Big Data in Transit

Public Transit systems manage the flow of millions of riders daily, but the big data collected from transit systems can be burdensome to analyze and understand. Case in point, New York City had 149 million train riders in the month of September 2014, and there were 6.1 million riders on September 23, 2014 alone. This no doubt provides a huge amount of data to analyze. The current research project analyzes the train rider data that are typically collected at turnstiles, and that would likely be date and time of day. Analysis of the time data could include grouping the data into time intervals (e.g., 8:00-9:00, 14:00-15:00) to reveal the peak periods of traffic flow as well as the nadirs. The goal of this research project is to develop a computer-based system to organize time data from multiple turnstiles, import those into a statistics/graphics program, and create a histogram of rider data based on a pre-defined time interval (e.g., 30 minutes). In lieu of using actual turnstile data, a homebrew, customized Visual Basic program was written to create a .csv file with dates and times. Then, using the language R—a language conducive to analyzing big data—data structures are created, and interval times are counted and saved to a different text file. Finally, the multidimensional Python language is used to read that text file and create a histogram showing the interval turnstile data. Additional research is being done to perform statistical analysis on the times data or inter-day data. When used by the Transit Authority, this approach will allow them to deploy trains and buses more efficiently.

Keywords: *public transit, metro passenger data, turnstile, Transit Authority, train, big data, Python programming language, Python, R programming language, R*