Datenbank-Spektrum manuscript No.
(will be inserted by the editor)

The Collaborative Research Center FONDA

Ulf Leser · Marcus Hilbrich · Claudia Draxl · Peter Eisert · Lars Grunske · Patrick Hostert · Dagmar Kainmüller · Odej Kao · Birte Kehr · Timo Kehrer · Christoph Koch · Volker Markl · Henning Meyerhenke · Tilmann Rabl · Alexander Reinefeld · Knut Reinert · Kerstin Ritter · Björn Scheuermann · Florian Schintke · Nicole Schweikardt · Matthias Weidlich

Received: date / Accepted: date

Abstract Today's scientific data analysis very often requires complex Data Analysis Workflows (DAWs) executed over distributed computational infrastructures, e.g., clusters. Much research effort is devoted to the tuning and performance optimization of specific workflows for specific clusters. However, an arguably even more important problem for accelerating research is the reduction of development, adaptation, and maintenance times of DAWs. We describe the design and setup of the Collaborative Research Center (CRC) 1404 "FONDA – Foundations of Workflows for Large-Scale Scientific Data Analysis", in which roughly 50 researchers jointly investigate new technologies, algorithms, and models to increase the portability, adaptability, and dependability of DAWs executed over distributed infrastructures. We describe the motivation behind our project, explain its underlying core concepts, introduce FONDA's internal structure, and sketch our vision for the future of workflow-based scientific data analysis. We also describe some lessons learned during the "making of" a CRC in Computer Science with strong interdisciplinary components, with the aim to foster similar endeavors.

Keywords Scientific Workflows; Data Science; Distributed Systems

1 Introduction

Data and its subsequent analysis are central ingredients of today's science, culminating in the emergence of concepts like "Data Science" or "Big Data". The analysis of Big Scientific Data today is typically achieved by chaining together a large set of independently developed tools that implement specific base functionalities. We call such chains, which actually may have a complex topology, Data Analysis Workflows (DAW). The fast execution of DAWs over large data sets, in turn, depends on state-of-the-art computational infrastructures, comprising hardware, such as compute clusters, multi-core servers, or high performance computing systems, and software, such as resource managers, schedulers, or file systems [10,8]. However, changes in the scientific questions to study, in the experimental setup producing the data, and in the definition of "state-of-the-art" computational infrastructure create the need for the continuous creation, adaptation, redesign, reuse, and maintenance of DAWs.

Ulf Leser, Marcus Hilbrich, Claudia Draxl, Peter Eisert, Lars Grunske, Patrick Hostert, Timo Kehrer, Christoph Koch, Henning Meyerhenke, Alexander Reinefeld, Björn Scheuermann, Nicole Schweikardt, Matthias Weidlich
Humboldt-Universität zu Berlin
Berlin, Germany
Tel.: ++49 (0)30-2093-3902
E-mail: leser@informatik.hu-berlin.de

Odej Kao, Volker Markl
Technische Universität Berlin

Knut Reinert
Freie Universität Berlin

Peter Eisert
Fraunhofer Heinrich-Hertz Institute

Dagmar Kainmüller
Max-Delbrück Center for Molecular Medicine

Birte Kehr
Universität Regensburg

Tilmann Rabl
Hasso Plattner Institute, University of Potsdam

Alexander Reinefeld, Florian Schintke
Zuse-Institute Berlin

Kerstin Ritter
Charité - Universitätsmedizin Berlin

DAWs therefore must be easy to develop, easy to adapt, easy to reuse and reproduce, and robust with respect to (slight) changes in input data or infrastructure. Current systems for specifying and executing DAWs are far from meeting these requirements. They are hardly usable by the growing number of Natural Science researchers who are computer-savvy but who are not experts in distributed systems, software engineering, or performance tuning. This creates an enormous bottleneck in times when data-driven research becomes ubiquitous and researchers call for a democratization of Data Science [3].

