A Cloud-based Strategy for Facilitating Adoption of Open-Source NLP in Applied Research Settings

David Carrell GHRI (carrell.d@ghc.org)

GHRI-IT Poster Session, Nov. 2010

GroupHealth

Introduction

Natural language processing (NLP) offers great potential to exploit information-rich clinical text widely available in electronic health records (EHRs), but the high total cost of operation, primarily in the form of salary costs for technical personnel, remains an impediment to widespread adoption. Cloud computing models have the potential to deliver NLP capacity to a wide variety of research settings securely and at relatively low cost. As depicted in the figure one or more **developer institutions** with broad expertise in open-source NLP deploy and maintain a **master copy** of the NLP system, including a web service that manages incoming traffic. **User sites** create private **clones** of the master system for exclusive use in a "virtual private cloud." A locally-deployed "I/O Manager," also provided by the developer institutions, encrypts and sends individual documents to the cloud and receives back annotated text and structured data. NLP processing occurs entirely in-memory, entirely avoiding security issues around cloud-based data storage. Updating and tweaking cloned NLP systems is done by re-cloning the enhanced/tweaked master system.

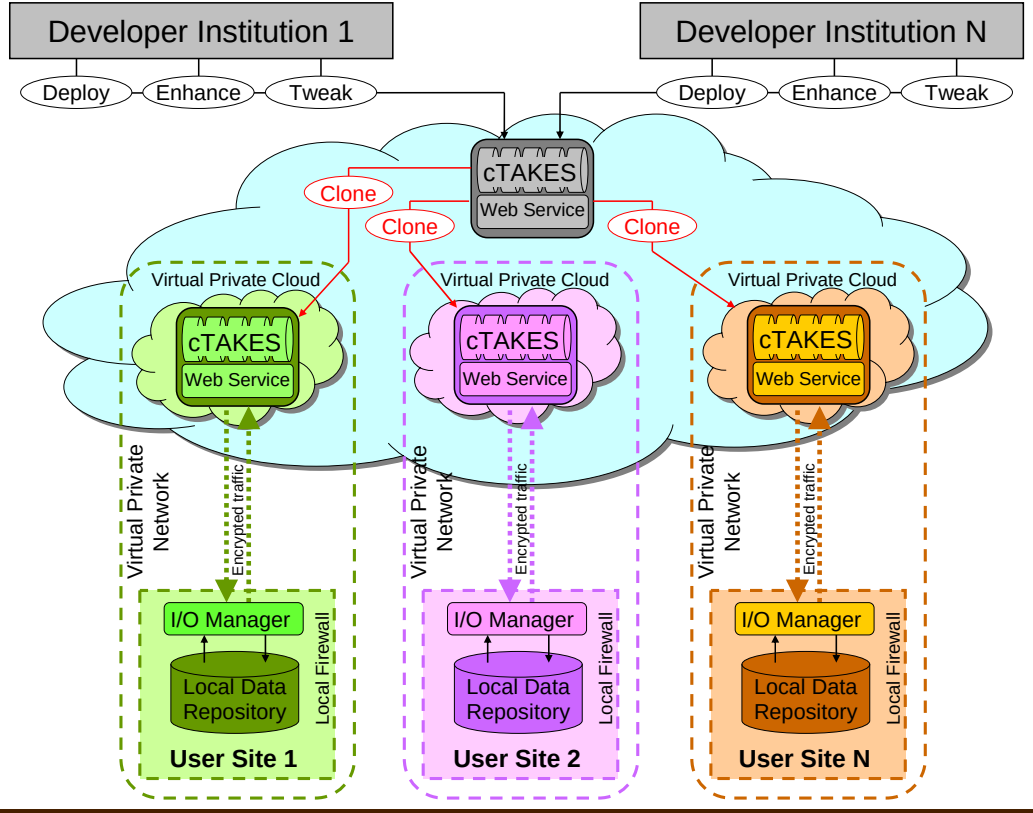
Developer Institutions

In this collaborative model one or more developer institutions assumes responsibility for:

- *Deploying* to the cloud the NLP system and a locally-deployed "I/O Manager" application that interfaces with the local text repository at the user site.
- *Updating/tweaking* the NLP pipeline as needed to address specific information extraction tasks, bugs.
- *Advising* the user sites on pre-processing, post-processing, and validation tasks performed locally at each user site.

Developer responsibilities may be coordinated and shared among multiple institutions.

Cloud Deployment Model



User Sites

User sites assume responsibility for:

- *Establishing comprehensive security* measures needed to satisfy local IRB requirements. This may require third-party cloud security consultation/assistance.
- *Implementing and securing* a cloud-based clone of the NLP system and an "I/O Manager" application used to manage document traffic.
- *Local data management,* pre-processing (including deidentification if desired) and post-processing.
- *Conducing validation analyses* to assure NLP algorithms are performing as intended or identify needed modifications.

The Security Challenge

Cloud computing solves several technical challenges and introduces a major new challenge: achieving security outside the local institutional firewall. Addressing this will require:

- *Local stakeholder participation* in security assessment, planning, implementation, auditing, etc.
- *Low-risk pilot opportunities* to prove the concept and build confidence in security measures (e.g., processing 100% de-identified text, initially).
- *Building a constituency* of local researchers who recognize the potential advantages of local NLP capacity.
- *Sacrificing NLP system performance* in the interest of reducing risk exposure (e.g., not persisting clinical text in the cloud).
- *Education* of stakeholders.