Analysis of Travel Behavior Big Data by Hadoop Ecosystem

Moein Hosseini
Northwestern University
Evanston, IL 60208
(773) 865-9971
Moeinhosseini2020@u.northwestern.edu

***Abstract*— Big Data has created many opportunities for different organizations to use the data produced by their activities and the data created in the surrounding environment of the organization. The organizations which are operating in the transportation area are also benefiting from Big Data analytics. Studying the behavior of travelers in choosing their transportation mode is an important factor in determining transportation-related policies. This paper represents how the data pertaining to the mode choice of the public could be analyzed by Big Data analytics tools.**

I. INTRODUCTION

TRAVEL behavior has become a hot topic in the transportation research stream. Analysis of the behavior of travelers could help transportation service providers offer a reliable transportation service that satisfies customers' needs. Although this provides a lot of benefits for the society, the complexity of human behavior imposes a lot of challenges throughout the analysis. Researchers try to uncover most of these complexities by a Stated Preference/Revealed Preference survey. Many questions are asked during the survey from customers that subsequently results in the development of a large dataset. The data set is then analyzed to formulate a travel demand model based on well-known travel demand model formulas.

There are many studies conducted in the area of travel demand modeling. Osman Idris et. al. [1] investigated the effect of psychological factors on the commuting mode choice in the city of Edmonton, Alberta. Habib et. al. [2] studied the activity-travel behavior of Canadian non-workers using a random utility maximization approach. Kamargianni et. al. [3] studied the influence of subjective and objective factors in school travel mode choice from a survey conducted on a sample of students living in Cyprus. Ji et. al. [4] examined the effect of several factors such as gender, age, income, and trip purpose among Nanjing (China) residents on choosing the public bicycle mode in order to access rail transit.

Big Data is a concept developed around a high volume of data. Due to the abundant of questions in a travel survey, the concept of Big Data is also applicable to a travel survey data. Therefore, tools developed to handle Big Data could be useful to analyze the data more effectively.

The current study analyzes the data collected over a survey using tools that are available in the Hadoop ecosystem. It is shown how a large data set could be quickly and accurately analyzed in Hadoop by a few lines of code.

In the following sections, first, an overview of Big Data characteristics and technologies along with a brief description of the Big Data ecosystem components are provided. More details are mentioned about the Hadoop component. Next, the steps of importing the data into the Hadoop platform and the analysis steps are provided. Results of the analysis for each block of the code is discussed. The paper ends with the concluding remarks.

II. BIG DATA

Big Data is a high volume of data with various complexity and ambiguity level, developed at different velocities. Traditional data analysis technologies are unable to precess Big Data efficiently and effectively. The concept of Big Data appears when the volume, velocity, or variety features of the data is more than the amount that could be handled by traditional technologies. Therefore, the 3V's is a basic characteristic of Big Data.[5]

Data volume is the generated amount of data. This could vary from a few kilobytes for a data collected from a small experiment to terabytes of data developed by sensors in an aircraft. Another example of data with a very large volume is clickstream logs from websites. Google is one

of the companies which is using this data set extensively. They capture the user behavior that would be helpful in determining their marketing policies. Traditional technologies cannot store and manipulate such a large amount of data.[5]

Data velocity is another important aspect of data. The amount of data transferred over time in each stage of the data lifecycle specifies the data velocity. Traditional technologies can analyze data when the flow of input data is at a fixed rate. However, in the case of Big Data, the continuity of the data stream and the necessity for minimizing processing delays reduces the effectiveness of traditional data analysis tools. Data generated by social media users, smartphone users, and sensors are examples of high-velocity data. The data generated by connected vehicles and autonomous vehicles can reach a velocity in order of gigabytes per second. The sensor network in an airplane flying from London to New York can generate 650 terabytes of data. [5]

