Load dataset to Distributed File System supported by Hadoop.

In this stage, we'll transfer the **Covid19_Testing.zip** file to HDFS. Initially, we'll create a directory within our HDFS and then proceed to upload the file into that specific folder.

- hdfs dfs -mkdir /user/yourUserName/Covid19_Testing
- hdfs dfs -mkdir /user/ yourUserName/tmp
- hdfs dfs -ls
- hdfs dfs -put Covid19_Testing.csv /user/ yourUserName/Covid19_Testing
- hdfs dfs -ls Covid19_Testing.csv/

```
100%[=====================================>] 1,599,756   8.14MB/s   in 0.2s

2023-11-30 05:30:12 (8.14 MB/s) - 'Covid19_Testing.zip' saved [1599756/1599756]

rm: cannot remove '/tmp/cookies.txt': Operation not permitted
-bash-4.2$ ls
Covid19_Testing.zip
-bash-4.2$ unzip Covid19_Testing.zip
Archive:  Covid19_Testing.zip
  inflating: Covid19_Testing.csv
-bash-4.2$
-bash-4.2$ du -h Covid19_Testing.csv
8.4M    Covid19_Testing.csv
-bash-4.2$ hdfs dfs -mkdir /user/wgiron2/Covid19_Testing
-bash-4.2$ hdfs dfs -mkdir /user/wgiron2/tmp
mkdir: `/user/wgiron2/tmp': File exists
-bash-4.2$ hdfs dfs -ls
Found 4 items
drwx------   - wgiron2 hdfs          0 2023-11-30 05:05 .Trash
drwxr-xr-x   - wgiron2 hdfs          0 2023-11-21 07:48 .hiveJars
drwxr-xr-x   - wgiron2 hdfs          0 2023-11-30 05:35 Covid19_Testing
drwxr-xr-x   - wgiron2 hdfs          0 2023-11-30 02:34 tmp


-bash-4.2$ hdfs dfs -put Covid19_Testing.csv /user/wgiron2/Covid19_Testing
-bash-4.2$
-bash-4.2$ hdfs dfs -ls Covid19_Testing/
Found 1 items
-rw-r--r--   3 wgiron2 hdfs    8739096 2023-11-30 05:38 Covid19_Testing/Covid19_
Testing.csv
-bash-4.2$
```

Step 2.5: Replicating Steps 2-9 Steps

In this phase, we initiate the creation of tables using Beeline. Our objective is to strategically define these tables, laying the groundwork for a smooth visualization of the extensive dataset we are handling. Additionally, we are redoing the step to establish and define tables through Beeline to allocate the datatable accordingly.-

1. Begin by launching beeline.

2. Create a new database named "Covid19_Testing":

```
create database covid19_testing;
```

3. To work with the table, switch to the newly created database:

```
Use covid19_testing;
```

4. Create an external table to facilitate table visualization.

```
DROP TABLE IF EXISTS Covid19_Testing;
CREATE EXTERNAL TABLE Covid19_Testing (
   specimen_collection_date STRING,
   area_type STRING,
   id STRING,
   acs_population STRING,
   new_Test STRING,
   new_positive_tests STRING,
   new_negative_tests STRING,
   new_indeterminate_tests STRING,
   cumulative_tests STRING,
   cumulative_positive_tests STRING,
   cumulative_negative_tests STRING,
   cumulative_indeterminate_tests STRING,
   cumulative_testing_rate STRING,
   data_as_of STRING,
   data_loaded_at STRING

)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ","
LOCATION "/user/wgiron2/Covid19_Testing"
TBLPROPERTIES ('skip.header.line.count' = '1');
```

After setting up the "covid19_testing" table, I opted to verify if I was in the correct database. I used the command:

show tables;

```
specimen_collection_date STRING,
area_type STRING,
id STRING,
acs_population STRING,
new_Test STRING,
new_positive_tests STRING,
new_negative_tests STRING,
new_indeterminate_tests STRING,
cumulative_tests STRING,
cumulative_positive_tests STRING,
cumulative_negative_tests STRING,
cumulative_indeterminate_tests STRING,
cumulative_testing_rate STRING,
data_as_of STRING,
data_loaded_at STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ","
LOCATION "/user/wgiron2/Covid19_Testing"
TBLPROPERTIES ('skip.header.line.count' = '1')
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231130064921_9db34d15-beaf-49d6-a771-917f02765ed8); Time taken: 0.159 secon
ds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.197 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> show tables;
INFO  : Compiling command(queryId=hive_20231130064927_77c990e5-e192-48b5-a6f2-db2855d70a31): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], prope
rties:null)
INFO  : Completed compiling command(queryId=hive_20231130064927_77c990e5-e192-48b5-a6f2-db2855d70a31); Time taken: 0.025 secon
ds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231130064927_77c990e5-e192-48b5-a6f2-db2855d70a31): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20231130064927_77c990e5-e192-48b5-a6f2-db2855d70a31); Time taken: 0.209 secon
ds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------------+
|     tab_name     |
+------------------+
| covid19_testing  |
+------------------+
1 row selected (0.271 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> |
```

Running the given command confirmed the successful integration of data into our pre-existing database.

