Milestone 4: Final Report

I- Introduction

In the ever-evolving landscape of the modern sharing economy, ride-sharing platforms have emerged as prominent players in the urban transportation paradigm. Services like Uber and Lyft have not only transformed the way people move from one point to another but also have created unique income-generating opportunities for countless individuals across the globe. At the heart of this transformation lies the intricate dance of supply and demand, where drivers aim to be in the right place at the right time to maximize their earnings while passengers seek timely, convenient, and cost-effective transportation.

In this dynamic ecosystem, an emerging field of research and practice has arisen, one that focuses on utilizing data analytics techniques to optimize the positioning of ride-sharing drivers during periods of low demand. To build upon the existing body of knowledge, our project commenced with a comprehensive literature review. We explored and synthesized the approaches employed by researchers and practitioners who have undertaken similar projects in the realm of ride-sharing analytics. This literature review not only informed our methodology but also provided valuable insights into the challenges and opportunities encountered by others in the field.

II- Related Work

The exploration of existing literature and research has played a pivotal role in shaping the methodologies, approaches, and algorithmic choices for our project. The literature review served as a foundation for understanding the current state of knowledge in optimizing the positioning of ride-sharing drivers during periods of low demand. In this section, we discuss key systems and papers that have significantly influenced our project.

The literature review revealed valuable insights into existing research and practices, offering a diverse range of data analytics techniques, algorithms, and innovative approaches. Notably, the page rank algorithm idea and K-means algorithm, central to our project, were derived from this comprehensive review.

The integration of the K-means algorithm, inspired by Papers 1 and 3, plays a central role in our project's objective of strategically positioning ride-sharing drivers. Uber's real-time data analysis and prediction of cab pickup locations using K-means clustering offer valuable insights into enhancing the accuracy of identifying optimal driver locations during periods of low demand.

Additionally, the utilization of the page rank algorithm, drawn from Paper 8, contributes to our approach in identifying interesting locations and travel sequences from GPS trajectories. The tree-based hierarchical graph introduced in this paper aligns with our objective of enhancing the efficiency of ride-sharing drivers by recommending strategic positions based on a comprehensive understanding of location dynamics.ive review.

While the primary focus has been on Papers 1, 3, and 8, additional papers offer valuable insights. Paper 2 emphasizes empirical data analytics and visualization for Uber services, providing crucial patterns for supply and demand. The use of R programming and machine learning models aligns with our goal of data-driven decision-making. Paper 4 introduces a methodology to identify high-activity areas in urban settings using taxi GPS trajectories, contributing to our project's objective of enhancing driver efficiency. Paper 5 provides insights into passenger-finding strategies from taxi datasets, employing

L1-Norm SVM for feature selection, offering general principles applicable to ride-sharing. Paper 6's use of the Sliding Window Ensemble Framework for recommending urban hotspots provides valuable reference for accurate forecasts. Finally, the trajectory data mining survey in Paper 7 offers insights into diverse data types and efficient storage, contributing to our understanding of handling trajectory data.

Collectively, these papers address challenges such as data analysis complexities, feature selection, and accurate forecasting. Integrating lessons from these works, our project aligns methodologies from real-time data analysis, predictive modeling, and trajectory data mining to optimize ride-sharing driver positioning during low-demand periods, ensuring continuous quality pickups throughout their shifts.

III- Project Design

Drawing from the insights gained through the literature review, we established a robust plan through the designing phase. The literature review served as a roadmap, guiding us through the complexities of big data analytics in the context of ride-sharing optimization. It delineated key milestones, from data acquisition and preprocessing to the application of algorithms and the subsequent visualization of results. Our project leveraged the power of Apache Hadoop as the foundational framework for distributed storage and processing of large-scale data. Hive, a data warehousing and SQL-like query language tool, was employed to facilitate efficient data management and retrieval. Apache NiFi and self-designed PySpark script played a pivotal role in the data ingestion process, ensuring seamless integration and flow of information within our analytics pipeline.

