Larry Gates
CM1004
CSSE494-08

Senior Thesis Fall Summary

Introduction:

The purpose of the project is to find areas of congestion using 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 the Hadoop Ecosystem, a common tool for big data analysis, the project will optimize the placement of advertisements for a given area based on traffic flow. Datasets we will take advantage of contain traffic flow data, demographics for a given area, and advertisement data for a city. By finding high congestion times for a certain area and combining the advertisement data should produce optimal areas to place advertisements along with the type of advertisement to place. A continuation of this project might have the potential of evaluating the effectiveness of the advertisements with machine learning.

Motivation:

Many cities and states collect traffic statistics for roads, to show traffic flow on display boards or for statistical information. Utilizing public domain data and finding patterns from the datasets are essential for users who are interested in finding trends in patterns that are related to each other. By completing this project successfully, this will be an example for other cities to attempt to record traffic data to find patterns and optimize selling locations.

Since the collecting and storing of data has become easier with the Internet of Things and the availability of cheaper storage, collecting traffic data within a well-funded location would not be a problem. The only limitation is the sensors used to collect the traffic data. Once a city has

the data, providing data to the algorithm should produce a result. If an algorithm is already prepared created, inputting traffic data will not be a hassle, and provide useful marketing data for a city.

Literature Review:

Currently, there appears to be no academic research into using big data and coordinating traffic flow with advertisements. However, plenty of academic research and articles involving big data and traffic datasets, along with many application articles on using big data and advertisement.

Smith and Demetsky (1997) discuss the importance of "intelligent transportation systems," at a time for Google Maps and traffic flows on the internet. Their academic paper discusses the need to have forecasting models, using several time prediction techniques, such as machine learning and historical average, to determine traffic forecasting. With the datasets being collected, the project will use historical averaging with the available data. Daas, Puts, Buelens, and van den Hurk (2013) highlight that at the time, big data was heavily "IT-perspective" and "focus on soft- and hardware issues." In the Big Data case study of traffic loop detection data, a successful plot was shown of peak hours and vehicle flow. The case study shows the high potential of finding a trend in the data.

The big data and advertisement side of academic articles talk about the collection and use of user data. Couldry and Turow (2014) elaborate on personalized advertising constantly mining personalized data. The article also looks "more broadly at the consequences of embedding big data use in advertising," which is not concern for the project. Bughin, Chui, and Manyika (2010) discuss the opportunities companies take with using the data available for a web-based company.

Allowing for better selective marketing for an area. The availability of data for marketing is not scarce due to the expanding amount of data, which will be useful in determining advertisements to display is a certain area.

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.

By working through the *Introduction to Hadoop* lesson plan, I was able to set up a cluster that runs Hadoop. 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.

Currently, a script is set up to run and download a set of 3 files for *Chicago Traffic Tracker – Congestion Estimates by Segments* and a set of 3 files for *Chicago Traffic Tracker – Congestion Estimates by Regions*. Since the Data Portal of Chicago does not have a reliable upload time, the script that downloads the files checks that the files downloaded are newer the previous set that was downloaded. Following is a description of the data and the planned use.

Data Description:

The two datasets currently collected are Segment Estimations and Regional Estimations, collected from *Chicago Traffic Tracker – Congestion Estimates by Segments* and *Chicago Traffic Tracker – Congestion Estimates by Regions*, respectably.

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 roughly collected every 20 minutes by the Data Portal. The important information from each row in the dataset are the location and area of the traffic flow, speed of traffic, and the timestamp the data was updated. Since the segments are sub-areas of a region, more options for finding optimal advertisement placement will be available, compared to having only 29 areas to find optimal advertisement placement. Refer to the appendix for information about the columns of the dataset.

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 important information from each row in the dataset are the boundary of a given region, speed of the traffic, and the timestamp the data was updated. Since the 29 regions each have a given number of segments, a given region will provide an overview of the average speed of all the segments. By having segments in each region, a pattern can be found about the given region. This can be used to find the best advertisement over the region, then finding a better fit of the advertisement in a smaller defined area, such as a segment. Refer to the appendix for information about the columns of the dataset.

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. This demographic data will help me determine some potential ideas of advertisements that would work. If the attempt to locate additional datasets about the specific billboards in Chicago and the success of advertisement is unsuccessful, I will simulate data that would give the appearance of advertisement data I have been looking for.

