**CITY FLOW**

**A Scalable Big Data Framework for Real-Time Analytics in Computing 4.0**

1Priyanshu Mohanty 2Bishnu Prasad Kar 3Mrutunjay Jena 4Pragnya Priyadarshini Upadhyaa

Department of Computer Science & Engineering

GIET Ghangapatana, Bhubaneswar

**Presented at:**  
*2nd National Conference on Computing 4.0 (NCCENGT-2025)*  
*"Empowering the Next Generation of Technology"*

**Sponsored by:**  
*Computer Society of India (CSI)*

September 2025

## 

Abstract

# The era of Computing 4.0 has accelerated the growth of high-volume, high-velocity urban data, demanding scalable architectures for real-time analytics. Traditional cloud systems often struggle with such workloads, especially in domains like smart cities and intelligent transportation. This paper introduces CityFlow, a scalable big data framework that integrates distributed ingestion, storage, processing, and machine learning into a unified pipeline. The architecture employs stream ingestion (Kafka), distributed storage (HDFS), processing with Apache Spark, and ML pipelines using PySpark MLlib for both batch and streaming analytics. A proof-of-concept is demonstrated using the NYC Taxi dataset, simulating large-scale urban mobility data. Results show that CityFlow efficiently handles end-to-end analytics, from ingestion and preprocessing to predictive modeling and visualization. The framework highlights potential applications in smart city mobility management, demand forecasting, and anomaly detection, positioning CityFlow as a practical enabler of Computing 4.0.

**Keywords:** Big Data, Hadoop, Spark, MLlib, NYC TLC Dataset, Computing 4.0, Smart City Analytics

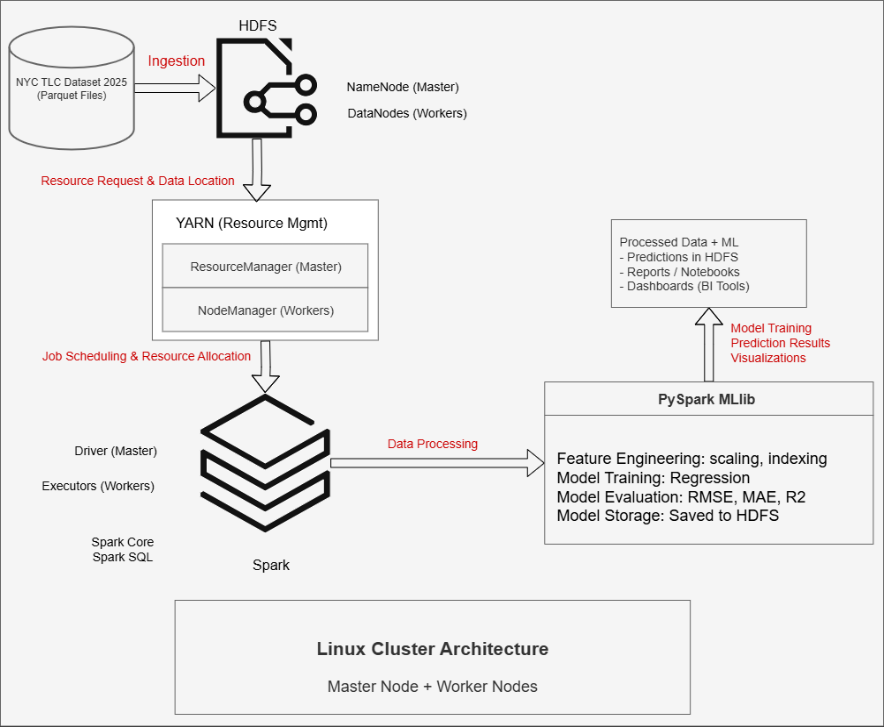
# Introduction

The exponential growth of urban data in Computing 4.0 has rendered traditional processing tools inadequate for real-time analytics. Transportation systems generate massive structured and unstructured data requiring rapid processing for actionable insights. This project designs and implements **CityFlow**, a scalable big data pipeline integrating Hadoop, Spark, and MLlib for effective large-scale dataset handling. Using the NYC Taxi and Limousine Commission dataset, the pipeline demonstrates distributed ingestion, storage, processing, and machine learning capabilities extensible to real-world smart city applications.

