Installing and using conda programs on AWS

Kim Dill-McFarland, kadm@uw.edu

version September 22, 2021

This is a bare bones tutorial of how to install anaconda and associated packages on an Amazon AMI EC2 instance. It assumes that you have followed the `1.AWS_setup_tutorial.pdf` to setup an AWS account and are familiar with the steps to create an EC2 instance.

Create EC2 instance

Online through AWS console, create an instance (https://us-west-2.console.aws.amazon.com/ec2).

Recommendations for genomics usage:

- Amazon Linux 2 AMI
- Memory ("m" or "r") or compute optimized type ("c"), depending if you have a few large processes vs many small ones
  - For example, RNAseq fastq processing should be run on a "c" type with approximately 1 processor per sample
- Increase base EBS volume to 32 GB to hold programs and packages
- Add additional 1TB EBS volume(s) up to the amount you will need for all your raw and processed data

Basic setup

In the Terminal, log-in into your instance and perform basic setup. More details are available in `1.AWS_setup_tutorial.pdf`.

This directory setup is slightly different than in the setup tutorial. We mount the EBS storage to `project/` and then fuse a subdirectory `project/data/` to our S3 bucket holding all our data. This is because a fused directory becomes *read-only*. Thus, if we fused to the main directory as in the setup tutorial, we would not be able to write any of our results to the EBS storage where we have the extra space.

```
## Updates if available
sudo yum upgrade -y
sudo yum update -y

## Install AWS command line client if not using Amazon OS
sudo yum install awscli -y

## Configure your account
aws configure
## FILL IN WITH YOUR KEYS ###

## Setup fuse
sudo amazon-linux-extras install -y epel
sudo yum install -y s3fs-fuse
### Fuse key
```

```
### FILL IN WITH YOUR KEYS ###
echo UserKey:SecretKey > ~/.passwd-s3fs
chmod 600  ~/.passwd-s3fs

## Setup EBS volumes
lsblk

sudo mkfs -t ext4 /dev/nvme1n1
sudo mkdir -p ~/project
sudo mount /dev/nvme1n1 ~/project/
### Change permissions
sudo chmod 777 -R ~/project/

## Mount S3 data
mkdir ~/project/data
sudo chmod 777 -R ~/project/data

s3fs kadm-data ~/project/data -o passwd_file=~/.passwd-s3fs \
    -o default_acl=public-read -o uid=1000 -o gid=1000 -o umask=0007
```

Install anaconda

Anaconda (often called conda) is a program management system for bioinformatic tools. For novice users, I recommend installing as many tools as possible through this system. As you gain experience, you may find that you move away from conda to get newer or more frequently up-dated tools.

Install python

Check if python 3 is installed.

```
python --version
```

If it is not, install it.

```
sudo yum install python3 -y
```

Download anaconda

```
#Make directory for programs
sudo mkdir -p ~/apps/anaconda
sudo chmod 777 -R ~/apps
cd ~/apps/anaconda

# Change to correct URL if not using Linux 64-bit
# and to update to latest version if needed
sudo curl -O https://repo.anaconda.com/archive/Anaconda3-2021.05-Linux-x86_64.sh
```

Compile and install anaconda

```
sudo bash Anaconda3-2021.05-Linux-x86_64.sh -b -p /home/ec2-user/apps/anaconda

# Set PATH and initialize
```

```
eval "$(/home/ec2-user/apps/anaconda/bin/conda shell.bash hook)"
conda init
```

You may need to exit and re-login for changes to take effect.

Install programs in anaconda

```
## Configure channel priority
conda config --add channels bioconda
conda config --add channels conda-forge
conda config --set channel_priority false
conda config --set allow_conda_downgrades true

## Install
## Example programs for RNA-seq data cleaning
conda install -c conda-forge -y fastqc samtools
conda install -y adapterremoval bedtools
conda install -c bioconda/label/cf201901 -y picard star subread

#Check installs
conda list
```
