phData does not believe in traditional interviews as they do not mimic the real world. In the real world, you'll be given project based work as part of a team and have time to perform research to solve the assigned task. As such, phData interviews, are project based.

This project is for a data engineer or solution architect role. If you are more interested in another role such as dev ops, please inform your contact you've been given the wrong project. After completing the assignment, you'll be requested to provide a short demo of the work you've completed and the thought process you used.

Overview

The customer runs a website and periodically is attacked by a botnet in a Distributed Denial of Service (DDOS) attack. You'll be given a log file in Apache log format from a given attack. Use this log to build a simple real-time detector of DDOS attacks.

Although you have tremendous amounts of freedom in designing this system, we don't want you spend too much time on the project. After all this is a replacement for an interview, so plan on spending somewhere between 3-12 hours depending on your familiarity with the technology you choose.

Non-Requirements

- Completing in a specific amount of time. Life is busy and chaotic. We understand you will not be able to work full time on the project.
- Machine learning. While machine learning **could** be used to solve this problem, but is in no way required.
- There is no need to run this code at scale. Everything should be done either a single node pseudo cluster or a small cluster.
- An exact end result. Two candidates given this assignment will find different solutions. Feel free to choose your own adventure as long as the base requirements are met.

Requirements

- Ingest
    - Read a file from local disk and write to a message system such as Kafka.

- Detection
    - Write an application which reads messages from the message system and detects whether the attacker is part of the DDOS attack
    - Once an attacker is found, the ip-address should be written to a results directory which could be used for further processing

    - An attack should be detected one to two minutes after starting

Recommendations

- Use the Cloudera Quickstart VM for developing the application.
- Time management
  (We cannot stress these two enough.)
    - Build the simplest possible solution first, utilizing your favorite programming language.
    - Don't get stuck on one aspect of the project.
- Ask questions and use the internet for research
- Think about how the solution will scale to hundreds or thousands of web servers
- Focus on your core strengths
- Give us feedback. We'd love to hear **any** feedback you have.