



**SCHOOL OF COMPUTER TECHNOLOGY
APPLIED A.I. SOLUTIONS DEVELOPMENT**

**BIG DATA TOOLS AND TECHNOLOGIES
FINAL REPORT
PROFESSOR - NIZAR ALI**

Submitted By:

Krutagna Joshi - 101513552

Vudit Suri - 101513592

Manik Kapil - 101447365

Riya Gupta - 101504556

Vithushan Vijayananthan - 101494734

Date: July 11, 2024

Our project, "Interpreting US Hospitalizations using Several Causal Factors," aims to identify and quantify key factors influencing hospitalization and fatality rates. By using postal codes as a common key, we seek to establish correlations with various factors, including fast food availability, housing conditions, and other socio-economic variables.

To achieve this, we merge all the datasets using PySpark to handle the large volume of data efficiently. We then create visualizations in Power BI to provide clear and actionable insights. The datasets are uploaded to various databases, including Hive, MySQL, SQL Server, and MariaDB, using a virtual machine for seamless data integration and accessibility. This approach ensures robust data handling and comprehensive analysis across multiple platforms.

Below are the datasets utilized in this analysis:

Database of Economic Incentives

The dataset contains information about various economic incentives provided to different recipients across New York State. Each entry includes details such as the recipient's name, project name, description, address, county, region, industry, and start date. The dataset also lists the type and amount of assistance awarded, the total public-private investment, disbursements to-date, project status, and compliance status. Additionally, it includes job creation and retention commitments and records the jobs created and retained to-date

Datafiniti Fast Food Restaurants

This dataset includes information on various fast food restaurants. It captures essential attributes such as the restaurant's name, address, city, state, postal code, latitude, longitude, and phone number. The dataset also provides details on the restaurant's type, cuisine, and menu, as well as ratings and reviews from customers. This comprehensive data is useful for analyzing the distribution, popularity, and customer satisfaction of fast food chains across different regions.

Home Fire Smoke Alarm Installs

This dataset documents smoke alarm installation activities conducted between September 2014 and March 2015. It includes the installation date, location (address, city, state, zip code), and the number of smoke alarms installed. The dataset also records the organization responsible for the installation and any additional comments or notes related to each installation event. This information is crucial for assessing the effectiveness and coverage of smoke alarm installation programs aimed at enhancing home safety.

Homes.com Properties Dataset

The dataset provides detailed information on properties listed on Homes.com. It includes attributes such as property ID, address, city, state, zip code, price, number of bedrooms and bathrooms, square footage, lot size, and year built. Additional details include property type, status (for sale, sold, etc.), and descriptions. This dataset is valuable for real estate market analysis, helping to identify trends in property values, availability, and other market dynamics.

Medical Services

This dataset comprises records of various medical services provided across different locations. Each entry contains information about the service provider, service type, address, city, state, zip code, and contact details. The dataset may also include details on the range of services offered, operational hours, and any specializations. This comprehensive data helps in understanding the distribution and accessibility of medical services, aiding in healthcare planning and resource allocation.

SITE HCC FCT DET

This dataset includes information relevant to healthcare facilities, focusing on factors such as facility type, location, and contact details. It may also capture specifics about the services offered, operational hours, and any accreditation or certification statuses. This dataset is crucial for analyzing the infrastructure of healthcare facilities, identifying gaps in service provision, and planning improvements to healthcare access and quality.

Crime Data

The dataset likely contains historical information related to various categories, potentially including demographics, economic indicators, or other time-sensitive data points. This dataset is essential for longitudinal studies, enabling researchers to track changes over time, compare past and present conditions, and predict future trends based on historical patterns.

The datasets were uploaded into different databases with the following configurations:

- **Hadoop/Hive:**
 - **Database Name:** health_centres
 - **Table Name:** my_table
 - **Details:** This table contains comprehensive data related to health centers, potentially including attributes like center names, addresses, services offered, and patient demographics. This dataset is crucial for analyzing healthcare accessibility and service distribution.