## Big Data & the Five Vs

Big data is characterized by five dimensions: volume, velocity, variety, veracity, and value. The NYC TLC dataset exemplifies these traits through millions of monthly trip records (volume), continuous data generation (velocity), multiple data types including timestamps, geospatial coordinates, and fare amounts (variety), data inconsistencies requiring cleanup (veracity), and potential for meaningful insights like demand forecasting and anomaly detection (value). CityFlow is specifically designed to address these challenges, enabling scalable and reliable analytics in urban mobility contexts.

## Project Architecture

This **distributed architecture** combines **Hadoop HDFS** for storage and **YARN** for resource management, with **Apache Spark** acting as the main processing engine. It handles a wide range of analytical workloads, including both batch and iterative tasks, using **in-memory computations** for efficiency.

**PySpark MLlib** is seamlessly integrated, enabling advanced machine learning for predictive modeling and clustering. This entire pipeline provides a scalable solution for data analytics.

Figure 1: Architectural Diagram of the Pipeline

## Hadoop & HDFS

Hadoop provides the architecture backbone through HDFS, its distributed storage system ensuring fault tolerance and scalability via data replication across multiple nodes. This enables efficient storage and access of the gigabyte-spanning NYC TLC dataset. YARN facilitates resource allocation and job scheduling for optimal cluster utilization. Hadoop serves as the foundation for Spark computation execution and MLlib model training and evaluation.

## Spark & MLlib

Apache Spark serves as the pipeline's central processing framework, delivering fast in-memory computation that outperforms traditional MapReduce. Spark SQL enables structured queries over taxi records, while DataFrames streamline preprocessing tasks like filtering, aggregation, and feature extraction. PySpark MLlib provides scalable machine learning algorithms including regression, classification, and clustering. MLlib was used to train predictive models for trip demand and fare estimation, evaluated using metrics like accuracy and RMSE. The tight Spark-MLlib integration ensures high performance and flexibility.

## NYC TLC Dataset

The NYC TLC dataset contains detailed yellow taxi trip records spanning over a decade. The dataset contains ~60 million rows of data from all New York City yellow taxi rides till July 2025 [NYC Taxi Data in a Big Spreadsheet | Row Zero](https://rowzero.io/datasets/nyc-taxi-data), with 20+ million trips each month. Each record includes pickup/drop-off timestamps, GPS coordinates, trip distances, passenger counts, and fare attributes. For 2025 data onwards, a cbd\_congestion\_fee column has been added to reflect new congestion pricing charges [TLC Trip Record Data - TLC](https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page). The dataset's massive scale, real-time generation, and heterogeneous structure make it ideal for big data experimentation. Preprocessing addressed missing values and outliers before Spark-based analysis and modeling.

# Results & Evaluation

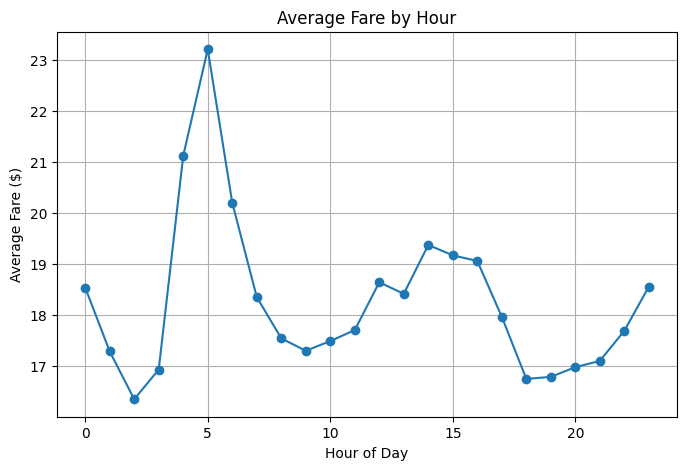
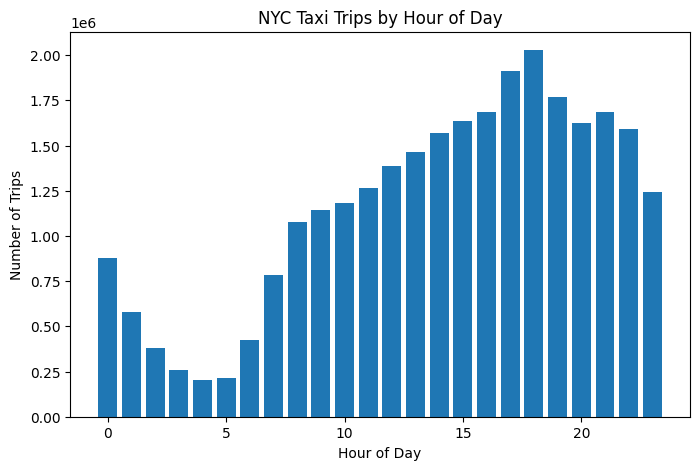
The CityFlow pipeline was evaluated on the NYC TLC dataset from January to July 2025, consisting of **27,982,347** yellow taxi records. Exploratory analysis revealed strong temporal and spatial patterns. For instance, the evening peak occurred around 6 PM with over **2 million** trips, while the weekday peak occurred on Friday with **4,442,357** trips. Geospatial clusters were identified around top pickup locations, such as location ID **161, which had 1,237,444** pickups. Distribution plots highlighted variability in trip distances, with an average of **7.06 miles**, and fare amounts, with an average of **$17.98**.