The inclusion of Apache Spark significantly enhanced our analytical capabilities. Spark, known for its in-memory processing and iterative computation, allowed for the rapid and efficient execution of complex algorithms on large datasets such as the k-means algorithm and page rank that were key algorithms in determining hotspots for uber drivers. This parallel processing framework played a crucial role in accelerating our data analytics tasks, contributing to the overall efficiency of our project.

Two sophisticated algorithms, K-means and Page Rank, were carefully chosen and implemented to analyze driver positioning during off-peak hours, forming a robust data-driven foundation for our optimization efforts.

In the K-means algorithm, the selection of the parameter "k" played a crucial role in achieving optimal results for cluster formation. Leveraging the elbow method, we systematically identified the appropriate number of clusters by evaluating the variance within each cluster. This meticulous approach ensured that the K-means algorithm efficiently categorized the data into meaningful clusters, allowing for a nuanced understanding of spatial patterns during low-demand periods.

Turning our attention to the Page Rank algorithm, our approach involved creating a graph where each node represented a distinct entity within the context of ride-sharing dynamics. Nodes represented each location id(an area). The Page Rank algorithm, renowned for its application in web link analysis, was adapted to assess the importance and influence of each node within the taxis movements. This adaptation allowed us to uncover critical hotspots and strategically significant locations based on the interconnectedness and relevance of nodes in the graph.

The fusion of K-means and Page Rank algorithms provided a comprehensive analysis of optimal driver positioning. The K-means algorithm identified spatial clusters, while the Page Rank algorithm quantified the significance of each location within the broader network. This combined approach not only

enriched our understanding of driver behavior during different hours during the week but also yielded valuable insights into the strategic deployment of drivers to maximize service efficiency.

In the subsequent phases of our project, these algorithmic analyses paved the way for the generation of heatmap through Tableau. This visual representation vividly showcased the identified hotspots and spatial clusters, providing an intuitive and actionable resource for ride-sharing platforms and drivers alike.

Tableau not only facilitated a comprehensive exploration of our findings but also empowered stakeholders to make informed decisions based on the insights derived from our analysis. See picture below for the system built for the project.

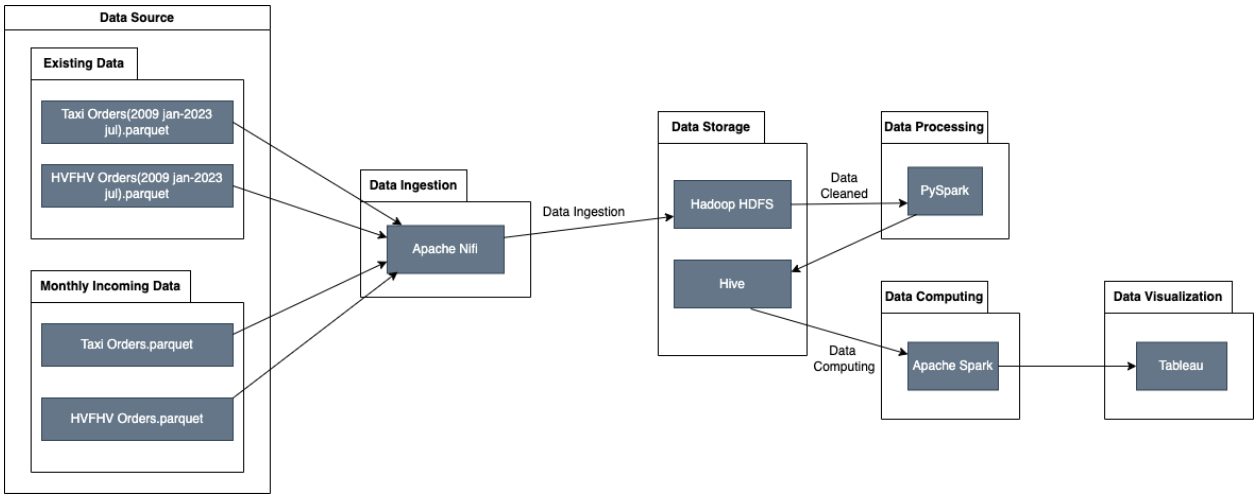


Figure. 1 Pipeline Diagram

IV- Big data Challenges

Our initial steps were marked by the formidable task of installing and configuring essential tools on Linux. Apache Flume, Apache Hadoop, Apache Hive, and Apache NiFi demanded meticulous setup, including modifications to configuration files. This process, though critical for the project's foundation, consumed significant time and highlighted the intricate nature of deploying a robust data analytics pipeline.

