## ****Dataset Summary – AWS COVID-19 Data Lake****

As part of our AWS-based data engineering project, we will utilize a curated **public COVID-19 data lake** hosted by AWS. This data lake provides a centralized repository of continuously updated datasets related to the global COVID-19 pandemic, enabling large-scale analytics and data warehousing solutions.

### **Data Exploration via AWS Glue Data Catalog**

We will explore and integrate data using the **AWS Glue Data Catalog**, which offers schema definitions and metadata for seamless access across AWS analytics services such as Amazon Athena and Amazon Redshift Spectrum.

## ****Data Sources and Structure****

This public data lake includes contributions from renowned institutions such as **Johns Hopkins University**, **The New York Times**, **Definitive Healthcare**, and the **Allen Institute for AI**. It includes datasets for confirmed cases, testing, hospital bed capacity, and supporting lookup data.

### **Primary Data Tables (Total: 10)**

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Description | Source | Provider |
| enigma\_jhu | Global confirmed COVID-19 cases by province/state/county | Johns Hopkins University | Enigma |
| nytimes\_states | COVID-19 confirmed cases & deaths at the U.S. state level | The New York Times | Rearc |
| nytimes\_counties | COVID-19 confirmed cases & deaths at the U.S. county level | The New York Times | Rearc |
| covid\_testing\_states\_daily | Daily COVID-19 testing stats by U.S. state | COVID Tracking Project | Rearc |
| covid\_testing\_us\_daily | Daily aggregated U.S. COVID-19 testing data | COVID Tracking Project | Rearc |
| covid\_testing\_us\_total | Cumulative U.S. COVID-19 testing totals | COVID Tracking Project | Rearc |
| hospital\_beds | Hospital bed availability and utilization across the U.S. | Definitive Healthcare | Rearc |
| country\_codes | Lookup table for international country codes | - | - |
| county\_populations | U.S. county-level population data based on census figures | - | - |
| us\_state\_abbreviations | Lookup for U.S. state abbreviations | - | - |

## ****Data Storage Location****

The raw data files are hosted in a public Amazon S3 bucket accessible via:  
[**https://dj2taa9i652rf.cloudfront.net/**](https://dj2taa9i652rf.cloudfront.net/)

**Key directories used:**

* enigma-jhud/
* enigma-nytimes-data-in-usa/
* rearc-covid-19-testing-data/
* rearc-usa-hospital-beds/
* static-datasets/

## ****Data Documentation & Metadata****

Comprehensive metadata and dataset documentation can be found on GitHub:  
 [**AWS Labs Open Data Registry**](https://github.com/awslabs/open-data-registry/blob/main/datasets/aws-covid19-lake.yaml)

**AWS Blog Post:**  
[**A Public Data Lake for Analysis of COVID-19 Data**](https://aws.amazon.com/blogs/big-data/a-public-data-lake-for-analysis-of-covid-19-data/)