VI Workshop de Dissertações de Mestrado do PPgSI
2019

Gerenciamento adaptativo de cache para a execução de *workflows* científicos

Autoria de:	Miguel Felipe Silva Vasconcelos			
Orientação de:	Prof. Dr. Daniel Cordeiro			
Coorientação de:				
Linha de pesquisa:	[x] Gestão e Desenvolvimento de Sistemas		[ ] Inteligência de Sistemas	
Área de pesquisa:	[ ] Banco de dados [ ] Gestão de tecnologia da informação	[x] Engenharia de software [ ] Interação humano computador	[ ] Inteligência artificial [ ] Reconhecimento de padrões	[ ] Processamento gráfico [ ] Otimização
Área de aplicação:	[ ] Ambientes corporativos / Processos de negócio [ ] Economia [ ] Jogos / Jogos sérios [ ] Saúde	[ ] Bioinformática [ ] Educação / Educação a distância [ ] Linguística / Língua natural [x] Outra Qual? Computação distribuída de alto desempenho	[ ] Biometria [ ] Governo eletrônico [ ] Quimioinformática	[ ] Dispositivos móveis [ ] Internet / Redes sociais [ ] Robótica [ ] Geral*
Semestre no curso (na data do workshop):	[ ] 2º semestre	[x] 3º semestre	[ ] 4º semestre	[ ] 5º semestre
Qualificação:	[ ] Qualificação já realizada em:		[x] Realização da qualificação planejada para: 01/11/2019	
Defesa:	Prazo máximo para depósito: 01/02/2021		Realização da defesa planejada para: 01/11/2020	
Publicações associadas ao projeto de mestrado:	**Sem publicações até o momento**.			

Resumo do projeto de pesquisa

Contexto:

Estamos na era da e-Ciência, em que o uso de ferramentas computacionais é amplamente disseminado. Uma dessas ferramentas são os fluxos de trabalhos científicos – ou *workflows* científicos – que permitem aos usuários focarem em seus experimentos ao invés de se preocuparem com o gerenciamento dos recursos computacionais.
Em relação ao desenvolvimento de *workflows* científicos, arcabouços para realização de computação distribuída de alto desempenho são amplamente utilizados, como o *Apache Spark*, que permite armazenar resultados de operações intermediárias em memória principal (RAM), obtendo desempenho superior aos arcabouços que só permite persistir em disco, por exemplo o *Apache Hadoop.* Armazenar resultados intermediários em cache é uma estratégia para evitar recálculos, visto que resultados de operações específicas podem ser utilizados inúmeras vezes durante uma execução do *workflow*. Entretanto, a decisão sobre quais operações persistir em memória – armazenar em cache – não é automática, é papel do pesquisador que irá desenvolver o *workflow* científico descobrir a melhor combinação de operações.
Esse é um problema de otimização combinatória, e deve levar em consideração variáveis como o custo de escrita e leitura das operações em cache, qual tipo de cache utilizar – RAM ou disco – e restrições como a memória principal ser um recurso finito e ser compartilhada para o armazenamento de cache e execução de outras operações.

Problema de pesquisa:

Algumas das soluções encontradas até o momento para o problema de decisão de cache usam heurísticas ou fazem uma busca exploratória, utilizando um modelo de custo, mas que potencialmente podem explorar todos os estados possíveis, o que não é computacionalmente viável, dependendo do número de operações que existem na aplicação. Não encontramos estudos que utilizam técnicas de aprendizado de máquina para tentar solucionar esse problema.

Objetivo de pesquisa:

O objetivo principal desse projeto é desenvolver um modelo que consiga identificar combinações de resultados de operações que devem ser armazenadas em cache para o arcabouço computacional *Apache Spark*, utilizado no desenvolvimento de *workflows* científicos, e que resulte em desempenho melhor do que abordagens propostas na literatura.

Caracterização da solução em desenvolvimento:

A hipótese considerada neste trabalho é a de que técnicas de aprendizado de máquina podem ser utilizadas para decidir quais resultados de operações devem ser armazenados em cache, e qual o tipo de cache, para arcabouços de computação distribuída de alto desempenho, utilizados na construção de *workflows* científicos.
Dado que *workflows* científicos são executados inúmeras vezes, pode ser possível utilizar dados de proveniência para descobrir padrões que auxiliem na identificação da melhor combinação de cache.

