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.

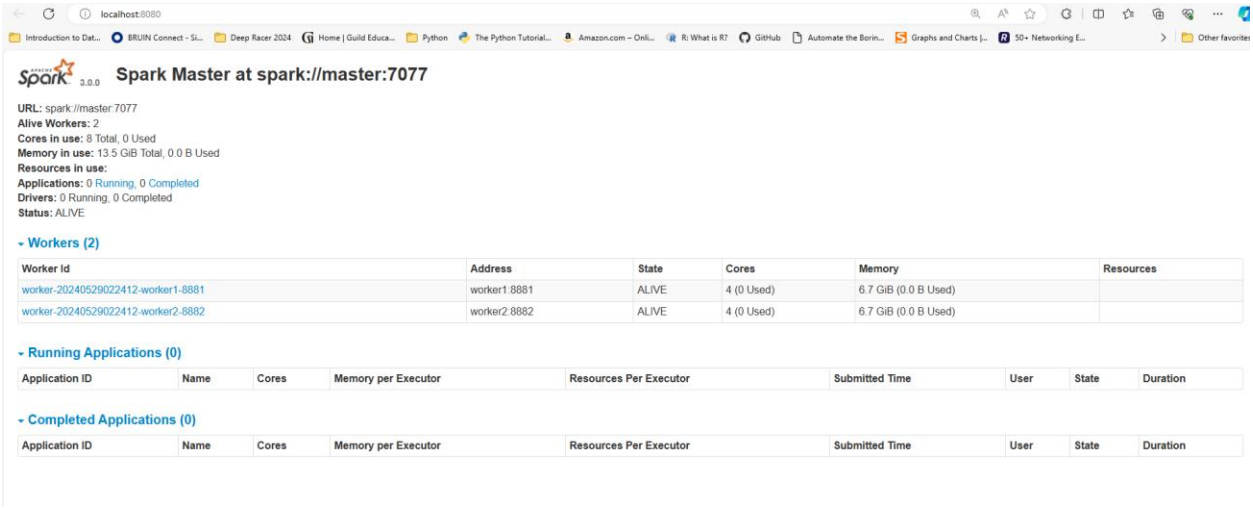
Methodology and results

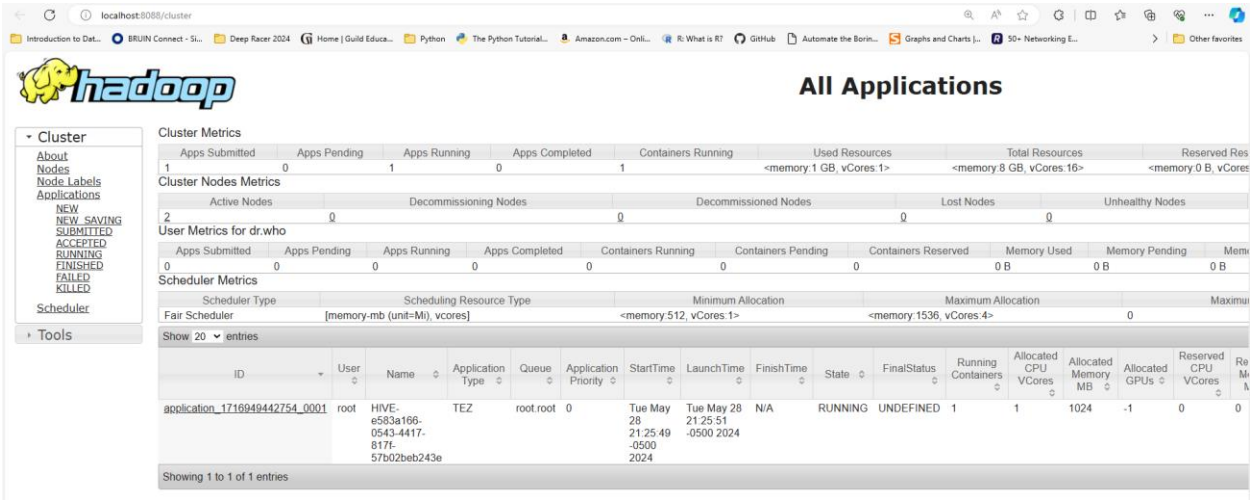
The methodology followed here is to store the dataset in HDFS (Hadoop Distributed File storage system), process the large data set with Spark, analyze the results, and build a Machine Learning model to identify which places need more facilities/services to improve people's health.

