

# HathiTrust Research Center Architecture Overview

---

**Robert H. McDonald | @mcdonald**

Executive Committee-HathiTrust Research Center (HTRC)

Deputy Director-Data to Insight Center

Associate Dean-University Libraries

**Indiana University**



# Follow Along



<http://slidesha.re/U4z1gW>

# HTRC Architecture Group

---

## Indiana University

- Beth Plale, Lead
- Yiming Sun
- Stacy Kowalczyk
- Aaron Todd
- Jiaan Zeng
- Guangchen Ruan
- Zong Peng
- Swati Nagde

## University of Illinois

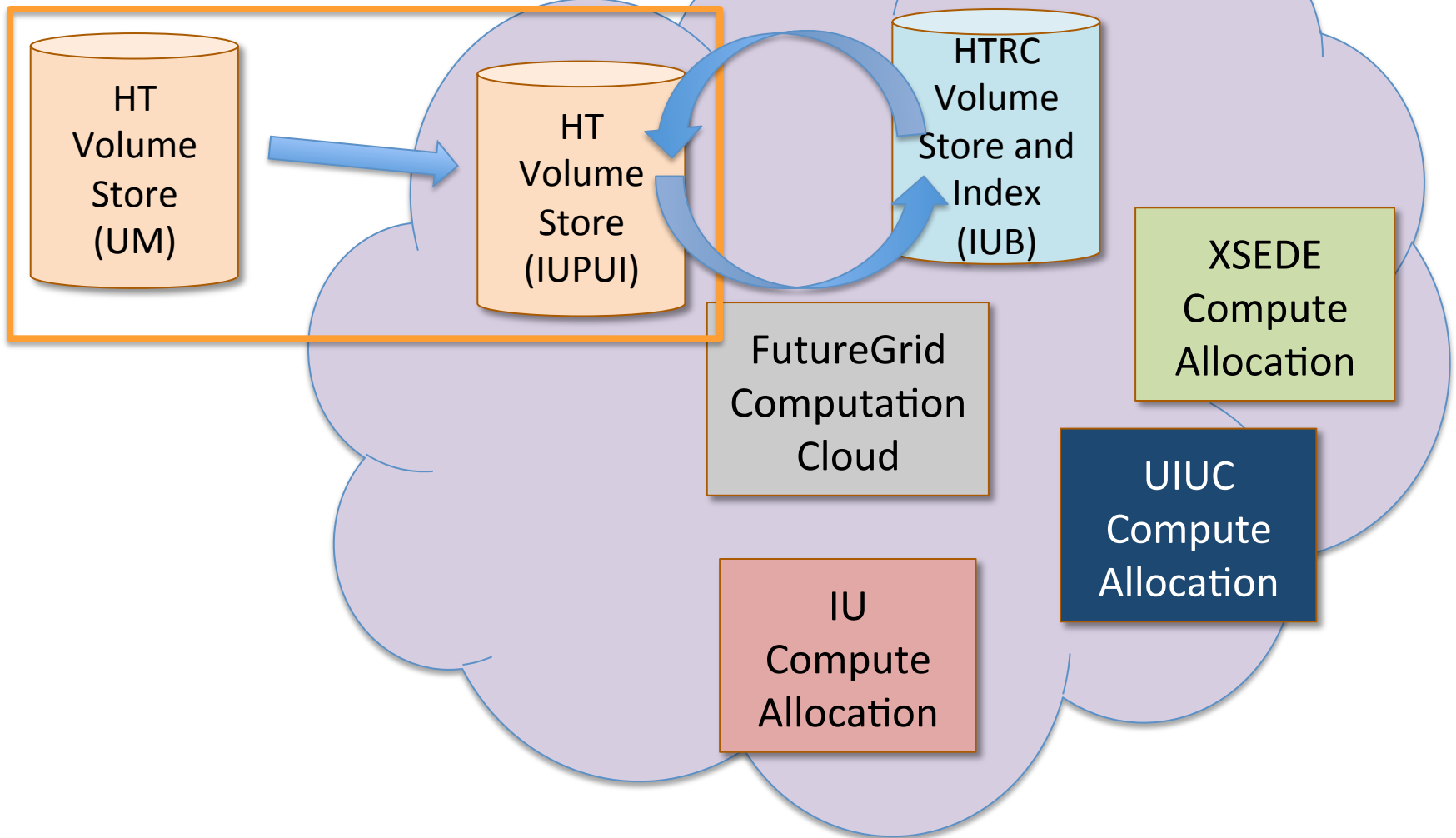
- J. Stephen Downie
- Loretta Auvil
- Boris Capitanu
- Kirk Hess
- Harriett Green

