

Scientific Data Analysis Using Data-Intensive Scalable Computing: the SciDISC Project*

Patrick Valduriez¹, Marta Mattoso², Reza Akbarinia¹, Heraldo Borges³,
José Camata², Alvaro Coutinho², Daniel Gaspar⁴, Noel Lemus⁴, Ji Liu¹,
Hermano Lustosa⁴, Florent Masegla¹, Fabricio Nogueira da Silva⁵, Vítor Silva²,
Renan Souza², Kary Ocaña⁴, Eduardo Ogasawara³, Daniel de Oliveira⁵,
Esther Pacitti¹, Fabio Porto⁴, Dennis Shasha⁶

¹Inria, LIRMM and University Montpellier - France

²COPPE/UFRJ, Rio de Janeiro – RJ – Brazil

³CEFET/RJ, Rio de Janeiro – RJ – Brazil

⁴LNCC, Rio de Janeiro – RJ – Brazil

⁵UFF, Rio de Janeiro – RJ – Brazil

⁶NYU, New York – NY – USA

Patrick.Valduriez@inria.fr, marta@cos.ufrj.br

Abstract. *Data-intensive science requires the integration of two fairly different paradigms: high-performance computing (HPC) and data-intensive scalable computing (DISC), as exemplified by frameworks such as Hadoop and Spark. In this context, the SciDISC project addresses the grand challenge of scientific data analysis using DISC, by developing architectures and methods to combine simulation and data analysis. SciDISC is an ongoing project between Inria, several research institutions in Rio de Janeiro and NYU. This paper introduces the motivations and objectives of the project, and reports on the first results achieved so far.*

1. Introduction

Modern science such as astronomy, biology, computational engineering and environmental science must deal with overwhelming amounts of data (*e.g.* coming from sensors and scientific instruments, or produced by simulation). Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore these massive datasets [Hey 2009].

* Invited paper, Latin American Data Science Workshop (LADaS), VLDB 2018, Rio de Janeiro, Brazil.

Such data-intensive science [Critchlow 2013] requires the integration of two fairly different paradigms: high-performance computing (HPC) and data-intensive scalable computing (DISC). HPC is compute-centric and focuses on high-performance of simulation applications, typically using powerful, yet expensive, supercomputers. DISC [Bryant 2011], on the other hand, is data-centric and focuses on fault-tolerance and scalability of web and cloud applications using cost-effective clusters of commodity hardware. Examples of DISC systems include big data processing frameworks such as Hadoop or Apache Spark or NoSQL systems (see [Bondiombouy 2016], which includes a survey of DISC systems). To harness parallel processing, HPC uses a low-level programming model (such as MPI or OpenMP) while DISC relies on powerful data processing operators (Map, Reduce, Filter, *etc.*). Data storage is also quite different: supercomputers typically rely on a shared disk infrastructure and data must be loaded in compute nodes before processing while DISC systems rely on a shared-nothing cluster (of disk-based nodes) and data partitioning.

Spurred by the growing need to analyze big scientific data, the convergence between HPC and DISC has been a recent topic of interest [Coutinho 2014, Valduriez 2015]. However, simply porting the Hadoop stack on a supercomputer [Fox 2016] is not cost-effective, and does not solve the scalability and fault-tolerance issues addressed by DISC. On the other hand, DISC systems have not been designed for scientific applications, which have different requirements in terms of data analysis and visualization.

This international project between Inria (France), several research institutions in Rio de Janeiro (Brazil) and NYU (USA), addresses the grand challenge of scientific data analysis using DISC (SciDISC), by developing architectures and methods to combine simulation and data analysis. We can distinguish between three main approaches depending on where analysis is done [Oldfield 2014]: post-processing, in-situ and in-transit. Post-processing analysis performs analysis after simulation, *e.g.* by loosely coupling a supercomputer and a SciDISC cluster (possibly in the cloud). This approach is the simplest but is restricted to batch analysis. In-situ analysis runs on the same compute resources as the simulation, *e.g.* a supercomputer, thus making it easy to perform interactive analysis. In-transit analysis offloads analysis to a separate partition of compute resources, *e.g.* using a single cluster with both compute nodes and data nodes that communicate through a high-speed network. Although less intrusive than in-situ, this approach requires careful synchronization of simulation and analysis.

