[](https://www.linkedin.com/in/alfredhull/)[](https://github.com/ahull002/SP_data_engineering_open-ended-capstone)

**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** 0](#_Toc58327547)

[**What is this?** 0](#_Toc58327548)

[**How to use this?** 0](#_Toc58327549)

[**How to test & develop this?** 0](#_Toc58327550)

[**Constraints, Limitations, and Assumptions** 1](#_Toc58327551)

[**Project Setup** 0](#_Toc58327552)

[**Artifact 2: Design Criteria** 0](#_Toc58327553)

[**Entity Relationship Diagram:** 0](#_Toc58327554)

[**Conceptual Diagram** 0](#_Toc58327555)

[**Functional Diagram** 0](#_Toc58327556)

[**Physical Diagram** 0](#_Toc58327557)

[**References:** 0](#_Toc58327558)

[**Appendix:** 1](#_Toc58327559)

~ 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.*

## **What is this?**

The NYC Taxi and Limousine Commission Data Pipeline is a series of data processing steps. The pipeline computationally ingests data at the pipeline's beginning. Then there are a series of steps in which each step delivers an output that is the input to the next step. This process continues until the pipeline is complete. In some cases, independent efforts run in parallel.

This data pipeline consists of three key elements: a source, a processing step or steps, and a destination (a sink). The data pipeline enables data to flow from an application to a data warehouse (MS Azure) for future analytics and report generations.

## **How to use this?**

This data pipeline is architected as a batch-based flow. In this architecture scenario, we have built an Extraction Transformation Load (ETL) flow from the New York Taxi & Limousine Commission data sets to push this data to a data warehouse and an analytics database.

## **How to test & develop this?**

This data pipeline is architected as a batch-based flow. In this architecture scenario, we have built an Extraction Transformation Load (ETL) flow from the New York Taxi & Limousine Commission data sets to push this data to a data warehouse and an analytics database.

## **Constraints, Limitations, and Assumptions**

**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 as the data is extensive and has large delta loads that are still active and will help people analyze the end product versus spending time wrangling the data itself.

**Project Pros & Cons:**

**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.

We will use this project management model to ensure that the client can save time and money throughout the process and have the flexibility to make changes anytime during the development process.

## **Project Setup**

The NYC Taxi & Limousine Commission, Data Engineering Project, will use the Agile Methodology as the project management process. The methodology involves breaking down each project into prioritized requirements and delivering each piece within an iterative cycle. Each iteration is evaluated and assessed (sprint retrospective) by the development team and the client (Springboard Mentor). The insights gained from the retrospective determines development next steps. We set clear goals in each iteration meeting during the retrospective, such as; expected changes, time estimates, priorities, and budgets.

The agile method gives high priority to client participation at the project's initiation and throughout its lifecycle. The intention is to keep the client involved at every step and to incrementally create a product that they are satisfied with at the end. This approach is essential in saving the client money and time because the client tests and approves the work at each development step, quickly making changes if needed.

**Project Setup Goals:**

* Clarifying needs and outcomes and connecting them to small-behaviors and straightforward
* Breaking down broad overwhelming goals into smaller objectives
* Troubleshooting what is not working well
* Learning from what is working well
* Finding ways to get started easier
* Diminishing the risk of harmful unintentional outcomes
* Illuminating the standards for success and expanding the definition of success to include a more comprehensive range of acceptability when learning.
* Examining your progress in a way that encourages continued progress and emphasizes recovering from setbacks over fearing and avoiding them (self-coaching and self-leading.)

~ This Page is intentionally left blank ~

# **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:**