30º Simpósio Brasileiro de Banco de Dados

Proposta de Tutorial

Reprodução de Experimentos Científicos: Teoria e Prática

Prof. Ary H. M. Oliveira

Ciência da Computação, Universidade Federal do Tocantins (UFT)
e-mail: aryhenrique@uft.edu.br
Lattes: http://lattes.cnpq.br/2481552882893652

Prof. Daniel de Oliveira

Instituto de Computação, Universidade Federal Fluminense (UFF)

e-mail: danielcmo@ic.uff.br
www.ic.uff.br/~danielcmo
Lattes: http://lattes.cnpq.br/0743793296062293

Profa. Marta Mattoso

COPPE, Universidade Federal do Rio de Janeiro (UFRJ)

e-mail: marta@cos.ufrj.br
www.cos.ufrj.br/~marta
Lattes: http://lattes.cnpq.br/1420784392366957

Junho de 2015

Tipo do Tutorial

O tutorial será introdutório e serão tratados diferentes aspectos da gerência de dados e processos para a reprodução de experimentos científicos computacionais e como a computação em nuvem pode ajudar.

Idioma de Apresentação

O idioma de apresentação será o português, entretanto, os slides estarão em inglês.

Resumo do Tutorial

A capacidade de reproduzir um experimento científico computacional (que chamaremos apenas de experimento) faz parte do princípio básico do método científico. Reproduzir um experimento passa pela tentativa de terceiros em reproduzir os procedimentos e os dados apresentados em um artigo científico. Por meio da reprodução, é possível comparar resultados, e avaliar os métodos empregados, para poder dar credibilidade e confiança à pesquisa científica. Neste tutorial, analisamos desafios em reproduzir resultados de experimentos e as abordagens existentes para apoiar tal reprodução. Apresentamos os conceitos, as principais ferramentas, o uso de nuvens computacionais e discutimos os problemas envolvendo reprodução de experimentos em larga-escala.

Organização do Tutorial

O tutorial está organizado em três partes. A primeira aborda os conceitos principais, a motivação, exemplos e os desafios da reprodução de experimentos científicos na computação. A segunda parte aborda um detalhamento dos conceitos da reprodução de experimentos, caracterizando o ciclo de vida da reprodução de experimentos, envolvendo a relação da publicação de artigos com a reprodução dos resultados publicados e as principais tecnologias que vêm servindo de base para o apoio à reprodução. Finalmente, a terceira parte apresenta as principais ferramentas existentes e mostra interativamente por meio do software *Reproescience*, alvo de desenvolvimento de pesquisas dos autores deste tutorial.

Parte I

a) Conceitos principais: Serão abordados os termos e definições relacionados com a reprodução de experimentos científicos na computação. .
b) Motivação: Serão apresentados exemplos de experimentos científicos, fazendo uso dos conceitos definidos. A motivação será mostrar o porquê de o desenvolvimento de um experimento ter de vir acompanhado de recursos que viabilizem a reprodução do experimento, para que seja possível verificar e validar os resultados gerados.
c) Desafios: Apresentação das dificuldades em prover os recursos de reprodução de resultados e perspectivas de como as áreas de banco de dados e computação em nuvem podem contribuir com soluções centradas em dados e infraestrutura computacional para permitir que um experimento científica seja passível de reprodução.

O progresso da ciência depende da efetiva disseminação e reprodução de pesquisas existentes, porém, é necessário ter acesso ao material usado na produção dos resultados dos experimentos para que seja possível validá-los [6]. A computação vêm auxiliando a ciência aumentando a produtividade dos pesquisadores e a qualidade da sua produção na análise de grande quantidade, complexidade e variedade de dados [7]. Entretanto, para que a comunidade científica possa dar prosseguimento às novas descobertas, tal desenvolvimento deve vir acompanhado de abordagens que viabilizem a reprodução do experimento, para que seja possível verificar e validar os resultados gerados.

Por meio da reprodução é possível comparar resultados, e consequentemente, avaliar os métodos empregados, e com isso, dar credibilidade e confiança à pesquisa científica. O termo reprodução está ligado ao conceito de experimento reprodutível que é a capacidade de se fornecer mecanismos que possibilitem a repetição e a reprodução de um experimento por outro cientista. O experimento reprodutível permite que novos trabalhos sejam produzidos a partir de pesquisas já consolidadas, de forma que uma nova solução seja desenvolvida a partir de um ponto mais avançado. Isto elimina a necessidade de se reconstruir teorias e/ou refazer experimentos já existentes a partir de um ponto inicial para se confirmar (ou refutar) uma determinada hipótese científica. J. Freire et al. [5] definem a reprodução na computação, como: "um experimento computacional desenvolvido no tempo *t* em hardware/sistema operacional *s* utilizando os dados *d* é reprodutível se ele puder ser executado no momento *t'* em um sistema *s'* com os dados *d'*, que são semelhantes (ou, potencialmente, os mesmos) que *d*".