In the SciDISC project, we study different architectures for SciDISC and their trade-offs. We address the following main steps of the data-intensive science process: (1) data preparation, including raw data ingestion (*e.g.* from sensors) and data cleaning, transformation and integration; (2) data processing and simulation execution; (3) exploratory data analysis and visualization; (4) data mining, knowledge discovery and recommendation. Note that these steps are not necessarily sequential, for instance, steps 2 and 3 need to be interleaved to perform real time analysis.

The expected results of SciDISC are: new data analysis methods for SciDISC systems; the integration of these methods as software libraries in popular DISC systems, such as Apache Spark; and extensive validation on real scientific applications, by working with our scientific partners such as INRA and IRD in France and Petrobras, the

National Research Institute (INCT) on e-medicine (MACC) and the e-astronomy laboratory LIneA in Brazil.

In the rest of this paper, we report on our first results. Section 2 discusses a generic SciDISC architecture that serves as a basis for developing new distributed and parallel techniques to deal with scientific data analysis. Section 3 deals with interactive analysis of simulation data and visualization. Section 4 addresses data mining of scientific data. Section 5 deals with the use of machine learning for recommendation in SciDISC. Section 6 concludes and gives our future research directions.

2. SciDISC Architecture

The first part of the project has been devoted to the definition of a SciDISC architecture that serves as a basis for developing new distributed and parallel techniques to deal with scientific data. We consider a generic architecture that features a high-performance computer (*e.g.* to perform data processing and simulation) with shared-disk and a shared-nothing cluster to perform data analysis. The high-performance computer can be a supercomputer (*e.g.* LNCC Santos Dumont supercomputer) or a large cluster of compute nodes (*e.g.* Grid5000), which yields different cost-performance trade-offs to be studied. Figure 1 illustrates an infrastructure of in-transit data analysis of simulation data, where simulation and computation of predictions are performed at a supercomputer while the analysis of results to evaluate the simulation quality and interpret the simulated phenomenon is done at a cluster.

This architecture allows us to design generic techniques for data transfer, partitioning and replication, as a basis for parallel data analysis and fault-tolerance in DISC [Liroz-Gistau 2013, Silva 2017, Souza 2017a, 2017b]. Additionally, envisioning an almost real-time data transfer between the HPC system and the analytics platform, an orchestrated and tuned set of components must be devised [Matheus 2018]. Security concerns, for instance, may restrict the exposure of simulation results through a single HPC entry node, which rapidly turns into a bottleneck at the HPC side.

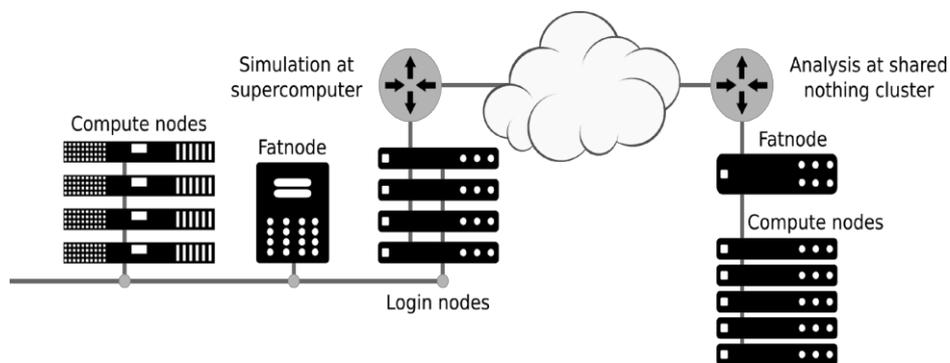


Figure 1. Infrastructure of in-transit data analysis of simulation data

3. From Simulation to Interactive Analysis and Visualization

In complex simulations, users must track quantities of interest (residuals, errors estimates, etc.) to control as much execution as possible. However, this tracking is typically done only after the simulation ends. We are designing techniques to extract, index and relate strategic simulation data for online queries while simulation is running.

