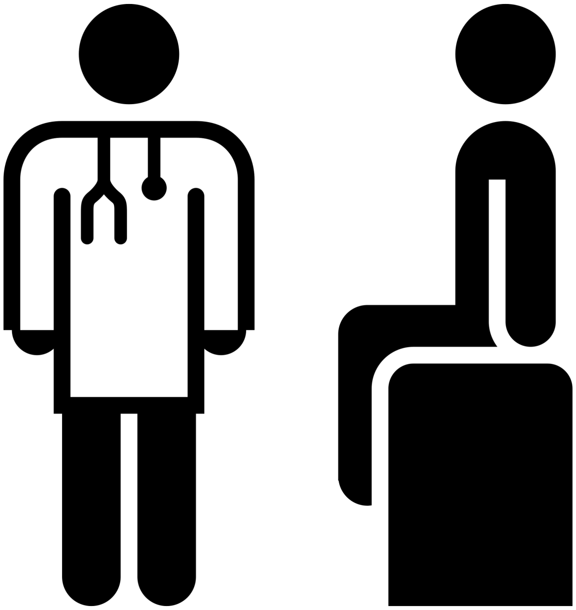
Healthcare Data Analysis



You've been provided access to a selected healthcare dataset, chosen for its relevance to public health and minimal preprocessing needs. Your challenge is to navigate through this dataset, utilizing data science techniques and AWS technologies to bring to light patterns and predictions that could shape future healthcare strategies.

For your data set, you are free to source the data from **Kaggle** or the **CDC** website. Make sure your chosen data set includes ***at least 2 columns of numerical data*** and ***at least 3,000 rows*** so that you can make meaningful statistical analysis. Some possible options from these websites are:

CDC Outpatient Respiratory Illness Activity:
https://data.cdc.gov/Public-Health-Surveillance/Outpatient-Respiratory-Illness-Activity-Map/6svj-q4zv/about_data

CDC Flu Vaccination provider locations information:
https://data.cdc.gov/Flu-Vaccinations/Vaccines-gov-Flu-vaccinating-provider-locations/bugr-bbfr/about_data

Cardiovascular disease dataset:
https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Functionality:

You will be expected to create the beginning of a data pipeline in a cloud environment. This will involve the use of S3 buckets, Lambda functions, RDS, and potentially an EC2. Your team will be using an S3 bucket as your initial endpoint for all data in this project. A lambda function will be needed to clean the data (if necessary) and store it in a MySQL RDS. Then you should connect to the RDS locally to generate at least 10 different meaningful insights from your cleaned data. This will be the minimum requirement. You may feel free to go beyond the bounds of this outline and improve the final product to the best of your abilities. Some suggestions on improvements have been listed below as extensions

- ➢ ***Checkpoint 1:*** Select your data set and set up your S3 bucket. Make sure that all members can upload to the s3 bucket programmatically using the boto3 library
- ➢ ***Checkpoint 2:*** Set up an RDS in a public VPC so that you and your team members can all access the RDS
- ➢ ***Checkpoint 3:*** Create a lambda function that will read files on upload, clean the data, and place it in the RDS
- ➢ ***Checkpoint 4:*** On your local machine, pull data from your RDS to create at least 5 meaningful insights on the data. These insights should be statistical observations worth highlighting in the data

NOTE: *There is a way for every team member to connect to the same AWS resources. Make sure you are only constructing a single Architecture*

Extensions

- ➢ ***Option 1:*** Set up an EC2 which will perform your statistical analysis (Checkpoint 4) and display it on a webpage. The webpage can be simple and contain no CSS. If you are working in VS code, open a new .html file and click {SHIFT + 1 + ENTER}. This will auto populate a document with the most basic HTML script. There are several websites that can guide you on hosting a static website on an EC2
- ➢ ***Option 2:*** Set up an EC2 with a python script that will perform some machine learning algorithm. Your goal is to see if you can find some meaningful insight using an appropriate algorithm (such as a classification algorithm or regression algorithm). Use the scikit-learn library to source the algorithms from: https://scikit-learn.org/stable/. Make sure to store any insights (such as tables

showing statistical norms/anomalies or any visualizations) in an s3 bucket

- ➢ ***Option 3:*** Any method you would like to add that will improve the overall comprehension of your analysis or deepen it. This may include visualizations, automation, or organization of work.

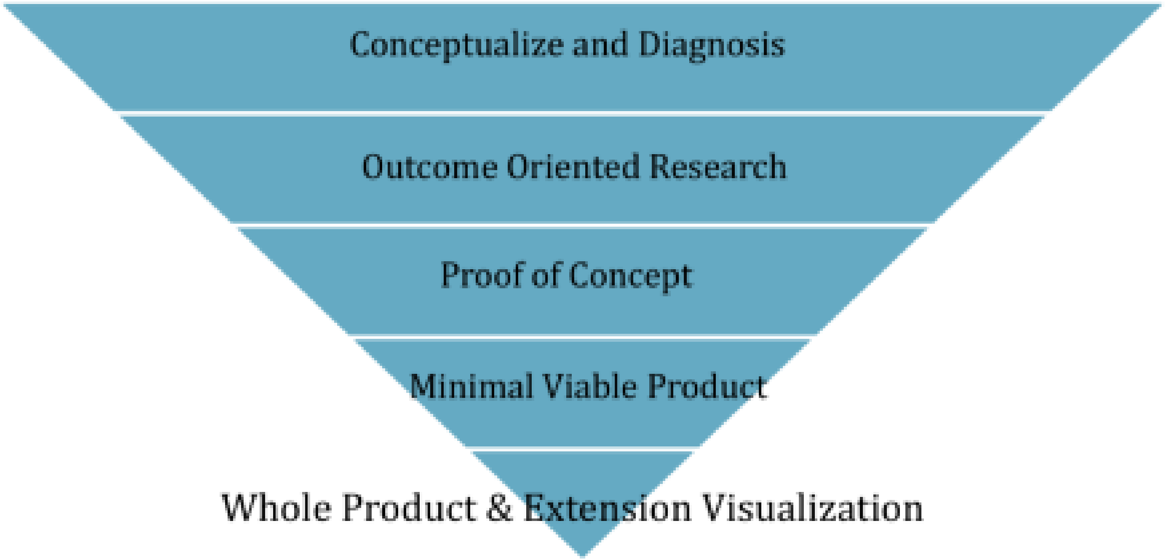
Requirements:

- Product Requirements Document (due end of Day 1)
- Python Scripts lambda function
- Database Schema for RDS (ER Diagram)
- Presentation Slides (due end of lunch Thursday)
- Jira Board (Should be shown and covered in presentation)

Considerations when Developing a PRD

A product requirements document (PRD) is an artifact used in the product development process to communicate what capabilities must be included in a product release to the development and testing teams.

The PRD will contain everything that must be included in a release to be considered complete, serving as a guide for subsequent documents in the release process. While PRDs may hint at a potential implementation to illustrate a use case, they may not dictate a specific implementation. The process diagram below showcases the steps considered when developing a PRD.



Conceptualize and Diagnosis

Outcome Oriented Research

Proof of Concept

Minimal Viable Product

Whole Product & Extension Visualization