HIGH-PERFORMANCE PARALLEL DATA PROCESSING FOR TIME SERIES AND GEOGRAPHIC ANALYSIS

Omkar Nagarkar
Purvil Patel
Atharva Jadhav
Sangram Jagtap

Introduction

Project showcases an approach to parallel data extraction using High Performance Computing (HPC) principles. Uses a distributed data processing system with MPI for parallel computations across multiple processors, augmented by OpenMP for intra-processor parallel processing.

Methods

- MPI (Message Passing Interface): For distributed parallel computations.
- OpenMP: For intra-processor parallel processing.
- C++ and Python: Implementation languages for processing and server management.
- Pandas and Matplotlib: For data congregation and
- Distributed File Processing Using MPI Parallel Data Cleaning and Processing with OpenMP

Dataset

NYC Parking Dataset: For the empirical dataset.
Size: 2.2 GB, 10 million rows x 43 columns

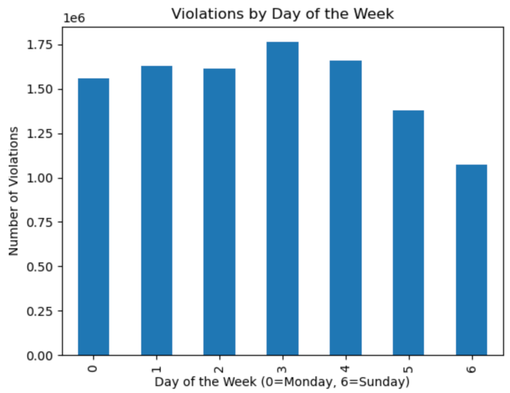
Relevant Features

- Issue Date
- Violation Location
- Street Name
- Violation County
- Violation Code

Procedure

- Data is divided and distributed across processors to minimize idle time and maximize read operations.
- Each processor works on its local data to reduce latency and network traffic for higher efficiency.
- Ensures even distribution of data among processors to prevent delays and achieve optimal utilization.
- Within each processor, OpenMP facilitates multi-threading, allowing multiple data operations to be performed concurrently.
- Multi-threading significantly decreases the time required for data cleaning and transformation.
- System scales effectively by utilizing processor cores to their full potential, processing records in parallel.

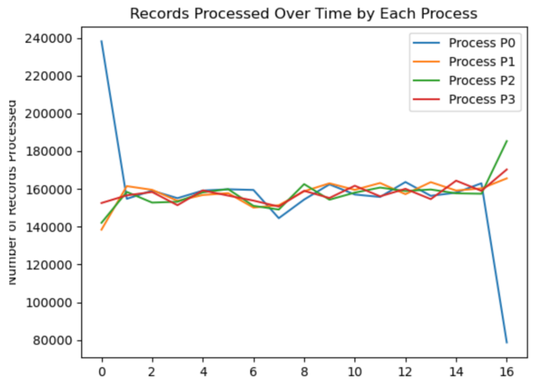
Results



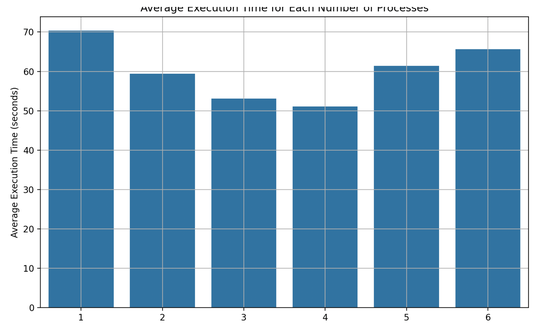
Time Analysis

Geographic Analysis

Performance 1



Performance 2



Conclusion

This project successfully integrates high-performance computing techniques to efficiently process large-scale datasets. The combination of MPI and OpenMP leads to a substantial reduction in processing time and improves the system's scalability and flexibility for complex analyses