We also ran into raw data inconsistencies and cleaning dilemmas. Indeed, handling raw data revealed inherent inconsistencies, including variations in schema and column names across different files. This required a comprehensive data cleaning approach to harmonize the data and address discrepancies. The intricacies of these inconsistencies added an extra layer of complexity, demanding careful attention to detail in the data preparation phase.

In addition to that, while the PageRank algorithm stood out as a potent analytical tool, it presented a unique challenge. The original algorithm, designed for diverse applications, lacked seamless alignment with our project's specifics. Notably, the algorithm did not account for self-loop occurrences, a prevalent phenomenon in our database reflecting Uber orders starting and ending in the same location area. To address this, we undertook the challenging task of customizing the PageRank algorithm to suit our project's nuanced requirements.

Our journey encountered limitations imposed by hardware capabilities, particularly in memory-intensive operations. To circumvent potential crashes due to memory constraints, strategic modifications were made to the PySpark script. This adaptation allowed the division of jobs into manageable batches, ensuring the efficient processing of substantial data files without compromising system stability.

Automating the pipeline for seamless, self-sustained operation brought its own set of challenges. Extensive directory and file monitoring were essential components, yet the automation process was not without hurdles. Small bugs occasionally surfaced during the writing process, underscoring the need for meticulous attention to detail and continuous refinement.

Finally, the absence of dedicated servers posed a significant challenge, especially when dealing with large datasets. The computational demands of processing extensive data without the support of dedicated servers added complexity to the project's execution, necessitating innovative solutions to optimize computational resources.

Overall, the challenges stemming from raw data intricacies encompassed installation hurdles, algorithmic customization, memory management, automation intricacies, and the meticulous handling of data inconsistencies. Addressing these challenges was paramount in establishing a robust foundation for our data analytics endeavors.

V- Project Outcomes and Reflection

The primary objective was to identify strategic locations for ride-sharing drivers during periods of low demand, aiming to enhance overall service efficiency and maximize drivers' earnings.

The data analytics pipeline, featuring Apache NiFi, Hadoop, Spark, Hive, and advanced algorithms, successfully processed and cleaned the data by removing null values, deleting unnecessary columns and making sure all the columns in different raw data files were consistent. The implementation of the K-means clustering algorithm categorized the data based on time and divided a week into several clusters. Subsequently, the Page Rank algorithm was applied to each cluster, resulting in the creation of detailed heatmaps through Tableau. The heatmap visually highlights hotspots, providing clear insights into optimal driver locations during different hours during the week.

The generated heatmap not only illustrates high-demand areas but also offers a nuanced understanding of time-sensitive patterns. This visualization serves as actionable tools, allowing ride-sharing drivers to make informed decisions about their positioning to maximize their earnings efficiently.

Despite the absence of the initially planned frontend interface due to time constraints that was factored by the complexity of the pipeline , the achieved outcomes significantly contribute to addressing the project's core challenge. The detailed heatmap represents a tangible and valuable asset, providing ride-sharing platforms and drivers with practical insights into optimizing service efficiency. The project underscores the importance of advanced data analytics techniques in deriving actionable information from large-scale datasets.

The outcomes, while meeting the fundamental project goal, also reveal the potential for future developments. The absence of the frontend component is acknowledged as an area for extension, and the project lays a solid foundation for further exploration in optimizing driver positioning and overall service quality.

