The Quest for Conformance Testing in the Cloud

Dylan Yaga
Computer Security Division
Information Technology Laboratory
National Institute of Standards and Technology
NIST/ITL Computer Security Division Efforts

• NIST/ITL Computer Security Division supports the development of biometric conformance testing methodology standards and other conformity assessment efforts through active technical participation in the development of these standards and the development of associated conformance test architectures and test suites.

• These test tools are developed to promote adoption of these standards and to support users that require conformance to selected biometric standards, product developers and testing labs.

• These test tools are collectively titled BioCTS
BioCTS Background

• Traditional Desktop-Based Application Developed in Microsoft® C# – either with the Installer Graphical User Interface, or Command Line, used to test conformance to Biometric Data Interchange Records

• Two Conformance Test Architectures (CTAs)
  • Tests Implementations of select ISO/IEC 19794-X Generation 1 & 2 Data Formats, ANSI/INCITS, and several PIV Profiles of Standards

• Tests 1000s+ of files in a single Batch Test
• Allows Editing of Implementations Under Test
• Provides Charts, and Detailed Test Results in Text and XML Format
BioCTS, like most Desktop-Based applications, is only as powerful as the computer it is being run on. While pursuing methods of increasing performance, several questions were asked:

- What if BioCTS could run on multiple computers at once?
- What if BioCTS could run on multiple computers at once, and that fact was transparent to the end-user?
- What if BioCTS could run on multiple computers at once, and that fact was transparent to the end-user, and the number of computers would scale up or down depending on the work load?

Finally, the question was asked:

- What if BioCTS ran in the cloud?
Benefits of Cloud Software

• Scalable, Powerful Computing
  • Many computers working together
  • The amount of computing power can increase, or decrease depending on the workload

• Redundancy and Work Load Balancing
  • Often Cloud Frameworks automate redundant distribution of data, in case one of the many computers fails
  • Often Cloud Frameworks ensure there are as little idle computers as possible

• On Demand Computing
  • Use it when needed, stop when work is done
Interests in Testing in the Cloud

• Several organizations have shown interest in Testing in the Cloud
  • BCC 2012 – Bojan Cukic, CITeR: Services and Computational Cloud: The Inevitable Future of Biometrics
  • BCC 2013 - Ravi Panchumarthy, Ravi Subramanian, and Sudeep Sarkar, University of South Florida: Biometric Evaluation on the Cloud: A Case Study with HumanID Gait Challenge
  • BCC 2013 – Prof. Young-Bin Kwon and Dr. Jason Kim - Chung-Ang Univ. and Korea Internet & Security Agency (KISA) - K-NBTC BIOMETRIC TEST SERVICES AND CERTIFICATION
Apache™ Hadoop® Background

- Open Source Framework for creating scalable distributed computing system (also known as a cluster)
- Runs on commodity computers (not specialized hardware) known as Nodes
- Linux & Java Based
- Processes Data using the MapReduce Programming Model
  - MapReduce originally referred to the proprietary Google technology, now is a general term for the technology
  - “Maps” portions of data out to Nodes, and processes it
  - “Reduces” the portions of processed data results into an aggregation
Setting the Stage for the Quest

- To perform this research with as little ramp up time as possible, a preconfigured Virtual Machine was used
  - This Virtual Machine was pre-setup with Apache™ Hadoop®, Java and several other helpful utilities
  - A Virtual Machine, available for download from Cloudera, Inc., was used
  - The Cloudera, Inc. “QuickStart” Virtual Machine image used was version 4.X (4.4 was initially used, but 4.7 has also been tested)
  - The Virtual Machine “QuickStart” images can be found: [http://www.cloudera.com/content/support/en/downloads/quickstart_vms/cdh-4-7-x.html](http://www.cloudera.com/content/support/en/downloads/quickstart_vms/cdh-4-7-x.html)
Benefits of using a Virtual Machine (VM)

• VMs are portable – The same VM can be used on several different computers.

• VMs can have their state saved – and restored. A VM can have the entire state saved to be restored if things become unstable later on; less time repeating work

• VMs are how a lot of cloud services work (e.g., Amazon Elastic Compute Cloud (E2C) works off VMs)
Problems Encountered on the Quest

• Platform Incompatibilities
  • Linux vs. Microsoft® C#
  • Microsoft® within a Java MapReduce Job

• How Apache™ Hadoop® Processes Data vs. How BioCTS Processes Data
  • Large Files, many splits vs. Small Discrete Files, no splitting
  • Linefeed Characters within Files
Problem 1: Platform Incompatibilities

- How can the Linux and Java-based Apache™ Hadoop® Framework work with the Microsoft® C# BioCTS Conformance Test Suites?

