**Project: Improve the Health of the Individuals**

**Name: Madhuri Basava**

**Date: 5/29/2024**

**Introduction**

People are suffering from many illnesses and may need more facilities/services.

This project’s objective is to find which city in the US has the most illnesses so we can build facilities and provide more services to improve people's health.

Components for this project: HDFS, Spark, and Yarn.

* HDFS is selected for the following reasons:

Primary Data Storage: HDFS is designed to store vast amounts of data across multiple nodes in a distributed manner. It handles large datasets efficiently, providing fault tolerance and high throughput.

Scalability: HDFS can scale horizontally by adding more nodes to the cluster, allowing it to accommodate growing volumes of transactional data.

Foundation for Analytics: As a core component of the Hadoop ecosystem, HDFS serves as the foundational storage layer for other tools like Hive, Spark, and HBase, facilitating various types of data processing and analysis.

* Spark is selected because it is a distributed data processing engine that can perform high-speed data querying, analysis, and transformations with large data sets.
* Yarn acts as a Job Scheduler and Resource Manager.

PySpark is selected since PySpark, with its Python-friendly ecosystem excels in data analysis and machine learning integration.

The big city health dataset is uploaded into HDFS and seamlessly integrated into Spark.

The data Source is from the Kaggle website:

<https://www.kaggle.com/datasets/noordeen/big-city-health-data>.

**Methodology and results**

The methodology followed here is to store the dataset in HDFS (Hadoop Distributed File storage system), process the large data set with Spark, analyze the results, and build a Machine Learning model to identify which places need more facilities/services to improve people's health.

**HDFS is running:**A screenshot of a computer program

Description automatically generated

Spark is running with 2 processors and Yarn is running with 2 active nodes:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

BigDataCitiesData.csv is uploaded from the Kaggle website into the git repo and loaded into the Hadoop cluster with wget.

A screenshot of a computer screen

Description automatically generated

Ran PySpark and data is loaded into a data frame:A screenshot of a computer screen

Description automatically generated

Display the data frame:

A screen shot of a computer

Description automatically generated

Transform the data frame into a table:

A computer screen shot of a computer code

Description automatically generated

Describe the table to understand the columns and their data types:

A screenshot of a computer program

Description automatically generated

Display the various important fields' unique values:

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

Pick the top 6 Illnesses of the individuals:

spark.sql("select `Indicator Category`, count(`Indicator Category`) from df\_healthData df group by `Indicator Category` ORDER BY count(`Indicator Category`) desc") .show();

A screenshot of a computer

Description automatically generated

Perform transformations to filter by Place not equal to 'U.S. Total' as we want to know the particular place.

spark.sql("select DISTINCT(`Indicator Category`), Place, Value from df\_healthData where `Indicator Category` in ('HIV/AIDS', 'Injury and Violence', 'Nutrition, Physical Activity, & Obesity', 'Infectious Disease', 'Cancer', 'Maternal and Child Health') and Place <> 'U.S. Total' ORDER BY cast(Value as int) desc ").show()

A screenshot of a computer

Description automatically generated

Here, from the above results, we see that the number one disease in the USA is HIV/AIDS. So, we need to concentrate more on helping them live their lives by providing them with the services they need.

Performed transformation to display the issues suffered by the individuals belonging to various cities so that we can provide the facilities/services in those cities. Gender or Race of the individuals is not taken into account, as it is not necessary in our case.

spark.sql("select DISTINCT(`Indicator Category`), Place from df\_healthData where `Indicator Category` in ('HIV/AIDS', 'Injury and Violence’, ‘Nutrition, Physical Activity, & Obesity', 'Infectious Disease', 'Cancer', 'Maternal and Child Health') and Place <> 'U.S. Total' ").show()

A screenshot of a computer

Description automatically generated

**Conclusion**

By seeing the above results, we can conclude that most people are suffering from HIV/AIDS. and more cases are seen in Washington, San Francisco, Baltimore, and Miami. Nutrition and violence are more common in San Jose, Oakland, Philadelphia, San Diago, and Houston cities. Nutrition, Physical Activity, & Obesity related issues are found more in Atlanta and Long Beach, CA. More people are suffering from Cancer in Atlanta and Washington, DC. Infectious diseases are more common in Boston and Washington, DC. Maternal and Child Health services are more needed in Fort Worth, Boston, Chicago, and Baltimore. More facilities and services can be provided in these cities to help improve the health of the individuals.

This project can be enhanced by incorporating additional data sources related to healthcare professionals who can be moved to the facilities that need them. The classic use case can be attributed to the pandemic (in the past, during COVID-19, more healthcare professionals were needed in New York City and other highly affected cities).