NEW YORK INSTITUTE OF TECHNOLOGY

Big Data Project Summary Report

Team Members

Dhruvil Patel (1324368)

Karan Patel (1213302)

Yueyuan Xu(1319488)

Professor: Liangwen Wu

- **Project Goal:** To what extend can survey responses be used for predicting heart disease risk?
  Can a subset of questions be used for preventative health screening for diseases like heart disease?

- **Data Lake vs Data Warehouse:**

  Schema-on-Read Flexibility: Data lakes allow you to store raw, unstructured, or semi-structured data without enforcing a schema upfront.

  Scalability: Data lakes, especially on cloud platforms like AWS, offer virtually unlimited scalability.

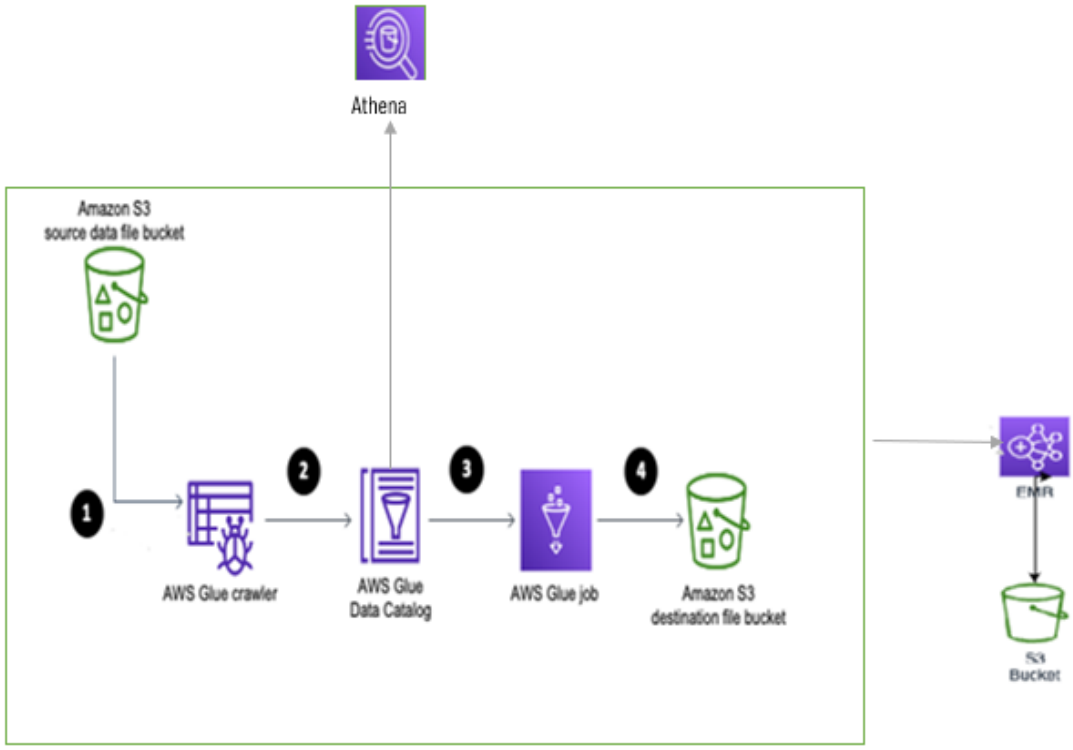
  Cost-Effective Storage: AWS, for example, provides services like Amazon S3, which is designed for scalable and durable storage at a lower cost compared to traditional data warehousing solutions.

  Support for Big Data Processing: Data lakes are well-suited for big data processing frameworks like Apache Spark

  Integration with Analytical Tools: AWS provides a range of analytical tools that can directly query and analyze data stored in a data lake, such as Amazon Athena, and AWS Glue. This integration simplifies the process of deriving insights from your dataset.

- **Dataset Link:** https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/discussion/284985

- **Data Pipeline**



Data Pipeline Overview:
This breakdown elucidates a data pipeline carefully architected for robust data management and analysis within the AWS ecosystem.

Source Data file bucket (finalproinput):
Commencing our journey, the raw data finds residence in finalproinput, an S3 repository serving as the primary source for our data endeavors.