SQL Server:

- **Database:** Datafiniti_Fast_Food_Restaurants
 - **Details:** This dataset includes detailed information on various fast food restaurants. It captures essential attributes such as the restaurant's name, address, city, state, postal code, latitude, longitude, and phone number. Additionally, it provides details on the restaurant's type, cuisine, menu, ratings, and customer reviews. This comprehensive data is useful for analyzing the distribution, popularity, and customer satisfaction of fast food chains across different regions.

```

SELECT id, dateAdded, dateUpdated, address, categories, city, country, keys, latitude, longitude, name, postalCode, province, sourceURLs, websites
FROM master.dbo.fast_food_table_new;

```

| id | dateAdded | dateUpdated | address | categories | city | country | keys |
|----|----------------------|----------------------|------------------------|--|---------------|---------|---|
| 1 | 2015-10-19T23:47:58Z | 2018-06-26T03:00:42Z | 800 N Canal Blvd | American Restaurant and Fast Food Restaurant | Thibodaux | US | us/latthibodaux/800ncanalblvd/1780593795 |
| 2 | 2015-10-19T23:47:58Z | 2018-06-26T03:00:42Z | 800 N Canal Blvd | Fast Food Restaurants | Thibodaux | US | us/latthibodaux/800ncanalblvd/1780593795 |
| 3 | 2016-03-29T05:06:36Z | 2018-06-26T02:59:52Z | 206 Wears Valley Rd | Fast Food Restaurant | Pigeon Forge | US | us/tn/pigeonforge/206wearsvalleyrd/864103396 |
| 4 | 2017-01-03T07:46:11Z | 2018-06-26T02:59:52Z | 3652 Parkway | Fast Food | Pigeon Forge | US | us/tn/pigeonforge/3652parkway/93075755 |
| 5 | 2018-06-26T02:59:43Z | 2018-06-26T02:59:43Z | 2118 Mt Zion Parkway | Fast Food Restaurant | Morrow | US | us/ga/morrow/2118mtzionparkway/1305117222 |
| 6 | 2015-10-23T23:59:49Z | 2018-06-26T02:59:43Z | 9768 Grand River Ave | Fast Food Restaurant | Detroit | US | us/m/detroit/9768grandriverave/791445730 |
| 7 | 2015-09-21T07:47:08Z | 2018-06-26T02:59:43Z | 13600 W McNichols Rd | Fast Food Restaurant | Detroit | US | us/m/detroit/13600wmcnicholsrd/2061630068 |
| 8 | 2016-06-07T16:20:41Z | 2018-06-26T02:59:41Z | 4111 Oceanside Blvd | Fast Food | Oceanside | US | us/ca/oceanside/4111oceanisdbldv/1143321601 |
| 9 | 2016-12-13T12:34:52Z | 2018-06-26T02:59:52Z | 162 Old Country Rd | Fast Food Restaurants | Riverhead | US | us/ny/riverhead/162oldcountryrd/67504952 |
| 10 | 2016-04-16T16:20:41Z | 2018-06-25T12:05:40Z | 1407 S Stockton Ave | Fast Food Restaurant | Monahans | US | us/tx/monahans/1407sstocktonave/1721138121 |
| 11 | 2016-04-03T01:24:16Z | 2018-06-25T12:05:39Z | 208 W Mason St | Fast Food Restaurant and Ice Cream Shop | Mabank | US | us/tx/mabank/208wmasonst/-1721138121 |
| 12 | 2015-04-23T05:22:01Z | 2018-06-26T03:00:52Z | 3105 Highway 80 E | Fast Food Restaurants | Oregon | US | us/oh/oregon/1927woodvilleroad/938337463 |
| 13 | 2018-01-26T05:22:01Z | 2018-06-26T03:00:52Z | 17620 Grand River Ave | Fast Food Restaurant | Pearl | US | us/ms/pearl/3105highway80e/938337463 |
| 14 | 2018-03-11T07:50:51Z | 2018-06-26T03:00:50Z | 5801 Highway 6 | Fast Food Restaurant | Detroit | US | us/m/detroit/17620grandriverave/1536894814 |
| 15 | 2016-06-04T20:48:06Z | 2018-06-25T12:05:39Z | 5801 Highway 6 | Ice Cream Shop and Fast Food Restaurant | Missouri City | US | us/tx/missouri/5801highway6/1721138121 |
| 16 | 2015-10-22T19:27:37Z | 2018-06-25T12:05:35Z | 3454 Manchester Rd | Fast Food Restaurant | Akron | US | us/oh/akron/3454manchesterrd/1065990473 |