Last but not least, variety is a key component in the 3V's attribute set. Data sets which are completely compatible with traditional data analysis technologies are in a structured format and the data type is predetermined. On the other hand, there is no control over the data format in Big Data and the data could be structured, semi-structured, or unstructured. A metadata associated with the data determines the type and content of the data set. This is usually seen in the tweets collected from the Twitter API where the collected data would be in JSON format showing the tweet content, sender name, time and date, etc. The photos collected from a social media website such as Flickr possess a metadata that mentions the time and date the photo was taken, the time and date it was upload on the website, the comment, and topic associated with the photo, location of the photo, resolution, etc. In order to analyze various formats of data the data analysis system should have the following features: scalability, distributed processing capabilities, image processing capabilities, graph processing capabilities, and video/audio processing capabilities. [5]

Other advantages of Big Data analytics tools over traditional tools are scalability, cost-efficiency, flexibility, high performance, and high availability. [5]

Because of the limited capabilities of traditional data analysis technologies, analysis of Big Data requires a completely new system. This new system is usually referred to as the "Big Data Ecosystem" consisting of many components, some of which are described below:

- Scale-out Databases

This is a strategy to increase the database capacity. The capacity is increased by adding database nodes and the capacity is increased proportionally. This technique supports the principles of distributed computing by providing a clustered horizontal-scale database solution. Several companies such as Mellanox and Clustrix offer scale-out databases.

- Horizontal Platforms

This type of platform exists in cases like Hadoop where a cluster of inexpensive commodity hardware hold the data and perform the relevant analysis. Horizontal scalability, cost-effectiveness, and fault tolerance are the most important features of this type of platform.

- Vertical Platforms

Vertical platforms allow customers to run data analysis on their own service. This platform could serve as many as hundreds of thousands of customers. Examples of this type of platform include transactional history reports requests with a customized frequency in online banking systems, queries of statistics and cross relate statistics across the life cycle of a game in online gaming systems. These platforms almost operate in real time. They are generally faster than horizontal platforms.

- BI/Visualization tools

Data visualization involves effective, efficient, and explicit representation of data using plots and graphics. This would facilitate the communication of information. Therefore, visualization improves the usability, understandability, and accessibility of complex data. Tableau, Infogram, ChartBlocks, Datawrapper, and D3 are a few examples of data visualization tools that are helpful in the area of Big Data.

- Security

Many organizations have started using Big Data without considering the security perspective of this technology. Data ownership and responsibility toward data protection are some examples of the security concern. There is a specialist in the security domain of Big Data which offer fraud detection systems, security incident and event management (SIEM) systems, and threats detection and prevention systems.

- Hadoop

Apache Hadoop is an open source software framework that can store data in a distributed configuration and can processes the very large amount of data within the software framework. Clusters of commodity hardware constitute the physical infrastructure of Hadoop. More information about the projects and components of Hadoop is available in the next section.

- Data Integration

The volume and variety of data in Big Data necessitate the development of data integration systems that could join data collected from various sources and produce a unified view of the data. A data set which has been preprocessed through a data integration system would create more business value for an organization. Talend is one of the companies that performs data integration on Big Data.

- Hardware

Physical infrastructure compatible with Big Data is a key component that could be defined in the set of hardware used for data storage and processing. The data processing architecture determines the type of hardware which is needed. For example, the hardware requirements for a horizontal platform is different from the ones on a vertical platform. Companies such as IBM and INTEL are active players in the Big Data infrastructure industry.

- Services

There are many companies that offer services that are highly dependent on Big Data. They collect data from the various organization and various section of the society. After performing detailed analysis at a sophisticated level, they sell the results of analysis on this large amount of data to organizations that could benefit from incorporating the results into their business policies. In order to maximize the profit, Big Data service providers try to use sources which are free of charge. For example, there are many websites that use the free data collected from the Google API to produce location-based services. As another example, there are healthcare analysis companies combine crowdsourced information available in online websites and social media with the information received from healthcare service providers to come up with insights that could be beneficial for the healthcare system.

- Cloud Providers

Cloud providers offer an internet-based computing framework by allowing other people and organizations use their computational power. They provide shared computing resources and data. The cloud computing service is usually free up to a certain limit. Customers should pay a subscription fee for an over limit computing service. Amazon's Public Elastic Compute Cloud, Google Big Data services (Google compute Engine, Google Big Query, and Google Prediction API), and Microsoft Azure are a few example of famous cloud providers.