SELECT * FROM covid19_testing LIMIT 3;

```
0: jdbc:hive2://bigdaiun0.sub03291929060.trai> SELECT * FROM covid19_testing LIMIT 3;
INFO  : Compiling command(queryId=hive_20231211045907_d1a63211-11ea-42b0-96d1-3a28991b7d4b): SELECT * FROM covid19_testing LIMIT 3
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:covid19_testing.specimen_collection_date, type:string, comment:nul
19_testing.cumulative_tests, type:int, comment:null), FieldSchema(name:covid19_testing.cumulative_positive_tests, type:int, comment:nul
INFO  : Completed compiling command(queryId=hive_20231211045907_d1a63211-11ea-42b0-96d1-3a28991b7d4b); Time taken: 0.343 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231211045907_d1a63211-11ea-42b0-96d1-3a28991b7d4b): SELECT * FROM covid19_testing LIMIT 3
INFO  : Completed executing command(queryId=hive_20231211045907_d1a63211-11ea-42b0-96d1-3a28991b7d4b); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------------------+---------------------------+-----------------------+-------------------------------+--
| covid19_testing.specimen_collection_date | covid19_testing.area_type |   covid19_testing.id  | covid19_testing.acs_population | c
+-----------------------------------------+---------------------------+-----------------------+-------------------------------+--
| 03/01/2020 12:00:00 AM                  | Analysis Neighborhood     | Bayview Hunters Point | 38480                         | N
| 03/01/2020 12:00:00 AM                  | Analysis Neighborhood     | Bernal Heights        | 26149                         | N
| 03/01/2020 12:00:00 AM                  | Analysis Neighborhood     | Castro/Upper Market   | 23138                         | N
+-----------------------------------------+---------------------------+-----------------------+-------------------------------+--
3 rows selected (0.407 seconds)
```

Step 2.6:Replicating Steps 2-9 Steps Cleandata using MapReduce

In Beeline, switch to the database you've previously created, such as "covid19_testing"

use covid19_testing;

Next, establish a view to define specific columns for analysis:

CREATE VIEW Covid19_Testing_reduced AS SELECT id, specimen_collection_date, new_positive_tests, area_type, acs_population FROM Covid19_Testing;

To view the new data table , select Covid19_Testing.

SELECT * FROM COALESCE Covid19_Testing_reduced limit 10;

```
jdbc:hive2://bigdaiun0.sub03291929060.trai> SELECT * FROM Covid19_Testing_reduced limit 10;
FO  : Compiling command(queryId=hive_20231211053229_872dc796-96e3-421c-96cb-ff99687db1d6): SELECT * FROM Covid19_Testing_reduced limit 10
FO  : Concurrency mode is disabled, not creating a lock manager
FO  : Semantic Analysis Completed (retrial = false)
FO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:covid19_testing_reduced.id, type:string, comment:null), FieldSchema(name:covid19_testing_reduced.specimen_collection_date, type:string, comment:null), Fi
FO  : Completed compiling command(queryId=hive_20231211053229_872dc796-96e3-421c-96cb-ff99687db1d6); Time taken: 0.37 seconds
FO  : Concurrency mode is disabled, not creating a lock manager
FO  : Executing command(queryId=hive_20231211053229_872dc796-96e3-421c-96cb-ff99687db1d6): SELECT * FROM Covid19_Testing_reduced limit 10
FO  : Completed executing command(queryId=hive_20231211053229_872dc796-96e3-421c-96cb-ff99687db1d6); Time taken: 0.001 seconds
FO  : OK
FO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------------+-----------------------------------------+-----------------------------------+-----------------------------+-----------------------------------+
| covid19_testing_reduced.id  | covid19_testing_reduced.specimen_collection_date | covid19_testing_reduced.new_positive_tests | covid19_testing_reduced.area_type | covid19_testing_reduced.acs_population |
+-----------------------------+-----------------------------------------+-----------------------------------+-----------------------------+-----------------------------------+
| Bayview Hunters Point       | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 38480                             |
| Bernal Heights              | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 26149                             |
| Castro/Upper Market         | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 23138                             |
| Chinatown                   | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 14310                             |
| Excelsior                   | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 40980                             |
| Financial District/South Beach | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 22963                             |
| Glen Park                   | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 8654                              |
| Golden Gate Park            | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 32                                |
| Haight Ashbury              | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 19181                             |
| Hayes Valley                | 03/01/2020 12:00:00 AM                  | NULL                              | Analysis Neighborhood       | 19818                             |
+-----------------------------+-----------------------------------------+-----------------------------------+-----------------------------+-----------------------------------+
 rows selected (0.426 seconds)
 jdbc:hive2://bigdaiun0.sub03291929060.trai>
```

```
INFO  : Compiling command(queryId=hive_20231202085058_34086b2d-7c5b-46e6-9f73-023a1d74aae2): INSERT OVERWRITE DIRECTORY '/user/wgiron2/tmp/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM Covid19_Testing_reduced
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:covid19_testing_reduced.id, type:string, comment:null), FieldSchema(name:covid19_testing_reduced.specimen_collection_date, type:string, comment:null), FieldSchema(name:covid19_
me:covid19_testing_reduced.acs_population, type:int, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20231202085058_34086b2d-7c5b-46e6-9f73-023a1d74aae2); Time taken: 0.479 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20231202085058_34086b2d-7c5b-46e6-9f73-023a1d74aae2): INSERT OVERWRITE DIRECTORY '/user/wgiron2/tmp/' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM Covid19_Testing_reduced
INFO  : Query ID = hive_20231202085058_34086b2d-7c5b-46e6-9f73-023a1d74aae2
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20231202085058_34086b2d-7c5b-46e6-9f73-023a1d74aae2
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: INSERT OVERWRITE DIRE...id19_Testing_reduced (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1698804105104_0583)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.92 s
--------------------------------------------------------------------------------
INFO  : Status: DAG finished successfully in 4.89 seconds
INFO  :
INFO  : Query Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  : OPERATION                            DURATION
INFO  : ----------------------------------------------------------------------------------------------
INFO  : Compile Query                          0.48s
INFO  : Prepare Plan                           4.42s
INFO  : Get Query Coordinator (AM)             0.00s
INFO  : Submit Plan                            0.46s
INFO  : Start DAG                              0.53s
INFO  : Run DAG                                4.89s
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : Task Execution Summary
INFO  : ----------------------------------------------------------------------------------------------
INFO  :   VERTICES      DURATION(ms)   CPU_TIME(ms)    GC_TIME(ms)   INPUT_RECORDS   OUTPUT_RECORDS
INFO  : ----------------------------------------------------------------------------------------------
INFO  :     Map 1           [illegible]      5,[illegible]      [illegible]        55,924               0
INFO  : ----------------------------------------------------------------------------------------------
INFO  :
INFO  : org.apache.tez.common.counters.DAGCounter:
INFO  :    NUM_SUCCEEDED_TASKS: 1
INFO  :    TOTAL_LAUNCHED_TASKS: 1
INFO  :    AM_CPU_MILLISECONDS: 2020
INFO  :    WALL_CLOCK_MILLIS: 2071
INFO  :    AM_GC_TIME_MILLIS: 0
INFO  : File System Counters:
INFO  :    HDFS_BYTES_READ: 8739096
INFO  :    HDFS_BYTES_WRITT
INFO  :    HDFS_READ_OPS: 4
INFO  :    HDFS_LARGE_READ_
INFO  :    HDFS_WRITE_OPS:
INFO  : org.apache.tez.comm
INFO  :    GC_TIME_MILLIS:
INFO  :    CPU_MILLISECONDS
INFO  :    WALL_CLOCK_MILLI
INFO  :    PHYSICAL_MEMORY_
INFO  :    VIRTUAL_MEMORY_B
INFO  :    COMMITTED_HEAP_BYTES: 983564288
INFO  :    INPUT_RECORDS_PROCESSED: 55924
INFO  :    INPUT_SPLIT_LENGTH_BYTES: 8739096
INFO  :    OUTPUT_RECORDS: 0
INFO  : HIVE:
```

```
-bash-4.2$ hdfs dfs -ls tmp/
Found 1 items
-rw-r--r--   3 wgiron2 hdfs     3745309 2023-12-02 08:51 tmp/000000_0
-bash-4.2$
```

