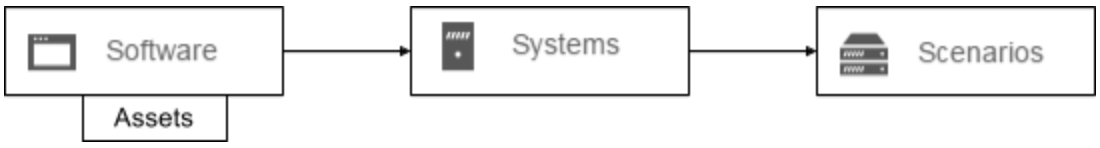
Cloud DevOps Resources

Project Proposal

In collaboration with Jackpine Technologies, we will be developing assets for use in their primary product, CONS3RT, a cloud orchestration and service management software suite. The suite allows a user to build a system by picking an operating system and software that he wants the system to run. Systems can be run individually or combined with other systems to make up a scenario. CONS3RT supports multiple cloud services so the user decides which cloud provider he wants to deploy the system/scenario to.

1. Vision and Goals Of The Project:

We will be creating a deployment of a Hadoop Cluster that will be included in CONS3RT. It will consist of master and slave systems running the Hadoop open source software, and it will be scalable to any number of machines.
The rest of the machines in the cluster act as both DataNode and NodeManager. These are the slaves.
We will write an installation script for installing necessary assets/applications on these systems, which will be platform agnostic. Once deployed, a user will be able to run parallel processes on the system for analyzing large amounts of data efficiently.



Users/Personas Of The Project

Users that require analysis of Big Data will be able to deploy a Hadoop cluster in any cloud and run their application on the cluster. Examples would

be corporations, whose size is between medium and large, since they will have large amounts of data to analyze. Applications that generate this kind of data would be social networking sites, financial data analytics, network traffic security analysis, etc.

2. Scope and Features Of The Project:

- Provide an easy installation process through CONS3RT.
- Include programs that will support parallel programming, like OpenMPI, OpenMP, SIMD Vector support, etc.
- Support deployment on multiple clouds with different hardware architectures and operating systems.
- Allow the cluster to be scalable, but also have limited resources, which will be configured by the user.
- Provide a secure deployment of the cluster to prevent data interception or corruption.

