Hello, I am Larry Gates and today I am presenting you my thesis “Analysis of Advertisement Data with Traffic Patterns using the Hadoop Ecosystem.” I was advised by Dr. Sriram Mohan.

**<Slide>**

My thesis is: using billboard advertisement data and congestion traffic, I am effectively able to find areas of congestion for a given region or segment with the best advertisement, for the region or segment.

**<Slide>**

My motivation for the project is based on: income avenue revenue, public domain data, and test of viability. For an advertisement company, the places that have notably higher congestions for various parts of the days can increase the price to place an advertisement at that location, place a billboard at that location, or ultimate remove a billboard from a location due to low success. For a city, billboard locations can cost more, overall or just at certain times of the seasons.

As we have seen, publicly available data is not hard to come by, and certain traffic data is easily available for some cities, such as Chicago. This thesis shows the success of using traffic data joined with advertisement data. The thesis was also a test of viability by getting a result of combining traffic data.

An original motivation was using electronic billboards, since those could be changed and use certain advertisements at a given day, due to the lack of availability for real-time traffic data, this aspect could not be explored.

**<Slide>**

Now, we have all heard of big data, so what is big data? From the definition found in “A formal definition of big data based on its essential features” by Mauro, Greco, and Grimaldi, the definition is as follows: “Big data is the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value.”

**<Slide>**

From the literature reviews I did during the thesis, I found no academic papers on the topic of using big data, advertisement data, and traffic data. However, I found papers involving big data and traffic data, and big data and advertisement data.

For big data and traffic datasets, Smith and Demetsky discussed the importance of intelligent transportation systems, at a time before Google Maps and traffic flows were available on the internet. Their paper discussed 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, this project is using historical averaging with the available data.

Daas, Puts, Buelens, van den Hurk highlight that at the time that Big Data was heavily viewed from an IT-perspective and forcues on software and hardware issues. In the Big Data case study done by Chen, pao, Lee about the study of traffic loop detection data, a successful plot of peak hours and vehicle flow was shown. The case study dhows the high potential of finding a trend in traffic data.

The big data and advertisement side of academic articles talked about the collection and use of user data. Couldry and turow elaborate on personalized advertisement constantly mining personalized data. The article also looks more broadly at the consequences of embedding Big Data use in advertising, which is not of concern for this thesis. Bughin, Chui, and Manyika discuss the opportunities companies take with using the data available for a web-based company. The availability of data for marketing is not scarce due to the expanding amount of data, which is to be useful in determine advertisements to display in a certain area.

**<Slide>**

To complete this thesis, I learned how to use the Hadoop Ecosystem, which takes advantage of using a distributed file system to handle storage needs of large files that need to be browsed. One method of handling the files stored on a Hadoop cluster is to make a MapReduce job, which will map a filter to all the files and sort the data into a desired output and ma the result to reducers to preform operations on the data to give a result. Tools available with Hadoop include: Apache Pig, and Apache Hive. Pig allows users to take MapReduce jobs to a higher level of abstractions when processing datasets. Pig can be used for cleaning and sorting data, as well as joining datasets together. Hive is a data warehouse framework that utilized the way data is stored in HDFS and accesses such data in SQL type format.

**<Slide>**

The data used from the City of Chicago’s Data Portal were *Chicago Traffic Tracker – Congestion Estimates by Segments*, and *Chicago Traffic Tracker – Congestion Estimates by Regions*. These two datasets shared traffic data collected at a closer to real time level. The files had the identification value of the segment and region, the description of the segment and region, the boundary, the speed and the time the data was last updated. Advertisement data had to be randomly generated for this project, due to the fact that specific advertisement data is hard to come by for academic purposes. A better description for how the data was created is described later. Census data was a possibility as a substitute for advertisement data. The map to the right shows the zipcodes, gray outlines that fall within the regions in the data set. The blips are the northeast and southwest corners of a region, matching color for the corners. The problem was easily finding the overlap of a zipcode inside of a region, the weight the zipcode would have on the demographics, and the overall calculation of area required an accurate and easy to use dataset to determine if a region was in a zipcode.

**<Slide>**

This is the first part of my architecture of the project. The downloading the data process started with a script that automatically downloaded 6 files, a *csv*, *json*, and *xml* of the segment and region estimates. Initially, I didn’t know what files I needed, so better to have different types, just in case. The files were then stored on a VM behind Rose’s firewall.

**<Slide>**

The second part of the architecture was the data analysis part. Using the cluster, I would take the *csv* data from the data collection VM and pull it over to my cluster of 5 VMs. The advertisement data also manually generated on the cluster and stored in the files system. Once there, a Pig script would remove any bad rows, such as the column headers, remove data outside of the dates of collection, then sorting the data and storing on file system. Afterwards, the cleaned traffic data is processed through a MapReduce job to find statistics about the congestion for a given day. The advertisement uses a Pig script to sort the advertisement data by segment and region and day. Afterwards, the advertisement data and traffic data is combined based on region and segment and the date. Then, MapReduce jobs are performed on the combined data, one to find the longest period of congestion, the number of times the advertisement was there with the rating and the rating. The four separate MapReduce jobs each had a different time length for finding the best advertisement for congestion. One MapReduce job took the 1st week of the data and found the longest advertisement and the rating for the region and segment during a week span. This would be done for the 2nd, 3rd up to the 52nd week of the year, then starting again with the next year. Another time split was by month, taking the best advertisement for a given month of the year. Same with a year, and all the data together. Afterwards, a Hive script was used to store the results into a SQL type structure for easily querying. The final Hive script was combining all the data together to compare.

**<Slide>**

The major algorithms I created for the thesis are as follows:

The advertisement data was a script that automatically downloaded the traffic data sets I needed. The traffic analysis took the segment identification or region identitication, speed and date to find the congestion statistics for traffic data for a given day. The grouping analysis was using different time frames to compare the traffic data. Querying the results was using hive scripts to combine the results together to compare and contrast.

**<Slide>**

The results here show a subset of the region output. To make the data easily comparable for humans, I had to take all the weeks which had outputs for every week that feel within the time frame and find the one that had the maximum theme count and rating compared to the rest of the weeks for a region. The same had to be done for month and year. This would give me a result of one row, which could be compared to the aggregation done on all the data. The green row shows that results can change as you use different time frames. The brown row shows that the algorithm checks congestion statistics to see if there is any worth placing an advertisement at that location. If not, the region is ignored. The gray shows that the theme can be consistent for all time frames. One thing to point out is that the range for an advertisement was 1 to 30 days, meaning that *travel* theme appeared a total of 54 times with a rating of 9.

**<Slide>**

The results here show a subset of the segment output. The data again went under the same aggregation to make the data human readable. The yellow highlighted rows are all values that were completely ignored. Segments had a higher chance of appearing with ignore since many had in the description that the segment was not being recorded any more or there was not enough congestion. The *Separated by Week* columns show that a theme count of 7 happened all the time, where the purple row has a high rating of 9 for a theme count of 5 weeks. The pink row also happens to show a different theme every time.

**<Slide>**

Future work on this project would be getting actual advertisement data or incorporating the census data into the project. Both would be interesting aspects to include in future work. I would have liked to use machine learning in processing both advertisement data and traffic data, but due to time and learning curve for the tools, I had to forgo the idea. This would be interesting to pursue in the future.