HDFS is running:

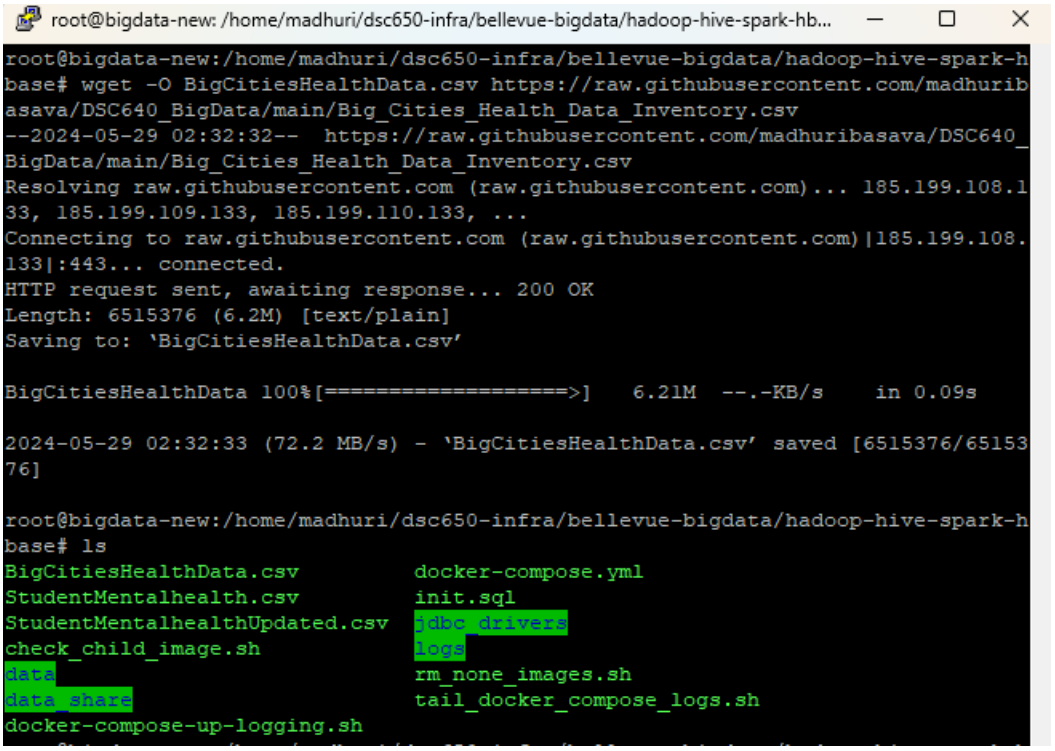
```
bash-5.0# hdfs dfsadmin -report
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/sl
f4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.
jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-05-29 22:27:51,061 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Configured Capacity: 103670202368 (96.55 GB)
Present Capacity: 75355073616 (70.18 GB)
DFS Remaining: 74974869385 (69.83 GB)
DFS Used: 380204231 (362.59 MB)
DFS Used%: 0.50%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (2):

Name: 172.28.1.2:9866 (worker1)
Hostname: worker1
Decommission Status : Normal
Configured Capacity: 51835101184 (48.28 GB)
DFS Used: 79024639 (75.36 MB)
Non DFS Used: 13983444481 (13.02 GB)
DFS Remaining: 37621637203 (35.04 GB)
DFS Used%: 0.15%
DFS Remaining%: 72.58%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 3
Last contact: Wed May 29 22:27:50 GMT 2024
Last Block Report: Wed May 29 22:19:44 GMT 2024
Num of Blocks: 119


Name: 172.28.1.3:9866 (worker2)
Hostname: worker2
Decommission Status : Normal
Configured Capacity: 51835101184 (48.28 GB)
DFS Used: 301179592 (287.23 MB)
Non DFS Used: 13761289528 (12.82 GB)
```

