John Blakkan Project Proposal

Rohit Nair W205

Andrea Pope

### Taxis in the Midst; Weather Impact on NYC Taxi Use

#### Objective:

Our project will analyze data provided by the New York City Taxi & Limousine Commission (TLC) covering over 1.1 billion individual taxi trips in the city between January 2009 and June 2016. We will complement this with historical weather data to determine how/if weather impacts taxi use, and use this data to predict upcoming taxi needs based on predicted weather for the next n days (n to be determined). Our primary research questions:

*Everyone knows it is hard to catch a taxi on a rainy day, but are there other trends based on weather less well known? How could understanding these trends better improve the number of taxis available and appropriate pricing? As a commuter, what are your best options on bad weather days?*

#### Background:

This is a rich set of data, and has captured great interest looking at everything from the most likely place to catch a cab to tipping habits of celebrities. City officials initially released the data in March of 2014, to a New York City analyst, Chris Whong. The analyst had requested the information under New York State’s Freedom of Information Law (FOIL). Due to the size of the data, Whong had to buy a hard drive and drop it off at the taxi commission’s headquarters. He later uploaded the database for others to download (which is what we will be using). Whong created a very interactive data visualization app called [NYC Taxis: A Day in the Life](http://nyctaxi.herokuapp.com/) that animated the 24-hour movements of each cab.

#### Project Goals:

Extract details from various trip data structures, match to historical weather (and other supporting information, as noted) and transform into a data structure that supports answering our primary research questions. Additionally, we will look at questions such as:

* How is taxi usage related to weather (temperature, precipitation, wind, …)?
* During inclement weather, are short trips or longer trips more prevalent?
* Is there a socio-economic impact?
* Or is there a type of weather that “evens” socio-economic impacts?
* How much is travel impacted by weather vs. money?
* What behaviors are most common for beautiful days?
* Does seasonality make a difference (e.g. taxis are used more on 50o days in fall, than 50o degree days in the Spring?)
* How does passenger count change in different weather?
* Social - are people more willing to share a cab?
* Financial - should cab companies be using their vans?

#### Primary Data:

Each section below describes the data sets included in our project.

##### Taxi & Limousine Commission - Trip Data

***Overview:*** The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

The For-Hire Vehicle (“FHV”) trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID. These records are generated from the FHV Trip Record submissions made by bases. As such, the accuracy of this data is less well known, and specifically it may not represent the total number of trips dispatched by all TLC-licensed bases.

***Data source:***http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml

***Total rows:*** 1.1 Billion; on average data for 14 million taxi trips per month[[1]](#footnote-1)

***Total Columns:*** ~ 60 columns across several trip data sources (yellow, green, and FHV), dependent on type of trip. Data dictionaries can be found here:

<http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>

***Planned Analysis:*** This data set will be the core of our analysis, on which we match weather (based on date/time) and other socio-economic factors we include. We will aggregate the data into neighborhoods and/or zones, to ease the reporting process, and help identify trends (pickup, drop-off, and b/t neighborhoods).

##### Weather Data

***Overview:*** Using a land based station at LaGuardia Airport in New York, we are collecting hourly statistics for:

* Air temperature
* Pressure and/or Precipitation
* Wind
* Sky cover & clouds

***Data source:*** http://www.ncdc.noaa.gov/cdo-web/datasets

***Total rows:*** TBD

***Total Columns:*** Focusing on 4 main statistics (noted above) for 1 station.

***Planned Analysis:*** With this data we will establish a weather scenario (temp, precipitation, wind, sky cover) for each taxi ride. Along with looking at the components separately, we will calculate a weather rating for aggregate analysis.

#### Supporting Data:

*Taxi & Zone Lookup:* Mapping table for zone ids in trip data (separate downloadable .csv file from <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>)

*Weather Forecast:* Aiming for free web-based source for LaGuardia Airport weather predictions (5-day). Options considering include:

* Web scrap from weather.com http://www.weather.com
* National Digital Weather Forecast: http://ndfd.weather.gov/technical.htm

Focus will be on same statistics as historical data (temperature, precipitation, wind, and sky cover).

*Weather Forecast:* Aiming for free web-based source for LaGuardia Airport weather predictions (5-day). Options considering include:

* Web scrap from weather.com http://www.weather.com
* National Digital Weather Forecast: http://ndfd.weather.gov/technical.htm

