**Tracking the COVID19 Outbreak**

**ETL Project Report By: Prajakta Galvankar & Lujane Abdelgadir**

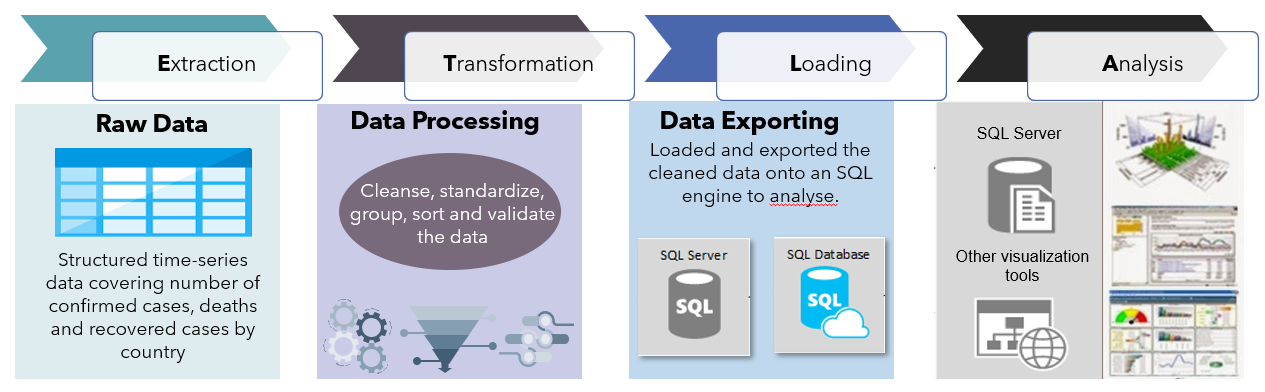
**Table of Content**

* Motivations & Summary
* ETL Process Summary
  + - **E**xtraction
    - **T**ransformation
    - **L**oading
* Analysis & Findings

Motivations & Summary:

ETL Process Summary:

The following diagram provides a summary of the ETL process to analyse the data:



**Extraction:**

* Selected time-series data covering number of confirmed cases, deaths and recovered cases by country
* The selected data is structured and in CSV format
* Used pandas/Jupitar Notebook to extract the data onto three separate pandas dataframes

**Transformation:**

* Removed unnecessary and inconsistent data:
* Checked column headers and rows to ensure that all countries and dates are consistent across the three databases
* A few countries were broken down by provinces/states
* Removed the provinces/states column because it had limited entries
* Grouped entries by aggregate value:
* Grouped the responses by sum per country so that there is one total value per day for each country for consistency
* Sorted the data in ascending order for ease of reference
* Added an index column for exporting to SQL
* **Loading:**
* Set up a database connection via SQL Alchemy and loaded the files to SQL for analysis
* We chose SQL because the data was structured, relational data with common columns between the files
* Also extracted the cleaned data into CSV for analysis on pandas