CIS4560 Term Project Tutorial

Authors: Sam Anteneh,
Gilbert Urbina,
Fatima Anammi,
Soliman Wahab,
Christian Najera
Instructor: Dr. Jongwook Woo
Date: 12/10/2024

Lab Tutorial

CoronaWhy Data Analysis using Hive

Objectives

- Get raw data files from GitHub
- Create directories in Hadoop Distributed File System
- Create tables using Hive Query Language
- Clean the raw data files using HQL for analysis
- Visualization

Platform Specifications

- Oracle Cluster / Hadoop 3.1.2
- Number of YARN nodes: 3
- Processors: AMD EPYC 7763
- Online CPUs: 6
- CPU Base Speed/Boost: 2.45 Ghz / 3.5 Ghz
- Memory: 32 GB
- Storage Capacity: 390.2 GB
- HDFS Allocated: 120 GB

Software:

- Excel
- Tableau
- OpenShot

Step 1: DOWNLOAD FILES

//launch git-bash or any terminal you use and login to your server
ssh username@your_server_ip

//download the zip file containing the dataset from GitHub, code should be one line
wget -O CoronaWhyDataset.zip https://github.com/G-Urbina/GroupProject-CoronaWhy/releases/download/CSV-files/CoronaWhy_Dataset.zip

//unzip the file
unzip CoronaWhyDataset.zip

//make sure you have the files
ls Public*

```
-bash-4.2$ ls Public*
Public_COVID-19_Canada_Cases.csv      Public_COVID-19_Canada_Recovered.csv
Public_COVID-19_Canada_Mortality.csv  Public_COVID-19_Canada_Testing.csv
```

Step 2: CREATE DIRECTORIES

//create CoronaWhy directory
hdfs dfs -mkdir CoronaWhy

//create four directories inside CoronaWhy for our tables, code should be one line
hdfs dfs -mkdir CoronaWhy/cases CoronaWhy/mortality CoronaWhy/recovered CoronaWhy/testing

//create five directories for our cleaned tables, code should be one line
hdfs dfs -mkdir CoronaWhy/cases_clean CoronaWhy/mortality_clean CoronaWhy/recovered_clean CoronaWhy/combined_death_recovery CoronaWhy/testing_clean

//make sure directories were created
hdfs dfs -ls CoronaWhy

```
-bash-4.2$ hdfs dfs -ls CoronaWhy
Found 9 items
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 04:42 CoronaWhy/cases
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 04:47 CoronaWhy/cases_clean
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/combined_death_recovery
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/mortality
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/mortality_clean
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/recovered
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/recovered_clean
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/testing
drwxr-xr-x   - gurbina6 hdfs          0 2024-12-10 21:35 CoronaWhy/testing_clean
```

//move unzipped csv files to the first four directories you created
hdfs dfs -put Public_COVID-19_Canada_Cases.csv CoronaWhy/cases

hdfs dfs -put Public_COVID-19_Canada_Mortality.csv CoronaWhy/mortality

hdfs dfs -put Public_COVID-19_Canada_Recovered.csv CoronaWhy/recovered

hdfs dfs -put Public_COVID-19_Canada_Testing.csv CoronaWhy/testing

//make sure files were put in the directories
hdfs dfs -ls CoronaWhy/cases

hdfs dfs -ls CoronaWhy/mortality

hdfs dfs -ls CoronaWhy/recovered

hdfs dfs -ls CoronaWhy/testing

Step 3: CREATE TABLES

//launch another git-bash or any terminal you use

//login to your server
ssh username@your_server_ip

//use HiveServer2
beeline

//use your directory
use username;

3.1 Cases Table

//drop "cases" table if it already exists and create table

!!!!replace username with yours using a text editor for all tables!!!!

```
DROP TABLE IF EXISTS cases;

CREATE EXTERNAL TABLE IF NOT EXISTS cases(case_id INT,
    provincial_case INT,
    age STRING,
    sex STRING,
    health_region STRING,
    province STRING,
    country STRING,
    date_report STRING,
    report_week STRING,
    travel_yn INT,
    travel_history_country STRING,
    locally_acquired STRING,
    case_source STRING,
    additional_info STRING,
    additional_source STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/username/CoronaWhy/cases'
TBLPROPERTIES ('skip.header.line.count'='1');
```

//make sure table was created

show tables;

//query should return 10 rows

select * from cases limit 10;

```
----------------------+-------------------------------------------------------------+--------------------------+
| cases.case_id  | cases.provincial_case  | cases.age  | cases.sex  | cases.health_region  | cases.province  | cases.country  | cases.date
_report  | cases.report_week  | cases.travel_yn  | cases.travel_history_country  | cases.locally_acquired  |                 cases.case_so
urce                  |                 cases.additional_info                 | cases.additional_source  |
+----------------+------------------------+------------+------------+----------------------+-----------------+----------------+-----------
---------+--------------------+------------------+-------------------------------+--------------------------+-------------------------
----------------------+-------------------------------------------------------------+--------------------------+
| 1              | 1                      | 50-59      | Male       | Toronto              | Ontario         | Canada         | 25-01-2020
         | 19-01-2020         | 1                | China                         |                          | (1) https://news.ontario.ca/m
ohltc/en/2020/01/ontario-confirms-first-case-of-wuhan-novel-coronavirus.html ;(2) https://globalnews.ca/news/6497313/coronavirus-timeine-c
ases-canada/; (3) https://globalnews.ca/news/6462626/coronavirus-toronto-hospital/ |                                                      |
                            |
| 2              | 2                      | 50-59      | Female     | Toronto              | Ontario         | Canada         | 27-01-2020
         | 26-01-2020         | 1                | China                         |                          | (1) https://news.ontario.ca/m
ohltc/en/2020/01/ontario-confirms-second-presumptive-case-of-wuhan-novel-coronavirus.html ;(2) https://globalnews.ca/news/6497313/coronavi
rus-timeine-cases-canada/ | Travel and Close Contact                                |                          |
| 3              | 1                      | 40-49      | Male       | Vancouver Coastal    | BC              | Canada         | 28-01-2020
         | 26-01-2020         | 1                | China                         |                          | https://news.gov.bc.ca/releas
es/2020HLTH0015-000151 |                                                             |                          |
| 4              | 3                      | 20-29      | Female     | Middlesex-London     | Ontario         | Canada         | 31-01-2020
         | 26-01-2020         | 1                | China                         |                          | (1) https://news.ontario.ca/m
ohltc/en/2020/01/ontario-confirms-third-case-of-2019-novel-coronavirus.html ;(2) https://globalnews.ca/news/6497313/coronavirus-timeine-ca
ses-canada/ |                                                             |                          |
| 5              | 2                      | 50-59      | Female     | Vancouver Coastal    | BC              | Canada         | 04-02-2020
         | 02-02-2020         | 0                |                               | Close Contact            | https://news.gov.bc.ca/releas
es/2020HLTH0023-000222 | The individual had close contact with family visitors from Wuhan city |                          |
| 6              | 3                      | 30-39      | Male       | Vancouver Coastal    | BC              | Canada         | 06-02-2020
         | 02-02-2020         | 1                | China                         |                          | https://news.gov.bc.ca/releas
es/2020HLTH0025-000236 |                                                             |                          |
| 7              | 4                      | 30-39      | Female     | Vancouver Coastal    | BC              | Canada         | 06-02-2020
         | 02-02-2020         | 1                | China                         |                          | https://news.gov.bc.ca/releas
es/2020HLTH0025-000236 |                                                             |                          |
| 8              | 5                      | 30-39      | Female     | Interior             | BC              | Canada         | 14-02-2020
         | 09-02-2020         | 1                | China                         |                          | (1) https://news.gov.bc.ca/re
```

