

Automated Malware Analysis using MapReduce and Virtualization

Alexis Galarza
 Universidad del Turabo
 Research Alliance in Math and Science
 Cyberspace Sciences and Information Intelligence Research Group
 Mentor: Dr. Stacy Prowell
<http://sites.google.com/site/galarzaatornl>

Abstract

A MapReduce architecture will be combined with virtualization to enable safe analysis of potentially malicious software. MapReduce will be used to distribute data within a cluster. Each node in the cluster will have a virtual machine that runs programs to analyze malware. The results of this analysis will facilitate malware analysis, classification, and detection.

MapReduce

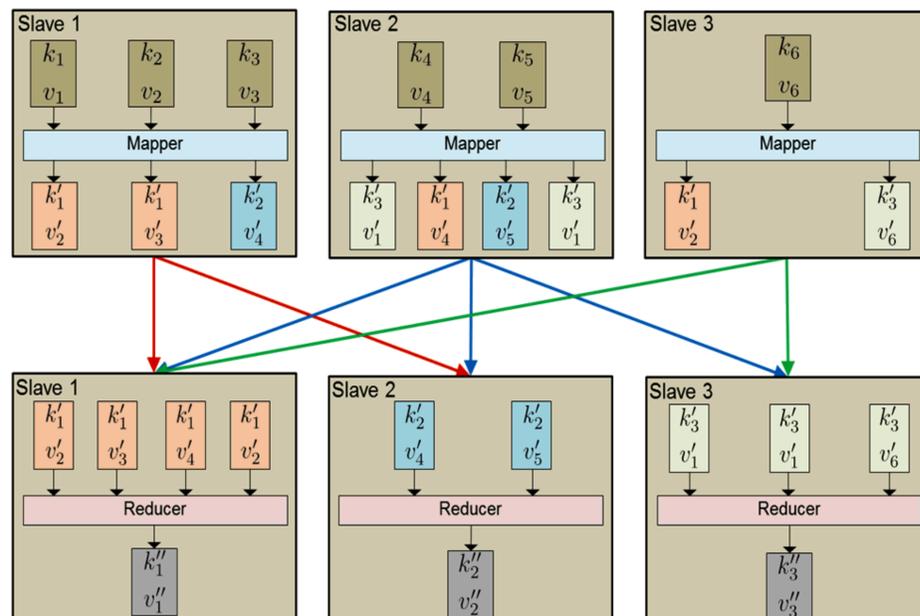
MapReduce is a framework for processing large data sets on a cluster. A master node distributes and organizes work among a set of slave nodes by splitting the data set and assigning the function the slave should use to transform the assigned data. Distributed work is done independently and simultaneously to increase overall performance. Each Map function takes a key-value pair and produces a set of key-value pairs.

$$\text{Map}(k, v) = \{ (k'_1, v'_1), (k'_2, v'_2), \dots \}$$

MapReduce organizes the output of each Map function: it partitions the output by key, and then each block of the partition is transformed into a single pair, the key unique to the block and the collection of values. The Reduce function transforms these new pairs into a final key-value pair.

