Workflow managers in high-energy physics

Enhancing analyses with Snakemake

Jamie Gooding, TU Dortmund University
FOSDEM24 Open Research DevRoom
3rd February 2024

We acknowledge funding from the European Union Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086
What are workflow managers?

Quite literally “tools to manage workflows”

Workflow managers help to...

- Define a workflow
- (Re-)run a workflow
- Organise rules
- Document workflow
What are workflow managers?

Quite literally “tools to manage workflows”

Workflow managers help to...
- Define a workflow
- (Re-)run a workflow
- Organise rules
- Document workflow

Data → Workflow → Results

Workflow managers in high-energy physics

Jamie Gooding  FOSDEM 2024 Open Research DevRoom
Snakemake background

- Evolved from GNU Make paradigm
  - Workflow defined from “rules”
  - Directed acyclic graph (DAG) links rules
  - Wildcards enable dynamic workflows
- Python-based language:
  - Shallow learning curve
- Significant ongoing development:
  - v8 released in Dec 2023
- Picked up in HEP over last ~5 years

https://snakemake.github.io/

**What is HEP?**

**HEP → High Energy Physics**

- Physics of the *very early* of universe
- Accelerate and collide particles
  - LHC built for this purpose
  - Experiments record collisions
- LHCb specialises in differences between *matter* and *anti-matter*

*Images: CERN*
What is HEP?

**HEP → High Energy Physics**

- Physics of the *very early* of universe
- Accelerate and collide particles
  - LHC built for this purpose
  - Experiments record collisions
- LHCb specialises in differences between *matter* and *anti-matter*
HEP analyses

- Analyses aim to measure something:
  - A particle’s mass, its lifetime, its possible decays
  - Look to contradict Standard Model
- Start with experimental data
- Extract measurement from data:
  - Dedicated scripts for processing
  - Shared, dynamic codebase
- Sizes of analyses can vary
HEP analyses

- Analyses aim to measure something:
  - A particle’s mass, its lifetime, its possible decays
  - Look to contradict *Standard Model*

- Start with experimental data

- Extract measurement from data:
  - Dedicated scripts for processing
  - Shared, dynamic codebase

- Sizes of analyses can vary

Results must be reproducible
May need to rerun analysis
HEP analyses

- Analyses aim to measure something:
  - A particle’s mass, its lifetime, its possible decays
  - Look to contradict Standard Model
- Start with experimental data
- Extract measurement from data:
  - Dedicated scripts for processing
  - Shared, dynamic codebase
- Sizes of analyses can vary

- Results must be reproducible
- May need to rerun analysis
- Large scales of data† (Several TB per analysis)
- Often stored remotely

†This will only get larger…
HEP analyses

- Analyses aim to measure something:
  - A particle's mass, its lifetime, its possible decays
  - Look to contradict Standard Model

- Start with experimental data

- Extract measurement from data:
  - Dedicated scripts for processing
  - Shared, dynamic codebase

- Sizes of analyses can vary

Results must be reproducible
May need to rerun analysis
Large scales of data†
(Several TB per analysis)
Often stored remotely
Analysis scripts can change frequently
Must support scripts of many languages/formats

†This will only get larger...
HEP analyses

- Analyses aim to measure something:
  - A particle’s mass, its lifetime, its possible decays
  - Look to contradict **Standard Model**
- Start with experimental data
- Extract measurement from data:
  - Dedicated scripts for processing
  - Shared, dynamic codebase
- Sizes of analyses can vary

Results must be reproducible
May need to rerun analysis
Large scales of data† (Several TB per analysis)
Often stored remotely
Analysis scripts can change frequently
Must support scripts of many languages/formats
Must be scalable and deployable

†This will only get larger...

