QuickStart

Get up and running fast

The fastest way may be to just install a pre-configured virtual Hadoop environment. Two such environments are:

- The Cloudera QuickStart Virtual Machine. This image runs within the free VMWare player, VirtualBox, or KVM and has Hadoop, Hive, Pig and examples pre-loaded. Video lectures and screencasts walk you through everything.
- The Hortonworks Sandbox. The sandbox is a pre-configured virtual machine that comes with a dozen interactive Hadoop tutorials.

Cloudera also provides their distribution for Hadoop (Apache 2.0 Licensed), including support for Hive and Pig and configuration management for various operating systems.

If you want to work exclusively with Hadoop code directly from Apache, the following articles from the website will be most useful:

- Single-Node Setup
- Cluster Setup

Note for the above Apache links, if you're having trouble getting "ssh localhost" to work on the following OS's:

Window Users To start ssh server, you need run "ssh-host-config -y" in cygwin environment. If he ask for CYGWIN environment value, set it to "ntsec tty". After you can run server from cygwin "cygrunsrv --start sshd" or from Windows command line "net start sshd".

Mac Users In recent versions of OSX, ssh-agent is already set up with launchd and keychain. This can be verified by executing "echo $SSH_AUTH_SOCK" in your favorite shell. You can use ssh-add -k and -K to add your keys and passphrases to your keychain.

Multi-node cluster setup is largely similar to single-node (pseudo-distributed) setup, except for the following:

1. The hostname or IP address of your master server in the value for fs.default.name, as hdfs://master.example.com/ in conf/core-site.xml.
2. The host and port of the your master server in the value of mapred.job.tracker as master.example.com:port in conf/mapred-site.xml.
3. Directories for dfs.name.dir and dfs.data.dir in conf/hdfs-site.xml. These are local directories used to hold distributed filesystem data on the master node and slave nodes respectively. Note that dfs.data.dir may contain a space- or comma-separated list of directory names, so that data may be stored on multiple local devices.
4. mapred.local.dir in conf/mapred-site.xml, the local directory where temporary MapReduce data is stored. It also may be a list of directories.
5. mapred.map.tasks and mapred.reduce.tasks in conf/mapred-site.xml. As a rule of thumb, use 10x the number of slave processors for mapred.map.tasks, and 2x the number of slave processors for mapred.reduce.tasks.
6. Finally, list all slave hostnames or IP addresses in your conf/slaves file, one per line. Then format your filesystem and start your cluster on your master node, as above.

See Hadoop Cluster Setup/Configuration for details.