Data Crawling (AWS Glue Crawler - myCrawler):
The automated myCrawler, part of AWS Glue, diligently examines finalproinput(S3 bucket). It discerns metadata intricacies and catalogs the underlying data structure.

Metadata Repository (AWS Data Catalog - DataCatalog):

The gleaned metadata assumes a pivotal role within the AWS Data Catalog, known as DataCatalog, a centralized repository ensuring meticulous metadata management.

ETL Processing (AWS Glue Job - nameee):
The transformative journey begins with nameee(etl job), an AWS Glue Job. It meticulously transforms raw data using predefined rules, depositing the processed data into a distinct S3 output bucket named outputfinal (s3 Bucket). This sets the stage for advanced analytics.

Intermediate Storage (outputfinal):
Outputfinal(s3 bucket) serves as an interim abode for the transformed data. It functions as a temporary storage nexus, facilitating a seamless transition to subsequent stages in the pipeline.

Big Data Processing (EMR - Cluster):
The introduction of Amazon EMR, configured to execute a Spark application, enhances the pipeline's efficiency through distributed and parallel processing.

Final Output Storage (creteee):
Culminating in FinalDataS3Bucket, the results from the Spark application encapsulate refined and analyzed data. This final repository marks the conclusion of the pipeline, presenting data in a state ready for consumption or downstream processes.

This methodically designed approach ensures a balanced amalgamation of efficiency and precision in handling complex data workflows.

- **Workflow Explanation: Exploratory Data Analysis and Visualization**

1. **Data Ingestion:** The process starts with reading a CSV file from an S3 bucket using Apache Spark for distributed data processing. Two Python scripts (spark.py and sql_visualization.py) are used for data analysis and visualization.

2. **Spark Analysis and Count Plots (spark.py):** The Spark session is initiated to read and process the data. The PySpark DataFrame is converted to a Pandas DataFrame for easier visualization. Count plots are created to explore relationships between different features and the target variable (heart disease or attack). The count plots are dynamically arranged based on the number of columns, and each subplot is saved to a BytesIO buffer.

3. **Upload Count Plots to S3 (spark.py):** The count plot images are saved in a buffer and uploaded to an S3 bucket (cretee). Two count plot images (image1.png and image2.png) are stored in the "store" folder within the bucket.

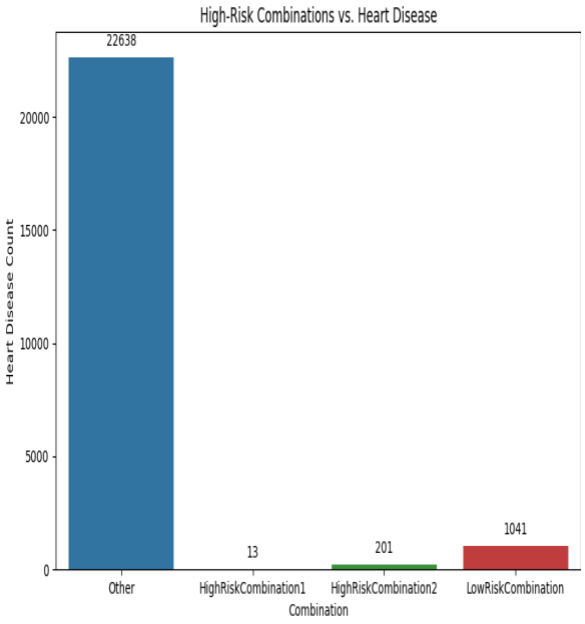
4. **SQL Analysis and Bar Plots (sql_visualization.py):** Another Spark session is initiated to perform SQL queries on the same dataset. Two SQL queries are executed to analyze specific combinations related to heart disease risk and healthy lifestyle. Bar plots are created to visualize the results of the SQL queries, and count numbers are annotated on each bar. The visualizations are saved to BytesIO buffers.

