Avaliando uma Estratégia Computacional Baseada em Workflows Científicos Apoiados por Placas Gráficas Genéricas

Fábio da Silva Cardozo[1], Ulisses Roque Tomaz[1], Sergio Manuel Serra da Cruz[1,2]

[1] Universidade Federal Rural do Rio de Janeiro (UFRRJ)
[2] Programa de Educação Tutorial (PET-SI/UFRRJ)
BR-465, Km 7 Seropédica-Rio de Janeiro-RJ-Brazil

{ulisses.rtomaz, fcardozzo}@gmail.com, serra@pet-si.ufrrj.br

Resumo. *O crescente volume de dados necessários para a realização de pesquisas atmosféricas e climáticas oferecem desafios. Este trabalho tem como objetivo apresentar uma estratégia computacional baseada em workflows científicos e técnicas de consistência dados destinada ao tratamento de longas séries de dados pluviométricos, para isso utiliza-se recursos de computação paralela em placas gráficas de propósito geral com coleta de dados de proveniência retrospectiva. Os primeiros resultados mostram que a estratégia apresenta altas taxas de preenchimento de falhas e ganhos de desempenho quando comparada a abordagem sequencial tradicional.*

Abstract: *The increasing volume of meteorological data needed to conduct atmospheric and climate studies offer new challenges for data analysis area. This work aims to present the initial steps of a computational approach called MetFlow. It is tailored to compute long series of rainfall data using parallel computing capabilities of low cost and general purpose graphics cards. The approach uses retrospective provenance metadata to enrich the quality of the raw rainfall data. Our initial results show that the approach not only augment the quality of the data but also offer performance gains when compared to the traditional sequential approach.*

1. Introdução

As ciências da terra têm avançado velozmente graças a experimentação computacional, tendo como um ponto importante os estudos baseados em simulações numéricas dos fenômenos meteorológicos (BATCHELOR, 2000). As mudanças climáticas em curso podem resultar em impactos agroambientais e econômicos severos para a humanidade e para o Brasil. Aumentos nas temperaturas do ar são esperados em nível mundial e uma das consequências decorrentes serão as variações nos ciclos hidrológicos (IPCC, 2013). Acredita-se que a maior ameaça para os seres humanos será manifestada em nível local, através de mudanças em eventos regionais de tempo e clima extremos. O Brasil é vulnerável a mudanças na frequência e intensidade de eventos extremos, como ondas de calor, secas, enchentes e chuvas extremas, como ocorrido nos últimos anos nas regiões Sul e Sudeste.

Estudos meteorológicos são caracterizados por manipularem grandes quantidades de dados em longas séries contínuas e consistentes. Contudo, obter séries com essas características ainda é um grande desafio nesta área de estudo. As falhas e perdas dos dados ocorrem desde o momento da coleta na estação meteorológica até sua disponibilidade em repositórios de dados (LEMOS FILHO *et al*., 2013). Além dessas dificuldades, os dados se encontram em diferentes estruturas, formatos, descontinuidades cronológicas e sem os descritores de proveniência.

Esse trabalho tem como objetivo apresentar uma estratégia computacional centrada no uso técnicas de consistência de dados pluviométricos apoiadas por *workflows* científicos executados em ambientes paralelos que utilizam placas gráficas que são capazes de transformar dados pluviométricos brutos em dados curados de qualidade enriquecidos por proveniência (FREIRE *et al.*, 2008). A estratégia apresentada é capaz de processar tais dados em arquiteturas *Compute Unified Device Architecture* (CUDA), que envolvem computação paralela em placas de processamento gráfico (GPGPU) de baixo custo e de uso genérico.