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

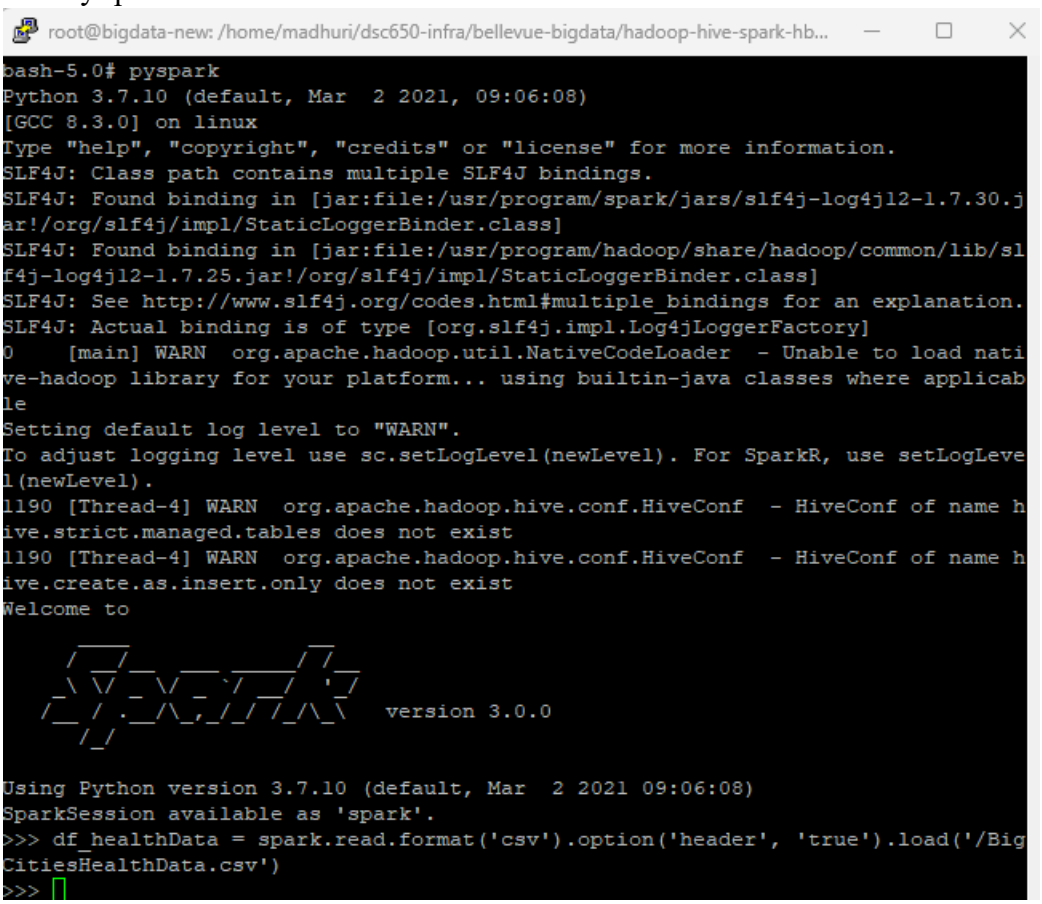




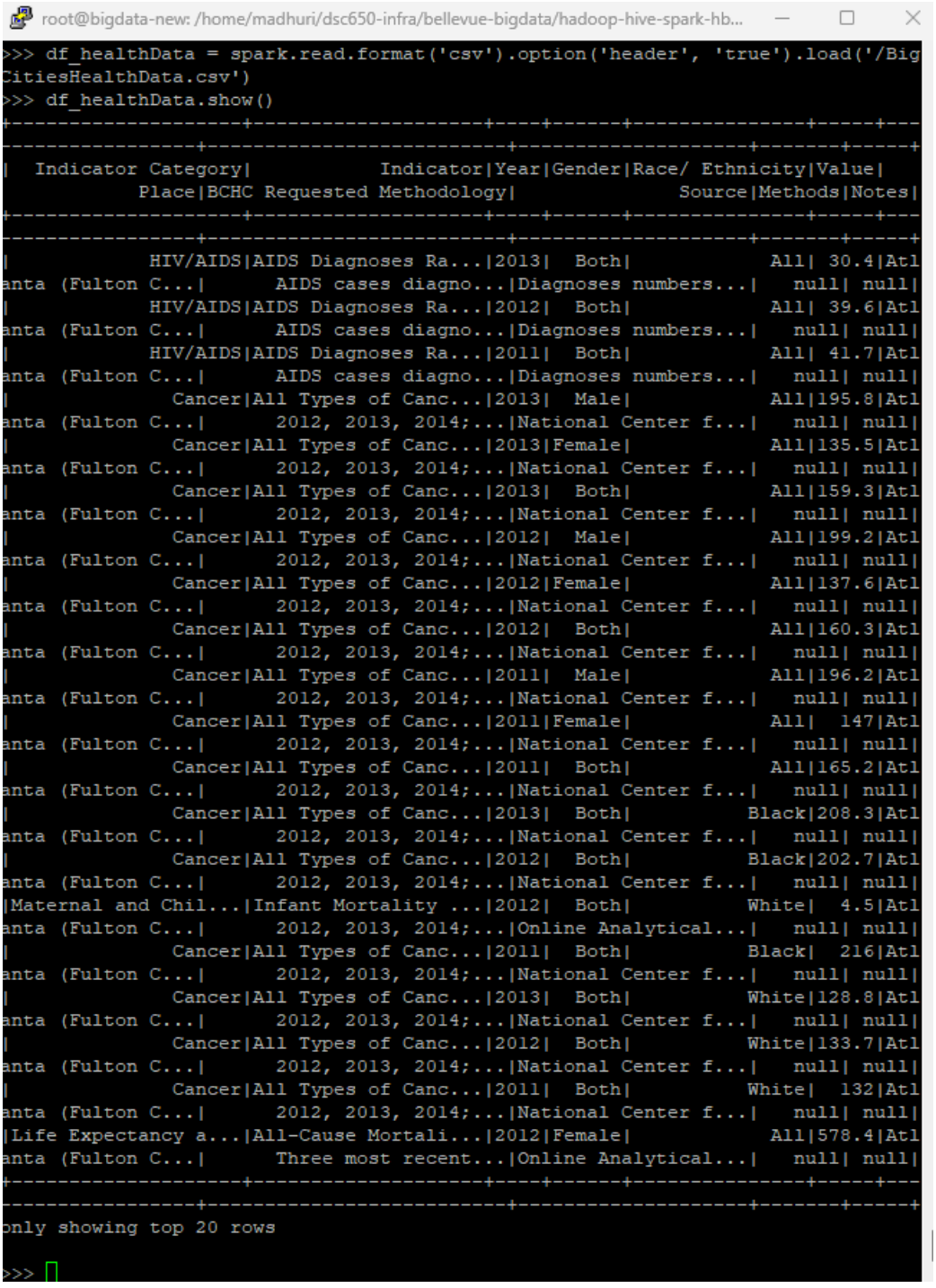
BigDataCitiesData.csv is uploaded from the Kaggle website into the git repo and loaded into the Hadoop cluster with wget.



