Executive Summary

By: Andy Cheng, Rheona Mehta, Nikhita Singh, Josh Nahm

Table of Contents:

The Scope

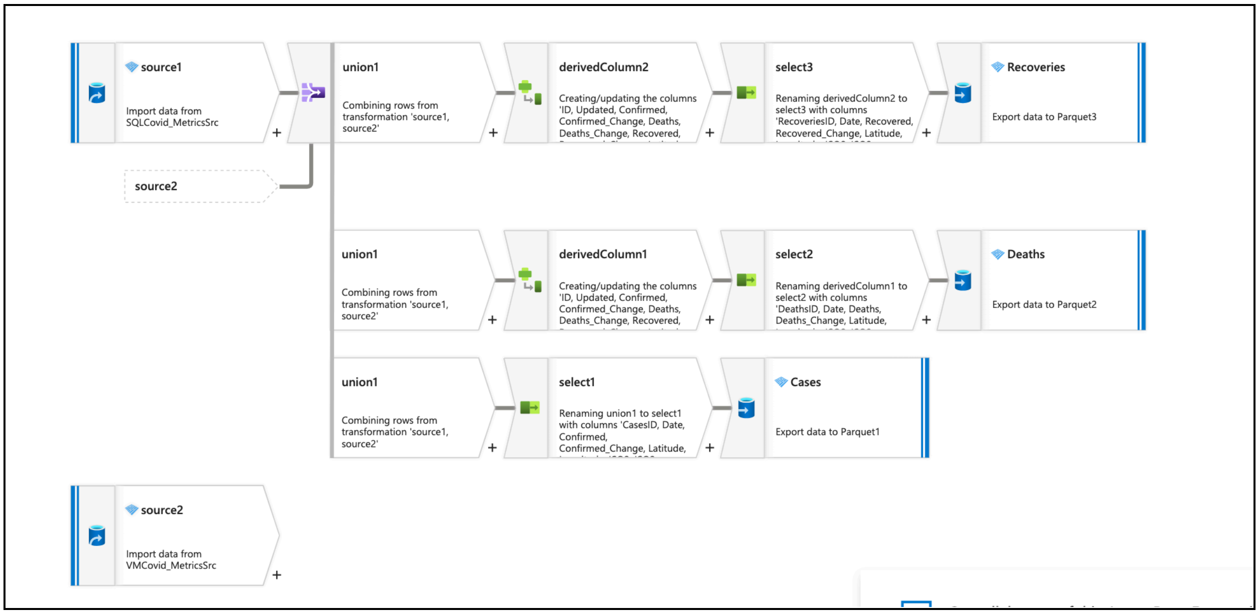
In this executive summary, we present the scope, methodology, and findings of our project for DS310. Our objective is clear: to identify the most effective, yet least restrictive policies that the government of Caladan can implement to keep the growth rates of COVID-19 deaths below 1% and new cases below 3% on a 30-day rolling average. The project progressed in four phases: Challenge 1, Challenge 2, Challenge 3, and Challenge 4.

Extracting and Loading Raw Data

Challenge 1 aimed to create a solution for extracting, cleaning, loading, and analyzing the raw data collected from various countries around the world. This data was dispersed across different sources. Policy data resided in a Cosmos DB (SQL API), while data for half of the countries, including cases, recoveries, and deaths, were stored on a Virtual Machine within our environment to simulate data retrieval from an on-premises source. Additionally, the remaining half of our data was located in Azure SQL DB. After extracting the data, we further organized it by creating folders for the collected data and storing each in Parquet file format.

Transforming Raw Data

With all the necessary data now securely stored in our Data Lake, our focus shifts to creating a usable Operational Data Store (ODS), which was done in Challenge 2. The aim was to transform the source data in the Data Lake into a more accessible and unified dataset, eliminating the need for downstream consumers to navigate between different data formats and sources. To meet the success criteria, we created a normalized ODS that combines metric and policy data, ensuring consistent data types and formats across all sources.



Loading Processed Data

In Challenge 3, we consolidated all the newly transformed data and loaded it into our big data processing engine: Azure Synapse Analytics. Once all the raw data were centralized and normalized data were available in the ODS, our focus shifted to designing a Data Warehouse. We accomplished this task by creating a Snowflake Schema. The Snowflake Schema we created had a fact table called 'Policies,' which was connected to many dimension tables such as 'Geographic,' 'Dates,' 'Cases,' 'Deaths,' and 'Recoveries.' We aimed to provide a generalized answer, but if we had more information about Caladan, we could have added further restrictions to the decomposition tree, such as including additional dimension tables in the 'Geographic' category, like 'Population,' 'Temperature,' 'Income,' and 'Area'.

Analyzing and Visualizing Data

Challenge 4, the final and most crucial phase, focused on analyzing and visualizing the gathered data to draw conclusions. To gather statistical evidence, we utilized Power BI to visualize various graphs, including the decomposition tree. This involved organizing policies based on their unique combinations of different policy levels. We achieved this by concatenating different columns to create a unique key for each policy combination used by countries. We then filtered them based on their success rates by percentage.

Narrowing It Down

For instance, 125 combinations of policies had 0 days where they were successful, meaning they did not meet the criteria of having a rolling average of cases growth rate under 3% and a rolling average of death growth rate under 1%. Conversely, 52 combinations of policies had a 100% success rate, and among those, 4 were active for 20 or more days. We chose this threshold as it provided a substantial amount of data to support our findings.

Additionally, since we were selecting a policy for implementation in Caladan over the spring, we ensured that the chosen policy was suitable for that time frame. Of the four, two were implemented during the spring, and we opted for the one with the lower stringency index to prioritize the least restrictive policies. Our research indicated that a combination of policies implemented by South Korea correlated with this criterion.

Our Recommendation

While examining this set of South Korea's policies, we identified the different levels of policies that contributed to their success. These included School Closing: 3, Workplace Closing: 2, Canceling Public Events: 1, Restriction on Gatherings: 0, Close Public Transport: 0, Stay-at-Home Requirements: 0, Restrictions on Internal Movement: 0, International Travel Controls: 3, Income Support: 1, Debt Contract Relief: 1, Public Information Campaigns: 2, Testing Policy: 3, Contact Tracing: 2, Facial Coverings: 1, and Vaccination Policy: 0, Protection of the Elderly People: 3.

Each level of policy has a value ranging from 0 to 3, representing its restrictiveness. Higher values indicate stricter measures, while lower values indicate less restrictive policies. For example, a level 3 policy, such as School Closing, mandates the closure of all schools. Conversely, a level 2 policy, like Workplace Closing, entails closing workplaces for certain sectors or categories of workers. Moreover, a level 1 policy, such as Canceling Public Events, recommends canceling such events, while a level 0 policy, like Restriction on Gatherings, implies no restrictions are necessary. Combining these policy levels, we identified a set of policies suitable for Caladan.

Work done:

The tasks were evenly distributed, but each person had their own area of expertise that they excelled in. These are the areas where we sparked.

Andy Cheng: ELT Process, External Data Research and Addition, Schema Design and Implementation, Power BI design (Visualization and Variable Creation), Power BI Report.

Rheona Mehta: ELT Process, Modeling Data (Data Model, Architecture Diagram), and presentation, Formatting of Executive Summary

Joshua Nahm: ELT Process, Executive Summary, and Deliverable Organization.

Nikhita Singh: ELT Process, External Data Research and Addition, Presentation, and Modeling Data (Data Model, Architecture Diagram)