We consider coupling these techniques with largely adopted libraries such as libMesh (for numerical solvers) and ParaView (for visualization), so that queries on quantities of interest are enhanced by visualization and provenance data. Interactive data analysis support is planned for post-simulation and runtime as in-situ and in-transit, taking advantage of memory access at runtime.

In [Silva 2017], we propose a solution (architecture and algorithms) to combine the advantages of a dataflow-aware SWMS and the raw data file analysis techniques to allow for queries on raw data file elements that are related but reside in separate files. Armful (<https://hpcdb.github.io/armful/>) is the name of the architecture and its main components are a raw data extractor, a provenance gatherer and a query processing interface, which are all dataflow aware. In [Silva 2017], we instantiate Armful with the Chiron SWMS [Ogasawara 2011]. In [Silva 2018], we remove the SWMS and instantiate Armful as DfAnalyzer, a library of components to support online in-situ and in-transit data analysis. DfAnalyzer components are plugged directly in the simulation code of highly optimized parallel applications with negligible overhead. With support of sophisticated online data analysis, scientists get a detailed view of the execution, providing insights to determine when and how to tune parameters [Souza 2017a, Camata 2018, Silva 2018]. In [Souza 2017b] we evaluate a parameter sweep workflow also in the Oil and Gas domain, this time using Spark to understand its scalability when having to execute legacy black-box code with a DISC system. The source code of the dataflow implementation for Spark is available on github (github.com/hpcdb/RFA-Spark).

We started investigating the combination of in-transit analysis and visualization, with the development of SAVIME (Scientific Analysis and Visualization In-Memory). The system adopts a multi-dimensional data model TARS (Typed Array Schema) [Lustosa 2017] that enables the representation of simulation output data, the topology mesh and simulation metadata. Data produced by the simulation in the HPC is ingested without any transformation as blocks of a Typed Array (TAR) in real-time into SAVIME, running in a Big Data cluster system. The communication between the two systems is implemented using an extended RDMA protocol that bridges the HPC computing nodes memory with a cluster receiver *fatnode* memory. SAVIME offers a set of high-level operators that manipulate in-memory multi-dimensional arrays, split into blocks. Query results can be streamlined into Paraview for visualization in the cluster, saving the HPC system from this extra load.

Finally, we have devised techniques to efficiently assess the uncertainty in simulation's output. Our approach uses probabilistic distribution functions (PDF) to fit the output of a parameter sweep study and replaces data by the best fitting PFD at each point. Next, we may answer uncertainty quantification queries on spatio-time regions of the simulation output using the PDFs instead of the replaced data. The PDF computing strategy has been implemented using different approaches in Apache Spark [Liu 2018].

4. Data Mining of Scientific Data

The current data deluge produced in scientific applications has fostered the development of new knowledge discovery techniques. In this context, an interesting problem raises when the studied phenomenon can be modeled as spatial-time series. The investigation of spatial-time series may shed light on patterns (motifs) [Mueen 2014] and can be used in predicting future series behavior [Dhar 2013]. In this context, we focus on the design of new algorithms to harvest large datasets of space-time series looking patterns that are relevant for the scientific domain studied (seismic, astronomy, and sensor data sources). Such datasets can even appear distributed on different sites [Allard 2015]. This work capitalizes on our previous results in data transformations [Ogasawara 2010] and sequence mining [Campisano 2016].

In [Campisano 2017], we tackle the problem of finding tight space-time sequences, *i.e.*, find within the same process: frequent sequences constrained in space and time that may not be frequent in the entire dataset, and the time interval and space range where these sequences are frequent. The discovery of such patterns along with their constraints may lead to extract valuable knowledge that can remain hidden using traditional methods since their support is extremely low over the entire dataset.

