BIG DATA (CSGY-6513)

SPRING 2025

SECTION-D

PROJECT PROPOSAL

Group members:

a) Rushabh Bhatt- rb5726
b) Utkarsh Mittal – um2113
c) Aryan Ajmera – aa12904

Project Abstract:

The Rodent Inspection dataset, provided by the NYC Department of Health and Mental Hygiene (DOHMH), contains detailed information on rat inspections across New York City. This project aims to analyze and extract meaningful insights from this dataset using Big Data technologies. By leveraging scalable data processing frameworks, we intend to identify patterns in rodent infestations, assess inspection trends, and provide data-driven recommendations for urban pest control measures.

Problem Statement:

Rodent infestations pose significant health risks and economic challenges in urban environments. Identifying high-risk areas and trends in rodent inspection data is crucial for targeted mitigation strategies. However, the large volume of data presents challenges in efficient processing, analysis, and visualization.

Objectives:

1. **Data Processing:** Efficiently handle and preprocess a large dataset using Big Data tools.
2. **Pattern Identification:** Analyze trends in rodent infestations based on geographic, temporal, and environmental factors.
3. **Visualization & Reporting:** Create interactive dashboards to communicate findings effectively.

Data Source Information:

- **Dataset Name:** Rodent Inspection
- **Data Source:** NYC Department of Health and Mental Hygiene (DOHMH)
- **Data Link:** Rodent Inspection Dataset
- **No. of datapoints:** Approx. 2.72 million records
- **Dataset file size:** Approx. 1GB
- **Number of Columns:** 25
- **Description:** The dataset contains detailed information about rodent inspections, including property locations, inspection dates, infestation status, and other environmental factors.

Proposed Technologies & Programming Language:

Programming Language:

- Python (for data processing and machine learning)
- SQL (for data querying and management)

Big Data Processing Frameworks:

- **Apache Spark (PySpark):** For distributed data processing and scalable analytics.
- **Hadoop HDFS:** For efficient storage and retrieval of large datasets.
- **HiveSQL:** For querying and managing large-scale tabular data.
- **Dask:** For parallel computing and efficient in-memory data processing.

Geospatial Analysis:

- **GeoPandas & Folium:** For geospatial analysis and visualization of rodent infestation hotspots.

Visualization Tools:

- **Matplotlib & Seaborn:** For statistical data visualization.

Reporting:

- **Tableau:** For creating dynamic, interactive visualizations with filters and geographic mapping.
- **Dash (Plotly):** A Python framework for building interactive web-based dashboards for real-time data visualization.
- **Power BI:** For integrating and presenting data insights in an intuitive and user-friendly format.
- **Streamlit:** For developing lightweight, interactive web applications to present data-driven insights.