Parte II

a) Experimentos reprodutíveis: Apresentação dos principais conceitos do ciclo de vida da reprodução de resultados de experimentos, descrevendo as características e funções, para orientar na seleção dos mecanismos necessários para armazenar, gerenciar, recuperar e reproduzir os elementos de um experimento.
b) Verificação e validação de resultados: Apresentar as inciativas de submissão de artigos e artigos executáveis.
c) Tecnologias de apoio: Apresentação dos conceitos de workflows científicos, proveniência de dados e o uso de nuvens de computadores.

Um experimento reprodutível é aquele cujo produto final é composto pelo artigo publicado, no formato digital ou físico, acompanhado de todos os objetos de pesquisa utilizados para a sua concepção. Os objetos de pesquisa são compostos pelos diferentes artefatos usados ou gerados através de um experimento científico [1], dentre eles: o artigo científico (manuscrito), hipóteses, conjunto de dados usados e/ou produzidos (entrada, intermediários, finais, metadados e proveniência), anotações e documentação, infraestrutura de *hardware* e *software*, plataforma de software e os parâmetros de configuração. Os objetos de pesquisa são como uma abstração usada para a comunicação, compartilhamento e reuso de resultados de pesquisas [1]. O compartilhamento dos objetos de pesquisa facilita o desenvolvimento da ciência, não apenas no processo de validação dos resultados, mas também pela possibilidade de exploração de novas hipóteses e combinação com outros métodos. A gerência de objetos de pesquisa traz oportunidades de pesquisas e desenvolvimento na área de Bancos de Dados.

A verificação da reprodução de experimentos tem sido alvo de preocupação da comunidade científica e tem sido discutida em diversos periódicos e chamadas de trabalhos em diferentes áreas do conhecimento apoiadas pela computação. Trabalhos submetidos a eventos como o congresso SIGMOD da ACM são rotulados como reprodutíveis ou compartilhados [2]. O objetivo é que os trabalhos reprodutíveis sejam submetidos juntamente com o protocolo e os recursos necessários para reproduzir os resultados obtidos no artigo científico, englobando: (1) o protótipo do sistema, com o código-fonte, arquivos de configuração e o ambiente operacional, (2) o conjunto de experimentos, contendo os sistemas de configuração e inicialização, scripts e os protocolos de avaliação usados para produzir os resultados experimentais, e (3) os scripts necessários para transformar os dados em equações, gráficos e tabelas que são apresentados no artigo. O Grande Desafio de Artigos Executáveis, organizado pela Elsevier foi um evento realizado na forma de uma competição cujo objetivo foi selecionar as melhores abordagens de concepção e gerenciamento de artigos executáveis. Os artigos executáveis são definidos como um objeto digital publicado que possui um manuscrito na forma de artigo que descreve um estudo científico e seus resultados, bem

como todo o código necessário para reproduzir os cálculos e visualizar os resultados obtidos com dados e programas do estudo [10]. O objetivo é aumentar a compreensão e a reprodução das publicações eletrônicas permitindo que os leitores e avaliadores possam interagir, validar e explorar os experimentos [8].

Parte III

a) Estado da Arte: Apresentação das principais abordagens e ferramentas que apoiam a reprodução de experimentos científicos (por ex.: Collage [10], Paper Maché [6] e Reprozip [3]).
b) Parte prática: Exemplo da plataforma de compartilhamento e reprodução de experimentos utilizando a virtualização em nuvens computacionais e a ferramenta *ReproeScience*.

Currículo dos Autores

Ary H. M. Oliveira é professor assistente no Curso de Ciência da Computação da Universidade Federal do Tocantins desde 2.008. Cursa o doutorado no Programa de Engenharia de Sistemas e Computação da COPPE/Universidade Federal do Rio de Janeiro sob orientação da Profa. Marta Lima de Queirós Mattoso e do Prof. Daniel de Oliveira. Seus interesses de pesquisa incluem bancos de dados, computação em nuvem, gerência de *workflows* científicos, paralelismo de dados, bioinformática e mineração de dados.

