James Devine

Evaluating the Scalability of Hadoop in a Real and Virtual Environment

What is Hadoop?

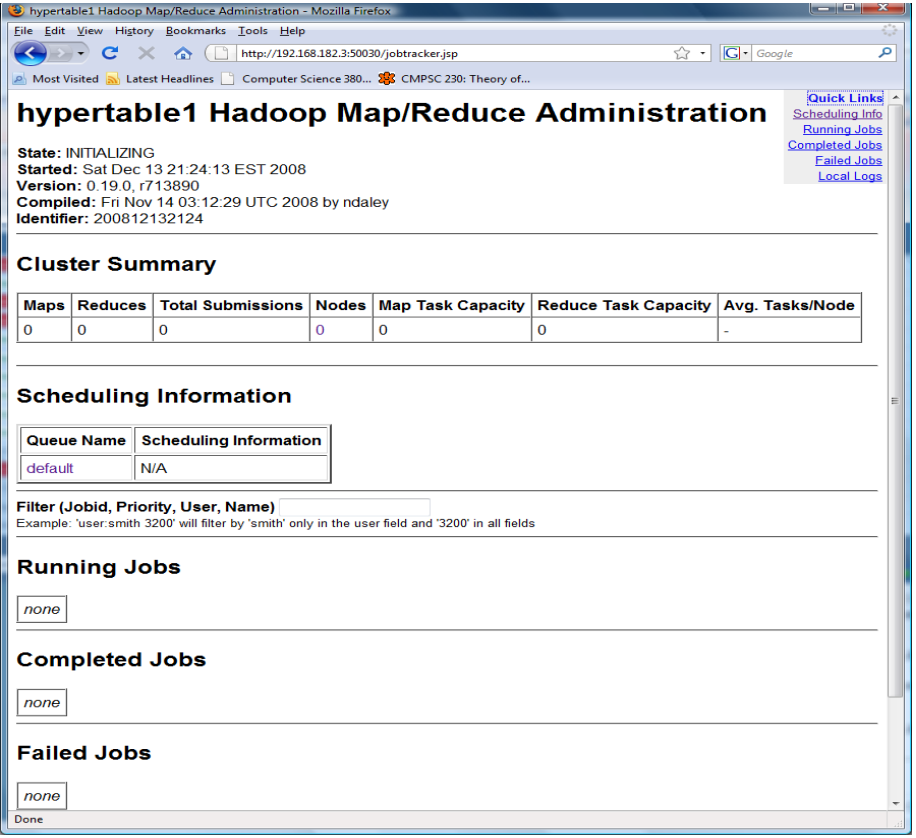
- Hadoop is a distributed file system written in Java that supports *MapReduce*
  - MapReduce breaks up an input file into many smaller pieces to be computed on individual nodes of a cluster and then reduces all of the results to a single file
- Default installation provides 3 way replication which stores each file in three locations

