E6760 DATA WAREHOUSING AND INTEGRATION

Behavioral Risk Factor Surveillance System: Leveraging Data for Public Health Insights
Project Report

Group 8

Student 1: Yash Bhadreshwara
Student 2: Siddhant Chavan

Percentage of Effort Contributed by Student 1: 50%
Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Yash Bhadreshwara
Signature of Student 2: Siddhant Chavan

- CONTENT

1. Summary

The Behavioral Risk Factor Surveillance System (BRFSS) is one of the most comprehensive public health surveillance programs globally, offering invaluable data on health-related risk behaviors. Despite its significance, the sheer volume and complexity of BRFSS data often present challenges for efficient processing and analysis. To address these challenges, this project leveraged AWS Cloud Services to design and implement a scalable, efficient, and data processing framework.

The project utilized a range of AWS services to create an ETL pipeline capable of handling massive datasets with ease. AWS S3 served as the central data storage layer, ensuring secure and scalable storage for raw BRFSS datasets. The AWS Glue service, combined with Crawlers, facilitated the automation of schema discovery and data cataloging, streamlining the transformation of raw data into a structured format. The transformed data was queried and analyzed in real-time using AWS Athena, enabling seamless interaction with the dataset.

A data warehouse was developed to store and manage the multidimensional BRFSS dataset effectively. The framework supports real-time querying and advanced analysis, leveraging Online Analytical Processing (OLAP)systems. The implementation also included mechanisms to ensure data consistency, historical tracking, and dynamic schema updates.

By leveraging the AWS ecosystem, this project successfully transformed the BRFSS data into an accessible, reliable, and actionable resource for public health stakeholders. Outcomes include improved data integration, reduced redundancies, enhanced analytics, and actionable insights—enabling decision-makers to devise strategies that address pressing public health issues effectively.

This cloud-based solution showcases the potential of modern technology in overcoming data challenges, offering a framework that bridges the gap between raw data and impactful health interventions.

2. Background

The BRFSS collects data through surveys to monitor health-related risk behaviors, chronic health conditions, and the use of preventive services. With over 400,000 interviews conducted annually, BRFSS is the largest continuously conducted health survey system. However, utilizing this wealth of data for actionable insights is challenging due to its size, complexity, and variability. This project addresses these challenges by creating a comprehensive data warehouse and OLAP system, transforming how health data is stored, managed, and analyzed.

The system enables efficient integration of diverse datasets, ensuring consistency and accessibility across various domains of public health. By implementing advanced ETL processes, it resolves common issues such as missing values and inconsistent formats, making the data more reliable for decision-making. Through this framework, public health officials can identify patterns in risk behaviors, compare state and national trends, and tailor interventions to address specific health challenges. Ultimately, the project bridges the gap between raw data and actionable insights, fostering data-driven strategies to improve health outcomes.

3. Problem Definition:

Despite its importance, the BRFSS dataset is notoriously difficult to work with due to several limitations. The primary issues include data fragmentation across multiple sources, inconsistency in formats, and a lack of historical tracking mechanisms. These challenges impede the ability of analysts to derive meaningful insights. For instance, monitoring trends such as mental health status or vaccination coverage over time requires robust data structures and analytical tools. Without such systems, identifying and addressing public health priorities becomes a daunting task. This project aims to resolve these issues by building an integrated data warehouse and analytical framework.

Along with the creation of a robust ETL process that extracts raw data on nutrition, physical activity, and obesity from the BRFSS dataset, cleans and transforms the data into an analyzable format, and loads it into a structured database for further analysis. The project will utilize tools and techniques like Python and MS Excel for data processing, AWS Cloud Services for database management, and visualizations on tools such as Tableau for presenting insights. The goal is to create a structured, accessible, and user-friendly data pipeline that allows public health stakeholders to easily query and visualize trends in mental health status and covid vacc status across various demographics and geographic regions.

4. Objectives

Data Extraction and Ingestion: Extracting raw data from the BRFSS dataset, focusing on health-related information such as Mental Health, Covid Vacc Status. Ensure consistency and reliable data import from national and state sources.

Data Cleaning and Transformation: Cleaning and standardizing the data by handling missing values and inconsistencies. Transforming the raw data into a structured format, including generating new calculated fields like Employment Status.