2. Trabalhos Relacionados

Atualmente, ambientes paralelos baseados em GPGPU representam uma alternativa viável para acelerar aplicações científicas baseadas em *workflows* (GOSWAMI *et al*., 2016, LIU *et al.*, 2016). Estes ambientes possibilitam que problemas complexos, de simulação computacional, sejam executados em tempos aceitáveis e a baixos custos quando comparados com *clusters* e nuvens de computadores. Para este fim, os sistemas heterogêneos compostos de processadores *multi-core* e aceleradores *many-core* como as GPGPUs representam uma tendência na atualidade. No entanto, o maior desafio está em identificar quais são as atividades do *workflow* ou os trechos de código adequados para cada tipo de arquitetura. Nesta seção apresentamos os fundamentos utilizados pela estratégia MetFlow concebida por (CARDOZO, 2014).

2.1 Técnicas de Consistência e Preenchimento de Falhas em Dados Pluviométricos

A técnica de controle de qualidade de dados meteorológicos adota sequencias de filtros de dados (FENG *et al.*, 2004), que são aplicados na detecção e identificação de erros nos dados brutos coletados pelos sensores das estações meteorológicas. Segundo Magina (2007) os filtros utilizados são capazes de identificar registros extremos reais e registros espúrios, os últimos devendo ser excluídos, pois se mantidos na série histórica aumentariam a frequência de casos extremos, e distorceriam a estimativa de parâmetros que produzem a função de probabilidade de extremos e os tempos de retorno.

Por adequação ao trabalho proposto as regras estabelecidas por Feng *et al.* (2004) foram adotadas nessa pesquisa, as quais foram aplicadas às séries de dados produzidas pelas estações meteorológicas na detecção de valores suspeitos. As regras utilizadas são: (i) Detecção de Valores Extremos Máximos e Mínimos; (ii) Verificação de Eventos Temporalmente Isolados; (iii) Verificação de Eventos Espacialmente Isolados.

Outra etapa do controle de qualidade dos dados meteorológicos consiste em realizar o preenchimento dos dados ausentes da série por meio de métodos estatísticos. Os métodos mais utilizados são: (i) Regressão Linear Simples (RL); (ii) Vizinho mais

Próximo (VP); (iii) Ponderação Regional com Bases em Regressões Lineares (PRRL); (iv) Ponderação Regional (PR) e (v) Inverso da Potência da Distância (IPD). Precinoto *et al.* (2013), em estudos anteriores, verificaram que o método RL é um dos mais adequados para preenchimento de falhas sobre dados de precipitação na região Sudeste do estado do Rio de Janeiro. Portanto, este será o adotado neste trabalho.

2.2 Tecnologias GPGPU-CUDA

Nos últimos anos as placas gráficas GPGPU tiveram seu uso difundido. Verificou-se que seu poder computacional, incialmente disponível para jogos, era potencialmente aplicável para resolução de diversas categorias de problemas científicos (CHAKRABARTI *et al.*, 2012, GOSWAMI *et al.*, 2016).

Adotou-se para este trabalho a arquitetura CUDA da NVIDIA. Este ambiente de programação que permite realizar computação de propósito geral utilizando a GPU e fornece acesso aos recursos do hardware através de comandos semelhantes aos das linguagens C. Os módulos CUDA implementam *threads* paralelas para executar as atividades do *workflow* relacionadas com as análises de dados e preenchimento de falhas baseados no método RL de preenchimento de falhas.

2.3 MetFlow

O MetFlow, cuja primeira versão foi concebida por Cardozo (2014), é um *workflow* científico (DEELMAN *et al.*, 2009) que realiza tratamento de dados e prognósticos quantitativos sobre dados de precipitação pluvial. Tais prognósticos utilizam grandes volumes de dados representados por longas séries temporais.

Por intermédio do MetFlow se executam experimentos do tipo simulações numéricas dos fenômenos meteorológicos distribuídos em ambientes de computação sequenciais e paralelos e também se armazenam dados curados juntamente com seus metadados de proveniência de cada execução.