# Presentation Overview

---

- Considerations for Current Architecture
- Architecture - Use Case Methodology
- Technical Overview
- UnCamp Sessions for Further Review

# Main Case – Data Near Computation



# Non-Consumptive Research Paradigm

---

- *No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.*
- Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.

# Amicus Brief and NCR

---

- Jockers, Sag, Schultz –
- <http://tinyurl.com/cy34hhr>

# Use Cases for Phase 1 Architecture

---

- Use Case #1 - Previously registered user submitted algorithm retrieved and run with results set
- Use Case #2 - HTRC applications/portal access (SEASR)
- Use Case #3 – Blacklight Lucene/Solr faceted access
- Use Case #4 - Direct programmatic access through Secure Data API (right now only for UnCamp and open content)



# HTRC Current Infrastructure

---

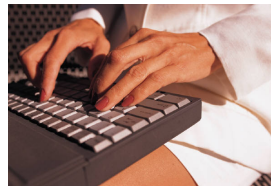
- Servers
  - 14 production-level quad-core servers
    - 16 – 32GB of memory
    - 250 – 500GB of local disk each
  - 6-node Cassandra cluster for volume store
  - Ingest service and secure Data API access point
- Storage (IU University Infrastructure)
  - 13TB of 15,000 RPM SAS disk storage
  - Increase up to 17TB by end of 2012
  - 500TB available in late year 2-year 3

# Key Components of Architecture

---

- Portal Access
- Blacklight Access
- Agent
- Registry
- Secured Data API Access
- Solr Proxy

# HTRC Architecture



## Portal Access

Blacklight

## Agent

Job Submission

Collection building

Direct programmatic access (by programs running on HTRC machines)