III. HADOOP ECOSYSTEM

As mentioned in the previous section, Hadoop is a software framework in the Big Data ecosystem. Nowadays this component has become very popular in the industry as well as among researchers. Hadoop is also composed of various components some of which are described below[6]:

- Hadoop Distributed File System (HDFS)

It is a distributed, scalable, and portable file system which is programmed in Java for the Hadoop framework. High-throughput, scalability, and high availability are the main features of this system.

- MapReduce

It is a framework for large-scale computations that performs parallel distributed algorithms on a cluster. MapReduce is also scalable and the reliability is achieved through job resubmission.

- Pig Latin

It is a high-level parallel data flow language. It utilizes the extract, transform, load technique and stores data at any point during the analysis.

- HBase

It is an open-source, non-relational, distributed database which is programmed by Java. HBase has scalability and compression features, and it performs the in-memory operation.

- Hive

It is a data warehouse in Hadoop that performs data summarization and accesses data by various queries. The processes in Hive are accelerated by the indexing feature. Data is stored in different formats by their relevant metadata.

- ZooKeeper

It is a centralized service which offers distributed synchronization. Moreover, it serves as a distributed configuration service and a naming registry for distributed systems. High availability is an important attribute of ZooKeeper.

- Ganglia

It is a distributed monitoring system which is specifically designed for high-performance computing systems such as clusters and networks. Ganglia is a scalable Hadoop project.

- Sqoop

The main operation of this Hadoop component is to transfer data between Apache Hadoop and structured databases (RDBMS).

- Hama

It is a distributed engine for massive scientific computations such as matrix, graph and network algorithm (BSP)

- HCatalog

This is a table management layer for Hive metadata to other Hadoop applications.

- Mahout

It is an Apache software package with a scalable machine learning library. Its machine learning algorithms are more focused on collaborating filtering, clustering, and classification.

- Ambari

It is an Apache Software Foundations project that performs provisioning, managing, and monitoring of a Hadoop cluster. Ambari is used by famous companies such as IBM, eBay, Kayak, and Samsung.

- Flume

It is a distributed service that can effectively collect, aggregate, and move a massive amount of log data. Tunable reliability, robustness, high availability, and fault tolerance are key features of this Hadoop component.

- Giraph

It is an interactive graph processing system developed in the Hadoop platform.

- DRILL

This system is a low latency SQL query engine for Hadoop.

IV. DATA DESCRIPTION AND PREPARATION

The data was collected in Tehran, the capital city of Iran. The city is 700 square kilometers with 22 municipalities and over 600 traffic analysis zones. The data was collected throughout a survey in December 2010. Educational trips constituted 27% of the 7.5 million trips per day that happened in the city in 2010. The number of kindergarten through grade 12 students on that specific day was 1,119,571. Around 70% of these students studied in middle school or high school. The survey questionnaire was distributed among 4,700 students based on a randomly selected stratified sampling method. The stratification was performed based on the population of the municipalities and the gender distribution in the society. After preprocessing the 3,441 returned questionnaires, 3,274 responses remained for further analysis. [7]

The questionnaire contained 19 questions which were mainly composed of two parts: socioeconomic and demographic characteristics, school trip mode-specific characteristics. [7]

Figure 1 shows a screenshot of the survey data.

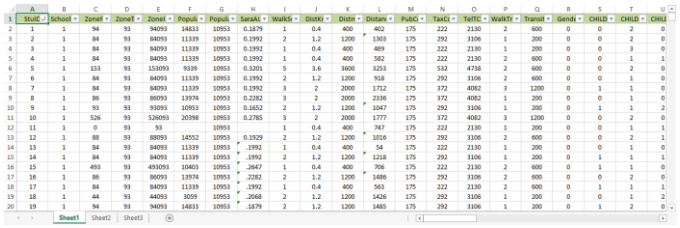

Figure 1) Screenshot of the survey data

As the first step of the analysis, the data was imported in the file browser of Hadoop (Figure 2).