To retrive the new file that was cleanded with mapreduce

hdfs dfs -get /user/wgiron2/tmp/000000_0

Verifing the file in bash

```
-bash-4.2$ hdfs dfs -ls tmp/
Found 1 items
-rw-r--r--   3 wgiron2 hdfs    3745309 2023-12-02 08:51 tmp/000000_0
-bash-4.2$ hdfs dfs -get /user/wgiron2/tmp/000000_0
-bash-4.2$ ls
000000_0  Covid19_Testing.csv  Covid19_Testing.zip
-bash-4.2$ du -h 000000_0
3.6M    000000_0
-bash-4.2$
```

Step 2.7 Downloading new cleaned data to PC

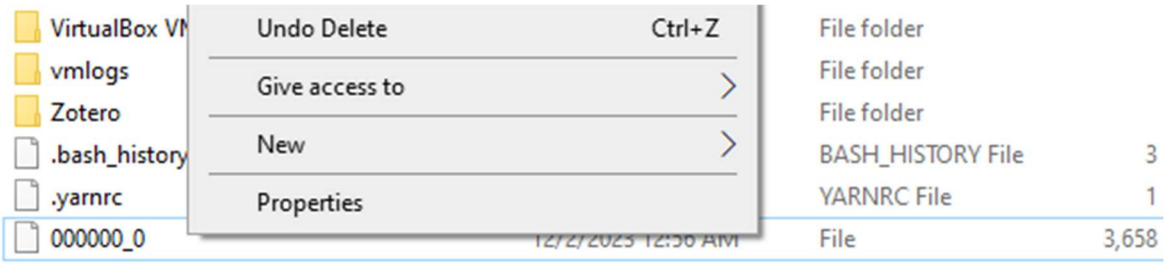
Next, we proceed to download the recently refined dataset, which will enable the creation of clearer and more visually engaging data representations. This step is crucial for improving the accessibility and overall clarity of the visualizations we intend to develop.

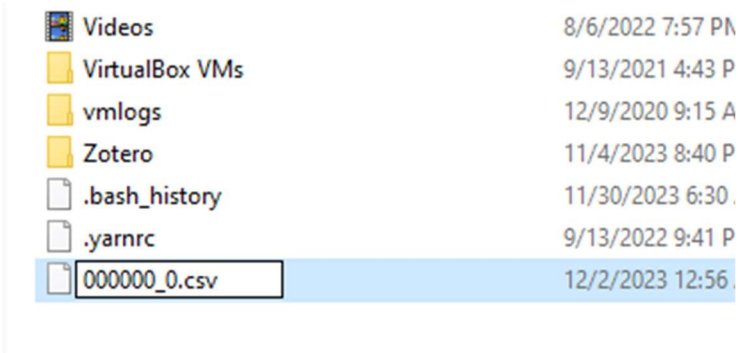
scp wgiron2@129.153.66.218:/home/wgiron2/000000_0 .

