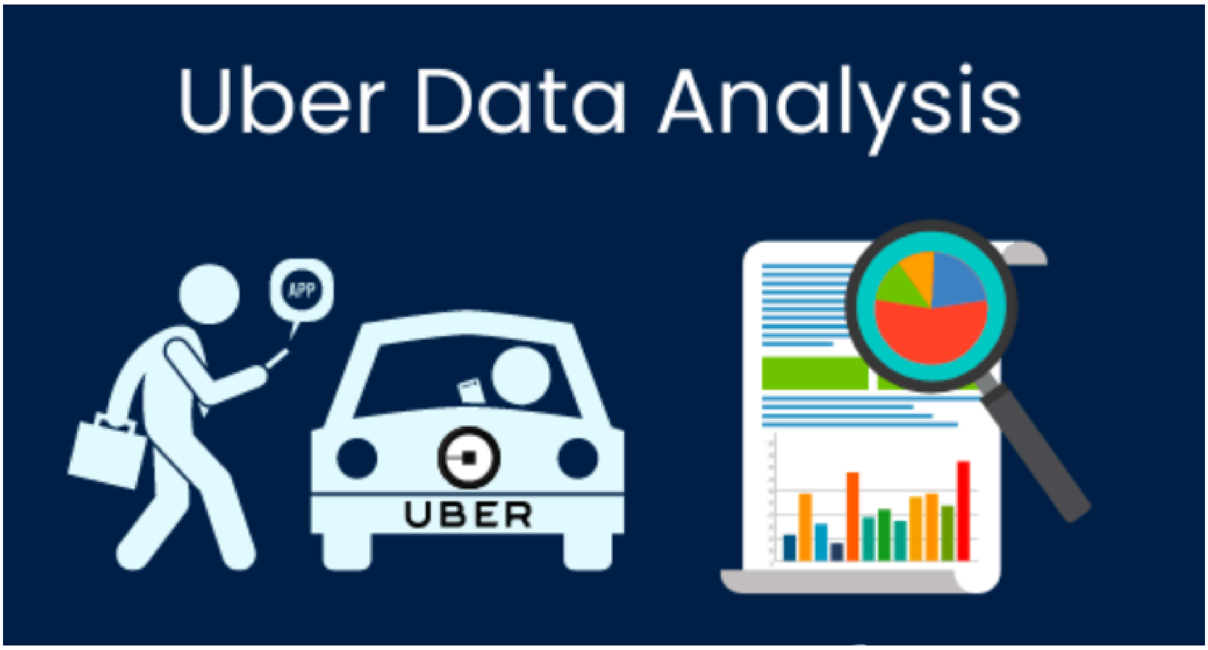
Introduction

PySpark’s ability to handle large datasets makes it a valuable tool for data processing and analysis in every industry. In this project, we will utilize PySpark to analyze uber data and gain insights from city supply and demand data by cleaning, and querying the data to feature engineering, we will explore the various capabilities of PySpark and how it can be applied to real-world problems in the demand & supply chain.

Insights from City Supply and Demand Data.

This data project can be used as a take-home assignment to learn Pyspark and Data Engineering. In this project, we will be using PySpark, a powerful Python library for big data processing, to analyze a dataset from **Uber**. The goal of this project is to gain insights into the demand and supply industry by cleaning, transforming, and analyzing the data using PySpark.



Uber Data Analysis

Assignment:

1. Which date had the most completed trips during the two-week period?
2. What was the highest number of completed trips within a 24-hour period?
3. Which hour of the day had the most requests during the two-week period?
4. What percentages of all zeroes during the two-week period occurred on weekend (Friday at 5 pm to Sunday at 3 am)?

   Tip: The local time value is the start of the hour (e.g. `15` is the hour from 3:00 pm - 4:00 pm)

5. What is the weighted average ratio of completed trips per driver during the two-week period? **Tip**: “Weighted average” means your answer should account for the total trip volume in each hour to determine the most accurate number in the whole period.
6. In drafting a driver schedule in terms of 8 hours shifts, when are the busiest 8 consecutive hours over the two-week period in terms of unique requests? A new shift starts every 8 hours. Assume that a driver will work the same shift each day.
7. **True or False**: Driver supply always increases when demand increases during the two-week period. **Tip**: Visualize the data to confirm your answer if needed.
8. In which 72-hour period is the ratio of Zeroes to Eyeballs the highest?
9. If you could add 5 drivers to any single hour of every day during the two-week period, which hour should you add them to? **Hint**: Consider both rider eyeballs and driver supply when choosing

10. Looking at the data from all two weeks, which time might make the most sense to consider a true “end day” instead of midnight? (i.e when are supply and demand at both their natural minimums)

-