ANÁLISE E PROCESSAMENTO AUTOMÁTICO DE GRANDES VOLUMES DE DADOS AMBIENTAIS (*BIG EARTH OBSERVATION DATA SETS)*

John Elton de Brito Leite Cunha [1]
Iana Alexandra Alves Rufino[1]
Carlos de Oliveira Galvão[1]
Thiago Emmanuel Pereira[2]
Francisco Vilar Brasileiro[2]
Esdras Vidal Pereira[2]
[1] Universidade Federal de Campina Grande, Centro de Tecnologia e Recursos Naturais
{john.brito, iana.alexandra, carlos.galvao}@ufcg.edu.br

[2] Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática
thiago.manel@gmail.com
fubica@computação.edu.br
esdras015@gmail.com

Abstract. Hydrology and water resources demand monitoring land use and cover, related to the impacts of climate and human action. However, very often data for such monitoring and sequent analysis are from spatial scales that cannot be fully collected by field survey. Remote sensing techniques and data are suitable to those needs, since include land use/land cover changes detection in different scales (from local to continental landscapes). This paper presents an intercontinental initiative: the EUBrazil Cloud Connect project, developed by European and Brazilian partners. The main goal is to provide a cloud computing infrastructure to use tools for multi-temporal analysis and trend analysis of huge remote sensing databases to understand the main current drivers of land use changes. SEBAL (Surface Energy Balance Algorithm for Land) algorithm has been processed for a long time series (more than 30 years of satellite images) covering the whole Brazilian semi-arid area. Web services for visualization, analysis and deployment for decision makers and researchers are used.

Palavras-chave: image processing, *cloud computing*, multitemporal analysis, semiarid, vegetation, processamento de imagens, computação em nuvem, análise multitemporal, semiárido, vegetação.

1. Introdução

As pesquisas ambientais utilizando SRO estão provendo dados em grandes escalas, mas estão tendo dificuldades em encontrar arquiteturas de *software* para acomodar o grande volume de dados (Li *et al.*, 2010). As plataformas *Amazon's Elastic Compute Cloud*, *Google App Engine* e *Microsoft's Windows Azure* estão entre as recentes tecnologias que permitem maior capacidade computacional para análise e processamento de grandes volumes de dados (*Big Data*) ambientais advindos do SRO (Wang *et al*., 2015; Ma *et al*., 2015; Yang *et al*., 2015).

No entanto, há desafios na conversão dos métodos e algoritmos utilizados tradicionalmente pela comunidade de usuários de SRO: algumas análises visuais e interativas, realizadas pelo operador, precisam ser substituídas por procedimentos computacionais automáticos; em muitos casos, há a necessidade de trabalhar com frações da área de estudo; automatizar os processos de coleta de dados em banco de dados, muitas vezes heterogêneo e

inconsistente; e a necessidade de implementação de procedimentos computacionais para o gerenciamento dos processos, desde a coleta de dados à geração do produto final.

Neste sentido, este artigo apresenta as iniciativas do projeto *EU-Brazil Cloud Connect* para difusão dos dados obtidos por sensores orbitais e ferramentas para análise multitemporal e análise de tendências de mudanças na superfície terrestre. Para tanto, são utilizadas ferramentas de computação em nuvem (*cloud computing*) para o processamento do Algoritmo SEBAL (*Surface Energy Balance Algorithm for Land*) para uma série superior a 30 anos de imagens orbitais para todo semiárido brasileiro e serviços Web para visualização, análise e disponibilização dos produtos para tomadores de decisão e comunidade científica.

2. Metodologia do Trabalho

O sistema, desenvolvido no âmbito do projeto *European Union - Brazil Cloud Connect,* prevê processamentos que ocorrem em segundo plano e em tempo real. Os processamentos que ocorrem em segundo plano utilizam o sistema *Blowout* para aquisição de dados climáticos e imagens de satélite e execução dos algoritmos Fmask - *Function of Mask* (Zhu e Woodcock, 2012) e SEBAL (Bastiaanssen, 2000). Os processamentos em tempo real são realizados no *BioClimate Scientific Gateway* (BSG) e são destinados aos usuários finais, permitindo análise de tendências, geração de mapas e extração de informações em diferentes formatos (CSV, GeoTIFF e PNG).