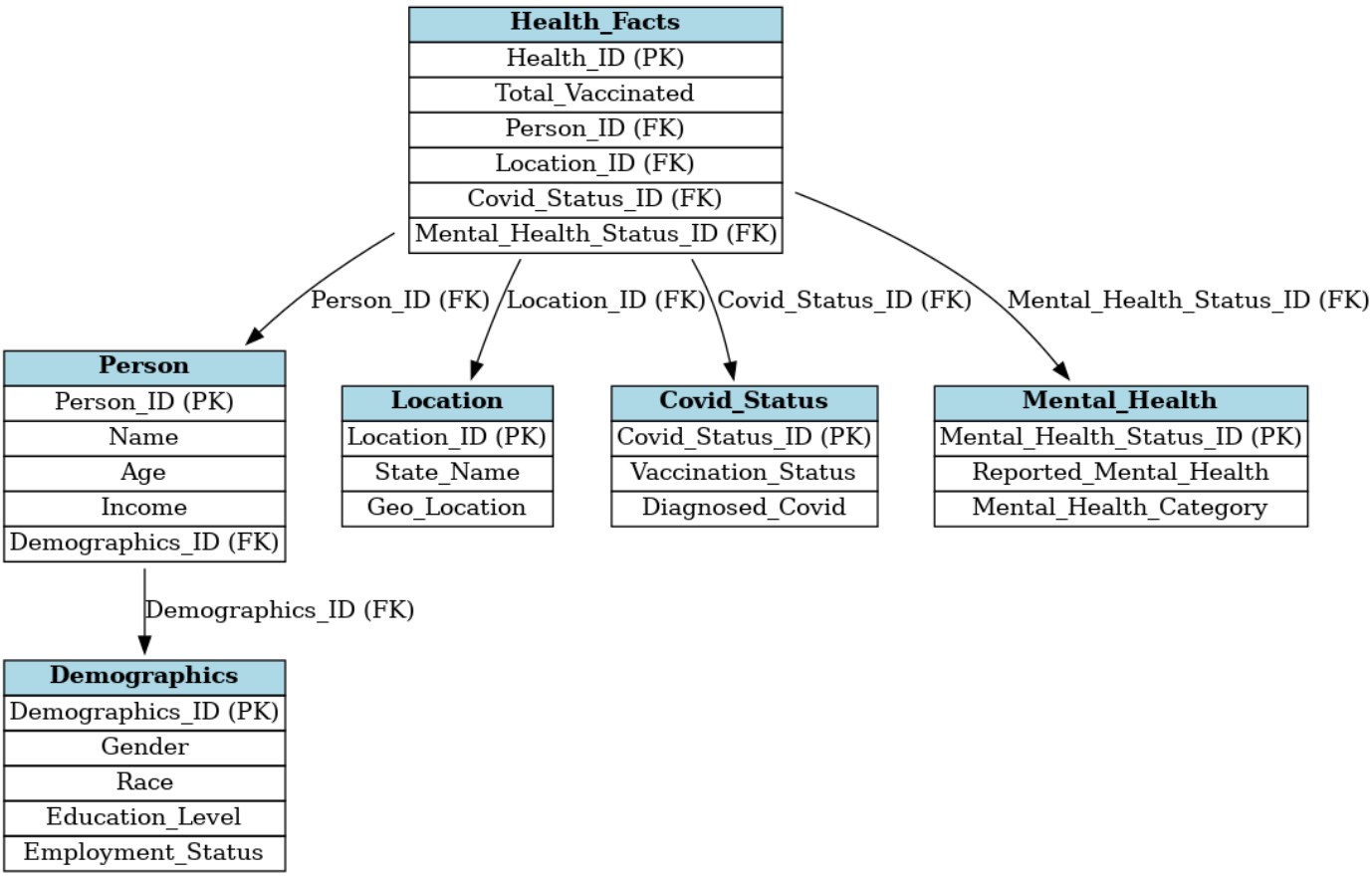
Trend and Comparative Analysis: Analyze the cleaned data to identify trends and patterns in Mental Health across different demographics (e.g., Mental Health Status, Diagnosed, state)

Data Loading and Reporting: Loading the cleaned data into a structured database or data warehouse for further analysis. Develop reports and visualizations that summarize key findings and health insights.

Geospatial and Demographic Visualization: Creating interactive dashboards and maps that allow public health officials to visualize data trends by location, demographics, and health behaviors, supporting data-driven decisions.

5. Model Architecture:

This model illustrates a data warehouse structure that integrates health-related data from the Behavioral Risk Factor Surveillance System. It shows how various dimensions and fact tables interact to provide a robust framework for analyzing public health metrics. The primary goal of this model is to enable multidimensional analysis, offering insights into public health behaviors, lifestyle factors, and environmental impacts on health.



