The dataset we used in the project is the COVID-19 epidemiological database, which is found in the BigQuery marketplace and then directly imported to the BigQuery platform for processing.

The dataset combines over 20000+ open, publicly sourced, licensed data and it includes many areas such as demographics, economy, epidemiology, geography, health, etc. Also, the data merge daily time series and use a consistent set of region keys where All regions are assigned a unique location key.

Here is a small piece example of the dataset, each column represents an attribute of the data, and we are going to use these attributes to retrieve the data we want.

To conclude, we can see that the Google BigQuery is a serverless and fully managed platform, it manages the underlying software for clients as well as the infrastructure which provides high scalability as well as the high availability.

Apart from that, it also provides extremely high cost-effectiveness and full-scan performance for data queries as it uniquely incorporates a massively parallel query engine (Dremel) that enables it to efficiently analyze large-scale datasets such as the COVID-19 datasets in our project.

In addition, BigQuery has native integrations with many data visualization tools such as Data Studio, Excel sheet, Power BI and so on. Which means that after processing the data, BigQuery can visualize and perform real-time data analysis of streaming data without to manually save and import the queried data itself.