Cited Work

Literature review: Milestone 1

[1] T. M. Gunawardena and K. P. N. Jayasena, "Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment," 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2020, pp. 1-6, doi: 10.1109/ICITR51448.2020.9310851.

[2] K. M. Sudar, P. Nagaraj, V. Muneeswaran, S. K. Jeevana Swetha, K. M. Nikhila and R. Venkatesh, "An Empirical Data Analytics and Visualization for UBER Services: A Data Analysis Based Web Search Engine," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-6, doi: 10.1109/ICCCI54379.2022.9741016.

[3] R. Srinivas, B. Ankayarkanni and R. S. B. Krishna, "Uber Related Data Analysis using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1148-1153, doi: 10.1109/ICICCS51141.2021.9432347.

[4] F. Bai, H. Feng and Y. Xu, "Identifying the Hotspots in Urban Areas Using Taxi GPS Trajectories," 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Huangshan, China, 2018, pp. 900-904, doi: 10.1109/FSKD.2018.8686932.

[5] B. Li et al., "Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset," 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), Seattle, WA, USA, 2011, pp. 63-68, doi: 10.1109/PERCOMW.2011.5766967.

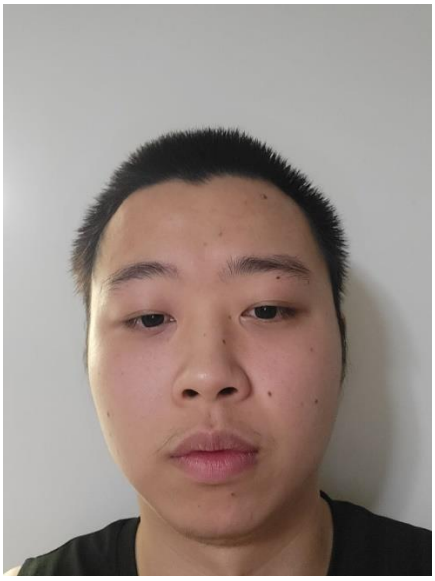
[6] K. Zhao, D. Khryashchev and H. Vo, "Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2723-2736, 1 June 2021, doi: 10.1109/TKDE.2019.2955686.

[7] Z. Feng and Y. Zhu, "A Survey on Trajectory Data Mining: Techniques and Applications," in IEEE Access, vol. 4, pp. 2056-2067, 2016, doi: 10.1109/ACCESS.2016.2553681.

[8] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web (WWW '09). Association for Computing Machinery, New York, NY, USA, 791–800. https://doi.org/10.1145/1526709.1526816

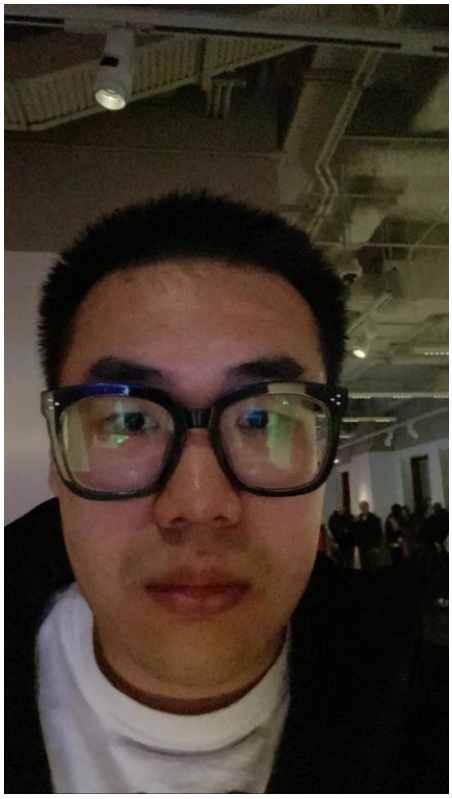
[9] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web (WWW '09). Association for Computing Machinery, New York, NY, USA, 791–800. https://doi.org/10.1145/1526709.1526816

Members



Chuwen Sun

Animan Yann



Gaofeng Zhu

Vanday Bundu