PROJECT REPORT
BIG DATA MANAGEMENT

POST GRADUATE DIPLOMA
IN DATA ENGINEERING

PROJECT TITLE
URBAN MOBILITY AND INFRASTRUCTURE OPTIMIZATION USING OPENSTREETMAP DATA

SUBMITTED BY:
NIRAJ BHAGCHANDANI [G23AI2087]
SHUBHAM RAJ [G23AI2028]
BHAVESH ARORA [G23AI2126]

SUBMISSION DATE: 25TH DEC, 2024
DEPARTMENT OF AIDE
INDIAN INSTITUTE OF TECHNOLOGY, JODHPUR

MAPPING TOMORROW: INNOVATIVE APPROACHES TO URBAN INFRASTRUCTURE WITH OPENSTREETMAP DATA

Team Composition:

NIRAJ BHAGCHANDANI [G23AI2087]
SHUBHAM RAJ [G23AI2028]
BHAVESH ARORA [G23AI2126]

1. Introduction

The problem of urbanization is one of the most critical challenges faced by modern society. Cities are growing rapidly, leading to an increasing strain on existing infrastructure, a rise in traffic congestion, inefficiencies in public transportation systems, and challenges in emergency response management. Addressing these issues is essential not only to improve the quality of life for residents but also to ensure the sustainable development of urban areas.

This project, titled **"Urban Mobility and Infrastructure Optimization Using OpenStreetMap Data,"** was conceptualized as a practical response to these challenges. It leverages the power of **geospatial data analytics, big data technologies, and machine learning** to create an innovative platform designed to provide actionable insights for city planners, policymakers, and decision-makers.

The project is built on **OpenStreetMap (OSM)**, a crowd-sourced geospatial database that provides detailed information about roads, buildings, transport routes, and other infrastructure elements. OpenStreetMap is a preferred data source because of its open nature, global coverage, and constant updates by contributors worldwide. By combining this data with advanced technologies, this project sought to create a scalable, interactive, and real-time solution for urban infrastructure planning.

The importance of this project lies in its ability to deliver insights dynamically. Traditional urban planning tools often rely on static models or expensive proprietary datasets. In contrast, this project aims to use **real-time data processing** and **machine learning models** to create a living, adaptive platform. The insights generated by the platform can be used to make immediate decisions, whether to address traffic bottlenecks, optimize emergency vehicle routes, or plan public transport expansions.

2. Motivation

Urban areas worldwide face increasingly complex challenges. The rapid growth of cities has led to issues such as:

1. **Traffic Congestion**: Commuters spend hours stuck in traffic daily, leading to lost productivity and increased pollution levels.
2. **Inefficient Public Transportation**: Many cities lack adequate public transport systems, leaving large populations underserved.
3. **Delayed Emergency Responses**: Congested roads delay emergency services, potentially costing lives.
4. **Resource Allocation**: City planners struggle to prioritize infrastructure investments due to insufficient or outdated data.

This project was motivated by a desire to address these pressing issues using **data-driven approaches**. By analyzing real-world geospatial data, the project sought to uncover hidden patterns, identify inefficiencies, and recommend improvements.

Another key motivation was the potential for **cost savings** and **efficiency gains**. OpenStreetMap, as an open-source dataset, eliminates the need for expensive proprietary geospatial data. Moreover, by automating data analysis and visualization, the project reduces the manual effort required by city planners.

From an academic perspective, the project also served as an opportunity to apply **advanced technologies** in a real-world context. It allowed the team to explore the intersection of big data management, machine learning, and geospatial analytics.

3. Team Composition and Roles

This project was executed by a team of three members, each bringing unique skills and expertise to the table:

3.1 Niraj Bhagchandani [G23AI2087]

- **Role**: Backend Developer and Data Engineer
- **Responsibilities**:
  - Designed and implemented the data pipeline for extracting and processing OSM data.
  - Developed backend APIs using **Node.js** and integrated them with the frontend.
  - Managed database architecture, including the setup and optimization of **PostgreSQL with PostGIS**.