3.2 Cleaned Cases Table

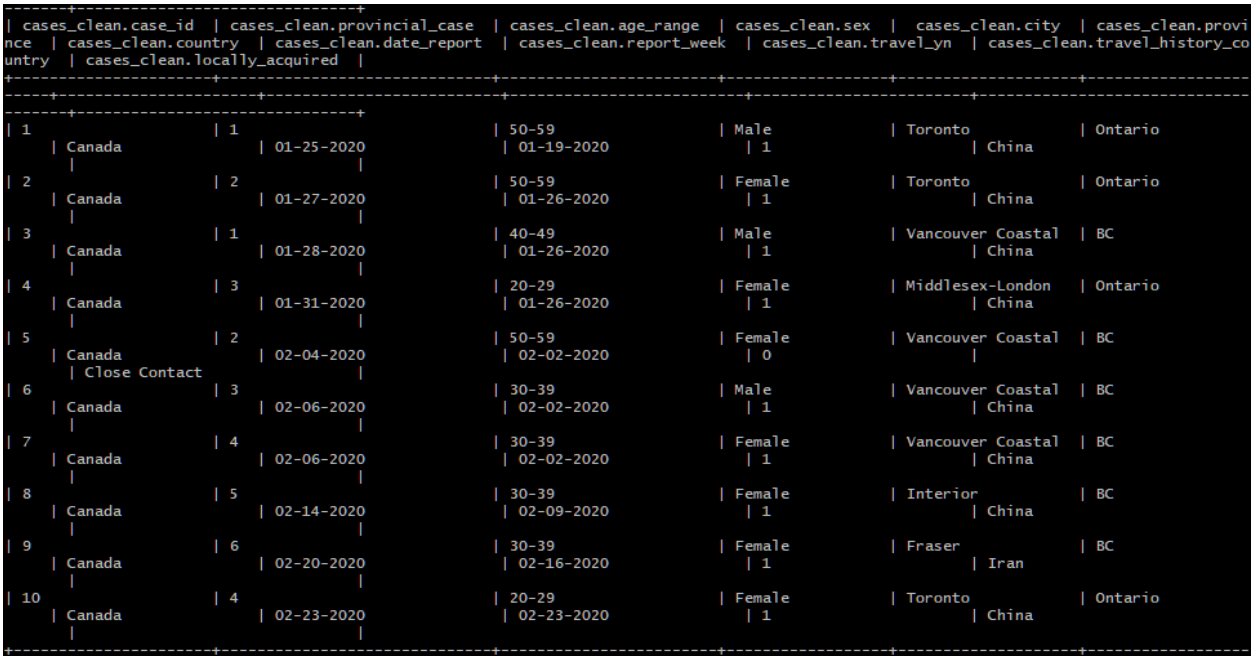
//create clean cases table, remove redundant fields, rename fields, and format date

```
DROP TABLE IF EXISTS cases_clean;

CREATE TABLE IF NOT EXISTS cases_clean
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/username/CoronaWhy/cases_clean/'
AS
SELECT
    case_id,
    provincial_case,
    age AS age_range,
    sex,
    health_region AS city,
    province,
    country,
    from_unixtime(unix_timestamp(date_report ,'dd-MM-yyyy'), 'MM-dd-yyyy')
date_report,
    from_unixtime(unix_timestamp(report_week ,'dd-MM-yyyy'), 'MM-dd-yyyy')
report_week,
    travel_yn,
    travel_history_country,
    locally_acquired
FROM cases;
```

//make sure table was created

show tables;

//query should return 10 rows



cases_clean.case_id	cases_clean.provincial_case	cases_clean.age_range	cases_clean.sex	cases_clean.city	cases_clean.province	cases_clean.country	cases_clean.date_report	cases_clean.report_week	cases_clean.travel_yn	cases_clean.travel_history_country	cases_clean.locally_acquired
1	1	50-59	Male	Toronto	Ontario	Canada	01-25-2020	01-19-2020	1	China	
2	2	50-59	Female	Toronto	Ontario	Canada	01-27-2020	01-26-2020	1	China	
3	1	40-49	Male	Vancouver Coastal	BC	Canada	01-28-2020	01-26-2020	1	China	
4	3	20-29	Female	Middlesex-London	Ontario	Canada	01-31-2020	01-26-2020	1	China	
5	2	50-59	Female	Vancouver Coastal	BC	Canada	02-04-2020	02-02-2020	0		Close Contact
6	3	30-39	Male	Vancouver Coastal	BC	Canada	02-06-2020	02-02-2020	1	China	
7	4	30-39	Female	Vancouver Coastal	BC	Canada	02-06-2020	02-02-2020	1	China	
8	5	30-39	Female	Interior	BC	Canada	02-14-2020	02-09-2020	1	China	
9	6	30-39	Female	Fraser	BC	Canada	02-20-2020	02-16-2020	1	Iran	
10	4	20-29	Female	Toronto	Ontario	Canada	02-23-2020	02-23-2020	1	China	

3.3 Mortality Table

//create mortality table

```
DROP TABLE IF EXISTS mortality;

CREATE EXTERNAL TABLE IF NOT EXISTS mortality (
    death_id INT,
    province_death INT,
    case_id INT,
    age STRING,
    sex STRING,
    health_region STRING,
    province STRING,
    country STRING,
    date_death_report STRING,
    death_source STRING,
    additional_info STRING,
    additional_source STRING
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/username/CoronaWhy/mortality'
TBLPROPERTIES ('skip.header.line.count'='1');
```

//make sure table was created

show tables;

//query should return 10 rows

select * from mortality limit 10;

```
+--------------------+-------------------------+--------------------+---------------+---------------+--------------------------+-----
---------------+-------------------+-----------------------------+-----------------------------------------------+----------------
-----------------------------------+-------------------------------------------------+
| mortality.death_id  | mortality.province_death  | mortality.case_id  | mortality.age  | mortality.sex  | mortality.health_region  | mort
ality.province  | mortality.country  | mortality.date_death_report  |              mortality.death_source             |            mor
tality.additional_info              |           mortality.additional_source           |
+--------------------+-------------------------+--------------------+---------------+---------------+--------------------------+-----
---------------+-------------------+-----------------------------+-----------------------------------------------+----------------
-----------------------------------+-------------------------------------------------+
| 1                  | 1                       | 60                 | 80-89         | Male          | Vancouver Coastal        | BC
             | Canada            | 08-03-2020                  | https://news.gov.bc.ca/releases/2020HLTH0068-000423 | Lynn Valley Re
sident                             |                                                 |
| 2                  | 1                       | 477                | 70-79         | Male          | Simcoe Muskoka           | Onta
rio            | Canada            | 11-03-2020                  | https://www.nationalobserver.com/2020/03/17/news/unconfirmed-covid-1
9-death-reported-ontarios-muskoka | Was being treated at Royal Victoria Regional Health Centre |
        |
| 3                  | 2                       | NULL               |               |               | Vancouver Coastal        | BC
             | Canada            | 16-03-2020                  | https://news.gov.bc.ca/releases/2020HLTH0086-000499 | Lynn Valley Re
sident                             |                                                 |
| 4                  | 3                       | NULL               |               |               | Vancouver Coastal        | BC
             | Canada            | 16-03-2020                  | https://news.gov.bc.ca/releases/2020HLTH0086-000499 | Lynn Valley Re
sident                             |                                                 |
| 5                  | 4                       | NULL               |               |               | Vancouver Coastal        | BC
             | Canada            | 16-03-2020                  | https://news.gov.bc.ca/releases/2020HLTH0086-000499 | Lynn Valley Re
sident                             |                                                 |
| 6                  | 5                       | NULL               |               |               | Vancouver Coastal        | BC
             | Canada            | 17-03-2020                  | https://vancouverisland.ctvnews.ca/b-c-declares-public-health-emerge
ncy-with-12-cases-of-covid-19-on-vancouver-island-1.4857080 | Lynn Valley Resident                            |
        |
| 7                  | 6                       | NULL               |               |               | Vancouver Coastal        | BC
             | Canada            | 17-03-2020                  | https://vancouverisland.ctvnews.ca/b-c-declares-public-health-emerge
ncy-with-12-cases-of-covid-19-on-vancouver-island-1.4857080 | Lynn Valley Resident                            |
        |
| 8                  | 7                       | NULL               | 80-89         | Male          | Fraser                   | BC
             | Canada            | 17-03-2020                  | https://vancouverisland.ctvnews.ca/b-c-declares-public-health-emerge
ncy-with-12-cases-of-covid-19-on-vancouver-island-1.4857080 | The other death is a man in his 80s in the Fraser Health region in the lower
 mainland. |                                                 |
| 9                  | 1                       | NULL               | 80-89         | Female        | Lanaudi?re               | Queb
ec             | Canada            | 18-03-2020                  | https://montreal.ctvnews.ca/covid-19-quebec-has-confirmed-its-first-
death-due-to-coronavirus-1.4858180 | Lived in senior's residence                     | https://globalnews.ca/news/6705211/granddaughter
-quebec-covid-19-victim-plea-social-media/ |
| 10                 | 2                       | 806                | 50-59         | Male          | Halton                   | Onta
rio            | Canada            | 19-03-2020                  | https://globalnews.ca/news/6701911/coronavirus-second-death-ontario/
 |                                                 |                                                 |
+--------------------+-------------------------+--------------------+---------------+---------------+--------------------------+-----
---------------+-------------------+-----------------------------+-----------------------------------------------+----------------
-----------------------------------+-------------------------------------------------+
```