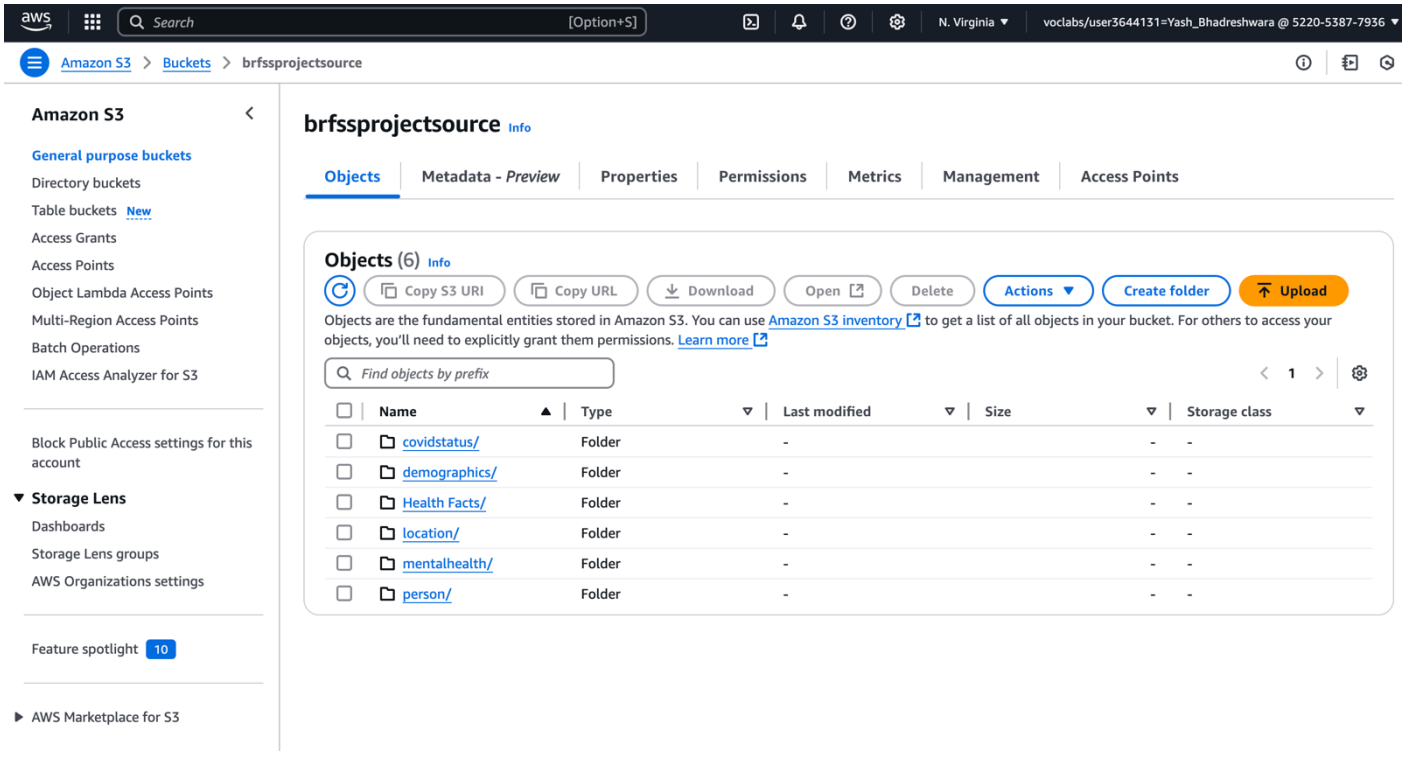
At the core of the model lies the Health_Facts table, which acts as the fact table. It contains key quantitative metrics such as Total Vaccinated. These metrics are the central data points analyzed in the system. Surrounding the fact table are several dimension tables including Covid_Status, Location, Person, Mental_Health, Demographics. Additionally, the Person dimension includes a hierarchy of Demographics _ID → Geo_Location, facilitating geographic analysis at various levels.

6. Loading Data in S3 Buckets

To efficiently manage and process the extensive data from the BRFSS dataset, we implemented a structured cloud storage solution using AWS S3. Two distinct buckets were created to handle the workflow: a source bucket and a target bucket. The source bucket served as the entry point for raw, unprocessed data. This bucket was designed to securely store the diverse and fragmented datasets collected from various sources, ensuring centralized access for further processing. By organizing the raw data in the source bucket, the project tackled challenges such as data fragmentation and inconsistencies, laying the groundwork for streamlined transformations.

Once the source bucket was established, raw BRFSS data files were uploaded to it. These files contained critical health-related data, such as information on person, Mental Health, Covid Status, Locations and Demographics which were to be standardized and analyzed in subsequent steps. The use of the source bucket provided a clean separation of concerns, ensuring that raw data was preserved in its original format before undergoing any modifications. This structure not only enhanced the organization of the ETL pipeline but also supported scalability and reliability for large-scale public health datasets.

The target bucket, on the other hand, was designed to store processed and transformed data. After the raw data in the source bucket was cleaned, standardized, and enriched, it was transferred to the target bucket, where it was readily available for downstream analysis. This separation of raw and processed data across two buckets ensures an efficient and secure workflow, enabling seamless integration with other AWS services, such as Glue and Athena, for advanced querying and analytics. This architecture highlights the robustness of the cloud-based solution in handling large and complex datasets, ultimately transforming raw data into actionable public health insights.

