Apache Hadoop MapReduce

Provisioning Your Cluster and Getting Started

As a recap to what we did in Project 1, please provision a Hadoop cluster using Amazon's Elastic MapReduce using the following instructions. The main difference here is that we will be connecting to the Master Instance over SSH and running some of the sample programs that are included with the Hadoop installation.

1. Launch a 5-node cluster (1 Master + 4 Core) (with `m1.small` for all roles) using Amazon EMR, following the instructions outlined in Project 1. You may use spot pricing if the savings are significant compared to on-demand pricing.
2. Once provisioned, login to the Master node of your cluster. The EMR console should tell you the DNS name of the master instance. If you have trouble connecting, make sure the security group associated with the master instance has SSH (Port 22) open to all IPs. Remember that EMR clusters use a special AMI which allows you to log in as the `hadoop` user.
3. List the entries in the home directory of the `hadoop` user, you should find multiple jar files, including `hadoop-examples.jar`.
4. Run `hadoop jar hadoop-examples.jar` to see a list of sample programs that are available with Hadoop.
5. Run the sample pi application (using 8 maps and 10,000 samples per map)to ensure that your Hadoop cluster is working correctly.

Click through to the next page when you are ready.