No projeto *EUBrazilCC* optou-se por utilizar imagens dos satélites LANDSAT 5, 7 e 8, que apresentam resolução espacial de 30 metros e temporal de 16 dias. O conjunto de imagens disponíveis para o semiárido reúne cerca de 42 mil imagens, divididas em 57 cenas. As imagens do satélite LANDSAT são obtidas diretamente do repositório de imagens na *National Aeronautics and Space Administration* (NASA).

As informações sobre o clima necessárias para a aplicação do algoritmo SEBAL são selecionadas com base na localização das imagens de satélite e horário da passagem do sensor na região de interesse. São necessárias também as informações sobre a velocidade e temperatura do ar (últimos 30 anos) e para as imagens disponíveis a cada nova passagem do satélite pelo mesmo local. Os dados climáticos utilizados nesta aplicação são obtidos no Instituto Nacional de Meteorologia (INMET).

Para detecção das nuvens e sombra de nuvens nas imagens LANDSAT é utilizado o algoritmo *Fmask*. Em testes com imagens de referência globalmente distribuídas, foi verificado

que a acurácia do método foi de 96,4%, sendo esse algoritmo uma boa opção para identificação de nuvens em imagens LANDSAT (Zhu e Woodcock, 2012).

Tradicionalmente, a aplicação do algoritmo SEBAL envolve a interação com o usuário para identificação dos pixels de referência, procedimento conhecido como o CIMEC - *Calibration Using Inverse Modeling at Extreme Conditions*, usado pelos algoritmos SEBAL e METRIC- *Mapping EvapoTranspiration at high Resolution with Internalized Calibration* (ALLEN *et al*., 2011). A automatização do processo de escolha dos pixels de referência foi realizada com base em Allen *et al*. (2013); algumas adaptações foram incorporadas para melhor representação e adequação aos aspectos locais do semiárido brasileiro.

3.1. Arquitetura do Sistema *Blowout*

O sistema *Blowout* usado nesse estudo usa recursos de computação de uma federação de provedores de IaaS (*Infrastructure as a Service*) para execução dos algoritmos *Fmask* e SEBAL. Esta federação é formada pela agregação dos recursos ociosos disponíveis em cada provedor membro da federação. Provedores IaaS geralmente mantêm uma fração de seus recursos computacionais ociosos para lidar com variações da demanda por seus recursos, bem como para tratar falhas temporárias na infraestrutura. Do ponto de vista de um provedor, juntar-se a uma federação é bastante conveniente por uma razão simples: é possível ter acesso a um conjunto de recursos computacionais maior do que os recursos locais.

Nesta seção, descrevemos em linhas gerais a arquitetura do sistema *Blowout*. Essa arquitetura, ilustrada na Figura 1, tem seis componentes principais: *submission service, task catalog, crawler, scheduler, worker node* e *fetcher.* Todos esses componentes são implantados em máquinas virtuais providas em recursos da federação. Além desses componentes, o *Blowout* utiliza o *middleware* Fogbow (Barros *et al*., 2015) para alocar e acessar os recursos federados.

O componente *submission service* atende requisições dos usuários do *Blowout* de um membro da federação. Entre as requisições atendidas estão a criação de novas unidades de trabalho, o monitoramento da execução dessas unidades de trabalho e o expurgo de dados já processados (para os quais não se tem mais interesse).

O componente *task catalog* mantém registros que descrevem as unidades de trabalho em processamento em uma instalação *Blowout*. Cada unidade corresponde a uma imagem LANDSAT. Novos registros são adicionados ao *task catalog* como resultado de uma requisição de criação atendida pelo *submission service*; a requisição considera datas que correspondem ao momento de criação das imagens LANDSAT.

