Analyzer of very large Web server log files

Kate Trofimova, Miro Banovic

November 23, 2017

Project Description: Analyzer of very large Web server log files

1 Introduction

For the course Software Engineering for Economists we received a team assignment. This assignment entails the coding, documentation and description of a computer program. Our team consists of Kate Trofimova and Miro Banovic, and we set out to make a program that is actually useful and needed. As there are millions of different expert forums in the internet, finding a way to analyze these from the outset could be important. Therefore, we are aiming to program an Analyzer of very large Web server log files. In the following, the goal of this program will be described, and why it is needed. Afterwards, we will go into detail with regard to what programming methods and libraries will be used.

2 Goal

Our program is inspired by the Udacity course "Intro to Hadoop and MapReduce" in the sense that we will be using the same tools. Our project specifically, however, should work in a way that it uses discussion forum data as its input and gives several outputs:

1. Information of the most active users
2. Correlation between the length of a post and the length of the answers
3. Top 10 tags, ordered by the number of question they appear in

3 What will be used?

As discussed above, we will be using MapReduce and Hadoop for this. MapReduce is a programming model and an associated implementation for processing and generating big data set with a parallel, distributed algorithm on a cluster (Wikipedia, 2017). It works by taking unstructured data blocks, sorting each block individually, and then forming new blocks from the sorted parts of the initial blocks. Hadoop, on the other hand, is a flexible and available architecture for large scale computation and data processing on a network of commodity hardware, which can be used for several purposes:

- Searching
- Log processing
- Recommendation systems
- Analytics
- Video and image analysis
- Data retention

In this exercise, the analytics aspect of the architecture is best fit to reach the goal of our program.

4 Way of working

While this document marks the beginning of our documentation/description process, we will start programming in the upcoming weeks. In this, we expect to encounter several challenges. Most importantly, the level of expertise in our team varies wildly, which is why we aim to achieve a fair distribution of tasks by splitting up work according to named expertise. In this, it is likely that Miro Banovic will take over most of the documentation of the project, while Kate Trofimova is expected to do the major share of programming work. As two of our team members dropped the course and thus left our team, we think of this to be an approximately even distribution of work.