Best practises in designing, building and deploying Dataverse EU on CESSDA Technical Infrastructure.

Wilko Steinhoff - DANS
CESSDA Expert Seminar: 26 September 2018
wilko.steinhoff@dans.knaw.nl
Introduction

» What is Dataverse
» Dataverse EU project
» Dataverse application
» CESSDA Technical Infrastructure
» Deploy Dataverse EU on Google Cloud Platform (GCP)
» Issues we ran into
» Continuous deployment
» Lessons learned so far
Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data.

Developed by Harvard University (https://dataverse.org).

It is a container that holds datasets and sub-dataverses.

» Data Files
» Documentation
» Code
» Metadata
Dataverse**EU** project

Dataverse**EU** is a one-year project in which we try to realise to a data repository service for CESSDA Service Providers who have limited technical resources.

Deploy Dataverse repository software on CESSDA Technical Infrastructure

Project is funded by the CESSDA 2018 workplan

Dataverse**EU** partners:

- ADP (Slovenia), AUSSDA (Austria), GESIS (Germany), SND (Sweden), TARKI (Hungary), SiencePro (France), UKDA (UK), UniData (Italy), SODA (Belgium), LSZDA (Latvia), DANS (Netherlands)

DANS team:

- Marion Wittenberg, Laura Huis in 't Veld, Vyacheslav Tykhonov, Eko Indarto, Wilko Steinhoff
Dataverse components

Web based Java application (**Web application ARchive**)

- PostgreSQL database
- Solr search-index
- Java EE Servlet container (Glassfish)
Dataverse architecture

Glassfish application server

Dataverse Web Application (WAR)

PostgreSQL DB

Solr index

TCP

HTTP
CESSDA Technical Infrastructure

CESSDA Technical Infrastructure “requirements”
» BitBucket Git
» Docker containers
» Kubernetes (K8S) @ Google cloud (Docker orchestrator)
» Use descriptor files (YAML) to define K8S Services and Deployments
» Jenkins; Continuous Deployment pipeline
» Infrastructure as code (IaC)

Benefits:
» Scalable cluster (horizontal and vertical)
» Docker containers are managed by K8S cluster manager
» Portable applications (IaC)
Dataverse EU resources needed

» GCP VM-instance(s)
» Kubernetes (K8S) cluster
» Docker images
» K8S Services
» K8S Deployments (workloads)
» Load Balancer or Ingress (SSL certificates)
» Mail relay (default port 25 is blocked)
» Persistent volumes (storage) for the Solr, Postgres, Dataverse and SSL services
Interaction with Google Cloud Platform

Google Cloud Platform

» **GCloud**
  
  a command-line (CLI) tool for running commands against Google Cloud Platform. GCloud is part of the Google Cloud SDK

Kubernetes cluster

» **Kubectl**
  
  A command line interface (CLI) for running commands against Kubernetes clusters

» **Use of .yaml configuration files to describe the desired resources**
Dataverse EU - Best practices

Persistent Storage

Docker containers write files to disk (I/O) for state or storage

Problem:
If a Docker container is restarted for some reason, all data will be lost.

Solution:
Mount Persistent storage into the container on external disk.
Internationalization (i18n)

Language - Internationalization (i18n)

Problem:
No native in-app language switch

Solution:
Run multiple localized Dataverse docker containers on the cluster
Switch redirects to domain running a localized docker image

Drawback:
» One Dataverse container per language needed, instead of just one
» Large set of rules for the Ingress / Load Balancer
Internationalization (i18n)
The root dataverse.
Create Health Checks and ReadinessProbes on your container

**Problem:**

**Error: Server Error**

The server encountered a temporary error and could not complete your request.

Please try again in 30 seconds.

**Solution:**

K8S will create a default Health Check for you on the Ingress/Load balancer if you do not supply one.