The Collaborative Research Center FONDA sets out to improve this situation considerably. It focuses on researching new techniques, algorithms, and models to increase the productivity of building, adapting, monitoring, and (re-)using DAWs. FONDA addresses the entire software stack necessary for modern large-scale DAW executions, ranging from specification languages to execution engines to the underlying resource managers and file systems. It takes a holistic view on DAWs, covering the entire DAW life cycle from specification through execution to monitoring and adaptation. FONDA will approach its long-term goals in different phases. In its first phase, FONDA investigates three particularly important aspects of productivity for DAW systems, namely portability, adaptability, and dependability (PAD):

- **Portability (P):** A DAW is portable when it can be executed on a variety of different computational infrastructures. Portability often is achieved by using declarative specifications or compiler techniques.
- **Adaptability (A):** A DAW is adaptable when it can self-adapt to different input data or different infrastructures as much as possible. Input data may, for instance, differ in its format or value distribution. Adaptations may be triggered automatically or by means of developer-provided elements in the DAW specification.
- **Dependability (D):** A DAW is dependable when it specifies and controls constraints that must be obeyed to make it run correctly, such as available main memory or size ranges of input files. Such constraints may be specified manually or be inferred automatically.

These properties were not chosen arbitrarily; instead, they directly correspond to key characteristics of scientific experimentations in general. To illustrate this point, consider a typical wet-lab protocol measuring the expression of some marker in a human sample (e.g., expression of a gene, presence of an antibody, existence of mRNA from a COVID-19 infection). Any such protocol is expected to be (1) portable, i.e., it must be precise enough to be implementable also by different laboratories. It should not use ingredients or techniques that are not available anywhere else. When followed in a different lab with the same sample, it should produce essentially the same results; (2) adaptable, i.e., it should also work for slightly changing circumstances, e.g., smaller variations in sample extraction. It should exhibit a certain tolerance to the amount of available material, for instance by prescribing relationships between the sample weight and the volumes of chemicals to be used for the measurements; (3) dependable, i.e., it should have built-in controls to detect failures or wrong measurements. It could for instance, measure expression values using multiple probes, such that a divergence of results hint to problems in the experiment, or it could include synthetic probes that, when creating a signal, indicate the presence of problems.

2 FONDA Core Concepts

Data analysis is a broad term that generally refers to methods for analyzing some data sets to answer a certain question. In scientific data analysis, the data sets typically are generated by experiments and are analyzed to research a given scientific problem. If the data sets are either very big, arrive at the point of analysis with high frequency, or are highly heterogeneous, we denote the analysis as “large-scale”. Extreme-scale problems cope with data in the petabyte range, such as in High-Energy Physics, and are often CPU-heavy and challenging to parallelize. However, there also exists a large and quickly growing number of problems whose study requires an analysis that is IO-heavy and in which large parts, often especially the data intensive ones, are rather simple (”embarrassingly”) to parallelize [1]. The scale of these problems still calls for distributed resources, while their complexity requires the combination of a large number of different base programs into structured DAWs. DAWs may take the form of linear pipelines or form complex producer-consumer networks. In many applications, they even include iterative or recursive structures. Figure 1 shows two exemplary (yet idealized) DAWs.

2.1 Scientific Data Analysis at Large

Scientific data analysis has a number of properties that makes this field particularly and increasingly dependent on technologies for the efficient creation and adaptation of large-scale DAWs.

First, the analyses being performed change rapidly as research progresses, novel types of experiments be-