3.4 Cleaned Mortality Table

//create clean mortality table, remove redundant fields, and format date

```
DROP TABLE IF EXISTS mortality_clean;

CREATE TABLE IF NOT EXISTS mortality_clean
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/username/CoronaWhy/mortality_clean'
AS
SELECT
    death_id,
    province,
    country,
    from_unixtime(unix_timestamp(date_death_report, 'dd-MM-yyyy'), 'MM-dd-
yyyy') AS date_death_report
FROM mortality;
```

//make sure table was created

show tables;

//query should return 10 rows

select * from mortality_clean limit 10;

mortality_clean.death_id	mortality_clean.province	mortality_clean.country	mortality_clean.date_death_report
1	BC	Canada	03-08-2020
2	Ontario	Canada	03-11-2020
3	BC	Canada	03-16-2020
4	BC	Canada	03-16-2020
5	BC	Canada	03-16-2020
6	BC	Canada	03-17-2020
7	BC	Canada	03-17-2020
8	BC	Canada	03-17-2020
9	Quebec	Canada	03-18-2020
10	Ontario	Canada	03-19-2020

3.5 Recovered Table

//create recovered table

```
DROP TABLE IF EXISTS recovered;

CREATE EXTERNAL TABLE IF NOT EXISTS recovered (
    date_recovered STRING,
    province STRING,
    cumulative_recovered STRING,
    province_source STRING,
    source STRING,
    additional_source STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/username/CoronaWhy/recovered'
TBLPROPERTIES ('skip.header.line.count'='1');
```

//make sure table was created
show tables;

//query should return 10 rows
select * from recovered limit 10;

recovered.date_recovered	recovered.province	recovered.cumulative_recovered	recovered.province_source	recovered.source	recovered.additional_source
12-02-2020	Alberta	NA	Alberta	https://www.alberta.ca/covid-19-alberta-data.aspx	NULL
13-02-2020	Alberta	NA	BC	http://www.bccdc.ca/health-info/diseases-conditions/covid-19/case-counts-press-statements	NULL
14-02-2020	Alberta	NA	Manitoba	https://www.gov.mb.ca/covid19/index.html	NULL
15-02-2020	Alberta	NA	New Brunswick	https://www2.gnb.ca/content/gnb/en/departments/ocmoh/cdc/content/respiratory_diseases/coronavirus.html	NULL
16-02-2020	Alberta	NA	NL	https://www.gov.nl.ca/covid-19/	NULL
17-02-2020	Alberta	NA	Nova Scotia	https://novascotia.ca/coronavirus/#cases	NULL
18-02-2020	Alberta	NA	Nunavut	https://www.gov.nu.ca/	NULL
19-02-2020	Alberta	NA	NWT	https://www.hss.gov.nt.ca/en/services/coronavirus-disease-covid-19	NULL
20-02-2020	Alberta	NA	Ontario	https://www.ontario.ca/page/2019-novel-coronavirus#section-0	NULL
21-02-2020	Alberta	NA	PEI	https://www.princeedwardisland.ca/en/topic/covid-19	NULL

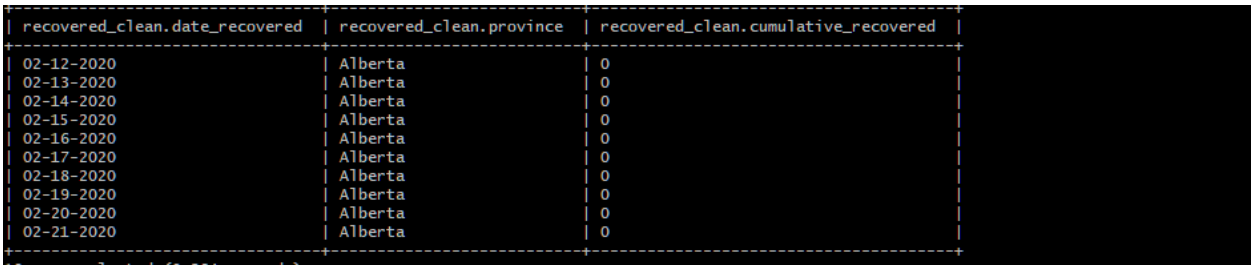
3.6 Cleaned Recovered Table

//create clean recovered table, remove redundant fields, format dates, and replace "NA"

```
DROP TABLE IF EXISTS recovered_clean;

CREATE TABLE IF NOT EXISTS recovered_clean
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/username/CoronaWhy/recovered_clean/'
AS
SELECT
    FROM_UNIXTIME(UNIX_TIMESTAMP(date_recovered, 'dd-MM-yyyy'), 'MM-dd-yyyy') AS date_recovered,
    province,
    CASE
        WHEN cumulative_recovered = 'NA' THEN 0
        ELSE cumulative_recovered
    END AS cumulative_recovered
FROM recovered;
```

//make sure table was created
show tables;

//query should return 10 rows
select * from recovered_clean limit 10;



recovered_clean.date_recovered	recovered_clean.province	recovered_clean.cumulative_recovered
02-12-2020	Alberta	0
02-13-2020	Alberta	0
02-14-2020	Alberta	0
02-15-2020	Alberta	0
02-16-2020	Alberta	0
02-17-2020	Alberta	0
02-18-2020	Alberta	0
02-19-2020	Alberta	0
02-20-2020	Alberta	0
02-21-2020	Alberta	0

3.7 Combine Mortality and Recovered Table

//create combined_death_recovery table

```
DROP TABLE IF EXISTS combined_death_recovery;

CREATE TABLE IF NOT EXISTS combined_death_recovery (
    event_date STRING,
    event_id INT,
    event_type STRING,
    event_province STRING,
    event_country STRING,
    cumulative_recovered INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/username/CoronaWhy/combined_death_recovery/';
```

//make sure table was created

show tables;

//insert fields from the mortality_clean and recovered_clean tables into combined_death_recovered using UNION ALL

```
INSERT INTO TABLE combined_death_recovery
SELECT

    from_unixtime(unix_timestamp(date_death_report, 'dd-MM-yyyy'), 'MM-dd-
yyyy') AS event_date,
    death_id AS event_id,
    'Death' AS event_type,
    province AS event_province,
    country AS event_country,
    1 AS cumulative_recovered

FROM mortality_clean

UNION ALL

SELECT

    from_unixtime(unix_timestamp(date_recovered, 'dd-MM-yyyy'), 'MM-dd-
yyyy') AS event_date,
    ROW_NUMBER() OVER (ORDER BY date_recovered) + 35 AS event_id,
    'Recovery' AS event_type,
    province AS event_province,
    'Canada' AS event_country,
    CASE
        WHEN cumulative_recovered = 'NA' THEN 0
        ELSE CAST(cumulative_recovered AS INT)
    END AS cumulative_recovered
FROM recovered_clean;
```