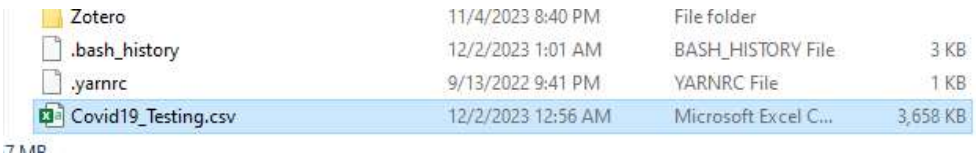
```
scp wgiron2@129.153.66.218:/home/wgiron2/000000_0 .
giron2@129.153.66.218's password:
00000_0                                       100% 3658KB    5.6MB/s   00:00

oben@DESKTOP-OC5D51E MINGW64 ~
```

Renaming file 00000_0 to Covid19_Testing



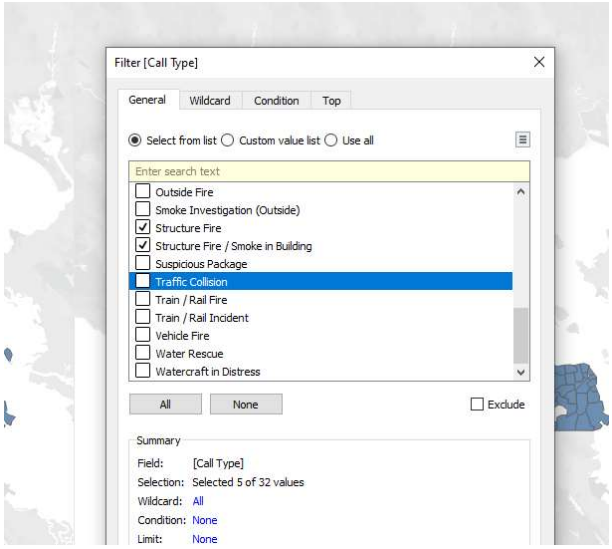




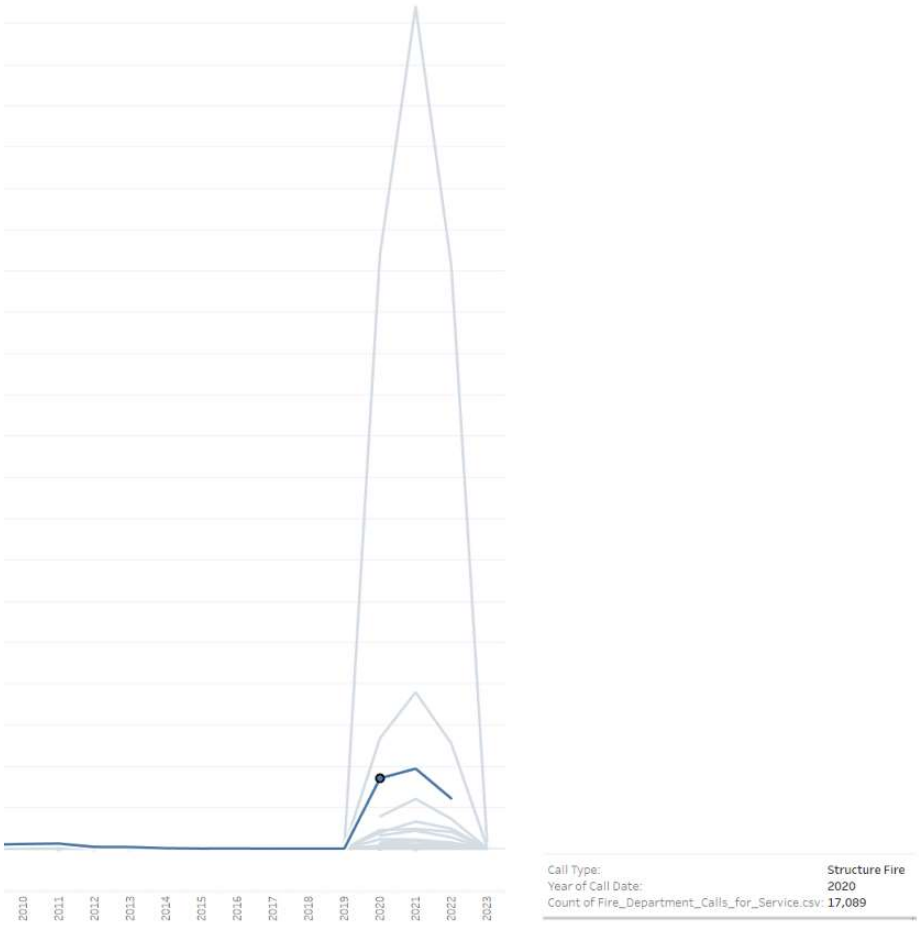
Step 11: Using our new data to display Visuals

To create compelling visuals, we'll utilize Tableau for its powerful visualization capabilities. Additionally, we'll leverage Excel 3D to enhance our data representation, ensuring a comprehensive and insightful view of key aspects.