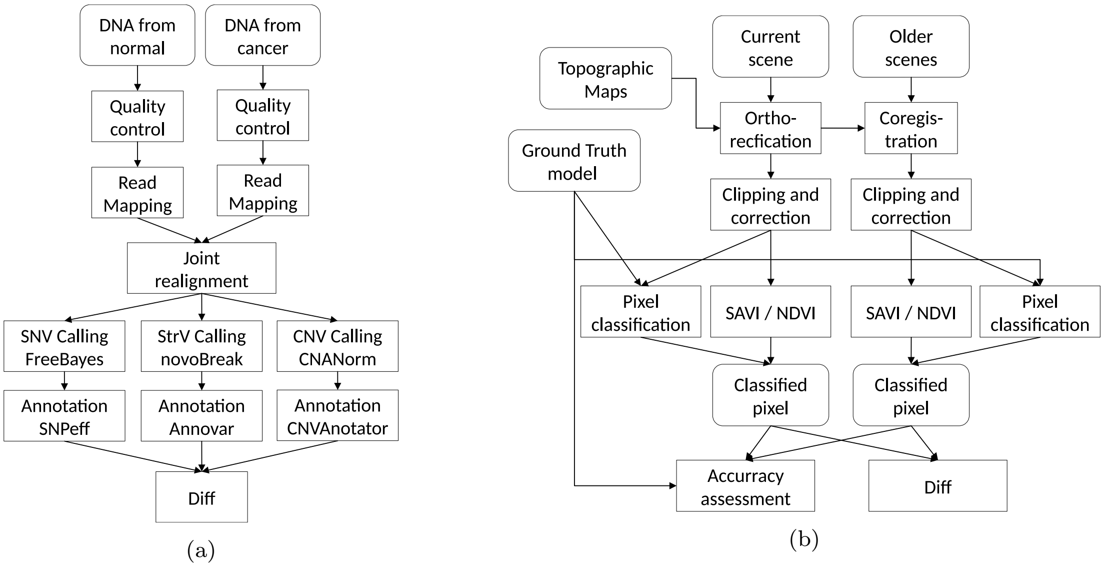

Fig. 1: Two exemplary DAWs (informal representations). Rounded boxes denote input data, square boxes computational tasks, edges the flow of data between computations. (a) DAW for somatic variant calling in cancer (inspired by `https://wabi-wiki.scilifelab.se`). The DAW takes two DNA read sets as input, aligns them to a reference genome, identifies and annotates different types of genomic variations in parallel, and eventually identifies cancer-specific variations. (b) DAW for land cover studies based on satellite images (inspired by [5]). Current and older images are aligned, clipped, corrected for different forms of noise, and classified at a per-pixel basis using different methods in parallel. Results are quality-controlled and compared to detect differences. NDVI - Normalized Difference Vegetation Index, SAVI - Soil-Adjusted Vegetation Index.

come available, and new research questions emerge. A typical scientific DAW is not executed as-is over long periods of time; instead, DAWs are continuously adapted (slowly or disruptively) [2]. For instance, using the DAW from Figure 1a) for DNA from other species, for DNA produced by different sequencing machines, or for searching other types of variations would require changing the DAW by replacing, adding, or deleting certain steps. Therefore, it is important to support the adaptation of DAWs to changing circumstances.

Second, DAWs are often developed for data and research problems studied by many groups in the world (e.g., [6,12]). Furthermore, for every step in a given DAW there often exist multiple tools with comparable functionality. The typical task of DAW designers is to find the best ways to connect and customize such tools. However, as all these tools come with their own implicit pre-requisites on formats, size of input data, available execution infrastructure, etc., it is important that developers have a means to lift these restrictions at the DAW level to create dependable, easy to (re)use DAWs.

Third, DAWs are often exchanged or shared between groups [4,11]. This sharing is important for many reasons, such as saving time and money or reproduction and validation of results. However, sharing is challenging when distributed infrastructures are involved, as different labs or sites virtually never have the same cluster hardware operated by the same software stack. Thus, it is very important that DAWs are portable to different infrastructures.

2.2 Architectures for Data Analysis Workflow Systems

To specify and execute a DAW on a distributed system, a number of components need to exist (see Figure 2). The most important ones are: A distributed file system (DFS, different nodes may access the same data), a resource manager (RM, to run tasks on different nodes and manage available resources), a runtime environment (RE, programs/tools must be installed prior to usage), a scheduler (to decide which program to start when on which node), a specification language (SL, to specify the analysis and interdependencies of its parts), and an execution engine (EE, to execute a DAW specification by steering the other components). There exists a variety of architectures for implementing these abstract components in real systems [7,3], and they very often result in strong mutual dependencies. For instance, resource managers often incorporate their own schedulers, and DAW languages are often bound to a specific exe-

