INTEGRAÇÃO ENTRE PRÉ-PROCESSADORES E W*ORKFLOWS* CIENTÍFICOS PARA AUXILIAR A GERÊNCIA DE DADOS METEOROLÓGICOS NO ESTADO DO RIO DE JANEIRO

Fillipe Souza da Silva Dornelas[1]; Sergio Manuel Serra da Cruz[2]; Gustavo Bastos Lyra[4]& Ednaldo Oliveira dos Santos[4,5]

[1]BolsistaPROIC/UFRRJ, Discente do Curso de Sistemas de Informação/UFRRJ; [2]Professor do DEMAT/ICE/UFRRJ, [4]Professor do DCA/IF/UFRRJ; [5]Professor orientador.

Palavras-Chave: chuva, qualidade de dados, workflows, proveniência.

Introdução

A busca pelo conhecimento científico faz com que instituições acadêmicas e cientificas procurem não só novas formas de gerenciar crescentes volumes de dados em seus experimentos científicos, como também reduzir o tempo e os custos necessários para seu tratamento (FENG *et al.*, 2004). A adoção de procedimentos computacionais apoiados por *workflows* científicos permite atingir ganhos de produtividade e qualidade na condução desses experimentos. Desta forma, a gerência adequada dos descritores de proveniência representam um diferencial (CRUZ, 2011).

No entanto, a gerência efetiva de experimentos científicos que manipulam grandes volumes de dados na área de Meteorologia, requer novas abordagens, como por exemplo, técnicas de distribuição de dados e processos, mecanismos de coleta de descritores de proveniência, entre outros. Portanto, mapear, caracterizar e adicionar qualidades aos diferentes tipos de dados meteorológicos e agregar descritores de proveniências correlacionados com os experimentos científicos dessa área é um problema ainda em aberto que deve ser pesquisado.

Em vista disso, este trabalho tem como objetivo desenvolver *workflows e* integrar seus resultados com o sistema *Meteoro* (LEMOS FILHO *et. al.*, 2013) para serem utilizados em experimentos científicos na área de Meteorologia que usam grandes massas de dados coletados a partir de estações meteorológicas existentes no Rio de Janeiro e disponibilizados em bases públicas de dados.

Metodologia

Os dados utilizados nesta pesquisa fazem parte de séries de dados meteorológicos coletados em diversas estações pluviométricas distribuídas no estado do Rio de Janeiro (latitudes 20º 45’ e 23º 21’ S e longitudes 40º 57’ e 44º 53’ W). As séries climáticas de precipitação pluvial foram obtidas em arquivos de textos disponíveis no banco de dados da Agência Nacional de Águas (ANA), com auxílio da ferramenta Hidroweb (http://hidroweb.ana.gov.br). Esses dados são obtidos de arquivos textos estruturados, e incorporados ao sistema de pré-processadores, previamente desenvolvido pelo grupo de pesquisa, denominado *Meteoro* (LEMOS FILHO *et. al.*, 2013).

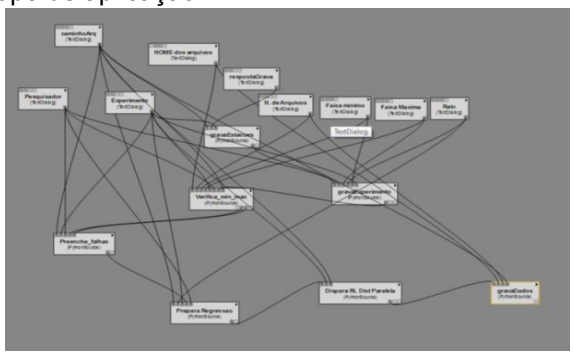
A metodologia de trabalho a ser aplicada nesse estudo está baseada na proposta de desenvolvimento de modelos semânticos bem fundamentados para a representação de descritores de proveniência de dados de experimentos científicos desenvolvida por Cruz (2011) e Cruz *et al.* (2011, 2012). Em vista disso, para o desenvolvimento dos *workflows* científicos está sendo usado o VisTrails (BAVOIL *et al.*, 2005) para operar sobre os dados meteorológicos obtidos no estado do Rio de Janeiro, para gerar dados livres de falhas. Escolheu-se este sistema por ser um sistema aberto, de amplo uso e sem custos de licenciamento.