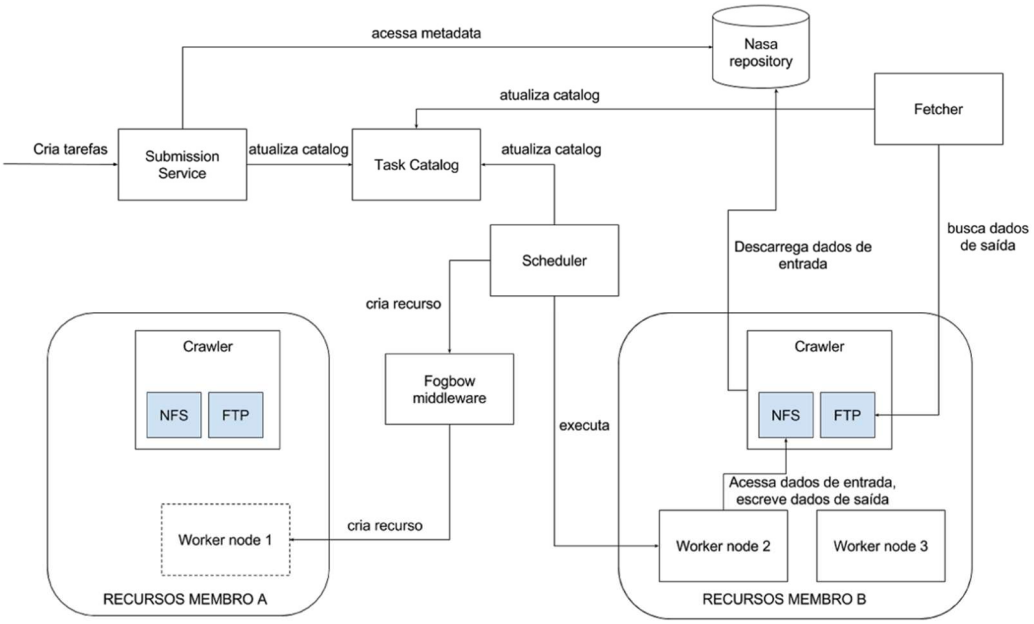

Figura 1 – Sistema *Blowout*

O componente *crawler* é responsável por obter as imagens LANDSAT da base de dados NASA relacionadas com as unidades de trabalho registradas. Após obter uma imagem, o *crawler* interage com o componente *task catalog* para indicar sua obtenção no registro correspondente. As imagens obtidas são compartilhadas com os demais componentes do *Blowout* através do sistema de arquivos distribuídos NFS (Pawlowski *et al*., 1994). Além das imagens LANDSAT, que são os dados de entrada usados pelo SEBAL, o componente *crawler* armazena também os dados resultantes da execução do SEBAL. O acesso para esses dados é disponibilizado através de um serviço FTP em execução no *crawler*.

O componente *scheduler* é responsável por definir o *worker node* que processará uma determinada imagem obtida pelo *crawler*. É também função do *scheduler* interagir como o middleware Fogbow para alocar máquinas virtuais nas quais serão implantados novos componentes *worker nodes*. Isso pode ser feito tanto em resposta para falhas em recursos que tenham sido alocados antes quanto para aumentar o conjunto de *worker nodes* disponível; por exemplo, para aumentar a taxa de processamento de imagens. Ainda, o *scheduler* monitora a execução dos algoritmos *Fmask* e SEBAL no *worker node*. Uma vez terminada a execução (após os dados de saída tenham sido copiados para o *crawler*), o *scheduler* indica ao *task catalog* para indicar que o registro relacionado indique que a imagem já foi processada.

O componente *fetch*er é responsável por agregar o resultado do processo dos algoritmos *Fmask* e SEBAL, após identificar no *task catalog* as unidades de trabalho que já foram processadas. Os dados de saída do algoritmo são coletados dos componentes *crawler* correspondentes através do serviço FTP disponibilizado por esses. Após obter esses dados de

saída, o componente *fetcher* também modifica os registros correspondentes no *task catalog* indicando que os resultados foram agregados. Essa indicação permite que o *crawler* remova os dados que mantinha relacionados com a imagem processada, uma vez que o *fetcher* os copiou.

3.2. BioClimate Scientific Gateway

O *BioClimate Scientific Gateway* (BSG) proporciona ao usuário uma interface de análise de alto nível, permitindo o acesso aos dados, análise e visualização através de múltiplas fontes de dados heterogêneos, expondo uma visão integrada. Além disso, suporta vários recursos, tais como análise de séries temporais univariadas, e comparação entre as séries de diferentes variáveis. É possível acessar o BSG a partir da página do projeto EUBrazilCC (www.eubrasilcloudconnect.eu).

Para execução das tarefas em tempo real, o sistema utiliza a técnica de *Parallel Data Analysis Service* (PDAS) e permite agendar as tarefas de acordo com os recursos disponíveis e ainda tem elasticidade para explorar os recursos de nuvens subjacentes. Elia *et al*. (2016) apresentam o sistema *BioClimate Scientific Gateway*.