It will check for HTTP 200 OK at your applications root ‘/’ or the path you supplied to your custom readinessprobe.
Email

Sending Email by a mail relay from your application

*Problem:*  
SMTP outgoing port 25 (smtp) is blocked on GCP

*Solution:*  
Deploy a mail-relay container on the cluster, that listens (internally) on port 25, that relays your email to G Suite server over HTTPS, that can send the email

From DVeu container — TCP/25 —> Mail Relay container — Transport Layer Security (TLS) —> G Suite server —> Receiver inbox

The mail-relay container is available for CESSDA projects from the CESSDA docker Container Registry
SSL Certification

Problem:
Let’s Encrypt SSL certificates expire (90 days)

Solution:
Automatic Let's Encrypt certification renewal with ‘Certbot’ docker image in a Kubernetes cronjob
SSL Certification

Daily Let’s Encrypt cronjob:

SSL certificates in a Kubernetes secret:
## Workloads

Workloads are deployable units of computing that can be created and managed in a cluster.

### DataSets

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Type</th>
<th>Pods</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataverse-de</td>
<td>✓ OK</td>
<td>Deployment</td>
<td>1/1</td>
<td>dataview</td>
</tr>
<tr>
<td>dataverse-en</td>
<td>✓ OK</td>
<td>Deployment</td>
<td>1/1</td>
<td>dataview</td>
</tr>
<tr>
<td>dataverse-sl</td>
<td>✓ OK</td>
<td>Deployment</td>
<td>1/1</td>
<td>dataview</td>
</tr>
<tr>
<td>letsencrypt-cronjob</td>
<td>✓ OK</td>
<td>Cron Job</td>
<td>0/3</td>
<td>dataview</td>
</tr>
<tr>
<td>mailrelay</td>
<td>✓ OK</td>
<td>Deployment</td>
<td>1/1</td>
<td>dataview</td>
</tr>
<tr>
<td>postgres</td>
<td>✓ OK</td>
<td>Deployment</td>
<td>1/1</td>
<td>dataview</td>
</tr>
<tr>
<td>solr</td>
<td>✓ OK</td>
<td>Deployment</td>
<td>1/1</td>
<td>dataview</td>
</tr>
</tbody>
</table>
Continuous deployment

CESSDA offers Continuous Deployment with Jenkins, to build, test and deploy code.

https://cit.cessda.eu

“Pipeline as Code” with Jenkins

Jenkinsfile in project root
1. Developer pushes code to Bitbucket
2. Jenkins receives notification - build trigger
3. Jenkins clones the workspace
4. Runs tests
5. Creates docker image
6. Pushes the docker image to GCP container registry
7. Updates the kubernetes deployment
Continuous delivery

Stage View

Average stage times:
(Average full run time: ~49s)

<table>
<thead>
<tr>
<th></th>
<th>Declarative: Checkout SCM</th>
<th>Check environment</th>
<th>Build Docker image</th>
<th>Push Docker image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15s</td>
<td>102ms</td>
<td>56ms</td>
<td>47ms</td>
</tr>
<tr>
<td>#8</td>
<td>15s</td>
<td>179ms</td>
<td>55ms</td>
<td>52ms</td>
</tr>
<tr>
<td>Sep 25 19:19</td>
<td>5 commits</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

cessda eric
Lessons learned

» Re-use of (Docker) images (versioning!)
» Use readinessProbe (K8S)
» Infrastructure as code (IaC)
» Use automated deployment Pipelines (Jenkins)
Links

Dataverses:
https://uk.dataverse-dev.cessda.eu/
https://de.dataverse-dev.cessda.eu/
https://sl.dataverse-dev.cessda.eu/

Bitbucket:
https://bitbucket.org/cessda/cessda.dataverseeu/src/develop/

Jenkins CI/CD:
https://cit.cessda.eu

GCP Container Registry:
eu.gcr.io/cessda-development/cessda-mail-relay:latest
eu.gcr.io/cessda-development/dataverse-eu-certbot:latest
eu.gcr.io/cessda-development/dataverse-eu-solr:4.9.2.4
eu.gcr.io/cessda-development/dataverse-eu-postgres:4.9.2.4
eu.gcr.io/cessda-development/dataverse-eu-dv:4.9.2.4
Thanks for listening

Any Questions?