Objective

The objective of this initiative was to extract market data (here, reflected as the Dow Jones Industrial Average closing price) as well as disease outbreak data for notable pandemics/epidemics over the past two decades, transform the data to enable downstream analyses, and load the transformed data into a target location for use.

Data Sources

This initiative utilized five [5] distinct data sources as noted within the table below. All data was available in CSV format.

|  |  |  |  |
| --- | --- | --- | --- |
| **Description of Dataset** | **Date Range** | **Data Source** | **Reference Link** |
| Dow Jones Industrial Average Data | 1/3/2000 to 3/3/2020 | Yahoo Finance | <https://finance.yahoo.com/quote/%5EDJI/> |
| SARS Outbreak Data (scraped from CDC) | 3/17/2003 to 7/11/2003 | Kaggle | <https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset> |
| Ebola Outbreak Data (scraped from CDC) | 8/29/2014 to 3/23/2016 | Kaggle | <https://www.kaggle.com/imdevskp/ebola-outbreak-20142016-complete-dataset> |
| H1N1 Outbreak Data (scraped from CDC) | 5/23/2009 to 7/6/2009 | Kaggle | <https://www.kaggle.com/de5d5fe61fcaa6ad7a66/pandemic-2009-h1n1-swine-flu-influenza-a-dataset> |
| COVID-19 Data (scraped from CDC) | 1/22/2020 to 3/2/2020 | Kaggle | <https://www.kaggle.com/imdevskp/corona-virus-report> |

Transformations Required and Performed via Python

This initiative utilized nine [9] distinct transformations prior to loading the data to its target location:

|  |  |
| --- | --- |
| **Description of Transformation** | **Module** |
| Index set/reset | Pandas |
| Column Renames | Pandas |
| Column Deletes | Pandas |
| Groupby (with sum() method) | Pandas |
| Datetime object conversions | Datetime |
| .loc date constraint | Pandas |
| Dataframe Outer Joins | Pandas |
| FillNA (forward fills) | Pandas |
| Percent Change Calculations | Pandas |

Final Production Database and Loaded Tables

This initiative resulted in four [4] tables being loaded to a PostgreSQL (relational) database.

|  |  |
| --- | --- |
| **Table Name** | **Description of Table** |
| sars\_market\_data | Table of datapoints within the range of dates for which SARS outbreak data was available (reference ‘Data Sources’ section above). Columns include:   * Date * Market (Price) * Deaths (Cumulative to-date) * Pct\_Chng\_Market (daily) * Pct\_Chng\_Death (daily) |
| h1n1\_market\_data | Table of datapoints within the range of dates for which H1N1 outbreak data was available (reference ‘Data Sources’ section above). Columns include:   * Date * Market (Price) * Deaths (Cumulative to-date) * Pct\_Chng\_Market (daily) * Pct\_Chng\_Death (daily) |
| ebola\_market\_data | Table of datapoints within the range of dates for which Ebola outbreak data was available (reference ‘Data Sources’ section above). Columns include:   * Date * Market (Price) * Deaths (Cumulative to-date) * Pct\_Chng\_Market (daily) * Pct\_Chng\_Death (daily) |
| covid\_market\_data | Table of datapoints within the range of dates for which Covid outbreak data was available (reference ‘Data Sources’ section above). Columns include:   * Date * Market (Price) * Deaths (Cumulative to-date) * Pct\_Chng\_Market (daily) * Pct\_Chng\_Death (daily) |