- Possible Solutions:
  - Rewrite the Software in Java
    - No: Long process, could introduce bugs, may not provide 100% exact compatibility with existing Conformance Test Suite
  - Change Platforms
    - Not yet: Initial implementation targets existing systems, which were already running Linux & Apache™ Hadoop®
    - Possible Future Work: Expanding to other Platforms
  - Investigate a method for running Microsoft® C# Under Linux
    - Quickest method, if possible
Problem 1: Platform Incompatibilities - Solved

• Step 1: Find a way for Microsoft® C# to Run under Linux
  • The open source implementation of C#, known as Mono, is cross-platform, and runs well under Linux
  • The BioCTS software compiles perfectly with Mono and can then run in Linux

• Step 2: Find a way for the Java-Based Apache™ Hadoop® MapReduce to run C#
  • A method called “Hadoop Streaming” enables Apache™ Hadoop® to use any program/programming language that Linux supports to be embedded within a MapReduce Job
    • This is where the information was sparse, the API is very informative, but literature and publications simply mention the capability in passing
    • During research, was unable to locate a real world example of doing this exact procedure
Hadoop Streaming

• Hadoop Streaming is a JAR (Java Archive) file that is distributed with Apache™ Hadoop®

• This JAR file allows a developer to create MapReduce jobs with any script or executable file, allowing the specification of
  • Program for the Map function
  • Program for the Reduce Function
  • Input File Type
  • Additional Files needed for distribution

• Allows Apache™ Hadoop® to use any program that can run on Linux
Next Step...

- Now that BioCTS could run on Linux within a Java-based MapReduce Job, the focus shifted on processing data.
- This is where the second problem occurred.
- Apache™ Hadoop® and BioCTS process data in two fundamentally different ways.
- At least, initially...
Problem 2: Small File Problem

- Apache™ Hadoop® processes large amounts of data by splitting it up, and often times from a single file.
- BioCTS processes data by testing a large number of discrete files, as a whole – with many inter-relationships needed to be tested within a file – therefore it CANNOT be split.
- A simple solution would be to concatenate the files – but Apache™ Hadoop® could still split them arbitrarily and not on file boundaries.
- So, how to process many files at once, but have Apache™ Hadoop® think it is one giant file?
Problem 2: Small File Problem – Solved

- Sequence Files – a file that is comprised of many key-value pairs
- The Key is the file path
- The value is the Data for that file
- Apache™ Hadoop® accepts Sequence Files as input, and knows to split the file based on key-values
- Each Key, Value pair is recorded on its own line
- So the data does not get arbitrarily split!
Sub-Problem 2: Converting Files with Internal New Line Characters

- When converting the individual files into a Sequence File a problem was found:
  - When the files-to-convert contained internal new line characters, the process would not work
  - The Line Feed characters were placed in the Sequence File
- The solution? Eliminate the new line characters, but preserve the data – because the complete file is needed for testing
- The Implementation: A quick conversion of the files-to-convert to a base-64 encoded file, would preserve the data, but not contain new line characters
- The base-64 encoded data would then have to be decoded by the BioCTS software before processing
base64 benefit: Single Line of Data

Single ANSI/NIST-ITL 1-2011 Transaction, which is multiple lines (due to internal linefeed characters)

Same ANSI/NIST-ITL 1-2011 Transaction, encoded as base64 – a single line
BioCTS on Apache™ Hadoop® - Overview

• Six Phase Process
  • Some processing performed on the Client Computer before uploading files to Apache™ Hadoop®
  • Two MapReduce Jobs
  • Additional processing performed on the Client Computer after processing has been performed within Apache™ Hadoop®

• All of this has been automated with a script
BioCTS on Apache™ Hadoop® – Phase 1

Client Computer Converts Files to base64

Files to Test

base64
BioCTS on Apache™ Hadoop® – Phase 3

Sequence File Generator

MapReduce Job to create Sequence File

Sequence File Header
<Key, Value>
<Key, Value>
<Key, Value>

Apache™ Hadoop®
BioCTS on Apache™ Hadoop® – Phase 4

MapReduce Job to test individual Key, Value pairs within Sequence File

Apache™ Hadoop®
BioCTS on Apache™ Hadoop® – Phase 5
BioCTS on Apache™ Hadoop® – Phase 6
BioCTS on Apache™ Hadoop® – All Together

Files to Test

Sequence File Generator

MapReduce Job to create Sequence File

Sequence File Header
- <Key, Value>
- <Key, Value>
- <Key, Value>

MapReduce Job to test individual Key, Value pairs within Sequence File

Apache™ Hadoop®

Client Computer Splits output Files into Individual Log Files

output

single log files

Client Computer Converts Files to base64

base64
BioCTS on Apache™ Hadoop® Screen Shot
Success

• BioCTS on Apache™ Hadoop® has successfully performed conformance testing on sample test data (e.g., ANSI/NIST-ITL 1-2011)!
• Utilizing cloud computing technologies allows testing many files in parallel
• Have been able to utilize the existing Conformance Test Suites, leveraging the years of development, within an Apache™ Hadoop® environment
  • Successful mixture of technologies to create a solution
• The process could be used for other types of testing
  • Performance
  • Quality
  • Image analysis
  • Anything that runs on discrete files that can be done in parallel
• What was originally a research project resulted in a working implementation
Planned Next Steps...

• Expanding the work to a multi-node cluster
• Ensuring that this process works on the latest versions of the required software
• Develop a user interface, so users are not using command lines
• Expand to all available Conformance Test Suites
Questions?
Contact

• BioCTS Team
  • All Available BioCTS Software Downloads are available from:
    • http://www.nist.gov/itl/csd/biometrics/biocta_download.cfm
  • BioCTS Team Email:
    • biocts@nist.gov

• Dylan Yaga
  • dylan.yaga@nist.gov