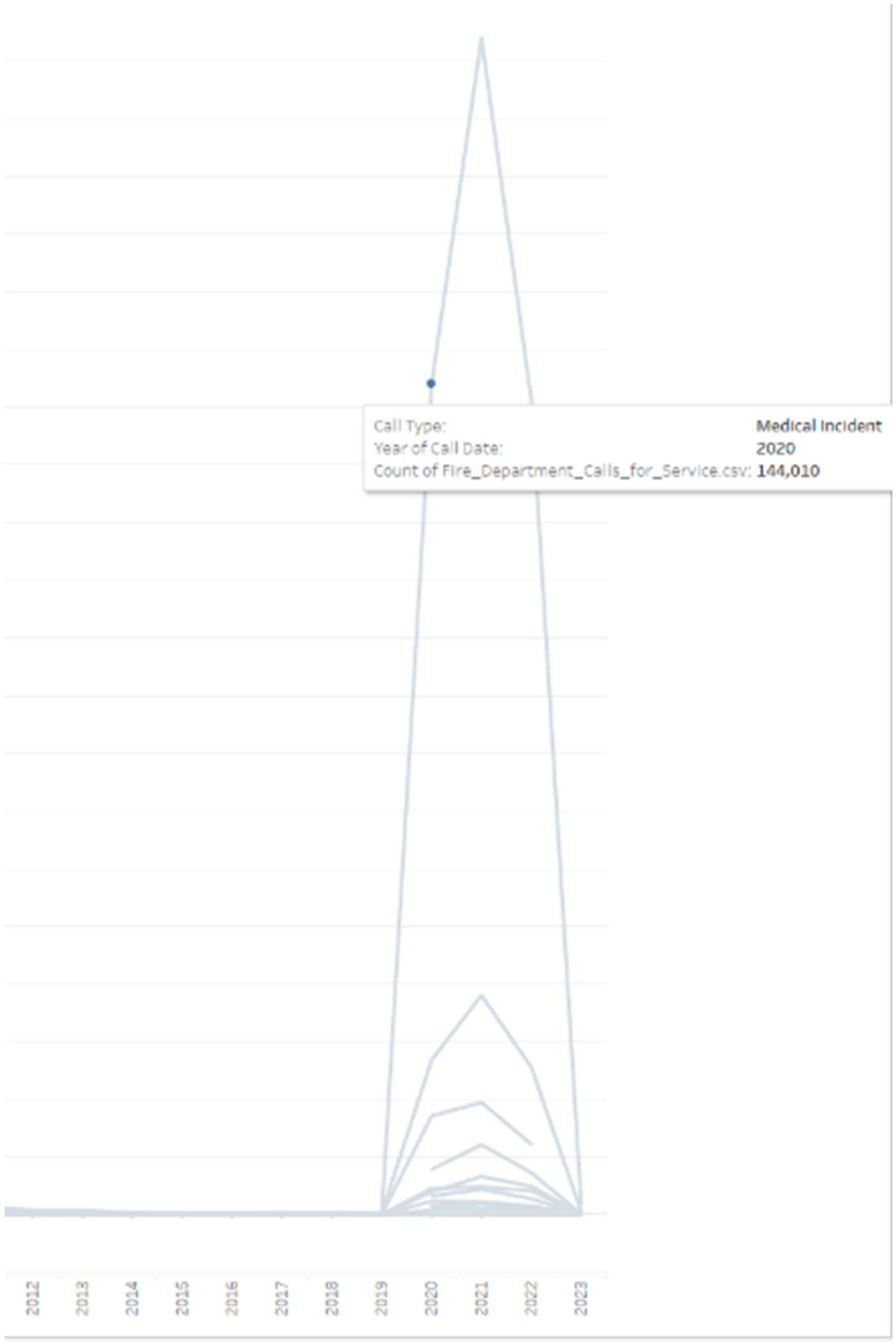
In our refined dataset, we've opted to filter out everything except fire calls and potentially life-threatening incidents. The following image dipicts that action.



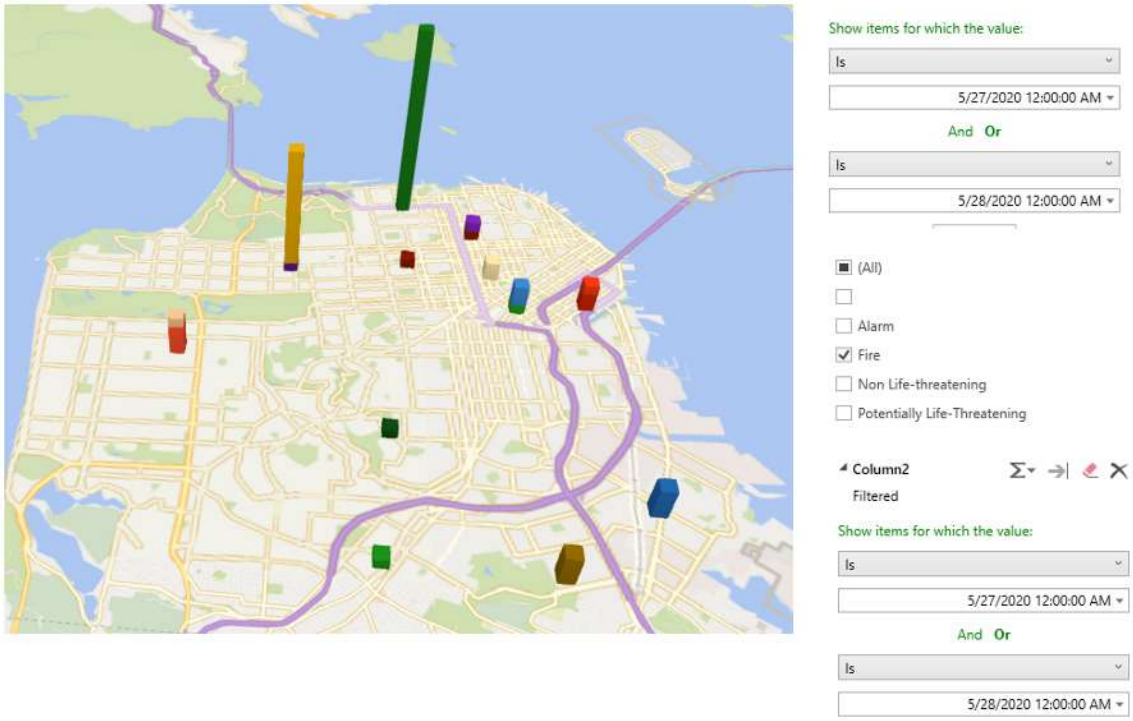
In the following visual we conducted a line chart in order to visualise the amounts of fires that started in 2020 due to an incident that occuured in that time line.



