### Covid-19 Analysis

### Data description

A centralised repository of up-to-date and curated datasets on or related to the spread and characteristics of the novel coronavirus (SARS-CoV-2) and its associated illness, COVID-19. Globally, there are several efforts underway to gather this data, and we are working with partners to make this crucial data freely available and keep it up-to-date. Hosted on the AWS cloud, we have seeded our curated data lake with COVID-19 case tracking data from Johns Hopkins and The New York Times, hospital bed availability from Definitive Healthcare, and over 45,000 research articles about COVID-19 and related coronaviruses from the Allen Institute for AI.

Note - Data is available here

<https://registry.opendata.aws/aws-covid19-lake/>

Data links

<https://covid19-lake.s3.us-east-2.amazonaws.com/enigma-jhu/>

<https://covid19-lake.s3.us-east-2.amazonaws.com/enigma-nytimes-data-in-usa/>

<https://covid19-lake.s3.us-east-2.amazonaws.com/rearc-covid-19-testing-data/>

<https://covid19-lake.s3.us-east-2.amazonaws.com/rearc-usa-hospital-beds/>

<https://covid19-lake.s3.us-east-2.amazonaws.com/static-datasets/>

# Problem Statement

As a data engineer you are supposed to do following

* Write an spark pipeline either local or emr to clean the source data and upload the cleaned data into an staging dir on s3
* Create glue tables in the glue catalog on cleaned data.
* Use athena for preview and validate the created tables.
* Draw an UML diagram for all the entities, You can use <https://app.diagrams.net/> for drawing.
* Write pyspark EMR pipeline with jupyter notebook to convert the relational model to dimensional model and publish the output for dimensions and facts into redshift tables for analysis.

# Dimensional Model

