Dataverse Scaling

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

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

Boston University College of Engineering

Project Overview

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.

The Dataverse Project

- 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

Screenshot of Cloud Dataverse

Dataverse

User Guide Support Sign Up Log In

Massachusetts Open Cloud - Cloud Dataverse

Metrics 4 Downloads

Contact Share

Search this dataverse... Find Advanced Search

Add Data

Dataverses (0)

Datasets (38)

Files (280)

Author Name

Eduardo Somarriba (5)

Rolando Cerda (5)

(2)

Ledesma Larrea, Pablo (2)

A. Degrande (1)

More...

Subject

Agricultural Sciences (1)

Other (1)

Social Sciences (1)

Keyword Term

1 to 10 of 38 Results

Sort

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

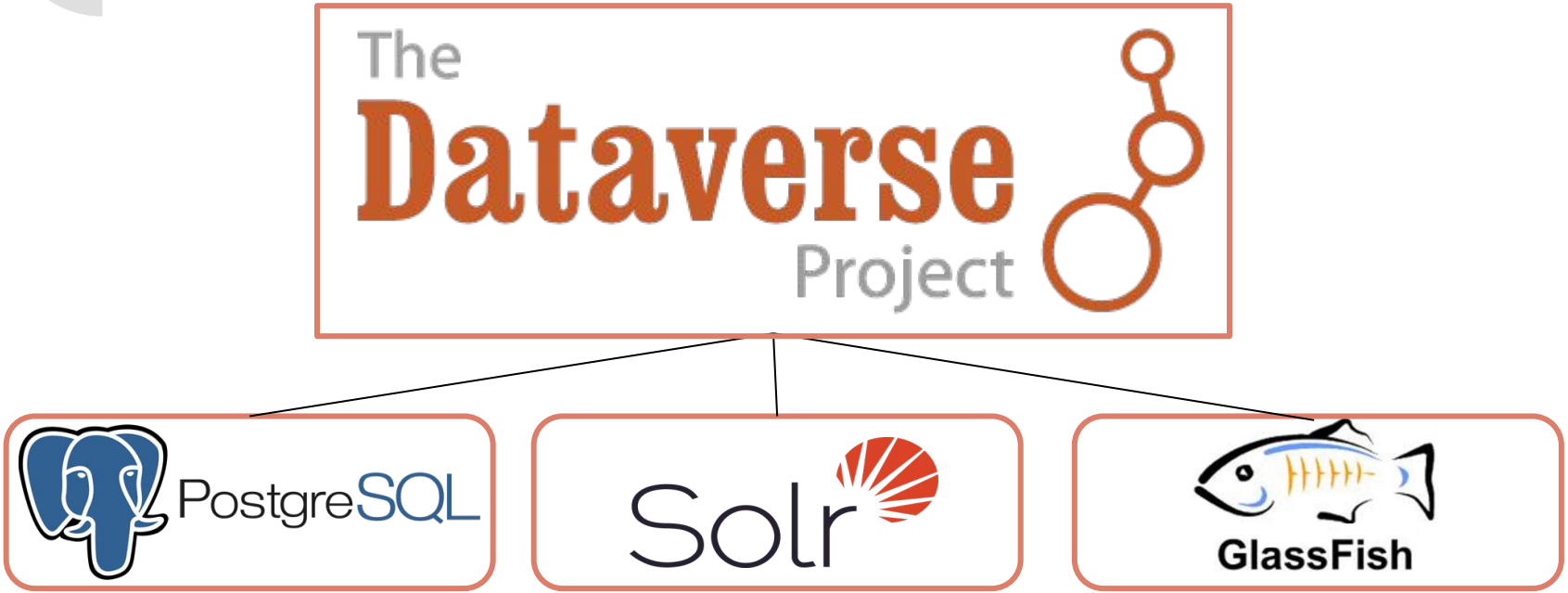
Aug 29, 2017

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

- Application database

- Application server

- Indexing / search engine

OPENSHIFT

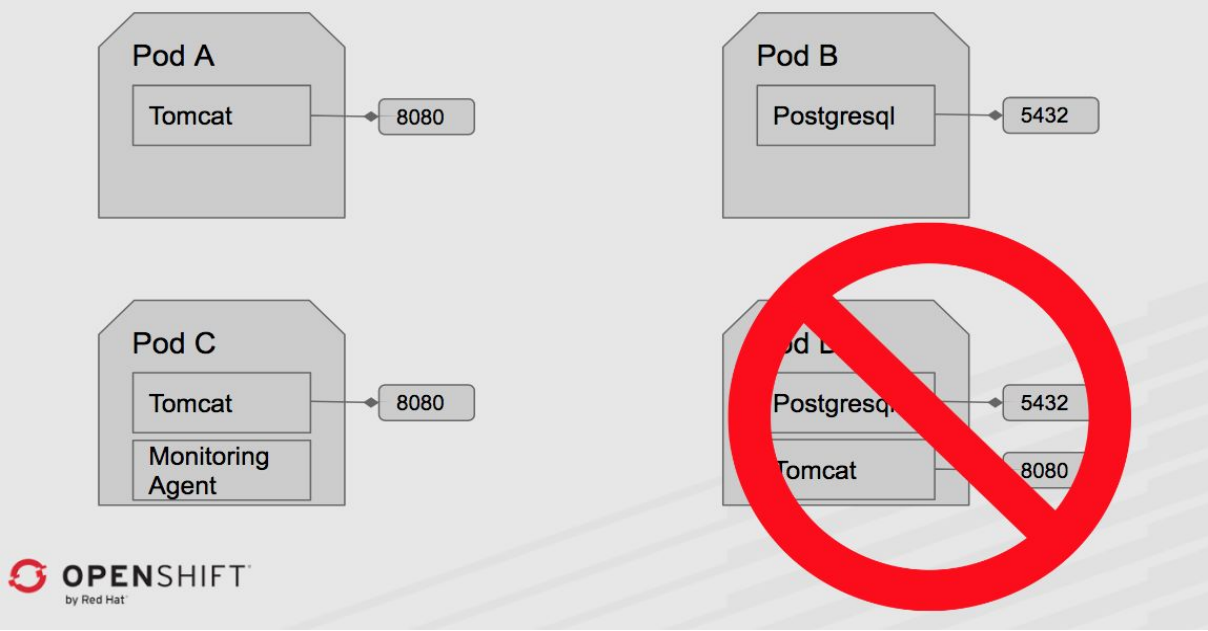
- Red Hat's container orchestration software built on top of Kubernetes.
- Each component has a Docker image and has been configured to run on OpenShift

Single deployment vs containerization/OpenShift

The original project consists of a single instance of each

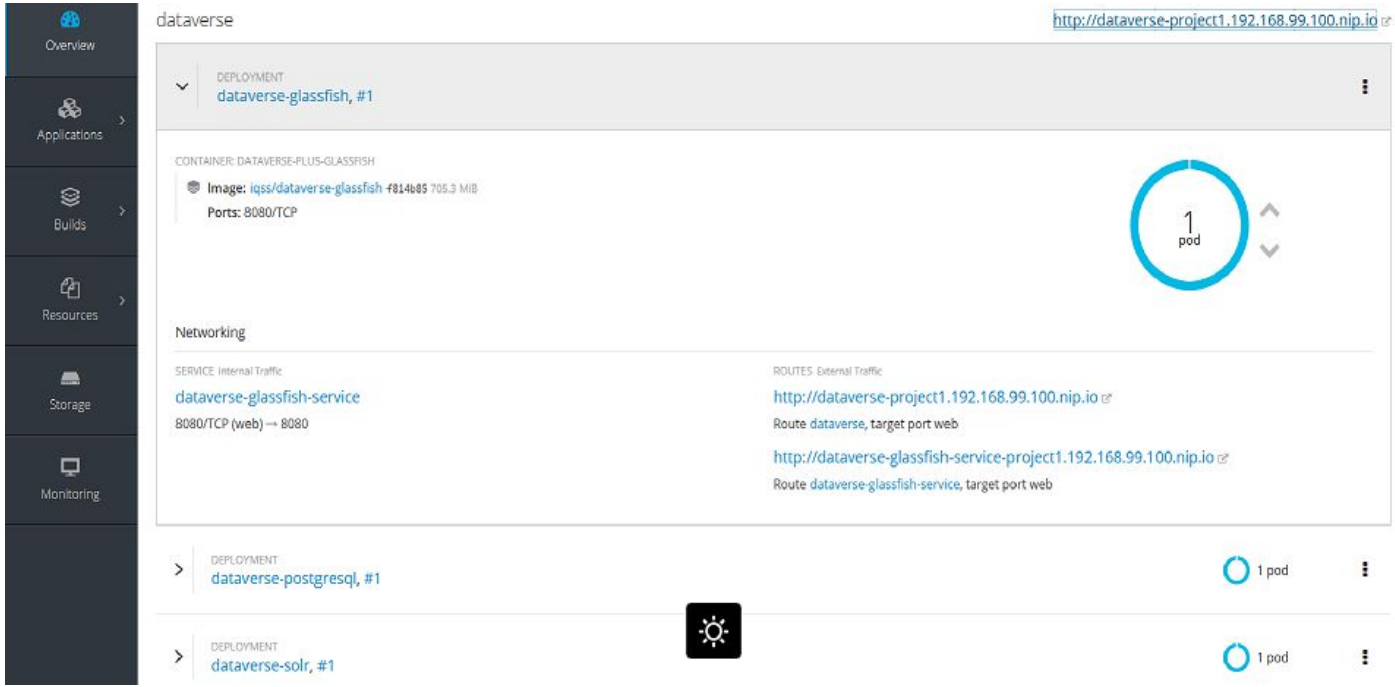
A pull request has been accepted to containerize each component and run on Openshift, but the pods do not scale