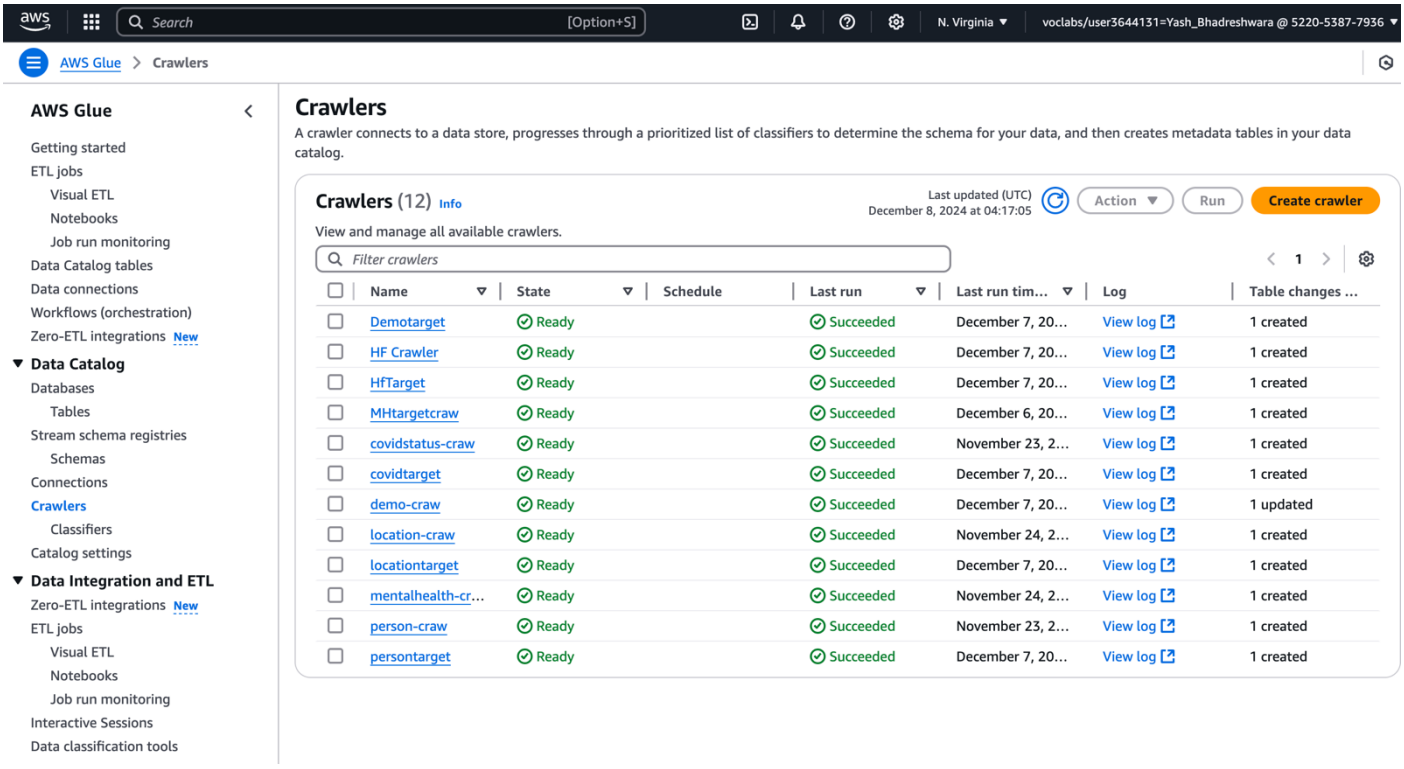


7. Crawlers

To efficiently automate the process of schema discovery and data cataloging, the project utilized AWS Glue Crawlers, creating two distinct crawlers for each data dimension: one for the source bucket and another for the target bucket. The source crawlers were configured to scan the raw data stored in the source bucket, automatically inferring the schema and cataloging the metadata in the AWS Glue Data Catalog. This setup ensured that the structure and attributes of the raw data were readily accessible, even for complex and diverse datasets. The source crawlers played a vital role in streamlining the extraction process by providing a standardized view of the unprocessed data for transformation.

In parallel, target crawlers were established to handle the processed data in the target bucket. After the data from the source bucket was cleaned, transformed, and loaded into the target bucket, the target crawlers were run to update the Glue Data Catalog with the new schema and metadata of the processed datasets. These crawlers ensured that the transformed data dimensions such as demographics, location, mental health, and COVID status were properly cataloged for advanced analytics and querying using AWS Athena.

By creating separate crawlers for the source and target buckets, the project maintained a clear distinction between raw and processed data while enabling seamless data exploration and integration. This dual-crawler approach automated schema discovery, reduced manual effort, and ensured the data catalog remained up-to-date as the pipeline evolved. This setup demonstrated the power of AWS Glue in creating an efficient and scalable data pipeline, transforming raw datasets into well-structured resources ready for multidimensional analysis.



8. ETL Implementation

The ETL implementation for this project was carried out using AWS Cloud Services to extract, transform, and load BRFSS data into the target bucket. The process leveraged the scalability and efficiency of AWS Glue, Athena, and S3 to create a seamless and robust ETL pipeline.

Extraction

The first step in the ETL process was extracting the source data. Using AWS Glue Crawlers, data stored in the source bucket was automatically scanned and cataloged in the AWS Glue Data Catalog. This automated process identified schemas for each dataset, which included health-related metrics, demographics, geographic data, and COVID-related information. For example:

- The Person dimension extracted demographic data such as age, income, and gender.
- The Location dimension sourced geographic attributes such as state name and geolocation.
- The Covid_Status dimension retrieved information on vaccination status and COVID diagnosis.

