COLUMBIA UNIVERSITY
School of Professional Studies

Ride-sharing Data Platform

Group 11:
Jiahao Xu | Sheng Luo | Yanting Wu |
Yixuan Wang | Zhiyuan Tang `

Background and Definition of the Business Use

- In recent years, the most significantly evolutions in the transportation industry has been the rise of ride-sharing companies.
- Due to the ease of use and intuitiveness, ride-sharing companies expanded their services in popularity and now offering services across the globe.
- All of the success of the ride-sharing companies is driven by data analytics. With the empowerment of data analytics, ridesharing companies use big data for everything from surge pricing strategy to assign drivers to commute.

Business Plan & Customer interactions

- **We are aiming to develop an up-to-date NYC yellow cab & green cab data platform with customized features to provide services for data analytics needs in ride-sharing industries.**
- Customers emphasis the features they value and easily access those features by implementing our data security agreement
- With the help of our website, yellow taxi & green taxi companies can utilize data analytics to optimize their pricing strategies, enhance their operating efficiency, and improve user experience.

Data Source Specification and Procurement Details

- The NYC taxi and Limousine Commission (TLC) has released its Trip Record data to the public for research and study purposes.
- The data used in the analysis includes information on yellow and green taxi trips in NYC.
- The data includes fields capturing pick-up and drop-off dates/times, locations, trip distances, itemized fares, rate types, payment types, passenger counts, dispatching base license number, pick-up date, time, and taxi zone location ID.
- The data was collected and provided to the NYC taxi and Limousine Commission (TLC) by authorized technology providers for yellow and green taxi trips.
- The data is publicly available and can be accessed through the TLC website.
- The TLC does not guarantee the accuracy of the data for yellow and green taxi trips, but performs routine reviews and enforcement actions to ensure complete and accurate information.

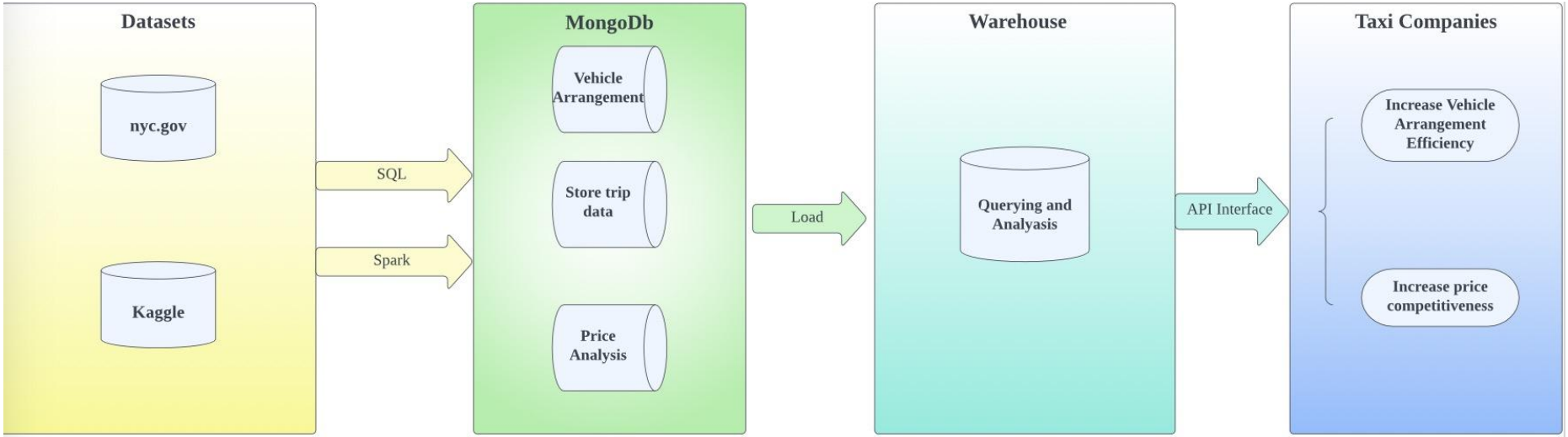
Project Design & Rationale

SQL: Group by companies, we can query the total number of order in each time and each location. We can also query the percentage of passengers for giving tips within each condition.

Spark: The Taxi trip data can be transformed using Spark's data processing API. Then the data can be aggregated and analyzed using Spark's SQL query interface to obtain information about the relationship between the tip amount and the number of riders.

MongoDB: Query language that can be used to search trip data and gain insights into ride patterns, peak hours, and popular routes. We can perform geospatial querying based on pick up and drop-off points.

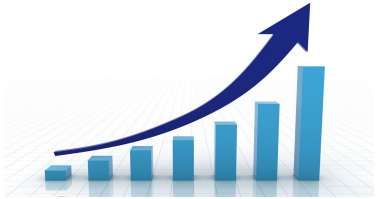
ETL Process Pipeline Diagram



Scalability and Cost Implications

- **Scalability:**

  1. Increase brand awareness and operational efficiency in a competitive market.
  2. Meet customers satisfaction including lessening the waiting time and informing customers tipping policy.
  3. Improve drivers experience by informing them locations of peak hours.
  4. Providing both drivers and customers on weather related information.



- **Cost implications:**

By utilizing AWS, we estimate that the monthly cost for companies is **6559** dollars, including the implementation of API Gateway, Aurora PostgresSQL, DocumentDB, EC2, and Simple Storage Service.

Data Quality Dimensions and Licensing

Data Quality Dimensions

- The platform aims to increase vehicle arrangement efficiency and price competitiveness.
- Includes extensive amount of data to ensure the accuracy and reliability of the platform.
- Durability is high since our team can add current data into the platform. Also, we provide personalized technical support.
- Reliable storage of data

Licensing

- Obtain a license to use the platform
- Negotiate the term of license agreement
- Sign written contract

Conclusion and Future Recommendations

Conclusion:

- With the help of three technologies, we are able to successfully design a platform that can be used globally by taxi related companies.
- Companies are able to perform price analysis, increase operating efficiency and vehicle arrangement efficiency
- Cloud services are utilized as we aim to provide secure and reliable storage data
- Companies are able to gain customer insights including customers' behavior, needs, and preferences.

Recommendation:

- For companies that concerns about data securities, our platform put securities in a priority position. We will implement certain protocols and access control.
- Companies are able to use our platform to perform marketing analysis by creating visualizations of our database.
- Competitive analysis can also be provided by using publicly available information to ensure that the companies stay ahead of the game.