| 17 | 2016-04-02T05:03:56Z | 2018-06-25T12:05:41Z | 1500 E Main St | Fast Food Restaurants | Eastland | US | us/tx/eastland/1500emainst/1797342045 |
| 18 | 2016-10-23T02:52:53Z | 2018-06-25T09:32:43Z | 1190 W Foothill Blvd | Fast Food Restaurant | Azusa | US | us/ca/azusa/1190wfoothillblvd/554191587 |
| 19 | 2016-03-23T04:12:21Z | 2018-06-25T09:32:32Z | 601 N Main St | Fast Food Restaurants | Summerville | US | us/sc/summerville/601mainst/1536894814 |
| 20 | 2016-09-14T19:13:09Z | 2018-06-25T09:31:52Z | 6316 W 89th St | Fast Food | Los Angeles | US | us/ca/losangeles/6316w89thst/10537776 |
| 21 | 2015-10-23T02:59:03Z | 2018-06-25T09:30:45Z | 2860 E Paris Ave SE | Fast Food Restaurant | Grand Rapids | US | us/m/grandrapids/2860eparisave/1536894814 |
| 22 | 2015-10-23T02:26:56Z | 2018-06-25T09:30:52Z | 2424 E State Road 44 | Fast Food Restaurant | Shelbyville | US | us/in/shelbyville/2424estateroad44/938337463 |
| 23 | 2015-11-27T18:40:25Z | 2018-06-25T09:30:50Z | 1611 Broadway | Fast Food | Brooklyn | US | us/ny/brooklyn/1611broadway/1536894814 |
| 24 | 2015-09-19T09:30:06Z | 2018-06-25T09:30:50Z | 1720 Walton Way | Fast Food Restaurants | Augusta | US | us/ga/augusta/1720waltonway/1536894814 |
| 25 | 2017-09-01T07:20:32Z | 2018-06-22T18:21:55Z | 1500 South Willow St | Fast Food Restaurant | Manchester | US | us/nh/manchester/1500southwillowst/93075755 |
| 26 | 2016-03-31T22:58:20Z | 2018-06-21T17:18:51Z | 1260 William D Tate | Fast Food Restaurant | Grapevine | US | us/tx/grapevine/1260williamdtate/93075755 |
| 27 | 2017-07-02T03:09:30Z | 2018-06-21T17:12:48Z | 324 Highland Crossing | Fast Food Restaurant | Ellijay | US | us/ga/ellijay/124highlandcrossing/442344130 |
| 28 | 2016-05-16T23:17:04Z | 2018-06-25T09:30:22Z | 196 Massachusetts Ave | Fast Food Restaurants | Arlington | US | us/ma/arlington/196massachusettsave/19768189 |
| 29 | 2017-06-03D1eR_xDlU | 2018-06-26T02:34:32Z | 20124 State St | Fast Food | Schenectady | US | us/ny/schenectady/1224statest/1161002137 |
| 30 | 2017-06-29T02:03:32Z | 2018-06-25T09:29:45Z | 1232 Ulster Ave | Fast Food Restaurant | Kingston | US | us/ny/kingston/1232ulsterave/1161002137 |
| 31 | 2016-09-19T14:50:03Z | 2018-06-25T09:29:44Z | 118th Avenue of The Am | Fast Food Restaurant | New York | US | us/ny/newyork/118thavenueoftheamericas/116100 |
| 32 | 2017-06-19T14:53:25Z | 2018-06-25T09:29:43Z | 114 Delancey St | Fast Food Restaurant | New York | US | us/ny/newyork/114delaneyst/1161002137 |
| 33 | 2016-09-19T14:52:42Z | 2018-06-25T09:29:41Z | 1095 Front St | Fast Food Restaurants | Uniondale | US | us/ny/uniondale/1050frontst/1161002137 |
| 34 | 2016-03-21T01:37:59Z | 2018-06-25T09:29:41Z | 105 Jericho Tpke | Fast Food Restaurant | Jericho | US | us/ny/jericho/105jerichotpke/1161002137 |
| 35 | 2018-03-21T01:37:59Z | 2018-06-25T09:29:41Z | 105 Jericho Tpke | Fast Food | Jericho | US | us/ny/jericho/105jerichotpke/1161002137 |
| 36 | 2016-09-19T15:24:34Z | 2018-06-25T09:29:41Z | 106 Wolf Rd | Fast Food Restaurants | Albany | US | us/ny/albany/106wolfrd/1161002137 |
| 37 | 2016-02-26T13:47:28Z | 2018-06-11T09:01:04Z | 1291 Boston Rd | Fast Food Restaurants | Springfield | US | us/ma/springfield/1291bostonrd/1251271227 |
| 38 | 2017-07-18T03:09:25Z | 2018-06-25T09:04:12Z | 18025 Ventura Blvd | Fast Food | Encino | US | us/ca/encino/18025venturabld/554191587 |
| 39 | 2017-09-01T21:46:56Z | 2018-06-25T09:46:57Z | 345 W Winthrop Rd | Fast Food Restaurants | Winona | US | us/na/winona/345wwinthroprd/14049191587 |