//query should return 10 rows

select * from combined_death_recovery limit 10;

combined_death_recovery.event_date	combined_death_recovery.event_id	combined_death_recovery.event_type	combined_death_recovery.event_province	combined_death_recovery.event_country	combined_death_recovery.cumulative_recovered
08-03-2020	1	Death	BC	Canada	1
11-03-2020	2	Death	Ontario	Canada	1
04-03-2021	3	Death	BC	Canada	1
04-03-2021	4	Death	BC	Canada	1
04-03-2021	5	Death	BC	Canada	1
05-03-2021	6	Death	BC	Canada	1
05-03-2021	7	Death	BC	Canada	1
05-03-2021	8	Death	BC	Canada	1
06-03-2021	9	Death	Quebec	Canada	1
07-03-2021	10	Death	Ontario	Canada	1

3.8 Testing Table

//create testing table

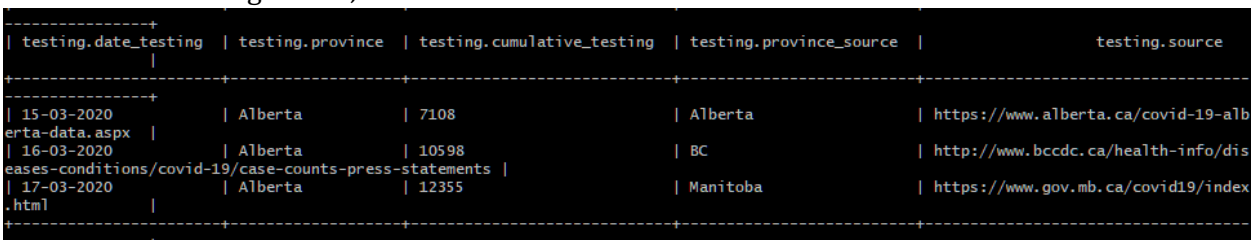
```
DROP TABLE IF EXISTS testing;

CREATE EXTERNAL TABLE IF NOT EXISTS testing (
    date_testing STRING,
    province STRING,
    cumulative_testing INT,
    province_source STRING,
    source STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/username/CoronaWhy/testing'
TBLPROPERTIES ('skip.header.line.count'='1');
```

//make sure table was created

show tables;

//query should return 3 rows

select * from testing limit 3;



```
+----------------+------------------+-----------------------------+--------------------------+-----------------------------------
----------------+
| testing.date_testing  | testing.province  | testing.cumulative_testing  | testing.province_source  |                  testing.source
          |
+-----------------------+-------------------+-----------------------------+--------------------------+-----------------------------------
----------------+
| 15-03-2020            | Alberta           | 7108                        | Alberta                  | https://www.alberta.ca/covid-19-alb
erta-data.aspx  |
| 16-03-2020            | Alberta           | 10598                       | BC                       | http://www.bccdc.ca/health-info/dis
eases-conditions/covid-19/case-counts-press-statements |
| 17-03-2020            | Alberta           | 12355                       | Manitoba                 | https://www.gov.mb.ca/covid19/index
.html           |
+-----------------------+-------------------+-----------------------------+--------------------------+-----------------------------------
```

3.9 Cleaned Testing Table

//create clean testing table, format date, correct province names, replace null values

```
DROP TABLE IF EXISTS testing_clean;

CREATE TABLE testing_clean
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/username/CoronaWhy/testing_clean/'

AS
SELECT
    CASE
        WHEN date_testing IS NULL OR TRIM(date_testing) = '' THEN ' '
        ELSE
            CASE
                WHEN unix_timestamp(date_testing, 'dd-MM-yyyy') IS NOT NULL
                THEN from_unixtime(unix_timestamp(date_testing, 'dd-MM-
yyyy'), 'MM-dd-yyyy')
                ELSE ''
            END
    END AS date_testing,

    CASE
        WHEN province = 'BC' THEN 'British Columbia'
        WHEN province = 'NL' THEN 'Newfoundland and Labrador'
        WHEN province = 'PEI' THEN 'Prince Edward Island'
        WHEN province = 'NWT' THEN 'Northwest Territories'
        ELSE province
    END AS province,

    CASE
        WHEN cumulative_testing IS NULL AND (date_testing IS NULL OR
TRIM(date_testing) = '') THEN ' '
        WHEN cumulative_testing IS NULL AND date_testing IS NOT NULL THEN
'0'
        ELSE cumulative_testing
    END AS cumulative_testing
FROM testing;
```

//make sure table was created

show tables;

//query should return 3 rows

select * from testing_clean limit 3;

testing_clean.date_testing	testing_clean.province	testing_clean.cumulative_testing
03-15-2020	Alberta	7108
03-16-2020	Alberta	10598
03-17-2020	Alberta	12355

Step 4: Create Visualizations

4.1 DOWNLOAD CASES_CLEAN CSV FILE

//launch git-bash or any terminal you use to login

//make sure the cases_clean file 000000_0 is in the directory
hdfs dfs -ls CoronaWhy/cases_clean

//get file and rename it
hdfs dfs -get CoronaWhy/cases_clean/000000_0 CanadaCases.csv

//open another terminal and download file to your pc, replace username and server Ip with yours
scp username@your_server_ip:/home/username/CanadaCases.csv CanadaCases.csv

