

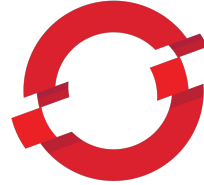
# Dataverse Scaling

Students: Michael Clifford, Patrick Dillon, Ryan Morano & Ashwin Pillai

Mentors: Phil Durbin (Harvard), Dan McPherson & Solly Ross (both Red Hat)



## Project Overview



**OPENSIFT**

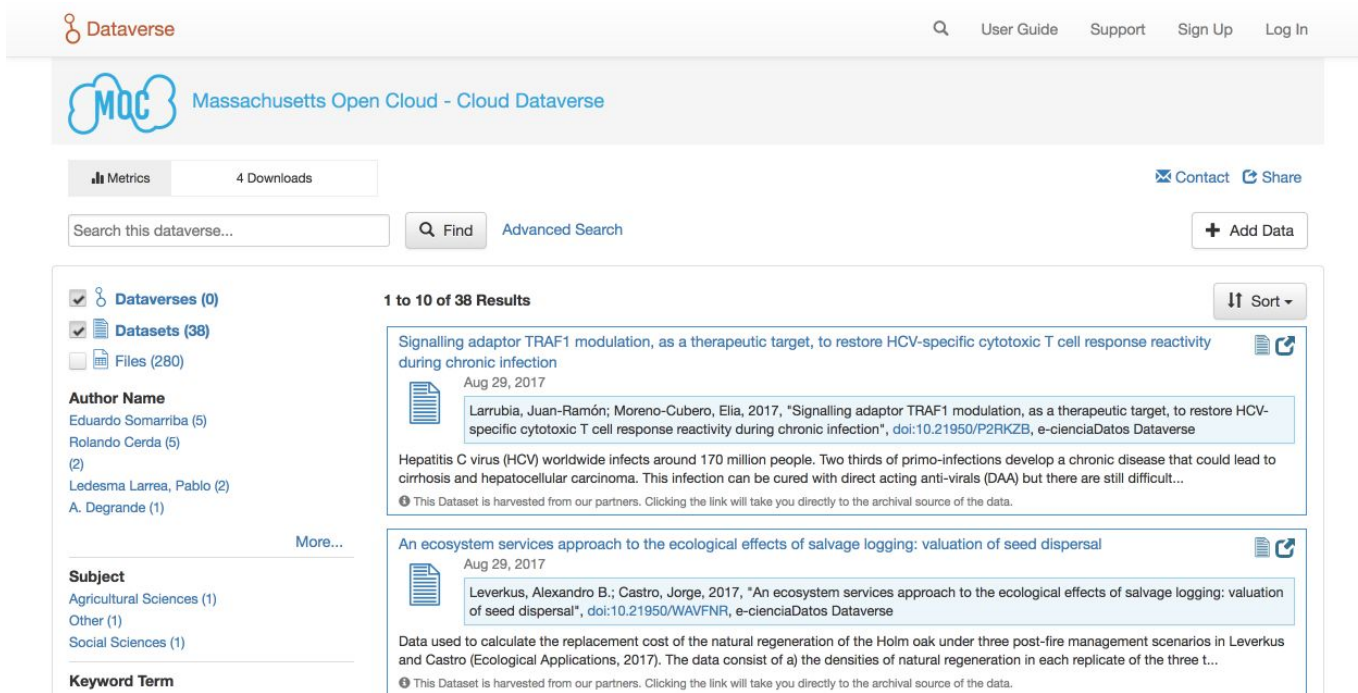
Problem: The open-source Dataverse project was originally built as a single deployment n-tier app

Our project is to continue recent efforts to containerize Dataverse and deploy on OpenShift.



