Title

- Healthcare Data Analysis - PEP Project 2
- By Team 2

Change history

- Document created 2/5/2024

Overview

The purpose of our product is to establish a streamlined data pipeline in a cloud environment, addressing challenges in data management and analysis. The solution caters to data engineers, analysts, and stakeholders seeking efficient data processing and insightful analytics. By integrating Amazon S3 buckets, Lambda functions, RDS, and potentially EC2, we offer a comprehensive platform for data cleaning, storage and analysis. This is crucial for organizations aiming to derive actionable insights from their data assets effectively, enabling informed decision-making and fostering innovation.

Dataset

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

Roles

Chi Asangwe (Scrum Master/Developer)
Alejandro Otaola (Developer)
Gabrielle Glasgow (Developer)

Objectives

Checkpoint 1: Create S3 bucket
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

- Select the dataset, we will be utilizing the cardiovascular disease dataset.
- Create an S3 Bucket with a name which clearly describes its data contents.
- Setup boto3 library and see if any connectivity steps are required to connect to s3 bucket.
    - Go through those steps if needed.
- Create access keys for team members to edit buckets and upload data points.
- Troubleshoot as needed.

Checkpoint 2: Create RDS

Set up an RDS in a public VPC allowing you and your team members can all access the RDS

- Set up a VPC on AWS
- Create an RDS the VPC
- Provide access to the RDS
    - Ensure that teammates IP's can access the RDS
    - Provide the rds endpoints, username, and password

Checkpoint 3: Create a lambda function

Create a lambda function that will read files on upload, clean the data, and place it in the RDS)

- Go to lamba and create a new function.
- Add a trigger which acts on the uploaded data.
- Add a script which cleans the data (gets rid of null values, duplicates, and creates data normalization).
- Push clean data onto project RDS.

Checkpoint 4: Local Data Analysis

To uncover patterns, correlations, and potential risk factors related to cardiovascular disease we will use Pandas and Matplotlib to delve into the data and extract meaningful insights.

- **Impact of Lifestyle:**
    - We will create a bar plot to illustrate the proportion of individuals with and without cardiovascular issues based on their lifestyle choices including smoking, alcohol intake and physical activity.
- **Blood Pressure Analysis:**
    - To assess the relationship between blood pressure levels and cardiovascular issues we can generate a box plot to compare systolic and diastolic blood pressure distributions.
- **Age Distribution:**
    - For a deeper understanding of the age demographics of the patients in the study we will use a histogram for visualization.
- **BMI Calculation:**
    - Using the provided height and weight data, we will calculate the BMI for each patient. Plotting the BMI distribution will help identify any trends between BMI and the likelihood of cardiovascular disease.
- **Gender Disparities:**
    - To potentially uncover any gender disparities or differences in risk factors for cardiovascular disease, we will calculate the percentage among male and females.

Extension: EC2 Machine Learning Analysis

- We will set up an EC2 instance with a Python script to execute machine learning algorithms sourced from scikit-learn.
- Store generated results or visualizations in a S3 bucket for further analysis. This will enhance our analytical capabilities, enabling deeper exploration of the dataset.

Success Metrics

- No bugs
- Successful infrastructure created (AWS)
- Reliable display of information
- Accurate data analysis

Timeline and Release Planning

- 4 days for MVP to be ready for demonstration

Scenarios

- **Public Health Campaign:**
  - A public health organization analyzes lifestyle choices 'impact on cardiovascular health, shaping targeted interventions in their awareness campaign.
- **Clinical Decision Support:**
  - A clinic uses the dataset to correlate blood pressure with cardiovascular disease in a clinical decision support tool.
- **Geriatric Care Program:**
  - A facility tailors intervention for elderly cardiovascular health based on age demographics from dataset.
- **Wellness Program:**
  - A corporate wellness program adjusts initiatives based on BMI insights from the dataset to mitigate risk.
- **Gender Specific Health Initiative:**
  - Company develops gender-sensitive strategies to reduce cardiovascular disease, guided by dataset analysis of gender disparities in risk factors

Open Issues

- None

Q&A

- Feel free to ask any questions.