In the same chart, a notable increase in medical incidents is evident, closely linked to the onset of COVID. Additionally, it is observable that medical incidents experienced a sharp rise in 2020 during the occurrence of riots.

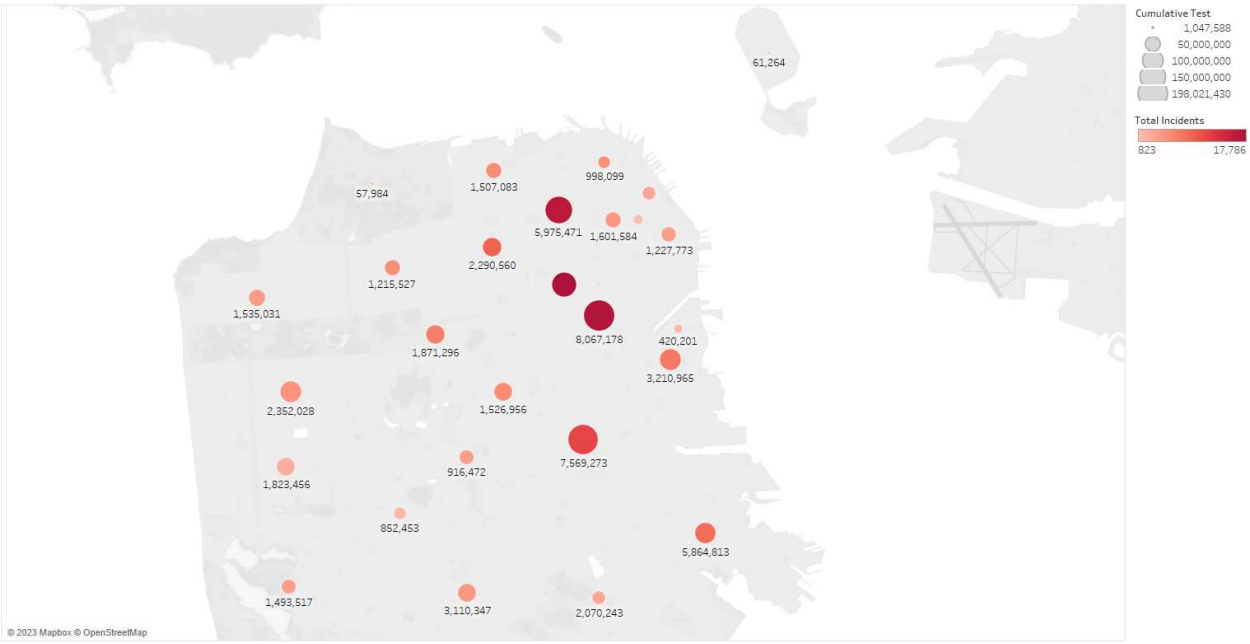


We use Excel 3D to plot data points of fires that occurred during the Floyd riots. By examining the timeline from 5/27/2020 to 5/28/2020, we depict the areas affected by fires calls made during these riots.