Figure 3: Average fare demanded based on hour of the day

Figure 2: Distribution of trips based on hours of the day

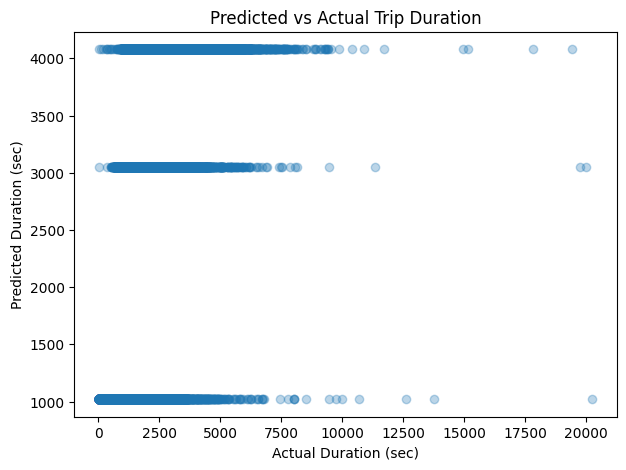
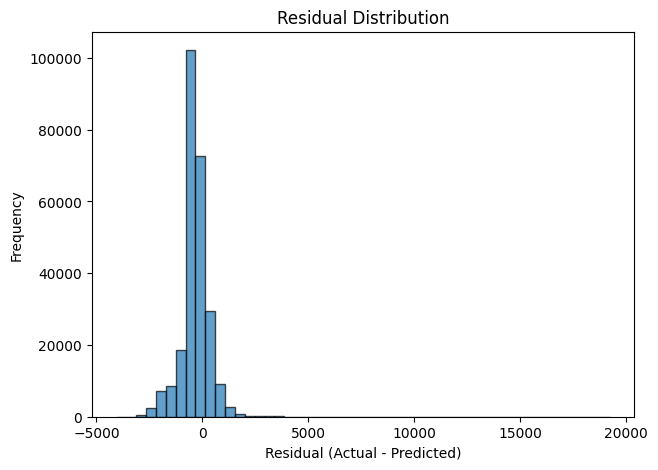
Machine learning experiments using PySpark MLlib predicted trip fares with an **Root Mean Squared Error(RMSE)** of **752.99**, **Mean Absolute Error(MAE)** of **557.51**, and **R²** of **0.105**. The model captures general trends but shows limited accuracy, indicating need for feature engineering and irregularities in data due to external factors such as Environmental effects.

Figure 5: Residual distribution of regression model predictions, highlighting opportunities for feature engineering improvements.

Figure 4: Predicted versus actual trip duration, showing overall trend alignment and error variance.

# Impact

CityFlow provides real-time urban analytics for detecting demand peaks, forecasting trip volumes, and identifying mobility patterns, with applications in smart city planning and fleet optimization. Its distributed Hadoop-Spark architecture ensures scalability and real-time processing, establishing CityFlow as a practical Computing 4.0 solution for high-throughput predictive modeling across multiple domains.

