

# 15-319 Introduction to Cloud Computing

## Project 1 Introduction to the Hadoop Environment

**Assigned Date:** January 14<sup>th</sup>, 2010  
**Deadline:** January 28<sup>th</sup>, 2010 at 11:59 p.m.

Attached Document: Instructions to Complete Project 1

### Goals:

1. Understand the Cloud infrastructure at CMUQ (Qloud).
2. Request your own Cloud system (4 nodes) from Qloud.
3. Configure Hadoop on your cloud in order to submit jobs remotely.
4. Running Hadoop example code to get familiar with the command line interface
5. Understand the HDFS file system and essential HDFS commands.
6. Setup eclipse and get familiar with writing Hadoop code and running jobs through it on your Cloud.
7. Gain experience in processing large amounts of data by running a word count application on a large text file on both a single-threaded application as well as a distributed Hadoop job.

### Background:

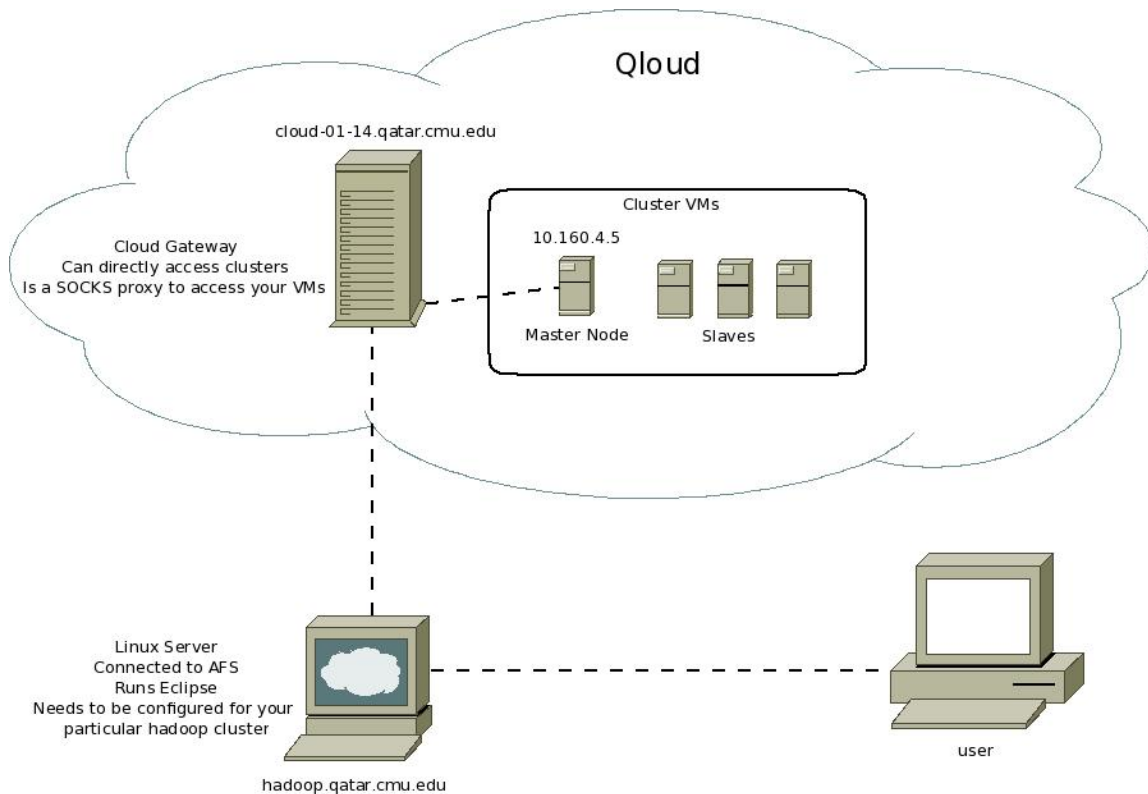
#### Cloud Infrastructure in CMUQ (The Qloud)

CMUQ has dedicated cloud infrastructure running on a 14-blade IBM Bladecenter with a total of 112 physical cores and 7 TB storage. The cloud is managed using advanced management software (IBM Cloud 1.4) with Tivoli provisioning manager which allows for the cloud resources to be provisioned via software.

