HTRC Data Capsule

Non-Consumptive Use of Texts

Jiaan Zeng[1]; Guangchen Ruan[1]; Alexander Crowell[2]; Atul Prakash[2]; Beth Plale[1]

[1]Indiana University Bloomington, [2]University of Michigan

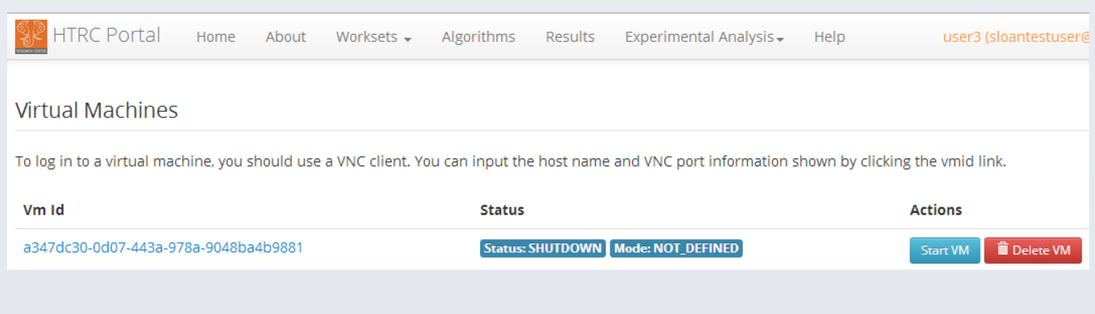
Background

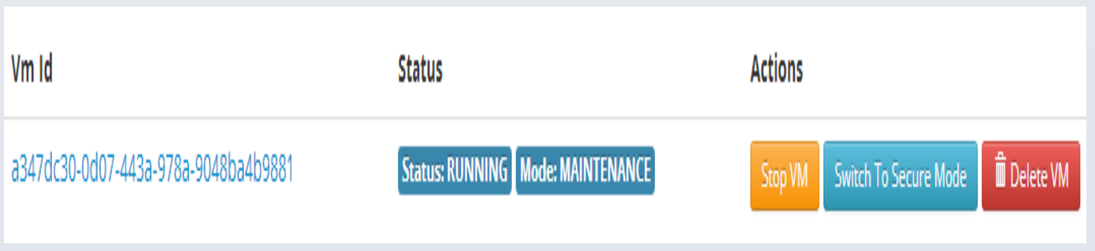
- The HathiTrust Research Center (HTRC) was recently established to provision for automated analytical techniques on the text data and images of the HathiTrust digital repository.
- HTRC is about **Big Data**!
  - 11,158,214 books; 3,905,374,900 pages; 500 terabytes
- Most of the data are **copyrighted**.
  - It suggests need for new forms of access that preserves intimate nature of interaction with texts *while* at same time *honoring restrictions on access.*