## Security (OAuth2)

Data API access interface

Solr Proxy

## Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

## Audit

Cassandra cluster

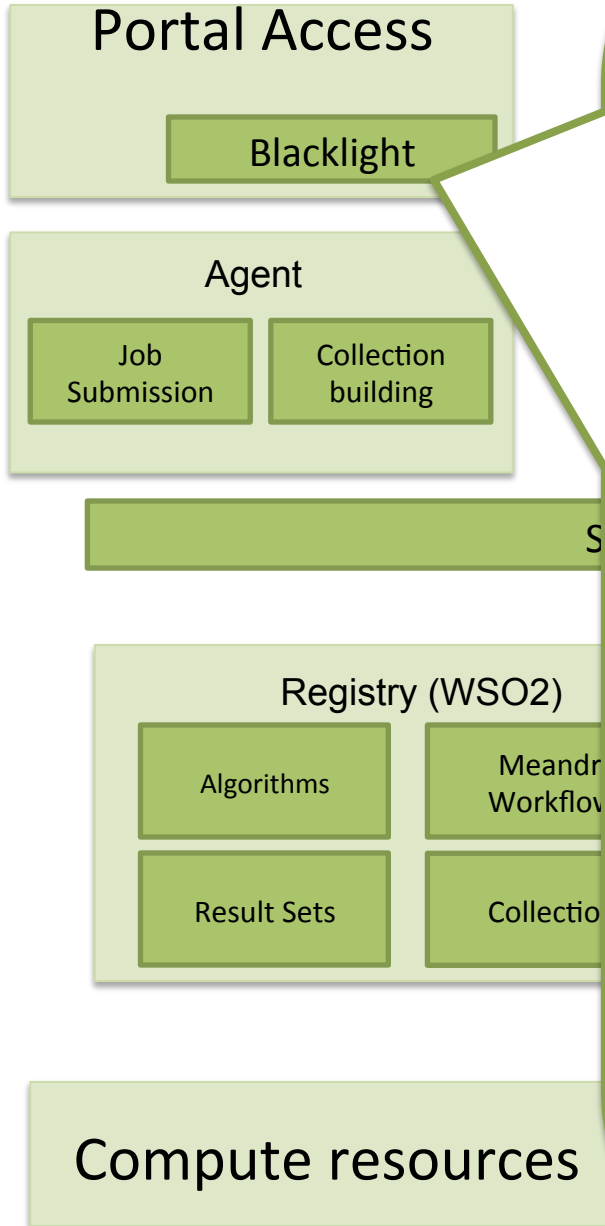
volume store

Solr index

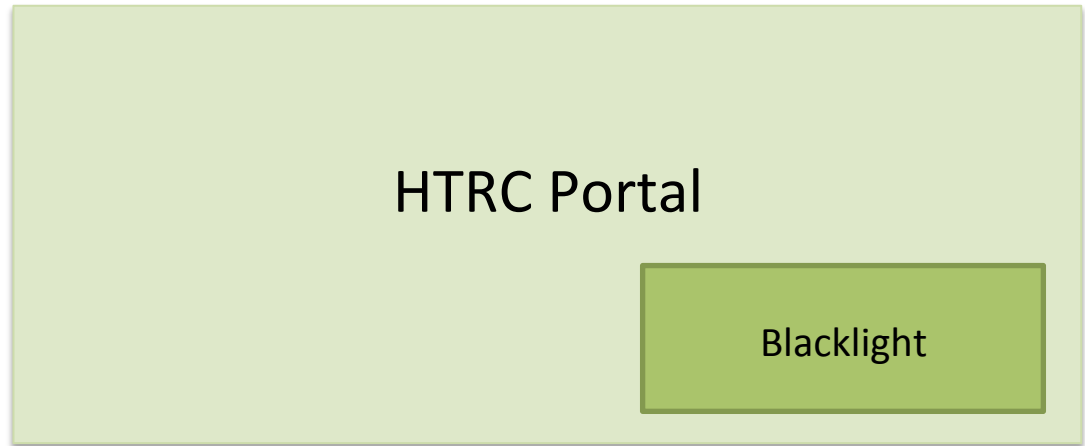
Compute resources

Storage resources

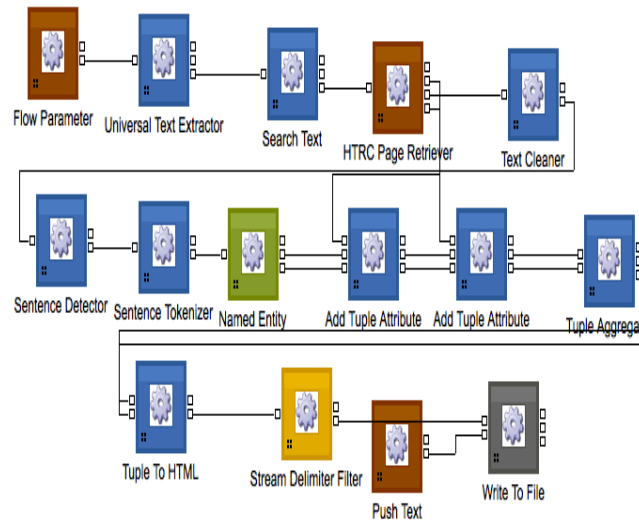
# HTRC Architecture



## Portal Access



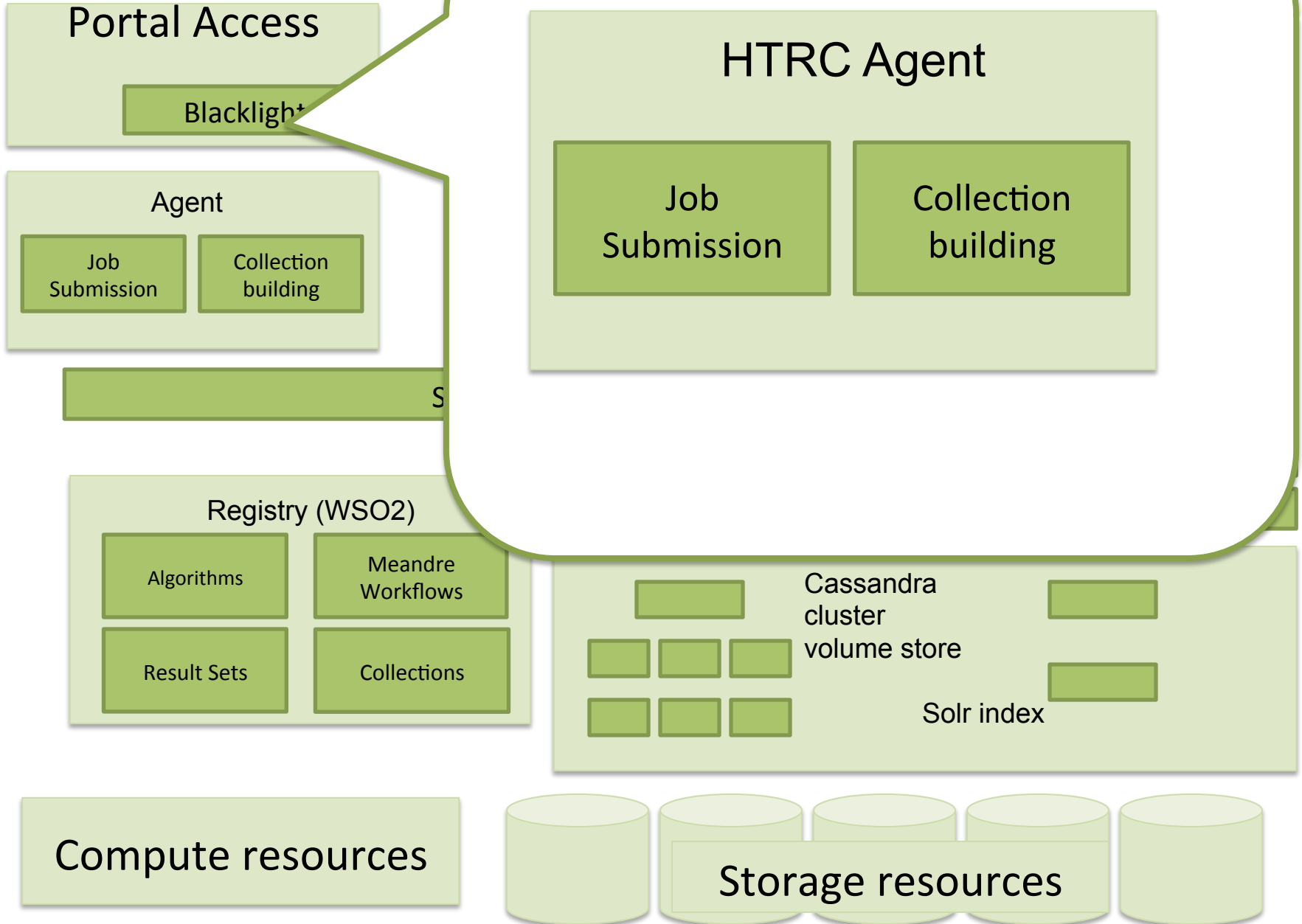
