HathiTrust Research Notebooks

{ Milinda Pathirage and Beth Plale } Data to Insight Center, Indiana University

Motivation

HathiTrust Research Center (HTRC) provisions for computational analysis to the millions of digitized books in the HathiTrust digital library but analysis is limited to pre-defined algorithms executed via a Web UI.

The IPython Notebook (now called Jupyter) is a highly popular text analysis tool that would greatly increase the kinds of analysis a researcher can carry out. Hosting one IPython Notebook is one thing; hosting 1000 is another. We explore a parallel environment for lightweight virtualization that utilizes Docker. The effort is part of the HTRC advanced research group, and could be available as early as 1 year.

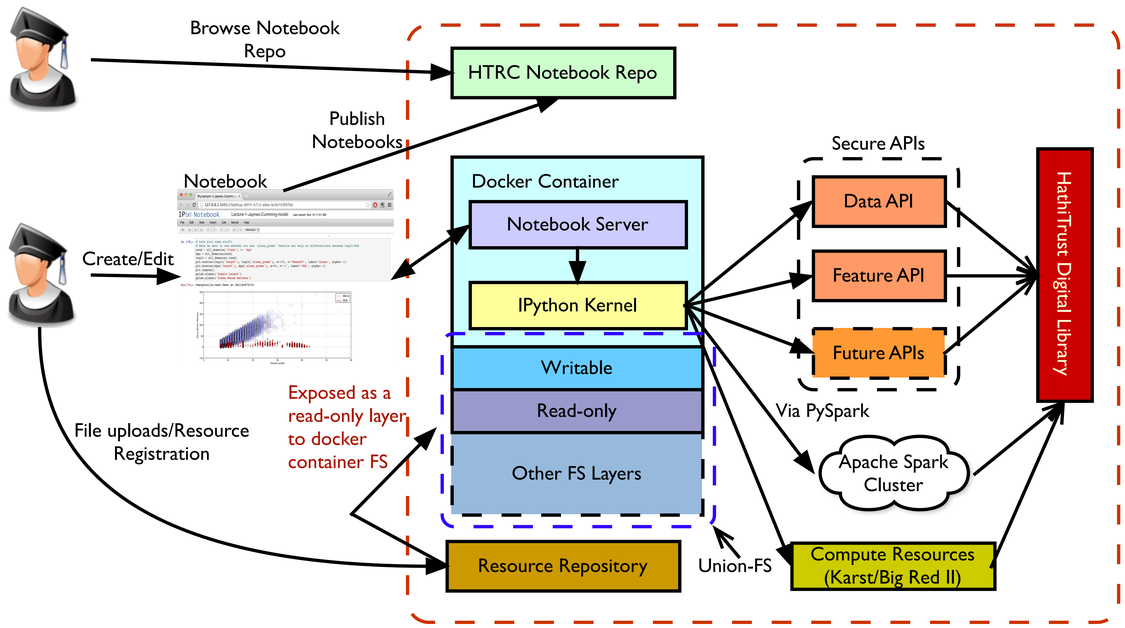
Research Questions

- How to provide a efficient and scalable isolated development environments for individual users
- How to expose HathiTrust corpus to interactive data analysis environment in a secure manner with data access level restrictions
- How to enable integration of user's data sets and existing code to the analysis workflow while preserving the security model defined by HTRC
- How to enable interactive access to parallel computing resources and data analytics frameworks such as Apache Spark

Research Notebooks

- Based on **IPython** [1] Notebooks
- Rich architecture for interactive computing and data visualization
- Supports code, text, mathematical expressions, inline plots and other rich media
- Supports multiple programming languages and frameworks such as **Python**, **R**, **Julia** and **Spark** [2]

Proposed Architecture



- Multi-user IPython environment based on Docker [3]. This provides container level isolation for user's development environment.
- Resource repository for uploading files or registering external resources that will be exposed to container environment as a read-only file system layer [4].
- Notebook repository hosted by HTRC for publishing research notebooks
- REST APIs for exposing data, features and text analysis tools to the notebook environment with necessary security restrictions
- Research notebook infrastructure runs inside a firewalled environment which implements the HTRC security model.

References

[1] Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007.

[2] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.

[3] Docker. `https://www.docker.com/`.

[4] Docker. File System Layers. `https://docs.docker.com/terms/layer/`.

[5] Dirk Merkel. Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, 2014(239), March 2014.

Docker

- Provides lightweight virtualization based on containers [5]
- Docker is fast and has minimal overhead/resource usage
- Can scale to thousands of containers
- Open platform with lots of tools for management, monitoring and networking

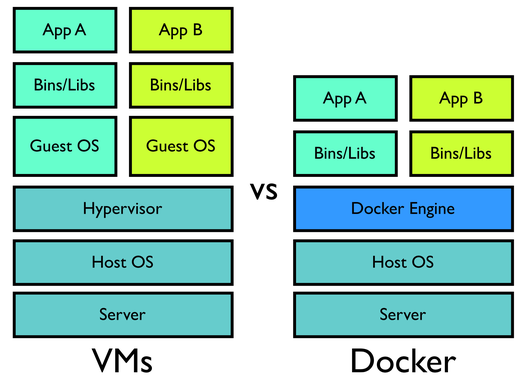


Figure 1: Virtual Machines vs Docker Containers

Roadmap

1. Multi-user research notebooks implementation and scalability experiment
2. Secure access to HTRC APIs (for data, features and tools) with necessary data access level restrictions
3. Resource repository implementation and integration of resource repository to research notebook runtime based on dockers union file system as a read-only layer
4. HTRC notebook repository for publishing research notebooks
5. Integration of HTRC hosted Apache Spark [2] cluster to research notebook runtime for large scale text analysis on HTRC corpus
6. Support for multiple programming languages in notebooks