Dataverse Scale

Project Logistics:

Mentors: Dan McPherson email: dmcphers@redhat.com; Phil Durbin email: philip_durbin@harvard.edu;
Min-max team size: 2-4
Expected project hours per week (per team member): 6-8
Will the project be open source: yes

Preferred Past Experience:

Docker Nice to have
Kubernetes/OpenShift Nice to have
Git Valuable
Java Nice to have

Project Overview:

Background:
Dataverse is a popular Open Source project that has built a very vibrant community of users. Surprisingly though, Dataverse hasn't really been built for scale or high availability. Its backlog of work primarily consists of adding more features to the project to make users happy and consists less of how to make the product more resilient and easier to host. This is a really great state for a product to be in because it means that users are happy and engaged. Even though the community isn't demanding it, there is a great need to make Dataverse easier to operationalize at scale. In particular, the MOC has an offering called Open Dataverse where it's using the project to front its 20 petabyte research oriented datalake. This project and other projects like it are only a few outages away from making the current situation an emergency.

Project Specifics:
Dataverse was recently containerized to run on top of OpenShift (https://github.com/IQSS/dataverse/pull/4168). The goal of this project is to continue that work and make each of the components of Dataverse function at scale on OpenShift. This includes glassfish, postgresql, and solr.

Some Technologies you will learn/use:

Containers/Docker/Kubernetes/OpenShift
Software Engineering (Agile, Scrum, Git)

Cloud Computing (OpenShift, OpenStack)
Cloud Scale (Web Frameworks, Databases, Search Engines)
CI/CD with Jenkins and OpenShift