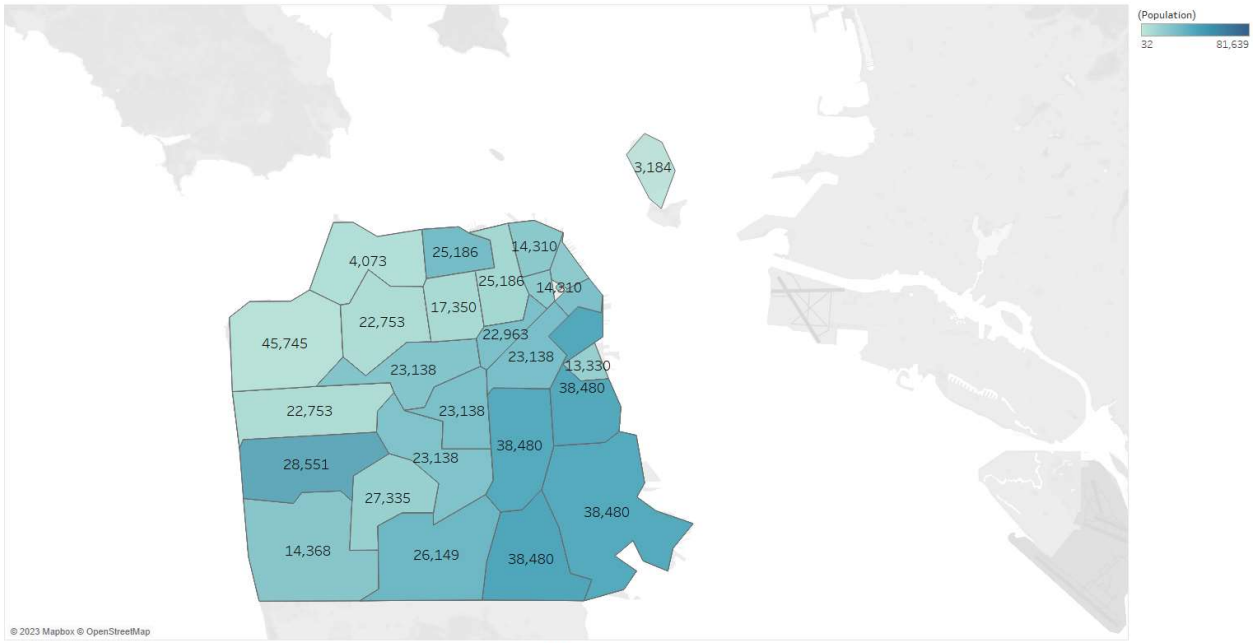
Covid19_Testing.CSV visuals on Tableau

Spatial Analysis of Fire Incidents and Number of test taken



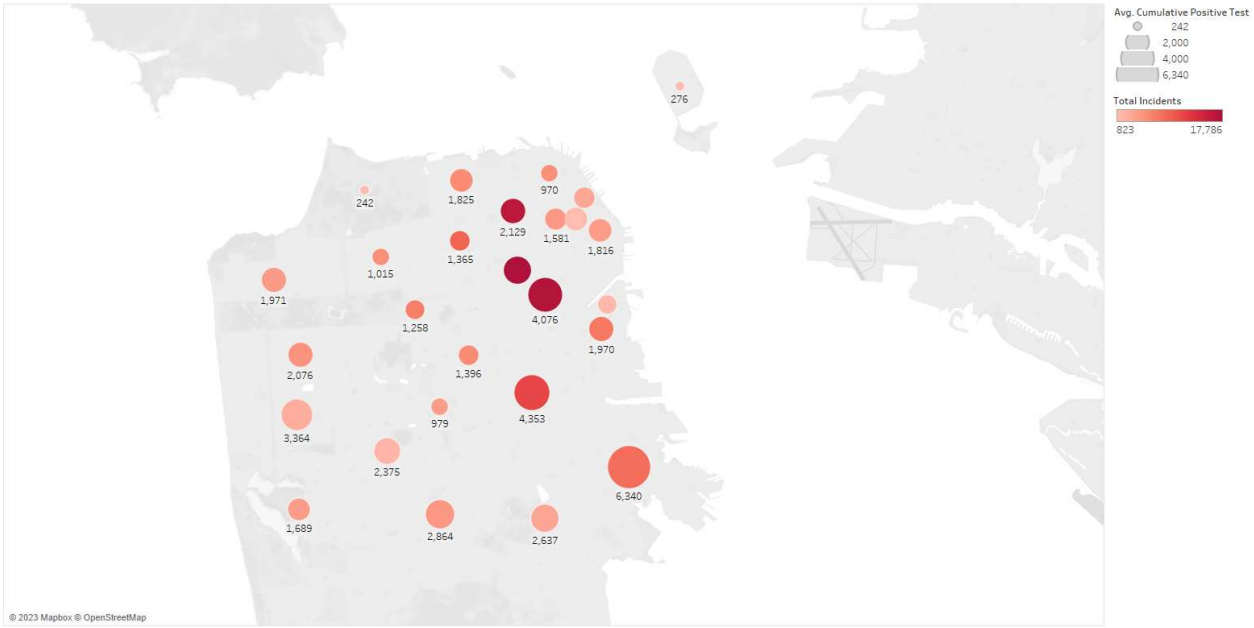
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Total Incidents. Size shows sum of Cumulative Test. The marks are labeled by sum of Cumulative Positive Test. Details are shown for Zip Code. The data is filtered on Incident Type and Neighborhood (COVID-19 Testing SF). The Incident Type filter keeps 11 of 33 members. The Neighborhood (COVID-19 Testing SF) filter excludes Null.

Population Density Of San Fransico



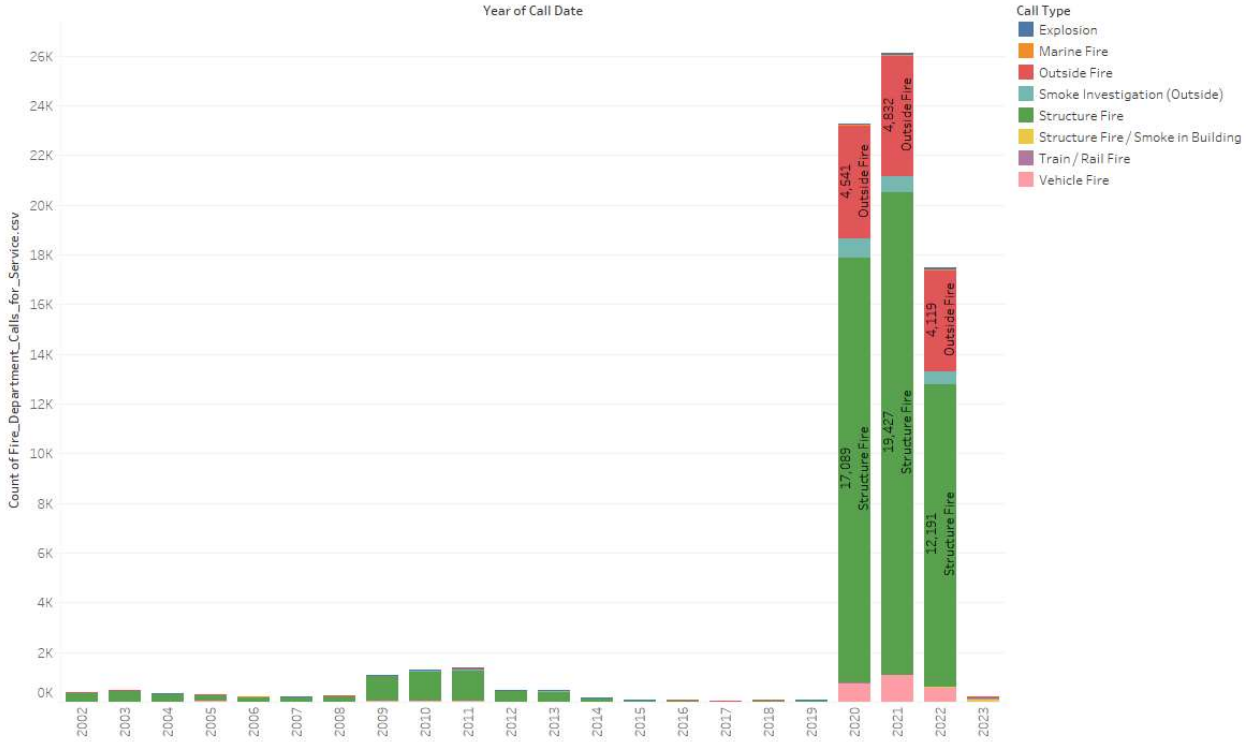
Map based on Longitude (generated) and Latitude (generated). Color shows average of Population. The marks are labeled by average of Population. Details are shown for Zip Code and Neighborhood.

