# Linking provenance with system logs: a context aware information integration and exploration framework for analyzing workflow execution

Elias el Khaldi Ahanach, Spiros Koulouzis, Zhiming Zhao,
Informatics Institute,
University of Amsterdam,
Amsterdam, The Netherlands,
elias.el.khaldi@gmail.com, {S.Koulouzis|Z.Zhao}@uva.nl

## ABSTRACT

When executing scientific workflows in a distributed environment, anomalies of the workflow behavior are often caused by a mixture of different issues, e.g., careless design of the workflow logic, buggy workflow components, unexpected performance bottlenecks or resource failure at the underlying infrastructure. The provenance information only defines data evolution at the workflow level, which does not have an explicit connection with the system logs provided by the underlying infrastructure. Analyzing provenance information and apposite system metrics requires expertise and a considerable amount of manual effort. Moreover, it is often time-consuming to aggregate this information and correlate events occurring at different levels in the infrastructure. In this paper, we propose an architecture to automate the integration among the workflow provenance information with the performance information collected from infrastructure nodes running workflow tasks. Our architecture enables workflow developers or domain scientists to effectively browse workflow execution information together with the system metrics, and analyze contextual information for possible anomalies.

## REFERENCES

[1] L. Candela, D. Castelli, and P. Pagano, "Virtual research environments: an overview and a research agenda," *Data Science Journal*, vol. 12, no. 0, pp. GRDI75–GRDI81, 2013.

[2] Z. Zhao, A. Belloum, C. De Laat, P. Adriaans, and B. Hertzberger, "Distributed execution of aggregated multi domain workflows using an agent framework," *Services, 2007 IEEE Congress on*, pp. 183–190, 2007.

[3] M. A. Miller, W. Pfeiffer, and T. Schwartz, "The cipres science gateway: enabling high-impact science for phylogenetics researchers with limited resources," *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*, p. 39, 2012.

[4] S. Koulouzis, A. S. Belloum, M. T. Bubak, Z. Zhao, M. ivkovi, and C. T. de Laat, "Sdn-aware federation of distributed data," *Future Generation Computer Systems*, vol. 56, pp. 64 – 76, 2016.

[5] K. Evans, A. Jones, A. Preece, F. Quevedo, D. Rogers, I. Spasic,´ I. Taylor, V. Stankovski, S. Taherizadeh, J. Trnkoczy, G. Suciu, V. Suciu, P. Martin, J. Wang, and Z. Zhao, "Dynamically reconfigurable workflows for time-critical applications," *Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science*, pp. 7:1–7:10, 2015.

[6] P. Groth and L. Moreau, "Prov-overview. an overview of the prov family of documents," 2013.

[7] R. Cushing, S. Koulouzis, A. Belloum, and M. Bubak, "Applying workflow as a service paradigm to application farming," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 6, pp. 1297–1312, 2014.

[8] R. F. da Silva, R. Filgueira, I. Pietri, M. Jiang, R. Sakellariou, and E. Deelman, "A characterization of workflow management systems for extreme-scale applications," *Future Generation Computer Systems*, vol. 75, pp. 228–238, 2017.

[9] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. De Laat, "Addressing big data challenges for scientific data infrastructure," in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, pp. 614–617, IEEE, 2012.

[10] S. Koulouzis, D. Vasyunin, R. Cushing, A. Belloum, and M. Bubak, "Cloud data federation for scientific applications," in *Euro-Par 2013: Parallel Processing Workshops* (D. an Mey, M. Alexander, P. Bientinesi, M. Cannataro, C. Clauss, A. Costan, G. Kecskemeti, C. Morin, L. Ricci, J. Sahuquillo, M. Schulz, V. Scarano, S. L. Scott, and J. Weidendorfer, eds.), (Berlin, Heidelberg), pp. 13–22, Springer Berlin Heidelberg, 2014.

[11] R. Prodan and T. Fahringer, "Overhead analysis of scientific workflows in grid environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, pp. 378–393, March 2008.

[12] I. Foster, "Globus toolkit version 4: Software for service-oriented systems," *Journal of computer science and technology*, vol. 21, no. 4, p. 513, 2006.

[13] R. Ferreira da Silva, T. Glatard, and F. Desprez, "Self-healing of operational workflow incidents on distributed computing infrastructures," in *2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, pp. 318–325, May 2012.

[14] S. Madougou, S. Shahand, M. Santcroos, B. van Schaik, A. Benabdelkader, A. van Kampen, and S. Olabarriaga, "Characterizing workflowbased activity on a production e-infrastructure using provenance data," *Future Generation Computer Systems*, vol. 29, no. 8, pp. 1931 – 1942, 2013. Including Special sections: Advanced Cloud Monitoring Systems & The fourth IEEE International Conference on e-Science 2011 eScience Applications and Tools & Cluster, Grid, and Cloud Computing.

[15] P. Gaikwad, A. Mandal, P. Ruth, G. Juve, D. Krl, and E. Deelman, "Anomaly detection for scientific workflow applications on networked clouds," in *2016 International Conference on High Performance Computing Simulation (HPCS)*, pp. 645–652, July 2016.

[16] "cadvisor (container advisor), official github page." https://github.com/google/cadvisor. Accessed: 2019-03-28.

[17] "Prometheus, an open-source systems monitoring and alerting toolkit.." https://prometheus.io/docs/introduction/overview/. Accessed: 2019-0328.

[18] G. M. Kurtzer, "Singularity 2.1. 2-linux application and environment containers for science, 2016," *Available from Internet:¡ https://doi. org/10.5281/zenodo*, vol. 60736.