## App SEAR



## App Blacklight



# HTRC Architecture



# HTRC Architecture

**Portal Access**

Blacklight

**Agent**


Job Submission    Collection building

## HTRC Registry

**Registry (WSO2)**

Algorithms    Meandre Workflows

Result Sets    Collections



**Registry (WSO2)**

Algorithms    Meandre Workflows

Result Sets    Collections

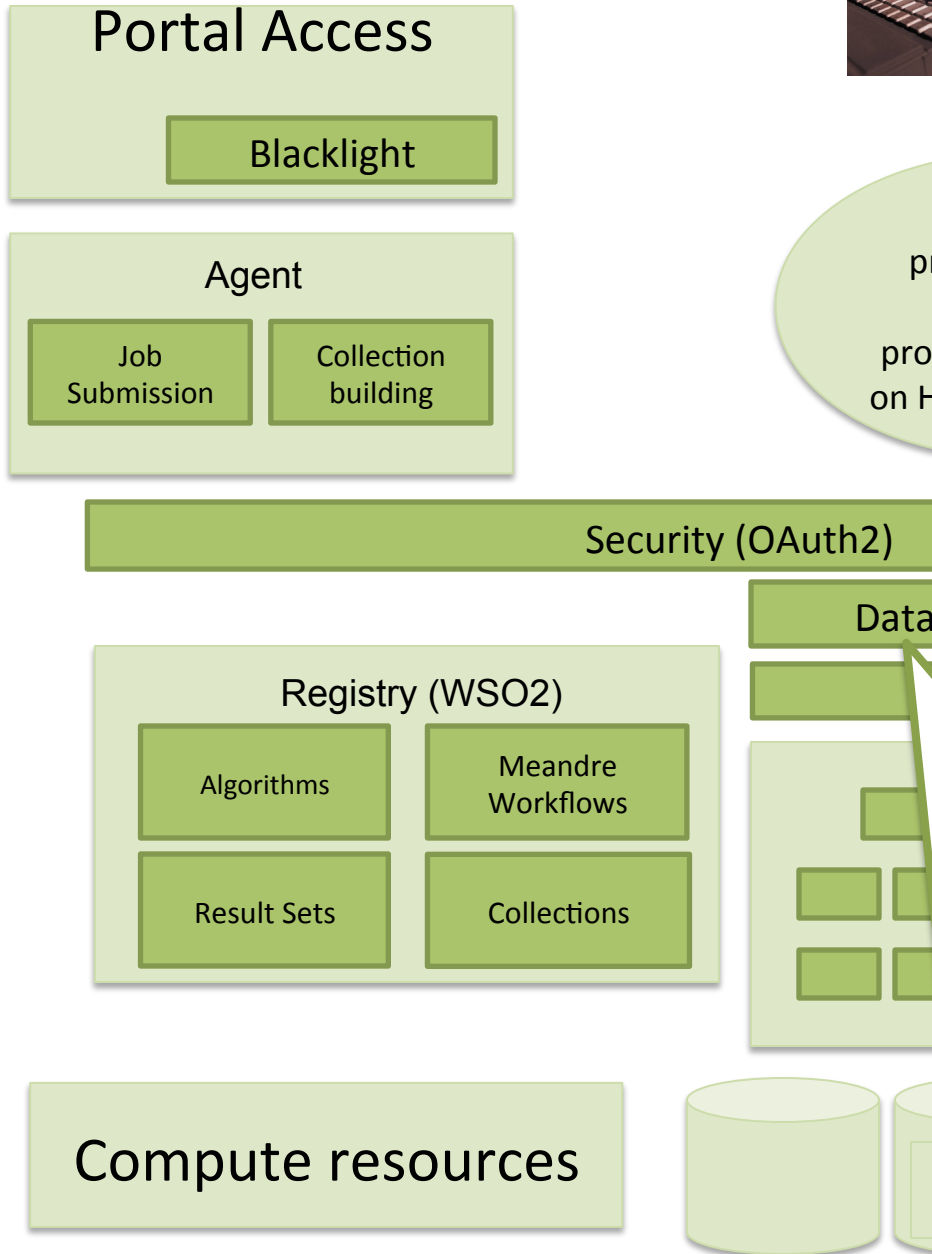
volume store

Solr index

**Compute resources**

**Storage resources**

# HTRC Architecture



## Secure Data API

- RESTful Web Service
  - Language agnostic
  - Clients don't have to deal with Cassandra
- Simple OAuth2 authentication
- HTTP over SSL
- Audits client access
- Protected behind firewall, accessible only to authorized IPs



H



Solr Proxy

Solr proxy

Solr service

RFS distributed file system

Direct programmatic access (by programs running on TRC machines)