MariaDB:

- Database:** Medical Services
- Details:** This dataset comprises records of various medical services provided across different locations. Each entry contains information about the service provider, service type, address, city, state, zip code, and contact details. The dataset may also include details on the range of services offered, operational hours, and any specializations. This comprehensive data helps in understanding the distribution and accessibility of medical services, aiding in healthcare planning and resource allocation.

| reviews_count | postal_code | general_info | category | locality |
|---------------|-------------|--|------------------------|----------------------------|
| 11.201 | 11209 | Heights Aesthetic Laser Center, located in the Brooklyn Heights neighbor | Health & Wellness | New York, Brooklyn |
| 11.209 | 11209 | Hair Removal/Beauty Salons/Physicians & Surgeons, Dermatology/Skin Ca | Health & Wellness | New York, Brooklyn |
| 44.136 | 11209 | Skin Care/Beauty Salons/Health & Wellness Products/Skin Care/Beauty Sal | Health & Wellness | New York, Brooklyn |
| 76.118 | 11209 | Optical Goods/Contact Lenses/Eyeglasses/Optical Goods/Contact Lenses/Ey | Optical Goods | New York, Brooklyn |
| 85.123 | 11209 | Strongsville High School, located in the Strongsville, OH neighbor | Education & University | Ohio, Strongsville |
| 19.134 | 11209 | College & University/High Schools (K-12)/Schools | Education & University | Ohio, Strongsville |
| 28.204 | 11209 | V. Sattui Winery, located in the Napa Valley neighbor | Food & Beverage | California, Napa Valley |
| 76.009 | 11209 | Optical Goods/Eyeglasses/Optometrists | Optical Goods | New York, Brooklyn |
| 45.227 | 11209 | Philadelphia, located in the Philadelphia, PA neighbor | Health & Wellness | Pennsylvania, Philadelphia |
| 85.051 | 11209 | Optical Goods/Optometrists | Optical Goods | New York, Brooklyn |
| 32.803 | 11209 | Health & Fitness Program Consultants/Massage Therapists/Physical Therap | Health & Fitness | New York, Brooklyn |
| 76.009 | 11209 | Notaries Public/Hospital/Health Estate Law Processing/Notaries Public/Hospital | Health & Fitness | New York, Brooklyn |
| 33.134 | 11209 | Healthcare & Medical Services/Hospital/Healthcare & Medical Services/B | Health & Fitness | New York, Brooklyn |
| 79.912 | 11209 | Healthcare & Medical Services/Hospital/Healthcare & Medical Services/ | Health & Fitness | New York, Brooklyn |
| 32.073 | 11209 | Services include Botox/Xenonin, Dermat/Files, Laser Hair and Tattoo remov | Health & Fitness | New York, Brooklyn |
| 33.144 | 11209 | Physicians & Surgeons/Dermatology/Skin Care/Beauty Salons/Physici | Health & Fitness | New York, Brooklyn |
| 44.203 | 11209 | Services include Botox/Xenonin, Dermat/Files, Laser Hair and Tattoo remov | Health & Fitness | New York, Brooklyn |
| 48.139 | 11209 | Lymphatic Drainage Message Post-Surgery and Fibrosis after Liposuction | Health & Fitness | New York, Brooklyn |
| 10.003 | 11209 | Healthcare & Medical Services/Hospital/Healthcare & Medical Services/B | Health & Fitness | New York, Brooklyn |
| 90.048 | 11209 | Healthcare & Medical Services/Hospital/Healthcare & Medical Services/B | Health & Fitness | New York, Brooklyn |
| 41.037 | 11209 | Healthcare & Medical Services/Hospital/Healthcare & Medical Services/B | Health & Fitness | New York, Brooklyn |
| 46.227 | 11209 | Our mission at Wellness from Within: Massage and Healing is to create a | Health & Fitness | New York, Brooklyn |
| 33.301 | 11209 | Massage Therapists/Massage Therapists/Physical Therapists/Physical Ther | Health & Fitness | New York, Brooklyn |
| 60.646 | 11209 | Dr. Zoran Reparic is an experienced board-certified plastic surgeon specia | Health & Fitness | New York, Brooklyn |
| 53.005 | 11209 | Skin Care/Physicians & Surgeons/Physicians & Surgeons, Cosmetic Sur | Health & Fitness | New York, Brooklyn |
| 27 | 11209 | gical Surgeon/Optometrists/Optical Goods/Eyeglasses/Optical Goods/Repair | Health & Fitness | New York, Brooklyn |
| 27 | 11209 | Optical Goods/Optometrists/Physicians & Surgeons, Ophthalmology/Contac | Health & Fitness | New York, Brooklyn |
| 60.646 | 11209 | Optical Goods/Repair/Optical Goods/Repair/Optical Goods/Repair | Health & Fitness | New York, Brooklyn |
| 2 | 11209 | At LensCrafters located at 16970A W Biscayne Rd, we believe vision car | Health & Fitness | Florida, Miami |