We introduce a new spatiotemporal Sequence Miner (STSM) algorithm to discover tight space-time sequences. We evaluate STSM using a seismic use case and illustrate its ability to detect frequent sequences constrained in space and time. When compared with general spatial-time sequence mining algorithms, STSM allows for new insights by detecting maximal space-time areas where each pattern is frequent. Additionally, in [Cruz 2017], we started studying sensor data sources using spatial-temporal aggregations from trajectories of the buses of Rio de Janeiro. As a preliminary work on this subject, we established a baseline for anomaly identification in urban mobility, which may be useful for developing new approaches that help better discover patterns and understand urban mobility systems.

5. Machine Learning and Recommendation

Scientists commonly explore several input data files and parameter values in different executions of scientific workflows. These workflows can execute for days in DISC environments and they are costly both in terms of execution time and financial cost [Liu 2016, 2017, Pineda-Morales 2016]. It is fundamental that input data files and parameter values chosen for a specific workflow execution do not produce undesired results. In addition, depending on how parameters are set, the workflow execution may present a better performance. Today, scientists spend much time choosing appropriate parameter values and data files based on their experience, but this is an error-prone task since many of these parameters are not independent of each other, *i.e.*, if one parameter is modified, it may imply on changing the value of many other parameters of the workflow. It is worth noticing that this parameter space opens room for parameter fine tuning and consequently improvements both in performance and quality of results. However, due to the (very) large parameter space, this parameter recommendation is an open problem. Our proposal is to use provenance data captured during previous workflow executions to recommend data files and parameters values for future

executions. We use Machine Learning algorithms (ML) [Raedt 2008] to predict which data files and parameters are more suitable for an execution.

We have developed a series of predictive models [Silva Jr 2018] in order to identify which combinations of data files and parameters values produce results with more quality and in less time. We use as input datasets provenance traces from SciPhy (bioinformatics) and Montage (astronomy) workflows (workflows that we have access to specialists that can inform how to measure quality of results). This way, we are able to suggest “ideal” parameter values and data files for scientists that will produce results with more quality and/or less time. These predictive models are based on traditional ML algorithms such as Classification Trees, Support Vector Machines (SVM), One Class SVM and Inductive Logic Programming (ILP). Each predictive model presents different precision and accuracy, and it may be required to choose the best one before recommending parameter values and data files to use. Thus, we plan to use user feedback to fine-tune the recommendation [Servajean 2015], *i.e.*, we have a 2-level recommendation scenario. First, we have to recommend which predictive model to use and then run this model with new data to finally recommend the parameter values and data files for workflow executions. This combination of ML and feedback is novel when compared with existing approaches [Ferro 2011, Huang 2013].

6. Conclusion

The SciDISC project addresses the grand challenge of scientific data analysis using DISC, by developing architectures and methods to combine simulation and data analysis. In this paper, we introduced the motivations and objectives of the project, and reported on the first results achieved so far in terms of generic architecture, interactive analysis of simulation data and visualization, data mining of scientific data, and machine learning and recommendation.

The first results are quite encouraging and lead to exiting future work. Based on in-situ data extraction and analysis, we plan to improve our dataflow monitoring, debugging and extend our support for adaptation at runtime like parameter fine-tuning and data reduction. We will also continue the development of the SAVIME system. The aim is to compute almost in real-time simulation output analysis and ready to be consumed visualization output. Regarding post-processing of simulation data, we will continue to study Spark, one of the most popular DISC systems, and explore it as a platform for efficiently computing probability distribution functions on numerical simulation output during a parameter sweep exploration. We will pursue our work on data mining of spatial-time series in two main areas: compare motif identification techniques with sequence mining techniques and explore spatial-temporal aggregation techniques of sensor data to enable spatiotemporal pattern mining. Regarding ML and recommendation, we have developed a series of predictive models to suggest parameter values and data files for workflow executions. Since these models present different accuracy and precision, it may be difficult to choose a specific model to predict parameters and data files. Thus, we propose to develop a recommendation system that will allow for users to choose the best predictive model based on opinions of colleagues and other users, and on the performance of such predictive models on previous recommendations. This recommendation process is being implemented within the SciManager system (www.scimanager.ic.uff.br).

