Working with Hadoop MapReduce: My experiences by Bina Ramamurthy (bina@buffalo.edu)

There are several approaches to working with Hadoop Mapreduce. Here are some of them and my experience with each. You may choose one of the many possibilities based on your background and expertise.

Hadoop Distributed File system (HDFS) provides the big data infrastructure and MapReduce(MR) provides the programming model or the algorithm to process the data stored in the HDFS. This is the simplest model. There are other models involving HBASE, Hive etc. But we will cover only the HDFS/MR in this document.

1. You can implement HDFS/MR from the scratch using the tutorial provided Michael Noll. It is available here: http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

This will require Ubuntu installation. You can follow up single-node cluster with multiple node cluster tutorial available from Michael Noll.

My experience: I have done both: single as well as multiple cluster; this is how I got started. I have used this setup to demo at conference and at tutorials. This is a very good place to start. I still use the single node version on (Ubuntu) on my laptop.

2. The second one I tried and still use is the VMware virtualized image from Cloudera Inc. (www.cloudera.com). You can download the CDH3 or the latest CDH4 and install only the components you need.

My experience with this: There are many videos supporting the concepts. Documentation is very good. I was fortunate to have taken an in-person full day tutorial with Aaron Kimball, co-founder of Cloudera.

3. Of course, Hadoop is an Apache project now: The details are available from the source at: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/SingleCluster.html

You can also look at all the documentation.

My experience: I have used this sometimes. People who are used to working Apache project, especially ones who plan to contribute to the source etc. can use this download.

4. And then there is Yahoo: a prolific contributor and supporter of the Hadoop project. See http://developer.yahoo.com/hadoop/

My experience: An older version of HDFS was supported when I used it. This site explains the concepts very nicely. This site is well-liked by students.

5. Last but not least, both amazon web services (www.aws.amazon.com) and Google App Engine (GAE) support MR workflows.

My experience: These are good but you will not learn the internal working (code) of Hadoop Mapreduce. But you get things done faster than any of the solutions above. You may have to pay for some of the services.