Caleb Post

Lab 1 : Introductuion to MapReduce & HDFS

Question 1: Explain the MapReduce concept in your own words using a simple example.

The map reduce concept is in essence decoupling the processing of the data from the data itself. You write a method which has no knowledge of the amount of data it is running on, it could be one line, it could be terabytes of data. This map function breaks the data apart and processes it into whatever form you want. The power of this is that Hadoop can then break the data into smaller sub problems, distributing them among the nodes connected, which can then in turn distributed if needed and the network is set up that way. The answers are passed back to the master node then, and the reduce step takes place. The reduce step is the combination of all the smaller sub problems into a final output, whatever that output may be.

The power of this approach is that all the map steps can be done in parallel, breaking large computations down into relatively small processes on many machines at once. The data from this then needs to be combined into the final answer, the reduce step.

An example of this would be calculating the maximum temperature from a weather satellite network. This satellite network is broadcasting gigabytes of temperature data to the NSF for study each day. This data is added in text files to their Hadoop cluster.

When the max temperature job is run, the map function is run on each node containing pertinent temperature files. We don’t know how many times this will run, but it will be run over all the temperature data.

The intermediate temperatures are then passed into a reducer. The reducer pulls the maximum temperature from all the processed temperatures and writes it as output.