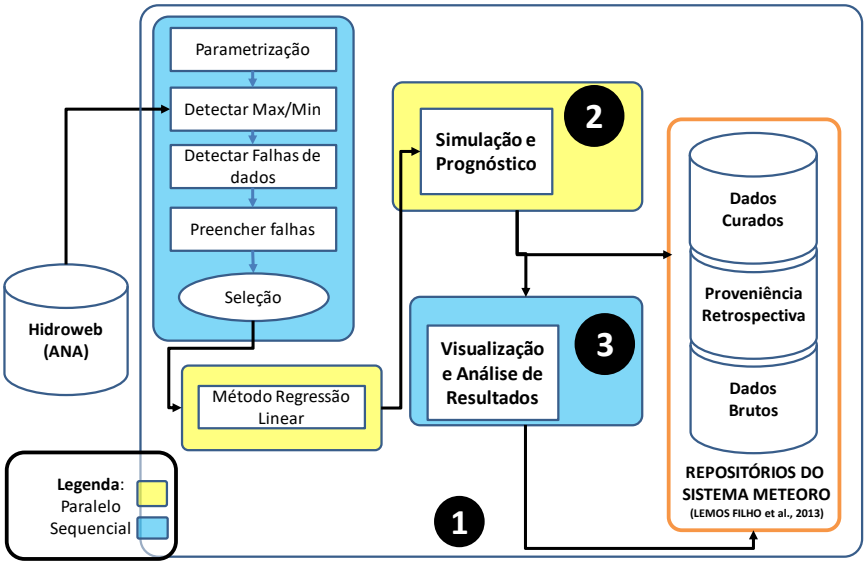

Figura 1 – Representação conceitual das fases de experimentos computacional em Pluviometria apoiada pelo MetFlow nos ambientes sequencial e paralelo (CGPGPU-CUDA).

A atual versão do MetFlow utiliza módulos CUDA e Python, apoiada pelo método de regressão linear (RL) para efetuar preenchimento de falhas das séries temporais.

2.4 Trabalhos Relacionados

Atualmente, existem diversos trabalhos na área de pré-processamento de dados meteorológicos. Magina (2007) propõe, de forma semiautomática, formatos de tratamento estatísticos para series históricas meteorológicas, mas conta somente com o suporte de macros em planilhas MS Excel® para aplicação desse tratamento. Esta abordagem torna o trabalho humano massivo e sujeito a erros. Além disso, não incorpora as questões de proveniência de dados. Lemos Filho *et al.* (2013), propõem um sistema baseados em proveniência e *pipelines* de pré-processamento de dados pluviométricos em uma plataforma Web, que, no entanto, não traz uma solução ou aplicação paralela de alto desempenho, nem se utiliza de *workflows* científicos. A Tabela 1 apresenta uma comparação entre as funcionalidades de trabalhos correlatos.

Asvija *et al.* (2010) sugerem o uso de *workflows* científicos para implementar modelo numérico meteorológico *fifth-generation Model Mesoscale*, que trata do prognóstico em Meteorologia de mesoescala de fenômenos atmosféricos, como brisas, tempestades de convecção, não contemplando o tratamento de dados e coleta de dados de proveniência.

Horta *et al.* (2013) trata do uso de *workflows* científicos empregados em ambiente de *clusters*, mas não apresentam solução desenvolvida que se aplique ao problema que norteia este trabalho. Os autores se restringem ao campo da investigação teórica, indicando o que é possível realizar juntando *workflow* científico e computação massiva paralela com GPU.

Tabela 1 – Comparativo entre os trabalhos relacionados.

	MetFlow (2014)	Filho et al. (2013)	Magina (2007)	Horta, et al. (2013)	Asvija et al. (2010)	Whang e Shi (2014)
Uso de Workflows científicos	X	-	-	X	X	-
Coleta de proveniência retrospectiva	X	X	-	-	-	-
Técnicas de detecção de falhas e preenchimento	X	X	X	-	-	X
Uso de técnicas de paralelismo de dados	X	-	-	-	X	X
Uso de GPGPU	X	-	-	X	-	X
Programação em CUDA	X	-	-	X	-	X
Prevê validação cruzada de dados	X	-	-	-	-	-
Uso de dados de pluviometria	X	X	X	-	-	-
Compartilhamento de esquemas relacionais	X	X	-	-	-	-
Compatibilidade com *schema* PROV	X	-	-	-	-	-

Whang e Shi (2014) reconhecem o problema da qualidade de dados meteorológicos e apresentam uma metodologia e um *workflow* (sem uso de SGWfC)

baseado em MPI executado em placas GPGPU para tratar dados sumarizados de temperatura da superfície da Terra coletados pela WMO/NOAA, a iniciativa gera dados curados em formato texto estruturado facilitando simulações computacionais. No entanto, o trabalho não considera a importância da proveniência de dados.