MongoDB:

- **Database:** Economic Incentives
- **Format:** JSON
- **Details:** The dataset contains information about various economic incentives provided to different recipients across New York State. Each entry includes details such as the recipient's name, project name, description, address, county, region, industry, and start date. The dataset also lists the type and amount of assistance awarded, the total public-private investment, disbursements to-date, project status, and compliance status. Additionally, it includes job creation and retention commitments and records the jobs created and retained to-date. This information is essential for analyzing the impact of economic incentives on local economies and job markets.

The screenshot shows the MongoDB Atlas Data Services interface. On the left, a sidebar lists various project components: Overview, Deployment (selected), Database (Data Lake, Economic_incentives), Services (Device & Edge Sync, Triggers, Data API, Data Federation, Atlas Search, Stream Processing, Migration), and Security (Quickstart, Backup, Database Access, Network Access, Advanced). The 'Economic_incentives' database is selected. The main panel displays the 'Economic_incentives' collection with a single document highlighted. The document details a grant for Cree Inc. in Marcy, NY, worth \$500,000. The document structure is as follows:

```

_id: ObjectId('668b37f6451c4ccb16ad78ed')
Project ID # : "133,153"
Recipient Name : "Cree"
Project Name: "Cree - Marcy Nanocenter Capital"
Project Description: "Grant, Manufacturing, Mohawk Valley, Cree"
Project Address: "5737 Marcy Suny Parkway"
County : "Oneida"
Postal Code : "13483"
Region : "Mohawk Valley"
Industry : "Manufacturing"
Start Date : "11/21/2019"
End Date : null
Assistance Type : "Grant"
Total ESD Assistance Awarded : "500,000,000"
Total Public-Private Investment : "1,005,000,000"

```

SQL Server:

- **Database:** smoke alarm
- **Details:** This dataset documents smoke alarm installation activities conducted between September 2014 and March 2015. It includes the installation date, location (address, city, state, zip code), and the number of smoke alarms installed. The dataset also records the organization responsible for the installation and any additional comments or notes related to each installation event. This information is crucial for assessing the effectiveness and coverage of smoke alarm installation programs aimed at enhancing home safety.