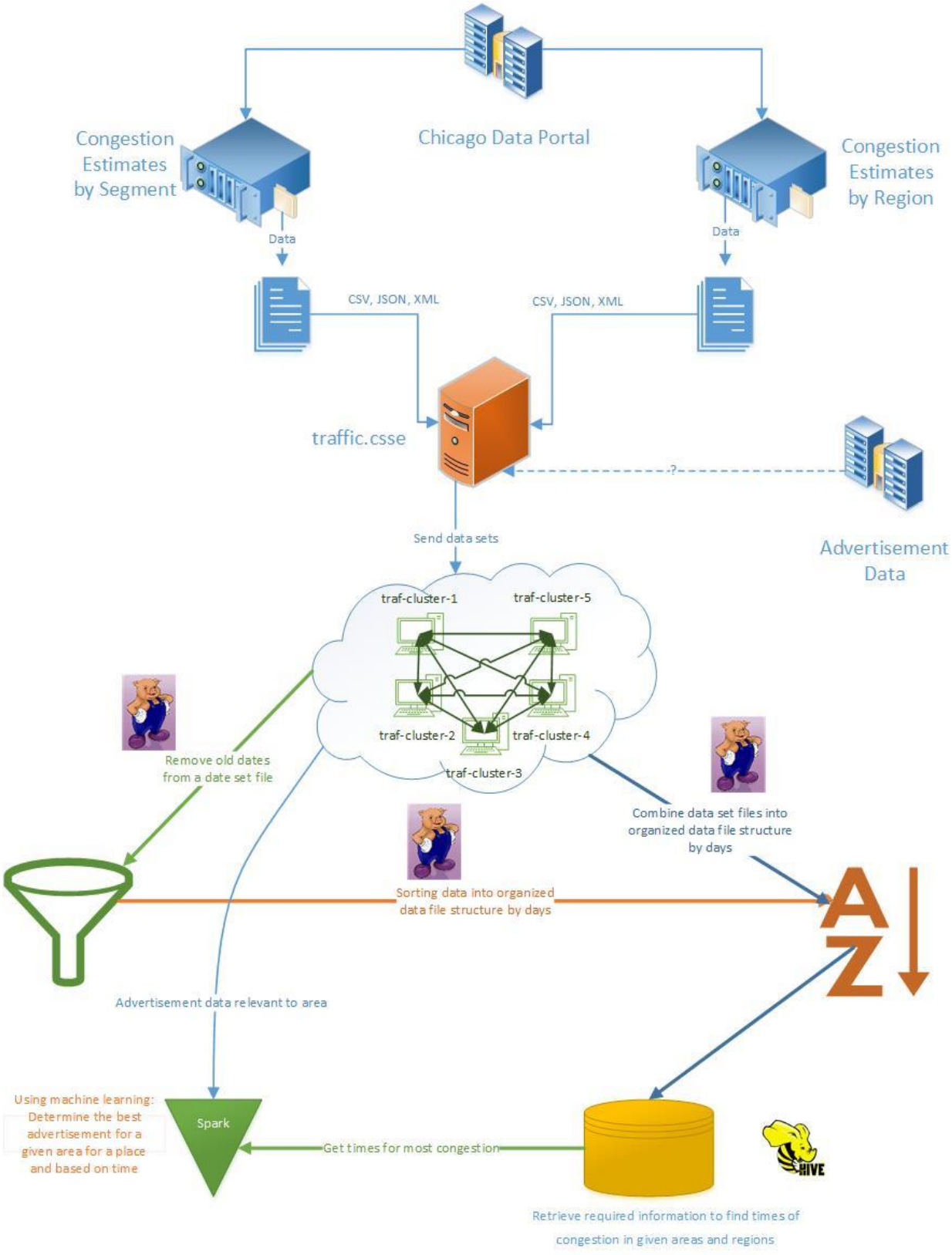
The data will need to be sorted per segment and region, since there are data points for almost every hour of every day in a segment and region. Data will also have to be cleaned, as some rows have not been updated for years, i.e. one segment has the LAST_UPDATED for 2010-07-21. The data can be sorted 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:



Sources:

Jacques Bughin, Michael Chui, and James Manyika. 2010. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly* (2010), 14.

Anon. Chicago Traffic Tracker - Congestion Estimates by Segments . Retrieved November 1, 2016 from https://data.cityofchicago.org/transportation/chicago-traffic-tracker-congestion-estimates-by-se/n4j6-wkkf/about (S1)

Anon. Chicago Traffic Tracker - Congestion Estimations by Traffic Segments. *Chicago Traffic Tracker - Congestion Estimations by Traffic Segments*. (S2)

Anon. City of Chicago | Data Portal. Retrieved November 7, 2016 from https://data.cityofchicago.org/transportation/chicago-traffic-tracker-congestion-estimates-by-re/t2qc-9pjd/about (S3)

Nick Couldry and Joseph Turow. 2014. Advertising, big data and the clearance of the public realm: marketers' new approaches to the content subsidy. *International Journal of Communication* 8 (2014), 1710–1726.

Piet J.h. Daas, Marco J. Puts, Bart Buelens, and Paul A.m. Van Den Hurk. 2015. Big Data as a Source for Official Statistics. *Journal of Official Statistics* 31, 2 (January 2015). DOI:http://dx.doi.org/10.1515/jos-2015-0016

Brian L. Smith and Michael J. Demetsky. 1997. Traffic Flow Forecasting: Comparison of Modeling Approaches. *Journal of Transportation Engineering* 123, 4 (1997), 261–266. DOI:http://dx.doi.org/10.1061/(asce)0733-947x(1997)123:4(261)

Appendix:

Chicago Traffic Tracker – Congestion Estimates by Segments:

Spreadsheet Information:

- 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

Chicago Traffic Tracker – Congestion Estimates by Regions:

Spreadsheet Information:

- 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.