Future Generation Computer Systems 75 (2017) 216–227

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Editorial

Scientific workflows: Past, present and future

Malcolm Atkinson [a], Sandra Gesing [b,*], Johan Montagnat [c], Ian Taylor [b,d]

[a] University of Edinburgh, School of Informatics, Edinburgh EH8 9AB, UK
[b] University of Notre Dame, Center for Research Computing, Notre Dame, IN 46556, USA
[c] Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France
[d] Cardiff University, School of Computer Science & Informatics, 5 The Parade, Cardiff CF24 3AA, UK

ARTICLE INFO

Article history:
Received 17 December 2015
Received in revised form 20 April 2017
Accepted 24 April 2017

Keywords:
Scientific workflows
Scientific methods
Optimisation
Performance
Usability

ABSTRACT

This special issue and our editorial celebrate 10 years of progress with data-intensive or scientific workflows. There have been very substantial advances in the representation of workflows and in the engineering of workflow management systems (WMS). The creation and refinement stages are now well supported, with a significant improvement in usability. Improved abstraction supports cross-fertilisation between different workflow communities and consistent interpretation as WMS evolve. Through such re-engineering the WMS deliver much improved performance, significantly increased scale and sophisticated reliability mechanisms. Further improvement is anticipated from substantial advances in optimisation. We invited papers from those who have delivered these advances and selected 14 to represent today's achievements and representative plans for future progress. This editorial introduces those contributions with an overview and categorisation of the papers. Furthermore, it elucidates responses from a survey of major workflow systems, which provides evidence of substantial progress and a structured index of related papers. We conclude with suggestions on areas where further research and development is needed and offer a vision of future research directions.

1. Introduction

Data-intensive Workflows (a.k.a. scientific workflows) are routinely used in the majority of data-driven research disciplines today, often exploiting rich and diverse data resources and parallel and distributed computing platforms. Workflows provide a systematic way of describing the methods needed and provide the interface between domain specialists and computing infrastructures. Workflow management systems (WMS) perform the complex analyses on a variety of distributed resources. With the dramatic increase of primary data volumes and diversity in every domain, workflows play an ever more significant role, enabling researchers to formulate processing and analysis methods to extract latent information from multiple data sources and to exploit a very broad range of data and computational platforms.

2. Highlights over the past 10 years

This special issue celebrates significant progress over the past ten years that has greatly increased the use of data-intensive workflows, built on substantial improvements in their usability, capabilities, architecture and reliability. Ten years ago there were diverse scientific workflow systems showing promise and early use [1]. We asked leaders of workflow groups, "*What was the most significant result from using your workflow system in the last 10 years?*" Their replies are collated in Table A.1; it offers an extensive body of evidence with comprehensive coverage and a structured index into the literature. Some highlights are presented here.[1] Pegasus played a key role in the detection of gravitational waves by offering sustained and reliable data handling and computation for the long-running research campaign.[2] Increased capacity, scale and reliability is reported in almost every case. The scientific workflow community spawned new technology developments in workflow-based data provenance. Kepler made early progress in this work and after a number of grand challenges, Gill started activity at W3C

* Corresponding author.
E-mail addresses: malcolm.atkinson@ed.ac.uk (M. Atkinson), sandra.gesing@nd.edu (S. Gesing), johan.montagnat@cnrs.fr (J. Montagnat), ian.j.taylor@nd.edu (I. Taylor).

http://dx.doi.org/10.1016/j.future.2017.05.041
0167-739X/

[1] References to Table A.1 refer to the rows associated with the named workflow system.
[2] https://pegasus.isi.edu/application-showcase/ligo/.

with an incubator[3] that proposed a core vocabulary[4] and which led to a working group (Moreau and Groth) who saw through the process all the way to W3C PROV standard.[5] Building on that standard, Taverna and WINGS have advanced workflow description significantly, initiating a foundation for reasoning about multiple workflow languages consistently. Building long-term relationships with communities, including tuning access via well-crafted science gateways and composing extensive libraries of workflows and workflow fragments has proved particularly productive, pioneered by Taverna but now almost universal. A striking example is Galaxy, with thousands of users on its public site and 4,300 publications citing Galaxy's contribution to their results in the last 10 years. That progress means that using workflows has become routine, *e.g.* KNIME. Table A.1 contains many examples of delivering success to others using the power of data-driven workflows.

There remains a diversity of systems, many with their own investments, culture and committed communities; some have remained leaders while others have been replaced, but the quality, support and maturity has increased across the board [2]. The scientific workflow community has educated the world; 10 years ago very few researchers had heard of workflows, today virtually every domain uses them. Some aspects of the consolidated progress are presented.