- Open-source web app
- Platform to share research data & replicate work
- Community of researchers, academic institutions, journals, and software developers
- 30 institutional dataverse repositories worldwide
- MOC: [dataverse.massopen.cloud](https://dataverse.massopen.cloud)

# Screenshot of Cloud Dataverse



The screenshot displays the Cloud Dataverse web interface. At the top, the Dataverse logo is on the left, and navigation links for 'User Guide', 'Support', 'Sign Up', and 'Log In' are on the right. Below this, the 'Massachusetts Open Cloud - Cloud Dataverse' banner is visible. The main content area shows search results for '1 to 10 of 38 Results'. On the left sidebar, there are filters for 'Dataverses (0)', 'Datasets (38)', and 'Files (280)'. Under 'Author Name', several authors are listed with their respective counts: Eduardo Somarriba (5), Rolando Cerda (5), Ledesma Larrea, Pablo (2), and A. Degrande (1). Under 'Subject', 'Agricultural Sciences (1)', 'Other (1)', and 'Social Sciences (1)' are listed. Under 'Keyword Term', there are no results. The search results list two datasets. The first dataset is titled 'Signalling adaptor TRAF1 modulation, as a therapeutic target, to restore HCV-specific cytotoxic T cell response reactivity during chronic infection' by Larrubia, Juan-Ramón; Moreno-Cubero, Elia, 2017. The second dataset is titled 'An ecosystem services approach to the ecological effects of salvage logging: valuation of seed dispersal' by Leverkus, Alexandro B.; Castro, Jorge, 2017. Both datasets include a brief description and a link to the archival source.

**Dataverse**

Search this dataverse... **Find** [Advanced Search](#) **+ Add Data**

**1 to 10 of 38 Results** **Sort**

**Dataverses (0)**

**Datasets (38)**

**Files (280)**

**Author Name**

- Eduardo Somarriba (5)
- Rolando Cerda (5)
- Ledesma Larrea, Pablo (2)
- A. Degrande (1)

**Subject**

- Agricultural Sciences (1)
- Other (1)
- Social Sciences (1)

**Keyword Term**

**Signalling adaptor TRAF1 modulation, as a therapeutic target, to restore HCV-specific cytotoxic T cell response reactivity during chronic infection**

Aug 29, 2017

Larrubia, Juan-Ramón; Moreno-Cubero, Elia, 2017, "Signalling adaptor TRAF1 modulation, as a therapeutic target, to restore HCV-specific cytotoxic T cell response reactivity during chronic infection", doi:10.21950/P2RKZB, e-cienciaDatos Dataverse

Hepatitis C virus (HCV) worldwide infects around 170 million people. Two thirds of primo-infections develop a chronic disease that could lead to cirrhosis and hepatocellular carcinoma. This infection can be cured with direct acting anti-virals (DAA) but there are still difficult...

This Dataset is harvested from our partners. Clicking the link will take you directly to the archival source of the data.

