Larry Gates

CM1004

CSSE494-08

# Senior Thesis Fall Summary

## Introduction:

The purpose of the project is to find areas of congestion usign traffic flow data and determine an appropriate advertisement to be placed at a given location, and if possible, at a given time of day. By using Hadoop Ecosystem, a common tool for big data analysis. I will retrieve datasets containing traffic flow data, demographics for a given area, and advertisement data for a city.

## Motivation:

I was motivated to do the project to gain experience in researching a topic as preparation for graduate school. The idea of learning how to use a tool unfamiliar to me and applying to a topic is difficult but a useful skill to obtain. Currently, congestion in traffic is unavoidable now a day with the increasing number of automobiles on the road, especially during rush hour. During the summer between my sophomore and junior year, I noticed an electronic billboard on my hour drive home. Due to the traffic during rush hour, I could see the billboard clearly for a couple of minutes. If a company could request time to make sure the advertisement was seen by the maximum number of people, then advertisement agencies can adjust prices for a given time, to maximize profits. A thought later occurred to me that if an electronic billboard sold times that lined up with traffic lights in a city, then when the light was red, traffic facing the billboard could display that advertisement on the billboard during the red light. With that, the billboard company could charge for the prime time.

## Literature Review:

Currently I cannot find any article that portrayed exactly to my thesis topic.

TODO: Determined this could be done later after initial draft. I do have literature, just need to summary the essential ones.

## Current Progress:

Using Dr. Mohan’s lesson plans for *Introduction to Hadoop*, I have been teaching myself a condensed version of using Hadoop Ecosystem. Through the lesson plans, I am strongly familiar with the workings of Hadoop’s internal systems, along with several of the services that work with Hadoop. I did get experience with various services, such as Pig and Hive, through some of the labs for the course.

Using the instruction manual from Apache, I set up a cluster that runs Ambari-Server, which is a Hadoop Ecosystem. The cluster runs with 5 nodes and can run various Hadoop labs.

The server holds all available services that were initially provided by Apache and has been operational.

Since the beginning of the fall term, I have been collecting data from the Data Portal of Chicago. Using a bash script, every 5 minutes the script downloads 3 different layouts of traffic data for a given region of Chicago, along with traffic data for a given segment of Chicago. Since the website does not upload a new spreadsheet every 5 minutes, the script checks if the spreadsheet that was downloaded is not the same as the previous file that was downloaded. If the set of 3 files is the same, the new 3 files will be deleted. Currently, there are over 4,000 spreadsheets for both segment and region data from September 18th, 2016 to October 31st, 2016, with about 24 per day.

**Data Description:**

The two datasets currently collected are Segment Estimations and Regional Estimations.

Segment Estimations:

The segment estimations "contains the current estimated [vehicle] speed for about 1250 segments covering 200 miles of arterial roads" [S1] in Chicago. The spreadsheet apparently does not update every segment, which is due to an uploading problem on the data portal's end. The segment estimations are supposedly collected every 20 minutes by the Data Portal. The fields in the spreadsheet are[S2]:

- SEGMENT_ID: Unique arbitrary number to represent each segment
- STREET: Street name of the traffic segment
- DIRECTION: Traffic flow direction for the segment
- FROM_STREET: Start street for the segment in the direction of traffic flow
- TO_STREET: End street for the segment in the direction of traffic flow
- LENGTH: Length of segment in miles
- STREET_HEADING: Direction of the "STREET" from the origin point
- START_LATITUDE, START_LONGITUDE, END_LATITUDE, END_LONGITUDE: These four points represent the start and end points of the segment in the direction of traffic flow
- CURRENT_SPEED: Real-time estimated speed in miles per hour
- LAST_UPDATED: Date and time of update to spreadsheet

The columns that are of high importance are SEGMENT_ID, STREET, LENGTH, START_LATITUDE, START_LONGITUDE, END_LATITUDE, END_LONGITUDE, CURRENT_SPEED, and LAST_UPDATED. These sets of data for each row will be essential to finding traffic patterns. As time progresses, additional columns could be potentially included.

