EasyBuild.io

building software with ease

VUB VRIJE UNIVERSITEIT BRUSSEL

VUB HPC

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

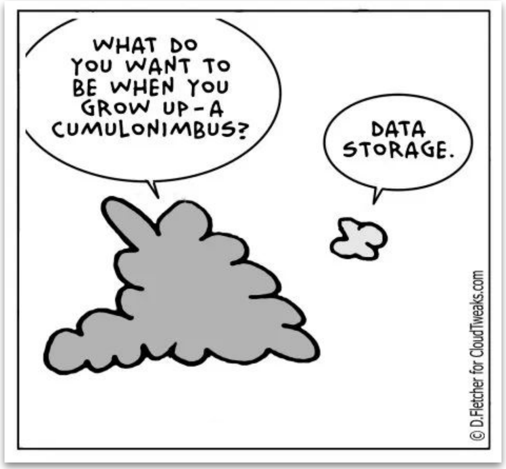
Installing datasets with EasyBuild

EasyBuild User Meeting 2024

Sam Moors (@smoors)

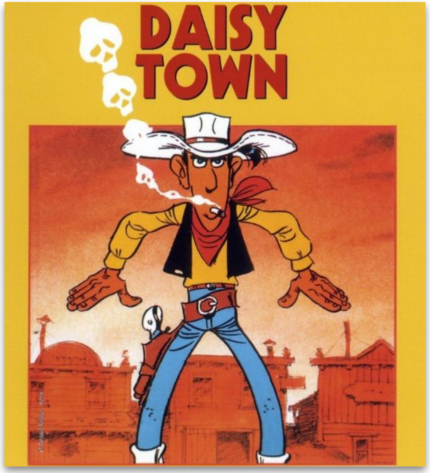
Outline

- Motivation
- Wishlist
- Leveraging EasyBuild
- Leveraging environment modules
- Optimizing storage
- Example use case
- More potential use cases
- Random hallucinations


WHAT DO YOU WANT TO BE WHEN YOU GROW UP - A CUMULONIMBUS?
DATA STORAGE.
© D.Fletcher for CloudTweaks.com

Motivation

- Datasets for research software
  - More fields: Machine learning, Biology, Data science
  - More datasets
  - Bigger datasets
- Current situation
  - Installed by site admins
    - Custom scripts
    - Manual downloads
  - Installed by researchers individually
  - Checksums? Versioning?



Wishlist

- ❓ Checksums
- ❓ Versioning
- ❓ Reproducible
- ❓ Use as dependency for software
- ❓ Share installation recipes
- ❓ Easy discovery, easy loading
- ❓ Easy swapping between versions
- ❓ Minimize data duplication
- ❓ Dedicated fast storage

Leveraging EasyBuild

✅ Checksums

✅ Versioning

✅ Reproducible

✅ Use as dependency for software

✅ Share installation recipes

Leveraging environment modules

- ✅ Easy discovery, easy loading
  - `module spider/av my-super-data`
  - `module whatis my-super-dataset/123`
  - `module load my-super-dataset/123`
    - Sets environment variables: paths to datasets
- ✅ Easy swapping between versions
  - `module swap my-super-dataset/456`
    - Recommendation: no software dependencies

Optimizing storage

- ✅ Minimize data duplication
  - Central installation
- ✅ Dedicated storage
  - Custom installation location for datasets

Implementation

- Generic `Dataset` easyblock (inherits from `Binary`)
    - Parameters:

```
'extract_sources': [True, "Whether to extract data sources", CUSTOM],
'data_install_path': [None, "Custom installation path for datasets", CUSTOM],
'cleanup_data_sources': [False, "Whether to delete the data sources after installation", CUSTOM]
```

    - Post-processing via `postinstallcmds`

https://github.com/easybuilders/easybuild-easyblocks/pull/3246

Implementation

- First-class support in framework
  - `--subdir-data` similar to `--subdir-software`
    - Default = `‘data’`
  - `--installpath-data` similar to `--installpath-software`
    - Default = `--installpath` + `--subdir-data`
  - `--sourcepath-data` similar to `--sourcepath`
    - Default = same as `--sourcepath`

https://github.com/easybuilders/easybuild-framework/pull/4474

Can I create a module for a preexisting dataset?

- Yes!
  - **`eb my-super-dataset/123 --module-only --installpath-data /path/to/my-super-dataset`**
  - Or set parameter `data_install_path` in easyconfig

Example use case: RFdiffusion

- RFdiffusion = protein structure generation
- Depends on 2 datasets: models (3.9GB), schedules (33M)
- Problem:
  - Reinstall RFdiffusion without re-downloading models
  - Don’t force storing the models twice
- Solution: separate easyconfigs for the datasets
  - RFdiffusion-models
  - RFdiffusion-schedules

https://github.com/easybuilders/easybuild-easyconfigs/pull/20019

Potential use case: AlphaFold Database

- AlphaFold = ML-based Protein folding
- AlphaFold DB (~2.5 TB) is a collection of datasets
- Problem: DB is regularly updated, but not each dataset
  - Lots of data duplication between versions
- Solution: custom `Dataset`-derived EasyBlock
  - All datasets stored in a single location with checksums
  - Reuse datasets across versions using symlinks

Potential use case: ESM-2

- ESM-2 = Language model for proteins
- Uses pretrained PyTorch models (37GB)
- Problem: models downloaded on first use:
    - `torch.hub.load("facebookresearch/esm:main", "esm2_t33_650M_UR50D")`
    - Models stored in `$TORCH_HOME` (default = `~/.cache/torch`)
    - Data is not checked with checksums
- Solution: custom `Dataset`-derived EasyBlock
    - Use PyTorch as build dep
    - Set `$TORCH_HOME` to central storage
- Other software also using PyTorch models:
    - RELION-5
    - EvoDIFF

Random hallucinations (1)

- User-initiated automated central dataset installation?
  - cfr. EasyBob (Jörg), EESSI bot (Pedro), Gitlab auto installation/deployment (Alexander)
  - Custom EasyBlocks for data repos that require custom download procedures
    - Huggingface, Kaggle, …
  - User provides:
    - A supported dataset repo
    - Dataset name (+ version)

Random hallucinations (2)

- Loading datasets via EESSI?

EESSI

EUROPEAN ENVIRONMENT FOR
SCIENTIFIC SOFTWARE INSTALLATIONS

Random hallucinations (3)

- How to get a list of installed *datasets* but not *software*?
  - Hierarchical module system
  - Other solutions?

Thank you!

More Ideas, suggestions, questions?

"Using EasyBuild to install datasets is like taking a stroll in a well-manicured park. It's smooth, effortless, and before you know it, you're surrounded by the beauty of organized data without breaking a sweat!"

ChatGPT