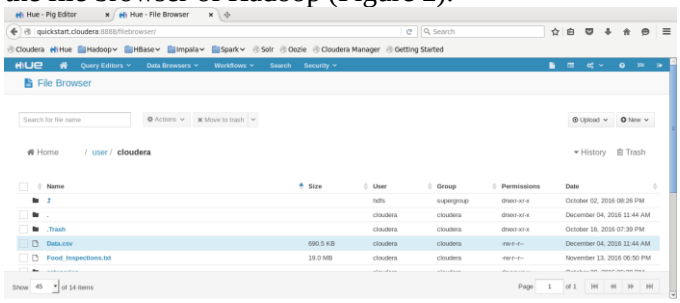

Figure 2) Screenshot of the data stored in File Browser

A metadata was created to enable various Hadoop projects to access the data.

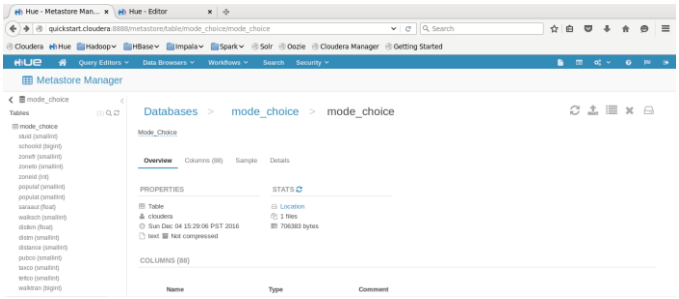

Figure 3) Screenshot of the metadata created in Hadoop

Figure 4 shows a screenshot of the code written in Hive. The complete program is listed in the next section.

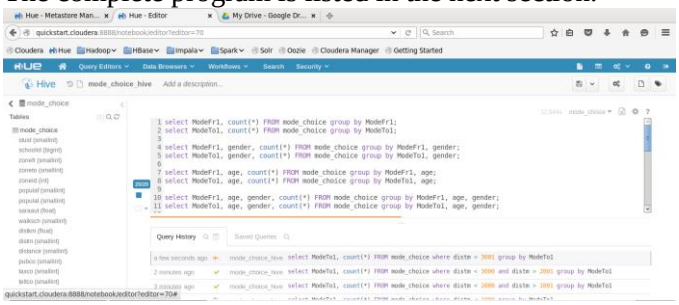

Figure 4) Screenshot of the Hive program

V. DATA PROCESSING IN HADOOP

As mentioned earlier, Hive was the selected Hadoop project to perform the analysis. The simplicity of coding and quick development of the results were the main reasons for choosing this tool among other tools available in Hadoop.The complete code is list below:

```
select ModeFr1, count(*) FROM mode_choice group by ModeFr1;
select ModeTo1, count(*) FROM mode_choice group by ModeTo1;

select ModeFr1, gender, count(*) FROM mode_choice group by ModeFr1, gender;
select ModeTo1, gender, count(*) FROM mode_choice group by ModeTo1, gender;

select ModeFr1, age, count(*) FROM mode_choice group by ModeFr1, age;
select ModeTo1, age, count(*) FROM mode_choice group by ModeTo1, age;

select ModeFr1, age, gender, count(*) FROM mode_choice group by ModeFr1, age, gender;
select ModeTo1, age, gender, count(*) FROM mode_choice group by ModeTo1, age, gender;

select ModeFr1, auto, lic, count(*) FROM mode_choice group by ModeFr1, auto, lic;
select ModeTo1, auto, lic, count(*) FROM mode_choice group by ModeTo1, auto, lic;
```

```
select ModeFr1, income, count(*) FROM mode_choice group by ModeFr1, income;
select ModeTo1, income, count(*) FROM mode_choice group by ModeTo1, income;

select ModeFr1, count(*) FROM mode_choice group by ModeFr1 where distm < 1000;
select ModeFr1, count(*) FROM mode_choice group by ModeFr1 where distm < 2000 and distm > 1001;
select ModeFr1, count(*) FROM mode_choice group by ModeFr1 where distm < 3000 and distm > 2001;
select ModeFr1, count(*) FROM mode_choice group by ModeFr1 where distm > 3001;

select ModeTo1, count(*) FROM mode_choice group by ModeTo1 where distm < 1000;
select ModeTo1, count(*) FROM mode_choice group by ModeTo1 where distm < 2000 and distm > 1001;
select ModeTo1, count(*) FROM mode_choice group by ModeTo1 where distm < 3000 and distm > 2001;
select ModeTo1, count(*) FROM mode_choice group by ModeTo1 where distm > 3001;
```

