**Contributions to Apache Hadoop Automated Installation Using Chef**

**Proposal:**

*My proposal for my project will be making contributions to Hadoop Cookbook on github (*[*https://github.com/caskdata/hadoop\_cookbook*](https://github.com/caskdata/hadoop_cookbook)*), fulfilling option 1 requirements for the project*.

**Background:**

Apache Hadoop is a framework that enables efficient distributed processing of large data among a cluster or several clusters of computers. A master node assumes charges of several slave nodes where large data can be divided and processed individually before being recombined to produce a specified output. Chef, on the other hand, is a configuration management used to orchestrate the setup and configuring of virtual and physical machines. Chef is primarily written in Ruby, and utilizes files called “recipes” which instruct how a machine or group of machines should be setup for deployment purposes. Combining the two creates a structure which can be self-healing in nature and also spun-up on the fly for quick use and application. This is the foundation for the Hadoop Cookbook project, which is in need of several immediate modifications including:

* Move storm-env.sh creation to storm recipe
* Switch to new Ruby *hash syntax*
* Add non-cookbook files to chefignore

**New Concepts:**

I expect to learn many new concepts while contributing to this project. While I have some basic knowledge in configuration management, I will be using a completely new tool that I have never worked with previously, Chef. Chef is almost entirely based off of Ruby, and will require me to learn this language as well. This project will also help solidify my knowledge of Hadoop which I am currently researching as a part of Mentored Research.

**Comparative Analysis of Licensing:**

Apache License 2.0

* Allows for free use of the software for any intended purpose, distribution, modification, under specified terms of the license. Does not acknowledge any royalties. In every licensed file changed, a notification must be added stating that changes have been made to that file.

GNU General Public License (GPL)

* Allows for modification as long as changes from the original are tracked and dated, and as long as the resulting product is still GPL licensed. Additional licenses may not be used.

MIT license

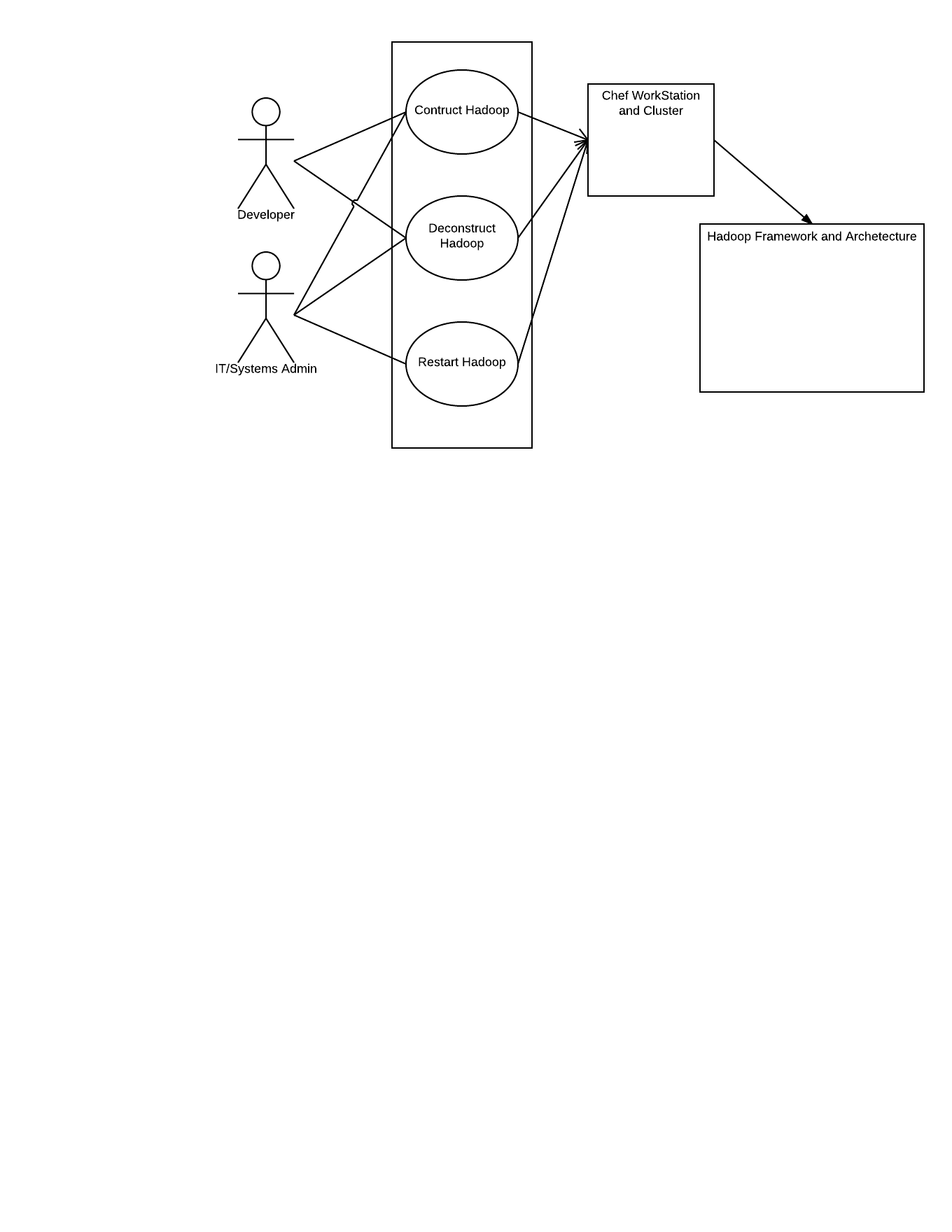
* Allows for any modification and use of the software, including for commercial use, as long as the modified version retains the license.

Mozilla Public License

* Attempts to balance between proprietary development and open source development. Can be modified and distributed, yet separate licenses can be included for proprietary software elements, as long as the proprietary code is kept in separate files.

**Choice of licensing**

Apache License 2.0 was chosen for my project’s license since the original license of Hadoop\_Cookbook is also an Apache License. This also gives a large degree of freedom for using my code as it is not intended for proprietary use.



My Project GitHub URL (Hadoop\_Cookbooks\_KS): https://github.com/ksayers91/Hadoop\_Cookbooks\_KS