**An ecosystem services approach to the ecological effects of salvage logging: valuation of seed dispersal**

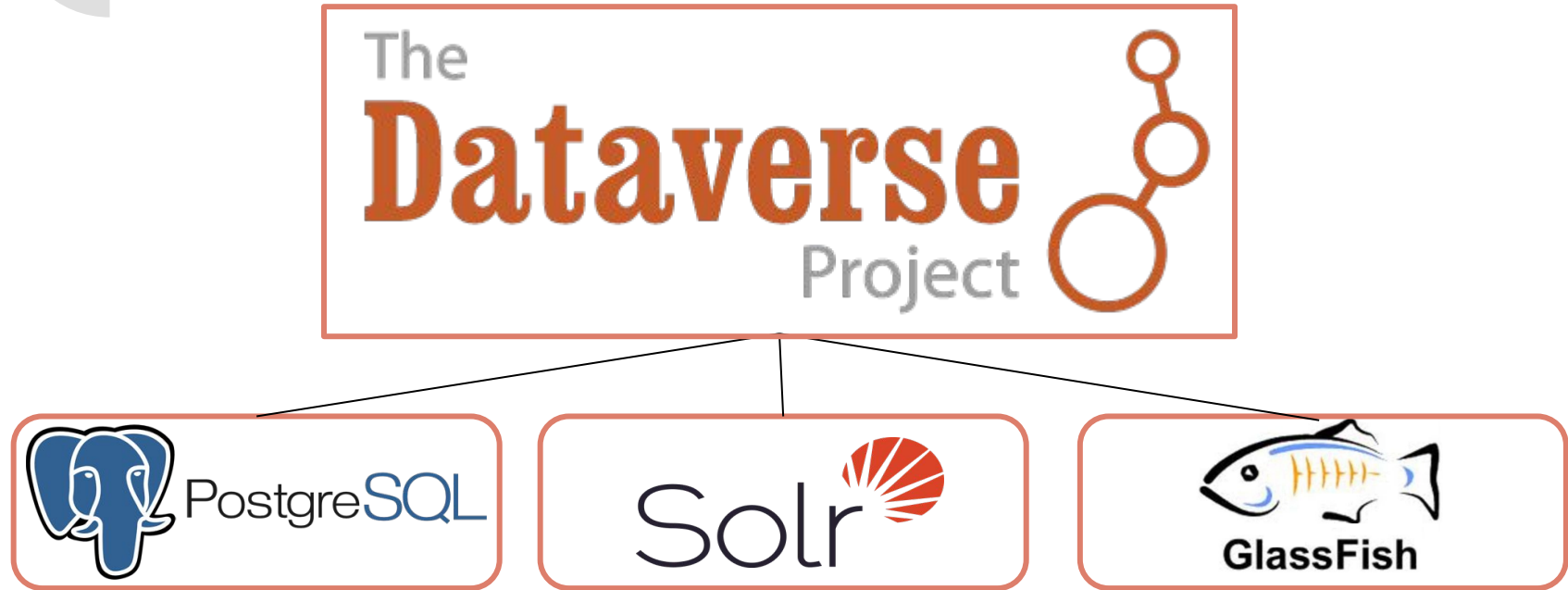
Aug 29, 2017

Leverkus, Alexandro B.; Castro, Jorge, 2017, "An ecosystem services approach to the ecological effects of salvage logging: valuation of seed dispersal", doi:10.21950/WAVFNR, e-cienciaDatos Dataverse

Data used to calculate the replacement cost of the natural regeneration of the Holm oak under three post-fire management scenarios in Leverkus and Castro (Ecological Applications, 2017). The data consist of a) the densities of natural regeneration in each replicate of the three t...

This Dataset is harvested from our partners. Clicking the link will take you directly to the archival source of the data.

## The main components of Dataverse



## Primary Components of Dataverse



PostgreSQL

- Application database



GlassFish

- Application server

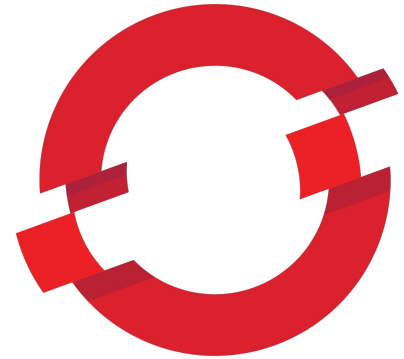


- Indexing / search engine



# OPENSIFT

- Red Hat's container orchestration software built on top of Kubernetes.
- Each component has a Docker image and has been configured to run on OpenShift