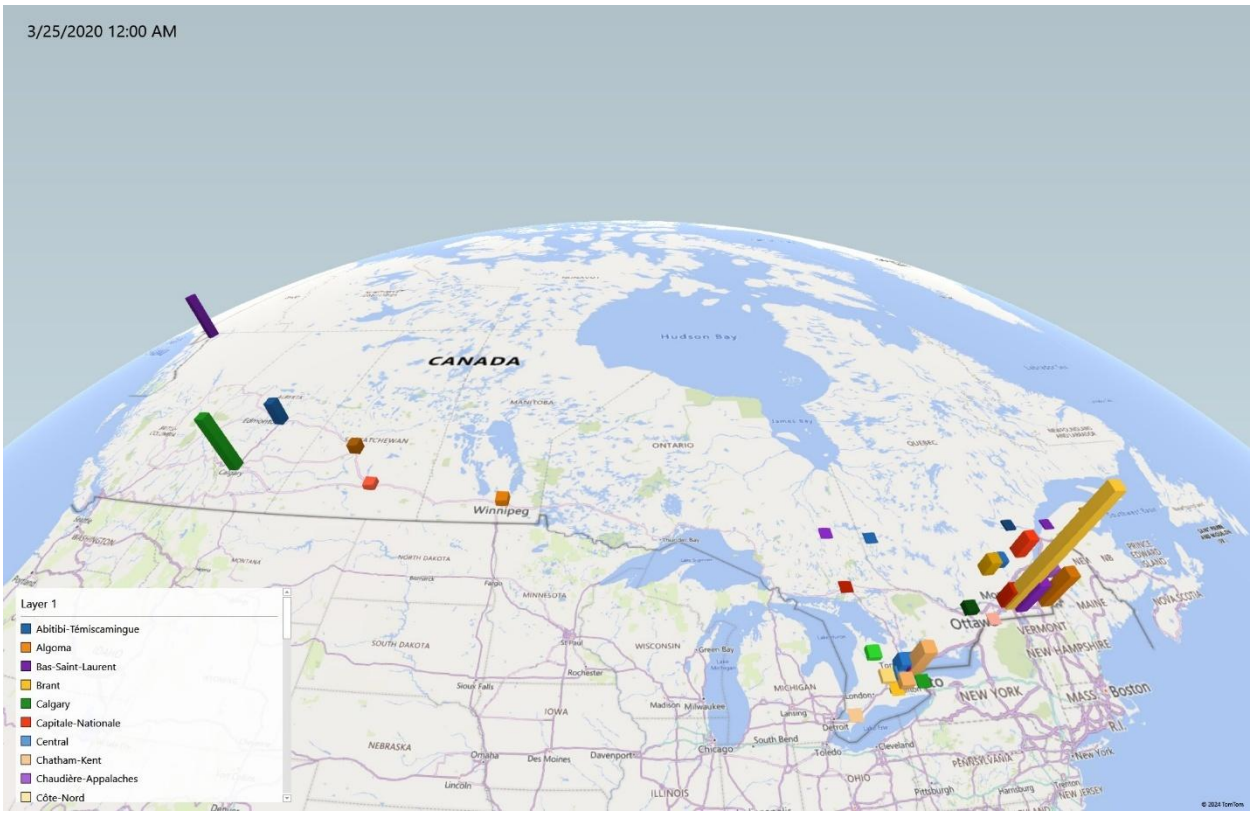
4.2 CREATE XLSX FILE WITH EXCEL

1. Open Excel
2. Open a new blank workbook
3. Click on the **Data** tab
4. Click on **Get Data > From File > From Text/CSV**
5. Select the **CanadaCases.csv** file you downloaded, click import
6. When you see a preview of the data, click on load
7. Rename the columns from Column1 to Column12 with case_id, provincial_case, age_range, sex, city, province, country, date_report, report_week, travel_yn, travel_history_country, locally_acquired
8. Save the workbook as Canada_Covid19_Report.xlsx

4.3 CREATE 3D MAP IN EXCEL – Cities

1. Open Canada_Covid19_Report.xlsx
2. Click on the **Insert** tab and then **3D Map**
3. With the 3D Map window open, enable the **Map Labels**
4. For **Location**, add the city field as City
5. For **Height** select case_id and change it to (Count - Not Blank)
6. For **Category** select city

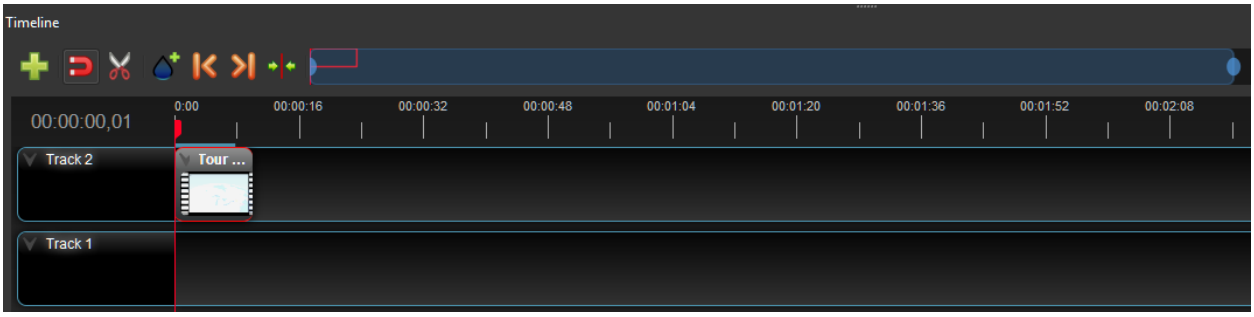
7. For **Time** select date_report
8. For **filters** add city, select all and then uncheck "Not Reported"
9. Hover mouse over the bars to view total cases per city (case_id - Count)
10. (Optional) To get a better view of the data click on the 3D Map down arrow until you see Canada and the bars. If you don't see the legends click on the Legends tab



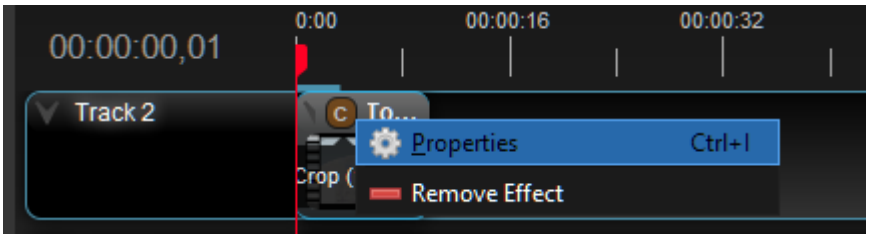
4.4 CREATE 3D MAP VIDEO – Cities

1. Download and install the open-source video editor OpenShot from one of the links below
   Website: https://www.openshot.org/
   GitHub: https://github.com/OpenShot/openshot-qt/releases
2. With the 3D Map created in section **4.3** try to align the map as the image above (Hint: use the arrows on the map to get a better view)
3. In Excel **3D Map**, click on **Create Video** with **Presentations & HD Displays** selected, click on the Create button and wait for Excel to finish
4. Run the **OpenShot** program
5. Click on **File > Import Files** and select the 3D Map video you created

6. In the **Timeline** interface ensure two tracks are available, if not delete any extra tracks by right clicking on them until **Track 2** and **Track 1** are left
7. Drag and drop the video file in **Track 2** at the 0:00 mark



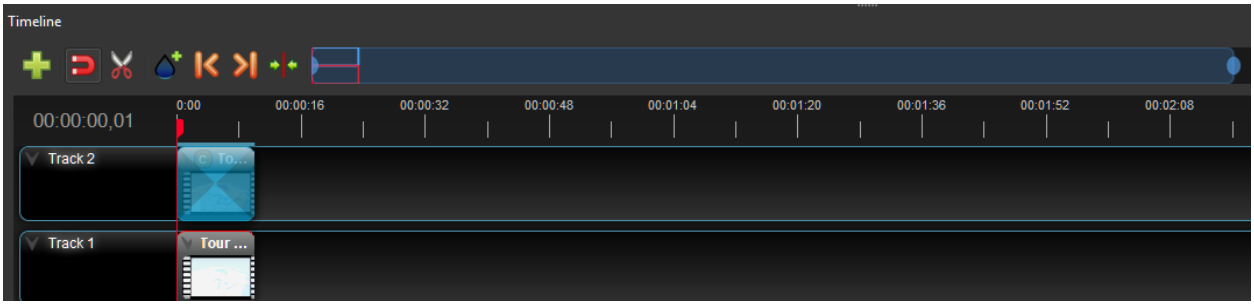
8. Click on the **Effects** tab above the **Timeline**
9. Drag and drop the **Crop** effect over the video in **Track 2**, an icon with the letter "c" will appear
10. With the video at the starting point (0:00), right click the crop effect icon in the video timeline and select **properties**, change **X Offset** to 0.55 and **Y Offset** to 0.35, close the properties window



11. Click on the **Project Files** tab above the **Timeline**, drag and drop the video into **Track 1** at the 0:00 mark
12. Select the video in **Track 2** by clicking on it
13. With the video you just cropped selected, move the video in the **Video Preview** by clicking and holding down the left mouse button, move it to the bottom right corner until it matches with the background uncropped video

    Hint: keep an eye on the trademark on the bottom right corner and the edge of the globe
14. Click on the play button until the video stops at the end, then right click the video in the **Timeline** and select **properties**.
15. Change **Location X** to 0.65, **Location Y** to 0.42, **Scale X** to 1.47, and **Scale Y** to 1.47
16. OpenShot will automatically apply a smooth transition from the starting position of the cropped video to the position applied at the end of the video. Hit play to make sure the transition is working, if needed, adjust the position at the end of the video by dragging it