Resultados e Discussão

Após análise dos dois sistemas gerenciadores de workflow científico conforme item anterior, foi verificado que o melhor sistema para o desenvolvimento do workflow científico em nosso trabalho foi o *VisTrails*. Isso se justifica porque ele facilita o sistema e permite a visualização da proveniência retrospectiva dos *workflows* e execução de suas versões, sendo capaz de expressar resultados de modo gráfico. Além disso, ele possibilita definir um *workflow* concreto e controlar alterações, ou seja, diferentes versões, de modo visual através de árvores.

Um dos desafios dessa pesquisa foi a integração de um workflow desenvolvido em *VisTrails* e a execução de pelo menos parte do modelo estatístico usando outros recursos, uma vez que o *VisTrails* não possui uma ferramenta específica para este tipo de aplicação.

Com isso foi necessário criar um sub-workflow de integração em código *python* usando o componente "*python source*" do próprio *VisTrails*, tanto para a execução local quanto para a execução em uma máquina remota. Naturalmente que após a aplicação dos métodos estatísticos são gerados novos arquivos, que por sua vez, servem de entrada para o workflow, pois contém dados curados, que servem como resultado e fazem parte também da coleta de proveniência dos dados do experimento, conforme figura ao lado.



Conclusão

A abordagem proposta apresenta o paradigma dos *workflows* científicos voltados para tratar problemas que manipulam grandes volumes de dados meteorológicos e capturar informações de proveniência relacionadas às operações de criação, modificação e exclusão de arquivos, além de relacioná-las aos dados de proveniência e retrospectiva do *workflow*.

Atualmente a integração entre os sistemas Meteoro e Workflow são feitas através de um arquivo de texto que após ser gerado pelo Workflow, o mesmo é carregado no sistema *Meteoro*. Através dos testes e avaliações verificou-se que o melhor sistema a ser usado no desenvolvimento dos *workflows* científicos sobre os dados meteorológicos no estado do Rio de Janeiro é o *VisTrails*. Para a integração entre os sistemas, o arquivo gerado é devidamente identificado pelo sistema Meteoro e para fins de proveniência é feita a identificação de cada arquivo usado para a geração dos resultados dos métodos estatísticos.

A próxima etapa será a integração no nível de banco de dados, onde ambos os sistemas compartilhem a mesma base de dados e possam assim ter mais autonomia na execução de seus procedimentos.

Agradecimentos e Auxílio Financeiro

Agradecemos à FAPERJ pelo financiamento do Projeto "Uso de *Workflows* Científicos e Ontologias em Experimentos Científicos baseados em Grandes Massas de Dados Meteorológicos" e ao Programa PROIC/UFRRJ pela bolsa de iniciação científica do autor principal deste trabalho. Agradecemos também a Fábio Cardozo da Silva, Mestrando do PPGMMC/UFRRJ, em relação ao desenvolvimento do workflow cientifico.

Referências Bibliográficas

BAVOIL, L.; CALLAHAN, S. P.; CROSSNO, P. J. *et al.*, 2005. VisTrails: Enabling Interactive Multiple-View Visualizations. Proceedings of IEEE Visualization, pp. 135-142.

CRUZ, S.M.S. *Uma Estratégia de Apoio à Gerência de Dados de Proveniência em Experimentos Científicos*. Tese de Doutorado. PESC/COPPE-UFRJ, 2011.

CRUZ, S.M.S.; SILVA, C.E.P.; OLIVEIRA, D. *et al.*., 2011. Capturing Distributed Provenance Metadata from Cloud-Based Scientic Workflows. J. Inform. and Data Management, v. 2, n. 1, pp. 43-50.

CRUZ, S.M.S.; MACHADO, M.L.M.; MATTOSO, M., 2012.A Foundational Ontology to Support Scientific Experiments. In: Proceedings of Joint V Seminar on Ontology Research in Brazil and VII International Workshop on Metamodels, Ontologies and Semantic Technologies, pp. 144-155, Recife, Pernambuco.

FENG, S.; HU, Q.; QIAN, W., 2004. Quality Control of Daily Meteorological Data in China, 1951–2000: A New Dataset. International Journal of Climatology, p. 853-870. 14 abr. 2004

LEMOS FILHO, G. R.; PRECINOTO, R. S.; CORREIA, T. P.; SANTOS, E. O.; LYRA, G. B. & CRUZ, S. M. S., 2013. Assimilação, Controle de Qualidade e Análise de Dados de Meteorológicos Apoiados por Proveniência. XXXIII CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO/BreSci – VII Brazilian e-Science Workshop, Maceió/AL, Julho/2013.