Fundamentos:

- *Workflows* científicos: Permitem que pesquisadores expressem atividades computacionais em várias etapas de processamento de dados, por exemplo extrair dados de um sensor, executar análises e realizar processamento dos resultados;
- Dados de proveniência: são informações sobre a execução de *workflows* científicos e permitem a reprodutibilidade dos experimentos e reexecução do *workflow* em caso de falha;
- Decisão de cache: problema combinatório de escolher a melhor combinação de operações para serem armazenadas em cache, qual o tipo de cache (RAM ou disco), e que possui restrições como a memória RAM ser finita e seu uso ser compartilhado para o armazenamento de cache e execução de outras operações.

Trabalhos relacionados:

GOTTIN, V. M.; PACHECO, E.; DIAS, J.; CIARLINI, A. E.; COSTA, B.; VIEIRA, W.; SOUTO, Y. M.; PIRES, P.; PORTO, F.; RITTMEYER, J. G. Automatic caching decision for scientific dataflow execution in Apache Spark. In: Proceedings of the 5th ACM SIGMOD Workshop on Algorithms and Systems for MapReduce and Beyond. Houston, TX, Estados Unidos: ACM, 2018. p. 2.
Yang, Z.; Jia, D.; Ioannidis, S.; Mi, N.; Sheng, B. Intermediate data caching optimization for multi-stage and parallel big data frameworks. In: 2018 IEEE 11th International Conference on Cloud Computing (CLOUD). San Francisco, CA , Estados Unidos: IEEE, 2018. p. 277–284. ISSN 2159-6190.
U, E.; SAXENA, M.; CHIU, L. Neutrino: Revisiting memory caching for iterative data analytics. In: 8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 16). Denver, CO, Estados Unidos: USENIX Association, 2016.

Validação:

A avaliação do modelo proposto será intrínseca, e levará em consideração as seguintes métricas:

- utilização de memória RAM
- utilização de CPU
- tempo de execução

- tempo necessário para identificar melhor combinação de operações para serem armazenadas em cache

Essas métricas serão obtidas por meio da execução de *workflows* científicos desenvolvidos utilizando o *Apache Spark*.

Limitações, riscos e ameaças:
Nesse trabalho iremos considerar apenas a aplicação do modelo proposto para *workflows* científicos, portanto, outras categorias de aplicações que sejam desenvolvidas utilizando o arcabouço *Apache Spark* podem não ter ganhos de desempenho ao utilizar nossa solução.

Contribuição científica:
- um modelo que utiliza técnicas de aprendizado de máquina para decidir quais resultados intermediários devem ser armazenados em cache para arcabouços de computação distribuída de alto desempenho

Contribuição tecnológica (se pertinente):
- uma implementação do modelo proposto no arcabouço computacional *Apache Spark*

Método de pesquisa			
Gênero (escolha UM)	[ ] Pesquisa teórica	[x] Pesquisa prática	[ ] Pesquisa empírica [ ] Pesquisa metodológica
Natureza (escolha UMA)	[ ] Pesquisa básica	[x] Pesquisa aplicada	
Abordagem (escolha UMA)	[x] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Revisão de literatura* (você pode escolher mais de uma)	[ ] Revisão narrativa [ ] Revisão descritiva [x] Revisão de escopo	[ ] Meta-análise [x] Revisão sistemática qualitativa [ ] Revisão *guarda-chuva*	[ ] Revisão teórica [ ] Revisão realística [ ] Revisão crítica
Procedimento técnico principal (escolha UM)	[x] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto*	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação	[ ] Pesquisa etnográfica [ ] Teoria fundamentada em dados [ ] Ciência do projeto [ ] Outra Qual? _____
Análise de dados (você pode escolher mais de uma)	[x] Estatística descritiva [ ] Estatística inferencial	[ ] Teste estatístico [ ] Análise de conteúdo	[ ] Análise do discurso [ ] Outros: _____

* Definição de tipos de revisões de literatura estabelecida por Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

Próximas atividades:
Na Tabela 1 é possível visualizar o cronograma de atividades com a lista de atividades planejadas para serem realizadas durante esse projeto de pesquisa.

Atividade		2019		2020								
Num.	Descrição	11	12	1	2	3	4	5	6	7	8	9	10
1	Revisão Bibliográfica	x	x	x									
2	Levantamento de Workflows Científicos	x	x	x									
3	Estudo arquitetura do Apache Spark	x	x	x	x	x	x	x	x				
4	Elaboração do algoritmo adaptativo de cache		x	x	x	x	x	x					
5	Implementação do algoritmo no Apache Spark				x	x	x	x	x				
6	Adaptação no Apache Spark para coletar métricas				x	x	x	x	x				
7	Testes da implementação				x	x	x	x	x				
8	Análise de desempenho de execução em workflows científicos						x	x	x	x			
9	Redação da Dissertação								x	x	x	x	x
10	Redação de artigos									x	x	x	x

Tabela 1 – Cronograma de Atividades