7. Acknowledgements

This work was partially funded by CNPq, FAPERJ and Inria (SciDISC project), EU H2020 Programme and MCTI/RNP-Brazil (HPC4E grant no. 689772), and performed (for Inria) in the context of the Computational Biology Institute (www.ibc-montpellier.fr). The experiments in SciDISC are carried out using the Inria Grid'5000 testbed (www.grid5000.fr), NACAD/COPPE supercomputers and LNCC SINAPAD Santos Dumont supercomputer (sdumont.lncc.br).

References

- T. Allard, G. Hébrail, F. Masegla, E. Pacitti. Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering. SIGMOD Conference, 779-794, 2015.
- C. Bondiombouy, P. Valduriez. Query Processing in Multistore Systems. *Int. Journal of Cloud Computing*, 5(4): 309-346, 2016.
- R. Bryant. Data-Intensive Scalable Computing for Scientific Applications. *Computing in Science & Engineering*, 13(6):25-33, 2011.
- J. Camata, V. Silva, P. Valduriez, M. Mattoso, A. Coutinho. In Situ Visualization and Data Analysis for Turbidity Currents Simulation. *Computers & Geosciences*, 110, pp.23-31, 2018.
- R. Campisano, F. Porto, E. Pacitti, F. Masegla, and E. Ogasawara. Spatial Sequential Pattern Mining for Seismic Data. SBBB Conference, 2016.
- R. Campisano, Sequence Mining in Spatial-Time Series (Master Degree Dissertation), CEFET/RJ, 2017.
- A. Coutinho. Computational Science and Big Data: Where are We Now? XLDB Workshop, <http://xldb-rio2014.linea.gov.br/program>, 2014.
- T. Critchlow, K. Kleese van Dam. *Data-Intensive Science*. Chapman and Hall/CRC, 2013.
- A.B. Cruz, J. Ferreira, B. Monteiro, R. Coutinho, F. Porto, E. Ogasawara, Detecção de Anomalias no Transporte Rodoviário Urbano, Brazilian Symposium on Databases (SBBB), 2017.
- V. Dhar, Data Science and Prediction. *Comm. of ACM*, 56(12):64-73, 2013.
- M. Ferro, A. R. Mury, and B. Schulze. A proposal to apply inductive logic programming to self-healing problem in grid computing: How will it work? *Concurrency and Computation Practice and Experience*, 23: 2118-2135, 2011.
- G. Fox, J. Qiu, S. Jha, S. Ekanayake, S. Kamburugamuve. Big Data, Simulations and HPC Convergence. <https://www.researchgate.net/publication/301231174>, 2016.
- D. Gaspar, F. Porto, R. Akbarinia, E. Pacitti, TARDIS: Optimal Execution of Scientific Workflows in Apache Spark. *Int. Conf. on Big Data Analytics and Knowledge Discovery (DaWaK)*, 74-87, 2017.
- T. Hey. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
- X. Huang, T. Lu, X. Ding, and N. Gu. Enabling Data Recommendation in Scientific Workflow Based on Provenance. 8th China Grid Annual Conference 1-8, 2013.
- A. Khatibi, F. Porto, J. Rittmeyer, E. Ogasawara, P. Valduriez, D. Shasha. Pre-processing and Indexing Techniques for Constellation Queries in Big Data. *Int. Conf. on Big Data Analytics and Knowledge Discovery (DaWaK)*, 164-172, 2017.
- M. Liroz-Gistau, R. Akbarinia, E. Pacitti, F. Porto, P. Valduriez. Dynamic Workload-based Partitioning Algorithms for Continuously Growing Databases. *Trans on Large-Scale Data and Knowledge-Centered Systems*, Springer, 12:105-128, 2013.
- J. Liu, E. Pacitti, P. Valduriez, D. de Oliveira, M. Mattoso. Multi-Objective Scheduling of Scientific Workflows in Multisite Clouds. *Future Generation Computer Systems*, Elsevier, 63: 76-95, 2016.