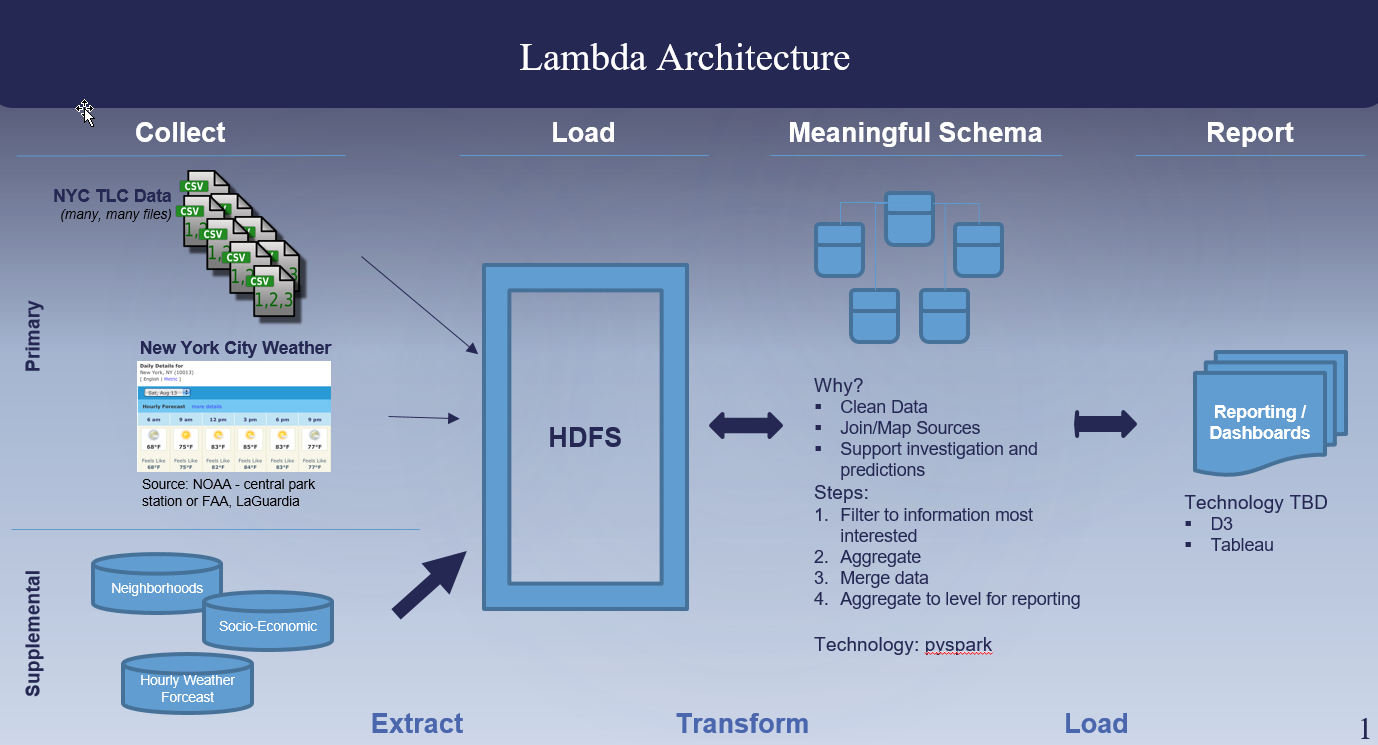
Focus will be on same statistics as historical data (temperature, precipitation, wind, and sky cover).

*Continuous Taxi Trip details:* Pending a reliable source, we will supplement the data with real-time taxi data to continuously improve our taxi/weather statistics. This possibility is still under research.

Additionally, the team maintains the right to include additional supporting information as found necessary to support our analysis, e.g. different neighborhood classifications.

#### Technology Plan

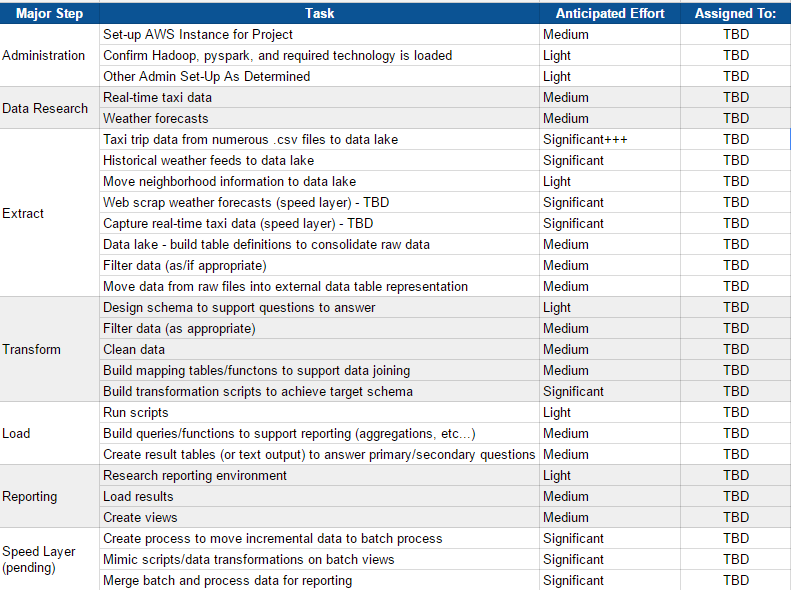
Our technology approach will follow the diagram below.



In addition to this, we may add a speed layer, in the event we can continue to capture additional taxi trip information, along with the real-time weather reports. Along with predictions, this would allow us to continue to enhance our data with the best data.

#### Division of Work

Project work will be split among the following tasks. These are likely to change/increase as we move through the project. Assignment will be dependent on finalized working environment. Additionally, if needed, due dates will be added to keep team members on track, but that is to be determined.



1. https://www.ocf.berkeley.edu/~dlevitt/2015/12/13/final-project-nyc-taxi-and-uber-data/ [↑](#footnote-ref-1)