# OPENSIFT



## Single deployment vs containerization/OpenShift

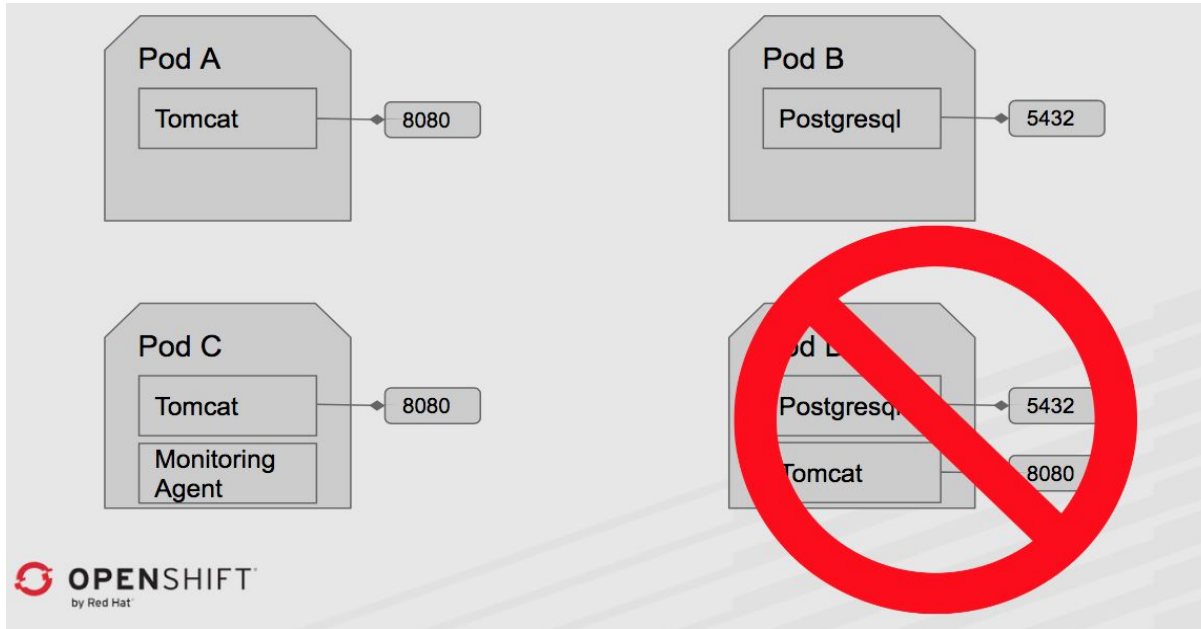


The original project consists of a single instance of each

A pull request has been accepted to containerize each component and run on Openshift, but the pods do not scale



# Containers, pods, scalability



Source: OpenShift Tutorial slides shared by mentor

# OpenShift Dashboard - Scaling

The screenshot displays the OpenShift Dashboard interface. On the left is a vertical sidebar with navigation links: Overview, Applications, Builds, Resources, Storage, and Monitoring. The main content area is titled 'dataverse' and shows a deployment named 'dataverse-glassfish, #1'. The deployment details include the container name 'DATAVERSE-PLUS-GLASSFISH', the image 'iqss/dataverse-glassfish-f814b85 705.3 MiB', and ports '8080/TCP'. A large blue circle with the number '1' and the word 'pod' indicates the current pod count. Below this, the 'Networking' section shows the service 'dataverse-glassfish-service' with port '8080/TCP (web) → 8080'. It also lists two routes: 'http://dataverse-project1.192.168.99.100.nip.io' and 'http://dataverse-glassfish-service-project1.192.168.99.100.nip.io'. At the bottom, there are two more deployment entries: 'dataverse-postgresql, #1' and 'dataverse-solr, #1', each with a pod count of 1. A gear icon is visible between the deployment entries.

# OpenShift Dashboard - Scaling

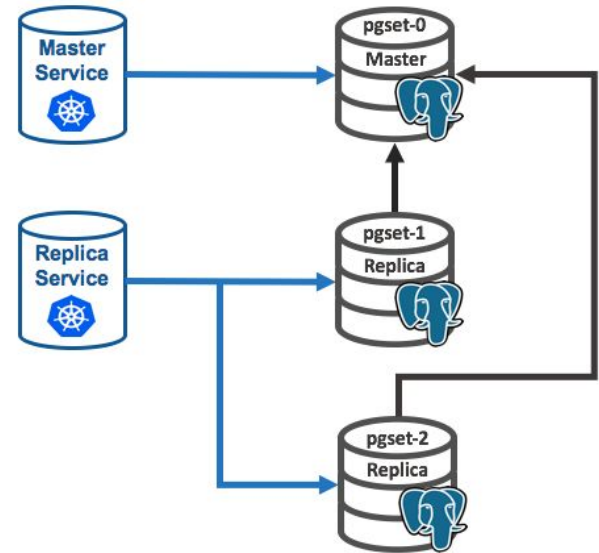
The screenshot displays the OpenShift Dashboard interface. On the left is a sidebar with navigation links: Overview, Applications, Builds, Resources, Storage, and Monitoring. The main content area shows the 'dataverse' deployment details. At the top, there's a URL: <http://dataverse-project1.192.168.99.100.nip.io>. Below this, the 'DEPLOYMENT' section for 'dataverse-glassfish, #1' is expanded. It shows container details for 'DATAVERSE-PLUS-GLASSFISH', including the image 'iqss/dataverse-glassfish-f814b85' and ports '8080/TCP'. A circular progress indicator shows '1 pod'. To the right of this indicator, a red arrow points to it with the text 'Manually Scale Pods'. Below the deployment details, the 'Networking' section shows 'SERVICE Internal Traffic' for 'dataverse-glassfish-service' with port '8080/TCP (web) -> 8080'. It also shows 'ROUTES External Traffic' for 'http://dataverse-project1.192.168.99.100.nip.io' and 'http://dataverse-glassfish-service-project1.192.168.99.100.nip.io'. At the bottom, there are two more deployment entries: 'dataverse-postgresql, #1' and 'dataverse-solr, #1', each with a '1 pod' indicator.