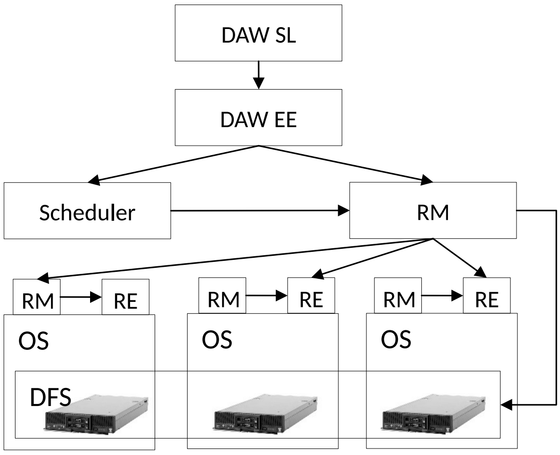

Fig. 2: Components of a distributed DAW infrastructure (highly idealized). Abbreviations are explained in the text. For simplicity, the figure disregards many alternatives and extensions, such as virtualization, multicore nodes, or network channels.

cution engine and vice versa. A prominent example of such an infrastructure is a Hadoop cluster [9].

The development of such infrastructures or infrastructure components requires large efforts, often taking hundreds or thousands of person years. Data centers running an established infrastructure are very reluctant to change it, as this typically affects many aspects of systems operations (accounting, responsibilities, backup, etc.) and puts running applications at risk. Therefore, FONDA does not aim to develop entirely new DAW systems or components thereof. Instead, it pursues its research goals by enhancing existing systems, for instance by exchanging scheduling algorithms, by factoring out new architectural components, or by extending specification languages.

3 FONDA Structure and Vision

FONDA encompasses 11 research subprojects supported by one infrastructure project and an Integrated Research Training Group. Primarily, research subprojects are structured according to the component of a DAW infrastructure they address: Specification language, execution engine, or computational infrastructure. These are grouped in two research areas (see Figure 2), augmented by two auxiliary areas:

- Area A, “Abstractions for DAW Specification Languages“, targets PAD properties of DAW specification languages and execution engines from the viewpoint of the language layer.
- Area B, “Abstractions for DAW Execution Infrastructures”, targets PAD properties of execution engines and computational infrastructures for DAWs from the viewpoint of the infrastructure layer.
- Auxiliary Area S (service) consists of two service projects, one for providing testbeds, benchmarks, and repositories, and one for organizing an integrated Research Training Group for all FONDA PhD students.
- Auxiliary Area T (teams) is devoted to teams (see below).

3.1 Teams

Teams are a special structural element of FONDA. Each team has a specific thematic focus that represents an overlap in the research interests of several of FONDA’s subprojects. Within the team, this overlap is addressed in a joint manner to exploit synergies, to reduce redundant work, and to avoid heterogeneity of solutions. The teams thus also achieve a continuous synthesis of results from multiple subprojects and perform preparatory work for future phases of FONDA. Four of the five teams (foundations of DAW specifications; provenance management; validity constraints; visual monitoring) develop a software library, whereas a fifth team creates a comprehensive benchmark of scientific DAWs including different input data sets.

Teams are organized as virtual research subprojects. As all other subprojects, they have a project leader, a work program, a time schedule, and a set of assigned resources. However, the resources of a team are not requested as independent funding; instead, all subprojects participating in a team delegate staff partly to a team and foster synergies of subprojects. As the total amount of work planned within a team is only between 24 and 30 person months and since teams are staffed from five to seven subprojects, the overall load per participating subproject is moderate. At the same time, each participating subproject (and many other of FONDA’s subprojects) directly benefits from the work done in a team, as it, by design, works on a topic of high relevance for the subproject itself.

3.2 Interdisciplinarity

FONDA’s central aim is the investigation of new Computer Science methods, tools, and systems to support urgent needs in the Natural Sciences. We are convinced that such an interdisciplinary setting can only be approached successfully by an equally interdisciplinary team. Therefore, 2/3 of the 19 Principal Investigators

(PIs) of FONDA are Computer Scientists by education, while 1/3 work as Natural Scientists in the Life Sciences (genomics and biomedical image analysis), the Geosciences (remote sensing), and Physics (material science). These were selected mostly based on their direct dependence in current research on the analysis of large and complex data sets, the possibility to exploit synergies in terms of the types of data that are analyzed, and the free availability of large data sets at project start. Also important were established co-operations between the fields and a critical mass of excellent research groups in the Berlin/Brandenburg area. The integration of Computer Science and Natural Sciences already happens at the subproject level. Six of the 11 subprojects in FONDA are interdisciplinary, i.e., they have one PI from Computer Science and one from a Natural Science. We find such close cooperations indispensable for ensuring the suitability of solutions and for fostering adoption of new methods.

