Brown Dog

Background

The DCIC is pursuing the development of cyberinfrastructure for the curation and management of digital assets at scale. "Brown Dog" is a CIC Big-10 $10.5M NSF/DIBBs-funded collaboration with U. of Illinois NCSA Supercomputing Center and industry partners (NetApp and Archive Analytics Solutions). DIBBs is the NSF Data Infrastructure Building Blocks program, launched in 2013. *Brown Dog* is the largest funded DIBBs project to date, and Maryland's role is to: (1) create a big records observatory that can scale to very large sizes, and (2) continuously test these emerging digital services on this big archive.

Towards this goal, the Maryland iSchool had partnered with a commercial company, Archive Analytics Solutions Ltd. which invested some $2M of software developments to date, to produce a highly scalable digital repository called *Indigo*. AAS has now offered a redistributable, open source license on the Indigo code to the DCIC at the U. Maryland and Indigo has to-date demonstrated promising scalability on the DCIC Big Data Archive of 100 million data objects and 72 terabytes of data.

Indigo is based on:

- **NoSQL Apache Cassandra** database. Originally created for Facebook, it is now an open source distributed database system for deployment of large numbers of nodes across multiple distributed data centers. It is in use with CERN, eBay, GitHUB, Hulu, Instagram, Netflix, Twitter, and at 1500 other companies, and scales to petabytes of storage and billions of objects.
- **CDMI (Cloud Data Management Interface) industry standard**, a RESTful HTTP industry cloud storage interface from the *SNIA*, the Storage Networking Industry Association. *SNIA* is made up of some 400-member companies and has developed this consortia-based alternative to the Amazon S3 API.

Objectives

Explore the massively scalable digital archives capabilities of Indigo and Brown Dog. Provide missing software infrastructure to the cultural heritage and archives communities in the form of a repository, model architecture, and model workflows.

- Horizontal Scaling - Scales up and down with predictable and efficient use of resources
- Interoperability - Flexibly makes use of external software and services
- Rich Metadata - Supports extensible metadata fields and datafication of materials
- Computational Infrastructure - Support for Map-Reduce and other compute jobs
- Data-driven Catalog - Hierarchical catalog with data visualization

Student Project Ideas:

1. Archival Data Visualization

   This project aims to create meaningful graphics and dashboards which facilitate preservation and data-driven research over archival collection. The student team will explore the data we have for the collections in DARRA using visualization tools built around an Elasticsearch index. Some of the visualizations developed by students will later be integrated into the DARRA web interface.

   The CI-BER archives, consisting of many different data sets, makes up our largest research collection at 100 million files and 72 Terabytes of data. CI-BER includes many collections from US government agencies, transferred from the National Archives and Records Administration (NARA). These include large scientific datasets, recent administrative records, and historical materials. The data and metadata is primarily stored in the DARRA catalog, which is an Apache Cassandra distributed database. However, the majority of the metadata is also indexed in Elasticsearch. It is this Elasticsearch index of all the CI-BER collections that we will use to create visualizations.

2. Quality Assurance Testing of Clowder Web UI
3. Develop an Extractor for CI-BER Data
4. Develop a Converter for CI-BER Data
5. GAP Analysis of the extractor and converter services

Themes

Cyberinfrastructure

Stakeholders

- DCIC Affiliated Researchers
- Digital Preservation Community
- Researchers in Computational Archival Science

Data

- CI-BER Testbed Collections

Final Product

For Project 1. [other project details will have to be explored in greater detail], all visualizations will be delivered with an accompanying blog post that describes the technique, any relevant queries, and an explanation of how the visualization supports research or preservation. Individual students or teams can build any visualization that seems meaningful, but are encouraged to explore these areas:

- Visualizations that show an overview of archival collections
- Visualizations that facilitate comparison between neighboring folders
- Visualizations of metadata statistics (keys/values commonly used)
- Queries and results for photos of interest:
  - Portraits of people
  - Large group photographs
- Finding large data sets, such as large spreadsheets, tab/comma separated files
- Finding large earth observation imaging (large satellite images)
- Finding materials related to a place or geographic area

Skills

This project calls for students with any of the following skill sets:

- Data visualization
- JavaScript and JSON
- Digital archives and digital preservation
- Search indexing
- Using RESTful APIs especially the Elasticsearch API

Software

- Elasticsearch
- Kibana
- Jupyter Notebook

Professional Development

Students who work on DARRA projects may obtain experience in these areas:

- Digital archives processing
- Datafication or collection enhancement
- Approaches to software integration
- Data visualization
- Distributed systems

Deadlines

- Mid semester objectives.
- Late November objectives.