No BSG estão disponíveis nove variáveis obtidas pela aplicação do algoritmo SEBAL: *Normalized Different Vegetation Index* (NDVI), *Leaf Area Index* (LAI), *Enhanced Vegetation Index* (EVI), Albedo de superfície (α), Temperatura de superfície (TS), Saldo de Radiação (Rn), Fluxo de calor no Solo (G), Fração evapotranspirativa (EF) e Evapotranspiração de 24 horas (ET_{24h}). Essas variáveis poderão ser analisadas conjuntamente com variáveis: precipitação e temperatura do ar disponibilizadas pelo *Climate Research Unit* (CRU), distribuição de ocorrências de plantas obtidos a partir de modelagem de nicho ecológico e métricas florestais obtidas a partir de imagens LiDAR 3D.

3. Resultados Preliminares

A Figura 2 apresenta as cenas que são submetidas ao processamento do algoritmo SEBAL usando o sistema *Blowout* desenvolvido no âmbito do projeto EUBrazilCC e a área escolhida para as análises usando o BSG. Para apresentação dos resultados preliminares foram escolhidas duas variáveis biofísicas obtidas pela aplicação do algoritmo SEBAL: *Enhanced Vegetation Index* (EVI) e Saldo da radiação (Rn), obtidos para a cena de órbita 215 e ponto 65.

Figura 2 – Detalhe do recorte utilizado para aplicação no *BioClimate Scientific Gateway*

Para testes, utilizou-se o *Bounding Box* indicado na Figura 2. Observa-se nas Figuras 3, 4, 5 e 6 algumas funcionalidades do BSG para análise de tendência temporal das variáveis obtidas pelo processamento do algoritmo SEBAL. O BSG permite que as análises possam ser comparadas com dados de precipitação e a temperatura do ar.

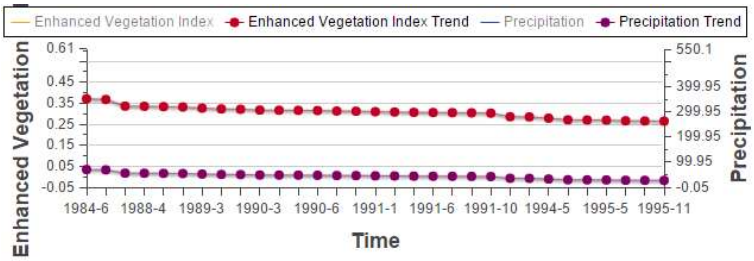

Figura 3 – Análise de tendência com EVI e precipitação

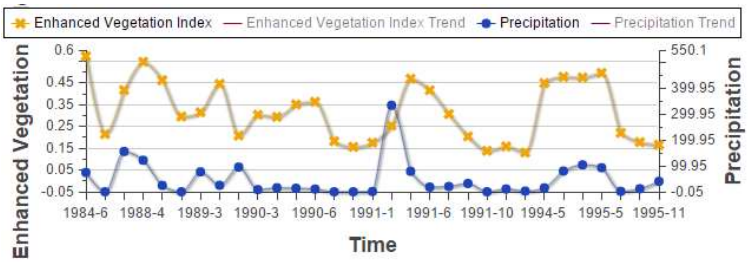

Figura 4 – Série temporal de EVI e precipitação

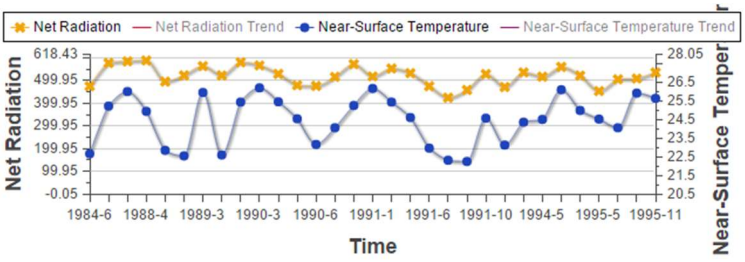

Figura 5 – Série temporal de Rn e temperatura do ar

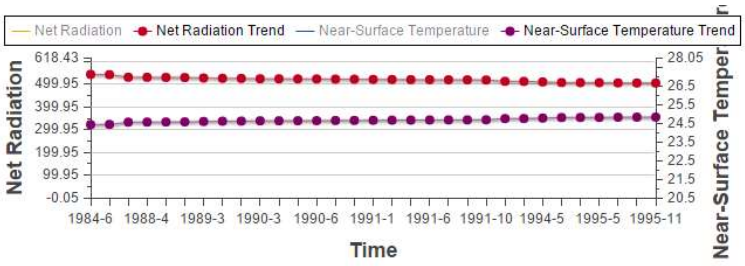

Figura 6 – Análise de tendência de Rn e temperatura do ar

4. Conclusões

Ferramentas e algoritmos para processamento de grandes massas de dados de imagens de satélite através de computação em nuvem permitem expandir a capacidade de análise de dados de mudança de uso e cobertura do solo atualmente disponível.