3.3 Long-Term Vision

Intuitively speaking, FONDA aims to lay the scientific basis for the development of Integrated Development Environments (IDE) for large-scale scientific data analysis. Such an IDE should include support for multiple powerful languages and methods to develop, share, and maintain DAWs, should have a plug-in architecture for solutions to requirements of different scientific domains, should contain built-in features for DAW execution, validation, and debugging, and should feature configurable deployment methods over a wide range of computational infrastructures. It should be so easy to use that a typical data scientist from whatever scientific domain can apply it to solve efficiently the in-silico parts of her respective research question also on very large data sets. Such tool support will be indispensable to speed-up the data-driven part of the Natural Sciences and beyond.

4 Lessons Learned

Computer Science research in Germany rarely organizes in Collaborative Research Projects (compared to disciplines like Physics, Medicine, Biology, or Engineering subjects), although these are among the largest types of projects funded by the German Research Foundation. We do not want to speculate why this is the case; however, we will in the following report on a few lessons learned during the onset of FONDA that might help to overcome the barrier for starting such initiatives in other places of Germany as well.

First, it takes an astonishing amount of time. Discussions on FONDA started in October 2016, almost four years before the project eventually started. The two-phase review process is only one reason for this fact. It also requires considerable effort with numerous meetings, workshops, draft papers, etc. to create a strong and unified team that gathers behind a common idea and to convince university bodies to support it in an adequate manner. Finally, it also requires significant time to write, improve, review, re-write, etc. the grant text. The page limit of DFG for CRC proposals is 400 pages, and FONDA almost reached this limit. This grant text, however, also builds a central backbone of a CRC and eases the establishment of communication structures, organization, activities, and monitoring of the running project.

Second, a careful process for designing the overall project goals and for agreeing on central concepts is of utmost importance. Projects of this size probably cannot be created in a purely top-down fashion, following a single, homogeneous, and undisputed plan. Instead, they emerge from discussions of many interested and dedicated people around a common theme. This process also involves many excellent researchers entering and leaving the team, when it turns out that the project focus moves into a direction not suitable for their individual idea. Clear communication and transparent election processes are important to avoid disappointments as much as possible. For instance, FONDA had two internal rounds of proposing, evaluating, and selecting subprojects at an increasingly more elaborated state, starting from single page sketches to six page drafts. This process was accompanied by two two-day workshops and numerous meetings in smaller groups to find and discuss ideas. Altogether, FONDA discussed roughly 20 ideas for subprojects involving roughly 30 researchers, of which eventually 12 subprojects and 21 researchers were selected for the final proposal.

Third, it requires a critical mass to start an endeavor of the size of a CRC. Here, Berlin is in an excellent position through its three universities (plus Potsdam nearby) and the rich and large set of non-university research institutions. Researchers from these different institutions have worked closely together in the past in various occasions, like common DFG-funded Research Training Groups and Research Units, joint coordinated BMBF projects, Helmholtz-funded structured research initiative, and the Einstein Center for the Digital Future supported by the Berlin senate. The recently established Berlin University Alliance, funded by the “Exzellenzstrategie” of the German government, gave a further push to university-spanning projects. This by no means implies that it takes a city of the size of Berlin to establish a CRC – many CRCs emerge from a single university (e.g., CRC 912, "Highly Adaptive Energy-Efficient Computing" in Dresden, or CRC 901, "On-The-Fly Computing" in Paderborn). Nevertheless, we believe it is very difficult to find a common topic among a group of researchers to which all can subscribe with enthusiasm when there is only a limited amount of candidates available, such as a single faculty. It is much easier to find such a group when candidates may come from different organizations, and universities.