Region Estimations:

The region estimations "contains the current estimated congestion for the 29 traffic regions" [S3] in Chicago. The spreadsheet has updated all 29 regions. The region estimations are supposed to be collected every 10 minutes by the Data Portal. The fields in the spreadsheet are[S4]:

- REGION: Name of the region – made up of the names of the community areas within the region
- REGION_ID: Unique arbitrary number to represent each region
- WEST: Lowest longitude values on the regions boundary
- EAST: Highest longitude values on the regions boundary
- NORTH: Highest latitude value on the regions boundary
- SOUTH: Lowest latitude value on the regions boundary
- DESCRIPTION: Describes the streets that demark the region's boundary
- CURRENT_SPEED: Real-time estimated congested level
- LAST_UPDATE: Time stamp for the latest congestion estimation run.

The columns that are of high importance would be the REGION_ID, WEST, EAST, NORTH, SOUTH, CURRENT_SPEED, and LAST_UPDATE. These sets of data for each will be essential to find traffic patterns. The WEST, EAST, NORTH, and SOUTH data will be combined to be more efficiently handled, since the regions should not change every time. As time progresses, additional columns could be potentially included.

**Future Progress:**

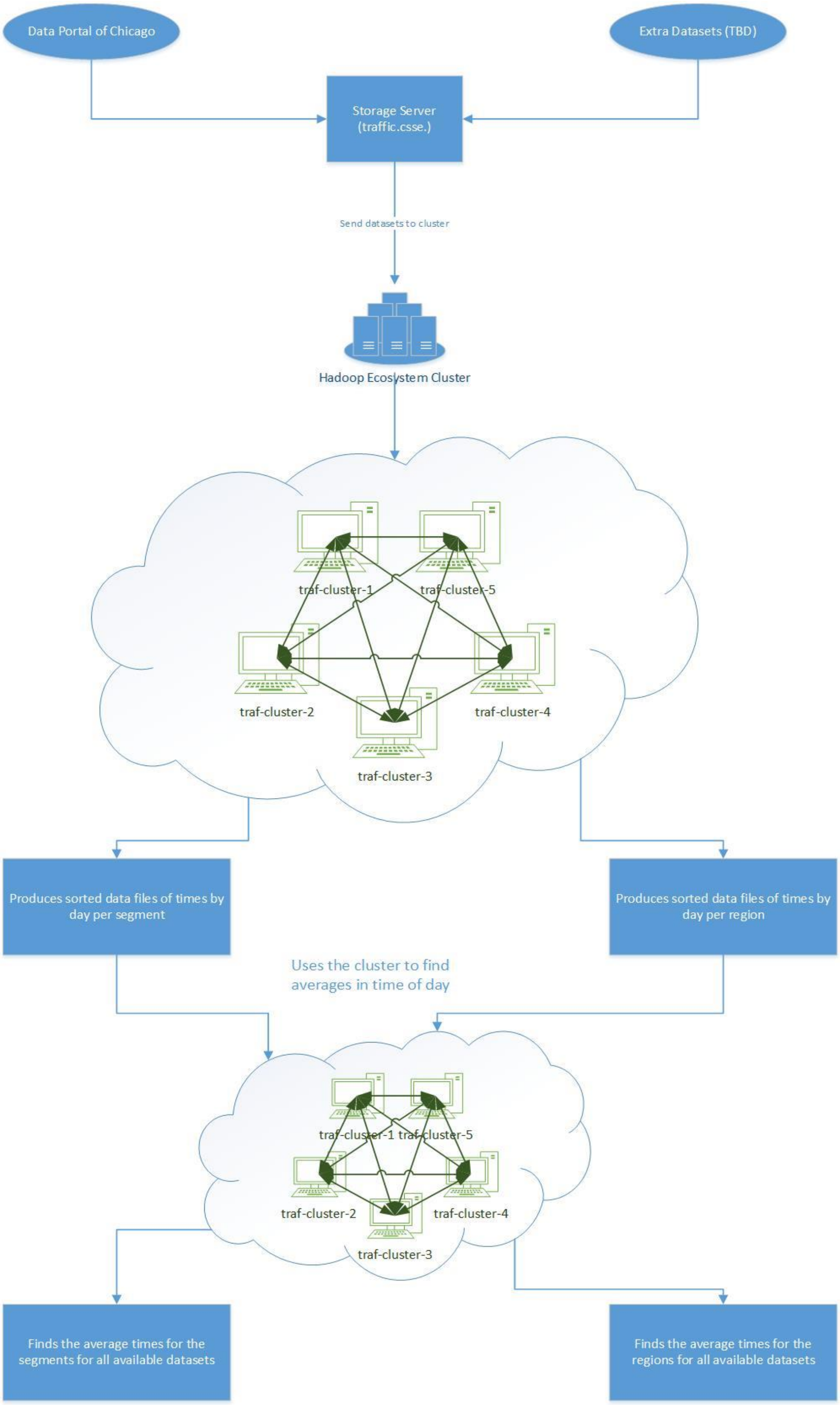
The next step for this project begins with using the services of Hadoop and the actual data. I will continue to collect traffic data from the Data Portal of Chicago throughout the year, and I will look for additional sites that contain datasets pertaining to traffic data. Along with traffic, I will locate additional datasets to help determine locations of billboards in Chicago, along with the advertisements displayed on the billboard. The Data Portal of Chicago has many datasets that appear to have some relevant data about advertisements. Determining if the information is valid for the project along with the accuracy will be determine as the project progresses. Outside of the Data Portal of Chicago, I can use U.S. Census data to find demographics of a given area and overlay the information into segments and regions defined in the datasets I have acquired. These demographic data will help me determine some potential ideas of advertisements that would work.

