SMU - Data Science

ETL Project – Group 6

Requirements Document

Authors –

Stephanie Smith

Cenez Tindoc

Monica Moussavi

Henry Greyner

R. Shane Livingston

2020

SMU University, Dallas, Texas 75205

TABLE OF CONTENTS

I. 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.

II. SYSTEM INFORMATION

a. Development Tools

    a. Python

    b. Matplotlib (Seaborn)

    c. PostgreSQL

b. Data Sources

    a. Kaggle

    b. Worlddata.com

    c. Twitter

c. Data File Types

    a. Input

        i. .csv

    b. Output

        i. .csv

    c. Plots and Graphs

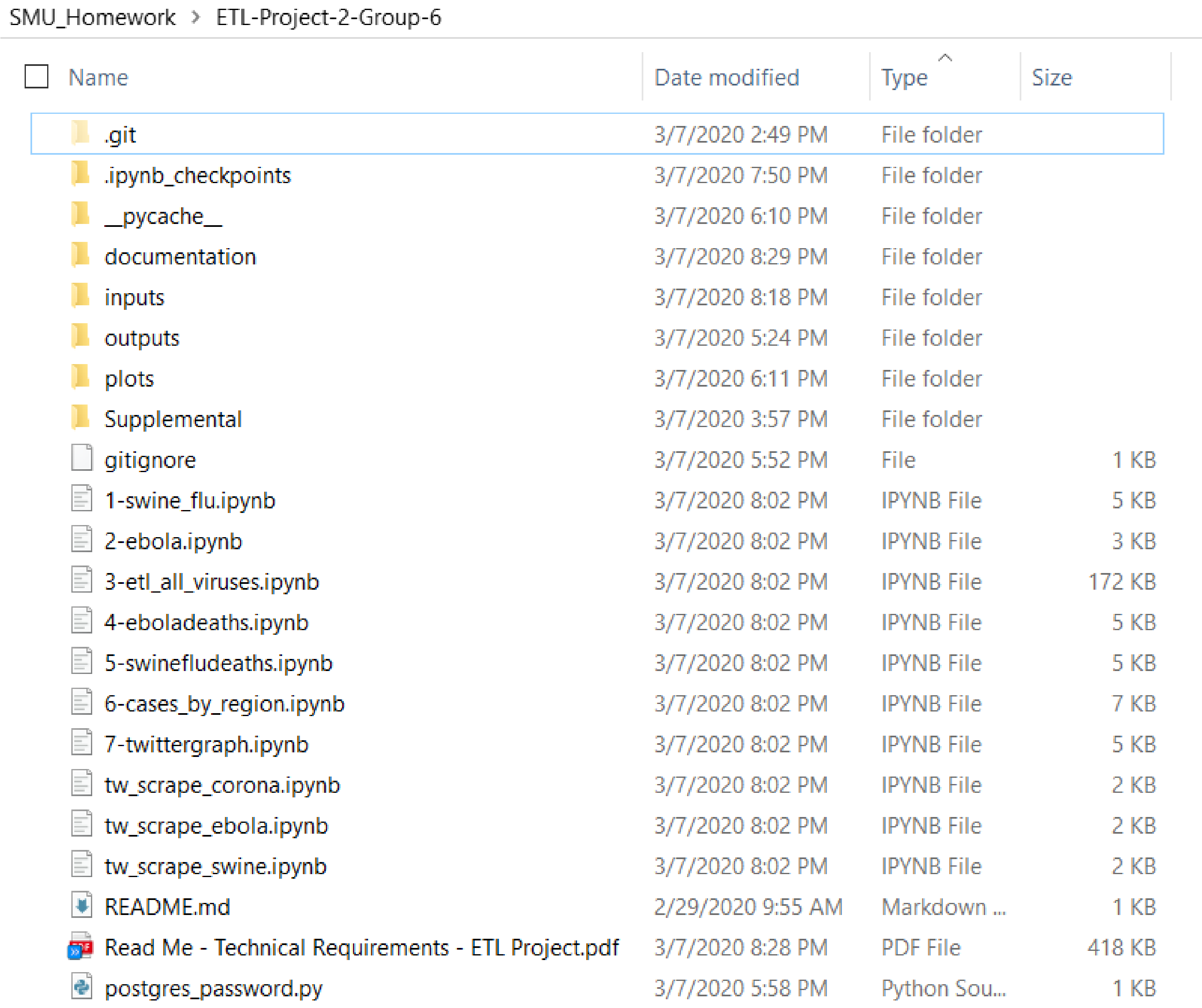
d. Data Transmission Type

    a. Download

    b. Scrape (Twitter – GetsOldTweets3)

III. 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.
-



SMU_Homework › ETL-Project-2-Group-6

☐ Name	Date modified	Type	Size
.git	3/7/2020 2:49 PM	File folder	
.ipynb_checkpoints	3/7/2020 7:50 PM	File folder	
__pycache__	3/7/2020 6:10 PM	File folder	
documentation	3/7/2020 8:29 PM	File folder	