A integração de análises estatísticas e análises cruzadas do BSG e a capacidade de processamento do algoritmo SEBAL em séries ainda mais longas no Sistema *Blowout* representam um avanço significativo nos estudos de mudanças do uso e cobertura do solo no semiárido brasileiro.

Agradecimentos

Este projeto é resultante do Edital MCTI/CNPq 13/2012 – Programa de Cooperação Brasil – União Europeia na Área de Tecnologias da Informação e Comunicação – TIC.

Referências Bibliográficas

ALLEN, R. G.; PEREIRA, L. S.; HOWELL, T. A.; JENSEN, M. E. Evapotranspiration information reporting: I. Factors governing measurement accuracy. **Agricultural Water Management**, 98, pp.899–920, 2011.

ALLEN, R.; BURNETT, B.; KRAMBER, W.; HUNTINGTON, J.; KJAERSGAARD, J.; KILIC, A; TREZZA, R. Automated calibration of the METRIC-Landsat evapotranspiration process. **Journal of the American Water Resources Association**, 49(3), pp. 563-576, 2013.

BARROS, A.; BRASILEIRO, F.; FARIAS, G.; GERMANO, F.; NÓBREGA, M.; RIBEIRO, A.; SILVA, I.; TEXEIRA, L. Using Fogbow to federate private clouds in **XXXIII Brazilian Symposium on Networks and Distributed Systems** – Tools track. Vitória-ES, Maio-2015.

BASTIAANSSEN, W. G. M. SEBAL - based sensible and latent heat fluxes in the irrigated Gediz Basin, Turkey. **Journal of Hydrology**, 229, pp. 87-100, 2000.

ELIA, D.; NUZZO, A.; NASSISI, P.; FIORE, S.; BLANQUER, I.; BRASILEIRO, F. V.; RUFINO, I.; SEIJMONSBERGEN, A.; ANDERS, N.; GALVÃO, C.; CUNHA, J.; SOUSA-BAENA, M., CANHOS, V.;

ALOISIO, G.A Science Gateway for Biodiversity and Climate Change Research. **8th International Workshop on Science Gateways (IWSG 2016)**, June-2016.

LI, J.; HUMPHREY, M.; AGARWAL, D.; JACKSON, K.; VAN INGEN, C.; RYU, Y. E-Science in the cloud: a Modis satellite data reprojection and reduction pipeline in the windows azure platform. **Parallel & Distributed Processing (IPDPS) IEEE International Symposium**. pp.1-10. 2010.

MA, YAN; WU, H.; WANG, L.; HUANG, B.; RANJAN, R.; ZOMAYA, A.; JIE, W. Remote sensing big data computing: Challenges and opportunities. **Future Generation Computer Systems**, 51, pp. 47-60, 2015.

PAWLOWSKI, B.; JUSZCZAK, C.; STAUBACH, P.; SMITH, C.; LEBEL, D.; HITZ, D. NFS Version 3 **Design and Implementation.** USENIX, 1994.

WANG, X. Z.; ZHANGA, H. M.; ZHAO, J. H.; LIN, Q. H.; ZHOU, Y. C.; LI, J. An Interactive Web-Based Analysis Framework for Remote Sensing Cloud Computing. ISPRS **Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences**, 1, pp. 43-50, 2015.

YANG, C.; SUN, M.; LIU, K.; HUANG, Q.; LI, Z.; GUI, Z.; LOSTRITTO, P.HAOWEI. Contemporary computing technologies for processing big spatiotemporal data. In: **Space-Time Integration in Geography and GIScience**. Springer Netherlands, pp. 327-351, 2015.

ZHU, Z.; WOODCOCK, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. **Remote Sensing of Environment**, 118, pp. 83-94, 2012.