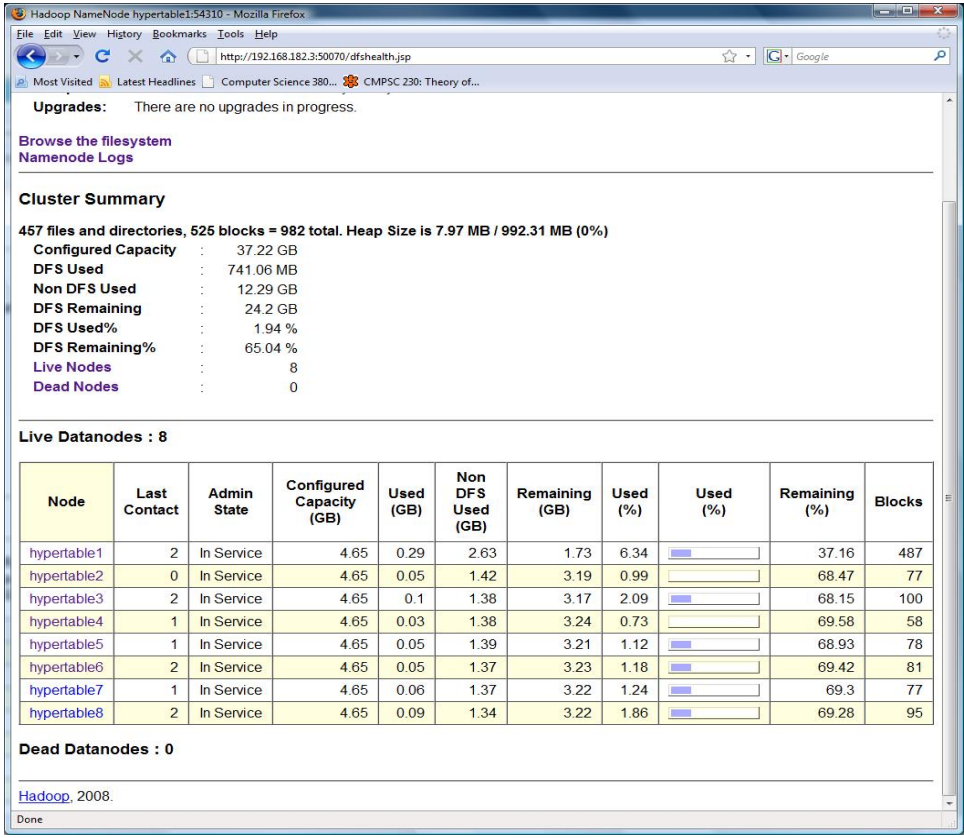
Installation of Hadoop

- Download package from the Hadoop website
- Extract package
- Set up an ssh key
- Edit the hadoop-site.xml, slaves, and masters files
- Copy the Hadoop files to each nodes in the cluster
- Start the dfs and mapred services

Managing Hadoop

- Hadoop has two main web GUI's for the cluster





•The job tracker webpage shows the status of Jobs that are running on the cluster

•The DFS health webpage shows the status of The file system and provides access to the HDFS

Sending a Job to Hadoop

- The hadoop executable is used to send a job to the cluster
  - *install_dir/bin/hadoop jar program.jar org.myorg.ProgramName /input /output*
- One a job is sent to the cluster the input is broken up into many small pieces and sent to different nodes in the cluster to process