The data will need to be sorted per segment and region, since there are data points for almost every hour of every day in a segment and region. Data will also have to be cleaned, as some rows have not been updated for years, i.e. one segment has the LAST_UPDATED for 2010-07-21. The data can be sorted by Pig, allowing the data to be aggregated and condensed. Once the data is sorted, I can find averages for a given period. I can use a service that will find the average for various stages along with finding common patterns in a segment or region among the data, or write a custom MapReduce application to do the work for me.

With advice from Dr. Wollowski for incorporating machine learning into my senior thesis, I could potentially use a simple form of machine learning to find typical traffic flow during certain times of day for a given direction. This can be done with finding the averages of speeds for times. If I am successful in finding data that contains location and type of advertisement, I can use the cluster to determine the type of advertisements to display for a segments or regions. The latter part of the suggestions for machine learning from Dr. Wollowski will be applicable if advertisement data is easily accessible.

The proposed architecture flow is displayed below, showing where I collect the data, tools used at what stages, along with stages and flow.

## Proposed Architecture:


Data Portal of Chicago
Extra Datasets (TBD)
Storage Server (traffic.csse.)
Send datasets to cluster
Hadoop Ecosystem Cluster
traf-cluster-1
traf-cluster-5
traf-cluster-2
traf-cluster-4
traf-cluster-3
Produces sorted data files of times by day per segment
Produces sorted data files of times by day per region
Uses the cluster to find averages in time of day
traf-cluster-1
traf-cluster-5
traf-cluster-2
traf-cluster-4
traf-cluster-3
Finds the average times for the segments for all available datasets
Finds the average times for the regions for all available datasets

**Sources: (What style formatting are the source?)**

S1. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Se/n4j6-wkkf/about

S2: https://data.cityofchicago.org/api/assets/3F039704-BD76-4E6E-8E42-5F2BB01F0AF8?download=true

S3: https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Re/t2qc-9pjd/about

S4: https://data.cityofchicago.org/api/assets/88B2ABA5-BF4C-4A41-949C-2B11D725ADAB