## Replica pods

How does an application use multiple servers or databases?

In general we will use the concept of StatefulSets from Kubernetes to create master-slave relationships between replicas



Source: <http://blog.kubernetes.io/>



## So what are we doing?

Dataverse is already running on OpenShift, but we need to configure each component

- PostgreSQL - configuring with StatefulSets is documented
- Glassfish - similar concept but specific configuration is novel
- Solr - configuring search indexer is a reach goal



# Users

**Researchers** - Want to publish data and analyse code on a reliable platform that can handle a high volume of traffic if research findings see a spike in popularity. Also open to be able to reexamine existing data in a new way, for example use machine learning.

**Journals** - Want to verify and publish author's research findings and data using dataverse repositories to increase the impact of journals and preserve data and make it citable.

**Institutions** - Need a place to host research data using customized dataverses for researchers, departments, and faculty to share their data. Deploying scalable Dataverse on OpenShift to production should be simple.

**Developers** - Develop Dataverse on a local version of OpenShift and easily deploy changes to production

**Companies** - Want to track the running instances of dataverse and collect results from tests.



# Users Stories

## Sprint #1

- As a student I want to read a project description that will help me understand this project.
- As a developer I want to deploy dataverse locally so that I can further develop it
- As an audience member I want to see slides that will help me understand the project
- As a developer I want to be able to incorporate my changes into production environment easily

## Sprint #2

- As a developer I want to be able to employ Glassfish as a Kubernetes statefulset.
- As a developer I want to be able to employ PostgreSQL as a Kubernetes statefulset..



# Release Planning

## **Release #1**(Feb 8) - Initial

- Stand up Dataverse on local OpenShift (using minishift)
- Develop user stories
- Prepare presentation

## **Release #2** (Feb 22) - PostgreSQL

- Modify OpenShift config in Dataverse to allow a scaled PostgreSQL

## **Release #3** (Mar 15) - PostgreSQL

- Finish PostgreSQL

## **Release #4** (Mar 29) - Glassfish

- Modify OpenShift config in Dataverse to allow a scaled Glassfish

## **Release #5** (Apr 12) - Glassfish

- Finish PostgreSQL

## **Release #6** (Apr 26) - Solr

- Deploy Dataverse into the MOC's OpenShift deployment
- Run the load test against Dataverse in the MOC.





## Release Planning



# TAIGA

<https://tree.taiga.io/project/msdisme-2018-bucs528-template-6/>



THANKS!!



GlassFish

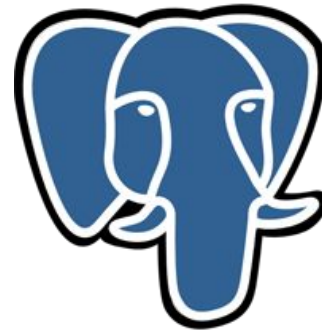
The  
**Dataverse**  
Project



Solr 



redhat®



PostgreSQL

Boston University College of Engineering