inputs	3/7/2020 8:18 PM	File folder	
outputs	3/7/2020 5:24 PM	File folder	
plots	3/7/2020 6:11 PM	File folder	
Supplemental	3/7/2020 3:57 PM	File folder	
gitignore	3/7/2020 5:52 PM	File	1 KB
1-swine_flu.ipynb	3/7/2020 8:02 PM	IPYNB File	5 KB
2-ebola.ipynb	3/7/2020 8:02 PM	IPYNB File	3 KB
3-etl_all_viruses.ipynb	3/7/2020 8:02 PM	IPYNB File	172 KB
4-eboladeaths.ipynb	3/7/2020 8:02 PM	IPYNB File	5 KB
5-swinefludeaths.ipynb	3/7/2020 8:02 PM	IPYNB File	5 KB
6-cases_by_region.ipynb	3/7/2020 8:02 PM	IPYNB File	7 KB
7-twittergraph.ipynb	3/7/2020 8:02 PM	IPYNB File	5 KB
tw_scrape_corona.ipynb	3/7/2020 8:02 PM	IPYNB File	2 KB
tw_scrape_ebola.ipynb	3/7/2020 8:02 PM	IPYNB File	2 KB
tw_scrape_swine.ipynb	3/7/2020 8:02 PM	IPYNB File	2 KB
README.md	2/29/2020 9:55 AM	Markdown ...	1 KB
Read Me - Technical Requirements - ETL Project.pdf	3/7/2020 8:28 PM	PDF File	418 KB
postgres_password.py	3/7/2020 5:58 PM	Python Sou...	1 KB

