Enabling Use of High Performance Computing for Secure Processing of Protected Health Information

Robert Bill[1], Genevieve B. Melton, MD, MA[1, 3], Bart Gottschalk[2], Andrew Ring[2], Justin L. Dale[4], Jorge Viñals, PhD[2], Anne-Francoise Lamblin, PhD[2], Serguei Pakhomov, PhD[1, 5]

[1]Institute for Health Informatics, [2]Minnesota Supercomputing Institute, [3]Department of Surgery, [4]Academic Health Center Information Services and [5]College of Pharmacy, University of Minnesota, Minneapolis, MN

Abstract

Analyzing large clinical note collections using natural language processing (NLP) requires computational resources in excess of that typically available internal to academic health centers. External high performance computing (HPC) centers have sufficient resources; however, strict HIPAA compliance requirements limit their use for processing Protected Health Information (PHI). We have developed a protocol for using HPC resources while minimizing security risks to PHI by using encrypted tunnels, ensuring that PHI data is never written to shared disk space, and by isolating PHI processing from other users. Formal independent evaluation confirmed the protocol poses minimal risk to PHI and is compliant with HIPAA. This protocol and approach can significantly expand the ability of medical centers to leverage computationally intensive processing securely and cost-effectively.

Introduction

Our goal is to establish a protocol for utilizing high performance computing (HPC) resources in conjunction with a HIPAA-compliant clinical data center to develop a cost effective environment for analyzing protected health information (PHI) that maintains HIPAA-compliant protection without de novo hardware investment. HPC resources available at the Minnesota Supercomputing Institute (MSI) have a shared data space, concurrent users and services that typically pose unacceptable risks to PHI. Automatic de-identification previous to processing notes on the MSI HPC cluster helps minimize risk to PHI; however, it does not completely eliminate the risk and poses a challenge to information extraction by removing and/or obfuscating data needed for accurate processing (1,2). We developed an alternative protocol for secure PHI data processing without de-identification that uses the Unstructured Information Management Architecture Asynchronous Scaleout (UIMA-AS) framework (3) with security constraints.

Methods

By using the UIMA-AS framework, PHI servers place clinical notes into a secure messaging server where approved clients on remote, single-user MSI HPC compute nodes retrieve the clinical notes. Remote analysis engines return results to the same secured messaging service so that all interaction with long-term storage remains within the protected, HIPAA-compliant data center (Figure 1). Security constraints include the following: a) encryption of all data transfers (SSH v2); b) preventing processes on the HPC compute nodes from writing data to long-term storage (disk); c) immediate process termination and node restart upon concurrent user detection; d) restarting each compute node on task completion to remove any PHI from active memory. We ran experiments to test concurrent logins and detect PHI, and Secure Digital Solutions (SDS), a professional security assessment provider, performed a HIPAA and ISO risk assessment using 111 control points to produce technical, physical and administrative metrics. We also benchmarked distributed processing of 100,000 clinical notes beginning with 16 clients on 2 nodes of an HPC cluster, then doubling nodes each subsequent trial to identify scaling limits. The analysis applied a BioMedICUS (4) NLP pipeline consisting of the tokenizer (5), parser (6), acronym detector, and concept mapper (7).

Results

The SDS security assessment determined the software and security protocol pose minimal risk to PHI and are approved for the processing of clinical data containing PHI. The benchmarking results (Figure 2) show that the NLP pipeline processing scaled predictably up to 512 client agents running on HPC compute nodes for one secure PHI server node. The notes-per-hour processing rate declined above 512 client processes due to the UIMA-AS messaging server rapidly dropping connections once its capacity was exceeded, thus limiting further scaleout.

Conclusion

Experiments and the SDS external evaluation confirm that sufficient protection of PHI exists when using UIMA-AS in combination with the security constraints defined in this abstract. Computational capacity available in academic HPC centers can therefore be incorporated into analytics engines without requiring additional hardware investments or extensive security modifications making this a potentially appealing approach for other centers. This protocol may also generalize to types of PHI data other than clinical notes (e.g. lab values, administrative claims, genotype).