3.2 Shubham Raj [G23AI2028]

- **Role**: Full-Stack Developer
- **Responsibilities**:
  - Developed the frontend application using **React.js** and integrated it with backend APIs.
  - Designed and implemented interactive map visualizations using **Leaflet.js**.
  - Deployed the application on **AWS EC2** and configured Kubernetes for scaling.

3.3 Bhavesh Arora [G23AI2126]

- **Role**: Data Analyst and Machine Learning Specialist
- **Responsibilities**:
  - Conducted exploratory data analysis (EDA) on OSM datasets.
  - Built predictive models using **Scikit-learn** to forecast traffic congestion trends.
  - Generated data visualizations, including heatmaps, scatter plots, and road network graphs.

The collaborative effort and division of responsibilities ensured that the project was completed efficiently and effectively.

4. Objectives

The project had clearly defined objectives aimed at addressing urban mobility and infrastructure challenges:

4.1 Traffic Analysis

- **Goal**: Identify traffic congestion hotspots and visualize them through heatmaps.
- **Approach**: Analyze geospatial data to determine node density and traffic flow patterns.

4.2 Route Optimization

- **Goal**: Provide the shortest and most efficient routes for emergency vehicles.
- **Approach**: Develop algorithms that adapt dynamically to real-time traffic data.

4.3 Public Transport Assessment

- **Goal**: Evaluate existing public transport networks and identify underserved areas.
- **Approach**: Map bus and metro routes to visualize coverage and recommend new routes.

4.4 Infrastructure Planning

- **Goal**: Assist policymakers in prioritizing infrastructure investments.
- **Approach**: Highlight areas with inadequate road networks, insufficient public amenities, or high congestion.

These objectives were designed to create a comprehensive solution that addressed multiple facets of urban planning.

5. Scope

The project's scope encompassed several dimensions of data acquisition, processing, analysis, and visualization:

5.1 Data Acquisition

- Extract large-scale geospatial data from OpenStreetMap using APIs.
- Include data on roads, intersections, public amenities, and transport routes.

5.2 Data Processing

- Preprocess raw OSM data to clean and structure it for analysis.
- Use big data tools like **Hadoop** and **Spark** to handle large datasets efficiently.

5.3 Data Analysis

- Perform exploratory data analysis to uncover trends and patterns.
- Use machine learning models to predict future traffic conditions and identify potential bottlenecks.

5.4 Visualization

- Develop interactive maps using **Leaflet.js** to display insights visually.
- Include layers for traffic heatmaps, route suggestions, and public transport networks.

5.5 Deployment

- Host the application on AWS EC2 instances for scalability and reliability.
- Use Kubernetes for containerized deployment and dynamic scaling.

---

6. Challenges and Solutions

6.1 Managing Large Datasets

- **Challenge**: OSM datasets contain millions of records, making them difficult to process and analyze.
- **Solution**: Implemented distributed processing frameworks like Hadoop and Spark to parallelize tasks and handle large volumes of data.

6.2 Real-Time Data Processing

- **Challenge**: Ensuring low latency when processing and displaying real-time data.
- **Solution**: Used Dask for parallel processing and optimized backend API performance.

6.3 User Interface Design

- **Challenge**: Creating a simple and intuitive interface for non-technical users.
- **Solution**: Incorporated user feedback to refine the design and used Material-UI components for a professional appearance.

6.4 Scalability

- **Challenge**: Ensuring the application could handle increasing user traffic without downtime.
- **Solution**: Deployed the app on AWS Kubernetes, enabling automatic scaling based on demand.

Project Deliverables

1. **Interactive Web Application**:
    - Provides an interface to explore urban data via maps, heatmaps, and route suggestions.
    - Integrates real-time updates for traffic patterns and transport analysis.
    - Demonstrated via Live Frontend URL and Backend API URL.
2. **Benchmarking Study**:
    - Compared geospatial data processing tools like **PostGIS**, **MongoDB**, and **Hadoop/Spark**.
3. **GitHub Repository**:
    - Documented and open-source codebase for transparency and future enhancements.
    - Link: GitHub Repository

Technological Stack

Frontend:

- **React.js**: Built the responsive and interactive user interface.
- **Leaflet.js**: Rendered dynamic maps for visualization.
- **Material-UI**: Provided accessible and modern UI components.

