Capstone Ideas - NYC Taxi Ridership

**Problem Statement:**

Covid 19 disrupted transportation worldwide. New York city was hit particularly hard, due to the population density of the city. Taxi ridership was undoubtedly affected by the virus, but just how much? This data should give us insight into the level of taxi ridership disruption due to the coronavirus.

**Context:**

New York City currently has 13,587 yellow taxis. Yellow taxis are allowed anywhere within Staten Island, Manhattan, Brooklyn, Queens, and the Bronx. They are available for street-hailing and pre-arranged pickup via an app. The total aggregated fairs per day of yellow taxis are measured in millions of dollars. This provides income for thousands of drivers.

**Criteria for Success:**

Success will include creating a queryable data frame of taxi trips from January 2009 to the most recent data available of July 2021. This data could be used to evaluate the presumed decrease in taxi ridership during the covid 19 pandemic, and subsequent increase in ridership as the coronavirus becomes less impactful. Additionally, weather data can be evaluated to see how ridership changes with weather.

**Scope of Solution:**

Create a data warehouse that will include yellow taxi, green taxi, for hire, high volume for hire ridership data, as well as weather data.

**Data Source and type:**

Ridership data = <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.

Data will be in CSV format to start then converted to an Apache Spark dataframe.

Weathe data = <https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA/date/>

Data will originally be in HTML, but will be converted to Parquet, then as an Apache Spark dataframe.

**Proposed architecture for the solution and rationale behind it:**

A database utilized by Apache Spark will be the backend. Spark is optimized for big data workloads, and will be able to handle queries quickly with the given amount of data.

**Methodology:**

To start, I’ll utilize a python script to scrape the weather and TLC trip record data page for the taxi ridership data.Then, we’ll create the dataframes using Apache Spark. I’ll use Spark to transform, query, and show findings from the data.

**Deliverables:**

A GitHub repo containing the work you complete for each step of the project, including:

■ A slide deck

■ A readme markdown file with descriptions of the project and its working