Introduction

According to research, over 90% of the world's data was generated within the last two years[1]. Following an exponential trend that has exceeded even Moore's famous law for processors, our ability to store data per-capita has reached levels that were largely inconceivable just a decade ago. This spectacular increase in the amount of data we generate and store has essentially broken many of the traditional systems and algorithms we had relied upon to manage our data. The emergence of the "Big Data" field was an effort to find solutions to the ever-growing challenges of working effectively with data sets that simultaneously are growing ever-more massive and ever-more important to our daily lives. Though there are many different projects under heavy development, the Hadoop framework occupies center stage in our present Big Data toolset.

The goal of this project was to focus on a particular segment of this growing avalanche of data: natural text. The explosive rise of social media over the past decade combined with the increased technological resources at the disposal of free "text aggregation" services such as Project Gutenberg has resulted in a proliferation of text-based opinions, knowledge, and nonsense. Gaining insight from such massive and often disorganized data poses serious challenges but is a task well-fitted for a framework such as Hadoop.

Throughout the course of handling the data sets (rationale and setup instructions described in-depth in Appendix C), several Hadoop "ecosystem" projects were a particularly good fit to assist in solving aspects of the problem. These projects include Hive, Pig, Oozie, Hue, and most prominently, Mahout.

Problem Description and Specification

Two different data-sets were utilized during the course of the project: millions of sentences in various languages from the Leipzig Corpora collection and hundreds of thousands of full-text books from Project Gutenberg. One of the primary reasons for the initial choice of the Leipzig Corpora data was the availability of previously-computed analysis done on subsets of the data that makes it possible to verify some of the initial results of our operations on larger segments of the data. In addition to this analysis (easily accessible via the CorpusBrowser interface discussed in Appendix C), individual partitions of the massive data-set can be downloaded as traditional SQL databases and interacted with through the familiar (and fast!) RDBMS syntax.

The overall size of the Leipzig Corpora data (approximately 120 GB) is vastly out of any of our cluster machines' ability to store in-memory. Having to spill to disk means that

[1] SINTEF. (2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. Retrieved January 27, 2014 from www.sciencedaily.com/releases/2013/05/130522085217.htm

answers will not be as immediately attainable as is common with RDBMS queries, but what Hadoop may sacrifice in speed it more than makes up for in size and power. Using Hadoop, we want to aggregate the disparate conventional databases that comprise the Leipzig Corpora and perform queries and analytics on all of them at once. The data warehousing and broad querying capabilities of Hadoop are often what initially attracts firms to the framework; being able to aggregate various separate databases or denormalized tables from conventional, unitary databases is of tremendous and growing utility to a variety of businesses.