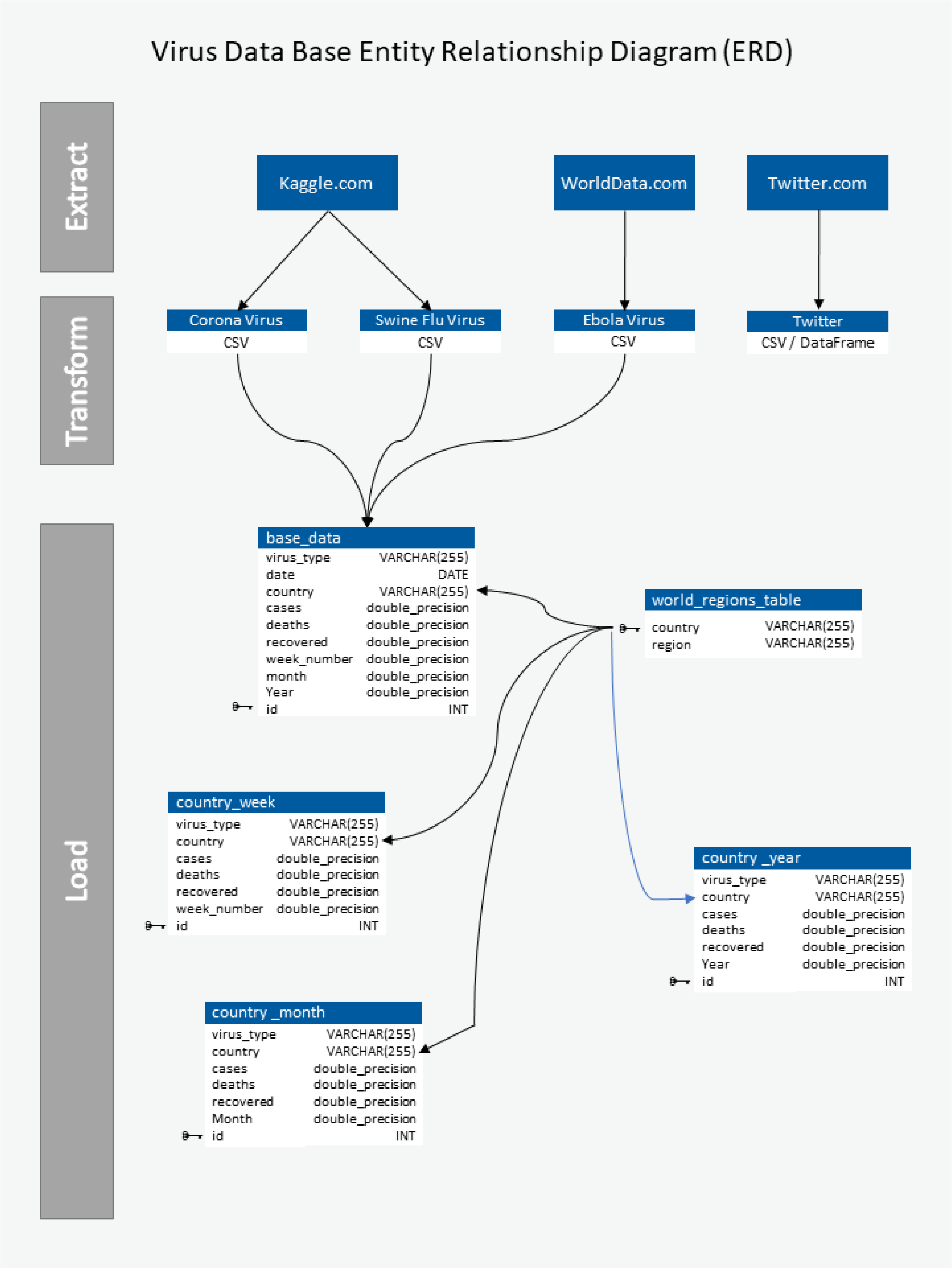
IV. DETAILED REQUIREMENTS

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.

V. IMAGES AND MOCKUPS

Entity Relationship Diagram – ETL Segments



Virus Data Base Entity Relationship Diagram (ERD)

base_data

virus_type	VARCHAR(255)
date	DATE
country	VARCHAR(255)
cases	double_precision
deaths	double_precision
recovered	double_precision
week_number	double_precision
month	double_precision
Year	double_precision
id	INT

world_regions_table

country	VARCHAR(255)
region	VARCHAR(255)

country_week

virus_type	VARCHAR(255)
country	VARCHAR(255)
cases	double_precision
deaths	double_precision
recovered	double_precision
week_number	double_precision
id	INT

country _year

virus_type	VARCHAR(255)
country	VARCHAR(255)
cases	double_precision
deaths	double_precision
recovered	double_precision
Year	double_precision
id	INT

country _month

virus_type	VARCHAR(255)
country	VARCHAR(255)
cases	double_precision
deaths	double_precision
recovered	double_precision
Month	double_precision
id	INT