The screenshot shows a database management interface with a left sidebar for navigating databases and objects. The main area displays a table titled 'Results 3' with the following schema:

| | rec_Loc_name | rec_Status | rec_Score | rec_Match_Type | rec_ID | rec_GeoID | rec_Match_dBID | rec_Desp_Lon | rec_Desp_Lat | rec_Side | rec_ARC_Address | rec_ARC_City |
|----|----------------|------------|-----------|----------------|----------|-----------|----------------|--------------|--------------|----------|-------------------------------------|--------------|
| 1 | Zipcode | M | 100 | A | 66.09594 | 18.443217 | | | | | PO Box 778 | MANATI |
| 2 | Zipcode | M | 100 | A | 66.0912 | 18.45487 | 0093 | | | | Almirante Norte, Carrera #140 Km. 1 | VEGA BAJA |
| 3 | Zipcode | M | 100 | A | 66.09382 | 18.21571 | 0093 | | | | Almirante Norte, Carrera #140 Km. 1 | VEGA BAJA |
| 4 | Zipcode | M | 100 | A | 66.10293 | 18.21571 | 00703 | | | | Bo. Serranito car #792 | AGUAS BUENAS |
| 5 | Zipcode | M | 100 | A | 66.10293 | 18.21571 | 00703 | | | | CAR 797 KM. 6.4, BQ. JAUQUETAS | AGUAS BUENAS |
| 6 | Zipcode | M | 100 | A | 66.10293 | 18.21571 | 00703 | | | | PO Box 1000 | AGUAS BUENAS |
| 7 | Zipcode | M | 100 | A | 66.09398 | 18.14093 | 00709 | | | | Bo. Asuncion car #162 | ABRINTO |
| 8 | Zipcode | M | 100 | A | 66.09398 | 18.14093 | 00709 | | | | El Bosque car #162 | ABRINTO |
| 9 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | 2 Aven Periferal 1202 | CAGUAS |
| 10 | Street Address | T | 100 | A | 66.09171 | 18.241103 | | | | | Av Rafael Cordero # 124 | CAGUAS |
| 11 | Street Address | T | 100 | A | 66.09382 | 18.21571 | | | | | Av Rafael Cordero # 124 | CAGUAS |
| 12 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Canabon Carr 156 Km.3 | CAGUAS |
| 13 | Zipcode | M | 100 | A | 66.09382 | 18.21571 | 00725 | | | | Bo. Canabon Carr 156 Km.3 | CAGUAS |
| 14 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Pueblo Cofre de Perote 2.7 | CAGUAS |
| 15 | Zipcode | M | 100 | A | 66.09342 | 18.213195 | 00725 | | | | Bo. Pueblo Cofre de Perote 2.7 | CAGUAS |
| 16 | Zipcode | M | 100 | A | 66.09316 | 18.213195 | 00725 | | | | Bo. Pueblo Urb. Saverana | CAGUAS |
| 17 | Zipcode | M | 100 | A | 66.09316 | 18.213195 | 00725 | | | | Bo. Pueblo Urb. Saverana | CAGUAS |
| 18 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. San Antonio sector Buena Vista | CAGUAS |
| 19 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Santo Domingo calle c #3 | CAGUAS |
| 20 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Tomás de Castro II | CAGUAS |
| 21 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Tomás de Castro II | CAGUAS |
| 22 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Turabo Arriba | CAGUAS |
| 23 | Zipcode | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Bo. Turabo Km. 40 #2 | CAGUAS |
| 24 | Street Address | M | 100 | A | 66.09382 | 18.21571 | | | | | Calle 31 Oeste Km.4 | CAGUAS |
| 25 | Street Address | M | 100 | A | 66.09162 | 18.213195 | 00725 | | | | Calle 31 Oeste Km.4 | CAGUAS |