- J. Liu, E. Pacitti, P. Valduriez, M. Mattoso. Scientific Workflow Scheduling with Provenance Data in a Multisite Cloud. *Trans. on Large-Scale Data- and Knowledge-Centered Systems (TLDKS)*, 33: 80-112, 2017.
- J. Liu, E. N. Lemus, E. Pacitti, F. Porto, P. Valduriez, Parallel Computation of PDFs on Big Spatial Data Using Spark, arXiv:1805.03141, 2018
- H. Lustosa, F. Porto, N. Lemus, P. Valduriez, TARS: Na Extension of the Multi-dimensional Array Model, *ER FORUM – Conceptual Modeling: Research In Progress*, Valencia, 2017.
- A. Matheus, H. Lustosa, F. Porto, B. Schulze, Towards In-transit Analysis on Supercomputing Environments, arXiv:1805.06425, 2018.
- A. Mueen. Time series motif discovery: Dimensions and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(2):152-159, 2014.
- E. Ogasawara, L.C. Martinez, D. de Oliveira, G. Zimbrao, G.L. Pappa, and M. Mattoso. Adaptive Normalization: A novel data normalization approach for non-stationary time series. *Int. Joint Conf. on Neural Networks (IJCNN)*, 1-8, 2010.
- E. Ogasawara, J. Dias, D. Oliveira, F. Porto, P. Valduriez, M. Mattoso. An Algebraic Approach for Data-centric Scientific Workflows. *Proceedings of the VLDB Endowment (PVLDB)*, 4(12):1328-1339, 2011.
- R. Oldfield, K. Moreland, N. Fabian, D. Rogers. Evaluation of Methods to Integrate Analysis into a Large-Scale Shock Physics Code. *ACM Int. Conf. on Supercomputing*, 83-92, 2014.
- T. Özsu, P. Valduriez. *Principles of Distributed Database Systems – Third Edition*. Springer, 850 p, 2011.
- E. Pacitti, R. Akbarinia, M. El Dick: P2P Techniques for Decentralized Applications. *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2012.
- L. Pineda-Morales, J. Liu, A. Costany, E. Pacitti, G. Antoniu, P. Valduriez, M. Mattoso. Managing hot metadata for scientific workflows on multisite clouds. *IEEE BigData Conf*, 390-397, 2016.
- F. Porto, A. Khatibi, J. Nobre, E. Ogasawara, P. Valduriez, D. Shasha. Point Pattern Search in Big Data. *Int. Conf. on Scientific and Statistical Database Management (SSDBM)*, 2018.
- L. D. Raedt. *Logical and Relational Learning: From ILP to MRDM (Cognitive Technologies)*. Springer, New York, 2008.
- M. Servajean, R. Akbarinia, E. Pacitti, S. Amer-Yahia. Profile Diversity for Query Processing using User Recommendations. *Information Systems*, 48: 44-63, 2015.
- V. Silva, J. Leite, J. Camata, D. de Oliveira, A. Coutinho, P. Valduriez, M. Mattoso. Raw data queries during data-intensive parallel workflow execution. *Future Generation Computer Systems*, Elsevier, 75: 402-422, 2017.
- V. Silva, D. de Oliveira, P. Valduriez, M. Mattoso. DfAnalyzer: Runtime Dataflow Analysis of Scientific Applications using Provenance, *Proceedings of the VLDB Endowment*, 2018.
- D. Silva Jr., A. Paes, E. Pacitti, D. Oliveira. Data Quality Prediction in Scientific Workflows. In preparation, 2018
- R. Souza, V. Silva, P. Miranda, A. Lima, P. Valduriez, M. Mattoso. Spark Scalability Analysis in a Scientific Workflow. *Brazilian Symposium on Databases (SBBDB)*, Best Paper Award, 2017.
- R. Souza, V. Silva, J. Camata, A. Coutinho, P. Valduriez, M. Mattoso. Tracking of Online Parameter Fine-tuning in Scientific Workflows. *Workshop on Workflows in Support of Large-Scale Science (WORKS)*, ACM/IEEE Supercomputing Conference, 2017.
- P. Valduriez. Data-intensive HPC: opportunities and challenges. *Big Data and Extreme-scale computing (BDEC)*, <http://hal-lirmm.ccsd.cnrs.fr/lirmm-01184018>, 2015.