Backend:

- **Node.js** with **Express.js**: Handled server-side API logic.
- **PostgreSQL + PostGIS**: Stored and queried geospatial data.
- **Python**:
    - **Pandas** and **Dask**: Real-time data processing.
    - **Scikit-learn**: Traffic prediction models.

Big Data Frameworks:

- **Hadoop/Spark**: Managed and processed large-scale OSM datasets.

Cloud Infrastructure:

- **AWS Services**:
    - **EC2**: Hosted the frontend and backend.
    - **S3**: Used for scalable data storage.
    - **Kubernetes**: Containerized deployment for scalability.

Features Delivered

1. Traffic Heatmaps

- Highlighted high-density traffic areas.
- Generated using **node density analysis** and **spatial clustering**.

2. Route Optimization

- Suggested shortest and safest paths for emergency services.
- Routes were calculated using algorithms that analyzed real-time congestion data.

3. Public Transport Recommendations

- Visualized gaps in public transport coverage.
- Provided suggestions for new bus routes and metro expansions.

4. Infrastructure Visualization

- Node density maps showcased urban planning inefficiencies.
- Included overlays for road networks, public amenities, and natural zones.

5. Real-Time Analytics

- Used live data for on-the-fly traffic predictions and visual updates.

Challenges and Solutions

1. Managing Large Datasets

- **Challenge**: OpenStreetMap data includes millions of records of roads, nodes, and buildings.
- **Solution**: Leveraged Hadoop and Spark for distributed data processing.

2. Real-Time Processing

- **Challenge**: Ensuring low latency while processing traffic data.
- **Solution**: Used Dask for parallel computation and efficient updates.

3. Scalability

- **Challenge**: Handling increasing user demand.
- **Solution**: Deployed AWS Kubernetes for auto-scaling.

4. Intuitive UI

- **Challenge**: Simplifying complex geospatial data for non-technical users.
- **Solution**: Built an easy-to-navigate interface using Leaflet.js and Material-UI.

Data Analysis and Insights

Exploratory Data Analysis (EDA)

1. **Node Density Heatmap**:
   - Revealed areas with high node concentrations, representing urban hubs.
2. **Road Length Distribution**:
   - Analyzed variations in road lengths to identify underdeveloped areas.
3. **Scatter Plots**:
   - Visualized geographic clustering of nodes.

Geospatial Data Visualizations

- **Road Networks**: Highlighted connectivity and intersections.
- **Feature Frequency**: Categorized roads (e.g., highways, residential).

Key Findings

1. Urban centers have significant congestion patterns around peak hours.
2. Certain public transport routes showed coverage gaps in suburban areas.
3. Emergency response times could be reduced by optimizing road usage during critical times.

7. Results

The project successfully achieved its objectives and delivered the following outcomes:

7.1 Fully Functional Application

- Provides dynamic visualizations of traffic patterns and public transport networks.
- Offers real-time route optimization for emergency vehicles.

7.2 Comprehensive Insights

- Generates actionable data visualizations, including heatmaps and node density graphs.
- Identifies underserved areas in public transport networks.

7.3 Scalable Infrastructure

- Hosted on AWS with Kubernetes for high availability and performance.

7.4 User Feedback

- Received positive feedback for its intuitive interface and practical insights.

---

8. Future Enhancements

The team identified several areas for future improvement:

8.1 IoT Integration

- Incorporate live data from traffic sensors to improve real-time accuracy.

8.2 Deep Learning Models

- Use neural networks for advanced traffic prediction and route optimization.

8.3 Global Adaptation

- Expand the platform to support multiple cities and regions.

8.4 Simulation Tools

- Add features for simulating infrastructure changes and assessing their impact.

Conclusion

This project demonstrated the transformative potential of open geospatial data when combined with advanced technologies. It serves as a blueprint for future initiatives aimed at building smarter cities and optimizing urban infrastructure. The success of this project highlights the importance of collaboration, innovation, and the practical application of technology in addressing real-world challenges.