This code contains 20 queries. Each query counts the number of students who used each mode based on their socio-economic characteristics and the transportation network features. Gender, age, auto ownership of the household, driver's license acquisition of the parents, and household income are the socio-economic characteristics which were investigated. The only transportation network characteristics which were considered here is the distance from home to school.

VI. RESULTS AND DISCUSSION

The result of analysis on the mode chosen by middle school and high school students is represented in figure 5 (going to school) and figure 6 (returning to home). The analysis shows that 49.97% of students walk to school and 43.65% walk from school to home. Therefore, students of this age range are active travelers in the society. The second mostly used mode of transportation is the school bus system (around 23%). This system is administered by each school independently. Public transit only constitute approximately 12 percent of the trips since most of the students live near their school, or the high travel time of the public transit mode makes it less attractive. Slightly less than 10 percent of the students ride their household car from home to school, but this number jumps to 16.83% when the students want to commute from school to home. There are three modes of transportation which are not used during the morning but are used for the afternoon commute: tele-taxi, taxi, and cycle. The data shows that the demand for walking as an active mode of transportation and public transit via bus is shifted to more private modes such as automobile, taxi, and tele-taxi. The appearance of cycling as a one-way commute mode of transportation (only from school to home) for a very small number of respondents requires a thorough investigation to identify if it is a survey error, or there is a special behavior behind this choice.

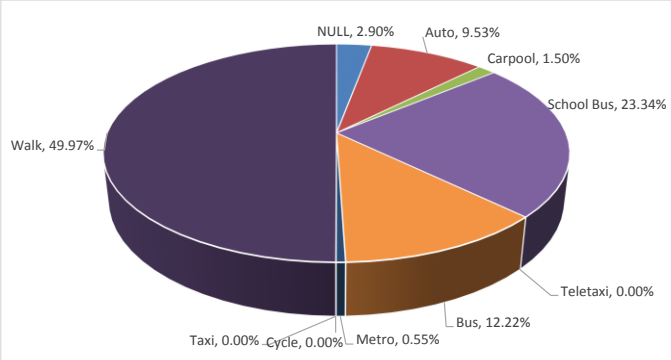

Figure 5) Mode choice of students from home to school

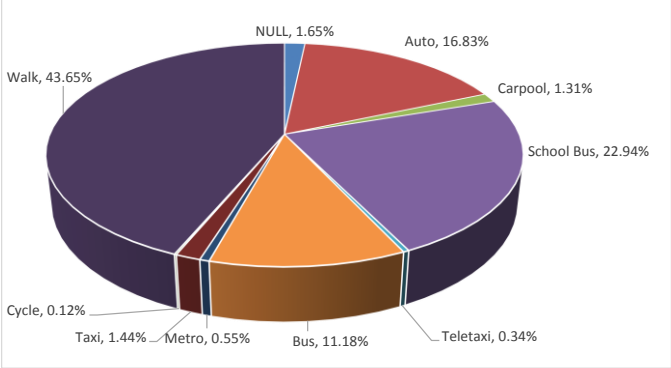

Figure 6) Mode choice of students from school to home

Since the education system in Iran is sex-segregated, mode choice of each gender category could be analyzed separately and decisions could be made accordingly for each gender group. Female students are slightly more active than male students. The results show that 51.12% of girls walk to school while 48.15% of boys walk to school. For the school-to-home trip, the walking mode share increase by 4% for female while the one for boys decreases by 16%.

