**Healthcare Case study**

****

| **Data sources:** Flat files , RDB , S3 ,JSON,parquet  Scheduler : Task  Streaming : Streams / Snowpipes  **Visualization :** PowerBI  **Language :** Javascript (Stored procedure) |
| --- |

**Problem Statement :**

Recent Covid-19 Pandemic has raised alarms over one of the most overlooked areas to focus: Healthcare Management. While healthcare management has various use cases for using data science, patient length of stay is one critical parameter to observe and predict if one wants to improve the efficiency of the healthcare management in a hospital.

This parameter helps hospitals to identify patients of high LOS risk (patients who will stay longer) at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimized to minimize LOS and lower the chance of staff/visitor infection. Also, prior knowledge of LOS can aid in logistics such as room and bed allocation planning.

Suppose you have been hired as Data Scientist of HealthMan – a not for profit organization dedicated to manage the functioning of Hospitals in a professional and optimal manner.

As a snowflake developer/ consultant, you are helping the data analytical team to explore the dataset using snowflake , history data is loaded and current data is loaded with snowpipes into database , Merge all the datasets based on the common key and create a single table and Check for noise or missing data.

The task is to accurately predict the Length of Stay for each patient on a case by case basis so that the Hospitals can use this information for optimal resource allocation and better functioning. The length of stay is divided into 11 different classes ranging from 0-10 days to more than 100 days. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University

#### Datasets :

<https://github.com/akgeoinsys/COVID-19>

**Tasks :**

1. **Create database covid\_db and schema schema**
2. **Create a covid\_history table to load all the historical datasets and apply cluster key on the most used column for querying.**
3. **Create internal stage and load every 3 months data which has different schema**
4. **Perform basic transformation (cast, change date format) during data load**
5. **Upload all current data into AWS s3 bucket snow\_extstage and perform load**
6. **Load the data from external stage**
7. **Perform bad records filtering on loading ,file format and compression**
8. **Create snowpipes to continuously load new US covid data from external stage from s3 bucket to the staging table ,maintain the change data capture and merge the data to the consumption table**
9. **Share your table to new non-snowflake user**
10. **create clone of the table with time travel before one day and write query to get history data using particular timestamp**
11. **Create stored procedure to insert the data into table after typecasting date format**
12. **Create a visualization for patient history analysis using powerBI**

**And prepare reports /dashboard for business queries**

**Business Questions :**

What is the daily number of confirmed cases?

Which countries with the highest number of confirmed cases have the most per capita?

where are confirmed cases increasing most rapidly?

More questions will be added during project phase

Reference data :

<https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>