API access interface

Solr Proxy



Storage resources





# NoSQL Methodology

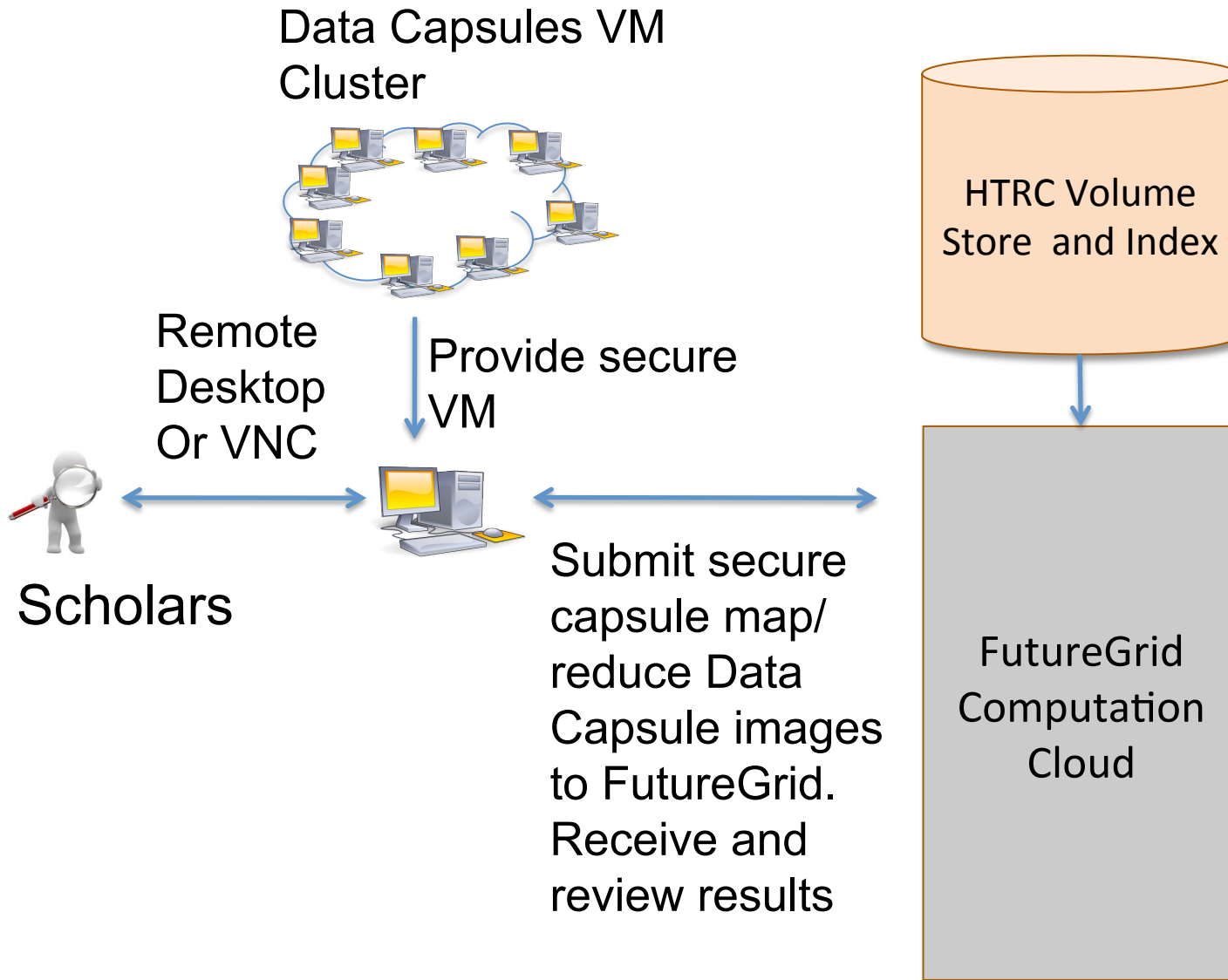
---

- Currently HT content is stored in a pair-tree file system convention (CDL)
- Moving these files into a NoSQL store like Cassandra enabled HTRC to aggregate them into larger sets of files for use in retrieval
- Use of Cassandra enabled HTRC to share content over a commodity based Cassandra cluster of virtual machines
- Originally investigated use of MongoDB, CouchDB, Hbase and Cassandra

# HTRC Solr Proxy + Solr Service

---

- Preserves all query syntax of original Solr
- Prevents user from modification
- Hides the host machine and port number HTRC Solr is actually running on
- Creates audit log of requests
- Provides filtered term vector for words starting with user-specified letter
- Filters out “dangerous” requests to Solr
- Adds additional features to Solr
  - E.g. Term Vectors



# Non-Consumptive Research-Secure Data Capsule

# Sessions for Further Review

---

- For more on Secure Data API – Tues Topic I/II (Yiming Sun)
- For more on Portal/SEASR – Tues Topic II (Loretta Auvil)
- For more on Portal/Blacklight – Tues Topic III (Stacy Kowalczyk)

# Contact Information

---

- Robert H. McDonald
  - Email – [robert@indiana.edu](mailto:robert@indiana.edu)
  - Chat – rhmcdonald on googletalk | skype
  - Twitter - @mcdonald
  - Blog – <http://www.rmcdonald.net>
  - Twitter Hashtag: #HTRC12