Spatial Analysis of Fire Incidents and Number of Positive Test results

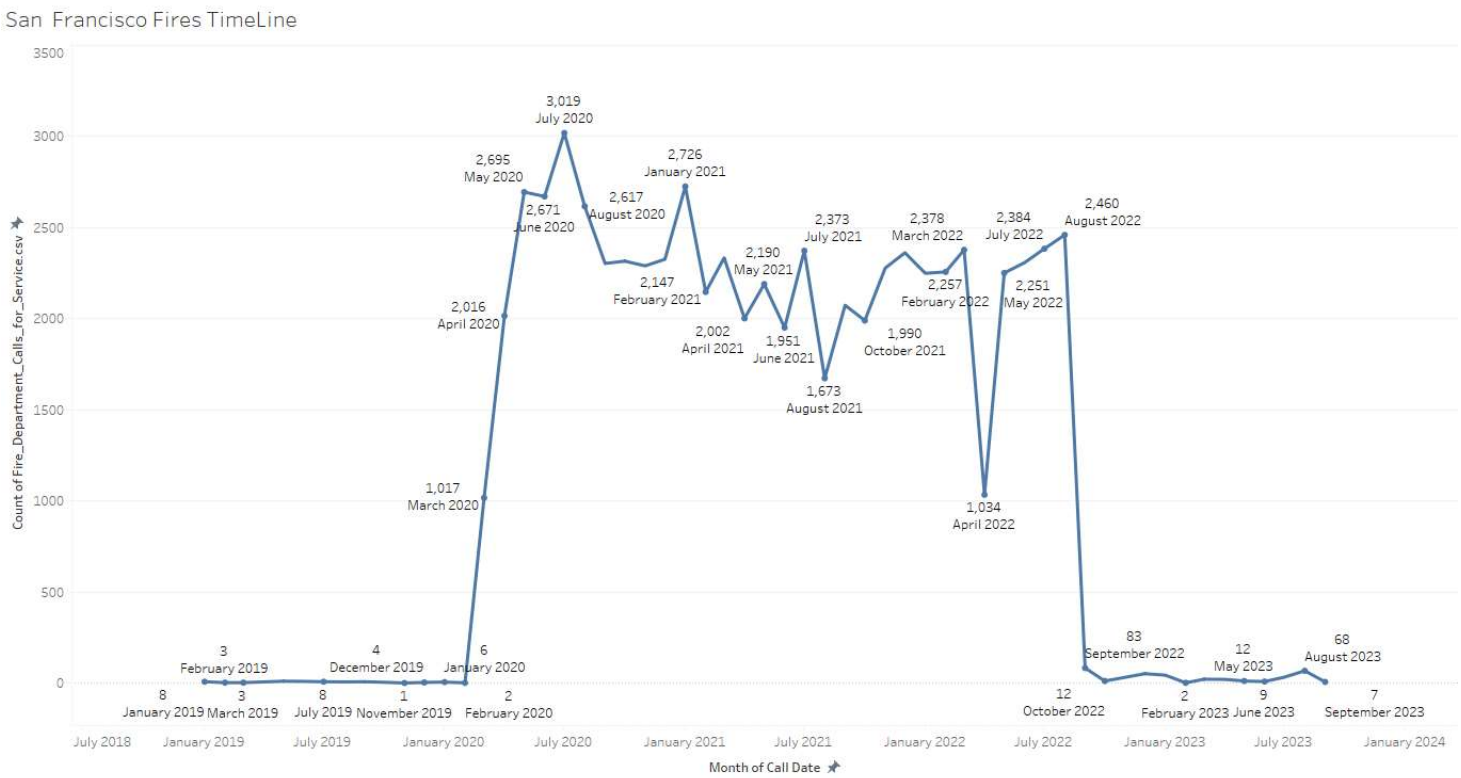


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Total Incidents. Size shows average of Cumulative Positive Test. The marks are labeled by average of Cumulative Positive Test. Details are shown for Zip Code. The data is filtered on Incident Type and Neighborhood (COVID-19 Testing SF). The Incident Type filter keeps 11 of 33 members. The Neighborhood (COVID-19 Testing SF) filter excludes Null.

San Francisco Fires



Between 2020 and 2022, there was a notable surge in fires in San Francisco, followed by a subsequent decline in 2023.



There was a noticeable increase in fires coinciding with the lockdown of San Francisco in March 2020.

REFERENCE

Datasets were sourced from the official website:

https://datasf.org/opendata/.

COVID-19 Testing by Geography Over Time

https://data.sfgov.org/COVID-19/COVID-19-Testing-by-Geography-Over-Time/qhc5-mubk/about_data

Fire Department Calls for Service

https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3/about_data