The Glue Data Catalog made the extracted data readily available for further transformations and ensured data consistency throughout the process.

Transformation

Transformation was carried out using SQL queries in AWS Athena and AWS Glue transformations, wherever necessary, to clean, map, and enrich the data. Key transformation steps included:'

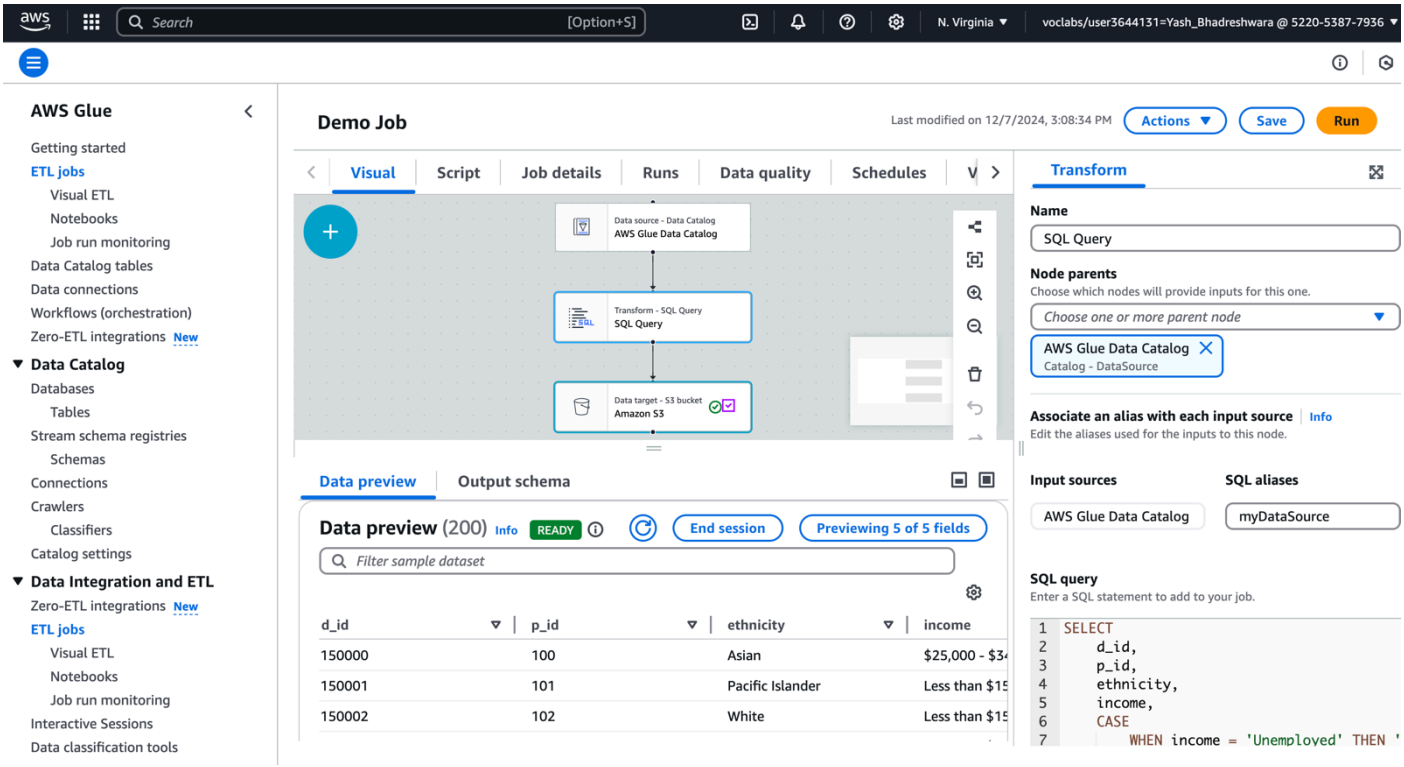
- Data Mapping: Data fields from the source bucket were mapped to their corresponding attributes in the target schema. For example, employment status was added to the Demographics dimension, and state IDs were mapped to geographic attributes in the Health_Facts table.
- Data Cleaning: Missing values were imputed using default values or averages, and inconsistent formats, such as date fields, were standardized to ensure uniformity.
- Data Enrichment: Certain fields, such as employment status, were added to provide more granular insights.

