

**Springboard**

**Ste 1100 22 Battery Street**

**San Francisco, CA 94111-5525**

*Prepared By: Alfred D. Hull*

Data Engineering

Open-Ended Capstone

December 2020



2020

**Contents**

[**Artifact 1: Scope** 1](#_Toc57574036)

[**Project Idea & Proposal** 1](#_Toc57574037)

[**Artifact 2: Design Criteria** 0](#_Toc57574038)

[**Entity Relationship Diagram:** 0](#_Toc57574039)

[**Conceptual Diagram** 0](#_Toc57574040)

[**Functional Diagram** 0](#_Toc57574041)

[**Physical Diagram** 0](#_Toc57574042)

[**References:** 0](#_Toc57574043)

[**Appendix:** 1](#_Toc57574044)

~ This Page is intentionally left blank ~

# **Artifact 1: Scope**

*This phase defines the opportunity this project intends to address, clarifies expected outcomes, and highlights the necessary resources required for optimal success.*

## **Project Idea & Proposal**

**Question 1. Data sources you are considering for your open-ended capstone**

***Data****: New York Taxi & Limousine Commission data set.*

***Link****: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page*

**Question 2. Volume available of data for each source (historic and delta)**

The historical load comprises the below-archived file sets for each month of the year since 2009, totaling around 288 gigabytes. The complementary delta loads occur each month and increment the data set by at least 2 gigabytes in total.

***Delta Load (2 gigabytes per month).***

*Yellow Taxi Trip Records (CSV): file size – 579,698 KB*

*Green Taxi Trip Records (CSV): file size – 39,744 KB*

*For-Hire Vehicle Trip Records (CSV) – 108,186 KB*

*High Volume For-Hire Vehicle Trip Records (CSV) – 1,273,342 KB*

***Historical Load (288 gigabytes).***

*Yellow Taxi Trip Records (CSV): file size – 83,476,512 KB*

*Green Taxi Trip Records (CSV): file size – 5,723,136 KB*

*For-Hire Vehicle Trip Records (CSV) – 15,578,784 KB*

*High Volume For-Hire Vehicle Trip Records (CSV) – 183,361,248 KB*

**Question 3. The volume you will be using for your capstone**

As of December 2020, I plan on utilizing all 288 gigabytes of the Historical Load, and I will delta load the new storage location by the number of months from December to the month when the capstone is due.

**Question 4. Why do you think this is a good data source to be used for the capstone**

I believe this is a good data source for the Springboard Data Engineering Open-Ended capstone because the data is extensive and has large delta loads that are still active.

**Question 5. Any pros and cons that you can think of**

Pros

* As of December 2020, the NYC Taxi and Limousine Commission's actively maintains the file sets and protect them under the Freedom Of Information Law (FOIL).
* The U.S. NYC Government is transparent in sharing this information with the general public.

Cons

* There is no API built for the data sets, and several folders partition the data sets out by year and month
* The file format is in .csv, and there may be risks associated with file corruption

# **Artifact 2: Design Criteria**

The design criteria phase will help lay the theoretical engineering solution and a foundational way-forward to establish the data pipeline. This phase is critical as it walks through the conceptual diagram, functional diagram, and physical diagram. This approach is essential so that engineers understand the scope of the build-out and why.

## **Entity Relationship Diagram:**

### **Conceptual Diagram**

The subject for-hire vehicle (fhv) data utilizes four entities:

* fhv\_tripdata
* fhvhv\_tripdata
* green\_tripdata
* yellow\_tripdata

Engineering a data pipeline to aggregate this data will allow ease of handling for individuals who would like to analyze this information from one fact table in a database: theoretically named dispatches and shown below.

Diagram

Description automatically generated

**Figure. DB Conceptual Diagram**

### **Functional Diagram**

### **Physical Diagram**

# **References:**

# **Appendix:**