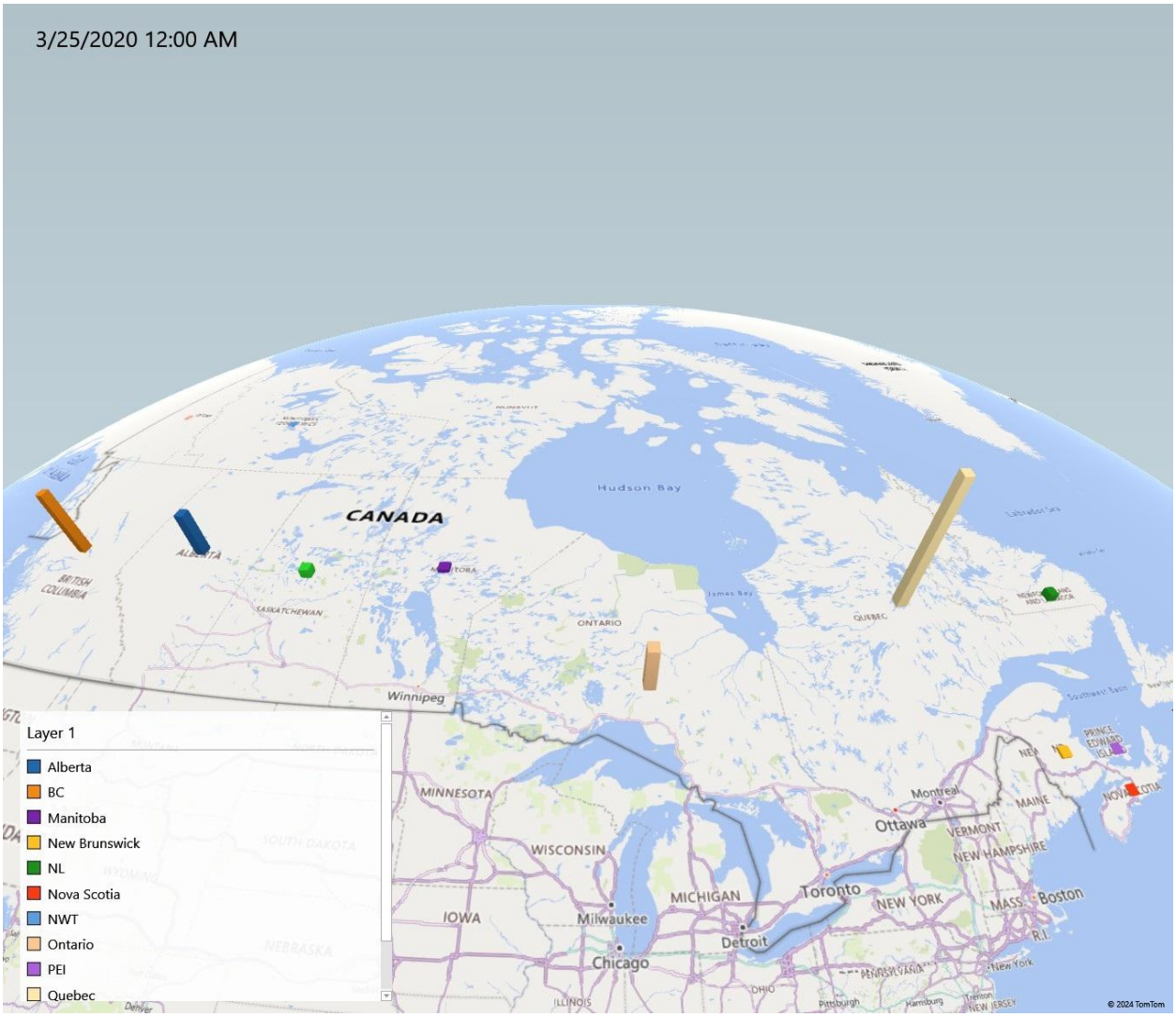
17. Click on the jump to start button **|<<** (0:00 mark)
18. Select the **Transitions** tab above the **Timeline** and drag and drop the **Circle in to out** effect over the cropped video in **Track 2**



19. Click on **File > Export Project > Export Video**
20. Choose the **Video Profile** HD 720p 60fps (1280x720) or any resolution you want. On the **advanced** tab, subtract 1 frame from the **End Frame** property, for example if it shows 605 change it to 604, this will remove the black frame at the end of the video.
21. Click **Export Video**

4.5 CREATE 3D MAP IN EXCEL – Provinces

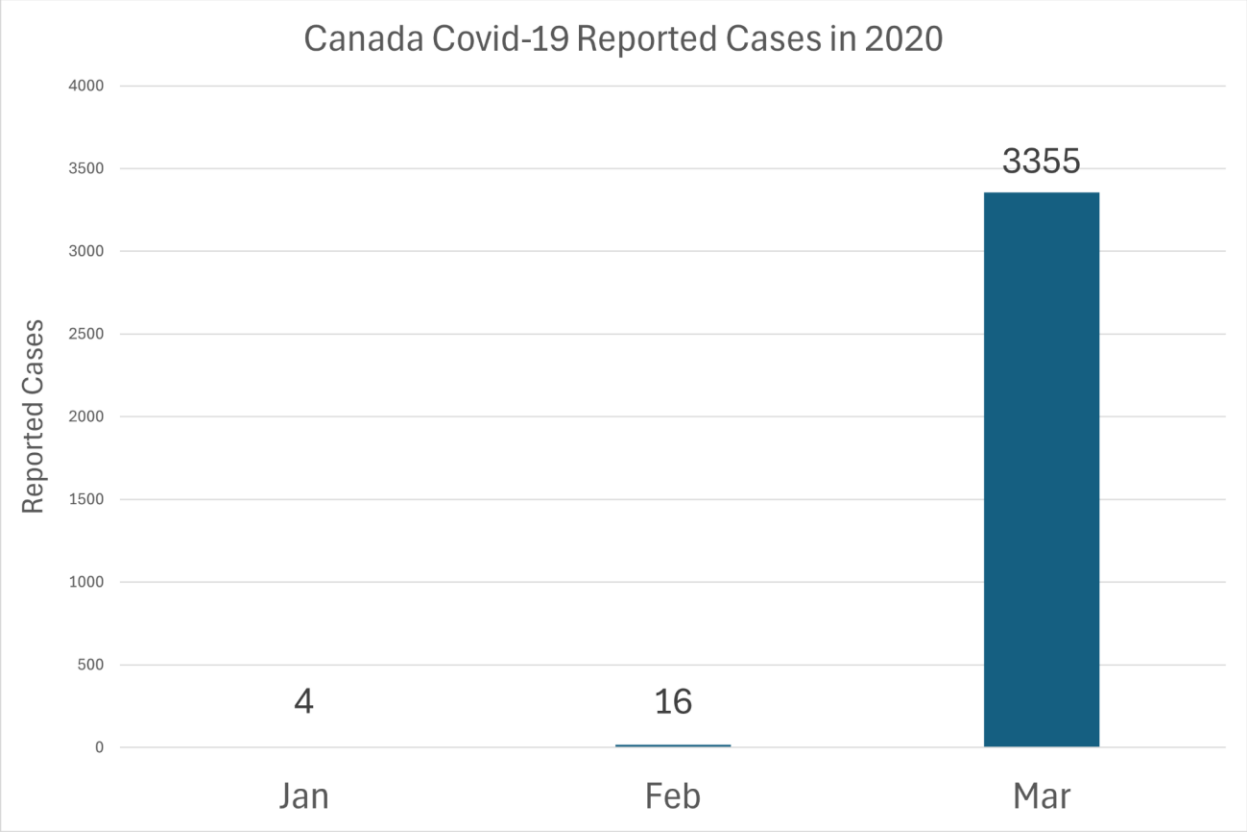
1. Open Canada_Covid19_Report.xlsx
2. Click on the **Insert** tab and then **3D Map**
3. For **Location** select province
4. For **Height** select case_id (Count – Not Blank)
5. For **Category** select province
6. For **Time** select date_report
7. Hover mouse over the bars to view total cases per province (case_id - Count)