Loading

The final step in the ETL process was loading the transformed data into the target bucket. The processed datasets were written back to the target bucket in structured formats, ready for querying and analysis. Each table was populated according to the relationships defined in the schema. For instance:

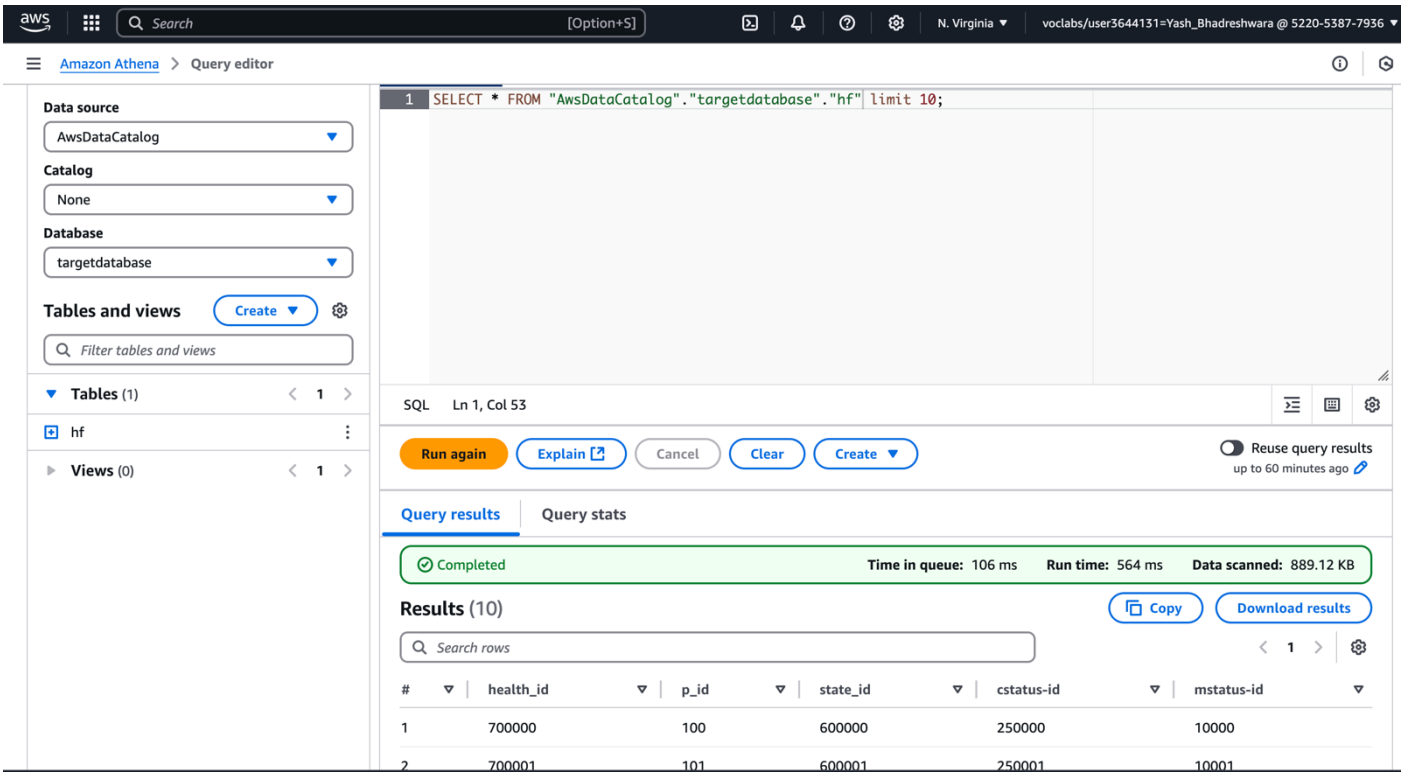
- The Health_Facts table was populated with metrics such as Total Vaccinated and mental health status.
- Dimension tables like Person, Location, and Covid_Status were updated with the latest demographic, geographic, and health data.

This cloud-based ETL pipeline not only streamlined data processing but also ensured scalability, consistency, and reliability in handling the vast and complex BRFSS datasets. The seamless integration of AWS services allowed public health stakeholders to extract actionable insights efficiently, transforming raw data into meaningful public health interventions.