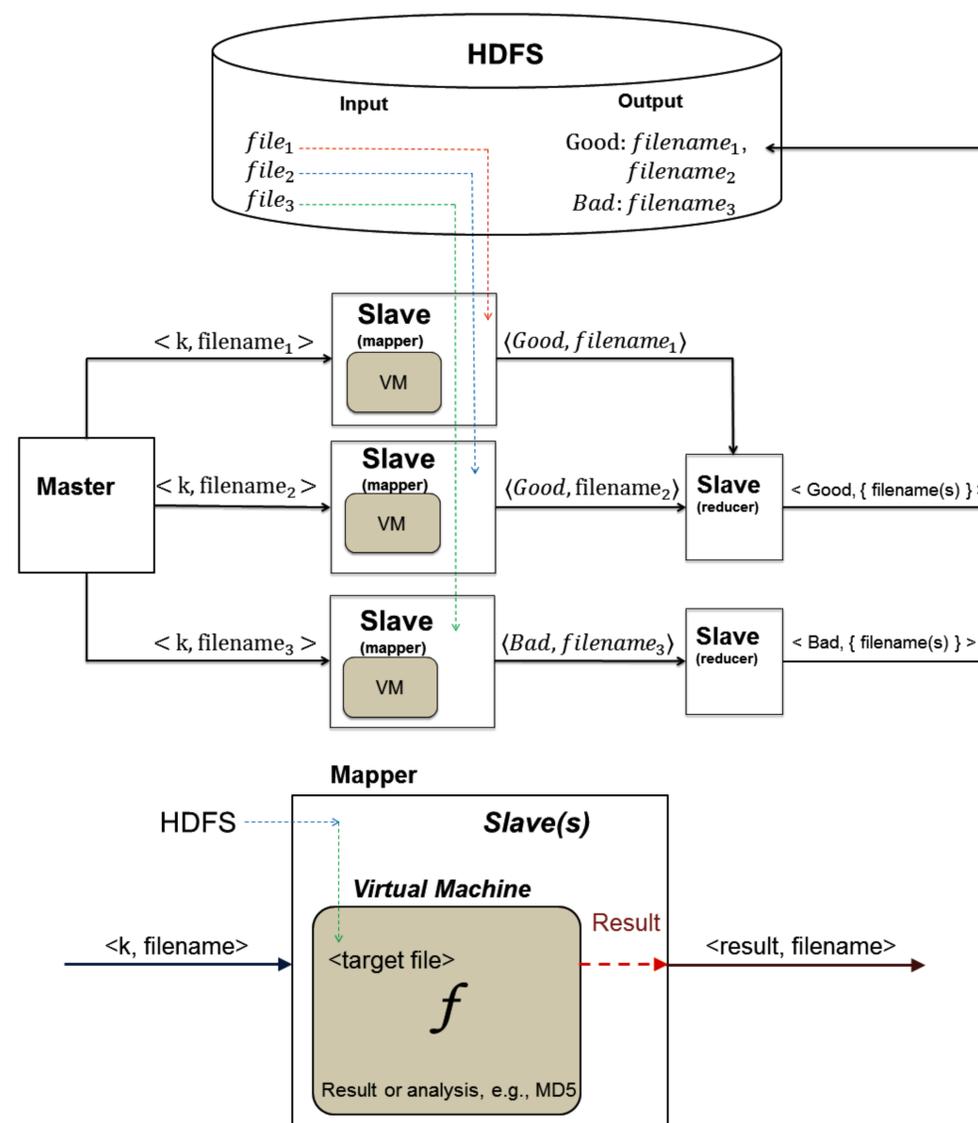
$$\text{Reduce}(k', \{ v'_1, v'_2, \dots \}) = (k'', v'')$$

MapReduce's parallel processing and distributed data architecture provide the advantage of failure resistance and efficiency. The implementation used is Apache's Hadoop.



Virtualization

Software can be executed within an environment called a virtual machine (VM) that simulates real hardware. The real machine's operating system (host) cooperates with the VM's operating system (guest). VMware Workstation, a commercial virtualization platform, was used to analyze the data distributed by Hadoop inside virtual machines. The VMware VIX API was used in the map function to control the VMs so operations like powering them on, and transferring data from and to the slaves could be done in the background without user interaction. The advantage of this approach is that the data will be analyzed in a safe environment. In the case that a VM becomes infected, the infection is contained inside the virtual computer.



Methodology

1. Setup a Hadoop cluster, with each computer having a Windows virtual machine
2. Put the files that are going to be analyzed in HDFS
3. Start the map:
 - a. Mapper receives filename as a value
 - b. System transfers file to the virtual machine with VIX API
 - c. Malware detection applications run when file is in the virtual machine
 - d. System passes result to the host
 - e. Mapper produces the result of the analysis as key and the name of the file as value
4. Start the reduce: writes the results of the analysis to HDFS
5. Display and analyze results

Results

By combining the VIX API with the Java Native Interface (JNI) and Hadoop, all the VMs in the cluster can be started in the background and files can be transferred back and forth between the slaves and the VMs inside them. A program that computes the MD5 checksum of the files stored in HDFS was used to test the architecture. The analysis was performed successfully and the results were stored in HDFS, indicating that the architecture was completed successfully.

Future Work

- Perform deep malware analysis by adding malware detection methods like file header inspection and virus scanners to the VMs
- Change MapReduce input method to a catalog of files instead of input directory
- Manipulate data with XOR before copying to local file system to add more security
- Return data to normal once inside the VMs

References

- Dean, J. & Ghemawat, S. (2004) *MapReduce: Simplified Data Processing on Large Clusters*. Retrieved July 26, 2011 from <http://labs.google.com/papers/mapreduce.html>.
- White, T. (2010). *Hadoop: The definitive guide* (2nd ed). California: O'Reilly Media.
- Yahoo! Hadoop Tutorial. Available at <http://developer.yahoo.com/hadoop/tutorial/>
- VIX API Reference Documentation. Available at http://www.vmware.com/support/developer/vix-api/vix110_reference/index2.html