Fourth, it is vital that appropriate supporting structures are built and involved early on. FONDA over the entire time was supported by two secretaries, one part time researcher, and dedicated staff at the SFB department. We early on established a so-called "inner circle" of five PIs that discussed and took strategic decisions regarding funding structure, selection of subprojects, and preparation of reviews. Both the pre-proposal and the final proposal were reviewed externally by three colleagues who invested considerable time to provide helpful feedback.

Fifth, one should have a very good understanding of the DFG review process, its different phases and persons involved, and the time framework it obeys on any applicant. DFG offers great support during this process; already in the first meeting, based on a rough draft and not yet involving scientific reviewers, feedback was intensive and important. The pre-proposal ("Skizze") was reviewed by six colleagues who gave detailed and constructive feedback which led to a multitude of changes both in the CRC structure, the participating PIs, the structure of the text, and the overall goals. It is important to be prepared for and embrace such feedback to make the proposal stronger.

Sixth, one should always keep in mind that, no matter how excellent the specific research ideas might appear to the PIs and their communities, any CRC proposal must convince three groups of reviewers: In the pre-proposal phase, at the on-site review of the full proposal, and the grants committees for CRCs at the DFG senate. These groups are increasingly large and increasingly diverse regarding their scientific backgrounds. This is especially the case for interdisciplinary projects like FONDA; however, the final decision is taken by the committee at the DFG senate, which is composed of researchers of all disciplines. It is a delicate process to consider this heterogeneity during the writing of the grant proposal. Parts of a CRC proposal must be convincing for the very expert, while other parts must also be appealing to people from completely different disciplines. Finding the right balance for the different parts of a proposal is a challenge.

Eventually, we want to confirm that all the effort is worth it. A CRC not only requires a critical mass to be formed; it also creates a critical mass to tackle research problems at a depth and breadth that is impossible to achieve in smaller settings. We are very much looking forward to the interesting results this will yield in the future.

Acknowledgements FONDA is supported by the German Research Foundation (DFG) as CRC 1404. We acknowledge substantial contributions from Humboldt-Universität zu Berlin, Technische Universität Berlin, the Max-Delbrück-Center for Molecular Medicine, and Charité - Universitätsmedizin Berlin.

We are grateful to all excellent and inspiring colleagues that helped over more than two years in shaping our plans. We also thank the experts involved in reviewing FONDA at the different stages for their constructive and always fair evaluations.

References

1. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Communications of the ACM **51**(1) (2008)
2. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-science: An overview of workflow system features and capabilities. Future Generation Computer Systems **25**(5), 528–540 (2009)
3. Deelman, E., Peterka, T., Altintas, I., Carothers, C.D., van Dam, K.K., Moreland, K., Parashar, M., Ramakrishnan, L., Taufer, M., Vetter, J.: The future of scientific workflows. Int. J. of High Performance Computing Applications (2017)
4. Goderis, A., De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Fisher, P., Michaelides, D., Tanoh, F.: Discovering scientific workflows: the myexperiment benchmarks. IEEE Transactions on Automation Science and Engineering (2008)
5. Goerner, A., Gloaguen, R., Makeschin, F.: Monitoring of the ecuadorian mountain rainforest with remote sensing,. Journal of Applied Remote Sensing **1**(1) (2007)
6. Leipzig, J.: A review of bioinformatic pipeline frameworks. Brief Bioinform **18**(3), 530–536 (2017)
7. Ling, L.: Computing infrastructure for big data processing. Frontiers of Computer Science **7**(2) (2013)
8. Liu, J., Pacitti, E., Valduriez, P., Mattoso, M.: A survey of data-intensive scientific workflow management. Journal of Grid Computing **1-37** (2015)
9. Ren, K., Kwon, Y., Balazinska, M., Howe, B.: Hadoop's adolescence: An analysis of hadoop usage in scientific workloads. PVLDB (2013)
10. da Silva, R., Filgueira, R., Pietri, I., Jiang, M., Sakellariou, R., Deelman, E.: A characterization of workflow management systems for extreme-scale applications. Future Generation Computer Systems (2017)
11. Starlinger, J., Brancotte, B., Cohen-Boulakia, S., Leser, U.: Similarity search for scientific workflows. PVLDB (2014)
12. Zhu, Z.: Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. SPRS Journal of Photogrammetry and Remote Sensing **130**(370-384) (2017)