CSC Project: Distributed Web Scraping

Group: Michail Roesli (V00853253), Quinn Gieseke (V00884671), Blake Smith (V00850827)

Problem

We'll provide distributed web scraping as a service which utilises idle mobile phones. This will allow citizens to support organizations by providing their phones or computers to improve the performance of the system.

Plan

Our initial plan consists of several basic steps:

Phase 1

1. Simulate a RAFT cluster within a single linux environment
2. Communicate between the RAFT cluster and Android VMs
3. Distribute fixed list of URLs to web-scraping nodes in an efficient manner

Phase 2

1. Incorporating user-provided plugins into the web-scraping code dynamically
2. Dynamically generating urls to scrape
3. Integrate a distributed write database
4. Meaningful data reduction
5. Multiple user-provided plugins running concurrently

Design and Implementation

We'll use RAFT to connect computers and mobile devices into a reliable distributed network (See figure 1 and 2). The leader node will be used to communicate with the phones which scrape the urls collected by the followers which are crawling a website. The implementation will be done using Kotlin and Java.

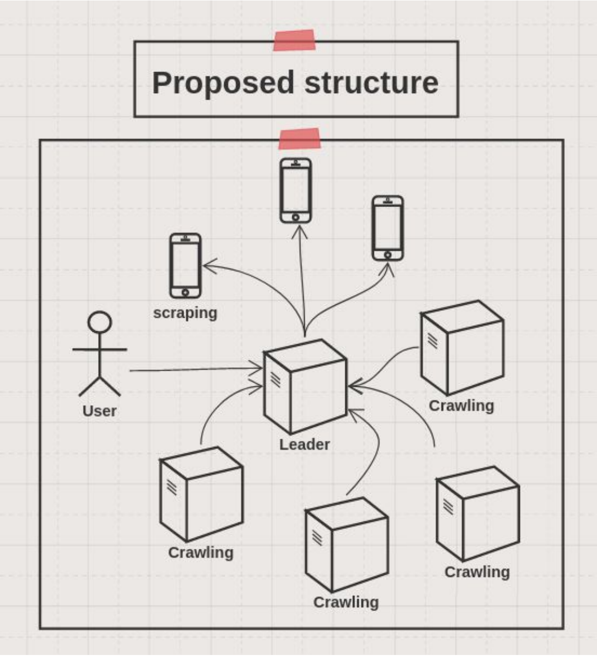

Figure 1. Proposed Structure

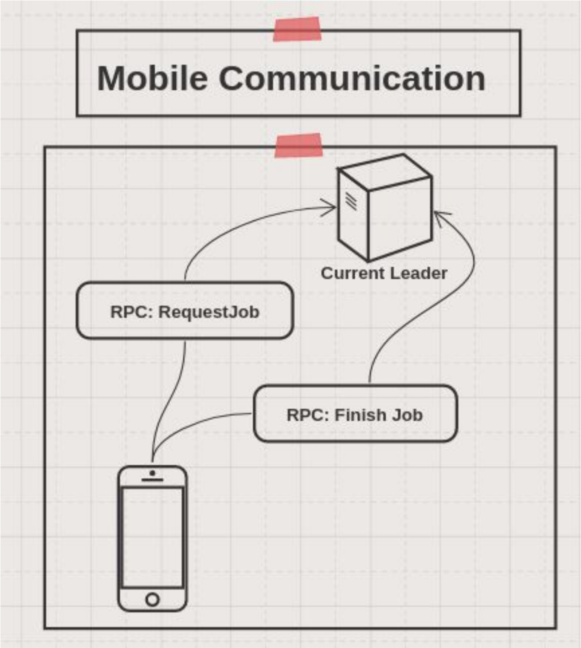

Figure 2. Mobile Communication