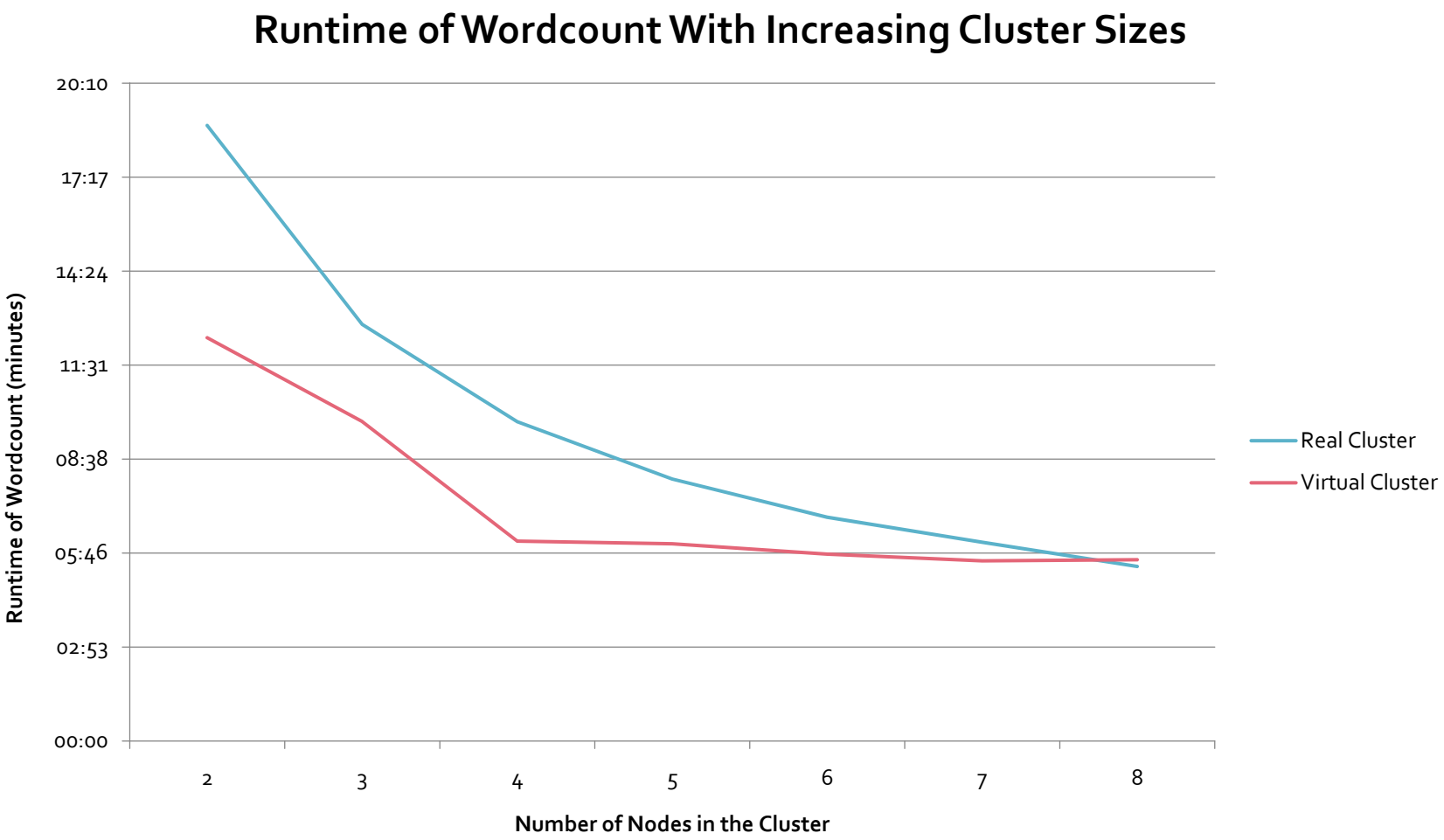
Cluster Architecture

- The real and virtual cluster initially consisted of 8 nodes (the size was automatically decreased during testing)
- One node was specified as the master
  - Ran the NameNode and SecondaryNameNode services along with a DataNode and TaskTracker service
- Each addition node served as a slave
  - Ran the DataNode and TaskTracker services

Testing the Scalability

- Hadoop wordcount example program was used as a benchmark
  - Wordcount counts the occurrences of each word is a set of files
- 438 ebooks totaling 289MB were randomly download from Project Gutenberg to form the test data
- Bash script written to run the experiments
  - Copied books into the dfs
  - Ran the wordcount program 5 times
  - Reduced the cluster size by one node and then ran again (until only 2 nodes left in the cluster)

Results


Runtime of Wordcount With Increasing Cluster Sizes
Runtime of Wordcount (minutes)
20:10
17:17
14:24
11:31
08:38
05:46
02:53
00:00
2
3
4
5
6
7
8
Number of Nodes in the Cluster
Real Cluster
Virtual Cluster

Results (cont.)

- Increasing the cluster size decreased the time required to run the wordcount program
- The graph begins to show signs of diminishing return
- Surprisingly adding more virtual machines to the virtual environment improved runtime
- In the real cluster there was nearly a 75% decrease in the runtime of wordcount when going from 2 to 8 nodes in the cluster

Future Research

- Experiment more with the customizable parameters of Hadoop
- Conduct a more comprehensive study into the benefits of using Hadoop
- Explore the MapReduce process in much more depth and write programs to utilize it
  - Use MapReduce to try to write solutions for NP-complete problems
- Explore the possibility of creating a permanent Hadoop Cluster in Alden Hall