

**SMU - Data Science**

2020

ETL Project – Group 6

Requirements Document

Authors –

Stephanie Smith

Cenez Tindoc

Monica Moussavi

Henry Greyner

R. Shane Livingston

SMU University, Dallas, Texas 75205

Table of Contents

[I. Overview………………………………………………………………………………………………………………………………………………………………... 3](#_Toc381186303)

[II. System Information………………………………………………………………………………………………………………………………………………… 3](#_Toc381186305)

[III. GitHub Folder Structure and Run Order …………………………………………………………………………………………………………………. 5](#_Toc381186306)

[IV. Detailed Requirements…………………………………………………………………………………………………………………………………………… 5](#_Toc381186306)

V. Images and Mockups ……………………………………………………………………………………………………………………………………………….6

ERD………………………………………………………………………………………………………………………………………………………………………7

# Overview

This team project demonstrates the abilities learned concerning Extract, Transform, and Load (ETL) gained in recent classes. The Group 6 team decided to work with three Virus’ data - Coronavirus, Swine, and Ebola. We sourced data sets from Kaggle, WHO, and CDC concerning Virus Type, Country, Cases, Deaths, Recovery, and Open Cases for the time- period the virus occurred:

* Coronavirus – 2020
* Ebola – 2015
* Swine Flu – 2009

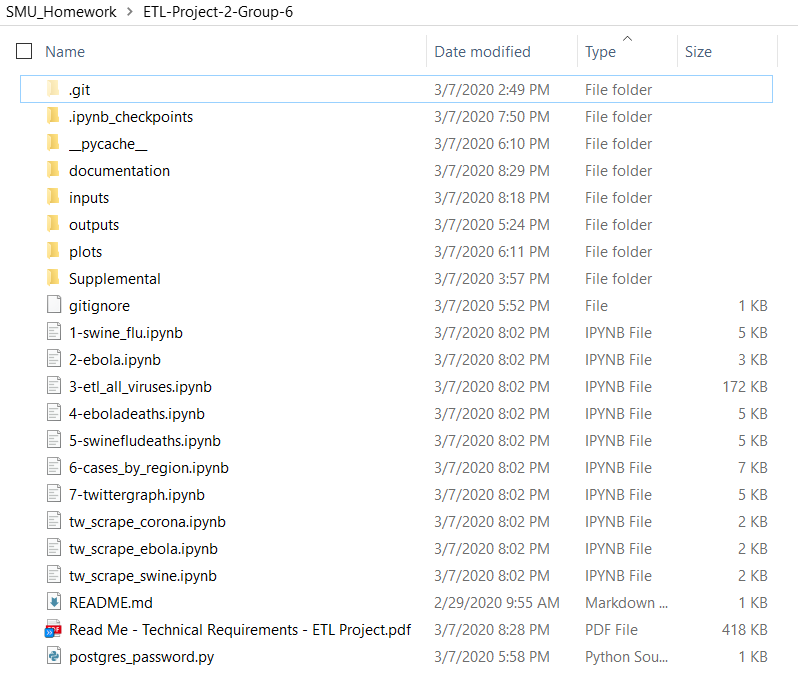
The process was to analyze the data elements available in each virus type dataset, cleanse the data have all data elements standardized into a final base table. Then we could then create aggregate tables for plotting and graphing as part of a bonus for the project.

# System Information

1. Development Tools
   1. Python
   2. Matplotlib (Seaborn)
   3. PostgreSQL
2. Data Sources
   1. Kaggle
   2. Worlddata.com
   3. Twitter
3. Data File Types
   1. Input
      1. .csv
   2. Output
      1. .csv
   3. Plots and Graphs
4. Data Transmission Type
   1. Download
   2. Scrape (Twitter – GetsOldTweets3)

# GitHUb FOlder Structure and rUN oRDER

* Below is the folder structure with the .ipynd files labeled in numerical order to run. Items 3 and 6 in the run process has the PostgreSQL DB connection The tw\_scrape\_(virus\_name) files do not need to be run as they take 20+- minutes each – unless you want to have them in the background. Please remember an error may come up ‘Too many request’ – if this happens just rerun the job or lower the count to 10000 tweets or less.



# Detailed Requirements

|  |  |  |  |
| --- | --- | --- | --- |
| Req. ID | Requirement Name | Priority | Description |
| 0 | Requirements Documents | 0 | Requirements Documents |
| 1 | Decide Project Topic | 1 | Comparison of Viruses |
| 2 | Decide on Virus Types | 2 | Coronavirus, Swine and Ebola |
| 3 | Analyze Data | 3 | Review various data sets and decide on which sets to use |
| 4 | **Extract** Source Data | 4 | Download or scrape data |
| 4a | Coronavirus | 5 | Download data |
| 4b | Swine | 6 | Download data |
| 4c | Ebola | 7 | Download data |
| 4d | Twitter | 8 | Scrape data |
| 5 | **Transform** Data – Corona | 9 |  |
| 5a | Remove Columns | 10 |  |
| 5b | Split Date Timestamp | 11 |  |
| 5c | Normalize | 12 |  |
| 5d | Modify data to Incremental | 13 |  |
| 6 | **Transform** Data – Swine | 14 |  |
| 6a | Remove Columns | 15 |  |
| 6b | Split Date Timestamp | 16 |  |
| 6c | Normalize | 17 |  |
| 6d | Modify data to Incremental | 18 |  |
| 7 | **Transform** Data – Ebola | 19 |  |
| 7a | Remove Columns | 20 |  |
| 7b | Split Date Timestamp | 21 |  |
| 7c | Normalize | 22 |  |
| 7d | Modify data to Incremental | 23 |  |
| 8 | **Transform** Data – Twitter | 24 |  |
| 8a | Add virus column to datasets | 25 | Add virus column to .csv data datetime and text. |
| 9 | **Create PostgreSQL** | 26 | Create DB and Table Structure |
| 10 | Create DB | 27 |  |
| 11 | Create Base Tbl | 28 |  |
| 12 | Create Aggregate Tbl | 29 |  |
| 13 | Create Twitter Tbl | 30 |  |
| 14 | Create Country Region Ref | 31 |  |
| 15 | Create ERD | 32 |  |
| 16 | **Load** Base Table | 33 |  |
| 17 | **Load** Aggregate Tables | 34 |  |
| 18 | Country – Weekly Table | 35 |  |
| 19 | Country – Monthly Table | 36 |  |
| 20 | Country – Year Table | 37 |  |
| 21 | **Load** Country Region Table | 38 | Load Country Region Reference Table |
| 22 | Create Plots and Graphs | 39 | Create Plots and Graphs |
| 23 | Country – Weekly Plot | 40 | Create line chart for top five countries. |
| 24 | Region – Cases Trend | 41 |  |
| 25 | Region – Deaths Trend | 42 |  |
| 26 | Region – Recovered Trend | 43 |  |
| 27 | Twitter Bar Chart | 44 | Create Bar Chart showing timeframe for the three viruses to accumulate 10000 tweets. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Images and Mockups

**Entity Relationship Diagram – ETL Segments**