Containers, pods, scalability

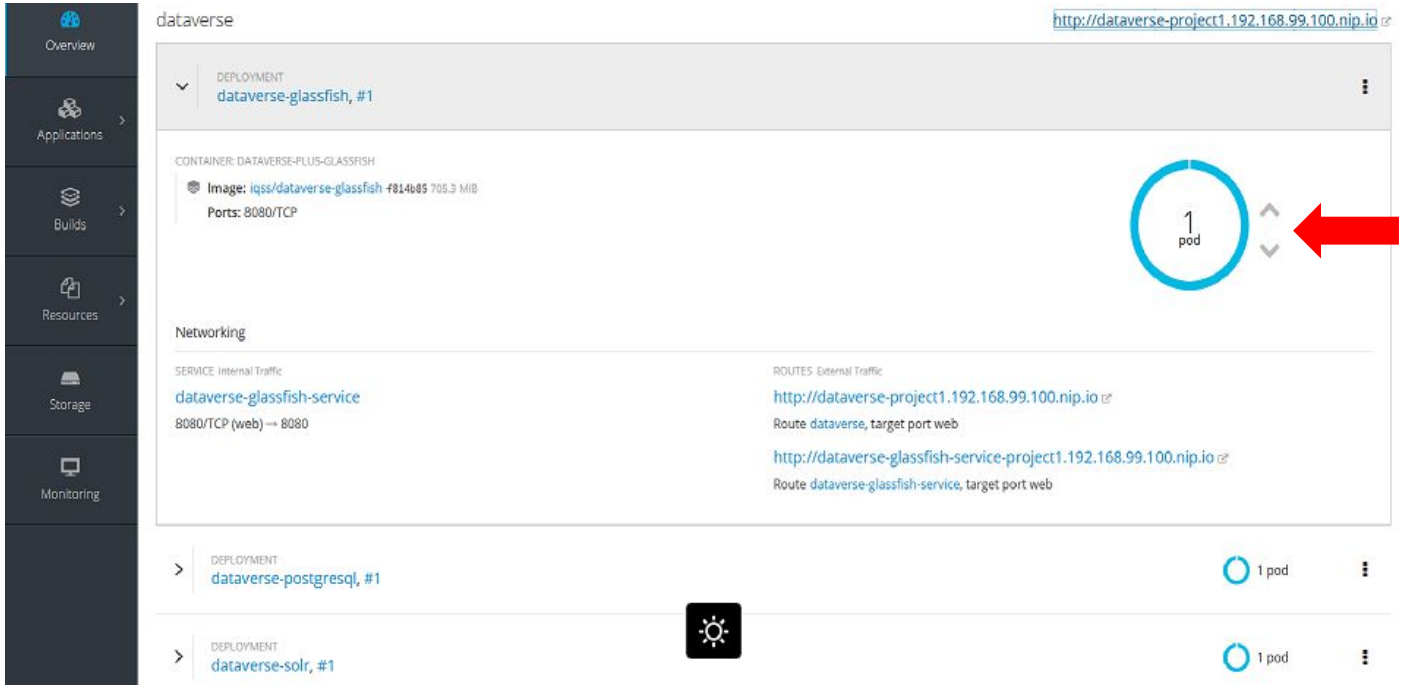


Source: OpenShift Tutorial slides shared by mentor

OpenShift Dashboard - Scaling



OpenShift Dashboard - Scaling

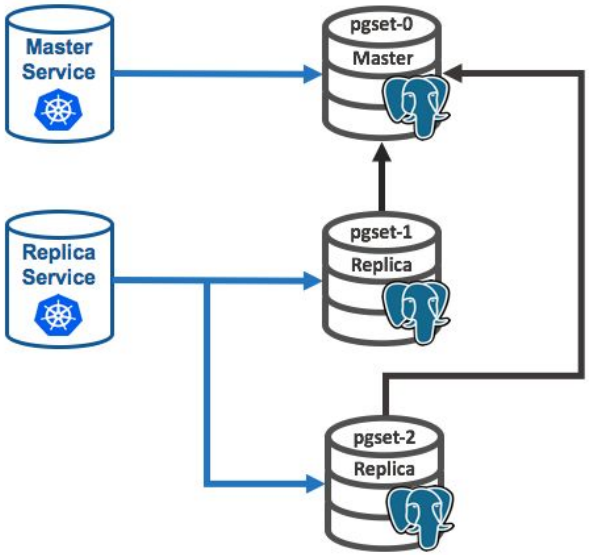


Manually Scale Pods

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

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

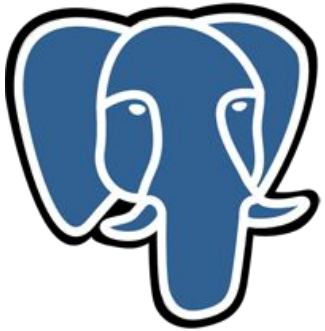
THANKS!!

GlassFish

The Dataverse Project

Solr

redhat



PostgreSQL