4.6 CREATE BAR CHART USING PivotTable – Reported Cases Per Month

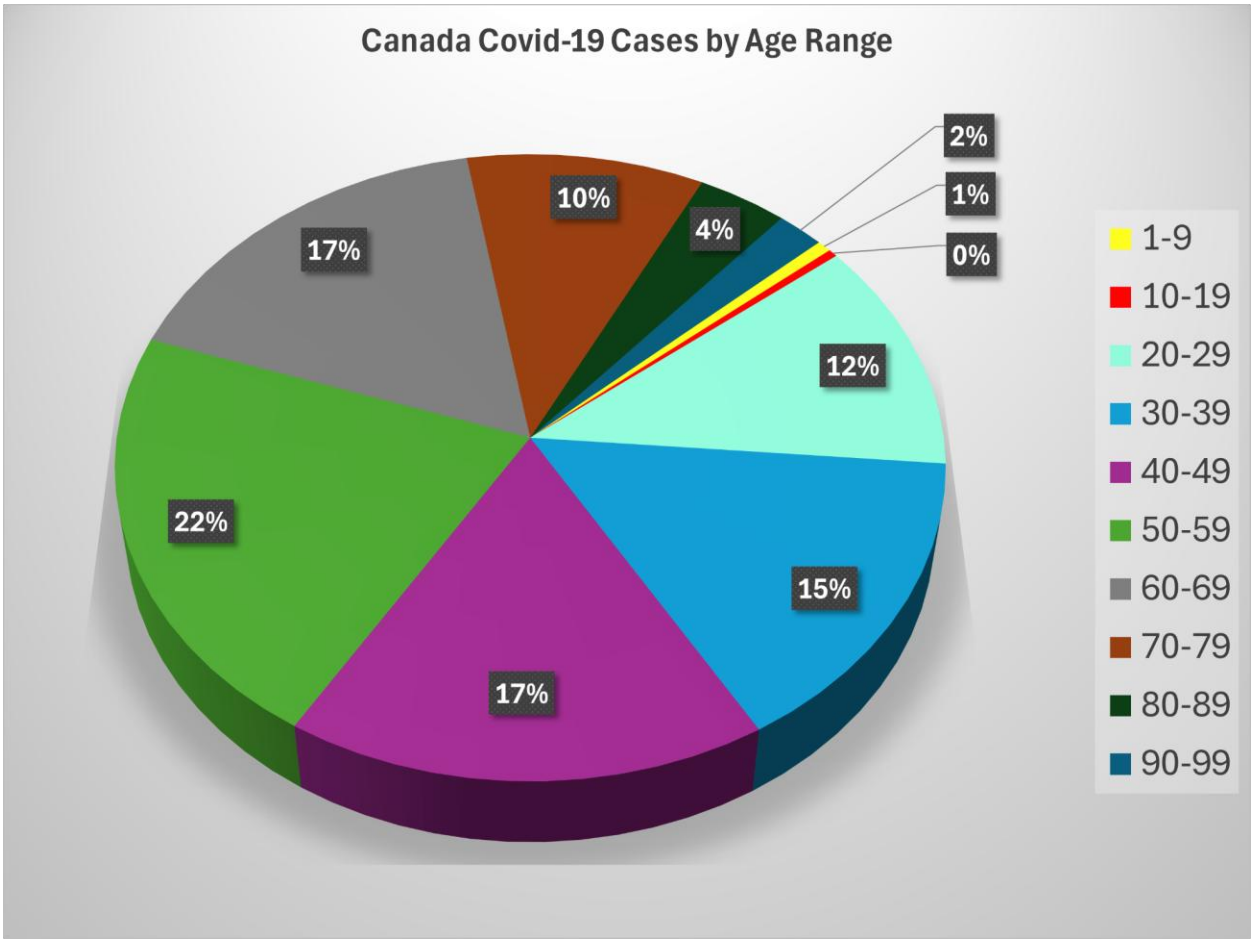
1. Open Canada_Covid19_Report.xlsx
2. Click on the **Insert** tab and then on **PivotTable**
3. On the right, the **Field List** should be visible. If not click on the **PivotChart Analyze** tab and select **Field List**
4. Drag and drop the case_id field into **Values** and then click the drop-down arrow, select **Value Field Settings**, and change it from Sum to **Count**, click OK.
5. Drag and drop the date_report field into the **Axis (Categories)** box

6. On the **PivotTable**, click on Row Labels or anywhere on the table
7. Click on the **Insert** tab and then select the first **2-D bar** chart
8. Right click on the "**Count of case_id**" button on the top left corner and select **Hide All Field Buttons on Chart**
9. Click on the plus icon (**Chart Elements**) and select **Axis Titles > Primary Vertical**, select **Data Labels**, and unselect **Legend**
10. Rename the vertical **Axis Title** to Reported Cases
11. Rename the title to Canada Covid-19 Reported Cases in 2020



4.7 CREATE PIE CHART USING PivotTable – Cases Per Age Range

1. Open Canada_Covid19_Report.xlsx
2. With the CanadaCases sheet selected, click on the **Insert** tab and then on **PivotTable**
3. On the right, the **Field List** should be visible. If not click on the **PivotChart Analyze** tab and select **Field List**
4. Drag and drop the case_id field into **Values** and then click the drop-down arrow, select **Value Field Settings**, and change it from Sum to **Count**, click OK.
5. Drag and drop the age_range field into **Rows**
6. On the **PivotTable** click on the **Row Labels** drop down menu uncheck all except, <10, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
7. Click on "<10" and in the **Formula Bar** rename it to "0-9"
8. On the **PivotTable**, click on Row Labels or anywhere on the table
9. Click on the **Insert** tab and then select **3-D Pie** chart
10. Right click on the "**Count of case_id**" button on the top left corner and select **Hide All Field Buttons on Chart**
11. Click on the plus icon (**Chart Elements**) and select **Data Labels**
12. Click on the **Styles** icon and select the third option with a silver background
13. Click on the **Styles** icon and change the color to your preference
14. Right click any of the **Data Labels (percentages)** and select **Format Data Labels**, on the right panel uncheck **Value**
15. Right click the pie and select **Format Data Series**, change **Angle of first slice** to 50 degrees
16. Adjust the position of the **Data Labels (percentages)** withing their slice. For 0%, 1%, and 2%, place them just outside their slices
17. Select the **Legends (age ranges)** box and click on the **Home tab > change Font size to 14**
18. Rename the title to "Canada Covid-19 Cases by Age Range"



4.8. DOWNLOAD COMBINED_DEATH_RECOVERY CSV FILE

//launch git-bash or any terminal you use to login

//make sure files 000000_0 and 000001_0 is in the directories

```
hdfs dfs -ls CoronaWhy/combined_death_recovery
hdfs dfs -ls CoronaWhy/combined_death_recovery/HIVE_UNION_SUBDIR_1
hdfs dfs -ls CoronaWhy/combined_death_recovery/HIVE_UNION_SUBDIR_2
```

//get 000000_0 and 000001_0 files

```
hdfs dfs -get CoronaWhy/combined_death_recovery/HIVE_UNION_SUBDIR_1/000000_0
hdfs dfs -get CoronaWhy/combined_death_recovery/HIVE_UNION_SUBDIR_2/000001_0
```