Figure 1- Model for secure remote processing of PHI using UIMA Asynchronous Scaleout

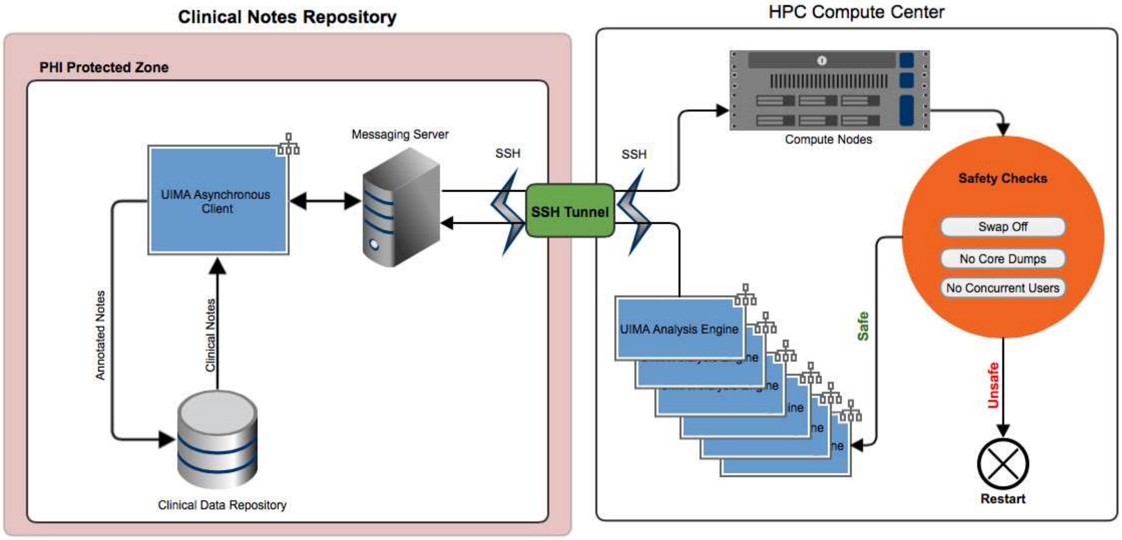
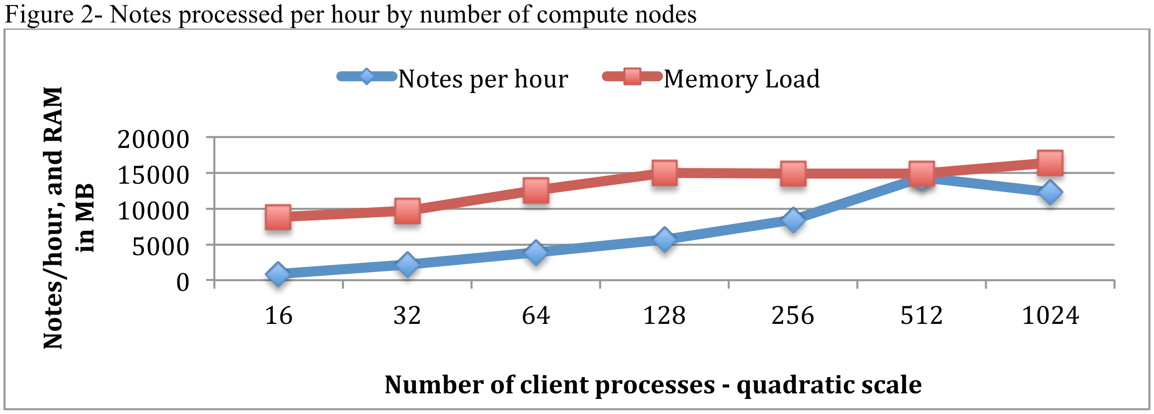


Figure 2- Notes processed per hour by number of compute nodes



Acknowledgments

This research was supported by Agency for Healthcare Research & Quality Grant (#R01HS022085) (GM) and National Institutes of Health (#R01LM011364 (GM), #R01GM102282 (SP) and #8UL1TR000114 (Blazer)).

References

1. Meystre SM, Ferrández Ó, Friedlin FJ, South BR, Shen S, Samore MH. Text de-identification for privacy protection: A study of its impact on clinical text information content. J Biomed Inform. Elsevier; 2014.
2. Meystre SM, Ferrández Ó, South BR, Shen S, Samore MH. How Much Does Automatic Text De-Identification Impact Clinical Problems, Tests, and Treatments? AMIA Summits Transl Sci Proc. American Medical Informatics Association;2013:177.
3. Epstein EA, Schor MI, Iyer BS, Lally A, Brown EW, Cwiklik J. Making Watson fast. IBM J Res Dev. IBM; 2012;56(3.4):11–5.
4. BioMedICUS - BioMedical Information Collection and Understanding System [Internet]. UMN NLP-IE Program [cited 2014 Sept 25]. Available from https://bitbucket.org/nlpie/biomedicus/
5. Apache OpenNLP [Internet]. Apache Software Foundation [cited 2014 Sept 25]. Available from: http://opennlp.apache.org/
6. Klein D, Manning CD. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. 2003. p. 423–30.
7. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium. 2001. p. 17.