3. Metodologia

Os dados utilizados nesse trabalho são oriundos do sistema HidroWeb mantido pela Agência Nacional de Águas (HIDROWEB, 2014). Eles são séries temporais pluviométricas não consistentes sobre: (i) históricos diários de chuva; (ii) inventário de bacias, rios, estados, municípios, estações (pluviométricas e fluviométricas) e suas respectivas coordenadas geográficas.

Para a execução dos experimentos foi utilizado um notebook com processador Intel Core i5 dual core de 2,9 GHz, com 4GB de RAM DDR3 de 533 MHz e placa de vídeo (GPGPU) NVIDIA GeForce GT 335M de 1080 MHz, com 1GB e 72 cores CUDA. Neste estudo utilizamos dados de chuva da região Sul Fluminense, pois apresentam importância industrial, agropecuária e também por ser a principal fonte de captação de águas do estado do Rio de Janeiro. Dentre todas as estações da região, foram selecionadas todas as 77 estações meteorológicas dentro da área de maior pluviosidade e de altimetria próximas. As coordenadas dessas estações variam entre as latitudes são 22º 03’ e 23º 21’ S e longitudes 43º 25’ e 44º 54’ W.

Consideraram-se apenas séries de chuvas superiores ou iguais a 20 anos de dados e com início a partir de janeiro de 1960 até dezembro de 2013. Ou seja, nossos experimentos utilizaram séries de dados pluviométricos reais com 53 anos. Como prova de conceito utilizou-se o *workflow* MetFlow composto por um conjunto de atividades (módulos) que fazem parte do *workflow* científico paralelo que é capaz de processar longas séries de dados pluviométricos.

De acordo com a Figura 1, os experimentos realizados por intermédio do MetFlow podem ser divididos em três fases distintas: 1) *pré-processamento, 2) prognóstico e 3) visualização* dos dados meteorológicos. Os dados brutos processados pelo *workflow* e os dados curados são armazenados em um repositório de dados do tipo relacional capaz de armazenar os metadados de proveniência retrospectiva gerados a cada execução do mesmo.

Este tipo de abordagem é muito importante, pois permite consultar e compartilhar conjuntamente os dados e metadados dos experimentos, ampliando sua transparência e confiabilidade dos experimentos. Ressalta-se que o repositório de dados utilizado na pesquisa compartilha o mesmo esquema originalmente proposto por Lemos Filho *et al.* (2013).

4. Experimentos em Sequencias e Paralelos com Regressão Linear

Inicialmente, os experimentos carregam os dados pluviométricos brutos do HidroWeb através do MetFlow, que é composto por um conjunto de tarefas configuradas pelo pesquisador [parametrização, preenchimento de falhas, preparação de dados para o ambiente paralelo (vetorização) e processamento de dados] executados em ambientes sequencial e paralelo. O modelo estatístico RL que efetua o preenchimento de falhas das séries pluviométricas é processado em paralelo por meio de código CUDA, gerando

novos repositórios de dados enriquecidos por metadados de proveniência que etiquetam cada item de dado analisado resultante desse processo.

Em especial, a proveniência (FREIRE *et al.*, 2008) atua como um certificado de qualidade e autenticidade de cada item de dado meteorológico, o que permite o compartilhamento e reuso dos dados sem falhas com descritores detalhados, que assim, explicitam a lógica da geração de cada item de dado do banco.