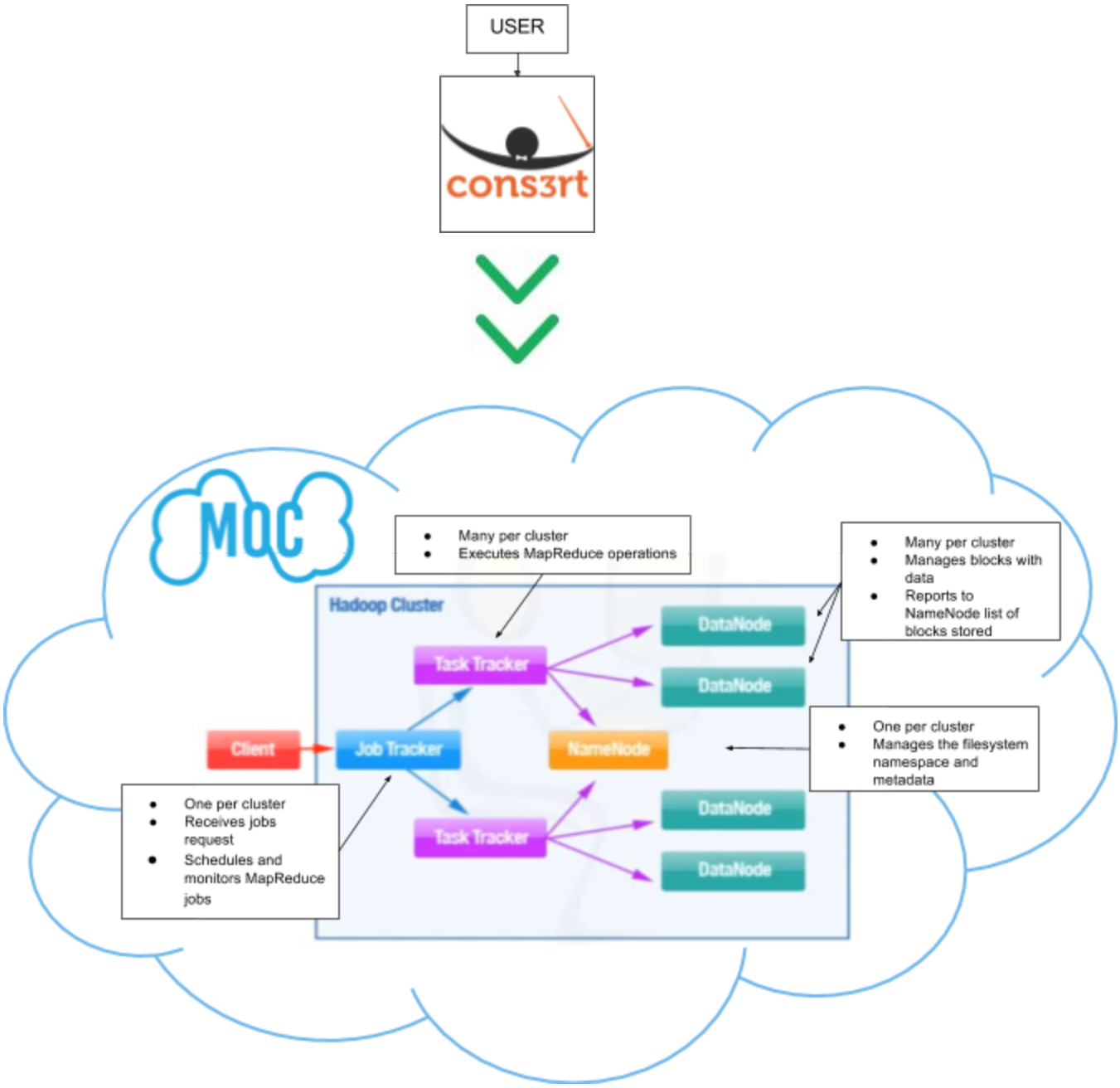
3. Solution Concept

Global Architectural Structure Of the Project:

The project consists of installing a multitude of assets in several systems. The systems will be divided into two groups, one master and the rest slaves. Aside from the Hadoop software, the additional programs installed to enable parallel execution will be the same on each system. However, the systems could be running different hardware, which is usually the case in private clouds with commodity hardware. Our script will allow the installation to be successful on a wide range of platforms.

The following is a list of the different programs aside from the Hadoop cluster that we will include:
MPI - API that allows distributed computing across nodes. Open source implementations include OpenMPI and MPICH2.
Openmp - API that allows parallel processes to execute on multiple cores simultaneously.
Pthreads - Allow multiple concurrent processing threads to execute in a CPU.
GPU - Allow for programming of GPUs. Programs include Cuda for Nvidia GPUs or OpenCL for other GPUs.

The programs (assets) like Hadoop will each have a dedicated script that will gather the information about the system it is on and then download and install the required libraries and dependencies for the programs to run on that system. However, the complex part will be for the systems to cooperate so that only one is designated the master and the rest as slaves. The scripts will most likely be written in python, although it might be better to do it in bash since there are convenient system commands already available.



Design Implications and Discussion:

Installing a Hadoop cluster involves installing it via a packaging system applicable to the operating system (Linux). It is vital that we split the components onto separate machines. One machine in the cluster would act as the NameNode and another machine as the JobTracker, these are the masters. Depending on the workload, the TaskTrackers and DataNodes (the slaves), are either run on dedicated machines or on shared infrastructure.

4. Acceptance Criteria

Minimum acceptance criteria would be a scenario of a master machine and multiple slave machines running the Hadoop software. The machines would be homogeneous and running the same version of the operating system, most likely the latest version of Ubuntu, which is 16.04.1. In addition, the hardware on the system will be the same, most likely intel x86_64 architecture. Also, we will have only two nodes, a master and a slave. Once we are able to do that, it should be fairly easy to scale to more slave machines.

5. Release Planning

Release #1 (Week 5):

Hadoop Cluster and Programs Setup:

Hadoop cluster deployed in a test environment with all the necessary software.

Release #2 (Week 7):

Automated Setup in OS and Hardware

Create a script that will install all the necessary software on a single operating system with limited hardware variability.

Release #3 (Week 9):

Platform Independent Extension

Extend the script to work with multiple distributions of Linux.

Release #4 (Week 11):

Platform Independent Extension

Extend the script to work on a wider range of hardware architectures.

Release #5 (Week 13):

Asset Addition

Add more assets in the realm of GPUs and FPGAs. This is a stretch goal but would be amazing since Amazon just recently announced that they will include FPGAs in their cloud. Programs that might be included are Cuda or OpenCL and Xilinx.