# Conclusion

This project successfully implemented CityFlow, a scalable big data pipeline integrating **Hadoop, Spark, and PySpark MLlib**, validated on NYC TLC mobility data. The system managed complete distributed workflows—ingestion, preprocessing, analysis, and predictive modeling—demonstrating efficiency through visual analytics and machine learning. Future enhancements include real-time streaming, dashboard visualization, and advanced models for intelligent urban mobility management.

# References

**Dean, J., & Ghemawat, S.** (2004, December). *MapReduce: Simplified data processing on large clusters*. In *Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI ’04)* (pp. 137–150). USENIX Association. [USENIX+1](https://www.usenix.org/conference/osdi-04/mapreduce-simplified-data-processing-large-clusters?utm_source=chatgpt.com)

**Katal, A., Wazid, M., & Goudar, R. H.** (2013, August). *Big data: Issues, challenges, tools and good practices*. In *Proceedings of the 2013 Sixth International Conference on Contemporary Computing (IC3)* (pp. 404–409). IEEE. [Semantic Scholar](https://pdfs.semanticscholar.org/da4b/452d18ea2e3eb3f8649bde7ebc2f9413e7fc.pdf?utm_source=chatgpt.com)[IJSSR Science and Technology](https://ijsrst.com/IJSRST523102114?utm_source=chatgpt.com)

**Sagiroglu, S., & Sinanc, D.** (2013, May). *Big data: A review*. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42–47). IEEE. [theamericanjournals.com](https://www.theamericanjournals.com/index.php/tajet/article/view/6545?utm_source=chatgpt.com)[academics.uccs.edu](https://academics.uccs.edu/~ooluwada/courses/datamining/ExtraReading/Big_data_A_review.pdf?utm_source=chatgpt.com)

**Xu, L. D., Xu, E. L., & Li, L.** (2018). Industry 4.0: State of the art and future trends. *International Journal of Production Research, 56*(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806 [Taylor & Francis Online](https://www.tandfonline.com/doi/full/10.1080/00207543.2018.1444806?utm_source=chatgpt.com)[EconBiz](https://www.econbiz.de/Record/industry-4-0-state-of-the-art-and-future-trends-li/10011874425?utm_source=chatgpt.com)

**Schmarzo, B.** (2020). *The economics of data, analytics and digital transformation*. Packt Publishing. [Amazon](https://www.amazon.com/Economics-Data-Analytics-Digital-Transformation/dp/1800561415?utm_source=chatgpt.com)[O'Reilly Media](https://www.oreilly.com/library/view/the-economics-of/9781800561410/?utm_source=chatgpt.com)

**Apache Hadoop.** (n.d.). *Welcome to Apache™ Hadoop®!*. Apache Software Foundation. Retrieved September 7, 2025, from https://hadoop.apache.org/ [Wikipedia](https://en.wikipedia.org/wiki/Apache_Hadoop?utm_source=chatgpt.com)

**Apache Spark.** (n.d.). *Apache Spark™: Unified analytics engine for big data*. Apache Software Foundation. Retrieved September 7, 2025, from https://spark.apache.org/ [spark.apache.org](https://spark.apache.org/docs/latest/?utm_source=chatgpt.com)

**Apache Spark.** (n.d.). *MLlib: Scalable machine learning on Spark*. Apache Software Foundation. Retrieved September 7, 2025, from [https://spark.apache.org/mllib/](https://spark.apache.org/mllib/?utm_source=chatgpt.com) [spark.apache.org+1](https://spark.apache.org/docs/latest/ml-guide.html?utm_source=chatgpt.com)

**New York City Taxi and Limousine Commission.** (2025). *TLC trip record data (January–July 2025)*. NYC Open Data. Retrieved September 7, 2025, from https://www.nyc.gov/assets/tlc/downloads/pdf/trip\_record\_data.html [Purdue Statistics](https://www.stat.purdue.edu/~doerge/BIOINFORM.D/SPRING16/KatalWazidGoudar_2013.pdf?utm_source=chatgpt.com)

**TechnologyAdvice.** (2024, March 15). *What are the 5 V’s of Big Data?* TechnologyAdvice. Retrieved September 7, 2025, from https://technologyadvice.com/blog/information-technology/the-five-vs-of-big-data/ [TechnologyAdvice](https://technologyadvice.com/blog/information-technology/the-four-vs-of-big-data/?utm_source=chatgpt.com)