Neste trabalho foram utilizadas duas versões do *workflow* MetFlow no VisTrails. A versão sequencial possui módulos de processamento desenvolvidos em Python e a versão paralela possui módulos adicionais de processamento paralelo desenvolvidos em linguagens CUDA e Python. Ambas utilizaram o MySQL acoplado ao VisTrails. Os dados, assim como, a proveniência retrospectiva são armazenados na base de dados do sistema Meteoro (LEMOS FILHO *et al.*, 2013) e podem ser armazenados em granularidades distintas.

Os módulos sequenciais são codificados apenas em Python e os paralelos em linguagem CUDA (seus códigos estão disponíveis mediante solicitação aos autores). Os módulos paralelos são aqueles que consomem mais recursos de processamento, ou seja, vetorização e preparação da regressão linear, e por esse motivo são executados em ambientes de maior capacidade de processamento. O módulo coletor de proveniência utilizado no protótipo foi o disponível internamente pelo VisTrails (CALLAHAN *et al.*, 2006).

Uma das principais características do MetFlow são os elevados percentuais de correção e preenchimentos de falhas e baixo tempo de execução quando comparados com a tradicional abordagem manual. Em termos de quantificação de correções de falhas, os resultados estão apresentados na Tabela 2, onde é possível observar os números de meses com falhas e os percentuais de ajustes dos dados. Por exemplo, nos experimentos com 36 estações (dados desde 1960 até 2013, ou seja, 53 anos de dados e um total de 22.896 meses) havia 1.461 meses que apresentavam algum tipo de falhas, sendo corrigidos 1.310 meses.

O MetFlow em sua versão paralela utilizando apenas três *threads* e o método RL com 36 estações dentro de um raio de 20Km de distância foi capaz de corrigir automaticamente 89,6% dos casos (1.310 meses) com um tempo médio de processamento de 5:57min, enquanto que o percentual de acerto para 77 estações foi de 72,17% com um tempo médio de processamento de 16:35min. Os valores percentuais crescentes de correções se devem ao maior número de falhas corrigidas nas séries de dados e ao maior número de estações. As diferenças de tempos de execução entre as versões sequencial e paralela são pequenas, porém crescentes e ligeiramente melhores na versão paralela.

Tabela 2 – Variações percentuais de correções de falhas executadas pelo MetFlow em sua versão paralela usando o método regressão com raio de distância de 20Km e número de execuções de cada versão do MetFlow = 5.

	Meses avaliados (= 12 * nº. estações * 53 anos)	Meses com falhas	Meses com correções realizadas	Tempo médio de execução sequencial	Tempo médio de execução paralelo	% de acertos dos dados
17 estações	10.812	552	327	2:53min	2:51min	59,24
36 estações	22.896	1.461	1.310	6:10min	5:57min	89,66
77 estações	48.972	2.479	1.789	17:51min	16:35min	72,17

5. Conclusão

Neste trabalho ficou clara a adequação da modelagem de um experimento em Pluviometria para o tratamento e preenchimento de falhas usando o MetFlow em ambiente baseado em GPGPU. Porém, nossos primeiros resultados mostram que para atingir uma solução paralela mais genérica se necessitam de investigações complementares.

As pesquisas com *workflows* científicos em execução paralela em placas genéricas do tipo GPGPU ainda estão se desenvolvendo em todo o mundo, ainda não há um referencial teórico estabelecido. No entanto, estimamos que o uso de placas gráficas para acelerar a execução de *workflows* será amplamente desenvolvido futuramente.

As principais contribuições deste trabalho são o expressivo percentual de correção e preenchimento de falhas pelo MetFlow aliados com geração de anotações de proveniência enriquecendo os dados. No que tange à correção das falhas obtivemos acertos expressivos com valores superiores a 59 % mesmo com pequeno número de estações utilizadas. Este valor pode ser ampliado a medida que se aumentam o número de estações e se variam as distâncias entre eles, os quais são estabelecidos nos parâmetros utilizados pelos cientistas ao executar o MetFlow.