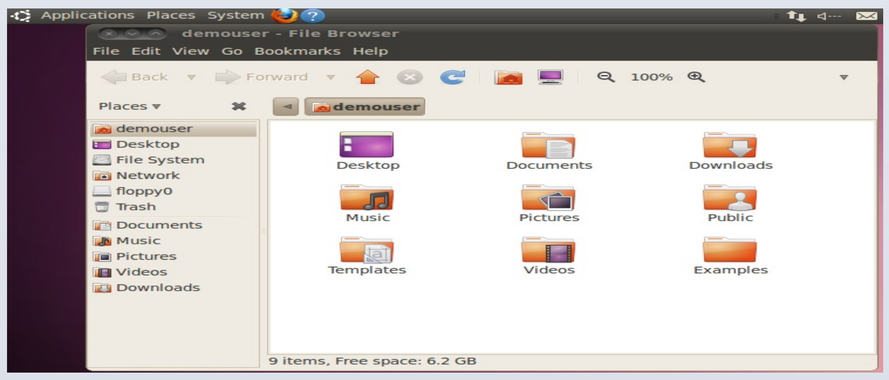
Research Questions

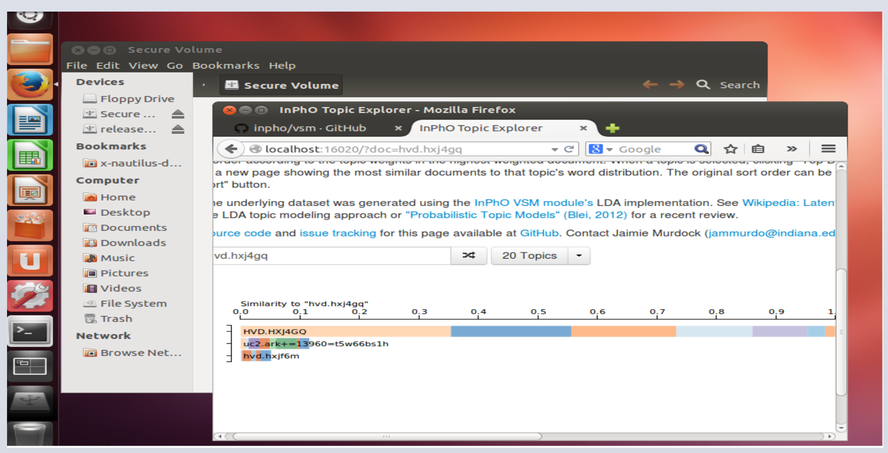
- **Non-consumptive use:** can framework provide safe handling of large amounts of protected data?
- **Openness:** can framework support user-contributed analysis without resorting to code walkthroughs prior to acceptance?
- **Large-scale** and **low cost:** can protections be extended to utilization of large-scale national (public) computational resources?