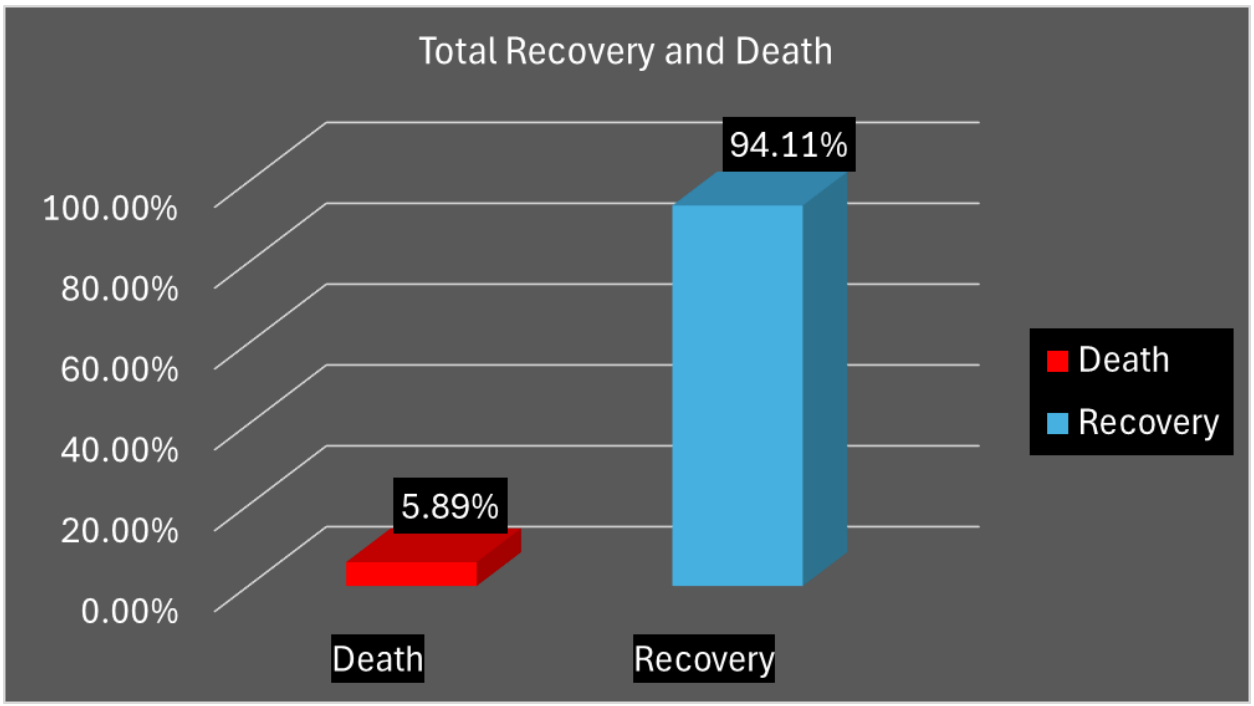
//now merge the two output files to combinedDeathRecovery.csv

```
cat 000000_0 000001_0 > combinedDeathRecovery.csv
```

//open another terminal and download file to your pc, replace username and server Ip with yours
scp username@your_server_ip:/home/username/combinedDeathRecovered.csv
combinedDeathRecovered.csv

4.8. OPEN CSV FILE IN EXCEL & CREATE BAR CHART – Recovered vs Deaths

1. Insert a new row above your data (select the top row and right-click to choose "Insert").
2. In the first row of your new columns, label each column as follows:
    a. Event_Date
    b. Event_id
    c. Event_type
    d. Event_province
    e. Event_country
    f. Cumulative_recovered
3. Save the Workbook as Recovered_Death_Table.xlsx
    a. **Insert Pivot Table**:
4. Select all data and go to **Insert** → **PivotTable**.
5. Choose **New Worksheet** and click **OK**.
    a. **Configure Pivot Table**:
6. Drag Event_type to **Rows**.
7. Drag Cumulative_recovered to **Values**.
8. Right-click on **Cumulative_recovered** in **Values**, select **Value Field Settings**, choose **Count**, then go to **Show Values As** → **% of Grand Total**.
    a. **Insert 3D Column Chart**:
9. Select the Pivot Table and go to **Insert** → **3-D Column** chart.
    a. **Customize Chart**:
10. Change the bar colors:
    a. Red for "Death" events.
    b. Blue for "Recovery" events.
11. Right-click bars → **Format Data Series** → **Fill** → Select colors.
    a. **Add Data Labels**:
12. Right-click on bars in the chart → **Add Data Labels** to display counts/percentages.



4.9 DOWNLOAD TESTING CSV FILE

//launch git-bash or any terminal you use to login

//check the 000000_0 file is in the testing_clean directory
hdfs dfs -ls CoronaWhy/testing_clean

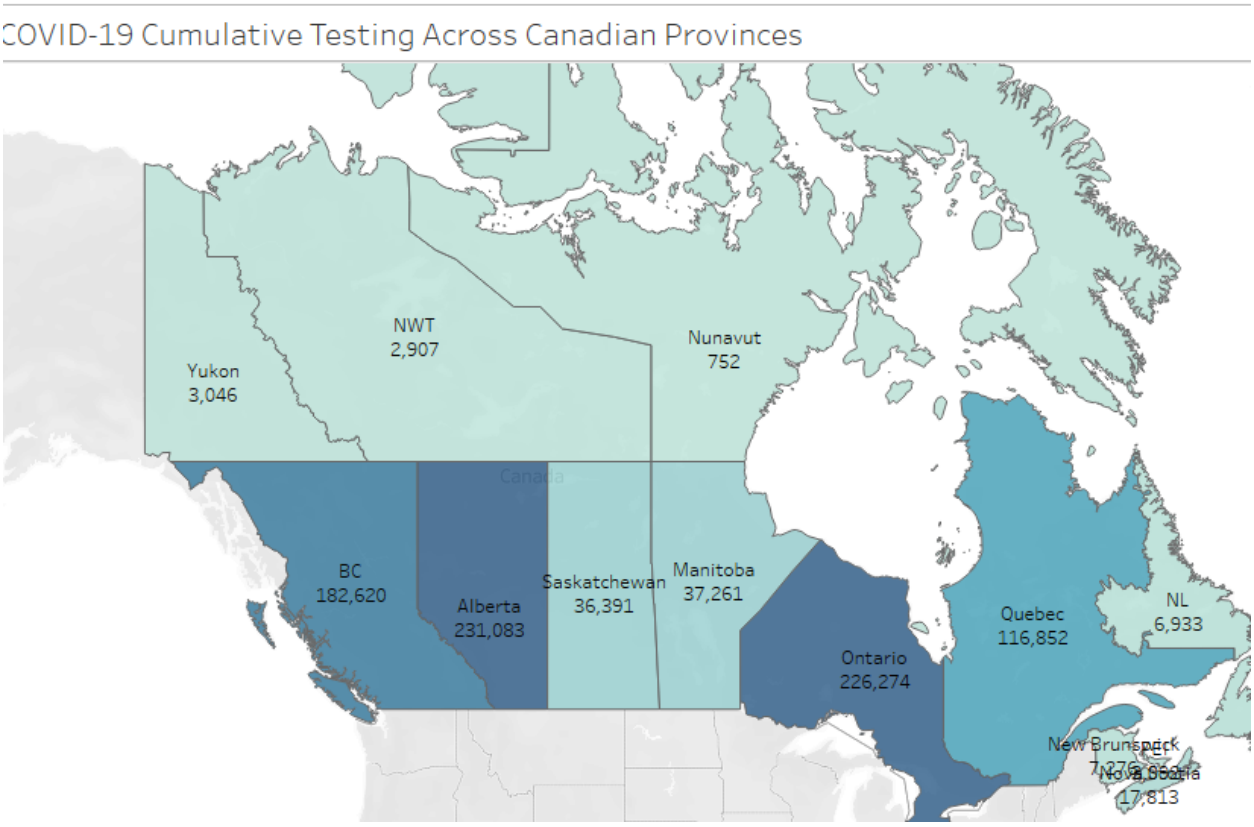
//get the 000000_0 file and rename it to Cumulative_testing.csv
hdfs dfs -get CoronaWhy/testing_clean/000000_0 Cumulative_testing.csv

//open another terminal and download file to your pc, replace username and server Ip with yours
scp username@your_server_ip:/home/username/Cumulative_testing.csv Cumulative_testing.csv

4.10 CREATE GEOGRAPHIC MAP USING TABLEAU – Cumulative Testing

1. Open Tableau and click on **Text File** to load the dataset.
2. Drag **Province** to Columns and **Cumulative Testing** to Rows.

3. Select **Map** from the visualization options.
    - If the data does not display, right-click on the map, select **Edit Locations**, and set:
        - **Country/Region** to "Canada."
        - **State/Province** to the field **Province.**
4. Drag the **Cumulative Testing** field to **Color** and the **Province** field to **Details** to display the province names and cumulative testing numbers directly on the map.



References

1. https://github.com/G-Urbina/GroupProject-CoronaWhy
2. https://www.kaggle.com/datasets/skylord/coronawhy
3. https://www.calstatela.edu/centers/hipic