9. Athena

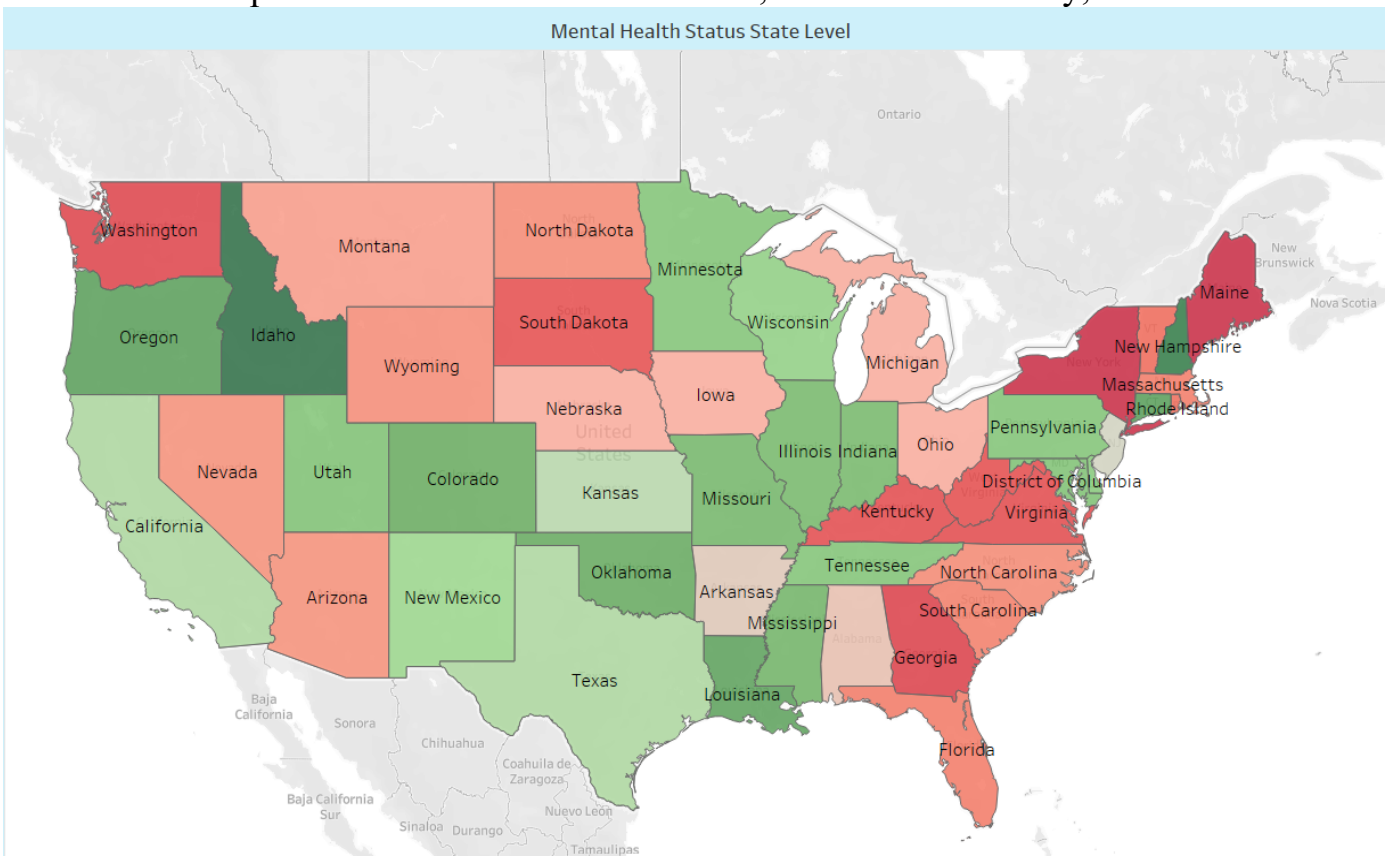
In this project, **AWS Athena** was utilized as a powerful tool for querying and analyzing the processed data stored in the target bucket. By directly interacting with the datasets cataloged in AWS Glue, Athena enabled seamless and efficient execution of SQL queries on the output data without the need for managing underlying infrastructure. Using Athena, I conducted detailed analyses on key metrics such as vaccination coverage, mental health trends, and demographic distributions. The ability to perform real-time queries allowed for swift identification of patterns and insights, facilitating a deeper understanding of health behaviors and trends. Athena's serverless nature ensured cost efficiency and scalability, making it an ideal choice for handling the large and complex BRFSS dataset. This streamlined analytical capability proved instrumental in transforming the processed data into actionable insights, empowering data-driven public health strategies.



10. Insights & Recommendations

1. **Percentage of People Diagnosed with COVID:**
   - The average COVID diagnosis rate across the population is approximately **49.89%**.
   - This suggests nearly half the population has faced exposure, which could be attributed to factors like vaccination availability, public health measures, or population density in different states.

2. **Mental Health Score (1-4):**
   - The overall mental health score is low, averaging **2.497**.
   - This indicates a moderate level of mental health struggles, potentially linked to the pandemic's stressors like isolation, economic instability, or health concerns.