The tools for the whole workflow lifecycle have much improved, eliminating many impediments to use, and greatly improving the productivity of all those, from domain experts to data engineers, who work with workflows. Ten years ago the tooling was focused on authoring and adapting to distributed computing interfaces. It now extends to managing research campaigns using workflows, with automation of data identification, exploitation of provenance, support for curation and oversight of progress, processes and performance [3]. The combination of improved abstraction and full-lifecycle tools has made it much easier for end users to understand, select, reproduce, adapt and use previously developed workflows. This builds on the sustained investment in building libraries of workflows, workflow fragments, re-usable components, accessible services, and established models for describing and cataloging these resources so they are easily found [4,5].

The abstract definition and representation of workflows has advanced significantly, *e.g.* as reported in this issue by Garijo et al. [6]. This yields four benefits:

1. The meaning of scientific methods encoded in workflows is less dependent on implementations and therefore can be sustained as the digital environment evolves, thereby extending the benefits from investing in developing scientific methods as workflows.
2. This independent definition permits mappings to diverse platforms, enabling WMS to exploit the latest advances in hardware and software platforms.
3. Scientific workflows are more easily reused and repurposed, lowering experimental design cost and speeding-up discoveries.
4. Abstract definitions also facilitate sharing ideas and effective methods across discipline boundaries.

The development, deployment and use of pioneering data-intensive workflows that cope with the scale of modern data, the rates of demand, and the diversity of enactment contexts continues to require effective alliances between innovating domain experts, adept data scientists, and experienced systems engineers. However, substantial advances have been made to accelerate their work and improve productivity. Practitioners can work with moderate scale test data sets on their personal devices or local facilities, and then use exactly the same workflows on production platforms. Several factors contribute to this achievement:

1. Improvement in virtualised infrastructures, abstractions, tools and monitoring introduced above.
2. Automation and optimisation of data handling, coping with diverse sources, eliminating unnecessary transfers, and discarding unneeded storage.
3. Planning and optimisation that automatically adapts to resource availability, scale and load.
4. Minimised recovery costs after partial failures.

The software stacks that provide parallelised multistage distributed heterogeneous target environments for workflows have increased capacity, elasticity and recoverability. These platforms have and will continue to evolve thereby increasing the importance of abstraction and automated mappings as a means of preserving the meaning of scientific methods. However, this can pose challenging set up requirements, to provide the initial enactment context or to rebuild an earlier enactment context for scientific reproducibility – some early experience tackling such issues has been reported [7,8]. Many of these data-intensive platforms also have their own languages for creating data-driven methods, that are intimately integrated and well presented. These are emerging as new data-intensive workflow systems that appeal to communities who are not constrained by prior investments. The usability from such integrated systems has to be weighed against the potential for lock in.

Data distribution and data streaming are of growing significance. Over the past decade, the emergence of the Web of Data[6] and the Open Linked Data initiative[7] led to a much increased environment of remotely accessible and interoperable data sources for scientific analysis. In addition, data streaming occurs in the monitoring and exploitation of data from connected instruments, worn devices and the Internet of Things (IoT). Initially, it was the domain of signal processing communities and was treated differently; today the handling of data units flowing in streams from any source merges with the task-oriented traditional workflow approach. The Node-red system is a prime example.[8] The developers of several stream-based workflow enactment models show that this can also achieve substantial performance gains by reducing disk traffic [9].

There is a pervasive drive by funders, governments and researchers to increase the openness of research and the accessibility of data, motivated by the desire to increase equality, stimulate use and cross-fertilisation, and to improve quality by facilitating challenging review. This is captured as the FAIR principles [10]. The related workflows and their enactment context should be included, but workflows should also help implement those principles.

The diversity of data-intensive workflow languages will persist, partly because of prior investment but also because they are tuned to meet different requirements. Part of that investment is intellectual: learning, developing skills and understanding, and part is cultural, the interaction with other users. These factors, as well as wanting to continue practices that have proved effective, weigh against change. We may anticipate further languages emerging, from the interplay of programming language research, workflow system research and new application requirements. These will continue to yield benefits but the multiplicity of languages has inherent costs:

3 https://www.w3.org/2005/Incubator/prov/charter.
4 https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/.
5 https://www.w3.org/TR/2013/REC-prov-dm-20130430/.
6 https://www.w3.org/standards/semanticweb/data.
7 http://linkeddata.org/.
8 http://nodered.org.