Table: Smoke_alarm_data

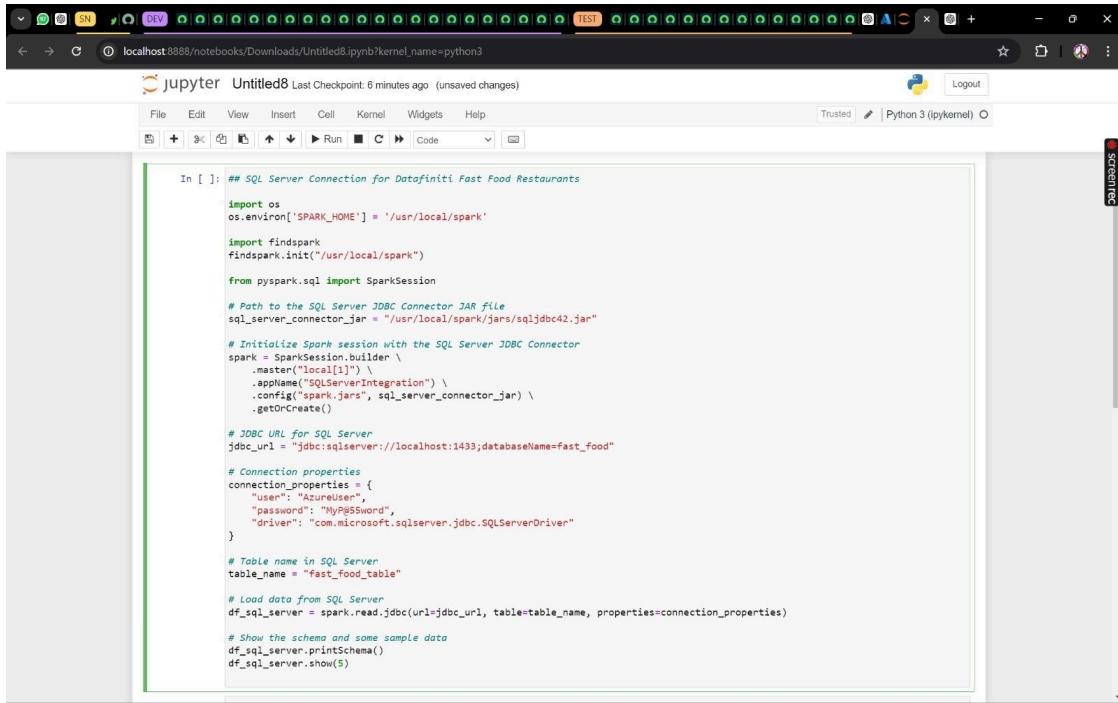
Below we have connected to our databases using jupyter notebook. We set up a spark session to connect to each of the databases.

Connecting to Hive Database:

- Set the Spark environment with `findspark` and initialize a `SparkSession` with Hive integration.
- Specify the Hive JDBC connector JAR file and the JDBC URL for the Hive database (`health_centres`).
- Load data from the Hive table (`my_table`), print its schema, and display sample data.

Connecting to SQL Server Database (Datafiniti Fast Food Restaurants):

- Set the Spark environment with `findspark` and initialize a `SparkSession` with SQL Server integration.
 - Specify the SQL Server JDBC connector JAR file and the JDBC URL for the SQL Server database (`fast_food`).
 - Load data from the SQL Server table (`fast_food_table`), print its schema, and display sample data.



