4. CONCLUSÃO

Os *workflows* científicos apresentam-se como uma abordagem fundamental para a modelagem de experimentos científicos baseados em simulações. A gerência da proveniência desses *workflows* é uma questão fundamental, pois além da análise do experimento, auxilia a reprodução, peça fundamental do processo científico. Além da reprodução, os dados de proveniência podem ajudar a máquina de execução do workflow no escalonamento, detecção de falhas, dimensionamento do ambiente, etc., pois contêm informações acerca das atividades executadas, as diferentes ativações utilizadas, os tempos de execução, os resultados gerados, possíveis erros ocorridos dentre outros. Obter essas informações traz grandes benefícios para os cientistas, pois a partir delas é possível abortar uma execução que não caminha na direção almejada e, em caso de sucesso, reproduzir a execução de um *workflow*, além de realizar um melhor gerenciamento dos recursos necessários e análises de desempenho, bem como detectar erros com maior facilidade e em tempo real. Como os dados de proveniência são utilizados frequentemente para tomar decisões em tempo real, é importante ter consultas com um tempo de resposta pequeno, para que essas não afetem o desempenho do *workflow*. Entretanto, as máquinas de *workflows* de hoje utilizam repositórios de proveniência centralizados, o que impõe pontos de falha e problemas de segurança e desempenho. Tendo isso em mente, foi proposto neste artigo uma estratégia de fragmentação e alocação dos dados de proveniência de maneira a minimizar o tempo necessário para que as consultas sejam realizadas. Essa estratégia levou em consideração dois fatores principais: o fato de que as consultas realizadas por um cientista em geral giram em torno do *workflow* em que ele está trabalhando; e que, no geral, os dados gerados por aquele cientista são armazenados nos sítios mais próximos a ele. Assim, foi realizada uma fragmentação horizontal sobre os atributos *tag* e localidade, que representam respectivamente o rótulo de identificação do *workflow* e o local onde o mesmo foi executado. Uma série de consultas, frequentemente utilizadas por cientistas, foi selecionada e avaliada tanto no ambiente distribuído, como no ambiente centralizado. Os resultados obtidos indicam uma melhora significativa no ambiente distribuído em consultas realizadas em um único fragmento, mostrando ganhos significativos sobre o tempo de execução da consulta na base centralizada mesmo considerando sítios distantes fisicamente. Como trabalho futuro será realizado uma análise da fragmentação vertical e híbrida.

REFERENCES

Allen, M., Chapman, A., Blaustein, B., Seligman, L., (2011), "Getting It Together: Enabling Multi-organization Provenance Exchange". In: *TaPP 2011*, Athens, Greece.

Costa, F., Oliveira, D., Ocaña, K., Ogasawara, E., Mattoso, M., (2012), "Enabling Re-Executions of Parallel Scientific Workflows Using Runtime Provenance Data. In: 4th International Provenance and Annotation Workshop"

Costa, F., Silva, V., Oliveira, D., Ocaña, K., Dias, J., Ogasawara, E., Mattoso, M., (2013), "Capturing and Querying Workflow Runtime Provenance with PROV: a Practical Approach". In: *BigProv'13*, Genova, Italy.

Deelman, E., Gannon, D., Shields, M., Taylor, I., (2009), "Workflows and e-Science: An overview of workflow system features and capabilities", *Future Generation Computer Systems*, v. 25, n. 5, p. 528–540.

Dias, J., Ogasawara, E., Oliveira, D., Porto, F., Coutinho, A., Mattoso, M., (2011), "Supporting Dynamic Parameter Sweep in Adaptive and User-Steered Workflow". In: *6th WORKS*, p. 31–36, Seattle, WA, USA.

Freire, J., Koop, D., Santos, E., Silva, C. T., (2008), "Provenance for Computational Tasks: A Survey", *Computing in Science and Engineering, v.10*, n. 3, p. 11–21.

Missier, P., Belhajjame, K., Cheney, J., (2013), "The W3C PROV family of specifications for modelling provenance metadata". In: *Proceedings of the 16th EDBT*, p. 773–776, New York, NY, USA.

Ocaña, K. A. C. S., Oliveira, D., Ogasawara, E., Dávila, A. M. R., Lima, A. A. B., Mattoso, M., (2011), "SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes", BSB 2011: Springer, p. 66–70.

Oliveira, D., Ogasawara, E., Ocaña, K., Baião, F., Mattoso, M., (2011), "An Adaptive Parallel Execution Strategy for Cloud-based Scientific Workflows", *Concurrency and Computation: Practice and Experience*, v. (online)

Özsu, M. T., Valduriez, P., (2011), *Principles of Distributed Database Systems*. 3 ed. New York, Springer.

Vaquero, L. M., Rodero-Merino, L., Caceres, J., Lindner, M., (2009), "A break in the clouds: towards a cloud definition", *SIGCOMM Comput. Commun. Rev.*, v. 39, n. 1, p. 50–55.

Zhou, W., Ding, L., Haerberlen, A., Ives, Z. G., Loo, B. T., (2011), "TAP: Time-aware Provenance for Distributed Systems". In: *Proc. of TaPP'11*, Greece.