Executive Summary

Big data is everywhere and businesses that can access and analyze it have a huge advantage over those who can't. One option for leveraging big data to make more informed decisions is to hire a big data consulting company to take over the entire project. This method requires the least effort, but is also the least cost effective. The problem is that the know-how for starting a big data project is not commonly known and the consulting alternative is not very cost effective. This creates the need for a cost effective approach that businesses can use to start and manage big data projects. This report details the development of an advisory tool to cut down on consulting costs of big data projects by taking an active role in the project yourself. The tool is not a set of standard operating procedures, but simply a guide for someone to follow when embarking on a big data project. The advisory tools has three steps that consist of data wrangling, statistical analysis, and data engineering.

Data wrangling is the process of cleaning and organizing data into a format that is ready for statistical analysis. The guide recommends using the open source software and programming language of R. The next step is the statistical analysis portion of the process which takes the form of exploratory data analysis and the use of existing models and algorithms. The use of existing methods should always be attempted to the highest performance before justifying the costs to pay for big data analytics and the development of new algorithms. Data engineering consists of creating and applying statistical algorithms, utilizing cloud infrastructure to distribute processing, and the development of a complete platform solution.

The experimentation for the design of our advisory toolwas carried out through analysis of many large data sets. The data sets were analyzed to determine the best explanatory variables

to predict a selected response. The iterative process of data wrangling, statistical analysis, and model building was carried out for all the data sets. The experience gained, through the iterations of data wrangling and exploratory analysis, was extremely valuable in evaluating the usefulness of the design. The statistical analysis improved every time the iterative loop of wrangling and analysis was navigated.

In house data wrangling, before submission to a data scientist, is the primary cost justification of using the advisory tool. Data wrangling typically occupies 80% of data scientist's time in big data projects. So, if data wrangling is self-performed before a data scientist receives the data, then less time will be spent wrangling by the data scientist. Since data scientists are paid very high hourly wages, extra time saved wrangling equates to direct cost savings. This is assuming that the data wrangling performed before a data scientist takes over is of adequate quality.

The results of applying the advisory tool may vary from case to case, depending on the critical skills the user possesses and the development of such skills. The critical skills begin with coding in R and Python as well as knowledge in the statistical methods of choice. Basic knowledge of statistics, and any programming language is a must to begin utilizing this guide. Statistical proficiency is the limiting factor in the advisory tool. The best start for doing a big data project on one's own is to first learn R and become familiar with the statistical libraries it contains. This allows data wrangling and exploratory analysis to be performed at a high level. This project pushed the boundaries of what can be done with big data using traditional computer framework without cloud usage. Storage and processing limits of traditional computers were tested and in some cases reached, which verified the eventual need to operate in the cloud environment.