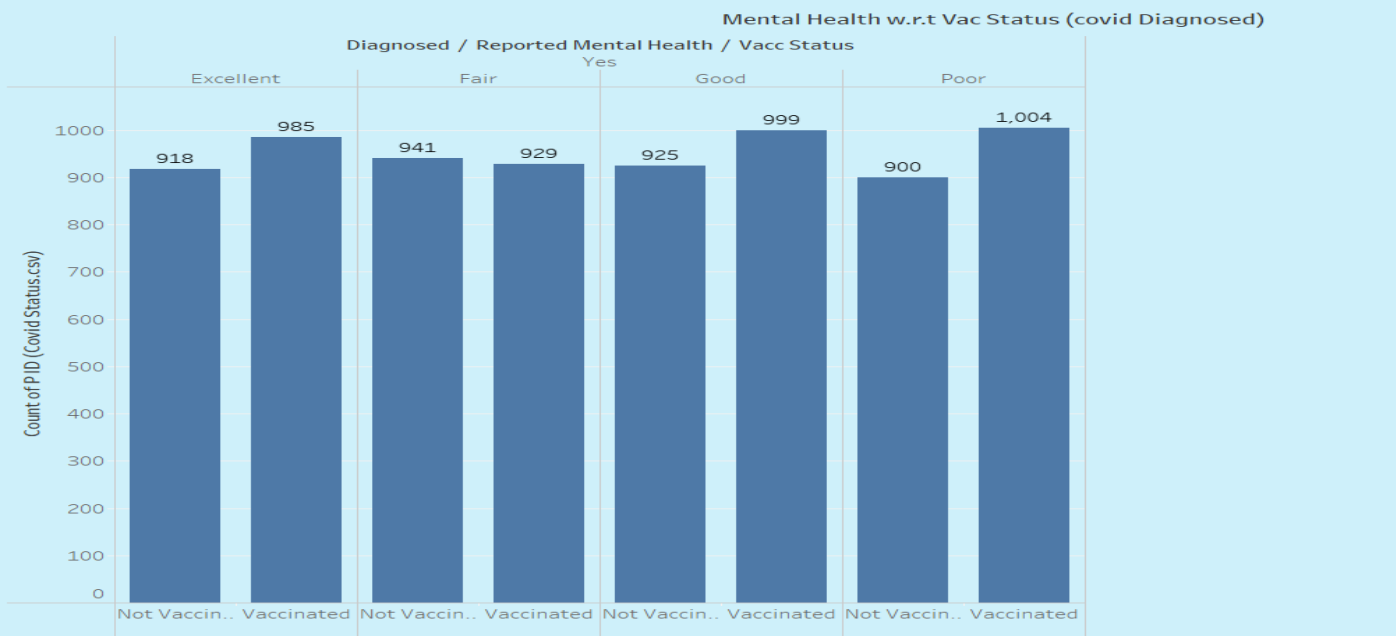
3. **Vaccination Trends:**
   - The count of people partially vaccinated is relatively small (7,457).
   - Partial vaccination may reflect hesitation, misinformation, or challenges in vaccine distribution during the analyzed time.

4. **State-Level Mental Health Status:**
   - States like South Carolina and Florida appear to have poor mental health outcomes (indicated in red on the map), while states like Montana and North Dakota show better mental health outcomes.
   - These differences might stem from factors such as healthcare access, state-specific pandemic responses, or sociocultural differences.

5. **Top Vaccinated & Diagnosed States:**
    - States like California, Alabama, and Delaware lead in vaccination and diagnosed rates. This could indicate better vaccination outreach and also higher reporting or testing rates for COVID cases.

6. **Mental Health Timeline in 2021:**
    - Mental health deteriorated significantly in the first few months of 2021 and then started stabilizing.
    - The sharp decline could align with peak COVID waves, social restrictions, or vaccination anxieties.

7. **Mental Health with Respect to Vaccination & Diagnosis Status:**
    - Those vaccinated and diagnosed report better mental health scores compared to those unvaccinated and diagnosed.
    - Vaccination likely provides psychological relief, fostering a sense of security and reducing health-related stress.



Recommendation:

Focus on Vaccination Drives:

- Increase outreach efforts in states with low vaccination counts, leveraging community programs and addressing vaccine hesitancy through education.

Address Mental Health Crisis:

- States with poor mental health scores (e.g., South Carolina, Florida) should prioritize mental health programs.
- Invest in telehealth services, counseling, and campaigns to destigmatize mental health issues.

Monitor High COVID-Diagnosed States:

- States like California and Alabama should maintain robust health infrastructure to manage higher diagnosis rates and potential future waves.

Tackle Pandemic-Induced Mental Stress:

- Implement policies focusing on work-life balance, economic support, and recreational activities to mitigate long-term pandemic stress.

11.Conclusion:

This project successfully addresses the challenges associated with analyzing large-scale public health datasets, particularly the Behavioral Risk Factor Surveillance System (BRFSS). By leveraging AWS Cloud Services, it provided a scalable, efficient, and reliable framework to transform raw health data into actionable insights, emphasizing key trends in mental health, COVID-19 diagnoses, and vaccination patterns.

- **Key Achievements:**

1. **AWS-Based ETL Pipeline:**

- Data Storage: Raw and processed data were managed efficiently using AWS S3 buckets, ensuring secure and scalable storage.

- Schema Automation: AWS Glue crawlers automated schema discovery and data cataloging, reducing manual efforts.

- Data Transformation: AWS Athena and SQL queries cleaned and structured data, enabling advanced analysis.

- Real-Time Analysis: A data warehouse structure supported real-time querying using OLAP systems.

2. **Insights on Mental Health and COVID-19:**

- Mental Health Trends: The pandemic has had a significant negative impact on mental health, particularly in states like South Carolina and Florida.

- Vaccination Status: Vaccinated individuals reported better mental health scores, highlighting the psychological benefits of vaccination.

- COVID Exposure: Nearly 50% of the population has been diagnosed with COVID-19, showcasing the widespread impact of the pandemic.

- Geographic Insights: States with high vaccination rates and COVID cases, such as California and Alabama, were able to report better health metrics due to proactive public health measures.

This project highlights the potential of cloud-based solutions like AWS in transforming raw, fragmented data into structured, actionable resources for public health stakeholders. By combining robust data warehousing techniques with advanced analytics, it bridges the gap between data and impactful health interventions.