Dentre as principais limitações destacamos que os ganhos de desempenho da versão paralela comparada com a versão sequencial foram pouco significativos, indicando que maiores estudos devem ser realizados para verificar a adequação do uso ou da escolha do módulo paralelizável ou mesmo a escolha de melhores equipamentos para a realização dos experimentos.

Agradecimentos

Agradecemos ao FNDE e ao MEC/SeSU pelo financiamento do programa PET-SI/UFRRJ e ao programa PPGMMC/UFRRJ pelas bolsas concedidas.

Referências

Asvija, B.; Shamjith, K.V.; Sridharan, R.; Chattopadhyay, S. Provisioning the MM5 meteorological model as Grid Scientific Workflow. 2010.

Batchelor, G. K. An Introduction to Fluid Dynamics. Cambridge University Press. 2000.

Cardozo, F. S. Tratamento e preenchimento de falhas de séries de dados meteorológicos utilizando *workflows* científicos paralelos em ambientes de GPU. Dissertação Mestrado – UFRRJ, 2014.

Callahan, S. P.; Freire, J.; Santos, E.; Scheidegger, C. E.; Silva, C. T.; Vo, H. T. VisTrails: visualization meets data management. In: Proceedings of the, 2006 ACM SIGMOD, p. 745-747, 2006.

Chakrabarti, G.; Grover, V.; Aarts, B. et al. CUDA: Compiling and optimizing for a GPU platform. Procedia Computer Science, v. 9, p. 1910–1919. 2012.

Deelman, E.; Gannon, D.; Shields, M.; Taylor, I. Workflows and e-Science: An overview of workflow system features and capabilities, FGCS, v. 25, n. 5, p. 528-540, 2009.

Freire, J.; Koop, D.; Santos, E.; Silva, C.L. Provenance for computational tasks: A Survey. Computing in Science & Engineering, v. 10, n. 3, p. 11–21, 2008.

Feng, S.; Hu, Q.; Qian, W. Quality control of daily meteorological data in China, 1951-2000: a new dataset. Int. Journal of Climatology, v. 24, n. 7, p. 853–870. 2004.

Goswami, A. et al., Landrush: Rethinking In-Situ Analysis for GPGPU Workflows, 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Cartagena, pp. 32-41, 2016.

Lemos Filho, G. R.; Precinoto, R. S.; Correia, T. P.; Santos, E. O.; Lyra, G. B.; Cruz, S. M. S., Assimilação, Controle de Qualidade e Análise de Dados de Meteorológicos Apoiados por Proveniência, VII e-science Workshop, XXXIII CSBC. 2013;

Liu, J. et al. An efficient geosciences workflow on multi-core processors and GPUs: a case study for aerosol optical depth retrieval from MODIS satellite data. 2016. http://dx.doi.org/10.1080/17538947.2015.1130087.

HidroWeb - Sistema HidroWeb. Disponível em: <http://hidroweb.ana.gov.br>. Acesso em 14 de maio de 2014.

Horta, F.; Dias, J.; Elias, R. et al. Prov-Vis: Large-Scale Scientific Data Visualization Using Provenance. 2013.

IPCC Climate Change 2013: The Physical Science Basis. Disponível em: <http://www.ipcc.ch/report/ar5/wg1/>. Acesso em 14 de março de 2016.

Magina, F. C. Aquisição Automática e Tratamento de Dados Meteorológicos Aplicáveis ao Projeto e Operação de Linhas Aéreas de Transmissão de Energia Elétrica. Dissertação de Mestrado. 2007.

Precinoto, R. S.; Lemos Filho. G. R.; Correia, T. P.; Santos, E. O.; Lyra. G. B.; Cruz, S. M. S. 2013. Uso De Sistema De Pré-Processadores Para Obtenção De Séries Pluviométricas De Qualidade. Congresso Brasileiro de Agrometeorologia 2013.

Shi, X.; Wang, D. Processing NOAA Observation Data over Hybrid Computer Systems for Comparative Climate Change Analysis. In: WorldComp, 2014. p. 1.