Ran PySpark and data is loaded into a data frame:



Display the data frame:



Transform the data frame into a table:

```
>>> df_healthData.createOrReplaceTempView('df_healthData')
>>> spark.sql('SHOW TABLES').show()
1006710 [Thread-4] WARN  org.apache.hadoop.hive.conf.HiveConf  - HiveConf of nam
e hive.strict.managed.tables does not exist
1006711 [Thread-4] WARN  org.apache.hadoop.hive.conf.HiveConf  - HiveConf of nam
e hive.create.as.insert.only does not exist
1006749 [Thread-4] WARN  org.apache.spark.sql.hive.client.HiveClientImpl  - Dete
cted HiveConf hive.execution.engine is 'tez' and will be reset to 'mr' to disabl
e useless hive logic
+--------+-------------+-----------+
|database|    tableName|isTemporary|
+--------+-------------+-----------+
|        |df_healthdata|       true|
+--------+-------------+-----------+

>>>
```

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

```
>>> spark.sql("describe df_healthData").show()
+--------------------+---------+-------+
|            col_name|data_type|comment|
+--------------------+---------+-------+
|  Indicator Category|   string|   null|
|           Indicator|   string|   null|
|                Year|   string|   null|
|              Gender|   string|   null|
|     Race/ Ethnicity|   string|   null|
|               Value|   string|   null|
|               Place|   string|   null|
|BCHC Requested Me...|   string|   null|
|              Source|   string|   null|
|             Methods|   string|   null|
|               Notes|   string|   null|
+--------------------+---------+-------+

>>>
```

Display the various important fields' unique values:

```
>>> spark.sql("Select DISTINCT `Place` from df_healthData").show();spark.sql("Se
lect DISTINCT `Place` from df_healthData").show();
+--------------------+
|               Place|
+--------------------+
|Fort Worth (Tarra...|
|Miami (Miami-Dade...|
|          U.S. Total|
|       Cleveland, OH|
|     Kansas City, MO|
|      Sacramento, CA|
|                null|
|         Seattle, WA|
|          Boston, MA|
|         Houston, TX|
|   San Francisco, CA|
|     Los Angeles, CA|
|          Dallas, TX|
|        San Jose, CA|
|Portland (Multnom...|
|         Chicago, IL|
|         Detroit, MI|
|Atlanta (Fulton C...|
|      Washington, DC|
|     Minneapolis, MN|
+--------------------+
only showing top 20 rows

+--------------------+
```

```
KeyboardInterrupt
>>> spark.sql("Select DISTINCT `Gender` from df_healthData").show();
+------+
|Gender|
+------+
|  null|
|Female|
|  Both|
|  Male|
+------+
```

```
>>> spark.sql("Select DISTINCT `Race/ Ethnicity` from df_healthData").show();
+--------------------+
|     Race/ Ethnicity|
+--------------------+
|     Native American|
|                null|
|Does not include ...|
|         Multiracial|
|               Other|
|                 All|
|Native Hawaiian o...|
|               White|
|               Black|
|            Hispanic|
|            Asian/PI|
|American Indian/A...|
+--------------------+
```