The screenshot shows a Jupyter Notebook interface running on a local host. The notebook has a single cell (In [1]) containing Python code for connecting to a SQL Server database. The code uses the PySpark library to set up a SparkSession and read data from a SQL Server table named 'fast_food_table'. The code includes comments explaining the steps: setting the SPARK_HOME environment variable, initializing the SparkSession with the SQL Server JDBC connector, specifying the JDBC URL and connection properties (username, password, driver), and finally loading the data from the SQL Server table.

```
## SQL Server Connection for Datainiti Fast Food Restaurants
import os
os.environ['SPARK_HOME'] = '/usr/local/spark'

import findspark
findspark.init('/usr/local/spark')

from pyspark.sql import SparkSession

# Path to the SQL Server JDBC Connector JAR file
sql_server_connector_jar = "/usr/local/spark/jars/sqljdbc42.jar"

# Initialize Spark session with the SQL Server JDBC Connector
spark = SparkSession.builder \
    .master("local[1]") \
    .appName("SQLServerIntegration") \
    .config("spark.jars", sql_server_connector_jar) \
    .getOrCreate()

# JDBC URL for SQL Server
jdbc_url = "jdbc:sqlserver://localhost:1433;databaseName=fast_food"

# Connection properties
connection_properties = {
    "user": "AzureUser",
    "password": "MyP@ssw0rd",
    "driver": "com.microsoft.sqlserver.jdbc.SQLServerDriver"
}

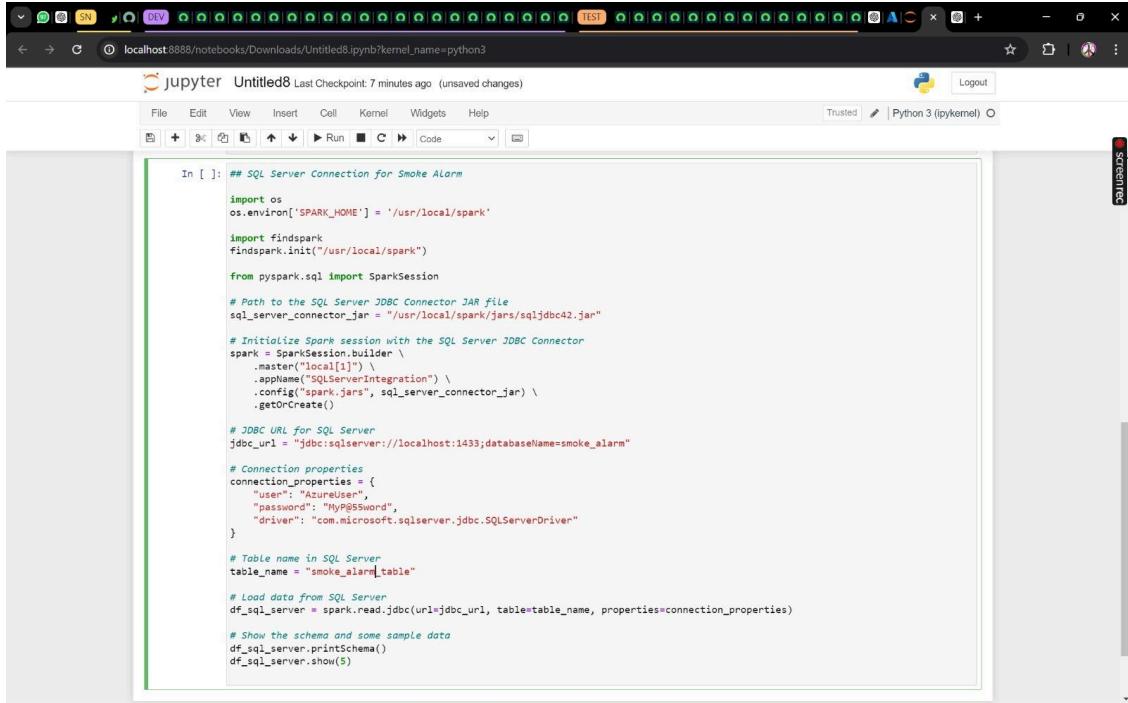
# Table name in SQL Server
table_name = "fast_food_table"

# Load data from SQL Server
df_sql_server = spark.read.jdbc(url=jdbc_url, table=table_name, properties=connection_properties)

# Show the schema and some sample data
df_sql_server.printSchema()
df_sql_server.show(5)
```

Connecting to SQL Server Database (Smoke Alarm):

- Set the Spark environment with `findspark` and initialize a `SparkSession` with SQL Server integration.
- Specify the SQL Server JDBC connector JAR file and the JDBC URL for the SQL Server database (`smoke_alarm`).
- Load data from the SQL Server table (`smoke_alarm_table`), print its schema, and display sample data.



```

In [ ]: ## SQL Server Connection for Smoke Alarm

import os
os.environ['SPARK_HOME'] = '/usr/local/spark'

import findspark
findspark.init("/usr/local/spark")

from pyspark.sql import SparkSession

# Path to the SQL Server JDBC Connector JAR file
sql_server_connector_jar = "/usr/local/spark/jars/sqljdbc42.jar"

# Initialize Spark session with the SQL Server JDBC Connector
spark = SparkSession.builder \
    .master("local[1]") \
    .appName("SQLServerIntegration") \
    .config("spark.jars", sql_server_connector_jar) \
    .getOrCreate()

# JDBC URL for SQL Server
jdbc_url = "jdbc:sqlserver://localhost:1433;databaseName=smoke_alarm"

# Connection properties
connection_properties = {
    "user": "AzureUser",
    "password": "My$85Word",
    "driver": "com.microsoft.sqlserver.jdbc.SQLServerDriver"
}

# Table name in SQL Server
table_name = "smoke_alarm_table"

# Load data from SQL Server
df_sql_server = spark.read.jdbc(url=jdbc_url, table=table_name, properties=connection_properties)

# Show the schema and some sample data
df_sql_server.printSchema()
df_sql_server.show(5)

```

From our merged dataset, we came up with the following visuals:

