Virtual Earthquake and seismology Research Community e-science environment in Europe
Project 283543 – FP7-INFRASTRUCTURES-2011-2 – www.verce.eu – info@verce.eu

The VERCE Training at LRZ: Data-intensive methods using dispel4py

(dispel4py)

9-11 March 2015

Introduction dispel4py

The Fourth paradigm — data-intensive research methods

What is a workflow?

Why use a workflow?

There are many workflow languages — why invent dispel?

What is dispel4py good for?

Malcolm Atkinson

Your Scientific Methods

- Do you automate them?
- How do you automate them? formalise them?
- How often do you repeat them?
  - With new data or new parameters or …?
- How do you improve them?
- How do you collaborate? Gain recognition & credit?
- Would you gain from handling more data? Faster?

What is a workflow?

http://dispel4py.org/

- A composition of steps: data-handling + data-analysis+simulation journey

- Many ways of forming steps
  - Require good libraries of ready made steps
  - *Learn to add your own* tomorrow

- *Many ways of combining steps* tomorrow

- Running in many computing environments

- Engineers take care of efficiency and reliability

- Recursive — a journey can be a step in another journey

A script or a program

Why use a workflow

http://dispel4py.org/

- Rapid prototyping and experiment
- Saving you labour and repeated drudgery
- Reducing error rates
- Saving you from doing your own housekeeping
  - Returning resources such as file space
  - Gathering all your results
- Acceleration due to workflow optimisation, e.g. parallelisation
- Sharing & getting credit for methods
- Incrementally improving methods
- Combining methods developed by different experts
- Combining computation and data handling steps written in different languages

What makes dispel4py different

Look out for old rival workflow systems from a fault zone!

Wings ⇒ Pegasus ⇒ HTCondor

Kepler

http://dispel4py.org/

Building on the strength and familiarity of Python
gain from many scientific libraries like ObsPy
use the tools you normally use to program

Operations on data units in data streams rather than tasks on files

Processes run continuously and concurrently coupled by streams

Why add dispel4py

http://dispel4py.org/

The DATA Bonanza - Improving Knowledge Discovery in Science, Engineering and Business, M Atkinson, R Baxter, P Brezany, O Corcho, M Galea, M Parsons, D Snelling & J van Hemert, Wiley 2013

Free download from http://onlinelibrary.wiley.com/book/10.1002/9781118540343

Raising the level of discourse
- Removing much technology specific information as technology changes
- Relieving users from concerns about optimisation

Improving the logical description
- Streams of data with auto-iteration over data units
- Multiple streams in & multiple streams out
- Behaviour, data interpretation & data representation

Kryder's Law & energy costs, …

Covering existing models
- Distributed query
  - Optimisation based on avoiding IO & characterising operators
- Real-time processing
- Task-based batch processing

What is dispel4py good for

http://dispel4py.org/

- **Everything** ….
  - but investment in libraries is needed for each *new* topic
  - plus common libraries for shared activities, such as data handling

- **Everything** ….
  - but the dispel4py *engineering team* need to
    - make it perform at the scales you need
    - make it excel on the DCIs you use
      - laptop to cloud via supercomputers & clusters
  - make it reliable

- So I will hand you over to their tender mercies