**Dataset:**

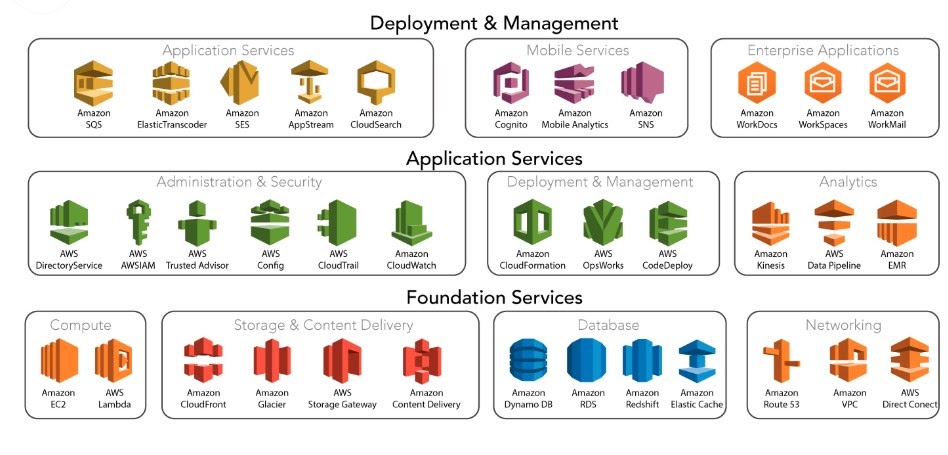
The year 2020 is marked with the red line in the history of modern human evolution on the earth. The reason behind that is the virus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is also called covid-19 as a short term. World health organization announced this as a new disease on 11th Feb 2020. It affected all the countries in the world. That’s why WHO announced it as a pandemic. At that time no medical science helped with the correct curable medicine for the disease. That’s why, within a month, almost all the countries went under lockdown for grappling with the disease. The repercussion of this disease is continuing throughout the world for more than 1 year. In our project, the dataset is taken from the Kaggle website.

All kinds of sectors are affected by this pandemic. In the process of defending against the disease, data analytics started playing a major role. For example, for determining the severity of the disease, data is collected from the patients and analyzed for getting better information about the symptoms and effects. The major symptoms, the ratio of severe cases, the ratio of deaths, ratio of recovering cases is determined with the help of data analytics. Applying data science algorithms, we can get good insights that help to make good decisions for defending the situation.

**Source:** <https://www.kaggle.com/kimjihoo/coronavirusdataset>

**What is AWS?**

Amazon Web Services is a cloud computing platform which was launched in 2006. Amazon Web Services includes storages, servers, networking, remote computing, email, mobile development, and security. AWS continues to add more tools as we progress into the future. Many companies use Amazon Web Services, some reputable companies include Sony and Samsung. AWS helps these business scale and grow by providing high functionality and delivery of their services.



**Data Loading and Cleaning**

* For our project, we have taken the dataset from Kaggle which we mentioned before.
* We applied Data Modelling Techniques to the raw dataset and made a data model.
* We prepared the data files according to the modelled diagram and uploaded the same to S3 bucket.
* We used the Glue Service in the AWS to make a Crawler for fetching the data from S3 bucket. After that, applied some techniques to clean the data like null values, etc.
* After Cleaning it we loaded it to Amazon Redshift.

**Data Processing Using ETL:**

* For our project we fetched the data with a crawler and stored it into our glue database that we made.
* The crawler fetched 12 tables from the S3 Bucket so we could perform ETL operations on it.
* We then made ETL Jobs for 12 tables in glue that we fetched from our crawler.
* Glue Jobs had Python generated script that applied mapping and dropping null values from all our tables.
* After all the 12 tables were processed, we used the crawler again to get the data from the S3 Bucket.
* Now all our data was processed inside our glue database so we could use it wherever we need it like in Athena and Redshift.

**Working with Glue:**

* AWS Glue is a ETL service which is serverless, glue is the most important service for our project as it helps us connect to all services to access the data in glue.
* Glue helps crawl data to store it into a glue database.
* It also helps build data catalogs and perform data preparation of big amounts of data.
* We can even transform data in glue by using its python generated scripts and built-in functionalities

**Athena and Analytics:**

* Amazon Athena is query interactive service to help analyze big data from S3 and Glue
* Athena uses standard Structured Query Language (SQL).
* We used Athena to perform to queries on our processed data in the Glue Database
* We can perform multiple queries on our data in Athena
* We gained many analytical insights from our data in Athena and used those insights to perform visualizations in Tableau

**Data Warehouse and Redshift:**

* Amazon Redshift is a Data Warehouse which can perform analytics on structured and semi-structured data
* This analysis can happen across data lakes, operational databases, and data warehouses
* Amazon Redshift has clusters which you can make, and each cluster holds one database
* We created a cluster and used the query editor in Redshift to create an external schema for our dataset
* We used that clusters API endpoint to connect Tableau to Redshift to perform visualizations on our processed data

**Data Visualization:**

This graph shows the information related to the number of confirmed cases and deceased cases of a particular age group.

Chart, bar chart

Description automatically generated

The below visual represents the number of confirmed cases in each province.

Chart, bubble chart

Description automatically generated

**Technical Difficulties:**

Initially, we faced many issues in modeling the dataset. The dataset contains many tables with redundancy. So, applied some techniques and modeled the data with some meaningful tables and relationships between them.

After modeling the dataset. We again prepared the data files according to the modeled diagram. After that, we started working on the AWS platform. We came across many issues while working on the AWS platform and those were resolved with the mutual help of the members of the group. One technical difficulty was that our lambda function wasn’t working, the crawler wouldn’t start everything we uploaded files into S3. Later we found that we hadn’t had the correct libraries installed and had some syntax issues, so after installing the libraries and fixing syntax it worked.

**Key Learnings:**

* Modeling the data of the real world is not that  easy and it is a very important step to maintain flexibility in the database, to make more meaningful information from the data.
* We learned to work on AWS using the services like S3, Glue, Redshift, Athena and solved many issues while working on them. While working on those issues we understood the working of those services more precisely.
* Working on Tableau is very easy, we learned how to connect the tableau with the AWS redshift and to get meaningful information from the Data warehouse. Worked on maintaining blogs and making videos for the project.

**Conclusion:**

The analysis of the Korean data set is useful to prevent the flow of cases by extracting the major factors that affect the number of cases. This analysis will be useful to understand the situation better and to make the strategies to prevent worse situations. From our analysis, according to the available data, people of age group 20s and older age groups are having more chance of getting covid 19 but people in the age group 20s are can easily recover whereas older age group people are hard to recover when compared to the 20s. In the pattern of the number of cases of covid 19, there is a sudden jump in the number and after that, it is slowly increasing the number. So, this type of analysis will be helpful for governments to prepare their strategies.