Analysis of the car ownership and driver's license acquisition shows that these factors are not significant in the transportation mode of middle school and high school students. In the household of the students, there are two persons with a driver's license and the household owns a single car. This is normal among the Iranian society.

The survey respondents were categorized in 6 household income categories:

- less than 5 million rials per month
- between 5 and 10 million rials per month
- between 10 and 15 million rials per month
- between 15 and 20 million rials per month
- between 20 and 25 million rials per month
- more than 20 million rials per month

Investigating the relationship between mode choice and income categories shows that as the household income increases, the walking mode is more preferred by students belonging to low-income categories. In other words, moving from a lower income category to the higher one, decreases the propensity to choose walking as the travel mode.

Distance is one of the most important factors in the transportation mode chosen from home to school and vice versa. As the travel distance increases, students become reluctant to choose the walking mode. The walking mode share for home-to-school and school-to-home trips are 81% and 77% for those residing in 1 kilometer of the school, respectively. The walking mode share for home-to-school and school-to-home trips are 65% and 55% for those residing in a distance between 1 and 2 kilometers of the school, respectively. The walking mode share drops to less than 11% for both commuting trips when the distance becomes more than 3 kilometers. Based on this analysis the government could determine the optimized distance between schools in order to promote the walking mode.

VII. CONCLUSION

The analysis shows that Big Data Analysis provides useful insights in the transportation field. The dataset used in this paper contains travel mode choice of middle school and high school students along with their socio-economic characteristics and the transportation mode-specific features. The relationships between the chosen mode and various characteristics, in terms of conditional distributions, were tested. The result shows how a small piece of code written in a Hadoop project, Hive, could extract knowledge from such a large dataset.

As the next step, further analysis of the dataset through more sophisticated algorithms will be considered. The author will generate a travel demand model to predict the mode choice of a student given their socio-economic characteristics and transportation-specific features of each mode. Big Data-related tools and high-level programming in Python and R will be performed for the analysis.

REFERENCE

[1] A. Osman Idris, K. M. N. Habib, A. Tudela, and A. Shalaby. "Investigating the Effects of Psychological Factors on Commuting Mode Choice Behaviour." *Transportation Planning and Technology*, 38, no. 3 (2015): 265-76.

[2] K. N. Habib, W. El-Assi, M. S. Hasnine, and James Lamers. "Activity-Travel Behaviour of Non-Workers in the National Capital Region of Canada: Application of a Comprehensive Utility Maximizing System of Travel Option Modelling." *Paper presented at the Transportation Research Board 95th Annual Meeting*, 2016.

[3] M. Kamargianni, S. Dubey, A. Polydoropoulou, and C. Bhat, "Investigating the Subjective and Objective Factors Influencing Teenagers' School Travel Mode Choice–an Integrated Choice and Latent Variable Model." *Transportation Research Part A: Policy and Practice* 78 (2015): 473-88.

[4] Y. Ji, Y. Fan, A. Ermagun, X. Cao, W. Wang, and K. Das. "Public Bicycle as a Feeder Mode to Rail Transit in China: The Role of Gender, Age, Income, Trip Purpose, and Bicycle Theft Experience." *International Journal of Sustainable Transportation*, (2016).

[5] Krishnan, Krish. *Data Warehousing in the Age of Big Data*. Boston: Morgan Kaufmann, 2013.

[6] M. Kerzner, S. Maniyam, "Hadoop Illuminated", Elephant Scale LLC, 2013, ch. 12, pp. 47-54

[7] A. Samimi, and A. Ermagun. "Students' Tendency to Walk to School: Case Study of Tehran." *Journal of Urban Planning and Development* 139, no. 2 (2012): 144-52.

Moein Hosseini was born in Tehran, Iran, in 1991. He received a B.Sc. degree in Civil Engineering and a B.Sc. degree in Industrial Engineering from Sharif University of Technology, in 2014. He received an M.Sc. in Civil Engineering specializing in Transportation Engineering from the University of Toronto, in 2016. He is currently a Ph.D. student at Northwestern University. His research interests include Big Data Analysis in Transportation, Machine Learning in Transportation, Social Network Analysis, and Operation Research.