Table of Contents

[Introduction 1](#_Toc790384234)

[Statement of the Problem 1](#_Toc1255036610)

[Data Sources and Accessibility 2](#_Toc1250912288)

[Data Issues 2](#_Toc1549006390)

[Possible Data Architecture 4](#_Toc764543985)

[Possible Visualizations 4](#_Toc537165809)

# Introduction

Link to project: <https://github.com/microsoft/WhatTheHack/tree/master/038-MDWCovid19>

Caladan is a medium-sized Commonwealth country with a population of 3.2 million, and two urban centers – Duncan and Stillgard. Although the Coronavirus has not yet severely affected the Commonwealth, leaders are worried that a new wave will arrive as spring approaches. Therefore, the government of Caladan wants to develop a policy plan to mitigate the impact of the upcoming wave.

This project proposal discusses the problem to be addressed, data sources and accessibility, possible data issues that will have to be solved, a suggested data architecture for the project, and possible visualizations to derive and present insights from the project.

# Statement of the Problem

While Caladan hasn’t experienced severe impact from the Coronavirus thus far, its leaders are concerned about how the country might be affected by a possible resurgence as spring approaches. The leaders are, therefore, seeking to identify and implement the most unrestrictive policies that would limit the Coronavirus mortality growth rate to less than one percent and the growth rate of new cases to less than three percent on a 30-day average.

This project will address the officials’ concerns and help them achieve their goals by building a modern data pipeline solution that will not only extract, clean, load, and analyze Coronavirus metrics from five different countries, but also make it easier to generate reports that will provide the officials with better and updated insight on the Coronavirus situation. Using the insights obtained from the reports, the project will also entail performing calculations and an analysis to evaluate the relationship between the Coronavirus growth rate and policies enacted by different governments. That is, looking at how the implementation of various global policies affects the growth of Coronavirus across different countries. The project will conclude with the identification of the most effective policies that Caladan can implement to meet the targeted growth rate thresholds.

# Data Sources and Accessibility

The data that will be used during the project can be grouped into two parts – Coronavirus metrics, which are the case, death, and recovery metrics from five sample countries, and policy data that includes the global policies that have been implemented by ten countries. All the data is located in an Azure environment. Specifically, data for the Coronavirus metrics is located in the Azure SQL Database and Azure virtual machine, while the policy data is available in Cosmos Database (SQL API) via the Azure portal.

In order to access the data, an Azure subscription with owner access control is required. From there, the following credentials will be used to log in to the needed resources from the Azure environment

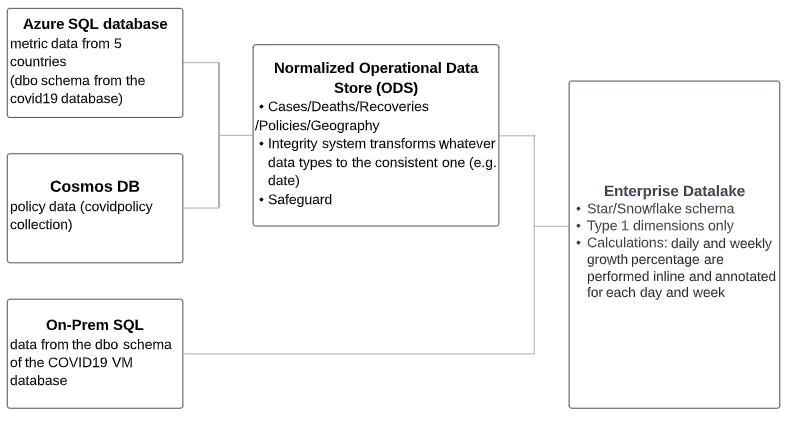
* **Azure virtual machine:**
  + Username: vmadmin
  + Password: Password.1!!
* **Azure SQL**
  + Username: sqladmin
  + Password: Password.1!!

# Data Issues

Given the problem, the type of data, and the goals Caladan officials wish to achieve to address the problem, there are a number of data issues that it may face and/or will face:

1. Data inconsistency: Since Caladan has access to various sources of COVID-19 data and from different geographies, the data formatting isn’t consistent across all sources, which makes it hard to process, analyze, and generate insights/reports.
2. Connected to the above-mentioned issue, having connections to various COVID-19 data sources makes it difficult to achieve data synchronization – consolidating data from multiple sources and achieving data harmonization.
3. Establishing connections to different COVID-19 data sources, as is this case for Caladan, makes it difficult to keep track of the data activity across all the data sources, resulting in the country’s inability to derive better and more accurate analysis of the data.
4. Dirty or unclean data: This can take various forms including:
   1. The COVID-19 metrics data from the five different countries might have missing values for one or more of the metrics.
   2. Data from different countries might come in different shapes. For example, “1/26/2023” might show as “1/26/2023”, “26/01/2023”, “2023/01/26” in different countries
5. Data fragmentation: some of the data might be fragmented due to limitation of the original resources. For example, data for certain countries might be divided into several regions; data on fatality rate might be divided into various age groups
6. Data security: Although the current data sources don’t contain sensitive information, it is expected to ingest sensitive information from Covid19 and other infectious diseases in the future. Additionally, strict access is expected to be given only to those who truly require it for business reasons.

# Possible Data Architecture



# Possible Visualizations

* Drop down menu: for daily/weekly data;
* Line chart: This can be used to show how the total number or average number of cases, deaths, or recoveries changed over time, making it easier to identify the trend and pattern in the data, and thus evaluate the effectiveness of the implemented global policies.
  + For a quick snapshot, the sparklines in Excel can also be used to see the trend in the data.
* Grouped bar charts: These can be used to show and compare the number of cases, deaths, or recoveries across the five different sample countries.
* Time series graphs for all the sample countries indicating the number of cases, deaths, and recoveries for each of the sample countries.
* Tableau map with circle size being determined by total number of cases, deaths, or recoveries to identify the country with the highest number of cases. Using the map to identify the countries with the highest/lowest numbers can make it easier to know which policies have been more effective than the others since we can then look at the kind of policies that the countries with the highest/lowest numbers have implemented.
  + Heatmaps can also be created to visualize the hotspots.
* Pie chart: The pie chart can be used to identify and compare the policies that were the most implemented across different countries.
* Dashboard: Various visualizations can be put together in a dashboard, making it easier to filter the visualizations based on certain conditions such as the number of deaths in a specific country.