Hadoop Architecture Labs – Amazon Web Services Elastic Map Reduce Version

The labs in this directory are designed for working within an Amazon Web Services (AWS) Elastic Map Reduce (EMR) Cluster, which is currently at version 5.8.1. While the general principles are the same when using an Amazon Web Services Elastic Map Reduce cluster instead of Think Big Academy's vagrant instances, the specific details and file locations are different.

These labs familiarize you with four major components of a Hadoop cluster and how to move data among them:

- The local file system
- The Hadoop Distributed File System (HDFS)
- YARN, the cluster operating system
- The Hive Warehouse

Preparing Your Environment

Tools You'll Need On Your Local Computer

Much of the work in this course will be done at the Linux command line on your cluster. That means you will require a three pieces of software on your local laptop or PC:

1. A Unix or Linux terminal emulator.
2. An ssh client.
3. A modern web browser.

   piece of secure software that allows you to connect to your cluster over the Internet. The standard tool we use for that is the Unix or Linix Secure Shell or `ssh`.

Macintosh users already have `ssh` installed as part of their operating systems.

Saving Information About Your Cluster

Your instructor has given you an IP address for your Elastic Map Reduce cluster. You should write that down and keep it handy. Your instructor should also have given you a cryptographic key for your cluster.

For the purposes of the labs in this course, we're going create a shell environment variable called `cluster` that has your cluster's IP address. We're also going to assume you have been given a cryptographic private key called `crypto.pem` which you can use to access your cluster.

First, you

The majority of free disk space on your EMR cluster is in two mounted file systems, `/mnt` and `/mnt1`. We'll be doing most of our work in the local file system on `/mnt`.

Unarchiving Your Data

You should have been given a zip file of the class materials, which we will assume here is named `hadoop-class.zip.` This file should have been placed in one of your AWS Elastic Map Reduce partitions, `/mnt`. You should unzip that file into a new directory called `hadoop-class` and list its contents using the following commands

```
   cd /mnt
   unzip sparkclass.zip
   cd sparkclass
   ls
```

You should see the following files in the file listing.

```
data        exercises   handouts    images      slides
```

You are now ready to start Lab 1 on AWS