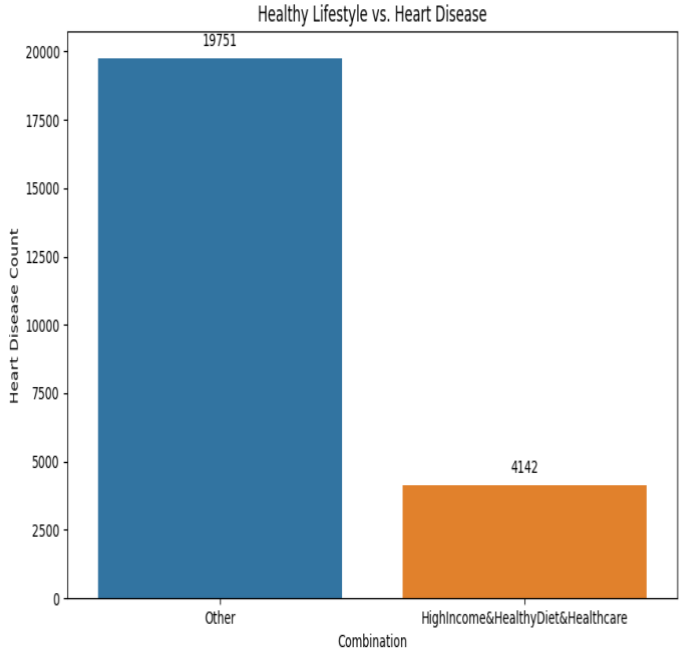
5. **Upload Bar Plots to S3 (sql_visualization.py):** The bar plot images for high-risk combinations, healthy lifestyle, and gender combos are uploaded to the same S3 bucket (cretee) in the "store" folder.

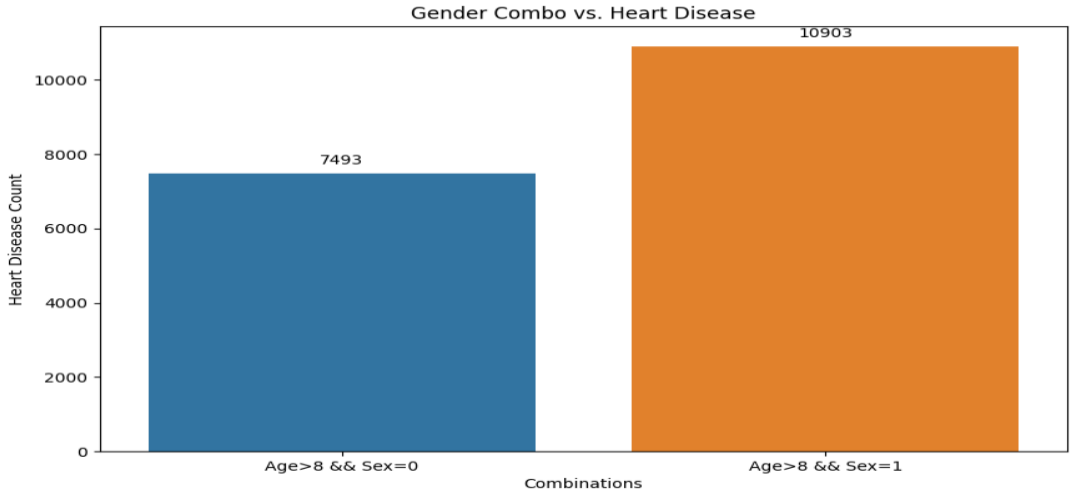
6. **Conclusion and Report Generation:** The generated visualizations provide insights into high-risk combinations, healthy lifestyle factors,

and gender-related heart disease patterns. These visualizations can be included in a comprehensive report to communicate findings and support decision-making processes. Users can access the S3 bucket to retrieve and share the visualizations with relevant stakeholders.

Workflow Completion: The Spark sesions are stopped to release







- **End Summary/Context:** In summary, the project successfully explored the use of survey responses to predict heart disease risk. Utilizing a Kaggle dataset and an AWS-based data pipeline, we demonstrated the flexibility and scalability of a Schema-on-Read approach. The EDA and visualization workflow, powered by Apache Spark, revealed insights into high-risk combinations, healthy lifestyle factors, and gender-related patterns associated with heart disease.

- The visualizations stored in the S3 bucket provide a valuable resource for informed decision-making and discussions on preventative health measures. This project showcases the potential of survey data in predictive modeling for heart disease risk, emphasizing the importance of well-architected data pipelines and analytical tools in extracting meaningful insights from health-related datasets.