Jamie Gooding  FOSDEM 2024 Open Research DevRoom
Workflow managers in high-energy physics  5
Snakemake in Analysis

- Snakemake meets these needs!
- Well-established user base in LHCb:
  - Internal expertise → internal training (right)
- Features and functionality suit analyses well:
  - Interface with HPC resources
  - Remote protocol integration

Snakemake pipelines @ the LHCb experiment at CERN

Workshop on Basic Computing Services in the Physics Department - subMIT
2024-02-02 @ MIT

https://github.com/reallyblaised/snakemake-tutorial
Scalable, deployable workflows

- Include/sub-workflows/modules/wrappers break into smaller files
- Checkpoints for flexible workflow definitions, re-evaluating DAG
- `--batch` flag divides many jobs from a rule into batches
- Conda environment package requirements
Distributed computing

- Large data scales require large computing scales!
- Use of clusters for processing, fitting, etc., common
- Snakemake supports common interfaces (see right)
- Submitting rules as cluster jobs is straightforward:
  - Define profile, run with `--profile {profile}` flag
  - Resource limits can be set globally/per rule
  - Rules can be specified as local to run locally

Supported frameworks†:

†list is not exhaustive!
Remote file access

› Files usually stored away from institutes:
  - CERN EOS/Worldwide LHC Computing Grid
› remote module provides easy implementation
  - Simply initialise provider and wrap
    `{provider}.remote(path)`
  - `glob_wildcards` and `keep_local`
What do analysts need?

- **Scalability**
  - Data scales will skyrocket (see right)
  - Experiments growing by O(100) authors each year

- **Usability**
  - Analysts not software devs by trade

- **Functionality**
  - Closer collaboration between devs and HEP users

ATLAS Collab., 2022 (CERN-LHCC-2022-005)

Implement Ganga as an executor for snakemake
#2095

https://github.com/ganga-devs/ganga/issues/2095
Conclusions

- Workflow managers (e.g., Snakemake) deeply useful for research
- These tools meet HEP needs!
  - Functionality in place to leverage HEP resources
  - Use will become unavoidable in very near future (next few years)
- Should capitalise on field-specific user base
  - Room to collaborate on development/training

Useful papers/links

https://snakemake.readthedocs.io/
https://github.com/reallyblaised/snakemake-tutorial
https://hsf-training.github.io/analysis-essentials/snakemake/README.html


Get in touch

@goodingjamie
in/goodingjamie
GoodingJamie
jamie.gooding@cern.ch

Workflow managers in high-energy physics
Backup
Anatomy of a Snakemake rule

Let’s deconstruct a typical Snakemake rule

```python
rule rule_A:
    input:
        script = "{script_dir}analyse.py",
        infiles = expand("file{n}.csv", n=range(3)),
        config = rules.rule_B.output.config
    resources:
        mem_mb=200
    threads: 4
    output:
        results = "results.txt"
    shell:
        "python {input.script} --input {input.infiles} --config {input.config} --cores {threads} --output {output.results}""
```

Workflow managers in high-energy physics 13
Anatomy of a Snakemake rule

Let’s deconstruct a typical Snakemake rule

```python
rule rule_A:
    input:
        script = "\{script_dir\}analyse.py",
        infiles = expand("file{n}.csv", n=range(3)),
        config = rules.rule_B.output.config
    resources:
        mem_mb=200
    threads: 4
    output:
        results = "results.txt"
    shell:
        "python {input.script} --input {input.infiles}
        --config {input.config} --cores {threads}
        --output {output.results}""
```

- **Path defined as variable**
- **Expand method generates list of files**
- **Direct reference to rule output**
- **Specify memory requirement**
- **Number of threads per job**
- **Scaled down if fewer available**
The LHCb Experiment

LHCb Collab., 2014
(LHCBTDR-015)
Analysis reproducibility

- Recent push for reproducibility in HEP
- Many platforms/frameworks
  - Highlight: REANA (right)
    - Collation of FOSS tools and frameworks for reusable pipelines
    - Tools common between experiments
    - Uses shared CERN infrastructure
- Preservation of analyses is a current hot topic

https://reanahub.io/