Users can request *virtual machines* (VMs) on the cloud with customized CPUs, RAM and disk space, as well as custom OS/software images. These machines run Red Hat Linux on top of the Xen virtualization platform and will be pre-configured with Java and the Hadoop SDKs. Since these resources are virtualized and provisioned through the management server, cloud resources can be accessed only through the *Cloud Gateway* (see Figure 1).

From your computer, you will need to log on to `hadoop.qatar.cmu.edu` server which is where you will run Eclipse and communicate to the provisioned cloud. This machine does not run any of the Hadoop code, it just acts as a liaison with your provisioned cloud.

For the purpose of this project, you will request a pre-set Cloud configuration of 4 nodes. In this project you will learn how you can request a Cloud from the cloud management software and work with those machines once provisioned.



**Figure 1 - CMU Cloud Infrastructure Logical View**

### **What is Hadoop?**

Hadoop is an open-source software framework by the Apache Software Foundation that consists of several projects that are useful to cloud computing. Among other things, it contains an implementation of MapReduce (a distributed programming framework), as well as a distributed file system, the Hadoop Distributed File System (HDFS). Hadoop is fast-becoming the industry standard implementation of MapReduce, and companies like Yahoo!, IBM, Facebook, LinkedIn, Last.fm etc. are using Hadoop to manage large-scale data computation.

### **Find more information at:**

<http://wiki.apache.org/hadoop/>

## **Part I: Setting up the system (see Part I in the instructions document)**

1. Login to `hadoop.qatar.cmu.edu` to set up Firefox for accessing the cloud management server.
2. Login to the Cloud management system and request a Cloud. Your Cloud should be provisioned in about 24 hours.
3. **Deliverable:** Take a screenshot of the cloud management system webpage detailing your Cloud and the IPs, and name it as “part1.jpg” and submit it.

## **Part II: Hadoop command line interface (see Part II in the instructions document)**

1. Once you have your cloud provisioned, login to the Master node as root.
2. Deploy Hadoop and start the required daemons on the Master node.  
Run the PiEstimation example code provided in the Hadoop package.
3. **Deliverable:** Submit the run logs and the final value of Pi obtained through this method “part2a.txt”.
4. **Deliverable:** Login to the JobTracker web interface on your master node and take a screenshot of your job in progress. Submit the screenshot taken as “part2b.jpg”.

## **Part III: HDFS command line interface (see Part III in the instructions document)**

1. Configure the HDFS file system of your Cloud and learn the basic HDFS commands.
2. Practice adding files to your Cloud HDFS file system.

## **Part IV: Configuring Hadoop through eclipse (see Part IV in the instructions document)**

1. An eclipse environment is pre-installed on (`hadoop.qatar.cmu.edu`), launch the environment.
2. Follow the instructions to configure Eclipse to use hadoop and MapReduce plugin.

## **Part V: WordCount program as a single-threaded application**

In `/afs/qatar.cmu.edu/course/15/319/data/`, there is an 11 GB text file.

1. Write a simple application in the language of your choice to count the number of times each word occurs in this document. Your program should output the time it takes to complete the task as well as the number of times each word occurs in the file.

2. **Deliverable:** submit the runtime as well as a copy of your program. Create a folder called “part5” and place your source code. You do not have to submit the output text file. Be advised that the output text file might be large.

## **Part VI: WordCount on Hadoop (see Part V in the instructions document)**

1. Use the HDFS put command to copy the 11 GB text file from the AFS store to your Cloud’s HDFS system.
2. Now create an eclipse Hadoop project, copy the WordCount Hadoop example to your project and run this project on your Cloud to completion.
3. **Deliverable:** Submit a screenshot of eclipse running your project to completion, as well as the eclipse console log. Take a screenshot of your eclipse screen after the job ran successfully, and submit it as “part6a.jpg”, copy the run log from the console as “part6b.txt”.

### **Submission:**

Follow the instructions in the attached document. Add all the deliverable files from each part into a single zip file (project1.zip) and place it in:

`/afs/qatar.cmu.edu/course/15/319/handins/username/`

This file is to be submitted once and the final timestamp on the server will determine your submission time.

### **Grading:**

As mentioned in the syllabus, this project is worth 10% of your final grade. You have two weeks to finish the project.