Daniel de Oliveira é professor do Instituto de Computação da Universidade Federal Fluminense (UFF) desde 2013. Recebeu o grau de Doutor em Engenharia de Sistemas e Computação pela COPPE/UFRJ em 2012 sob a orientação da Profa. Marta Lima de Queirós Mattoso. Seus interesses de pesquisa incluem bancos de dados, computação em nuvem, gerência de *workflows* científicos, paralelismo de dados, bioinformática e mineração de dados. Publicou mais de 70 artigos em periódicos indexados e em congressos nacionais e internacionais. É membro da ACM, IEEE e SBC. Vem publicando com frequência em eventos de prestígio internacional de computação em nuvem como o IEEE *Cloud* e o IEEE *e-Science*, além de ter recebido o prêmio de melhor artigo do 2nd *International Workshop on Cloud Computing and Scientific Applications* (CCSA) realizado em conjunto com o IEEE/ACM *International Symposium on Cluster, Cloud and Grid Computing* (CCGrid 2012).

Marta Mattoso é professora titular no Programa de Engenharia de Sistemas e Computação da COPPE/Universidade Federal do Rio de Janeiro e bolsista de produtividade em pesquisa do CNPq nível 1B. Concluiu o doutorado na COPPE/UFRJ em 1993. Orientou 15 teses de doutorado e 55 de mestrado. Publicou mais de 200 artigos completos em revistas e congressos internacionais e nacionais. Vem participando de Comitês de Programa de congressos nacionais e internacionais, revisora ad-hoc de revistas nacionais e internacionais. É a *Scientific Chair da Provenance Week* 2016, reunindo os eventos IPAW e TaPP. Sócia da SBC, IEEE e ACM. Foi diretora de publicações da SBC no período de 2005 a 2007. Coordena diversos projetos de pesquisa com financiamento do CNPq, Capes/Cofecub, INRIA, FAPERJ e Finep, nas áreas de distribuição e paralelismo em bancos de dados, *workflows* científicos em ambientes de paralelismo e gerência de dados de proveniência. Trabalha de modo interdisciplinar com pesquisadores de áreas como bioinformática e engenharia do petróleo.

Apresentação do Tutorial

A apresentação do tutorial será realizada pelos professores Marta Mattoso (COPPE/UFRJ), Ary H. M. Oliveira (UFT) e Daniel de Oliveira (IC/UFF).

Apoio Audiovisual Requerido

O apresentador necessitará de computador e data show para a apresentação do tutorial. Além disso, será necessária a conexão com a internet para apresentar um repositório de dados na nuvem e os serviços correspondentes.

Principais Referências Bibliográficas

[1] BELHAJJAME, K., ROURE, D. D., GOBLE, C. A. "Research Object Management: Opportunities and Challenges". February 2012.

[2] BONNET, P., MANEGOLD, S., BJORLING, M., et al. "Repeatability and Workability Evaluation of SIGMOD 2011", SIGMOD Rec., v. 40, n. 2, pp. 45-48, set. 2011.

[3] CHIRIGATI, F., SHASHA, D., FREIRE, J. "Packing Experiments for Sharing and Publication". In: Proceedings of the 2013 ACM SIGMOD Internat. Conference on Management of Data, pp. 977-980, 2013.

[4] GOBLE, C. "The Reality of Reproducibility in Computational Science: reproduce? repeat? rerun? and does it matter. Keynotes and Panels", 8th IEEE International Conference on e-Science, v. 327, n. 5964, pp. 415-416, October 2012.

[5] FREIRE, J., BONNET, P., SHASHA, D. "Computational Reproducibility: State-of-the-art, Challenges, and Database Research Opportunities". In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12, pp. 593-596, New York, NY, USA, 2012.

[6] BRAMMER, G. R., CROSBY, R. W., MATTHEWS, S. J., et al. "Paper Mâché: Creating Dynamic Reproducible Science", Procedia Computer Science, v. 4, n. 0, pp. 658-667, 2011.

[7] KARPATHIOTAKIS, M., BRANCO, M. ALAGIANNIS, I, AILAMAKI, A. Adaptive Query Processing on RAW Data”. In: Proocedings of VLDB Endowment, vol. 7, n. 32, p. 1119–1130, sep. 2014.

[8] KOOP, D., SANTOS, E., MATES, P., et al. "A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers", Procedia Computer Science, v. 4, n. 0, pp. 648-657, 2011.

[9] KLINGINSMITH, J., MAHOUI, M., WU, Y. "Towards Reproducible e-Science in the Cloud". In: Cloud Computing Technology and Science, 2011 IEEE Third International Conference on, pp. 582-586, Nov 2011.

[10] NOWAKOWSKI, P., CIEPIELA, E., HAREZLAK, D., et al. "The Collage Authoring Environment", Executable Paper Grand Challenge - ICCS 2011, n. 4, pp. 608-617, 2011.