DivSeek Canada Portal Documentation

Lacey-Anne Sanderson

Oct 11, 2020

GUIDES:

The DivSeek Canada Portal is a web-based platform to implement association genetics workflows supporting plant breeding and crop research focusing on large scale plant genetic resources / crop genotype-phenotype data sets whose access is brokered / managed by the project.

CHAPTER

ONE

CREATE DIVSEEK CANADA PORTAL GUIDE

This guide will walk you through how to create a copy of the DivSeek Portal. The DivSeek portal has been designed to encapsulate the functionality in a collection of Docker containers so that you can **easily create your own portal with the same functionality to house your data**.

1.1 Finding a place to Host your Portal

The DivSeek Canada Portal is being designed to run within a Docker Compose deployed system of Docker containers when the application is run on a Linux server or virtual machine.

Platform Requirements:

- An operating system able to install docker and run instances; Linux preferred.
- The ability to expose ports to the outside world.
- **Enough physical space to house the website containing your data. You can choose to keep all of the following on a single p**
  - Root partition: at least 20Gb
  - Docker partition: at least 200Gb for docker
  - Each Crop partition: at least 500Gb
- Enough processing power to quickly deliver pages (at least 4 core, 6 GB RAM) with additional processing power required for analysis completed via Galaxy.

1.1.1 ComputeCanada OpenStack Cloud setup

We have chosen to host the main DivSeek Canada portal on the ComputeCanada Cloud. To do so for your portal, first request an allocation which based on the platform requirements listed above. The following steps indicate how to prepare your allocation for your own portal.

Create the Cloud Instance

We start by creating a persistent p4-6gb (4 core, 6 GB RAM) flavour of compute instance. The security group should open up the TCP/IP ports exposed by the various docker instances, as specified in the project's docker-compose.yml file.

Note: It is important to ensure that the "root" partition gets a sufficiently large disk volume size - typically, at least 20 gigabytes - generally larger than the default image size - which is sometimes as small as 2 gigabytes. This root volume size must be explicitly set during cloud instance launch since it is not generally changeable after the instance has been launched.

Docker Image and Volume Storage

By default, the Docker image/volume cache (and other metadata) resides under /var/lib/docker which will end up being hosted on the root volume of a cloud image, which may be relatively modest in size. To avoid "out of file storage" messages, which related to limits in inode and actual byte storage, it is advised that you remap (and copy the default contents of) the /var/lib/docker directory onto an extra mounted storage volume (which should be configured to be automounted by fstab configuration).

In effect, it is generally useful to host the entire portal and its associated docker storage volumes on such an extra mounted volume. We generally use the /opt subdirectory as the target of the mount, then directly install various code and related subdirectories there, including the physical target of a symbolic link to the /var/lib/docker subdirectory. You will generally wish to set this latter symbolic link first before installing Docker itself (here we assume that docker has not yet been installed (let alone running).

In Compute Canada, using the OpenStack dashboard, a cloud "Volume" can be created and attached to a running DivSeek Canada Portal cloud server instance. We suggest creating a volume at least 200 GB in size (to allow for significant genomic data storage). After attaching the volume to the instance, the volume is initialized and mounted from within an SSH terminal session, as follows (where '$' is the Linux Bash CLI terminal prompt):

```
# Before starting, make sure that the new volume (here, 'vdb') is visible (should be!)
# NAME    MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
# vda     254:0    0  2.2G  0 disk
# ├─vda1  254:1    0  2.1G  0 part /
# ├─vda14 254:14   0    4M  0 part
# └─vda15 254:15   0  106M  0 part /boot/efi
# vdb     254:16   0  200G  0 disk

# First, initialize the filing system on the new, empty, raw volume (assumed here to␣
↪be on /dev/vdb)
sudo mkfs -t ext4 /dev/vdb

# Mount the new volume in its place (we assume that the folder '/opt' already exists)
sudo mount /dev/vdb /opt

# Provide a symbolic link to the future home of the docker storage subdirectories
sudo mkdir /opt/docker
sudo chmod go-r /opt/docker

# It is assumed that /var/lib/docker doesn't already exist.
# Otherwise, you'll need to delete it first, then create the symlink
sudo ln -s /opt/docker /var/lib
```

It is also recommended that a separate additional volume for each crop dataset deployed be created, following the above instructions. A 500 GB volume is likely to be needed for this given the genomic large data sets involved. Also,

you can attach volumes containing the raw input data (e.g. VCF files) as you need (Note: you should obviously not run the mkfs command afresh on any volumes which already have data!) It is recommended first that you create a subdirectory /opt/divseekcanada then mount these additional volumes in that subdirectory, namely something like the following:

```
# Create a master folder for the DivSeek Canada code
sudo mkdir -p /opt/divseekcanada
# ensuring easy $USER access to these resources
sudo chown ubuntu:ubuntu /opt/divseekcanada
# Add the data volumes
sudo mkdir -p /opt/divseekcanada/data/downy-mildew
sudo mount /dev/vdc /opt/divseekcanada/data/downy-mildew
sudo mkdir -p /opt/divseekcanada/Sunflower
sudo mount /dev/vdd /opt/divseekcanada/Sunflower

After completing the above steps, you should configure /etc/fstab file for system␣
↪boot up mounting of the new volumes:

# These volumes need to be auto mounted upon each reboot of the system
# so you should (carefully) add them to the Linux /etc/fstab file
# of the server, something like the following text entries (customize for your crop):
/dev/vdb        /opt    ext4    rw,relatime     0       0
/dev/vdc        /opt/divseekcanada/data/downy-mildew    ext4    rw,relatime     0     ␣
↪  0
/dev/vdd        /opt/divseekcanada/Sunflower    ext4    rw,relatime     0       0

# test the fstab mount with a 'fake' mounting
sudo mount -vf
```

Now, you can proceed to install Docker and Docker Compose.

CHAPTER

TWO

GENOME CANADA PILOT PROJECT

The first iteration of the platform is funded under a Genome Canada Project with co-funding from other partners.

> Growing populations, a changing climate and increasing constraints on land, water and fertilizer together translate into increased risks to global food security and pressure to dramatically expand agricultural productivity in Canada – and quickly. This can't happen, though, without accelerated plant breeding programs to develop high-yielding, climate-friendly and "earth-friendly" plant varieties. Further, Canada is required by the terms of international treaties and agreements to develop mechanisms for sharing these plants and the genetic information underlying them.
>
> DivSeek Canada is a project that will offer a way forward on both fronts. DivSeek Canada is the Canadian arm of an international initiative, DivSeek – a community driven effort involving a diverse set of partners who have voluntarily come together to unlock the potential of crop diversity to enhance the productivity, sustainability and resilience of crops and agricultural systems. This new Canadian-based project will accelerate plant breeding by leveraging the genetic diversity in the world's live collections and seed banks to create a unified, coordinated and cohesive information management platform. Canadian stakeholders will be consulted to guide the development of the platform, establish it on Compute Canada infrastructure and populate it with genomic information for three Canadian crops (lentils, flax and sunflower). It will also make mapping, breeding and visualization tools available on the platform for plant breeders, develop training resources and develop a long-term plan for its continued enhancement, as well as sustainable hosting, outreach and stakeholder support in Canada.
>
> The platform not only provides an expandable database for Canadian crop information, but also offers a model for the DivSeek initiative globally. It is expected to galvanize the use of genomic information by plant breeders to accelerate crop breeding in Canada, particularly in small to medium-sized crop communities who have not previously had the financial resources or bioinformatics skill set to exploit the genomic information available.
>
> —DivSeek Canada: Harnessing Genomics to Accelerate Crop Improvement in Canada