Snapshots

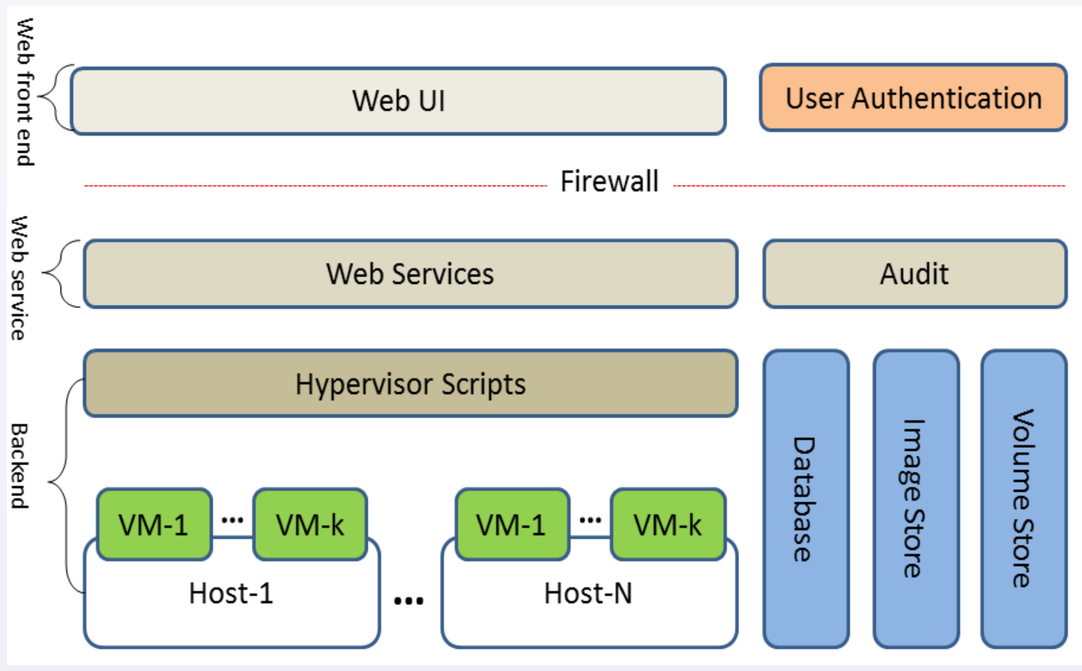




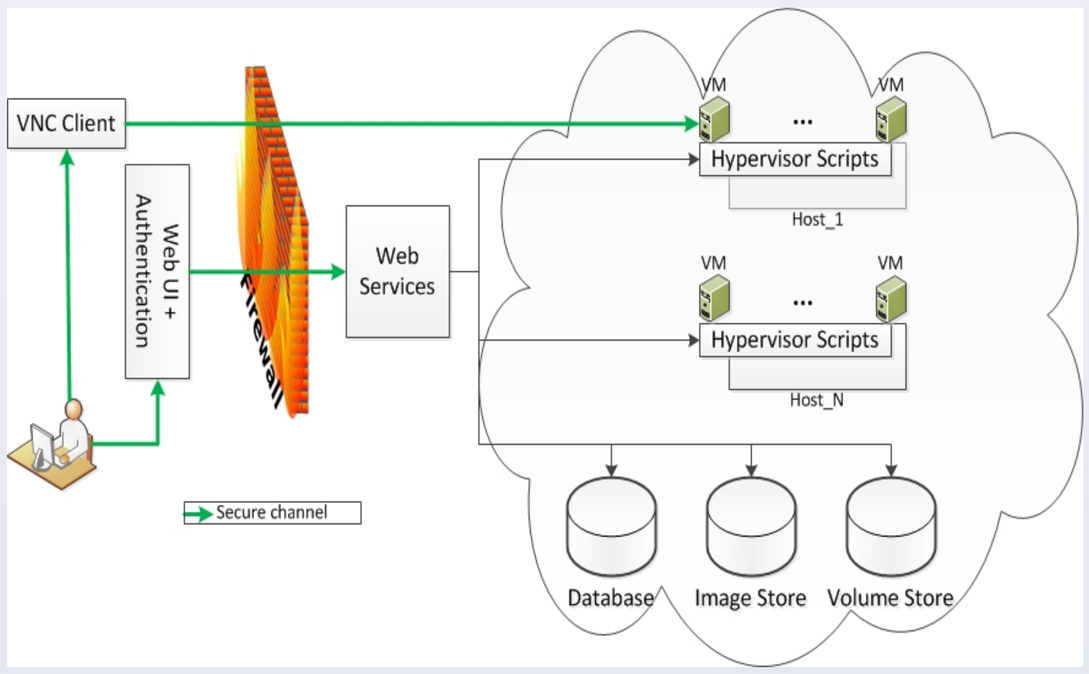




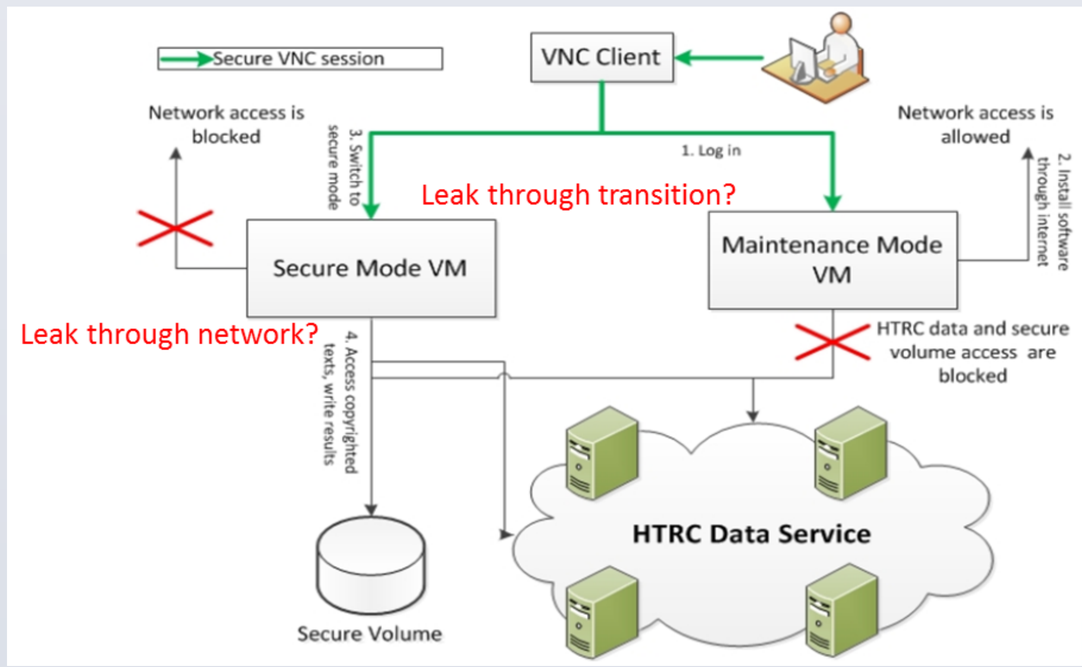
HTRC Data Capsules Architecture



HTRC Data Capsules Workflow



HTRC Data Capsules Access



Future Work

- A user may leak the data through the VNC channel or encode the final results. We plan to analyze the traffic on both channels.
- A user may leak through the covert channel between VMs. We plan to place the VMs on different hosts.
- We plan to enhance the data service for HTRC Data Capsules to accommodate multiple sources.