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();

```
>>> spark.sql("select `Indicator Category`, count(`Indicator Category`) from df_healthData
df  group by `Indicator Category` ORDER BY count(`Indicator Category`) desc  ") .show();
+--------------------+------------------------+
|  Indicator Category|count(Indicator Category)|
+--------------------+------------------------+
|            HIV/AIDS|                    2177|
| Injury and Violence|                    1916|
|Nutrition, Physic...|                    1841|
|  Infectious Disease|                    1486|
|              Cancer|                    1432|
|Maternal and Chil...|                    1323|
|Behavioral Health...|                     983|
|         Food Safety|                     874|
|Life Expectancy a...|                     544|
|        Demographics|                     504|
|             Tobacco|                     432|
|2009-2013 America...|                     161|
|       United States|                      27|
|from the flu shot...|                       7|
|      your nose?"" "|                       7|
|(see note above a...|                       6|
|(percent of respo...|                       4|
|Age Group: United...|                       2|
|            (S1701)"|                       2|
| FOR THE POPULATI...|                       1|
+--------------------+------------------------+
only showing top 20 rows

>>>
```

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()

```
>>> 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. Tota
l' ORDER BY cast(Value as int) desc  ").show()
+------------------+--------------------+------+
|Indicator Category|               Place| Value|
+------------------+--------------------+------+
|          HIV/AIDS|      Washington, DC|4199.6|
|          HIV/AIDS|   San Francisco, CA|4125.6|
|          HIV/AIDS|   San Francisco, CA|4104.2|
|          HIV/AIDS|      Washington, DC|4094.6|
|          HIV/AIDS|      Washington, DC|4045.8|
|          HIV/AIDS|      Washington, DC|3990.8|
|          HIV/AIDS|      Washington, DC|3961.8|
|          HIV/AIDS|      Washington, DC|3889.7|
|          HIV/AIDS|   San Francisco, CA|  3568|
|          HIV/AIDS|   San Francisco, CA|3535.2|
|          HIV/AIDS|   San Francisco, CA|3492.4|
|          HIV/AIDS|       Baltimore, MD|3454.1|
|          HIV/AIDS|       Baltimore, MD|  3449|
|          HIV/AIDS|       Baltimore, MD|  3395|
|          HIV/AIDS|       Baltimore, MD|3380.5|
|          HIV/AIDS|       Baltimore, MD|3171.8|
|          HIV/AIDS|       Baltimore, MD|3118.8|
|          HIV/AIDS|   San Francisco, CA|2794.8|
|          HIV/AIDS|Miami (Miami-Dade...|2783.9|
|          HIV/AIDS|   San Francisco, CA|2781.7|
+------------------+--------------------+------+
only showing top 20 rows

>>>
```

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()

```
>>> spark.sql("select DISTINCT(`Indicator Category`), Place from df_healthData where `Indic
ator Category` in ('HIV/AIDS','Injury and Violence','Nutrition, Physical Activity, & Obesit
y','Infectious Disease','Cancer','Maternal and Child Health') and Place <> 'U.S. Total' ").
show()
+--------------------+--------------------+
|  Indicator Category|               Place|
+--------------------+--------------------+
| Injury and Violence|        San Jose, CA|
|Nutrition, Physic...|Atlanta (Fulton C...|
|           HIV/AIDS|         Seattle, WA|
|Maternal and Chil...|Fort Worth (Tarra...|
|Nutrition, Physic...|    Philadelphia, PA|
|Maternal and Chil...|          Boston, MA|
|Maternal and Chil...|         Chicago, IL|
|  Infectious Disease|          Boston, MA|
|             Cancer|Atlanta (Fulton C...|
| Injury and Violence|         Oakland, CA|
|             Cancer|      Washington, DC|
|Nutrition, Physic...|      Long Beach, CA|
| Injury and Violence|    Philadelphia, PA|
|           HIV/AIDS|Miami (Miami-Dade...|
|Maternal and Chil...|       Baltimore, MD|
|           HIV/AIDS|       Cleveland, OH|
|  Infectious Disease|      Washington, DC|
|           HIV/AIDS|Atlanta (Fulton C...|
| Injury and Violence|San Diego County, CA|
| Injury and Violence|         Houston, TX|
+--------------------+--------------------+
only showing top 20 rows

>>>
```

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).