**Speech**

The purpose of the thesis is to find the areas of congestion using traffic flow data and determine an appropriate advertisement to be placed at a given location. By using the Hadoop Ecosystem, a common tool for big data analysis. I optimize the placement of advertisements for a given area based on traffic flow. By finding high congestion times for a certain area and combining the advertisement data relative to the area, the system produces optimal ranges to place advertisements, along with the type of advertisement to place. A majority of the work of cleaning and organizing the traffic and advertisement data is done by Pig, and the aggregation MapReduce jobs. Many additional tools were explored as I learned to use Hadoop independently. Tools I used were MapReduce in Java, and Apache Pig. The MapReduce is used to aggregate the advertisement data to find the best advertisement for a segment. Pig is used to clean the old data and sort into multiple output. Some of the tools I attempted to use were Apache Spark and jsoup. I attempted to use Apache Spark for the machine learning applications, but due to time constraints and limited experience, I had to forgo using it. I did use jsoup to download demographical data for zip codes in Chicago. But once I came to the realization of easily determining what segments are in each zip code. The tool was very useful but unable to use the data I collected. Some of the future modifications I would like to implement would be the idea would be able to apply a machine learning algorithm on the traffic data, comparing different algorithms with different outcomes. I would also like to incorporate machine learning into advertisement, finding the best advertisement for a given area. This would come along with purchasing more advanced advertisement data, such as billboard providers or additional demographical data. I can collect some through various websites, but using the traffic data available, determining the weight of zip codes for a given segment or region was beyond my depth for the project. To allow for various input from other locations, I would like to make it easy to standardize the input. If I had various locations to gather input from besides Chicago, I could find create a suitable API to easily allow the input of data. I would have liked to incorporate more Hadoop tools, such as Hive, to handle data management and aggregation. If this project was done over several years, additional data would have been gathered and used to test a larger scale Hadoop file system. I could also compare the outcome and performance of the Hadoop Ecosystem to other big data management software.

I currently have 14801 Region Estimate Files with a total of 56.3 MB. Each region dataset contains 29 rows for each pre-defines Region in Chicago. I currently have 14830 Segment Files with a total of 2.2 GB. Each have 1,257 rows for each pre-defined Segment in Chicago.

**Potential Questions**

* Can you walk through one of your diagrams?
  + TODO
* What other Big Data management systems are available?
  + TODO
* What type of machine learning algorithms could have potentially been used?
  + TODO
* What do you have remaining to do?
  + TODO
* How do you collect the data?
  + TODO
* What does that data look like?
  + TODO